REPUBLIKA E SHQIPËRISË
UNIVERSITETI "ISMAIL QEMALI", VLORË
FAKULTETI I SHKENCAVE TEKNIKE
DEPARTAMENTI I MATEMATIKËS

DISERTACION

PËR GRADËN SHKENCORE

"DOKTOR"

KLASIFIKIMI DHE REGRESI ME ANË TË PEMËS

DOKTORANT: Msc. ADEM META

UDHËHEQËS SHKENCOR: PROF. ASOC. DR. LUELA PRIFTI

VLORË, 2019

REPUBLIKA E SHQIPËRISË
UNIVERSITETI "ISMAIL QEMALI", VLORË
FAKULTETI I SHKENCAVE TEKNIKE
DEPARTAMENTI I MATEMATIKËS

DISERTACION

i

Paraqitur nga

ADEM META MA, MSC

Për marrjen e gradës shkencore "DOKTOR"

PROGRAMI I STUDIMIT: MATEMATIKË

DREJTIMI: STATISTIKË

Tema:

KLASIFIKIMI DHE REGRESI ME ANË TË PEMËS

Udhëheqës Shkencor

Prof. Asoc. DR. LUELA PRIFTI

Mbrohet me date: 20 Dhjetor, 2019 para jurisë

1. _____ Kryetar
2. _____ Antar(Oponent)
3. _____ Anëtar(Oponent)
4. _____ Anëtar
5. _____ Anëtar

VLORË, 2019

Tabela e përmbajtjes

Faqe

Lista e figurave

Lista e tabelave

Mirënjohje

Së pari unë dua të falenderoj udhëheqësen time të disertacionit Prof. Asoc. Dr. LUELA PRIFTI. Ky disertacion do të ishte i pa mundur pa ndihmën, kontributin dhe konsulencën e saj shkencore. Një mirënjohje për drejtuesit e Universitetit duke filluar nga Rektorati, Dekanati, si dhe për Katedrën e matematikës pranë universitetit “Ismail Qemali” Vlorë, të cilët më dhanë mundësinë që të punoj edhe në këtë moshë për të arritur në nivele, të cilat janë ëndërrime për çdo person që kërkon gjithmonë e më shumë nga vetja e tij. Një mirënjohje të veçantë dhe për familjen time, bashkëshorten dhe dy djemtë e mi, të cilët më kanë inkurajuar mua që të mos ndalem në ambicjen e vazhdueshme për të arritur në nivele sa më të larta.

PËRMBLEDHJE

Proçesi i të mësuarit është një nga proçeset më të gjatë për të arritur qellime të caktuara. Ai kërkon një alternim të inteligjencës natyrale dhe një punë sistematike dhe këmbëngulëse. Në rastin e studimit të një baze të dhënash duke përdorur "Klasifikimin dhe Regresin me anë të pemës" duhet një impenjim dhe këmbëngulje e jashtëzakonshme, pasi duhet të bëhet një lidhje organike midis atributeve të një baze të dhënash dhe parashikimeve që mund të bëhen me to. Në këtë studim për të bërë parashikimet e duhura përdoret modeli i strukturës së një peme. Klasifikimi dhe regresi me anë të pemës, ndryshe metoda e ndarjes së vazhdueshme qe ndërton një pemë klasifikuese për variablat parashikuese, të cilat janë kategorike si "po", "jo": etj dhe pemën e regresit në rastin kur variablat parashikues janë të vazhdueshme. Algoritmi klasik për këtë teori është propozuar së pari nga Breiman i cili se bashku me tre autorë të tjerë si Olshen, Stonne dhe Friedman publikuan të parin libër në këtë fushë më 1984, i cili u pasurua më vonë nga studjues të tjerë si Ripley, Kass apo Quilin. Dy janë algoritmet kryesore që përdoren për të ndërtuar pemën klasifikuese të specifikuar QUEST (Quick, Unbiased, Efficent Statistical Tree) algoritm, i cili e paraqet në konteksin e analizës së pemës klasifikuese, algoritmi CHAID (Chi-square Autentic Interaction Detector Kass 1980).

Metoda e klasifikimit dhe regresit me anë të pemës (CART) është një metodë, që në përgjithesi për të zgjedhur variablat e shpërndarë përdorë një fushë përcaktimi të gjerë. Kontributi im në këtë studim është një përgjithësim teorik për preferencat që e karakterizojnë në përgjithësi këtë metodë, kur aplikohet metoda selektive e shpërndarjes, si dhe analiza dhe krahasimi i prashikimeve për tre bazë të dhënash të ndryshme: nga Clëveland clinic, Ohio, USA, nga South Africa dhe nga baza e të dhënave "Boston House Market", në dy të parat aplikohet pema klasifikuese dhe së fundi një bazë te dhënash (Boston House Market) ku aplikohet pema e regresit. Në përfundim të analizës së secilës bazë të dhënash realizohen dhe krahasimet midis tyre. Për një bazë të dhënash me shumë elemente, rendimenti i të mësuarit të algoritmeve, në lidhje me përpjekjet për të bërë veprimet e duhura, kërkon një kujtesë të fuqishme, gjë e cila realizohet në ditët e sotme nga kompjuteri. Një nga kontributet e mia në këtë studim eshte dhe sistemimi i disa koncepteve bazë që përdoren në proçesin e ndërtimit dhe strukturimit të pemës klasifikuse për gjitha variablat dhe përfundimet lidhur me konceptet bazë që janë përdorur për klasifikimin dhe regresin klasik me anë të pemës, duke përmirësuar në mënyrë të rëndësishme saktësinë, kur ndërtojmë këto lloje modelesh. Këto veprime shumë të ndërlikuara, në ditët e sotme kryhen nga kompjuteri. Ne kapitujt e ketij punimi doktorature trajtohen idetë bazë për vendimet që merren me anë të një peme përfundimtare klasifikuese. Ato lidhen me :

Tre elementet bazë të ndertimit të një peme klasifikuese.

Paraqitjen e funksionit të papastërtisë dhe disa shembuj të tij.

Vlerësimin e probabilitetit të çdo klase pasardhëse në çdo nyje të pemës.

Avantazhet e strukturës së pemës duke përdorur metodën e klasifikimit.

Trajtimin e konceptit të rizëvëndësimit të shkallës së gabimit dhe masës së kostos së përgjithshme.

Pikat e dobëta të krasitjes së pemës klasifikuese si dhe avantazhet dhe disavantazhet e kësaj metode.

Nënpemët më të mira të krasitura janë të mbivendosura aty dhe mund të përftohen në se ne vazhdojmë një proçes të pandërprerë ndarjeje dhe krasitjeje.

Metoda e bazuar në vlerësimin e kryqëzuar (cros-validation) për të zgjedhur parametrin e kompleksitetit për të shkuar te nënpema përfundimtare.

Qëllimin e modelit të mesatarizimit, proçedura e ndarjes.
Metodën e e katrorëve më të vegjël.
Proçedurën e vlerave absolute të diferencave të mesatareve të devijimeve.
Zgjedhjen e algoritmit të përshtatshëm dhe përdorimin e tij.
Nëpërmjet vërtetimeve trajtohet fuksionimi në tërësi i procedurës CART.
Në pjesën e fundit të çdo kapitulli janë paraqitur dhe përfundimet e arritura.

ABSTRAKTI

Klasifikimi dhe Regresi me anë të pemës është një model i të mësuarit që paraqitet si një makinë për të ndërtuar modele parashikuese te pemëve nisur nga një bazë të dhënash. Këto modele merren duke e ndarë bazën e të dhënave në pjesë të vogla, në të cilat modelet parashikuese janë pjesë e secilës pjesëz. Këto pjesë mund të parqiten grafikisht si një pemë e cila jep perfundime. Klasifikimi me anë të pemës është i ndërtuar si një pemë klasifikuese për variablat kategorikë në vartësi të variablave të varura të cilat marrin një vlerë numerike të fundme për vlerat të cilat nuk janë vendosur në një renditje të caktuar me një parashikim gabimi. Pema e Regresit është një pemë klasifikuese me një varibël të varur të vazhdueshëm në të cilën variablat e e pavarura marrin vlera të vazhdushme ose vlerat diskrete, me një parashikim gabimi i cili njehsohet me katrorët e diferencave midis vlerave të vrojtuara dhe atyre të parashikuara. Në kapitullin e parë të këtij punimi, paraqiten disa parime baze në ndërtimin e një peme klasifikimi/regresi, ndërsa ne kapitullin e dytë bëhet një përshkrim i detajuar i shpërndarjes, duke bëre përgjithesimet e duhura teorike, dhe jepen në menyre te detajuar algoritmet që përdoren në shpërndarje, si dhe duke perdorur nje shmbull konkret, ku aplikohen algoritmet e ndryshme për të bërë shperndarjen në një baze të dhënash. Ne kapitullin e tretë zë një vënd të rëndesishëm krasitja e pemës së klasifikimit apo regresit duke berë përgjithesimet e duhura teorike, si dhe duke dhënë dhe një informacion të hollësishëm se si do të përdoren algoritmet e ndryshme për të krasitur pemën e mbingarkuar, për të arritur te pema përfundimtare. Në vazhdim, përse duhet të përdorim metodën e klasifikimit dhe regresit me anë të pemës. Një vend të rëndësishëm në këtë punim disertacioni zë dhe vertetimi i disa teoremave për pemën e klasifikimit dhe të regresit, si dhe avantazhet dhe disavantazhet e kësaj metode.

Në kapitullin e katërt të këtij punimi realizohet nje analizë e hollesishme per të ndërtuar pemën klasifikuese , duke përdorur dy baza të dhënash. Ne këtë pjes një vënd të rëndësishëm zë analiza e dy bazave të të dhënave, shperndarja, krasitja, deri në marrjen e pemës optimale, duke bërë dhe disa adaptime të algoritmeve. Në këtë kapitull i kushtohet një vëmëndje e veçant, proçedurave që ndiqen për të marrë pemen me saktësi sa më të lartë, duke përdorur metodat me efikase, përfshir këtu dhe grafikët për të arritur në krasitjen më të sakt të pemës së klasifikimit. Ne këtë kapitull një vënd të vecant zë dhe paraqitja dhe interpretimi i disa fakteve nga ana grafike si dhe interpretohen përfundimet e marra duke përdorur softwarin R .

Në kapitullin e pest, bëhet një përshkrim i detajuar i pemës së regresit, një përshkrim i proçesit të shpërndarjes, teknikat që përdoren, si dhe një vështrim i përgjithshm i krasitjes, duke përdorur disa metoda të cilat konkretizohen duke përdorur nje bazë të dhënash me një variabël parashikues të vazhdueshëm sikurse është “Boston House Market”.

ABSTRACT

Tree classification and regression is a model of learning that appears as a machine to build predictive tree models from a database. These models are taken by dividing the database into small parts, in which predictive models are part of each particle. These pieces can be graphically parsed as a tree that gives predictions and we can write some conclusions. Tree classification is constructed as a classification tree for categorical variables subordinated to subordinate variables that receive a finite numeric value for values that are not set in a given order with a prediction of error. The Regression Tree is a classification tree with a continuous dependent variable in which independent variables receive continuous values or discrete values with an error prediction that is computed with squares of differences between the observed and predicted values. In the first chapter of this paper, some basic principles are presented in the construction of a classification / regression tree, while in the second chapter a detailed description of the distribution is made, making the theoretical generalization, and giving in detail the algorithms that are used in distribution as well as using a concrete examples, where different algorithms apply to distribute to a database. In the third chapter, there is an important place in classifying or regressing trees by making appropriate theoretical generalizations and providing detailed information on how to use different algorithms to prune the overcrowded tree to reach to the final tree. In the following, why should we use the classification and regression method by means of a tree. An important place in this thesis is the authentication of some theorems for the classification and regression tree as well as the advantages and disadvantages of the method.

In the fourth chapter of this paper, a detailed analysis is carried out to construct the classification tree, using two databases. This is an important part of the analysis of two databases, distribution, pruning, optimum tree making, and some adaptations of algorithms. In this chapter, special attention is paid to the procedures followed to get the tree with the highest precision, using the most efficient methods, including graphs, to get the most accurate classification of the classification tree. In this chapter, a special voice is given and the presentation and interpretation of some facts graphically as well as interpreted the conclusions obtained using the R software. In the fifth chapter, a detailed description of the regression tree is made, where a description of the distribution process, the techniques used, and a general overview of the pruning, is made using some methods that are concretized using a database with a constant predictive variable like “Boston House Market”.

KAPITULLI I

VËSHTRIM I PËRGJITHSHËM MBI KLASIFIKIMIN DHE REGRESIN ME ANË TË PEMËS

1.1 Elementet e CART

Metodologjia që përdoret te pema e klasifikimit dhe e regresit(CART) është e njohur teknikisht si ndarje binare rekursive. Procesi është binar sepse nyjet e prindërve janë të ndarë gjithmonë në dy nyjet pasardhëse dhe gjithkund rekursive për shkak se proçesi mund të përsëritet duke e trajtuar çdo nyje pasardhëse (fëmijë) si një prind. Elementet kryesore të CART janë:

1. Ndarje e çdo nyje në një pemë.
2. Vendos, kur një pemë është e plotë.
3. Shëno si nyje fundore çdo përfundim të klasës.

Metodologjia, Klasifikimi përbëhet nga tri pjesë:

a. Ndërtimi i pemës maksimale.
b. Zgjedhja e madhësisë së duhur për pemën.
c. Klasifikimi i të dhënave të reja duke përdorur dhe ndërtuar pemën.

Ndërtimi i një peme nuk është dhe shumë i komplikuar dhe është i lehtë për ta bërë me dorë, kur kemi një numër të vogël të variablave parashikuese. Megjithatë, është shumë e vështirë dhe e komplikuar kur kemi shumë variabla parashikuese. Në shumicën e rasteve, studiuesit merren me më shumë se dhjetë variabla dhe kjo kërkon teknologji për të realizuar qëllimin tonë. Zgjedhja e metodes së duhur është një nga hapat më të rëndësishme për ndërtimin e një peme të klasifikimit.

Për çdo nyje t, supozohet se ndodhet një kandidat s i cili mund të shpërndahet në dy nën nyje t_L dhe t_R të tila që janë propocionale respektivisht me P_L dhe P_R (Figura 1).

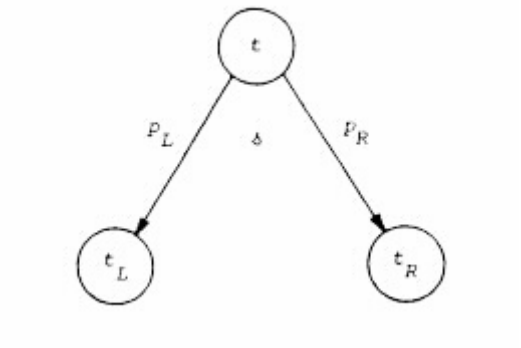


Figura 1: *Nyja "t" dhe dy nënnyjet*

Atëherë mirësia e shpërndarjes është :

$$\Delta i(s,t) = i(t) - P_L i(t_L) - P_R i(t_R).$$

Kështu që kandidati s ndjek një shpërndarje binare për çdo nyje. Shpërndarja s në çdo nyje i dërgon të gjitha x_n në t dhe ato që kanë përgjigjen "yes" shkojnë në t_L dhe ato që kanë përgjigjen "no" shkojnë në t_R.

1.2 Hapat që përdoren në CART

Analiza e CART përbëhet nga katër hapa themelore. Hapi i parë është ndërtimi i pemës, duke përdorur ndarjen rekursive të të gjitha nyjeve. Gjatë ndarjes çdo nyjë që përfitohet përcaktohet si një klasë parashikuse, bazuar në shpërndarjen e klasave në këtë bazë të dhënash, ku duhet të zgjedhim atë nyje fundore e cila na jep vendimin më të mirë të cilin do ta klasifikojmë në bazë të algoritmeve të caktuara. Caktimi i klasës parashikuse për çdo nyje ndodh në atë moment kur kjo nyje do të shpërndahet në nyje fundore të cilat i quajmë dhe nyje fundore. Në hapin e dytë ndalohet procesi i ndërtimit të pemës. Në këtë pikë një pemë "maksimale" është prodhuar, e cila ndoshta në masë të madhe mbipërshtat informacionin e përfshirë në këtë bazë të dhënash.

Hapi i tretë pema " do të krasitet", pra jemi në krijimin e një sekuence të pemëve të cilat duhet të vijnë duke u thjeshtuar deri sa të arrijnë në kulmin e nyjeve dhe duke u bërë gjithnjë e më të rëndësishme.

Hapi i katërt përzgjedhja optimale e pemës, gjatë së cilës pema e cila i përshtatet të dhënave në këtë bazë të dhënash nuk mbivlerëson informacionin dhe është zgjedhur nga radhët e pemëve të renditura të pemëve të krasitura.

1.3 Problemi i klasifikimit dhe i regresit

Në përgjithësi janë shumë algoritme për parashikimin e një madhësie të ndryshueshme e cila mund të jetë e vazhdueshme ose kategorike për të cilat gjithashtu perdorim variabla të pavaruara të vazhdueshme apo kategorike dhe mbërrijëmë në përfundime për ndikimin e tyre. Si shëmbuj për këtë kemi Modelin e përgjithshem linear (GLM, General Linear Model) dhe modeli i përgjithshëm i regresit (GRM, General Regression Model). Mund të specifikojmë një kombinim linear për parashikuesit e vazhdueshëm nga efektet e variablave kategorike me dy ose tre mënyra të efekteve të tyre vepruese dhe të parashikojmë një variabël të vazhdueshëm të varur. Një tjetër shembull ku përdoret si parashikues një variabël i vazhdueshëm është GDA (General Discriminant Function Analyses).

Një nga format më të vjetra të klasifikimit është e njohur si analiza lineare e diskriminantit. Kjo metode lidhet me formimin e kombinimeve lineare të variablave parashikues (të ngjashme me një model linear të regresit) në mënyrë të tillë që vlerat mesatare e këtyre kombinimeve lineare të jenë të ndryshme dhe të jetë e mundur për nivele të ndryshme të variablit klasifikues. Bazuar në vlerat e kombinimeve lineare, analiza lineare e diskriminantit paraqet një sërë probabilitetesh të nyjeve pasardhëse për çdo nivel të klasifikimit, për çdo vrojtim, së bashku me nivelin e variablit të klasifikimit të parashikuar në këtë analizë.

Supozojmë se kemi një ndryshore që duhet ta klasifikojmë dhe që mund të marrë një nga tre vlerat: pas një analize linearë të diskriminantit, do të kemi tre shanset (duke shtuar deri në një) për çdo variabël dhe duhet të tregojmë se sa e mundshme është që vrojtimi të kategorizohet në secilin nga tre kategori; klasifikimi i parashikuar është ai që ka probabilitetin më të lartë. Mund të marrim njohuri për cilësinë e klasifikimit duke parë vlerat e probabilitetit.

Shembuj tipikë të klasifikimit janë në përgjithësi ato kur duhet të parashikojmë një variabël të varur kategorik nga një ose me shumë variabla të pavarura të cilat mund të jenë të vazhdueshme apo kategorike. Në raste të tjera mund të jemi të interesuar në parashikimin e një apo shumë alternativave. Në këto raste kemi të bëjmë me disa katagori apo klasa për kategoritë e ndryshme të variablave të varura. Janë disa mënyra për të bërë analizën e tipeve të pemës klasifikuese, regresi binomial ose regresi mulimonomial duke përdorur logaritmin e

saj, ku analiza lineare e logaritmit të tabelave me shumë denduri, si ANCOVA, apo analizat të tipit CHAID të cilat japin rezultate të ngjashme me ato të marra nga CART.

Universiteti i Kalifornisë, Qëndra Mjeksore e San Diegos, ka bërë një studim për pacientët të cilët janë shtruar në këtë qëndër me atak në zemër. Për 24 orët e para ku janë shtruar 215 persona është mbajtur një informacion statistikor me 19 variabla për pacientet të cilët mbijetuan në 24 orët e para. Per te identifikuar pacientet me rrezik të lartë (që nuk munden të mbijetonin ne 30 ditët në vazhdim) u ndertua pema (Figura 2) për tu përgjigjur "jo" apo "po".

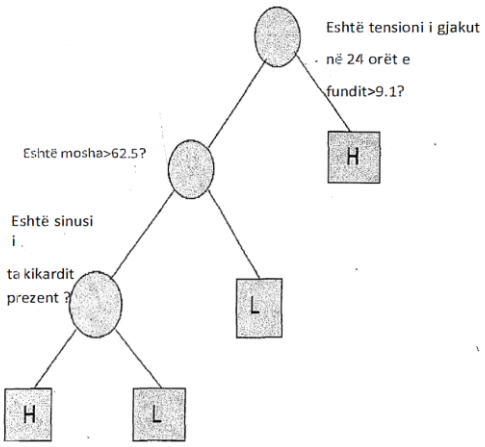


***Figura 2**: Pema klasifikuese për të identifikuar pacientët me rrezik të lartë*

Nga Figura 2, nëse vlera e tensionit të ulët të gjakut është më e madhe ose e barabartë me 9.1 për gjatë gjithë 24 orëve pacienti duhet të klasifikohet si në rrezik të lart, por në se është më i madh ose i barabartë me 9.1 dhe mosha është më vogël se 62,5 vjec rreziku është i vogël, por nëse mosha është më e madhe se 62.5 vjec dhe atje është prezent sinus tachycadrida është përsëri rrezik i lartë dhe kur nuk është prezent rreziku është i vogël. (Breiman)

1.4 Pema e Klasifikimit

Pemët e klasifikimit përdoren për të parashikuar të gjithë elementet ose objektet që i përkasin një klase të caktuar për një variabël kategorik të varur nga madhësitë e një apo disa parashikueseve për një bazë të dhënash.

Qëllimi i pemës së klasifikimit është që të parashikojmë ose të shpjegojmë përgjigjet për një variabël të varur kategorike, dhe si i tillë, teknikat e disponueshme kanë shumë anë të përbashkëta me teknikat e përdorura në metodat më tradicionale si analiza diskriminante, vleresimet jolineare etj. Fleksibiliteti i pemës së klasifikimit i ka bërë ato një opsion analize. Pemët e klasifikimit, sipas mendimit të shumë studiuesve janë metoda efikase dhe tërheqëse por kjo nuk do të thotë se përdorimi i tyre është i rekomanduar duke përjashtuar metodat tradicionale. Në përgjithësi, kur supozimet teorike dhe kushtet për shpërndarjen normale të të dhënave plotësohen. Metodat tradicionale mund të jenë të preferueshme dhe të aplikueshme. Por si një teknikë paraprake, ose si një teknikë e fundit, kur metodat tradicionale nuk japin rezultat ndërtojmë pemën e klasifikimit.

Cilat janë pemët e klasifikimit? Mendojmë se duam të ndërtojmë një sistem për klasifikimin e një grupi të monedhave në klasa të ndryshme (ndoshta një centëshe, pesë centëshe, 10 centëshe dhe 25 centëshe). Supozojme se ka një matje në të cilën monedhat

ndryshojnë nga njëra tjetra që është diametri, e cila mund të përdoret për të ndertuar një sistem për klasifikimin hierarkik të monedhave. Nëse mund të rrokullisim monedhat poshtë në një kanal të ngushtë në të cilin janë hapur katër vrima me diametra përkatësisht sa të një centëshi, pesë centëshe, dhjetë centëshi apo të një 25 centëshi, atehere nëse monedhat bien nëpërmjet kësaje rrugice, ato që futen tek vrima e parë i klasifikojmë si një centëshe, ato që futen tek e dyta i klasifikojmë si pesë centëshe, ato që futen tek vrima e tretë i klasifikojmë si dhjetë centëshe dhe ato që futen tek vrima e fundit i klasifikojmë si njëzet e pesë centëshe. Në këtë mënyrë ne kemi ndërtuar një pemë klasifikimi. Proçesi i përdorur për pemën tonë të klasifikimit ofron një metodë efikase për klasifikimin e një grupi të monedhave, dhe më në përgjithësi, mund të zbatohet në një shumëllojshmëri më të gjerë të problemeve të klasifikimit.

Studimi dhe përdorimi i pemëve të klasifikimit nuk është aq i përhapur në fushat e probabilitetit dhe në njohjen e modeleve statistikore (Ripley, 1996), por pemët e klasifikimit janë përdorur gjerësisht në fusha të ndryshme si në fushën e mjekësisë, shkenca kompjuterike (strukturat e të dhënave), botanikë (klasifikimi), dhe psikologji (teoria e vendimit). Pema klasifikuese me shfaqjen e saj grafike ka ndihmuar për të bërë më të lehtë interpretimin, krahasuar me ato numerike. Pemët klasifikuese mund të jenë mjaft komplekse. Megjithatë, procedurat grafike mund të zhvillohen për të ndihmuar që ta bëjmë sa më të thjeshtë pemën me qellim që të na e lehtësojnë interpretimin e pemes komplekse.

Një nga problemet kryesore ku ne duhet të parashikojmë me anë të regresit është kur një variabel i vazhduar është një nga variablat e varur (të cilet mund të jenë të vazhduar ose kategorike). Një shembull tipik është të parashikosh çmimin e një shtëpie, (i cili është një variabël i varur dhe i vazhdueshëm) i cili është i varur nga madhësi të tilla si sipërfaqja e saj (e vazhdueshme), katet(diskrete), apo qyteti ku ndodhet (kategorike Zip= numër që kodon qytetet). Ne përdorim regresin e thjeshtë ose modelet e përgjithshme lineare (GLM) që të parashikojmë se sa mund të shitet një shtëpi, duke nxjerrë një ekuacion linear me të cilin mund të bëjmë parashikimet e duhura. Janë disa modele lineare dhe jolineare për të bërë parashikimet e duhura. CHAID gjithashtu është një mënyrë për të analizuar problemet e regresit, i cili jep rezultate të ngjashme me ato të CART.

1.5 Historia e pemës së klasifikimit dhe regresit

Klasifikimi dhe regresi me anë të pemës është publikuar për herë të parë nga: Breiman, Friedman, Olshen dhe Stone ne vitin 1984. Ata paraqiten modelin bazë të pemës së përdorur në statistikë. Sipas tyre, pema binare jep një mënyrë shumë interesante dhe paraqet te baza e të dhënave dicka të rëndësishme dhe specifike të cilën ata e quajtën problemi i pemës klasifikuese apo problemi i regresit. Sipas tyre kjo metodë nuk mund të marrë përsipër dhe të përgjithësojë se gjithmonë ka një saktësi maksimale, por me punë të kujdesshme gjatë gjithë procesit mund ta bëjmë atë gjithmon dhe më efektive. Në ditët e sotme pas disa dekadash nga publikimi i këtij libri dhe sidomos me zhvillimet teknologjike është bërë e mundur që ky proces të përsoset dhe të ketë më tepër aftësi për të bërë parashikime me saktësi më të lartë. Pemët moderne të klasifikimit mund ta ndajnë bazën e të dhënave në shpërndarje lineare në nënbashkësi të cilat janë shumë të përshtatshme për të pasur një saktësi të lartë në parashikime. Po ashtu dhe pema e regresit mund të përshtatet pothuajse në të gjitha modelet që ne njohim si metoda e katrorëve më të vegjël, kuantilet, regresi logjistik i Puasonit dhe modelet propocionale sikurse dhe modelet shumë dimensionale. Nje rol të vecantë në rritjen e saktësisë së parashikimeve kanë luajtur dhe zhvillimet që janë bërë në zhvillimin e softwareve, si dhe zhvillimet në përmirsimin e mëtejshëm të algoritmeve bazë që përdoren në këtë proces.

1.6 Zbatimet e CART

Zbatimet e CART janë të shumta si në fushën e mjekësisë, në shkencat polikike, në probleme te ekonomise, në shkencat natyrore etj.

Një nga fushat ku gjen zbatim klasifikimi dhe regresi me anë të pemës është fusha e mjeksisë dhe sidomos ajo e sëmundjeve të zemrës. Një nga objektivat kryesore të shumë kërkimeve shkencore që bëhen sot në shumë klinika të botës është të zbulojënë një metodë sa më të besueshme dhe efektive i cila duhet të ketë cilësinë të klasifikojë pacientët që kanë simtoma të reja të sëmundjeve kardiovaskulare në kategori të caktuara të cilat japin mundësinë te mjeku se në çfarë treguesish ata duhet ti klasifikojnë si të rëndësishëm domethënë me të cilët duhet të tregohet kujdes mjekësor. Në bazë të rregullave që parashikojnë këto modele, mjekët pasi të marrin informacionin e përpunuar nga statisticienet,mund të kategorizojnë pacientët në pacientë me rrezikshmëri të lartë, të mesëm ose të ulët dhe në bazë tëkësaj ndertojnë linjën se në çfarë niveli duhet të jetë kujdesi mjekësor (të shtrohet në spital apo lloje të tjera kujdesi). Metodat tradicionale statistikore janë relativisht të vështira të përdoren për të adresuar pemën klasifikuese. Arsyet përse nuk mund të përdoren janë të ndryshme, se pari janë disa mundësi parashikimi për cdo variabël që shërben si parashikues në datën që ne përdorim, së dyti selektimi dhe ndarja e variablave është tepër e veshtirë. Metodat tradicionale statistikore në përgjithësi janë metoda të varfëra dhe jo efektive në krahasimet shumë dimensionale. Një tjetër arsye është se në shumicën e rasteve variblat parashikues janë shumë rrallë të shpërndare mire. Nga vrojtimet e bëra është vene re se variablat që përdoren në studimet klinike nuk janë me shpërndarje normale. Është parë se grupet e ndryshme të pacientëve kanë një diferencë të madhe në devijimin standart. Po ashtu një faktor tjetër qe lidhet me ndërveprime komplekse midis variablave për disa pacientë që mund të ekzistojnë në bazen e te dhenave. Si një rast tipik është historia e familjes që mund të ketë një ndikim në disa madhësi të tjera. Këto ndërveprime të variablave të ndryshme janë përgjithësisht të vështira të modelohen sidomos kur ndërveprimi është substancial. Dhe së fundi rezultatet e metodave tradicionale janë të vështira të përdoren. Pamvaresisht nga metodat statistikore të përdorura, në nxjerrjen e vendimeve nga klinikat kërkohen rregulla të cilat do të përdoren për një bazë të dhënave të cilat duhet të jenë relativisht të mëdha. Për çdo pacient në bazën e të dhënave duhet të parashikohet me anë të një variabli të varur. Me saktësi të parashikohet nëse pacienti në të ardhmen do të ketë problem kardiake apo jo, kjo në vartësi të disa variablave të tjerë si mosha, mbipesha, pirja e duhanit, pirja e alkolit apo historia familjare me këtë sëmundje? Në këto 20 vjetët e fundit ka pasur një rritje të interesit të studjuesve të ndryshëm për të përdorur analizën me anë të CART si dhe avancim në teknikat e përdorura. Kjo e ka bërë këtë metodë e cila është e ndryshme nga metodat tradicinale të jetë më e aplikuar. Meqënëse analiza në këtë metodë është e ndryshme nga metodat e tjera kjo eshte pranuar me vonesë. Analiza në CART është një analizë komplekse dhe së dyti përdorimi i sofwareit në CART ka pasur vështirësi.

Por tani është e mundëshme që të performojmë një analizë të CART pa një kuptim të thellë për çdo hap kompleks të përdorur në software. Tani është provuar se CART është një metodë efektive për të krijuar rregulla të caktuara nga klinikat e ndryshme të cilat i aplikojnë ato dhe është më efektive se metodat tradicinale. Klasifikimi dhe regresi me anë të pemës është një metodë që shpjegon korelacionet komplekse që ekzistojnë midis parashikuesve të ndryshëm të cilën nuk e kanë bërë metodat tradicinale. Një nga qellimet kryesore në këtë punim është një përshkrim i përgjithshëm i metodologjisë së përdorur në CART, duke e shoqëruar këtë me përdorime praktike si dhe duke bërë dhe përgjithësimet teorike.

CART nuk sjell ndryshime për ndryshime të vogla të një bazë të dhënash dhe analiza e saj kërkon automatikisht modele të rëndësishme në të cilat me një farë mënyre zbulohen struktura të fshehta në një bazë të dhënash me të vërtetë komplekse.

Informacionet që jep CART janë shumë të vlefshme, efektive dhe të besueshme në gjenerimin me një saktësi shumë të mirë të parashikimeve duke përdorur modele të ndryshme në fusha të ndryshme të jetës dhe shoqërisë së sotme njerëzore. Aplikimet e CART janë sot të përdorëshme në të gjitha fushat e jetës, duke filluar në mjekësi, ekonomi, shkenca sociale etj.

Pema e klasifikimit shpjegon në mënyrë të qartë për një variabël përgjegjëse si është e varur nga madhësitë e tjera, të cilat mund të jenë kategorike, diskrete apo të vazhdueshme.

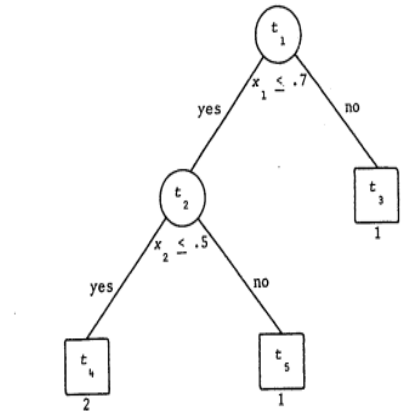
1.7 Disa pyetje standarte

Në se një bazë të dhënash ka një strukturë standarte dhe pyetjet që do të marrin përgjigje do të jenë standarte, supozojmë se vektorët e bazës së të dhënave kanë këtë formë: x=($x_1, x_2 ..... x_M$), ku M është dimensioni i fiksuar dhe variabëlat $x_1, x_2 ..... x_M$ ndjekin rend të caktuar së bashku me variablat kategorike.

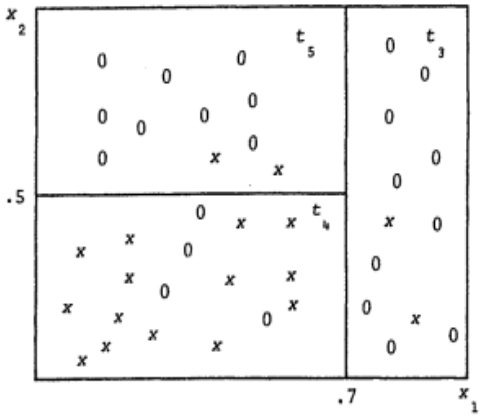
1. Ne secilen nga shpërndarjet nje variabël i caktuar merr një vlerë të vetme.
2. Për çdo variabël të vendosur në një renditje të caktuar x_m, do të kemi një shpërndarje që kënaq: $\{x_m \le c ?\}$ për të gjitha c e renditura nga $(-\infty, \infty)$.
3. Nëse x_m është variabël kategorik, duke marrë vlerat në $\{b_1, b_2, ... b_L\}$ Secila përfshin madhësitë e formës $\{x_m \in S\}$? ku S është nënbashkësi e $\{b_1, b_2, ... b_L\}$

Procesi i mëtejshëm i shpërndarjes për të gjitha variablat M konsiston në bashkësi standarte, kështu nëse M=4 dhe x_1, x_2, x_3 janë renditur dhe $x_4 \in (b_1, b_2, b_3)$, atëherë natyrshëm lindin pyetjet e mëposhtëme: është $x_1 \le 3.2,$ është, $x_3 \le -6.8,$ është, $x_4 \in (b_1, b_2)$? e kështu me radhë. Në këtë proces të shpërndarjes për një bazë të dhënash nuk është një numër i pafundëm ndarjesh. Për shembull nëse x_1 është renditur atëhere baza e të dhënave ka e shumta N vlera të çfardoshme $x_{1,1}, x_{1,2}, ... x_{1,N}$, kështu ne kemi N shpërndarje të ndryshme të cilat gjenerohen nga bashkësitë e formës $\{x_1 \le c\}$? ku është dhënë se $\{x_1 \le c_n\}$?, n=1.... $N' \le N$, ku c_n janë marrë në mënyrë të tillë që të jenë në mes të dy vlerave të njëpasnjëshme të x_1.

Për ndryshoret kategorike x_m në të gjenerohet shpërndarja t_L, dhe, t_R nëse x_m merr L vlera të caktuara, atëherë kemi $2^{L-1} - 1$ shpërndarje që janë të përkufizuara si vlera të x_m. Tek këto nyje aplikohet algoritmi i shpërndarjes i cili gjen shpërndarjen më të mirë duke filluar me x_1 duke vazhduar deri te x_m dhe pastaj bën një krahasim midis shpërndarjeve dhe selekton më të mirën e tyre. Nëse e konsiderojmë një pemë me dy klasa si në figurën e mëposhtme dhe marrim dy ndryshore të renditura $x_1, x_2, ku, 0 \le x_i \le 1, i = 1,2.$ Një mënyrë ekuivalente për të parë këtë proçes të pemës është edhe ndarja me kuadrate si në figurën e mëposhtme dhe kjo ndarje në drejtëkëndësha të vegjël vazhdon dhe e bën përbërjen e elementeve në këto katrorë homogjenë.



Figurë 3: Ndarja e një peme në dy klasa



0=klas 1, x=klas 2

Figurë 4*: Ndarja në klasa homogjene*

1.8 Metodologjia që përdoret në CART

Metodologjia që përdor CART duke e ndarë variablin përgjegjës në mënyrë të vazhdueshme në grupe homogjene duke përdorur një kombinim të variablave të pavarura të cilat mund të jenë kategorike apo numerike. Çdo grup është i karakterizuar nga vlera tipike në lidhje me variablin që do të parashikojmë, dhe numrin e vrojtimeve. Gjithashtu kjo metodë kërkon dhe intuitë në krijimin e grupeve dhe shpërndarjen e vazhdueshme të varaiblave të varura.

1.9 Klasifikimi dhe zgjidhja e problemit vendimmarrës

Metodologjia e klasifikimit dhe regresit me anë të pemës teknikisht është e njohur si ndarje vazhdueshme binare. Ky proçes është quajtur binare pasi çdo nyje paraardhëse ndahet në dy nyje pasardhëse të cilat i quajmë egzaktësisht të vazhdueshëm (recursive) pasi përsëritja e proçesit të shpërndarjes e trajton çdo nyje që në fillim është si fëmijë, të

konsiderohet pas kësaj si prind. Disa rregulla kyce për të analizuar klasifikimin dhe regresin me anë të pemës:

a. Ndahet çdo nyje në pemë
b. Vendoset kur pema është përfundimtare
c. Shënohet secila nyje përfundimtare si vlerë parashikuese.

Metodologjia klasifikuese konsiston në tre pjesë

1. Ndertohet nje pemë maksimale
2. Zgjidhet përmasa e duhur e pemës
3. Klasifikohet baza e të dhënave të reja duke përdorur pemën e ndërtuar.

Ndertimi i një peme nuk është i komplikuar dhe është i lehtë per tu bërë me dorë vetëm në rastet kur numri i variablave parashikues në të është i vogël. Ky proçes është tepër i komplikuar në mënyrë manuale kur kemi shumë variabla dhe zakonisht, statisticienët apo mjekët i kushtojnë vëmëndje rasteve kur baza e të dhënave ka më shumë se dhjetë variabla. Si në shembullin e paraqitur në Tabela 1.

row.names	**sbp**	**tobacco**	**ldl**	**adiposity**	**famhist**	**Type a**	**obesity**	**Alcohol**	**age**	**chd**
1	160	12	5.73	23.11	Present	49	25.3	97.2	52	Y
2	144	0.01	4.41	28.61	Absent	55	28.87	2.06	63	Y
3	118	0.08	3.48	32.28	Present	52	29.14	3.81	46	N
4	170	7.5	6.41	38.03	Present	51	31.99	24.26	58	Y
5	134	13.6	3.5	27.78	Present	60	25.99	57.34	49	Y
6	132	6.2	6.47	36.21	Present	62	30.77	14.14	45	N
7	142	4.05	3.38	16.2	Absent	59	20.81	2.62	38	N
8	114	4.08	4.59	14.6	Present	62	23.11	6.72	58	Y
9	114	0	3.83	19.4	Present	49	24.86	2.49	29	N
10	132	0	5.8	30.96	Present	69	30.11	0	53	Y
11	206	6	2.95	32.27	Absent	72	26.81	56.06	60	Y
12	134	14.1	4.44	22.39	Present	65	23.09	0	40	Y
13	118	0	1.88	10.05	Absent	59	21.57	0	17	N
14	132	0	1.87	17.21	Absent	49	23.63	0.97	15	N
15	112	9.65	2.29	17.2	Present	54	23.53	0.68	53	N
16	117	1.53	2.44	28.95	Present	35	25.89	30.03	46	N
17	120	7.5	15.33	22	Absent	60	25.31	34.49	49	N

***Tabela 1**: Një bazë të dhënash*

1.10 Përfundime

Në këtë kapitull paraqiten elementet bazë të pemës së klasifikimit dhe regresit si dhe hapat që ndiqen në ndërtimin e kësaj peme dhe metodologjia që përdoret. Krahas përshkrimit të pemës së klasifikimt dhe pemës së regresit jane paraqitur dhe anët e përbashkëta dhe ndryshimet midis tyre, si dhe një përshkrim i shkurtër i historisë së zhvillimit dhe përdorimit të kësaj metode në fushat e ndryshme, të jetës shoqërore, ekonomike apo shkencore. Shembujt e paraqitur ilustrojnë këtë metodë për përdorimin e saj në fusha të ndryshme të jetës.

KAPITULLI 2

SHPËRNDARJA E TË DHËNAVE

2.1 Vështrim mbi ndarjen

Duke përdorur softuerët statistikor mund të realizojmë një studim të tillë. E rëndësishme është që të gjejmë një metodë se si ta ndertojmë pemën klasifikuese. Për një nyje të cfardoshme t, supozojmë se kemi një kandidat për të shpërndarë në dy nyje përkatësisht t_R dhe t_L të tilla që këto të jenë raporte propocionale sikurse t_R me P_R sikurse t_L me P_L (P_L, P_R probabiliteti i majtë i djathtë). Mirësia e shpërndarjes është një mjet që zvogëlon papastërtinë dhe njehsohet si:

$$\Delta i(s,t) = i(t) - P_L i(t_L) - P_R i(t_R).$$

Një problem klasifikues konsiston në katër komponente: Komponenti i pare është një variabël kategorik (categorical outcome) ose variabël i varur. Ky variabël parashikon të ardhmen, bazuar në "parashikuesit" ose variablat e pavarura. Një variabël tipik në këtë llojë është i mbijetuari, ka nevojë për operacion apo jo, do të ketë problem të enëve të zemrës etj. Komponenti i dytë i pemës së klasifikimit është "parashikuesi" ose ndryshorja e pavarur si mbipesha(obesity), duhanpirja apo pirja e alkolit etj. Këto janë karakteristika të cilat janë të lidhura fuqishem me variablat përgjegjëse për të cilat jemi të interesuar. Në përgjithësi, në bazën e te dhenave janë disa mundësi të cilat mund të arrihen me anë të variablave parashikuese. Komponenti i tretë i pemës klasifikuese është e gjithë bashkësia në bazën e të dhënave. Kjo bashkësi e bazës së të dhënave përfshin të dyja vlerat e variablave të pavarura (ose outcomes) dhe variablat parashikuese të cilat parashikojnë të ardhmen e një pacienti.
Komponenti i katërt i problemit të klasifikimit është si të parashikojë të ardhmen e bazës së të dhënave, i cili përbëhet nga të dhënat për pacientët për të cilët duhet të jemi në gjendje të realizojme parashikime të sakta. Është një besueshmëri e përbashkët se vlefshmëria e një bazë të dhënash është e nevojshme për tu vërtetuar, gjithashtu nuk është e nevojshme të verifikojmë performancën e një rregulli të tillë që na çon në marrjen e vendimit. Vendimi përfundimtar për një problem përfshin dy komponentë përveç ato që gjenden në një problem të klasifikimit. Këto komponente janë probabiliteti "para" për secilin rezultat, i cili përfaqëson probabilitetin që një pacient rastësisht i zgjedhur për të parashikuar të ardhmene tij, nëse do të ketë një rezultat të veçantë, si dhe probabiliteti i pasëm (posterior probability). Probabiliteti i pasëm është llogaritur normalisht nga përditësimi paraprak i probabilitetit duke përdorur teoremën e Bayes . Në terma statistikore, probabiliteti i pasëm është probabiliteti i ndodhje se ngjarjes A duke pasur parasysh se ngjarje B ka ndodhur.

2.2 Rregulli i shpërndarjes dhe strukturimi i pemës kualifikuese

Para se të shpjegojmë proçedurën e ndarjes le të japim disa informacione dhe përkufizimet për nyjet dhe gjethet. Nyja është një pikë me lidhje, ose një pikë me rishpërndarje, ose një pikë në fund për transmetimin e të dhënave. Nyja ka të programuar ose ka aftësinë për të njohur këtë proçes, po ashtu të transmetojë përpara për në nyjet e tjera. Një pemë mund të përkufizohet në mënyrë rekursive, si një grupim i nyjeve (duke filluar në një

nyje rrënjë), ku secila nyje është një strukturë e të dhënave e përbërë nga një vlerë, së bashku me një listë të nyjeve (e nyjeve fundore ose të ashtuquajtura"nyje fëmijë"), me kufizimet që ka çdo nyje që të mos ripërsëritet. Një pemë mund të përcaktohet në mënyrë abstrakte, si një e tërë (globalisht) si një pemë e renditur, me një vlerë të caktuar për çdo nyje. Të dyja këto perspektiva janë të dobishme: ndërsa një pemë mund të analizohet matematikisht si një e tërë, si një strukturë e të dhënave punuar më vete. Për shëmbull, duke kërkuar në një pemë si një e tërë, mund të flasim për "nyjen mëmë" të një nyje të caktuar, por në përgjithësi si një strukturë e të dhënave një nyje jepet vetëm kur përmban listën e fëmijëve të saj, por nuk përmban një referencë ndaj prindit te saj (nëse ka).
Të gjitha nyjet mund të arrihen nëpërmjet një peme të caktuar. Përmbajtja e tyre mund të modifikohet ose të fshihet, dhe elemente të reja mund të krijohen. Pema nyje tregon një bashkësi nyjesh që lidhen mes tyre. Pema fillon në nyje rrënjë dhe degët te cilat dalin jashtë për çdo nyje dhe fillon në nivelin më të ulët të pemës. Nyjet në pemë kanë një marrëdhënie hierarkike me njëri-tjetrin. Termat prind, fëmijë, vëlla janë përdorur për të përshkruar marrëdhëniet nyjet prind që kanë dhe fëmijë. Fëmijët në të njëjtin nivel quhen vëllezërit e motrat (vëllezër apo motra).
Në një pemë te nyjeve, nyja e parë quhet rrënja. Çdo nyje, përveç rrënjës, ka saktësisht një nyje mëmë. Një nyjë mund të ketë nje numër caktuar të fëmijëve. Një fletë është një nyje pa fëmijë.
Vëllai dhe motra janë nyje me të njëjtin prind.

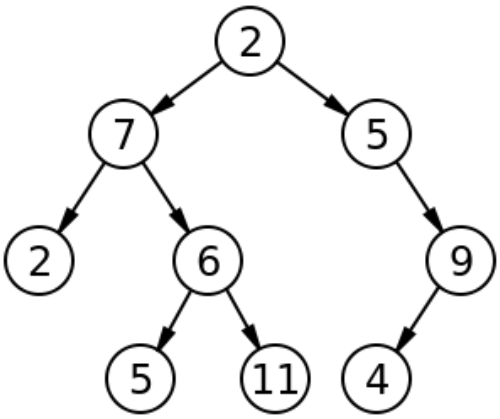


Figura 5: Shembull peme

Në këtë diagram, nyja e etiketuar 6 ka dy fëmijë, etiketuar 5 dhe 11, dhe një prind, etiketuar 7. Nyja rrënjë, në krye nuk ka prind. Për të ndarë një nyje në dy nyje pasardhëse, CART gjithmonë bën pyetje me pergjigje "po" ose "jo" përgjigje. Për shembull, pyetja "si janë të lidhura me sëmundjet e zemrës me historinë e familjes?"
Ndërtimi i një peme përfshin tre zgjedhje të rëndësishme që duhen bërë gjatë ndërtimit të pemës së klasifikimit. Zgjedhja e parë është si do të realizohet proçesi i ndarjes, cilat variabla shpjeguese do të përdoren dhe ku do të imponohet që të fillojë ndarja. Këto janë të përcaktuara nga rregullat e ndarjes. Zgjedhja e dytë përfshin përcaktimin e madhësisë së duhur te pemës, dhe pas kësaj duke përdorur një proçes krasitjeje arrijmë në pemën optimale që duam të gjejmë. Zgjedhja e tretë është për të përcaktuar si duhet të përfshihen kostot e aplikimit specifik. Kjo mund të përfshijë vendimet për caktimin e kostove të ndryshme. Ndarja binare dhe e vazhdueshme, siç përshkruhet më sipër, vlen për gjetur pemën e klasifikimit apo pemën e regresit. Megjithatë, kriteret për minimizimin e papastërtise së nyjes (domethënë, maksimizimin e homogjenitetit) janë të ndryshme për të dy metodat, si për klasifikimin dhe për regresin.

Strukturimi i pemës së klasifikimit e quajtur ndryshe dhe strukturimi i klasifikuesit të pemës binare ndërtohet si rezultat i një proçesi që përsëritet vazhdimisht në shpërndarjen që bëhet bazës së të dhënave në nënbashkësi, duke filluar nga një në dy e kështu me radhë. Figura 6 paraqet një peme me gjashtë klasa.

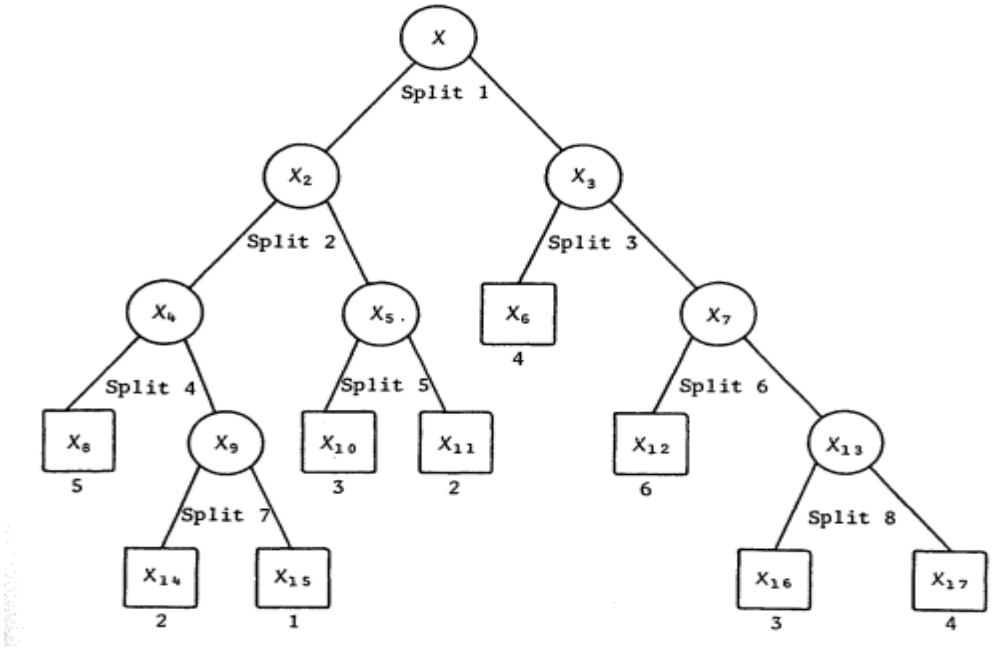


Figura 6: Pema me gjashtë klasa

Ne figurën 6 vëmë re se nyja e parë (X), të dhënat ndahen në dy nënbashkësi X_2 dhe X_3 të cilat janë të papajtushme me njëra tjetrën pra $X_2 \cap X_3 = \varnothing, dhe, X_2 \cup X_3 = X$, në mënyrë të ngjashme për nënbashkësitë e tjera X_4, X_5 janë të papajtueshme, po ashtu X_6 me X_7 e kështu me radhë dhe po të shikojmë nënbashkësitë

$X_8, X_{14}, X_{15}, X_{10}, X_{11}, X_6, X_{12}, X_{16}, dhe, X_{17}$ janë nyje fundore ku proçesi i shpërndarjes ka mbaruar. Këto nyje fondore janë pjesë të bashkësisë kryesore X. Të gjitha këto nyje fundore janë të etiketuara në një farë mënyre si klasa të rëndësishme të objektit të studimit tonë. Sikurse shihet mund të ketë më shumë se dy nyje fundore për nivele klasash të ndryshme. Nyjet fundore janë të shënuara me katror në këtë figurë dhe ato janë: $X_8, X_{14}, X_{15}, X_{10}, X_{11}, X_6, X_{12}, X_{16}, dhe, X_{17}$ të cilat gjithashtu janë pjesë të X-it. Në përgjithësi mund të jenë të paktën dy ose më shumë nyje fundore. Çdo nën pjesë te nyjet fundore është një klasifikues për rastin e figures 6 . Kemi gjashtë shpërndarje dhe nëntë nyje fundore ose klasifikues. Shpërndarjet janë bërë në bazë të kushteve që janë vënë për koordinatat e X= $(x_1, x_2, ...)$.

Një mundësi për shpërndarjen e parë në dy nënbashkësitë X_2 dhe X_3 është:

$X_2 = \{x; x_4 \le 7\}, X_3 = \{x; x_4 > 7\}.$

Për shpërndarjen e tretë në nënbashkësitë X_3, X_6, dhe, X_7 mund të jetë:

$X_6 = \{x \in X_3; x_3 + x_5 \le -2\}, X_7 = \{x \in X_3; x_3 + x_5 > -2\}$

Klasifikuesi i pemës parashikon klasat me përmasat e vektorit x në këtë mënyrë: Nga përkufizimi i shpërndarjes së parë duhet të përcaktohet egzaktësisht nëse x shkon në X_2 apo në X_3, si në rastin tonë nëse x shkon në X_2 nëse $X_4 \le 7$ dhe në X_3 nëse është më e madhe

se 7 dhe kështu japim përkufizimin e shpërndarjes së tretë të katërt e me radhë. Kur vlerat e x kanë arritur te nyjet fundore, atëherë klasat parashikuese janë të etiketuara saktësisht në këto nyje. Në tërësi ndertimi i pemës se klasifikimit ka tre elemente kryesore që janë:

a. Zgjedhja e shpërndarjes

b. Vendosja se kur duhet që të deklarohet që nyja është përfundimtare dhe nëse duhet të vazhdohet më tej.

c. Përcaktimi i çdo nyje fundore si një klasë.

Pika më e vështirë e problemit është si ta përdorim bazën e të dhënave dhe të përcaktohet kur dhe si do të bëhet shpërndarja, cilat do të konsiderohen dhe si do të përcaktohen nyjet fundore. Pra të përcaktojmë saktësisht se si do të bëhet një shpërndarje e mirë dhe kur duhet ta ndërpresim këtë proçes.

2.3 Ndërtimi i pemës së klasifikimit

Në fillim renditim disa përkufizime për klasifikuesin

Përkufizim 2.1 : Një klasifikues ose një rregull klasifikues është një funksion y(x) i përkufizuar në X, ku për cdo x, y(x) është i barabartë me një nga numrat 1……n.

Një mënyrë tjetër është; nëse përkufizojmë një nënbashkësi A_n të X në të cilën y(x)=n; kështu që $A_n = \{x; y(x) = n\}$. Bashkësitë A_1,..... A_n janë të papajtusheme dhe $X = \mathop{\mathrm{Y}}_{n} A_n$.

Përkufizim 2.2: Një klasifikues është një nenbashkesi e X në nënbashkëshkësitë Y , A_1,..... A_n të cilat janë të papajtueshme ku $X = \mathop{\mathrm{Y}}_{n} A_n$ të tilla që për çdo $x = \in A_n$ klasa parashikuse është n.

Për klasifikuesin e dhënë Y paraqesim funksionin $R_P(Y) = P[Y(X_1 \ldots X_n) \neq Y]$ dhe e quajmë këtë klasifikuesi i përgjithshëm i gabimit. Duhet që për një bazë të dhënash α dhe probabilitetin e dhënë P të gjejmë një funksion Y i tillë që të zvoglojmë në maksimum funksionin $R_P(Y)$. Ky veprim është relativisht i vështirë nëse ne lejojmë që klasifikuesi të jetë arbitrar. Nga eksperienca është gjetur se duke bërë disa kufizime për klasat ose më konkretisht duke i vendosur ato në renditje mund të zgjidhim problemin tonë dhe kjo është mënyra për të ndërtuar pemën e klasifikimit.

Së pari që të ndertojmë klasifikuesin duhet të jemi në gjëndje të përcaktojmë mënyrën si ta bëjmë shpërndarjen binare të nje bazë te dhënash në nën pjesë më të vogla. Idea bazë në këtë shpërndarje në nyje të ndryshme është që çdo nyje të ndahet në nyje pasardhëse dhe nyjet pasardhëse të jenë më të pastra se në nyjet parardhëse.

2.4 Pema fillestare dhe metodologjia e rritjes

Para se të diskutojmë metodologjinë e zhvillimit të pemës, duhet të formulojmë në mënyrë të kompletuar se çfarë metode do të përdorim për ta filluar dhe ndërtuar një pemë klasifikuese. Në një bazë të dhënash L që e përdorim si shembull për një klasë të caktuar j, le të kemi Nj numrin e klasave në klasën N. Shpesh probabilitetet paraprake $\{\pi(j)\}$duhet të jenë propocional me {Nj/N}. Në disa shembuj këto porpocione ndoshta nuk reflektohen, por mundet që një pjesë e saj të mund ta plotësojë këtë gjë.

Në se marrim një nyje T, le të kemi N(T) numrin total të rasteve në bazen e te dhenave L, ku $x_n \in T$ dhe $N_j(T)$ numri i klasave j në T. Raporti i rasteve të klasave j në L është $N_j(T)/N_j$. Për një bashkësi të dhënë probabiliteti $\pi(j)$ të interpretohet si probabiliteti

që një klase j të jetë prezent në pemë. Kështu që kemi: $p(j,T) = \pi(j)N_j(T)/N_j$ dhe këtë e konsiderojmë si një vlerësues zëvendësues për probabilitetin që ky rast të jetë në klasën j dhe në nyjen T. Ky ri vlerësim p(T) i probabilitetit që në çdo rast bie në nyjen T është: $P(T) = \sum_j p(T) = 1$ dhe ky probabilitet në rastin kur është në klasën j dhe bie në nyjet T është: $p(j/T) = p(j,T)/p(T)$ dhe kënaq këtë kusht $\sum_j p(j/T) = 1$, kur $\{\pi(j) = \left\{\frac{N_j}{N}\right\}\ atehere\ p\left(\frac{j}{T}\right) = \frac{N_j(T)}{N(T)}\ keshtu\ qe\ \{p(\frac{j}{T})\}$ janë relativisht porpocional për klasën j dhe nyjen T.

Katër elementet që janë të nevojshëm për proçedurën e rritjes së një peme fillestare janë:

1. Në njëbashkësi binare të kemi këtë formë $\{\text{është}, x \in A?\}, A \subset X$
2. Kriteri i shpërndarjes mirësia e $\Phi(s,T)$ duhet të vlerësohet për çdo shpërndarje s që duhet të bëjmë në çdo nyje T.
3. Një rregull për të ndaluar shpërndarjen e mëtejshme.
4. Një rregull për të shënuar apo filluar çdo nyje përfundimtare në një klasë të caktuar.

Në çdo ndarje binare që ne i bëjmë nyjes T nëpërmjet një shpërndarje s duhet të marrim dy nëndegë në të cilat të kemi në njërin krah T_L "po" dhe në tjetrin T_R "jo". Në fakt nëse pyetja është $\{x \in A\}, atëhere, T_L = T \text{ I } A, dhe, T_R = T \text{ I } A^c$, atëherë plotesi i A-së është plotesi i A-së në X. Në çdo nyje të ndërmjetme T shpërndarja e selektuar është shpërndarja s^* e cila e maksimizon $\Phi(s,T)$.

Të përcaktojmë një bashkësi te pershtashme për një shpërndarje binare s te sejcila nyje. Në përgjithësi është e thjeshtë të konceptohet që bashkësia S të shpërndahet duke u përmbajtur parimit që për vlerat duhet të kemi $x \in A?, A \subset X$ dhe çdo shpërndarje s shoqëron apo dërgon të gjitha x_n në përgjigjet po apo jo në t_R apo t_L dhe në shëmbullin që diskutuam më lartë papastërtia e nyjes është përkufizuar si më poshtë:

$$i(t) = -\sum_1^6 P(j/t)\log P(j/t)$$

Pema rritet në këtë mënyrë: te nyja e parë t_1 aplikojmë një shpërndarje s^* e cila jep zvogëlimin më të madh të papastërtisë; $\Delta i(s^*, t_1) = \max_{s \in S} \Delta(s, t_1)$, ku t_1 është shpërndarë në t_2 dhe t_3 duke përdorur shpërndarjen s^* dhe të njëjtën proçedurë kërkimore për më të mirën $s \in S$ në të dyja nyjet t_2 dhe t_3 të para si të vecuara. Që te arrihet te nyjet përfundimtare gjatë rritjes së pemës duhet të përdorim një mënyrë kerkuese në bazë të rregullave të përcaktuara. Kur të arrihet në një nyje ku papastërtia nuk ka ndonjë zvogëlim të rëndësishëm me nyjen e mëparshme atëhere këtë nyje e konsiderojmë si përfundimtare. Karakteri i klasës së nyjes përfundimtare është i përcaktuar nga rregulli i papastërtisë i specifikuar si më poshtë: $P(j_0/t) = \max_j P(j/t)$ ku t është përshtatur si një klasë j_0 e nyjes fundore.

Kriteri i shpërndarjës së mire është si nënprodukt i funksionit të papastërtisë.

Përkufizim 2.3: Funksion i papastërtisë do të quhet funksioni Φ i përkufizuar në një bashkësi ku të gjithë elementet janë vendosur në një renditje të caktuar $p_1, p_2, ... p_k$ duke kënaqur kushtin $p_j \geq 0$, ku j=1,2,3,…K dhe $\sum_j p_j = 1$.

Funksioni i papastërtisë mund të përcaktohet në mënyra të ndryshme, por ai duhet të gezojnë tre vetitë e mëposhtme:

a. Φ arrin maksimumin vetëm atëhere kur kemi shpërndarje uniforme, domethënë të gjitha p_j janë të barabarta.

b. Φ arrin minimumin vetëm te pikat (1,0,0….0),(0,1,0,0….),(0,0,1,0,0…),……., (0,0,0,0,…..1), kur probabiliteti i të qënurit në klasë të çfardoshme është 1 dhe 0 për klasat e tjera.

c. Φ është funksion simetrik per $p_1, p_2, ... p_k$, edhe nëse përkëmbejmë p_j, Φ qendron konstant.

Funksion i papastërtisë është një funksion Φ i cili është i përkufizuar si lidhje e renditur numrash $(p_1, p_2, ...p_j)$ që plotësojnë këtë kusht $p_j \geq 0$, ku j=1,2,3….J dhe $\sum_j p_j = 1$, gjithashtu kënaq vetitë e mësiperme

Me tu dhënë funksioni i papastërtisë Φ, masë e papastërtisë për ndonjë nyje t me shpërndarje s që çon një raport p_R të bazës së të dhënave në drejtimin t_R te nyja t. dhe raportin p_L, te, t_L.

Njehsojmë uljen e papastërtisë si më poshtë:

$$\Delta i(s,t) = i(t) - p_R i(t_R) - p_L i(t_L)$$

Marrim shpërndarjen e mirë $\phi(s,t)$ të jetë $\Delta i(s,t)$.Supozojmë se kemi bërë disa shpërndarje deri sa të arrijmë në një nyje përfundimtare. Bashkësinë e shpërndarjeve që është përdorur së bashku me renditjen e nyjeve që kemi përdorur e quajmë shpërndarje binare të pemës T.

Shënojmë bashkësinë e nyjeve përfundimtare me $\tilde{T}$, dhe bashkësinë I(t)=i(t)p(t) , atëherë papastërtia e pemës mund te paraqitet si më poshtë:

$$I(T) = \sum_{t \in \tilde{T}} I(t) = \sum_{t \in \tilde{T}} i(t)\, p(t)$$

Eshtë e qartë se selektimi i shpërndarjes që maksimizon $\Delta i(s,t)$ është ekuivalent me selektimin e atyre shpërndarjeve që minimizojnë papastërtinë e pemës në tërësi I(T). Në se marrim një nyje $t \in \tilde{T}$ dhe bëjmë një shpërndarje s në nyjet dhe t_R, dhe, t_L, atëherë pema e re T' ka papastërti $I(T') = \sum_{\tilde{T}-t} I(t) + I(t_L) + I(t_R)$ dhe rënia e papastërtisë së pemës është:

$$I(T) - I(T') = I(t) - I(t_L) - I(t_R). \qquad *$$

Kjo varet vetëm në nyjen t dhe shpërndarjen s. Rrjedhimisht maksimizimi dhe zvogëlimi i papastërtisë gjatë shpërndarjes “t” është ekuivalent me maksimizimin e madhësisë $\Delta I(s,t) = I(t) - I(t_L) - I(t_R)$

Përcaktojmë raportin p_L, p_R për nyjen t të selektuar nga bashkësia e të gjitha nyjeve me t_L, dhe, t_R e përcaktuar si më poshtë.

$p_L = p(t_L)/p(t), p_R = p(t_R)/p(t)$, atëherë $p_L + p_R = 1$ dhe barazimi (*) mund të shkruhet si:

$\Delta I(s,t) = [i(t) - p_L i(t_L) - p_R i(t_R)]p(t) = \Delta i(s,t)p(t)$

Sikurse shihet $\Delta I(s,t) \neq \Delta i(s,t)$ me faktorin p(t) dhe kështu përsëritja e vazhdueshme e shpërndarjes selektive të pemës në tërësi çon në minimizimin e papastërtisë së pemës. Fiksimi i pikës se kur do të ndalohet shpërndarja e mëtejshme bëhet në këtë mënyrë fikson një vlerë $\beta > 0$ dhe konsiderojmë nyje përfundimtare nyjen që plotëson kushtin $\max_{s \in S} \Delta I(s,t) < \beta$.

Një shpërndarje natyrale e nyjeve do të quhet e mirë nëse përdoret kriteri në të cilën për çdo nyje të reduktojmë koston e joklasifikimit, për këtë së pari duhet të japim përkufizimin e funksionit të papastërtisë (të cilin e dhamë më lartë). Në çdo nyje fundore zgjedhim që të bëjmë shpërndarjen në të cilën të reduktojmë I(T) ose në mënyrë ekuivalente të maksimizojmë $\Delta I(s,t) = I(t) - I(t_L) - I(t_R)$ ose $\Delta i(s,t) = i(t) - p_L i(t_L) - p_R i(t_R)$.

Brënda kesaj peme duket më shumë natyrale që të marrim si papastërti të pemës R(T) dhe rizëvëndësimi është një vleresim i përafërt i pritshmërisë së raportit te misklasifikimit. Kjo është e njëjtë sikurse të përkufizojmë i(t) si të barabartë me r(t), ku $r(t) = \min_i \sum_j c(i/j)p(j/t) = 1 - mak_j\, p(j/t)$, atëherë shpërndarja më e mirë t maksimizon $r(t) - p_L r(t_L) - p_R r(t_R)$, gjë e cila është e njëjtë me maksimizimin e $R(t) - R(t_L) - R(t_R)$ dhe funksioni i papastërtisë së nyjes është $\phi(p_1,.....p_j) = 1 - mak_j\, p_j$.

Ky funksion kënaq të gjitha vetitë që ne dhamë në përkufizimet e mësipërme. Ky kriter paraqet disa vështiresi së pari kriteri i përcaktuar më sipër ndoshta është zero për të gjitha nyjet e shpërndarjeve të mësipërme S, për këtë kemi teoremën e mëposhtme.

Teorema 2.1: Për çdo shpërndarje të t-së në t_L, dhe, t_R, $R(t) \geq R(t_L) + R(t_R)$ nëse $j^*(t) = j^*(t_L) = j^*(t_R)$.

Vertetim: Shënojmë $R(t) = \sum_j C(j^*(t) \mid j)p(j,t) = \sum_j C(j^*(t) \mid j)[p(j,t_L) + p(j,t_R)]$ ose

$R(t) - R(t_L) - R(t_R) =$

$\sum_j C(j^*(t) \mid j)p(j,t_L) - \min_i \sum_j C(i/j)p(j,t_L) + \sum_j C(j^*(t) \mid j)p(j \mid t_R) - \min_i \sum_j C(i \mid j)p(j,t_R).$

Krahu i djathtë shihet qartë se nuk është negativ dhe është i barabartë me zero nëse $j^*(t) = j^*(t_L) = j^*(t_R)$ dhe tani supozojmë se e kemi atëherë vështirësia e dytë është të përcaktohet sasia e nyjes përkatëse.

2.5 Pema e Klasifikimit

Nëse kemi një bazë të dhënash me shumë variabla si në Tabela 1

Le të shënojmë me $X_1, X_2, ... X_n$, Y variablat e çdo shtylle, të cilat janë variabla të rastit, ku secili variabël ka një fushë vlerash të caktuar. Variabli Y ka një fushë vlerash ={1……..m). Variablat $X_1, X_2, ... X_n$ i quajmë variabla atribuese të cilat marrin vlera të ndryshme numerike ose dhe kategorike dhe variablin Y e quajmë variabël parashikues (i varur nga

atributet). Klasifikuesi Y është një funksion Y me fushë

përcaktimi $FP(x_1)xFP(x_2).x....xFP(x_n)$.

Në se marrim α= $FP(x_1)xFP(x_2).x....xFP(x_n)xFP(Y)$ si një bashkësi ngjarjesh.

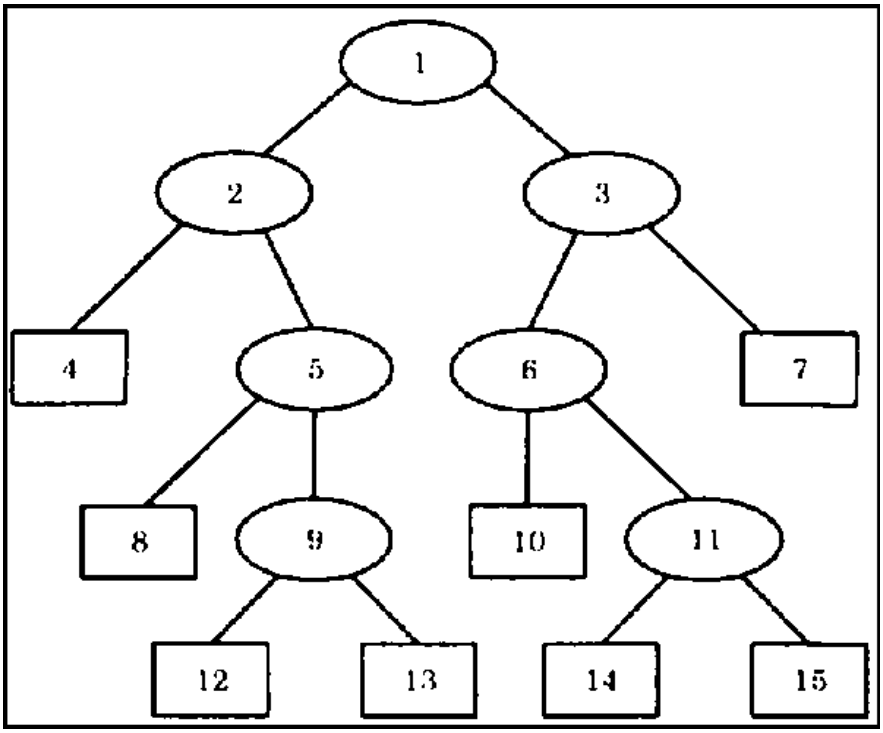
Do të supozojmë se klasifikimi i kësaj baze të dhënash bëhet në mënyrë probabilitare duke gjeneruar bashkesitë e bazës së të dhënave në lidhje me një shpërndarje probabilitare të panjohur P në lidhje me bashkësinë e ngjarjeve α(alfa). Për klasifikuesin Y dhe për probabilitetin e dhënë P në α duhet të ndertojmë një klasifikues i cili në mënyrë sistematike të parashikojë për çdo element të bazës së të dhënave sipas një rregulli të caktuar për çdo element që është te Y dhe për vlerë nga bashkësia e vektorit x në X të ketë një relacion të caktuar.

Një pemë e klasifikimit është një cikël i caktuar grafikësh $t(\tau)$ në formën e një peme. Rrënja kryesore e kësaj peme t(τ) nuk ka ndonjë rrënjë tjetër parardhëse. Çdo nyje tjetër ka vetëm një rrënjë dhe mund të ketë 0 deri në dy degë që dalin dhe nyjen e fundit që nuk ka më dalje do ta quajmë gjethe dhe e shënojmë me gërmën T ose e quajmë nyje fundore. Çdo gjethe nyje është në nivelin e një klase të caktuar. Çdo nyje fundore ose gjethe ka një atribut të caktuar të cilin e shënojmë me $_{X_T}$ dhe e quajmë atribut i shpërndarjes. Në çdo linjë të brendshme kemi dy nyje, të cilat i quajmë njërën si prind dhe tjetrën si fëmijë. Nyja pasardhëse ose fëmija ka në brëndësi të saj edhe parashikuesin(atributin) të cilin e shënojmë q(T, T'). Çdo degë (T, T') nga një nyje e brëndëshme T ne nyjen fëmijv T' ka një atribues, parashikues $q(T, T')$ që e shoqeron, ku $q(T, T')$ përfshin vetem atributin $_{X_T}$ te nyjes T. Bashkësia e të gjitha nyjeve parashikuse Q_T që del nga të gjitha nyjet e brendëshme T duhet domosdoshmërisht të përmbajë parashikime të papajtueshme, parashikime të cilat me gjithë atributet e shpërndarjes të japin parashikime të cilat janë të vërteta. Duke pasur parasysh pemën klasifikuese τ, mund të përkufizojmë klasifikuesin $Y_\tau(x_1 ..... x_n)$ në kete mënyrë:

$$Y(x_1 ... x_m, T) = \left\{ \begin{array}{c} Klasifikohe(T)\ ne\ se\ T\ eshte\ nyje\ fundore \\ Y(x_1 ... x_m, T_j)\ ne\ se\ T\ eshte\ nyje\ e\ brendeshme,\ X_i\ eshte\ shenuar\ si\ T, \\ dhe\ q(T, T_j) = i\ vertet \end{array} \right\}$$

$Y_T(x_1 ... x_m) = Y(x_1 ... x_m, Rrenja(T))$, ku rrënja(T) është nyja fillestare.

Kështu që të bëjmë një parashikim fillojmë te nyja fillestare te rrënja dhe ndërtojmë pemën me parashikime të vërteta deri sa nyja përfundimtare ose te gjethja. Nëse pema τ është e mirëformuar, atëherë pema klasifikuese e përkufizuar mësipër do të japë një funksion $Y\tau()$, cili është një klasifikues i mirë përkufizuar .



Legjenda
Ovals = nyjet
Katrorët = Nyjet fundore(gjethet)
1 = Nyja rrënjë
Vijat bashkuese= degët

Figura 7: **Paraqitja e nyjeve të ndërmjetme dhe fundore të një peme**

Secila nyje e brëndëshme korrespondon me një madhësi të një variabli të caktuar. Vijat bashkuese të një nyje që quhet prind dhe nyja pasardhëse që konsiderohet si fëmijë. Dy variante janë propozuar për pemën klasifikuese. Nëse lejojmë e shumta dy degë për çdo nyje të brendshme marrim pemën klasifikuse binare ose në rast të kundërt marrim pemën klasifikuese me k-dalje.

Pema binare është prezantuar për herë të parë nga Breiman në vitin 1984, ndërsa pema klasifikuse me k-dalje është prezantuar për herë të parë nga Quinlan në vitin 1986. Diferenca kryesore midis këtyre dy pemëve lidhet me atribute diskrete apo atribute të vazhdueshme. Të dyja lejojnë parashikues të formës X>c, ku c është një konstante. Për pemën klasifikuese binare, parashikuesit e formës $X \in S$, ku S është një nënbashkësi e vlerave të mundëshme të atributeve të lejuara. Në këtë kuptim për çdo nyje duhet të përcaktojmë, atributin e ndarjes dhe bashkësinë e ndarjes. Për atributet diskrete në k-daljet e pemës kualifikuese, mund të jenë aq shumë parashikues shpërndarëse sa dhe vlerat e atributeve të variablave dhe të gjitha janë të formës $X = x_i$ ku x_i është një nga vlerat e mundëshme të X. Për variablat e vazhdueshme, të dy tipet e pemës klasifikuese, te nyja shpërndarëse në dy pjesë e formës $X \le s$ dhe X>s, ku numri real s është quajtur pikë e shpërndarjes.

Tani formalisht do të përkufizojmë pemën e klasifikimit me ndertimin e një problemi duke e ilustruar me klasifikuesin e përgjithshëm.

Në se është dhënë një bazë te dhënash D me N grupe identike të pavarura nga α, grupuar në lidhje me shpërndarjet probabilitare P, duhet të gjejmë një pemë kualifikuese τ e tillë që përqindja e gabimit kualifikues e përcaktuar nga funksioni $R_P(C_T)$ dhe klasifikuesi

korespondues C_T të minimizohet. Çështja kryesore për tu zgjidhur në pemën kualifikuese dhe në veçanti në problemin kualifikues në përgjithësi, është fakti që kualifikuesi duhet të jetë një parashikues i mirë për shpërndarjen, por jo për grupet që dalin nga shpërndarjet. Kjo do të thotë se nuk mund që thjeshtë të ndërtojmë një klasifikues që të jetësa më i mirë që të jetë e mundur duke respektuar grupet që do të krijohen, kështu që duhet të theksojmë se nuk mund të arrijmë të kemi një pemë kualifikuese ku gabimi të jetë zero me një pemë kualifikuese arbitrare nëse nuk kemi ndonjë kontradiksion me shëmbujt tanë. Një tjetër koncept është dhe "zhurma". Fenomeni i zhurmës është i quajtur ndryshe overfitting është një nga çështjet e rëndësishme të klasifikimeve. Për këtë arsye pema kualifikuse është ndërtuar në dy faza. Në fazën e parë ndertojmë një pemë aq të madhe sa është e mundur, në një mënyrë që të minimizojmë gabimin duke respektuar disa nënbashkësi të variablave të bashkësisë së basës së të dhënave. Në fazën e dytë kemi krasitjen e kësaj baze të dhënash. Kjo kraasitje bëhet duke lëvizur, pra duke hequr disa nën-peme duke reduktuar dhe vleresuar gabimin e përgjithshëm gjatë gjithë proçesit të krasitjes.

Disa nga fazat për ndërtimin e një peme klasifikuese janë: Ndërtimi i një peme optimale, me kosto minimale të pemës me një funksion të thjeshtë dhe ndërtimi i një peme klasifikuese optimale me përmasa të tilla që të përfshijë të gjithë informacionin e dhënë në nje bazë të dhënash. Për kete, shumica e algoritmeve të përdorura për pemën klasifikuese duhet të përdorim;

Input node T, metoda e shpërndarjes seleksionuse V.

Output: Pema klasifikuese τ për D me rrënjë të T.

Ndërtojmë një pemë (Nyja T, ndarja e të dhënave D, metoda e selektimit dhe e shpërndarjes së atributeve V).

1. Aplikojmë metodën selektive të shpërndarjes V në D të gjejmë atributet e shpërndarjes X për nyjen T.
2. Le të jetë n numri i fëmijëve të nyjes T
3. nëse T shpërndahet.
4. Ndahet D në $D_1, D_2, \ldots\ldots D_n$ dhe etiketojmë me shënimin T me atributet e shpërndarjes X.
5. Krijojmë nyjet fëmijë $T_1, T_2, \ldots\ldots T_n$ për T dhe etiketojmë (T, T_1) me parashikuesin $q(T, T_1)$
6. Për çdo $i \in \{1,2,3 \ldots n\}$
7. Ndërtojmë pemën (T_i, D_i, V)
8. Fund për çdo nyje
9. Tjetër
10. Etiketohet T në klasë kryesore të D
11. Fund nëse

Prezantimi i skemës për pemën klasifikuese duke përdorur algoritmin Greedy, konsiston në vendimin që për çdo hap të përdoret atributi i shpërndarjes dhe shpërndarja e bashkësisë ose e pikave, nëse është e nevojëshme, ndarjen e bazës së të dhënave në lidhje me shpërndarjet e reja determinuese, duke vazhduar me parashikime pasi të kemi zbatuar shpërndarje të njëpasnjëshme dhe duke e përsëritur këtë proces për cdo nyje pasardhëse në këtë pemë. Proçesi i ndërtimit në një nyje është i përfunduar kur një kusht përfundimtar është arritur. Diferenca midis dy metodave klasifikuese në pemë është se në rastin e k-daljeve nuk zbatohet shpërndarja e bashkësisë për madhësitë diskrete. Te paraqesim në mënyrë më të detajuar si të aplikojmë atributin e shpërndarjes dhe shpërndarjen e bashkësisë ose të pikës që realizohen në çdo hap në mënyrë të vazhdueshme në procesin e ndertimit të pemës dhe do të tregojmë disa kushte të cilat duhet të kënaqen në përfundim të këtij proçesi. Sikurse u theksua

kemi dy tipe të pemëve klasifikuese binare dhe me k-dalje dhe ndryshimi midis tyre është se në tipin me k-dalje nuk ka nevojë të bëhet shpërndarje e bashkësisë për atributet diskrete. Do të diskutojmë se si do të bëhet shpërndarja e atributeve dhe shpërndarja e bashkësive ose e pikave, që do të aplikohet në çdo hap në proçesin e vazhdueshëm që do të përdoret në ndertimin e pemës.
Në fillim maksimizojmë pemën e cila mund të jetë në të vërtetë shumë komplekse. Optimizimi i pemës na ndihmon në gjetjen e përmasave të pemës së duhur e cila do të na jape zgjidhjen për problemin që duhet të studjojmë. Ky proçes ka dy anë, nga njëra anë atë të rritjes maksimale dhe nga ana tjetër atë të krasitjes së kësaj peme për të gjetur pemën e duhur. Për të realizuar këtë proçes duhet të përdorim dy algoritme, atë të optimizimit të pemës dhe vlerësimit të kryqëzuar.
Katër elementet e nevojshme në proçedurën e rritjes se pemës fillestare:
1. Një bashkësi me dy pyetje binare të formulara $\{\text{është}, x \in A?\}, A \subset X$
2. Mirësia e kriterit të shpërndarjes $\phi(s,t)$ e cila mund të vlerësohet për çdo shpërndarje s dhe çdo nyje t.
3. Një rregull i caktuar se kur duhet të qëndrojmë.
4. Një rregull për caktimin e çdo nyje fundore në një klasë të caktuar.
Si rezultat i përdorimit të dy pyetjeve binare e cila gjeneron një S të caktuar dhe ndan s në nyjet e ndryshme ku një nyje të caktuar ti jap dy vlera "Po" ose "Jo" dhe nëse merr vlerën e parë ajo duhet të shkojë në t_L, dalja e majtë dhe lëvizje e dytë në të djathtë tr. Në fakt, nëse pyetja është $\{\text{është } x \in A?\}$, atëherë $t_L = t \cap A$ dhe $t_R = t \cap A^c$, ku A^c është plotesi i A në X.
Së pari zgjedhim një nga atributet si rrënjë duke marrë parasysh të gjitha vlerat e saj si degë. Ne mënyrë rekursive, zgjedhim nyjet e tjera të brendshme me vlerat e tyre si degë. Pastaj duke përsëritur këtë proces deri sa të gjitha subjektet janë të së njëjtës klasë, nyja bëhet gjethe etiketuar me atë klasë. Ndodh që, duhet të ndalojmë proçesin, kur nuk ka më shumë subjekte të mbetura ose kur atributet më të reja janë për tu përdorur si nyje. Më në fund, klasifikimi i vlerës së synuar (subjekt) është i bazuar në atë klasë e cila ka numrin më të madh të elementeve.

2.6 Shpërndarja e atributeve dhe selektimi i tyre

Në çdo hap që aplikojm algoritmin e ndërtimit të vazhdueshëm(rekursive), duhet të vendosim se cilën nga variablat duhet të shpërndajmë. Qellimi i shpërndarjes është që ta ndajmë aq shumë sa është e mundshme në klasa të ndryshme, të cilat do të jenë me etiketime të ndryshme. Që ta bëjmë këtë në mënyrë intuitive dhe të dobishme, duhet të përdorim sistemin metrik që të vlerësojmë me një afërsi se sa ka ndikuar ndarja e klasave dhe sa është përmirësuar kur një shpërndarje e veçantë është zbatuar. Një sistem të tillë metrik ku zbatojmë ndryshe metodat selektuese të shpërndarjes. Një nga metodat kërkuese është kriteri i shpërndarjes selektive i cili prodhon një pemë me produktivitet dhe saktësi të lartë (Murthy 1997). Një nga metodat më popullore të shpërndarjes selektive është ajo e bazuar te papastërtia (Breiman et al, 1984; Quinlan 1986). Studimet dhe zbatimet e ndryshme kanë treguar se kjo metodë ka një saktësi shumë të mirë parashikuse dhe është e thjeshtë në zbatim. Secila nga metodat e mesiperme e zbatuar në shpërndarjet selektive bazohet në funksionin e papastertise $\phi(p_1 \ldots . p_k)$, ku p_j duhet te interpretohet si probabiliteti i të parit të një klase të etiketuar si y_j. Intuitivisht, funksioni përcakton masën e papastërtisë së bazës së të dhënave. Disa nga vetitë të cilat duhet të kënaqë ky funksion janë:

1. Të jetë i luget, domethene $\frac{\partial^2 \phi(p_1 \ldots . p_k)}{\partial p_i^2} > 0$
2. Të jetë simetrik në të gjitha argumentat, në se π është një përkëmbim i tillë që: $\phi(p_1 \ldots\ldots p_k) = \phi(p_{\pi_1} \ldots . p_{\pi_k})$.
3. Të ketë një maksimum të vetëm të (1/k……1/k) kur përzierja e klasave të etiketuara është në kulmin e papastërtisë.
4. Të arrihet minimumi te (1,0,……..0),(0,1,…….o), (0,…….1), kur përzierja e klasave të etiketuar është në kulmin e pastërtisë.

Papastërtia për nyjen T të pemës klasifikuse që ne filluam të ndërtojmë është: $i(T) = \phi(P[Y = y_i \mid T], \ldots, P[Y = y_k \mid T]$ ku $P[Y = y_j \mid T]$ është probabiliteti që çdo klasë e etiketuar si y_j mund të arrijë te nyja T.

Për një bashkësi të dhënë Q të predikateve të shpërndara për atributet e variblave X, që shpërndan një nyje T në nyje të tjera T1,……Tn, do të përkufizojmë reduktimin në papastërti si më poshtë:

$$\Delta i(T, X, Q) = i(T) - \sum_{i=1}^{n} P[T_i \mid T] \bullet i(T_i) = i(T) - \sum_{i=1}^{n} P[q(T, T_i)(X) \mid T] \bullet i(T_i), \quad (1)$$

intuitivisht, reduktimi i papastërtisë që në sasi është sa sasia e pastërtisë e fituar nga shpërndarja, ku papastërtia pas shpërndarjes është e barabartë me shumën e të gjitha papastërtive të nyjeve të dala nga nga çdo nyje prind. Duke u nisur nga ilustrime të ndryshme me funksionin e papastërtisë ne mund të formulojmë dy kriteret seleksionuese të shpërndarjes:

GINI GAIN. Ky kriter shpërndarje është praqitur për herë të parë nga Breiman (1984), dhe me funksionin e papastërtisë si Gini index: $gini(T) = 1 - \sum_{j=1}^{k} P[Y = y_j \mid T]$. Duke zvëndësuar në (1) gjejmë përfitimin Gini të kriterit të shpërndarjes:

$$GG(T, X, Q) = gini(T) - \sum_{i=1}^{n} P[q_{(T,T_i)}(X) \mid T] * gini(T_i) \qquad (2)$$

Për dy klasat e etiketuara, fitimi Gini merr një formë me kompakte:

$$GG_b(T, X, Q) = P[Y = y_0 \mid T]^2 \frac{(PY = Y_0 \mid T] - P[T \mid T])^2}{P[T_1 \mid T](1 - P[T_1 \mid T]} \qquad (3)$$

Kriteri i shpërndarjes është praqitur për herë të parë nga Quinlan (1986), i cili e konsideroi funksionin e papastërtisë si një rastësi(entropy) të një baze të dhënash ku entropia është:

$entropy(T) = -\sum_{j-1}^{k} P[Y = y_j \mid T] \bullet \log P[Y = Y_j \mid T]$. Duke e zvëndësuar te (1) gjejmë kriterin e fitimit si më poshtë:

$$IG(T, X, Q) = entropy(T) - \sum_{j=1}^{n} P[q_j(X) \mid T] \bullet entropy(T_j)$$

Raporti Gain. Quinlan prezantoi versionin e tij të thjeshtuar për fitimin dhe lëvizi fitimin nga atributet e variablave me fushë përcaktimi të gjerë (Quinlain 1986).

$$GR(T, X, Q) = \frac{IG(T, X, Q)}{-\sum_{j=1}^{[Dom(X)} P[X = x_j \mid T] \log P[X = x_j \mid T]}$$

Gjithashtu kemi dy metoda të tjera të shpërndarjes të cilat janë të njohura në statistikë si: χ^2

$$\chi^2(T,X) = \sum_{i=1}^{[Dom(X)]} \sum_{j=1}^{k} \frac{(P[X = x_j \mid T] - P[X = x_i, Y = y_j \mid T])^2}{P[X = x_i \mid T] \bullet P[Y = y_j \mid T]}$$

Kjo statistikë vlerëson se sa një klasë e etiketuar varet nga vlerat e atributeve të shpërndarjes. Testi χ^2 nuk varet nga bashkësia Q e parashikueseve shpërndarës. Sipas Shao 1999 testi χ^2 ka asimptotikisht një shpërndarje χ^2 me gradë lirie [Dom(X)|(k-1).

Statistika G^2

$$G^2(T,X,Q) = 2 \bullet N_T \bullet IG(T) \log_e 2$$

N_T është numri i rekordeve në nyjen T. Asimptotikisht statistika G^2 ka një shpërndarje χ^2 (Mingers, 1987). Për madhësitë me atribute diskrete me k-dalje në pemën klasifikuese, bashkësia e parashikueseve është përcaktuar duke specifikuar variablat atribute. Duhet të përcaktojmë bashkësinë e ndarjeve më të mirë, pikat ti vendosim në një renditje të caktuar që të vlerësojmë se sa e mirë është një ndarje në variabla të veçanta.

2.7 Selektimi i bashkësisë së ndarjes për atributet diskrete

Shumica e metodave të selektimit të bashkësive përdorin të njëjtin kriter ndarje të variablave dhe vlerësojnë se cila ndarje është më e mira. Në përgjithësi proçesi i gjetjes së bashkësisë së ndarjes është një llogaritje intensive përveç rastit kur fusha e përcaktimit e atributeve të ndarjes është etiketuar në klasa të vogla. Këtë e ka trajtuar Breiman (1984) përderisa ky algoritëm përsëri është duke u përdorur për rastet kur kemi të bëjmë me dy klasa, atëherë kur kriteri selektiv i papastërtisë përdoret, si më poshtë:

Teorem 2.2(Breiman 1984). Le të kemi i një bashkësi e fundme, ku $q_i, r_i, i \in I$ janë elemente pozitive dhe funksioni $\Phi(x)$ të jetë një funksion i mysët. Për I_1, I_2 pjesë e I një optimum i problemit:

$$\arg\min_{I_1,I_2} \sum_{i \in I_1} q_i \Phi\left(\frac{\sum_{i \in I_1} q_i r_i}{\sum_{i \in I_1} q_i}\right) + \sum_{i \in I_2} q_i \Phi\left(\frac{\sum_{i \in I_2} q_i r_i}{\sum_{i \in I_2} q_i}\right)$$ ka vetinë që $\forall i \in I_1, \forall j \in I_2, r_1 < r_2$

Nga teorema rrjedh një algoritëm efiçent që zgjidh këtë problem optimizimi i cili rendit elementet r_i nga I në rendin rritës. Kjo quhet shpërndarja optimale dhe konsiderohet një shpërndarje normale. Në këtë mënyrë kemi që:

$I = Dom(X), q_i = P[X = x_i \mid T], r_i = P[C = c_0 \mid X = x_i, T]$ dhe $\Phi(x)$ është indeksi Gini ose entropia për dy klasat e etiketuara, ku të dyja janë konkave:

$$gini(T) = 2P[C = c_0 \mid T](1 - P[C = c_0 \mid T])$$

$$entropy(T) = -P[C = c_0 \mid T] \ln(P[C = c_0 \mid T] - (1 - P[C = c_0 \mid T]) \ln(1 - P[C = c_0 \mid T])$$

Kriteri i optimizimit deri në një faktor konstant është Gini Gain. Per te gjetur shpërndarjen më të mirë dhe më efektive duhet ti renditim në rendin rritës gjithë elementet e bazës së të dhënave(DomX), ku $r_i = P[C = c_0 \mid X = x_i, T]$, për të bërë shpërndarjen. Në këtë studim, të gjitha kriteret e shpërndarjes janë paraqitur, por më shumë do të përdorin indeksin Gini ose informacionin e shumëzimit të ashtuquajturit "gain" fitim me një faktor i cili nuk varet nga bashkësia e shpërndarjes. Me zhvillimet e mëtejshme në vitin 1997 Loh dhe Shih kanë propozuar teknika te ndryshme të cilat konsistojnë në trasformimin e vlerave diskrete në vlera të vazhdueshme dhe duke përdorur shpërndarjen e cila quhet metoda "split point"

shpërndarja e pikave me atribute të vazhdueshme të fitojmë shpërndarjen me atribute diskrete.

2.8 Selektimi i ndarjes së pikës për atributet e vazhdueshme

Dy janë metodat që janë propozuar për shpërndarjen e pikës që të gjendet zgjidhja më e mirë për shpërndarjet e atributeve të vazhdueshme:

Analiza kuadratike e diskiminantit dhe i ashtuquajturi "exhaustive search" kërkimi dobësues. Kërkimi dobësues përdor të njëjtin selektim të kriterit të shpërndarjes që përdor metoda e shpërndarjes së atributeve dhe konsiston në vlerësimin e të gjitha mënyrave të mundshme të shpërndarjes të fushës së përcaktimit për atributet e vazhdueshme duke i ndarë në dy pjesë ose klasa. Për ta bërë proçesin më efektiv, baza e të dhënave është atributi i cili duhet vlerësuar dhe vendosur në një renditje të caktuar. Pas kësaj duhet të përdorim statistikat e mjaftueshme për të krijuar grupet dhe të zgjedhim kriterin për të llogaritur çdo shpërndarje të pikave. Kjo tregon se në tërësi procesi kërkon kapërcimin linear duke shumëzuar me një konstante çdo vlerë. Shumica e algoritmeve të ndertimit të pemës klasifikuese të propozuara në literaturat e ndryshme janë ato të cilat i quajmë kërkime dobësuese. Loh dhe Shih kanë propozuar Analizën e katroreve te Diskriminantit (QDA) për të gjetur shpërndarjen e pikave për atributet e vazhdueshme duke e treguar këtë nga pamja e një pike dhe me një saktësi të caktuar për ndërtimin e pemës. Edhe kjo mënyrë është po aq e mirë sa mënyra e kërkimit dobësues.

Këto dy mënyra sygjerojnë se për këtë situatë një zgjidhje: grupo klasat e etiketuara në dy superklasa duke u bazuar në disa ngjashmëri të këtyre klasave dhe përkufizo QDA dhe shpërndaje këto bashkësi në këto superklasa. Kjo metodë mund të përdoret në shpërndarjen e të dhënave kur kemi të bëjmë me elementë kategorikë të të dhënave dhe numri i klasave është më shumë se dy. Idea e përafrimit së të shpërndarjes së të dhënave-pikës me të njëjtën klasë të etiketuar dhe me një shpërndarje normale. Për këtë le të marrim si pikë shpërndarje një pikë midis qëndrës së dy shpërndarjeve dhe me të njëjtin probabilitet për të qenë në çdo klasë. Për atributet e vazhdueshme X, dhe për parametrat e dy shpërndarjeve normale, probabiliteti që ti përkasin një shpërndarje α_i ka mesataren μ_i dhe variancën σ_i^2

$$\alpha_i = P[C = c_i \mid T]$$
$$\mu_i = E[X \mid C = c_i, T]$$
$$\sigma_i^2 = E[X^2 \mid C = c_i, T]$$

Ekuacioni i pikës së shpërndarjes μ është:

$$\alpha_1 \frac{1}{\sigma_1\sqrt{2\pi}} e^{\frac{-(\mu-\mu_1)^2}{2\sigma_1^2}} = \alpha_2 \frac{1}{\sigma_2\sqrt{2\pi}} e^{\frac{-(\mu-\mu_2)^2}{2\sigma_2^2}}$$

Ky ekuacion mund të reduktohet në ekuacionin e fuqisë së dytë si më poshtë për shpërndarjen e pikës:

$$\mu^2\left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2}\right) - 2\mu\left(\frac{\mu_1}{\sigma_1^2} - \frac{\mu_2}{\sigma_2^2}\right) + \frac{\mu_1}{\sigma_1^2} - \frac{\mu_2}{\sigma_2^2} = 2\ln\frac{\alpha_1}{\alpha_2} - \ln\frac{\sigma_1^2}{\sigma_2^2}$$

Nëse σ_1^2 është shumë afër vlerës së σ_2^2, zgjidhja e ekuacionit të fuqisë së dytë nuk është numerikisht stabël dhe në këtë rast, mënyra më preferuar është të zgjidhet ekuacioni linear:

$2\mu(\mu_1-\mu_2)=\mu_1^2-\mu_2^2-2\sigma_1^2\ln\frac{\alpha_1}{\alpha_2}$ që numerikisht është i zgjidhshëm përderisa $\mu_1\neq\mu_2$. Për të njehsuar fitimin Gini për ndryshoren X me pikë shpërndarje μ është e nevojshme të njehsojmë statistikat e mjaftueshme:

$$P[C=c_i \mid X\le\mu,T], P[C=c_i \mid X\le\mu,T], and P[X\le\mu \mid T]=$$
$$=P[C=c_0 \mid T]P[C=c_0 \mid X\le\mu,T]+P[C=c_1 \mid T]P[C=c_1 \mid X\le\mu,T]$$

Duke zëvëndësuar në (3), atëhere probabiliteti $P[x\in C_1 \mid x\le\mu,T]$ nuk është gjë tjetër veçse një përmbledhje e funksionit të shpërndarjes (c.d.f), si një shpërndarje normale me një mesatare μ_i dhe variancë σ_i^2:

$$P[C=c_0 \mid X\le\mu,T]=\int_{x\le\mu}\frac{1}{\sigma_1\sqrt{2\pi}}e^{-(x-\mu)^2/2\sigma_1^2dx}=\frac{1}{2}\left(1+E_{rf}\left(\frac{\mu_1-\mu}{\sigma_1\sqrt{2}}\right)\right)$$

$P[C=c_1 \mid X\le\mu]$ fitohet në mënyrë të ngjashme. Një nga avantazhet e kësaj metode është se nuk kërkohet klasifikimi apo ndarja e një bazë të dhënash njehsimi i statistikave të mjaftueshme mund të bëhet lehtësisht dhe gjetja e pikës së shpërndarjes.

Proçesi i rritjes së vazhdueshme të pemës ka dhe një proçes përfundimtar. Kriteri themelor i mosrritjes së mëtejshme të pemës klasifikuese është kur të ashtuquajturat pika të shpërndarjes janë të përshkruara në minimumin e vetë. Duke e ndaluar proçesin kur një sasi e vogël e një bazë të dhënash është në dispoizicion, shmangim marrjen e një vendimi statistikor i cili është i parëndësishëm dhe është i zhurmshëm dhe i gabuar. Mundësi të tjera të mbarimit të proçesit të shpërndarjes janë kur nuk gjejmë më atribute parashikuese dhe kjo zakonisht arrihet kur pema është rritur në maksimumin e saj.

Proçesi i ndertimit të pemës klasifikuese është një proçes që mund të reduktohet në një proçes të njehsimit të statistikave të mjaftushme në çdo nyje të pemës. Ideja kryesore për të bërë këto llogaritje është ajo e vlerësimeve empirike.

1. Për probabilitetet e formës P[p(Xj)|T] me disa prashikues p(Xj) të ndryshoreve Xj dhe vlerësimi është i thjeshtë për një numër të caktuar të pikave të bazës së të dhënave në një nyje të caktuar T të bazës së të dhënave Dt, për të cilat prashikuesi p(Xj) përmban disa pika të bazës së të dhënave në tërësi në Dt.

$$P[p(X_j)\mid T]\overset{e}{=}\frac{|\{(x,c)\in D_T \mid X_j=x_j\}|}{|D_T|}$$

2. Për probabilitetin me kusht të formës $P[p(X_{j)} \mid C=c_0,T]$, është përllogaritur si më poshtë:

$$P[p(X_j)\mid C=c_0,T]\overset{e}{=}\frac{|\{x,c_0)\in D_T \mid X=x_j\}|}{|\{x,c_0)\in D_T\}|}$$

Për funksionin e pritshëm të atributeve si $E[f(X_j)\mid T]$, vlerësimi i përafërt është i thjeshtë si një përllogaritje mesatare e vlerave të funksionit i cili zbatohet në vlerat e atributeve për pikat e të dhenave në D_T: $E[f(X_j)\mid T]\overset{e}{=}\frac{\sum_{(x,c)\in D_T}f(x_j)}{|D_T|}$, ku f(x) është funksioni vlerat e të cilit pritet të përafrohen.

3. Për $E[f(X_j) \mid C = c_0, T]$ vlerësimi i përafërt është:

$$E[f(X_j) \mid C = c_0, T] \overset{e}{=} \frac{\sum_{(x,C_0)\in D_T} f(x_j)}{|\{(x, c_0) \in D_T\}|}$$

2.9 Natyra Hierarkike e pemës klasifikuese

Breiman (1984) ka paraqitur shëmbuj të përdorimit të pemëve te klasifikimit. Një shembull tipik është, kur pacientët me probleme në zemër janë pranuar në spital, pas dhjetra testeve që janë kryer shpesh për të marrë informacion lidhur me probleme psikologjike si dhe matje te numrit të rrahjeve të zemrës, sa është tensioni i gjakut, dhe kështu me radhë. Informacione të tjera, meren nga mosha e pacientit dhe historia mjekësore e trashigimisë familjare. Pacientët më pas janë ndjekur në vazhdimësi për të parë nëse ata mbijetojnë nga ataku në zemër, për të paktën 30 ditë. A do të jetë i dobishëm trajtimi që u bëhet pacientëve për ti mbrojtur nga ataku në zemër, dhe në perparimin e teorisë mjekësore për rastet kur kemi një mos përcaktim të saktë, ose e thënë ndryshe nuk ka mbijetesë nga zemra, në qoftë se matjet e marra menjëherë pas pranimit në spital mund të përdore për të identifikuar pacientët me rrezik të lartë (ata të cilët nuk kanë gjasa për të mbijetuar të paktën 30 ditë). Një pemë klasifikimi që Breiman (1984) e ka zhvilluar për të adresuar këtë problem është i thjeshtë. Tre pyetje duhet të bëhen deri sa të arrijmë te pema vendimëmarrëse.

Pema vendimtare e zhvilluar nga Breiman (1984) paraqitet në Figuren 8:

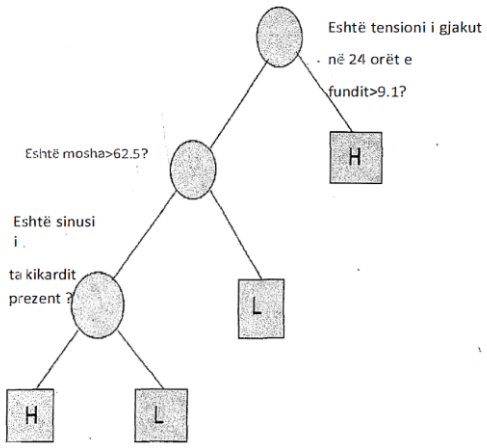


***Figura 8**: Struktura e një peme vendimmarrëse*

Ku P (presioni i gjakut), A (mosha), dhe T (nëse sinusi i takikardit eshte i pranishem ose jo,(me vlerat 1ose 0) do të kishin këto vlera P= 9.1, M= 62,5, dhe 0, p,a dhe t jane kofiçentet linear te funksionit te diskriminatit dhe respektivisht, "Nëse p + P është më pak se ose e barabartë me zero, pacienti është me rrezik të ulët, ndryshe në qoftë se një + a është më pak se ose e barabartë me zero, pacienti është rrezik me të ulët, në qoftë se t + T është më pak se ose e barabartë me zero, pacienti është me rrezik të ulët, ndryshe pacienti është me rrezik të lartë. " Sipërfaqësisht, analiza dalluese dhe proçeset e pemës klasifikuese vendimtare mund të duken të ngjashme, për shkak se të dyja përfshijnë koefiçientet dhe ekuacionin vendimmarrës. Por ndryshimi i vendimeve të njëkohshme të analizes se diskriminantit nga vendimet hierarkike te pemëve të klasifikimit duhet theksuar se nuk mund të quhen të mjaftueshme.

Dallimi ndërmjet këtyre dy qasjeve ndoshta mund të bëhet më i qartë duke marrë parasysh se si çdo analizë do të kryhet me anë të regresit. Për shkak të rrezikut që egziston në shembullin e Breiman (1984) atje është një variabël i varur i ekspozuar, parashikimet në analizën diskriminuese mund të riprodhohen nga një regres i shumëfishtë i rrezikut në tre variablat parashikues për të gjithë pacientët. Pemët parashikuese të klasifikimit mund të riprodhohen vetëm me anë të një analize te veçantë dhe të thjeshtë të regresit, ku rreziku së pari është i varur nga P për të gjithë pacientët, atëherë rreziku është i varur në një variabel për pacientët te cilet nuk klasifikohen si me rrezik të ulët në regresin e pare. Kjo ilustron qartë natyrën e njëkohshme të vendimeve analizuese dhe dalluese në krahasim me natyrën rekursive, hierarkike të pemëve përfundimtare kualifikuese. Një nga karakteristikat e pemëve të klasifikimit është që ka shume nderlikime.

Pema e regresit është ndertuar duke përdorur të njëtin algoritëm, atë të ndarjes së vazhdueshme nga bashkësia e madhe në nënbashkësi të vogla. Ky algoritëm i cili ka n impute të tilla si $D_i = \{\langle x_i, y_i \rangle\}_{i=1}^{n}$, dhe nëse kriteret e caktuara për të përfunduar këtë proçes nuk arrihen atëhere duhet të bëhet testi i nyjes t, në të cilën dy degët janë marrë duke aplikuar të njëjtin algoritmin me dy nënbashkësitë e imputeve të kësaj baze të dhënash. Të gjitha teknikat e regresit përmbajne një output përgjegjës të vetme dhe një ose më shumë të dhëna ose variabla parashikues. Variabli përgjegjës i regresit është numerikë. Metodologjia e përgjithshme e ndërtimit të pemës i lejon variablat hyrëse të jenë një përzierje e variablave të vazhdueshme dhe kategorike. Një pemë perfundimtare është prodhuar kur çdo nyje fundore në pemë përmban një test mbi vlerën e ndonjë variabli input. Nyjet fundore të pemës përmbajnë vlerat dalëse parashikuse të cilat janë të ndryshueshme. Një pemë e regresit mund të konsiderohet si një variant i pemëve vendimtare, e projektuar për të përafruar funksionet e vlerave reale, në vend që të përdorim metodat e klasifikimit. Një pemë regresit është e ndërtuar nëpërmjet një proçesi të njohur si ndarje binare kudo rekursive, i cili është një proçes përsëritës që ndan të dhënat në ndarëse ose degët, dhe pastaj vazhdon ndarjen, çdo ndarje në grupe të vogla si metoda që lëviz lart çdo degë. Fillimisht, të gjitha të dhënat në bashkësinë e trajnimit janë grupuar në të njëjtën ndarje. Duke përdorur algoritmet fillojmë shpërndarjen e të dhënave në dy ndarëse ose degët, duke përdorur çdo ndarje të mundshme binare në çdo fushë. Algoritmi zgjedh ndarjen që minimizon shumën e devijimeve nga mesatarja në katror në dy ndarëse të veçanta. Ky rregull zbaton ndarjen për secilën prej degëve të reja.

Ky proçes vazhdon derisa çdo nyje të arrijë një madhësi minimale dhe të bëhet një nyje fundore. (Në qoftë se shuma e devijimeve ne katrore në një nyje është zero, atëherë kjo nyje është konsideruar si një nyje fundore edhe nëse ajo nuk ka arritur madhësinë minimale.

Për ndërtimin e pemës se regresit përdorim dy algoritme: atë të minimizimit të shumës së katrorëve të distancave dhe atë të minimizimit të vlerave absolute të devijimit, i cili është përdorur edhe nga autori i të parit libër Breman. Më poshtë paraqesim një përshkrim për të dyja këto metoda. Së pari atë të minimizimit të shumës së distancave ne katrorëve të dhe pas kësaj atë të minimizimit të shumës së vlerave absolute të devijimeve.

Kjo pemë është paraqitur për here të parë nga Breiman me 1984 dhe është zbatuar si pjesë e CART. Pema e regresit është gjithashtu një pemë binare, e cila ka një vlerë numerike konstante në çdo nyje dhe përdor variancën për të matur papastërtinë. Kështu që kriteri i shpërndarjes matet : $E_{rr}(T) = \sum_{i=1}^{N_T} (y_i - \bar{y}_i)^2$,

$$\Delta E_{rr}(T) = E_{rr}(T) - E_{rr}(T_1) - E_{rr}(T_2)$$

Arsyeja për të përdorur variancën si masë të papastërtisë lidhet me faktin se parashikuesi më i mirë në një nyje është mesatarja e vlerave të variablave parashikuese në çdo test që duhet të

bëjmë në çdo nyje. Një alternativë për kriterin e shpërndarjes e propozuar nga Breiman është bazuar në variancën e zgjedhjes si masë e papastërtisë.

$$Err_s(T) = Var(Y \mid T) \overset{e}{=} \frac{1}{N_T} Err(T)$$

$$\Delta Err_s(T) = Err_s(T) - P[T_1 \mid T] \bullet Err_s(T_1) - P[T_2 \mid T] \bullet Err_s(T_2)$$

Nëse pergjasia maksimale është përdorur për të gjitha propabilitetet dhe pritshmëritë të cilat janë parë praktikisht. Pas kësaj ne kemi këtë lidhje midis variancës se popullimit dhe variancës së zgjedhjes: $\Delta Err_s(T) = \overset{e}{=} \frac{Err(T)}{N_T} - \frac{N_{T_1} Err(T_1)}{N_T N_{T_1}} - \frac{N(T_2) Err(T_2)}{N_T N(T_2)} = \frac{\Delta Err(T)}{N_T}$, në varësi të kësaj lidhjeje dhe nëse në një bazë të dhënash nuk mungon ndonjë element, përdorimi i kriterit të minimizimit të rezultateve në një pikë do të çonte detyrimisht në minimizimin e të tjerave.

Për një atribut kategorik X, minimizimi i $Err_s(T)$ mund të bëhet në mënyrë shumë efiçente duke përdorur kushtet e mësipërme:

$$\Phi(x) = -x^2$$
$$q_i = P[X = x_i \mid T]$$
$$r_i = P[Y \mid X = x_i, T]n$$

Kjo ka kuptimin që në mënyrë të thjeshtë mund ti vendosim elemetet e bazës së të dhënave në rendin rritës. $P[Y \mid X = x_i, T]$ duke realizuar shpërndarjen sipas renditjes. Përafrimi empirik që përdoret për

$$q_i = P[X = x_i \mid T]$$
dhe
$$r_i = P[Y \mid X = x_i, T]$$

është kriteri që $Err_s(T)$ merr vlera maksimale.

Në rastin e pemës klasifikuese, parashikimi është bërë me një mënyrë të caktuar drejtimi të pemës për secilën degë deri sa të arrijmë në nyjet përfundimtare të ashtuquajturat gjethe. Kuptohet se vlerat të cilat shoqërojnë gjethet janë vlerat e modelit parashikues që duam. Krasitja është një mjet që ndihmon të përmisojmë saktësinë e pemës klasifikuese. Metodat e krasitjes do të shikohen në mënyrë të detajuar më poshtë. Këto metoda janë të njëjta dhe për pemën e regresit.

Përkufizim 2.4: Mediana e shpërndarjes së një variabli të rastësishëm Y për të gjitha vlerat e popullimit është një vlerë k e tillë që gjysma e vlerave të këtij popullimi Y është më e vogël se k dhe gjysma e vlerave të Y është më e madhe se k, atëhere kjo vlerë k kënaq këtë ekuacion $\int_{-\infty}^{k} p(y)dy = \frac{1}{2}$, ku p(y) është funksioni i densitetit.

Teoremë 2.3: Një konstante k që minimizon vlerat e pritura të gabimit mesatar të katrorëve të distancave është vlera mesatare e variablit përgjegjës.

$k_l = \frac{1}{n_l} \sum_{D_{li}} y_i$, ku n_l është kardinali i bashkësisë D_l e cila përmban rastin me nyje fundore l dhe $n_l = me, numerin, e, D_l$.

Vërtetim: Në se Y është variabël rasti i vazhdueshëm me densitet të probabilitetit funksionin f(y), atëherë funksioni që duhet të minimizojmë në lidhje me k është: $\phi(k) = E[(Y-k)^2] = \int_{-\infty}^{\infty}(y-k)^2 f(y)dy$, ku $E[Y] = \int_{-\infty}^{\infty} yf(y)dy =$

$\int_{-\infty}^{\infty}(y^2 - 2yk + k^2) f(y)dy =$

$\int_{-\infty}^{\infty} y^2 f(y)dy - 2k\int_{-\infty}^{\infty} yf(y0dy + k^2, ku, \int_{-\infty}^{\infty} f(y)dy = 1$

Minimizimi në lidhje me k:

$\frac{\partial}{\partial k}\phi(k) = 0 \Leftrightarrow 0 - 2\int_{-\infty}^{\infty} yf(y)dy + 2k = 0 \Leftrightarrow k = \int_{-\infty}^{\infty} yf(y)dy$, pra k=E(y).

Breiman dhe bashkautorët kanë theksuar se përdorimi i kriterit të minimizimit gabimit të shumës së vlerave absolute të devijimeve mund të na japë shpërndarjen më të mirë për pemën e regresit. Kjo metodë përdor kriterin e selektimit të minimumit të shumës së vlerave absolute të devijimit midis modelit parashikues dhe vlerave të Y. Përdorimi i këtij kriteri çon në pemë të cilat janë më të qëndrueshme ndaj vlerave të huaja (outliers). Në ndryshim nga minimizimi i shumës së katrorëve të distancave i cili mund të na shkaktojë dhe ndonjë gabim në rastet kur kemi vlere jo normale, pasi prezenca e tyre natyrisht që ndikon fuqimisht në vlerën mesatare. Ndërtimi i pemës duke përdorur këtë metodë është i bazuar në rastin kur kemi një bazë të dhënash më element $\{\langle x_i, y_i \rangle\}_{j=1}^{n}$ i cili e minimizon vlerën absolute të devijimit mesatar $\frac{1}{n}\sum_{i=1}^{n}| y_i - r(\beta, x_i)|, ku, r(\beta, x_i)$ është modeli parashikues i modelit $r(\beta, x)$ për rastin $\langle x_i, y_i \rangle$. Konstantja k e cila minimizon mesataren absolute të vlerësuar të devijimeve të vrojtuara në lidhje me k, është mesorja e vlerave të Y. Minimizimi i diferencës së mesatares së devijimeve me këtë konstante korrespondon me minimumin e pritshmërisë statistikore të $| y_i - k |$

Teoremë 2.4: Konstantja k e cila minimizon vlerën e pritshme të devijimeve absolute me një variabël të vazhdueshëm dhe të rastit Y, me densitet të probabilitetit f(y), është mediana e variablit Y.

Vertetim

Funksioni që duam te minimizojmë në lidhje me k është:

$\phi(k) = E(| y-k |) = \int_{-\infty}^{\infty}| y-k | f(y)dy = \int_{-\infty}^{k}| k-y | f(y)dy + \int_{k}^{\infty}| y-k | f(y)dy =$

$k\int_{-\infty}^{k} f(y)dy - \int_{-\infty}^{k} yf(y)dy + \int_{k}^{\infty} yf(y)dy - k\int_{k}^{\infty} f(y)dy$ duke zëvëndësuar

$\int_{k}^{\infty} f(y0dy - 1 - \int_{-\infty}^{k} f(y)dy$ marrim:

$\phi(k) = k\int_{-\infty}^{k} f(y)dy - k + k\int_{-\infty}^{k} f(y)dy - \int_{-\infty}^{k} yf(y)dy + \int_{k}^{\infty} yf9y)dy =$

$$2k\int_{-\infty}^{k} f(y)dy - k - \int_{-\infty}^{k} yf(y)dy + \int_{k}^{\infty} yf(y)dy =$$

$2kF(k) - k - \int_{-\infty}^{k} yf(y)dy + \int_{k}^{\infty} yf(y)dy$, ku F(y) është funksioni progresiv i shpërndarjes së variablit Y.

Nëse marrim derivatin e pjesëshëm të këtij funksioni në lidhje me k dhe duke e barazuar me zero gjejme:

$$\frac{\partial}{\partial k}\phi(k) = 2F(k) + 2kf(k) - 1 - kf(k) - kf(k) = 2F(k) - 1$$

Kështu që : $\frac{\partial}{\partial k}\phi(k) = 0 \Leftrightarrow F(k) = \frac{1}{2}$ dhe sipas përkufizimit të funksionit progresiv të devijimit ky funksion duhet të jetëi barabartë me ½ për çdo medianë të shpërndarjes.

2.10 Reduktimi i papastërtisë si masë e mirësisë së shpërndarjes

Në softwarë të ndryshme mund të zgjedhim të maksimizojmë reduktimin e papastërtisë si një alternative e cila maksimizon shkallën e shpërndarjes në proçesin e selektimit dhe të shpërndarjes së imputeve dhe në zgjedhjen e imputeve më të mira. Papastërtia e një nyje është shkalla e heterogjenitetit duke respektuar kompozimin e niveleve për variablat të cilat janë si objektivi ynë. Nëse një nyje t e cila shpërndahet në dy degë në të majtë dhe në të djathtë përkatësisht në t_L dhe t_R të tilla që t_R janë porpocionale me P_R dhe janë porpocionale me P_L "Mirësia e shpërndarje"(Goodness of split) është e përkufizuar si zvogëlim i papastërtisë dhe matematikisht është si më poshtë:

$$\Delta i(s,t) = i(t) - P_L i(t_L) - P_R i(t_R).$$

ku i(t) është indeksi i papstërtisë për nyjen t dhe dy pjesët e tjera të formulës $P_L i(t_L), dhe, P_R i(t_R).$ janë përkatësisht indeksi i papastërtisë së nyjes së majtë dhe të djathtë (të marra nga Entropy). Shpërndarja e nyjes t në dy nyjet e tjera e bazuar në shpërndarjen e imputit X1, algoritmi i pemës egzaminon të gjithë kandidatët të cilët duhet të shpërndahen dhe që kanë formën X1<Xj dhe $X_1 \geq X_j$ ku Xj janë numra realë midis vlerave minimale dhe maksimale të X1. Ato vlera të cilat janë më të vogla kalojnë në të majtë dhe të tjerat kalojnë në të djathtë. Për shembull per të shpërndarë 200 kandidatë në input-in X1, atëherë kandidatët të cilët duhet të shpërndahen kanë vlerat Xj = 1,2,3….200. Algoritmi krahason reduktimin e papastërtisë për këto 200 shpërndarje dhe selekton ato të cilat arrijnë reduktimin më të mirë të papastërtisë e cila kosiderohet dhe si shpërndarja më e mirë. Papastërtia apo pastërtia si mase përdoret në ndërtimin e pemës vendimtare në CART është Gini Index. Pema vendimtare që ndërtohet në CART bëhet gjithmone duke përdorur algoritmin i cili përdor pemën binare, ku çdo nyje ka dy nyje pasardhëse.

Masa Gini është masë e papastërtisë së një nyje dhe është më e përdorur veçanërisht kur variabli i varur është variabël kategorik dhe është i përkufizuar si më poshtë:

$$g(t) = \sum_{j \neq i} p(j/t)\, p(i/t)$$

Nëse kosto e mosklasifikimit nuk është përcaktuar atëherë kemi:

$$\sum_{j \neq i} C(i/j)\, p(j/i)\, p(i/t)$$

Nëse kosto e joklasifikimit është përcaktuar, ku shuma i kalon të gjitha kategoritë k të $p(j / t)$ i cili është probabiliteti i një kategorie j në nyjen t dhe $C(i/ j)$ është probabiliteti i mosklasifikimit të kategorisë j në raport me një kategori tjetër i.

Një pemë mund të përcaktohet në mënyrë abstrakte, si një e tërë si një pemë me një renditje të caktuar, me një vlerë të caktuar për çdo nyje. Të dyja këto perspektiva janë të dobishme: ndërsa një pemë mund të analizohet matematikisht si një e tërë, kur në fakt ajo është përfaqësuar si një strukturë e të dhënave ku ajo është e përfaqësuar dhe ka punuar më vete për çdo nyje. Për shembull, duke e parë pemën si një të tërë, mund të flasim për "nyjen mëmë" të një nyje të caktuar, por në përgjithësi si një strukturë e të dhënave një nyje dhënë përmban vetëm listën e fëmijëve të saj.

Le të konsiderojmë një kompani A, si një pemë që ka shumë furnizues të cilët përbëjnë nyjet e çdo furnizuesi, apo shërbime të ndryshme. Vlerësimi i pastërtise së nyjeve të njohur si Gini, mat shkallën e pastërtisë për një rajon që përmban pika të të dhënave nga klasa ndoshta të ndryshme. Ideja kryesore është se nuk ka "fëmijë" qe të bëjnë punë të përsëritura, për këtë arsye do të përcaktojmë papastërtinë e nyjes. Masa Gini do të ndihmojë kompaninë A për të vendosur se sa nyjet do të mbahen si të papastërta ose sa furnizues do të ofrojnë shërbime , produkte të ngjashme ose që kanë punë të përsëritura, të cilat mund të reduktohen.

Një nyje e pastër ka devijim 0; ndryshe devijimi është pozitiv. Një nyje me vetëm një klasë (një nyje e pastër) ka indeks Gini 0; ndryshe indeksi Gini është pozitiv. Nëse do të zbatohen në praktikë, një nyje të pastër nuk do të ketë ndonjë punë ose shërbim të përsëritur dhe secila nyje do të jetë krejtësisht e ndryshme nga të tjerat.

Problemi eshte sa e realizueshme është kjo metodë e tillë në nje mjedis në mesin e furnizuesve. Një opsion është se nuk ka nje qendër komanduese që ruan të gjitha rolet brenda pemës dhe për fëmijët e saj. Një tjetër funksion do të jetë një lloj funksioni inxhinierik i cili ka cilësite ku për secilin do të bëhen veprime të ndara dhe të pavarura nga njëra-tjetra. Një mundësi tjetër është që brenda çdo "fëmijë" nuk ka punë të përsëritura.

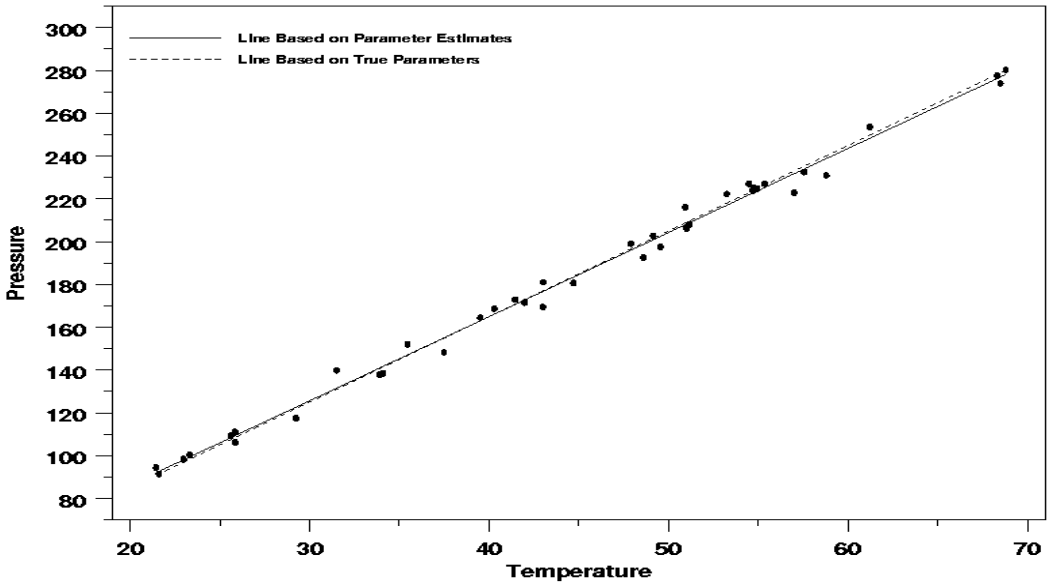


Figura 9: *Grafiku real dhe i përafruar i të dhënave*

2.11 Funksioni i papastërtisë

Funksioni i papastërtisë mat shkallën e pastërtisë për një rajon që përmban pika të të dhënave nga baza e të dhënave, e cila është e mundshme që këto klasa ndoshta të jenë të ndryshme. Supozojmë se numri i klasave është K. Atëhere funksioni papastërtise është një funksion i $p_1, p_2, ... p_k$, ku probabiliteti për çdo pikë të të dhënave në rajon i përket klasës 1, 2, ..., K. Gjatë këtij procesi, nuk i dimë probabilitetet e vërteta. Ajo që do të përdorim është përqindja e pikave në klasë 1, klasa 2, klasën 3, dhe kështu me radhë, kjo sipas të dhënave që kemi në këtë bazë të dhënash.

Funksion i papastërtisë do të quhet funksioni Φ i përkufizuar në një bashkësi ku të gjithë elementet janë vendosur në një renditje të caktuar $p_1, p_2, ... p_k$ duke kënaqur kushtin $p_j \geq 0$, ku j=1,2,3,…K dhe $\sum_j p_j = 1$.

Funksioni i papastërtisë mund të përcaktohet në mënyra të ndryshme, por ai duhet të gezojë tre vetite e meposhtme:

Φ arrin maksimumin vetëm atëhere kur kemi shpërndarje uniforme, domethënë. të gjitha p_j janë të barabarta.

Φ arrin minimumin vetëm te pikat (1,0,0….0),(0,1,0,0….),(0,0,1,0,0…),……., (0,0,0,0,…..1), kur propabiliteti i të qënurit në klasë të çfardoshme është 1 dhe 0 për klasat e tjera.

d. Φ është funksion simetrik per $p_1, p_2, ... p_k$, edhe nëse përkëmbejmë p_j, Φ qëndron konstant.

Përkufizim 2.5: Nëse njihet funksioni i papastërtisë Φ, masën e papastërtisë të në një nyje të caktuar t është: $i(t0 = \phi(p(1/t), p(2/t), ...p(k/t))$ ku p(j/t) është një vlerësues i perafërt i probabilitetit të pasëm të klasës j për një pikë të dhënë në nyjen t.

Ky do të quhet funksion i papastërtisë i matur në nyjen t. Kur kemi i(t) do të përkufizojmë shpërndarjen e mirë(goodness of split) të një nyje të dhënë nga funksioni $\Phi(s/t)$:

$\Phi(s/t) = \Delta i(s,t) = i(t) - p_R i(t_R) - p_L i(t_L)$ ku $\Delta i(s,t)$ është diferenca e masës së papastërtisë në nyjen t dhe shumës së papastërtisë së nyjes së majtë dhe të djathtë. P janë propabilitetet ku p_R, p_L janë të shpërndara në mënyrë porpocinale në nyjen e djathtë t_R dhe në nyjen e majtë t_L, të shikojmë grafikun e mëposhtëm.

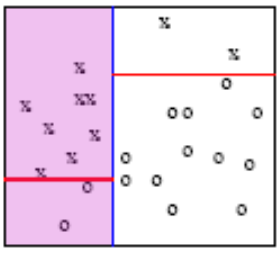


Figura 10: *Ndarja e bazës së të dhënave në grupe*

Supozojmë se zona në të majtë me ngjyrë lejla është nyja që është shpërndarë, pjesa e sipërme është nyja pasardhëse që del në të majtë dhe pjesa e poshtme është nyja pasardhëse

në krah të djathtë dhe qartësisht shihet se proporcionaliteti i pikave të dërguara në nyjen e majtë është $p_L = 8/10, dhe, p_R = 2/10$.

Algoritmi i klasifikimit përçon të gjithë kandidatët duke selektuar më të mirin në të cilin $\Delta i(s,t)$ është maksimizuar.

Le të përkufizojmë I(t)=i(t)p(t), që është funksioni i papastërtisë i nyjes t pesha e të cilës është vlerësuar të jetë në porporcion i të dhënave që shkon në nyjen t me probabilitetin që ndodhet në zonën e nyjes t. Një mënyrë thjesht për të bërë këtë vlerësim është që të numërojmëtë gjitha pikat që janë në nyjen t dhe ta pjestojmë me numrin total të pikave të gjithë datës. Masa agregate e funksionit të papastërtisë për një pemë T, të cilën e shënojmë I(T) është:

$I(T) = \sum_{t \in \tilde{T}} I(t) = \sum_{t \in \tilde{T}} i(t)\, p(t)$, kjo është një shumë e të gjitha gjetheve(ose nyjeve fundore) të çdo nyje. Për një nyje të çfardoshme kemi që:

$$p(t_L) + p(t_R) = p(t)$$
$$p_L = p(t_L) / p(t)$$
$$p_R = p(t_R) / p(t)$$
$$p_R + p_L = 1$$

Zona e mbuluar nga nyja pasardhëse e majtë t_L, dhe nga nyja pasardhëse e djathtë t_R janë të papajtueshme dhe në se bëjmë kombinimin e zonave nga më të mëdhatë të prindërve të çdo nyje, atëhere shuma e probabiliteteve të bashkësive të papajtueshme është e barabartë me bashkimin e dy bashkësive, atëhere p_L bëhet raporti relativ midis nyjes se majtë fëmijë duke respektuar nyjen prind. Le të përkufizojmë diferencën e peshës së masës së papastërtisë së nyjes prind me dy nënnyjet fëmijë:

$$\begin{aligned}\Delta I(s,t) &= I(t) - I(t_L) - I(t_R) \\ &= p(t)i(t) - p(t_L)i(t_L) - p(t_R)i(t_R) \\ &= p(t)i(t) - p_L i(t_L) - p_R i(t_R) \\ &= p(t)\Delta(s,t)\end{aligned}$$

2.12 Funksionet e papastërtisë

1. Funksioni i entropisë $\sum_{j=1}^{K} p_j \log \frac{1}{p_j}, nese, p_j = 0$, duke përdorur limitin $\lim p_j \to \log p_j = 0$.
2. Mosklasifikimi : $1 - \max_j p_j$.
3. Indeksi Gini $\sum_{j=1}^{K} p_j (1 - p_j) = 1 - \sum_{j=1}^{K} p_j^2$.

2.13 Devijimi i katrorëve më të vegjël

Devijimi i katrorëve më të vegjël të distancave (LSD) është përdorur si masë e papastërtisë së një nyje kur variabli përgjegjës është i vazhdueshëm, dhe është llogaritur si:

$$R(t) = \frac{1}{N_w(t)} \sum w_i f_i (y_i - \bar{y}(t))^2$$

Ku $N_w(t)$ është numri i peshës në çdo rast në një nyje të caktuar t, w_i është vlera e peshës së një variabli në një rast *i*, f_i është vlera e një variabli me denduri të ndryshme, y_i është vlera e variablit përgjegjës, dhe $y(t)$ është pesha mesatare për nyjen *t*.

$$i(p) = \sum_{i \neq j} p_i p_j = 1 - \sum_j p_j^2$$

Në këtë rast kemi zgjedhur ndarjen që të shumtën ul Indeksin Gini (domethënë rrit pastërtinë). Pas këtij proçesi përsëritje i cili gjeneron ndarje të reja nga ndarjet e vjetra që tashmë kemi. Në këtë mënyrë për të bërë këtë kemi nevojë për të përsëritur të njëjtat hapa kur kemi ndarë nyjen e parë. Pra, kemi nevojë për shpërndarje për çdo proçes të ri të ndarjes. Kjo është shumë e vështirë për të bërë me dorë. Kjo është shumë më e lehtë për tu realizuar me R.

Proçesi i ndarjes merr vetëm një ndryshore në një kohë dhe rezultati i kësaj është ndarja e dy variablave dhe kështu me radhë. Dhe kështu do të shikojmë se për (pacientët) shëmbull në pemën fillestare ose Tmax do të jetë zakonisht e vështirë për të lexuar pasi do të jetë e mbingarkuar nga të dhënat. Zgjidhja është që duhet të krasitim pemën fillestare për të marrë një pemë të re që ka një numër më të vogël dhe që është më e lehtë për tu lexuar, dhe më e rëndësishmja i prezanton të dhënat shumë më mirë. Papastërtia Gini është një masë që shpesh zgjedh rastësisht një element nga grupi që do të etiketohen gabimisht nëse do të ishte etiketuar rastësisht në përputhje me shpërndarjen e etiketave në këtë bashkësi. Ajo mund të llogaritet duke mbledhur të gjitha probabilitet e çdo nyje të cilat janë zgjedhur dhe shumëzohet me gabimet probabilitare të këtyre nyjeve. Ajo arrin minimumin e saj (zero), kur të gjitha rastet e këtyre nyjeve tentojnë në një nyje të vetme. Për të llogaritur papastërtinë për një bashkësi të caktuar me vlera , {1, 2, ..., m}, dhenëse p_i = një pjesë e nyjeve të etiketuar me vlerë në një grup.

$$I(p) = \sum_{i=1}^{m} p_i (1 - p_i) = \sum_{i=1}^{m} (p_i - p_i^2) = \sum_{i=1}^{m} p_i - \sum_{i=1}^{m} p_i^2 = 1 - \sum_{i=1}^{m} p_i^2$$

Pastërtia Gini e një nyje është: $p(1\text{-}p)$

- **<u>Entropia e një nyje</u>**

Nje nga menyrat më të përdorura për të matur papastërtinë e një nyje është llogaritja e entropisë:

$$-\sum_{i=1}^{m} p_i \log_2 p_i ,$$

ku p_i është probabiliteti i klasës qe llogaritet si një raport proporcional i klasave në këtë bashkësi -[p*log(p) + (1-p)*log(1-p)]

Entropia maksimale/Gini kur p=.5

Entropia minimale /Gini kur p=0 ose 1

Gini mund të prodhojë nyje të pastra. Shpërndarja ndalohet kur përmiresimi i pastërtisë nuk është statistikisht i rëndesishëm. Ndryshimi midis pemës së regresit dhe asaj të klasifikimit është se në pemen e regresit parshikimin e njëhsojmë si një mesatare të vlerave numerike te

objektetit në studim. *Masa e pastertisë matet me rrënjën katerore të shumës së katrorëve të devijimeve nga mesatarja e gjetheve.*

Meqënëse variabli parashikues i modelit të regresit është numerik lehtësisht mund të gjejmë diferencen midis realit dhe parashikuesit. Vlera absolute mesatare e devijimit e mat dhe e klasifikon gabimin në çdo model duke mesatarizuar vlerën absolute të gabimit mesatar të parashikimeve:

$MAD(r) = \frac{1}{n}\sum_{i=1}^{n} |(y_i - r(\beta, x_i)|$ ku $\{\langle x_i, y_i \rangle\}_{i=1}^{n}$ është baza e dhënë, $r(\beta, x_i)$ është parashikusi i modelit të regresit të cilin duam ta vlerësojmë për rastin $\langle x_i\, y_i \rangle$. Dhe në këtë situatë do të shikojmë për modelin i cili jep gabimin më të vogël dhe matësi më i mirë i kësaj është metoda e katrorëve më të vegjël. Një gabim tjetër i përbashkët është dhe gabimi mesatar relativ i katrorëve RMSE, që jepet si më poshtë:

$$RMSE(r) = (\frac{1}{n}\sum_{i=1}^{n}(y_i - r(\beta, x_i))^2 / (\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{\bar{y}})^2) = \frac{MSE(r)}{MSE(\bar{\bar{y}})}$$

ku $\bar{\bar{y}}$ është mesatarja e vlerave të Y. Kjo jep vlerën relative të gabimit.

2.14 Përdorimi i Algoritmeve në shpërndarje

Algoritmet bazë të pemes klasifikuese konsiderohen të jenë një nga metodat më të mira të të mësuarit dhe të përdorura më së shumti. Metodat e bazuara në pemën klasifikuese paraqesin modele parashikuese me saktësi shumë të mirë, stabilitet dhe shumë lehtësi interpretimi. Ato paraqesin lidhjet jo-lineare mjaft mirë dhe janë të përshtatshme në zgjidhjen e çdo problemi të klasifikimit ose të regresit. Pemët vendimtare përdorin algoritme të shumta për të vendosur se kur duhet ndarë një nyje në dy ose më shumë nën-nyje. Krijimi i nën-nyjeve rrit homogjenitetin e nënnyjeve rezultuese. Pra, pastërtia e nyjes rritet në lidhje me variablin e synuar. Pema vendimtare ndan nyjet në të gjitha variablat e disponueshëm dhe pastaj zgjedh ndarjen që rezulton me nënnyjet më homogjene.

Zgjedhja e algoritmeve bazohet gjithashtu në llojin e variablave përgjegjës. *Le të shohim katër algoritmet më të përdorura në pemën e vendimit duke përdorur një shembull si më poshtë:*

Le të marrim një klasë prej 36 studentë me tre variabla Gjinia (Djalë / Vajzë), Klasa (XI / XII) dhe gjatësia (160 cm deri në 180 cm, (160,170) dhe (170,180)), 18 nga këta luajnë basketboll në kohën e lirë. Kërkojmë të krijojmë një model për të parashikuar se kush do të luajë basketboll gjatë kohës së lirë? Në këtë problem, ne duhet të veçojmë studentët që luajnë basketboll në kohën e tyre të lirë bazuar në gjininë, klasën dhe gjatësinë.

Kjo është struktura ku pema vendimtare na ndihmon, të veçojmë studentët në bazë të të gjitha vlerave të tre variablave dhe do të identifikojme variablin, i cili krijon grupet më të mira homogjene të studentëve (që janë heterogjene me njëri-tjetrin). Më poshtë, mund të shikojmë se klasa të ndryshueshme janë në gjendje të identifikojënë grupet më të mira homogjene krahasuar me dy variablat e tjerë.

a. Ndarja sipas gjinisë

Gjinia M/F			
Gjinia	Numri i studentëve	Luajnë basketboll	Përqindja
Femra	16	4	25%
Meshkuj	20	14	70%
Totali	36	18	50%

Tabela 2 : *Ndarja sipas gjinisë*

b. Ndarja sipas gjatesisë

Gjatesia(>170 ose<170)			
Gjatesia	Numri i studentëve	Luajnë basketboll	Përqindja
>170cm	20	12	60%
<170 cm	16	6	37.5%
Totali	36	18	50%

Tabela 3 : *Ndarja sipas gjatesisë*

a. Ndarja sipas klasave

Klasat(XI ose XII)			
Klasat	Numri i studentëve	Luajnë basketboll	Përqindja
XI	16	6	37.5%
XII	20	12	60%
Totali	36	18	50%

Tabela 4 : *Ndarja sipas klasave*

Siç u përmend më lart, pema e vendimmarrjes identifikon variablin më të rëndësishëm dhe cila është vlera që jep grupet më të mira homogjene të popullimit. Si identifikohet ndryshueshmëria dhe ndarja? Për ta bërë këtë, pema e vendimmarrjes përdor algoritme të ndryshme, të cilat ne do të diskutojmë në vijim.

Si te vendosim se kur një pemë duhet të shpërndahet?

Vendimi për të bërë ndarje strategjike ndikon shumë në saktësinë e një peme. Kriteri i vendimit është i ndryshëm për pemët e klasifikimit dhe regresit.

Pemët Vendimtare përdorin algoritme të shumta për të ndarë një nyje në dy ose më shumë nën-nyje. Krijimi i nën-nyjeve rrit homogjenitetin e nën-nyjeve rezultuese. Pastërtia e nyjes rritet në lidhje me variablin përgjegjës. Pema vendimmarrëse ndan nyjet në të gjitha variablat e disponueshëm dhe pastaj zgjedh ndarjen që rezulton në nën-nyjet më homogjene.

Zgjedhja e algoritmeve bazohet gjithashtu në llojin e variablave përgjegjës. Le të shohim katër algoritmet më të përdorura në pemën e vendimit:

Indeksi Gini

Indeksi i Gini thotë, nëse zgjedhim dy madhesi nga një popullim në mënyrë të rastësishme atëherë ata duhet të jenë në të njëjtën klasë dhe probabiliteti për këtë është 1 nëse popullimi është i pastër.

1. Ne rastin e variablave kategorik, objektivi yne mund te jete "Suksesi" ose "Mos suksesi".

2. Kryen vetëm ndarjet Binare

3. Më e lartë vlera e Gini-t, më i lartë homogjeniteti.

4. CART (Klasifikimi dhe Regresi me anë të pemës) përdor metodën Gini për të krijuar ndarje binare.

Hapat për të llogaritur indeksin Gini për një ndarje

1. Si të llogarisim Ginin për një nën-nyje, duke përdorur shumën e formulës per probabilitetin e poshtem për sukses dhe dështim ($Gini = p^2 + (1-p)^2$),ku (p-sukses dhe 1-p-deshtim)

2. Llogarisim Ginin për një ndarje duke përdorur rezultatin e ponderuar Gini të secilës nyjë të kesaj ndarjeje.

Duke u referuar shembullit të përdorur më lart, ku duam të veçojmë nxenësit bazuar në madhesine e synuar (duke luajtur basketboll ose jo). Në tabelën e mësipërme, e ndajmë popullimin duke përdorur dy variablat e dhëna, si Gjinia dhe Klasa. Kërkojëmë të identifikojmë se cila ndarje prodhon nën-nyje më homogjene duke përdorur indeksin Gini.

a. Llogarisim Ginin për shpërndarjen në nyjen gjinia

1. Llogarit, Gini për nën-nyjen Femra = $0.25^2 + 0.75^2 = 0.625$

2. Gini për nën-nyjen Mashkull =$0.70^2 + 0.30^2 = 0.58$

3. Llogarisim Ginin e ponderuar për shpërndarjen

Gjinia = $\frac{20}{36} * 0.58 + \frac{16}{36} * 0.625 = 0.62$

b. Në mënyrë të njëjtë për shpërndarjen në Klasa:

1. Gini për nën-nyjen Klasa XI = $0.375^2 + 0.625^2 = 0.53$.
2. Gini për nën-nyjen Klasa XII =$0.6^2 + 0.4^2 = 0.52$

3. Llogarisim Ginin e ponderuar për shpërndarjen klasa = $\frac{16}{36} * 0.53 + \frac{20}{36} * 0.52 = 0.52$

c. Në mënyrë të njëjtë për shpërndarjen sipas gjatësisë

1. Gini për nën-nyjen lartesia me <170cm = $0.6^2 + 0.4^2 = 0.52$.

2.Gini për nën-nyjen lartesia >170 cm =$0.0.375^2 + 0.625^2 = 0.53$

3. Llogaritim Ginin e ponderuar për shperndarjen klasa = $\frac{16}{36} * 0.52 + \frac{20}{36} * 0.53 = 0.53$

Nga llogaritjet e mesiperme vëme re se rezultati i Ginit për gjinine është më i lartë se i shperndarjes në klasa dhe gjatësisë, prandaj ndarja e nyjeve do të bëhet për gjininë.

Hi-katror χ^2

Është një algoritëm që zbulon rëndësinë statistikore midis dallimeve të një nën-nyje dhe nyjes prind. Ne e matim atë me shumën e katroreve të diferencave të vlerave të vrojtuara me vlerat e pritura duke e pjestuar me vlerat e pritura të variablave të synuara.

1. Punon me variablin kategorik objektiv "Suksesi" ose "Mos suksesi".

2. Mund të kryejë dy ose më shumë ndarje.

3. Më e lartë vlera e Hi-katror është më e lartë është rëndësia statistikore e dallimeve midis nën-nyjeve dhe nyjes prindërore.

4. Hi-katror i secilës nyje llogaritet duke përdorur formulën: Hi-katror $\chi^2(n) = \frac{(O_i - E_i)^2}{E_i}$, ku Oi vlerat e vrojuara dhe Ei vlerat e pritura

6. Gjeneron pemën e quajtur CHAID (Chi-squared Automatic Interaction Detector). Ky lloj testi është një teknikë që përdoret në gjetjen e pemës vendimtare bazuar në përshtatshmërin, ose në rregullimin e rëndesise së testit. CHAID është një teknikë e klasifikimit të pemës jo vetëm që vlerëson bashkëveprimet komplekse midis parashikuesve, por gjithashtu tregon modelimin përfundimtar në një diagramë peme të lehtë për t'u interpretuar. "Trungu" i pemës përfaqëson modelimin përfundimtar të bazës së të dhënave. CHAID pastaj krijon një shtresë të parë të "degëve" duke shfaqur vlerat e variablit te varur parashikues më te fortë. CHAID përcakton automatikisht se si të grupohen vlerat e këtij parashikuesi në numrin e kategorive të menaxhueshme.

Hapat si të llogarisim Hi- katror për shpërndarjen

1. Llogarisim Hi-katror për cdo nyje individuale duke llogaritur devijimin mesatar kuadratik për Suksesi dhe Mos suksesin (luajnë dhe nuk luajnë basketboll).

2. Llogarisim Hi-katror të shperndarjes duke përdorur shumen e Hi-katroreve për sukses apo dështim për sejcilën nyje të ndarjes.

3. Së pari shikojmë dhe llogaritim vlerën për nyjen Femra, konkretisht llogarisim vlerën aktuale për "Luaj Basketboll" dhe "Nuk luajne Basketboll", këtu janë respektivisht 4 dhe 14.

4. Llogarisim vlerën e pritur për "Luaj basketboll" dhe "Nuk luaj Basketboll", këtu do të ishte 4 dhe 14 për të dyja, sepse nyja prind ka probabilitet 50% dhe ne kemi aplikuar të njëjtën probabilitet në numërimin e Femrave (16).

5. Llogarisim devijimet mesatare kuadratike duke përdorur formulën e mësiperme.

6. Llogarisim Hi-katrorin e nyjes për "Luajne basketboll" dhe "Nuk luajne basketboll" duke përdorur formulën e mesipërme. Këtë e shikojmë në tabelën e mëposhtme

7. Ndjekim hapa te njejta hapa për llogaritjen e vlerës Hi-katror për nyjen Meshkuj.

8. Shtojëmë të gjitha vlerat Hi-katror për të llogaritur Hi-katror për gjininë e ndarë.

Nyje	Luajnë Bask	Nuk luajnë basket	Totali	Pritshmeria te luajnë basketboll	Pritshmeria nuk luajnë basketboll	Devijimi Luajnë Basket	Devijimi nuk luajnë basketboll	Hi- Kateror	
								Luajnë basketboll	Nuk luajnë basketboll
Femra	6	12	16	8	8	-2	4	0.5	2
Meshkuj	12	6	20	10	4	2	-4	0.4	1.6
							shuma	0.9	3.6
							Totali	*4.5*	

Tabela 5: *Hi-katror për gjininë*

Shpërndarja sipas klasave:

Kryen hapa të ngjashëm të llogaritjes për ndarje në Klasa dhe do të marrim tabelën e mëposhtme.

Nyje	Luajne Bask	Nuk luajnë basket	Totali	Pritshmeria të luajnë basketboll	Pritshmeria nuk luajnë basketboll	Devijimi Luajnë Basket	Devijimi nuk luajnë basketboll	Hi- Kateror	
								Luajnë basketboll	Nuk luajnë basketboll
XI	6	10	16	8	8	-2	2	0.5	0.5
XII	12	8	20	10	10	2	-2	0.4	0.4
							shuma	0.9	0.9
							Totali	1.8	

Tabela 6: *Hi-katror për ndarjen sipas klasave*

Nga tabela e mësipërme vihet re se vlera e Hi-katror gjithashtu identifikon ndarjen në gjinia është më e rëndësishme krahasuar me ndarjen në klasa.

Entropia

Po te shohim figurën e më poshtme dhe le të mendojmë se cila nyje mund të përshkruhet me lehtësi. Unë jam i sigurt, përgjigjja do te mendohet se është C sepse kërkon më pak informacion pasi të gjitha vlerat janë të ngjashme. Nga ana tjetër, B kërkon më shumë informacion për ta përshkruar atë dhe A kërkon informacionin maksimal. Me fjalë të tjera, mund të themi se C është një nyje e pastër, B është pak e papastër dhe A është më e papastër.

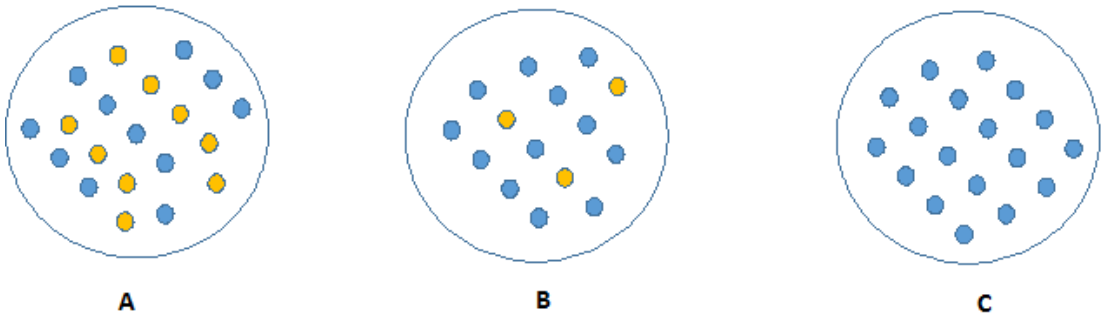


***Figura 11** : Imazhi A,B,C*

Mund të mberrijmë në përfundimin se nyja më pak e papastër kërkon më pak informacion për ta përshkruar atë. Nyja më e papastër kërkon më shumë informacion. Teoria e informacionit është një masë për të përcaktuar këtë shkallë të çorganizimit në një sistem të njohur si Entropy. Nëse shembulli është krejtësisht homogjen, atëherë entropia është zero dhe nëse shembulli është e ndarë në mënyrë të barabartë (50% - 50%), entropia është një. Entropia mund të llogaritet duke përdorur formulën: $Entopy = -plog_2 p - (1-p)log_2(1-p)$.

Këtu p dhe 1-p janë probabiliteti i suksesit dhe mos suksesit përkatësisht në atë nyje. Entropia përdoret gjithashtu me variablat kategorike të targetuar. Ajo zgjedh ndarjen që ka entropinë më të ulët në krahasim me nyjen prindore dhe ndarjet e tjera. Sa më e vogël është entropia, aq më mirë është shpërndarja.

1. Llogarisim entropinë e nyjes prind.

2. Llogarisim entropinë e çdo nyje individuale të ndarjes dhe llogarisim mesataren e ponderuar të të gjitha nën-nyjeve që janë në këtë ndarje.

Le të përdorim këtë metodë për të identifikuar ndarjen më të mirë për shembullin e mësiperm.

1. Entropi e nyjes prind - (18/36) log2 (18/36) - (18/36) log2 (18/36) = 1. Kjo tregon se ajo është një nyje e papastër.

2. Entropi për nyjen femra = - (4/16) log2 (4/16) - (12/16) log2 (12/16) = 0.81 dhe nyjen me gjininë mashkullore, - (14/20) log2 (14/20) - (6/20) log2 (6/20) = 0.88

3. Entropia për ndarjen Gjinia = Entropia e ponderuar e nën-nyjeve = (16/36) * 0. 81+ (20/36) * 0.88 = 0.85

4. Entropi për nyje Klasa XI, - (6/16) log2 (6/16) - (10/16) log2 (10/16) =0.95 dhe nyjen Klasa XII, - (12/20) log2 (12 / 20) - (8/20) log2 (8/20) = 0.970

5. Entropia për ndarje Klasa = (16/36) * 0.95+ (20/36) * 0.97 = 0.96

Nga mësipër mund të shikojmë se entropia për ndarjen ne gjini është më e ulta midis të gjithëve, kështu që pema do të ndahet në gjinia. Ne mund të marrim informacion nga entropia si 1- Entropia.

Reduktimi i Variances

Deri tani, kemi diskutuar algoritmet për variablin përgjegjës kategorike. Reduktimi i variancës është një algoritëm që përdoret për variablin e vazhdueshëm në problemet e regresit. Ky algoritëm përdor formulën standarde të ndryshimit për të zgjedhur ndarjen më të mirë. Ndarja me variancë të ulët zgjidhet si kriter për ndarjen e popullsisë:

$$Variance = \frac{\sum_{i=1}^{n}(X - \bar{X})^2}{n}$$

Hapat në llogaritjen e variancës

Llogarisim variancën për secilën ndarje si mesatare të ponderuar të çdo vargu të nyjeve.

Le të caktojë vlerën numerike 1 për ata që luajnë basketboll dhe 0 për ata qe nuk luajnë basketboll. Tani ndjekim hapat për të identifikuar ndarjen e duhur:

1. **Varianca e nyjes rrënjë, vlera mesatare është:** (18*1 + 18*0)/36 = 0.5 dhe në këtë rast në bazë të shënimit të mësiperm kemi 18 njësha dhe 18 zero.

   Varianca është: ((1-0.5)^2+(1-0.5)^2+….10 here +(0-0.5)^2+(0-0.5)^2+…8 here) / 36, te cilën mund ta shkruajmë: (18*(1-0.55)^2+18*(0-0.5)^2) / 36 = **0.25**

2. Mesatarja e nyjes femra = (4*1+12*0)/16=0.25 dhe Varianca = (4*(1-0.25)^2+12*(0-0.25)^2) / 16 = 0.19
3. Mesatarja e nyjes meshkuj = (14*1+6*0)/20=0.7 dhe Varianca = (14*(1-0.7)^2+6*(0-0.7)^2) / 20 = 0.21
4. Varianca për shpërndarjen gjinia = Variancën e ponderuar të nën-nyjeve = (16/36)*0.19 + (20/36) *0.21 = **0.21**
5. Mesatarja per nyjen e klases XI = (6*1+10*0)/16=0.375 dhe Varianca = (6*(1-0.375)^2+10*(0-0.375)^2) / 16= 0.23

6. Mestarja për nyjen e klasës XII = (12*1+8*0)/20=0.6 dhe Varianca = (12*(1-0.6)^2+8*(0-0.6)^2) / 20 = 0.24
7. Varianca për shpërndarjen klasa = (16/36)*0.23 + (20/36) *0.24 = **0.24**

   Nga llogaritjet e mësiperme shikojmë se ndarja gjinia ka variancë më të ulët krahasuar me nyjen prind, kështu që ndarja do të ndodh në variablin gjinia.

2.15 Përfundime

Në këtë kapitull jepet një përshkrim i shpërndarjes së bazës së të dhënave duke dhënë rregullat dhe kriteret që përdoren për të arritur në një pemë maksimale. Paraqitet mënyra për të ndërtuar kualifikuesin e pemës si dhe metodologjia që përdoret për selektimin e atributeve të një baze të dhënash, kjo jepet për variablat e vazhdushme dhe ato diskrete. Nje vendë të rëndesishvm zë reduktimi i papastërtisë për të arritur dhe realizuar një shpërndarje sa më të mirë. Një nga idetë kryesore të shpërndarjes është që të përdorim shpërndarjet probabilitare në vënd të një ndarje fikse dhe të përcaktojmë keto probabilitete duke analizuar sjelljet e shpërndarjes nën të ashtuquajturën zhurma"noise". Në këtë kapitull adresohen mënyrat themelore të shpërndarjes së variablave me anë të selektimit të tyre për të ndërtuar pemën klasifikuese duke përdorur dhe paragjykimet e indeksit Gini në selektimin e variablave, gjithashtu në rastet kur p-vlera ndryshon dhe ndikon fuqimisht në cilësin e shpërndarjes si në rastet kur ndërvartesia midis variablave është e dobet apo e fort. Ne pjesën e fundit nëpërmjet një shëmbulli me të dhëna reale zbatohen katër algoritme
të ndryshme duke treguar se si duhet të realizohet ndarja për një nyje të caktuar.

KAPUITULLI 3

KRASITJA

3.1 Krasitja

Një nga pyetjet që lind në algoritmin për pemën vendimmarrëse është madhësia optimale e pemës përfundimtare. Një pemë e madhe ka shumë rreziqe, me të dhëna mbiperputhese dhe ka cilësi të dobëta për të bërë përgjithësime. Një pemë e vogël nuk mund të japë informacion të rëndësishëm strukturor në lidhje me të dhënat në studim. Megjithatë, është e vështirë për të të treguar me një algoritëm se kur pema duhet të ndalet, sepse është e pamundur për të të treguar nëse shtimi i një nyje të vetme shtesë do të ulë në mënyrë dramatike gabimin. Ky problem është i njohur si efekt horizont. Një strategji e përbashkët është që të rritet kjo pemë derisa çdo nyje të përmbajë një numër të vogël të rasteve dhe mbas kësaj duhet që të heqim nyjet që nuk japin informacion shtesë.
Shkurtimi duhet të zvogëlojë madhësinë e një peme pa ulur saktësinë parashikuese të matur nga vlerësimi i kryqëzuar. Ka shumë teknika për krasitjen e një peme të cilat ndryshojnë nga matjet që janë përdorur për të optimizuar performancën.

Pemët vendimmarrëse dhe listat që do të shpërndahen në copa të shkallëzuara në pjesë që janë të papajtueshme dhe të ndara në pjesë të ndryshme ku secila të jetë etiketuar si një klasë e caktuar. Përshkrimi i pjesë që i përket një klase të veçantë mund të shndërrohet në formë te papajtueshem normale duke përdorur standartin në operacionet logjike. Në këtë formë çdo klasë është përshkruar nga një pohim premisa e të cilit përbëhet nga një shperndarje, duke e përkufizuar çdo seksion si dhe kujt klase i përkasin. Komponentët individualë janë quajtur të papajtueshme. Në pemën vendimtare edhe nyjet, janë të shpërndara dhe reciprokisht të papajtueshme, që do të thotë se ato nuk mbivendosen në çdo cep të hapësirës ku ato shtrihen.

Një nga problemet që kërkon një vëmëndje të veçant është dhe gjetja e një pemë përfundimtare e cila duhet të jetë një pemë e thjesht e lexueshme dhe e interpretueshme. Për të aritur këtë së pari ne duhet të rrisim një pemë të cfardoshme dhe pas kësaje duhet të bëjmë të ashtuquajturin proçes të krasitjes. Për të realizuar krasitjen përdoren disa metoda dhe një nga metodat kryesore është "kosto e përgjithshme".
Minimumi i një peme T do të quhet rrënjë e pemës. Kjo rrënjë do të shpërndahet në dy degë të cilat i quajmë degë e majtë dhe e djathtë dhe i shënojmë me t=e majtë (s) dhe t= e djathtë(s) dhe s e quajmë prind të t. $\widetilde{T}$ jep një bashkesi të nyjeve përfundimtare dhe elementet e T — $\widetilde{T}$ i quajmë nyje jofundore. Një pemë do të quhet e parëndësishme nëse plotësohet një nga këto kushte |T|=1; $|T|=1; T=\{rrenj(T)\}, T-\tilde{T}, eshte, bosh$ ndryshe T është e rëndësishme. Për një pemë të dhënë të rëndësishme T marrim $t_1 = rrenj(T), t_L = majt(t_1), t_R = djatht(t_1), T_L = T_{t_L}, dhe, T_R = T_{t_R}$ athërë T_R, dhe, T_L i quajmë degët kryesore të djathtë dhe të majtë të pemës T. Këto dy degë janë bashkësi të papajtueshme dhe jo boshe, ku bashkimi i të cilave jep T dhe po ashtu $\tilde{T}_L, dhe, \tilde{T}_R$ janë bashkësi të papajtueshme jo boshe bashkimi i të cilave jep $\tilde{T}$ dhe në veçanti,

$$|T| = 1 + |T_L| + |T_R| \qquad (3.1.3)$$

$$|\tilde{T}| = |\tilde{T}_L| + |\tilde{T}_R| \qquad (3.2.3)$$

Vetitë e pemëve përgjithsisht vërtetohen duke përdorur induksionet matematike, bazuar në vrojtimet e degëve primare te pemëve te rëndësishme dhe që kanë me pak nyje fundore se pema origjinale. Për shembull nga barazimi (3.1.3) dhe (3.2.3) me induksion matematikë provohet se $|T|=2|\tilde{T}|-1$.

Në përgjithësi pema përfundimtare është një parashikues i fuqishëm që në menyrë eksplicite paraqet një stukturë të caktuar të bazës të dhënave. Saktësia dhe kuptueshmëria varet se sa koncizë jemi në të mësuarit e algoritmeve për të gjetur strukturën përfundimtare të pemës. Modeli perfundimtar nuk duhet të ndërthuret me modelet negative të strukturës përfundimtare të cilat nuk përgjithësojnë vlerat positive. Mekanizmi i krasitjes kërkon një instrument të ndjeshëm që të përdoret në këtë bazë të dhënash dhe të zbulojë nëse marrëdhënia midis komponenteve të bashkesise se parcaktimit është autentike. Proçesi i krasitjes thjeshton klasifikuesin dhe përmirëson përformancën e tij duke eleminuar disa komponente. Gjithashtu ky proçes lehtëson analizën e mëtejshme të modelit tonë përfundimtar. Sigurisht që krasitja duhet të mos eleminojë pjesët parashikuese të klasifikuesit. Rrjedhimisht procesi i krasitjes së pemës klasifikuese kërkon një mekanizëm që të vendosi nëse një bashkësi e caktuar është parashikuese apo jo dhe të bëjë lidhjen e çdo elementi me të gjithë elementet e te dhenave. Algoritmi i krasitjes gjithashtu do të përdore dhe testin statistikor i cili ndihmon ne krahasimin e hipotezës bazë dhe hipotezën alternative. Qëllimi kryesor është që të maksimizojmë saktësinë e parashikimit. Për të gjetur pemën e cila të jetë sa më e thjeshtë me një llojë saktësie, për të bërë krasitjen metoda që përdoret është kosto e pergjithshme. Kjo metode konsiston në rritjen e vazhdueshme të parametrit kompleks gjatë proçesit të krasitjes. Duke filluar nga nyja përfundimtare këto nyje mund të krasiten nëse rezultati ndryshon parashikimin e kostos se mosklasifikimit dhe ky ndryshim është më i paktë se terësia e pemës. Parametri i përgjithshëm është masë se sa shumë është shtuar saktësia e shpërndarjes në të gjithë pemën për të garantuar kompleksitetin shtesë. Nëse parametri kompleks është rritur atëherë më tepër nyje janë dhe duhet të krasiten dhe si rezultat i kësaj pema vjen duke u thjeshtuar. Kërkuesit dhe përdoruesit e shumtë të kësaj metode kanë arritur në përfundimin se pema më mirë dhe më e thjeshtë është pema që ka përmasa të arsyeshme dhe qartësisht të lexueshme dhe të interpretueshme e cila në esencë arrihet pas një krasitje të kujdesëshme dhe e bazuar në kritere të sakta. Në rrisim pemën mjaft të madhe dhe e shënojmë këtë pemë fillestare $T_{\max}$ dhe paskësaj duhet të fillojmë proçesin e krasitjes nga nyjet fundore dhe të vazhdojmë deri te nyjet rrënjë. Së pari duhet të përkufizojmë krasitjen.

Përkufizim 3.6: Një degë T_t e T me një nyje rrënjë $t \in T$ konsiston në nyjen t dhe të gjitha pasardhësit e t në T.

Përkufizim 3.7: Krasitja e një dege T_t e T nga një pemë T konsiston në fshirjen nga T të gjithë pasardhësve të t dhe që është duke krasitur të gjitha ato T_t përveç nyjes rrënjë.

Pëkufizim 3.8: Nëse T' është marrë nga T pas një krasitje të suksesëshme të degëve, atëherë T' do të quhet një nënpemë e krasitur e pemës T dhe është e tillë që $T' < T$ (dhe të dyja këto pemë kanë të njëjtën nyje rrënjë).

Madje në rastin kur një peme ka më shumë se 40 deri ne 50 nyje atje është një numër shumë i madh i nën-pemëve, bile edhe një numër shumë i madh mënyrash për të krasitur këtë pemë deri sa të arrijmë te një pemë optimale dhe që ti shërbejë qellimit tonë të cilën e

shënojmë (t_1). Për këtë duhet të selektojmë proçedurën më të mirë për të arritur te nënpema që bënë një përshkrim më të mirë të dhënave tona. Nga kriteret më të mira është kriteri i vlerësimit të raportit të mosklasifikimit $R^*(T)$ për pemët e ndryshme gjatë këtij proçesi. Pavarësisht se sa e madhe është ndërtuar pema maksimum $T_{\max}$, çfarë kriteri shpërndarje kemi përdorur, çfarë proçesi i përzgjedhjes kemi përdorur për krasitjen, vlerësimi i R(T) për cdo nyje $t \in T_{\max}$ bëhet në mënyrë progresive dhe e krasitim pemën maksimale duke filluar nga nyjet fundore dhe duke vazhduar te nyja rrënjë me kushtin se R(T) të jetë sa më e vogël që të jetë e mundur.

Dan Steinberg 2004 CAS P.M. thotë: "Brenda çdo peme të madhe është një pemë e vogël perfekte e cila është duke pritur për tu gjetur".

Le të supozojmë se një pemë e cila është rritur në maksimum ka L nyje fundore, atëherë ndertojmë një varg të tillë që të jetë në zvogëlim dhe të gjejmë gjithmon një pemë më të vogël ose e quajtur ndryshe më e thjeshtë. $T_{\max}, T_1, T_2, ...\{t\}$ e kështu me radhë. Për çdo vlerë H, ku $1 \le H < L$ le të marrim në konsiderate klasën T_H për të gjitha nënpemët e pemës maksimum $T_{\max}$ do të kemi L-H nyje fundore të lëna. Duke selektuar T_H si një nënpemë e cila maksimizon R(T), atëherë kjo jep $R(T_H) \min_{T \in T_H} R(T)$ ose ndryshe T_H është kosto minimale e pemës me L-H nyje fundore. Ky është një proçes që intuitivisht duhet të zbatohet duke përdorur algoritme.

3.2 Krasitja duke minimizuar koston e përgjithshme

Le të përkufizojeme koston e përgjithshme

Përkufizim 3.9: Për çdo nën pemë $T < T_{\max}$, përkufizojmë kompleksitetin e nyjeve fundore $|\tilde{T}|$ në T. Le të kemi $\alpha \ge 0$, një numër real të cilin e quajmë parametrin e kompleksitetit dhe përkufizojmë si masë të kostos së përgjithshme dhe e shënojmë me $R_\alpha(T)$ të dhënë si më poshtë: $R_\alpha(T) = R(T) + \alpha |\tilde{T}|$

Kështu shihet se $R_\alpha(T)$ është një kombinim linear i kostos së pemës dhe kopleksitetit të saj. Tani për një vlerë të caktuar të α ne gjejmë një nën-pemë $T_\alpha < T_{\max}$ e cila ka një minimum $R_\alpha(T)$ kështu që kemi: $R_\alpha(T(\alpha)) = \min_{T < T_{\max}} R_\alpha(T)$, në se α është e vogël, atëherë mundësia për të pasur numër të madh të nyjeve fundore është i vogël dhe $T(\alpha)$ është e madhe. Për shembull, nëse kur $T_{\max}$ është shumë e madhe dhe ku çdo nyje fundore ka një element nga të dhënat atëherë çdo rast është i klasifikuar korrektësisht kur $R(T_{\max}) = 0$ dhe $T_{\max}$ minimizon $R_0(T)$, nëse mundësia për paramatrin e kompleksitetit alfa në nyjet fundore rritet, atëherë zvogëlimi i nënpemëve $T(\alpha)$ dhe këto pemë do të kenë më pak nyje fundore. Pra, për një vlere te α shumë të madhe minimizimi i një nënpeme T_α konsiston në vetëm një nyje rrënjë dhe pema maksimale $T_{\max}$ është komplet e krasitur. Gjithashtu nëse α gjatë gjithë kohës është një madhësi e vazhduar atëherë janë e shumta një numër i kufizuar nënpemësh nga pema maksimale $T_{\max}$. Proçesi i krasitjes do të na japë një numër të kufizuar vargjesh të pemëve të

ndryshme T_1, T_2, T_3..... të cilat në menyrë progresive japin më pak nyje. Përfundimisht ky proçes në të cilin pema e duhur $T(\alpha)$ e cila është një pemë minimale që arrihet për një vlerë të caktuar të alfës gjatë zmadhimt të saj, vlerë të cilën e shënojmë α' për pemën e $T(\alpha')$, që është pemë e duhur. Në këtë proçes të krahasitjes ne duhet tu përgjigjemi disa pyetjeve si: është atje vetëm një pemë unike $T < T_{\max}$ e cila minimizon $R_\alpha(T)$?

Në minimizimin e vargut të pemëve T_1, T_2, T_3..... është çdo pemë pasardhëse e marrë pas një proçesi krasitje të pemës parardhëse dhe të plotësohet ky kusht $T_1 > T_2 > T_3$,..... $> \{t\}$?

Më praktikisht më se i rëndësishëm është të gjejmë një algoritëm për të zbatuar këtë proçes krasitje për tëmbrritur në një minimum të $R_\alpha(T)$.

Përkufizim 3.10 : Nënpema më e vogël $T(\alpha)$ për parametrin kompleks α është e përkufizuar nga këto kushte: $R_\alpha(T(\alpha)) = \min_{T<T_{\max}} R_\alpha(T)$

$R_\alpha(T) = R_\alpha(T(\alpha))$ *atehere* $T(\alpha) < T$

Ky përkufizim çon në minimum të kostos së përgjithshme duke selektuar gjithashtu dhe vlerën më të vogël të R_α . Qartësisht nëse një pemë e tillë ekziston ajo është dhe unike. Por pyetja kryesore është nëse ekziston dhe më konkretisht nëse ne supozojmë se kemi dy pemë minimale T, dhe, T' të $R(\alpha)$ dhe le të supozojmë se është dhe një pemë tjetër. Atëherë $T(\alpha)$ është përkufizuar si më sipër dhe kështu që pemë tjetër nuk egziston.

Rrjedhim 3.1: Për çdo vlerë të α , egziston një vlerë më e vogël që minimizon nënpemën që ne përkufizuam më sipër,vertetimi i të cilit është në fund të këtij materiali. Pika për të cilën duhet të fillojmë krasitjen në përgjithësi nuk është $T_{\max}$ por më tepër $T_1 = T(0)$. Kështu që nënpema më e vogël që kënaq kushtin $R(T_1) = R(T_{\max})$ është T_1, për të gjetur këtë pemë T_1 nga $T_{\max}$, le të marrim t_L, t_R nga një nyje fundore të pemës maksimale $T\max$ të cilat merren nga një shpërndarje e një nyje të çfardoshme t. Nga ku $R(t) \geq R(t_L) + R(t_R)$. Nëse kemi $R(t) = R(t_L) + R(t_R)$, atëherë krasitim t_L, dhe, t_R, e vazhdojmë këtë proçes deri sa të mos bëhen më krasitje. Për një degë të çfardoshme T_t nga pema T_1 përcaktojmë $R(T_t)$ si $R(T_t) = \sum_{t' \in \tilde{T}_t} R(t')$ ku $\tilde{T}_t$ është bashkësia e nyjeve fundore të T_t .

Rrjedhim 3.2: Për ndonjë t të një nyje jo fundore nga pema T_1 kemi $R(t) > R(T_t)$, duke filluar me T_1, kryesore në minimizimin e kostos së përgjithshme, gjatë krasitjes është në të kuptuarit se ajo punon sipas parimit që të krasitet lidhja më e dobët në pemë. Për një nyje të çfardoshme $t \in T_1$, e dhuruar nga {t}, një nëndegë e T_t konsiston në një nyje të vetme të përcaktojemë {t}. Marrim, ose vendosim $R_\alpha(\{t\}) = R(t) + \alpha$ për çdo degë T_t dhe përkufizojmë $R_\alpha(T_t) = R(T_t) + \alpha \mid \tilde{T}_t \mid me$ kusht që $R_\alpha(T_t) < R_\alpha(\{t\})$, atëherë dega T_t ka vlerën minimale të kostos së përgjithshme nga një nyje e vetme e marrë nga bashkësisa e {t}. Por në disa pika kritike të α kemi që dy vlera të kostos së përgjithshme bëhen të barabarta. Në këtë pikë nëndega e {t} është më e vogël se T_t dhe ka të njëtin kosto të përgjithshme dhe kjo është pema e preferuar. Për të gjetur këtë pikë kritike duhet të zgjidhim inekuacionin

$R_\alpha(T_t) < R_\alpha(\{t\})$ dhe gjejmë $\alpha < \dfrac{R(t) \_ R(T_t)}{\tilde{T}_t - 1}$ nga rrjedhimi i mësipërm pika kritike në krahun e djathtë të mosbarazimit të mëspërm është pozitive. Përkufizojmë funksionin $g_1(t)$, ku $t \in T_1$ si më posht: $g_1(t) = \left\{ \begin{matrix} \dfrac{R(t) - R(T_t)}{|T_t| - 1}, t \notin \bar{T}_1 \\ +\infty, t \in \bar{T}_1 \end{matrix} \right\}$, gjithashtu përkufizojmë lidhjen më të dobta $\bar{t}_1$ në T_1 nyje të tillë që $g_1(\bar{t}_1) = \min_{t \in T_1} g_1(t)$ dhe vendosim $\alpha_2 = g_1(\bar{t}_1)$. Nyja $\bar{t}_1$ është më e dobëta lidhje në kuptimin se nëse parametri alfa rritet, ajo është nyja e parë që vlera e $R_\alpha(\{t\})$ bëhet e barabartë me $R_\alpha(T_t)$, ku rrjedhimisht $\{\bar{t}_1\}$ është preferuara e $T_{\bar{t}_1}$ dhe α_2 është vlera e alfës në të cilën barazimi realizohet. Përkufzojmë një pemë të re $T_2 < T_1$ duke kryer krasitjen në degën $T_{\bar{\bar{t}}_1}$, dhe kjo është: $T_2 = T_1 - T_{\bar{\bar{t}}_1}$. Duke përdorur T_2 në vënd të T_1, gjejmë lidhjen më të dobët në T_2. Më saktësisht duke marrë në konsideratë T_{2t} si një nëndegë e degës T_t e cila nga ana e sajë ndodhet në T_2, përkufizojmë $g_1(t) = \left\{ \begin{matrix} \dfrac{R(t) - R(T_{2t})}{|\tilde{T}_{2t}| - 1}, t \in T_2, t \notin \tilde{T}_2 \\ +\infty, t \in \tilde{T}_2 \end{matrix} \right\}$ ku

$\bar{t}_2 \in T_2, \alpha_3, eshte, dhene,$

$g_2(t_2) = \min_{t \in T_2} g_2(t)$

$\alpha_3 = g_2(\bar{t}_2)$

Duke e përsëritur këtë proçes dhe duke përkufizuar $T_3 = T_2 - T_{\bar{t}_2}$ dhe duke gjetur lidhjen më të dobët $\bar{\bar{t}}_3$ në T_3 dhe parametrin korespondues me vlerë α_4, dhe kështu në T_4 e përsëritim sërisht procesin. Nëse në ndonjë hap gjejmë një shumëfish të lidhjeve të dobëta domethënë. nëse $g_k(\bar{t}_k) = g_k(\bar{t}^{++}_k)$, atëherë përkufizojmë: $T_{K+1} = T_K - T_{\bar{t}_k} - T_{\bar{t}}$. Duke vazhduar në këtë mënyrë marrim një varg zvogëlues të nënpemëve $T_1 > T_2 > T_3 > ..... > \{t_1\}$ dhe përgjigjen për minimizimin e e kosto së përgjithshme e jep teorema e mëposhtme.

Teoremë 3.5: Nëse $\{\alpha_k\}$ është një varg në rritje atëherë atje është $\alpha_k < \alpha_{k+1}, per, k > 1, ku, \alpha_1 = 0. Per, k \geq 1, \alpha_k \leq \alpha < \alpha_{k+1}, T(\alpha) = T(\alpha_k) = T_k$.

Kjo teoremë jep informacion se si minimizimi i kostos së përgjithshmese punon. Ne e fillojmë me T_1, gjejmë degën me lidhjen më të dobët $T_{\bar{t}_1}$ dhe e krasitim që të gjejmë pemën T_2 kur α arrin α_2. Tani gjejmë degën me lidhjen më të dobët në T_2 e cila është $T_{\bar{\bar{t}}_2}$ dhe e

krasitim atë të gjejmë T_3 kur α arrin α_3 e kështu me radhë vazhdojmë këtë proces. Ky proçes krasitje i cili përsëritet disa herë, matematikisht arrihet me një llogaritje të shpejtë dhe kërkon një kohë shumë më të shkurtër se koha e ndërtimit të pemës. Duke filluar me T_1, ky algoritëm fillimisht tenton të krasiti nënpemën e cila ka shumë nyje fundore. Pasi pema është duke u zvogëluar, proçesi gjatë kësaj kohe tenton të krasiti më pak. Përfundimisht vargu i minimizimit të kostos së përgjithshme te pema është një nënvarg i vargut të pemëve të ndërtuara ku gjejmë një pemë me një numër të reduktuar të nyjeve fundore dhe me një kosto minimale.

$Pema: T_1, T_2, T_3, T_4, T_5, T_6, T_7, T_8, T_9, T_{10}, T_{11}, T_{12}, T_{13}$
$\overline{T_k}$: 71 63 58 40 34 19 10 9 7 6 5 2 1

Tabela 7: *Numeri i nyjeve per çdo peme*

Në tabelën 7 paraqitet një rast konkret se si punon kjo proçedurë.
Nga tabela 7 , kur $T(\alpha)$ ka shtatë nyje fundore nuk ka më nyje të tjera pas kësaj që të kenë më të vogël R(T), kështu që $R_\alpha(T) = R(T) + 7\alpha < R_\alpha(T(\alpha))$ i cili sipas përkufizimit është i pamundur. Në rastin e bazës së të dhënave "Boston House Market" tabela 8 e mëposhtëme paraqet se si arrihet kosto e përgjithshme.
Kjo metodë e krasitjes që u dikutua më sipër në një varg zvogëlues të nënpemëve $T_1 > T_2 > T_3 > ..... > \{t_1\}$ ku $T_k = T(\alpha_k), \alpha_1 = 0$ dhe në këtë situate problemi reduktohet në zgjedhjen e një peme me përmasa optimale. Nëse rizëvëndësimi vlerëson $R(T_k)$, atëherë kjo do të përdoret si kriter për të përzgjedhur pemën më të madhe T_1. Por nëse në një pemë është bërë vlerësimin e kostos se mosklasifikimit $\hat{R}(T_k)$ vlerë e cila është mëe vogla, atëherë kjo nënpemë është pema e duhur të cilën e shënojmë T_{k_0} ku $\hat{R}(T_{k0}) = \min_k \hat{R}(T_k)$.

Kjo metodë e krasitjes që u dikutua më sipër në një varg zvogëlues të nënpemëve $T_1 > T_2 > T_3 > ..... > \{t_1\}$ ku $T_k = T(\alpha_k), \alpha_1 = 0$ dhe në këtë situate problemi reduktohet në zgjedhjen e një peme me përmasa optimale. Nëse rizëvëndësimi vlerëson $R(T_k)$, atëherë kjo do të përdoret si kriter për të përzgjedhur pemën më të madhe T_1. Por nëse në një pemë është bërë vlerësimi i kostos se mosklasifikimit $\hat{R}(T_k)$ vlerë e cila është më e vogla, atëherë kjo nënpemë është pema e duhur të cilën e shënojmë T_{k_0} ku $\hat{R}(T_{k0}) = \min_k \hat{R}(T_k)$.

	CP	nsp	litrelerror	xerror	xstd
1	0.1250000	0	1.0000	1.00000	0.063918
2	0.1000000	1	0.8750	0.97500	0.063530
3	0.0625000	2	0.7750	0.96250	0.063328
4	0.0250000	3	0.7125	0.88750	0.061984
5	0.0187500	5	0.6625	0.85625	0.061357
6	**0.0125000**	**7**	**0.6250**	**0.95000**	**0.063119**
7	0.0093750	10	0.5875	0.99375	0.063823
8	0.0083333	32	0.3375	1.00625	0.064011
9	0.0062500	35	0.3125	0.97500	0.063530
10	0.0031250	53	0.2000	0.98750	0.0637271
11	0.0000000	57	0.1875	0.98750	0.063727

Tabela 8: *Kostua e përgjithshme e një baze të dhënash*

3.3 Nënpema më e mirë e krasitur

Në këtë proçes të krasitjes dhe të përzgjedhjes së pemës më të mirë, janë dy përshtatje për të gjetur më të mirën:

Së pari përdorim një provë nga testimi i shembullit të zgjedhur dhe së dyti vlefshmerine e kryqezuar.

Nëse kemi një bazë të dhënash me shumë elementë, mund të njehsojmë përqindjen e gabimit duke provuar të gjitha pemët se cila ka gabimin më të vogël. Sidoqoftë në praktikë shumë rrallë kemi një bazë të dhënash shumë të madhe, por edhe në raste se kemi një bazë të dhënash shumë të madhe mund të përdorim këtë bazë të dhënash si zgjedhje për të gjetur pemën më të mirë. Le të marrim një shëmbull, i cili ka dy nivele, të cilat mund të paraqiten si përgjegjës dhe jo pergjëgjës ose 0 dhe 1. Probabiliteti i pasëm përgjegjës në një nyje është raporti i regjistrimit me nivelin e caktuar që është i barabartë me nivelin përgjegjës ose 1, brënda kësaj nyje. Në mënyrë të ngjashme, probabiliteti i pasëm për nivelin jo përgjegjës të nyjes është raport i regjistrimit me nivelin e caktuar e cila ndryshe është si jo përgjegjës ose 0 brënda kësaj nyje. Këto probabilitete të pasme janë të përcaktuara gjatë proçesit dhe ato bëhen pjesë e vendimit për të gjetur modelin përfundimtar të pemës.

Qëllimi i krasitjes është të heqë disa pjesë të një modeli të pemës klasifikuese që duam të përshkruajnë duke përdorur ndryshimet e rastit në shembullin që përdorim si model për trajnim, në vënd të tipareve themelore të fushës së përcaktimit. Kjo e bën modelin më të kuptueshëm për përdoruesit, dhe potencialisht më të saktë në të dhëna e reja që nuk i kemi përdorur për trajnimin e klasifikuesit. Kur krasitim, një mekanizëm efikas i cili është i nevojshëm,është që të dallojmë pjesët e një klasifikuesi që janë për shkak të efekteve të

rastësishme nga pjesë që përshkruajnë strukturën përkatëse. Testet statistikore të rëndësishme për të përcaktuar nëse një efekt i vërejtur është një tipar i vërtetë i një fushe përcaktimi ose eshte aty vetëm për shkak të luhatjeve të rastit. Në këtë mënyrë ato mund të përdoren për të marrë vendimet e krasitjes në modelet e klasifikuesit. Gabimi i reduktuar i krasitjes (Quinlan, 1987A), është një algoritëm standard për pas-krasitjen për të gjetur pemën përfundimtare, e cila nuk merr në konsideratë nivelin statistikor. Ky është i njohur si një nga algoritmat e krasitjes së shpejtë. Ai prodhon nje pemë me dy cilësi që është e sakta dhe më e vogla (Esposito et al., 1997). Kjo pjesë shqyrton nëse testet e rëndësishme mund të përdoren për të përmirësuar këtë proçedurë të njohur si krasitja. Siç do të shohim, problemi kryesor është për të zgjedhur një nivel të rëndësishëm dhe të përshtatshëm të nivelit statistikor i cili duhet të jetë i përshtatëshëm për çdo rast krasitje.

3.4 Testi statistikor

Hipoteza 3.1. Gabimi i reduktuar i krasitjes gjeneron pemë më të vogla dhe më të sakta të pemëve përfundimtare, nëse hapat e krasitjes janë bërë duke përdorur teste ku niveli i testit statistikor është i zgjedhur në mënyrë të përshtatur në çdo baze të dhënash që mund të kemi. Testet e rëndësishme statistikore mund të ndahen në të ashtuquajturat "testet parametrike" që bëjnë disa supozime matematikore në lidhje me funksionin themelor të shpërndarjes, dhe ato të ashtuquajtura "teste jo-parametrike" (Good, 1994) që janë në thelb teste në të cilat nuk njihet shpërndarja probabilitare e të dhënave. Testet e bazuara në shpërndarjen Hi-katror i përkasin grupit të parë. Në këto teste supozojmë se testi statistikor ndjek shpërndarjen Hi-katror. Përdorimi i tyre është i diskutueshm për rastet kur vellimi i zgjedhjes që ne studjojmë është i vogël, sepse atëherë supozimet e kërkuara për zbatimin e shpërndarjes Hi-katror nuk janë të vlefshme. Testi i përkëmbimit, nga ana tjetër, nuk ka të bëjë me supozimet që kanë lidhje me shpërndarjet themelore, dhe i përkasin grupit të dytë të testeve. Si pasojë, ata mund të aplikohen me çdo bazë të dhënash, pavarësisht madhësisë së saj. Sjellja për një bazë të dhënash me vëllim të vogël është veçanërisht e rëndësishme në algoritme dhe konsiderohen si indikuese të pemës përfundimtare, ku duhet të merren vendime të tjera për të krasitur pemën, për të gjetur nënpemën më të mirë. Duke pasur parasysh këto konsiderata, është e mundur, që për një sasi të caktuar të krasitjes së pemës duke përdorur një test të përkëmbimit e cila e bën atë më të saktë duke përdorur teste parametrike për krasitjen e pemës. Me të dy llojet e testeve, sasia e krasitjes që duhet ti bëhet pemës dhe për rrjedhim dhe përmasat e pemës përfundimtare janë të lidhura me nivelin statistikor të testit, gjë e cila na çon dhe në hipotezën e mëposhtëme:

Hipoteza 3.2. Nëse pema vendimtare A është rezultat i krasitjes duke përdorur testin e përkëmbimit një dhe pema vendimtare B është rezultat i krasitjes duke përdorur një test parametrik, dhe të dy këto pemë kanë të njëjtën madhësi, atëherë A do të jetë mesatarisht më e saktë se B. Me poshtë do të shpjegohet pse është e rëndësishme të marrin në konsideratë rëndësinë statistikore kur vendimet e krasitjes janë bërë.

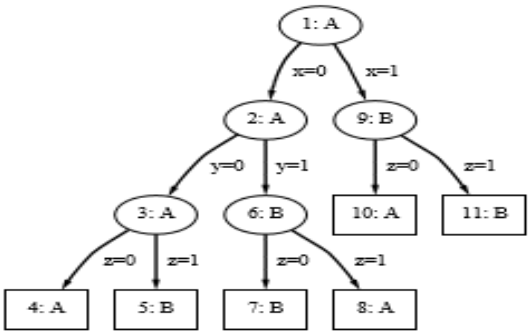


Figura 12: *Një pemë përfundimtare e krasitur*

Supozojmë se çdo klasë është emëruar dhe është e lidhur me çdo nyje në këtë pemë, duke e marrë shumicën e klasave në modelin që shikojmë dhe duke e arritur te çdo nyje e veçantë. Në Figura 12 kemi dy klasa: A dhe B. Pema e paraqitur në këtë figurë mund të përdoret për të parashikuar klasën, ku duke filtruar arrijmë në nyjen fundore të cilën e quajmë gjethe. Megjithatë, duke përdorur një pemë vendimtare të pa krasitur për klasifikuesin i cili potencialisht i mbipërshtatet të dhënave të modelit tonë të trajnimit. Në përgjithësi është e kështillueshme para se pema të përdoret. "Një metodë e përgjithshme, e shpejtë, dhe lehtë për tu zbatuar ne krasitje është "shkurtimi i gabimit te reduktuar" (Quinlan, 1987A). Ideja është që të mbajë disa nga rastet që kemi në dispozicion nga bashkësia e pemëve të krasitura, kur pema është ndërtuar, dhe për të krasitur pemën derisa gabimi i klasifikimit në këtë rast të pavarur fillon të rritet. Për arsye se disa kërkesa në këtë proçes të krasitjes nuk janë përdorur për ndërtimin e pemës përfundimtare, na krijohet nje situate ku kemi një vlerësim të njëanshëm të normës së gabimit të saj dhe në këto raste do ta konsiderojmë atë si një proçes që ka më pak vlerësim real të përqindjes së gabimit. Reduktimi i gabimit të krasitjes do të jetë si një udhëzues për funksionimin e tij.
Figura 15 tregon një shembull të krasitjes së pemës të marrë nga pema e figurës 14.

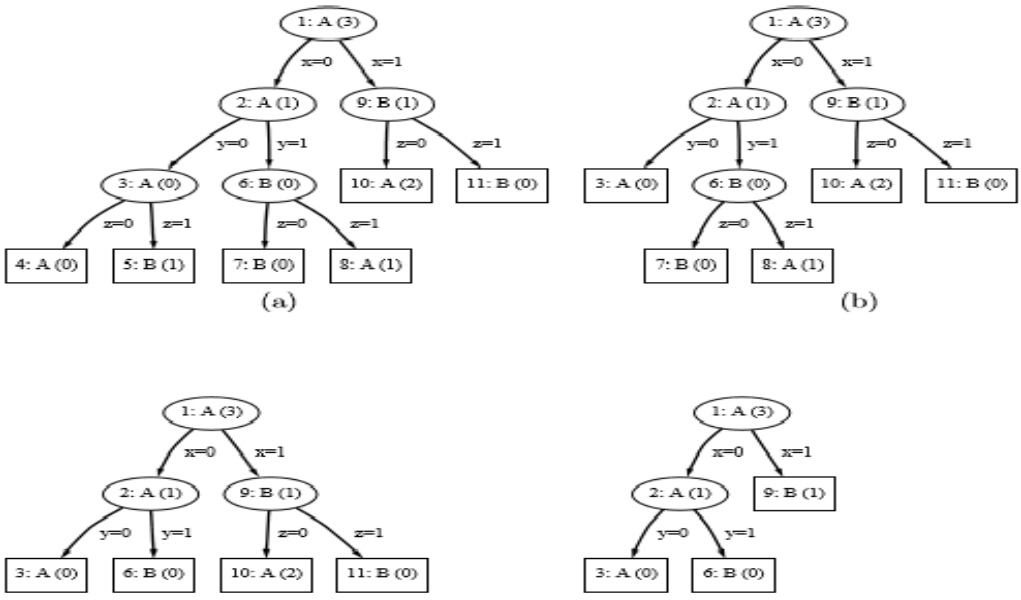


Figura 13: *Pema e krasitur*

Reduktimi i gabimit të krasitjes në shembullin e dhënë do të shfaqet në se nuk do të rrisim numrin e përgjithshëm të gabimeve të klasifikimit. Për të rregulluar këtë pemë duke filluar nga poshtë-lartë duhet të sigurohemi që rezultati te pema më e vogël e krasitur ka gabim minimal mbi të dhënat e krasitjes (Esposito et al., 1995). Kjo strategji e rregullimit është një rezultat i drejtpërdrejtë me kusht që një nyje mund të konvertohet vetëm në një nyje fundore e cila quhet ndryshe gjethe për të gjithë nënpemën e cila tashmë është konsideruar se duhet të krasitet. Duke supozuar se pema është e përshkuar nga e majta në të djathtë, për proçedurën e krasitjes së pari le të marrim në konsideratë largimin e nënpemës së lidhur me nyjen 3 të figurës 13a. Për arsye se gabimi në këtë nënpemë është më madh se gabimi në nyjen tre duhet të konvertojmë nyjen tre si nyje fundore. Nga ana tjetër nyja 6 është zëvëndësuar me nje nyje fundore për të njëjtën arsye, Figura 13c. Duke përpunuar të dy pasardhësit e tij, proçedura e krasitjes pastaj konsideron nyjen 2 për fshirje. Megjithatë, për shkak se nënpema e bashkëngjitur me nyjen 2 e bën atë me më pak gabime (0) se sa gabimi nënyjen 2 i cili është (1 gabim), dhe kështu nënpema mbetet në vend. Nënpema tjetër e zgjedhur që nga nyja 9 konsiderohet si pemë që duhet krasitur, duke rezultuar në një nyje fundore figura 13d . Në hapin e fundit, nyje 1 konsiderohet për shkurtim, duke e lënë këtë pemë të pandryshuar. Për fat të keq, ka një problem me këtë procedurë të thjeshtë dhe elegante të krasitjes: ajo përfshin të dhënat gjatë krasitjes. Oates dhe Jensen (1997). Pasoja është e njëjtë si për overfitting nëse te dhënat që përdorim për trajnim, e cila konsiderohet si një pemë tepër komplekse përfundimtare. Ky është një shëmbull i thjeshtë që tregon se pse mbivendosja ndodh. Për një bazë të dhënash me 10 atributet e rastit me vlera binare të shpërndara në mënyrë uniforme në 0 dhe 1. Supozohet se klasat janë gjithashtu binare, me një numër të barabartë të rasteve për çdo klasë, ku klasat janë etiketuar A dhe B. Sigurisht, norma e pritur e gabimit për këtë fushë është e njëjtë për çdo klasifikues të mundshëm, pritja matematike e gabimt konsiderohet përkatësisht 50%, dhe pema më e thjeshtë e mundshme për këtë problem, duke parashikuar të gjitha klasat të cilat në shumicën e rasteve përbëhet nga nyje fundore. Ne do të donim që të gjejmë këtë pemë të parëndësishëme, sepse mund të nxjerrm një përfundim të saktë ku asnjë nga atributet në këtë rast nuk jep ndonjë informacion në lidhje me klasat e etiketura. Duke aplikuar reduktimin e gabimit për këtë problem, duke përdorur një shembull të krijuar rastësisht prej shumë rastesh, mund të përfitojmë një tjetër kriter i cili u përdor për herë të parë nga (Quinlan, 1986). Në këtë rast dy të tretat e të dhënave janë përdorur për të rritur pemën fillestare e cila është e pa krasitur dhe pjesa e tretë e mbetur është e vendosur mënjanë për proçedurat standarte të krasitjes për të gjetur klasifikuesin duke përdorur hold-out set (Cohen, 1995; Furnkranz, ¨ 1997; Oates & Jensen, 1999).

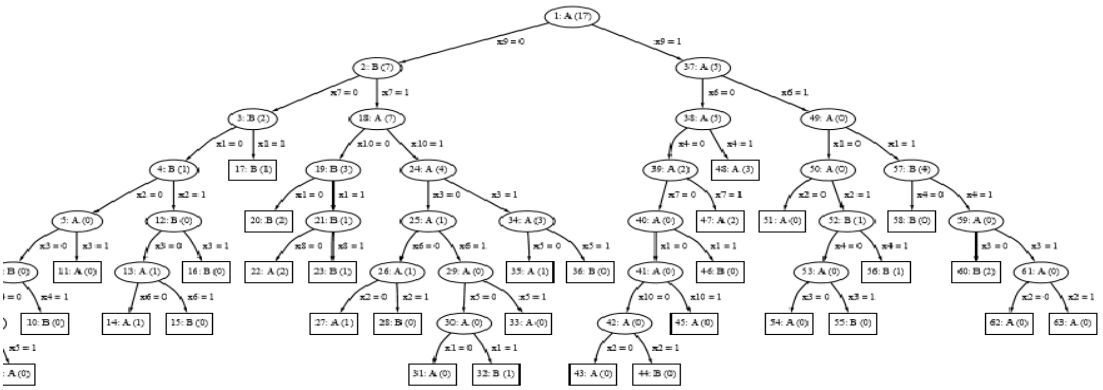


Figura 14: *Pema pra krasitjes*

Figura 14 tregon pemën e pakraasitur. Numri i rasteve në të dhënat e krasitjes që janë të këqija klasifikohen nga nyjet e pemës që janë dhënë në kllapa. Figura 14 tregon të njëjtën pemë pas krasitjes. Figura 14 tregon se, edhe pse kemi reduktuar gabimn e krasitjes me sukses kemi reduktuar dhe madhësinë e pemës se pakrasitur, kjo sigurisht nuk do të gjenerojë një pemë minimale përfundimtare. Kjo hipotezë e lehtë mund të konfirmohet duke përsëritur eksperimentin me grupe të të dhënave të ndryshme të krijuara rastësisht (Jensen & Schmill, 1997). Figura 15 përmbledh rezultatet e arritura duke përsëritur atë shumë herë për secilën nga 10 madhësitë e ndryshme të caktuara në këtë shembull. Nivelet e rendësise janë 95% për marrjen e pemës përfundimtare. Ato tregojnë se duke reduktuar gabimin dhe një shkurtim te vërtetë gjenerohet gjithmonë një pemë tepër komplekse.

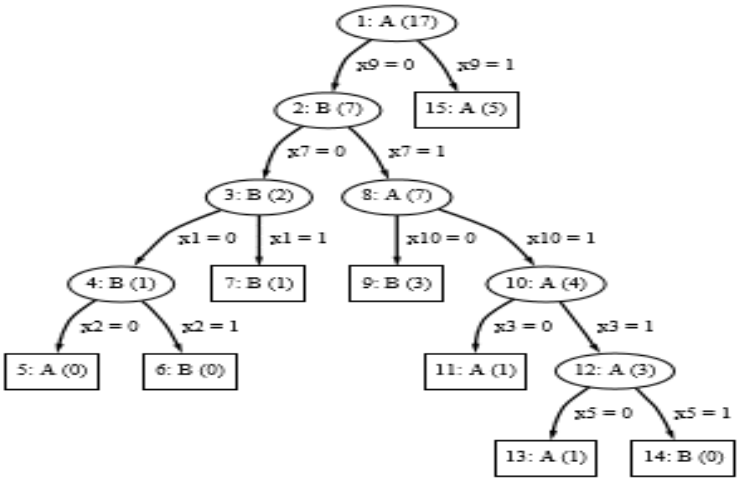


Figura 15: **Pema pas krasitjes**

Figura 14 tregon se kjo pemë është e madhe dhe në një farë menyre e vështirë për tu lexuar. Për shkak të numrit të madh të nënpemëve që duhet të konsiderohen për krasitje, atje ka gjithmonë disa pemë që mund të përshtaten me të dhënat të cilat mund të gjenden vetëm rastësisht. Proçedura e krasitjes së gabur mund ti ruajë këto pemë. Kjo gjithashtu shpjegon se sa më e madhe të jetë pema e pakrasitur, aq më shumë ka të ngjarë që ndonjë nga nënpemët të përshtaten me të dhënat tona e cila mund të merret rastësisht. Problemi lind sepse reduktimi i gabimit në krasitje nuk merr parasysh faktin se mospërputhja e shembujve mund të shkaktojë klasa ku shumica në një nyje të veçantë të jenë të pasakta edhe nëse të dhënat nuk janë përdorur gjatë trajnimit. Shpërndarja e vlerave të klasës në nyjet e një pemë përfundimtare nuk pasqyron domosdoshmërisht shpërndarjen e vërtetë, dhe ky efekt është veçanërisht i theksuar, nëse të dhënat në nyjet e shembullit janë të vogla. Proçedura e krasitjes nuk teston nëse lidhja midis parashikimeve dhe vlerave të vrojtuara te klasës në të dhënat e krasitjes është statistikisht e rëndësishëme, ose të mospërputhet vetëm për shkak të ndryshim të shembullit. Testet e rëndësishme statistikore veçanërisht testet e tabelave te kontigjences janë shumë të rendesishme, pasi një nënpemë është me vlerë dhe do të mbahet vetëm nëse ka një saktësi të konsiderueshëm në mes të parashikimeve të saj dhe etiketimeve të çdo klase në pemën e krasitur.

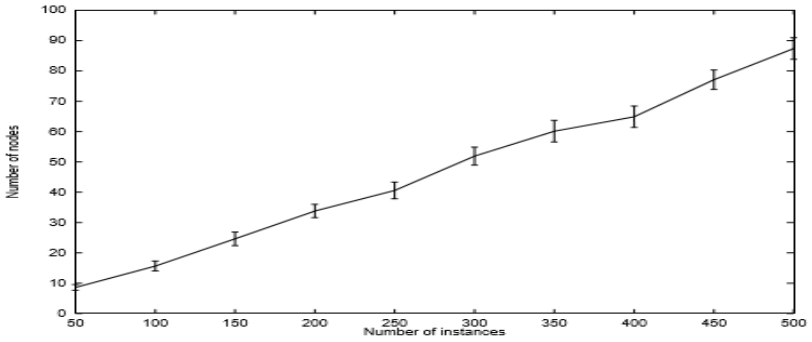


Figura 16: *Madhësia relative e një peme të krasitur duke përdorur gabimin e reduktuar të krasitjes*

3.5 Modelet e pemëve përfundimtare

Një pemë përfundimtare është një kompozim i disa pjesëve:
Përkufizimi i nyjes, ose i rregullave se si të përkufizojmë që cilët elementë të bazës së të dhënave të jenë te nyja përfundimtare, se si të gjejmë probabilitetet pasardhëse të nyjeve fundore, caktimi i nivelit të synuar për çdo nyje fundore.
Probabilitetet pasardhëse janë njehsuar për çdo nyje duke përdorur bazën e të dhënave si shembull i trajningut. Caktimi i nivelit të synuar për çdo nyje fundore është bërë gjithashtu te te dhenat e përdorur. Probabilitetet pasardhëse janë të vrojtuara në raportetet e nivelit të piketuar brënda secilës nyje te të dhenat që përdorim. Caktimi i një niveli të synuar për një rregjistrim individual ose te një nyje si e tëra do të quhet pema përfundimtare. Dhe tani le të shikojmë konkretisht një pemë përfundimtare ku si qëllim kryesor kemi që të maksimizojmë fitimin të minimizojmë shpenzimet ose të minimizojmë gabimin e mosklasifikimit. Për shembull le të shikojmë nëse mund të marrim një pemë përfundimtare kur duam të maksimizojmë fitimin. Le të marrim matricën e mëposhtme në të cilën objektivi është binar.

Objektivi	**Vendimi 1**	**Vendimi 2**
1	$20	0
0	-$1	0

Tabela 9: *Matrica e një shembulli*

Me vendimin 1 është caktuar një nivel i caktuar si objektiv ose përgjegjësi cili është 1.
Me vendimi 2 është caktuar një nivel si objektiv i cili është 0. Matrica e fitimit tregon se në përgjithësi do të kosiderohet si e vërtetë nëse në mënyrë korrekte, atëherë fitimi është $20. Nëse ne nuk kemi një përgjegjës të vërtetë atëhere kemi humbur $1 dhe në këtë rast fitimi është 0.

3.6 Llogaritja e vlerës së një peme

Gjatë proçesit të ndërtimit të pemës klasifikuese ndeshemi me rastet kur duhet të njehsojmë vlerën e pemës. Vlera e një peme mund të njesohet duke përdorur vlershmërinë e një bazë të dhënash. Baza e të dhënave të cilën përdorim për testim, në të cilat nivelet e objektivit janë të njohura për të gjitha hyrjet, gjithashtu kemi të përkufizuar të gjitha nyjet përfundimtare. Për të njehsuar vlerën e pemës përdorim nyjet fundore ose të ashtuquajturat gjethe. Në softuere të ndryshme përdoren metoda të ndryshme kalkulimi dhe një nga këto është ajo e vlefshmërisë së bazës të dhënave duke krahasuar pemët e ndryshme të cilat kanë numër të ndryshëm nyjesh. Gjithashtu vlera e një peme mund të njehsohet duke përdorur bazën e të dhënave që tëstojmë dhe duke krahasuar përformancën e secilës pemë përfundimtare. Në të dy rastet metoda e njëhsimit të vlerës së pemës është e njëjtë.

Per rastin binar kemi dy nivele klasash, përgjegjës dhe jo përgjegjës (për të cilat përdorim shënimin 1 ose 0). Në rastin tonë do të përdorim fitimin të cilin e përdorëm edhe më lartë si masë të vlefshmërisë. Do të tregojmë se si fitimi në një nyje fundore do të njehsohet duke përdorur matricën fituese qe e kemi në tabelën 9. Ky njehsim ka një proçedurë me dy hapa. Së pari, cdo rregjistrim nga vlefshmëria e një baze të dhënash është shënuar në nyjen përfundimtare. Bazuar në rregullat që kemi përcaktuar për çdo nyje përfundimtare të cilën e përdorim në atë pjesë të bazë së të dhënave që përdorim për trajnim.

Të gjitha regjistrimet që janë vendosur në çdo nyje janë të shënuara duke pasur të njëjtat probabilitete të pasme për çdo gjethe, gjatë fazës që punojmë me te dhenat në studim. Ngjashmërisht, të gjitha rregjistrimet që bien në çdo nyje fundore janë të shënuara si nivele target ose klasa që janë përgjegjës ose jo të cilat janë përcaktuar për çdo nyje fundore.

Se dyti, fitimi është njehsuar për çdo nyje fundore të pemës duke u bazuar në vlerën aktuale të qellimit në çdo rregjistrim të vlershmërisë së te dhenave. Nëse një nyje fundore është e klasifikuar si nyje përgjegjëse domethënë niveli 1 dhe duke pasur n_1 të tilla dhe duke shënuar me n_0 rregjistrimet ku nivelet e të cilave janë jo përgjegjëse domethënë 0, atëherë fitimi i sejcilës nyje është: $n_1 * \$20 + n_0 * (-\$1)$. Nëse në anën tjetër nyja fundore është klasifikuar si jo përgjegjëse, atëherë fitimi llogaritet; $n_1 * \$0 + n_0 * (\$0)$. Duke ndjekur këtë procedurë fitimi duhet të njehsojmë fitimin e çdo nyje fundore dhe pastaj të gjejmë shumën e tyre për të gjetur fitimin total të pemës. Fitimi mesatar gjendet duke pjesëtuar fitimin total me numrin total të rregjistrimeve në këtë pemë. Gjithashtu mund të njehsojmë fitimin total dhe atë mesatar duke përdor bazën e të dhënave që kemi për testim. Zakonisht e bëjmë këtë kur duam të krahasojmë përformancën e secilit model të pemëve përfundimtare.

Vlefshmeria e kryqezuar është një raport optimal midis kompleksitetit të pemës dhe gabimit të mosklasifikimit. Kur përmasat e pemës janë duke u rritur, gabimi i mosklasifikimit është duke u zvogëluar dhe nëse marrim njëpemë maksimale, atëherë gabimi i mosklasifikimit është zero. Por në krahun tjetër pema komplekse vendimtare performon në mënyrë të keqe në të dhënat e pavarura. Qëllimi ynë kryesor këtu është të gjejmë një pemë me raporte optimale midis kompleksitetit të pemës dhe gabimit të mosklasifikimit. Kjo arrihet përmes funksionit të kostos komplekse:

$$R_\alpha(T) = R(T) + \alpha(\tilde{T}) \rightarrow \min_T$$

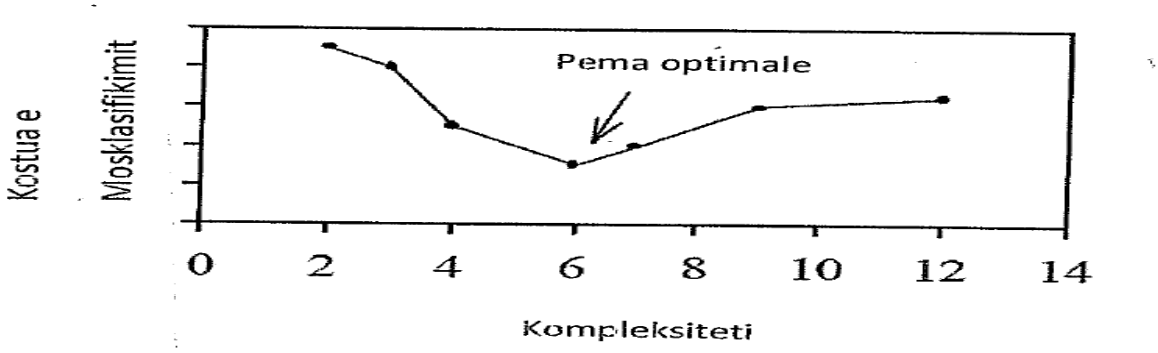


Figura 17: *Zgjedhja e një peme optimale*

ku R(T) –gabimi i mosklasifikimit të pemës, $\alpha(\tilde{T})$ -masa e kompleksitetit e cila varet nga $\tilde{T}$ - shuma totale e nyjeve fundore të pemës, α - një parametër që është gjetur përmes një pjese të zgjedhjes së testimit, kur një pjesë e të dhënave është marrë si model specifik i testimit. Ky proçes duhet të përsëritet disa herë për modelet specifike te përzgjedhura rastësisht për proçesin e testimit të këtij modeli specifik.

Në palosjen me pesë të vlefshmerise së kryqëzuar, për shembull, të dhënat janë ndarë, rastësisht, në pesë nënbashkësi që kanë madhësi të barabarta. Pas kësaj, pema është rritur duke përjashtuar një nga nënbashkësitë, dhe pastaj performanca është vlerësuar në mesin e përjashtuar. Përsëritim hapat në mënyrë të njëjtë në të pesë nënbashkësitë. Së fundi, njehsojmë performimin mesatar për të pesë nënbashkësitë. Kjo bëhet me lehtësi duke përdorur paketën "ripart" në programin RGui dhe rezultatet mund të merren nga komandat "print cp" dhe "plotcp". Madhësia e cp, e cilat do të përfshihen në rezultatet, është përdorur për të përcaktuar një madhësi të përshtatshme për pemët ose për një krasitje sa më të mirë të pemës. Në këtë pjesë, do të përpiqemi për të minimizuar gabimin relativë të vlefshmërisë së kryqezuar të cilin e shënojmë“x gabim" nga një (5-foldefault) vlersim i kryqëzuar cp, ku xstd është gabimi standard i gabimit relativ, edhe përdorim "rregullin 1-SE", e cila e përdor vlerën më të madhe të cp ku cp është 0.05 me "x gabim" brenda një devijimi standard të minimumit. [Breiman1984].

Shkalla e klasifikimit të gabuar $P[Err \mid X]$ e një peme, duke trajtuar $P[Err \mid X]$ si një variabël të rastësishëm. Ne do të kemi dy faktorë të rastësishëm, së pari imputi i rastësishëm i atributeve të vektorit X dhe të gjitha probabiliteteve të panjohura P, klasave të probabiliteteve $P[C_j \mid A_t]$ dhe për nyjet pasardhëse $P[A_c \mid A_d]$ e cila është për çdo variabël të rastit Q.

$$E_x[Q] = \int Qf(X)dX$$
$$E_p[Q] = \int Qf(P)dP$$
$$E[Q] = E_x[E_p[Q]] = E_p[E_x[Q]]$$

Ku f(X) dhe f(P) janë aktualisht funksione të nyjeve pasardhëse dhe pa kushte. Devijimi nga standarti për nyjet më pasardhese llogaritet si më poshtë:

$$\sigma_p[Q] = \sqrt{E_p[Q^2] - E_p[Q]^2}$$

$$\sigma[Q] = \sqrt{E[Q^2] - E[Q]^2}$$

Nëse kushtet janë të përfshira, atëherë vendosim ti pranojmë kushtet. Për shembull nëse $E_{X|A_t}[Q] = \int Qf(X \mid A_t)dX$, ku P është kompozim i të gjitha probabiliteteve, duke përfshirë $P[C_j \mid A_t]$ për çdo nyje t. Rrjedhimisht e trajtojmë P si të pavarur nga A_t dhe llogaritim vlerat e pritshme të variablave të rastit brënda një nyje t. Vlera të cilat janë si më poshtë $E^t[Q] = E_p[E_{X|A_t}[Q]$, ku indeksi i sipërm të prezanton kushtin "$\mid A_t$" nëse Q ka të njëjtin indeks të sipërm t në $E^t[Q]$. Për këtë shkalla e gabimit të pritur për një nënpemë me nyje rrënjë t është të paraqitet nga $E[r_t]$ më mirë se $E^t[r_t]$ sepse vlera e pritur duhet që sigurisht të njehsohet me supozimin që A_t është e vërtetë. Sikurse dihet, në studimin e një baze te dhenash me anë të pemës klsifikuese dhe regresit kërkohet një bazë të dhënash me sa më shumë elemente dhe duke përdorur shpërndarjen e kësaj baze të dhënash në grupe sa më të vogla te nyjet përfundimtare ose të ashtuquajturat gjethe ndodhen pak grupe. Kjo e bën këtë proçes të pabesueshëm, rrjedhimisht është e arsyeshme që të bëjmë një kombinim të vlerave të pritura dhe shmangieve standarte që të vlerësojmë përqindjen e gabimit duke përdorur formulën e mëposhtme:

$$E[r^2] = E[r]^2 + \sigma[r]^2$$

$$ku$$

$$\sqrt{E[r^2]}$$

është përqindja e vlerësuar e gabimit në nyjen fundore. Në përgjithësi, përdorim vlësimin k-norm $\| r \|_k = \sqrt[k]{E[| r |^k]}$ i cili është i barabartë me $\sqrt[k]{E \mid r \mid^k}$ pasi e kemi konsideruar $r \geq 0$, qartësisht shihet se përqindja e gabimit në pemë është r= r_{rrenj}. Në teoremat dhe supozimet e mësipërme për një nyje vendimtare marrim,

$E[r_d^k] = \sum_{c \in d} E[r_c^k]P^*[A_c \mid A_d]$, për një nyje fundore T, përkufizojmë J si numër të klasave dhe duke përdorur supozimet e bëra në këtë material kemi që:

$$E[r_T^k] = E[(1 - P[C_{label(T)} \mid A_T, X])^k]$$

$$\approx E[(1 - P[C_{label(T)} \mid A_T])^k]$$

$$= \prod_{i=0}^{k-1} \frac{n_T - n_{j,label(T)} + (J-1)\lambda + i}{n_T + J\lambda + i}$$

$$\prod_{i=0}^{k-1} \frac{n_T - mak_j n_{j,T} + (J-1)\lambda + i}{n_T + J\lambda + i}$$

Duke përdorur k-norm në vlerësimin e përqindjes së gabimit dhe renditjen për të gjetur pemën optimale të krasitur në nyjen t, së pari gjejmë një pemë optimale e cila ndodhet poshtë

nyjes t dhe e kosiderojmë pemën e krasitur t me një vlerë më të ulët të përqindjes së gabimit në këtë k-norm. Më poshtë po shikojmë Algoritmin për këtë;

Për 2-norm përqindja e gabimit është $|| r ||_2 = \sqrt{E[r]^2 + \sigma[r]^2}$ e cila përfshin vlerat e pritura dhe devijimin standart.

Algorithm: R=Prune Tree(t)
Input: a tree rooted at node t
Output: the optimal pruned tree(modified from input), and its k-th moment error rate R(returned value of this function)

Compute $R_{leaf} = \prod_{m=0}^{k-1} \frac{n_t - n_{Label(t)} + \lambda(J-1) + m}{n_t + \lambda J + m}$;

If t is a decision node, then

Compute $R_{tree} = \sum_{c \in Children(t)} \frac{n_c + \eta}{n_t + \eta K_t} \Pr uneTree(c)$;

If $R_{tree} < R_{leaf} - \varepsilon (or \sqrt[k]{R_{tree}} < \sqrt[k]{R_{leaf} - \varepsilon}) then \mid return, R_{tree}$;

end
Replace the subtree rooted at t with leaf;
end
return R_{leaf};

3.7 Testet e pavarësisë

Testet për pavarësinë duke përdorur tabelat e kontigjencës përcaktojnë nëse ka një varësi të rëndësishme statistikore në mes të vlerave të dy variablave nominalë. Në problemin e mësipërm të krasitjes, të dy variablat janë (a) vlerat e klasës aktuale në të dhënat e krahasitjes së bazës të dhënave dhe (b) vlerat e klasës parashikuese të nënpemës. Kërkojmë të dimë nëse ka në të vërtetë një varësi të konsiderueshme të vlerave të vërteta të klasës dhe atyre të parashikuara, apo nëse është e mundur që korrelacioni i vrojtuar është i rastit. Një shembulli të veçantë i cili në një farë mënyre shkaktohet nga proçesi i krasitjes që zbatojmë.

n_{11}	n_{12}	...	n_{1J}	N_{1+}
n_{21}	n_{22}	...	n_{2J}	N_{2+}
...	...	...	...	...
n_{I1}	n_{I2}	...	n_{IJ}	N_{I+}
N_{+1}	N_{+2}	...	N_{+J}	N

Tabela 10: *Tabela e kontigjences*

Tabela 10, tregon një tabelë të kontigjencës. Rreshtat i dhe shtyllat j korrespondojnë me vlerat e dy variablave të cilat i konsiderojmë si nominalë. Çdo qelizë e tabelës përmban numrin e n_{ij} herë të kombinimeve përkatëse i vlerave që janë vrojtuar në N raste. Rrjeshtat dhe shtyllat përfundojnë në Ni + dhe N + j janë shumat e hyra në çdo rresht dhe shtylle respektive. Në vlerësimin e algoritmeve të klasifikimit, tabelat e kontigjencës krahasojnë vlerat parashikuese të matricës (të cilën e quajmë "confusion matix").

x	y	z	actual class	predicted class
0	0	1	A	A
0	1	1	B	B
1	1	0	B	B
1	0	0	B	B
1	1	1	A	B

Tabela 11: *Matrica e pemës përfundimtare*

Tabela 11 paraqet një matrice të tillë për një pemë të krasitur të cilën e quajmë pemë përfundimtare. Ajo përmbledh vlerat e vrojtuara dhe teorike në çdo klasë si dhe për të dhënat për proçesin e krasitjes në këto të dhëna të paraqitura në tabelën 10. Shuma e elementeve në kolona përfaqësojnë numrin e rasteve për pemët e krasitura për çdo klasë që arrijnë në nyjen përkatëse,në këtë rast për nyjen 1. Shuma e elementeve të cdo rreshti korespondon me numrin e rasteve për pemët e krasitura që do të caktohet për çdo klase përkatëse të nënpemëve që janë në këtë rast, pemët e plota që do të përdoren për klasifikimin. Një matricë e tille e bën të lehtë për të parë se sa shumë raste të krasitjes do të klasifikohen në mënyrë korrekte nga një nënpemë: numri i rasteve të pemëve të klasifikuara është sa shuma e elementeve të diagonales së matricës. Matrica e tille, është një lloj i veçantë i tabelës kontigjencës , janë baza e testeve statistikore të shqyrtuara në këtë pjesë. Hipotezat e teseve për dy variabla, të tilla si për vlerat e vrojtuara dhe teorike të çdo klase, nëse këto variabla janë të pavarura është quajtur "hipoteza bazë," dhe një test i rëndësishëm që përcakton nëse ka prova të mjaftueshme për të hedhur poshtë këtë hipotezë. Kur krasitja e pemës përfundimtare, duke hedhur poshtë hipotezën baze korrespondon me mbajtjen e një nënpeme në vend të krasitjes. Shkalla e shoqërimit mes dy variablave matet nga "nje test statistikor." Testi statistikor llogarit probabilitetin që e njëjta ose një vlerë më ekstreme e statistikës do të ndodhë rastësisht në qoftë se hipoteza zero është e saktë. Kjo sasi është quajtur "p-vlera" të testit. Në qoftë se p-vlera është e ulët, hipoteza zero mund të hidhet poshtë, që është shkalla e re e vartësisë, gjë e cila nuk ka gjasa të jetë për shkak të fatit. Zakonisht, kjo është bërë duke krahasuar p-vlerën me α=0,05, gjë e cila mund të bëjë të mundur të gjejmë informacion të mjaftueshëm për të hedhur poshtë hipotezën zero. Nëse α është të paktën po aq i madh sa vlera p. Një test statistikor i rëndësishëm mund të zbatohet për problemin e krasitjes duke njehsuar vlerën e p-së për vartësinë e vërejtur dhe krahasuar atë me vlerën α, mbajtjen ose hedhjen e nënpemës në përputhje me rrethanat. Dy vlerat si ajo α dhe p janë të rëndësishme kur të vlerësohet një test i rëndësishëm statistikor.

Fuqia e një testi rritet dhe është e provuar se kur baza e të dhënave rritet atëherë dhe vartësia e variablave ka shanse më të larta për të qënë e disponueshme. Fuqia e efektit është duke u testuar: Një vartësi e fortë është më shumë gjasa të jetë prezente se sa një vartësi e dobët.

Kjo mundet të arrihet duke mbledhur një sasi të mjaftueshme dhe më të madhe të të dhënave. Detyra e testimit të hipotezave shkencore ndryshon rrënjësisht nga problemi i modeleve të klasifikimit ne krasitje. Kur testet e rëndësishme përdoren për krasitje, qëllimi është për të maksimizuar saktësinë në të dhënat. Dy llojet e gabimit janë njësoj të rëndësishme. Problemi është për të gjetur ekuilibrin e duhur në mes të α dhe ß për të shmangur nënkrasitjen apo mbikrahasitjen. Bilanci i saktë varet nga tre faktorë të listuara më sipër. Kjo do të thotë se niveli optimal i testit statistikor varet ndër të tjera nga sasia e të dhënave në dispozicion për problemin në studim. Duke përdorur të njëjtën llogjikë do të kemi që një vlerë e fiksuar për testin statistikor nuk është gjithmon gjëja e saktë dhe e duhur që duhet të bëjmë.

Kjo çështje është e pavarur nga fakti i njohur se α gjithashtu duhet të rregullohet për teste të rëndësishme (Jensen & Schmill, 1997). Rregullimet për teste të shumta janë të nevojshme për shkak të një krasitje e cila zakonisht kërkon më shumë se një provë që do të kryhet dhe gjasat për zbulimin e një varësie, ka shanse të rritet kur numri i elementeve të përfshira në test rritet. Duket se nevoja e balancimit α dhe ß është anashkaluar vazhdimisht në qasjet e mëparshme që zbatohen teste rëndësie në algoritme të mësuarit. Për fat të keq ka pak shpresa për të gjetur një zgjidhje analitike për këtë problem, sepse ajo varet nga forca e efektit bazë, e cila është zakonisht e panjohur. Ka informacione në studime të ndryshme ku gjenden mënyra të tjera për zgjedhjen me optimale të vlerës së α dhe për këtë si bazë teorike përdoret vleresimi i kryqëzuar . Vihet re se në të gjithë këtë teori, supozojmë se e njëjta vlerë e alfës është e përshtatshme për çdo zonë të kësaj hapësire të shembullit që studiojmë. Është e besueshme që përmirësime të mëtejshme janë të mundshme për të rregulluar nivelin statistikor lokal për secilin shembull , për shkak se zona të ndryshme zakonisht përmbajnë sasi të ndryshme të të dhënave. Megjithatë, është shumë e vështirë për të zgjedhur një vlerë të alfës që të jetë vlera e duhur në mënyrë të pavarur për çdo zonë. Për më tepër, kjo zgjedhje është domosdoshmërisht e bazuar në të dhënat më pak informacion dhe për këtë arsye ka të ngjarë të jetë më pak e besueshme. Kështu kufizojmë vëmendjen tonë në qasjen globale. Testet statistikore janë të bazuara në shpërndarjen e testit statistikor i bazuar në hipotezën zero. Siç u përmend më lart, ato mund të ndahen në dy grupe: testet parametrike, të cilat mbështeten në supozimin se shpërndarja takon një klasë të veçantë të funksioneve parametrike, dhe teste jo-parametrike, të cilat nuk kërkojnë që në funksionin e shpërndarjes të ketë ndonjë formë të veçantë. Në seksionin pasues diskutohen testet parametrike bazuar në shpërndarjen Hi-katror, dhe më pas kemi paraqitur një grup të testeve jo-parametrike të njohur si "teste të përkëmbimeve."

3.8 Testet parametrike dhe joparametrike

Testet më të perdorura për pavarësi bazohen në tabelat e kontigjencës të cilat janë të bazuara në faktin se disa teste statistikore pothuajse ndjekin një shpërndarje Hi-katror me (I - 1) (J - 1) gradë të lirisë në qoftë se hipoteza zero është e saktë. I tillë është testi statistikor Hi-katror (Agresti, 1990) $\chi^2 = \sum_i \sum_j \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$, ku e_{ij} janë qelizat me vlerat e pritshme nën hipotezën bazë të llogaritura sipas $e_{ij} = N \hat{p}_i \hat{p}_j = N \frac{N_{i+}}{N} \frac{N_{+j}}{N} = \frac{N_{i+} + N_{+j}}{N}$, ku $\hat{p}_i$ është probabiliteti i vlerësuar pas një vrojtimi në rreshtin i, dhe $\hat{p}_j$ është probabiliteti korrespondues për shtyllen j. Për shkak se këto dy këto probabilitete janë të pavarur nën hipotezën bazë, prodhimi i tyre përbën mundësinë që një vrojtim do të jete në qelizën (i, j). Një alternativë për testin statistikor Hi-katror, e cila gjithashtu ka një shpërndarje Hi -katror, është "raporti log i likelihood " (Agresti, 1990).

$$G^2 = 2\sum_i \sum_j n_{ij} \log(n_{ij} / e_{ij})$$

Një disavantazh i testeve bazuar në shpërndarjen Hi-katror është se ato janë statistikisht të pavlefshme kur vëllimi i zgjedhjes është i vogël (Agresti, 1990).
Shpërndarja Hi-katror është një përafrim i shpërndarjes së vërtetë të testeve Statistikore nën hipotezën bazë, dhe ky përafrim është i saktë kur vëllimi i zgjedhjes është i madh. Për fat të

keq, nuk ka asnjë rregull të vetëm që mund të përdoret për të përcaktuar se kur përafrimi është i vlefshëm (Agresti, 1990, faqe 247). Në Cochran (1954), sugjerohet se një test i bazuar në statistiken χ 2 mund të përdoret në qoftë se asnjë nga vlerat e pritshme te qelizave është më i vogël se 1, dhe jo më shumë se 20% e tyre kanë pritshmeri te vlerave nën 20 Agresti (1990, faqe 247). Testi Hi-katror përafërsisht ka tendencë të jetë i dobët për tabelat e kontigjencës për të dy rastet për vëllimin e vogël dhe për ato të cilat janë me të vërtet shumë të mëdha. "Megjithatë eshte provuar në mënyrë empirike se testi χ 2 për rastet me vellim të vogël punon me mire se testi G^2 (Agresti, 1990, faqe 246). Në vartësi të qelizave që presim të numërojmë, duke përdorur shpërndarjen Hi-katror në raport me G^2 mund të rezultojë si një test që është ose shumë konservative ose shumë liberal (Agresti, 1990, faqe 247). Një test që është shumë konservator prodhon p-vlera që janë shumë të mëdha, ndërsa ai që është shumë liberal prodhon p-vlera që janë shumë të vogla.

b. Testet jo-parametrike

Testet jo-parametrike kanë avantazhin se ato nuk bëjnë supozime në lidhje me shpërndarjen e statistikës që duhet të provojme. Testet e përkëmbimeve janë një klasë e testeve jo-parametrike që llogaritin shpërndarjen statistikore që duhet të provojmë nën hipotezën bazë e cila shprehet në mënyrë eksplicite, duke numëruar të gjitha permutacionet e mundshme të të dhënave në bazën e të dhënave. Më i njohuri i këtij grupi të testeve është testi i Fisherit, i tabelave të kontigjencës (Agresti, 1990). Ndryshe nga testet parametrike duke përdorur shpërndarjen Hi-katror, testet me përkëmbime janë statistikisht të vlefshme në situata ku vëllimi është i vogël (Mira, 1994). Ato janë të bazuara në faktin se, në bazë të hipotezës bazë, të gjitha permutatacionet e mundshme të të dhënave kanë shanse të barabarta të ndodhin. Vlera e p-së e një testi të përkëmbimeve është një rrjedhim i këtyre përkëmbimeve për të cilat testi statististikor ka një vlerë në mënyrë të barabartë ose më ekstreme se sa për të dhënat origjinale (Good, 1994). Në rastin e problemit të klasifikimit, përkëmbimi i të dhënave përkon me përkëmbimet e klasave të etiketuara të të gjtha rasteve (Jensen, 1992). Çdo përkëmbim mund të shkruhet si një tabelë e kontigjencës duke i çiftëzuar të gjitha vlerat e klasave të parashikuara.Tre tabelat në Tabela 12 ndajnë të njëjtat vlera anësore. Duke përkëmbyer klasat e etiketuara nuk ndryshojnë numrin e rasteve që i përkasin çdo klase, gjithashtu as nuk ndryshojnë numrin e rasteve të caktuara për çdo klasë nga klasifikuesi.

2	0	2	1	1	2	0	2	2
0	2	2	1	1	2	2	0	2
2	2	4	2	2	4	2	2	4
p-value=1/3			p-value=1			p-value=1/3		

Tabela 12: *Tabelat e disa përkembimeve*

Në terma statistikorë, testet e përkëmbimit në tabelat e kontigjencës nxjerrin një vlerë p e cila është e kushtëzuar me totalin e dhënë në vlerat anësore të tabelave. Tabelat e kontigjencës në mënyrë identike rezultojnë me të njëjtën vlerë për testet statistikore. Kështu, vlera p e një testi të përkëmbimit mund të llogaritet duke mbledhur probabilitet e të gjitha tabelave të paparashikuara me një vlerë të barabartë ose me më shumë ekstremitet për testet statistikore. Probabiliteti i një tabele kontigjence të paparashikuar si p_f është ekuivalent me raportin e testit të përkëmbimit që merret në mënyrë të rastësishme. Ai mund të shkruhet në këtë formë:

$$p_f = \frac{\prod_i n_{i+}! \prod_j n_{j+}!}{n! \prod_i \prod_j n_{ij}!}$$

Ky funksion është i njohur si shpërndarje e shumëfishtë hipergjeometrik (Agresti, 1990). Nëse s_f është vlera e testit statistikor në tabelën kontigjencës f për bazën e të dhënave fillestare dhe s_0 vlera e saj për të dhënat origjinale, atëhere vlera p mund të shkruhet si më poshtë: $p = \sum I(s_f \le s_0) p_f$ ku I () është funksioni tregues dhe shuma është mbi të gjitha tabelat kontigjencës për të njëjtën anë. Për fat të keq, të dyja metodat e llogaritjes së saktë të p-vleres nuk janë shumë të sakta në rastin kur studiojmë një numër të vogël elementesh. Për disa statistika në të cilat duhet të kryhen teste të sofistikuara dhe kemi të bëjmë me modele ku numri i elementeve është i vogël janë zhvilluar dhe përdoren algoritma të tjera të cilat ndryshe i quajmë "network algorithem". Këto modele japin një vlerë të saktë të p-vlerës (Good, 1994). Ata bëjnë të mundur përdorimin e vetive matematikore për testet statistikore, duke shkurtuar hapësirën. Megjithatë, edhe këto algoritme të sofistikuara janë ende me njëhsime shumë të shtrenjta dhe të aplikueshme vetëm në qoftë se vellimi i zgjedhjes është i vogël.

3.9 Testet Statistikore

Dy statistikat të mundshme tashmë janë diskutuar në kontekstin e testeve parametrike: χ^2, dhe, G^2. Të dyja mund të përdoren për të realizuar testin e përkëmbimit (Good, 1994). Niveli i testit statistikor është i thjeshtë dhe është si pjesë e përkëmbimeve të rastit, për të cilat vlera e statistikës është të paktën po aq e madh sa për të dhënat origjinale, sepse të dy statistikat rriten në mënyrë monotone duke qënë të lidhura me shkallët e lirisëqë është i pranishëm në nje bazë te dhenash. Shpërndarja Hi-katror, e cila është një bazë për testet parametrike e diskutuar më parë, është në fakt vetëm një përafrim për përkëmbimin e shpërndarjeve të dy statistikave dhe sikurse u tha më sipër, ky përafrim është i pabesueshëm për rastet kur numri i elementeve të shëmbullit është i vogël (Agresti, 1990). Duhet të theksojmë se megjithëse probabiliteti i raportit të testit është vetëm një përafrim i vlerës p(p-vlerës)si një provë e saktë ku është e garantuar se kjo shërben për të përafruar nga afër vlerën e vërtetë të p-vlerës, por ky rast nuk është test që bazohet në shpërndarjen Hi-katror. Një tjetër test statistikor potencial i cili është përmendur më lart, edhe pse nuk ka luajtur ndonjë rol në testet statistikore. Probabiliteti p_f i një tabele të kontigjencës të paparashikuara në hipotezën zero e dhënë nga shpërndarjet e shumëfishta hipergjeometrike është një alternativë për χ^2, ose, G^2 (Good, 1994). Rrallë herë tabelat kontigjencës të cilat kanë një p të vogël të p_f, tregojnë një lidhje të fortë në mes të dy variablave të përfshira. Niveli i testit statistikor është një raport i përkëmbimeve të rastësishme për sejcilën p_f të cilat nuk janë më të mëdha me bazën fillestare të dhënave sepse sa më e madhe është lidhja aq më i vogël është probabiliteti.

Kur të dy variablat në tabelën e paparashikuar janë binare, ky ndryshim i testit është i njohur si versioni me dy anë të testit të saktë te Fisherit (Agresti, 1990). Në rastin e përgjithshëm është nganjëherë i quajtur testi i Freeman dhe Halton (Good, 1994). Të gjitha testet e përkëmbimeve kanë disavantazhin se shpërndarja e p-vlera është shumë e rrallë, kur vellimi i bazës së të dhënave është jashtëzakonisht i vogël (Agresti, 1990). Kjo është për

shkak të numrit të vogël të tabelave të kontigjences që janë të mundshme, domethënë kur ka shumë pak raste.
Për të marrë një pemë përfundimtare është e domosdoshme që të kemi një bazë të dhënash të cilën e përdorim për trajnim. Në bazën e të dhënave që përdorim si të vlefshme për të realizuar qëllimin tonë duhet të bëjmë një krasitje para se të arrijmë në modelin përfundimtar.

Qëllimi kryesor i përdorimit të kësaj baze të dhënash është:
Të zhvillojmë rregullat dhe të shënojmë regjistrimet e çdo nyje ose përkufizimet e çdo nyje.
Të njehsojmë probabilitetet e pasme (raportin e rasteve ose rekordeve në çdo nivel të targetit) për secilën nyje. Duke shënuar nivelin e targetit të çdo nyje.
Baza e të dhënave e vlefshme përdoret për të krasitur një pemë duke selektuar përmasën e saktë të kësaj peme ose për të gjetur nënpemen optimale. Zakonisht pema fillestare që ndërtojmë është shumë e madhe. Këtë pemë zakonisht e quajmë pemë maksimale. Duke hequr disa degë të kësaj pemë maksimale krijojmë pemë më të vogla dhe po ashtu në vartësi nga numri i degëve që heqim mund të krijojmë pemë të ndryshme. Natyrisht pema më e vogël ka vetëm një nyje fundore apo gjethe, e cila gjithashtu është dhe nyje rrënjë. Pema më madhe natyrisht që ka shumë nyje fundore. Prerja e degëve të ndryshme na jep nënpemë të ndryshme, ku duhet të selektojmë një nga ato dhe të dhënat e vlefëshme do të shërbejë për të zgjedhur atë më të mirën. Vlefshmëria e çdo nënpeme me përmasa të ndryshme njehsohet duke përdorur të dhënat e vlefshme dhe natyrisht një nga cilësitë që përdoret është fitimi për rastin tonë. Sikurse dihet fitimi njehsohet duke perdorur regjistrimet e të dhënave të vlefshme që përdorim. Fitimi i vlefshëm apo i sanksionuar në këtë rast do të përdoret për të zgjedhur pemën optimale. Përmasa e një peme përcaktohet nga numri nyjeve fundore që ka pema. Një pemë do të konsiderohet optimale në këtë rast nëse jep fitim më të lartë se çdo pemë dhe se nga përmasat do të kosiderohet si më e vogla.

3.10 Matja e vlefshmërisë së një shpërndarjeje

Metoda e shpërndarjes së nyjeve bëhet duke përdorur algoritme të ndryshme të cilat do të diskutohen në këtë material. Nëse qellimi ynë është një variabël nominal, me dy variabël me dy vlera përgjegjës dhe jo përgjegjës, por ka dhe raste kur variabli nominal ka dhe më shumë se dy mundësi si p.sh ngjyra e cila mund të jetë e bardhë, e verdhë, e kuqe, etj. Nëse variabli është kategorik dhe ka cilësinë që e vendosim në një renditje të caktuar i quajmë variabla ordinal.

Një i tillë është rreziku i cili mund të kosiderohet i lartë ose i ulët. Nëse qëllimi ynë është një variabël ordinal, atëherë metodat që do të përdorim për të bërë shpërndarjen janë Entropy dhe Gini. Kur qëllimi është një variabël i vazhdueshëm atëherë për të përcaktuar vlefshmërinë e shpërndarjes përdoret, testi Fisher. Për të përcaktuar vefshmërinë e shpërndarjes përdorim reduktimin e variancës.

3.11 Kontrolli i rritjes së pemës realizohet nëpërmjet:

a. Vetisë së përshtatjes së shpërndarjes nepermjet:

Nëse përdorim vetinë e përshtatjes së shpërndarjes për pemën përfundimtare, atëherë vlerat e p-së janë përshtatur për një numër të caktuar të shpërndarjes së nyjeve për nivele të mëparshme dhe në veçanti nëse niveli i α është i specifikuar në bazë të vetisë se nivelit të rëndësisë së α, atëherë çdo shpërndarje që ka një vlerë të p-së mbi këtë vlerë duhet ta refuzojmë ose ta pranojëmë.

b. Vetisë së nyjeve fundore

Mund të kontrollojmë rritjen e pemës duke fiksuar vetinë e numrit të nyjeve fundore, për shembull nëse fiksojmë 100 nyje fundore dhe nëse rezultati i shpërndarjes në një situatë të vecantë me më pak se 100 regjistrime, atëherë shpërndarja e mëtejshme duhet të mos bëhet duke supozuar se rritja e pemës është ndaluar në këtë nyje të caktuar.

c. Vetisë së përmasës së shpërndarjes

Nëse vlera e p-së të vëllimit të shpërndarjes është e fiksuar për shembull në një numër 300 rregjistrime dhe nëse një nyje ka më pak se 300 rregjistrime, atëherë nuk duhet të marrim në konsideratë shpërndarje të tjera. Vlera e parazgjedhur e kësaj vetie duhet të jetë sa dyfishi i nyjeve fundore ose e gjetheve, e cila specifikohet nga vetia e nyjeve fundore.

Teoremë 3.6: Nëse T_1, dhe, T_2 janë nënpemët e krasitura të pemës T. dhe T_2 është një nënpemë e krasitur e pemës T_1 atëherë dhe vetëm atëherë, kur çdo nyje jo përfundimtare T_2 është gjithashtu nyje jo përfundimtare dhe për T_1.

Vertetim:

Nëse $\#(T)$ përkohësisht jep numrin e nënpemëve të krasitura të pemës T. Nëse T është e vogël dhe e parëndësishme, atëherë $\#(T)=1$. Në rast të kundërt $\#(T)=\#(T_L)\#(T_R)+1$. Shihet qartë nga barazimi i mësipërm që maksimumi i nënpemëve të krasitura për një pemë që ka m nyje fundore rritet në mënyrë të shpejtë me m. Nënjë rast të veçant le të marrim T_n si njëpemë ku nyjet fundore të së cilës kanë egzaktësisht n paraardhës, kështu që $|\tilde{T}|=2^n$. Duke ndjekur barazimin e mësipërm kemi $\#(T_{n+1})=(\#(T_n))^2+1$. Si rezultat i kësaj $\#(T_1)=2, \#(T_2)=5, \#(T_3)=26, \#(T_4)=677$ dhe kështu me radhë.

Tani $(\#(T))^{\frac{1}{|\tilde{T}_n|}}$ është e lehtë të shihet se kur rritet n-ja kjo konvergjon në një numër $b=1.5028368$, gjithashtu është e lehtë të gjendet se nga zgjidhja e barazimit $b=(\#(T))^{\frac{1}{|\tilde{T}_n|}}$ kemi $\#(T)=[b^{|\tilde{T}_n|}]$ për çdo n>1.

Le të fiksojmë një pemë të çfardoshme T_0, konkretisht mund të marrim pemën maksimale $T_{\max}$ dhe le të kemi R(t), ku $t \in T_0$ dhe një numër real i fiksuar α. Për një numër të dhënë α marrim $R_\alpha(t)=R(t)+\alpha, per, t \in T_0$. Për një nënpemë të dhënë T nga T_0, bashkësia $R(T)=\sum_T R(t)$ dhe $R_\alpha(T)=\sum_T R_\alpha(t)=R(T)+\alpha|\tilde{T}|$. Nëse T nuk është pemë e rëndësishme me rrënjë t_1, kur $R(T)=R(t_1), dhe, R_\alpha(T)=R_\alpha(t_1)$.

Një nënpemë e krasitur $T_{1,}$ e Tse do të quhet një nënpemë optimale në respekt të α edhe nëse kemi: $R_\alpha(T_1)=\min_{T' \prec T} R_\alpha(T')$.

Sikurse dihet kemi një numër të caktuar të nënpemëve te krasitura nga pema T, dhe natyrisht që midis këtyre nënpemëve të krasitura ndodhet dhe pema optimale e cila nuk është e vetme.

Le te shënojmë me T_1 një nga nënpemët e krasitura të pemës T e cila do të konsiderohet si nënpema optimale më vogla nga pema T, nëse $T' > T$ për çdo pemë optimale T' nga pema T. Atje ndodhet e shumta një nënpemëe krasitur e cila konsiderohet si më vogla nga T në respekt të alfës dhe kur ajo ekziston është dhënë nga $T(\alpha)$.

Le të jetë T' një nënpemë e vogël dhe e parëndësishme nga T dhe T'_L, dhe, T'_R dy degët kryesore të saj. Atëherë kemi

$$R_\alpha(T') = R_\alpha(T'_L) + R_\alpha(T'_R)$$

Ky fakt na jep mundësinë të vertetojmë teoremën 3.1 duke përdorur induksionin matematik.

Teoremë 3.7: Çdo pemë T ka një nënpemë e cila është unike dhe konsiderohet si nënpema optimale më e vogël të cilën e shënojmë $T(\alpha)$. Le të kemi një pemë T, jo të rëndësishme e cila ka si rrënjë t_1 dhe si degë kryesore T_L, dhe, T_R,

atëherë $R_\alpha(T(\alpha)) = \min[R_\alpha(t_1), R_\alpha(T_L(\alpha) + R_\alpha(T_R(\alpha))]$ nëse

$R_\alpha(t_1) \le R_\alpha(T_L(\alpha) + R_\alpha(T_R(\alpha)), atehere, T(\alpha) = \{t_1\}, ndryshe, T(\alpha) = \{t_1\} \cup T_L(\alpha) \cup T_R(\alpha)$

dhe rezultati tjetër vjen apriori nga vetia transitive e shenjës <.

Teormë 3.8: Nëse $T(\alpha) < T' < T, athere, T(\alpha) = T'(\alpha)$ rritja e alfës çon në rritjen e ndërshkimit për pemën komplekse dhe me sa duket në të njëjtën kohë në $T(\alpha)$ më të vogël. Vlefshmëria e këtij rezultati varet maksimalisht nga struktura e bashkësisë të nënpemëve të krasitura.

Teoremë 3.9: 1. Nëse $\alpha_2 \ge \alpha_1, atëhere, T(\alpha_2) < T(\alpha_1)$

2. Nëse: $\alpha_2 > \alpha_1, dhe, T(\alpha_2) < T(\alpha_1), atëhere$

$$\alpha_1 < \frac{R(T(\alpha_2)) - R(T(\alpha_1))}{|T(\alpha_1)| - |\tilde{T}(\alpha_2)|} \le \alpha_2$$

Vërtetim: Qartësisht shihet se nëse $T(\alpha_1)$ është e parëndësishme, atëherë edhe $T(\alpha_2)$ është e parëndësishme për $\alpha_2 \ge \alpha_1$, (1) është vertetuar në teoremën 3.7 me induksion matematik, kështu që nëse

$\alpha_2 \ge \alpha_1, dhe, T(\alpha_2) < T(\alpha_1) \Rightarrow R(T(\alpha_2) + \alpha_2 |\tilde{T}(\alpha_2)| \le R(T(\alpha_1) + \alpha_2 |\tilde{T}(\alpha_1)|$

dhe

$R(T(\alpha_1) + \alpha_1 |\tilde{T}(\alpha_1)| < R(T(\alpha_2) + \alpha_1 |\tilde{T}(\alpha_2)|$

Dhe nga të dy mosbarazimet e mësipërme arrijmë në përfundimin e pikës së dytë të teoremës.

Teoremë 3.10: Në qoftë se $R_\alpha(t) \ge R(T_t)$ për të gjitha $t \in T - \tilde{T}$, atëherë $R(T(\alpha)) = R_\alpha(T), dhe, T(\alpha) = \{t \in T : R_\alpha(s) > R_\alpha(T_s)$ për të gjithë pasardhësit s të t-së}

Vertetim: Kjo teoremë vertetohet me metodën e induksionit matematik.

Për $\tilde{T} = 1$ është e vertetë.

Supozojmë se është e vërtetë për të gjitha pemët që kanë më pak se n nyje fundore, ku $n \ge 2$.

Le të kemi T një pemë e cila ka n nyje fundore, me nyje rrënjë t_1 dhe me degë primare

T_L, dhe, T_R. Nga hipoteza kemi që $T_L(\alpha) = \{t \in T_L : R_\alpha(s) > R_\alpha(T_s)$ për të gjithë pasardhësit $s \in T_L, te, t\}$ dhe
$T_R(\alpha) = \{t \in T_R : R_\alpha(s) > R_\alpha(T_s)$ për të gjithë pasardhësit $s \in T_R, te, t\}$, gjithashtu
$R_\alpha(T_L(\alpha) = R_\alpha(T_L)$ dhe $R_\alpha(T_R(\alpha) = R_\alpha(T_R)$.
Si rezultat kemi : $R_\alpha(T) = R_\alpha(T_L) + R_\alpha(T_R) = R_\alpha(T_L(\alpha)) + R_\alpha(T_R(\alpha))$.
Kështu duke ndjekur teoremën 3.7 nëse $R_\alpha(t_1) = R_\alpha(T)$, atëherë $T(\alpha) = \{t_1\}$ dhe nëse
$R_\alpha(t_1) > R_\alpha(T)$, atëhere
$T(\alpha) = \{t_1\} \cup T_L(\alpha) \cup T_R(\alpha), dhe, R_\alpha(T(\alpha)) = R_\alpha(T_L(\alpha)) + R_\alpha(T_R(\alpha))$
pra, në të gjitha rastet përfundimi i teoremës është arritur.

Nëse është dhënë një pemë jo e rëndësishme T, le të marrim
$g(t,T) = \dfrac{R(t) - R(T_t)}{|\tilde{T}_t| - 1}, per, t \in T - \tilde{T}$, lehtësisht shihet se për çdo $t \in T - \tilde{T}$ dhe për α numër real atëhere janë të vërteta:

$g(t,T) \geq \alpha$, kusht i nevojshëm dhe i mjaftueshëm është që $R_\alpha(t) \geq R_\alpha(T_t)$.
$g(t,T) > \alpha$, kusht i nevojshëm dhe i mjaftueshëm është që $R_\alpha(t) > R_\alpha(T_t)$.

Teoremë 3.11:
Nëse është dhënë një pemë e rëndësishme T, dhe marrim $\alpha = \min_{t \in T - \tilde{T}} g(t,T)$, atëherë T është e vetme nënpema optimale e krasitur të cilën e shënojmë $\alpha, per, \alpha < \alpha_1$; T është një nënpemë optimale e krasitur në respekt të α_1, por jo më e vogla; dhe T nuk është pema optimale e krasitur në vetveten e sajë në respekt të α për $\alpha > \alpha_1$.
Marrim $T_1 = T(\alpha_1)$.
Atëherë $T_1 = \{t \in T : g(t,T) > \alpha_1$ për të gjitha nyjet fundore s të t-së}.
(3.3.3)
Le të kemi $t \in T - \tilde{T}$, atëherë $g(t,T_1) > g(t,T), nese, T_{1t} < T_t$ (3.4.3)dhe g(t,T) ndryshe.
Vertetim: Kjo teoremë ndjek në mënyrë të menjëherëshme teoremën 3.10 pasi T është një nënpemë e krasitur unike në vetveten e saj në respekt të $\alpha, per, \alpha < \alpha_1$, po kështu T është një nënpemë optimale e krasitur në respekt të α_1, por kjo nuk është më e vogla, dhe kjo përmbahet në mosbarazimin e mësipërm (3.3.3).
Në veçanti $R(t_1) + \alpha_1 |\tilde{T}_1| = R(T) + \alpha_1 |\tilde{T}|$, por
$|\tilde{T}_1| < |\tilde{T}|$, le të kemi $\alpha > \alpha_1$, atëherë $R(T_1) - R(T) = \alpha_1(|\tilde{T}| - |\tilde{T}_1^{!}|) < \alpha(|\tilde{T}| - |\tilde{T}_1|)$ nga ku rrjedh që $R_\alpha(T(\alpha)) \leq R_\alpha(T_1) < R_\alpha(T)$.
Rrjedhimisht, T nuk është një nënpemë optimale e krasitur e cila në vetveten e saj është shënuar α. Le të kemi $t \in T - \tilde{T}$, nëse $T_{1t} = T_t$ atëherë $g(t,T_1) g(t,T)$ nga përkufizimi. Tani supozojmë se $T_{1t} < T_t$ nga teorema 3.11, T_{1t} është nënpema optimale e krasitur nga T_t e shënuar si α_1, nëse marrim $\alpha_2 > \alpha_1$ atëherë {t} është nënpema optimale e krasitur në respekt të α_2. Meqënëse T_t është një nënpemë optimale dhe unike e cila nga ana e vetë është në

respekt të α, që për $\alpha < \alpha_1$ atëherë në bazë të teoremës 3.8 kemi që: $\frac{R(t)-R(T_{1t})}{|\tilde{T}_{1t}|-1} > \alpha_1 \geq \frac{R(T_{1t})-R(T_t)}{|\tilde{T}_t|-|\tilde{T}_{1t}|}$,

Rrjedhimisht:

$$R(t)-R(T_t)=R(t)-R(T_{1t})+R(T_{1t})-R(T_t)<(R(t)-R(T_t))\left(1+\frac{|\tilde{T}_t|-|T_{1t}|}{|\tilde{T}_{1t}|-1}\right)=(R(t)-R(T_{1t}))\left(\frac{|\tilde{T}_t|-1}{|\tilde{T}_{1t}|-1}\right)$$

Kështu që $g(t,T_1)=\frac{R(t)-R(T_{1t})}{|\tilde{T}_{1t}|-1} > \frac{R(t)-R(T_t)}{|\tilde{T}_t|-1}=g(t,T)$

Nga teorema 3.10 arrihet në vertetimin e plotë të teoremës. Le të quajmë T_0 një pemë jo të rëndësishme. Marrim $\alpha_1=\min_{t\in T_0-\tilde{T}_0} g(t,T)$ dhe $T_1=\{t\in T_0 : g(s,T_0)>\alpha_1$, për të gjithë parardhësit s nga t}(3.10), kur $T_1 < T_0$ në bazë të teoremës 3.11,

$T_0(\alpha)=T_0, per, \alpha<\alpha_1, dhe, T_0(\alpha_1)=T_1$. Nëse T_1 është një pemë jo e rëndësishme, atëherë $T_0(\alpha)=T_1, per, \alpha\geq\alpha_1$ sipas teoremës 3.10, Supozojmë që në të vërtetë që T_1 është një pemë jo e rëndësishme dhe marrim $\alpha_2=\min_{t\in T_1-\tilde{T}_1} g(t,T_1)$ dhe $T_2=\{t\in T_1 : g(s,T_1)>\alpha_2$, për të gjithë paraardhësit s nga t}, atëherë $T_2 < T_1, dhe, \alpha_2>\alpha_1$ e cila rrjedh në bazë të 3.10. Nga teorema 3.10 rrjedh që $T_1(\alpha)=T_1, per, \alpha<\alpha_2, dhe, T_1(\alpha_2)=T_2$.

Nëse $\alpha_1\leq\alpha<\alpha_2, atëhere, T_0(\alpha)<T_0(\alpha_1)=T_1<T_0$ gjë e cila rrjedh në bazë të teoremës 3.9 dhe në bazë të teoremës 3.8 marrim $T_0(\alpha)=T_1(\alpha)=T_1$ dhe në mënyrë të ngjashme marrim $T_0(\alpha_2)<T_0(\alpha_1)=T_1<T_0$, nga e cila rrjedh se. Në se pema T_2 është e parëndësishme atëherë $T_0(\alpha)=T_2, per, \alpha\geq\alpha_2$. Ndryshe proçesi i proçedimit mund të përsëritet disa herë. Në të vërtetë ky proçes mund të përsëritet aq herë sa një pemë e vogël është arritur. Kështu që atje është një numr i plotë pozitive k dhe një numër real $\alpha_k, 1\leq k\leq K$ dhe pemët $T_k, \alpha_k, 1\leq k\leq K$ të tilla që:

$$-\infty<\alpha_1,\alpha_2<.....\alpha_k<\infty;$$
$$T_0>T_1,.....T_k=\{rrenjen(T_0)\};$$
$$\alpha_{k+1}=\min_{t\in T_k-\tilde{T}_k} : g(t,T_k), 0\leq k<K :$$

$$T_{k+1}=\{t_{0\leq k<K}\in T_k : g(s,T_k)>\alpha_{k+1}$$

për të gjithë paraardhësit s nga t dhe

$$T_0(\alpha)=\begin{Bmatrix} T_0, \alpha<\alpha_1, \\ T_k, 1\leq k<K, dhe, \alpha_k\leq\alpha<\alpha_{k+1}, \\ T_K, \alpha\geq\alpha_K \end{Bmatrix} \quad (3.5.3)$$

Në bazë të përkufizimit kemi $g(t,T_k)=\frac{R(t)-R(T_{kt})}{|\tilde{T}_{kt}|-1}, t\in T_k-\tilde{T}_k$ (3.6.3)

Formulat e mësipërme së bashku na çojnë në një algoritëm për të përcaktuar K, α_k, dhe, T_k duke e marrë të mirëqënë $T_0(\alpha), ku, -\infty<\alpha<\infty$. Le të kemi $0\leq k<K$, atëherë T_k është nënpema optimale e krasitur e shënuar si $\alpha_{k+1}, por, T_k(\alpha_{k+1})=T_{k+1}$ dhe në

veçanti $R_{\alpha_{k+1}}(T_k) = R_{\alpha_{k+1}}(T_{k+1})$ dhe në bazë të disa veprimeve të thjeshta algjebrike marrim barazimin e mëposhtëm:

$$\alpha_{k+1} = \frac{R(T_{k+1}) - R(T_k)}{|\tilde{T}_k| - |\tilde{T}_{k+1}|}, 0 \le k < K \qquad (3.7.3)$$

Teoremë 3.12. Le të kemi $g_k(t), ku, t \in T_0 - \tilde{T}_0, dhe, 0 \le k \le K-1,$ I përkufizuar në mënyrë rekursive si më poshtë:

$$g_0(t) = g(t, T_0), ku, t \in T_0 - \tilde{T}_0, dhe, per, 1 \le k \le K-1,$$

$$g_k(t) = \begin{Bmatrix} g(t, T_k), t \in T_k - \tilde{T}_k \\ g_{k-1}(t), ndryshe \end{Bmatrix} \qquad (3.8.3)$$

Për ndryshe për $-\infty < \alpha < \infty,$

$T_0(\alpha) = \{t \in T_0 : g_{k-1}(s) > \alpha$

për të gjithë paraardhësit e s nga t} (3.9.3)

Vertetim: Për këtë është e mjaftueshme të tregojmë se nëse $0 \le k \le K-1, dhe, \alpha \le \alpha_{k+1},$ atëherë $T_0(\alpha) = \{t \in T_0 : g_k(s) > \alpha,$ për të gjithë paraardhësit s nga t}

Nga 3:11 dimë që $\alpha \le \alpha_k$; pasi $T_0(\alpha_k)$ është një pemë e vogël e pa rëndësishme dhe $T_0(\alpha) < T_0(\alpha_k), per, \alpha \ge \alpha_k$ në bazë të teoremës 3.9 dhe 3:10 për çdo $-\infty < \alpha < \infty$, duke pasur parasysh teoremën 3.10 dhe 3.11 është vlefshme për k=0 dhe $\alpha < \alpha_1$.

Dhe tani supozojmë se $1 \le k \le K-1$ dhe për

$T_0(\alpha) = \{t \in T_0 : g_{k-1}(s) > \alpha$

$\alpha \le \alpha_k$ për të gjithë paraardhësit nga t} (3.10.3)

Atëherë në veçanti, $T_k = T_0(a_k) = \{t \in T_0 : g_{k-1}(s) > a\}$ për të gjithë paraardhësit s të t-së }

dhe rrjedhimisht $g_{k-1}(s) > a_k, s \in T_k - \tilde{T}_k$. tani $T_k - T_k \subset T_{k-1} - \tilde{T}_{k-1},$ kjo sipas (3.10) dhe

$g_k(s) = \{g(s, T_k) \ge g_{k-1}(s), per\ s \epsilon T_k - \tilde{T}_k,\ g_{k-1}(s) ndryshe;\}$

gjë e cila rrjedh nga (3.11) që për $s \in T_0 - \tilde{T}_0$ dhe $a \le a_k, g_k(s) > a$ si kusht i nevojshëm dhe i mjaftueshëm, $g_{k-1}(s) > a.$ Kështu që nga (3.9.3) rrjedh 3.10.3 për $a < a_k$. Dhe tani suppozojmë se $a_k < a < a_{k+1}$. Atëherë

$T_0(a) = T_k(a) = \{t \epsilon T_k; g(s, T_k) > a$ *për të gjithë paraardhësit e s dhe* $t -$ *së*$\}$

gjë e cila rrjedh nga (3.8.3) që:

$T_0(a) = \{t \epsilon T: g(s) > a$ *për të gjithë paraardhësit e s dhe* $t -$ *së*$\}$

Supozojmë se $t \notin T_k$. Nga Nga (3.10.3) atje është një tjetër paraardhës s i t-së i tillë që $g_{k-1}(s) \le a_k; s \notin T_k - \tilde{T}_k$ nga(7.27), po kështu $g_k(s) = g_{k-1}(s) \le a_k < a$ nga mesiper. Gjë e

cila rrjedh nga (3.9.3) po kështu rrjedh që $a_k < a \le a_{k+1}$ dhe përfundimisht rrjedh $a \le a_{k+1}$. Dhe me induksion, provohet dhe tregohet fusha e ndryshimit të k.

Teorem 3.13: Le të kemi $t \in \tilde{T}_0$ ku $-\infty < a < \infty$. If $g_{k-1}(s) > a$ për të gjihë parardhësit s të t-së, atëherë $t \in T_0(a)$. Ndryshe, nëse s është nyja eparë nga T_0 e vetme në rrugën e saj nga rrënja e T_o deri te t për të cilën $g_{k-1}(s) \le a$. Atëherë kjo është një para ardhës unik i t-së në $\tilde{T}_0(a)$.

Le ta risjellim këtë dhe nëse s është një pasardhës nga t, atëherë l(s,t) është gjatësia e rrugicës nga s në t. Kur është dhënë $-\infty < \alpha < \infty$ dhe $t \in T_0$, marrim $S_\alpha(t) = \min[\, R(s) - \alpha(l(s,t)) : s = t, ose$ s është një paraardhës i t-së].

Teoremë 3.14: Supozojmë se $R(t) \ge 0$ për të gjitha $t \in T_0$, atëherë $S_\alpha(t) > 0, per, -\infty < \alpha < \infty$ dhe për çdo nyje jo fundore t nga $T_0(\alpha)$.

Vertetim: Le të kemi t një nyje jo fundore nga $T_0(\alpha)$ dhe le të kemi s një para ardhës i t-së. Tani $T_0(\alpha)$. është një nënpemë optimale dhe unike e krasitur e cila në vetvete është në respekt të α kështu që: $\dfrac{R(s) - R(T_{0s}(\alpha))}{|\tilde{T}_{0s}| - 1} > \alpha$, në bazë të teoremës 3.10 dhe duke marrë $R(s) > \alpha(|\tilde{T}_{0s}| - 1)$, është lehtësisht e dukshme se duke përdorur metoden e induksionit matematik që $|\tilde{T}_{0s}(\alpha)| \ge l(s,t) + 2$ dhe rrjedhimisht që $R(s) > \alpha(l(s,t) + 1)$.

Teoremë 3.15: Supozojmë se $R(t) \ge 0$ për të gjitha $t \in T_0$, dhe për një numër real α marim: $T_{suff}(\alpha) = \{u \in T_0 : S_\alpha(t) > 0$ për të gjithë para ardhësit t të u-së}, atëherë $T_0(\alpha) < T_{suff}(\alpha) < T_0$.

Vertetim: Qartësisht shihet se $T_{suff}(\alpha)$ është një nënpemë e T_0 dhe ajo përmban rrënjën e T_0, kështu që $T_{suff}(\alpha) < T_0$, gjithashtu le të jetë v një nyje jo fundore nga $T_0(\alpha)$, atëherë $S_\alpha(v) > 0$ në bazë të teoremës 3.14, për më tepër, nëse t është një para ardhës i v, ajo është një nyje jo fundore e $T_0(\alpha)$ kështu që $S_\alpha(t) > 0$, në bazë të së njëtës teoremë. Perfundimisht v është një nyje jo fundore e $T_{suff}(\alpha)$. Rrjedhimisht $T_0(\alpha) < T_{suff}(\alpha)$.

3.12 Një algoritëm eksplicit i krasitjes

Le të konsiderojmë një pemë fillestare $T_0 = \{1, ... m\}$ ku m-ja është një numër specifik, sikurse janë $\alpha_{\min}$ dhe madhësitë r(t)=dega e majtë (t), r(t)= dega e djathtë (t), dhe R(t) për $1 \le t \le m$. Duke perdorur barazimin $|T| = 1 + |T_L| + |T_R|$, gjejmë $T_0(\alpha)$ për $\alpha > \alpha_{\min}$. Në këtë algoritëm, "k=1" ka kuptimin marrim k=1 dhe ∞ është një numër i madh dhe positive.

$$N(t) = |\tilde{T}_{kt}|; S(t) = R(T_{kt}); g(t) = g_k(t), dhe, G(t) = \min[\, g_k(s) : s \in T_{kt} - \tilde{T}_{kt}]$$

Procesi i përsëritjes realizohet deri sa të arrihet kushti N(1)=1 dhe T_k është e parëndësishme dhe është provuar se kënaq kushtet tona, në këtë moment algoritmi është i përfunduar. Në barazimin e mësipërm "R(T)" është shkalla e klasifikimit të gabuar, e cila

është relative në krahasim me numrin e klasifikimit të gabuar të nyjes rrënjë, ku "$\tilde{T}$" është numri i nyjeve fundore ose të ashtuquajtura gjethe. Ne në këtë pjesë shikojmë që të kemi një vlerë sa më të vogël të R(T). Por sidoqoftë nuk duam që të gjejmë një numër të madh gjethesh ose nyjesh fundore. Kështu që qëllimi ynë është të gjejmë një nënpemë të cilën e shënojmë simbolikisht me T_α dhe që minimizon $R_\alpha(T)$. Cilën vlerë të α ne duhet të zgjedhim duke filluar nga zero në maksimum në mënyrë të tillëqë α=0 pema maksimale T_{max} është më e mira, α=elarte, pa dyshim që ne nuk duhet të shkojmë në nyjen rrënjë, por më e mira gjendet midis tyre. Ideja e krasitjes së pemës nuk është edhe aq e komplikuar, por në të njëjtën kohë nuk është edhe e thjesht. Degët të cilat në mënyrë direkte reflektojnë zhurmën ose vlera të huaja në pemën fillestare duhet të largohen të parat. Krasitja është e bazuar kryesisht në dy koncepte bazë të mosklasifikimit dhe të vlerësimit të kryqëzuar dhe një nga metodat më të përdorëshme është metoda me 5 apo 10 palosje të vlerësimit të kryqëzuar.

3.13 Perfundime

Në këtë kapitull në pjesën e parë paraqitet një përshkrim i procesit të krasitjes dhe i marrjes së një peme optimale, një pemë e cila duhet të jetë më e mira, të jetë e besueshme dhe e vlefshme për tu përdorur në fushat e ndryshme te jetës. Per më tepër bëhet një paraqitje e detajuar si në aspektin numerik dhe atë grafik se si te zgjedhim numrin e duhur te nyjeve fundore per te arritur te pema optimale, nëpërmjet përdorimit efikas të softwarit R. Përdorimi i testeve stistikore është i rëndësishëm për të gjetur permën përfundimtare më të mirën e të mirave, pasi në bashkësin e pemeve përfundimtare gjendet një e cila konsiderohet më mira.

Në pjesën e dytë janë paraqitur gjitheashtu në mënyrë të përmbledhur testet statistikore, si ato parametrike dhe jo parametrike. Pavarësisht se klasifikimi dhe regresi me ane te pemes është një metodë jo parametrike, për të shmangur ndonjë gabim të llojit të parë apo të dytë ne duhet të bëjëmë matjen e vlefshmërisë së shpërndarjes me qellim që të gjëmë më të mirën. Gjatë proçesit të rritjes së pemës duhet të kemi një kontroll të vazhdueshëm për të arritur te pema maksimale. Në këtë kapitull një vënd të rendesishem zë dhe vërtetimi i disa teoremave si një element thelbësor për strukturimin e pemës së klasifikimit dhe regresit. Një algoritem eksplicit i krasitjes si një nga proçeset më të rëndësishme për të arritur te pema optimale është paraqitur në pjesën e fundit të këtij kapitulli.

KAPITULLI 4

DISKUTIME, KUFIZIMET DHE RASTET E STUDIUARA

4.1 Supozimet e CART

Pema e klasifikimit dhe regresit është një metodë jo parametrike që përdoret si një teknikë për pemën përfundimtare të klasifikimit dhe regresit, si rezultat nuk është e nevojshme të bëhet dhe verifikohet ndonjë supozim dhe të shikohet nëse është me shpërndarje normale baza e të dhënave. Kjo është një nga përparsitë e përdorimit të CART.

- **Avantazhet dhe disavantazhet e përdorimit të pemës së klasifikimit dhe të regresit**

a. Avantazhet e CART

1. CART është një metodë jo parametrike.
2. CART është efektive me çdo llojë baze të dhënash dhe nuk kërkon që variablat të selektohen më përpara.
3. Algoritmet e CART identifikojnë variablat më të rëndësishëm dhe gjithashtu eleminojnë ato të cilat janë të parëndësishme.
4. CART është e lehtë të kuptohet dhe interpretohet kur kemi marrë pemën përfundimtare.
5. Kur përdorim CART nuk është e nevojshme të transformohet baza e të dhënave.
6. Nëse ndryshojmë një ose disa variabla me logaritmet e tyre ose në rrënjët katrore kjo nuk e ndryshon strukturën e pemës, vetëm mënyra e shpërndarjes do të jetë e ndryshme kështu duke zëvëndësuar vlerat fillestare me log(x+100) do të shikojmë se struktura e pemës nuk ndryshon.
7. CART është rezistente ndaj vlerave ekstreme.
8. Kjo metodë me shumë lehtësi kontrollon vlerat ekstreme, pasi sikurse e dimë vlerat ekstreme kanë efekte negative në marrjen e rezultateve përfundimtare në disa metoda të tjera statistikore. Algoritmi i shpërndarjes në CART me lehtësi do të udhëheqë zhurmën në bazën e të dhënave.
9. CART nuk ka nevojë për supozimet dhe njehsohet shumë shpejt.
10. CART është fleksibël dhe ka aftësinë të përshtatet në kohë.
11. Shkalla e gabimit të klasifikimit është e dhënë në CART.

b. Disavantazhet e CART

Si çdo metodë dhe CART ka disa te meta.

1. CART nuk jep gjithmonë të njëjtën pemë.
2. CART nuk ndihmon kur përdorim kombinimin e variablave.
3. Pema mund të jetë jo e selektuar, një variabël nuk mund të përfshihet në qoftë se është mbuluar nga një variabël tjetër.
4. Struktura e pemës mund të jetë e paqëndrueshme, por një ndryshim në shëmbull mund të japë pemë të ndryshme.
5. Pema është optimale në çdo ndarje.
6. CART është shumë e komplikuar për tu lexuar, kur variablat përbëhen nga shumë kategori.

7. Ka një numër të kufizuar të programeve e software që mund të përdoren për të bërë analizën e pemës me anë të klasifikimit dhe regresit.

Një nga programet më të njohura software në statistika është RGui ose R. Ka dy paketa të përbashkëta për modelet e klasifikimin dhe regresit me anë të një peme në R: "tree" dhe "rpart". Në përgjithësi, dy paketa janë të ngjashme; së pari duhet të rritim një pemë dhe pas kësaj duhet ta krasitim atë në mënyrë që të gjejmë pemën me të mirë që të paraqesë të dhënat tona në mënyrën sa më të mirë. Megjithatë, rezultatet e çdo paketë mund të jenë të ndryshme. Më parë do të përdorim paketën rpart sepse rezultati është më lehtë për të interpretuar.

4.2 Vlerat e munguara

Në disa raste në bazën e të dhënave që përdorim ndoshta mund të mungojnë vlerat e disa ndryshoreve. Supozojmë se secila variabël ka 5% shansin të ketë mungesa në mënyrë të pavarur. Atëherë për një bazë të dhënash që ka 50 variabla, probabiliteti i mungesës së disa vlerave të variablave mund të jetë aq i lartë sa të ketë vlerat rreth 92.3%, pra 90% e tyre do të ketë të paktën një vlerë të munguar. Kështu që nuk mund ta hedhim poshtë këtë baze te dhenash për çka mund të shkaktojnë këto mungesa. Ne proceset kualifikuese shpesh ndeshemi me mungesa të vlerave të caktuara. Metoda klasifikuese për të udhëhequr këtë proces përdorë një metodë tjetër të përshtatshme (surrogate split).

Supozojmë se shpërndarja më e mirë për një nyje t është s. Tani le të mendojmë se çfarë mund të bëjmë në se kjo vlerë mungon. Metoda klasifikuese me anë të pemës e trajton këtë mungesë duke bërë një rizëvëndësim te shpërndarja.

Për të gjetur një ndarje duke u bazuar në një variabël tjetër, pema klasifikuese shikon te të gjitha shpërndarjet e pikave të të dhënave të cilat përdorin variablat e tjerë dhe zgjedh atë shpërndarje e cila është e ngjashme dhe që na jep pemën optimale. Së bashku me të njëjtën linjë mendimi, ndarja e dyte e përshtatshme më e mirë mund të gjendet në rast se të dyja, variablat më të mira dhe surrogate mungojnë, e kështu me radhë. Pema e klasifikimit nuk do të përdoret për të gjetur nje ndarje të dytë më të mirë. Këtu, qëllimi është për të ndarë të dhënat sa më të ngjashme të jetë e mundur pas ndarjes më të mirë në mënyrë që për të kryer vendimet e ardhshme poshtë pemës, që zbresin pas ndarjes më të mirë. Nuk ka asnjë garanci ndarja dytë më e mirë i ndan të dhënat në mënyrë të ngjashme si ndarje më të mirë, edhe pse matjet e tyre në mirësi janë të afërta.

4.3 Rastet e studiuara

Në këtë studim, përdoren tre baza të dhënash, të marr nga një spital i Afrikës së Jugut, një nga Cleveland Clinic, Ohio USA për të cilat do të përdoret pema e klasifikimit dhe baza tjeter e te dhenave është për pemën e regresit dhe për këtë do të përdor bazën e të dhënave “Boston House Market” në të cilën variabli përgjegjës është i vazhdueshëm. Do të përdoret softuari R për analizat statistikore.

Baza e të dhënave të meshkujve me probleme kardiovaskulare në rajonin e Western Cape, South Africa. Shumica e burrave që kanë rezultuar positive me CHD, kanë pasur një trajtim mjekësor për të ulur tensionin e gjakut, gjithë ashtu kanë pasur dhe trajtime të tjera për të reduktuar dhe faktorë të tjerë të sëmundjeve kardiovaskulare. Në shumicën e rasteve matjet janë bërë pasi janë bërë këto trajtime. Kjo bazë e të dhënave është shkëputur nga baza e të dhënave shumë e madhe që është përshkruar në Rousseau me 1983, South African Medical Journal.

a. Analiza e bazës së të dhënave në spitalin e Afrikës së Jugut

Në këtë punim, qëllimi im është:

1. Të studioj dhe të kuptoj lidhjen midis CHD (Coronary Heart Diseases) dhe faktorëve të tjerë si mosha, historia familjare, pirja e duhanit apo faktorëve të tjerë që do të përshkruhen me poshtë.
2. Të studiojë dhe analizojë lidhjen midis historisë familjare dhe Coronary Heart Diseases, duke përdorur tabelën e kontigjences.
3. Te studioj efektin e nëntë faktorëve me Coronary Heart Diseases, qe paraqet një interes të veçantë.

- **Variabli përgjegjës eshte CHD (coronary heart disease)**

Duhet të shohim se si janë të lidhura të gjitha variablat dhe ndryshoret përgjegjëse të sëmundjeve koronare të zemrës. Qëllimi i këtij studimi është që të zbatohet një metodë analitike për të gjithë pacientët që janë vërejtur në këtë rast studimor për: (a) të identifikojë nivelin e ndikimit për të gjithë faktorëve; (b) të shqyrtojë ndërveprimet ndërmjet variablave klinike dhe ndikimi i tyre në sëmundjet koronare të zemrës; dhe (c) për të ilustruar në mënyrë të qartë se si këto variabla bashkëveprojnë, në 462 pacientë të analizuar të cilët i janë referuar Spitalit në Afrikën e Jugut. Analizat e shumëllojshme të sëmundjeve koronare të zemrës do të kryhen duke përdorur ndarjen ne pjesë dhe në mënyrë rekursive për të gjithë pacientët e referuar, ku do të përdorim pemën e klasifikimit dhe të regresit (CART).

Variablat shpjegues

Do të fillojmë analizën duke ndërtuar një pemë që na ndihmon për të klasifikuar (pacientët) sipas të dhënave në studim të zgjedhura nga ky spital, që përmban 9 matjet në 462 pacientë ku variabli përgjegjës është CHD. Matjet ose Variablat e pavarur parashikues janë:

1. sbp (systolic blood pressure, continues variable). E vazhdueshme
2. Tobacco (cumulative tobacco (kg), continues variable).E vazhdueshme
3. Ldl (low density lipoprotein cholesterol, continues variable). E vazhdueshme
4. Adiposity (, continues variable).E vazhdueshme
5. Famhist (family history of heart disease (Present, Absent), categorical variable). Kategorike
6. Typea (type-A behavior, continues variable). E vazhdueshme
7. Alcohol (current alcohol consumption, continues variable). E vazhdueshme
8. Age (age at onset, discrete variable, range from 15 years old to 64). Diskrete
9. Obesity (continues variable).E vazhdueshme

Së pari lexojmë të dhënat në studim në software

```
"row.names" "sbp" "tobacco" "ldl" "adiposity" "famhist"
"typea" "obesity" "alcohol" "age" "chd"
```

Le të marrim një informacion numerik për te dhenat:

str(y)

```
'data.frame':   462 obs. of  11 variables:
 $ row.names: int  1 2 3 4 5 6 7 8 9 10 ...
 $ sbp      : int  160 144 118 170 134 132 142 114 114 132 ...
 $ tobacco  : num  12 0.01 0.08 7.5 13.6 6.2 4.05 4.08 0 0 ...
 $ ldl      : num  5.73 4.41 3.48 6.41 3.5 6.47 3.38 4.59 3.83 5.8 ...
 $ adiposity: num  23.1 28.6 32.3 38 27.8 ...
 $ famhist  : Factor w/ 2 levels "Absent","Present": 2 1 2 2 2 2 1 2 2 2 ...
```

```
$ typea    : int  49 55 52 51 60 62 59 62 49 69 ...
$ obesity  : num  25.3 28.9 29.1 32 26 ...
$ alcohol  : num  97.2 2.06 3.81 24.26 57.34 ...
$ age      : int  52 63 46 58 49 45 38 58 29 53 ...
$ chd      : Factor w/ 2 levels "N","Y": 2 2 1 2 2 1 1 2 1 2 ...
```

Po ashtu ne mund të marrim një përmbledhje numerike të datës bazë e cila është si më poshtë.

Le të lexojmë te dhenat në softwarin R:

row.	Names	sbp	tobacco	ldl	adiposity	famhist	typea	obesity	alcohol	age	chd
1	1	160	12	5.73	23.11	present	49	25.3	97.2	50	Y
2	2	144	0.01	4.41	28.61	absent	55	28.87	2.06	63	Y
3	3	118	0.08	3.48	32.28	present	52	29.14	3.81	46	N
4	4	170	7.50	6.41	38.03	Present	51	31.99	24.26	58	Y
5	5	134	13.60	3.50	27.78	Present	60	25.99	57.34	49	Y
6	6	132	6.20	6.47	36.21	present	62	30.77	14.14	45	N

Tabela 13: *Baza e të dhënave nga spitali i Afrikës së Jugut*

Për të parë shpërndarjen e të dhënave nëse është normale apo jo, mund të përdorim paraqitjen grafike të funksionit të densitetit. Gjithashtu mund të shikojmë funksionin e densitetit të çdo variabli me variablin përgjegjës.

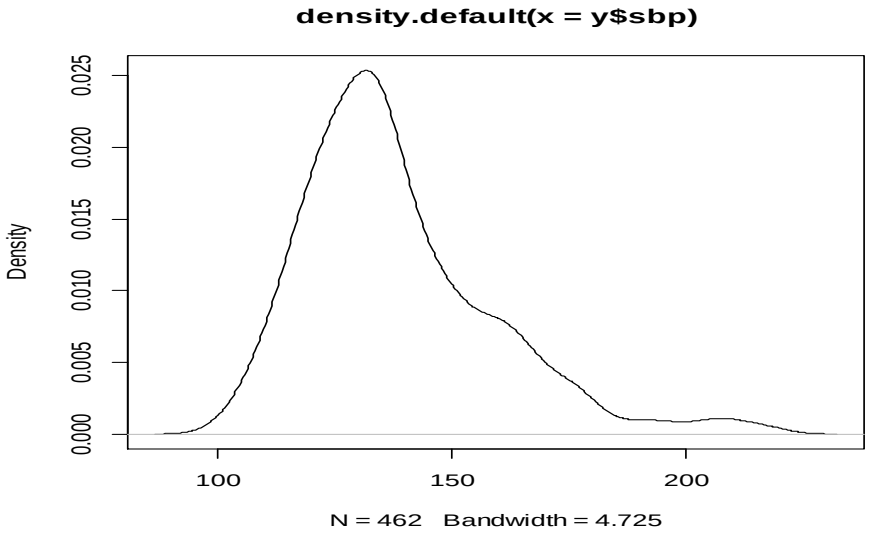


Figura 18: *Grafiku i densitetit të bazës së të dhënave*

Në grafikët e mëposhtëm do të shikojmë shpërndarjen tre dimensionale të të dhënave për variablat e ndryshme, qartësisht shikohet se si janë shpërndarjet e të dhënave dhe si ndryshojnë vlerat e saj.

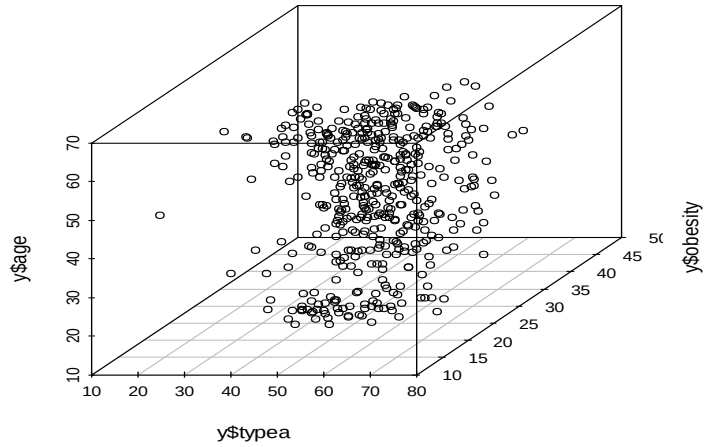


Figura 19: *Shpërndarja tredimensionale e age, obesity dhe type në lidhje me variablin përgjegjës*

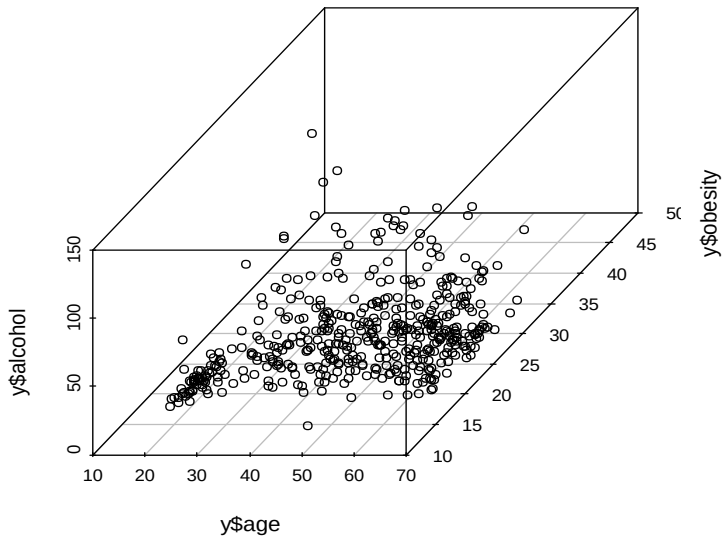


Figura 20: *Shpërndarja tredimensionale e age, obesity dhe alcohol në lidhje me variablin përgjegjës*

Ne figuren 21 paraqitet një formë tjetër grafike e kësaj baze të dhënash për të parë shpërndarjen e disa variablave.

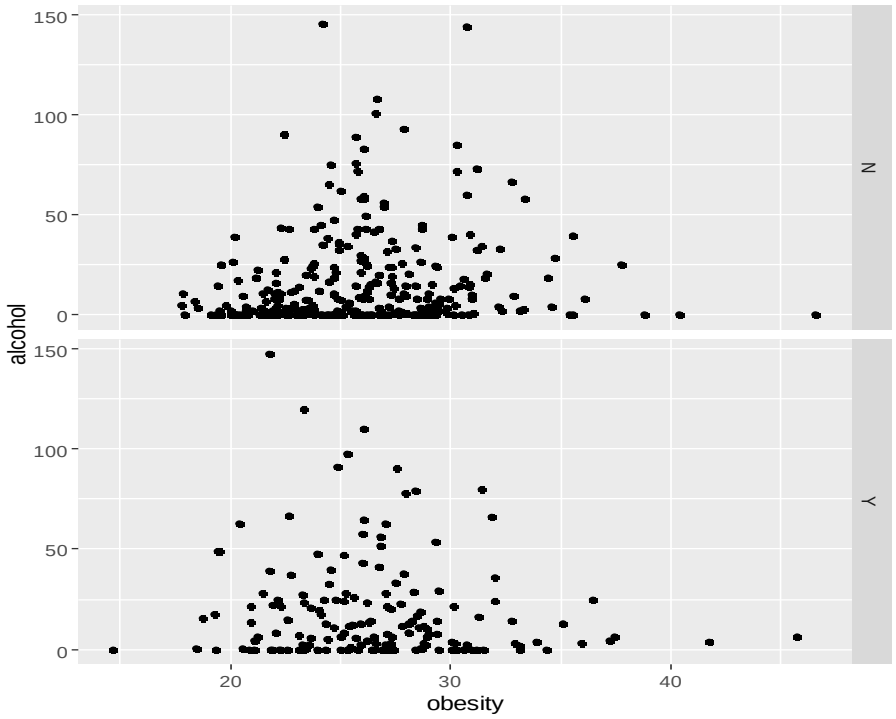


Figura 21: *Shpërndarja e të dhënave alcohol dhe obesity*

a. Së pari të studiojemë varësinë midis sëmundjeve koronare të zëmrës dhe historisë familjare.
b. Duke përdorur paraqitjet grafike mund të krahasojmë madhësinë e vlerave për të gjithë variablat në rastin kur sëmundjet koronare të zemrës janë prezent dhe në rastin kur nuk janë prezent.

b. Vlerësimi i varësisë midis CHD me secilin prej nëntë faktorëve.

Duhet të shikojmë shpërndarjen e bazës së të dhënave nëse është normale apo jo, megjithëse CART është një test jo parametrik gjë e cila e shmang nevojën e të parit nëse është apo nuk është normale. Duke e kontrolluar nuk bëjmë ndonjë gabim përveç se sigurohemi për rezultatet tona. Për të parë këtë mund të përdorim grafikët si box plot, histogramet dhe qqplot për sejcilën nga variablat tona, gjithashtu do të krahasojmë dhe madhësinë e vlerave të çdo variabli kur sëmundjet koronare të zemrës janë prezente dhe në rastin kur ato nuk janë prezente duke përdorur box plot si mjet krahasuese.
Nga grafiket e Box plot qartësisht shikojmë se baza e të dhënave ka vlera ekstreme të vogla apo te medha(outliers), gjë e cila tregon se baza e të dhënave nuk është me shpërndarje normale.

Nga grafikët e mësipërm shikojmë se mosha e vjeter eshtë më shumë prezente se sa mosha e re. Po ashtu pothuajse nga të gjitha Box plots shikojmë se në shumicën e variablave kanë vlera të huaja, në disa nga ato kemi një shpërndarje jo normale dhe së fundi vlerat minimale dhe maksimale të gjithë variablave janë pothuaj të njëjta si në rastin kur CHD është

prezente apo nuk është prezente. Ne disa raste shikojmë se disa variabla kane vlera pothuaj te njëjta si rastin kur CHD është prezente dhe kur nuk është prezente.

Shikojmë se disa variabla shpjeguese janë me shpërndarje normale dhe disa me shpërndarje jo normale, por do të përdorim pemën e klasifikimit dhe të regresit e cila është analizë statistkore jo parametrike, rrjedhimisht do të bëjmë analizën e të dhënave.

Grafikët e Boxplots të paraqitura ne Figurën 22 dhe Figurën 23 tregojnë se shpërndarja e te dhënave është jo normale. Më poshtë janë dy grupe për CHD ku respektivisht me histori familjare po dhe jo.

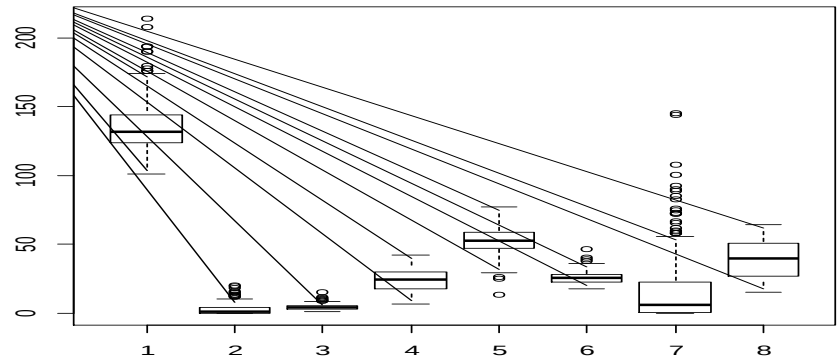


Figura 22: *Boxplot kur historia familjare është present. CHD(po)*

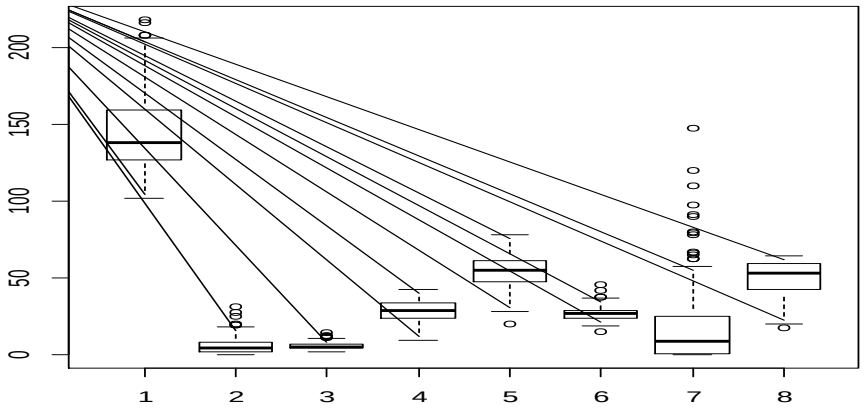


Figura 23: *Boxplot kur historia familjare nuk është present. CHD(jo)*

Dhe një paraqitje tjetër dy dimensionale e shpërndarjes së bazës së të dhënave për variablat e ndryshme.

With (y, plot (tobacco, ldl, col=chd, pch=as.numeric(chd))) > with (y, plot (adiposity, typea, col=chd, pch=as.numeric(chd)))

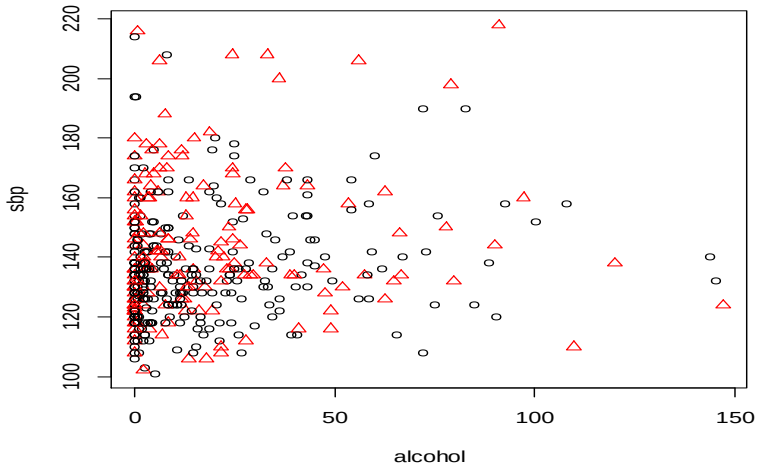


Figura 24: *Shpërndarja dy dimensionale e variablave alcohol dhe sbp*

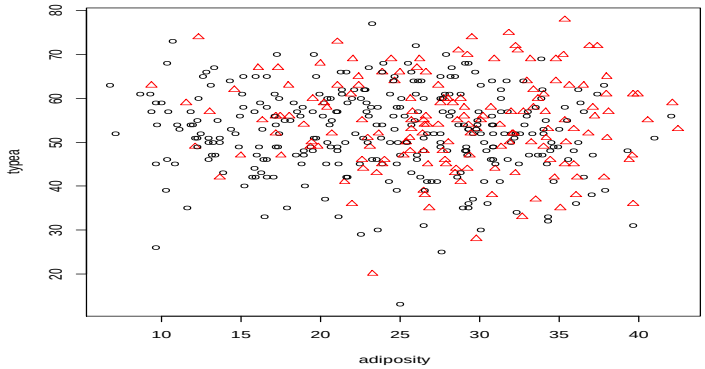


Figura 25: *Shpërndarja dy dimensionale e variablave adiposity dhe typea*

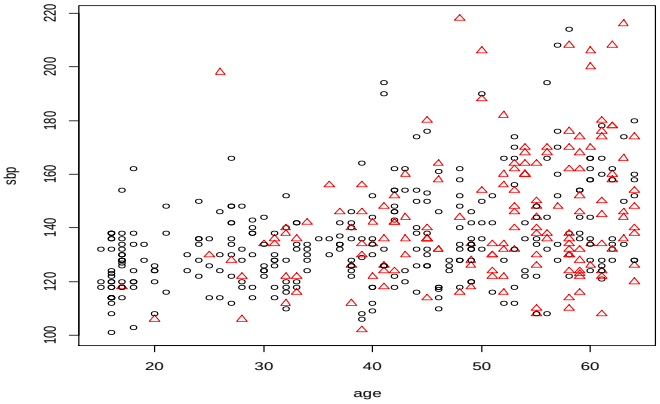
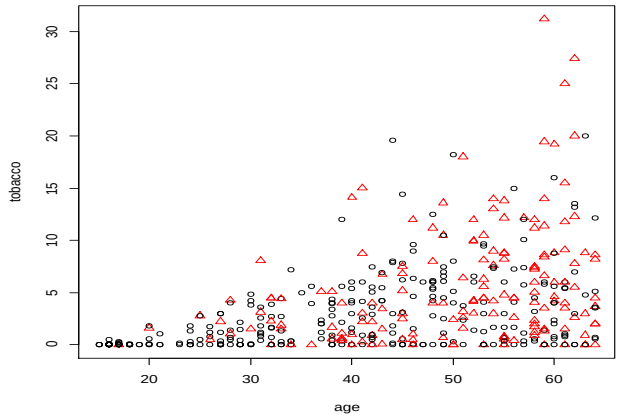


Figura 26: *Shpërndarja dy dimensionale e variablave age dhe sbp*



Figurë 27: *Shpërndarja dy dimensionale e variablave age dhe tabaco*

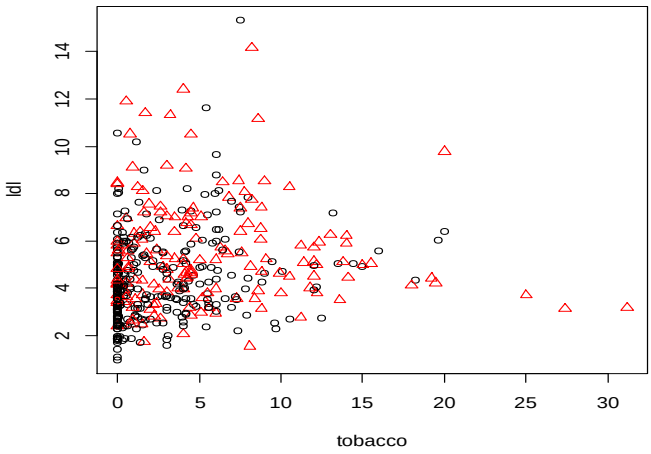


Figura 28: *Shpërndarja dy dimensionale e variablave tabaco dhe idl*

Nga grafikët e mësipërm vihet re se të dhënat kanë një shpërndarje jo normale dhe në mënyrë të specifikuar shikojmë se si janë të përqëndruara elementet e ndryshëm të te dhënave duke e parë në planin dy dimensional.

4.4 Varësia midis variablave

Le të studjojme varësinë ndërmjet variablave të ndryshme duke përdorur testin Hi-katror.

Së pari të dhënat janë marrë nga një zgjedhje e rastësishme.
Ne kemi me pritshmëri me më shumë se 5 pika në çdo qelizë.
Së pari marrim dy variablat dhe performojmë hipotezat bazë dhe alternative të cilat janë dhënë më poshtë**:**
Variables: CHD versus Family History (të dyja kategorike).
H_0: CHD status dhe Family History janë të pavarura.
H_1: CHD status dhe Family History janë të varura.
Niveli i rëndësisë është = 0.05.

Rows:	chd	Columns:	famhist
Absent	Present	All	
N	206	96	302
Y	64	96	160
All	270	192	462

Tabela 14: *Tabela e varësise per variablat CHD dhe famhis*

Chi-Square Test: Absent, Present

Expected counts are printed below observed counts
Chi-Square contributions are printed below expected counts

	Absent	Present	Total
1	206	96	302
	176.49	125.51	
	4.933	6.937	
2	64	96	160
	93.51	66.49	
	9.311	13.094	
Total	270	192	462

Chi-Sq = 34.274, DF = 1, P-Value = 0.000

Tabela 15: *Tabela Hi-kateror*

Nga tabela 14 dhe 15 shikojmë se vlera p-value (0.000) është më e vogël se 0.05(5%). Kjo tregon se statusi CHD dhe historia familjare (Family History) janë të varura. Rrjedhimisht besojmë se ka varësi midis këtyre dy variablave. Këtë kontroll hipotezash e bëjmë dhe për variablat e tjerë të bazës së të dhënave dhe arrijmë në të njëjtin përfundim se të gjitha kombinimet e mundëshme kanë varësi me njëra tjetrën.

Një përmbledhje statistikore numerike e bazës së të dhënave

1. Përmbledhje statistikore

Pese numrat per sbp(systolic blood pressure), tobacco and ldl.

row.	namessbp	tobacco	ldl
Min.: 1.0	Min.:101.0	Min. : 0.0000	Min. : 0.980
1st Qu.:116.2	1st Qu.:124.0	1st Qu.: 0.0525	1st Qu.: 3.283
Median :231.5	Median :134.0	Median : 2.0000	Median : 4.340
Mean :231.9	Mean :138.3	Mean : 3.6356	Mean : 4.740
3rd Qu.:347.8	3rd Qu.:148.0	3rd Qu.: 5.5000	3rd Qu.: 5.790
Max. :463.0	Max. :218.0	Max. :31.2000	Max. :15.330

Tabela 16: *Permbledhje statistikore për bazën e të dhënave*

Nga tabela 17 vihet re se presioni i gjakut (SBP) lëviz nga 101 në 218, për duhanin kemi pacient që nuk e përdorin atë në masën max 31.2 kg dhe për ata që kanë densitet të ulët të lipoprotein dhe të kolesterolit, variabli i vazhdueshëm (LDL) varion nga 0.980 në vlerën maksimale 15.330. Nga tabela 16 shikojmë se adiposity varion nga 6.74 ne 42. 49, dhe typea varion nga 13 to 78, obesity i cili nuk është shumë i lartë dhe varion nga 14.70 to 46.58 dhe alkooli është nga një vlerë minimale nga zero deri në 147.19.
Nga tabela 16 dhe17 shikojmë se mosha e pacientëve varion nga 15 ne 64 vjec dhe pika e mesit pra mediana është 45 vjeç.
Para se të fillojmë analizën duhet të bëjmë disa ndryshime në bazën e të dhënave filestare në rregullimet që duhet të bëjmë në variablat kategorike me qëllim që të jetë e lexueshme nga software si faktorë të ndryshëm. Përndryshe software R nuk i trajton ata si faktorë të ndryshëm e cila e ndryshon pemën në pemë të regresit!
Do të përdorim funksionin rpart.control për të kontrolluar përmasat fillestare të pemës, pemës maksimaleTmax, numrin e palosjeve të vlerësimi i kryqzuar, dhe parametrin e kompleksitetit "cp or α .

Alcohol	**Diposity**	**famhist**	**typea**	**obesity**
Min. : 6.74	Absent :270	Min.:13.0	Min.:14.70	Min.: 0.00
1st Qu.:19.77	Present:192	1st Qu.:47.0	1st Qu.:22.98	1st Qu.: 0.51
Median :26.11		Median :53.0	Median :25.80	Median : 7.51
Mean :25.41		Mean:53.1	Mean :26.04	Mean : 17.04
3rd Qu.:31.23		3rd Qu.:60.0	3rd Qu.:28.50	3rd Qu.: 23.89
Max.: 42.49		Max.:78.0	Max. :46.58	Max.:147.19
Age Min.:15.00 1st Qu.:31.00 Median :45.00 Mean :42.82 3rd Qu.:55.00 Max. :64.00	chd N:302 Y:160			

Tabela 17: *Përmbledhje statistikore për adiposity, typea, obesity dhe alcohol.*

Funksioni print cp jep një vlerësim të përafërt të vlerësim i kryqëzuar dhe të gabimit të mosklasifikimit (xerror), gabimit standart(xstd) për këto gabime dhe për rizëvëndësimin e gabimit të vlerërsuar përafërsisht:
Rrite pemën në maksimum (Tmax)(shiko Apendix 1).
Ne qartësisht shikojmë se gabimi zvogëlohet kur pema bëhet më e madhe, por gabimi vlefshmeri e kryqëzuar zvogëlohet në fillim dhe arrin minimumin kur (xstd=0.063823) dhe kur pema ka 10 shpërndarje, po ashtu (α = cp = 0.009375), dhe pas këtij momenti fillon rritet, kështu që vleresimi i kryqëzuar sygjeron se përmasa optimale e pemës është pema me 10 shpërndarje. Pas kësaj përkufizojmë variablin CAD1 me anë të të cilit mund të bëjmë përmbledhjen e cila na jep më shumë informacion te rëndësishëm për secilin variabël dhe për çdo nyje.
Gjithashtu mund të gjejmë informacion për vlerat e munguara të të dhënave tona.

```
Call:cart<-rpart(PRONO~.,data=MYOCARDE)
```

rpart (formula = chd ~ sbp + tobacco + ldl + adiposity + famhist + typea + obesity + alcohol + age, data = x, method = "class", control = my.control)

n= 462

Variablat e rëndësishëm

age	adiposity	tobacco	ldl	sbp	obesity	typea	alcohol	famhist
16	15	14	12	12	12	9	8	3

Node number 1: 462 observations, complexity param=0.125
predicted class=N expected loss=0.3463203 P(node) =1
class counts: 302 160
probabilities: 0.654 0.346

- Shperndarja primare:

age< 50.5 to the left, improve=24.58856, (0 missing)
tobacco< 0.49 to the left, improve=19.42366, (0 missing)
famhist splits as LR, improve=15.51823, (0 missing)
ldl< 4.315 to the left, improve=12.58910, (0 missing)
adiposity< 25.16 to the left, improve=10.38739, (0 missing)

- Surrogate splits:

adiposity< 31.34 to the left, agree=0.721, adj=0.250, (0 split)
sbp< 155 to the left, agree=0.710, adj=0.221, (0 split)
tobacco< 7.24 to the left, agree=0.695, adj=0.180, (0 split)
typea< 38.5 to the right, agree=0.649, adj=0.058, (0 split)
ldl< 8.25 to the left, agree=0.645, adj=0.047, (0 split)
Nga softwari R marrim informacionin për sejcilën nyje e cila numerikisht është e ngjashme me atë që gjejmë me pemën. Këto rezultate japin informacion të detajuar për sejcilën nyje.

Le të ndërtojmë pemën.

Në figurën 30 paraqesim grafikun e pemës së mbingarkuar $T_{\max}$ e cila është e vështirë të lexohet, por kjo është normale kur kemi të bëjmë me një pemë maksimale $T_{\max}$, kjo pemë nuk është një pemë optimale e cila ka nevojë të krasitet. Duke përdorur funksionin rpart.control i cili na lejon të kontrollojmë përmasën e pemës fillestare, Tmax, dhe duke përdorur vlefshmeria e kryqezuar, dhe procedurat e tjera ne mund të bëjmë krasitjen dhe marrjen e pemës optimale. Kjo procedurë ka disa opsione , duke parë këtë funksion në tabelat dhe grafikët e mësipërm "cp" i cili është (complexity parameter) Në fillim rritim pemën në maksimum T_{max}. Më poshtë kemi skemën e pemës duke përdorur të 9 variablat dhe pa përdorur testin e duhur për të klasifikuar pacientët:

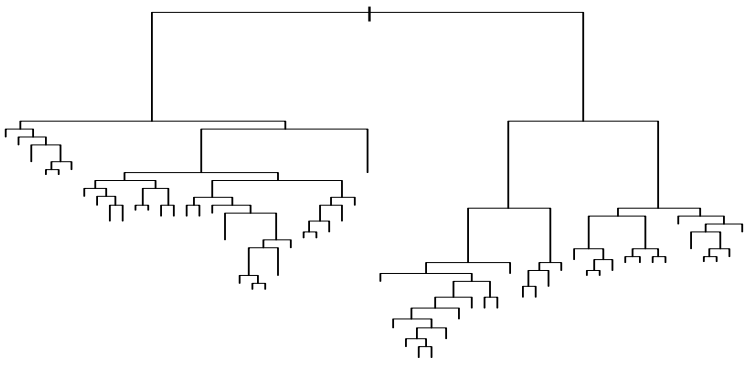


Figura 29: *Pema maksimal*e

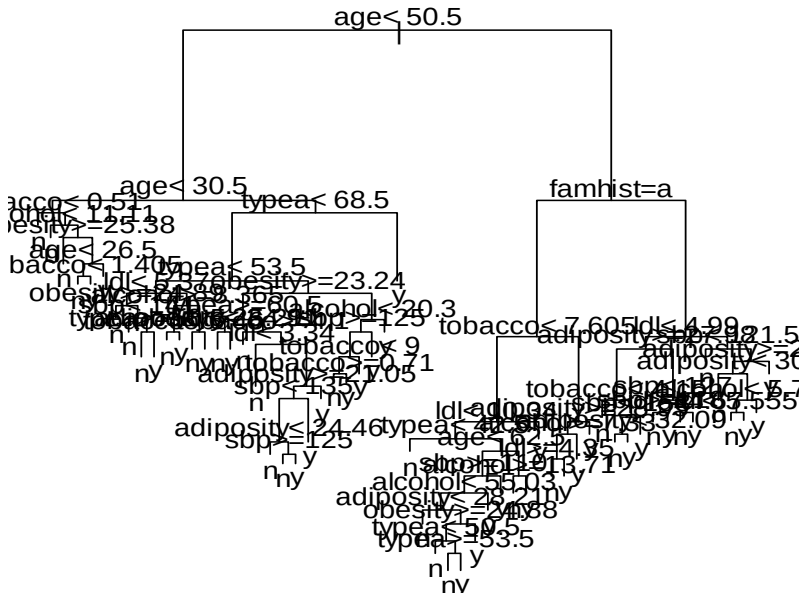


Figura 30: *Pema maksimale me tekstin*

Nga figura 30 shohim se pema është e mbingarkuar dhe nuk mund të nxjerrim ndonjë përfundim për të dhënat tona, për këtë arsye është e nevojshme të krasitim pemën dhe për të gjetur pemën më të mirë. Për të gjetur koeficientin e kompleksitetit duhet të bëjmë grafikun e cp.

Ky grafik dhe vlerësimi i gabimit të mosklasifikimit në bazën e të dhënave dhe vlerësimet, kundrejt kompleksitetit të pemës paraqitur në Figura 32.

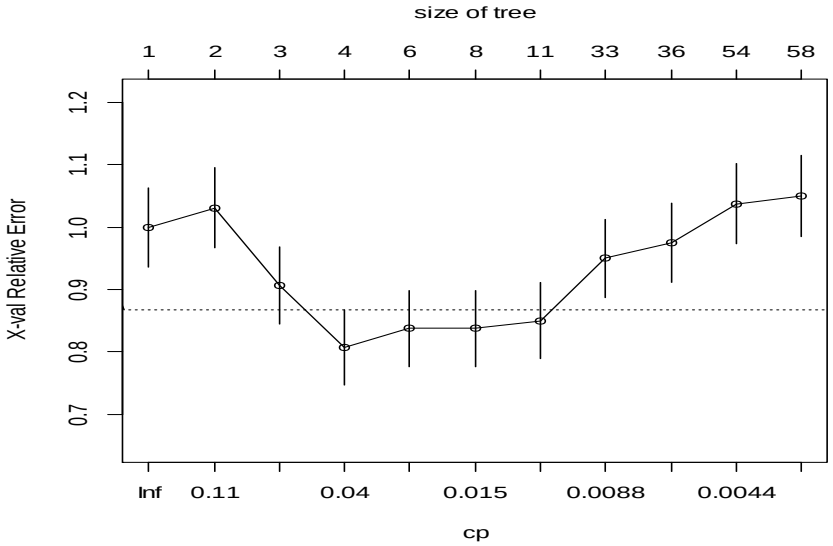


Figura 31: *Grafiku i kompleksitetit për krasitjen me anë te vlefshmërisë së kryqëzuar*

Figura 32 tregon se vleresimi i kryqëzuar sygjeron një pemë optimale të madhësisë që varion nga tetë në njëmbëdhjetë nyje fundore. Duke zgjedhur një pemë me njëmbëdhjetë nyje fundore, kështu që ky është një model i përshtatshëm për rastin tonë. Një pemë mund të krasitet në mënyrë interaktive në disa mënyra. Kodin e mëposhtëm të krasitjes së pemës e cila do të ketë vetëm 11 nyje fundore, pemën të cilën e marrim për cp = 0.009375. Hapi tjetër është se për të krasitur pemën e cila do jetë pema më optimale siç përcaktohet nga vlerësimi i kryqëzuar.

```
 Classification tree:
rpart(formula = chd ~ sbp + tobacco + ldl + adiposity + famhist +
    typea + obesity + alcohol + age, data = x, method = "class",
    control = my.control)

Variables actually used in tree construction:
[1] adiposity age      alcohol  famhist  ldl      obesity  sbp
[8] tobacco  typea

Root node error: 160/462 = 0.34632

n= 462

         CP nsplit rel error  xerror     xstd
1  0.1250000      0    1.0000 1.00000 0.063918
2  0.1000000      1    0.8750 0.96875 0.063430
3  0.0625000      2    0.7750 0.91875 0.062571
4  0.0250000      3    0.7125 0.86875 0.061612
5  0.0187500      5    0.6625 0.88750 0.061984
6  0.0125000      7    0.6250 0.89375 0.062104
7  0.0093750     10    0.5875 0.91875 0.062571
8  0.0083333     32    0.3375 1.01875 0.064193
9  0.0062500     35    0.3125 1.05625 0.064705
10 0.0031250     53    0.2000 1.11875 0.065445
11 0.0000000     57    0.1875 1.15000 0.065764
```

4.5 Krasitja e pemës me selektim

Duke pasur parasysh një pemë tepër të madhe të cilën e shënojmë Tmax, atëherë të gjitha nënpemët e këtij modeli do të jenë gjithashtu të mëdha dhe duhet të bëjmë kërkime të mëtejshme për të gjetur pemën e cila e thënë ndryshe do të jetë një moderim i kësaj peme. Së pari e konsiderojmë proçesin e krasitjes si një proçes me dy faza. Në fazën e parë krijojmë një grup të pemëve të krasitura të marra nga Tmax duke e bërë këtë në bazë të disa kritereve të caktuara, ndërsa në fazën e dytë një prej pemëve të tilla është zgjedhur si modeli përfundimtar. Kjo është qasja e ndjekur në CART (Breimanet al., 1984). Lloji i dytë i metodave për krasitjen përdor një proçedurë me një hap të vetëm dhe është më e shpeshtë në përdorim. Algoritmi i fundit vepron nëpër nyjet e pemës nga lart poshtë apo nga poshtë lart, duke vendosurne baze te kritereve te vlersimit se cilën nyje do të krasiti dhe cilin nyje do të mbajmë.

Këto dy forma të dallueshme të krasitjes së një peme kanë një ndikim në vlerësimin e metodës së përdorur në procesin e krasitjes. Kur e konsiderojmë këtë një metodë me dy hapa, vlerësimi i pemëve mund të shihet si një problem i modelit te përzgjedhjes, për arsye se duam

të krahsojmë pemët alternative të krasitura me qëllim të përzgjedhjes së një peme më të mirë. Në rastin tjeter , metoda me një hap përdor vlerësimin në nivel lokal, pra duhet të vendosim në çdo nyje nëse duhet krasitur apo jo. Për më tepër, metoda me dy hapa ka një shkallë të fleksibilitetit që eshte e pershtatshme nga ana e përdorimit praktik të pemës bazë të regresit. Në fakt, ato mund të prodhojnë sekuenca të modeleve alternative tëpemëve të krijuara në fazën e parë së bashku me vlerësimin e tyre (ose një vlerësim të gabimit të tyre). Këto pemë mund të konsiderohen si modele alternative që do të shkëmbehen ndërmjet modeleve komplekse dhe rezultatittë vlerësimit. Sistemi zgjedh një nga këto pemë bazë duke përdorur disa paragjykimeve (psh vlersimi i gabimit me i vogel), por pa asnjë kosto shtesë llogaritje mund të lejojmë përdoruesin të zgjedhë çdo pemë tjetër që i përshtatet më mirë nevojave të tij të aplikimit.

Mund të shohim në Figurën 32 se kur gabimi zvogëlohet se si pemët bëhen më të mëdha, por gabimi vlersimit te kryqëzuar arrin minimumin kur pema ka 11 ndarje (= CP = 0.009375), dhe pastaj fillon të rritet përsëri. Kështu vleresimi i kryqëzuar sygjeron se pema optimale është pema me 11 nyje. Me poshtë është pema me 11 nyje e cila merret pasi kemi krasitur pemën maksimale Tmax. Kjo është pema me përmasa optimale.

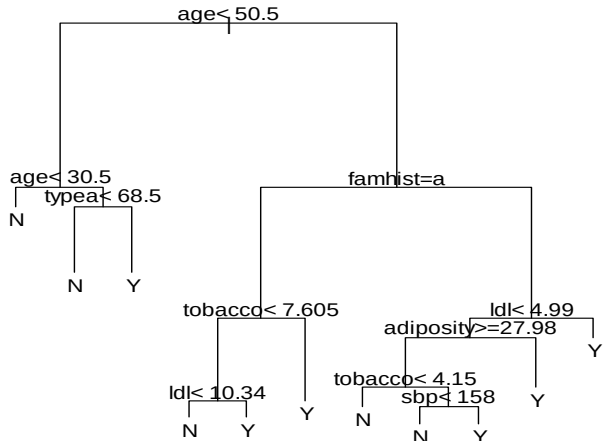


Figura 32: *Nënpema me e mirë e krasitur*

Përfundime për këtë rast studimi

Figura 33 tregon shpërndarjen primare. Ne mund të ndajmë këtë pemë në dy pjesë në të majtë ku historia familjare e kësaj semundjeje nuk është prezente dhe me një moshë më të vogël se 50.5 dhe në të djathtë ku historia familjare është prezente dhe me një moshë më të madhe se 50.5. Në të djathtë kur historia familjare është prezente dhe nëse ldl është më i madh se 4.99, atëherë përgjigja për mundësinë sëmundjeve të zemrës është po, kur ldl është më pak se 4.99 dhe adiposity është më i madh apo i barabartë me 27.98 atëherë sëmundja kardiovaskulare është present domethene përgjigja është po, nëse adiposity është më pak ose i barabartë me 27.98 dhe nëse tobacco është më pakë se 4.15 atëherë përgjigja është jo, por nëse tobacco është më madhe se 4.15 dhe sbp është më shumë se 158 atëherë përgjigja është po dhe për sbp më pak se 158 përgjigja është jo. Nëse historia familjare nuk është prezente për moshën <50.5 dmth nëse nuk ka histori familjare dhe për moshën nën 30.5 përgjigja është jo, dhe nëse type a është më shumë se 68.5 përgjigja është podhe nëse typea është më pak se 68.5përgjigja është nuk do të ketë sëmundje koronare të zëmrës. Është e lehtë të kuptohet se cilat janë variablat e rëndësishme për të bërë parashikime. Nga paraqitja e pemës

përfundimtare qartësisht shikojmë se variablat si alkoli apo obeziteti nuk kanë të njëjtin ndikim në sëmundjet koronare të zëmrës.

Analiza Statistikore e te dhenave marre nga spitali i Klevelandit, Ohio USA

Së pari: Variabli përgjegjës ALLCAD

Përshkrimi i bazes së të dhënave dhe konkluzionet statistikore:

Do të fillojë me analizën duke ndërtuar një pemë që ndihmon për të klasifikuar pacientët tanë, duke u bazuar në emrat që janë dhënë në bazën e të dhënave e cila përmban 11 matjet në 5017 pacientë duke përjashtuar ketu vlerat e munguara të cilat hiqen pasi duke patur në konsideratë që dhe numrin e pacientëve në këtë bazë të dhënash ku ky numër është i vogël. Për këtë bazë të dhënash do të përdorim një emër të caktuar të cilin unë e kam quajtur Y, ku kemi dy variabla përgjegjës të cilat janë CAD dhe AllCAD. Matjet (variablat e pavarura ose parashikues) janë:

1. BNP (variabël i vazhdueshëm).
2. CRP16(variabël i vazhdueshëm).
3. DLDL (variabël i vazhdueshëm).
4. UHDL (variabël i vazhdueshëm).
5. DIABETICS(variabël kategorik), ku; ND=jo=0 dhe YD=po=1.
6. Smoking(variabël kategorik), ku; NS=jo=0 dhe YS=po=1.
7. AGE(variabël i vazhdueshëm).
8. GENDER (variabël kategorik), ku; M=mashkull dhe F=femër.
9. CRECLR (variabël i vazhdueshëm).
10. HTN(variabël kategorik), ku; NH=jo=0 dhe YH=po=1.
11. CVDNY (variabël kategorik), ku; N=jo=0 dhe Y=po=1.

Pacientët mund të klasifikohen në dy klasa dhe më poshtë janë dhënë disa detaje se si variabli ALLCAD është klasifikuar: Të gjitha rastet e CAD, duhet të ketë të paktën një nga këto:

1. RPROC6_S/P_RecentMI
2. RPROC7_HxPCI/CABG
3. HxCabg
4. HxPci
5. HxMI
6. MAXLAD ≥ 50% stenosis in LAD (angiographic)
7. MAXRCA ≥ 50% stenosis in RCA (angiographic)
8. MAXLCX ≥ 50% stenosis in LCX (angiographic)
9. MAXLMT ≥ 50% stenosis in LMT (angiographic)
10. MaxStenosis ≥ 50% stenosis (angiographic)

Pasi të lexojmë datën në R marrim një pamje të saj si më poshtë:

```
'data.frame':   5017 obs. of  12 variables:
 $ ALLCAD   : Factor w/ 2 levels "N","Y": 2 2 2 2 2 2 2 2 2 2 ...
 $ BNP      : num  102.7 74.4 34.9 115.1 121.3 ...
 $ CRP16    : num  0.92 5.72 0.45 3.63 2.62 ...
```

```
 $ DLDL     : int  94 66 62 88 57 107 121 81 117 74 ...
 $ UHDL     : num  40.3 31 37.8 28.9 31.6 30.6 38.8 41.8 35.2 28.4 ...
 $ DIABETICS: Factor w/ 2 levels "ND","YD": 1 2 1 1 1 1 2 2 1 2 ...
 $ smoking  : Factor w/ 2 levels "NS","YS": 1 2 2 2 1 2 1 1 2 1 ...
 $ CVDYN    : Factor w/ 2 levels "N","Y": 2 2 2 2 2 2 2 2 2 2 ...
 $ AGE      : num  55.9 59.5 63.5 78.3 75.3 ...
 $ GENDER   : Factor w/ 2 levels "F","M": 2 2 2 2 1 2 1 1 2 2 ...
 $ CRECLR   : num  111.8 100.9 99.2 100.1 65.7 ...
 $ HTN      : Factor w/ 2 levels "NH","YH": 2 1 2 2 2 1 1 2 1 2 ...
The following is a summary statistics Tabela:
'data.frame':   5018 obs. of  12 variables:
 $ V1 : Factor w/ 3 levels "ALLCAD","N","Y": 1 3 3 3 3 3 3 3 3 3 ...
 $ V2 : Factor w/ 2816 levels "10","10.1","10.2",..: 2816 42 2441 1554 198 272 2211 2619
2535 2390 ...
 $ V3 : Factor w/ 1800 levels "0.05","0.1","0.11",..: 1800 157 1403 64 1059 782 981 414 538
1077 ...
 $ V4 : Factor w/ 203 levels "100","101","102",..: 203 197 169 165 191 160 8 22 184 18 ...
 $ V5 : Factor w/ 546 levels "10.8","11.3",..: 546 253 160 228 139 166 156 238 268 202 ...
 $ V6 : Factor w/ 3 levels "DIABETICS","ND",..: 1 2 3 2 2 2 2 3 3 2 ...
 $ V7 : Factor w/ 3 levels "NS","smoking",..: 2 1 3 3 3 1 3 1 1 3 ...
 $ V8 : Factor w/ 3 levels "CVDYN","N","Y": 1 3 3 3 3 3 3 3 3 3 ...
 $ V9 : Factor w/ 4256 levels "22.01232","22.080766",..: 4256 1183 1616 2108 3866 3563
2472 3465 3737 357 ...
 $ V10: Factor w/ 3 levels "F","GENDER","M": 2 3 3 3 3 1 3 1 1 3 ...
 $ V11: Factor w/ 5012 levels "10.3427","10.39755",..: 5012 579 40 4978 12 3288 4505 3065
3018 2292 ...
 $ V12: Factor w/ 3 levels "HTN","NH","YH": 1 3 2 3 3 3 2 2 3 2 ...
names(x)
 [1] "ALLCAD"    "BNP"       "CRP16"     "DLDL"      "UHDL"      "DIABETICS"
 [7] "smoking"   "CVDYN"     "AGE"       "GENDER"    "CRECLR"    "HTN"
 attach(x)
Tabela(x$ALLCAD)

   N    Y
1106 3911
Vlerat e munguara
sum(complete.cases(x))
[1] 5017
 Table(GENDER)
GENDER
   F    M
1680 3337
chisq.test(Table(GENDER))
library(mvpart)
out1=rpart(BETUPAP~MAT+MWMT+MCMT+MAP+MSP, dat1, xv="p", all.leaves=T)
summary(out1)
Chi-squared test for given probabilities

data:  Table(GENDER)
X-squared = 547.27, df = 1, p-value < 2.2e-16
```

Komanda chisq.test(Tabela(SEX)) bën të mundur që të bëjmë testin Hi-katror dhe testi i mirësisë për gjashtë variablat . Për tu siguruar duhet të testojmë për një pritshmëri të vlerave të barabarta në çdo qelizë, por në këtë kur kemi një vlerë shumë të vogë $l2.2X10^{-16}$ = 0.00000000000000022 është tepër e vështirë. Nëse nuk duam raporte të barabarta të porporcioneve, kemi nevojë të japim një bashkësi të raporteve për çdo qelizë dhe në rastin tonë raportet e arsyeshme për atakun ne zemër në bazën e të dhënave janë 60/40, nëse nuk duam vlera të barabarta të raporteve duhet të japim një bashkësi të porpocioneve për të gjitha qelizat. Një raport i arsyeshëm është (60/40) për bazën e të dhënave të atakut kardiak të zemrës.
Mund të përdorim cros-tabualation për dy variablat kategorike me tabelat dhe të bëjme testin Hi-katror për të parë pavarsinë e variablave

```
Table (GENDER, ALLCAD)
ALLCAD
GENDER    N    Y
    F  559 1121
    M  547 2790
chisq.test(Table(GENDER,ALLCAD))
Pearson's Chi-squared test with Yates' continuity correction
data:  Table(GENDER, ALLCAD)
X-squared = 184.3321, df = 1, p-value < 2.2e-16
Table(GENDER,ALLCAD)
ALLCAD
GENDER    N    Y
    F  559 1121
    M  547 2790
Pearson's Chi-squared test with Yates' continuity correction
data:  Table(GENDER, ALLCAD)
X-squared = 184.33, df = 1, p-value < 2.2e-16
Table(HTN,smoking)
   smoking
HTN    NS   YS
  NH  487  953
  YH 1227 2350
chisq.test(Tabela(HTN,smoking))
Pearson's Chi-squared test with Yates' continuity correction
data:  Tabela(HTN, smoking)
X-squared = 0.086114, df = 1, p-value = 0.7692
Table(ALLCAD,smoking)
     smoking
ALLCAD   NS   YS
    N  502  604
    Y 1212 2699
chisq.test(Tabela(ALLCAD,smoking))

Pearson's Chi-squared test with Yates' continuity correction
data:  Table(ALLCAD, smoking)
X-squared = 78.839, df = 1, p-value < 2.2e-16
summary(x)
```

Nga informacioni i mësipërm shikojmë se vlerat p-values (0.000) janë më të vogëla se 0.05(5%). Kjo tregon se ALLCAD është i varur dhe variablat e tjere janë të varura. Kjo eshte

e dukshme duke perdorur dhe Testin e Parson's Hi-kateror. Rrjedhimisht besojmë se ka varësi midis këtyre variablave. Këtë studim statistikor e bëjmë dhe për variablat e tjerë të bazës së të dhënave dhe arrijmë në të njëjtin përfundim se për të gjitha kombinimet e mundëshme kanë varësi me njëra tjetrën

```
ALLCAD        BNP              CRP16             DLDL              UHDL
N:1106   Min.   :    4.5   Min.   :  0.050   Min.   : 19.00   Min.   : 7.10
Y:3911   1st Qu.:   36.8   1st Qu.:  1.050   1st Qu.: 77.00   1st Qu.:27.80
         Median :   93.8   Median :  2.460   Median : 95.00   Median :33.50
         Mean   :  251.4   Mean   :  7.049   Mean   : 99.17   Mean   :35.22
         3rd Qu.:  238.8   3rd Qu.:  6.060   3rd Qu.:116.00   3rd Qu.:40.70
         Max.   :25000.0   Max.   :225.250   Max.   :313.00   Max.   :94.70
DIABETICS smoking   CVDYN         AGE         GENDER       CRECLR
ND:3361   NS:1714   N: 984   Min.   :22.01   F:1680   Min.   :  6.335
YD:1656   YS:3303   Y:4033   1st Qu.:55.25   M:3337   1st Qu.: 75.315
                             Median :63.59            Median : 98.662
                             Mean   :63.32            Mean   :102.914
                             3rd Qu.:71.99            3rd Qu.:126.040
                             Max.   :92.06            Max.   :349.432
HTN
NH:1440
YH:3577
```

Tabela18: *Përmbledhje statistikore për bazën e të dhënave nga spitali Kleveland, Ohio, USA.*

```
1) root 5017 1106 Y (0.22045047 0.77954953)
  2) CVDYN=N 984    0 N (1.00000000 0.00000000) *
  3) CVDYN=Y 4033  122 Y (0.03025043 0.96974957)
    6) GENDER=F 1190   69 Y (0.05798319 0.94201681)
     12) BNP< 11.32 44    9 Y (0.20454545 0.79545455) *
     13) BNP>=11.32 1146   60 Y (0.05235602 0.94764398)
       26) UHDL>=55.55 90   11 Y (0.12222222 0.87777778)
         52) BNP< 54.55 23    6 Y (0.26086957 0.73913043)
          104) BNP>=47.9 5    1 N (0.80000000 0.20000000) *
          105) BNP< 47.9 18    2 Y (0.11111111 0.88888889) *
         53) BNP>=54.55 67    5 Y (0.07462687 0.92537313) *
       27) UHDL< 55.55 1056   49 Y (0.04640152 0.95359848)
         54) CRP16< 0.535 53    6 Y (0.11320755 0.88679245) *
         55) CRP16>=0.535 1003   43 Y (0.04287139 0.95712861)
          110) BNP< 82.35 324   23 Y (0.07098765 0.92901235)
            220) BNP>=81.25 5    2 Y (0.40000000 0.60000000) *
            221) BNP< 81.25 319   21 Y (0.06583072 0.93416928)
              442) UHDL>=34.35 183   17 Y (0.09289617 0.90710383)
                884) CRECLR>=178.8365 5    2 Y (0.40000000 0.60000000) *
                885) CRECLR< 178.8365 178   15 Y (0.08426966 0.91573034)
                 1770) UHDL< 45.85 122   15 Y (0.12295082 0.87704918)
                   3540) UHDL>=41.85 38    9 Y (0.23684211 0.76315789)
                     7080) smoking=YS 15    6 Y (0.40000000 0.60000000)
                      14160) AGE>=63.22108 5    1 N (0.80000000 0.20000000) *
                      14161) AGE< 63.22108 10    2 Y (0.20000000 0.80000000) *
                     7081) smoking=NS 23    3 Y (0.13043478 0.86956522) *
                   3541) UHDL< 41.85 84    6 Y (0.07142857 0.92857143) *
                 1771) UHDL>=45.85 56    0 Y (0.00000000 1.00000000) *
              443) UHDL< 34.35 136    4 Y (0.02941176 0.97058824) *
          111) BNP>=82.35 679   20 Y (0.02945508 0.97054492) *
    7) GENDER=M 2843   53 Y (0.01864228 0.98135772)
     14) UHDL>=46.05 165   11 Y (0.06666667 0.93333333)
       28) UHDL< 48.55 59    8 Y (0.13559322 0.86440678)
         56) CRECLR< 82.97053 20    6 Y (0.30000000 0.70000000)
          112) CRECLR>=75.45391 7    3 N (0.57142857 0.42857143) *
          113) CRECLR< 75.45391 13    2 Y (0.15384615 0.84615385) *
         57) CRECLR>=82.97053 39    2 Y (0.05128205 0.94871795) *
       29) UHDL>=48.55 106    3 Y (0.02830189 0.97169811) *
     15) UHDL< 46.05 2678   42 Y (0.01568335 0.98431665) *
```

Tabela 19: *Një informacion numerik për pemen me variabël përgjegjës ALLCAD.*

library(rpart)
set.seed(18)
Do të përdorim funksionin rpart.control për të kontrolluar këto gjëra:

1. Parametrin e kompleksitetit "α", i cili jepet nga cp.
2. Minimumin e pemës fillestare e cila jepet nga minsplit.
3. Numri i palosjeve që do të përdoren në vleresimin e kryqezuar, i cili jepet nga xval.

my.control=rpart.control(cp = 0.00001, minsplit=15, xval=5)
Le të shikojmë pemën fillestare të outputeve tona se si duket? Jemi duke përdorur të 11 variablat për të klasifikuar pacientët.
Ne do të përdorim modelimin me anë të funksionit **rpart** Së pari variabli përgjegjës "ALLCAD" do të ndiqet nga simboli~dhe pastaj i vendosim të gjitha variablat parashikues të cilat do të na ndihmojnë të bëjmë parashikimet për të klasifikuar ato në cdo pacient e cila

pasqyrohet te variabli përgjegjës që në rastin tonë është "ALLCAD. Simboli i mësipërm që përdorëm tregon se duam ti përdorim të 11 variablat në këtë model. Pas kësaj, zgjedhim metodën që do të përdorim që në rastin tonë është Klasifikimi**'class'**. Përfundimisht shënojmë funksionin e kontrollit që krijuam më parë.
Në figurën 34, objektet pemë janë listuar si formula të tekstuara. Në fillim emërojmë variablat dhe mënyrat se si ato janë shpërndarë, numri i subjekteve të çdo nyje dhe pas kësaj numri i subjekteve të mosklasifikimit(humbjet) dhe klasa aktuale e çdo nyje. Simboli***** është përdorur për të dhënë nyjet fundore në këtë pemë. Aktualisht në këtë hap nuk gjejmë ndonjë gjë shumë të rëndësishme pasi duhet të kalojmë në hapin tjetër.
Më poshte po japim komandën e ndërtimit dhe tekstit të kësaj peme maksimale.
plot(ALLCAD1)
text(ALLCAD1)

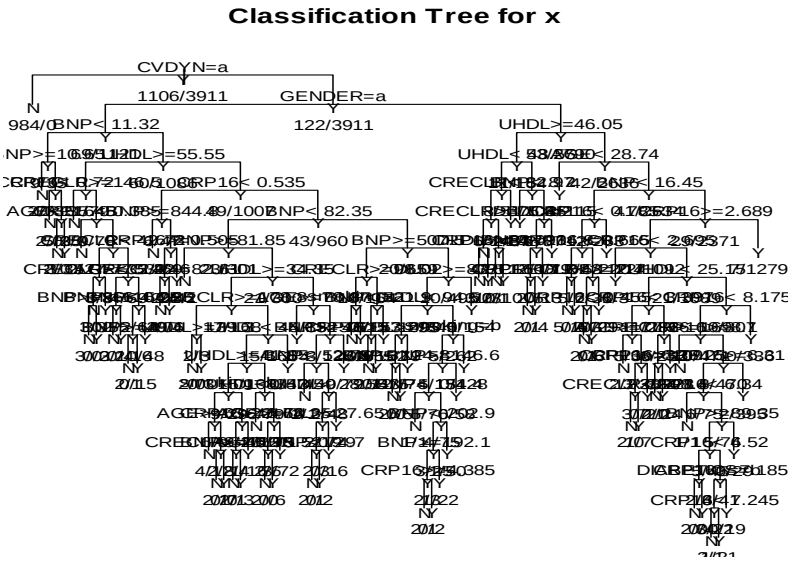


Figura 33: *Pema maksimale për variablin ALLCAD.*

Në figurën 34 pema fillestare duket se është e vështirë të lexohet pasi është shumë e madhe dhe kjo mund të konsiderohet si dicka normale pasi pema maksimale nuk mund të konsiderohet si pema optimale, pasi kjo pemë ka nevojë të krasitet për të arritur në pemën optimale.
Funksioni printcp jep tabelën e parametrit të kompleksitetit i cili është i konsideruar si një përmbledhje e modelit tonë. Ky ndihmon që të vendosim për përmasën e pemës optimale, e cila do të konsiderohet si pema më e mirë dhe ky proces kërkon që të kalojmë në një proces duke listuar të gjitha pemët nga përmasat më të vogla në pemët më të mëdhaja duke i parë këto të shoqëruara me parametrin e kompleksitetit, vlerësimi i kryqëzuar i cili vlerëson gabimin e mosklasifikimit , gabimin e marginimit dhe gabimin real **.**
Tabela=printcp(ALLCAD1).
table1 = printcp(CAD1)

```
Classification tree:
rpart(formula = ALLCAD ~ ., data = x, method = "class", control = my.control)

Variables actually used in tree construction:
[1] AGE     BNP     CRECLR  CRP16   CVDYN   GENDER  smoking UHDL

Root node error: 1106/5017 = 0.22045

n= 5017

          CP nsplit rel error  xerror      xstd
1 0.88969259      0   1.00000 1.00000 0.0265488
2 0.00054250      1   0.11031 0.11031 0.0098646
3 0.00030139      6   0.10759 0.11844 0.0102126
4 0.00022604     15   0.10488 0.12297 0.0104003
5 0.00001000     19   0.10398 0.13201 0.0107649
```

Tabela 20: *Tabela e parametrit të kompleksitetit për variablin ALLCAD.*

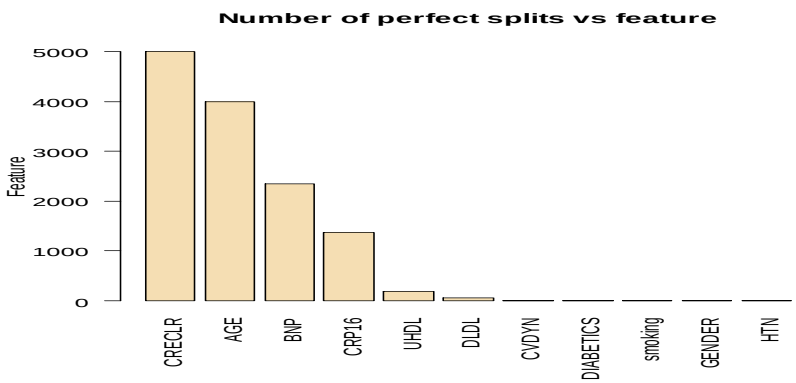


Tabela 21: *Renditja e variablave sipas rëndësisë.*

Nga tabela 21 shohim se jo të gjithë variablat luajë te njëjtin rol. Vëme re se renditja e variablave të treguar në grafikun e mësipërm nuk është domosdoshmërisht e njëte, gjë e cila do te pasqyrohet dhe ne modelin përfundimtar nga përzgjedhja e variablave për të ndërtuar pemën përfundimtare.

Nga tabela 20 shikojmë se gabimi në këte rast zvogëlohet kur përmasat e pemës rriten. Gabimi i vlerësimit të kryqëzuar në fillim fillon të zvogëlohet deri sa arrin minimumin kur pema ka një shpërndarje kur kur ka 6 nyje, dhe pastaje fillon të rritet në mënyrë të menjëhershme. Nga tabla 21 shikojmë se variabli parashikues CRECRL ka nje rendesi me te madhe dhe vjen renditja Age, BNP e keshtu me radhe. Mund të ndertojeme grafikun per parametrin e kompleksitetit duke përdorur vlerat e dhëna në tabelën 20 për funksionin plotcp. Ai gjithashtu ndihmon të vendosim për përmasën e pemës optimale duke vizualizuar vlerën e parametrit të kompleksitetit (x axes) përballë vlerësimit të gabimit të vlefshmërisë së kryqëzuar (y axes):

plotcp(ALLCAD1)

fit$cptable[which.min(fit$cptable[,"xerror"]),"CP"]

pfit<- prune(fit, cp=) # from cptable

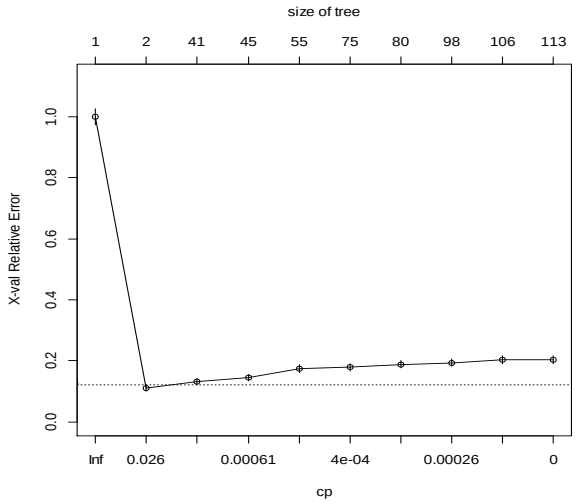


Figura 34: *Parametri i kompleksitetit për variablin ALLCAD.*

Nga figura 35 shikojmë se sygjerimi për pemën optimale është për pemën me 6-10 nyje, e cila arrihet kur a= cp≃0.0003.
Mund të marrim një vendim dhe të bëjmë krasitjen e pemës maksimale fillestare që morëm në fillim. Kjo mund të bëhet duke zgjedhur pemën me gabimin e vleresimit te kryqëzuar më të vogël dhe duke përdorur rregullin 1-SE i cili ka qënë i preferuar nga Breiman (1984) në librin e tij të parë për CART. Rregulli 1-SE mund të përdoret duke marrë vlerën më të vogël të vlersimi i kryqzuar, duke shtuar gabimin standart të veprimit duke gjetur kështu gabimin më të vogël të vlerësimit të kryqëzuar që e zvogëlon këtë numër. Në rastin tonë duke përdorur këtë rregull marrim(1-SE): 0.1101 + 0.0098 = 0.1199, kështu që rregulli 1-SE sygjeron se përmasa e pemës optimale është me gjashtë shpërndarje domethene me 7 nyje gjë cila arrihet te $a = cp = 0.00030139$.Kjo gjë është e ngjashme dhe me tabelën për parametrin e kompleksitetit.

Hapi tjetër është krasitja e pemës deri sa të marrim pemën optimale e cila është e përcaktuar nga 1-SE dhe rregulli që përdorëm më parë për vlersimin e kryqëzuar:

CAD2 =prune.rpart(CAD1, cp=0.000300139)
plot(CAD2)
text(CAD2)

Më poshtë është dhënë pema më e mirë e cila është një nënpemë e pemës tonë fillestare:

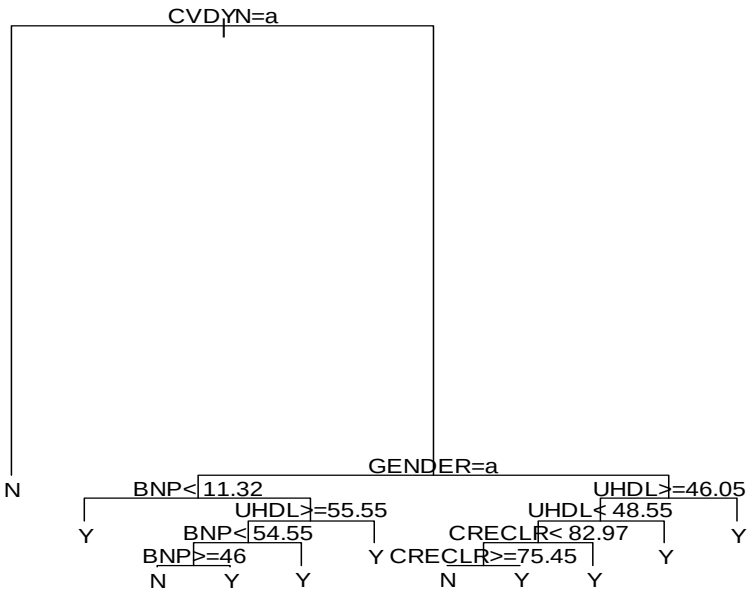


Figura 35: *Nënpema më e mirë e krasitur për variablin përgjegjës ALLCAD.*

Në figurën e mësipërme lehtësisht interpretohen faktet e kësaj baze të dhënash për variablin përgjegjs ALLCAD. Meqense interpretimi i pemës u bë në bazën e të dhënave nga spitali i Afrikes se Jugut(fq 103), lehtesisht dhe në të njëjtën mënyrë bëhet dhe interpretimi i informacionit të marrë dhe në këtë rast. Vlen të theksohet se meqënse baza e të dhënave nga spitali i Kleveland Klinik është më madhe, gabimi në nyjen rrënjë është më i vogël se në rasin e bazës së të dhënave nga spitali i Afrikës së Jugut.

Se dyti: Variabli përgjegjës CAD

Përshkrimi i bazës së të dhënave dhe një përmbledhje statistikore
Për marrjen e një peme për variablin përgjegjës CAD do të përdorim të njëjtat hapa si për të parën. Kështu fillojmë të ndërtojmë një pemë klasifikuese fillestare duke u bazuar në bazën e të dhënave e cila përmban 11 variabla dhe 5017 pacient dhe të ndara në dy kategori për variablin përgjegjës CVD (YN). Variablat përshkruhen si më poshtë:

1. BNP (variabël i vazhdueshëm).
2. CRP16 (variabël i vazhdueshëm).
3. DLDL (variabël i vazhdueshëm).
4. UHDL (variabël i vazhdueshëm).
5. DIABETICS (variabël kategorik), ku; ND=no=0 dhe YD=yes=1.
6. Smoking (variabël kategorik), ku; NS=no=0 dhe YS=yes=1.
7. AGE (variabël i vazhdueshëm).
8. GENDER (variabël kategorik), ku; M=mashkul dhe F=femër.
9. CRECLR (variabël i vazhdueshëm).
10. HTN (variabël kategorik), ku; NH=no=0 dhe YH=yes=1.
11. ALLCAD (variabël kategorik), ku; N=no=0 dhe Y=yes=1.

Pacientët mund të klasifikohen në dy klasa, dhe në vazhdim do të paraqitet një informacion për variablin CVDYN: ky variabël është etiketuar si: Cardio Vascular Anti-HTN Alpha-Blocker, Y=ypo=1, N=jo=0

Baza e të dhënave është etiketuar me y dhe është ruajtur nën faillin csv.

Një përmbledhje statistikore është dhënë në tabelen 11 kur filluam të analizojmë këtë bazë të dhënash. Mund të modifikojmë këtë komandë rpart.control function në mënyrë që të gjejmë rrugën për të ndërtuar këtë pemë si më poshtë:
my.control=rpart.control(cp = 0.00001, minsplit=15, xval=5)

Shënim: Ne do të përdorim të njëjtën bazë të dhënash dhe të njëtat library të sofwarit që përdorëm në pjesën e parë për variablin përgjegjës CAD. Nuk është e nevojshme të lexohet baza e të dhënave pasi është bërë më parë. Po ashtu do të përdorim pothuajse të njëtat kode si për pjesën e pare(duke ndryshuar variablin pergjegjes).

Hapi tjetër është që duhet të specifikojmë modelin që do të përdorim të fitojmë pemën maksimale për variablin përgjegjës CVD. Dhe tani do të përdorim të 11 ndryshoret të cilat i sqaruam më sipër se çfar përfaqësojnë për të fituar pemën klasifikuese me anën e tëcilës ne do tëklasifikojmë pacientët:
CVD1 =rpart(CVDYN ~ ., data=x, method='class',control=my.control)
Më poshtë është pema maksimale
plot(CVD1)
text(CVD1)

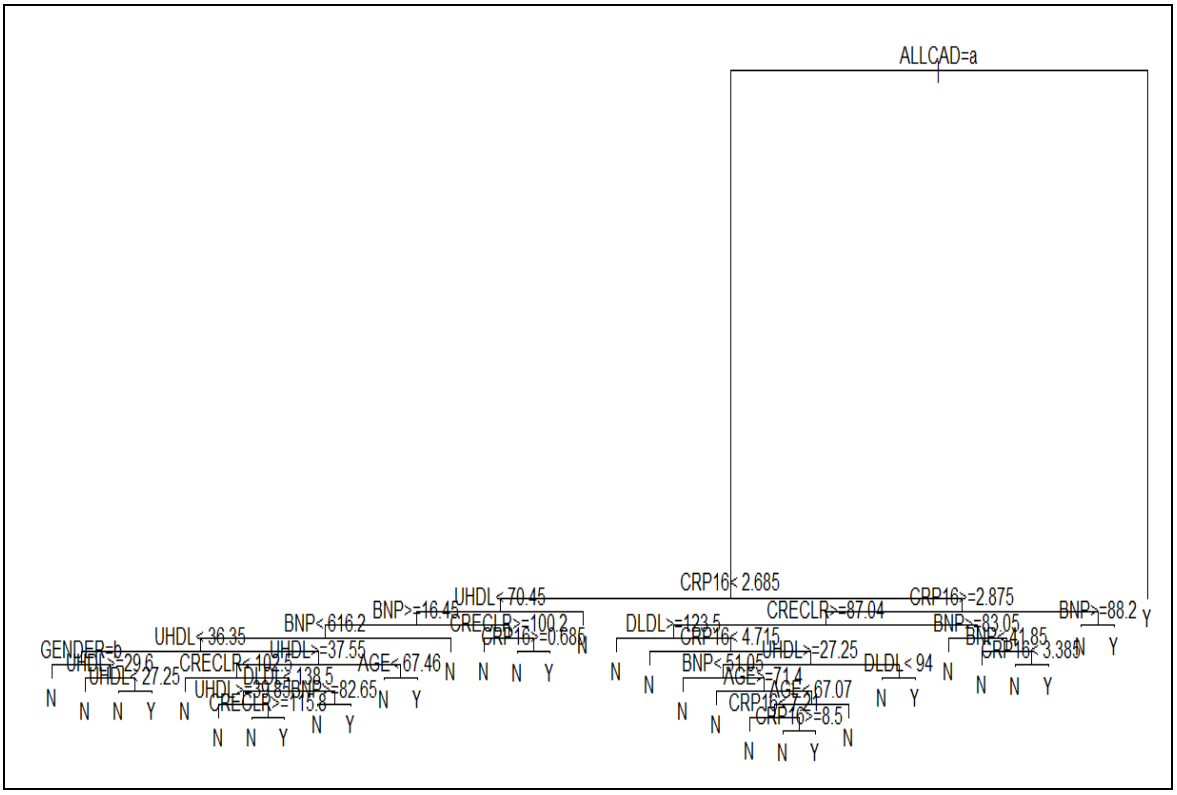


Figura 36: *Pema fillestare maksimale për variablin përgjegjës CVD*

Nga Figura 37 pema fillestare është një pemë shumë e madhe dhe është e vështirë ta lexojmë atë, por sikurse thamë më sipër kjo në një farë mënyre është normale pasi pema maksimale nuk është pema optimale të cilën do ta arrijmë pasi të bëjmë procesin e krasitjes. Le të

shikojmë tabelën e parametrit të kompleksitetit se çfarë duhet të kenë këto përmasa. Për të gjetur pemën më të mirë duke përdorur R kemi:

Table1 <- printcp(CVD1)

```
Classification tree:
rpart(formula = CVDYN ~ ., data = x, method = "class", control = my.control)

Variables actually used in tree construction:
[1] AGE    ALLCAD BNP    CRECLR CRP16  DLDL   GENDER UHDL

Root node error: 984/5017 = 0.19613

n= 5017

          CP nsplit rel error  xerror     xstd
1 0.87601626      0   1.00000 1.00000 0.028582
2 0.00101626      1   0.12398 0.12398 0.011088
3 0.00076220      8   0.11585 0.15142 0.012219
4 0.00045167     16   0.10976 0.19309 0.013740
5 0.00040650     25   0.10569 0.19309 0.013740
6 0.00033875     30   0.10366 0.19411 0.013775
7 0.00001000     33   0.10264 0.19715 0.013878
```

Tabela e parametrit të kompleksitetit për variablin përgjegjës CVD.
summary(x.rp)

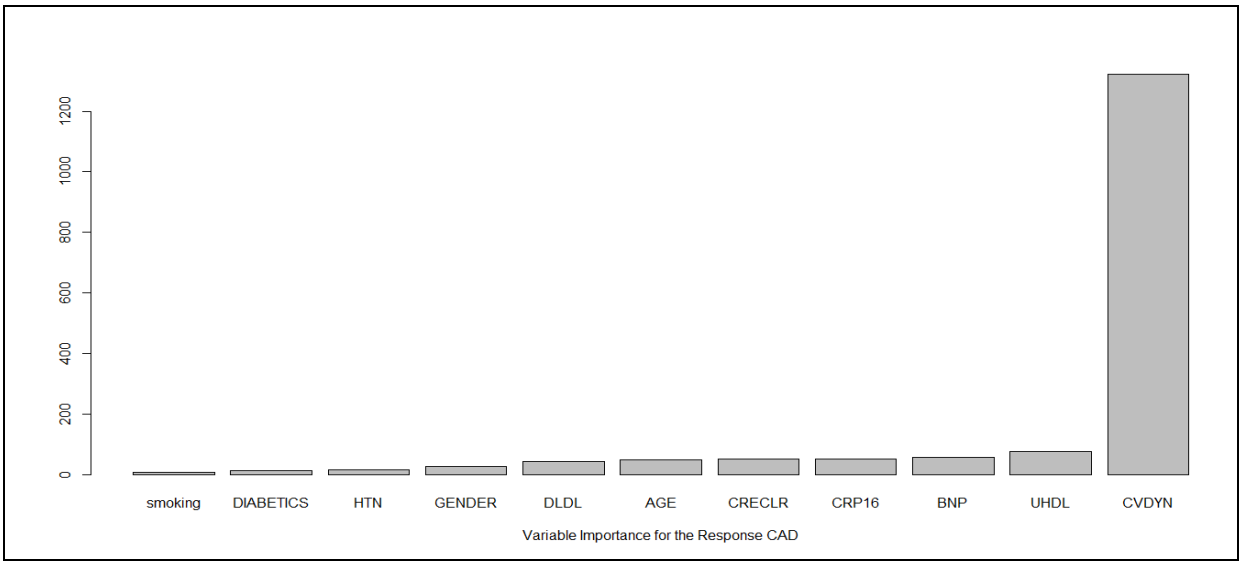


Tabela 22: *Renditja e variablave sipas rëndësisë për variablin CVD*

Call:
rpart(formula = ALLCAD ~ BNP + CRP16 + DLDL + UHDL + DIABETICS + smoking + CVDYN + AGE + GENDER + CRECLR, data = x, method = "anova", control = rpart.control(cp = 0.001))
n= 5017
CP nsplit rel error xerror xstd
1 0.862778999 0 1.0000000 1.0001378 0.01904362

2 0.001505864 1 0.1372210 0.1372879 0.01189652
3 0.001138310 2 0.1357151 0.1387293 0.01181936
4 0.001000000 3 0.1345768 0.1403248 0.01176509

Variable importance
CVDYN 99
Node number 1: 5017 observations, complexity param=0.862779
mean=1.77955, MSE=0.1718521
left son=2 (984 obs) right son=3 (4033 obs)
Primary splits:

CVDYN	splits as LR,	improve=0.86277900,	(0 missing)
UHDL	< 37.45 to the right,	improve=0.03777798,	(0 missing)
GENDER	splits as LR,	improve=0.03693704,	(0 missing)
DIABETICS	splits as LR,	improve=0.01964155,	(0 missing)
Smoking	splits as LR,	improve=0.01584180,	(0 missing)

Surrogate splits:
AGE < 32.34223 to the left, agree=0.804, adj=0.003, (0 split)
UHDL < 78.5 to the right, agree=0.804, adj=0.002, (0 split)

Node number 2: 984 observations
mean=1, MSE=0

Node number 3: 4033 observations, complexity param=0.001505864
mean=1.96975, MSE=0.02933535
left son=6 (1190 obs) right son=7 (2843 obs)
Primary splits:

GENDER	splits as LR,	improve=0.010974000,	(0 missing)
UHDL	< 35.75 to the right,	improve=0.009135748,	(0 missing)
BNP	< 12.95 to the left,	improve=0.005801406,	(0 missing)
AGE	< 35.41273 to the left,	improve=0.004035337,	(0 missing)
DLDL	< 90.5 to the right,	improve=0.002723678,	(0 missing)

Surrogate splits:

UHDL	< 45.95 to the right,	agree=0.735,	adj=0.101,	(0 split)
CRECLR	< 50.22452 to the left,	agree=0.712,	adj=0.024,	(0 split)
DLDL	< 179.5 to the right,	agree=0.707,	adj=0.006,	(0 split)
BNP	< 15152.45 to the right,	agree=0.705,	adj=0.002,	(0 split)
AGE	< 31.99589 to the left,	agree=0.705,	adj=0.001,	(0 split)

Node number 6: 1190 observations, complexity param=0.00113831
mean=1.942017, MSE=0.05462114
left son=12 (44 obs) right son=13 (1146 obs)
Primary splits:

BNP	< 11.32	to the left,	improve=0.015099120,	(0 missing)
UHDL	< 75	to the right,	improve=0.012608830,	(0 missing)
AGE	< 35.44832	to the left,	improve=0.012453070,	(0 missing)
CRP16	< 0.535	to the left,	improve=0.008230881,	(0 missing)
DIABETICS	splits as LR,	improve=0.006086626,		(0 missing)

Node number 7: 2843 observations

mean=1.981358, MSE=0.01829474

Node number 12: 44 observations
mean=1.795455, MSE=0.1627066

Node number 13: 1146 observations
mean=1.947644, MSE=0.04961487
Vihet re se vlera e gabimit të trajnimit zvogëlohet kur pema rritet, por vlerësimi i kryqëzuar zvogëlohet në fillim, arrin vlerën minimale kur pema ka një shpërndarje ku (a = cp = 0.00101626), dhe menjëherë fillon të rritet në mënyrë të menjëhershme.
Grafiku i parametrit të kompleksitetit paraqitur në Figurën 38 duke përdorur funksionin plotcp**.** Ai na ndihmon të vendosim se çfarë përmase për pemën do të zgjedhim për të marrë pemën optimale e cila është dhe më e mira.

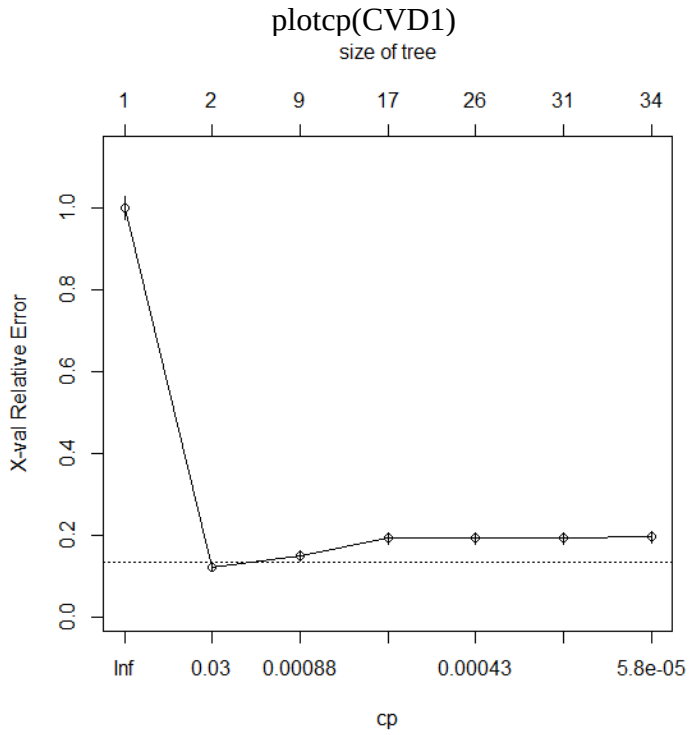


Figura 37: *Parametri i kompleksitetit për variablin CVD*

Kështu që vlerësimi i kryqëzuar sygjeron se përmasat e pemës optimale janë te pema me 8 shpërndarje ku vlera e a= cp≃0.0007.

Mund të marrim një vendim dhe të bëjmë krasitjen e pemës fillestare dhe maksimale. Për këtë do të përdorim rregullin 1-SE ku: 0.1239 + 0.011 = 0.1349. kështu që rregulli 1-SE sugjeron se pema me përmasa optimale është me 8 shpërndarje dhe ka nëntë nyje fundore dhe kjo arrihet për $a = cp = 0.0007622$. Kjo është e ngjashme me atë se çfare sygjeron dhe parametri i kompleksitetit.

Hapi tjetër është të krasitim pemën e figurës 37 për të arritur te pema me përmasa optimale duke përdorur rregullin 1-SE dhe rregullin e vlefshmërisë së kryqëzuar.
CVD2 <- prune. rpart(CVD1, cp=0.00076220)
plot(CVD2)

text(CVD2)

Figura 38 është një konfigurim për pemën optimale më të mirë të krasitur:

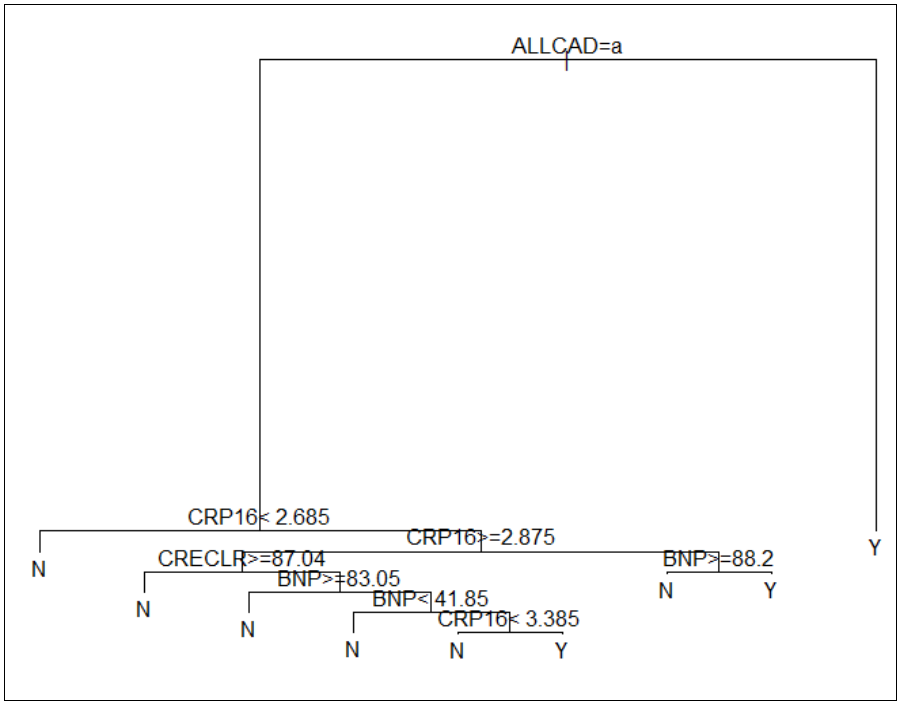


Figura 38: *Nen-pema më e mirë për variablin përgjegjës CVD*

Interpretimi i pemës përfundimtare të dhënë në figurën 38 është qartësisht i lehtë për tu bërë, pasi informacioni që kjo përmban nuk është i mbingarkur.

4.6 Përfundime

Megjithëse se CART mund të tregojë statistikisht se cilët faktorë janë veçanërisht të rëndësishëm në një model ose marrëdhënie në kuptimin e fuqisë shpjeguese dhe ndryshueshmërisë. Ky proces është matematik dhe është identik me disa teknika të regresionit të njohur, por paraqet të dhënat në një mënyrë që është me e lehtë për tu interpretuar nga ata që nuk janë të përgatitur mirë në analizat statistikore. Në këtë mënyrë, CART paraqet një pamje në formën e një peme e cila tregon marrëdhëniet e sofistikuara të variablave nga baza e të dhënave dhe mund të përdoret si një hap i parë në ndërtimin të një modeli informativ përfundimtar për disa të dhëna të rëndësishme, në rastin tonë faktorët që duhet të kontrollojmë në sëmundjet kardiovaskulare.
Në të ardhmen për të përmirsuar problemet e shëndetit publik, statisticienët mund të përdorin CART për të furnizuar mjekët me të dhëna paraprake, te cilat mund ti përdorin në parandalimin e përparimit të mëtejshëm të sëmundjeve dhe në marrjen e disa masave për çdo pacient që të parandalohet çdo e keqe për ata të cilvt kan këto probleme. Ky proces na jep një lidhje midis elementeve bazë si kolesteroli, duhanpirja, diabeti, historia familjare, alkoli dhe element te tjer klinikë të cilat duke u interpertuar statistikisht për të ndihmuar personelin mjeksor për të marrë masat paraprake dhe parandaluese për të ruajtur jetën e pacientëve dhe

për ta përmirësuar atë. Në shëndetin publik, megjithatë, kjo metodë e prezantimit nuk motivon praktikuesit pa një ekspertizë statistikore të cilët duhet të njohin mekanizmin e efektit shëndetësor për të përcaktuar klinikisht sa të rëndësishme janë dhe çfarë ndërhyrjesh efektive duhen bërë. Nga ana tjetër, nëse të dhënat janë shpjeguar thjesht pa një thellësi lidhjesh analitike ose duke bere përjashtime te variablave do të na conte në një drejtim i cili nuk është me rigorozitet shkencor, që mund të ketë pasoja negative për shumë pacientë. Nga kjo analizë statistikore (duke përdorur CART) duhet të vizualizojmë dhe të bëjmë një ndërlidhje dhe interpretim rigoroz statistikor duke paraqitur një model i cili është i vlefshëm dhe i interpretueshëm. Pema e Klasifikimit dhe Regresionit, dhe shembujt nga praktika klinike që ne kemi studiuar në këtë punim na jep mundësinë që të identifikojë pacientët me rrezik të lartë brenda 24 orëve nga pranimi në spital për një infarkt miokardi. Ky shembull provon se sa të rëndësishëm janë studimet që kanë dalë duke përdorur analizën e CART në mjediset klinike që vëzhgojnë infarktin e miokardit. Në disa raste këto studime kanë shumë variabla të cilat e komplikojnë situatën dhe ne nuk mund të parashikojmë saktësisht dhe në mënyrë të pavarur një rezultat të caktuar, të tillë si sulmi në zemër. Analiza CART mund ti drejtojë hulumtuesit mjekësorë për të izoluar cili nga këto variabla është më i rëndësishëm si një vend i mundshëm i ndërhyrjes.
Megjithëse CART është një metodë që po gjen një zbatim sa vjen dhe më të madh, përsëri kjo metodë ka avantazhet dhe disavantazhet e saj, të cilat janë renditur në këtë studim. Si në cdo punim statistikor një nga problemet që ndeshet është dhe ai i vlerave të munguara, e cila në këtë metodë zëvendësohet me vlera "surrogate". Një vënd të rëndësishëm në këtë kapitull zënë dhe testet për të parë në se variablat e ndryshme të bazës se të dhënave janë të varura apo të pavarura nga njëra tjetra, kanë shpërndarje normale dhe për këtë përdor testin Hi-katror si dhe përdorimin e disa paraqitjeve grafike. Në të gjitha testet e ndërvartësisë midis variablave, qartësisht shihet se ka një lidhje funksionale midis tyre, e cila pasqyrohet nga testet statistikore. Në këtë punim nuk janë pasqyruar të gjitha rezultatet e këtyre testeve, por është parë se përfundimet janë të njëjta për të gjithë variablat për të dy bazat e të dhënave.

Në shkencë, asnjë model nuk pranohet derisa të provohet vërtetësia e saj në botën reale. Shkencëtarët përdorin modele për të bërë parashikime dhe pastaj kryejnë teste kritike për të kontrolluar nëse këto parashikime ishin të sakta. Secili model duhet të specifikojë se cilat rrethana fizike janë të nevojshme dhe të parashikojnë se cilat të dhëna duhet të gjenden si rezultat. Modelet shkencore testohen duke bërë parashikime dhe duke kontrolluar ato, saktësia është një mase për vlerësimin e modeleve të klasifikimit. Informalisht, saktësia është pjesë e parashikimeve që ne nxjerrim në jetën reale. Formalisht, saktësia ka përkufizimin e mëposhtëm: Saktësia = Numri i parashikimeve të sakta/Numri i përgjithshëm i parashikimeve. Nga kjo formulë shikojmë se saktësia e pemës së klasifikimit dhe regresit duhet të provohet në jetën reale dhe nga studimet e deri tanishme është parë se rezultatet e CART janë relativisht të larta, por duhet theksuar se duhet punuar me kujdes, me intuitë të lartë dhe me një bashkëpunim të ngushtë midis statistikantit dhe mjekut specialist. Nje aspekt i rënëdesishëm është dhe numri i variablave nga baza e të dhënave të përdorura nga pema për të bërë parashikimet e duhura, sa më shumë variabla të përdoren aq më i mirë është parashikimi. Por duhet theksuar se rëndësia e variablave nuk është e njëjtë për çdo bazë të dhënash, të cilën e pamë edhe në tabelat 21 dhe 22. Në rastet e studiuara në këtë punim, për variablin CHD janë përdorur shtatë nga nëntë variabla parashikues te pema përfundimtare. Në shëmbullin e dytë janë përdorur 6 nga nëntë variablat parashikues. Natyrshëm lind pyetja përse nuk përdoren të gjitha variablat parashikues dhe në cilin rast saktesia është më lartë? Së pari disa variabla si mbipesha apo kolesteroli i mirë nuk kanë të njëjtën influencë në këto lloj sëmundjesh dhe së dyti në rastin e bazës së të dhënave të marra nga Klinika e spitalit të

Klevelandit ka informacion për pesë mijë pacientë, gjë e cila padyshim jep një informacion më të gjerë në softwarin R, po ashtu varet dhe çfarë madhësish janë matur në bazën e parë dhe cfarë janë matur në bazën e të dhënave, sa është vlera numerike mesatare për çdo madhësi në secilën bazë të dhënash. Një ndryshim esencial midis dy bazave të të dhënave është dhe gabimi që bëhet në nyjen rrënjë, sikurse shihet për bazën e të dhënave nga spitali i Kleveland Klinikes janë respektivisht 22% dhe 19% për dy variblat përgjegjëës dhe 34% për bazën e të dhënave nga spitali i Afrikës se Jugut, shihet se ato kanë relativisht ndryshime , gjë e cila çon në përfundimin se dhe saktësia në këtë bazë më të madhe të dhënash është më e madhe.

KAPITULLI 5

NJË VËSHTRIM I PËRGJITHSHËM I PEMËS SË REGRESIT

5.1 Pema e Regresit

Analiza e regresit me shumë variabla është një problem sa i njohur aq dhe i përdorur. Analiza e regresit mund të klasifikohet si një aplikim i metodave investigative të marrëdhënieve të variablave të varura dhe variablave të pavaruara me anë të të cilave bëjmë parashikimet e duhura. Një nga rastet me të cilin do të merremi në këtë studim është Tregui i shtëpive në Boston SHBA i cili paraqet pemën e regresit. Kohët e fundit janë duke u përdorur shëmbujt e ndryshëm të pemës së regresit në të cilat kemi më shumë se një variable si objektive, ka raste që mund të kemi më shumë se 5 apo 6 variabla si objektive. Një shembull interesant që përballet me problemin e regresit me shumë variabla përgjegjës është përdorur dhe te libri i parë i botuar në vitet 80 nga Breiman dhe bashkautorët e tjerë. Në këtë studim do të analizojme vetëm rastin kur kemi një variabël përgjegjës.

Si objektiv i metodës së regresit është që të fitojmë një model bazë me të dhënat në studim. Në bazën e të dhënave kemi një çift të renditur të formës $\langle x_i, y_i \rangle$ ku x_i është një vektor ku vlerat e të cilit do të përdoren si atribute parashikuese për variablin përgjegjës y_i. Në konteksin e analizës së regresit, matrica është përdorur për të thjeshtuar disa formulime. Le të marrim matricën e imputeve në të cilën një nga vlerat që ndodhet në rreshtin e i-të të vektorit x_i, nëse atje janë n vektor, X është matrica me dimensione n x a , ku a është numri i atributeve në bazen e te dhenave. Do të mbledhim vlerat e objektivit në dalje me vektor të formës matricore n x 1, Y. Në gjithashtu mund të prezantojmë bashkësinë e bazes se te dhenave D si matric D me përmasa n x (a+1). Mund të shikojmë se sistemi i regresit si një funksion që lidh bashkësinë e të dhënave D me një model regresii cili na jepet në këtë formë $r_D(\bullet)$. Modeli i regresit është një funksion që lidh vektorin hyrës $x_i \in X$ me numrin real $y \in Y$. Analiza e regresit ka si një nga shqetësimet kryesore vlerësimin ose parashikimin e vlerës mesatare të variablit të varur Y duke u bazuar në vlerat e variablit apo variablve të pavarur X_i $E(Y / X_1, X_2, ... X_a)$, ku E(.) jep pritshmërinë statistikore. Lidhja regresive e atributeve dhe vlerave të variablit të targetit e cila zakonisht përshkruhet nga relacioni i mëposhtëm: $y_i = r(\beta, x_i) + \varepsilon_i$ ku $r(\beta, x_i)$ është modeli i regresit me variabla hyrëse $\{X_i\}_{j=1}^{a}$ ku β është parametrik (sllopa) dhe ε_i gabimi i vrojtimit. Qëllimi kryesor i modelit të regresit është që të gjejë modelin me parametrin më të mirë β duke përdorur kriterin e selektimit. Në përgjithësi modelet e ndryshme të regresit kërkojnë një vlerësim sa më të mirë të parametrit β.

5.2 Matja e saktasisë së modeleve të regresit

Do të përdorim modelet e regresit për të përfituar një parashikues numerik për varaiablat përgjegjës. Kjo është e mundur nëse dimë vlerën e etiketuar të variablit të varur. Duke përdorur këtë vlerë mund ta krahasojmë atë me modelin parashikues dhe e klasifikojmë se si e bën paraqitjen. Meqënëse valiabli parashikues i modelit të regresit është numerik lehtësisht mund të gjejmë diferencen midis vlerës reale dhe parashikuesit. Vlera absolute

mesatare e devijimit e mat dhe e klasifikon gabimin në çdo model duke mesatarizuar vlerën absolute të gabimit mesatar të parashikimeve:

$$MAD(r) = \frac{1}{n}\sum_{i=1}^{n} |(y_i - r(\beta, x_i)|$$

ku $\{\langle x_i, y_i \rangle\}_{i=1}^{n}$ është data e dhënë, $r(\beta, x_i)$ është parashikusi i modelit të regresit të cilin duam ta vleresojmë për rastin $\langle x_i\, y_i \rangle$. Dhe në këtë situate do të shikojmë për modelin i cili jep gabimin më të vogël dhe matësi më i mirë i kesaj është metoda e katrorëve me te vegjel. Një gabim tjetër i përbashkët është dhe Gabimi mesatar relativ i katrorëve RMSE, që jepet si më poshtë: $RMSE(r) = (\frac{1}{n}\sum_{i=1}^{n}(y_i - r(\beta, x_i))^2 / (\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{\bar{y}})^2) = \frac{MSE(r)}{MSE(\bar{\bar{y}})}$ ku $\bar{\bar{y}}$ është mesatarja e vlerave të Y.

Kjo jep vlerën relative të gabimit. Një vlerë midis zeros dhe njëshit është një tregues i mirë i r-së gjë që tregon se është më mirë se sa parashimi i vlerës mesatare të Y.

Modeli i pemës së regresit është ndarje e vazhdueshme në nënbashkësi ku variabli përgjegjës ruan një marrëdhënie të caktuar me variablat e pavarura. Për variablat parashikuese mund të bëhet një kombinim i atyre të vazhdueshme me ato kategorike. Baza e te dhenave ndahet në mënyrë të vazhdueshme në nënbashkësi më të vogla deri sa modelet më të vogla mund të (e.g. $Y = \beta_0 + \beta_1 X + \epsilon$) kënaqin çdo pjesë sa do e vogël qoftë ajo. Ky është njëjti proces që u paraqit në pemën klasifikuese. Teknikat e përdorura janë të ngjashme me ato të përdorura në C*RT*.

Pema e regresit është një model jo linear e cila bën parashikimet e duhura duke bërë një kombinim të të gjitha variablave që jane dhënë në bazën e te dhenave, të cilat mund të jenë të vazhdueshme, diskrete dhe kategorike. Në këtë punim do të ndërtojme dhe analizojme pemën e regresit për bazën e të dhënave: Tregu e shtëpive në zonën e Bostonit ku variabli përgjegjës është i vazhdueshëm.

Baza e të dhënave Boston House Market ka 506 vrojtime me 14 variabla të cilat përshkruhen si më poshtë.

Qëllimi i studimit në këtë kapitull është të parashikojme çmimin e shtëpive në Boston (variabli i varur) me anë të pemës së regresit duke përdorur softwarin R.

Variablat	**Pershkrimi**
crim	per capita (crime rate by town). renditia e qyteteve sipas perqindjes se krimeve
zn	proportion of residential land zoned for lots over 25,000 sq.ft. Raporti i zonave rezidenciale per shtepite me mbi 25000 sq ft
indus	proportion of non-retail business acres per town. Perqindja e siperfaqes se pa shitshme per bizneset
chas	Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)

	Variacioni i varfërisë Charles River (= 1 nëse trakti i lumitt kufizohet; 0 ndryshe)
nox	nitrogen oxides concentration (parts per 10 million) përqendrimi i oksideve të azotit (pjesë për 10 milionë)
rm	average number of rooms per dwelling numri mesatar i dhomave për banesë
age	proportion of owner-occupied units built prior to 1940 përqindja e njësive të okupuara nga pronarët e ndërtuar para vitit 1940
dis	weighted mean of distances to five Boston employment centers mesataren e ponderuar e distancave për pesë qendrat e punësimit në Boston
rad	index of accessibility to radial highways indeksi i hyrjes në autostradat duke u nisur nda qendra
tax	full-value property-tax rate per $10,000 norma e pasurisë së taksës me vlerë të plotë për 10,000 dollarë
ptratio	pupil-teacher ratio by town raporti nxënës-mësues nper qytetin
black	*1000(Bk - 0.63)*^2where *Bk* is the proportion of blacks by town 1000 (Bk - 0.63) ^ 2 ku Bk është përqindja e zezakëve në qytet
lstat	lower status of the population (percent) statusi më i ulët i popullsisë (në përqindje)
medv	median value of owner-occupied homes in $1000s vlera e mesores se shtëpive të zëna nga pronaret në $ 1000s

Tabela 23: *Variablat për bazën e të dhënave "Boston House Market".*

1. Në fillim duhet të instalojmë disa nënprograme të Rit si MASS dhe rpart.
install.packages("MASS")
install.packages("rpart")
require(MASS)
require(rpart)
Note: rpart është i nevojshëm pasi mat inekuacionin statistikor i cili quhet keficenti Gini.
Duke lexuar datën "Boston House Market" marrim tablonë e mëposhtëme për të gjithë variablat e kësaj baze të dhënash.
Names (Boston)

```
[1] "crim"    "zn"      "indus"   "chas"    "nox"     "rm"      "age"
[8] "dis"     "rad"     "tax"     "ptratio" "black"   "lstat"   "medv"
```

Pas këtij proçesi vazhdojmë punën me softwarin për të ndërtuar modele të regresit duke përdorur në këtë model të rpart formulën e anoves, e cila jep pemën e regresit.

boston.rp=rpart (medv~., method="anova", data=Boston, control=rpart.control (cp=0.0001))
summary(boston.rp)
Dhe si rezultat i kësaj marrim tabelën komplekse numerike që ndihmon në prodhimin e pemës.

Complexity Table

	CP	nsplit	rel error	xerror	xstd
1	0.4527442007	0	1.0000000	1.0008441	0.08279329
2	0.1711724363	1	0.5472558	0.6415611	0.05826556
3	0.0716578409	2	0.3760834	0.4282302	0.04558908
4	0.0361642808	3	0.3044255	0.3524707	0.04251598
5	0.0333692301	4	0.2682612	0.3352451	0.04269041
6	0.0266129999	5	0.2348920	0.3201398	0.04284871
7	0.0158511574	6	0.2082790	0.2833742	0.04035666
8	0.0082454484	7	0.1924279	0.2716571	0.04230165
9	0.0072653855	8	0.1841824	0.2594427	0.03907483
10	0.0069310873	9	0.1769170	0.2556601	0.03828815
11	0.0061263349	10	0.1699859	0.2541351	0.03861434
12	0.0048053197	11	0.1638596	0.2392679	0.03424519
13	0.0045609248	12	0.1590543	0.2332381	0.03402908
14	0.0039410233	13	0.1544934	0.2264541	0.03388449
15	0.0033161209	14	0.1505523	0.2285377	0.03485257
16	0.0031206492	15	0.1472362	0.2311663	0.03650942
17	0.0022459421	16	0.1441156	0.2284981	0.03631381
18	0.0022354038	18	0.1396237	0.2279692	0.03615274
19	0.0021720940	19	0.1373883	0.2279692	0.03615274
20	0.0019335691	20	0.1352162	0.2300202	0.03624752
21	0.0017169079	21	0.1332826	0.2295729	0.03616074
22	0.0014440359	22	0.1315657	0.2281099	0.03614753
23	0.0014098099	23	0.1301217	0.2291503	0.03615862
24	0.0013635419	24	0.1287119	0.2282475	0.03615277
25	0.0012778421	25	0.1273483	0.2279400	0.03613419
26	0.0012473645	26	0.1260705	0.2274864	0.03614715
27	0.0011372540	28	0.1235757	0.2279233	0.03615827
28	0.0009633247	29	0.1224385	0.2272594	0.03615926
29	0.0008486865	30	0.1214752	0.2269299	0.03612045
30	0.0007098930	31	0.1206265	0.2260585	0.03614413
31	0.0005879326	32	0.1199166	0.2274213	0.03612924
32	0.0005115280	33	0.1193287	0.2299455	0.03615310
33	0.0003794039	34	0.1188171	0.2306533	0.03615437
34	0.0003719398	35	0.1184377	0.2303838	0.03613785
35	0.0003438561	36	0.1180658	0.2302506	0.03613847
36	0.0003311450	37	0.1177219	0.2302972	0.03613873
37	0.0002274363	39	0.1170596	0.2300330	0.03614484
38	0.0001955788	40	0.1168322	0.2301039	0.03614407
39	0.0001000000	41	0.1166366	0.2304656	0.03615669

Tabela 24: *Parametri i kompleksitetit të bazës së të dhënave.*

Nga tabela e meposhteme veme re se niveli i rrënjës numri 1, para se të bëhet shpërndarja atje janë 506 vrojtime. Gjithashtu gabimi i katrorëve me te vegjel është 84.42 dhe mesatarja për të gjithë bazen e te dhenave është 22.53.

```
Node number 1: 506 observations,    complexity param=0.4527442
  mean=22.53281, MSE=84.41956
  left son=2 (430 obs) right son=3 (76 obs)
  Primary splits:
      rm      < 6.941    to the left,  improve=0.4527442, (0 missing)
      lstat   < 9.725    to the right, improve=0.4423650, (0 missing)
      indus   < 6.66     to the right, improve=0.2594613, (0 missing)
      ptratio < 19.9     to the right, improve=0.2443727, (0 missing)
      nox     < 0.6695   to the right, improve=0.2232456, (0 missing)
  Surrogate splits:
      lstat   < 4.83     to the right, agree=0.891, adj=0.276, (0 split)
      ptratio < 14.55    to the right, agree=0.875, adj=0.171, (0 split)
```

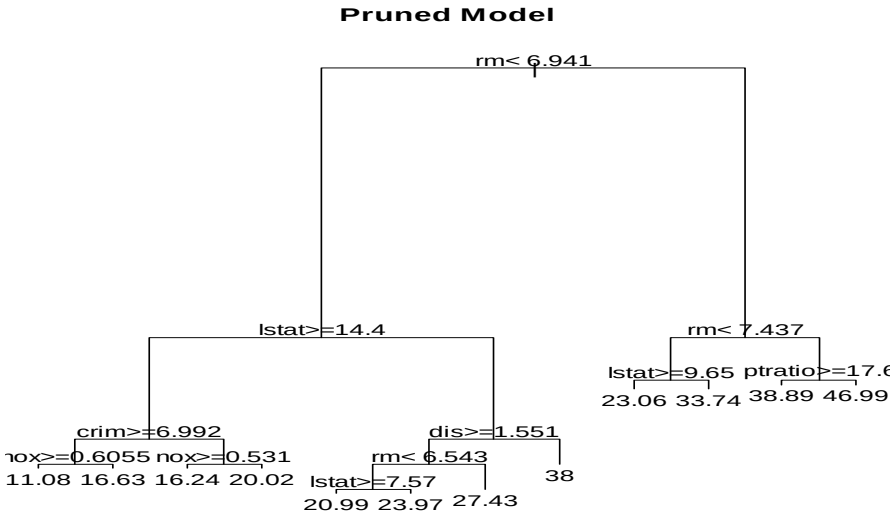
Pema më madhe me 39 nyje fundore ka gabimin më të vogël të raportit të vlerësimit të kryqëzuar. Sidoqoftë kjo pemë është shumë e madhe që të bëjmë parashikime, prandaj për të përmiresuar këtë në fillim bëjmë një shpërndarje të parë duke u bazuar në mesataren e numerit te dhomave. Nëse një shtëpi ka më pak se 7 dhoma, atëherë vrojtimi shkon në të majtë, ndryshe shkon në të djathtë të pemës. Së dyti bëjmë një shpërndarje duke u bazuar në statusin e ulët të popullsisë. Nëse numri i dhomave është i panjohur atëherë statusi ulët i popullsisë mund të përdoret për shpërndarjen dhe në këtë rast në vlerën 9.725.

5.3 Krasitja

Duke u kthyer në tabelën 21, norma më e ulët e gabimit është në një pemë me 27 nyje, por për shkak se pema me 12 nyje fundore është brenda një gabimi me standarde minimale, pema më e vogla me 12 nyje fundore është e mjaftueshme. Krasitja e kësaj peme mund të bëhet duke zgjedhur një vlerë në tabelën komplekse që është më e madhe se ajo e prodhuar për pemën optimale (pemë me 12 nyje) por më pak se vlera e kompleksitetit të pemës mbi atë (pemë me 11 nyje). Këtu, kemi nevojë për një pemë me parametër kompleksiteti nga 0.0048 në 0.0061.

Bëjmë krasitjen e pemës duke përdorur kodin e mëposhtëm:

```
boston.prune=prune(boston.rp,cp=0.005)
plot(boston.prune,main=main="Pruned Model")
text(boston.prune)
```



Figurë 39: *Pema bazë e krasitur duke u bazuar në rregullin SE*

Figura 40 tregon shpërndarjen primare në numrin e dhomave për çdo shtëpi (rm < 6.941). Ndarja e dytë në të majtë duket e rëndësishme për sa i përket aftësisë së modeleve të ndarjes e të dhënave për të reduktuar shumat e mbetura të shesheve. Shtëpitë e shtrenjta kanë tendencë që të kenë një numër mesatar të dhomave më të madh. Shtëpitë më të lira kanë më pak dhoma (<7 në mesatare) dhe një status të ulët në popullatën me një shkallë të lartë të krimit.

5.4 Krasitja interaktive

Një tabelë komplekse e cila mund të ndihmojë në përcaktimin e madhësisë së pemës së krasitur, duke marrë në kosideratë raportin e të gjitha nyjeve me numrin e nyjeve fundore.

```
plotcp(boston.rp, minline=TRUE, lty=3, col=1, upper="size")
```

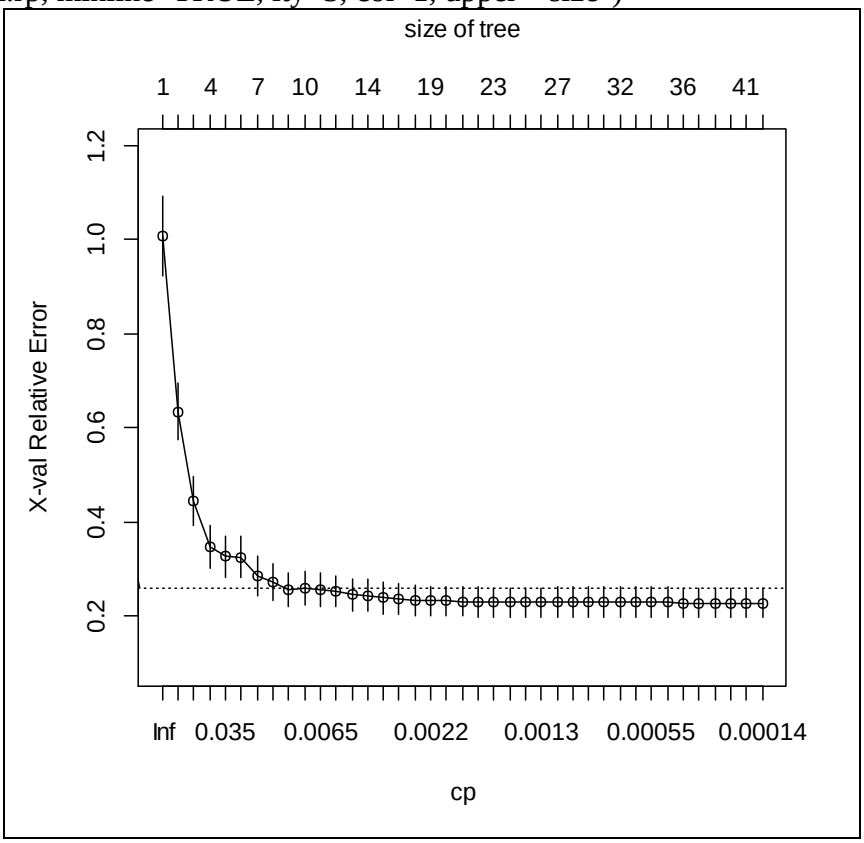


Figura 40: *Grafiku i kompleksitetit për të bërë krasitjen me vlerësimin e kryqëzuar*

Figura 40 tregon se vlefshmeria e kryqëzuar sygjeron një pemë optimale të madhësisë në mes të shtatë dhe të katërmbëdhjetë nyjeve fundore. Zgjidhet një pemë me nëntë nyje fundore, kështu që kjo mund të përshtatet me këtë model.

Një pemë mund të krasitet në mënyrë interaktive në disa metoda. Më poshtë japim kodin që duhet të përdorim për të bërë krasitjen e kesaj peme dhe me këtë numër të caktuar të nyjeve fundore i cili i plotëson kushtet që duhen në modelin tonë.

```
boston.prune.int=snip.rpart(boston.prune,toss=c(8,9,20))
plot(boston.prune.int,uniform=T,branch=0.1,main= "Interactive Pruning")
text(boston.prune.int,pretty=1,use.n=T)
```

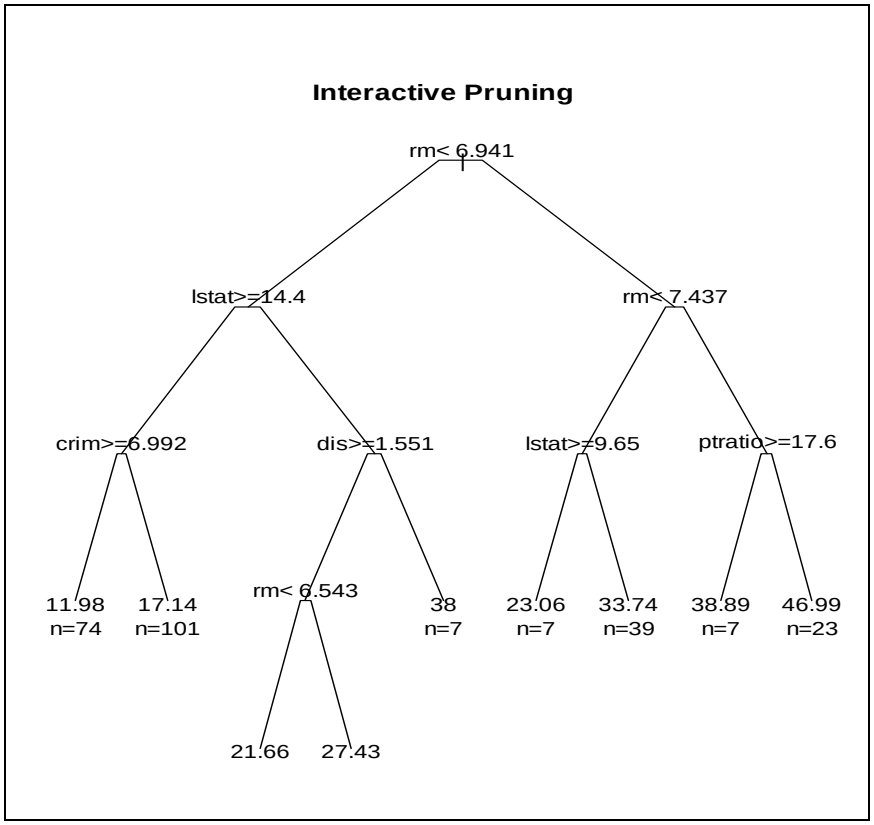


Figura 41: *Pema B – Rezultati i nje krasitje interaktive*

Krasitja interaktive e pemës më poshtë përdor variablat rm, lstat, crim, dis, dhe ptratio për të përcaktuar shpërndarjen
meanvar(boston.prune.int)

Ne Figura 43 paraqitet grafiku i Mesatare-Variancë në boshtin e x-ve është vendosur mesatareja e variablit përgjegjës dhe në boshtin e y-ve mesatarja e devijimit.

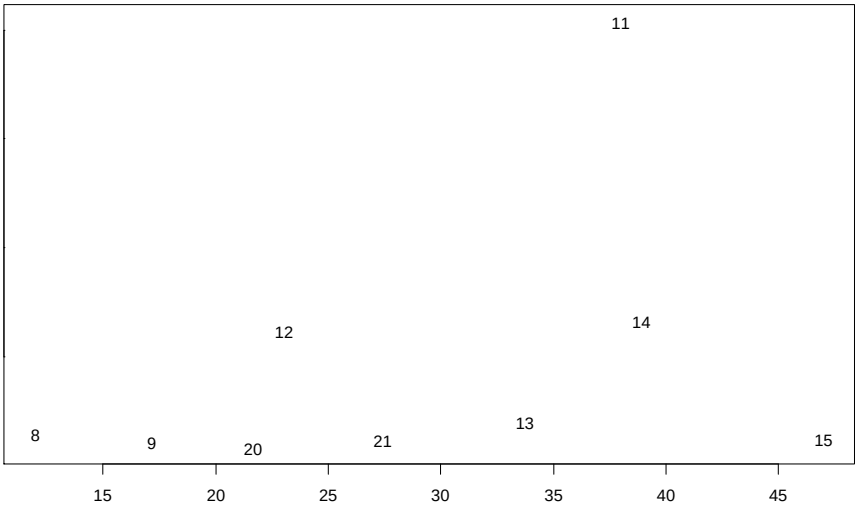


Figura 42: *Mesatare -Variancë*

Parashikimet

Examine the predictions from both tree models using the predict function.

Model 1: for Tree A

```
boston.pred1=predict(boston.prune)
```

Model 2: for Tree B

```
boston.pred2=predict(boston.prune.int)
```

Compute the correlation matrix of predictions with the actual response.

```
boston.mat.pred=cbind(Boston$medv,boston.pred1,boston.pred2)
boston.mat.pred=data.frame(boston.mat.pred)
names(boston.mat.pred)=c("medv","pred.m1","pred.m2")
cor(boston.mat.pred)
```

	medv	pred.m1	pred.m2
medv	1.0000000	0.9144071	0.9032262
pred.m1	0.9144071	1.0000000	0.9877725
pred.m2	0.9032262	0.9877725	1.0000000

Matrica e korrelacionit e mësipërme tregon se parashikimet në mes të modeleve 1 dhe 2 janë të lidhura shumë me përgjigjen.

Model 1 tregon se parashikimet janë pak më të mirë se parashikimet në modelin 2 .

Parashikimet mund të gjenerohen duke përdorur kodin e mëposhtëm:

```
par (mfrow=c(1,2),pty="s")
with(boston.mat.pred, {
eqscplot(pred.m1,      medv,  xlim=range(pred.m1,pred.m2),ylab="Observed",
xlab="Predicted", main="Model 1")
abline(0,1,col="blue",lty=5)
eqscplot(pred.m2,medv,xlim=range(pred.m1,pred.m2),ylab="Observed",     xlab="Predicted",
main="Model 2")
abline(0,1,col="blue",lty=5)
par(mfrow=c(1,1))
})
```

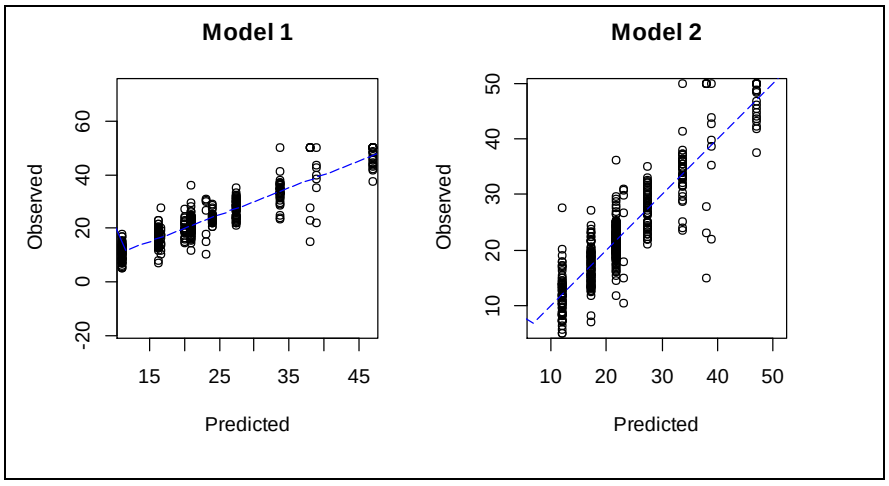


Figura 43: *Modeli i vrojtuar vs Modeli i parashikuar*

Figura 44 tregon se të dyja modelet janë shumë të mira për të bërë parashikimet e duhura për vlerën e medianes për çmimin e shtëpive në Boston. Por nëse e shikojmë me kujdes mund të themi se modeli 1 është pak më i mirë për të bërë parashikime duke u krahasuar me modelin

2. Çfarë ndodh nëse në bazën e të dhënave na mungon variabli rm dhe duam të parashikojmë çmimin e shtëpive? Mund të krijojmë një pemë të regresit duke përdorur mesataren e dhomave duke e konsideruar (rm) si një variabël të hequr.

```
boston.rp.omitRM=update(boston.rp,~.-rm)
summary(boston.rp.omitRM)
```

```
           CP nsplit rel error    xerror       xstd
1  0.4423649998      0 1.0000000 1.0040176 0.08319221
2  0.1528339955      1 0.5576350 0.6009111 0.04900741
3  0.0627501370      2 0.4048010 0.4382420 0.04057083
4  0.0400760532      3 0.3420509 0.4188052 0.04137352
```
…

Examine the first node.
Node number 1: 506 observations, complexity param=0.442365
mean=22.53281, MSE=84.41956
left son=2 (294 obs) right son=3 (212 obs)
Primary splits:

lstat	< 9.725	to the right,	improve=0.4423650,	(0 missing)
indus	< 6.66	to the right,	improve=0.2594613,	(0 missing)
ptratio	< 19.9	to the right,	improve=0.2443727,	(0 missing)
nox	< 0.6695	to the right,	improve=0.2232456,	(0 missing)
tax	< 416.5	to the right,	improve=0.2017517,	(0 missing)

Surrogate splits:

Indus	< 7.625	to the right,	agree=0.822,	adj=0.575,	(0 split)
nox	< 0.519	to the right,	agree=0.802,	adj=0.528,	(0 split)

Qellimi kryesor i shpërndarjes tani është në lstat dhe shpërndarja e dorës së dytë, shpërndarjet janë indus dhe nox. Kurr m është harruar atëherë modeli i ri i përdorur në kompletimin e shpërndarjes nga modeli origjinal do të bëjë shpërndarjen e parë.

5.5 Testimi i paraqitjes

Për të gjetur një vlerësim sa më real të modelit të paraqitur, në mënyrë rastësore e ndajmë bazën e të dhënave në bashkësi, të cilat do ti përdorim për trajnim dhe pas kësaj, përdorim këtë bashkësi për të krijuar modelin të cilin duhet të vlerësojmë.

```
set.seed(1234)
n=nrow(Boston)
```

Për shembullin tonë 80% të bazës së të dhënave do ta përdorim si material për trajnim dhe pjesën tjetër prej 20% do të jetë bashkësia e testit.

```
boston.samp=sample(n,round(n*.8))
bostonTrain=Boston[boston.samp,]
bostonTest=Boston[-boston.samp,]
```

Më poshtë është funksioni i cili do të prodhojë MSE për modelin tone.

```
testPred=function(fit,data=bostonTest){
#MSE for performance of predictor on test data
testVals=data[,"medv"]
```

```
predVals=predict(fit,data[,])
sqrt(sum((testVals - predVals)^2)/nrow(data))
}
```

Vlera eMSE për modelin e mëparshëm të krasitur është 3.719.

```
testPred(boston.prune,Boston)
[1] 3.719268
```

Duke llogaritur MSE për modelin tone, ku baza fillestare e të dhënave që kemi përdorur është Bostonë. Vlerësimi MSE është 3.719268, e cila është një normë e rizëvëndësimit të gabimit.

Montojmë përsëri modelin në bashkësinë e bazës së trajnimit dhe duke shqyrtuar tabelën e kompleksitetit e cila tregon se modeli më i mirë i bazuar në një rregull të gabimit standart është një pemë me shtatë nyje terminal. Vija e kuqe në të gjithë figurën e mëposhtme paraqet rregullin 1 -SE.

```
bostonTrain.rp=rpart(medv~.,data=bostonTrain,method="anova",cp=0.0001)
plot(bostonTrain.rp)
```

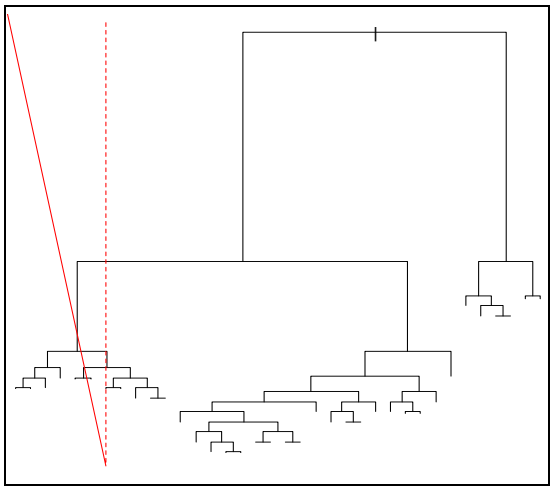


Figura 44: *Pema e regresit duke përdorur rregullin 1 -SE*

Dhe tani mund të bëjmë krasitjen e pemës së tranimit.

```
bostonTrain.prune=prune(bostonTrain.rp, cp=0.01)
plot(bostonTrain.prune, main= "Boston Train Pruning Tree")
text(bostonTrain.prune)
```

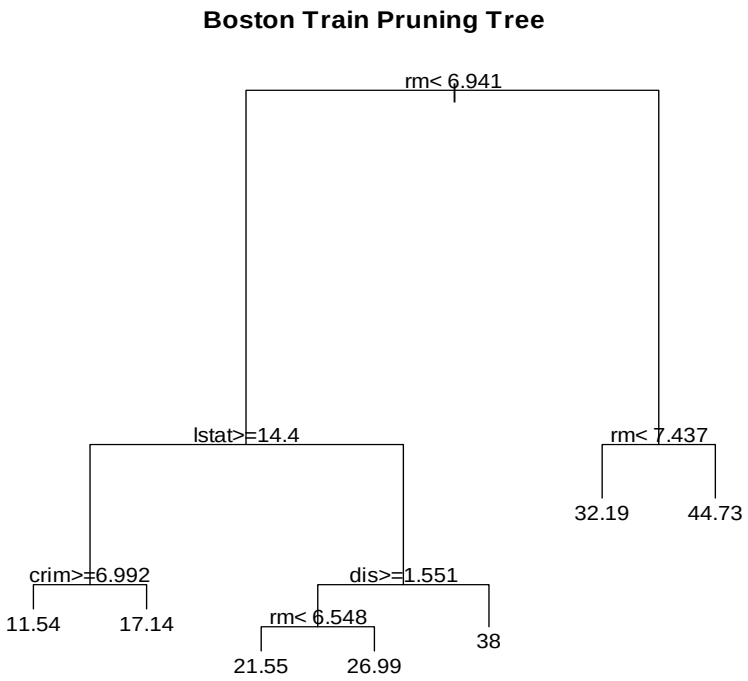


Figura 45: *Pema e krasitur e regresit për bazën e të dhënave "Boston House Market".*

Nga pema e më siperme shikojmë se: Shtëpitë me më shumë dhoma do të vlejënë më shumë. Çmimet e shtëpive janë në varësi porpocionale me numerin e dhomave. Lagjet me më shumë punëtorë të klasës më të ulët (vlera më e lartë "LSTAT") do të vlejnë më pak. Nëse përqindja e studentëve ndaj mësuesve është në raport me njerëzit është më e lartë, është e mundur që në këto lagje të ketë më pak shkolla, kjo mund të jetë sepse ka më pak të ardhura tatimore që mund të jenë sepse në atë lagje njerëzit fitojnë më pak para. Nëse njerëzit fitojnë më pak para atehere edhe shtëpitë e tyre të vlejnë më pak.

Gabimi mesatar i katrorëve për bazën e të dhënave është 4.06 dhe vlera e MSE për këtë bazë të dhënash është 4.78. Kjo vlerë e MSE është e perafert me gabimin mesatar të katrorëve.

```
testPred(bostonTrain.prune, bostonTrain)
[1] 4.059407
testPred(bostonTrain.prune, bostonTest)
[1] 4.782395
```

Parashikimi përformancës së modelit mund të testohet përmes grafikut të vlerave të vrojtuara me vlerat e parashikuara.

```
bostonTest.pred=predict(bostonTrain.prune, bostonTest)
with(bostonTest,{
cr=range(bostonTest.pred, medv)
eqscplot(bostonTest.pred, medv, xlim=cr, ylim=cr, ylab="Observed", xlab="Predicted", main="Test Dataset")
abline(0,1,col="blue", lty=5)
})
```

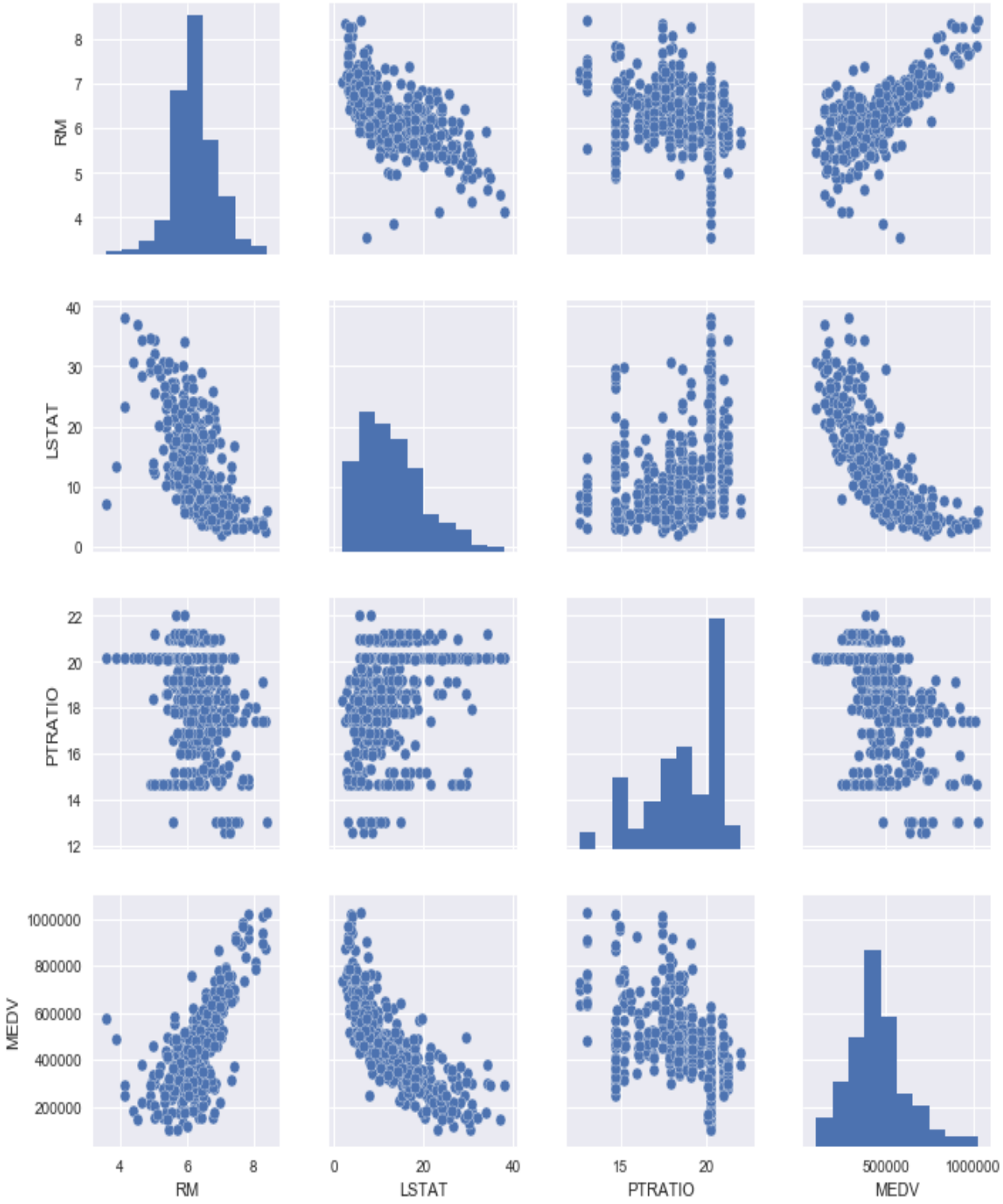


Figura 46: Skaterplot dhe Histogram

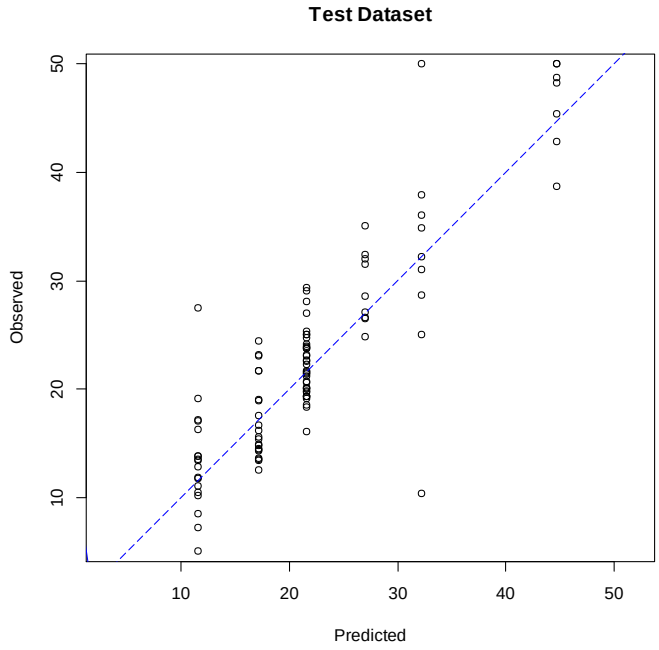


Figura 47: *Skaterplot për çmimimet e vrojtuara vs. te parashikuara.*

Figura 46 tregon se përformimimi i modelit parashikues është një tregues i mirë për çmimin e shtëpive në tregun e shtëpive në Boston, pasi elementët e saj shtrihen pothuaj se afër kësaj vije.

5.6 Përfundime

Klasifikimi dhe regresi me anë të pemës (CART), përdor një kombinim të kërkimeve dhe teknikave kompjuterike të testimit të cilat zbulojnë modele të rëndësishme dhe marrëdhëniet e fshehura në këto të dhëna. Ai mund të zbatohet pothuajse për çdo bazë të dhënash. Për një bazë të dhënash , kur nuk kemi asnjë ide se si vazhdohet me analizën, thjesht mund të përdorim metodën CART dhe ky program do të ndihmoj që të marren përfundimet e duhura. A mundet me të vërtetë që CART të japë rezultate të dobishme dhe te besueshme? Përgjigja befasuese është po. Në këtë studim marrim rezultate të dobishme për variablat që janë të rëndësishme dhe me nivelin e rëndësisë α=0.05 për sëmundjet e zemrës. Kur analiza automatike CART krahasohet me regresin logjistik ose me analizën e përcaktorit, CART zakonisht punon rreth 10% deri në 15% më mirë në shembujt që përdorim për të mësuar? Paraqitja e CART në rastet që përdorim për testim është shumë e rëndësishme. CART nuk varet nga mangësitë statistikore që kanë teknikat konvencionale hap pas hapi. Aanaliza automatike e CART krahasohet me modelet më të mira parametrike të ekipeve të sofistikuara të statisticienëve, CART është ende konkurruese. CART shpesh mund të gjenerojë modele në një orë ose dy që janë vetëm më pak te sakta krahasuar me modele që kërkojnë disa ditë për tu ndërtuar. Klasifikimi dhe regresi me anë të pemës pasqyron këto dy anët, duke mbuluar

përdorimin e pemëve si një metodë e analizës së të dhënave, dhe në një kuadër më matematikor, duke dëshmuar dhe provuar disa nga teorite themelore matematikore.
Gjithashtu në këtë punim, paraqesim paketa dhe algoritme të ndryshme për ndërtimin e pemës se klasifikimi dhe regresit, të cilat zbatohen si për pemët e klasifikimit dhe regresit. Për momentin algoritmet që përdoren nuk e mbështesin paralelizmin. Megjithatë, synohet të zgjerohet gama e algoritmeve të përdoruara për të arritur në të njëtat përfundime. Kjo jep një garanci dhe siguri më të lartë për efektivitetin e kësaj metodologjie në jetën e përditshme. Krahasimet me metodat e ndarjes rekursive "rpart", "tree" tregojnë se rpart përformon shumë mirë në një shumëllojshmëri të gjerë të cilësimeve, shpesh duke balancuar saktësinë parashikuese dhe kompleksitetin më mirë se metodat e kërkimit te përdorura në periudha të ndryshme. Në kapitullin 2, krahasohen parametrat të ndryshëm për algoritmet e ndryshme. Mund të vërehet se zgjedhja e veçantë e probabiliteteve të operatorit të variacionit është mjaft e fuqishme, me kusht që vëllimi i zgjedhjes të jetë mjaft i madhe. Cilësimet e parazgjedhura në numrin e iteracioneve dhe madhësisë së popullsisë janë të mjaftueshme për shumicën e grupeve të të dhënave me kompleksitet të mesëm. Megjithatë, për skema shumë komplekse të të dhënave, një rritje në numrin e iteracioneve ose vëllimi i zgjedhjes, mund të përmirësojë dukshëm performancën parashikuese të funksioneve të ndryshme. Qëllimi i përdorimit të algoritmeve të ndryshme, nuk është të zëvendësojë algoritme të mirë-përcaktuara për rpart apo tree, por më tepër të plotësojë me një gamë më të gjerë mënyrat për ndërtimin pemëve me një metodë alternative e cila mund të kryejë me një kohë të mjaftueshme. Nga natyra e algoritmit jemi në gjendje të zbulojme modele të cilat mund të modelohen nga një algoritëm i cili ka saktësi më të lartë. Ndërsa modelet mund të jenë në thelb të ndryshme nga modelet e pajisura në mënyrë te vazhdueshme, ku në përgjithësi mund të jetë më e dobishme të përdoren të dy qasjet, pasi kjo mund të zbulojë lidhje të reja midis të dhënave. Një përfundim i rëndesishem është se sa më madhe të jetë baza e të dhënave aq më të mira janë rezultatet përfundimtare për gjetjen e pemës më të mirë, e cila gëzon dhe cilësi më të lartë në parashikimet e bëra.
Ndryshimi në strukturën e pemëve të vendimit mund të çojë në dallime në klasifikim, edhe kur sigurohet me inpute të barabartë. Kjo tregon se gjetja e strukturës optimale për një pemë vendimi mund të jetë një hap i rëndësishëm në krijimin një algoritmi të klasifikimit. Gjithashtu, edhe pse përformanca në të dy rastet, duke përdorur të dy paketat, pemët janë identike në bashkësinë e testimit, por ka akoma shumë që mund të thuhet në lidhje me dallimet e algoritmeve duke analizuar klasifikimet e tyre.
Përformanca e dy pemëve vendimtare ishte e barabartë në të dy bazat e të dhënave të përdorura në këtë material studimi për pemën e klasifikimit. Nga kjo mund të konkludohet se struktura nuk ka rëndësi. Megjithatë, kur rezultatet e testimit krahasohen midis tyre për dy bazat e të dhënave, pa dyshim që sa më e madhe të jetë baza e të dhënave, sa më kujdesshëme të zgjidhet vlera e cp aq më të sakta do të jenë rezultatet përfundimtare të pemes optimale, gjë e cila kërkon kujdes, kembëngulje ne aplikimin e kujdesshem të alogoritmeve dhe paketave të ndryshme. Kur saktesia e vleres së cp është e lartë në mund të shikojemë se përformanca e pemës së vendimit është me e saktë dhe ka një përformanc shumë më të mirë në gjetjen e pemës më të mirë. Në përfundim: Pema e vendimit me bazë më të madhe të dhënash do të ishte zgjedhja më e mirë për të bërë parashikime me një saktësi me të lartë.

Biblografia

1. L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, Classification and Regression Trees, 1984, Chapman & Hall.
2. Applied Multivariate Statistical Analysis by Richard A. Johnson and Dean W. Wichern.
3. http://www.r-project.org/
4. http://www.statmethods.net/advstats/cart.html
5. http://www.youtube.com/watch?v=_RxqyvRK0Rw&feature=bf_prev&list=PL50858E6E9391F981
6. http://www.youtube.com/watch?v=m3mLNpeke0I
7. http://www.youtube.com/watch?v=f0eCYQY4gcQ&feature=related
8. http://plantecology.syr.edu/fridley/bio793/cart.html
9. http://www.statsoft.com/textbook/classification-and-regression-trees/
10. http://www.redbrick.dcu.ie/~noel/R_classification.html
11. Kuhnert, Perta, and Bill Venables. " Tree-based Models II." An Introduction to R: Software for Statistical Modeling & Computing. Cleveland, Australia: CSIRO Mathematical and Information Sciences. 283-296. Scribd.Web. 18 Apr. 2012. <http://www.scribd.com/doc/18226026/An-Introduction-to-RSoftware-for-Statistical-Modelling-and-Computing-Course-Notes>.\
12. "Classification and Regression Trees (CART)." Electronic Textbook StatSoft.StatSoft, Inc., 2002. Web. 20 Apr. 2012. <http://www.obgyn.cam.ac.uk/cam-only/statsbook/stcart.html>.
13. Stine, Robert. "Lecture 8: Classification & Regression Trees." Spring 2011.University of Pennsylvania Data Mining.Web. 19 Apr. 2012.<http://www-stat.wharton.upenn.edu/~stine/mich/ DM08.pdf>.
14. "Lesson 10: Classification/Decision Trees ." File last modified on 2012. Penn State STAT 557 Data Mining. The Pennsylvania State University.Drupal.Web. 19 Apr. 2012. <https://onlinecourses.science.psu.edu/stat557/book/export/html/83>.
15. "Classification and Regression Trees (C&RT)." Electronic Textbook.StatSoft, Inc., 2002. Web. 20 Apr. 2012. <http://www.obgyn.cam.ac.uk/cam-only/statsbook/stcart.html>.
16. http://www.statsoft.com/textbook/classification-and-regression-trees/
17. http://artax.karlin.mff.cuni.cz/~smetp0am/odkazy/CLASSFINAL.PPT#260,6,Classification of Patients as High or No risk group.

18. "Classification and Regression Trees (C&RT)." *Electronic Textbook*. StatSoft, Inc., 2002. Web. 20 Apr. 2012. <http://www.obgyn.cam.ac.uk/cam-only/statsbook/stcart.html>.
19. *CRAN R Project*. Vers. 3.1-52. N.p., Mar.-Apr. 2012. Web. 20 Apr. 2012. <http://cran.r-project.org/web/packages/rpart/rpart.pdf>.
20. Kuhnert, Perta, and Bill Venables. "Tree-based Models II." *An Introduction to R: Software for Statistical Modeling & Computing*. Cleveland, Australia: CSIRO Mathematical and Information Sciences. 283-296. *Scribd*. Web. 18 Apr. 2012. <http://www.scribd.com/doc/18226026/An-Introduction-to-RSoftware-for-Statistical-Modelling-and-Computing-Course-Notes>.
21. "Lesson 10: Classification/Decision Trees." File last modified on 2012. Penn State STAT 557 Data Mining. The Pennsylvania State University. *Drupal*. Web. 19 Apr. 2012. <https://onlinecourses.science.psu.edu/stat557/book/export/html/83>.
22. "Regression Trees: An Overview." *New Zealand Digital Library: food and nutrition*. University of Waikato Department of Computer Science, Sept. 2003. Web. 21 Apr. 2012. <http://www.greenstone.org/greenstone3/nzdl?a=d&d=HASH01b184c9bb619e754e65efdc.8.pp&c=fnl2.2&sib=1&dt=&ec=&et=&p.a=b&p.s=ClassifierBrowse&p.sa= >.
23. Ripley, Brian, Terry M Therneau, and Beth Atkinson. "Package 'rpart' Recursive Partitioning." *Classification and Regression Trees* by L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, Chapman & Hall, 1984.
24. Stine, Robert. "Lecture 8: Classification & Regression Trees." Spring 2011. *University of Pennsylvania Data Mining*. Web. 19 Apr. 2012. <http://www-stat.wharton.upenn.edu/~stine/mich/ DM08.pdf>.
25. https://www.bu.edu/sph/files/2014/05/MorganCART.pdf
26. https://www.stat.wisc.edu/~loh/treeprogs/guide/wires11.pdf
27. https://www.sciencedirect.com/science/article/pii/S2212567115007972
28. https://www.google.com/search?safe=strict&q=Random+forest&stick=H4sIAAAAAAAAAD2OTQqDMBBGyc7eoYvgBTRF7F26kSEmZiQ_diYg7RE9UldNJXT35vHxmObaXrrQ9cO78_HXaewASGnKGfIIG0iqT0wo0UNGYuHOEsyC5kiy5nJGD7EGVEub2vlM3iI5sdO3-x_okZVtX9Pr4rDOvV1MSqe-CPa8koo_QDaYTTSG6CIcZHgl0SYXeAvJC8cN7wAAAA&sa=X&ved=2ahUKEwi-qL7n24jeAhUao4MKHW3YAYwQxA0wEXoECAUQBg&biw=1536&bih=683&dpr=1.25
29. https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd
30. http://www.stat.cmu.edu/~cshalizi/350-2006/lecture-10.pdf

31. http://www2.stat.duke.edu/~rcs46/lectures_2017/08-trees/08-tree-regression.pdf
32. https://www.google.com/search?safe=strict&sa=X&biw=1536&bih=683&q=Machine+Learning:+An+Artificial+Intelligence+Approach&stick=H4sIAAAAAAAAAONgFuLSz9U3MMuNz8kxV-LRT9c3zEpOK8-uLDLS4nHKz88OzkxJLU-sLAYAJAKiiioAAAA&npsic=0&ved=0ahUKEwitx7e8g4neAhXYqYMKHVxqBi0Q-BYINA
33. https://www.google.com/search?safe=strict&sa=X&biw=1536&bih=683&q=Data+Mining+Techniques:+For+Marketing,+Sales,+and+Customer+Relationship+Management&stick=H4sIAAAAAAAAAONgFuLSz9U3MMuNz8kxV-LRT9c3NEqqNDZKNy3W4nHKz88OzkxJLU-sLAYAlg9kbSoAAAA&npsic=0&ved=0ahUKEwihhNDAg4neAhVT1IMKHY59Dr0Q-BYIQw
34. https://www.amazon.com/Principles-Adaptive-Computation-Machine-Learning/dp/026208290X
35. https://www.amazon.com/Learning-Data-Yaser-S-Abu-Mostafa/dp/1600490069/ref=pd_lpo_sbs_14_img_2?_encoding=UTF8&psc=1&refRID=XP3NEGA8SNX49VHWBM3W

36.Academic Journal of Business, Administration, Law and Social Sciences E-ISSN 2410-8693 / ISSN 2410-3918 . Adem Meta**: Use of Distribution Algorithms, for the Construction of a Classification and Regression Tree.**

37. ICIS Ⅴ-2016, Vol 1 Fourth international Conference On: **"Interdisciplinary Studies- Global Challenge 2016" 17 December, 2016 Tirana – Bialistok(Poland): Sjelljet Kaotike dhe dimensioni I fraktaleve.**

38. ICIS IⅤ-2016, Vol 1 Fifth international Conference On: **"Interdisciplinary Studies- Global Challenge 2016" 1 October, 2016 Tirana – Bialistok(Poland). Adem Meta:**
1. "A summery of classification and regression tree with application".
2. "An overview for chaos fractals and applications".

39. Academic Journal of Business, Administration, Law and Social Sciences E-ISSN 2410-8693 / ISSN 2410-3918. A Meta: "An overview for Regerssion Tree". 2018

SHTOJCË

Aneksi A: Kodet në R software.

```
x<- read.csv("X.csv", header=T)
head(x)
boxplot(x[,2])
hist(x[,2])
qqnorm(x[,2])
qqline(x[,2]
```

We use similar codes for the other variables to do the histograms, box plots and qqplots.

```
# attach libraries:
library(MASS)
library(rpart)
summary(x)
my.control<- rpart.control(cp = 0, minsplit=5, xval=5)
CAD1 <- rpart(chd ~ sbp + tobacco + ldl + adiposity + famhist + typea + obesity + alcohol+
age , data=x, method='class',control=my.control)
CAD2 <- rpart(Num
~Age+Sex+ChestPain+RestBP+Chol+FBS+RestECG+Thalag+Exang+OldPeak+Slope+Ca+
Thal, data=x, method='class',control=my.control)
Summary(CAD1)
plot(CAD1)
text(CAD1)
Table1 <- printcp(CAD1)
plotcp(CAD1)
CAD2 <- prune.rpart(CAD1, cp= 0.0093750)
CAD2 <- prune.rpart(CAD1, cp=.0093750) post(CAD2, file="")
plot(CAD2)
text(CAD2)
```

Here I used all the 9 variables to classify ((patientsClassificationtree:rpart(formula = chd ~ sbp + tobacco + ldl + adiposity + famhist +
typea + obesity + alcohol + age, data = y, method = "class", control = my.control)

```
boxplot(x[,2],x[,3],x[,4],x[,5],x[,7],x[,8],x[,9],x[,10],x[11])
boxplot(x[,2],x[,3],x[,4],x[,5],x[,9},x[11])
> boxplot(x[,2],x[,3],x[,4],x[,5],x[,9},x[,11])
> boxplot(x[,2],x[,3],x[,4],x[,5],x[,9],x[,11])
> with(x, plot(BNP, CRP16, col=ALLCAD, pch=as.numeric(ALLCAD)))
> with(x, plot(BNP, CRP16,AGE col=ALLCAD, pch=as.numeric(ALLCAD)))
> with(x, plot(smoking,AGE, col=ALLCAD, pch=as.numeric(ALLCAD)))
> with(x, plot(BNP, CRP16,AGE col=ALLCAD, pch=as.numeric(ALLCAD)))
> with(x, plot(BNP, CRP16, col=ALLCAD, pch=as.numeric(ALLCAD)))
```

R := empty set of rules while not x empty split x into growing set and pruning set build decision tree on growing set and prune on pruning set r := best rule from decision tree R := add r to R remove instances from x that are covered by r return.

```
x<- read.csv("X.csv", header=T)
> names(x)
 [1] "ALLCAD"   "BNP"      "CRP16"    "DLDL"     "UHDL"     "DIABETICS"
```

```
 [7] "smoking"  "CVDYN"   "AGE"     "GENDER"  "CRECLR"  "HTN"
> with(x, plot(BNP, CRP16, col=ALLCAD, pch=as.numeric(ALLCAD)))
> R := empty set of rules while not x empty split x into growing set and pruning set build decision tree on growing set and prune on pruning set r := best rule from decision tree R := add r to R remove instances from x that are covered by r return R
Error: unexpected symbol in "R := empty set"
> xtabs( ~ ALLCAD, data = x.df)
> require(tree)
Loading required package: tree
Warning message:
In library(package, lib.loc = lib.loc, character.only = TRUE, logical.return = TRUE,  :
  there is no package called ‘tree’
> xtabs( ~ HTN, data = x.df)
Error in terms.formula(formula, data = data) : object 'x.df' not found
> x.df = read.csv("x.txt")
In addition: Warning message:
In file(file, "rt") : cannot open file 'x.txt': No such file or directory
> local({pkg <- select.list(sort(.packages(all.available = TRUE)),graphics=TRUE)
+ if(nchar(pkg)) library(pkg, character.only=TRUE)})
Warning message:
package ‘rpart’ was built under R version 3.1.3
> ecoli.df = read.csv("x.csv")
> head(ecoli.df)
ALLCAD     BNP CRP16 DLDL UHDL DIABETICS smoking CVDYN      AGE GENDER
1      Y 102.708  0.92   94 40.3        ND      NS     Y 55.92060      M
2      Y  74.439  5.72   66 31.0        YD      YS     Y 59.52361      M
3      Y  34.911  0.45   62 37.8        ND      YS     Y 63.46338      M
4      Y 115.101  3.63   88 28.9        ND      YS     Y 78.33812      M
5      Y 121.257  2.62   57 31.6        ND      NS     Y 75.34839      F
6      Y  60.021  3.03  107 30.6        ND      YS     Y 66.45311      M
CRECLR HTN
1 111.80770  YH
2 100.86810  NH
3  99.21414  YH
4 100.07540  YH
5  65.72415  YH
6  90.79862  NH
> xtabs( ~ ALLCAD, data = ecoli.df)
ALLCAD
 N    Y
1106 3911
> require(tree)
Loading required package: tree
Warning message:
In library(package, lib.loc = lib.loc, character.only = TRUE, logical.return = TRUE,
there is no package called ‘tree’
> ecoli.tree1 = tree(class ~ mcv + gvh + lip + chg + aac + alm1 + alm2,
+   data = ecoli.df)
Error: could not find function "tree"
ecoli.rpart1 = rpart(class ~ mcv + gvh + lip + chg + aac + alm1 + alm2, data = ecoli.df)
```

```
> ecoli.rpart1 = tree(ALLCAD ~ BNP + CRP16 + DLDL + UHDL + DIABETICS + smoking + CVDYN+ AGE+GENDER, data = ecoli.df)
> local({pkg <- select.list(sort(.packages(all.available = TRUE)),graphics=TRUE)
+ if(nchar(pkg)) library(pkg, character.only=TRUE)})
> utils:::menuInstallPkgs()
--- Please select a CRAN mirror for use in this session ---
trying URL 'http://cran.case.edu/bin/windows/contrib/3.1/tree_1.0-37.zip'
Content type 'application/zip' length 120391 bytes (117 Kb)
opened URL
downloaded 117 Kb
package ‘tree’ successfully unpacked and MD5 sums checked

The downloaded binary packages are in
C:\Users\metaad01\AppData\Local\Temp\RtmpyMbTZR\downloaded_packages
>  ecoli.tree1 = tree(ALLCAD ~ BNP + CRP16 + DLDL + UHDL + DIABETICS + smoking + CVDYN+ AGE+GENDER, data = ecoli.df)
> local({pkg <- select.list(sort(.packages(all.available = TRUE)),graphics=TRUE)
+ if(nchar(pkg)) library(pkg, character.only=TRUE)})
Warning message:
package ‘tree’ was built under R version 3.1.3
>  ecoli.tree1 = tree(ALLCAD ~ BNP + CRP16 + DLDL + UHDL + DIABETICS + smoking + CVDYN+ AGE+GENDER, data = ecoli.df)
> summary(ecoli.tree1)

Classification tree:
tree(formula = ALLCAD ~ BNP + CRP16 + DLDL + UHDL + DIABETICS + smoking + CVDYN + AGE + GENDER, data = ecoli.df)
Variables actually used in tree construction:
[1] "CVDYN"
Number of terminal nodes:  2
Residual mean deviance:  0.2181 = 1094 / 5015
Misclassification error rate: 0.02432 = 122 / 5017
> plot(ecoli.tree1)
> text(ecoli.tree1, all = T)
> cv.tree(ecoli.tree1)
$size
[1] 2 1
$dev
[1] 1095.499 5295.158
$k
[1]     -Inf 4198.866
$method
[1] "deviance"
attr(,"class")
[1] "prune"  "tree.sequence"
> ecoli.tree2 = prune.misclass(ecoli.tree1, best = 6)
Warning message:
In prune.tree(tree = ecoli.tree1, best = 6, method = "misclass") :best is bigger than tree size
> summary(ecoli.tree2)
```

```
Classification tree:
tree(formula = ALLCAD ~ BNP + CRP16 + DLDL + UHDL + DIABETICS + smoking +
CVDYN + AGE + GENDER, data = ecoli.df)
Variables actually used in tree construction:
[1] "CVDYN"
Number of terminal nodes:  2
Residual mean deviance:  0.2181 = 1094 / 5015
Misclassification error rate: 0.02432 = 122 / 5017
>> attach(x)
> Table(x$ALLCAD)

N    Y
1106 3911
> Table(GENDER)
GENDER
F    M
1680 3337
> chisq.test(Table(GENDER))

Chi-squared test for given probabilities
data:  Table(GENDER)
X-squared = 547.2691, df = 1, p-value < 2.2e-16
> Table(GENDER,ALLCAD)
ALLCAD
GENDER    N    Y
F  559 1121
M  547 2790
> chisq.test(Table(GENDER,ALLCAD))
Pearson's Chi-squared test with Yates' continuity correction
data:  Table(GENDER, ALLCAD)
X-squared = 184.3321, df = 1, p-value < 2.2e-16
> chisq.test(Table(AGE,ALLCAD))
Pearson's Chi-squared test
data:  Table(AGE, ALLCAD)
X-squared = 4231.441, df = 4254, p-value = 0.594

chisq.test(Table(smoking,ALLCAD))
Pearson's Chi-squared test with Yates' continuity correction
data:  Table(smoking, ALLCAD)
X-squared = 78.8394, df = 1, p-value < 2.2e-16
chisq.test(Table(DIABETICS,ALLCAD))
Pearson's Chi-squared test with Yates' continuity correction
data:  Table(DIABETICS, ALLCAD)
X-squared = 97.824, df = 1, p-value < 2.2e-16
with(x, {
scatterplot3d(DLDL,   # x axis
AGE,     # y axis
BNP,    # z axis
main="3-D Scatterplot shembull 1")
})
```

```
with(x, {
scatterplot3d(UHDL,   # x axis
BNP,     # y axis
CRP16,    # z axis
main="3-D Scatterplot shembull 1")
})
with(x, {
scatterplot3d(CRECLR,   # x axis
AGE,     # y axis
CVDYN,    # z axis
main="3-D Scatterplot shembull 1")
})
library(scatterplot3d)
with(x, {
s3d <- scatterplot3d(CRECLR,AGE, CVDYN,        # x y and z axis
color="blue", pch=19,        # filled blue circles
type="h",                    # vertical lines to the x-y plane
main="3-D Scatterplot shembull 1",
xlab="CRECLR",
ylab="AGE ",
zlab="CUDYN")
s3d.coords <- s3d$xyz.convert(CRECL,AGE, CVDYN) # convert 3D coords to 2D
projection
text(s3d.coords$x, s3d.coords$y,             # x and y coordinates
labels=row.names(mtcars),               # text to plot
cex=.5, pos=4)          # shrink text 50% and place to right of points)
})
with(x, {
scatterplot3d(BNP,   # x axis
CRP16,     # y axis
UHDL,    # z axis
 main="3-D Scatterplot shembull 1")
})
library(scatterplot3d)
with(x, {
s3d <- scatterplot3d(BNP,CRP16, UHDL,        # x y and z axis
color="blue", pch=19,        # filled blue circles
type="h",                    # vertical lines to the x-y plane
main="3-D Scatterplot shembull 1",
xlab="BNP",
ylab="CRP16",
zlab="UHDL"
s3d.coords <- s3d$xyz.convert(BNP,CRP16,UHDL) # convert 3D coords to 2D projection
text(s3d.coords$x, s3d.coords$y,             # x and y coordinates
labels=row.names(mtcars),               # text to plot
cex=.5, pos=4)          # shrink text 50% and place to right of points)
})
with(x, {
scatterplot3d(AGE,   # x axis
smoking,     # y axis
```

```
GENDER,    # z axis
main="3-D Scatterplot shembull 1")
})
library(scatterplot3d)
with(x, {
s3d <- scatterplot3d(AGE,smoking,GENDER,        # x y and z axis
color="blue", pch=19,        # filled blue circles
type="h", # vertical lines to the x-y plane
main="3-D Scatterplot shembull 1",
xlab="AGE",
ylab="smoking",
zlab="GENDER")
s3d.coords <- s3d$xyz.convert(AGE,smoking,GENDER) # convert 3D coords to 2D projection
text(s3d.coords$x, s3d.coords$y,             # x and y coordinates
labels=row.names(mtcars),              # text to plot
cex=.5, pos=4)          # shrink text 50% and place to right of points)
})


library(scatterplot3d)
with(mtcars, {
scatterplot3d(disp,   # x axis
wt,     # y axis
mpg,    # z axis
main="3-D Scatterplot Example 1")
})

y<- read.csv("Y.csv", header=T)
Eksplorimi i datës
names(y)
Nga grafikët e mësipërm shikojmë se kjo datë ka një shpërndarje jo normale.
with(y, plot(tobacco, ldl, col=chd, pch=as.numeric(chd)))
distMatrix <- as.matrix(dist(y[,2:3]))
>  heatmap(distMatrix)
distMatrix <- as.matrix(dist(y[,1:4]))
>  heatmap(distMatrix)

distMatrix <- as.matrix(dist(x[,2:4]))
>  heatmap(distMatrix)

distMatrix <- as.matrix(dist(y[,7:9]))
> heatmap(distMatrix)
> distMatrix <- as.matrix(dist(y[,7:10]))
> heatmap(distMatrix
> x<- read.csv("X.csv", header=T)
str(x)

x.rp=rpart(ALLCAD~BNP+CRP16+DLDL+UHDL+DIABETICS+smoking+CVDYN+AGE
+GENDER+CRECLR,method="anova",data=x,control=rpart.control(cp=0.001))
```

```
(medv=ALLCAD)(x)
summary(x.rp)
CAD1                                                    =rpart(ALLCAD
~BNP+CRP16+DLDL+UHDL+DIABETICS+smoking+CVDYN+AGE+GENDER+CRECL
R, data=x, method='class',control=my.control)

CAD1 =rpart(ALLCAD ~ ., data=x, method='class',control=my.control)
```

The function Print gives a text version of our tree. Figurë 3:

```
print(CAD1)
y.rp=rpart(chd~sbp+tabacco+Idl+adiposity+fmhistory+typea+obesity+alcohol+age,method="
anova",data=y,control=rpart.control(cp=0.001))
sbp"       "tobacco"   "ldl"       "adiposity" "famhist"
 [7] "typea"     "obesity"   "alcohol"   "age"       "chd"
```

Aneksi B
Disa grafikë për shpërndarjen e bazës së të dhënave

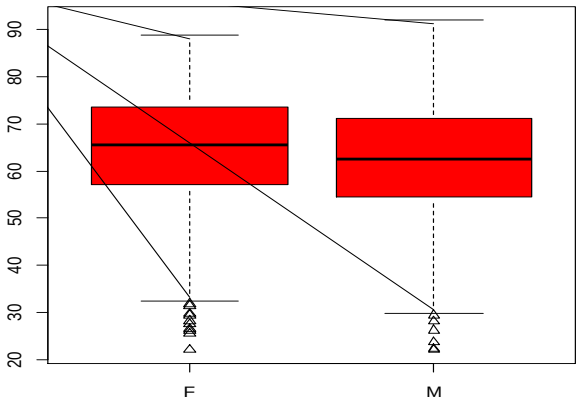


Figura 47: Boxplot **p**ër gjinitë femer Mashkull

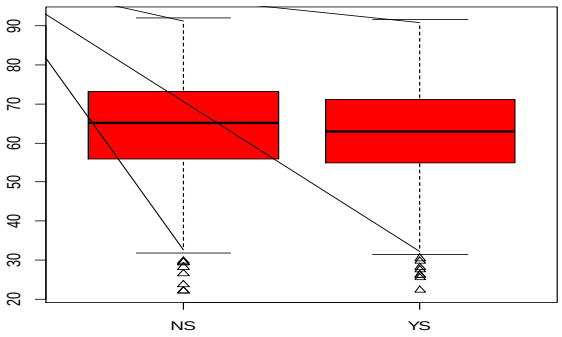


Figura 48: Boxplot për kur historia familjare nuk është prezente dhe kur është prezente.

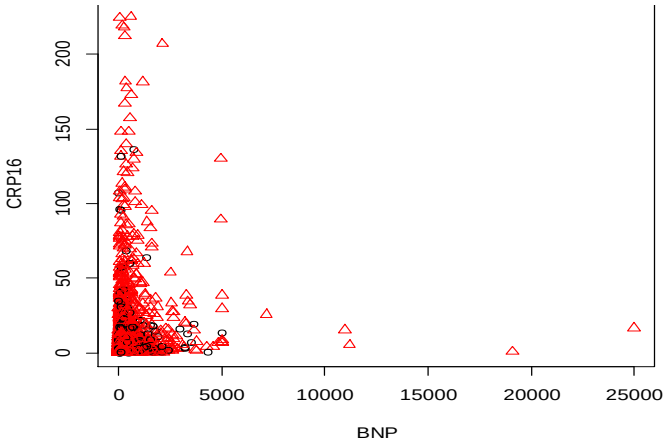


Figura 49: Shpërndarja dy dimensionale për BNP vs CRP16

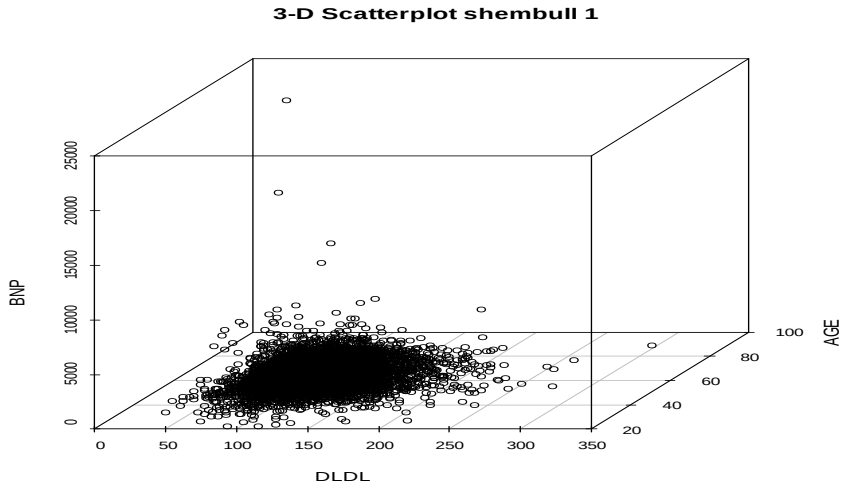


Figura 50: 3-D Scatterplot për DLDL, AGE dhe BNP

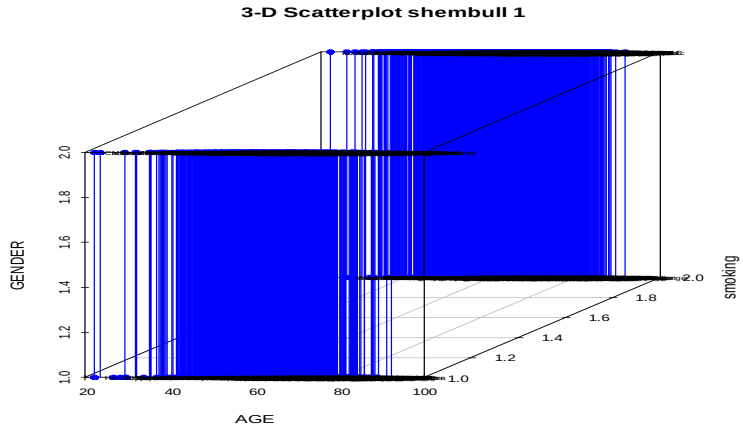


Figura 51: 3-D Scatterplot për GENDER, AGE dhe CRECLR

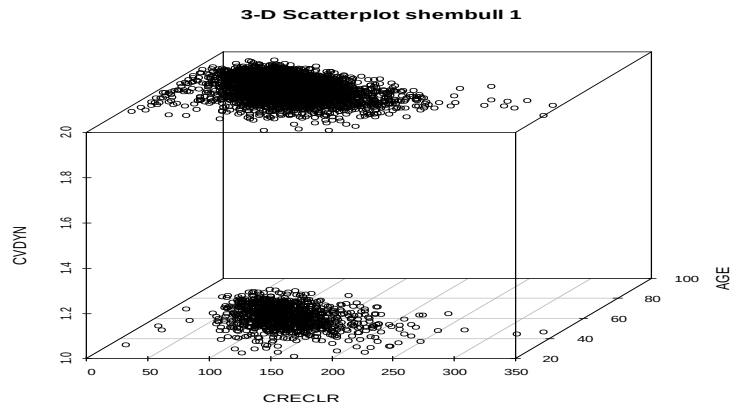


Figura 52**:** 3-D Scatterplot për CVDYN, AGE dhe CRECLR

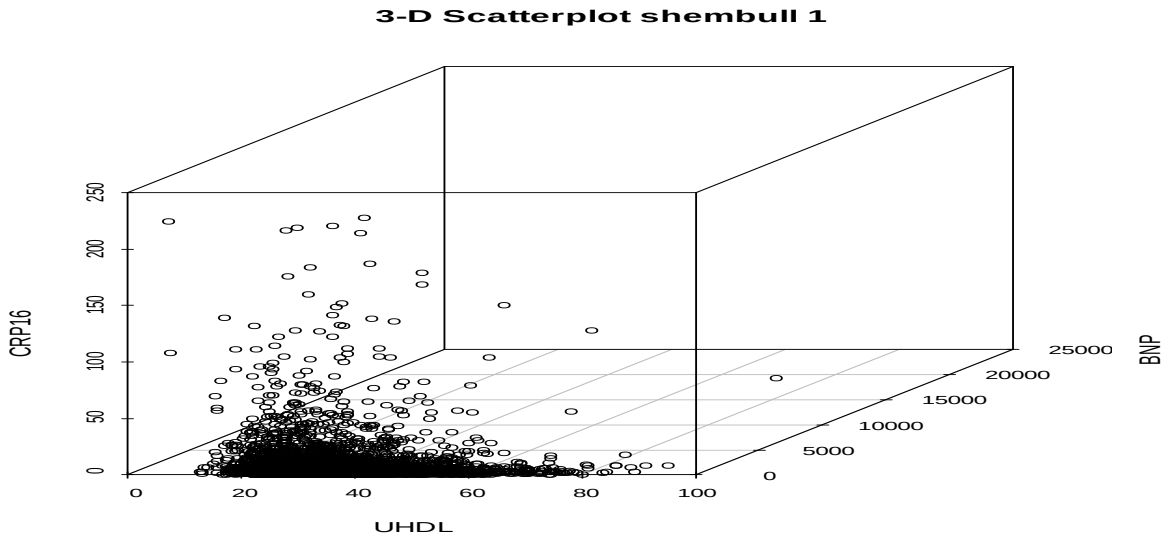


Figura 53: 3-D Scatterplot për UHDL, BNP dhe CRP16

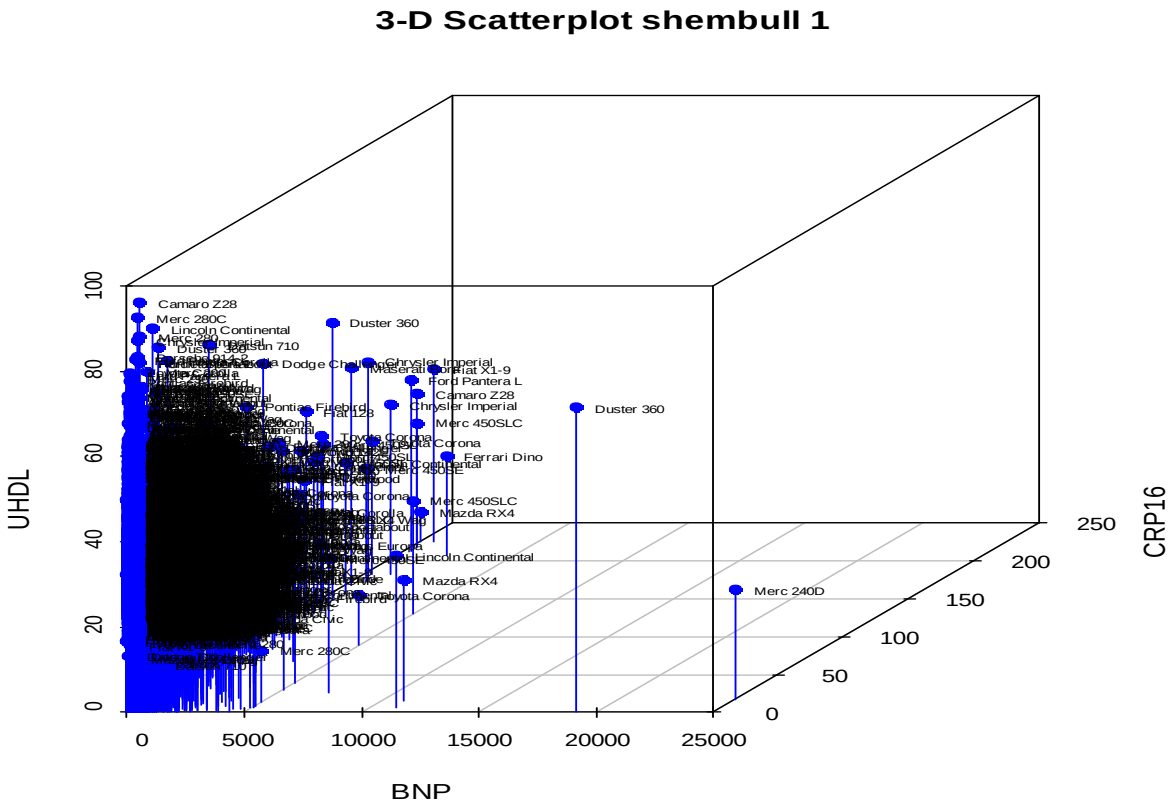


Figura 54: 3-D Scatterplot për BNP, UHDL dhe CRP16

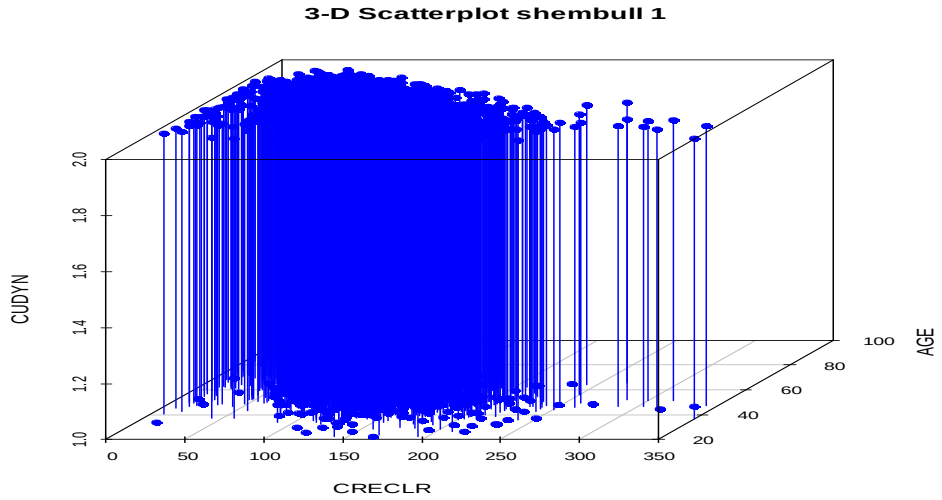


Figura 55: 3-D Scatterplot për CRECRL AGE dhe CLDN

Aneksi C: Disa tabela te llogaritjeve

Variables actually used in tree construction:
[1] adiposity age alcohol famhist ldl obesity sbpTobaccotypea
Root node error: 160/462 = 0.34632
n= 462

CP	nsp	litrelerror	xerror	xstd
1 0.1250000	0	1.0000	1.00000	0.063918
2 0.1000000	1	0.8750	0.97500	0.063530
3 0.0625000	2	0.7750	0.96250	0.063328
4 0.0250000	3	0.7125	0.88750	0.061984
5 0.0187500	5	0.6625	0.85625	0.061357
6 0.0125000	7	0.6250	0.95000	0.063119
7 0.0093750	10	0.5875	0.99375	0.063823
8 0.0083333	32	0.3375	1.00625	0.064011
9 0.0062500	35	0.3125	0.97500	0.063530
10 0.0031250	53	0.2000	0.98750	0.0637271
11 0.0000000	57	0.1875	0.98750	0.063727

y.rp=rpart(chd~sbp+tobacco+ldl+adiposity+famhist+typea+obesity+alcohol+age,method="a
nova",data=y,control=rpart.control(cp=0.001))
> summary(y.rp)
Call:
rpart(formula = chd ~ sbp + tobacco + ldl + adiposity + famhist +
typea + obesity + alcohol + age, data = y, method = "anova",
control = rpart.control(cp = 0.001))
n= 462

CP	nsplit	rel error	xerror	xstd
1 0.117548766	0	1.0000000	1.0038565	0.03019452
2 0.036324104	1	0.8824512	0.9176419	0.03994883
3 0.035369235	2	0.8461271	0.9540061	0.04627575
4 0.033938862	3	0.8107579	0.9512883	0.04686725
5 0.030356727	4	0.7768190	0.9432938	0.04729023
6 0.017171328	5	0.7464623	0.9204187	0.05090138
7 0.013941244	6	0.7292910	0.9635595	0.05520414
8 0.012843514	7	0.7153497	1.0122148	0.05863805
9 0.012316738	9	0.6896627	1.0200617	0.05902310
10 0.011951573	12	0.6527125	1.0218839	0.05970650
11 0.011712541	13	0.6407609	1.0365123	0.06164681
12 0.011125828	14	0.6290484	1.0530802	0.06221389
13 0.010903549	15	0.6179226	1.0435225	0.06180536
14 0.010847948	17	0.5961155	1.0415065	0.06167475
15 0.010586125	18	0.5852675	1.0511701	0.06195579
16 0.009232541	19	0.5746814	1.0479931	0.06241228
17 0.008947702	20	0.5654488	1.0616167	0.06359657
18 0.007815761	21	0.5565011	1.0684294	0.06387115
19 0.007455620	22	0.5486854	1.0810673	0.06460478
20 0.004767496	25	0.5260812	1.0665201	0.06448219
21 0.003309666	26	0.5213137	1.0580525	0.06418118

```
22 0.001246378     27 0.5180040 1.0627395 0.06440530
23 0.001138245     28 0.5167576 1.0617933 0.06411419
24 0.001000000     29 0.5156194 1.0620645 0.06410505

Variable importance
Age tobacco  adiposity typea  ldl  sbp obesity          famhist  alcohol
21       16       14       11       10       9        9                 5             5

Node number 1: 462 observations,    complexity param=0.1175488
mean=1.34632, MSE=0.2263826
left son=2 (290 obs) right son=3 (172 obs)
Primary splits:
age       < 50.5   to the left,  improve=0.11754880, (0 missing)
tobacco   < 0.49   to the left,  improve=0.09285731, (0 missing)
famhist   splits as  LR,         improve=0.07418690, (0 missing)
ldl       < 4.315  to the left,  improve=0.06018379, (0 missing)
adiposity < 25.16  to the left,  improve=0.04965827, (0 missing)
Surrogate splits:
adiposity < 31.34  to the left,  agree=0.721, adj=0.250, (0 split)
sbp       < 155    to the left,  agree=0.710, adj=0.221, (0 split)
tobacco   < 7.24   to the left,  agree=0.695, adj=0.180, (0 split)
typea     < 38.5   to the right, agree=0.649, adj=0.058, (0 split)
ldl       < 8.25   to the left,  agree=0.645, adj=0.047, (0 split)

Node number 2: 290 observations,    complexity param=0.03536924
mean=1.22069, MSE=0.1719857
left son=4 (108 obs) right son=5 (182 obs)
Primary splits:
age       < 30.5   to the left,  improve=0.07416862, (0 missing)
tobacco   < 0.49   to the left,  improve=0.06492540, (0 missing)
typea     < 68.5   to the left,  improve=0.05865142, (0 missing)
ldl       < 4.155  to the left,  improve=0.05085479, (0 missing)
adiposity < 25.16  to the left,  improve=0.03823596, (0 missing)
Surrogate splits:
adiposity < 21.27  to the left,  agree=0.779, adj=0.407, (0 split)
tobacco   < 0.17   to the left,  agree=0.762, adj=0.361, (0 split)
obesity   < 22.945 to the left,  agree=0.710, adj=0.222, (0 split)
ldl       < 3.12   to the left,  agree=0.707, adj=0.213, (0 split)
sbp       < 115    to the left,  agree=0.641, adj=0.037, (0 split)

Node number 3: 172 observations,    complexity param=0.0363241
mean=1.55814, MSE=0.2466198
left son=6 (82 obs) right son=7 (90 obs)
Primary splits:
famhist splits as  LR,         improve=0.08956194, (0 missing)
tobacco < 7.47   to the left,  improve=0.07632467, (0 missing)
ldl     < 2.44   to the left,  improve=0.06974491, (0 missing)
typea   < 67     to the left,  improve=0.03861789, (0 missing)
obesity < 25.115 to the left,  improve=0.01433518, (0 missing)
Surrogate splits:
```

ldl < 3.88 to the left, agree=0.593, adj=0.146, (0 split)
tobacco < 11.895 to the right, agree=0.570, adj=0.098, (0 split)
obesity < 24.91 to the left, agree=0.564, adj=0.085, (0 split)
typea < 48.5 to the left, agree=0.558, adj=0.073, (0 split)
sbp < 109 to the left, agree=0.547, adj=0.049, (0 split)

Node number 4: 108 observations, complexity param=0.01284351
mean=1.074074, MSE=0.06858711
left son=8 (80 obs) right son=9 (28 obs)
Primary splits:
tobacco < 0.51 to the left, improve=0.15793750, (0 missing)
alcohol < 11.105 to the left, improve=0.13514030, (0 missing)
obesity < 19.53 to the right, improve=0.08909091, (0 missing)
adiposity < 26.03 to the left, improve=0.04898785, (0 missing)
age < 24.5 to the left, improve=0.04659629, (0 missing)
Surrogate splits:
alcohol < 8.39 to the left, agree=0.778, adj=0.143, (0 split)
age < 24.5 to the left, agree=0.769, adj=0.107, (0 split)
adiposity < 7.89 to the right, agree=0.759, adj=0.071, (0 split)

Node number 5: 182 observations, complexity param=0.03393886
mean=1.307692, MSE=0.2130178
left son=10 (170 obs) right son=11 (12 obs)
Primary splits:
typea < 68.5 to the left, improve=0.09155773, (0 missing)
adiposity < 36.58 to the left, improve=0.03104308, (0 missing)
ldl < 3.34 to the left, improve=0.03030303, (0 missing)
famhist splits as LR, improve=0.02835648, (0 missing)
sbp < 133 to the left, improve=0.01696073, (0 missing)
Surrogate splits:
sbp < 192 to the left, agree=0.94, adj=0.083, (0 split)

Node number 6: 82 observations, complexity param=0.03035673
mean=1.402439, MSE=0.2404819
left son=12 (58 obs) right son=13 (24 obs)
Primary splits:
tobacco < 7.605 to the left, improve=0.16100660, (0 missing)
adiposity < 23.97 to the left, improve=0.06501237, (0 missing)
ldl < 2.78 to the left, improve=0.04351125, (0 missing)
typea < 41.5 to the left, improve=0.03815247, (0 missing)
obesity < 30.365 to the left, improve=0.02332656, (0 missing)
Surrogate splits:
obesity < 19.42 to the right, agree=0.72, adj=0.042, (0 split)

Node number 7: 90 observations, complexity param=0.01717133
mean=1.7, MSE=0.21
left son=14 (39 obs) right son=15 (51 obs)
Primary splits:
ldl < 4.99 to the left, improve=0.09502262, (0 missing)
tobacco < 1.375 to the left, improve=0.07259522, (0 missing)

```
sbp       < 167    to the left,  improve=0.04247501, (0 missing)
typea     < 66.5   to the left,  improve=0.03614458, (0 missing)
adiposity < 28     to the right, improve=0.03325123, (0 missing)
Surrogate splits:
adiposity < 28.65  to the left,  agree=0.667, adj=0.231, (0 split)
obesity   < 25.115 to the left,  agree=0.656, adj=0.205, (0 split)
sbp       < 175    to the right, agree=0.611, adj=0.103, (0 split)
tobacco   < 1.55   to the left,  agree=0.611, adj=0.103, (0 split)
alcohol   < 24.4   to the right, agree=0.611, adj=0.103, (0 split)

Node number 8: 80 observations,    complexity param=0.001246378
mean=1.0125, MSE=0.01234375
left son=16 (73 obs) right son=17 (7 obs)
Primary splits:
obesity   < 19.715 to the right, improve=0.13200720, (0 missing)
ldl       < 2.395  to the right, improve=0.05063291, (0 missing)
adiposity < 12.185 to the right, improve=0.04360056, (0 missing)
sbp       < 119    to the right, improve=0.03556359, (0 missing)
typea     < 49.5   to the right, improve=0.02226102, (0 missing)
Surrogate splits:
adiposity < 9.505  to the right, agree=0.938, adj=0.286, (0 split)

Node number 9: 28 observations,    complexity param=0.01284351
mean=1.25, MSE=0.1875
left son=18 (13 obs) right son=19 (15 obs)
Primary splits:
alcohol   < 11.105 to the left,  improve=0.28888890, (0 missing)
obesity   < 24.84  to the right, improve=0.13846150, (0 missing)
adiposity < 21.33  to the left,  improve=0.09551657, (0 missing)
ldl       < 4.18   to the left,  improve=0.06666667, (0 missing)
sbp       < 123    to the right, improve=0.05668934, (0 missing)
Surrogate splits:
adiposity < 11.965 to the left,  agree=0.643, adj=0.231, (0 split)
sbp       < 127    to the left,  agree=0.607, adj=0.154, (0 split)
tobacco   < 1.805  to the left,  agree=0.607, adj=0.154, (0 split)
ldl       < 5.14   to the right, agree=0.607, adj=0.154, (0 split)
obesity   < 22.635 to the left,  agree=0.607, adj=0.154, (0 split)

Node number 10: 170 observations,    complexity param=0.01231674
mean=1.270588, MSE=0.1973702
left son=20 (78 obs) right son=21 (92 obs)
Primary splits:
typea     < 53.5   to the left,  improve=0.03565102, (0 missing)
ldl       < 6.275  to the left,  improve=0.02881380, (0 missing)
obesity   < 25.635 to the right, improve=0.02547446, (0 missing)
tobacco   < 8.04   to the left,  improve=0.01885114, (0 missing)
famhist splits as  LR,           improve=0.01852334, (0 missing)
Surrogate splits:
ldl       < 3.305  to the left,  agree=0.594, adj=0.115, (0 split)
tobacco   < 0.23   to the left,  agree=0.588, adj=0.103, (0 split)
```

age < 42.5 to the right, agree=0.588, adj=0.103, (0 split)
adiposity < 24.16 to the right, agree=0.576, adj=0.077, (0 split)
obesity < 26.45 to the right, agree=0.576, adj=0.077, (0 split)

Node number 11: 12 observations
mean=1.833333, MSE=0.1388889

Node number 12: 58 observations, complexity param=0.01090355
mean=1.275862, MSE=0.1997622
left son=24 (11 obs) right son=25 (47 obs)
Primary splits:
typea < 42.5 to the left, improve=0.08915907, (0 missing)
adiposity < 24.435 to the left, improve=0.08681441, (0 missing)
tobacco < 4.95 to the right, improve=0.05723443, (0 missing)
age < 62.5 to the left, improve=0.05239335, (0 missing)
obesity < 30.365 to the left, improve=0.04023810, (0 missing)

Node number 13: 24 observations, complexity param=0.01171254
mean=1.708333, MSE=0.2065972
left son=26 (15 obs) right son=27 (9 obs)
Primary splits:
adiposity < 28.955 to the right, improve=0.24705880, (0 missing)
ldl < 4.565 to the right, improve=0.16955020, (0 missing)
tobacco < 12.7 to the right, improve=0.07563025, (0 missing)
alcohol < 7.33 to the right, improve=0.06722689, (0 missing)
typea < 54.5 to the left, improve=0.04413530, (0 missing)
Surrogate splits:
sbp < 133 to the right, agree=0.792, adj=0.444, (0 split)
obesity < 23.585 to the right, agree=0.792, adj=0.444, (0 split)
ldl < 3.68 to the right, agree=0.750, adj=0.333, (0 split)
alcohol < 21.55 to the left, agree=0.750, adj=0.333, (0 split)
tobacco < 23.7 to the left, agree=0.708, adj=0.222, (0 split)

Node number 14: 39 observations, complexity param=0.01394124
mean=1.538462, MSE=0.2485207
left son=28 (20 obs) right son=29 (19 obs)
Primary splits:
adiposity < 27.985 to the right, improve=0.15043860, (0 missing)
sbp < 129 to the left, improve=0.12072310, (0 missing)
typea < 51.5 to the right, improve=0.08715304, (0 missing)
tobacco < 0.75 to the left, improve=0.08640553, (0 missing)
obesity < 23.46 to the right, improve=0.08002646, (0 missing)
Surrogate splits:
obesity < 24.98 to the right, agree=0.795, adj=0.579, (0 split)
ldl < 3.645 to the right, agree=0.667, adj=0.316, (0 split)
age < 58.5 to the right, agree=0.667, adj=0.316, (0 split)
sbp < 135 to the right, agree=0.641, adj=0.263, (0 split)
tobacco < 0.75 to the left, agree=0.641, adj=0.263, (0 split)

Node number 15: 51 observations, complexity param=0.00745562

mean=1.823529, MSE=0.1453287
left son=30 (36 obs) right son=31 (15 obs)
Primary splits:
tobacco < 7.2 to the left, improve=0.08928571, (0 missing)
ldl < 6.705 to the left, improve=0.07100614, (0 missing)
typea < 55.5 to the left, improve=0.07100614, (0 missing)
age < 59.5 to the right, improve=0.07054674, (0 missing)
sbp < 154 to the left, improve=0.06265664, (0 missing)
Surrogate splits:
ldl < 5.105 to the right, agree=0.745, adj=0.133, (0 split)
typea < 68.5 to the left, agree=0.745, adj=0.133, (0 split)
obesity < 31.745 to the left, agree=0.745, adj=0.133, (0 split)
alcohol < 57.855 to the left, agree=0.745, adj=0.133, (0 split)

Node number 16: 73 observations
mean=1, MSE=0

Node number 17: 7 observations
mean=1.142857, MSE=0.122449

Node number 18: 13 observations
mean=1, MSE=0

Node number 19: 15 observations
mean=1.466667, MSE=0.2488889

Node number 20: 78 observations, complexity param=0.01231674
mean=1.179487, MSE=0.1472715
left son=40 (58 obs) right son=41 (20 obs)
Primary splits:
ldl < 5.37 to the left, improve=0.11385470, (0 missing)
alcohol < 7.12 to the right, improve=0.07600108, (0 missing)
tobacco < 6.46 to the left, improve=0.05714286, (0 missing)
sbp < 135 to the left, improve=0.04515977, (0 missing)
adiposity < 25.135 to the left, improve=0.03994514, (0 missing)
Surrogate splits:
adiposity < 31.55 to the left, agree=0.769, adj=0.10, (0 split)
alcohol < 91.775 to the left, agree=0.769, adj=0.10, (0 split)
obesity < 34.965 to the left, agree=0.756, adj=0.05, (0 split)
age < 49.5 to the left, agree=0.756, adj=0.05, (0 split)

Node number 21: 92 observations, complexity param=0.01231674
mean=1.347826, MSE=0.2268431
left son=42 (71 obs) right son=43 (21 obs)
Primary splits:
obesity < 23.24 to the right, improve=0.06519115, (0 missing)
adiposity < 20.72 to the right, improve=0.04444444, (0 missing)
famhist splits as LR, improve=0.04385334, (0 missing)
typea < 56.5 to the right, improve=0.03602941, (0 missing)
ldl < 2.83 to the left, improve=0.02084573, (0 missing)

Surrogate splits:
adiposity < 18.025 to the right, agree=0.848, adj=0.333, (0 split)

Node number 24: 11 observations
mean=1, MSE=0

Node number 25: 47 observations, complexity param=0.01090355
mean=1.340426, MSE=0.224536
left son=50 (13 obs) right son=51 (34 obs)
Primary splits:
adiposity < 24.435 to the left, improve=0.11823550, (0 missing)
age < 62.5 to the left, improve=0.10893580, (0 missing)
tobacco < 4.95 to the right, improve=0.05546713, (0 missing)
ldl < 6.355 to the left, improve=0.04881744, (0 missing)
obesity < 30.365 to the left, improve=0.04158986, (0 missing)
Surrogate splits:
obesity < 23.86 to the left, agree=0.894, adj=0.615, (0 split)
sbp < 129 to the left, agree=0.809, adj=0.308, (0 split)
ldl < 2.78 to the left, agree=0.809, adj=0.308, (0 split)

Node number 26: 15 observations
mean=1.533333, MSE=0.2488889

Node number 27: 9 observations
mean=2, MSE=0

Node number 28: 20 observations, complexity param=0.01195157
mean=1.35, MSE=0.2275
left son=56 (10 obs) right son=57 (10 obs)
Primary splits:
tobacco < 4.15 to the left, improve=0.2747253, (0 missing)
adiposity < 34.875 to the left, improve=0.2216117, (0 missing)
sbp < 161 to the left, improve=0.1160488, (0 missing)
typea < 55.5 to the right, improve=0.1015578, (0 missing)
obesity < 26.49 to the left, improve=0.1015578, (0 missing)
Surrogate splits:
age < 59.5 to the left, agree=0.75, adj=0.5, (0 split)
sbp < 133 to the left, agree=0.65, adj=0.3, (0 split)
ldl < 4.22 to the left, agree=0.65, adj=0.3, (0 split)
adiposity < 30.305 to the left, agree=0.65, adj=0.3, (0 split)
typea < 56.5 to the left, agree=0.65, adj=0.3, (0 split)

Node number 29: 19 observations
mean=1.736842, MSE=0.1939058

Node number 30: 36 observations, complexity param=0.00745562
mean=1.75, MSE=0.1875
left son=60 (7 obs) right son=61 (29 obs)
Primary splits:
obesity < 29.39 to the right, improve=0.13300490, (0 missing)

ldl < 6.705 to the left, improve=0.12804230, (0 missing)
typea < 55.5 to the left, improve=0.09030100, (0 missing)
sbp < 154 to the left, improve=0.07407407, (0 missing)
adiposity < 28.915 to the right, improve=0.05939394, (0 missing)
Surrogate splits:
adiposity < 37.62 to the right, agree=0.889, adj=0.429, (0 split)
typea < 38 to the left, agree=0.833, adj=0.143, (0 split)

Node number 31: 15 observations
mean=2, MSE=0

Node number 40: 58 observations, complexity param=0.009232541
mean=1.103448, MSE=0.09274673
left son=80 (42 obs) right son=81 (16 obs)
Primary splits:
sbp < 141 to the left, improve=0.17950630, (0 missing)
obesity < 24.89 to the right, improve=0.09198917, (0 missing)
adiposity < 25.72 to the right, improve=0.05679182, (0 missing)
ldl < 4.33 to the right, improve=0.04784240, (0 missing)
alcohol < 7.12 to the right, improve=0.03041730, (0 missing)
Surrogate splits:
tobacco < 10.4 to the left, agree=0.741, adj=0.062, (0 split)
alcohol < 43.35 to the left, agree=0.741, adj=0.062, (0 split)

Node number 41: 20 observations, complexity param=0.01112583
mean=1.4, MSE=0.24
left son=82 (11 obs) right son=83 (9 obs)
Primary splits:
alcohol < 8.365 to the right, improve=0.24242420, (0 missing)
typea < 48.5 to the left, improve=0.14062500, (0 missing)
ldl < 7.04 to the right, improve=0.10774410, (0 missing)
famhist splits as LR, improve=0.10774410, (0 missing)
tobacco < 6.08 to the left, improve=0.06593407, (0 missing)
Surrogate splits:
adiposity < 28.425 to the right, agree=0.75, adj=0.444, (0 split)
tobacco < 0.95 to the right, agree=0.70, adj=0.333, (0 split)
famhist splits as LR, agree=0.70, adj=0.333, (0 split)
obesity < 25.315 to the right, agree=0.70, adj=0.333, (0 split)
sbp < 138 to the right, agree=0.65, adj=0.222, (0 split)

Node number 42: 71 observations, complexity param=0.01084795
mean=1.28169, MSE=0.2023408
left son=84 (26 obs) right son=85 (45 obs)
Primary splits:
typea < 60.5 to the right, improve=0.07897520, (0 missing)
alcohol < 2.04 to the left, improve=0.04186851, (0 missing)
tobacco < 5.2 to the right, improve=0.04000840, (0 missing)
ldl < 3.39 to the left, improve=0.03954941, (0 missing)
adiposity < 20.72 to the right, improve=0.03861299, (0 missing)
Surrogate splits:

obesity < 31.375 to the right, agree=0.662, adj=0.077, (0 split)
age < 49.5 to the right, agree=0.648, adj=0.038, (0 split)

Node number 43: 21 observations, complexity param=0.004767496
mean=1.571429, MSE=0.244898
left son=86 (8 obs) right son=87 (13 obs)
Primary splits:
ldl < 4.035 to the left, improve=0.09695513, (0 missing)
typea < 61.5 to the left, improve=0.08012821, (0 missing)
obesity < 22.115 to the left, improve=0.08012821, (0 missing)
age < 41.5 to the left, improve=0.08012821, (0 missing)
famhist splits as LR, improve=0.06136364, (0 missing)
Surrogate splits:
adiposity < 16.29 to the left, agree=0.762, adj=0.375, (0 split)
age < 40.5 to the left, agree=0.762, adj=0.375, (0 split)
sbp < 123 to the left, agree=0.714, adj=0.250, (0 split)
tobacco < 0.26 to the left, agree=0.667, adj=0.125, (0 split)
famhist splits as LR, agree=0.667, adj=0.125, (0 split)

Node number 50: 13 observations
mean=1.076923, MSE=0.07100592

Node number 51: 34 observations, complexity param=0.007815761
mean=1.441176, MSE=0.2465398
left son=102 (15 obs) right son=103 (19 obs)
Primary splits:
typea < 50.5 to the left, improve=0.09751924, (0 missing)
age < 60.5 to the left, improve=0.07424561, (0 missing)
tobacco < 0.68 to the left, improve=0.07000390, (0 missing)
ldl < 5.49 to the right, improve=0.05969786, (0 missing)
sbp < 155 to the right, improve=0.04474006, (0 missing)
Surrogate splits:
sbp < 159 to the right, agree=0.706, adj=0.333, (0 split)
ldl < 5.49 to the right, agree=0.676, adj=0.267, (0 split)
age < 57.5 to the left, agree=0.676, adj=0.267, (0 split)
tobacco < 0.005 to the left, agree=0.647, adj=0.200, (0 split)
obesity < 29.015 to the right, agree=0.647, adj=0.200, (0 split)

Node number 56: 10 observations
mean=1.1, MSE=0.09

Node number 57: 10 observations
mean=1.6, MSE=0.24

Node number 60: 7 observations
mean=1.428571, MSE=0.244898

Node number 61: 29 observations, complexity param=0.00745562
mean=1.827586, MSE=0.1426873
left son=122 (15 obs) right son=123 (14 obs)

Primary splits:
sbp < 136 to the left, improve=0.1944444, (0 missing)
ldl < 6.82 to the left, improve=0.1470588, (0 missing)
typea < 55.5 to the left, improve=0.1096491, (0 missing)
obesity < 24.925 to the left, improve=0.1095734, (0 missing)
adiposity < 32.625 to the left, improve=0.0937500, (0 missing)
Surrogate splits:
tobacco < 3.75 to the right, agree=0.690, adj=0.357, (0 split)
typea < 50.5 to the right, agree=0.690, adj=0.357, (0 split)
adiposity < 27.18 to the right, agree=0.655, adj=0.286, (0 split)
alcohol < 0.625 to the left, agree=0.655, adj=0.286, (0 split)
age < 52.5 to the left, agree=0.621, adj=0.214, (0 split)

Node number 80: 42 observations, complexity param=0.001138245
mean=1.02381, MSE=0.02324263
left son=160 (35 obs) right son=161 (7 obs)
Primary splits:
obesity < 22.195 to the right, improve=0.12195120, (0 missing)
alcohol < 0.945 to the right, improve=0.12195120, (0 missing)
sbp < 135 to the left, improve=0.06873614, (0 missing)
typea < 50.5 to the left, improve=0.06873614, (0 missing)
tobacco < 2.51 to the left, improve=0.03586801, (0 missing)
Surrogate splits:
adiposity < 15.29 to the right, agree=0.929, adj=0.571, (0 split)

Node number 81: 16 observations
mean=1.3125, MSE=0.2148438

Node number 82: 11 observations
mean=1.181818, MSE=0.1487603

Node number 83: 9 observations
mean=1.666667, MSE=0.2222222

Node number 84: 26 observations, complexity param=0.003309666
mean=1.115385, MSE=0.102071
left son=168 (13 obs) right son=169 (13 obs)
Primary splits:
tobacco < 2 to the left, improve=0.13043480, (0 missing)
obesity < 30.815 to the left, improve=0.10471830, (0 missing)
sbp < 131 to the left, improve=0.08152174, (0 missing)
alcohol < 24.225 to the left, improve=0.05920242, (0 missing)
typea < 65.5 to the right, improve=0.05797101, (0 missing)
Surrogate splits:
sbp < 131 to the left, agree=0.731, adj=0.462, (0 split)
age < 34.5 to the left, agree=0.731, adj=0.462, (0 split)
alcohol < 13.19 to the left, agree=0.692, adj=0.385, (0 split)
ldl < 4.63 to the right, agree=0.654, adj=0.308, (0 split)
adiposity < 30.265 to the left, agree=0.654, adj=0.308, (0 split)

Node number 85: 45 observations, complexity param=0.01058613
mean=1.377778, MSE=0.2350617
left son=170 (17 obs) right son=171 (28 obs)
Primary splits:
tobacco < 4.1 to the right, improve=0.10467130, (0 missing)
ldl < 3.39 to the left, improve=0.05877385, (0 missing)
alcohol < 0.185 to the left, improve=0.05877385, (0 missing)
adiposity < 20.72 to the right, improve=0.05621877, (0 missing)
famhist splits as LR, improve=0.04236695, (0 missing)
Surrogate splits:
age < 42.5 to the right, agree=0.756, adj=0.353, (0 split)
ldl < 7.735 to the right, agree=0.667, adj=0.118, (0 split)
adiposity < 17.19 to the left, agree=0.667, adj=0.118, (0 split)
sbp < 142.5 to the right, agree=0.644, adj=0.059, (0 split)
typea < 59.5 to the right, agree=0.644, adj=0.059, (0 split)

Node number 86: 8 observations
mean=1.375, MSE=0.234375

Node number 87: 13 observations
mean=1.692308, MSE=0.2130178
Node number 102: 15 observations
mean=1.266667, MSE=0.1955556
Node number 103: 19 observations
mean=1.578947, MSE=0.2437673
Node number 122: 15 observations
mean=1.666667, MSE=0.2222222
Node number 123: 14 observations
mean=2, MSE=0
Node number 160: 35 observations
mean=1, MSE=0
Node number 161: 7 observations
mean=1.142857, MSE=0.122449
Node number 168: 13 observations
mean=1, MSE=0
Node number 169: 13 observations
mean=1.230769, MSE=0.1775148
Node number 170: 17 observations
mean=1.176471, MSE=0.1453287
Node number 171: 28 observations, complexity param=0.008947702
mean=1.5, MSE=0.25
left son=342 (11 obs) right son=343 (17 obs)
Primary splits:
ldl < 4.18 to the left, improve=0.13368980, (0 missing)
age < 37.5 to the left, improve=0.13368980, (0 missing)
tobacco < 0.55 to the left, improve=0.08888889, (0 missing)
famhist splits as LR, improve=0.08888889, (0 missing)
typea < 56.5 to the right, improve=0.08888889, (0 missing)
Surrogate splits:
adiposity < 24.8 to the left, agree=0.750, adj=0.364, (0 split)

obesity < 24.83 to the left, agree=0.750, adj=0.364, (0 split)
sbp < 133 to the left, agree=0.679, adj=0.182, (0 split)
age < 43.5 to the right, agree=0.679, adj=0.182, (0 split)
tobacco < 0.14 to the left, agree=0.643, adj=0.091, (0 split)
Node number 342: 11 observations
mean=1.272727, MSE=0.1983471
Node number 343: 17 observations
mean=1.647059, MSE=0.2283737

Aneksi D: Tre bazat e të dhënave (ruajtur ne Exel)

a. Baza e të dhënave nga Spitali i Afrikës se Jugut

row.names	sbp	tobacco	ldl	adiposity	famhist	typea	obesity	alcohol	age	chd
1	160	12	5.73	23.11	Present	49	25.3	97.2	52	Y
2	144	0.01	4.41	28.61	Absent	55	28.87	2.06	63	Y
3	118	0.08	3.48	32.28	Present	52	29.14	3.81	46	N
4	170	7.5	6.41	38.03	Present	51	31.99	24.26	58	Y
5	134	13.6	3.5	27.78	Present	60	25.99	57.34	49	Y
6	132	6.2	6.47	36.21	Present	62	30.77	14.14	45	N
7	142	4.05	3.38	16.2	Absent	59	20.81	2.62	38	N
8	114	4.08	4.59	14.6	Present	62	23.11	6.72	58	Y
9	114	0	3.83	19.4	Present	49	24.86	2.49	29	N
10	132	0	5.8	30.96	Present	69	30.11	0	53	Y
11	206	6	2.95	32.27	Absent	72	26.81	56.06	60	Y
12	134	14.1	4.44	22.39	Present	65	23.09	0	40	Y
13	118	0	1.88	10.05	Absent	59	21.57	0	17	N
14	132	0	1.87	17.21	Absent	49	23.63	0.97	15	N
15	112	9.65	2.29	17.2	Present	54	23.53	0.68	53	N
16	117	1.53	2.44	28.95	Present	35	25.89	30.03	46	N
17	120	7.5	15.33	22	Absent	60	25.31	34.49	49	N
18	146	10.5	8.29	35.36	Present	78	32.73	13.89	53	Y
19	158	2.6	7.46	34.07	Present	61	29.3	53.28	62	Y
20	124	14	6.23	35.96	Present	45	30.09	0	59	Y
21	106	1.61	1.74	12.32	Absent	74	20.92	13.37	20	Y
22	132	7.9	2.85	26.5	Present	51	26.16	25.71	44	N
23	150	0.3	6.38	33.99	Present	62	24.64	0	50	N
24	138	0.6	3.81	28.66	Absent	54	28.7	1.46	58	N
25	142	18.2	4.34	24.38	Absent	61	26.19	0	50	N
26	124	4	12.42	31.29	Present	54	23.23	2.06	42	Y
27	118	6	9.65	33.91	Absent	60	38.8	0	48	N
28	145	9.1	5.24	27.55	Absent	59	20.96	21.6	61	Y
29	144	4.09	5.55	31.4	Present	60	29.43	5.55	56	N
30	146	0	6.62	25.69	Absent	60	28.07	8.23	63	Y
31	136	2.52	3.95	25.63	Absent	51	21.86	0	45	Y
32	158	1.02	6.33	23.88	Absent	66	22.13	24.99	46	Y
33	122	6.6	5.58	35.95	Present	53	28.07	12.55	59	Y

b. Baza e të dhënave nga spitali i Cleveland Clinic, Ohio USA.

ALLCAD	BNP	CRP16	DLDL	UHDL	DIABETICS	smoking	CVDYN	AGE	GENDER	CRECLR	HTN
Y	102.708	0.92	94	40.3	ND	NS	Y	55.9206	M	111.8077	YH
Y	74.439	5.72	66	31	YD	YS	Y	59.52361	M	100.8681	NH
Y	34.911	0.45	62	37.8	ND	YS	Y	63.46338	M	99.21414	YH
Y	115.101	3.63	88	28.9	ND	YS	Y	78.33812	M	100.0754	YH
Y	121.257	2.62	57	31.6	ND	NS	Y	75.34839	F	65.72415	YH
Y	60.021	3.03	107	30.6	ND	YS	Y	66.45311	M	90.79862	NH
Y	85.374	11.59	121	38.8	YD	NS	Y	74.4449	F	59.88623	NH

Y	79.866	14.04	81	41.8	YD	NS	Y	76.95551	F	58.83319	YH
Y	71.442	3.79	117	35.2	ND	YS	Y	47.42231	M	183.6859	NH
Y	25.839	3.29	74	28.4	YD	NS	Y	46.71321	M	155.8404	YH
Y	218.052	6.77	73	39.6	ND	YS	Y	65.40178	M	111.0856	NH
Y	36.045	2.27	102	33.8	ND	YS	Y	72.10404	M	121.4273	NH
Y	521.883	2.33	74	35	ND	NS	Y	68.17248	M	105.4286	YH
Y	68.364	1.42	110	38.2	ND	YS	Y	63.50992	M	95.61259	NH
Y	1073.088	9.88	63	37.1	ND	YS	Y	74.96509	M	10.3427	YH
Y	16.929	1.39	96	30.9	YD	YS	Y	74.72416	M	87.21807	NH
Y	329.994	38.88	113	45	ND	NS	Y	61.06229	F	113.071	YH
Y	45.522	0.96	62	27	ND	NS	Y	58.98973	M	109.8031	YH
Y	34.506	4.14	58	26	ND	YS	Y	72.40794	F	58.40752	YH
Y	17.334	1.39	71	27.5	YD	NS	Y	48.53114	M	127.0401	YH
Y	59.535	3.06	113	40.4	ND	NS	Y	71.38672	M	93.17853	NH
Y	710.127	7.64	123	41.7	YD	YS	Y	75.12389	F	38.29493	YH
Y	93.312	0.79	115	41.3	YD	YS	Y	56.59138	M	129.8026	YH

c. Baza e të dhënave “Boston House Matket”, USA.

CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PT	B	LSTAT	MV
0.00632	18	2.31	0	0.538	6.575	65.2	4.09	1	296	15.3	396.9	4.98	24
0.02731	0	7.07	0	0.469	6.421	78.9	4.9671	2	242	17.8	396.9	9.14	21.6
0.02729	0	7.07	0	0.469	7.185	61.1	4.9671	2	242	17.8	392.83	4.03	34.7
0.03237	0	2.18	0	0.458	6.998	45.8	6.0622	3	222	18.7	394.63	2.94	33.4
0.06905	0	2.18	0	0.458	7.147	54.2	6.0622	3	222	18.7	396.9	5.33	36.2
0.02985	0	2.18	0	0.458	6.43	58.7	6.0622	3	222	18.7	394.12	5.21	28.7
0.08829	12.5	7.87	0	0.524	6.012	66.6	5.5605	5	311	15.2	395.6	12.43	22.9
0.14455	12.5	7.87	0	0.524	6.172	96.1	5.9505	5	311	15.2	396.9	19.15	27.1
0.21124	12.5	7.87	0	0.524	5.631	100	6.0821	5	311	15.2	386.63	29.93	16.5
0.17004	12.5	7.87	0	0.524	6.004	85.9	6.5921	5	311	15.2	386.71	17.1	18.9
0.22489	12.5	7.87	0	0.524	6.377	94.3	6.3467	5	311	15.2	392.52	20.45	15
0.11747	12.5	7.87	0	0.524	6.009	82.9	6.2267	5	311	15.2	396.9	13.27	18.9
0.09378	12.5	7.87	0	0.524	5.889	39	5.4509	5	311	15.2	390.5	15.71	21.7
0.62976	0	8.14	0	0.538	5.949	61.8	4.7075	4	307	21	396.9	8.26	20.4
0.63796	0	8.14	0	0.538	6.096	84.5	4.4619	4	307	21	380.02	10.26	18.2
0.62739	0	8.14	0	0.538	5.834	56.5	4.4986	4	307	21	395.62	8.47	19.9
1.05393	0	8.14	0	0.538	5.935	29.3	4.4986	4	307	21	386.85	6.58	23.1
0.7842	0	8.14	0	0.538	5.99	81.7	4.2579	4	307	21	386.75	14.67	17.5
0.80271	0	8.14	0	0.538	5.456	36.6	3.7965	4	307	21	288.99	11.69	20.2
0.7258	0	8.14	0	0.538	5.727	69.5	3.7965	4	307	21	390.95	11.28	18.2
1.25179	0	8.14	0	0.538	5.57	98.1	3.7979	4	307	21	376.57	21.02	13.6
0.85204	0	8.14	0	0.538	5.965	89.2	4.0123	4	307	21	392.53	13.83	19.6
1.23247	0	8.14	0	0.538	6.142	91.7	3.9769	4	307	21	396.9	18.72	15.2