Contents

[KAPITULLI 2 1](#_Toc44722643)

[2. Data Deluge dhe KDD 1](#_Toc44722644)

[2.1 Pwrmbytja nga tw dhwnat 2](#_Toc44722645)

[2.2 Procesi i KDD 3](#_Toc44722646)

[KAPITULLI 3 6](#_Toc44722647)

[3. Hyrje nw machine learning (nga Expert System, Machine Learning) 6](#_Toc44722648)

[3.1 Cfarw wshtw machine learning? 6](#_Toc44722649)

[3.2 Cilat janw tipet e ndryshme tw Machine Learning? 6](#_Toc44722650)

[3.3 Historiku i ML (nga dataversity.net a brief history of ML Keith D.Foote 2019 link : https://www.dataversity.net/a-brief-history-of-machine-learning/#) 7](#_Toc44722651)

# ABSTRAKT

Machine Learning wshtw njw nga fushat që po fiton rëndësi me kohën për shkak të fenomeneve të ndryshme si ai i data deluge(shikoje njer kte se pertoj). Qëllimi kryesor i këtij studimi, është dhënia e një qasjeje më të afërt në lidhje me algoritmat e machine learning dhe marrja nw shqyrtim e njw rasti praktik pwr aplikimin e tyre.

Në këtë studim, bwhet njw hyrje nw lidhje me koncepte teorike mbi procesin e machine learning, tipet e machine learning, historikun etj duke u ndalur kryesisht nw tre algoritma tw rwndwsishwm pwr tw bwrw njw prezantim teorik tw tyre.

Në fund, fokusi vendoset te praktika, duke studiuar njw rast me qwllim realizimin e njw parashikimi nwpwrmjet kwtyre algoritmave pwr tw parw se cili performon mw mirw, përmes gjuhws Python librarive dhe funksioneve të larmishme që kjo gjuhë ofron.

**Fjalë kyce :**

Machine Learning, Data mining, Knowledge Discovery in Databases, Klasifikimi, Parashikimi, Shoqërimi,Supervised, Unsupervised, Regresion Logjistik, Decision Tree, Support Vector Machine, Algoritëm, Matricw Konfuzioni, Raport Klasifikimi, Test, Trajnim, Rrjeta Neurale etj.

# LISTA E FIGURAVE

## KAPITULLI 2

## 2. Data Deluge dhe KDD

Në këtë kapitull do të paraqiten nocionet kryesore mbi data deluge dhe procesin e Knowledge Discovery in Database.

## 2.1 Pwrmbytja nga tw dhwnat

Zhvillimi I shpejtë dhe në masa të konsiderueshme teknologjik që ka pësuar shoqëria në vitet e fundit, shoqërohet me lindjen e një sërë problematikash të reja me të cilat tashmë kërkojnë zgjidhje inteligjente dhe me shpejtësi relativisht të njëjtë me atë të krijimit të tyre. Një nga këto probleme është edhe e ashtuquajtura “data deluge”, e shqipëruar do të ishte “përmbytje të dhënash”. Aktualisht, ne gjenerojmë sasi voluminoze të dhënash gjatë cdo sekonde të përdorimit të rrjeteve të cakutara.

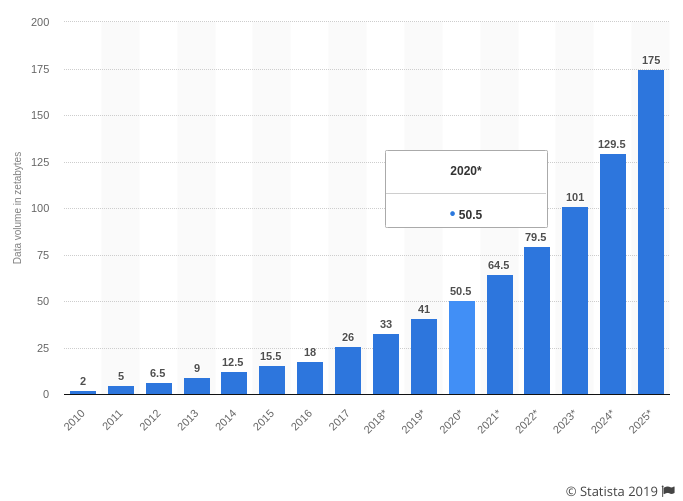
Sipas një artikulli nga Forbes, “Cdo ditë, ne gjenerojmë rreth 2.5 kuintilion byte të dhenash … vetëm në dy vitet e fundit kemi gjeneruar 90 përqind të nivelit të përgjithshëm botëror të të dhënave” (Bernard Marr, Enterprise Tech). Këto të dhëna rriten në mënyrë eksponenciale dhe vijnë nga burime të ndryshme, duke ju referuar përsëri artikullit të mësipërm:një nga burimet kryesore është përdorimi I internetit, vetëm web browser-I Google proceson rreth 3.5 bilion kërkime në ditë, ndërkohë që nësë përfshijmë këtu edhe browsera të tjerë shifrat rriten deri në 5 bilion.

Fig 2.1 Rritja eksponenciale e sasisë së të dhënave përgjatë viteve

Nga ana tjetër, një burim fondamental I të dhënave mbeten rrjetet sociale ku vetëm rrjeti social Facebook ka rreth 1.5 bilion përdoruesë aktivë të cilët gjenerojnë të dhëna në përditshmëri duke përfshirë këtu foto, komente, të dhëna personale, “pëlqime” etj. E nëse kalojmë më pas tek komunikimi, I cili kryhet edhe përmes rrjeteve sociale por edhe nëpërmjet aplikacioneve të posacme si ëhatsapp, Skype apo edhe rrjeteve të komunikimit elektronik e-mail. Gjithashtu, sasi marramendëse të dhënash gjenerohen përmes sektorit të financave, shërbimet, shëndetësia etj.

Pra në një situatë të tillë, volumi I madh I të dhënave të reja që gjenerohen është mbi kapacitetin menaxhues të institucioneve dhe të kërkuesve për ta përdorur. Rrjedhimisht, kemi lindjen e fushave të tilla të studimit sic është Big Data. Big Data është një fushë e cila përfshin zbulimin e mënyrave për të arritur në analizën, nxjerrjen sistematike të informacionit nga sasi të tilla voluminoze të dhënash të cilat janë shumë të mësha apo komplekse për tu analizuar nëpërmjet software-ve tradicionalë të procesimit të të dhënave.

## 2.2 Procesi i KDD

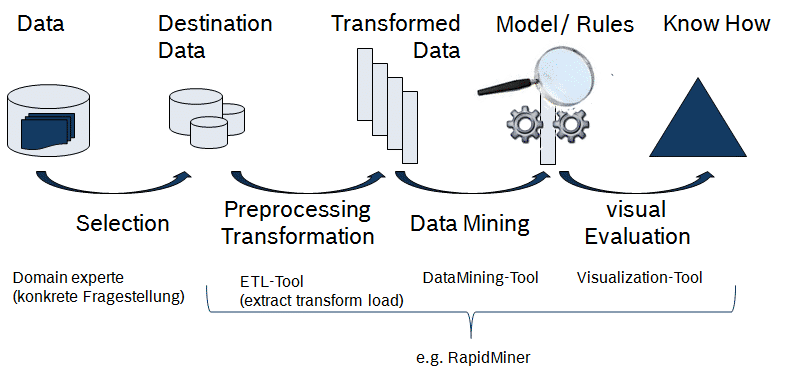
KDD është një shkurtim për Knowledge Discovery in Databases, ky term I referohet procesit të gjerë të gjetjes së “njohurive” brenda sasive të mëdha të të dhënave. Është procesi I përdorimit të metodata të Data Mining, për të arritur në ekstraktimin e informacionit të dobishëm .Ky process përfshin disa hapa të cilat ilustrohen në figurën më poshtë.

Fig 2.1 Skema e hapave të procesit të Knowledge Discovery in Databases

Fillimisht përcaktohen objektiva të qartë mbi mënyrën se si do të përdoren të dhënat, cfarë vendimesh duhen marrë në lidhje me transformimin e tyre, algoritmat që do të përdoren, mënyra e vizualizimit etj.

##### 2.2.1 Selektimi i tw dhwnave

Hapi I parë pas përcaktimit të qartë të qëllimeve, hapi që vjen është është “data selection”, pra selektimi I të dhënave. Gjatë këtij procesi, qe njihet edhe me ermin “Data Targeting”, determinohen të dhënat që do të përdoren. Zbulohen cilat të dhëna janëtë aksesueshme dhe grumbullohen të dhëna që sic e thamë më parë vijnë nga burime të ndryshmë, por njekohësisht, janw edhe të strukturuara në mënyra të ndryshme. .   
Aktualisht, të dhënat I gjejmë në disa formate :

* Të dhëna të strukturuara : përfshihen të dhëna të cilat kanë një format të qartë e të mirëpërcaktuar, nga ku edhe procesi i ekstraktimit të informacionit është shumë I thjeshtë, si një shembull këtu mund të përmendim formatin excel.
* Të dhëna të semi-sturkturuara : të dhënat kanë një format të dukshëm, me pak përpunim mund të strukturohen me qëllim nxjerrjen e informacionit, përfshihen formate të tipit XML.
* Të dhëna të quasi-strukturuara : kanë një format të crregullt, kërkojnë formatim më thelbësor për ekstraktim informacioni.
* Të dhëna të pastrukturuara : nuk kanë asnjë lloj formati rrjedhimisht edhe nxjerrja e informacionit është më e vështirë.

Pas zbulimit të këtij seti të dhënash të cilat janë të aksesueshme, teargetojmë ato të dhëna të cilat kanë probabilitet të lartë të gjetjes së informacionit të nevojshëm. Më pas, këto të dhëna të cilat vijnë nga burime të ndryshme dhe në formate të ndryshme për sa I përket sturkutimit, integorhen së bashku në një vendndodhje të vetme. Rezultati i këtij procesi janë “të dhënat desinacion”.

##### 2.2.2 Para-procesimi i tw dhwnave

Kjo fazë përfshin kuptimin e sturkturës së të dhënave ose lidhjet ndërmjet të dhënave. Të dhenat destinacion që ishin si rezultat i hapit të mëparshëm i nënshtrohen një procesi “pastrimi”. Gjatë pastrimit të të dhënave, dy janë veprimet kryesore :

1. Identifikimi i vlerave që mungojnë dhe nëse ka, gjetja e mënyrës më optimale për zëvëndësimin e tyre, qoftë ky eleminimi i vlerave apo mbushja e tyre me vlera të tjera sipas mesatares, modës apo përmes ndonjë algoritmi parashikues mbi përafrimin maksimal të vlerës origjiniale.
2. Zbulimi i të ashtuquajtrve “outliers” të cilët në thelb janë vlera ekstremisht të larta apo të ulëta, anomali të cilat mund të vijnë si rezultat i gabimeve gjatë marrjes së të dhënave.

Detektimi dhe zgjidhja e këtyre dy cështjeve realizohet përmes analizave statistikore, ndërtimit të grafikëve të shpërndarjes normale, përmes box-plotting apo edhe përdorimit të algoritmave të Data Mining që merren me parashikimin e mënyrave më optimale për zëvëndësimin, eleminimin apo normalizimin e të dhënave në varësi të rastit.

Disa nga operacionet që mund të realizohen mbi të dhënat në këtë fazë janë :

***Agregimi :*** mbi të dhënat aplikohen operime përmbledhjeje apo agregimi, psh të dhënat e shitjeve javore agregohen për të llogatitur totalin mujor dhe vjetor.

***Gjeneralizimi :*** në këtë hap, të dhënat e një niveli më të ulët zëvëndësohen nga koncepte të niveleve më të larta në bazë të konceptit të hierarkisë. Psh, qyteti zëvëndësohet nga shteti.

***Normalizimi :*** Normalizimi performohet kur disa të dhëna të caktuara rrumbullakosen (te numri shifra më e madhe e përafërt ose te shifra më e vogël).

***Ndërtimi i atributeve:*** këto atribute ndërtohen dhe përfshihen në set-it e atributeve ndihmues për data mining.

Rezultati i këtij procesi janë të dhënat përfundimtare, të cilat mund të përdoren në modelim.

##### 2.2.3 Data Mining

##### 2.2.3.1 Cfarë është Data Mining ?

Data mining është procesi llogatitës i zbulimit të “pattern”-ave në bashkësi voluminoze të të dhënave duke përfshirë metoda nga statistika, machine learning dhe të sistemeve të databazës. Shkurtimisht, kthimi i të dhënave të papërpunuara në informacion të përdorshëm.  
Data Mining është hapi më i rëndësishëm në KDD. Përfshin aplikimin e veprimeve inteligjente si: clustering, kalsifikimi, regresioni, për ekstraktimin e patterns.

##### 2.2.3.2 Cilat janë objektivat e data mining?

Objektivat kryesorë të data mining janë zbulimi i trendeve apo paternave dhe parashikimi i sjelljeve të së ardhmes. Kompanitë të cilat ekstraktojnë informacion të dobishëm në mënyrë të suksesshme e kanë më të thjeshtë të rrisin përfitimet e tyre duke u bazuar në sjelljet e mëparshme të klientëve. Shërben për rritjen e efektshmërisë së marketingut, rritjen e fitimit, ulen e kostove etj.

##### 2.2.3.3 Data mining si pjesë e KDD

Gjatë këtij hapi, të dhënat i nënshtrohen një transformimi që mund të përfshijë reduktimin e volumit të të dhënave përmes “sampling”, transformimit të atributeve përmes diskretizimit të atributeve numerike etj. Ky hap është i orientuar nga lloji i projektit që po zhvillohet dhe në varësi të objektivave të projektit përzgjidhen edhe strategjitë që do të ndiqen, algoritmet e Data Mining që do të përdoren si dhe teknikat më frytdhënëse, e më të përshtatshmë në varësi të llojit të projektit që do të zhvillohet dhe objektivave të tij. Rezultati i këtij hapi janë patternat, pra ekstraktimi i informacionit.

##### 2.2.4 Vlerwsimi i patternave

Gjatë këtij hapi, interpretohen paternat e ekstraktuar gjate procesit të data mining. Informacioni i nxjerrë nga të dhënat ekzistuese duhet të ketë tre karakteristika :

1. Duhet të jetë informacion i ri : nuk ka kuptim të shpenzosh kohë, para dhe energji për të zbuluar informacione ekzistuese.
2. Duhet të jetë i saktë : informacioni i ekstraktuar duhet vlerësuar para përdorimit sepse ai mund të jetë i pasaktë ose i pavlefshëm.
3. Informacioni duhet të jetë potencialisht i përdorshëm : Duhet të nxjerrim atë lloj informacioni që mund ta përdorim dhe që na hyn në punë.

KETU TE BEHET LIDHJA ME MACHINE LEARNING!!!

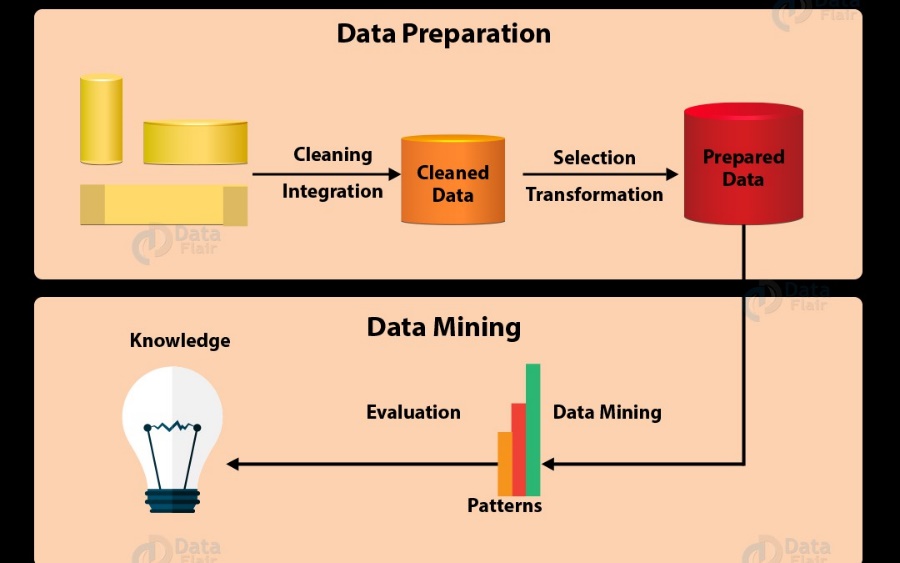


Fig 2.2 Procesi i ekstraktimit të njohurive

## KAPITULLI 3

## Hyrje nw machine learning (nga Expert System, Machine Learning)

## Cfarw wshtw machine learning?

Machine learning (ML) pwrkufizohet si njw aplikim i inteligjencws artigiciale (AI) i cili u jep sistemeve aftwsinw pwr tw mwsuar nw mwnyrw automatike si dhe pwrmirwsim i performancws duke pwrdorur eksperienca tw mwparshme pa qenw tw programuar nw mwnyrw eksplicite.

Mw specifikisht, machine learning wshtw njw qasje ndaj analizws sw tw dhwnave e cila pwrfshin ndwrtimin dhe pwrshtatjen e modeleve, tw cilat lejojnw mwsimin pwrmes eksperiencws. Procesi i mwsimit fillon me observimin e tw dhwnave, me qwllim gjetjen e patternave nw sasi voluminoze tw dhwnash pwr tw marrw vendime mw tw mira nw tw ardhmen bazuar nw shembujt e siguruar nga tw dhwnat.

ML pwrfshin ndwrtimin e algoritmave tw cilwt pwrshtatin modelet e tyre me qwllim pwrmirwsimin e aftwsive tw tyre pwr parashikimin e situatave. Pwrdoren algoritma tw cilwt nw mwnyrw itearative mwsojnw nga tw dhwnat se si tw pwrmirwsojnw, pwrshkruajnw tw dhwnat dhe parashikojnw rezultatet. Njw algoritwm i tillw parashikues krijon njw model parashikues, i cili do tw arrijw tw parashikojw rezultate pwr tw dhwna tw reja bazuar nw tw dhwnat e pwrdorura pwr trajnimin e modelit.

## Cilat janw tipet e ndryshme tw Machine Learning?

Algoritmat e ML kanw si qwllim mwsimin dhe pwrmirwsimin e saktwsisw sw tyre pwrmes procesimit tw sasive progresisht mw tw mwdha tw tw dhwnave. Pwr tw realizuar kwtw qwllim nw varwsi tw natyrws sw problemit, ka disa qasje tw ndryshme bazuar nw tipin dhe volumin e tw dhwnave qw po trajtohen. Mw poshtw jepet njw pwrshkrim i kategorive nw tw cilat klasifikohen kwto detyra.

##### Supervised learning (Tw mwsuarit e kontrolluar)

Ky lloj i tw mwsuarit zakonisht nis mw njw set tw konsoliduar tw dhwnash, me njohuri tw mira paraprake mbi mwnyrwn e klasifikimit tw kwtyre tw dhwnave. Qwllimi i surpervised learning wshtw ekstraktimi nga tw dhwnat i patternave tw cilat mund tw aplikohen nw njw proces analitik. Tw mwsuarit e kontrolluar zakonisht realizohet nw kontekstin e klasifikimit apo regresionit. Algoritma qw pwrdoren nw kwtw lloj tw mwsuari pwrfshijnw regresionin logjik, pemwt e vendimeve, support vector machines (SVM), rrjetat artificiale neurale etj. Synimi si nw rastin e klasifikimit si nw atw tw regresionit wshtw gjetja e lidhjeve specifike apo strukturave brenda tw dhwnave input tw cilat na lejojnw prodhimin korrekt tw tw dhwnave output, e thwnw ndryshe, parashikimin e saktw tw tyre. Kwtu vlen tw pwrmendet se, tw dhwnat e pasakta e ulin nw mwnyrw tw dukshmw eficencen e modelit.

Algoritmat trajnohen duke pwrdorur shembuj tw paraprocesuar dhe performanca e algoritmave vlerwsohet pwrmes tw dhwnave test. Nw raste tw caktuara, patternat e identifikuar nw njw pjesw tw tw dhwnave nuk mund tw detektohen nw njw sasi mw volumioze tw dhwnash. Raste tw tilla, ku modeli wshtw trajnuar vetwm pwr tw pwrfaqwsuar patternat e gjetur nw njw sasi mw tw vogwl tw dhwnash, gjenerojnw njw lloj problemi tw quajtur “overfitting”. Overfitting so tw thotw qw modeli i krijuar mund tw performojw me sukses pwr njw set tw pwrcaktuar tw dhwnash, por dwshton kur aplikohet mbi sete tw panjohura tw dhwnash voluminoze. Mbrojtja ndaj kwtij problemi realizohet pwrmes testimit tw vazhdueshwm.

##### Unsupervised Learning (Tw mwsuarit e pakontrolluar)

Ky lloj i tw mwsuarit, wshtw mw i pwrshtatshwm nw raste kur problemi kwrkon pwrdorimin e njw sasie shumw tw madhe tw dhwnash tw paemwrtuara, pra pa pasur njohuri paraprake mbi kto tw dhwna. Analizimi dhe kuptimi i kwtyre tw dhwnave kwrkon algoritma tw cilwt janw tw aftw tw klasifikojnw tw dhwnat bazuar nw patternat apo kllasterat e gjetur. Prandaj, tw mwsuarit e pakontrolluar kwrkon njw proces iterativ tw analizimit tw tw dhwnave pa ndwrhyrjen e njeriut.

Algoritmat qw bien nwn kwtw kategori segmentojnw tw dhwnat nw grupe shembujsh ose grupe cilwsish. Nw thelb, pwrmes kwtij procesi gjejmw emwrtime pwr tw dhwnat duke e shndwrruar mw pas nw tipin e mwsipwrm (tw mwsuarit e kontrolluar). Ngjashwmrisht me algoritmat e supervised learning, edhe kwtu algoritmat kwrkojnw pwr gjetjen e patternave por nw kwtw rast nuk kemi asnjw lloj kuptimi mbi tw dhwnat qw po analizohen.

##### Reinforcement learning (Tw mwsuarit e pwrforcuar)

Ky lloj i tw mwsuarit wshtw i bazuar mbi sjelljen, pra, algoritmi merr informacion nga analizimi i tw dhwnave nw mwnyrw qw pwrdoruesi tw drejtohet tek rezultati optimal. Ndryshe nga llojet e tjera pwr tw cilat kemi folur deri mw tani, sistemi kwtu nuk trajnohet me njw set tw dhwnash, por sistemi mwson pwrmes provws dhe gabimit. Pra, njw sekuencw vendimesh tw suksesshme do tw rezultojw nw pwrforcimin e procesit sepse e ka zgjidhur problemin nw mwnyrwn mw tw saktw.

##### Rrjetat neurale dhe deep learning (tw mwsuarit e thellw)

Deep learning wshtw njw metode specifike e ML e cila inkorporon rrjetat neurale nw shtresa tw njwpasnjwshme me qwllim mwsimin nga tw shwnat nw mwnyrw iterative. Zakonisht pwrdoret kur kemi tw bwjmw me tw dhwna tw pastrukturuara.

Rrjetat neurale komplekse krijohen pwr tw simuluar mwnyrwn e punimit tw trurit njerwzor me qwllim trajnimin e kompjuterave pwr zgjidhjen e problemave tw papwrkufizuar nw mwnyrwn e duhur apo abstraksioneve. Njw rrjet neural konsiston nw tre a mw shumw shtresa : njw shtresw inputi, njw apo disa shtresa tw fshehta (nw varwsi tw kompleksitetit tw problemit)dhe njw shtresw outputi. Tw dhwnat merren pwrmes shtresws sw inputit, mw pas tw dhwnat modifikohen nw shtresat e fshehura dhe nw fund rezultati merret pwrmes shtresws sw outputit.

Deep learning wshtw njw teknikw e machine learning e cila pwrdor njw hierarki rrjetash neurale pwr tw mwsuar nga njw kombinim i algoritmave tw kontrolluar dhe tw pakontrolluar dhe zakonisht ndwrverpron me tw dhwna tw paemwrtuara dhe tw pastrukturuara.

## Historiku i ML (nga dataversity.net a brief history of ML Keith D.Foote 2019 link : https://www.dataversity.net/a-brief-history-of-machine-learning/#)

Machine Learning wshtw pjswrisht e bazuar nw njw model tw interaksionin ndwrmjet qelizave tw trurit, njw model ky i krijuar nw 1949 nga Donald Hebb nw njw libwr tw titulluar “Organizimi i sjelljes” (nga origjinali : “The Organization of Behaviour”). Ky libwr parqet teoritw e Hebb mbi eksitimin e neuroneve dhe komunikimin mes tyre, pwr krijimin e rrjeteve neurale atrificiale. Mw pas, nw 1950 Arthur Samuel zhvilloi njw program kompjuterik pwr lojwn Checkers, duke iniciuar dicka tw njohur si “alpha-beta pruning”. Programi zgjedh lwvizjen e rradhws duke pwrdorur njw strategji “minimax”, pra minimizimi i humbjes maksimale qw mund tw rezultojw nga njw zgjedhje e gabuar. Strategji kjo, qw eventualisht solli krijimin e Algoritmit Minimax.

Nw 1957, Frank Rosenbalt kombinoi modelin e Donald Hebb me atw tw Arthur Samuel dhe krijoi “perceptronin”, qw wshtw njw njwsi e njw rrjeti neural, pra njw neuron artificial. Zhvillimi tjetwr vjen nw vitin 1967, kur u krijua algoritmi i fqinjit mw tw afwrt (nga origjinai : The Nearest Neighbor Algorithm), algoritwm ky i pwrdorur pwr zgjidhjen e problemeve tw klasifikimit apo regresionit. Nw vitet ’60 u arrit njw zhvillimi i ri nw lidhje mw rrjetat neruale si rrjedhim i zbulimit tw “multilayers” (nga anglishtja: shumw shtresa).

Nw fundin e viteve ’70, dhe fillimi i viteve ’80, ndodhi mw nw fund ndarja ndwrmjet ML dhe AI. Inteligjenca Artificiale u fokusua nw pwrdorimin e qasjeve logjike, knwledge-based dhe braktisi kwrkimet nw fushwn e rrjetave neurale. Kwshtu, ML pwsoi njw riorganizim duke u fokusuar mw shumw nw taktikat e pwrdorura nw teorinw e probabilitetit dhe statistikw duke ofruar zgjidhje pwr probleme praktike. Lulwzimi i industrisw sw ML u realizua nw vitet ’90, kryesisht si rezultat i zhvillimit tw Internetit duke pwrfituar nga disponueshmwria nw rritje e tw dhwnave digjitale.

Gjatw viteve ’90, u krijuan edhe tw ashtuquajturit “Boosting Algorithms” tw cilwt pwrdoren gjatw tw mwsuarit e kontrolluar pwr zvogwlimin e variancws dhe disbalancave, si dhe pwr rritjen e performancws sw pwrgjithshme tw modelit. Mw vonw, pwrmes njw teknike tw Deep Learning tw quajtur Long-Short Term Memory (LSTM), u punua pwr realizimin e speech recognition. Nw 2015, programi i Google pwr njohjen e zwrit pati njw pwrmirwsim rreth 49% nw performancwn e tij pwrmes pwrdorimit tw kwsaj teknike. Nw 2006, arriti tw bwhej realitet njohja e fytyrws me algoritma qw performonin rreth njw qind herw mw mirw se algoritma tw mwparshwm. Nw 2014, Facebook zhvilloi algoritmin DeepFace, i cili wshtw i aftw tw njohw dhe verifikojw fytyrat e individwve me tw njwjtwn saktwsi si tw njeriut.

Nw vitet e fundit, ML wshtw pwrgjegjwse pwr disa nga zhvillimet mw domethwnwse nw teknologji, si pwr shembull industria e re e makinave qw vetwdrejtohen. Algoritmat e ML tw kombinuar teknologjitw e reja tw llogaritjes promovojnw pwrshkallwzueshmwrinw dhe pwrmirwsojnw eficencwn.

KAPITULLI 4

Nw kwtw kapitull do tw flasim pwr disa nga algoritmat e ML, pwrkatwsisht Logistic Regression (LR), Decision Tree (DT) dhe Support Vector Machine (SVM).

## 4.1 Algoritmi Logistic Regression

Logistic regression është një mjet statistikor pwr modelimin e lidhjes ndwrmjet disa variablave “shpjeguese” dhe njw rezultati tw pritshwm me vlera reale. Në këtë model, probabilitetet që përshkruajnë përfundimet e mundshme të një seti të dhënash tw cilat modelohen duke përdorur një funksion logjistik.

Ndryshe nga regresioni linear i zakonshëm ku variabli rezultat wshtw zakonisht njw vlerw e vazhdueshme, nw regresionin logjistik rezultati wshtw njw event binar. Ka dy modele tw regresionit logjistik, regresioni logjistik binar dhe regresioni logjistik multinomial. E para pwrdoret zakonisht nw raste kur variabli i varur wshtw dikotomik (binar) dhe variablat e pavarura janw tw vazhdueshme ose kategorike, ndwrsa e dyta nw rastet kur variabli i pavarur nuk wshtw dikotomik dhe wshtw i pwrbwrw nga mw shumw se dy kategori.

Mundësia e ngjarjes për një event është raporti midis probabilitetit që ky event do të ndodhë mbi probabilitetin që eventi nuk do të ndodhë. Nëse probabileti që nje event të ndodhë është *p* dhe të mos ndodhë *p-1* atëherë:

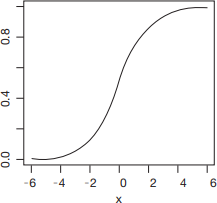
Duke qenë se LR llogari probabilitetin e ndodhjes të një eventi mbi probabilitetin e mos ndodhjes të tij, impakti i variablave të pavarura zakonisht sqarohet në terma të odds (mundwsive). Në LR mesatarja e response variables *p* në termat e një variable x modelohet duke lidhur p dhe x nëpërmjet ekuacionit p = α + βx (alpha & beta).

Fatekeqwsisht, ky nuk është një model i mirë pasi vlera të mëdha të x do të japin vlera tw p tw cilat nuk do tw ndodhen midis 0 dhe 1. Zgjidhja e LR ndaj kwtij problemi, është transformimi i odds duke përdorur logaritmin natyral (ln) (Peng Lee & Ingersoll, 2002). Atëherë do të kishim:

ku p është probabiliteti i rezultatit që na intereson dhe x ështe variabli shpjegues, ndwrsa alpha dhe beta janë parametrat e LR.

Një këndvështim i rëndësishëm në analizimin e LR është edhe Odds Ratio (OR). OR është raporti midis odds relative pwr dy evente të ndryshwm. Për dy evente A dhe B do të kishim:

LR është një metodë që përcakton një kurbë y = f(x) ku y konsiston nw tw dhwna tw koduara nw mwnyrw binare (0,1 – failure, success). Kur përgjigjia është një variabël binar dhe x është një variabël numerik, logistic regression përcakton një kurbë logjistike për marrëdhënien midis x dhe y. Kurba logjistike është në formë S-je ose e quajtur ndryshe sigmoida dhe zakonisht pwrdoret nw modelimin e rritjes sw popullsisw. Kjo kurbë fillon më një rritje të ngadaltë lineare, pasuar nga një rritje eksponenciale e cila mw pas ngadalwson swrish nw njw gjendje tw qwndrueshme.



**Fig. 4.1 Kurba logjistike**

##### 4.1.1 Pwrzgjedhja e variablave dhe selektimi i tyre

Një nga problemet më të mwdha në ndwrtimin e njw modeli logjistik është përzgjedhja e variablave që do të përfshihen. Zakonisht kwrkuesit përfshijnw e sa më shumë variablave të jetë e mundur, me idenw se njw sasi e konsiderueshme tw dhwnash mund tw sjellw nxjerrje mw tw madhe informacioni tw vlefshwm. Kjo është një qasje e gabuar, pasi përzgjedhja e të dhënave është më shumë e komplikuar se kaq. Futja e një numri të madh variablash në modelin në ndërtim, me qëllim gjetjen e një strukture sinjifikative, shpesh rezulton nw dy situata:

1. Një ose disa variabla janë sinjifikante nga ana statistikore, por studiuesi nuk mbështetet në teori për të lidhur variablat sinjifikantë me modelin e eventit për të cilin interesohet. Në këtë rast harrohet që punohet me shembuj dhe mund të ndodhë të përftohen rezultate të gabuar.
2. Një model me shumw variabla përfaqëson një fuqi të ulët statistikore. Prandaj, nëse ka një lidhje midis një variabli shpjegues dhe ngjarjes së një eventi, mund të ndodhë ta humbasim, sepse modelet e saturuara (ato që përmbajnë të gjitha variablat shpjeguese të mundshme) nuk janë mjaftueshëm të ndjeshme për ta dalluar.

Pra, kujdesi në përzgjedhjen e variablave është një hap shumë i rendësishëm para zhvillimit të modelit.

Zhvillimin e një modeli mund ta nisim ose me një model të saturuar ose me një model bosh, i cili fillon vetëm me termin “intercept”. Në rastin e parë, eleminojmë përfshirjen e variablave një nga një, duke nisur nga ato më pak sinjifikative dhe duke mbajtur ato më të rëndësishmet. Kjo strategji është e lehtë për tu menaxhuar.Nga ana tjetër, metoda e dytë kërkon që gjithë variablat të testohen në çdo hap dhe të përzgjidhet ajo më e mira në secilën fazë.

Nëse shumw variabla përfshihen përnjëheresh në një model mund të ndodhë që të eleminojmë variabla signifikativë.

Nëse kemi një madhësi të madhe mostre, le të themi kemi 10 individë për variabël, mudn të përpiqemi të përfshijmë gjithë variablat shpjeguese në modelin e plotë. Megjithatë, nëse kemi një madhësi të limituar në lidhje me numrin e variablave kandidate, është e këshillueshme të bëhet një selektim paraprak. Një mënyrë për të arritur këtë është testimi i variablave paraprakisht duek përdorur modele me vetëm një variabël në një kohë (modele univariabël) dhe më pas të përfshijmë në model ato variabla që kanë treguar një vlerë të arsyeshme të p (P <= 0.25). Nuk ka arsye të shqetësohemi për një kriter rigoroz për vlerën e *p* , sepse kjo është thjesht një fazë parapërgatitëse. Kriteri i vlerës se P lejon reduktimin e numrit fillestar të variablave në model duke ulur kështu rrezikun e mospërfshirjes së variablave sinjifikativë.

Ka një diskutim mbi strategjinë e duhur për përzgjehdjen e variablave[8]. Ka mënyra të ndryshme për të arritur këtë dhe është e rëndësishme të zgjidhet ajo më e përshtatshmja dhe të kuptohet paraprakisht, në mënyrë që të vazhdohet me zhvillimin e modelit.

##### 4.1.3 Validimi i algoritmit LR

Modelet LR përdoren zakonisht për të parashikuar një variabwl të varur nga një set variablash të pavarur. Një pyetje e rëndësishme është nëse përfundimet e analizës sw regresionit logjistik mbi një mostër mund të zgjerohen mbi popullatën nga e cila kjo mostër është marrë. Kjo pyetje referohet si validim i modelit. Në praktikë, një model mund të valiohet duke derivuar një model dhe duke vlerësuar koeficentët e tij në një dataset. Më pas përdorim këtë model për të parashikuar një dataset tjetër, dhe kontrollojmë mbetjet.

Kur një model validohet duke përdorur të dhënat mbi të cilat është zhvilluar, ka mundwsi që ai të jetë i mbivlerwsuar. Kështu, validimi i modelit duhet të vlerësohet duke zhvilluar teste të përshtatshmërisë në një dataset tjetër. Nëse një model është trajnuar me një nën set të një seti vëzhgimesh dhe validohet me pjesën e mbetur, quhet validim i brendshwm (internal). Metodat më të përdorura për një validim të mirë internal janë data-splitting, repeated data- spliting dhe bootstrapping.

Nëse validiteti testohet me një set të ri të dhënash të pavarur nga e njëjta popullsi apo në tjetër e ngjashme, quhet validim i jashtwm (external). Duke siguruar nje set të dhwnash të ri, bëjmë të mundur kontrollin e modelit në një kontekst tjetër. Nëse modeli i parë i përshtatet setit të dytë të të dhënave, kemi një farë sigurie mbi pergjithësimin e modelit. Megjithatë, nëse modeli nuk përshtatet me setin e dytë të të dhënave, mungesa e përshtatshmërisë mund të vijë si pasojë e konteksteve të ndryshme të dy seteve të të dhënave ose prej faktit që në të vërtetë modeli nuk është përshtatur si duhet.

Shumë nga avantazhet dhe diavantazhet e modelit linear regression aplikohen edhe për modelin e logistic regression. Logistic regression është përdorur gjerësisht, por lufton me kufizimet e tij dhe ekzistojnw modele tw tjera tw cilat mund të kenë performancë më të mirë parashikuese. Një tjetër disavantazh i modelit të regresionit logjistik është se interpretimi është më i vështirë.

Logistic regression ka probleme me ndarjen e plotë. Nëse ka një veçori që do t'i ndante në mënyrë të përkryer dy klasat, modeli i logistic regression nuk mund të trajnohet më. Kjo vjen për shkak se pesha për këtë funksion nuk do të konvergjojë, sepse pesha optimale do të ishte e pafundme. Kjo është me të vërtetë paksa për të ardhur keq, sepse një tipar i tillë është me të vërtetë i dobishëm. Problemi i ndarjes së plotw mund tw zgjidhet duke futur penalizimin e peshave ose duke pwrcaktuar njw shpwrndarje probabiliteti tw mwparshwm tw peshave.

Një avantazh tjetër është qw, modeli logistic regression nuk është vetëm një model klasifikimi, por gjithashtu jep dhe rezultate probabilitare. Ky është një avantazh i madh mbi modelet që mund të ofrojnë vetëm klasifikimin përfundimtar. Logistic regression mund gjithashtu të zgjerohet nga klasifikimi binar në klasifikimin e shumë klasaveduke marrw emwrtimin regresioni multinomial, sic pwrmendet edhe mw sipwr.

## 4.2 Algoritmi Decision Tree

Decision Trees konsiderohen si qasjet mw tw rwndwsishme pwr pwrfaqwsimin e klasifikimit, njw metodw e kontrolluar e cila rreket tw zbulojw lidhjet midis atributeve qw shwrbejnw si input dhe atributeve qw shwrbejnw si target. Një decision tree wshtw njw klasifikues i shprehur si njw ndarje rekursive e njw instance hapwsire. Ajo konsiston nw nyje tw cilat formojnë një pemë të drejtuar me nyjen kryesore që quhet rrënjë, e cila nuk ka asnjë degw hyrëse. Një nyje e cila ka degw dalëse quhet nyje e brendshme ose nyje test. Të gjitha nyjet e tjera quhen gjethe (të njohura ndryshe si nyje fundore ose nyje të vendimit). Në një decision tree (pemë vendimi), çdo nyje e brendshme e ndan hapësirën në dy nënhapësira të tjera sipas një funksioni diskret të vlerave të atributeve të inputeve. Në rastin më të thjeshtë dhe më të shpeshtë, çdo test konsideron një atribut të vetëm, të tillë që instanca e hapwsirws të ndahet sipas vlerës së atributit. Në rastin e atributeve numerike, kushti i referohet njw intervali.

Procesi i ndërtimit të decision tree bëhet në mënyrë rekursive. Fillimisht përzgjidhet një atribut për tu vendosur në nyjen rrënjë dhe ndërtohet degë për cdo vlerë të mundshme. Ky proces mund të përsëritet në mënyrë rekursive për cdo degë duke përdorur vetëm ato instanca që arrijnë degën. Nwse në një moment të caktuar të gjitha instancat në nyje kanë të njëjtin klasifikim ajo pjesë e pemës nuk do të zhvillohet. Kryesorja është të përcaktosh se cili atribut do të ndahet. Përvoja ka treguar se atribut në rrwnjw duhet vendosur ai që na jep sasinë më të madhe të informacionit në nyjet pasardhëse.

Çdo nyje i përket një klase, e cila përfaqëson vlerën më të përshtatshme të synuar. Gjethja mund të mbajë një vektor probabilitetesh që tregon probabilitetin e atributit të synuar kur ka një vlerë të caktuar. Instancat klasifikohen duke naviguar nga rrënja te gjethet, nw varwsi tw rezultateve gjatë rrugës.

Në rastin e atributeve numerike, DT mund të interpretohet gjeometrikisht si një grup hiperplanesh, secili ortogonal në lidhje me një bosht. Megjithatë, preferohen DT me pak komplekse. Kompleksiteti i pemës ka një efekt thelbwsor në saktësinë e saj. Zakonisht kompleksiteti i pemës matet me njw nga metrikat e mëposhtme:

* Numri total i nyjeve
* Numri total i gjetheve
* Thellësia e pemës
* Numri i atributeve të përdorura.

Çdo shteg nga rrënja e DT në një nga gjethet e saj mund të transformohet në një rregull vetëm duke bashkuar testet përgjatë shtegut për të formuar pjesën e mëparshme dhe për të marrë klasën e parashikimit të gjethes si vlerë.

##### 4.2.1 Framework-et e algoritmave pwr Decision Tree

DT ka disa forma algoritmi të cilwt ndërtojnw automatikisht një DT nga një set të dhënash. Qëllimi është të gjendet DT optimale duke minimizuar gabimin e përgjithshëm. Megjithatë, mund të përcaktohen edhe funksione të tjera, si për shembull minimizimi i numrit total të nyjeve apo thellësia mesatare. Përftimi i një DT optimale nga një set të dhënash konsiderohet një detyrë e vështirë.

Ka algoritma të ndryshwm që mund të zhvillojnë një pemë vendimi. Ato ndryshojnë në: strukturën e mundshme të pemës (p.sh. numri i ndarjeve për nyje), kriteret se si të gjejnë ndarjet, momentin e realizimit tw njw ndarjeje si dhe mwnyrwn se si do të vlerësohen modelet e thjeshta brenda nyjeve të fletës. Mund të përmendim:

* ID3
* C4.5
* CART

Algoritmi i klasifikimit dhe regresionit (CART) është ndoshta algoritmi më i popullarizuar për përftimin e pemëve.

Formula e mëposhtme përshkruan marrëdhënien ndërmjet rezultatit y dhe karakteristikave x.



Çdo instancë bie nën një nyje flete (nëngrupi Rm), ku {x∈Rm} është funksioni i identitetit që kthen 1 nëse x është në nëngrup Rm dhe 0 në të kundërtën. Nëse një shembull bie në një nyje fletësh Rl , rezultati i parashikuar është y=cl , ku cl është mesatarja e të gjitha instancave të trajnimit në nyjen e fletëve Rl .

Por nga vijnë këto subsset? Kjo është mjaft e thjeshtë: CART-i merr një tipar dhe përcakton se cila pikë e prerjes minimizon variancën e y për një detyrë regresioni ose indeksin Gini të shpërndarjes së klasës së y për detyrat e klasifikimit. Gjerësia tregon se sa vlerat y në një nyje janë përhapur rreth vlerës së tyre mesatare. Indeksi Gini na tregon se sa "e papastër" një nyje është, p.sh. nëse të gjitha klasat kanë të njëjtën frekuencë, nyja është e papastër, nëse vetëm një klasë është e pranishme, ajo është maksimalisht e pastër.

Varianca dhe indeksi Gini minimizohen kur pikat e të dhënave në nyjet kanë vlera shumë të ngjashme për y. Si pasojë, pika më e mirë e ndërprerjes i bën të dy nënkategoritë rezultuese sa më të ndryshme të mundshme në lidhje me rezultatin e synuar. Për karakteristikat kategorike, algoritmi përpiqet të krijojë nën-grupe duke provuar grupime të ndryshme të kategorive. Pas përcaktimit të ndërprerjes më të mirë për funksion, algoritmi përzgjedh tiparin e ndarjes që do të rezultonte në ndarjen më të mirë në kuptim të variancës ose indeksit Gini dhe shton këtë ndarje në pemë. Algoritmi vazhdon këtë kërkim-dhe-ndarje në mënyrë të përsëritur në të dy nyjet e reja derisa të arrihet një kriter i ndaluar. Kriteret e mundshme janë: Një numër minimal i rasteve që duhet të jenë në një nyje përpara ndarjes, ose numri minimal i rasteve që duhet të jenë në një nyje fundore.

Algoritmat e përdorur për përfitimin e DT kanë natyrë “lakmuese” dhe ndërtojnë pemën në një mwnyrë rekursive nga lart poshtë, e cila njihet ndryshe edhe si “përça e sundo”. Në çdo iteracion, algoritmi konsideron ndarjen e setit të trajnimit duke përdorur rezultatin e një funksioni diskret të atributeve të inputeve.

Përzgjedhja e funksionit më të përshtatshëm bëhet sipas disa përmasave ndarëse. Pas përzgjedhjes së ndarjes së duhur, secila nyje ndan setin trajnues më tej në subsete më të vogla deri sa ndarja nuk sjell më përfitim ose deri sa të arrihet një kusht ndalues.

##### Avantazhet dhe disavantazhet e DT

Disa nga avantazhet e pemwve tw vendimit si mjete klasifikimi jepen mw poshtw :

* Struktura e pemës është ideale për kapjen e ndërveprimeve ndërmjet karakteristikave në të dhëna. Të dhënat përfundojnë në grupe të dallueshme të cilat shpesh janë më të lehta për tu kuptuar se sa pikat në një hyperplane shumëdimensionale, si psh në regresionin linear. Interpretimi është padyshim shumë mw i thjeshtë. Struktura e pemës gjithashtu ka një vizualizim natyral, me nyjet dhe skajet e saj. Një pemë me një thellësi prej tre njësish kërkon një maksimum prej tre tiparesh dhe pika të ndara për të krijuar shpjegimin për parashikimin e një instance individuale. Pastërtia e parashikimit varet nga performanca parashikuese e pemës. Shpjegimet për pemët e shkurtra janë shumë të thjeshta dhe të përgjithshme.
* Tw lehta pwr tu kuptuar - Rezultati i decission tree është shumë i lehtë për t'u kuptuar. Nuk kërkon ndonjë njohuri statistikore për t'i lexuar dhe interpretuar pemwt e vendimeve. Kanw njw përfaqësim grafik shumë intuitiv.
* Tw dobishme nw eksplorimin e tw dhwnave. Decision Trees është janw ndwr mënyrat më të shpejta për të identifikuar variablat më të rëndësishëm dhe lidhjen ndërmjet dy ose më shumë ndryshoreve. Me ndihmën e decission trees, ne mund të krijojmë variabla të reja / karakteristika që kanë fuqi më të mirë për të parashikuar variablin e synuar. Gjithashtu mund të përdoret në fazën e eksplorimit të të dhënave. Për shembull, nwse jemi duke punuar për një problem ku kemi informacion pwr qindra variabla, nwpwrmjet decsion tres do të mund identifikojmw ndryshoren më të rëndësishme.
* Decision Trees në mënyrë implicite kryejnë përzgjedhjen e variantit dhe kërkojnë përpjekje relativisht të pakta nga përdoruesit për pregatitjen e të dhënave.
* Kërkohet pak pastrim i të dhënave krahasuar me disa teknika të tjera të modelimit. Nuk ndikohet nga outliers.
* Tipi i të dhënave nuk është një kufizim. Mund të trajtohen edhe variablat numerikë edhe kategorikë. Gjithashtu mund të trajtohen probleme me shumë rezultate.
* Metoda Jo-Parametrike: Pema e vendimit konsiderohet të jetë një metodë jo parametrike. Kjo do të thotë se pemët e vendimit nuk kanë supozime rreth shpërndarjes së hapësirës dhe strukturës së klasifikuesit.
* Marrëdhëniet jolineare midis parametrave nuk ndikojnë në performancën e pemës.

Ndwrsa pwr sa i pwrket disavantazheve:

* Pemët nuk merren me marrëdhëniet lineare. Çdo lidhje lineare midis një tipari të inputit dhe rezultatit duhet të përafrohet nga ndarjet, duke krijuar një funksion hap. Kjo nuk është aspak efikase, gjithashtu ndryshimet e lehta në tiparin e hyrjes mund të kenë një ndikim të madh në rezultatin e parashikuar, gjë që zakonisht nuk është e dëshirueshme.
* Pemët janë mjaft të paqëndrueshme - Disa ndryshime në dataset e trajnimit mund të krijojnë një pemë krejt tjetër. Kjo është për shkak se çdo ndarje varet nga ndarja e prindërve. Dhe nëse një veçori tjetër përzgjidhet si tipari i parë i ndarjes, struktura e tërë e pemës ndryshon. Ndryshimi i shpejtw i strukturws ndikon nw humbjen e besimit te modeli.

##### Algoritmi Support Vector Machine (Analytics Vidhya-Understanding Support Vector Machine Algorithm form examples 2017)

Support Vector Machine (SVM) wshtw njw algoritwm i kontrolluar i machine learning i cili mund tw pwrdoret pwr probleme tw klasifikimit dhe njwkohwsisht dhe sfida tw regresionit, por me pwrdorim mw tw pwrhapur nw fushwn e klasifikimit. Nw algoritmin SVM, cdo njwsi tw dhwnash paraqitet grafikisht si njw pikw nw hapwsirwn n-dimensionale (ku n wshtw numri i tipareve), duke pwrdorur vlerwn e cdo tipari si vlerwn e koordinatws specifike. Pas kwsaj, performojmw klasifikimin duke gjetur hiper-planin qw diferencon mire klasat.



Support Vectors (Vektorwt Mbwshtettws) janw thjesht koordinatat e njw observimi tw caktuar. Klasifikuesi SVM wshtw njw barrierw e ndan nw mwnyrwn mw optimale klasat.

Pwr sa i pwrket identifikimit tw hiperplanit tw duhur kemi disa skenarw tw ndryshwm :

1. Kemi tre hiperplane A, B dhe C



Pwr tw identifikuar hiperplanin mw tw mirw pwr ndarjen midis yjeve dhe rradhwve duhet tw kujtojmw njw rregull tw thjeshtw : “Zgjidhet ai hiperplan i cili segregon mw mirw dy klasat”, nw skenarin e mwsipwrm wshtw e qartw qw hiperplani B e kryen kwtw detyrw nw mwnyrwn mw optimale.

1. Nw kwtw skenar, tw tre hiperplanet A, B dhe C performojnw njw ndarje tw mirw tw hiperplaneve. Kwtu, pwr tw identifikuar hiperplanin e duhur pwrdorim distancwn ndwrmjet pikws mw tw afwrt tw tw dhenave dhe hiperplanit. Kjo distancw quhet Margin.



Sic shohim nga figura, Margin pwr hiperplanin C wshtw mw e madhe se pwr hiperplanet e tjera. Rrjedhimisht, hiperplani i duhur wshtw C.

1. Skenari i tretw paraqitet nw figurwn e mwposhtme :



Nw kwtw rast, B ka marxhinwn mw tw lartw, por A bwn njw klasifikimi mw tw saktw tw klasave. Algoritmi SVM, i jep prioritet segregimit tw klasave mbi maksimizimit tw marxhinws. Prandaj, hiperplani i duhur wshtw A.

1. Nw skenarin e rradhws wshtw e pamundur tw bwhet njw klasifikim i mirw duke pwrdorur njw vijw tw drejtw, sepse nw njwrwn prej klasave kemi njw outlier, pra njw pjestar tw klasws tjetwr. Por sic e kemi pwrmendur edhe mw lart, algoritmi SVM ka avantazhin qw injoron “outliers”, kwshtu qw mund tw realizojmw pwrswri njw ndarje tw tillw.



1. Dhe nw rastin e fundit, grafiku duket kwshtu :



Nw kwtw rast, SVM prezanton njw tipar shtesw z = x2 + y2, dhe harton edhe njwherw pikat e tw dhwnave sipas boshteve x dhe z :



Pas kwtij transformimi wshtw e thjeshtw tw kemi njw hiperplan linear ndwrmjet tw dhwnave. SVM ka njw teknikw tw quajtur trick-u i krenelit i cili shmang nevojwn e shtimit tw kwtij tipari manualisht pwr tw pasur njw hiperplan. Algoritmi SVM, bwn transformime shumw komplekse tw tw dhwnave dhe mw pas gjen procesin mw tw mirw pwr ndarjen e tw dhwnave nw varwsi tw emwrtimeve apo rezultateve tw definuara.

##### 4.3.1 Avantazhet dhe disavantazhet e SVM

Avantazhet qw sjell pwrdorimi i algoritmit SVM :

* Punon shumw mirw me njw marxhinw tw qartw tw ndarjes
* Wshtw njw algoritwm shumw efektiv nw hapwsirat shumwdimensionale
* Ka performancw tw mirw edhe nw rastet kur numri i dimensioneve wshtw mw i madh se numri i njwsive
* Pwrdor njw subset tw pikave tw trajnimit nw funksionin e vendimit dhe wshtw eficent nw lidhje me memorjen.

Pwr sa i pwrket disavantazheve tw asociuara me kwtw algoritwm :

* Nuk performon mirw kur dataset-i qw kemi nw dorw wshtw i madh sepse koha e trajnimit qw kwrkon wshtw mw e lartw
* Algoritmi nuk wshtw eficent edhe nw rastet kur tw dhwnat kanw shumw zhurma psh kur klasat e targetuara kanw mbivendosje me njwra-tjetrwn.
* Algoritmi SVM gjithashtu nuk jep vlerwsime probabilitare, ato llogariten duke pwrdorut “cross-validation”.

# Kapitulli 5

Nw kwtw kapitull do tw fillojmw tw studiojmw njw rast praktik, pwr tw aplikuar mw pas algoritmat pwr tw cilwt folwm mw sipwr.

##### 5.1 Targetimi dhe pwrzgjedhja e dataset-it qw do tw marrim nw shqyrtim

Fillimisht, gjatw punimit pwr realizimin e kwsaj pune u punua pwr gjetjen e njw objektivi apo synimi pwrfundimtar tw punws, me qwllim pwrzgjedhjen e dataset-it ku do tw aplikohen algoritmat e machine learning. Kwshtu u vendos pwr pwrzgjedhjen e njw set-i tw dhwnash mbi vlerwsimet e hoteleve, tw cilat do tw lejonin punimin pwr parashikimin e vlerwsimit.

Seti fillestar i tw dhwnave, u gjet nga kaggle.com, i cili wshtw komuniteti mw i madh nw botw nw lidhjw mw Data Science dhe pwrfshin mjete tw fuqishme pwr mbwshtetjen e studimeve nw fushwn e Data Science. Data set-i i marrw pwrfshinte rreth peswqind mijw rekorde mbi vlerwsimet e hoteleve. Njw sasi e tillw u konsiderua tepwr voluminoze, prandaj u reduktua nw dy mijw rekorde. Gjithashtu, u modifikua set-i duke pwrzgjedhur vetwm ato fusha tw dhwnash qw u konsideruan tw dobishme pwr qwllimin e studimit. Njwkohwsisht, set-i i tw dhwnave u pasurua me rreth 100 rekorde mbi hotele shqiptare, tw dhwnat pwr tw cilwt u morwn nga booking.al. Njw tjetwr ndryshim qw u performua mbi set-in e tw dhwnave ishte ndryshimi i vlerwsimit nga njw vlerwsim numerik 0-10, nw njw vlerwsim kategorik i specifikuar : “bad”, “good”, “excellent”. Nw pwrfundim, seti i targetuar i tw dhwnave kishte pamjen e mwposhtme :

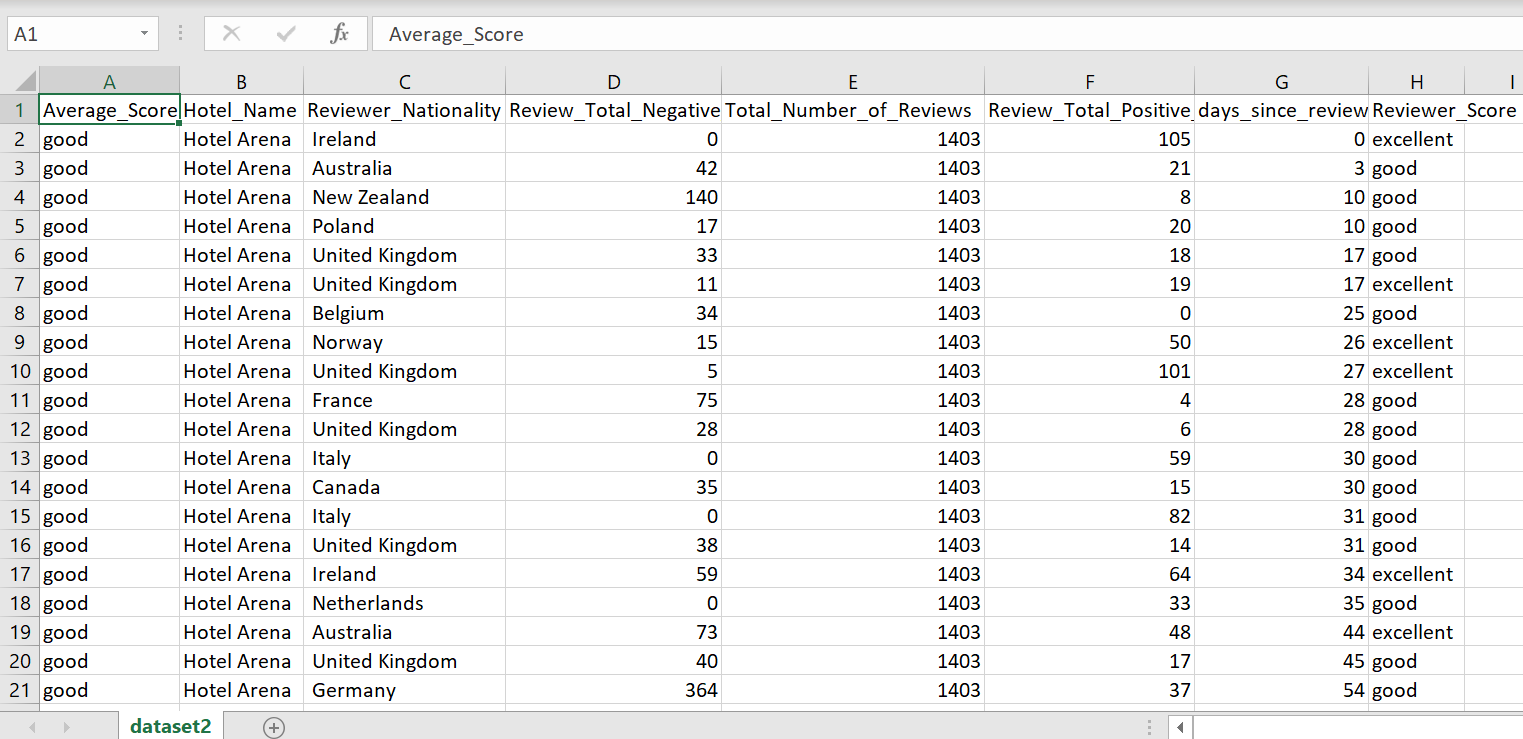
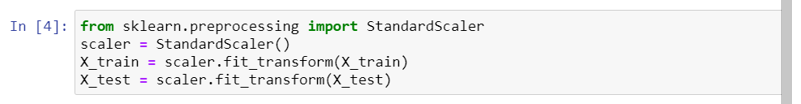


Fig 5.1 Pamja pwrfundimtare e data set-it me tw cilin do tw punohet gjatw studimit

##### 5.2 Paraprocesimi i tw dhwnave

Pwr realizimin e hapave mw tej, u pwrdor gjuha Python, si njw nga gjuhwt mw tw njohura nw botwn e Machine Learning, duke ofruar lehtwsi nw lidhje me funksionet apo libraritw e gatshme pwr aplikimin e algoritmave. Paraprocesimi i tw dhwnave mbetet njw nga hapat mw thelbwsorw nw mbarwvajtjen e ndwrtimit tw modelit. Ky hap nwnkupton ndwrhyrjen e mwtejshme nw setin e tw dhwnave pwr ta pregatitur atw pwr modelim. Njw nga paketat mw tw rwndwsishme qw wshtw pwrdorur nw kwtw hap wshtw sklearn dhe vecanwrisht pwr realizimin e paraprocesimit sklearn.preprocessing.

Veprimi i parw qw do tw ndwrmarrim wshtw standartizimi.Standartizimi nwnkupton transformimi i datasetit pwr tw pasur njw shpwrndarje normale, Gausiane, pasi njw shpwrndarje e ndryshme mund tw pwrkeqwsojw performancwn e algoritmit. Por gjithsesi, gjithcka wshtw nw varwsi tw algoritmit, sepse sic e pwrmendwm edhe mw lart, disa algoritma si SVM nuk kwrkojnw nw mwnyrw eksplicite realizimin e kwtij hapi. Pwr tw arritur zbatimin e standartizimit, gjuha Python na vjen nw ndihmw pwrmes klasws StandardScaler, klasw kjo e cila implementon njw API transformues pwr tw llogaritur mesataren dhe devijimin standart tw njw seti trajnimi. Mw poshtw jepet kodi i pwrdorimit tw kwtyre funksioneve :

  
Fig 5.2 Pwrdorimi i StandardScaler

Mw tej, vazhdojmw me njw tjetwr element qw i pwrket paraprocesimit tw tw dhwnave, qw wshtw menaxhimi i vlerave qw mungojnw. Pwr arsye tw ndryshme, sete te caktuara tw dhwnash pwrmbajnw vlera qw mungojnw (missing values). Kwto vlera ndikojnw negativisht nwperformancwn e algoritmit, si dhe nw rastin tonw duke qenw se do tw pwrdorim scikit-learn, kjo paketw pwrmban algoritma tw cilwt supozojnw se tw gjitha vlerat janw numerike dhe pwrmbajnw kuptim duke mos pranuar vlera Nan.

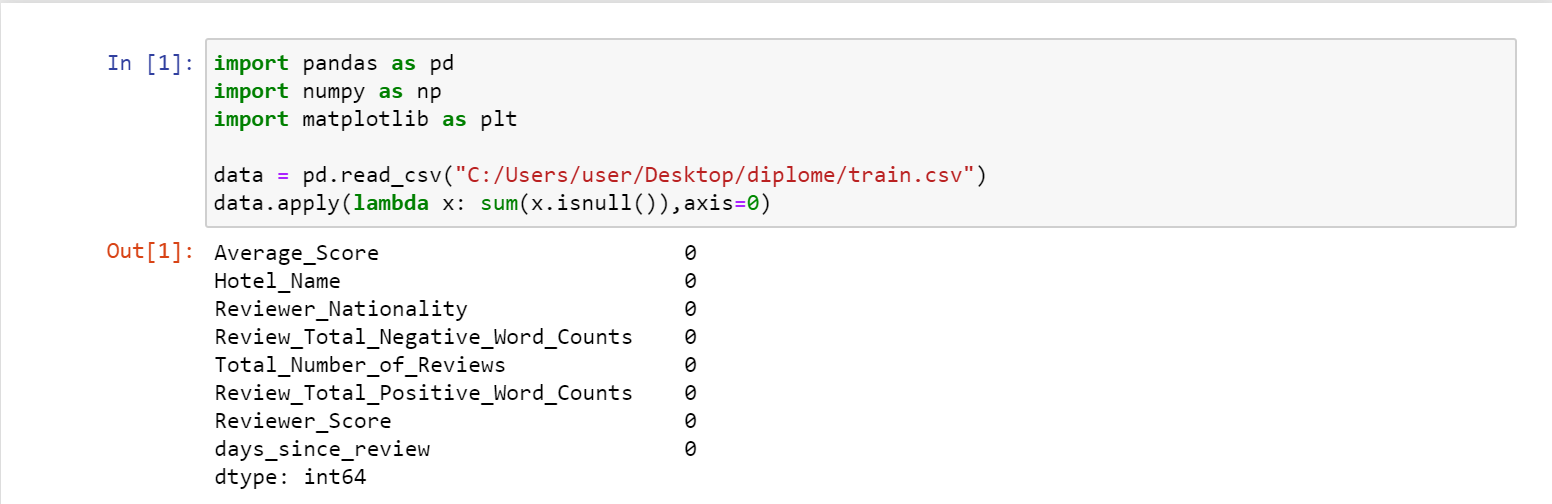
Pwrmes funksionit apply, identifikojme numrin e vlerave qw mungojnw nw mwnyrw qw tw kryejmw zwvwndwsimin e tyre.

Fig 5.3 Identifikimi i vlerave qw mugojnw

Por, nw rastin tonw sic e shohim, nuk kemi asnjw vlere null nw asnjerwn nga prej kategorive, prandaj edhe nuk ka nevojw qw tw pwrdorim njw funksion tjetwr pwr mbushjen e kwtyre vlerave.

Dhe sw fundmi, sic e pwrmendwm mw lart paketa scikit-learn pranon vetwm vlera numerike, kwshtu qw domosdoshmwrisht do tw na duhet tw kryejmw enkodimin e tw dhwnave. Pwr tw kryer enkodimin, pwrdorim klaswn labelEncoder si mw poshtw :

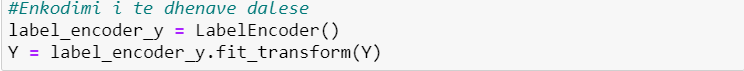


Fig 5.4 Enkodimi i tw dhwnave

Dy figurat nw vazhdim tregojnw formatin e tw dhwnave para dhe pas enkodimit :

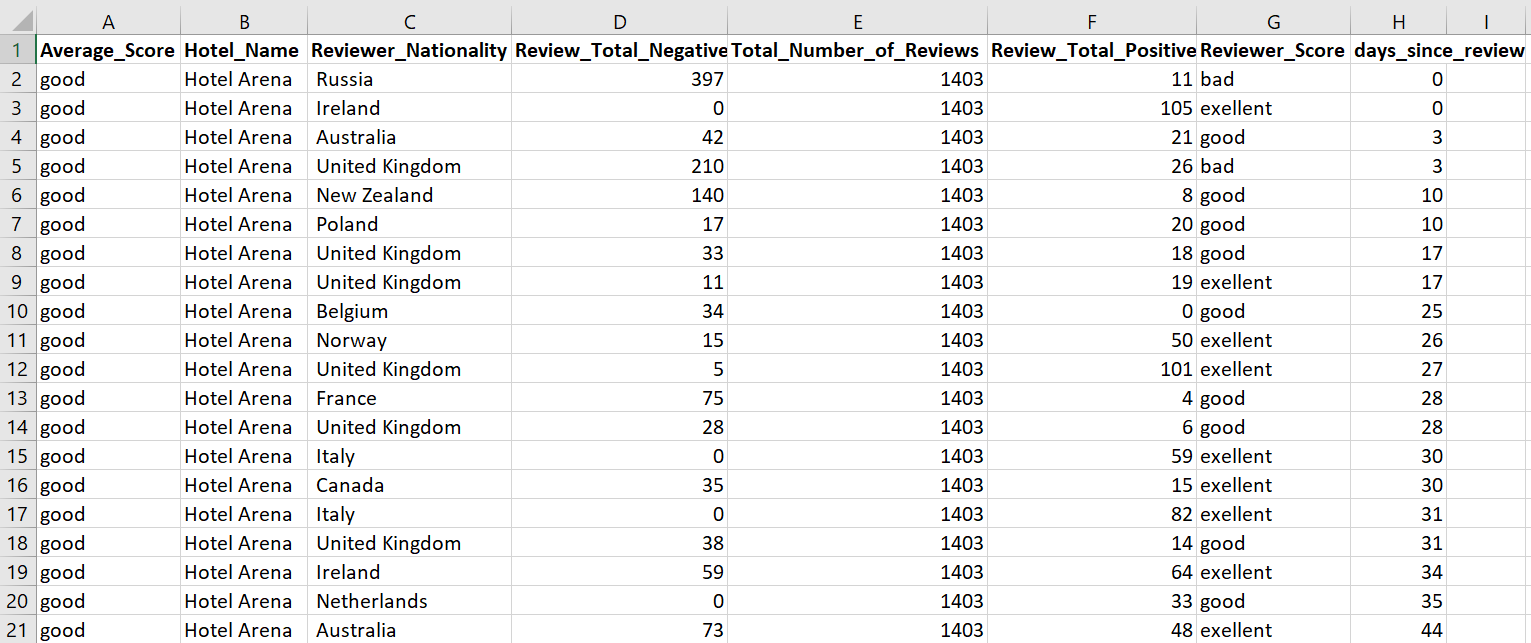


Fig 5.5 Tw dhwnat para enkodimit

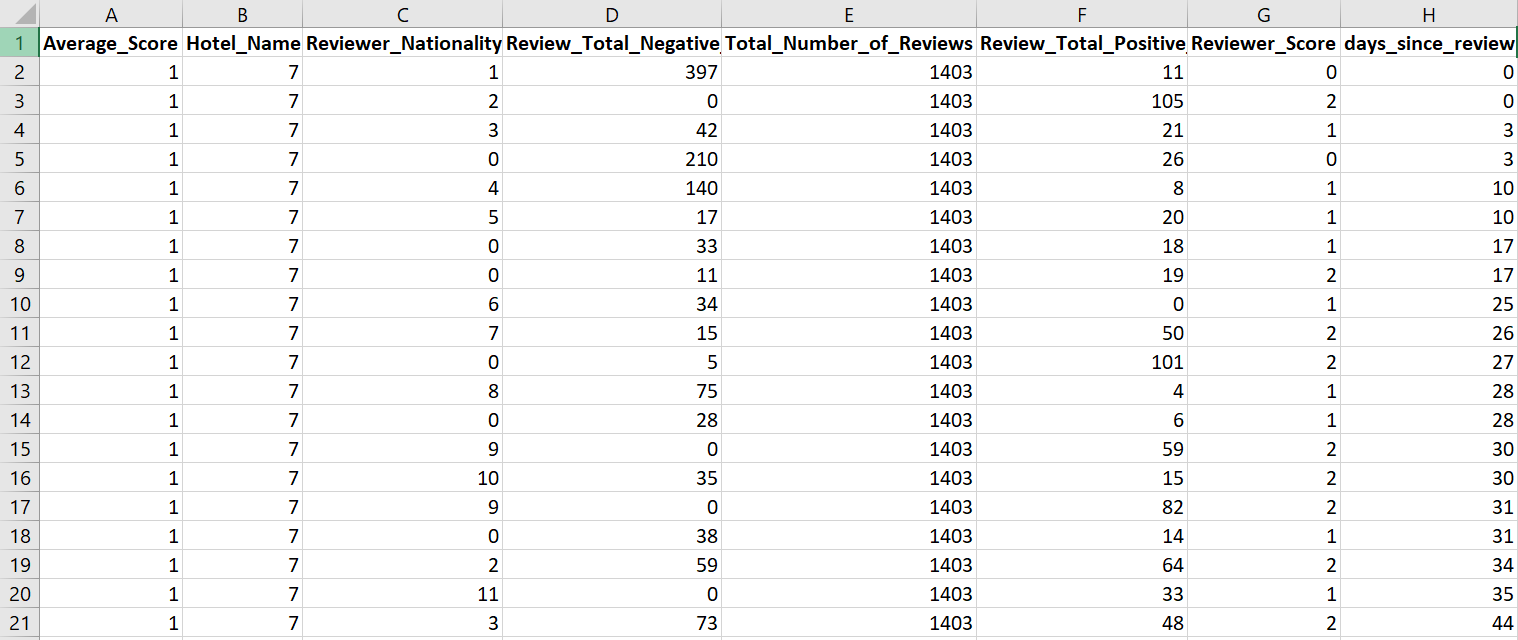
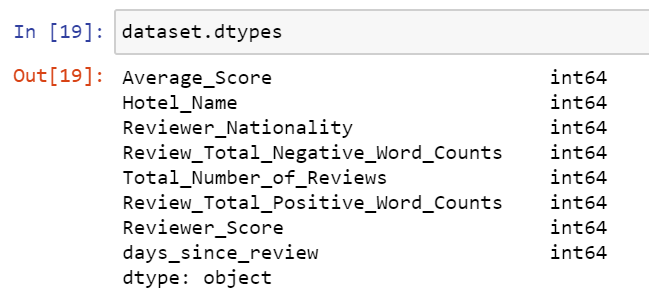


Fig 5.6 Tw dhwnat pas enkodimit

Nw kwtw pikw, vlen edhe tw kontrollojmw tipin e tw dhwnave pas kryerjes sw enkodimit, pwr tw kontrolluar nwse procesi i enkodimit ka shkuar si duhet dhe pwr tw pwrjashtuar rastin e ndodhjes sw ndonjw problemi qw rezulton nw fusha jonumerike :

  
Fig 5.7 Tipi i tw dhwnave pas enkodimit

Nga rezultati shohim qw tw gjitha fushat e tw dhwnave janw tw tipit Integer, pra procesi i enkodimi wshtw realizuar nw mwnyrwn e duhur pwr cdo fushw.

##### 5.3 Analizimi i tw dhwnave

Analizimi i tw dhwnave wshtw njw hap i domosdoshwp pwr tu njohur me dataset-in. Pwr tw kuptuar jo vetwm pwrciptazi, por edhe mw thellwsisht se me cfarw lloj tw dhwnash kemi tw bwjmw, cfarw marrwdhwnie kanw ato me njwra tjetrwn dhe kryesisht me fushwn qw do tw parashikohet, nw rastin tonw Reviewer Score.

##### 5.3.1 Tw dhwna paraprake mbi setin e tw dhwnave dhe ambientin e punws

Dataset-i pwrfundimtar pwrbwhet nga 2000 rekorde, prej tw cilwve 1900 janw rekorde tw hoteleve tw huaj dhe 100 tw hoteleve shqiptare. Ajo qw pwrmes kwtij studimi synohet, wshtw parashikimi i rezultatit tw vlerwsimit pwr hotelet shqiptare, duke pwrdorur si set trajnues atw tw hoteleve tw huaja.

Fushat e tw dhwnave janw si mw poshtw :

1. Average Score – e cila jep vlerwsimin mesatar pwr cdo hotel
2. Hotel Name – qw pwrcakton emrin e hotelit pwr tw cilin po flasim
3. Reviewer Nationality – e cila ka tw bwjw me vendin nga vjen vlerwsimi
4. Review Total Negative Word Counts – qw shwrben si njw numwrues i fjalwve tw komentit negativ tw lwnw nga vlerwsuesi
5. Total Number of Reviews – qw pwrcakton numrin total tw vlerwsimeve qw njw hotel i caktuar ka deri nw momentin e marrjes sw tw dhwnave
6. Review Total Postive Word Counts – qw shwrben si njw numwrues i fjalwve tw komentit negativ tw lwnw nga vlerwsuesi
7. Reviewer Score – wshtw fusha e pikwve tw vlerwsimit, njwkohwsisht edhe fusha qw do tw mundohemi tw parashikojmw
8. Days since review – e cila jep numrin e ditwve qw kanw kaluar nga dhwnia e review

MBASE ENKODIMI DUHET FUTUR KETUUU !!!

Ambienti nw tw cilin wshtw punuar me tw dhwnat si dhe ku do tw testohen algoritmat wshtw :

* Laptop HP EliteBook Folio
* Processor Intel(R) Core i7-4600U CPU @2.10GHz 2.7 GHz
* SO : Windows 10 Pro 64-bit operating system, x64-based processor

Pwr sa i pwrket gjuhws dhe ambientit tw zhvillimit :

* Python versioni 3.7.6
* Mjedisi : Jupyter Notebook 6.0.3 (wshtw njw ambient zhvillimi web-based, interaktiv i specifikuar pwr analizw tw dhwnash)

##### 5.3.2 Analizimi i variablave tw datasetit

Qwllimi i kwtij studimi wshtw trajnimi i modelit pwrmes rekordeve tw hoteleve tw huaja me synim aplikimin e algoritmit pwr parashikimin e fushws Reviewer Score pwr vlerwsimet e Shqipwrisw. Kwtu, do tw merremi me studimin e tw gjitha fushave tw setit tw tw dhwnave pwr tw kuptuar mw mirw peshwn e tyre nw pwrcaktimin e vlerwsimit.

Pwr arritjen e kwtij qwllimi, janw ndwrtuar histograma tw ndryshme duke pwrdorur libraritw matplotlib dhe seaborn tw gjuhws Python.

Nw figurwn e mwposhtme paraqitet histograma pwr fushwn Reviewer Nationality. Sic mund ta shohim qartwsisht nga figura, kemi njw numwr tw lartw kombwsish, pra reviews qw vijne nga vende tw ndryshme nga e gjithw bota, por me njw dominim tw dukshwm tw shtetit tw Britania e Madhe. Pra, kemi rreth 600 rekorde nga Britania e Madhe duke dhwnw dhe numrin mw tw madh tw vlerwsimeve, mw pas me njw numwr tw konsiderueshwm janw edhe shtete si Irlanda, Australia, Shtetet e Bashkuara tw Amerikws dhe Arabia Saudite.



Fig. 5.8 Histograma pwr Reviewer Nationality

Nw figurwn 5.9 shohim shpwrndarjen pwr fushwn Average Score, qw paraqet cili wshtw mesatarisht vlerwsimi pwr secilin hotel. Nw rastin tonw, shohim qw nuk kemi asnjw hotel me vlerwsimin bad, kemi njw pjesw tw hoteleve me vlerwsimin e pwrgjithshwm good, dhe dominojmw ata me vlerwsimin excellent, nw raportin : rreth 140 me 400. Njw nga vwrejtjet qw mund tw bwjmw nw kwtw moment wshtw mungesa e hoteleve me vlerwsimin mesatar bad, si dhe disbalanca ndwrmjet vlerwsimeve excellent qw dominojnw pwrkundrejt atyre good qw jane nw sasi shumw mw tw vogwl. Kjo mund tw na sjellw probleme mw vonw gjatw aplikimit tw algormitmit.

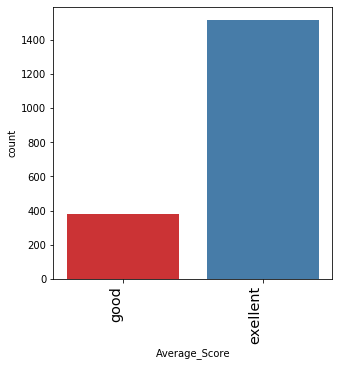
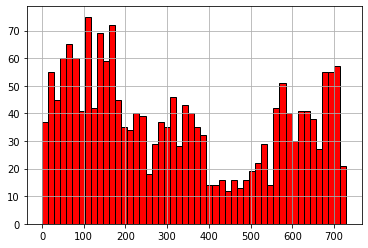


Fig. 5.9 Histograma pwr Average Score

Vazhdojmw mw tej me histogramwn pwr fushwn Days Since Reviews. Nga histograma mund tw shohim se kemi njw shpwrndarje nga 0 ne 700 ditw qw nga dhwnia e vlerwsimit, ku vlerat maksimale (mbi 80 reviews) arrihen afwr 100 - 200 ditw mw parw.

  
Fig. 5.10 Histograma pwr Days Since Review

Dhe sw fundmi, nw figurwn 5.11 tregohet histograma pwr fushwn Reviewer Score, nga histograma mund tw shohim se kemi njw numwr shumw tw vogwl tw vlerwsimeve bad, me dominim tw dukshwm tw vlerwsimeve nwn kategorinw excellent edhe kundrejt atyre me vlerwsimin good.

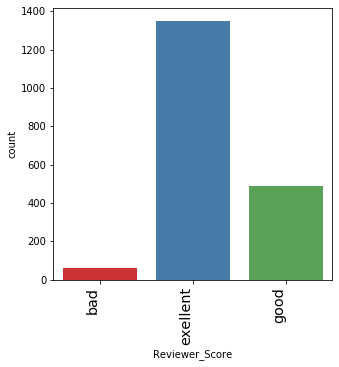


Fig. 5.11 Histograma pwr Reviewer Score

Njw disbalancw e tillw midis vlerwsimeve do tw sillte gabime tw mwdha nw parashikim, prandaj u vendos qw tw hiqeshin nga dataset ato vlerwsime qw binin nwn kategorinw bad, si dhe tw pasurohej seti i tw dhwnave me mw shumw hoteme me vlerwsim tw pwrgjithshwm si dhe vlerwsime individuale good.

Duke qwnw se kemi njw sasi mjaft tw vogwl tw kwtyre kategorive nw dataset rekordesh , ndryshimi i bwrw nuk i afekton nw masw tw madhe histogramat pwr days since reviews apo Reviewer Nationality, tw vetmet histograma qw pwsojnw ndryshim janw ajo pwr Reviewer Score dhe histograma pwr Average Score, tw cilat shndwrrohen si mw poshtw :





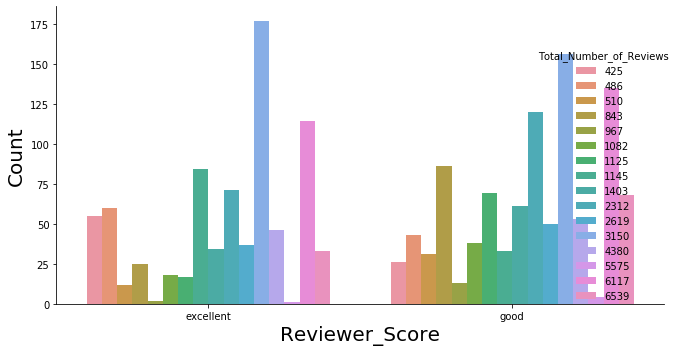
Fig. 5.12 Histogramat pas ndryshimeve

BEJ NJE KOMENT KTU PER KTO SE PERTOJ!

##### 5.3.3 Analizimi i fushws sw rezultatit

Njw nga hapat mw tw rwndwsishwm nw analizwn e tw dhwnave wshtw analiza e lidhjeve tw variablit qw do tw parashikohet, me variablat e tjerw nw mwnyrw qw tw vlerwsohen anomali apo pritshmwri tw mundshmw. Gjatw kwtij procesi, zbulohen patterna tw ndryshwm twk tw dhwnat dhe ky proces mund tw quhet si ekstaktimi i njohurive nga dataseti.

Fillimisht do tw diskutojmw lidhjen ndwrmjet fushws Reviewer Score dhe fushws Total Number of Reviews:

  
Fig. 5.13 Lidhja ndwrmjet fushave Reviewer Score dhe Total Number of Reviews

Kwtu mund tw shohim se kemi njw shpwrndarje tw ngjashme nw tw dyja rastet, si pwr kategorinw good edhe pwr atw bad, me njw dallim tw vogwl qw vlerat e numrit tw vlerwsimeve totale janw pak mw tw larta nw rastin e vlerwsimit excellent.

Kalojmw nw vlerwsimin e lidhjes ndwrmjet Reviewer Score dhe fushws Average Score. Histograma paraqitet si mw poshtw :

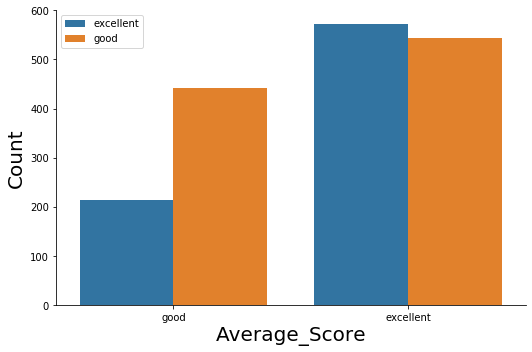


Fig. 5.14 Lidhja ndwrmjet fushave Reviewer Score dhe Average Score

Nw kwtw rast mund tw shohim njw pattern nw lidhje me tw dhwnat. Duket nga histograma se kur vlerwsimi mesatar i hotelit wshtw good, kemi njw dominim tw vlerwsimeve individuale me tw njejtwn kategori. Njwlloj edhe pwr rastin e kategorisw excellent.

Lidhja e rradhws qw do tw diskutojmw wshtw ajo ndwrmjet Reviewer Score dhe dy fushave tw ngjashme, Review Top Negative Word Counts dhe Review Top Positive Word Counts, qw pwrfaqwsojnw pwrkatwsisht numrin e fjalwve pwr vlerwsimin negativ dhe pwr atw pozitiv.

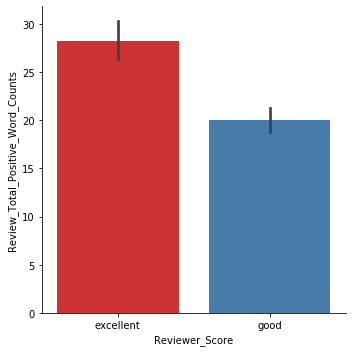
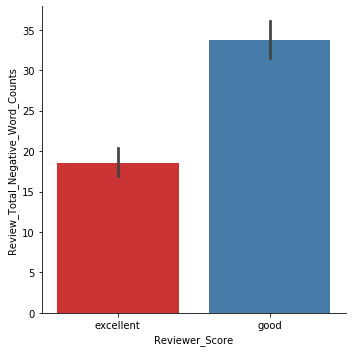
 

Fig. 5.15 Lidhja ndwrmjet fushws Reviewer Score me fushat Top Negatve Words dhe Top Positive Words

Nga histogramat edhe kwtu mund tw nxjerrim njw model mbi sjelljen e vlerwsuesve,kur kemi njw vlerwsim excellent, numri i fjalwve tw komentin pozitiv wshtw shume mw i madh se sa ai pwr komentin negativ. Dhe anasjelltas, kur kemi njw vlerwsim ne kategorinw good, numri i fjaleve negative wshtw mw imadh (megjithse diferenca nuk wshtw nw nw njwjtwn madhwsi si nw rastin e parw) se sa numri i fjalwve tw komentin pozitiv.

Dhe sw fundmi, do tw diskutojmw mbi lidhjen e fushws Reviewer Score me fushwn Days Since Review.

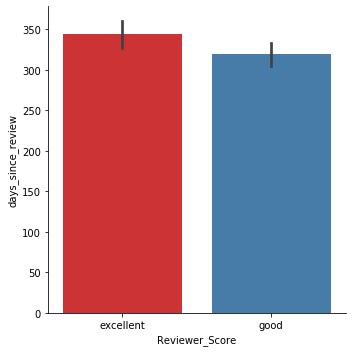


Fig. 5.16 Lidhja ndwrmjet Reviewer Score dhe Days Since Review

Nga histograma, wshtw e qartw se nuk kemi ndonjw ndryshim tw madh nw lidhje me numrin e ditwve qw kanw kaluar nga dhwnia e vlerwsimit pwrkundrejt llojit tw vlerwsimit. Prandaj, mund tw dalim nw pwrfundimin se kjo fushw nuk ka njw ndikim tw ndjeshwm nw parashikimin e rezultatit tw vlerwsimit.

Bazuar nw kwto tw dhwna, bwhet edhe trajnimi i modelit nw bazw tw algoritmave qw pwrmendwm mw parw.

KAPITULLI 6

Pasi u realizua zgjedhja e setit tw tw dhwnave, paraprocesimi e tw dhwnave dhe analiza e tyre, mbetwt tw trajnojmw modelin dhe tw realizojmw parashikimet. Algoritmat e pwrdorur, janw ata qw kemi pwrmendur nw Kapitullin e katwrt, pra :

* Logistic Regression
* Decision Tree
* Support Vector Machine

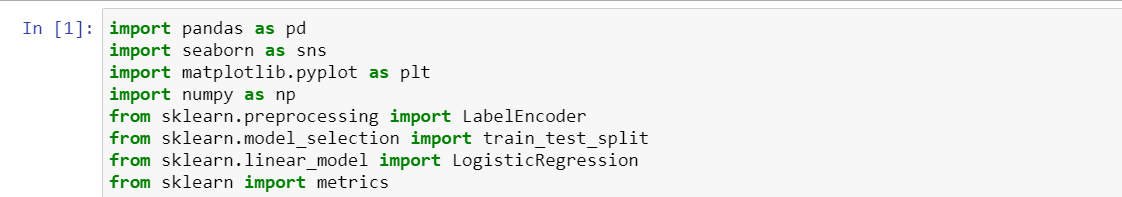
Pwr tw tre algoritmat, do tw krahasohet saktwsia e parashikimit dhe matrica e konfuzionit, duke bwrw mw pas edhe krahasimet pwr performancwn e pwrgjithshme tw algoritmave, pwr tw vlerwsuar cili wshtw mw i pwrshtatshmi pwr setin tonw tw tw dhwnave.

##### 6.1 Aplikimi i algoritmit Logistic Regression

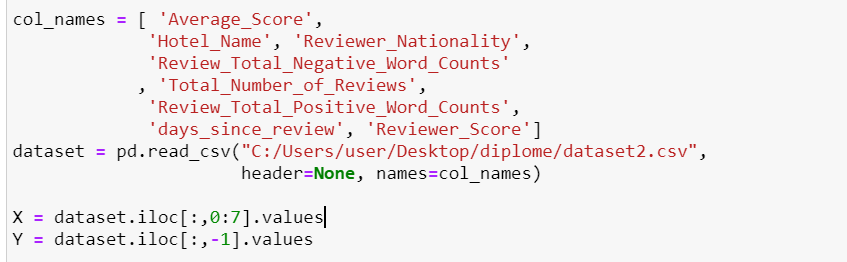
Regresioni logjik është një teknikë thelbësore e klasifikimit. Ajo i përket grupit të klasifikuesve linearë dhe ngjason me regresionin linear dhe polinomial. Procesi i regresionit logjik është i shpejtë dhe relativisht jo i komplikuar. Algoritmi bën pjesë në supervised learning dhe përdoret për të përshkruar dhe shpjeguar marrëdhëniet midis një variabli binar të varur me një ose më shumë ndryshore të pavarura.

##### 6.1.1 Trajnimi i modelit

Nw figurwn 6.1 shihen paketat e importuara gjatw trajnimit tw modelit :

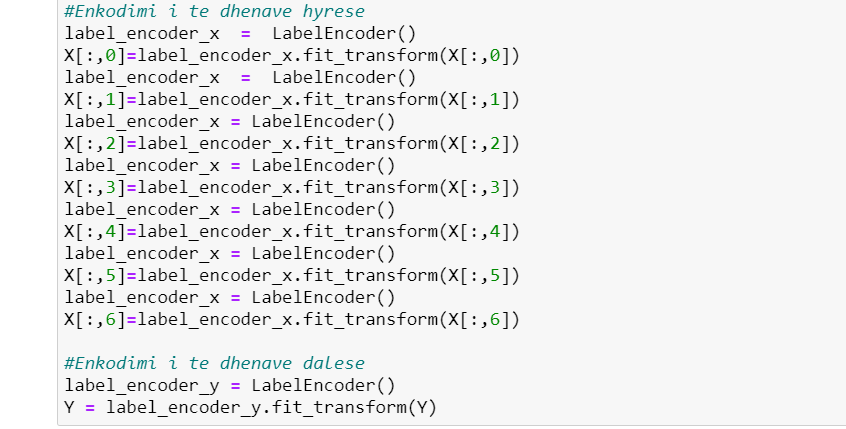
  
Fig. 6.1 Paketat e importuara

Mw poshtw jepet pjesa e leximit tw datasetit me anw tw Python si dhe pwrcaktimi i inputit dhe outputit. Pra do tw ndajmw tw dhwnat, fusha Reviewer Score do tw shwrbejw si output (shwnuar me Y), ndwrsa fushat e tjera do tw shwrbejnw si input i datasetit.

  
Fig. 6.2 Pwrcaktimi i inputit dhe outputit

Sic duket nga figura, fillimisht me anw tw variablit col\_names, bwjmw njw listw me emrat e fushave tw tw dhwnave. Tw dhwnat ndodhen nw njw file csv, dhe lexohen pwrmes librarisw Pandas, funksionit tw saj read\_csv, i cili bwn tw mundur edhe lidhjen me emwrtimet e kolonave. Rrjedhimisht, mw vonw mund t’i aksesojmw lehtwsisht ato mw vonw nw program.

Hapi i rradhws wshtw enkodimi i tw dhwnave. Enkodimi bwhet sipas kodit tw paraqitur nw figuren 6.3. Pwrmes kwtij kodi enkodojmw tw dhwnat pwr trajnimin e modelit, tw cilat do tw shwrbejnw si input dhe gjithashtu enkodojmw edhe tw dhwnat qw do tw shwrbejnw si rezultat.

  
Fig. 6.3 Enkodimi i tw dhwnave

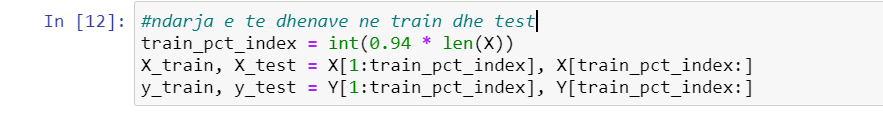
Enkodimi i tw dhwnkave wshtw bwrw pwr secilwn kolonw, nw mwnyrw qw tw dhwnat tw ishin tw pwrshtatshmw per t’iu servirur modelit. Pwr realizimin e enkodimit wshtw pwrdorur klasa LabelEcoder, pjesw e librarisw sklearn preprocessing dhe funksioni fit\_transform i cili kthen vlerat e enkoduara.

Pas enkodimit tw tw dhwnave, fillojmw me trajnimin e vwrtetw tw modelit pwrmes disa hapave :

1. Ndajmw tw dhwnat nw :

* Input trajnues (X\_train)
* Input testues (X\_test)
* Output trajnues (y\_train)
* Output testues (y\_test)

Pwr tw realizuar kwtw ndarje, sic shhet edhe nga kodi mw poshtw u pwrdor njw variabwl ndihmws train\_pct\_index, i cili shwrben si kufi ndarws. Duke qwnw se tw dhwnat qw do tw shwrbejnw si output ndodhen nw fund tw file-it, u pwrdor ky indeks pwrqindjeje. Nga llogaritjet, pas heqjes edhe tw rekordeve me vlerwsimin bad nga dataseti, u arrit nw pwrfundimin qw tw dhwnat hyrwse do tw pwrbenin 94% tw datasetit. Pra, nga kodi i mwposhtwm shohim se , X\_train merr nga rekordi i parw deri nw indeksin ndarws tw tw dhwnave hyrwse, ndwrsa X\_test merr pjeswn e mbetur, nga indeksi deri nw fund tw tw dhwnave hyrwse. i njwjti arsyetim vlen edhe pwr y\_train dhe y\_test.

  
Fig. 6.4 Ndarja e tw dhwnave nw tw dhwna trajnimi dhe testimi

1. Deklarimi i klasifikuesit

* Nw rastin tone klasifikuesi eshte Logistic Regression.

1. Pwrshtatja e modelit (Trajnimi i modelit)
2. Saktwsia e modelit
3. Parashikimi

Kwto hapa paraqiten nw kodin e mwposhtwm :

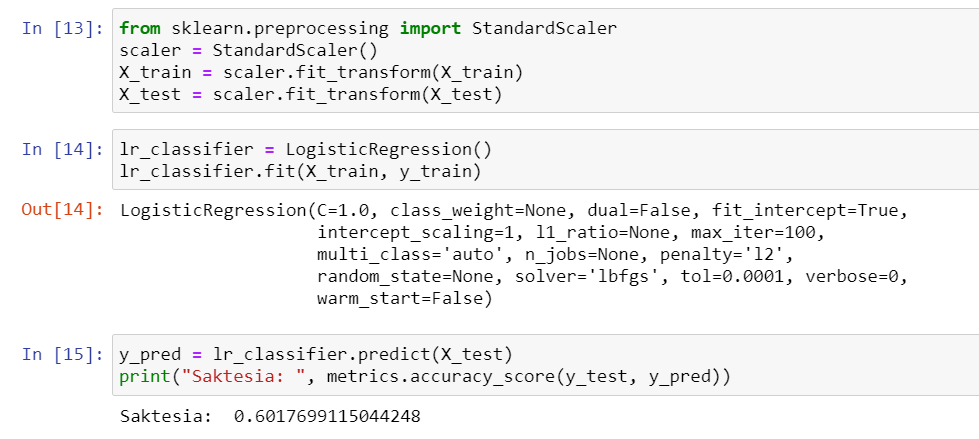
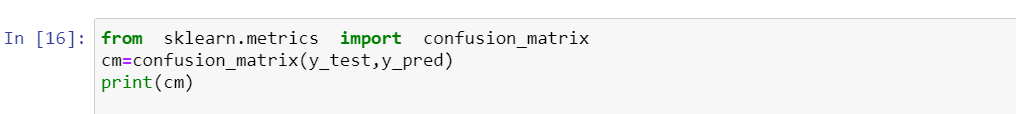


Fig. 6.5 Kodi pwr aplikimin e algoritmit dhe printimit tw saktwsisw

Sic shihet nga kodi, pas deklarimit tw klasifikuesit, kemi ndwrmarrw hapin e pwrshtatjes sw modelit. Ky hap, realizohet pwrmes funksionit fit, i cili merr si parametwr tw dhwnat e inputit dhe outputit trajnues. Mw pas kryejmw parashikimin pwrmes funksionit predict, qw merr si argument inputin testues. Nw fund printojmw saktesinw e modelit, duke pwrforur accuracy\_score nga klasa metrics.

Sic mund ta shohim, nw rastin e kwtij algoritmi arrijmw njw saktwsi pak mw shumw se 60%. Pra, algoritmi parashikon njw rezultat tw saktw nw mw shumw se 60% tw rasteve. Gjithsesi, pwr njw vizion mw tw qartw mbi parashikimin, mund tw printojmw matricwn e konfuzionit dhe raportin e klasifikimit.

##### 6.1.2 Matrica e konfuzionit

Fig. 6.6 Kodi pwr matricwn e konfuzionit

Nga ekzekutimi i kwtij kodi, pwrftojmw matricwn e konfuzionit. Matrica e konfuzionit wshtw njw tabelw qw shwrben pwr matjen e performancws sw algoritmit tw modelit qw do tw trajnohet. Zakonisht, matrica e konfuzionit pwrdoret nw algoritmat e supervised learning. Ajo wshtw njw tabelw e pwrbwrw nga dy rreshta dhe dy kolona, sic paraqitet mw poshtw :

|  |  |  |
| --- | --- | --- |
|  | **Parashikuar si jo** | **Parashikuar si po** |
| **Jo reale** | 30 | 33 |
| **Po reale** | 12 | 38 |

Tab. 5.2 Matrica e konfuzionit LR

Secili rresht i matricws sw konfuzionit pwrfaqwson tw dhwnat reale, tw sakta sipas mbledhjes sw tyre, ndwrsa kolonat pwrfaqwsojnw klasat qw i pwrkasin parashikimit. Sic mund tw shohim edhe nga emwrtimi i kolonave dhe rreshtave tw matricws, kemi 4 elementw qw shprehin 4 vlera tw ndryshme :

* Rezultati real negativ – >Parashikuar si negativ
* Rezultati real pozitiv – >Parashikuar si negativ
* Rezultati real pozitiv – >Parashikuar si pozitiv
* Rezultati real negativ – >Parashikuar si pozitiv

Matrica e konfuzionit, quhet e tillw sepse nw njw farw mwnyre shpreh edhe nivelin e konfuzionit me tw cilin wshtw pwrballur algoritmi. Konfuzioni zhvillohet nwpwrmjet kombinimeve tw rezultateve nga tw dhwnat reale dhe atyre qw ka parashikuar algoritmi.

Konkretisht nw kwtw rast, pas aplikimit tw Logistic Regression, janw arritur rezultatet e mwposhtme :

* True Negative -> Pwr 30 prej vlerwsimeve, wshtw parashikuar saktwsisht kategoria e parashikimit good. Pra, rezultati i parashikuar nga algoritmi pwrkon me vlerwsimin real tw hoteleve. Ne 30 raste wshtw parashikuar saktwsisht kategoria good.
* False Negative -> Pwr 12 prej vlerwsimeve, algoritmi ka parashikuar gabimisht kategorinw good ndwrkohw qw rezultati real ka qenw excellent. Pra, rezultati i parashikuar nga algoritmi nuk pwrkon me vlerwsimin aktual. Nw 12 raste wshtw parashikuar gabimisht kategoria good.
* False Positive -> Pwr 33 prej vlerwsimeve, nga algorimti wshtw parashikuar gabimisht vlerwsimi excellent ndwrkohw qw duhet tw ishte parashikuar vlerwsimi good. Nw 33 raste wshtw parashikuar gabimisht kategoria excellent.
* True Positive -> Pwr 38 prej vlerwsimeve, algoritmi ka parashikuar saktwsish njw vlerwsim excellent. Rezultati i parashikuar pwrkon me rezultatin real. Nw 78 raste wshtw parashikuar saktwsisht kategoria excellent.

Pwrfundimisht nga analiza kemi rezultatin e mwposhtwm :

* Numri i parashikimeve tw sakta :

True Negative + True Positive = 30 + 38 = 68 parashikime tw sakta.

* Numri i parashikimeve tw gabuara :

False Negative + False Positive = 12 + 33 = 45 parashikime tw gabuara.

##### 6.1.3 Raporti i klasifikimit

Njw tjetwr element i cili pwrdoret pwr tw matur nw njw parw mwnyre performancwn e algoritmit wshtw raporti qw vjen nga funksioni classification\_report i cili jep njw raport tw detajuar mbi performancwn e arritur.

Mw poshtw jepet kodi nw Python pwr shfaqjen e raportit tw klasifikimit qw ndodhet nw tabelwn 5.4. Edhe ky funksion wshtw pjesw e klasws metrics tw siguruar nga paketa sklearn dhe merr si argumenta rezultatet reale tw inputeve tw zgjedhura pwr testim di dhe rezultatet e parashikuara nga algoritmi.

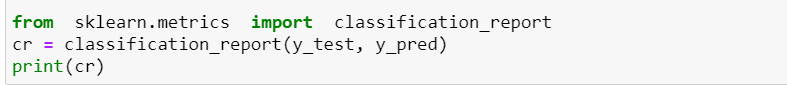
  
Fig. 6.7 Kodi pwr printimin e raportit tw klasifikimit

Tabela e raportit tw klasifikimit shfaqet si mw poshtw :

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-score | Support |
| 1 | 0.71 | 0.48 | 0.57 | 63 |
| 2 | 0.54 | 0.76 | 0.63 | 50 |
| Accurancy |  |  | 0.60 | 113 |
| Macro avg | 0.62 | 0.62 | 0.60 | 113 |
| Weighted  avg | 0.64 | 0.60 | 0.60 | 113 |

Tab. 5.4 Raporti pwr klasifikimin LR

Kolona support jep informacion mbi numrin e vlerwsimeve mbi tw cilat wshtw mbwshtetur algoritmi pwr tw parashikuar rezultatet e parashikuara.

Kolona Precision pwrfaqwson numrin e rezultateve good tw parshikuara qw janw realisht good si dhe numrin e rezultateve excellent tw parashikuara qw janw realisht excellent. Pwr rezultatin good kemi 71% tw pwrgjigjeve tw sakta dhe pwr rezultatin excellent 54%.

Formula për llogaritjen e precisionit:

ose

Kolona Recall (e njohur ndryshe edhe si sensitiviteti i algoritmit) përfaqëson numrin e eventeve pozitive (negative) të cilat u parashikuan si duhet sipas formulës:

ose

Kolona F1-score jep mesataren harmonike të precision dhe recall. Nëse bëjmë llogaritjet në varësi të vlerave që morëm për TP, FP,TN dhe FN nga confussion matrix shohim se do të arrijmë në të njëjtat vlera.

##### 6.1 Aplikimi i algoritmit Decision Tree

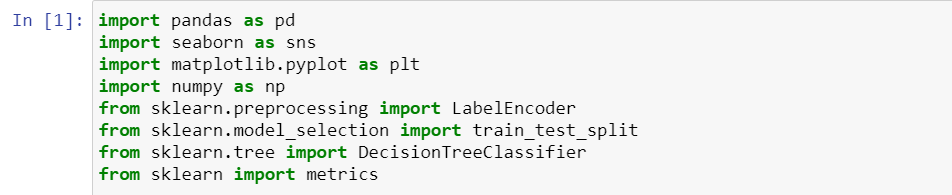
Një “decision tree” është një strukturë e ngjashme me një pemë ku cdo degë përfaqëson një rregull vendimi ndërsa cdo gjethe përfaqëson rezultatin e vendimit. Nyja rrënjë e pemës është nyja e sipërme e cila ndan vendimet në bazë të një atributi vlerë. Këto pemë janë të thjeshta për tu kuptuar dhe interpretuar sepse imitojnë mendimin njerëzor.



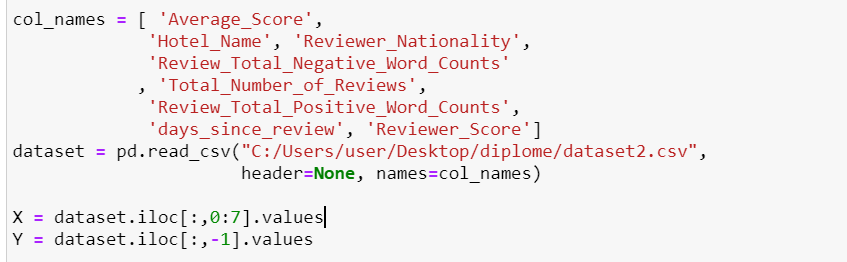
Fig 6.8 Skema e një decision tree

##### 6.1.1 Trajnimi i modelit

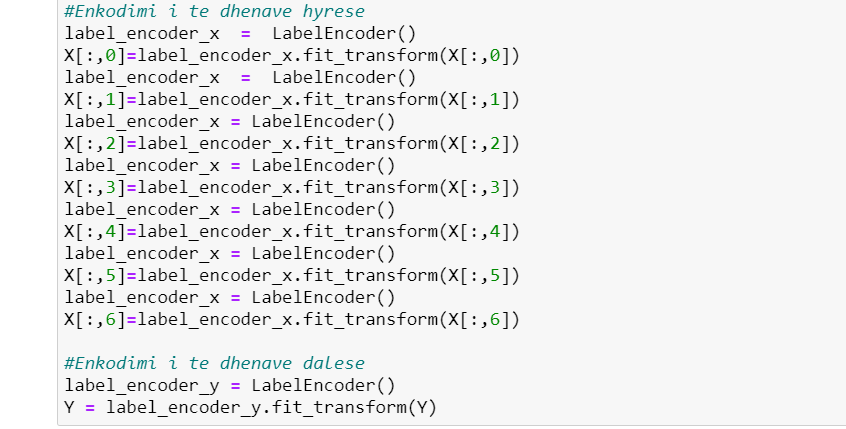
Nw figurwn 6.1 shihen paketat e importuara gjatw trajnimit tw modelit qw janw tw njwjta me ato tw Logistic regression me perjashtimin se nw vend tw logistic regression importojmw DecisionTreeClassifier :

  
Fig. 6.9 Paketat e importuara

Mw poshtw jepet pjesa e leximit tw datasetit me anw tw Python si dhe pwrcaktimi i inputit dhe outputit, njwlloj si nw algoritmin e mwparshwm :

  
Fig. 6.10 Pwrcaktimi i inputit dhe outputit

Enkodojmw tw dhwnat. Enkodimi bwhet sipas kodit tw paraqitur nw figuren 6.11. Pwrmes kwtij kodi enkodojmw tw dhwnat pwr trajnimin e modelit, tw cilat do tw shwrbejnw si input dhe gjithashtu enkodojmw edhe tw dhwnat qw do tw shwrbejnw si rezultat.

  
Fig. 6.11 Enkodimi i tw dhwnave

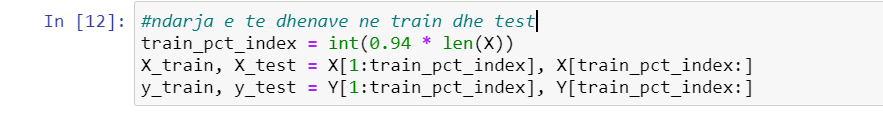
Enkodimi i tw dhwnave wshtw bwrw nw tw njwjtwn mwnyrw si nw algoritmin Logistic Regression.

Pas enkodimit tw tw dhwnave, fillojmw me trajnimin e modelit pwrmes hapave :

1. Ndajmw tw dhwnat nw :

* Input trajnues (X\_train)
* Input testues (X\_test)
* Output trajnues (y\_train)
* Output testues (y\_test)

Pwrswri edhe nw ndarjen e tw dhwnave kemi vepruar nw tw njwjtwn mwnyrw si nw algoritmin e mwparshwm.

  
Fig. 6.12 Ndarja e tw dhwnave nw tw dhwna trajnimi dhe testimi

1. Deklarimi i klasifikuesit

* Nw rastin tone klasifikuesi wshtw Decision Tree

1. Pwrshtatja e modelit (Trajnimi i modelit)
2. Saktwsia e modelit
3. Parashikimi

Kwto hapa paraqiten nw kodin e mwposhtwm :

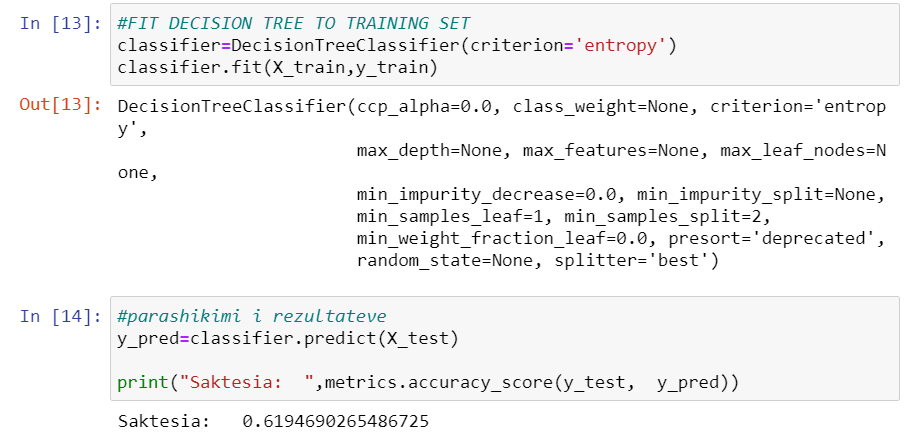
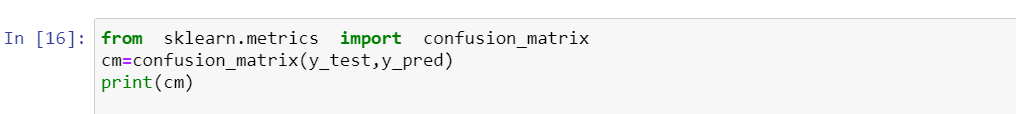


Fig. 6.13 Kodi pwr aplikimin e algoritmit dhe printimit tw saktwsisw

Sic shihet nga kodi, deklarojmw klasifikuesin dhe pwrshtasim modelin pwrmes funksionit fit, i cili merr si parametwr tw dhwnat e inputit dhe outputit trajnues. Mw pas kryejmw parashikimin pwrmes funksionit predict, qw merr si argument inputin testues. Nw fund printojmw saktesinw e modelit, duke pwrforur accuracy\_score nga klasa metrics.

Sic mund ta shohim, nw rastin e kwtij algoritmi arrijmw njw saktwsi thuajse 62%. Pra, algoritmi parashikon njw rezultat tw saktw nw 62% tw rasteve. Gjithsesi, pwr njw vizion mw tw qartw mbi parashikimin, printojmw matricwn e konfuzionit dhe raportin e klasifikimit.

##### 6.1.2 Matrica e konfuzionit

Fig. 6.14 Kodi pwr matricwn e konfuzionit

Nga ekzekutimi i kwtij kodi, pwrftojmw matricwn e konfuzionit. Matrica e konfuzionit wshtw njw tabelw qw shwrben pwr matjen e performancws sw algoritmit tw modelit qw do tw trajnohet. Zakonisht, matrica e konfuzionit pwrdoret nw algoritmat e supervised learning. Ajo wshtw njw tabelw e pwrbwrw nga dy rreshta dhe dy kolona, sic paraqitet mw poshtw :

|  |  |  |
| --- | --- | --- |
|  | **Parashikuar si jo** | **Parashikuar si po** |
| **Jo reale** | 38 | 25 |
| **Po reale** | 18 | 32 |

Tab. 5.2 Matrica e konfuzionit DT

Analizojmw matricwn e konfuzionit, nw kwtw rast, pas aplikimit tw Decision Tree, janw arritur rezultatet e mwposhtme :

* True Negative -> Pwr 38 prej vlerwsimeve, wshtw parashikuar saktwsisht kategoria e parashikimit good. Pra, rezultati i parashikuar nga algoritmi pwrkon me vlerwsimin real tw hoteleve. Ne 38 raste wshtw parashikuar saktwsisht kategoria good.
* False Negative -> Pwr 18 prej vlerwsimeve, algoritmi ka parashikuar gabimisht kategorinw good ndwrkohw qw rezultati real ka qenw excellent. Pra, rezultati i parashikuar nga algoritmi nuk pwrkon me vlerwsimin aktual. Nw 18 raste wshtw parashikuar gabimisht kategoria good.
* False Positive -> Pwr 25 prej vlerwsimeve, nga algoritmi wshtw parashikuar gabimisht vlerwsimi excellent ndwrkohw qw duhet tw ishte parashikuar vlerwsimi good. Nw 25 raste wshtw parashikuar gabimisht kategoria excellent.
* True Positive -> Pwr 32 prej vlerwsimeve, algoritmi ka parashikuar saktwsish njw vlerwsim excellent. Rezultati i parashikuar pwrkon me rezultatin real. Nw 32 raste wshtw parashikuar saktwsisht kategoria excellent.

Pwrfundimisht nga analiza kemi rezultatin e mwposhtwm :

* Numri i parashikimeve tw sakta :

True Negative + True Positive = 38 + 32 = 70 parashikime tw sakta.

* Numri i parashikimeve tw gabuara :

False Negative + False Positive = 18 + 25 = 43 parashikime tw gabuara.

##### 6.1.3 Raporti i klasifikimit

Mw poshtw jepet kodi nw Python pwr shfaqjen e raportit tw klasifikimit qw ndodhet nw tabelwn 5.4. Edhe ky funksion wshtw pjesw e klasws metrics tw siguruar nga paketa sklearn dhe merr si argumenta rezultatet reale tw inputeve tw zgjedhura pwr testim di dhe rezultatet e parashikuara nga algoritmi.

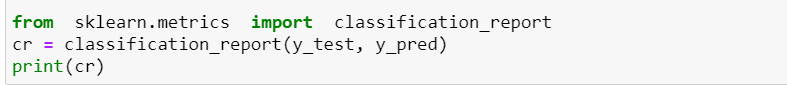
  
Fig. 6.15 Kodi pwr printimin e raportit tw klasifikimit pwr DT

Tabela e raportit tw klasifikimit shfaqet si mw poshtw :

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-score | Support |
| 1 | 0.68 | 0.60 | 0.64 | 63 |
| 2 | 0.56 | 0.64 | 0.60 | 50 |
| Accurancy |  |  | 0.62 | 113 |
| Macro avg | 0.62 | 0.62 | 0.62 | 113 |
| Weighted  avg | 0.63 | 0.62 | 0.62 | 113 |

Tab. 5.4 Raporti pwr klasifikimin LR

Kolona support jep informacion mbi numrin e vlerwsimeve mbi tw cilat wshtw mbwshtetur algoritmi pwr tw parashikuar rezultatet e parashikuara.

Kolona Precision pwrfaqwson numrin e rezultateve good tw parshikuara qw janw realisht good si dhe numrin e rezultateve excellent tw parashikuara qw janw realisht excellent. Pwr rezultatin good kemi 68% tw pwrgjigjeve tw sakta dhe pwr rezultatin excellent 56%.

Formula për llogaritjen e precisionit:

ose

Kolona Recall (e njohur ndryshe edhe si sensitiviteti i algoritmit) përfaqëson numrin e eventeve pozitive (negative) të cilat u parashikuan si duhet sipas formulës:

ose

Kolona F1-score jep mesataren harmonike të precision dhe recall. Nëse bëjmë llogaritjet në varësi të vlerave që morëm për TP, FP,TN dhe FN nga confussion matrix shohim se do të arrijmë në të njëjtat vlera.

##### 6.3 Aplikimi i algoritmit Support Vector Machine

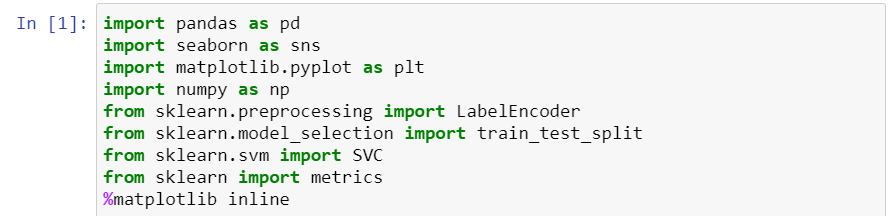
Një “decision tree” është një strukturë e ngjashme me një pemë ku cdo degë përfaqëson një rregull vendimi ndërsa cdo gjethe përfaqëson rezultatin e vendimit. Nyja rrënjë e pemës është nyja e sipërme e cila ndan vendimet në bazë të një atributi vlerë. Këto pemë janë të thjeshta për tu kuptuar dhe interpretuar sepse imitojnë mendimin njerëzor.



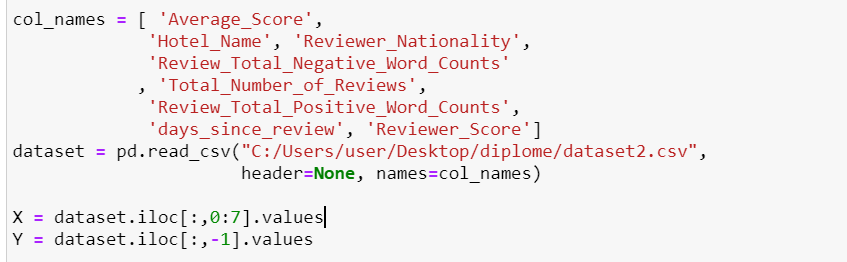
Fig 6.8 Skema e një decision tree

##### 6.3.1 Trajnimi i modelit

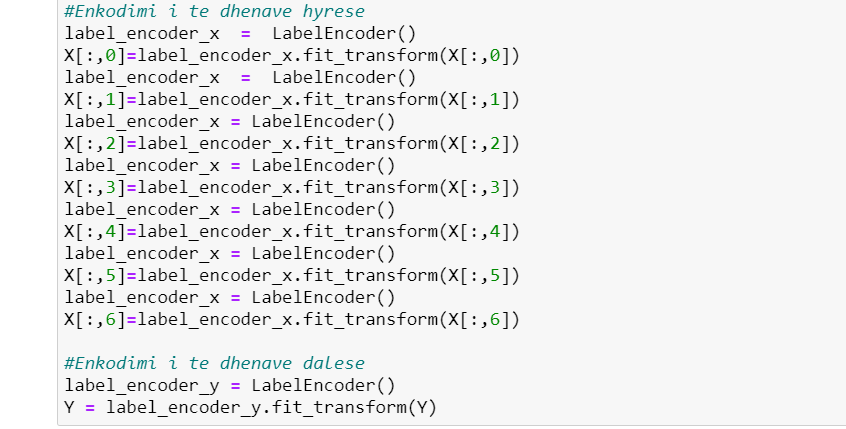
Nw figurwn 6.1 shihen paketat e importuara gjatw trajnimit tw modelit qw janw tw njwjta me ato tw Logistic regression me perjashtimin se kwtu importojmw SVC nga sklearn.svm :

  
Fig. 6.16 Paketat e importuara

Mw poshtw jepet pjesa e leximit tw datasetit me anw tw Python si dhe pwrcaktimi i inputit dhe outputit, njwlloj si nw dy algoritmet e mwparshwm :

  
Fig. 6.17 Pwrcaktimi i inputit dhe outputit

Enkodimi i tw dhwnave nw tw njwjtwn mwnyrw si nw dy rastet e para :

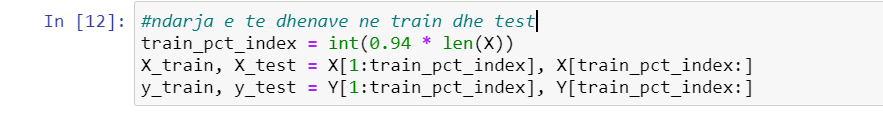
  
Fig. 6.18 Enkodimi i tw dhwnave

Pas enkodimit tw tw dhwnave, fillojmw me trajnimin e modelit duke ndjekr tw njwjtwt hapa si mw parw :

1. Ndajmw tw dhwnat nw :

* Input trajnues (X\_train)
* Input testues (X\_test)
* Output trajnues (y\_train)
* Output testues (y\_test)

Pwrswri edhe nw ndarjen e tw dhwnave kemi vepruar nw tw njwjtwn mwnyrw si nw algoritmin e mwparshwm.

  
Fig. 6.19 Ndarja e tw dhwnave nw tw dhwna trajnimi dhe testimi

1. Deklarimi i klasifikuesit

* Nw rastin tone klasifikuesi wshtw SVC

1. Pwrshtatja e modelit (Trajnimi i modelit)
2. Saktwsia e modelit
3. Parashikimi

Kwto hapa paraqiten nw kodin e mwposhtwm :

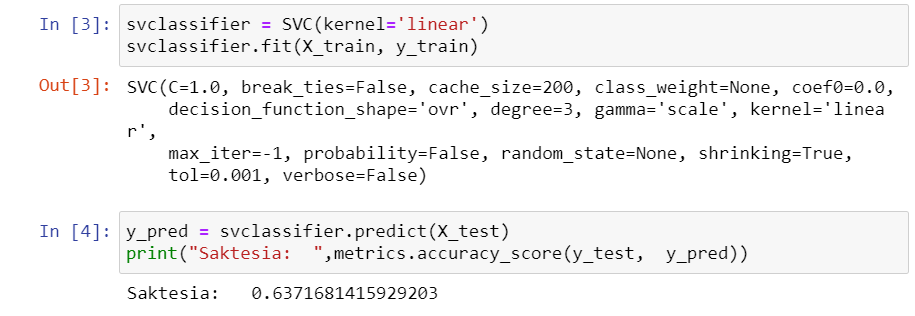
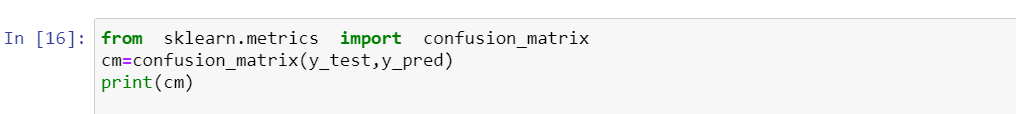


Fig. 6.20 Kodi pwr aplikimin e algoritmit dhe printimit tw saktwsisw

Sic mund ta shohim, nw rastin e kwtij algoritmi arrijmw njw saktwsi thuajse 64%. Pra, algoritmi parashikon njw rezultat tw saktw nw 64% tw rasteve. Printojmw matricwn e konfuzionit dhe raportin e klasifikimit.

##### 6.1.2 Matrica e konfuzionit

Fig. 6.21 Kodi pwr matricwn e konfuzionit

Nga ekzekutimi i kwtij kodi, pwrftojmw matricwn e konfuzionit. Matrica e konfuzionit wshtw njw tabelw qw shwrben pwr matjen e performancws sw algoritmit tw modelit qw do tw trajnohet. Zakonisht, matrica e konfuzionit pwrdoret nw algoritmat e supervised learning. Ajo wshtw njw tabelw e pwrbwrw nga dy rreshta dhe dy kolona, sic paraqitet mw poshtw :

|  |  |  |
| --- | --- | --- |
|  | **Parashikuar si jo** | **Parashikuar si po** |
| **Jo reale** | 34 | 29 |
| **Po reale** | 12 | 38 |

Tab. 5.2 Matrica e konfuzionit DT

Analizojmw matricwn e konfuzionit, nw kwtw rast, pas aplikimit tw SVC, janw arritur rezultatet e mwposhtme :

* True Negative -> Pwr 34 prej vlerwsimeve, wshtw parashikuar saktwsisht kategoria e parashikimit good. Pra, rezultati i parashikuar nga algoritmi pwrkon me vlerwsimin real tw hoteleve. Ne 34 raste wshtw parashikuar saktwsisht kategoria good.
* False Negative -> Pwr 12 prej vlerwsimeve, algoritmi ka parashikuar gabimisht kategorinw good ndwrkohw qw rezultati real ka qenw excellent. Pra, rezultati i parashikuar nga algoritmi nuk pwrkon me vlerwsimin aktual. Nw 12 raste wshtw parashikuar gabimisht kategoria good.
* False Positive -> Pwr 29 prej vlerwsimeve, nga algoritmi wshtw parashikuar gabimisht vlerwsimi excellent ndwrkohw qw duhet tw ishte parashikuar vlerwsimi good. Nw 29 raste wshtw parashikuar gabimisht kategoria excellent.
* True Positive -> Pwr 38 prej vlerwsimeve, algoritmi ka parashikuar saktwsish njw vlerwsim excellent. Rezultati i parashikuar pwrkon me rezultatin real. Nw 38 raste wshtw parashikuar saktwsisht kategoria excellent.

Pwrfundimisht nga analiza kemi rezultatin e mwposhtwm :

* Numri i parashikimeve tw sakta :

True Negative + True Positive = 34 + 38 = 72 parashikime tw sakta.

* Numri i parashikimeve tw gabuara :

False Negative + False Positive = 12 + 29 = 41 parashikime tw gabuara.

##### 6.1.3 Raporti i klasifikimit

Mw poshtw jepet kodi nw Python pwr shfaqjen e raportit tw klasifikimit qw ndodhet nw tabelwn 5.4. Edhe ky funksion wshtw pjesw e klasws metrics tw siguruar nga paketa sklearn dhe merr si argumenta rezultatet reale tw inputeve tw zgjedhura pwr testim di dhe rezultatet e parashikuara nga algoritmi.

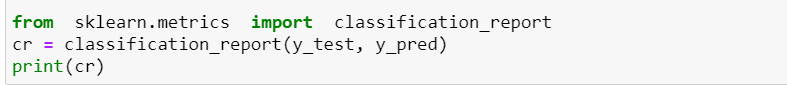
  
Fig. 5.3 Kodi pwr printimin e raportit tw klasifikimit pwr DT

Tabela e raportit tw klasifikimit shfaqet si mw poshtw :

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-score | Support |
| 1 | 0.74 | 0.54 | 0.62 | 63 |
| 2 | 0.57 | 0.76 | 0.65 | 50 |
| Accurancy |  |  | 0.64 | 113 |
| Macro avg | 0.65 | 0.65 | 0.64 | 113 |
| Weighted  avg | 0.66 | 0.64 | 0.64 | 113 |

Tab. 5.4 Raporti pwr klasifikimin LR

Kolona support jep informacion mbi numrin e vlerwsimeve mbi tw cilat wshtw mbwshtetur algoritmi pwr tw parashikuar rezultatet e parashikuara.

Kolona Precision pwrfaqwson numrin e rezultateve good tw parshikuara qw janw realisht good si dhe numrin e rezultateve excellent tw parashikuara qw janw realisht excellent. Pwr rezultatin good kemi 74% tw pwrgjigjeve tw sakta dhe pwr rezultatin excellent 57%.

Formula për llogaritjen e precisionit:

ose

Kolona Recall (e njohur ndryshe edhe si sensitiviteti i algoritmit) përfaqëson numrin e eventeve pozitive (negative) të cilat u parashikuan si duhet sipas formulës:

ose

Kolona F1-score jep mesataren harmonike të precision dhe recall. Nëse bëjmë llogaritjet në varësi të vlerave që morëm për TP, FP,TN dhe FN nga confussion matrix shohim se do të arrijmë në të njëjtat vlera

ML

AI

KDD

SVM

LSTM

LR

DT