
**UNIVERSITETI I TIRANËS**

**FAKULTETI I SHKENCAVE TË NATYRËS**

**DEPARTAMENTI I INFORMATIKËS**

# DISERTACION

**PËR MARRJEN E GRADËS "DOKTOR I SHKENCAVE"**

**ALBA ÇOMO**

# Data mining në Edukim.

**Udhëheqës Shkencor: Prof. Dr. Ilia NINKA**

**Tiranë, 2015**


**UNIVERSITETI I TIRANËS**

**FAKULTETI I SHKENCAVE TË NATYRËS**

**DEPARTAMENTI I INFORMATIKËS**

Disertacion

i

Paraqitur nga

MSc. Alba ÇOMO

Për marrjen e gradës shkencore

**DOKTOR**

**Specialiteti:** Informatikë

**Tema**: Data mining në Edukim.

**Mbrohet më dt ___.___._____ para jurisë:**

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

# Falënderime

*Realizimi i këtij punimi u bë i mundur pas një pune kërkimore disavjeçare, ku përveç kontributit tim personal nuk mund të lë pa përmendur mbështjen e pafund nga shumë njerëz. Është një kënaqësi për mua që të falenderoj të gjithë ata, edhe pse këtu mund të përmend vetëm disa prej tyre.*

*Mirënjohje dhe një falenderim shumë i veçantë për babain tim, shokun tim më të mirë, Mandin. Ku mbështetja dhe drejtimi i tij kanë qënë forca ime kryesore në përfundim e këtij punimi.*

*Një falenderim të veçantë ja dedikoj dhe udhëheqësit tim shkencor, Prof. Dr. Ilia NINKA i cili me mbështetjen, orientimet, sugjerimet dhe profesionalizmin e treguar kontribuoi pa kursim për kurorëzimin me sukses të kësaj pune kërkimore.*

*Një falenderim është për të gjithë miqtë dhe kolegët, të cilët më mbështetën, më ndihmuan dhe më nxitën në problematikat e këtij punimi.*

*Në fund doja të falendëroja nga zemra Bledin, Emën dhe të gjithë familjen time për mbështetjen dhe dashurinë që më kanë dhuruar në jetë.*

*Alba Çomo*

# Përmbledhje

*Për institucionet e arsimit të lartë, qëllimi i të cilave është të kontribuojnë në përmirësimin e cilësisë, suksesi i krijimit të kapitalit human është subjekti i një analize të vazhdueshme. Për këtë arsye parashikimi i suksesit të studentëve është thelbësor për institucionet e arsimit të lartë, sepse cilësia e procesit të të mësuarit është aftësia për të kënaqur dhe plotësuar nevojat e studentëve. Në këtë drejtim janë mbledhur të dhëna dhe informacione të rëndësishme, të cilat janë vënë në dispoziocion dhe të autoriteteve përkatëse, në mënyrë që të ruhet cilësia. Cilësia e institucioneve të arsimit të lartë nënkupton sigurimin e shërbimeve, të cilat plotësojnë më së shumti nevojat e studentëve, stafit akademik, dhe pjesëmarrësve të tjerë në sistemin arsimor. Pjesëmarrësit në procesin arsimor, duke përmbushur detyrimet e tyre nëpërmjet veprimtarive të duhura, krijojnë një sasi tepër të madhe të dhënash të cilat kanë nevojë të mblidhen dhe më pas të integrohen dhe të përdoren. Duke shndërruar këto të dhëna në njohuri, arrihet kënaqësia e të gjithë pjesëmarrësve : studentëve, profesorëve, administratës dhe komunitetit shoqëror.*

*Megjithëse për shumë kohë data mining është zbatuar me sukses në botën e biznesit, përdorimi i saj në arsimin e lartë është ende relativisht i ri. Përdorimi i data mining në arsim nënkupton identifikimin dhe nxjerrjen nga të dhënat të njohurive të reja dhe shumë të vlefshme. Qëllimi i përdorimit të data mining ishte zhvillimi i një modeli nga i cili mund të nxirren përfundime mbi suksesin akademik të studentëve. Metoda dhe teknika të ndryshme të data mining janë krahasuar gjatë këtij punimi duke shqyrtuar të gjithë variablat që mund të kenë ndikim në suksesin e studentit, si variablat shoqërore, demografike, etj.*

*Rezultatet e marra nga aplikimi i metodave të data mining, ndihmojnë institucionet e arsimit të lartë për të parashikuar sjelljen e studentëve. Kështu që i gjithë qëllimi i studimit do të përmblidhej në një pyetje të vetme, që do të ishte:*

- *Si mund të përdoren të dhënat e përftuara nga mjedisi arsimor dhe teknikat e data mining, për parashikimin e suksesit të studentit në arsimin e lartë?*

*Për të gjetur zgjidhjen më të mirë për problemin e përmendur, fillimisht mblidhen dhe përpunohen të dhënat nga mjedisi arsimor dhe më pas zbatohen teknika të caktuara të data mining sipas aftësisë që shfaq secila teknikë për të bërë parashikimin më të saktë dhe për të mundësuar krijimin e një modeli i cili do të shërbejë si një bazë për zhvillimin e një sistemi vendimesh mbështetës në arsimin e lartë.*

# Absctract

*For the institutions of higher education, whose focus is the improvement of education quality, the success in creating human capital is subject to a continuous analysis. For this reason, sudent success forecast is crucial per these institutions, because the quality of the learning process is the ability to satisfy and fulfill the students needs. With this in mind, important information and data has been gathered and has been given to the appropriate authorities so that quality could be preserved. The quality of the higher education institutions means offering the services that fulfill as well as possible the needs of students, academic staff and other partecipants in the education system. The actors in the education process, by fulfilling their obligations through everyday actions, create a big quanitity of data which needs to be collected, processed and used. By transforming this data into knowledge we can achieve the happiness of all partecipants in the education process: studens, professors, administration and social comunity.*

*Although datamining has been used for a long time in the business world, its use in the higher education institutions is relatively recent. The use of datamining in education means the identification and exctraction of new and useful knowledge from existing data. The purpose of the data mining usage was the development of a model that could be used to predict the academic success of the students.During this study, different methods and techiques have been compared by observing all the variables that might influence the sutdent succes, such as sociological variables, demografic ones etc.*

*The results from the aplication of data mining methods, help higher education institutions to predict the student behaviour. If the whole study could be concentrated into one question, this question would be:*

- *How can data mining techniques be used on data gathered from education institutions, to predict the success of sudents in higher education?*

*To find the best possible solution to the up-mentioned problem, data was gathered and prepared from the education system, then specified data mining techniques are applied according to the ability of each technique to predict as correctly as possible, and to enable the creation of a new model that will server as the foundation of a decision system to support high education institutions.*

## Deklarata e origjinalitetit të punimit

Nën përgjegjësinë time deklaroj se ky punim me titull: " ***Data Mining në Edukim.***", me udhëheqës shkencor **Prof.Dr. Ilia NINKA,** është rezultat i punës sime origjinale dhe është bërë personalisht nga vetë unë. Ky nuk është prezantuar asnjëherë para një institucioni tjetër për vlerësim dhe nuk është botuar i tëri. Punimi nuk përmban material të shkruar nga ndonjë person tjetër përveç rasteve të cituara dhe referuara.

**Me respekt**

**M.Sc. Alba ÇOMO**

# Tabela e përmbajtjes

# Lista e Figurave

## Lista e Grafikëve



## Lista e Tabelave

# 1. HYRJE

*Në këtë kapitull, jepet një përshkrim i përgjithshëm mbi studimin. Kapitulli fillon me një sfond në lidhje me përshkrimin e fushës së kërkimit. Më pas ndiqet nga objektivat dhe nga organizimi i kërkimit.*

## 1.1 Hyrje në studim

Për institucionet e arsimit të lartë, qëllimi i të cilëve është të kontribuojnë në përmirësimin e cilësisë së arsimit të lartë, suksesi i krijimit të kapitalit human është subjekti i një analize të vazhdueshme. Për këtë arsye parashikimi i suksesit të studentëve është thelbësor për institucionet e arsimit të lartë, sepse cilësia e procesit të të mësuarit është aftësia për të kënaqur dhe plotësuar nevojat e studentëve. Në këtë drejtim janë mbledhur të dhëna dhe informacione të rëndësishme, dhe janë vendosur në dispozicion dhe nën kujdesin e autoriteteve përkatëse, në mënyrë që të ruhet cilësia e kërkuar. Cilësia e institucioneve të arsimit të lartë implikon sigurimin e shërbimeve, të cilat plotësojnë më së shumti nevojat e studentëve, stafit akademik, dhe pjesëmarrësve të tjerë në sistemin arsimor. Pjesëmarrësit në procesin arsimor, duke përmbushur detyrimet e tyre nëpërmjet veprimtarive të duhura, krijojnë një sasi tepër të madhe të dhënash të cilat kanë nevojë të mblidhen dhe më pas të integrohen dhe të përdoren. Duke shndërruar këto të dhëna në njohuri, arrihet kënaqësia e të gjithë pjesëmarrësve : studentëve, profesorëve, administratës dhe komunitetit shoqëror.
Data mining përfaqëson procesin e përpunimit të të dhënave kompjuterike nga këndvështrime të ndryshme, me qëllim nxjerrjen e shembujve të nënkuptuar (implicit) dhe me interes (Witten, Frank dhe Hall, 2005). Informacioni i nxjerr nga të dhënat e përpunuara, ndihmon shumë çdo pjesëmarrës në procesin arsimor në mënyrë që të përmirësohet procesi i mësimdhënies/mësimnxënies, dhe të përqëndrohen në zbulimin, detektimin dhe shpjegimin e fenomeneve arsimore (El-Halees, 2008).
Të gjithë pjesëmarrësit në procesin arsimor mund të kenë përfitime nga aplikimi i data mining mbi të dhënat e arsimit të lartë,siç shihet edhe në Figurën 1.1.
Në këtë mënyrë me teknikat e data mining, ndërtohet një cikël në sistemin arsimor, cili konsiston në formimin e hipotezave, testimin dhe trajnimin, p.sh. përdorimi i tij mund të lidhet me veprat e shumta të procesit arsimor në lidhje me nevoja specifike (Romero and Ventura, 2007, pp. 136) të: studentëve, profesorëve dhe administratës.

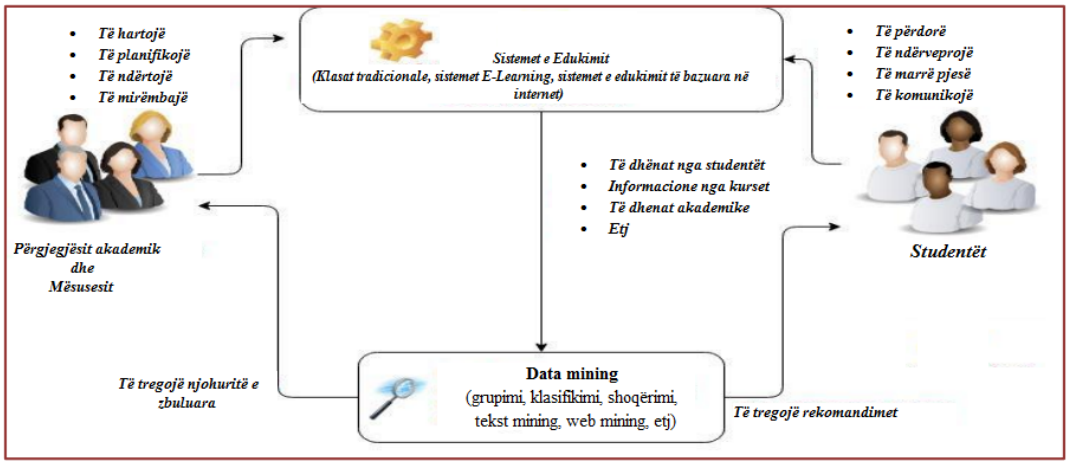



**Figura 1.1 Cikli i aplikimit të data mining në arsim**

Kështu, aplikimi i data mining në sistemin arsimor mund të drejtohet në mënyrë të tillë, që të mbështesë çdo nevojë për secilin nga pjesëmarrësit. Studentit i kërkohet të rekomandojë veprimtari shtesë, materiale mësimi dhe detyra që do të përmirësojnë të mësuarin e tij. Profesorët do të merrnin rezultatin e kërkuar, mundësinë për të klasifikuar studentët në grupe bazuar në nevojën që ata kanë për drejtim dhe monitorim në mënyrë që të gjejnë gabimet, të gjejnë veprimet efektive, etj. Administrimi dhe stafi administrativ, do të marrin parametrat që do të përmirësonin performancën e sistemit (Romero and Ventura, 2007, pp. 136).

Në vitet e fundit është vënë re një interes gjithnjë e më shumë në rritje në përdorimin e data mining për qëllime arsimore. Data mining përfaqëson fusha shumë premtuese të studiuesve në arsim, dhe ka kërkesa specifike në të cilat fushat e tjera kanë mungesa.

Një nga problemet në arsim që zgjidhet me data mining është parashikimi i performancës akademike të studentëve, qëllimi i të cilës është të parashikojë një variabël të panjohur (rezultat, nota, ose pikë) që përshkruan studentin. Vlerësimi i performancës së studentit përfshin monitorimin dhe udhëzimin e studentëve në procesin e të mësuarit dhe në procesin e vlerësimit. Vlerësimi, si proçedura kryesore për matjen e rezultateve të të mësuarit, tregon nivelin e performancës së studentëve, e cila shprehet në sasi dhe në cilësi. Kështu, provimet luajnë një rol mjaft të rëndësishëm në jetën e çdo studenti, duke përcaktuar të ardhmen e tyre. Minaei-Bidgolim, et al. (2003) ishte ndër autorët e parë i cili i klasifikoi studentët duke përdorur algoritmet gjenetik për të parashikuar notën e tyre finale. Superby, Vandamme dhe Meskens (2006) parashikuan suksesin akademik të studentit ( duke e klasifikuar: lëndët me rrezik të ulët,të mesëm dhe të lartë) nëpërmjet përdorimit të metodave të ndryshme të data mining (pemët e vendimit dhe rrjetat neurale). Romero et al. (2008) krahasoi metoda të ndryshme të data mining, në mënyrë që të parashikonte vlerësimin final bazuar në të dhëna e përftuara nga sistemi i të mësuarit elektronik. Zekić-Sušac, Frajman-Jakšić and Drvenkar (2009) krijuan një model për parashikimin e performancës së studentëve duke përdorur rrjetat neurale dhe pemët e klasifikimit për marrjen e vendimeve, dhe me analizën e faktorëve të cilët ndikojnë në suksesin e studentit. Kumar and Vijayalakshmi (2011), duke përdorur pemët e vendimit parashikoi rezultatin e provimit përfundimtar për të ndihmuar profesorët që të identifikonin studentët që kishin nevojë për ndihmë, në mënyrë që të përmirësohej performanca e tyre dhe të kalohej provimi. Suksesi i studimit në institucionet e arsimit të lartë në Shqipëri, deri në ditët e sotme është shqyrtuar vetëm për qëllimin e gjetjes

së notës mesatare, gjatësinë e studimit dhe tregues të ngjashëm, ndërkohë që faktorët që prekin rezultatet e studentëve në një degë të caktuar nuk janë studiuar mjaftushëm.
Në këtë studim janë krahasuar teknika të ndryshme të data mining si rrjetat neurale, pemët e vendimit dhe Naive Bayes. Këto teknika kanë qënë të sukseshme në parashikimin dhe klasifikimin. I gjithë studimi dhe e gjithë puna është bazuar në pyetësorët e aplikuar mbi studentët dhe gjithashtu dhe në të dhëna të marra nga sistemi i një fakulteti të Universitetit të Tiranës, në të cilët përveç të dhënave demografike janë shqyrtuar edhe të dhëna e studentëve para, gjatë dhe pas ciklit të studimeve në universitet.

## 1.2 Përcaktimi i problemit

Suksesi i studentëve është një faktor kyç në mbarëvatjen e arsimit në Shqipëri. Çdo ditë e më shumë, të rinjtë po fitojnë të drejtën për të studiuar në universitetet e vendit. Por, statistikat tregojnë se shumë prej tyre nuk kanë sukses. Në këtë pikë lind nevoja që të bëhet një kategorizim i veçorive dhe karakteristikave të këtyre studentëve në mënyrë që të meren masat e nevojshme. Ajo çfarë nevojitet është një studim i kompletuar mbi këtë fushë dhe një mënyrë që të arrihet parashikimi i performancës së studentëve në arsimin e lartë. Kjo fushë e studimit në vendin tonë është shumë e re dhe akoma e pa eksploruar, gjë që sjell edhe vështirësitë e saj, si për shembull në mbledhjen e të dhënave, duke qenë se edhe sistemi nuk është akoma plotësisht i informatizuar dhe sigurimi i të dhënave nga bazat e të dhënave të universitetit nuk është i mundur.

## 1.3 Objektivat e studimit

Objektivi i institucioneve arsimore, është performanca sa më e lartë e studentëve si edhe mënyrat me të cilat ata mund të ndihmojnë në përmirësimin e kësaj performance. Kështu që këto institucione janë shumë të interesuara për një mënyrë sipas së cilës mund të parashikojnë sjelljen e studentëve, dhe në bazë të kësaj të mund të marrin masa për të dhënë ndihmën në kohën e tyre. Pasi nga kjo do të varej vazhdimësia e arsimit në vendin tonë.
Parashikimet e marra nga aplikimi i metodave të data mining, ndihmojnë këto institucione për të arritur këtë objektiv. Kështu që i gjithë qëllimi i studimit do të përmblidhej në një pyetje të vetme, që do të ishte:

- *Si mund të përdoren të dhënat e përftuara nga mjedisi arsimor dhe teknikat e data mining, për parashikimin e suksesit të studentit në arsimin e lartë?*

Për të gjetur zgjidhjen më të mirë për problemin e përmendur, fillimisht mblidhen dhe përpunohen të dhënat nga mjedisi arsimor dhe më pas zbatohen teknika të caktuara të data mining sipas aftësisë që shfaq secila teknikë për të bërë parashikimin më të saktë.

## 1.4 Organizimi i studimit

Përpara se të aplikohen metodat e data mining, fillimisht krijohet një rrjedhë e punës sipas së cilës drejtohen të gjitha hapat në studim. Siç shihet në figurën 1.2, fillimisht pasi bëhet përcaktimi i problemit, zgjidhet dhe përpunohet një literaturë, nga e cila merren bazat për studimin. Më pas kemi të dhënat, mbi të cilat, pasi i nënshtrohen një procesi përpunimi, do të zbatohen teknikat data mining më të përshtatme për llojin e

problemit. Në fund kemi vlerësimin e rezultateve nga zbatimi i teknikave të data ming dhe shfaqjen e njohurive në mënyrë sa më të kuptueshme.


**Figura 1.2 Një paraqitje skematike e organizimit të studimit**

## 1.5 Plani i doktoraturës

Ky studim është i organizuar në pesë kapituj. Më poshtë po tregojmë një përmbledhje të secilit:

- **Kapitullin I**, paraqet hyrjen e studimit, diskutimin e problemit, objektivat e kërkimit si dhe organizimin i studimit.
- **Kapitulli II**, hedh një vështrim mbi studimet më të rëndësishme që janë zhvilluar në Data Mining në Edukim deri më sot. Fillimisht kapitulli jep një përshkrim të data mining në edukim dhe përshkruan grupe të ndryshme të përdoruesve, tipet e mjediseve arsimore dhe të dhënave që ato japin. Më tej vazhdon me një listë të detyrave më tipike/të zakonshme në mjedisin arsimor, që janë zgjidhur duke përdorur teknikat e data mining. Në fund mbyllet duke diskutuar disa nga drejtimet kërkimore më të rëndësishme.
- **Kapitulli III** përshkruan metodologjinë e studimit, në të do të diskutohet hartimi dhe procesi i këtij studimi. Përveç këtyre, do të diskutohen të gjitha metodat dhe teknikat të cilat janë përdorur për secilin studim të realizuar për qëllim të këtij punimi.
- **Kapitullin IV** paraqet në mënyrë të detajuar të gjitha studimet dhe analizat e realizuara. Duke paraqitur rezultatet e përftuara nga secili hap i parashikimeve, analizave dhe krahasimeve, të tilla si përpunimi i të dhënave, analizat shpjeguese, rezultatet e parashikimit dhe vlerësimi i modelit.
- **Kapitulli V** përmbledh përfundimet dhe konkluzionet e studimit. Në këtë kapitull do të jepen dhe disa rekomandime për pedagogët dhe studentët për të rritur performancën e këtyre të fundit në universitet.
- Në fund përmes dy anekseve do të jepen pyetësorët e përdorur(Aneksi 1) si dhe rezultatin e plotë të gjeneruar nga rastet e studimit.

# 2. DATA MINING NË FUSHËN E ARSIMIT

*Objektivi i këtij kapitulli është të ofrojë shpjegime dhe një përmbledhje të literaturës që lidhet me fushën e studimit. Kapitulli do të ofrojë përkufizime mbi data mining në arsim dhe zbulimin e njohurive, metodat më të përshtatshme të data mining në fushën e arsimit,aplikimet e tyre si dhe tendenca të kësaj fushe të re studimore.*

## 2.1 Data mining në arsim

### *2.1.1 Çfarë është data mining në arsim?*

Website i komunitetit të Data Mining (www.educationaldata mining.org) e përcakton data mining në edukim si më poshtë:

> *" Data Mining në Edukim është një disiplinë në zhvillim, e përqëndruar në zhvillimin e metodave për eksplorimin e llojeve unike të të dhënave që vijnë nga mjediset arsimore, dhe përdorimin e këtyre metodave për të kuptuar më mirë nxënësit,dhe parametrat me të cilat mësojnë ata."*

Data Mining, e quajtur gjithashtu "Zbulimi i Njohurive në Bazat e të dhënave" (KDD-Knowledge Discover Databases), është fusha e zbulimit të informacioneve shumë të dobishme nga sasi të mëdha të të dhënave. Nga studimet e shumta, është vënë re që metodat e data mining në arsim janë shpesh të ndryshme nga metodat standarte të Data Mining, për shkak të nevojës për të llogaritur në mënyrë eksplicite (dhe për mundësitë për të shfrytëzuar) hierarkinë shumë nivelëshe dhe mungesën e pavarësisë në të dhënat arsimore (Baker et al 2009). Për këtë arsye, është gjithnjë e më e zakonshme të shohësh përdorimin e modeleve të nxjerra nga literatura psikometrike në publikimet e data mining në arsim.

Data mining në edukim merret me zhvillimin, kërkimin dhe aplikimin e metodave kompjuterike për të përcaktuar modele të reja duke u nisur nga koleksione të mëdha të dhënash nga fusha e arsimit. Të dhënat që studiohen nga kjo fushë nuk kanë të bëjnë vetëm me të dhënat që merren nga ndërveprimi individual i studentëve e në sistemet e edukimit (psh kërkimi i informacionit, zgjidhja e pyetësorve dhe e ushtrimeve), por mund të përfshijnë gjithashtu nga bashkëpunimi/bashkëbisedimi i studentëve me njëri-tjetrin (psh chat-et), të dhënat administrative (psh shkolla, mësuesit, etj) apo dhe nga të dhënat demografike (psh mosha, gjinia, notat e shkollës).
Shumë nga metodat që përdoren sot në data mining mund të përdoren dhe për data mining në edukim. Karakteristikat që bëjnë dallimin e data mining në edukim është se kjo fushë fokusohet në të dhënat që merren nga sistemet e mësimdhënies/mësimnxënies.
Data mining në edukim merret më zhvillimin e metodave për të eksploruar tipet unike të të dhënave, që janë në mjedisin arsimor dhe përdor këto metoda për t'i kuptuar më mirë studentët dhe mënyrën se si ata mësojnë. Rritja e numrit të programeve kompjuterike për mësimdhënie, po ashtu edhe rritja e bazave të të dhënave me informacion mbi studentët, kanë krijuar një pirg të madh të dhënash që tregojnë se si

studentët mësojnë; përdorimi i internetit në edukim ka krijuar një kontekst të ri të njohurive si e-learning ose edukimi i bazuar në web, në të cilin mund të gjenden sasi të mëdha informacioni mbi ndërveprimin mësues student. Gjithë ky informacion është një minierë ari për studiuesit në data mining në edukim. Ata përdorin këto të dhëna për t'i kuptuar më mirë nxënësit dhe procesin e të mësuarit në mënyrë që të zhvillohen përafrime kompjuterike që përfshijnë të dhëna dhe teori për të shndërruar praktikën në dobi të studentëve. Data mining në edukim është shfaqur si një fushë kërkimi në vitet e fundit për kërkuesit nga e gjithë bota dhe lidhet me fusha kërkimi të tilla si (Romeo & Ventura 2010):

- **Edukimi offline** përpiqet të transmetojë njohuri dhe aftësi duke u bazuar në kontaktin e drejtpërdrejtë dhe gjithashtu studion psikologjikisht sesi njerëzit mësojnë. Mbi informacionin e mbledhur në klasë mund të aplikohen teknika statistikore dhe psikometrike.
- **E-learning dhe sistemi i menaxhimit të mësimit (LMS).** E-learning siguron edukim online dhe LMS siguron mjete komunikimi, bashkëpunimi, administrimi dhe raportimi. Të dhënat në këto metoda janë në formën e regjistrave ose bazave të të dhënave dhe mbi këto të dhëna mund të aplikohen teknika Web Mining (WM).
- **Tutori inteligjent (ITS) dhe sistemi hipermedial arsimor i përshtatshëm (AEHS)** janë alternativa ndaj përafrimit thjesht vendose në web, dhe përpiqen të përshtasin shpjegimin sipas nevojave të secilit student në veçanti. Të dhënat e gjeneruara nga këto sisteme janë në formën e regjistrave, modeleve të përdoruesve, dhe mbi këto të dhëna aplikohen teknika Data Mining.

Procesi i data mining në edukim konverton të dhënat e papërpunuara që vijnë nga sistemi i edukimit dhe i transformon ato në një informacion të dobishëm, i cili mund të këtë një impakt shumë të madh në kërkimin dhe praktikat në edukim. Ky proces nuk ndryshon shumë nga fushat e tjera të aplikimit në data mining si biznes, gjenetikë, mjekësi etj. pasi ajo ndjek të njëjtat hapa si në procesin e përgjithshëm të data mining (paraprocesim - data mining - pasprocesim). Megjithatë është e rëndësishme të vihet re që në këtë kapitull termi Data Mining është përdorur në kuptim më të gjerë sesa në përcaktimin e tij standart. Kjo do të thotë që nuk do të përshkruhen vetëm teknikat e data mining në edukim që përdorin teknika tipike si klasifikimi, grupimi, gërmimi me bashkim rregullash, gërmimi sekuencial, gërmimi tekst etj., por do të përdoren dhe përafrime të tjera si regresioni, korrelacioni, vizualizimi etj., që nuk konsiderohen të jenë teknika data mining në një kuptim të mirëfilltë. Për më tepër disa përmirësime metodologjike në data mining në edukim, si zbulimi me modele dhe integrimi i modelimit psikometrik, janë të pazakonta në kategoritë e data mining dhe nuk konsiderohen të jenë metoda tradicionale data mining.

Nga ana praktike data mining në edukim lejon të zbulohen njohuri të reja të bazuara mbi të dhënat e përdorimit nga studentët, për të vlerësuar sistemet arsimore dhe për të përmirësuar aspekte të cilësisë së edukimit dhe për të ndërtuar një proces më efektiv mësimdhënieje. Disa ide të ngjashme janë përdorur në mënyrë të sukseshme në sistemet e e-commerce, për të zbuluar interesin e klientëve në mënyrë që të rrisin shitjet online, megjithatë ka pasur më pak progres në drejtimin e edukimit deri më sot, megjithëse ky trend po ndryshon pasi po rritet interesi për të aplikuar data mining në mjedisin arsimor. Megjithatë ka disa pika të rëndësishme që diferencojnë aplikimin e data mining në edukim nga mënyra se si përdoren në fusha të tjera (Romoe & Ventura 2007):

- **Objektivi i data mining:** në secilën fushë ku ai aplikohet është i ndryshëm, për shembull në *biznes* objektivi kryesor është që të *rritet fitimi* i cili është i prekshëm dhe mund të matet më lekët e fituara, numrin e klientëve dhe besnikërinë e klientëve. *EDM*(Educational Data Mining) mund të përdoret si për të *përmirësuar procesin e të nxënit* duke drejtuar studimin e nxënësve, por mund të përdoret edhe për qëllime të qarta kërkimore si studimi më i thellë i fenomeneve arsimore.
- **Të dhënat në mjediset arsimore**: ka disa tipe të dhënash për t'u studiuar. Këto të dhëna janë specifike në fushën arsimore dhe kanë informacion semantik të komplikuar, kanë marrëdhënie me të dhëna të tjera dhe mund të gjenden në disa nivele hierarkie. Për shembull literatura e një kursi përbëhet nga disa kapituj, çdo kapitull ka disa mësime, dhe çdo mësim ka disa koncepte dhe nëpërmjet tyre mund të paraqiten marrëdhëniet ndërmjet pyetjeve të një testi dhe koncepteve të vlerësuara nga ai. Për më tëpër duhet që të merren në llogari, aspektet pedagogjike të mësuesit dhe të sistemit.
- **Teknikat të dhënat dhe problemet në arsim** kanë disa karakterisitika të veçanta që kërkojnë që data mining të trajtohet në një mënyrë të ndryshme, megjithëse disa teknika tradicionale mund të aplikohen direkt, të tjera duhet të përshtaten në problemin arsimor specifik që do të trajtojnë.

EDM përfshin grupe të ndryshëm përdoruesish apo pjesëmarrësish, grupet e ndryshme e shikojnë informacionin mbi arsimin nga pikëpamja specifike që ka të bëjë me misione, vizione dhe objektivat për të përdorur data mining. (Hanna 2004) për shembull, njohuria e zbuluar nga algoritmet EDM mund të përdoret jo vetëm për të ndihmuar mësuesit të menaxhojnë klasat e tyre, kuptojnë procesin e nxënies së studentëve dhe të korrigjojnë metodat e tyre shpjeguese, por gjithashtu mund të përdoren dhe nga pikëpamja e studentëve për të marrë mendime të tjera mbi situatën dhe për t'u dhënë atyre feedback (Merceron & Yacef 2008). Megjithëse nga pikëpamja e jashtme duket sikur janë vetëm 2 grupe që përfshihen, mësuesit dhe nxënësit, ka dhe shumë grupe të tjerë të përfshirë dhe këta mund të shihen te tabela 2.1.

Kohët e sotme ka një varietet të madh sistemesh edukimi siç janë klasat tradicionale e-learning, LMS, sistemet e edukimit, teste dhe pyetësorë, dhe të tjera si objekte mësimi, harta konceptesh, forume, lojëra edukative etj. Të gjitha të dhënat e ofruara nga mjediset e sipërme edukuese janë të ndryshme kështu që mund të përdoren në zgjidhjen e problemeve të ndryshme duke përdorur teknika data mining.

**Tabela 2.1 Përdoruesit/aksionerët e EDM-së**

| Përdoruesit/Aktorët | Qëllimet e përdorimit të data mining |
|---|---|
| Studentët /Nxënësit | ▪ Për të personalizuar e-learning<br>▪ Për ti rekomanduar nxënësve aktivitetet, burimet dhe mënyrat e të mësuarit ose eksperienca interesante të të mësuarit<br>▪ Për të përmirësuar më tej mësimin e tyre<br>▪ Për të sugjeruar rrugë më të shkurtra ose thjesht linqe për t'u ndjekur, të rekomandojë kurse, diskutime të rëndesishme, libra etj |

| | |
|---|---|
| Edukatorët/Mësuesit/ Instruktorët/Tutorët | ▪ Të marrin feedback për instruksionet<br>▪ Të analizojnë sjelljen dhe të mësuarin e studentëve<br>▪ Për të identifikuar studentët që kanë nevojë për ndihmë<br>▪ Për të parashikuar performancën e studentëve<br>▪ Për të klasifikuar nxënësit në grupe<br>▪ Të gjejë gabimet e bëra më shpesh<br>▪ Të përcaktoj aktivitetet më efektive<br>▪ Të përmirësoj dhe përshtasë kurset etj. |
| Zhvilluesit e kurseve/ Kërkuesit e edukimit | ▪ Të vlerësojnë dhe mirëmbajnë kurset dhe përmbatjen e tyre në procesin e mësimit<br>▪ Të përmirësojnë mësimin e nxënësve<br>▪ Për të ndërtuar automatikisht modelet e studentëve dhe tutorëve<br>▪ Të krahasojnë teknikat data mining në mënyrë që të mund të rekomandojnë më të mirën për secilin rast<br>▪ Të zhvillojnë mjete specifike data mining për qëllime edukuese, etj. |
| Organizatat/Universitetet / Qëndrat private të kurseve | ▪ Per të rritur procesin e marrjes së vendimeve në institucionet e shkollës së mesme<br>▪ Të arrijnë qëllime të caktuara<br>▪ Të sugjerojnë kurse të caktuara që mund të jenë të nevojshme për secilën klasë ose nxënës<br>▪ Të gjejë mënyrën më të mirë për përmiresimin e notave<br>▪ Të zgjedhë aplikantët më të kualifikuar për t'u diplomuar<br>▪ Të ndihmojë në pranimin e studentëve më të aftë në universitet, etj. |
| Administratorët/ Drejtoritë Arsimore/ Admisistratorët e rrjetit/ Administratorët e sistemit | ▪ Për të zhvilluar mënyrën më të mirë për të organizuar burimet e institucioneve (njerëzore dhe materiale) dhe ofrimet e tyre në edukim<br>▪ Për të përdorur me efiçencë burimet ekzistuese<br>▪ Për të rritur ofertat e programeve edukative dhe për të përcaktuar efektivitetin e qasjes së mësimit në distancë<br>▪ Për të vlerësuar mësuesit dhe kurikulat<br>▪ Për të vendosur parametra për të përmirësuar efektivitetin e website dhe për t'i përshtatur ato me përdoruesit (madhësi optimale e serverit, shpërndarje e ngarkesës së rrjetit, etj.). |

### *2.1.2 Metodat e data mining në arsim*

Metodat e data mining në edukimin shpesh ndryshojnë nga metodat e literaturave të gjera të data mining, në mënyrë të qartë duke shfrytëzuar nivelet e shumta të hierarkisë kuptimplotë të të dhënave edukative. Metodat kategorizohen në pesë kategoritë e dhën në figurën 2.1, ku tre kategoritë e para përdoren dhe nga fusha të tjera të data mining, ndërsa dy katëgoritë e tjera përdoren kryesisht në fushën e edukimit (Romero & Ventura (2007), Baker & Yacef (2009)).


**Figura 2.1 Metodat Data Mining në Edukim**

- ***Parashikimi***

Tek parashikimi qëllimi është të zhvillohet një model i cili mund të nxjerrë një aspekt të vetëm të të dhënave (variablave të parashikuara), nga disa kombinime të aspekteve të tjera të të dhënave (variablave parashikuese). Parashikimi ka dy përdorime kryesore në data mining në edukim. Në disa raste, metodat e parashikimit mund të përdoren për të studiuar se cilat veçori të një modeli janë të rëndësishme për parashikim, duke dhënë informacion për çështjen e caktuar. Kjo është një çështje e zakonshme për studimin e rezultateve të studentëve pa parashikuar fillimisht faktorë të ndërmjetëm. Në një tip të dytë përdorimi, metodat e parashikimit përdoren për të parashikuar sesi do të jetë variabla rezultat në një kontekst kur ajo nuk është e dukshme.

Janë tre tipe parashikimi që përdoren më së shumti:

- *Klasifikimi.* Në klasifikim, variabla e parshikuar është binare ose kategorike. Disa nga metodat më popullore të klasifikimit përfshijnë:
  - *Pemët e vendimit*
  - *Regresionin logjik* (për parashikime binare)
  - *Makinat e mbështetura në vektor*.
- *Regresioni*. Në regresion, variabla e parashikuar është variabël e vazhdueshme. Disa nga metodat më të përdorura të regresionit përfshijnë:
  - *Regresionin linear*
  - *Rrjetat neurale*
  - R*egresionin në makinat e mbështetura në vektorë*.
- *Vlerësimi i densitetit.* Në vlerësimin e densitet, variabla e parashikuar është një funksion densiteti probabilitar.

Për çdo tip parashikimi, variablat hyrëse mund të jenë kategorike ose të vazhdueshme; metoda të ndryshme parashikuese mund të jenë më efektive në varësi të variablave hyrëse të përdorura.

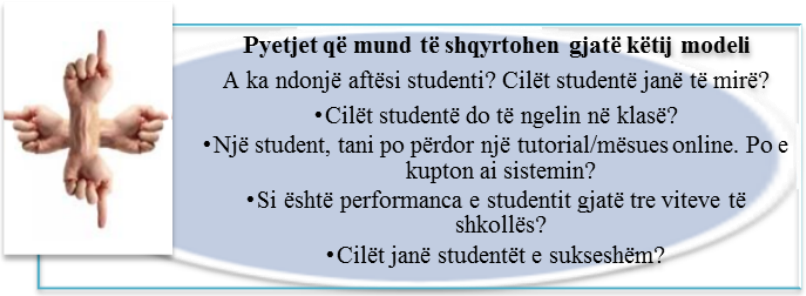



**Figura 2.2 Pyetje që mund të shqyrtohen gjatë modelit të parashikimit**

- **Grupimi**

Tek grupimi qëllimi është të gjenden veçori të dhënash që natyrshëm grupohen së bashku, duke e ndarë bashkësinë e të dhënave në bashkësi grupesh. Grupimi është veçanërisht i dobishëm në rastet ku kategoritë më të shpeshta të bashkësisë së të dhënave njihen paraprakisht. Nëse një bashkësi grupesh është optimale, në një kategori, çdo veçori e të dhënave do të jetë më e ngjashme me veçoritë e të dhënave të tjera brenda atij grupi sesa në grupet e tjera. Grupet mund të krijohen me madhësi të ndryshme (për shembull shiko figurën 5).

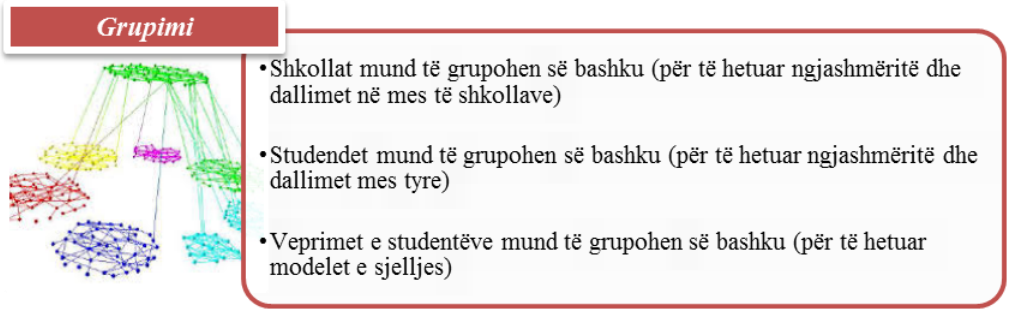



**Figura 2.3 Grupimet që mund të krijohen gjatë grumbullimit të të dhënave**

Algoritmet e grupimit ose mund të fillojnë pa hipoteza fillestare rreth grupeve në të dhënat (siç janë algoritmet k-means më një fillim random), ose fillojnë nga një hipotezë specifike, mundësisht të gjeneruar në një studim të mëparshëm më një bashkësi tjetër të dhënash. Një algoritëm grupimi mund të marrë si të mirëqenë që çdo veçori e të dhënave mund t'i përkasë ekzaktësisht një grupimi (si të algoritmi k-mean), ose mund të marrë si të mirëqenë që disa veçori mund t'i përkasin më shumë se një grupimi ose asnjë grupimi.

- **Metodat e marrëdhënieve**

Qëllimi i këtij modeli është të zbulojnë lidhje ndërmjet variablave, në një bashkësi të dhënash ku ndodhen shumë variabla. Kjo mund të duket si një përpjekje për të gjetur se cilat variabla janë më shumë të lidhura më një variabël të vetëm, për të cilën kemi interes, ose mund të marrë formën e përpjekjes për të zbuluar se cilat lidhje mes dy variablave janë më të forta.

Katër tipet më të përdorura të metodave të marrëdhënieve janë:

- *Rregullat e shoqërimit*, qëllimi është të gjendet rregulla të tipit *if-then* në mënyrë që nëse gjendet një bashkësi variablash, një variabël tjetër do të ketë një vlerë specifike.
- *Korelacioni*, qëllimi është të gjendet korelacione lineare (pozitive ose negative) mes variablave.

- *Modelet sekuenciale*, qëllimi është të gjejmë lidhje të përkohëshme ndërmjet ngjarjeve - për shembull të përcaktojmë cilat sjellje të studentëve çojnë drejt në një ngjarje të të mësuarit me interes.
- *Të dhënat rastësore*, qëllimi është që të gjendet nëse një ngjarje është shkaku i një ngjarje tjetër, ose duke analizuar kovariancën mes dy ngjarjeve, ose duke përdorur informacion mbi mënyrën se si është kushtëzuar një ngjarje.

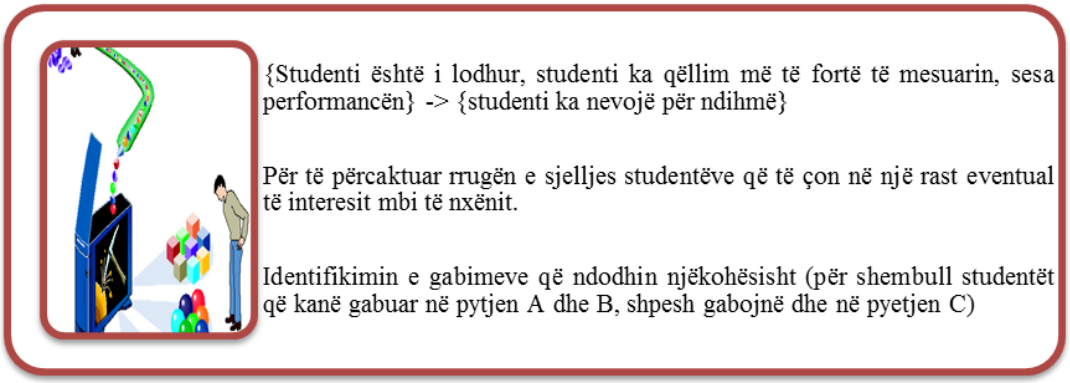



**Figura 2.4 Shembuj për relacionet ndërmjet tipareve të të dhënave nga sistemi i edukimit**

Lidhjet e gjetura mes marrëdhënieve duhet të kënaqin dy kritere:

- *Kuptimin statistikor* zakonisht vlerësohet nëpërmjet testeve statistikore standarte, siç janë F-tests.
- *Interesueshmërinë* e çdo gjetje vlerësohet në mënyrë që të reduktohet bashkësia e rregullave/korelacioneve/lidhjeve rastësore që i komunikohen gjurmimit të të dhënave. Në bashkësi shumë të mëdha të dhënash mund të gjenden qindra mijëra lidhje. Interesueshmëria mat përpjekjen për të përcaktuar se cilat gjetje janë më të dallueshmet dhe më të mirë mbështetura nga të dhënat.

- **Gjykimi njerëzor**

Destilimi i të dhënave për gjykim njerëzor është një tjetër fushë interesante në data mining në edukim. Në disa raste, qeniet njerëzore mund të nxjerrin konkluzione rreth të dhënave, kur ato shfaqen në mënyrën e duhur, të cilat jane përtej qëllimit të menjëhershëm të metodave plotësisht të automatizuara të data mining.
Metodat në këtë sferë të data mining në edukim janë metoda vizuale informimi. Në data mining në edukim të dhënat distilohen për gjykim njerëzor për dy arsye:

- *Identifikimin.* Kur të dhënat distilohen për identifikim, të dhënat shfaqen në mënyra që i mundësojnë qenies njerëzore që të dallojë lehtësisht modele të njohura por që mgjithatë janë të vështira për tu shprehur formalisht
- *Klasifikimin*. Të dhënat mund të destilohen për klasifikim, në mënyrë që të mund të mbështesin zhvillimin e mëvonshëm të një modeli parashikimi.

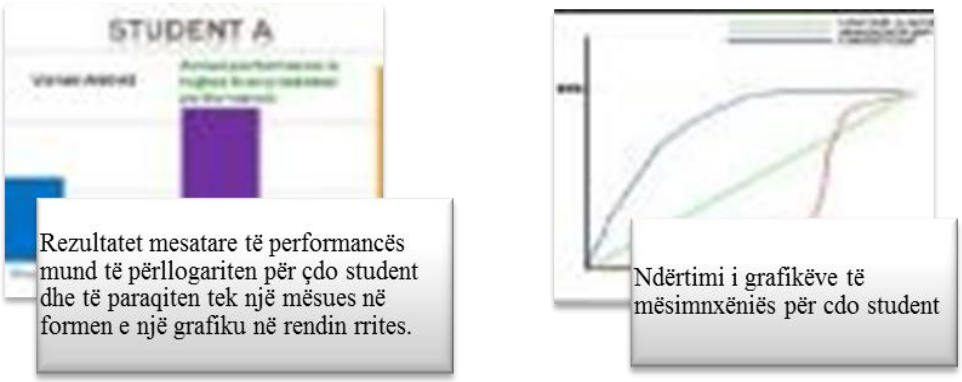
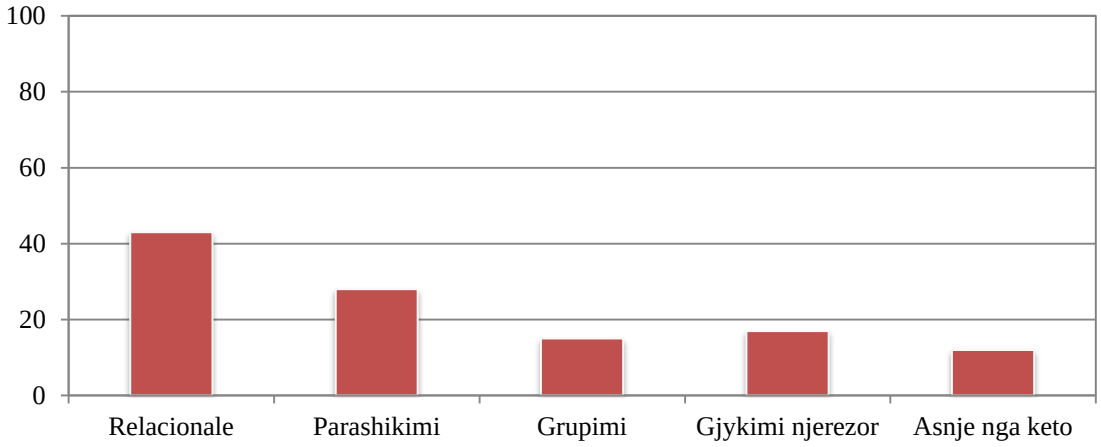



**Figura 2.5 Mënyra të ndryshme për paraqitjen e rezulatit në mënyrë të kuptueshme nga njeriu**

- **Zbulimi me modele**

Kategoria e pestë e taksonomisë së Baker për data mining në arsim, ndoshta është kategoria më e pazakontë, nga një perspektivë klasike e data mining. Në zbulimin me modele, një model i një fenomeni është zhvilluar përmes çdo procesi që mund të vlerësohet në ndonjë mënyrë (më së shumti, inxhinierimi i parashikimeve ose njohurive), dhe ky model më pas përdoret si komponent në një tjetër analizë, të tillë si parashikimi apo relacionet. Zbulimi me modele po bëhet një metodë gjithnjë e më e përhapur në kërkimet mbi data mining në edukim, duke mbështetur analizat e sofistikuara të tilla si: nga cilat materiale nën-kategori nxënësish do të kenë përfitim më të madh nga (Beck dhe Mostow 2008), si lloje të ndryshme të sjelljeve të nxënësve ndikojnë në mënyrat e ndryshme të të mësuarit (Cocea et al. 2009), dhe si ndikojnë variacionet e planeve të mësuesve në sjelljen e nxënësve me kalimin e kohës (Jeong dhe Biswas 2008).
Tek zbulimi në bazë të një modeli, një model i një fenomeni zhvillohet në bazë të parashikimit, grupimit, ose shpesh nëpërmjet inxhinierimit të njohurive. Ky model më pas përdoret si një komponent në analiza të tjera, siç janë parashikimi ose relacionet.
Në rastin e parashikimit, parashikimet e modelit të krijuar përdoren si variabla parashikuese në parashikimin e një variable të re. Në rastin e relacioneve, studiohen lidhjet mes parashikimeve të modelit të krijuar dhe variablave shtesë.

Historikisht, llojet e ndryshme të metodave relacionale, kanë qenë kategoria më e shquar në kërkimin mbi EDM. Në anketimin e Romero & Ventura për kërkimet rreth EDM-së (Data mining në Edukim) nga 1995-2005, 60 artikujt raportuan se shfrytëzuan metodat EDM-së për t'iu përgjigjur interesave të kërkimeve të tyre. 26 prej këtyre (43%) përfshinë metodat relacionale, 17 të tjera (28%) përfshinë metodat e parashikimit të llojeve të ndryshme. Metodat e tjera ishin më pak të zakonshme. Shpërndarja e plotë e të gjithë metodave të përftuara nga anketimi i bërë tregohet në grafikun e më poshtë.




**Grafiku 2.1 Proporcioni i artikujve që përdorin secilën nga metodat e data mining në edukim në anketimin e Romero dhe Ventura (2007) nga viti 1995-2005.**

***** *Vini re që artikujt mund të përdorin shumë metoda, kështuqë disa artikuj mund të gjenden në shumë kategori.*

### *2.1.3 Tendenca të rëndësishme në studimet e data mining në arsim*

Në këtë pjesë, do të shohim zhvillimin dhe ecurinë e data mining në vitet e fundit, dhe do të hetojmë cilat janë tendencat më të mëdha në studimin e EDM. Për të hetuar se çfarë tendencash janë aktualisht, ne analizojmë atë çfarë studiuesit ishin duke studiuar më parë, dhe çfarë janë duke studiuar tani, në këtë mënyrë kuptojmë se çfarë është e re dhe se çfarë atributesh kanë patur për disa kohë studimet e EDM.

Një mënyrë për të parë se ku ka qenë EDM në vitet e mëparshme është shqyrtimi i artikujve me ndikim më të madh në ato vite. Studimi i realizuar nga e Romero dhe e Ventura (2007) na jep një listë të plotë të artikujve, të botuar në vitet 1995 - 2005, të cilat shihen në data mining arsimor. Për të përcaktuar se cilët ishin artikujt me më shumë ndikim, ne shohim numrin e citateve qe ka marrë çdo artikull, një masë kjo që shpesh përdoret për të treguar ndikimin e artikujve, studiuesve, apo institucioneve. Tetë artikujt më të cituar të aplikuara në anketën e Romero-s dhe Ventur-ës janë të dhënë në Tabelën 2.2. Këto artikuj kanë patur shumë ndikim mbi studiuesit e data mining në arsimpor dhe në fusha të ngjashme; si të tillë, ata ilustrojnë me shembuj shumë nga tendencat kyçe në studimin tonë.

Siç kemi ilustruar dhe më lart (Grafiku 2.1), metodat relacionale të llojeve të ndryshme ishin tipi më i dalluar i studimeve të data mining në arsim ndërmjet viteve 1995 dhe 2005.

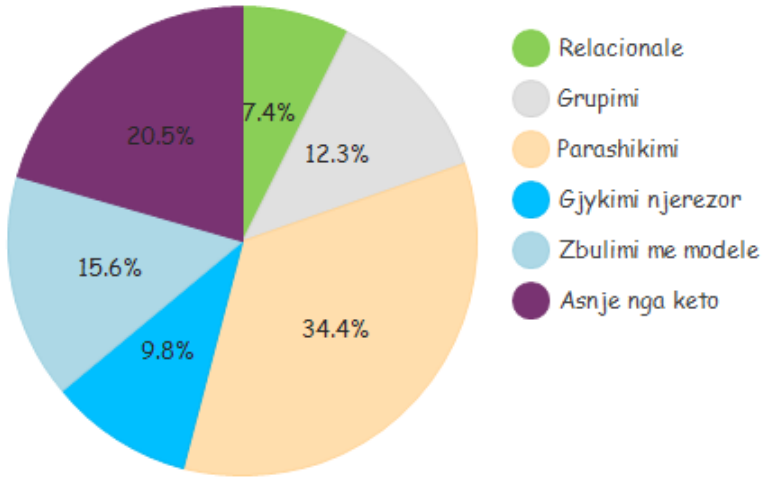
Ndërkohë që modeli relacional ishte mbizotërues nga vitit 1995 deri në vitin 2005, në vitet pasuese ai kaloi në vendin e pestë, me vetëm 9% të artikujve që e përfshinin. Parashikimi u zhvendos në pozicion dominus duke përfaqësuar kështu 42% të artikujve. Analiza e të dhënave sipas Gjykimit të Njeriut (Eksplorues) dhe Grupimi mbetën pothuajse në të njëjtat pozicione si më parë 12% dhe 15% të artikujve. Një metodë e re, që u dallua në mënyrë më të konsiderueshme në më shumë se sa në vitet e mëparshme, është zbulimi me modele. Edhe pse nuk u përfshi në studimet e Romero & Ventura, kjo metodë zë vendin e dytë për nga përdorimi duke përfaqësuar përkatësisht 19% të artikujve.

**Tabela 2.2 Tetë artikujt më të cituar në studimin e realizuar nga Romeo & Ventura në vitet 1995-2005**

| Artikulli | Numri i Citimeve |
|---|---|
| Zaïane, O. (2001). Web usage mining for a better web-based learning environment. Proceedings of Conference on Advanced Technology for Education, 60–64. | 110 |
| Zaïane, O. (2002). Building a recommender agent for e-learning systems. Proceedings of the International Conference on Computers in Education, 55–59. | 89 |
| Baker, R.S., Corbett, A.T., Koedinger, K.R. (2004) Detecting Student Misuse of Intelligent Tutoring Systems. Proceedings of the 7th International Conference on Intelligent Tutoring Systems, 531-540. | 83 |
| Tang, T., McCalla, G. (2005) Smart recommendation for an evolving elearning system: architecture and experiment, International Journal on E-Learning, 4 (1), 105–129. | 63 |
| Merceron, A., Yacef, K. (2003). A web-based tutoring tool with mining facilities to improve learning and teaching. Proceedings of the 11th International Conference on Artificial Intelligence in Education, 201–208. | 54 |
| Romero, C., Ventura, S., de Bra, P., & Castro, C. (2003). Discovering prediction rules in courses. Proceedings of the International Conference on User Modeling, 25–34. | 46 |
| Beck, J., & Woolf, B. (2000). High-level student modeling with machine learning. Proceedings of the 5th International Conference on Intelligent Tutoring Systems, 584–593. | 43 |
| Dringus, L.P., Ellis, T. (2005) Using data mining as a strategy for assessing asynchronous discussion forums, Computer and Education Journal , 45, 141–160. | 37 |

Një tendencë tjetër kyçe është rritja e ndikimit të modelimit të framework-eve nga Item Response Theory, Bayes Nets, dhe Markov Decision Processes. Këto metoda që ishin shumë të rralla në fillimet e data mining në arsim, filluan të dallohen rreth vitit 2005 (shfaqur, për shembull, në Barnes (2005) dhe (Desmarais dhe Pu 2005)), dhe u gjetën në 28% të. Rritja e përdorimit të këtyre metodave është e ngjashme me një reflektim të integrimit të studiuesve nga psikometria dhe modelimit të studentëve në mjedisin e Data Mining në Edukim.

Një ndryshim interesant ndërmjet punës në vitet e fundit, dhe punëve të mëparshme të data mining në arsim, është burimi nga vijnë të dhënat mësimore. Në vitin 1995 deri në vitin 2005, pothuajse të gjitha të dhënat vinin nga grupi i studimit që drejtonte procesin analizues - që do të thotë, në mënyrë që të bëheshin studime në data mining, një studiues fillimisht duhet të mblidhte të dhënat e veta arsimore.Kjo domosdoshmëri filloi të zhdukej në vitin 2008, për shkak të dy zhvillimeve. Së pari, Pittsburgh Science of Learning Center hapi një depo të të dhënave publike, të quajtur PSLC DataShop, i cili vë në dispozicion sasi të konsiderueshme të të dhënave nga një shumëllojshmëri mjedisesh të mësimit online, pa pagesë, për çdo studiues në mbarë botën. Së dyti, studiuesit po përdorin gjithnjë e më shpesh mjedise egzistuese për kurse online të përdorura nga një numër i madh studentësh në mbarë botën, të tilla si Moodle dhe WebCAT.



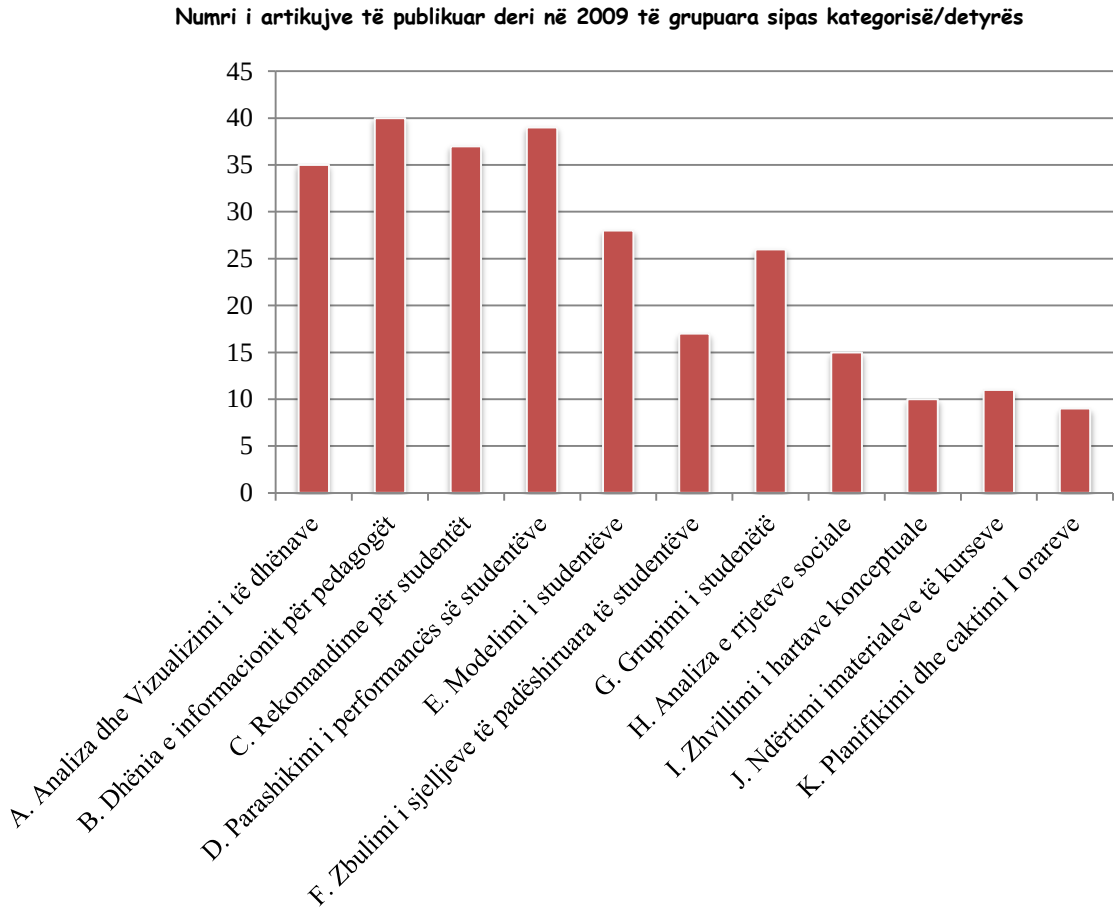

**Grafiku 2.2 Proporcionet e artikujve që përfshijnë secilën nga metodat e data mining në edukim vitet e fundit.**

Nëse kjo tendencë vazhdon, do të ketë përfitime të dukshme në fushën e data mining në edukim. Mes gjithë këtyre zhvillimeve, do të bëhet shumë më i lehtë vlerësimi nga jashtë i një analize. Nëse një studiues bën një analizë që prodhon rezultate që duken jo objektive ose "shumë të mira për të qenë të vërteta", një tjetër studiues mund të shkarkojë të dhënat dhe t'i kontrollojë personalisht. Një përfitim i dytë është se studiuesit do të kenë më shumë mundësi për të ndërtuar mbi përpjekjet e mëparshme të të tjerëve. Siç është e arsyeshme modele parashikuese të strukturës së fushave dhe njohuritë e një pas njëshme të studenti bëhen të disponueshme për grupe të dhënash publike, studiues të tjerë do të jenë në gjendje të testojnë modele të reja të këtyre fenomeneve në krahasim me një bazë të fortë. Rezultati është një shkencë e arsimit që është më konkrete, më e vlefshme, dhe më shumë progresive se sa ishte e mundur më parë.

## 2.2 Detyrat edukative dhe teknikat Data Mining

Ka shumë aplikime ose detyra në mjedisin edukues që janë zgjidhur me data mining (siç kemi folur dhe në paragrafin 2.1.3). Për shembull, *Baker* (Baker & Yacef (2009), Baker (2010)) sugjeron që katër fushat kryesore të aplikimit të EDM-se janë përmirësim i modeleve të studentit, përmirësim i modeleve të domain-it, studim i suportit pedagogjik nëpërmjet aplikacioneve edukues dhe kërkimit shkencor në nxënien e nxënësit. Nga ana tjetër ka pesë përafrime/metoda të tij, që janë parashikimi, grupimi, zbulimi i lidhjeve, distilimi i të dhënave për gjykim njerëzor dhe zbulimi me modele. *Castro (*Castro et al. 2007*)* nga ana tjetër sugjeron detyra/subjektet e mëposhtme të EDM: aplikime që merren më vlerësimin e performancës nxënëse të studentit, aplikime që ofrojnë rekomandime mbi përshtatjen e nxënies dhe lëndes në varesi të sjelljes së studentit, përafrime që merren më vlerësimin e materialit studimor dhe vlerësimin e kurseve edukative të bazuara në web, aplikime që përfshijnë feedback nga të dyja palët, mësues dhe student në kurset e-learning dhe zhvillimet për zbulimin e sjelljeve jotipike të nxënies së studentëve. Megjithatë ka akoma më shume aplikime kështu që janë krijuar disa kategoritë (shiko grafikun 2.3) për detyrat

kryesore në edukim të cilat përdorin teknikat data mining. Keto kategori vijnë nga komunitete të ndryshme kërkimore, të cilat përdorin teknika të ndryshme data mining.




**Grafiku 2.3 Numri i artikujve të publikuar deri në 2009 te grupuara sipas kategorisë/detyrës**

Siç mund të vihet re dhe nga figura 10 kategoritë ose linjat e kërkimit që kanë patur më shumë studime të komplikuara janë 8 të parat, nga A te G, secila me më shumë se 23 studime të publikuara, dhe kategoritë me më pak publikime janë te 4 të fundit nga H të K, me më pak se 15 referenca për secilën. Kjo ndodh ngaqë 8 kategoritë e para janë më të vjetra se të fundit(kështu që më shumë autorë kanë punuar në këto detyra), por gjithashtu mund të jetë edhe sepse studiuesit kanë interes të veçantë për secilën prej tyre. Për shembull, edhe pse analiza e rrjeteve sociale është një nga detyrat më të reja, ka më shume publikime për të sesa për 3 metodat e tjera më pak të preferuara. Gjithashtu vlen të përmendet që kategoritë janë grupuar sipas ngjashmërive që kanë më njëra-tjetrën, dhe sipas mendimit tonë ngjashmëritë janë, A dhe B që japin informacion mbi strukturën, C jep informacion mbi studentët, D, E, F, G studiojnë karakteristikat e studentëve, H dhe i studiojnë grafet dhe marrëdhëniet midis studentëve dhe koncepteve, dhe J dhe K ndihmojnë në krijimin/planifikimin respektivisht të materialit të kursit dhe kursin. Më poshtë do të shpjegohet me detaje secila nga këto kategori dhe studimi më i rëndësishëm i bërë për to. Por duke qënë se disa fusha janë të lidhura ngushtë më njëra-tjetren, disa tipare mund të shtrihen në disa kategori. (Romeo & Ventura 2010)

### *2.2.1 Analiza dhe vizualizimi i të dhënave*

Objektivi i analizës së vizualizimit të të dhënave është të vërë në dukje informacionin e dobishëm dhe të ndihmojë në marrjen e vendimeve. Në ambjentin edukativ për shëmbull, ajo mund të ndihmojë edukatorët dhe administruesit e kurseve që të analizojnë aktivitetet e kurseve dhe përdorimin e informacionit të tyre për të marrë një pamje të përgjithshme të nxënies së studentëve. Statistikat dhe vizualizmi i informacionit janë dy tenikat që janë përdorur më shumë për këtë detyrë.

Statistika është shkenca matematikore që merret më grumbullimin, analizën, interpretimin dhe paraqitjen e të dhënave. Është relativisht e thjeshtë për të marrë statistika bazë mbi të dhënat nga programe statistikore. E përdorur mbi të dhënat e edukimit kjo analizë përshkruese mund të na japë karakteristika të përgjithshme të të dhënave, si përmbledhjet dhe raportet mbi sjelljen e nxënësve. Nuk është e habitshme që mësuesit preferojnë statistika pedagogjike (përqindje kalueshmërie, keqkuptime tipike, përqindje të ushtrimeve të zgjidhura dhe materialeve të lexuara), të cilat janë të thjeshta për tu intrepretuar. Nga ana tjetër mësuesit mendojnë se statistika shumë e detajuar është e vështirë për tu analizuar dhe do shumë kohë për t'u interpretuar. Analiza statistikore e të dhënave edukuese, mund të na japë informacion të tillë si( Zorrilla et al. 2005, Heathcate & Dawson 2005, Wu & Leung 2002, Rahkila & Karjalainen 1999):

1. Sasia e materialit të cilin studentët duhet të lexojnë
2. Faqet e internetit me të vizituara. Orari, numri i hyrje/daljeve nga këto
3. Tregues statistikor mbi ndërveprimin e studentetit nëpër forume. Mesatarja totale e kontributeve në forumet e diskutimi. Numri i postimeve në raport me numrin e përgjigjeve
4. Sasia e ndërveprimeve nxënës- nxënës krahasuar më sasinë e ndërveprimeve nxënës mësues
5. Sasi e materialit të cilin studentët duhet të lexojnë. Rradha e leximit të kapitujve
6. Sjella e studentit dhe shpërndarja e kohës së tij.
7. Etj

Vizualizimi i informacionit përdor teknika grafike që ndihmojnë njerëzit të kuptojnë dhe të analizojnë të dhënat. Paraqitja vizuale dhe teknikat e ndërveprimit përfitojnë nga aftësia e syrit të njeriut që t'i transmetojë mendjes sasi informacioni të mëdha duke bërë të mundur për përdoruesit të shohin, eksplorojnë dhe kuptojnë sasi të mëdha informacioni, të gjitha përnjëherësh. Ekzistojnë studime të ndryshme të orientuara drejt vizualizimit të të dhënave arsimore për shumë probleme në fushën kërkimore dhe zhvilluese të edukimit(*Romeo et al (2010)):*

- *Teorite shkencore dhe evolimi i sistemeve.* Data mining në edukim mund të kontribojë në zhvillimin e sistemeve e-learning dhe zhvillimin dhe testimin e teorive shkencore mbi mësimdhënien e përmirësuar teknologjikisht. Mund të përdoren analiza eksploruese për të identifikuar modele të rregullta mbi të dhënat, për shembull strategji problem-zgjidhëse të studentëve apo modele të bashkëpunimit të sukseshëm.
- *Përcaktimi i parametrave të modeleve të studentëve.* Një *model studentor* është një strukturë të dhënash, e cila zakonisht përdoret në Sistemet Inteligjente të Tutorëve(ITS – Intelligent Tutoring Systems), e cila gjatë gjithë kohës ruan informacion rreth karakteristikave të rëndësishme të studentëve (për shembull sesi aftësitë e studentëve mund të zhvillohen dhe përmirësohen me praktikë) nëpërmjet veprimeve të përdoruesve (për shembull nga përgjigjet e testeve të studentëve). Modelet e studentëve bëjnë të mundur që sistemet të përshtaten sipas kërkesave të studentëve dhe situatave (për shembull zgjedhja e një ushtrimi me një nivel vështirësie të caktuar).
- *Modele bazë specifik.*Modelet studentore janë zakonisht të ndërtuar mbi një model specifik, për shembull një model i cili përshkruan fushën e veprimeve duke u bazuar në konceptet, aftësitë, mjetet mësimore dhe ndërlidhjet. Që një model studentor të parashikojë në mënyrë të saktë njohuritë, është e rëndësishme që modeli bazë të reflektojë aftësitë njohëse, si për shembull njohuritë dhe aftësia për të zgjidhur probleme. Data mining në edukim mund të ndihmojë në dizenjimin, përmirësimin dhe zhvillimin e një modeli specifik.
- *Krijimi i raporteve/njoftimeve per instrukturorët, studentët, etj.* Zakonisht mësuesit në skenarët e mësimit të bazuar me kompjuter e kanë të vështirë që të monitorojnë progresin e studentëve të tyre dhe problemeve e jetës reale, kjo për mungesë te bashkëbisedimit ballë për ballë, diferencave në kohë dhe në vend, etj. Data mining në edukim mund të përdoret për të ndërtuar mjete që mund të përdoren nga mësuesit. Këto mjete ofrojnë raporte statistikore apo dhe mënyra të tjera për të paraqitur një informacion në një mënyrë të lehtë. Për shembull, janë ndërtuar mjete të cilat sigurojnë statistika mbi përdorimin e sistemit, performanca e studentit dhe pjesëmarrja në aktivitet bashkëpunuese, etj. Informacione të ngjashme mund t'ju sigurohen dhe studentëve.
- *Rekomandimi i burimeve dhe aktiviteteve.* Data mining në edukim mund të përdoret për të siguruar suport për studentët/mësuesit duke përcaktuar burime dhe aktivitete të cilat janë të përshtatshme dhe përputhen me nevojat, interesat dhe preferencat e një studenti.

### *2.2.2 Dhënia e informacionit për instruktorët*

Objektivi është që të japë informacion për të ndihmuar mësuesit e lëndëve në procesin e marrjes së vendimeve (si të përmirësohet nxënia e studentëve, të organizohen burimet e kursit në mënyrë më efikase etj) dhe t'i mundësojë atyre të marrin vendimet e duhura në kohën e duhur. Është e rëndësishme të theksohet që kjo detyrë është e ndryshme nga analizimi dhe vizualizimi i të dhënave, e cila jep thjesht informacionin bazë për vizualizimin e të dhënave. Shumë teknika të ndryshme data mining janë përdorur për këtë detyrë, megjithëse nxjerrja e informacionit me grupim rregullash është teknika që është përdorur më tepër. Nxjerrja e informacionit me grupim rregullash vë në dukje lidhjet interesante ndërmjet variablave në databaza të mëdha dhe i paraqet ato në formë rregullash të forta sipas shkallës së interesit që ato shfaqin.

Ekstraktimi i informacionit me bashkim rregullash është përdorur(Merceron & Yacef 2004, Yu et al. 2008, Ramli 2005, Zheng et al. 2008 ):

- për të analizuar të dhënat e nxënies dhe të kuptohet nëse studentët përdorin materialet, dhe në qoftë se përdorimi i këtyre materialeve ka një ndikim pozitiv mbi notat e tyre;
- për të përcaktuar marrëdhënien mes çdo modeli sjelljeje – nxënien në mënyre të tillë që mësuesi të promovojë bashkëpunimin në mësim në web;
- për të nxjerrë informacion, i cili u jepet mësuesve për të analizuar, rafinuar dhe riorganizuar materialet e mësimit dhe testet në mjediset e nxënies përshtatëse;
- për të optimizuar përmbajtjen e portalit e-learning të universitetit;
- për të zbuluar lidhje interesante ndërmjet atributeve të studentëve, atë të problemeve dhe strategjive të zgjidhjes në mënyrë që të përmirësohen sistemet e edukimit online si për studentët dhe për mësuesit;
- për të identifikuar modele nxënieje interesantë dhe të papritura të cilat mund t'u japin informacion mësuesve se si të organizojnë më mirë strukturën e shpjegimit të tyre;
- për të përmirësuar një dizenjim kursi të përshtatur në mënyrë që të nxirren rekomandime si të përmirësohet struktura e kursit dhe përmbajtja e tij;
- për të gjetur marrëdhënie interesante mes karakteristikave, strategjisë së zgjidhjeve të përshtatur nga studentët nga pikëpamja e një sistemi mësimi të bazuar në web për celularë;
- të ndihmojë mësuesit të zbulojnë marrëdhënie të mira ose të keqija mes përdorimit të burimeve të bazuara në web dhe nxënies së studentëve; për të nxjerrë informacion mbi regjistrimet e studentëve në universitet; për të ndihmuar organizatat të përcakojnë stilet e mendimit të nxënësve dhe efektshmërinë e strukturës së web-it ; për të vlerësuar dizenjimin e website-it edukues dhe për të nxjerrë përgjigje duke analizuar sondazhet.

Një fushë kyçe, ka qenë në kërkimin e provave empirike për të përmirësuar dhe zgjeruar teoritë arsimore dhe fenomene të mirënjohura arsimore, drejt të kuptuarit më të thellë të faktorëve kryesorë që ndikojnë në të mësuarit, shpesh me synimin për të hartuar sisteme më të mira për mësim. Për shembull, Gong Y. Rai et al. (2009) hetuan ndikimin e vetë-disiplinës për të mësuar dhe zbuluan se, për aq kohë sa është në korelacion me njohuri të larta hyrëse dhe me pak gabime, ndikimi aktual në të mësuarit ishte i pakët. Perera et al. (2009) përdori teorinë e Big 5 për punën në grup, si një teori drejtuese për të kërkuar për modele të suksesshme të ndërveprimit brenda grupeve të studentëve. Madhyastha dhe Tanimoto (2009) hetuan lidhjen mes konsistencës dhe performancës së studentëve me qëllim që të sigurojnë udhëzues me udhëzime drejtuese, duke e bazuar punën e tyre në teorinë paraprake mbi implikimet e konsistencës në sjelljen e studentit.

### *2.2.3 Rekomandime për studentët*

Objektivi është që të bëhen rekomandime direkt te studentët, që kanë të bëjnë me aktivitete të përsonalizuara, linke për t'u vizituar, detyrën ose problemin e rradhës për t'u zgjidhur etj., dhe duhet të jetë në gjendje për të përshtatur materialin për t'u studiuar, ndëfaqet dhe sekuencat për çdo student. Teknika të ndryshme data mining janë përdorur për këtë detyrë, por më të përdorshmet janë nxjerrja e informacionit me bashkim rregullash, grupimi dhe nxjerrja e informacionit me modele sekuenciale. Ky i

fundit përpiqet të gjejë marrëdhënie mes ngjarjes së eventeve sekuenciale, për të parë në qoftë se ka ndonjë rradhë specifike në këto ngjarje.

Nxjerrja e informacionit me model sekuencial është zhvilluar (Zorrilla et al 2005, Zheng et al 2008):

- për të personalizuar rekomandimet mbi materialin për t'u studiuar, duke u bazuar në stilin e të nxënurit dhe në zakonet e përdorimit të web-it;
- për të identifikuar sekuenca domethënëse aktiviteti, të cilat tregojnë problem/sukses në mënyrë që t'i ndihmojnë studentët duke zbuluar problemet herët;
- për të rekomanduar linket më të përshtatshme për studentët, që ti vizitojnë në një sistem edukimi të përshtatshëm; për të përfshirë konceptin e itinerarit të rekomanduar në standartin SCORM (Sharable Content Object Reference Model) duke kombinuar profesionalizmin e mësuesve më eksperiencën e nxënësve;
- për të zgjedhur objekte studimi të ndryshëm për studentë të ndryshëm, duke u bazuar në profilet e nxënësve dhe në lidhjet e brendshme mes koncepteve; për përsonalizimin e pemëve të aktivitetit në një ambjent SCORM;

Nxerrja e informacionit me bashkim rregullash përdoret (Zheng et al. 2008, Gopalakrishnan et al 2014):

- për të rekomanduar aktivitete nxënie online ose materiale nxënie në sistemet e-learning;
- duke u bërë studentëve sugjerime të përsonalizaura të nxjerra nga analizat e rezultateve të testimeve dhe konceptet e përmendura në testim;
- për të ndërtuar një sistem të përsonalizuar për rekomandimin e materialeve që ndihmojnë studentët në gjetjen e materialeve të studimit;
- për të rekomanduar lëndët duke marrë parasysh lëndët më zgjedhje optimale;
- për të dizenjuar një sistem rekomandimi materialesh duke u bazuar në veprime nxënëse të studentëve të mëparshëm.

Grupimi është zhvilluar për (Romeo et al 2009, Dekson & Jaichandran 2015):

- të ndërtuar një model rekomandimi për studentet në situata të ngjashme në të ardhmen;
- për grupimin e dokumenteve në web duke përdurur metoda grupimi për të përsonalizuar e-learning duke u bazuar në bashkësitë e materialeve me frekuence përdorimi më të lartë;
- për të dhënë rekomandime materiali kursi të përsonalizuar duke u bazuar në aftësitë e nxënësit dhe duke u rekomanduar studentëve ato burime që ata nuk i kanë vizituar akoma po që do u hynin më shumë në punë.

Teknika të tjera data mining të përdorura janë:

- *rrjetat neurale dhe pemët e vendimit* ato japin një suport për nxënie të përsonalizuar dhe të përshtatshme;
- *rregullat e prodhimit* ndihmojnë studentët për të marrë vendime duke u bazuar në veprimtarinë e tyre akademike;
- *analiza e pemës së vendimit* rekomandon sekuencën optimale të nxënies për të thjeshtuar procesin e nxënies së studentëve dhe për të maksimizuar rezultatet e tyre;

- *teknologjia më agjentë inteligjent* dhe kurset e bazuara në SCORM ndërtojnë një sistem rekomandues të bazuar në agjent për planifikimin e sekuences së mësimeve në mësimin e bazuar në web; etj.

### *2.2.4 Parashikimi i përformances se studentëve*

Qëllimi i parashikimit është që të vlerësojë vlerën e panjohur të një variabli që përshkruan studentin. Në edukim vlerat e parashikuara zakonisht janë përformanca, njohuritë, pikët ose notat. Kjo vlerë mund të jetë numerike/e vazhduar (detyrë regresioni- detyrë në të cilën dataset-i i ka vlerat e mundshme të njohura) ose kategorike/të veçanta (detyrë klasifikimi- detyrë e cila fillon më një dataset ku përcaktimet e klasës janë të njohura). Analiza e regresionit bën lidhjen ndërmjet një variabli të varur me 1 ose më shumë variabla të pavarur. Klasifikimi është një proçedurë në të cilën elemente të veçantë janë vendosur në grupe në bazë të informancionit sasior që bazohet në një ose më shumë karakteristika të trashëguara të elementit dhe në një grup elementësh të trajtuar më parë.

Parashikimi i përformancës së një studenti është një nga aplikimet më të famshme të data mining në edukim, dhe në të janë aplikuar teknika dhe modele të ndryshme ( si, rrjetat neurale, rrjetat Baeziane, sistemet e bazuara në rregulla, analiza e regresionit dhe korrelacionit).

Një krahasim i metodave të bazuara mbi njohurinë e makinave është bërë për të parashikuar suksesin në një kurs/lëndë (në qoftë se është kaluar apo ka ngelur) në Sistemin e Mësimit Inteligjent. Krahasime të tjera të metodave të ndryshme të data mining janë bërë për të klasifikuar studentët(parashikuar notat e studentëve) të bazuar në përdorimin e të dhënave nga sistemi mësimor; për të parashikuar përformancën e studentëve (notën përfundimtare) bazuar në karakteristikat e nxjerra nga të dhënat e logjeve dhe për të parashikuar përformancën akademike të studentëve në Universitete (Çomo et al 2015, Cortez & Silva 2008 ).

Tipe të ndryshme të modeleve të data mining ( si për shembull pemët e vendimit, rrjetat baezine, rrjetat neruale, etj) janë përdorur:

- për të parashikuar notat përfundimtare të studentëve (duke përdorur rrjetat neurale të shumimit mbrapsht (back-propagation dhe feed forward)) ;
- për të parashikuar përformancën nga pikët e testeve (duke përdorur back propagation dhe counter propagation);
- për të parashikuar notat e studentëve (kalon ose ngel) nga logjet e e sistemit mësimor
- për të parashikuar përformancën e mundshme të kandidatëve që po merren në shqyrtim për t'u pranuar në universitet (duke përdorur topologjinë e percetimit me shumë shtresa).
- për të parashikuar mesataren që mendohet të ketë studenti duke u bazuar në njohuritë e aplikantit në kohën e pranimit;
- për të parashikuar performancën e nxënësit të bazuar në bagazhin e njohurive të formuar (duke përdorur rregullat kyçe te formimit vlerësues );
- për të parashikuar notat përfundimtare bazuar në karakteristikat e nxjerra nga të dhënat e logeve në një sistem edukues të bazuar (duke përdorur algoritme gjenetike për të gjetur rregulla bashkimi);
- për të parashikuar notat e studentëve në një universitet publik (duke përdorur pemët model, rrjetat neurale, regresionin linear, regresionin linear me peshë lokale dhe makinat më vektorë suporti);

- për të parashikuar kënaqësinë e studentëve të universiteteve (duke përdorur analizën e pemëve të vendimit dhe regresionin);
- për të parashikuar probabilitetin që një student do të ketë sukses në gjetjen e një pune në përfundim të universitetit

### *2.2.5 Modelimi i studentëvë*

Objektivi i modelimit të studentëve është që të zhvillojë modele njohës të njerëzve (përdorues/student), duke përfshirë një modelim të aftësive të tyre, dhe njohurive të deklaruara. Data mining është aplikuar që të marrë parasysh automatikisht karakteristikat e studentëve (motivimin, kënaqësinë, stilin e të mësuarit, statusin, etj.) dhe sjelljen e të mësuarit në mënyrë që të automatizojë ndërtimin e modeleve të studentëve. Teknika dhe algoritme të ndryshëm të data mining janë përdorur për këtë punë (kryesisht rrjetat Baeziane).

Algoritme të ndryshëm data mining (Naiv Bayes, rrjeti Bayes, makinat e bazuara në vektor, regresioni logjik dhe pemët e vendimit) janë krahasuar për të identifikuar modelet mendore të studentëve në sistemin e tutorimit inteligjent. Mësimi i bazuar mbi makinë i pakontrolluar (grupimi) dhe i kontrolluar (klasifikimi) janë propozuar për të ulur koston e zhvillimit në ndërtimin e modeleve të përdoruesve dhe për të thjeshtuar transferimin në mjediset e mësimit inteligjent (Baker et al. 2008, Chang et al. 2006, Johsson et al. 2005, Gong et al 2005):

- *Grupimi dhe klasifikimi* i variablave të mësimit janë përdorur për të matur motivimin e përdoruesit online.
- *Rrjetat Baeziane* janë përdorur për të bërë parashikime për njohuritë e studentëvë, për shembull, mundësia që një student ka që të zotërojë një aftësi në një kohë të caktuar; për të bazuar stilet e të mësuarit të studentëve në një sistem edukativ të bazuar në web; për të parashikuar kur një student do t'i përgjigjet saktë një problemi; për të modeluar ndryshimin e njohurive të studentëve gjatë marrjes së njohurive shtesë në ITS; të nxjerrë përfundime për variabla mësimi, të pavëzhgueshme nga sjellja e kërkimit për ndihmë të sudentëve në një sistem tutorimi të bazuar në web dhe për identifikimin e njohurive në mënyrë që të identifikojë ndikimin e vetëdisiplinës në njohuritë dhe të mësuarin e studentëve.
- *Modeli sekuencial i nxjerrjes së informacionit* është përdorur për të fituar automatikisht aftësinë për të ndërtuar modelet e studentëve; për të identifikuar karakteristika kuptimplota të studentëve dhe për të update-uar një model përdoruesi që reflekton njohuritë e fituara së fundmi, dhe për të parashikuar hapat e ndërmjetme të trurit të studentëve gjatë sekuencave të veprimeve të ruajtura.
- *Algoritmat e rregullave të unifikimit* janë aplikuar për nxjerrjen e të dhënave të përsonalitetit, të bazuar në modelet e edukimit të bazuar në web, në mënyrë që të nxjerrin karakteristikat e përsonalitetit të nxënësve dhe për modelimin e studentëve në sistemin e tutorimit inteligjent.
- *Teknika dhe modele të tjera* data mining janë përdorur gjithashtu për modelimin e studentëve.

Metodat e data mining në arsim, kanë bërë të mundur që studiuesit të modelojnë në kohë reale një gamë më të gjerë atributesh të lidhura me nxënësit, duke përfshirë ndërtime të niveleve më të larta nga çfarë kishte qenë e mundur më parë. Për shembull, në vitet e fundit, studiuesit kanë përdorur metodat EDM për të konkluduar nëse një student është duke luajtur me sistemin (Baker et al. 2008), duke përjetuar një vetë-efikasitet të varfër (McQuiggan etal. 2008), pa detyra (Baker 2007). Studiuesit

kanë qenë gjithashtu në gjendje për të zgjeruar modelimin e studentëve edhe përtej programeve arsimore, në drejtim të gjetjes së faktorëve që mund të parashikojnë dështimin e nxënësve në shkolla ( Romero et al. 2008; Superby et al. 2006).

### *2.2.6 Detyra të tjera*

- **Zbulimi i sjelljeve të padëshiruara të studentëve**

Objektivi i zbulimit të sjellejeve të padëshiruara të studentëve është që të zbulojë ata studentë që kanë ndonjë problem ose sjellje të pazakontë, të tilla si: veprime të palogjikshme, motivim i ulët, luajtje lojërash, keqpërdorim, braktisje shkolle etj. Shumë teknika data mining (kryesisht klasifikimi dhe grupimi) janë përdorur për të zbuluar këta tipe studentësh në mënyrë që t'u ofrohet ndihmë në kohën e duhur.

- Disa prej algoritmeve të klasifikimit që janë përdorur për të zbuluar sjelljen e studentëve problematikë janë pemë vendimi, rrjeta neurale, Naiv Bayes, regresioni logjik dhe makina të bazuara në vektor;
- për të parashikuar/parandaluar ndërprerjen e studimeve nga studentët përdoren rrjeta neurale, makina të bazuara në njohuri
- pemë vendimi, klasifikues Baezian, modele logjike, për të parashikuar lënien e universitetit nga studentët pas vitit të parë;
- algoritëm pemë vendimi C4.5, pemës së vendimit J48 dhe algoritmin e grupimit e më të largëtave më përpara (furthest first) për të parashikuar, kuptuar dhe parandaluar ngeliet nëpër provime për studentët universitarë.

- **Grupimi i studentëve**

Qëllimi këtu është që të krijohen grupe studentësh bazuar në tipare të caktuara, karakteristika personale, etj. Grupet e përftuara mund të përdoren nga instruktori/zhvilluesi i website-it për të zhvilluar një sistem nxënës të personalizuar, për të promovuar nxënie efektive në grup, për të dhënë përmbajtje adaptive etj. Teknika data mining të përdorura për këtë detyrë janë klasifikimi për nxënien e kontrolluar dhe grupimi për nxënien e pakontrolluar. Analiza e grupeve është ndarja e një bashkësie vëzhgimesh në nëngrupe në mënyrë të tillë që vëzhgimet në të njëjtin nëngrup të kenë disa karakterisitka të përbashkëta. Algoritme të ndryshme grupues janë përdorur për grupimin e studentëve, të tilla si:

- grupimi hierarkik aglomerativ, K-means dhe grupimi i bazuar në modele për të identifikuar studentët me aftësi të ngjashme;
- një algoritëm grupimi i paqartë për të gjetur grupe studentësh duke u bazuar në përsonalitetin e tyre dhe të dhënat mbi nxënien e tyre të marra nga kurset online;
- një algoritëm grupimi K-means për të grupuar në mënyrë efektive studentët që tregojnë aftësi të ngjashme (nota detyrash, nota provimesh dhe nota të nxënies online);
- një algoritëm grupimi K-means për të zbuluar tipare interesante që karakterizojnë punën e studentëve më të mirë dhe më të dobët;
- algoritëm grupimi gjenetik për të zgjidhur problemin e shpërndarjes së studentëve të rinj (ky algoritëm vendos studentët e rinj në klasa në mënyre të tillë që diferenca mes niveleve të nxënies në secilën klasë është në minimum dhe që numri i studentëve në secilën klasë të mos kalojë numrin maksimal që duhet të ketë çdo klasë).

Janë aplikuar algoritme të ndryshme klasifikimi për të grupuar studentët, të tilla si:

- analiza diskriminuese, rrjetat neurale, dhe pemët e vendimit për të klasifikuar studentët e universitetit në tre grupe (me rrezik ngelje të ulët, të mesëm dhe të lartë);
- pemë klasifikimi dhe regresioni për të krijuar një model më pemë vendimi për të ilustruar

### ➢ Analiza e rrjeteve sociale

Analiza e rrjeteve sociale ose analiza strukturore ka për qëllim të studiojë marrëdhëniet mes individëve në vend të studimit të karakteristikave individuale. Një rrjet social konsiderohet të jetë një grup njerëzish, një organizim personash, të cilët janë të lidhur në një marrëdhënie sociale të tillë si miqësia, marrëdhëniet bashkëpunuese ose shkëmbimi i informacionit. Teknika të ndryshme data mining janë përdorur për të nxjerrë informacion nga rrjetet sociale në ambjentet edukuese, por filtrimi bashkëpunues është ai më i përhapuri. Filtrimi bashkëpunues ose e thënë ndryshe filtrimi social është një metodë që bën parashikime automatike (filtrim), për interesat e një përdoruesi duke grupuar preferencat nga shumë përdorues të tjerë. Sistemet e filtrimit bashkëpunues mund të prodhojnë rekomandime duke llogaritur ngjashmëritë ndërmjet preferencave të studentëve, kështu që kjo detyrë lidhet më detyrën e rekomandimit të studentëve (të sqaruar në pikën F).

### ➢ Zhvillimi i hartave konceptuale

Objektivi i ndërtimit të hartave konceptuale është të ndihmojë instruktorët\edukatorët në procesin automatik të ndërtimit\zhvillimit të hartave të konceptit. Një hartë konceptuale është një graf, që tregon marrëdhënien mes koncepteve, dhe tregon strukturën hierarkike të njohurive. Disa teknika data mining (kryesisht rregulla shoqërimi dhe nxjerrje informacioni nga teksti) janë përdorur për të ndërtuar hartat konceptuale.

Nxjerrja e informacionit me rregulla shoqërimi është përdorur për të ndërtuar automatikisht harta konceptuale të bazuara në rekordet e testeve të bëra nga nxënësit; për të zhvilluar marrëdhëniet koncept-efekt për të diagnostikuar problemet në nxënien e studentëve dhe për diagnozë konceptuale të e-learning, përmes hartave konceptuale të ndërtuara automatikisht, që u japin mundësinë mësuesve të kapërcejnë kufirin e nxënies dhe konceptimet e gabuara të studentëve.

Nxjerrja e informacionit nga teksti është përdorur për të ndërtuar harta konceptuale nga artikujt akademikë në domainin e-learning; për të formuluar harta konceptuale nga bordet e diskutimit online duke përdorur ontologji të paqarta; për të gjetur marrëdhënie mes dokumenteve të tekstit dhe për të ndërtuar grafik indeksesh të dokumetit.

### ➢ Ndërtimi i materialeve të kursit

Qëllimi i ndërtimit të materialeve të kursit është që të ndihmojë instruktorët dhe zhvilluesit të ekzekutojnë procesin e ndërtimit/zhvillimit të ndërtimit dhe mësimit të materialeve të kursit automatikisht.Nga ana tjetër përpiqet dhe të ripërdorë\shkëmbejë burimet ekzistuese të mësimit mes përdoruesve dhe sistemeve të ndryshëm. Për

ndërtimin e materialeve të kursit janë përdorur teknika dhe modele të ndryshme Data Mining.

- **Planifikimi dhe caktimi i orareve**

Objektivi i planifikimit dhe ndërtimit të orarit është që të pasurojë procesin tradicional arsimor duke planifikuar kurse për të ardhmen, të ndihmojë studentët në zgjedhjen e kurseve të tyre, të planifikojë shpërndarjen e burimeve, të ndihmojë në procesin e pranimit dhe këshillimit, të zhvillojë kurrikulat etj. Teknika të ndryshme data mining përdoren për këtë detyrë (kryesisht rregulla shoqërimi).
Klasifikimi, kategorizimi, vlerësimi dhe vizualizimi janë përdorur në arsimin e mesëm për objektiva të ndryshme, të tilla si: planifikimi akademik, parashikimi i diplomimit të studentëve dhe krijimi i tipologjive të ndryshme të mësimit. Pemët e vendimit, analiza e linkeve dhe pyjet e vendimit janë përdorur në planifikimin e kurseve për të analizuar regjistrimet nëpër kurse të studentëvë.
Nxjerrja e informacionit më rregulla shoqërimi është përdorur për të dhënë impulse lidhur më kërkesën për kurse të reja, kjo për të ndihmuar në zhvillimin e kurseve të reja për bachelor dhe master. Rishikimi i kurrikulave është bërë duke përdorur nxjerrje informacioni me rregulla bashkimi në mënyrë që të identifikoheshin dhe të kuptoheshin, në qoftë se revizionet e kurrikulave mund të ndihmojnë studentët për pranimin e tyre në universitet. Një mjet vendimi (I bazuar në nxjerrjen e informacionit me rregulla shoqërimi) është ndërtuar për të ndihmuar marrjen e vendimeve që përmirësojnë cilësinë e shërbimit të ofruar nga universitetet, duke u bazuar në përqindjen e kalimit dhe ngelies së studentëve. Nxjerrja e informacionit me rregulla shoqërimi dhe algoritmet gjenetikë janë përdorur për të ndërtuar një sistem automatik oraresh për kurset, për të prodhuar një orar kursesh që plotëson më së miri nevojat e studentëve dhe të mësuesve.

- **Mbështetja pedagogjike**

Aplikimit tjetër i metodave data mining në arsim, ka qenë në mbështetjen pedagogjike (si në programet mësimore ashtu dhe në fusha të tjera, të tilla si sjelljet e të mësuarit bashkëpunues), drejt zbulimit të llojeve të përkrahjes pedagogjike që janë më të efektshme, qoftë në përgjithësi apo për grupe të ndryshme të nxënësve ose në situata të ndryshme. Një metodë e njohur për studimin e mbështetjes pedagogjike është dekompozimi (zbërthimi) i të mësuarit (Beck & Mostow 2008). Dekompozimi i të mësuarit përshtat kthesa eksponenciale të të mësuarit tek performanca e të dhënave, duke lidhur një sukses të mëvonshëm të studentit me sasinë e secilit lloj të përkrahjes pedagogjike që studenti ka marrë deri në atë pikë. Peshat relative për secilin lloj të mbështetjes pedagogjike, në modelin best-fit (përshtatja më e mirë), mund të përdoren për të treguar efikasitetin relativ të çdo lloji përkrahje që motivon të mësuarit.

## 2.3 Rëndësia e përzgjedhjes së variablave

Për parashikimin e rezultateve akademike të studentëve është e nevojshme të shqyrtohen shumë parametra. Për të bërë një parashikim efektiv të performancës së studentit nevojiten modelet e parashikimit që përfshijnë të gjitha variablat personale, sociale, psikologjike dhe variabla të tjera rrethuese. Gjithashtu një rol mjaft të

rëndësishëm do të luajnë edhe njohuritë e studentëve në lidhje me çështjen në fjalë, aftësia për të trajtuar një pyetje, aftësia për të përfunduar provimin në kohë etj.

### *2.3.1 Variablat demografik*

Variablat demografikë zakonisht përfshihen gjithmonë në studimin e suksesit të studentëve për një sërë arsyesh. Ndryshe nga shumë karakteristika të tjera subjektive (psh psikologjike dhe kulturore), karakteristikat demografike maten me lehtësi dhe dokumentohen në mënyrë të qëndrueshme. Ato lehtësojnë ndarjen e subjekteve në grupe të dallueshme me identitete të ndara. Edhe pse disa studime kanë raportuar se të dhënat demografike kanë një ndikim minimal në studim, raportet e përhapura gjerësisht kanë pohuar se rëndësia e këtyre variablave është aq e madhe saqë përjashtimi i tyre do të kërcënonte vlefshmërinë në çdo studim.

- **Mosha** është një nga tre përmasat për përcaktimin e kohëzgjatjes së studimeve të studentit. Mosha e ndërthurur edhe me faktorët e tjerë shoqërorë si për shembull familja, luajnë një rol mjaft të rendësishëm. Me kalimin e moshës lind edhe konflikti ndërmjet obligimeve në familje dhe atyre në punë, dhe për këtë arsye shpesh kalohet në ndjekjen e shkollës part time ose braktisjen e saj. Koha mesatare e përfundimit të ciklit të studimeve nga fillimi i vitit shkollor deri në marrjen e diplomës së nivelit të parë zakonisht varion deri në 5 vjet. Vështirësia dhe kompleksiteti i përfundimit të ciklit të parë të studimeve në një moshë më të madhe është se shumica prej tyre e bëjnë këtë të motivuar për një rritje në profesion apo ndërthurje të profesionit aktual me studimin e ri.

- **Gjinia**. Gjatë 20 viteve të fundit një ndërveprim kompleks i disa faktorëve legjislativ, shoqëror dhe institucional, kanë zgjeruar mundësitë e edukimit dhe punësimit për femrat. Nga studimet e kryera është vërtetuar se përqindjet e studentëve meshkuj që përfundojnë shkollën e lartë janë pothuajse të barabarta me përqindjen e femrave. Megjithatë në rastin e femrave mund të themi se janë një sërë faktorësh që ndikojnë në performancën e tyre ndryshe nga meshkujt. Obligimet familjare të gjinisë femërore ndryshojnë nga ato të gjinisë mashkullore dhe në përballje me to shumë prej tyre braktisin mësimin. Në studimet e kryera pohohet se për këtë arsye femrat braktisin studimet më shpesh se meshkujt. Prandaj gjinia shihet si një element i rëndësishëm në përcaktimin e suksesit të studentit në studimet universitare.

- **Qyteti/Fshati** nga vijnë. Ky është një tjetër element i rëndësishëm në parashikimin e performancës së studentëve. Studentët që vijnë nga fshati janë më të prirur të kenë vështirësi në studime. Kjo vjen si pasojë e disniveleve që ka në mënyrat e mësimdhënies dhe kontrollit të njohurive në fshat dhe në qytet. Faza e përshtatjes është e vështirë dhe kërkon një hapërsirë të caktuar kohore gjë që sjell rezultate të ulëta të studentëve në provime, dhe si pasojë sjell mos kalimin e studentit në vitin pasardhës.

### *2.3.2 Faktorët socio-ekonomik*

Shumë nga variablat e lidhur me suksesin e studentit kanë të bëjnë me gjendjen sociale dhe financiare. Faktorë të tillë si të ardhurat familjare, numri i anëtarëve të familjes, apo niveli arsimor dhe punësimi i prindërve luajnë një rol mjaft të rëndësishëm në ecurinë e suksesshme të studentit në shkollë ose në dështimin e tij.

Studentët me probleme familjare, qoftë këto ekonomike apo shoqërore janë më të prirurit për të braktisur mësimin.

- **Të ardhurat familjare.** Pjesë thelbësore e ciklit të studimeve të studentit janë të ardhurat mbi të cilat ai nis studimet e tij. Studentët që kanë vështirësi ekonomike janë të prirur të fillojnë punë njëkohësisht me ciklin e studimeve dhe kjo bën që ata të humbin interesin për mësimin, si pasojë nuk arrijnë të kalojnë në provime. Sipas studimeve studentët me të ardhura të konsoliduara janë më të prirur të kenë rezultate më të mira në mësime.

- **Profesioni i prindërve**. Profesioni i prindërve lidhet drejtpërdrejtë me nivelin e tyre arsimor. Ky faktor ndikon jo vetëm në të ardhurat ekonomike mbi të cilat mbështetet studenti por gjithashtu edhe në nivelin edukues dhe kulturor që prindërit i ofrojnë. Niveli arsimor dhe kulturor i prindërve i vjen gjithmonë në ndihmë fëmijëve dhe sipas studimeve të kryera janë pikërisht këta fëmijë që kanë një mbështetje dhe nxitje të tillë nga prindërit që kanë rezultate në studime.

- **Numri i anëtarëve të familjes**. Kryerja e studimeve të larta kërkon punë dhe përqëndrim nga ana e studentit. Një numër i madh anëtarësh në familjë shpesh bëhet shkak jo vetëm për probleme ekonomike që vështirësojnë studimet por gjithashtu edhe për mungesën e kushteve të duhura për studim. Gjatë studimit studenti ka nevojë për ambjentin e duhur si dhe për qetësinë e nevojshme për t'u përqendruar në studime. Zhurmat apo shpërqëndrimet e çdo lloj natyre sjellin rezultate të ulëta në punën e studentit

### *2.3.3 Eksperienca e mëparshme*

**Mesatarja e shkollës së mesme.** Rezultatet e studentit në shkollën e mesme luajnë një rol kyç në parashikimi e performancës së studentit. Nota mestare me të cilën studenti është vlerësuar në shkollën e mesme është një paraprirje për suksesin e tij në arsimin e lartë. Për këtë arsye ky variabël është konsideruar i rëndësishëm dhe kryesor për zhvillimin e studimit

# 3. METODOLOGJIA E STUDIMIT

*Në këtë kapitull, shpjegohet metodologjia e kërkimit që është përdorur për të arritur objektivin e caktuar dhe për t'iu përgjigjur pyetjes kryesore të studimit. Kapitulli fillon me dizenjimin e studimit; tregon qëllimin dhe strategjinë e studimit. Më pas ndiqet nga proçesi i studimit që është përdorur në këtë punim. Të gjitha metodat e aplikuara në secilën fazë të modelimit: mbledhje e të dhënave, përpunimi i të dhënave, analiza e tyre, janë të gjitha të deklaruara në këtë kapitull.*

## 3.1 Plani i studimit

Plani i studimit është një strukturë për përmbushjen e studimit dhe është themeli për kryerjen e punimit. Në përputhje me këtë, është një plan që drejton pjesët e mbledhjes së të dhënave dhe analizave të studimit. Ajo i identifikon të dhënat nga lloji i informacionit që duhet të mblidhet, burimi i të dhënave, dhe proçesi i mbledhjes së të dhënave.

> *"Një plan i mirë studimi do të sigurojë që informacioni i mbledhur të jetë konsistent me objektiat e studimit dhe që proçedurat në lidhje me mbledhjen e të dhënave të jënë të sakta dhe efiçente" (Javaheri, 2007).*

Plani kërkimor i këtij studimi jepet në Figurën 3.1 dhe shpjegohet ne mënyrë më të detajuar  më poshtë.

### *3.1.1 Qëllimi i studimit*

Studimet mund të kategorizohen duke iu referuar qëllimit të tyre. Për shembull, studimet më së shumti klasifikohen si kërkimore dhe konkluzive (përfundimtare). Studimet përfundimtare mund të kategorizohen si shkakore (shpjeguese) dhe përshkruese. Studimet shkakore ndërtojnë variabla shkakore ndërmjet variablave. Në këto studime, theksi vihet mbi analizën e situatës ose dilemës , për të qartësuar lidhjet me variablave. Qëllimi i studimeve përshkruese është të përshkruajnë karakteristka ose funksione. "Përshkrimi është për të bërë të kuptueshme gjërat e ndërlikuara duke i reduktuar ato në pjesët e tyre përbërëse". (Saunders et al., 2000)

Ndërsa studimet kërkimore, siç tregon dhe vetë emri, synojnë të eksplorojnë ose kërkojnë nëpër një problem ose situatë, në mënyrë që të sigurojnë depërtim dhe të kuptuarit e çështjes (Malhotra and Peterson, 2006).

Data Mining shpjegon proçedurën e zbulimit të njohurive ose informacion nga bazat e të dhënave. Data mining është një mjet shumë i vlefshëm, një qasje që ndërthur kërkimin dhe zbulimin, me studime statistikore konfirmuese, në mënyrë që të zbulojë dhe vërtetojë lidhjet.Megjithatë ne mund t'i ndajme detyrat e data mining në dy klasa:

- Data mining *parashikues* që përfshin përdorimin e variablave ose fushave në bashkësinë e të dhënave për të parashikuar vlera të panjohuara të variablave të tjera me interes.
- Data mining *përshkrues* që përqëndrohet në zbulimin e modeleve të interpretueshme nga njerëzit dhe që karakterizojnë veçoritë e përgjithshme të të dhënave në bashkësinë ë të dhënave.

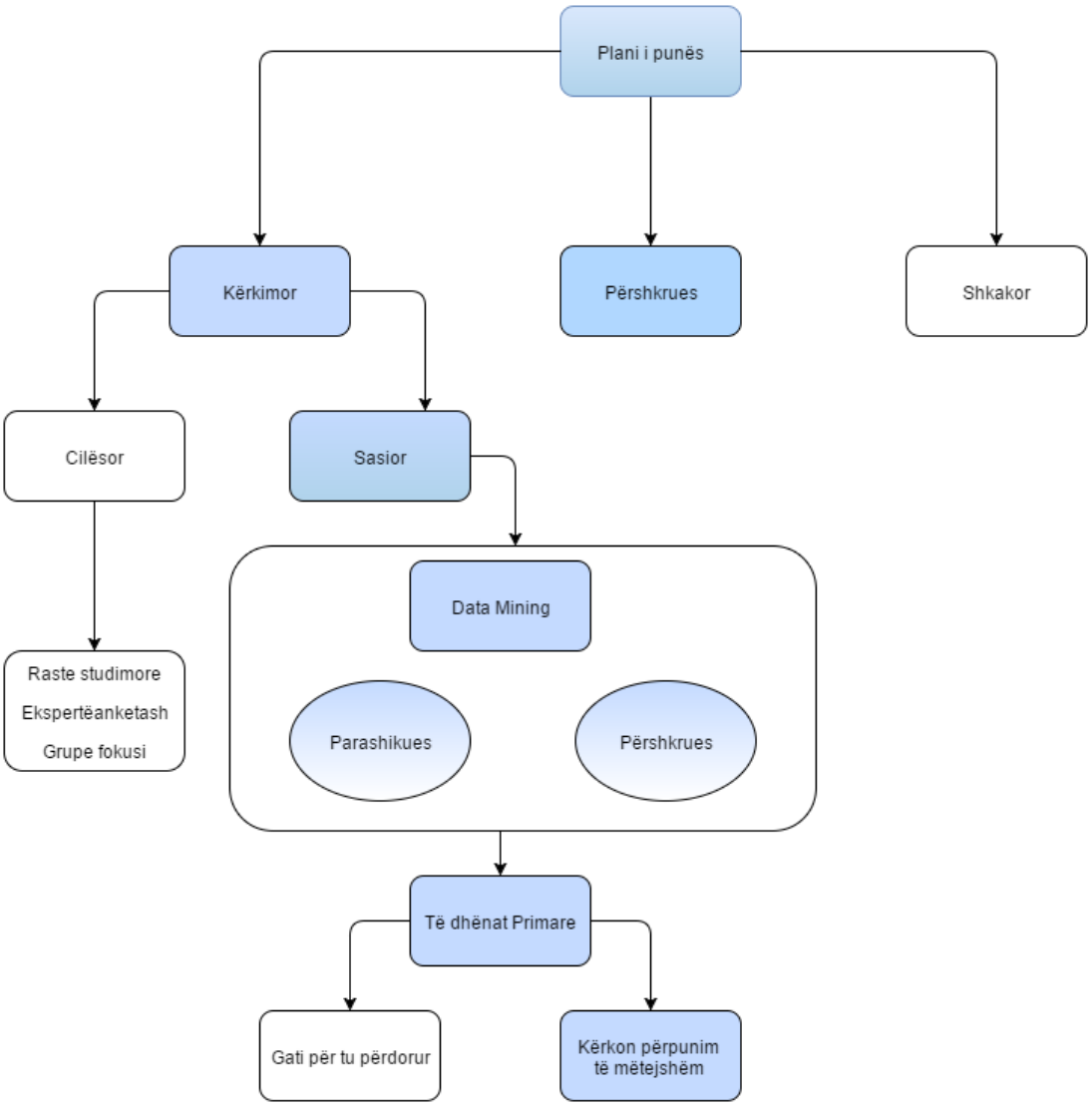



**Figura 3.1 Plani kërkimor i studimit**

Përshkrimi mundet të ndihmojë gjithashtu studiuesit në kërkimin e tyre si në rastin e studimve që zhvillohen në këtë punim. Qëllimi i këtij ponimi është kryesisht kërkimor, por me disa aspekte përshkruese që na ndihmojnë gjatë fazës së parashikimit.

### *3.1.2 Qasjet e studimit*

Qëllimet e studimit mund të arrihen nëpërmjet metodave të ndryshme: cilësore ose sasiore. Studimi cilësor ofron depërtime dhe detaje të problemit në studim. Karakteristikë të studimit cilësor është që ai bazohet në më së shumti në përshkrimin e vetë studiuesit, emocione dhe reagime. Nga ana tjetër, kërkimi sasior kërkon të përcaktojë sasinë e të dhënave. Ai kërkon fakte përfundimtare nëpërmjet zbatimit të disa analizave statistikore. Kjo do të thotë që, në qasjen sasiore, rezultatet bazohen në numra dhe statistika që paraqiten më pas në figura, ndërsa në atë qasjet cilësore, rezultate bazohen në shpjegimin e një ngjarje nëpërmjet fjalëve. (Sounders et al., 2000)

Koncepti i data mining i lejon ata që marrin vendime që të mbështeten në studimet kërkimore sasiore. Duke qenë se ky studim mbështetet në data mining, kërkimi mund të konsiderohet si një *kërkim sasior*.

### *3.1.3 Strategjia e studimit*

Strategjia e studimit, është një teknikë e përcaktuar sipas së cilës studiuesi dëshiron të mledhë të dhënat. Bazuar në karakterin e pyetjes së studimit, studiuesi mund të vendosë ndërmjet një eksperimenti, një pyetësori, analizën e të dhënave dytësore dhe një rasti studimi.

Aktualisht, janë dy klasa të dhënash të zbatuara në studime: të dhëna primare dhe sekondare (Zikmund, 2003). Një studiues gjeneron të dhëna primare për qëllimin e veçantë të trajtimit të problemit që ka në shqyrtim. Ndërsa të dhënat dytësore, janë të mbledhura paraprakisht për objektiva të ndryshme nga ato që po shqyrtohen në studim. Strategjia e këtij studimi është analizë e të dhënave primare të mbledhur nëpërmjet pyetësorëve ose sistemit të menaxhimit të studentëve.

Në përputhje me të gjitha këto, qëllimi kryesor i këtij studimi është kërkimor me disa aspekte përshkruese dhe qasja e tij është sasiore (data mining). *Strategjia e studimit është analiza e të dhënave primare.*

## 3.2 Proçesi i studimit

Në këtë punim, synohet të gjendet teknika më e mirë e klasifikimit që i përshtatet:

- Rasti I - Parashikimin e mundësive të punësimit të studentëve të diplomuar në vendin tonë.
- Rasti II – Parashikimin e performancës së studentëve, duke përdorur të dhënat e mbledhura nga sistemi menaxhimit të studentëve, të një fakulteti të Tiranës.
- Rasti III - Parashikimit të performancës së studentëve. Qëllimi i studimit është që të parashikohet suksesi i studentëve në arsimin e lartë, duke përdorur të dhënat e mbledhura nga pyetësorët.

Për të arritur qëllimin, janë zbatuar krahasuar teknika të ndryshme të klasifikimit. Rezulatet do të tregojnë se cila është teknika më e mirë për secilin problem (rezultatet jepen në kapitullin IV).

Pasi kemi të qartë problemin dhe qëllimet, për të arritur objektivin tonë, janë përzgjedhur metoda të klasifikimit për tu shqyrtuar dhe për të konkluduar se cila prej tyre është më e përshtatshmja për problemin tonë.

### *3.2.1 Mbledhja dhe përshkrimi i të dhënave*

***Rasti I: Parashikimi i mundësive të punësimit të studentëve të diplomuar në vendin tone duke përdorur të dhënat e mbledhura nga pyetësorët elektronikë***

Të dhënat janë marrë nga plotësimi i një pyetësori(Aneksi 1), i hartuar në mënyrë të tillë që të mund të merrej informacion sa më i dobishëm për studimin e të rinjve të diplomuar. Informacioni i kërkuar përmbante pyetje rreth sjelljes së studentëve në të kaluarën, faktorëve social, shoqëror dhe demografik si edhe të dhëna për gjendjen aktuale të punesimit. Target grupi i zgjedhur për të plotësuar pyetësorin ishin të rinjtë nga Tirana dhe Korça, mosha 18 vjeç deri në 27 vjeç që kishin mbyllur një cikël të caktuar studimesh dhe mund të ishin edhe në ndjekje të një cikli, trajnimi, apo formimi tjetër. Qëllimi i këtij studimi ishte që të bëhen parashikime sa më të sakta për mundësitë e punësimit të të rinjëve (studentëve të diplomuar) bazuar tek arsimi që ata kishin ndjekur, dega apo profili që kishin zgjedhur, rezultatet e arritura, trajnimet apo

kurset e ndryshme plotësuese te ndjekura, të numrit të gjuhëve të huaja që zotëronin, vendbanimit, moshës, etj. Pyetësori përmbante dhe pyejte për të marrë më shumë informacion mbi statusin e tyre të punësimit, si psh Eksperiencat e mëparshme të punësimit, puna që kryejnë aktualisht, prej sa kohësh punonin aty, arsyet pse e kishin zgjedhur atë punë dhe si e kishin fituar vendin e punës; ndërsa për të rinjtë e papunësuar ju kërkohej arsyeja përse ishin ende pa punë. Formulari u shpërnda në mënyrë elektronike online nëpërmjet rrjeteve sociale si facebook, google+, linkedin etj; pas eliminimit të të dhënave jo të plota, shembulli përfshiu gjithsej 200 te rinj.

U krijua modeli i suksesit të studentëve te diplomuar, ku suksesi si një variabli rezultat (punësuar), matet me suksesin të punësimit të tyre. Modeli i parashikimit të ndëtuar përmbante 10 variabla dhe klasën Punësuar. Variablat e përzgjedhura janë në Tabelën 3.1.

**Tabela 3.1 Variablat e përzgjedhur për parashikimin e punësimit të studentëve të punësuar**

| Variablat | Përshkrimi | Vlerat e mundshme |
|---|---|---|
| *Gjinia* | Gjinia e studentit | M - Mashkull,<br>F - Femër |
| *Mosha* | Mosha e studentit | 18-27 |
| *Vendbanimi* | Vendi ku jeton student | A---Tiranë,<br>B---Korçë |
| *Studimet e Përfunduara* | Ciklin e Studimeve të përfunduara ose që ndjekin studentët | A – Doktoraturë,<br>B- Bachelor,<br>C- Master,<br>D - Arsim i mesëm,<br>E- Arsim 8-Vjecar, |
| *Vendi i studimeve* | Vendi ku kanë ndjekur studimet | A--Shqipëri,<br>B--Jashtë Vendit |
| *Tipi i shkollës* | Tipi i shkollës së ndjekur | A---Shtetërore,<br>B---Private |
| *Dega* | Dega e Studimit | |
| *Mesatarja* | Mesatarja e ciklit të fundit të përfunduar<br>(ose të universitet nqs janë ende në vazhdim) | |
| *Gjuhë të huaja* | Numri i gjuhëve të huaja që zotëron | A---1,<br>B---2,<br>C ---3,<br>D--- >3 |
| *Trajnime* | Nëse kanë ndjekur trajnime shtesë | A---Po,<br>B---Jo |
| *Punësuar* | Nëse janë të punësuar aktualisht | A---Po,<br>B---Jo |

Ne figurat 3.2 (a) dhe 3.2(b) jepet informacion mbi variablat nominale dhe numerike per pjesemarresit ne studim.

Selected attribute

Name: Gjinia Type: Nominal
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

| No. | Label | Count | Weight |
|---|---|---|---|
| 1 | Femer | 107 | 107.0 |
| 2 | Mashkull | 73 | 73.0 |

a) Atributi Nominal Gjinia

Selected attribute

Name: Mosha Type: Numeric
Missing: 0 (0%) Distinct: 9 Unique: 0 (0%)

| Statistic | Value |
|---|---|
| Minimum | 18 |
| Maximum | 27 |
| Mean | 23.761 |
| StdDev | 1.732 |

b) Atribut Numerik Mosha

**Figura 3.2 Informacionet e shfaqura në seksionin Selected attribute**

Dataset-i "Mundësitë-e-punësimit"përmban të dhënat e 100 të rinjve që si vendbanim kanë Tiranën dhe 80 të rinjve që jetojnë në Korçë. Nga grafiku i parë mund të vlerësojmë që në Tiranë ka më shumë të rinj që janë të punësuar (pjesa e ngjyrosur me blu). Nëse vërejme grafikun e dytë që i referohet vlerës Korçë numri i të rinjve të punësuar duket më i ulët, të paktën në target grupin e anketuar. Në figurën 3.3 jepen grafikët e të gjithë atributeve në bazë të klasës 'Punësuar'.

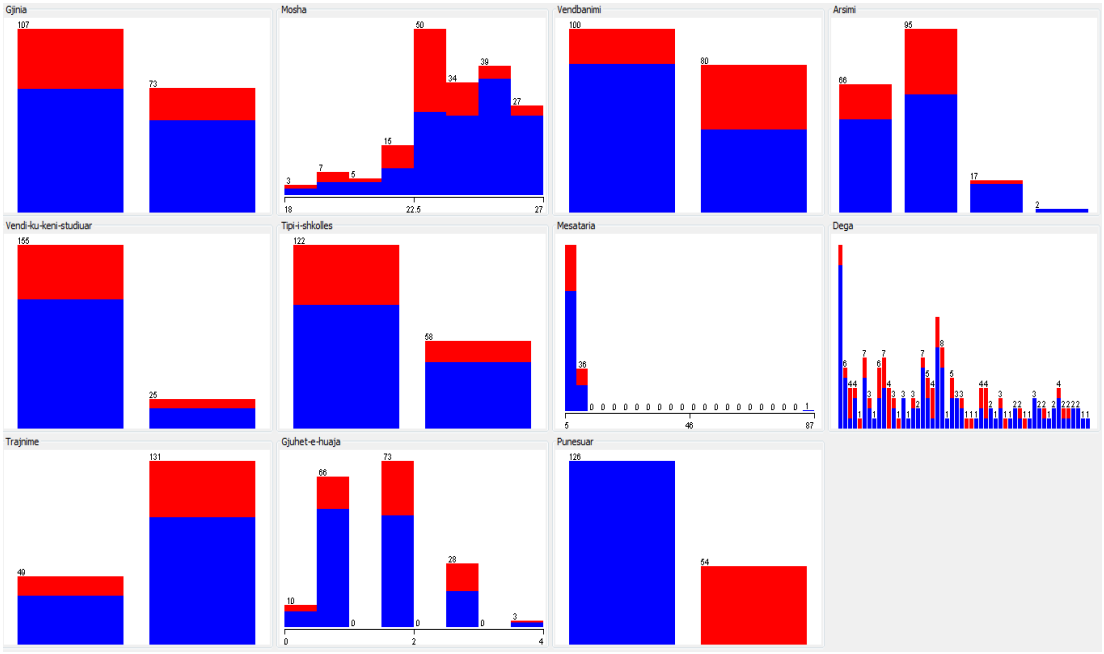



**Figura 3.3 Grafikët për të gjithë atributet dhe raportet e tyre me klasën punësuar.**

Kur analizojmë atributin Arsimi shohim se shumica e të rinjve sot kanë preferuar të ndjekin studimet vetëm 17 nga 180 të anketuarit kanë ndërprerë studimet pas mbarimit të shkolles së mesme ndërsa pjesa tjetër kanë preferuar ti ndjekin ato për më lart drejt universitetit (95 prej të anketuarve), ka të rinj që kanë studiuar master (66 prej 180 të anketuarve ) dhe shumë pak prej tyre kanë ndjekur doktoraturën (2 prej 180 të anketuarve). Duke studiuar raportet midis punësimi dhe arsimit të shfaqura me anë të grafikeve shohim se shumica e të rinjve që kanë ndjekur arsimin e mesëm janë të punësuar, kjo mbase ngaqë kanë kaluar shumë kohë që kur kanë mbaruar shkollën ose sepse kanë pranuar edhe alternativa të çfarëdoshme. Të gjithë të rinjtë që studiojnë doktoraturë janë të punësuar, ndërsa nga të rinjtë që kanë mbaruar universitetin, masterin ka më shumë të punësuar sesa të papunë, kjo sepse të rinjtë

kanë mësuar të jenë të pavarur dhe të angazhohen edhe me përgjegjësi të tjera përveç studimeve.
Nëse analizojmë atributin Vendi-ku-keni-studiuar vërehet se shumica e të rinjve kanë zgjedhur të ndjekin studimet në Shqiperi (155 nga 18) dhe vetëm një numër i vogël i të anketuarve (25) kanë vendosur të studiojnë jashtë vendit dhe siç mund të vihet re këta të fundit kanë patur mundësi më të mëdha punësimi kur janë kthyer në Shqipëri.
Nga atributi Tipi-i-shkollës vërehet se është rritur mjaft numri i të rinjve që preferojnë të frekuentojnë edhe shkollat private (58 prej 180) të cilët në krahasim me të rinjtë që kanë studiuar në shtet kanë patur mundësi më shumta punësimi.
Nga analizimi i të dhënave të mbledhura mesataria ka patur një ndikim shumë të vogël përsa i përket punësimit pas mbarimit të studimeve. Përsa i përket atributit Dega, ai ka marë 52 vlera të ndryshme por nga analiza shihet se të rinjtë që kanë studiuar Informatikë, Financë, Administrim Biznesi, Menaxhim, apo gjuhë të huaja kanë patur mundësi më të mëdha Punësimi.
Nga analizimi i atributit nominal Trajnime vërehet se pak të rinj në raport me numrin total kanë vendosur të ndjekin trajnime shtesë (49 prej 180) të cilët kanë patur mundësi plus përsa i përket punësimit. Gjuhët e huaja është atributi numerik që ka ndikuar pozitivisht për të rinjtë që kanë aplikuar për punë, të cilët kanë patur kualifikime shtesë në krahasim me të tjerët. Ndërsa nëse analizojmë klasën vërehet se pavarësisht periudhës të vështirë që po kalon vendi jonë të rinjve u eshtë kushtuar një prioritet i madh pasi 126 prej 180 të pyeturve e kanë gjetur aktualisht një punë.

### *Rasti II: Parashikimin e performancës së studentëve, duke përdorur të dhënat e mbledhura nga sistemi menaxhimit të studentëve, të një fakulteti të Tiranës*

Në sistemin arsimor sot, performanca e një studenti përcaktohet nga disa parametra (apo atribute). Për çdo student janë regjistruar disa të dhëna si ID, gjinia, vendlindja, vendbanimi, mesatarja e vitit të kaluar, dhe për çdo lëndë është regjistruar nota qëai ka marrë.
Sistemi i menaxhimit të studentëve nuk i siguron të dhënat në një format të gatshëm për një analizë të thjeshtë dhe direktë të tyre. Kështu që në fillim është e nevojshmë që të realizohet pastrimi dhe përpunimi e të dhënave.

Për të ndërtuar një model të mirë klasifikimi është ndjekur një metodologji e përbërë nga 5 hapa:

1. **Mbledhja e të dhënave.**Bashkësia e të dhënave që u përdorur në këtë studim është marrë nga sistemi i menaxhimit të studentëve të një fakulteti të Universitetit të Tiranës. Fillimisht kjo bashkësi përmante rreth 68300 rekorde me përsëritje dhe vlera të munguara, të cilat ju korrespondonin 2000 studentëve. Të dhënat janë marrë vetëm për dy degë që do ju referohemi si dega CS1 dhe ICT2 nga vitet akademik 2010-2013.
2. **Pastrimi dhe përpunimin e të dhënave**. Të dhënat e marra ishin të organizura në një tabelë (të ruajtura në një skedar Excel), format i cili na ndihmon në përpunimin dhe pastrimin e të dhënave. Në këtë hap të dhënat u ngarkuan në një bazë të dhënash të krijuar në MS SQL Server. Duke ekzekutuar disa query të ndryshmë mbi këtë bazë të dhënash u realizua:
    a. Proçesi i heqjes së dublikimeve
    b. Ndarja e studentëve sipas viteve dhe lëndëve përkatëse.
    c. Eleminimit të të dhënave jo të plota

d. Përllogaritja e mesatares vjetore të studenteve (Mesatarja e vitit të parë, mësatarja ë vitit të dytë dhë ajo e vitit të tretë)

Pas këtyre hapave studimi përfshiu rreth 1321 studentë në vitin e dytë ose tëtretë.

3. **Ndërtimi i modelit të klasifikimit**. Bashkësia e të dhënave përmbante 15 variabla, ku disa prej tyre janë eliminuar pasi ishin të parëndësishme për qëllimin e këtij studimi. Në këtë hap u zgjodhën vetëm ato fusha që ishin të nevojshme për burimet e të dhënave. Modeli i parashikimit të ndëtuar përmban 8 variabla dhe klasën ST_Mes_3 (Mesatarja e studentit në vitin e tretë). Variablat e përzgjedhura janë në Tabelën 3.2.Variabli ST_Mes_3, tregon mesataren e studentëve në përfundim të vitit të tretë. Rezultati është deklaruar si një variabël përgjigje. Ai është i ndarë në katër klasa: A, B, C dhe D (shiko tabelën 3.2). Në figurën3.4 tregohet shpërndarja e rezultateve të studentëve të studiuar.
4. **Vlerësimi i modelit**. ( shpjeguar në paragrafin 4.3)
5. **Përdorimi i modelit për punime të ardhshme**. (shpjeguar në paragrafin 4.3)

**Tabela 3.2 Variablat e përzgjedhur për studimin e studentëve në vitin e tretë**

| *Variablat* | *Përshkrimi* | *Vlerat* |
|---|---|---|
| ST_ID | Numër identifikues i studentit | |
| | Të dhënat demografike | |
| ST_Mosha | Mosha e studentit | 20,21,>22 |
| ST_Gjini | Gjinia e studentit | M, F |
| VendB | Vendbanimi i studentit | A ->Tiranë,<br>B-> Qytet tjetër,<br>C-> Fshat |
| ST_Shtet | Shtetësia e studentit | A -> Shqiptar,<br>B -> Tjetër |
| | Të dhëna akademike | |
| SP | Dega ku studion studenti | CS1, ICT2 |
| AK_Viti | Viti akademik i studentit | 3 |
| ST_Mes_1 | Mesatarjë e studentit në vitin e parë | A –> 900 - 10,<br>B -> 8.00 – 8.99,<br>C -> 7.00 – 7.99,<br>D ->< 6.99 |
| ST_Mes_2 | Mesatarja e studentit në vitin e dytë | |
| ST_Mes_3 | Mesatarja e studentit në vitin e tretë | |

Në studimin aktual, fushat e vlerave për disa nga variablat janë përcaktuar si më poshtë:

- **Student ID (ST_ID)**. Tregon kodin identifikues të studentit, i krijuar gjatë proçesimit të të dhënave, në mënyrë që të ruhej anonimati i tyre. Ky variablat, duke qënë se është unik për çdo student nuk është marrë në konsideratë gjatë aplikimit të metodave data mining.
- **Gjinia (ST_Gjini)**. Tregon gjininë e studentit e cila merr vlerat M (Mashkull) ose F (Femër).
- **Mosha (ST_Mosha)**. Tregon moshën e studentit të cilësi është vendosur emërtimi: A (20 vjeç), B(21 vjeç), C (mbi 21 vjeç).
- **Shtetësia e studentit (ST_Shtet)**. Tregon kombësinë e studentit, dhe i janë caktuar emërtimet: A (Shqiptar), B (Të tjerë, ku këtu përfshihen kombësitë Serbe, Kosovar, Maqedonas, etj).

- **Programi i studimit (SP)**. Tregon programin (degën) e studimit që po djeg student. Në rastin e studimit janë marrë vetëm dy programe studimi të cilave ju është vendosur emërtimi CS1 dhe ICT2.
- **Viti akademik i studentit (AK_viti).** Tregon vitin e studimit ku ndodhet student, në këtë studim janë marrë vetëm studentët e vitit të tretë, kështu që vlera e këtij variabli është konstante me vlerën 3.
- **Mesatarja (ST_MES_1, ST_MES2).** Tregon mesataren e ponderuar tëstudentit në vitin e parë(ST_MES_1) dhe në vitin e dytë(ST_MES_2), sipas emërtimeve: A (9.00 – 10.00), B (8.00 – 8.99), C (7.00 – 7.99), D (<= 6.99).
- **Mesatarja e vitit të tretë (ST_MES_3)**.Tregon mesataren e studentëve në përfundim të vitit të tretë. Rezultati është deklaruar si një variabël përgjigje. Ai është i ndarë në katër klasa: A, B, C dhe D. Kjo është dhe klasa mbi të cilën bazohet studimi, duke parashikuar mesataren e studentit në përfundim të vitit të tretë

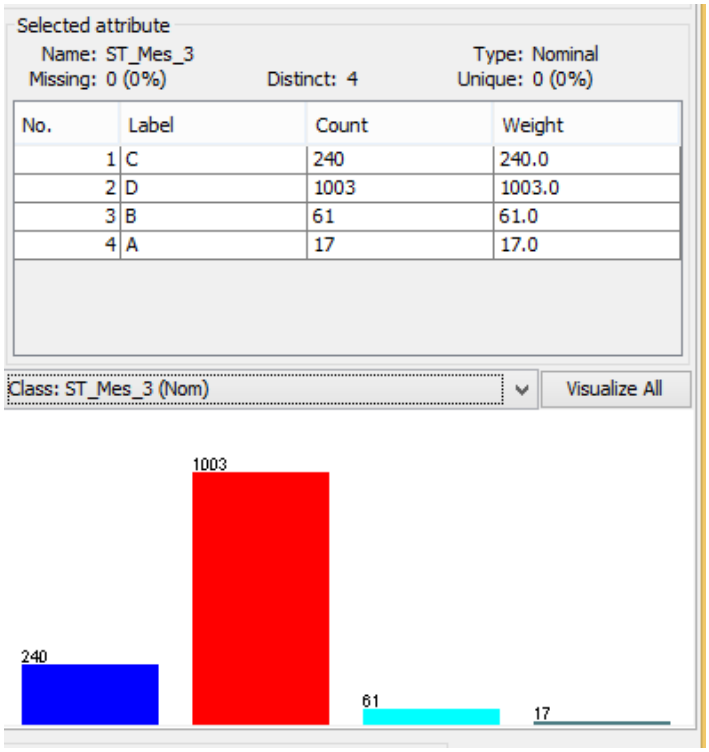



**Figura 3.4 Shpërndarja e rezultateve të studentëve bazuar në klasën ST_Mes_3**

### *Rasti III: Parashikimi i performances së studentëve duke përdorur të dhënat e mbledhura nga pyetësorët e shkruar*

Metoda dhe teknika të ndryshme të data mining u krahasuan gjatë studimit të parashikimit të performancës së studentëve, duke përdorur të dhënat e mbledhura nga pyetësorët(Aneksi 1) e shpërndarë gjatë semestrit të dytë në Universitetin e Tiranës, Fakulteti i Shkencave të Natyrës, viti akademik 2012-2013, tek studentët që kanë përfunduar vitin e parë në degën e Informatikës dhe Teknologjisë së Informacionit dhe Komunikimit.

Pyetësorët janë të hartuar në mënyrë të tillë që të mund të merrej një informacion sa më i dobishëm për studimin nga studentët. Informacioni i kërkuar ishte për sjelljen e

studentëve në të kaluarën (nota mesatare e shkollës së mesme), në lidhje me faktorë social e shoqëror, e demografik (përmendur në paragrafin 2.3), që ndikojnë në suksesin dhe performancën e studentit, si edhe të dhëna për gjendjen aktuale të provimeve të dhëna dhe krediteve të marra. Pra, ***suksesi i studentëve*** është vlerësuar në bazë të numrit të krediteve që ata kanë arritur të mbledhin në fund të vitit të parë shkollor. Në studim u shqyrtuan të gjithë variablat që mund të kenë ndikim në suksesin e studentit, si variablat shoqëror dhe demografik si dhe rezultatet e studentëve në shkollën e mesme. Pas eliminimit të të dhënave jo të plota, shembulli përfshiu gjithsej 200 studentë të pranishëm në kohën e studimit. U krijua modeli i suksesit të studentit, ku suksesi si një variabli rezultat, matet me suksesin ne degën e Informatikës dhe Teknologjisë së Iinformacionit dhe Komunikimit. Variablat e perzgjedhura janë dhënë tabelën 3.3.

Në studimin aktual, fushat e vlerave për disa nga variablat janë përcaktuar si më poshtë:

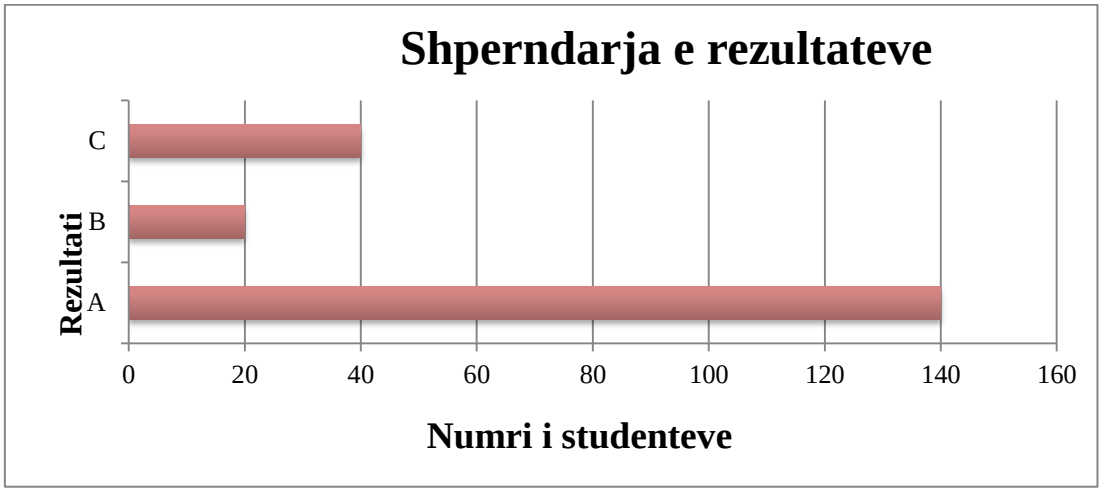
- **Dega**. Tregon kurset e ofruara nga Fakulteti i Shkencave të Natyrës në departamentin e Informatikës, të cilat janë Informatika (INF) dhe Teknologjia e Informacionit dhe Komunikimit (TIK).
- **Gjinia**. Tregon gjininë e studentit që do të jetë Mashkull (M) ose Femër (F).
- **Mosha**. Tregon moshën aktuale (mosha kur është plotësuar pyetësori) të studentit e cila ka emërtimin: A (18 vjeç), B 19 vjeç), C (20 vjeç), D (21 vjeç), E (mbi 21 vjeç).
- **Mesatarja e shkollës së mesme (MShM)**. Tregon notën mesatare me të cilën studenti ka përfunduar shkollën e mesme. Mesataret i janë caktuar të gjithë studentëve duke përdorur emërtimin: A (9.00-10.00), B (8.00-8.99), C (7.00-7.99), D (6.00-6.99), E (5.00-5.99).
- **Vendbanimi (VB)**. Tregon vendin nga ku vjen studenti, pra vendin ku ka gjendjen civile, pra dhe vendi ku mund të kryer studimet e shkollës së mesme.
- **Konvikti (Konv).** Tregon nëse studenti aktualisht banon në konvikt apo në një banesë me qera ose banesë të tij.
- **Përmasat e familjes (PF).** Tregon numrin e anëtarëve të familjes së studentit, sipas emërtimeve: A (1 anëtar), B (2 anëtar), C (3 anëtar), D (mbi 3 anëtar).
- **Të ardhurat familjare (AF)**. Tregon të ardhurat familjare të studentëve,sipas emërtimeve: A (Të ulëta), B (Të mesme), C (Të larta).
- **Profesioni i babait (PB).** Tregon profesionin e babait sipas emërtimeve:
  A (Buxhetor), B (Biznesmen), C (Bujqësi), D (Pensionist), E (Të tjera).
- **Profesioni i nënës (PN).** Tregon profesionin e nënës sipas emërtimeve:
  A (Buxhetore), B (Biznesmene), C (Shtëpiake), D (Pensioniste), E (Të tjera).
- **Rezultati (Rez)**. Rezultati i studentëve në përfundim të vitit të parë. Rezultati është deklaruar si një variabël përgjigje, ku student duhet të plotësonte numrin e krediteve. Ai është i ndarë në tre klasa: Kalon, Mbetet, KalonM. Nëse studenti ploteson të 60 kreditet kalon në vitin pasardhës (Kalon), nëse studenti plotëson 20 kredite kalon në vitin tjetër duke mbartur provimet e mbetura (KalonM) dhe nëse studenti nuk plotëson 20 kreditet nuk kalon në vitin pasardhës (Mbetet).

Duke qënë së rezultati është në rastin tonë të studimit është i bazuar në numrin e kreditive, përgjigjet që student ka dhënë për numrin e provimeve të hyrë dhe të marrë nuk janë shqyrtuar dhe janë eliminuar nga ky studim.

**Tabela 3.3 Variablat e lidhura me studentët**

| Variablat | Përshkrimi | Vlerat e mundshme |
|---|---|---|
| *Dega* | Dega ku bëjnë pjesë studentët | {INF,TIK} |
| *Gjinia* | Gjinia e studentit | {M,F} |
| *Mosha* | Mosha e studentit | { A--- 18,<br>B--- 19,<br>C--- 20 ,<br>D--- 21 ,<br>E--- >21 } |
| *Mesatarja e shkollës së mesme (MShM)* | Nota mesatare e shkollës së mesme | { A--- 9.00-10.00,<br>B--- 8.00-8.99,<br>C--- 7.00-7.99,<br>D--- 6.00-6.99,<br>E--- 5.00-5.99} |
| *Numri i provimeve dhënë (NPD)* | Numri i provimeve që student ka hyrë | { A --- 2 – 4<br>B --- 5 – 7<br>C --- 7 – 9<br>D --- > 9} |
| *Numri i provimeve marrë (NPM)* | Numri i provimeve që student ka kaluar | { A --- 2 – 4<br>B --- 5<br>C --- 6<br>D --- > 6} |
| *Numri i krediteve* | Numri i krediteve që ka marrë studentit | { A --- 20<br>B --- 21 – 30<br>C --- 31 – 40<br>D --- > 41} |
| *Vendbanimi (VB)* | Vendi ku jeton student | { A---Tiranë,<br>B---Qytet tjetër,<br>C---Fshat} |
| *Konvikti (Konv )* | Studenti jeton apo jo në konvikt | {Po,Jo} |
| *Përmasat e familjes (PF)* | Numri i pjestarëve të familjes së studentit | { A--1,<br>B--2,<br>C--3,<br>D-->3 } |
| *Të ardhurat (AF)* | Të ardhurat vjetore të familjes | { A---Të ulëta,<br>B---Të mesme,<br>C---Të larta} |
| *Profesioni i babait (PB)* | Profesioni që ka babai i studentit | { A---Buxhetor,<br>B---Biznesmen,<br>C---Bujqësi,<br>D---Pensionist,<br>E---Të tjera } |
| *Profesioni i nënës (PN)* | Profesioni që ka mamaja e studentit | {A---Buxhetore,<br>B---Biznesmene,<br>C---Shtëpiake,<br>D---Pensioniste,<br>E---Të tjera } |
| *Rezultati (Rez)* | Rezultati i studentit vitin e parë | { A---Kalon,<br>B---KalonM,<br>C---Mbetet} |

Në grafikun e 3.1 tregohet shpërndarja e rezultateve të studentëve të anketuar, sipas klasës Rez.




**Grafiku 3.1 Shpërndarja e rezultateve në degën e informatikës dhe tik**

Variabli rezultat – Vlerësimi i studentit në vitin akademik mund të grupohet në mënyra të ndryshme dy nga të cilat janë:

- Nëpërmjet tre klasave të përcaktuara në bazë të numrit të krediteve të studentëve: emërtimet janë të njëjta me ndarjet e caktuara në bazë të krediteve A→Kalon, B→ KalonM, C→Mbetet. % është siç tregohet në tabelën 3.4.

**Tabela 3.4 Tre klasat që i përkasin rezultateve të studentëve**

| Klasa | Rezultati | Studentët | Përqindja |
|---|---|---|---|
| 1 | A | 20 | 10% |
| 2 | B | 140 | 70% |
| 3 | C | 40 | 20% |

- Nëpërmjet dy klasave të koduara në këtë mënyrë: kategoria A-Mbetet, kategoria B-Kalon, siç tregohet në Tabelën 3.5.

**Tabela 3.5 Dy klasat që i përkasin rezultateve të studentëve**

| Klasa | Rezultati | Studentët | Përqindja |
|---|---|---|---|
| 1 | A | 40 | 10% |
| 2 | B | 160 | 90% |

Të dyja këto grupime kanë një shkallë të vogël gabimesh pasi përzgjedhja e rezultatit është bërë në mënyrë që të kishim një shpërndarje të tillë. Sa më i madh të ishte numri i kategorive në të cilat do të paraqitej rezultati,aq edhe më e lartë do të ishte mundësia për të gabuar.

### *3.2.2 Metodat e data mining të përdorura*

"*Data mining, është një metodë kompjuterike e proçesimit të të dhënave e cila është zbatuar me sukses në shumë fusha që kanë patur si qëllim përftimin e njohurive të dobishme nga të dhënat*"
(Klosgen and Zytkow, 2002).

Teknikat e data mining përdoren për të ndërtuar një model në lidhje me të cilin të dhënat e panjohura do të përpiqen të identifikojnë informacionin e ri. Pavarësisht nga origjina, të gjitha teknikat e data mining tregojnë një veçori të përbashkët: zbulimin automatik të lidhjeve të reja dhe varësinë e atributeve në të dhënat e vëzhguara. Në këtë mënyrë algoritmet janë të ndara në dy grupe:

- **Algoritme pa kontroll (të pa mbikqyrur)**

Kur të dhënat janë të padrejtuara ose të pambikqyrura, kushtet e rezultatit nuk paraqiten në mënyrë eksplicite në bashkësinë e të dhënave: detyra e algoritmit jo mbikqyrës është të zbulojë menjëherë modele të pandara në të dhënat, pa informacion fillestar mbi klasën të cilës mund t'i përkasin të dhënat, dhe nuk përfshin asnjë kontroll. (Cios, Pedrycz, Swiniarski and Kurgan, 2007). Shpesh modeli i prodhuar nga një algoritëm jo mbikqyrës mund të përdoret për detyra parashikimi, edhe pse nuk është i hartuar për punë të tilla. Grupimi dhe rregullat e shoqërimit i përkasin këtij grupi.

- **Algoritme me kontroll ( të mbikqyrur)**

Algoritmet e mbikqyrur, janë ata të cilët përdorin të dhëna paraprakisht me klasa të njohura, të cilave i përkasin të dhënat, për ndërtimin e modeleve, dhe më pas në bazë të modelit të ndërtuar parashikojnë klasën të cilës do ti përkasin të dhënat e panjohura. Metoda e klasifikimit i përket këtij grupi. Metoda të klasifikimit të të dhënave përfaqësojnë një proçes të mësimit të një funksioni që i cakton të dhënat në klasa të paracaktuara. Çdo algoritmi klasifikimi, që bazohet në të mësuarit induktiv, i caktohet bashkësia e të dhënave hyrëse, e cila konsiston në vektorë të vlerave të atributeve dhe klasave koresponduese.

Qëllimi i një teknike klasifikuese është që të ndërtojë një model që bën të mundur klasifikimin e veçorive të të dhënave të ardhshme, bazuar në një bashkësi karakteristikash specifike në një mënyrë të automatizuar. Sisteme të tilla marin një koleksion rastesh si input, ku secili i përket një numri të vogël klasash dhe përshkruhet nga vlera e tij për një bashkësi të dhënë atributesh. Si rezultat ata marrin një klasifikues që mund të parashikojë saktësisht klasën, të cilës i përket një rast i ri. Ka shumë llojë klasifikuesish dhe përzgjedhja e më të mirit është shumë e vështirë, sepse ata dallojnë në shumë aspekte si: shpejtësia e të mësuarit, sasia e të dhënave për kualifikim, shpejtësia e klasifikimit, fuqia (fortësia), etj.
Modelet e klasifikimit bëhen nga përdorimi i këtyre algoritmeve, qëllimi i të cilëve është parashikimi i klasës (suksesit të studentit) të cilëve do i përkasin disa shembuj të ri të pa emërtuar.

***Rasti I: Parashikimi e mundësive të punësimit të studentëve të diplomuar.***Në këtë studim është shqyrtuar ndikimi i tre algoritmeve për analizën inteligjente të të dhënave: LogitBoost, AdaBoost1 dhe NaiveBayesUpdateable. Synimi i këtij studimi ka qëne interpretimi rezultatit

të marrë dhë krahasimi i rezultateve të algoritmeve të klasifikimit që janë përdorur për të zbuluar mënyrën më të përshtashme për të parashikuar suksesin e studentëve të diplomuar.

***Rasti II: Parashikimi i performancës së studentëve, duke përdorur të dhënat e mbledhura nga sistemi menaxhimit të studentëve.*** Në këtë studim është shqyrtuar ndikimi i pemëve të vendimit në vendimamrrjen e zgjedhjes së një rrugë të përshtatshme për të rritur përformancën e studentëve, duke analizuar eksperiencat e studentëve me arritje akademike të ngjashme. Janë studiuar tre algoritme klasifikimi për ndërtimin e pemëve të vendimit RandomTree, J48 dhe REPTree. Më pas janë krahasuar këto algoritme për të zbuluar mënyrën më të përshtashme për të parashikuar suksesin e studentëve gjatë tre viteve të shkollës së lartë.

***Rasti III: Parashikimi i performances së studentëve.*** Në këtë studim është shqyrtuar ndikimi i tre algoritmeve për analizën inteligjente të të dhënave: C4.5, Multilayer Perceptron dhe Naïve Bayes. Tre teknikat e përzgjedhura të klasifikimit janë përdorur për të zbuluar mënyrën më të përshtashme për të parashikuar suksesin e studentëve.

## *Përshkrim i shkurtër i algoritmeve*

- **Algoritmi Naive Bayes (NB**) është një metodë shumë e thjeshtë klasifikimi e bazuar në teorinë e probabilitetit, teorema baeziane (Witten and Frank, 2000). Quhet naive sepse i thjeshtëzon problemet duke u bazuar në dy supozime të rëndësishme: supozon se atributet që marrin pjesë janë të pavarura me klasifikimet e njohura, dhe supozon që nuk ka atribute të fshehura që mund të ndikojnë në proçesin e parashikimit. Ky klasifikues përfaqëson çështjen e premtuar në zbulimin probabilitar të njohurive, dhe siguron një algoritëm shumë efikas për klasifikimin e të dhënave. Algoritmi ***NaiveBayesUpdateable*** (NBU) është një version i modifikueshëm i klasifikuesit NaiveBayes. Ky klasifikues përdor një preçision 0.1 kur klasifikuesi thërritet për të trajnuar 0 shembuj.
- **Algoritmi Multilayer Perceptron (MLP)** është një nga algoritmet gjerësisht të përdorur dhe më i përhapuri në rrjetat neurale. Rrjeti përbëhet nga një grup elementesh ndijor që ndërtojnë shtresën hyrëse, një ose më shumë shtresa të fshehura të përpunimit të elementeve, dhe shtresën rezultat të përpunimit të elementeve (Witten dhe Frank, 2000). MLP është veçanërisht i përshtatshëm për përafrimin e një funksioni klasifikimi (kur ne nuk jemi aq shumë të njohur me lidhjen që egziston ndërmjet atributeve hyrëse dhe dalëse) e cila e vendos shembullin e përcaktuar nga vlerat e vektorit të atributeve, brënda një ose më shumë klasave.
- **Algoritmi Meta LogitBoost (LB)**është zhvilluar nga Friedman në vitin 2000. Ai performon një regresion llogjik shtesë dhe gjeneron modele individuale në çdo interacion. Algoritmi LogitBoost përdor parametrin cross-validation për të përcaktuar një numër iteracionesh dhe për të mbaruar ekzekutimin para se performanca të bjerë.
- **Algoritmi AdaBoost1 (AB1)**është një algoritëm i cili është dizenjuar në mënyrë specifike për klasifikim dhe gjithashtu ai mund të përdoret si model në çdo algoritëm klasifikimi. Ky algoritëm mund të përballojë instancat e ponderuara (që kanë peshë), ku pesha e një instance është numër pozitiv. Trajtimi i instancave të ponderuara ndryshon mënyrën sesi llogaritet gabimi.
- **Algoritmi i pemëve të vendimit** më i zakonshëm në ditët e sotme në fushën e data mining, për të mos thënë mundësisht më i përdoruri, është C4.5. Në vitin

1993 profesor Ross Quinlan, zhvilloi një algoritëm të pemës së vendimit të njohur si C4.5; i cili përfaqëson rezultatin e studimit që të kthen tek algoritmi ID3(i cili gjithashtu u propozua nga Ross Quinlan në vitin 1986). Ky algoritëm trajton si atributet kategorik ashtu edhe ata të vazhdueshëm për të ndërtuar pemën e vendimit. Ai ka karakteristika shtesë siç janë: suportimi i vlerave që mungojnë, kategorizimin e atributeve të vazhdueshme, pruning (tëharrjen) e pemëve të vendimit, derivimin e rregullave, dhe të tjera. Baza e ndërtimit të algoritmit C4.5 është përdorimi i metodës së njohur me emrin përça dhe sundo; kjo është metoda që përdoret për të ndërtuar një pemë të përshtatshme nga bashkësia e të dhënave trajnuese S (Kumar & Vijayalakshmi 2011):

- Nëse të gjithë shembujt në S i përkasin të njëjtës klasë ose S është i vogël, pema është një gjethe e kategorizuar me klasën më të shpeshtë në S.
- Përndryshe, zgjidh një test të bazuar në një atrinut të vetëm me dy ose më shumë rezultate. Bëje këtë test rrënjën e pemës me një degë për secilin rezultat të testit, ndaje S në nëngrupe koresponduese S, S2,... në raport me rezultatin për secilin rast, dhe apliko të njëjtën proçedurë në mënyrë rekursive në çdo nëngrup.

Zakonisht ka shumë teste që mund të zgjidhen në hapin e fundit. C4.5 zgjedh dy kritere heuristikash për të renditur teste të mundshëm: dobinë e informacionit, e cila minimizon entropinë totale të nëngrupeve, dhe dobinë e ratios dfault që ndan dobinë e informacionit nga informacioni i marrë nga rezultatet e testit (Kumar and Vijayalakshmi, 2011).

Algoritmi J48 (që është përdorur) është një implementim i algoritmit të pemës së vendimit C4.5 në Weka. Paraqitja bëhet nga struktura e pemës. Në çdo nyje të brendshme, vlerësohet kushti i atributit, dhe secila degë e pemës përfaqëson një rezultat të studimit. Degëzimi i pemëve përfundon me gjethe të cilat përcaktojnë një klasë të cilës i përkasin shembujt. Në ditët e sotme, algoritmi i pemës së vendimit është një proçedurë mjaft e përhapur për shkak të implementimit të thjeshtë të saj, dhe në mënyrë të veçantë për faktin se rezultati i saj mund të shprehet në mënyrë grafike.

- **Algoritmi RandomTree** është një algoritëm klasifikimi (i kontrolluar) i zhvilluar nga Breiman dhe Cutler. Algoritmi mund të punojë dhë për teknikat e klasifikimit dhe te regresionit. RandomTree është një bashkësi e pemëve të parashikimit e cila njihet ndryshe si 'Pyll'(Forest). Algoritmi i klasifikimit punon si: klasifikuesi RandomTree merr vektorin e të dhënave hyrëse, e klasifikon atë në çdo pemë të pyllit dhe jep klasën e cila ka marrë shumicën e 'votave'. Algoritmi RandomTree është në thelb një kombinim i dy algoritmeve ekzistuese: Singel Model Tree kombinohen me RandomForest. (Witten and Frank, 2000)
- **Algoritmi REPTree**

Për të vlerësuar fuqinë(fortësinë) e klasifikuesit, metodologjia e zakonshme është kryerja e vlerësimit të kryqëzuar mbi klasifikuesin. Në rastet I dhe III të studimit është përdorur një vlerësim i kryqëzuar 3-fish (3-pjesësh): të dhënat u ndan në mënyrë random në tre nëngrupe me përmasa të barabarta. Dy nëngrupe u përdorën për trajnim, një për vlerësimin e kryqëzuar,dhe një për të matur saktësinë e rrjetit final të ndërtuar. Kjo proçedurë u performua tre herë në mënyrë të tillë që çdo nëngrup të testohej vetëm një herë. Rezulatet e testeve u mesatarizuan mbi 3 vlerësime të kryqëzuara. Weka bën llogaritjen e të gjitha këtyrë metrikave të performancës pasi ekzekutohet një vlerësim i kryqëzuar me k-pjesë. Në përfundim u krahasuan saktësitë

e modeleve (kapitulli IV). Ndërsa në rastin II të studimit janë përdorur 5 vlerësime të ndryshme Training Set, Vlerësimi i kryqëzuar i  3-fisht dhe 13-fisht.

## 3.3 Implementimi i modelit

Eksperienca tregon se asnjë skemë e të mësuarit makinë nuk është e përshtatshme për të gjithë problemet e Data Mining.
Për qëllimin e këtij punimi (për të tre rastet e studimit) është përdorur paketa e programit Weka, që u zhvillua në universitetin e Waikato-s në Zelandën e re. Kjo paketë është implementuar në gjuhën JAVA dhe në ditët e sotme mund të quhet si paketa më e përshtatshme dhe gjithpërfshirëse me algoritme të njohurisë makinë, në botën akademike dhe pa qëllime fitimi (Grupi i Njohurisë Makinë në Universitetin Waikato, 2015).
Paneli i Weka-s është një koleksion i algoritmave të të mësuarit makinë dhe mjeteve të preprocesimit të të dhënave. Ajo është dizenjuar që përdoruesit të provojnë në mënyrë të shpejtë metodat ekzistuese në dataset në mënyra mjaft fleksible.
Weka siguron suport për të gjithë procesin eksperimental të gjurmimit të të dhënave, duke përfshirë këtu përgatitjen e inputit, vlerësimin e skemave të të mësuarit sistematik, vizualizimin e të dhënave hyrëse dhe rezultatet e procesit të të mësuarit. Të gjithë mjetet janë të aksesueshme nëpermjet një ndërfaqeje të thjeshtë e cila u mundëson përdoruesve të krahasojnë metoda të ndryshme dhe të identifikojnë ato që janë më të përshtatshme për zgjidhjen e problemit.
Më të dhënat e mbledhura që u përmenden më lart, krijohet një skedar në exel me prapashtesën *.arff.(ose *.csv) më pas ky skedar, i cili përmban bashkësinë e të dhënave, ngarkohet në Weka Explorer.


**Figura 3.5 Ndërfaqja kryesore e programit WEKA**

Paneli "Classify" bën të mundur që të zbatohen algoritmet e klasifikimit mbi bashkësinë e të dhënave në mënyrë që të mund të vlerësohet saktësia e modeleve parashikuese. Modelet parashikuese të zbatuara sigurojnë mënyrën për të parashikuar nëse një student do të jetë ose jo i suksesshëm.

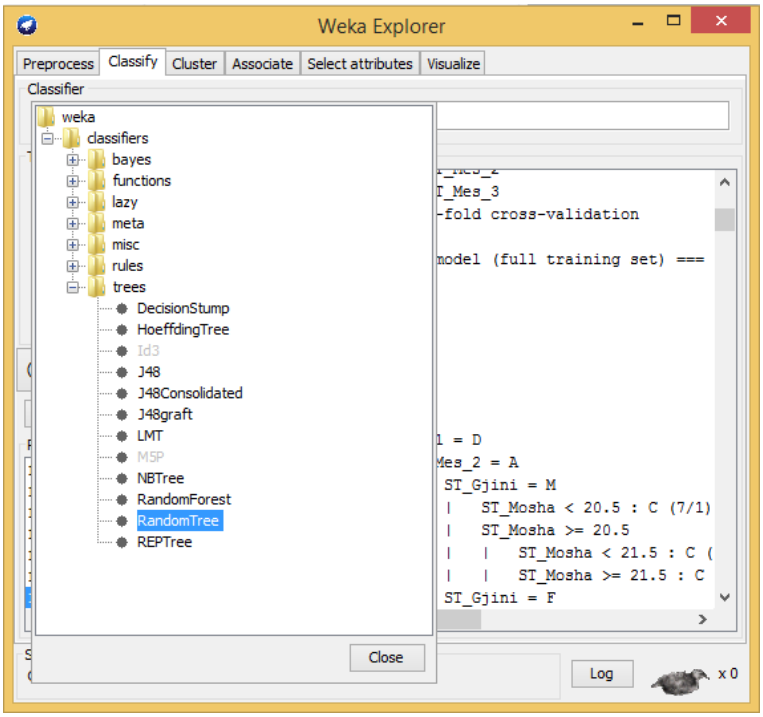



**Figura 3.6 Paneli "Clasify" në aplikacionin WEKA**

# 4. Rezultatet dhe Analiza e Studimit

*Ky kapitull nis me një analizë kërkimore në mënyrë që të kuptohet më mirë natyra e të dhënave. Pasi zbatohen algoritmet e përzgjedhura bëhet vlerësimi dhe analiza e rezultateve që jep secili prej tyre në mënyrë që të arrijmë në përfundimin se cili prej tyre është më i përshtashëm.*

## 4.1 Analiza e variablave hyrëse

Në mënyrë që të kemi një pamje më të mirë të rëndësisë së variablave hyrëse, është bërë zakon që të analizohet rëndësia e variablave hyrëse gjatë parashikimit të suksesit të studentit, në të cilin është analizuar ndikimi i disa variablave hyrëse të caktuara të modelit, në rezultat.

Testet u zhvilluan duke përdorur katër teste për vlerësimin e variablave hyrëse:

- Testin Chi-Square
- Testin OneR
- Testin dobia e informacionit (info gain)
- Testin raporti i dobisë (gain ratio)

Rezultatet e secilit test u monitoruan duke përdorur metrikat e mëposhtme:

- Atributi (emër atributi)
- Cilësia (Merit- përmasa e cilësisë)
- Merit dev (devijimi i cilësisë)
- Renditja (pozicioni mesatar i zënë nga atributi)
- Renditja dhe devijimi(devijimi, i cili zë vendin e pozicionit të atributit)

Algoritme të ndryshme sigurojnë rezultate të ndryshme, secili prej tyre e llogarit ndikimin e atributit në një mënyrë të ndryshme. Vlera mesatare e të gjithë algoritmeve merret si rezultati final i renditjes së atributeve, në vend që të zgjidhet vetëm një algoritëm dhe t'i besohet atij.
Rezultatet e përftuara nga këto vlera për rastin e parë dhe të dytë të studimit tregohen në Tabelën 4.1 dhe Tabelën 4.2. Në këtë tabelë agregate kolonat e "cilësië" nuk janë të zbatueshme, sepse algoritmet përdorin metrika që nuk përputhen. Qëllimi i kësaj analize është që të përcaktohet rëndësia individuale e secilit atribut.

Tabela 4.1 tregon që atributi MShM (nota mesatare e shkollës së mesme) ndikon më së shumti në rezultat, dhe ky atribut tregon performancën më të mirë në të katër testet e bëra. Më pas vijnë atributet: PB (profesioni i babait), AF (të ardhurat familjare). Ndërsa atributet me ndikim më të vogël janë: Gjinia dhe PF( përmasat e familjes).
Tabela 4.2 tregon që atributi ST_MES_1 (nota mesatare e viti të parë) ndikon më së shumti në rezultat, dhe ky atribut tregon performancën më të mirë në të katër testet e bëra. Më pas vijnë atributet: ST_MES2 (nota mesatare e viti të dytë), SP (programi

studimit). Ndërsa atributet me ndikim më të vogël janë: ST_Mosha dhe ST_Gjini (Mosha dhe gjinia e studentit).

**Tabela 4.1 Rezultatet e testeve dhe mesatarja e tyre e renditur, për rastin I të studimit**

| Atributet/Testet | Chi-squared | One R | Info Gain | Gain Ratio | MES Rang |
|---|---|---|---|---|---|
| MShM | 1 | 1 | 1 | 1 | 1 |
| PB | 3 | 3.3 | 2.7 | 6 | 3.75 |
| AF | 3.3 | 3 | 3.7 | 7.7 | 4.425 |
| Mosha | 2.7 | 2.7 | 2.7 | 9.7 | 4.45 |
| Konv | 7 | 5.3 | 7 | 3.7 | 5.75 |
| PN | 6 | 7.3 | 6.3 | 4.7 | 6.075 |
| VB | 5.7 | 6.7 | 6 | 6.3 | 6.175 |
| Dega | 8.7 | 8.3 | 8.3 | 4.7 | 7.5 |
| Gjinia | 9 | 9.3 | 9 | 4 | 7.825 |
| PF | 8.7 | 8 | 8.3 | 7.3 | 8.075 |

**Tabela 4.2 Rezultatet e testeve dhe mesatarja e tyre e renditur, për rastin III të studimit**

| Attributet/Testet | Gain Ratio | OneR | Info Gain | Chi Square | Mes Rang |
|---|---|---|---|---|---|
| ST_MES_1 | 1 | 1.7 | 2 | 1 | 1.425 |
| ST_MES_2 | 2 | 1.3 | 1 | 2 | 1.575 |
| ST_SHTET | 3.7 | 5.7 | 5.7 | 5.7 | 5.2 |
| SP | 4 | 6 | 3.3 | 3.3 | 4.15 |
| VendB | 4.3 | 3 | 3.7 | 3.7 | 3.675 |
| ST_Gjini | 6 | 6.3 | 5.3 | 5.3 | 5.725 |
| ST_Mosha | 7 | 4 | 7 | 7 | 6.25 |

## 4.2 Vlerësimi i performancës dhe dobisë së algoritmeve të klasifikimit

Në studim janë zhvilluar disa eksperimente në mënyrë që të vlerësojmë performancën dhe dobinë e algoritmeve të ndryshme të klasifikimit. Vlerësimi i një algoritmi klasifikimi është një nga pikat kryesore të data mining. Dy nga mjetet më të përdorura për analizimin e performancës së një algoritmi klasifikimi janë: Matrica e Çrregullimeve dhe ROC (Receiver Operating Characteristics).

Matrica e çrregullimeve paraqet numrin e intancave të klasifikuara ne rregull kundrejt atyre të klasifikuara jo në rregull. Kolonat përfaqësojnë parashikimet, dhe rreshtat paraqesin klasat aktuale. Për të vlerëuar fortësinë(fuqinë) e klasifikuesit, metodologjia e zakonshme është performimi i vlerësimit të kryqëzuar mbi klasifikuesin. Në tabelën e mëposhtme paraqitet një matricë e çrregullimeve me dy klasa ( Vërtetë dhe Gabuar).

**Tabela 4.3 Matrica e çrregullimeve për një klasifikues me dy klasa**

| | ***Klasat e parashikuara*** | | |
|---|---|---|---|
| ***Klasat Aktuale*** | | Klasa Vërtetë | Klasa Gabuar |
| | Klasa Vërtetë | Vërtetë Pozitive | Gabuar Pozitive |
| | Klasa Gabuar | Vërtetë Negative | Gabuar Negative |

Nga kjo tabelë evidentohet dhe numri i intancave të klasifikuara në saktë, që është e barabartë me vlerën e *Vërtetë Pozitive plus vlerën e Gabuar Negative* (qelizat me ngjyrën e gjelbër); ndërsa numri i instancave të klasifikuara gabim, është i barabartë me vlerën e *Gabuar Pozitive plus Vërtetë Negative (*qelizat me ngjyrën rozë*).*

Numri i instancave të parashikuara në rregull kundrejt vlerave të parashikuara specifikon vlerën e parametrit të *preçizionit* e cila merr vlera ndërmjet 0 dhe 1. Në rastet kur preçizioni është i barabartë me zero, nënkupton se modeli i ndërtuar nuk ka aftësi parashikuese.

Parametra të tjerë që përdoren për të analizuar ndjeshmërinë dhe saktësinë e parashikimit janë (Oprea 2014):

- **Shkalla e vertetë prozitive - TP** (True Positive) - tregon shembujt në klasën pozitive që janë parashikuar si pozitivë. Ajo paraqet raportin midis të dhënave të parashikuara si pozitive me të gjitha të dhënat që ndodhen në klasën X të modelit të ndërtuar.
- **Shkalla e vërtetë negative -TN** (True Negative) - tregon shembujt në klasën negative që janë parashikuar si negativë. Ajo paraqet raportin midis të dhënave të parashikuara si negative me të gjitha të dhënat që ndodhen në klasën X të modelit të ndërtuar.
- **Shkalla e gabuar positive - FP** (False Positive) - paraqet raportin midis të dhënave të klasifikuara gabimisht në klasën X me të dhënat e të gjithë klasave të tjera përveç klasës X.
- **Shkalla e gabuar negative - FN**(False Negative) - paraqet raportin e të dhënave të klasifikuara si negative por që në të vërtetë janë pozitive.
- **Preçizion** tregon raportin e të gjithë shembujve që vërtet i takojnë klasës X me numrin total të shembujve të klasifikuar që supozohet se i takojnë klasës X.
- **Recall** tregon shkallën e vërtetë pozitive (TP) dhe llogaritet me anë të formulës:

$$Recall = \frac{TP}{TP + FN}$$

- **F-measure** është një rregull i kombinuar i Precision dhe Recall sipas formulës:

$$F - Measuare = \frac{2 * Precision * Recall}{Precision + Recall}$$

- **MCC** është një matës i performancës për klasifikimet binare sidomos kur punohet me klasa të pabalancuara dhe llogaritet me anë të formulës :

$$MCC = \frac{(TP * TN) - (FP * FN)}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}}$$

- **Saktësia** (Accuracy) tregon raportin e numrit të instancave të parashikuara në rregull. Ajo llogaritet sipas formulës:

$$Saktësia = \frac{Rastet\ në\ diagonalen\ kryesore\ të\ matricës}{numrit\ të\ total\ të\ instancave} = \frac{TP + FN}{numri\ total\ i\ instancave}$$

- **Gabimi** tregon raportin e instancave të parashikuara gabim:
  - $\boldsymbol{Gabimi = 1 - saktësi}$

## 4.3 Rastet e studimit. Vlerësimi i algoritmeve të klasifikimit

Pasi përcaktohen algoritmet që do të zbatohen dhe mënyra e vlerësimit të performancës, kalohet më pas në ekzekutimin e algortimeve dhe vlerësimin e performacës së tyre.

### *4.3.1 Rasti I: Parashikimi i mundësive të punësimit të studentëve të diplomuar në vendin tonë duke përdorur të dhënat e mbledhura nga pyetësorët elektronikë*

Në këtë studim u shqyrtuan dhe u zbatuan tre algoritme të rëndësishme (LogitBoost - LB, AdaBoost1 – AB1 dhe NaiveBayesUpdateable -NBU) të DataMining në vlerësimin e preprocesimit të të dhënave dhe performancën e metodave të të mësuarit ku vlerësimi u bazua në saktësinë parashikuese të metodave, lehtësinë e të të mësuarit dhe karakteristikat e thjeshta në përdorim.

Synimi i këtij studimi është interpretimi i detajuar i rezultatit të marrë nga programi WEKA dhe gjetja e klasifikuesit i cili jep parashikimin më të mirë nga bashkësia e të dhënave 'Mundësi Punësimit'

***A. Interpretimi i rezultatit nga aplikimi WEKA***

*Pas përpunimit të bashkësisë së të dhënave, ngarkohet bashkësia në programin WEKA dhe ekzekutohen algoritmet LogitBoost, AdaBoostM1 dhe NaiveBaysUpdateable. Në këtë pjesë do të përshkruajmë rezultatin e shfaqur nga programi WEKA për secilin nga tre algoritmet*

**Algoritmin LogitBoost.**

.

Klasifikuesi LogitBoost është zhvilluar nga Friedman (2000).LogitBoost performon një regresion logjik shtesë dhe gjeneron modele individuale (fj) në çdo interacion. Ky algoritëm përshtat një model (fj) tek një version i ponderuar i dataset-it origjinal bazuar në disa vlera fallco (zi) të klasës dhe pesha (wi).

Kur përdoret ky algoritëm në një nyje, ai përdor opsionin cross-validation për të përcaktuar sa iteracione mund të ekzekutohen në një kohë dhe të njëjtën gjë bën kudo nëpër pemë. Kjo heuristikë përmirëson në mënyrë të konsiderueshme kohën e ekzekutimit me një ndikim shumë të vogël tek saktësia. Gjithashtu ky algoritëm mund të zvogëlojë edhe peshat për të përmirësuar kohën e ekzekutimit.
Në seksionin e outputit fillimisht shfaqet skema e përdorur, pastaj jepen informacione në lidhje me emrin e relacionit të përdorur, numrin e instancave të përdorura, numrin e atributeve që ka bashkësia e të dhënave si edhe renditen atributet. Më pas vjen modeli që ka ndërtuar klasifikuesi që tregon se si ai përdor atributet për të marë një vendim.

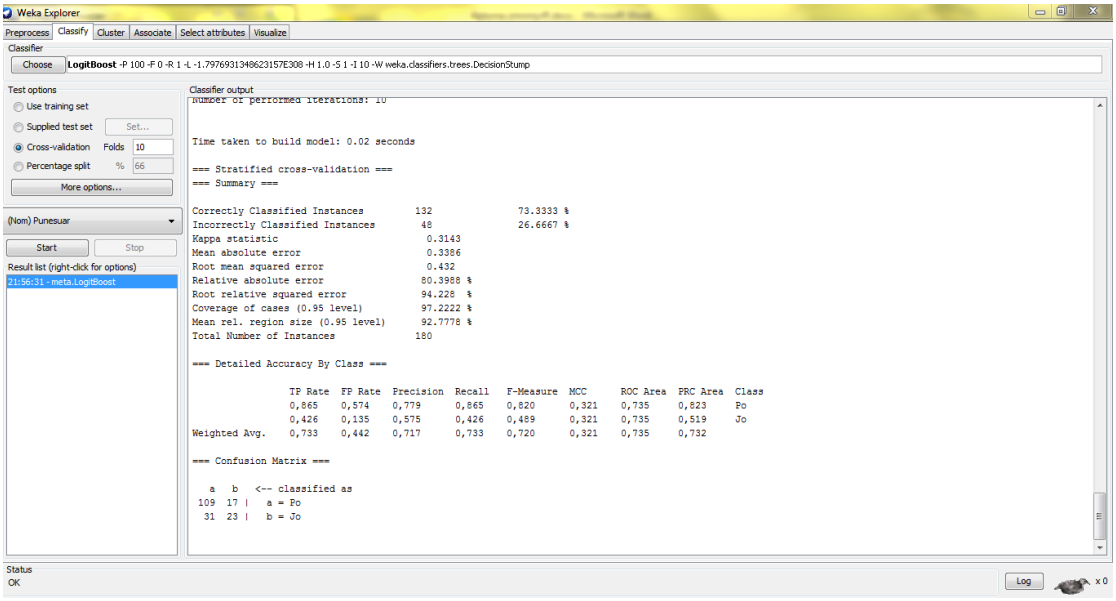



**Figura 4.1 Rezultati pas ekzekutimit të algoritmit LogitBoost**

Shihet se koha e shpenzuar për ndërtimin e modelit është 0.02 sekonda. Nga outputi shfaqen vlerat si më poshtë:

| | | |
|---|---|---|
| *=== Stratified cross-validation ===* | | |
| *=== Summary ===* | | |
| *Correctly Classified Instances* | *132* | *73.3333 %* |
| *Incorrectly Classified Instances* | *48* | *26.6667 %* |
| *Kappa statistic* | *0.3143* | |
| *Mean absolute error* | *0.3386* | |
| *Root mean squared error* | *0.432* | |
| *Relative absolute error* | *80.3988 %* | |
| *Root relative squared error* | *94.228 %* | |
| *Coverage of cases (0.95 level)* | *97.2222 %* | |
| *Mean rel. region size (0.95 level)* | *92.7778 %* | |
| *Total Number of Instances* | *180* | |

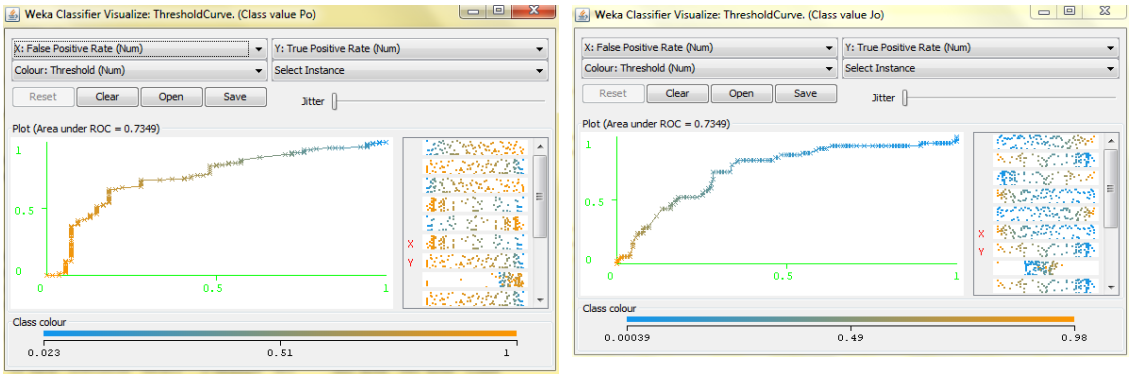
Dy rreshtat e parë tregojnë përkatësisht shkallën e saktësisë dhe shkallën e gabimit. Shohim që 132 instanca janë klasifikuar saktë dhe vetëm 48 prej tyre janë klasifikuar gabim. Shkalla e saktësisë së algoritmit është 73,3 % ndërsa shkalla e gabimit është 26,7 %. Më pas vjen Kappa Statistic që ndiqet nga katër rreshta që nuk janë shumë të dobishëm gjatë klasifikimit sepse ato janë masa që përdoren për të vlerësuar performancën e një parashikimi numerik dhe nuk kanë peshë në këtë rast kur kemi të bëjmë me një klasifikim.

Në output vihen re edhe vlerat e mëposhtme:

*=== Detailed Accuracy By Class ===*

| | *TP Rate* | *FP Rate* | *Precision* | *Recall* | *F-Measure* | *MCC* | *ROC Area* | *PRC Area* | *Class* |
|---|---|---|---|---|---|---|---|---|---|
| | *0,865* | *0,574* | *0,779* | *0,865* | *0,820* | *0,321* | *0,735* | *0,823* | *Po* |
| | *0,426* | *0,135* | *0,575* | *0,426* | *0,489* | *0,321* | *0,735* | *0,519* | *Jo* |
| *Weighted Avg.* | *0,733* | *0,442* | *0,717* | *0,733* | *0,720* | *0,321* | *0,735* | *0,732* | |

Roc (Receiver Operating Characteristic) area paraqet zonën nën një kurbë ROC. Për ta vizualizuar, klikohet tek rezultati i marë, zgjidhet opsioni **Visualize threshold curve** dhe më pas zgjedhim etiketën e klasës për të cilën duam të shohim shpërndarjen. Në figurat e mëposhtme tregohet aplikimi i këtij opsioni për të dy vlerat e klasës:




**(a) Shpërndarja për klasën Po** **(b) Shpërndarja për klasën Jo**

**Figura 4.2 Vizualizimi i parametrit ROC për klasën 'Punësuar"**

Në figurën 4.2 (a) jepet shpërndarja për klasën Po ndërsa në figurën 4.2(b) jepet shpërndarja për klasën Jo. Në aksin X (FP) tregohen vlerat e renditura gabimisht në atë klasë ndërsa në aksin Y tregohen vlerat e renditura saktë në atë klasë (TP). Kurbat ndërtohen duke renditur parashikimet e prodhuara nga klasifikuesi në rendin zbritës të probabilitetit që i shënohet klasës pozitive. Çdo pikë në listë korespondon me vendosjen e një pragu mbi probabilitetin që i është shënjuar klasës pozitive. Numri i pikave në kurbë mund të mos korrespondojë me numrin e instancave në dataset për shkak të probabiliteteve që u shënjohen disa instancave nga klasifikuesi. Duke klikuar një pikë në kurbë hapet një dritare tjetër si më poshtë:

Saktësia që raportohet në Explorer i korrespondon një pike të vetme të kurbës ROC aty ku pragu është përafërsisht 0,5. Kjo kurbë përdoret për të parë performancën e një klasifikuesi të përdorur dhe si ndryshon numri i shembujve të renditur pozitivisht përgjatë iteracioneve që kryen klasifikuesi i LogitBoost.

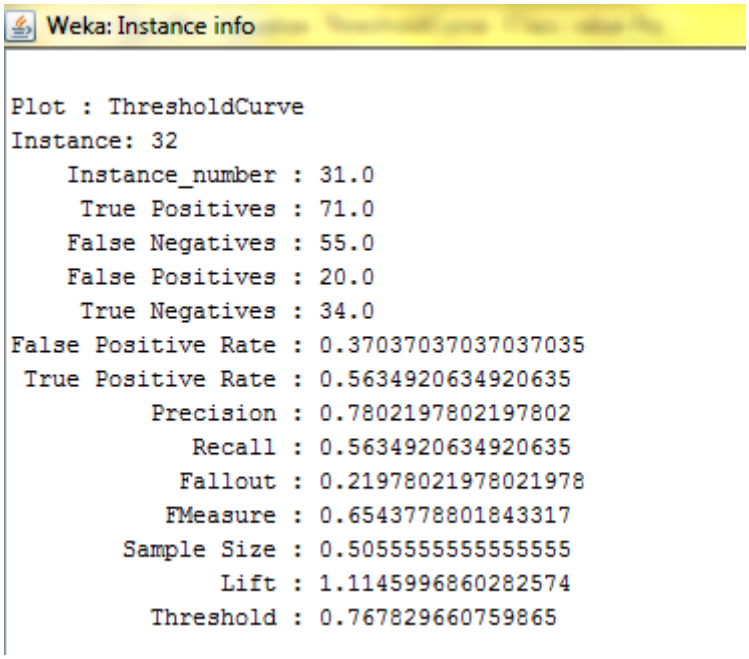



**Figura 4.3 Informacioni për një instancë të kurbës ROC**

PRC (Precision-Recall Curbe) area mund të vizualizohet në të njëjtën mënyrë dhe thjesht duke ndryshuar vlerat e X-ve në Precision dhe vlerat e Y-ve në Recall, për të parë shpërndarjen për një klasë të caktuar.

Pjesa më e rëndësishme e outputit është Matrica e çrregullimeve:

*=== Confusion Matrix ===*

| *a* | *b* | <-- | *classified as* |
|---|---|---|---|
| *109* | *17* | \| | *a = Po* |
| *31* | *23* | \| | *b = Jo* |

Nga këtu shohim që 109 nga 126 shembujt e klasës Po janë klasifikuar në mënyrë të saktë dhe vetëm 17 prej tyre janë klasifikuar gabimisht sikur i përkasin klasës Jo. Ndërsa nga 54 shembujt e klasës Jo, 31 prej tyre janë klasifikuar gabimisht dhe 23 janë klasifikuar saktë.

### Algoritmi AdaBoostM1

AdaBoostM1është një algoritëm që është dizenjuar në mënyrë specifike për klasifikim dhe gjithashtu ai mund të përdoret si model në çdo algoritëm klasifikimi. Ky algoritëm mund të përballojë instancat e ponderuara (që kanë peshë), ku pesha e një instance është numër pozitiv. Trajtimi i instancave të ponderuara ndryshon mënyrën se si llogaritet gabimi. Ai është shuma e peshave të instancave të klasifikuara gabimisht i ndarë me peshën e të gjitha instancave dhe jo nga fraksioni i instancave që janë të klasifikuara gabimisht.

Rezultatet e marra nga ekzekutimi i këtij algoritmi për modelin ‘Mundësi Punësimi’ janë si më poshtë :

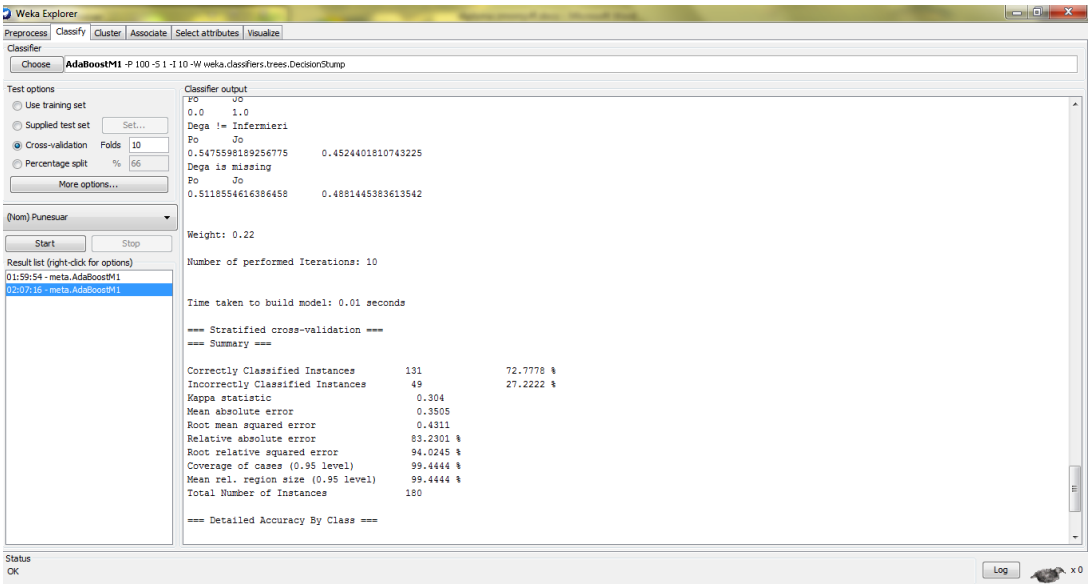



**Figura 4.4Rezultati i marë pas ekzekutimit të algoritmit AdaBoostM1**

Vihet re se algoritmi është ekzekutuar për një kohë prej 0,01 sekondash. 131 prej shembujve janë ekzekutuar në mënyrë të saktë ndërsa 49 janë klasifikuar apo ekzekutuar?? gabim. Numri i iteracioneve është 10 aq sa është edhe numri që kemi përzgjedhur përbri opsionit cross-validation.

Siç mund të vihet re nga matrica e çrregullimeve nga 108 nga 126 shembujt që i takojnë klasës Po vetëm 18 janë klasifikuar gabim, ndërsa nga 54 shembujt që i takojnë klasës Jo 31 janë klasifikuar gabim dhe vetëm 23 janë klasifikuar saktë.

*=== Confusion Matrix ===*

*a b <-- classified as*

*108 18 | a = Po*

*31 23 | b = Jo*

Për të zvogëluar shkallën e gabimit kemi eleminuar disa atribute të cilat mund të jenë të tepërta dhe të parëndësishme për të bërë një parashikim të saktë. Në panelin *Preprocess* përzgjidhen për t’u fshirë atributet: Gjinia, Vendi-ku-keni-studiuar, Tipi-i-shkolles, Trajnime dhe Gjuhet-e-huaja.

Riekzekutohet algoritmi me një nëngrup të dataset-it që do përmbajë vetëm pesë atribute: Mosha, Vendbanimi, Mesatarja, Dega, Arsimi dhe klasën Punësuar. Dataset-i i reduktuar tregohet në figurën 4.5. Rezultati i këtij riekzekutimi tregohet në figurën 4.6.

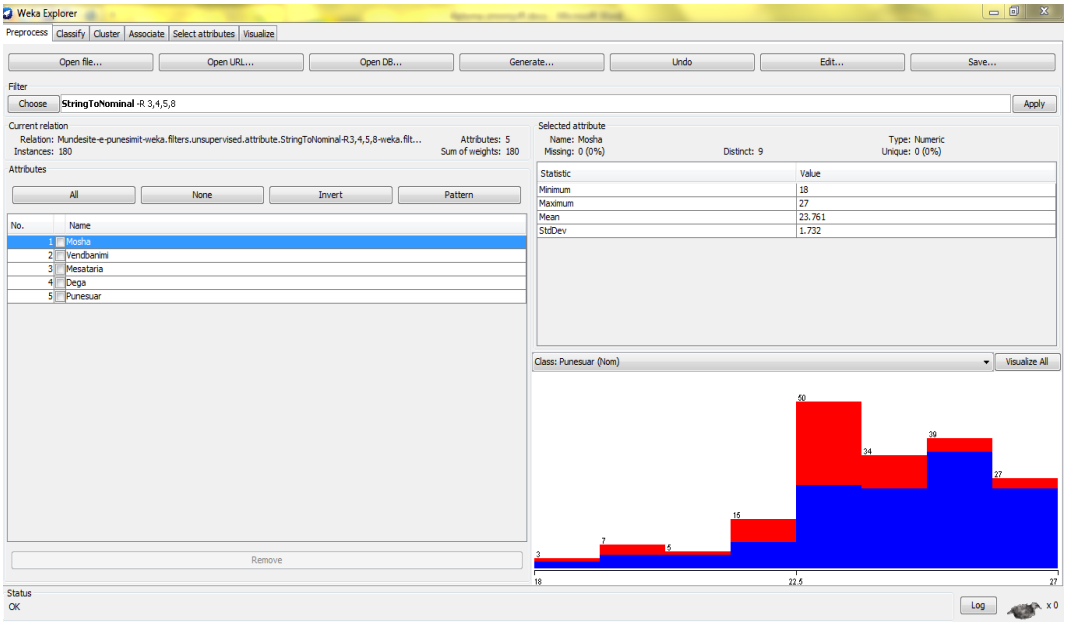



**Figura 4.5**

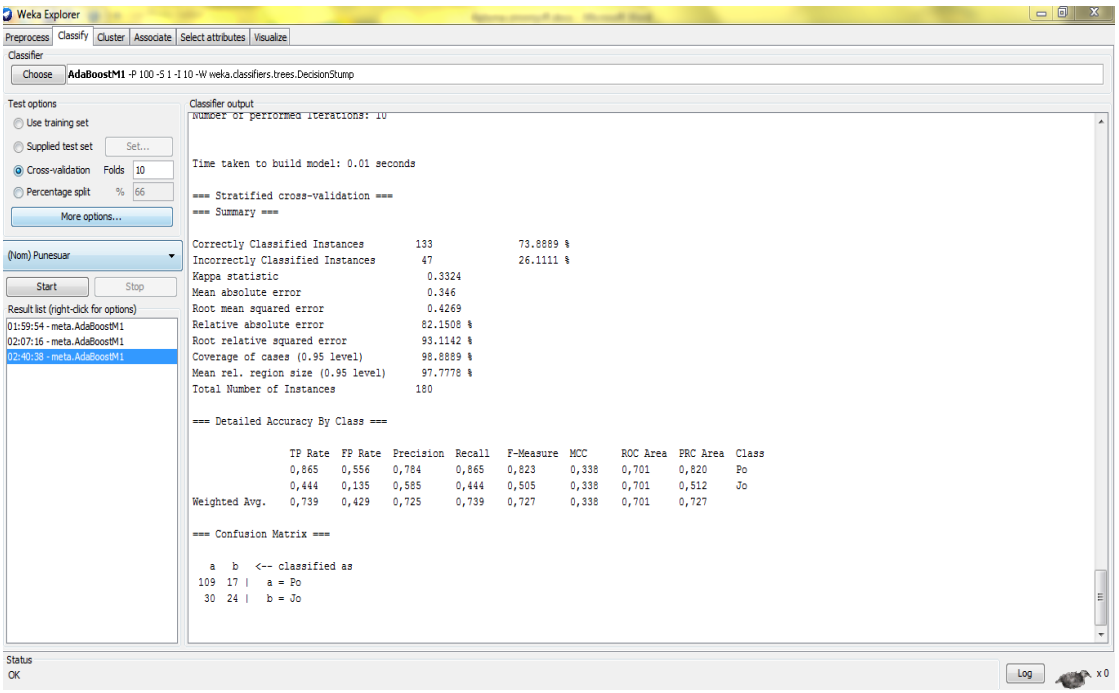



**Figura 4.6 Rezultatet e mara nga riekzekutimi i algoritmit AdaBoostM1 pas fshirjes së atributeve**

Në rastin kur përdoret një dataset i reduktuar vetëm me atributet më të rëndësishme të përzgjedhura shihet se saktësia rritet me afërsisht 2%. Numri i instancave të klasifikuara në mënyrë korrekte është më i madh (133) se më parë. Ndërsa koha për ndërtimin e modelit është sërish 0.01 sekonda, pra nuk ndryshon.

Nëse i hedhim një sy matricës së konfuzionit shohim se nga 126 instanca që i takojnë klasës Po 109 janë klasifikuar në mënyrë korrekte (është rritur me 1 numri i instancave të klasifikuara saktë për klasën Po në krahasim me rastin kur zbatohet algoritmi mbi dataset-in e plotë) dhe vetëm 17 prej tyre janë klasifikuar gabim sikur i përkasin klasës Jo.

Nga 54 instancat që i takojnë klasës Jo, 30 janë klasifikuar gabim (është reduktuar me 1 numri i shembujve të klasifikuar gabim sikur i takojnë klasës Jo) dhe 24 janë klasifikuar në mënyrë korrekte.

```
=== Confusion Matrix ===

  a   b   <-- classified as

109  17 |   a = Po

 30  24 |   b = Jo
```

Siç vihet re atributet që u eleminuan nuk ndikonin shumë për të realizuar parashikime të sakta. Atributet e përzgjedhura janë ato që kanë më shumë ndikim në rastin kur përdoret algoritmi AdaBoostM1. Nga kjo mund të derivojmë në përfundime se disa nga kriteret më të rëndësishme janë Mesataria dhe Dega.

### Algoritmi NaiveBayesUpdateable

NaiveBayesUpdateable është një version i modifikueshëm i klasifikuesit NaiveBayes. Ky klasifikues përdor një preçision 0.1 kur klasifikuesi thërritet për të trajnuar 0 shembuj. Klasifikuesi NaiveBayesUpdatable suporton atributet nominale, numerike, atributet nominale boshe, atributet binare si edhe atributet që mungojnë. Gjithashtu ai suporton vlerat binare, nominale dhe boshe të klasës. Ky klasifikues mund të ekzekutohet edhe kur numri minimal i instancave që ka klasa është 0.

Rezultati i ekzekutimit të algorimit është paraqitur në figurën 4.7.

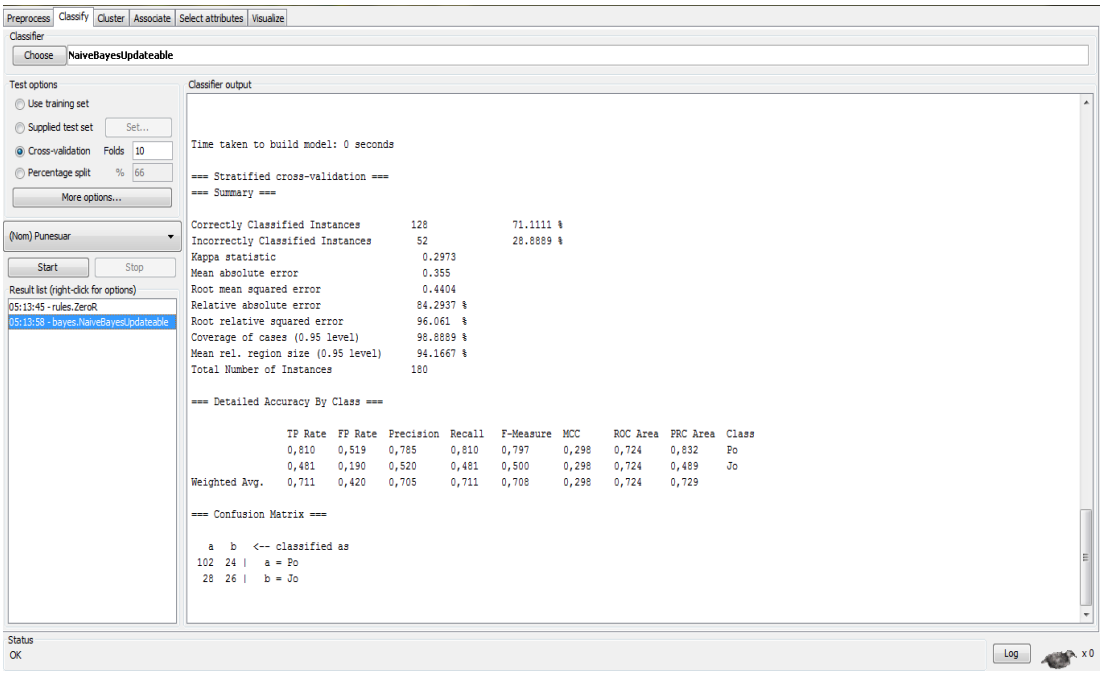



**Figura 4.7 Rezultati i marë nga ekzekutimi i algoritmit NaiveBayesUpdatable.**

Në rastin kur përdoret ky klasifikues koha për ndërtimin e modelit është shumë e vogël (0.01 sekonda). Instancat e klasifikuara në mënyrë të saktë janë 128 dhe vetëm 52 prej tyre janë klasifikuar gabim. Saktësia është 71% ndërsa shkalla e gabimit është 28%. Për informacione më të detajuara i referohemi matricës së çrregullimit të paraqitur më poshtë :

```
=== Confusion Matrix ===

  a   b   <-- classified as
102  24 |   a = Po
 28  26 |   b = Jo
```

Nga 128 instanca që i përkasin klasës Po, 102 janë klasifikuar në mënyrë korrekte dhe vetëm 24 janë klasifikuar gabim sikur ti përkisnin klasës Jo. Nga 54 instancat që i përkasin klasës Jo, 28 janë klasifikuar gabim sikur ti takonin klasës Po dhe 26 prej tyre janë klasifikuar në mënyrë korrekte.
Duke qënë se rezultatet nuk janë shumë të kënaqshme atëherë edhe në këtë rast është zgjedhur të selektohen vetëm disa atribute, duke elimininuar atributet Vendi-ku-keni-studiuar, Tipi-i-shkollës dhe Gjuhët-e-huaja.
Në figurën 4.8 është dhënë paraqitja grafike e rezultatit të marrë pasi është riekzekutuar algoritmi NaiveBayesUpdatable, për dataset-in e reduktuar. Në këtë rast vihet re se saktësia është rritur.

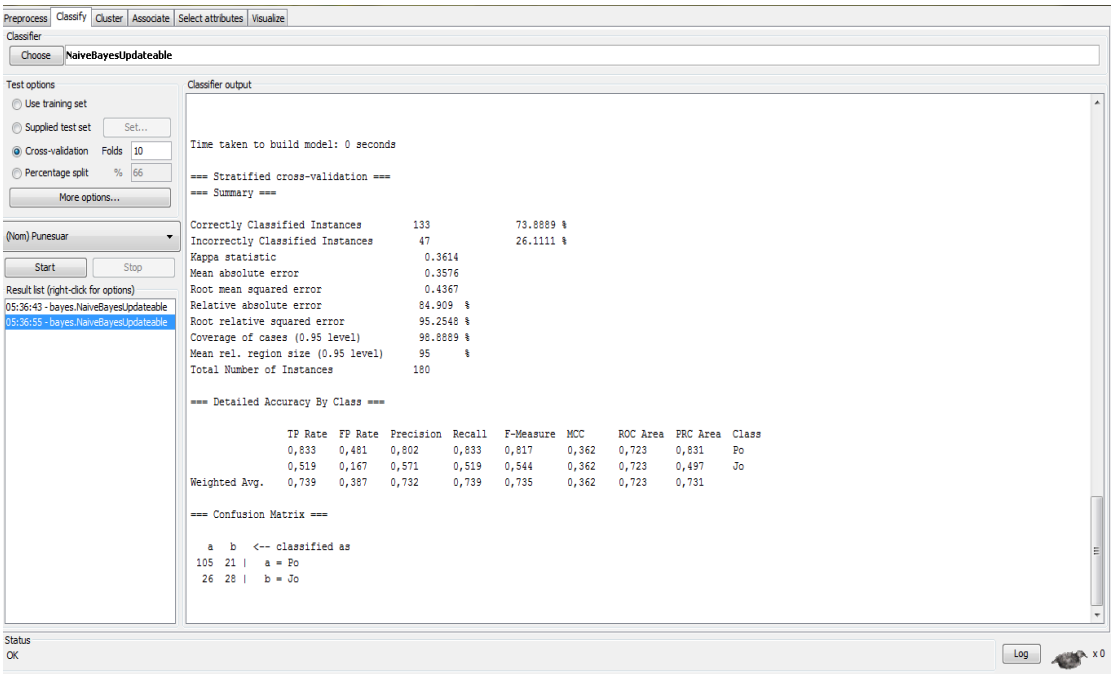



**Figura 4.8Rezultatet e mara pas riekzekutimit të klasifikuesit NaiveBayesUpdatable mbi dataset të reduktuar.**

Shkalla e saktësisë është 73,89%, ndërsa shkalla e gabimit është 26,11%. Nga 180 shembujt e përgjithshëm që përmban dataset-i, 133 prej tyre janë klasifikuar në mënyrë korrekte dhe vetëm 47 janë klasifikuar gabim. Në dallim nga rasti kur përdoreshin të gjithë atributet e dataset-it për klasifikim, kur ekzekutohet klasifikuesi

në një dataset me të dhëna të reduktuara saktësia është rritur me 2% dhe 5 instanca më tepër janë klasifikuar drejt. Koha e ekzekutimit është 0 sekonda. Nga matrica e konfuzionit shohim se 105 nga 126 instancat që i takojnë klasës Po janë klasifikuar saktë, ndërsa një numër i papërfillshëm prej 21 instancash të klasës Po është klasifikuar gabim sikur i takojnë klasës Jo. Nga 54 instancat e klasës Jo 26 prej tyre janë klasifikuar gabim ndërsa 28 janë klasifikuar drejt.

```
=== Confusion Matrix ===

  a   b   <-- classified as
105  21 |   a = Po
 26  28 |   b = Jo
```

Nga të tre algoritmat e përdorur, ky është algoritmi i vetëm që arrin të klasifikojë saktë një numër më të madh shembujsh të klasës Jo.

### *B. Analiza e Rezultateve të tre algoritmeve*

Në tabelat 4.4 dhe 4.5 jepen rezultatet e përmbledhura të ekzekutimit të tre algoritmeve të klasifikimit të data mining.

**Tabela 4.4 Krahasimi i performancës së algoritmeve LB, NBU dhe AB1**

| Kriteret e vlerësimit | Klasifikuesit | | |
|---|---|---|---|
| | **NBU** | **LB** | **AB1** |
| **Koha për ndërtimin e modelit (në sec)** | 0 | 0.03 | 0.01 |
| **Instancat e klasifikuara saktë** | 133 | 132 | 133 |
| **Instancat e klasifikuara josaktë** | 47 | 48 | 47 |
| **Saktësia e parashikimit** | 73.89% | 73.33% | 73.89% |

**Tabela 4.5 Krahasimi në bazë të klasës 'Punësuar' për algoritmet NBU, LB dhe AB1**

| Klasifikuesi | TP | FP | Preçizioni | Recall | F-Measure | Klasa |
|---|---|---|---|---|---|---|
| **NBU** | 0,833 | 0,481 | 0,802 | 0,833 | 0,817 | Po |
| | 0,519 | 0,167 | 0,571 | 0,519 | 0,544 | Jo |
| **LB** | 0,865 | 0,574 | 0,779 | 0,865 | 0,820 | Po |
| | 0,426 | 0,135 | 0,575 | 0,426 | 0,489 | Jo |
| **AB1** | 0,865 | 0,556 | 0,784 | 0,865 | 0,823 | Po |
| | 0,444 | 0,135 | 0,585 | 0,444 | 0,505 | Jo |

Performanca e tre modeleve u vlerësua në bazë të tre kritereve: saktësia e parashikimit, koha për ndërtimin e modelit dhe shkalla e gabimit, të cilat jepen në tabelën 4.4.
Nga matrica e çrregullimeve e dhënë në Tabelën 4.6 vihet re që NBU, LB dhe AB1 prodhojnë rezultate relativisht të mira. Rezultatet sugjerojnë fuqimisht që data mining mund të ndihmojë në parashikimin e suksesit të të rinjve në tregun e punës (punesuar: Po ose Jo).
Pas ekzekutimit të çdo algoritmi në të majtë të panelit të klasifikimit rreshtohen të gjithë algoritmat e përdorur dhe të gjithë rezultatet e mara. Kjo është një lehtësi që ofron Weka për të krahasuar algoritmet shumë thjesht, shpejt dhe për të përzgjedhur rastin më optimal.

**Tabela 4.6 Matrica e çrregullimeve për modelin "Mundësia e Punësimit"**

| | | **Klasat E Parashikuara** | | **Klasifikuesit** |
|---|---|---|---|---|
| | | **A – Po** | **B - Jo** | |
| ***Klasat Aktuale*** | **A – Po** | *105* | *21* | **NBU** |
| | **B - Jo** | *26* | *28* | |
| | | | | |
| | **A - Po** | *109* | *17* | **LB** |
| | **B - Jo** | *31* | *23* | |
| | | | | |
| | **A - Po** | *109* | *17* | **AB1** |
| | **B - Jo** | *30* | *24* | |

Qëllimi i këtij studimi është të gjendet algoritmi që bën parashikimin më të saktë. Nga të tre rezultatet e mara (Tabela 4.8) klasifikuesi *LogitBoost* është algoritmi që ka saktësinë më të vogël dhe kohën e ekzekutimit më të lartë në krahasim me dy të tjerët. Në këtë rast studimor ai nuk është shumë i përshtatshëm për t'u përdorur. Algoritmi *AdaBoostM1* ka një saktësi të njëjtë me algoritmin NaiveBayesUpdatable por koha e ekzekutmit të tij është 0,01 sekonda, pra më e lartë se koha e ekzekutimit të NaiveBayesUpdatable. Algoritmi AdaBoostM1 arrin të klasifikojë më shumë instanca të klasës Po në mënyrë të saktë (109) sesa algoritmi NaiveBayesUpdatable që arrin të klasifikojë saktë vetëm 106 instanca nga 126 shembujt që i takojnë klasës Po.
Avantazhi i algoritmit *NaiveBayesUpdatable* qëndron tek klasifikimi i shembujve të klasës Jo. Deri në këtë moment të dy algoritmat e përdorur paraprakisht siguronin një saktësi mjaft të lartë për parashikimin e klasës Po, ndërsa për klasën Jo gjithmonë numri i shembujve të klasifikuar saktë ishte më i vogël se numri i shembujve të klasifikuar gabim. Me algoritmin *NaiveBayesUpdatable* për herë të parë numri i shembujve që i takojnë klasës Jo dhe që ishin renditur saktë ishte më i madh se numri i shembujve të klasës Jo të renditur gabim. Ekzaktësisht nga 54 shembuj që i takojnë në të vertetë klasës Jo, 28 prej tyre u klasifikuan saktë dhe vetëm 26 u klasifikuan sikur i takonin klasës Po. Nëse i hedhim një sy tabelës së saktësisë (tabela 4.9) së marë si output nga klasifikuesit:

- *Vlera shkallës së vërtetë positive (TP)*

    - Për ***klasë "PO"*** - për klasifikuesin NaiveBayesUpdatable vlera e TP për klasën "Po" (0.841) është më e vogël se e njëjta vlerë e klasifikuesit AdaBoostM1 (0.865) dhe kjo pritej sepse AdaBoostM1 klasifikon saktë një numër më të madh shembujsh të klasës Po në krahasim me algoritmin NaiveBayesUpdatable.
    - Për ***klasën "JO"*** - E kundërta ndodh nëse i referohemi vlerave TP për klasën Jo. Vlera TP e klasës Jo për algoritmin NaiveBayesUpdatable është 0.519, pra më e lartë se TP e AdaBoostM1 që është vetëm 0.444 dhe kjo nënkuptohet sepse NaiveBayesUpdatable klasifikon saktë më shumë instanca të klasës Jo sesa AdaBoostM1 (raporti është 28 me 24 ).
- *Vlera shkallës së gabuar pozitive (FP)* - Vlera FP për klasën Po sipas algoritmit AdaBoostM1 është 0.556më e lartë se ajo e algoritmit NaiveBayesUpdatable 0.481, ndërsa për sa i takon klasës Jo AdaBoostM1 ofron një gabim fare pak më të vogël se algoritmi NaiveBayesUpdatable (0.135 në raport me 0.159). Vlera e FP duhet të jetë sa më e vogël sepse ajo tregon shkallën e gabimit në secilën klasë. Nëse i referohemi vlerës mesatare për të dy klasat kuptojmë që edhe në këtë rast algoritmi NaiveBayesUpdatable do të ishte zgjidhja më e mirë sepse ka një shkallë gabimi mestare më të ulët në krahasim me algoritmin AdaBoostM1 (0.385 në raport me 0.429), kjo jep më shumë siguri për të ardhmen.
- *Precision dhe Recall* - NaiveBayesUpdatable ka Precision dhe Recall më të lartë se klasifikuesi AdaBoostM1 (vlera mesatare e precisionit sipas NaiveBayesUpdatable është 0.737 dhe vlera e recall-it eshtë 0.744, ndërsa nga algoritmi AdaBoostM1 vlerat përkatësisht janë 0.725 dhe 0.739).

Nga krahasimet e bëra algoritmi NaiveBayesUpdatable jo vetëm që ka një kohë ekzekutimi më të vogël, ofron një saktësi më të madhe sidomos për shembujt e klasës "JO" me të cilët dy algoritmat e tjerë "dështuan", vlera mesatare të TP, të Precision-it, të Recall-it, të F-Measure dhe MCC më të larta se në rastin tjetër, ndërsa vlera e FP që nuk është një tregues pozitiv është më e vogël krahasuar me klasifikimin e marë nga AdaBoostM1. Edhe pse algoritmat kanë një saktësi të njëjtë prej 73.8889 % treguesit e tjerë të marë nga outputi bëjnë që klasifikuesi NaiveBayesUpdatable të jetë zgjedhja e duhur për të kryer parashikime në të ardhmen për raste të ngjashme me një saktësi dhe shpejtësi më të mirë.

Duke u bazuar në këto rezultate të rinjtë që janë duke nisur studimet duhet të jenë më të kujdesshëm në profilin që ndjekin të studiojnë, pasi i duhet kushtuar rëndësi edhe kërkesave të tregut të punës. Studentët që janë në vazhdim të studimeve duhet të përqëndrohen më shumë për të arritur rezultate sa më të kënaqshme sepse mesatarja është një nga kriteret që vleresohet sot në tregun e punës. Siç vihet re edhe vendbanimi ka një peshë të rëndësishme, sepse ndryshojnë mundësitë që u ofrohen të rinjve, ndryshojnë përgjegjësitë e punës, shanset për t'u rritur profesionalisht, pagesa etj. Po aq e rëndësishme është edhe mosha. Të rinjtë që sapo kanë mbaruar një universitet shpesh herë nuk kanë eksperiencën dhe as përgatitjen e duhur profesionale për të gjetur një punë menjëherë, që të paguhet mirë, apo të jetë e sigurtë për një kohë më të gjatë etj; ndërsa të rinjtë që kanë studiuar master apo doktoraturë normalisht që kanë edhe një përgatitje më të mirë profesionale apo eksperienca të vogla të akumuluara që i bën ata më të përshtatshëm për një punë më të mirë dhe më të sigurtë.

### *4.3.2 Rasti II: Parashikimin e performancës së studentëve, duke përdorur të dhënat e mbledhura nga sistemi menaxhimit të studentëve, të një fakulteti të Universitetit të Tiranës*

Synimi i këtij studimi është interpretimi i rezultatit të algoritmeve të klasifikmit, pemët e vendimit, duke u bazuar në dy algoritmet e klasifikimit C4.5, RepTree dhe RandomTree. Për të vlerësuar saktësinë e parashikimit këto algoritme janë krahasuar më pas me njëri-tjetrin.

#### *A. Pemët e Vendimit*

Pemët e vendimit janë një metodë klasifikimi e përdorur gjerësisht për klasifikimin e të dhënave. Ato janë të njohura për shkak se në ndërtimin e tyre nuk është e nevojshme vendosja e ndonjë parametri apo kriteri paraprak. Klasifikuesit pemë vendimi janë vetëshpjegues, trajtojnë atribute me vlerë numerike dhe nominale, bëjnë pjesë në grupin e klasifikuesve me vlerë diskrete, janë të afta të trajtojnë baza të dhënash me gabime dhe mungesë të dhënash dhe janë joparametrike (Murthy 1998). Një pemë vendimi është një strukturë peme si flowchart, ku çdo nyje e brendshme shënon një test në një atribut, çdo degë përfaqëson rezultatin e testit dhe çdo gjethe përmban një etiketë klase. Nyja e parë e pemës është rrënja. Disa algoritme pemë vendimi prodhojnë vetëm pemë binare (ku çdo nyje e brendshme degëzohet në vetëm dy nyje të tjera), ndërkohë që të tjera mund të prodhojnë pemë jo binare. Shumica e algoritmeve për pranimin e pemës së vendimit fillojnë me një bashkësi provë dhe më pas ndahet në mënyrë rekursive në bashkësi më të vogla ndërkohë që ndërtohet pema. Algoritme të zhvilluara sipas kësaj metode përfshijnë ID3, C4.5, CART, RandomTree, Random Forest, etj.

Në këtë studim janë shqyrtuar tre algoritme C4.5 (J48 nëWeka), RepTree dhe Random Tree për të ndërtuar një pemë vendimi nga një bashkësi të dhënash. Pema e vendimit e gjeneruar nga këto algoritme mund të përdoret për klasifikimin e rregullave, për këtë arsye ata njihen ndryshe dhe si *klasifikues statistikorë*.

*Algortimi C4.5*

Pas përpunimit të bashkësisë së të dhënave, ngarkohet bashkësia në programin WEKA dhe ekzekutohet algoritmit J48 (Versioni i algoritmit C4.5 në programim WEKA). (Rezultatin e këtij algoritmi e gjeni në Aneksin 2).

Për klasën ST_MES_3, atributi i cili ka ndikimin më të lartë sipas testit Gain Ratio më të ulët është atributi ST_MES_1 (tabela 4.2, për këtë arsye ky atribut do të jetë dhe rrënja e pemës së vendimit. Atributi pasardhës me ndikim më të lartë është atributi ST_MES_2, ky atribut përmban dhe bijtë e rrënjës së pemës së vendimit. Ky proçes do të realizohet dhe për të gjithë atributet e tjerë për të ndëtuat nivelet pasardhëse të pemës së vendimit. Rezultati i pemës së vendimit për këtë rast studimi është dhënë në aneksin 2, ndërsa në figurën e mëposhtme jepet një paraqitje grafike e pemës së vendimit të gjeneruar për këtë eksperiment:

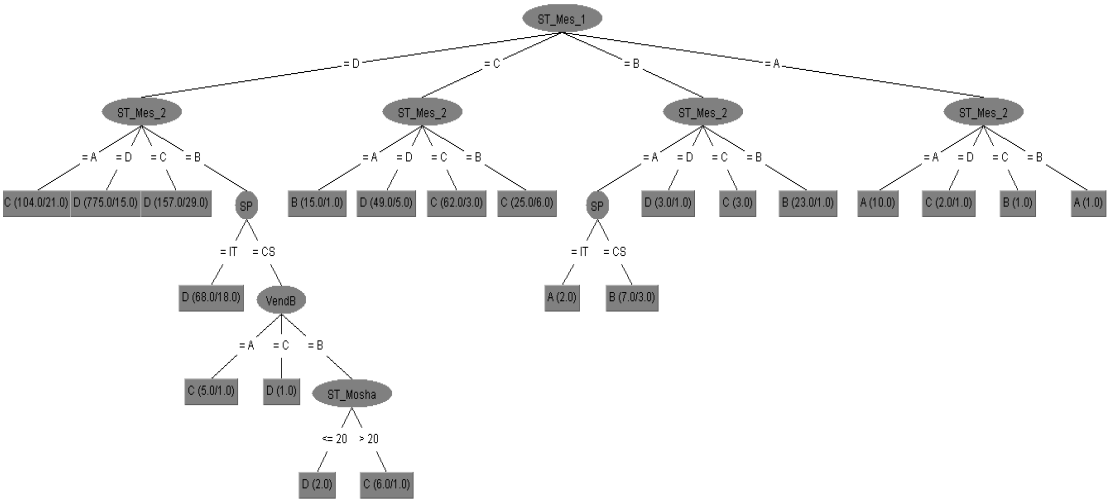



**Figura 4.9 Pema e vendimit për rastin II të studimit të gjeneruar nga ekzekutimi i algoritmit J48**

Gjethet e pemës përfundimtare përmbajnë vlerat e klasës së studimit, gjithashtu dhe numrin e instancave që janë klasifikuar në këtë klasë dhe numrin e instancave që janë klasifikuar gabim. Si për shembull gjethja me vlerën C në nivelin e fundit tregon se 6 instanca janë klasifikuar në në këtë klasë ku njëra prej tyre është klasifikuar gabim.

Më poshtë jepet vlerësimi i algoritmit J48, i cili është ekzekutuar tre herë, për tre vlerësime të ndryshme Training Set, dhe vlerësime të kryqëzuar të3 -fishtë dhe të 13-fishtë.
Vlerësimi i 3-fishtë është realizuar për të marrë një parashikim më të saktë. Mesatarisht ky algoritëm jep një saktësi parashikimi 91.40% dhe ekzekutohet mesatarisht prej 0.04 sekondash.

**Tabela 4.7 Vlerësimi i plotë i ekzekutimit të algoritmit J48.**

| Parametri/Testi | Training Set | Kryqëzimi i 3-fishtë | Kryqëzimi i 13-fishtë | Mesatarja |
|---|---|---|---|---|
| **Koha e ndërtimit të modelit (Sek)** | 0.08 | 0.02 | 0.02 | 0.04 |
| **Instanca të klasifikuara saktë** | 1215 | 1203 | 1204 | 1207.333 |
| **Instanca të klasifikuara gabim** | 106 | 118 | 117 | 113.6667 |
| **Saktësia e parashikimit** | 91.976% | 91.067% | 91.143% | 91.395% |
| **Gabimi absolut mesatar** | 0.6660 | 0.7130 | 0.7150 | 0.6980 |

**AlgortimiREPTree**

Pas përpunimit të bashkësisë së të dhënave, ngarkohet bashkësia në programin WEKA dhe ekzekutohet algoritmit REPTree (Rezultatin e këtij algoritmi e gjeni në Aneksin 2).

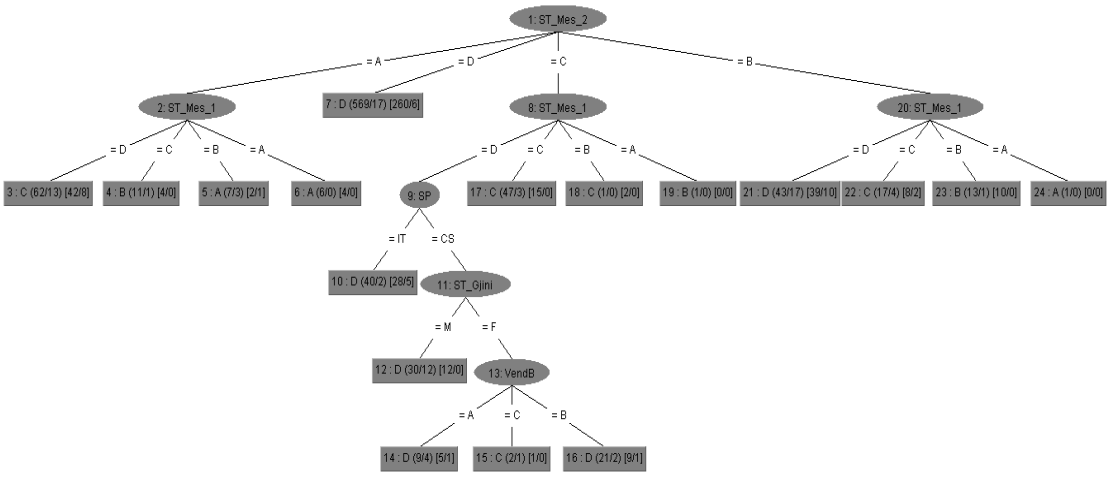



**Figura 4.10 Pema e vendimit për rastin II të studimit të gjeneruar nga ekzekutimi i algoritmit REPTree**

Më poshtë jepet vlerësimi i algoritmit REPTree,i cili është ekzekutuar tre herë, për tre vlerësime të ndryshme Training Set, dhe vlerësime të kryqëzuar të3-fishtë dhe të 13-fishtë.
Vlerësimi i 3fishtë është realizauar për të marrë një parashikim më të saktë. Mesatarisht ky algoritëm jep një saktësi parashikimi 90.89% dhe ekzekutohet mesatarisht prej 0.08 sekondash.

**Tabela 4.8 Vlerësimi i plotë i ekzekutimit të algoritmit REPTree.**

| Testi/Parametri | Training Set | Kryqëzimi i 3-fishtë | Kryqëzimi i 13-fishtë | Mesatarja |
|---|---|---|---|---|
| **Saktësia e parashikimit** | 91.370% | 90.462% | 90.840% | **90.89%** |
| **Instanca të klasifikuara saktë** | 1207 | 1195 | 1200 | **1,200.67** |
| **Instanca të klasifikuara gabim** | 114 | 126 | 121 | **120.33** |
| **Koha e ndërtimit të modelit (Sek)** | 0.11 | 0.11 | 0.03 | **0.08** |
| **Gabimi absolut mesatar** | 0.0694 | 0.0007 | 0.0719 | **0.0473** |

## Algoritmi Random Tree

Pas përpunimit të bashkësisë së të dhënave, ngarkohet bashkësia në programin WEKA dhe ekzekutohet algoritmi Random Tree (Rezultatin e këtij algoritmi e gjeni në Aneksin 2). Më poshtë jepet vlerësimi i algoritmit Random Tree, cili është ekzekutuar tre herë, për tre vlerësime të ndryshme Training Set, dhe vlerësim të kryqëzuar të3-fishtë dhe të 13-fishtë.

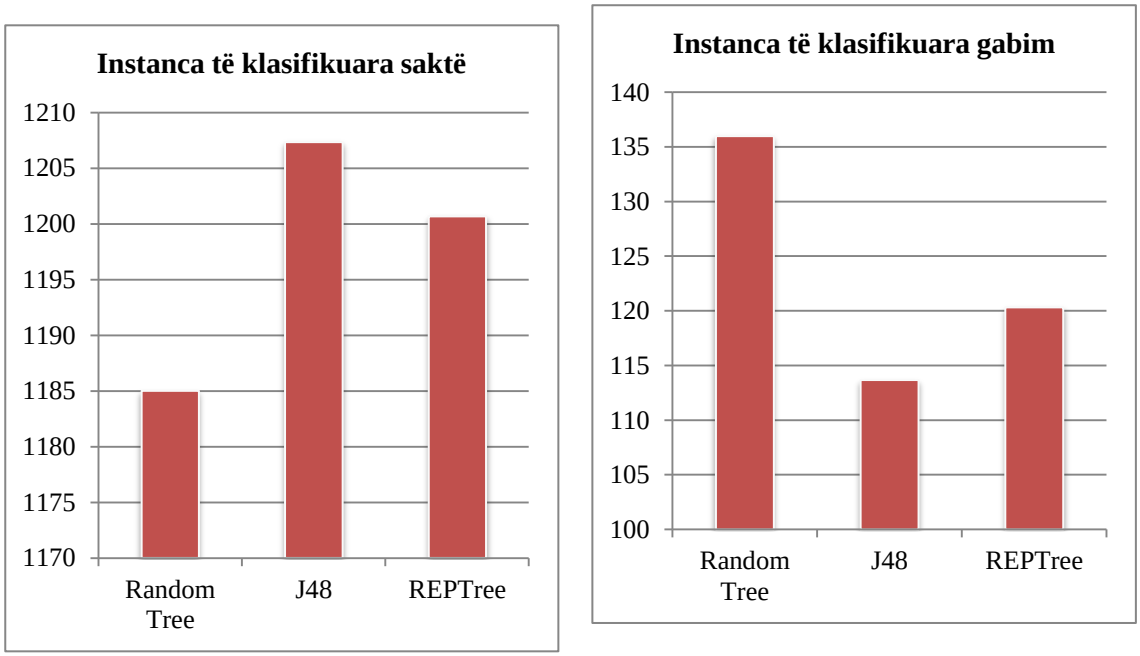
Vlerësimi i 3-fishtë është realizuar për të marrë një parashikim më të saktë. Mesatarisht ky algoritëm jep një saktësi parashikimi 89.70% dhe ekzekutohet mesatarisht prej 0.06 sekondash.

**Tabela 4.9 Vlerësimi i plotë i ekzekutimit të algoritmit Random Tree.**

| Testi/Parametri | Training Set | Kryqëzimi i 3-fishtë | Kryqëzimi i 13-fishtë | Mesatarja |
|---|---|---|---|---|
| **Saktësia e parashikimit** | 93.263% | 86.828% | 89.024% | **89.70%** |
| **Instanca të klasifikuara saktë** | 1232 | 1147 | 1176 | **1,185.00** |
| **Instanca të klasifikuara gabim** | 89 | 174 | 145 | **136.00** |
| **Koha e ndërtimit të modelit (Sek)** | 0.16 | 0 | 0.02 | **0.06** |
| **Gabimi absolut mesatar** | 0.0694 | 0.0007 | 0.0703 | **0.0468** |

### *B. Krahasimi i tre algoritmeve dhe analiza e rezultateve*

Në grafikët 4.1 dhe 4.2 dhe tabelën 4.10 është paraqitur një krahasim i vlerave mesatare të performancës së algoritmeve J48, REPTree dhe Random Tree.

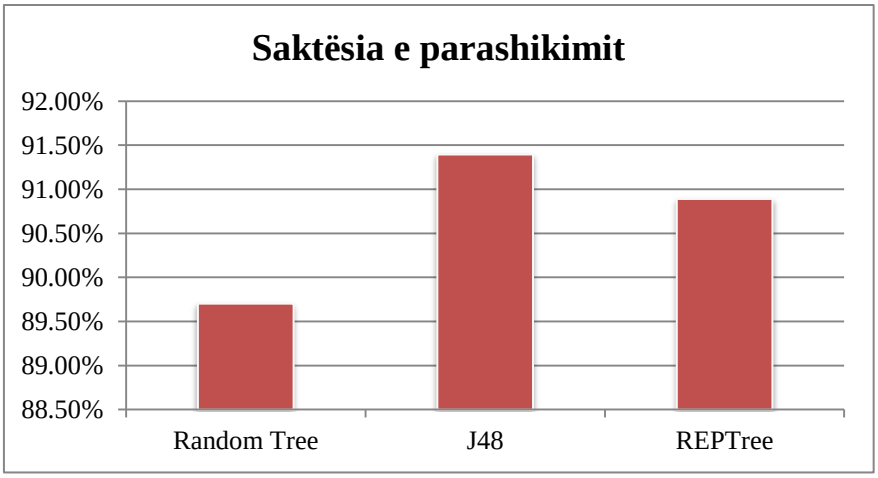



***(a) Instanca të klasifikuara saktë*** ***(b) Instanca të klasifikuara gabim***

**Grafiku 4.1 Instancat e klasifikuara sipas algoritmeve J48, RepTree dhe RandomTree**

Ashtu siç mund të vihet re nga grafiku 4.1 algoritmi J48 klasifikon më shumë instanca (1207 instanca) të sakta sesa dy të tjerët (repsektivish 1185 instanca algoritmi Random Tree dhe 1200 algoritmi REP Tree). Ky algoritëm ka dhe saktësinë më të lartë të parashikimit 91.34 % dhe kohën e ekzekutimit 0.04. Në tabelën 4.11 janë dhënë parametrat e tjerë të performancës së algoritmit J48 për tre testet e realizuara.

Pra mund të themi që për këtë rast studimi, algoritmi J48 jep parashikimin më të mirë dhe më të shpejtë. Në tabelën 4.11 kemi dhënë parametrat e tjerë për matjen e performancës, për të gjetur secili nga testet jep parashikimin më të mirë.




**Grafiku 4.2 Saktësia e parashikimit për tre algoritmet J48, REPTree dhe Random Tree**

**Tabela 4.10 Parametrat e krahasimit të algoritmeve J48, RepTree dhe Random Tree**

| ***Testi/Parametri*** | ***Random Tree*** | ***J48*** | ***REPTree*** |
|---|---|---|---|
| **Instanca të klasifikuara saktë** | 1185 | 1207.33 | 1200.67 |
| **Instanca të klasifikuara gabim** | 136 | 113.67 | 120.33 |
| **Koha e ndërtimit të modelit (Sek)** | 0.060 | 0.040 | 0.083 |
| **Gabimi absolut mesatar** | 0.047 | 0.070 | 0.047 |
| **Saktësia e parashikimit** | 89.70% | 91.40% | 90.89% |

**Tabela 4.11 Krahasimi në bazë të klasës ST_MES_3 për tre testet e zhvilluar**

| **Testi / Parametri** | **TP** | **FP** | **Preçizioni** | **Recall** | **Klasa** |
|---|---|---|---|---|---|
| **Training Set** | 0.725 | 0.031 | 0.841 | 0.725 | *A* |
| | 0.984 | 0.214 | 0.936 | 0.984 | *B* |
| | 0.672 | 0.004 | 0.891 | 0.672 | *C* |
| | 0.765 | 0 | 1 | 0.765 | *D* |
| **Vlerësimi i kryqëzuar 3-** | 0.704 | 0.033 | 0.824 | 0.704 | *A* |
| | 0.98 | 0.226 | 0.932 | 0.98 | *B* |

| | | | | | |
|---|---|---|---|---|---|
| **fishtë** | 0.607 | 0.003 | 0.902 | 0.607 | *C* |
| | 0.824 | 0.005 | 0.7 | 0.824 | *D* |
| **Vlerësimi i kryqëzuar 13-fishtë** | 0.717 | 0.034 | 0.823 | 0.717 | *A* |
| | 0.979 | 0.217 | 0.934 | 0.979 | *B* |
| | 0.607 | 0.004 | 0.881 | 0.607 | *C* |
| | 0.765 | 0.005 | 0.684 | 0.765 | *D* |

Ashtu siç mund të vihet re nga kjo tabele, vlerat TP, FP, preçizionit dhe Recall janë më të mira për testin Training Set. Kjo vihet re dhe nga tabela 4.7 ku mund të shohim se ky test ka saktësinë e parashikimit më të lartë 91.99 %, por ndryshe nga dy testet e tjera kërkon pak më shumë kohë se të tjerët 0.08 sekonda.

### *C. Nxjerrja e rregullave nga pema e vendimit*

Një klasifikues i bazuar në rregull përdor një bashkësi rregullash *NQS-ATËHERË* për klasifikimin. Një rregull if-then-else shprehet në formën:

*Nqs <kusht> Atëherë <realizo veprim>*

Klasifikuesit pemë vendimi janë një metodë e përhapur e klasifikimit për shkak se mënyrat se si pemët e vendimit funksionojnë është lehtësisht të kuptueshme dhe njihen për saktësinë e tyre. Megjithatë ato mund të bëhen të mëdha dhe të vështira për t'u interpretuar. Për të nxjerrë rregulla nga një pemë vendimi, një rregull krijohet për çdo degë nga rrënja në një nyje gjethe. çdo kriter ndarës gjatë një dege, futet në një *DHE* logjike për të formuar parakushtin (pjesën "*NQS*"). Gjethja përmban parashikimin e klasës duke formuar pasojën e rregullit (pjesën "*ATËHERË*"). Mes rregullave të nxjerra zbatohet një *OSE* logjike. Për shkak se rregullat nxirren direkt nga pema, kemi një rregull për gjethe dhe një tuple-i i shoqërohet vetëm një gjethe, gjithashtu kemi një rregull për çdo kombinim vlerë-atribut.

Në tabelën e mëposhtme kemi listuar disa nga rregullat që mund të nxirren nga pema e algoritmit J48 (figura 4.4). Kolona e parë e tabelës përmban rregullin, në kolonën e dytë është dhënë numri i instancave që klasifikohen në këtë rregull dhe në kolonën e tretë është dhënë numri i instancave të klasifikuara gabim.

Nisur nga këto rregulla, zhvilluesit dhe kërkuesit mund të ndërtojnë një aplikacion i cili mund t'ju japë profesorëve dhe studentëve një informacion mjaft interesant, i cili ju siguron një udhëzues, në mënyrë që të zgjedhin një rrugë të përshtatshme, duke analizuar eksperiencat e studentëve me arritje akademike të ngjashme.

**Tabela 4.12 Disa rregulla të nxjerrja nga pema e vendimit e gjeneruar nga algoritmi J48**

| **Rregulli** | **Nr_Inst#** | **Nr_Inst_gabim#** |
|---|---|---|
| *Nqs ST_MES_1 = D dhe ST_MES_2 = B dhe SP = 'IT' Atëherë ST_MES_3 = D* | 68 | 18 |
| *Nqs ST_MES_1 = D dhe ST_MES_2 = A atëherë ST_MES_3 = C* | 104 | 21 |
| *Nqs ST_MES_1 = A dhe ST_MES_2 = A atëherë ST_MES_3 = A* | 10 | 0 |
| *Nqs ST_MES_1 = D dhe ST_MES_2 = B dhe SP = CS dhe VENDB = B dhe ST_MOSHA>20 atëherë ST_MES_3 = C* | 6 | 1 |

Në mënyrë të ngjashme në mund të nxjerrim dhe rregulla nga pemët e vendimit që janë gjeneruar për dy algoritmet e tjera. Në Tabelën 4.13 kemi dhënë të algoritmit RepTree.

**Tabela 4.13Disa rregulla të nxjerra nga pema e vendimit e gjeneruar nga algoritmi REPTree**

| Rregulli | Nr_Inst# | Nr_Inst_gabim# |
|---|---|---|
| **Nqs ST_MES_2 = D Atëherë ST_MES_3 = D** | 569 | 17 |
| **Nqs ST_MES_2 = C dhe ST_MES_1 = D dhe (SP = IT ose (SP = CS dhe (ST_GJINI = 'M' ose (ST_GJINI = F dhe (VendB =A ose VendB = B)))) atëherë ST_MES_3 = D** | 100 | 20 |
| **Nqs ST_MES_2 = B dhe ST_MES_1 = C atëherë ST_MES_3 = C** | 17 | 4 |

### *4.3.3 Rasti III: Parashikimi i performancës së studentëve duke përdorur të dhënat e mbledhura nga pyetësorët e shkruar*

Në studim janë zhvilluar disa eksperimente në mënyrë që të vlerësojmë performancën dhe dobinë e algoritmeve të ndryshme të klasifikimit (NaiveBayes -NB, J48 dhe MultiLayer Perceptron –MLP) për të parashikuar suksesin e studentit.

#### *A. Analiza e rezultateve*

Synimi i këtij studimi është vlerësimi klasifikuesit dhe të gjëndet klasifikuesi i cili jep rezultatin më të mirë por dhe në formë të kuptueshme për përdoruesin. Rezultatet e eksperimenteve përmblidhen në Tabelat 4.14, 4.15, 4.16.

**Tabela 4.14 Performanca parashikuese e klasifikuesve NB, MLP dhe J48**

| Kriteret e vlerësimit | Klasifikuesit | | |
|---|---|---|---|
| | **NB** | **MLP** | **J48** |
| **Koha për ndërtimin e modelit (në sec)** | 0 | 2.92 | 0.01 |
| **Instancat e klasifikuara saktë** | 161 | 158 | 168 |
| **Instancat e klasifikuara josaktë** | 39 | 42 | 32 |
| **Saktësia e parashikimit ( në %)** | 80.5 | 79 | 84 |

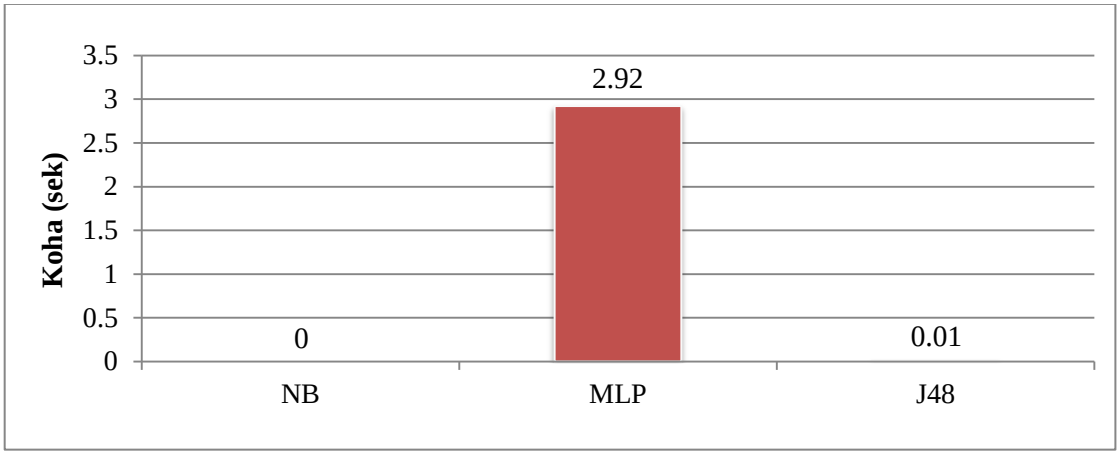
**Tabela 4.15 Krahasimi i vlerësimeve për klasifikuesit NB, MLP dhe J48**

| Kriteret e vlerësimit | Klasifikuesit | | |
|---|---|---|---|
| | **NB** | **MLP** | **J48** |
| **Statistika kappa** | 0.5512 | 0.5464 | 0.6158 |
| **Mean absolut error (MAE)** | 0.1657 | 0.1491 | 0.1725 |
| **Root mean squared error (RMSE)** | 0.3038 | 0.338 | 0.313 |
| **Gabimi absolut relativ (RAE)** | 53.51% | 48.13% | 55.69% |
| **Root relative squared error (RRSE)** | 77.55% | 86.28% | 79.91% |

**Tabela 4.16 Krahasimi i vlerësimeve në bazë të klasave, për klasifikuesit NB, MLP dhe J48**

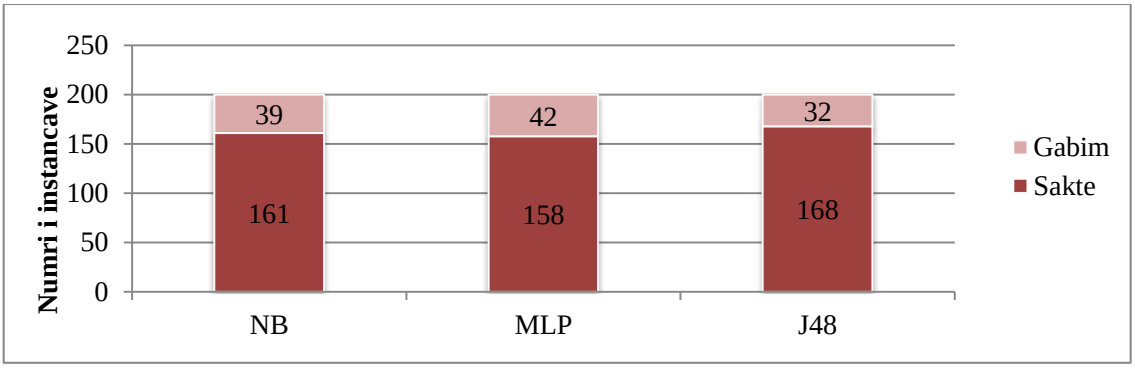
| Klasifikuesi | TP | FP | Preçizioni | Recall | Klasa |
|---|---|---|---|---|---|
| NB | 0.3 | 0.044 | 0.429 | 0.3 | Kalon |
| | 0.893 | 0.4 | 0.839 | 0.893 | KalonM |
| | 0.75 | 0.044 | 0.811 | 0.75 | Mbetet |
| MLP | 0.35 | 0.078 | 0.333 | 0.35 | Kalon |
| | 0.85 | 0.333 | 0.856 | 0.85 | KalonM |
| | 0.8 | 0.05 | 0.8 | 0.8 | Mbetet |
| J48 | 0.35 | 0.006 | 0.875 | 0.35 | Kalon |
| | 0.943 | 0.383 | 0.852 | 0.943 | KalonM |
| | 0.725 | 0.05 | 0.784 | 0.725 | Mbetet |

Grafiku 4.2 tregon kohën që do secila nga skemat e marra në konsideratë për ndërtimin e modelit. Multilayer perceptron, klasifikuesi i rrjetave neurale, merr më shumë kohë për të ndërtuar modelin. Klasifikuesi NaiveBayes dhe i pemës së vendimit, e ndërtojnë më shpejt në kohë modelin për bashkësinë e përcaktuar të të dhënave.




**Grafiku 4.3Koha për ndërtimin e modelit 'Parashikimi i performancës së studentit"**

Grafiku 4.3 tregon instancat e klasifiuara saktë në raport me ato të klasifikuara gabim.




**Grafiku 4.4 Shkalla e gabimit e modelit 'Parashikimi i performancës së studentit"**

Nga të tre rezultatet e marra klasifikuesi *Multilayer Perceptron* është algoritmi që ka saktësinë më të vogël (79%) dhe kohën e ekzekutimit më të lartë (2.92 sek) në krahasim me dy algoritmet e tjera. Algoritmi *NaiveBayes* ka një saktësi (80.5%) dhe koha e ekzekutmit të tij është shumë e ulët (0 sek). Algoritmi *J48* arrin të klasifikojë më shumë instanca të klasës në mënyrë të saktë (168) sesa algoritmi NaiveBayes që arrin të klasifikojë saktë vetëm 161 instanca nga 200 shembujt që i takojnë klasës. Pra, ashtu siç tregohet në grafikun 4.1 algoritmi J48 prashikon më mirë se algoritmet e tjerë të përdorur për eksperimentim.

Performanca e teknikave të të mësuarit është shumë e varur nga natyra e të dhënave trajnuese. Matricat e çrregullimeve janë shumë të vlefshme për vlerësimin e klasifikuesve. Për të vlerësuar fortësinë(fuqinë) e klasifikuesit, metodologjia e zakonshme është performimi i vlerësimit të kryqëzuar mbi klasifikuesin.
Në përgjithësi, vlerësimi i kryqëzuar është provuar të jetë statistikisht i mirë mjaftueshëm në vlerësimin e performancës së klasifikuesve. Nga matrica e çrregullimeve e dhënë në Tabelën 4.17, është vënë re që MLP, NB dhe J48 prodhojnë rezultate relativisht të mira. Rezultatet sugjerojnë fuqimisht që data mining mund të ndihmojë në parashikimin e suksesit të studentëve në një degë (kalojnë ose mbesin).

**Tabela 4.17 Matrica e çrregullimeve për modelin 'Parashikimi i performancës së studentit"**

| Klasifikuesit | A | B | C | Klasa |
|---|---|---|---|---|
| NB | 6 | 14 | 0 | A=Kalon |
| | 8 | 125 | 7 | B=KalonM |
| | 0 | 10 | 30 | C=Mbetet |
| MLP | 7 | 13 | 0 | A=Kalon |
| | 13 | 119 | 8 | B=KalonM |
| | 1 | 7 | 32 | C=Mbetet |
| J48 | 7 | 12 | 1 | A=Kalon |
| | 1 | 132 | 7 | B=KalonM |
| | 0 | 11 | 29 | C=Mbetet |

### *B. Rezultati i studimit*

Në problemin në arsim është shumë e rëndësishme për modelin e klasifikimit të përftuar, që të jetë shoqërisht i përdorshëm, në mënyrë që mësuesit të marrin vendime për të përmirësuar të mësuarit e studentit. Megjithatë, disa modele janë më të interpretueshme se të tjerat (Romero et al. 2008). Edhe pse rezultatet e klasifikuesve janë relativisht të mira, vlerësohet më shumë ai klasifikues i cili jep rezultatin në forma të kuptueshme për përdoruesin. Ky algoritëm në këtë rast është algoritmi J48, i

cili jo vetëm jep rezultatin më të saktë por gjithashtu arrin ta shprehë atë në formën më të kuptueshme për përdoruesin e cila është pema e vendimit.

- Pemët e vendimit konsiderohen si modele lehtësisht të kuptueshme sepse një proces arsyetues mund të jepet për secilin konkluzion. Modelet e njohurive nën këtë paradigëm mund të transformohen në një bashkësi rregullash if-then, që janë një nga metodat më të përhapura të prezantimit të njohurive, për shkak të thjeshtësisë së tyre dhe kuptueshmërisë të cilën profesorët mund ta kuptojnë dhe interpretojnë shumë lehtë.
- Metodat statistike dhe rrjetat neurale janë konsideruar të jenë më pak të përshtatshme për qëllimet e data mining. Modelet e njohurive të përftuara nën këto metoda, zakonisht konsiderohen të jenë si mekanizma të kutive të zeza, të afta për të përftuar saktësi shumë të mirë por japin rezultate shumë të vështira për t'u kuptuar nga njerëzit.

Figura 4.11 tregon pemën e vendimit që rezultoi nga zbatimi i algoritmit J48 mbi bashkësinë e të dhënave të mbledhura, e cila është shumë e lehtë të lexohet dhe të kuptohet.

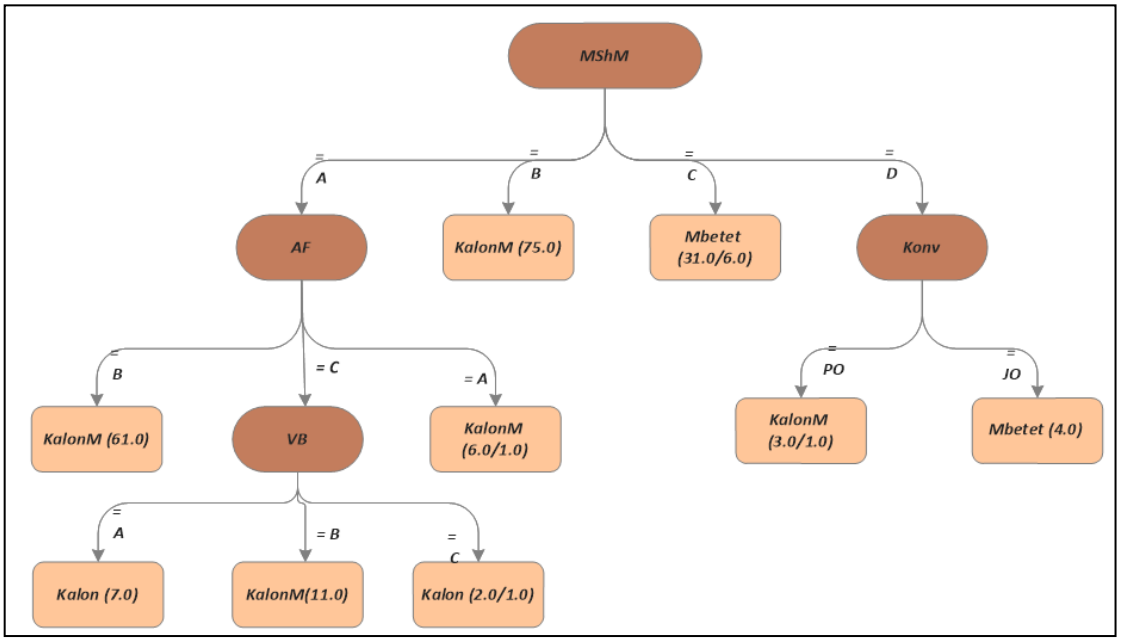



**Figura 4.11 Modeli i Pemës së vendimit të përftuar rasti I**

Edhe pse rezultatet e klasifikuesve janë të mira, vlerësohet më shumë ai klasifikues i cili jep rezultatin në forma të kuptueshme për përdoruesin. Figura 4.1 tregon pemën e vendimit që rezulton nga zbatimi i algoritmit J48 mbi bashkësinë e të dhënave të mbledhura. Ky model, mund t'u japë profesorëve informacion mjaft interesant rreth studentëve dhe u siguron një udhëzues, në mënyrë që të zgjedhin një rrugë të përshtatshme, duke analizuar eksperiencat e studentëve me arritje akademike të ngjashme. *Pra algoritmi J48, i cili përveçse jep rezultatin më të saktë, arrin gjithashtu ta shprehë atë në formën më të kuptueshme për përoduresin*.

# 5. Përfundime

*Në këtë kapitull, jepet një përmbledhje e shkurtër dhe përfundime të të gjithë rasteve të studimit. Në fund, ofrohen sugjerime për kërkime të mëtejshme në lidhje me këtë studim dhe rekomandime për pjesëmarrësit në këtë fushë kërkimore.*

## 5.1 Përfundime

Të gjitha çfarë janë trajtuar në këtë punim tregojnë që fusha e data mining në arsim po zgjerohet me shpejtësi, por gjithashtu që ka kaluar një periudhë tranzicioni. Numri i madh i konferencave të mbajtura rreth data mining në arsim ka çuar në një rritje të konsiderueshme të studimeve të botuara. Si rrjedhim i kësaj, bazat e të dhënave arsimore publike dhe mjetet për kryerjen e kurseve online rrisin aksesimin e të dhënave arsimore nga një grup më i gjerë individësh, duke bërë kështu të mundur uljen e pengesave për t'u bërë një studies i fushës së data mining, si rrjedhojë pritet një rritje e mëtejshme.
Vitet e fundit janë parë ndryshime të mëdha edhe në llojet e metodave data mining në edukim të cilat janë përdorur, me parashikimin dhe zbulimin e modeleve që janë rritur, ndërsa metodat relacionale janë bërë më të rrallë.
Në këtë pikë, metodat e data mining në arsim kanë pasur një nivel ndikimi në arsim dhe fusha të ngjashme ndërdisiplinore (të tilla si të inteligjencës artificiale në arsim, sistemet inteligjente tutoring, dhe modelimit të përdoruesit). Megjithatë, deri tani vetëm një pjesë e vogël e atikujve kanë arritur më shumë se 50 citate (siç kemi treguar në kapitullin 2), duke treguar se ka ende hapësirë të konsiderueshme për një rritje në ndikimin shkencor të data mining në arsim.

Gjithashtu në këtë punim u shqyrtuan dhe u zbatuan algoritme klasifikimi të rëndësishme të data ming në vlerësimin e preoperative të të dhënave rreth studentëve.

- **Rasti I** i studimit u fokusua në parashikime sa më të sakta për mundësitë e punësimit të të rinjve (studentëve të sapodiplomuar) bazuar tek arsimi që ata kishin ndjekur, dega apo profili që kishin zgjedhur, rezultatet e arritura, trajnimet apo kurset e ndryshme plotësuese të ndjekura, të numrit të gjuhëve të huaja që zotëronin, vendbanimit, moshës, etj. Duke krijuar një model parashikimi të përbërë nga 10 atribute dhe klasën "Punësuar" ( e cila kishte vetëm dy vlera Po/Jo).
  Në këtë studim u analizua rezultati i tre algoritmeve të klasifikimit NaiveBaye Updatable, MultiLayerPerceptron dhe LoogitBoost.
  Rezultatet treguan se algoritmi i cili jepte një saktësi parashikimi të saktë dhë të shpejtë të klasës së modelit "Punësuar" ishte algoritmi NaiveBayesUpdateable
- **Rasti II** i studimit u fokusua në shqyrtimin e performancës së studentëve gjatë tre viteve të shkollës së lartë, duke marrë të dhënat nga një fakultet i Inversitetit të Tiranës. Pasi u përpunuan të dhënat në ndërtimin e modelit u përfshinë të dhënat e rreth 1321 studentëve në përfundim të vitit të tretë dhe modeli i ndërtuar përmbante 8 atribute dhe klasën e parashkimit ST_Mes_3; e cila jepte vlerësimin mesatar të studentëve në vitin e tretë duke u bazuar në rezultatet që ai kishte marrë gjatë tre viteve të ndjekura.
  Në këtë studim u analizua roli i pemëve të vendimit në procesin e vendim marrjes nga pjesëmarrësit në këtë model dhe secili nga tre algoritmet (J48,

RepTree dhe RandomTree) e shqyrtuara jepte një parashikim më të saktë. Rezultatet treguan se algoritmi i cili jepte parashikimin më të saktë dhe më të shpejtë ishte algoritmi J48.

- **Rasti III** në këtë studim u shqyrtuan dhe u zbatuan tre algoritme klasifikimi të rëndësishme të data mining në vlerësimin e preoperative të të dhënave rreth studentëve në degë të caktuara ( kalon ose mbetet) dhe performancën e metodave të të mësuarit ku vlerësimi i tyre u bazua në saktësinë parashikuese të metodave, lehtësinë e të të mësuarit dhe karakteristikat user friendly.
  Rezultatet treguan që klasifikuesi Naive Bayes ???? në pemën e vendimit të parashikimit dhe në metodat e rrjeteve neurale. Gjithashtu u tregua që një klasifikues i mirë i modelit duhet të jetë njëkohësisht i saktë dhe i kuptueshëm për përdoruesit e tij, kushdo qofshin ata. Ky studim u bazua në mjediset tradicionale në klasa, duke qenë se teknikat e data mining u zbatuan pasi të dhënat u mblodhën.

Mund të konkludojmë se këto metodologji të përdorura, mund të përdoren për të ndihmuar studentët dhe profesorët për të përmirësuar performancën e studentëve; për shembull ndihmon në zvogëlimin e shkallës së mbetjes së studentëve duke ndjekur hapat e duhur në kohën e duhur për të përmirësuar cilësinë e të mësuarit.

Duke qenë se të mësuarit është një proces aktiv, interaktiviteti është një element bazë në këtë proces i cili ndikon në shkallën e kënaqësisë së studentëve dhe performancën. Është shumë e rëndësishme që t'i jepet përgjigje këtyre pyetjeve:

*- Si të bëjmë të mundur që modelet e paarshikimit të jenë user friendly për profesorët ose përdoruesit jo të specializuar në këtë fushë?*

*- Si të integrojmë sistemin e mbledhjes së të dhënave të universitetit dhe mjeteve të data mining?*

## 5.2 Rekomandime

Sipas studimit të realizuar nga Hamilton (2009) puna me të dhënat e studentëve mund të ndihmojë profesorët si të gjurmojnë avancimin akademik si dhe të kuptojnë se cilat praktika mesimdhënieje janë më praktike???. Ky studim shpjegon gjithashtu sesi studentët mund të shohin të dhënat e vlerësimit të tyre për të identifikuar pikat e tyre të forta dhe të dobta në mënyrë që të caktojnë se ku duan të arrijnë në arsimimin e tyre. Rekomandimet e këtij studimi janë që shkollat duhet të kenë një strategji të qartë për të zhvilluar një kulturë që udhëhiqet nga të dhënat dhe një fokus të përqëndruar në ndërtimin e infrakstrukturës që nevojitet për të mbledhur dhe vizualizuar të dhënat në mënyrë të shpejtë dhe domethënëse. Vizioni që të dhënat mund të përdoren nga mësuesit për të ndihmuar në përmirësimin e mësimdhënies dhe nga studentët për të monitoruar nxënien e tyre nuk është i re. Ka gjithashtu të dhëna të rëndësishme që tregojnë efektivitetin në fusha të tjera, të tilla si energjia dhe kujdesi shëndetësor.

Bizneset e internetit – qoftë ata që ofrojnë shërbimet dhe të mirat e përgjithshme, qoftë kopmanitë e mësimdhënies- kanë zbuluar fuqinë e përdorimit të të dhënave në përmirësimin e shpejtë të praktikave të tyre përmes eksperimentimit dhe matjes së ndryshimit që është i kuptueshëm dhe që con në marrjen e masave vepruese. Çelësi për konsumatorët e analizës së të dhënave, për shembull student, prindër, mësues, administratorë, është që të dhënat të paraqiten në mënyrë të tillë që t'i përgjigjen

qartësisht pyetjes që kërkohet dhe të drejtojnë drejt një veprimi për ta përmirësuar situatën. Më poshtë janë listuar disa rekomandime specifike për mësuesit, kërkuesit dhe zhvilluesit.

**A. Mësuesit -** Mësuesit e arsimit të lartë duhet të rrisin përdorimin e data mining në edukim dhe analizës së mësimit për të përmirësuar nxënien nga studentët. Rekomandimet e ekspertëve për të thjeshtuar këtë adaptim janë:

- **Mësuesit duhet të zhvillojnë një kulturë të përdorimit të të dhënave për marrjen e vendimeve lidhur me mësimdhënien.** Mësuesit duhet të kenë të dhëna të studentëve që u tregojnë atyre specifika të dobishme mbi të cilat mund të veprohet lidhur me mësimdhënien dhe mësimnxënien. Kjo do të thotë që mësuesit duhet të kenë akses pothuajse në kohë reale të paraqitjes vizuale të të dhënave lidhur me nxënien e studentëve, kjo e paraqitur në mënyrë të thjeshtë për t'u kuptuar dhe me nivel detajesh të tillë që të informojë me vendimet e tyre lidhur me mësimdhënien. Rezultatet e një testi të 6 muajve më parë nuk i tregojnë një mësuesi sesi ta ndihmojë një student nesër. Tipi i të dhënave që u jepet mësuesve duhet të jetë sa më ndihmues në marrjen e këtyre vendimeve dhe mësuesit do të duhet t'i shohin këto të dhëna me një pikëpamje ndryshe nga ajo që jepen nga sistemet e të dhënave, duke marrë parasysh qëllimet e llogari-dhënies.
- **Institucionet e arsimit të lartë duhet të kuptojnë që departamenti i teknologjisë së informacionit është pjesë e punës për përmirësimin e arsimimit por nuk është i vetmi departament përgjegjës.** Krijimi i një kulture që udhëhiqet nga të dhënat kërkon shumë më tepër sesa thjesht blerja/bërja e një sistemi kompjuterik. Gjithë departamentet duhet të bashkohen dhe të punojnë së bashku në mënyrë interaktive për të zhvilluar dhe përmirësuar mbledhjen e të dhënave, procesimin, analizimin dhe shpërndarjen e tyre. Studimi i realizuar nga Hamilton (2009) sugjeron që adoptimi i një kulture të përdorimit të të dhënave fillon me pyetje te tilla si:
  - *Cilat materiale mësimi apo teknika mësimdhënieje kanë qënë më efektive në përmirësimin e nxënies së studentëve në fusha të ndryshme?*
  - *A ka ndryshime ndërmjet suksesit të studentëve në shkollat tona krahasuar me shkolla të tjera?*
  - *A ka mësues të cilët janë veçanërisht të sukseshëm në mësimdhënie, praktikën e të cilëve mund ta përdorim si model për të tjerët?*
- **Kuptimi i të gjithë detajeve të një zgjidhje të propozuar.** Kur blihet/realizohet një sistem menaxhimi të mësimnxënies, institucionet duhet të pyesin deri në detaje për tipin e analizës së nxënies që sistemi do të gjenerojë si dhe duhet të sigurohen që sistemi do u japë mësuesve informacion lidhur me mënyrën se si mund të përmirësojnë mësimdhënien dhe mësimnxënien:
  - *Ku janë bazuar analizat?*
  - *A janë vlerësuar këto matje?*
  - *Kush i sheh të dhënat analitike dhe në cilin format dhe çfarë duhet të bëjnë ata për të patur akses?*

  Nqs studentët/mësuesit do të përdorin vizualizime ose raporte të tjera nga një paketë Data Mining, ata duhet të kenë mundësi të shohin paraprakisht këtë sistem që të sigurohen që të dhënat që do u paraqiten janë të kuptueshme. Çfarëdo lloj modeli parashikues i propozuar, duhet të jetë

transparent dhe të mbeshtetet mbi të dhëna të forta empirike bazuar mbi të dhëna të ngjashme.

- **Ndihmo studentët dhe prindërit të kuptojnë burimin dhe dobinë e të dhënave të mësimnxënies.** Gjatë kohës që institucionet ecin drejt përdorimit të të dhënve të sistemeve të mësimnxënies dhe bashkimit të këtyre të dhënave nga burime të shumëfishta, ata duhet të ndihmojnë studentët të kuptojnë se nga vijnë këto të dhëna, si përdoren këto të dhëna nga sistemet e mësimnxënies, dhe si ata mund të përdorin këto të dhëna për të marrë vendime lidhur më zgjedhjet dhe veprimet e tyre. Në mënyrë të ngjashme prindërit mund t'i ndihmojnë fëmijët e tyre të marrin vendime më të zgjuara, nqs ata kanë akses tek të dhënat dhe kuptojnë sesi janë gjenuar këto të dhëna dhe çfarë duan të thonë.

B. **Kërkuesit dhe zhvilluesi -** Kërkimi dhe zhvillimi në data mining në edukim dhe analizën e mësimdhënies ndodh si në organizata akademike dhe në ato tregtare. Kërkimi dhe zhvillimi janë të lidhura ngushtë me njëra-tjetrën pasi ato kërkojnë të kuptojnë proceset bazike të intrepetimit të të dhënave, marrjes së vendimeve dhe mësimnxënies dhe përdorin këto njohuri për të zhvilluar sisteme më të mira.

- **Zhvillim i kërkimit për përdorimin dhe impaktin e mënyrave alternative të prezantimit të të dhënave të detajuar, mësuesve, studentëve dhe prindërve.** Vizualimi i të dhënave ofron një urë të rëndësishme mes sistemeve teknologjike dhe analizës së të dhënave. Vendim marrja se si të dizenjohen vizualizimet, në mënyrë të tillë që aktorët të interpretojnë lehtësisht të dhënat, është një fushë aktive e kërkimit. Zgjidhja e këtij problemi kërkon identifikimin e zgjedhjeve ose vendimeve që mësuesit, studentët dhe prindërit do të duan të marrin bazuar në të dhënat e mësimnxënies si dhe faktorët që do të jenë prezentë kur tipe të ndryshme vendimesh do të duhet të merren.
- **Zhvillo mekanizma suporti dhe rekomandimi të cilat minimizojnë nivelin e analizimit të të dhënave nga mësuesit.** Mësuesit në një klasë duhet të kenë më shumë sesa thjesht akses lidhur me notat e studentëve. Mjete vlerësimi në kohë reale dhe sisteme që suportojnë vendime do të aftësonin mësuesin që të punojë me sisteme të automatizuara për të marrë vendime aty për aty, për të përmirësuar arsimimin e të gjithë studentëve. Për të ndihmuar mësuesit gjatë veprimtarisë së tyre nevojiten sisteme rekomandimi të cilat lidhin profilet e nxënies së studentëve me veprime të rekomanduara nxënie dhe burime mësimore.
- **Vazhdimi i perferksionimit të anonimizimit të të dhënave dhe mjeteve për bashkimin e të dhënave në mënyrë të tillë që t**ë **mbrohet privatësia individuale por në të njëjtën kohë të kemi përmirësime në pëdorimin e të dhënave të mësimnxënies.** Është akoma e paqartë pjesa ligjore që mbulon nxjerrjen e të dhënave për auditim, vlerësim dhe studim. Me gjithatë mbetet akoma shumë për të bërë për të kuptuar sesi mund të arrihet bashkimi ose ndarja e të dhënave të studentëve në nivele të ndryshme të sistemit arsimor në mënyrë të tillë që të jetë i mundur kombinimi i të dhënave nga burime të ndryshme por në të njëjtën kohë të mbrohet privatësia e studentëve.
- **Zhvillimi i modeleve për të bërë të mundur përshtatjen e sistemeve të zhvilluara në një kontekst dhe të përdoren në mënyrë efektive në**

**kontekste të tjera.** Ndryshimet në kontekstet arsimore e kanë bërë një sfidë transferimin e modeleve të parashikimit në sisteme të tjera arsimore. Për shkak se studentët, syllabuset dhe praktikat administrative ndryshojnë ndërmjet institucioneve të ndryshmeve, të dhënat që mund të grumbullohen nga ata ndryshojnë gjithashtu. Për këtë arsye shpesh një model i zhvilluar për një institucion nuk mund të aplikohet direkt në mënyrë efiçente në një instucion tjetër.

## 5.3 Punë në të ardhmen dhe kërkime të mëtejshme

Për punë në të ardhmen, eksperimentet mund të zgjerohen me më shumë atribute të veçanta në mënyrë që të meren rezultate më të sakta, të dobishme për të përmirësuar rezultatet e të mësuarit të studentëve.

Gjithashtu, eksperimetet mund të zhvillohen duke përdorur algoritme të tjera të data mining për të marrë një përafrim më të gjerë dhe rezultate akoma më të vlefshme dhe të sakta. Mund të përdoren programe të ndryshme ndërkohë që në të njëjtën kohë do të përdoren faktorë të ndryshëm.

Kërkuesit në bashkëpunim me programuesit e një instucioni arsimor mund të përdorin rezultatet e këtyre eksperimenteve ose të eksperimenteve të tjera për të ndërtuar një sistem parashikues i cili mund të ndihmojë studentët, mësuesit apo dhe përsona të tjerë për të marrë një parashikim mbi disa të dhëna kyçe në lidhje me përformancën e studentëve gjatë dhe pas përfundimit të ciklit mësimor.

## A. Aneksi 1

### PYETËSOR ( Mars 2013)

**Përshkrim i pyetësorit:**

*Pyetesori që ju keni në dorë shërben për të studiuar performancën e studentëve duke u bazuar në rezultatet tuaja në sistemin arsimor dhe në profilin tuaj demografik – social – shoqëror.*
*Përgjigjet tuaja në këtë pyetësor do të regjistrohen direkt në një bazë të dhënash dhe të trajtohen si tërësisht konfidenciale. Përgjigjet tuaja individuale nuk do të ndahen me fakultetin apo petagogët tuaj. Çdo e dhënë, duke përfshirë raca / përkatësisë etnike dhe gjinisë, që nuk është aktualisht në dispozicion për publikun do të përdoret vetëm në formë të përmbledhur. Pyetësori do të plotësohet në mënyrë anonime ( duke mos specfikuar emrin tuaj) në mënyrë që të ruajmë kofindencialitetin e të dhënave që ju do të ofroni.*
*Pjesëmarrja juaj në këtë pyetësor është vullnetare. Ju mund të refuzoni të përgjigjeni për çdo pyetje. Nuk ka asnjë rrezik personal për ju, në përgjigje të këtij pyetësori, që ju identifikojnë ju.*

*Ju lutemi që të dhënat të plotësohen në mënyrë sa më të saktë.*
***Ju Faleminderit!***

**Çështjet ku bazohet Pyetësori:**

*A. Profili demografik i studentit*
*B. Profili social- shoqëror i studentit*
*C. Rezultatet në sistemin arsimor (të tilla si arritjet e studentit gjatë vitit të parë(numri total i krediteve) dhë gjatë shkollës së mesme(mesatarja))*

## A. Profili i demografik:

1. **Cila është gjinia juaj:**

☐ Femër ☐ Mashkull

2. **Cila është mosha juaj:**

☐ 18 Vjeç
☐ 19 Vjeç
☐ 20 Vjeç
☐ 21 Vjeç
☐ mbi 21 Vjeç Ju lutem jepni moshën tuaj _____

3. **Cili është vendbanimi juaj ( ju lutem specifikoni se ku ju keni gjendjen civile tuaj):**

☐ Tiranë

☐ Qytet Tjetër Ju lutem shkruani emrin e qytetit_____

☐ Fshat Ju lutem shkruani emrin e fshatit_____

4. **Jetoni në konvikt:**

☐ Po ☐ Jo

Nqs përgjigjia është PO. Ju lutem specifikoni tipin e konviktit:

☐ Privat ☐ Shtetëror

## B. Profili social- shoqëror i studentit:

1. **Sa është numri i pjesëtarëve në familjen tuaj:**

☐ 1
☐ 2
☐ 3
☐ mbi 3 Ju lutem specifikoni numrin e saktë _____

2. **Sa janë të ardhurat vjetore të familjes suaj:**

☐ Të ulëta ☐ Të mesme ☐ Të larta

3. **Cili është profesioni i babait tuaj:**

☐ Buxhetor
☐ Biznesmen
☐ Bujqësi
☐ Pensionist
☐ Tjetër Ju lutem specifikoni profesionin e babait_____

4. **Cili është profesioni i mamasë tuaj:**

☐ Buxhetore
☐ Biznesmene
☐ Shtëpiake

☐ Pensioniste
☐ Tjetër Ju lutem specifikoni profesionin e mamasë_____

## C. Profili social- shoqëror i studentit:

**1. Me çfarë mesatare ju keni përfunduar shkollën e mesme:**

☐ 10:00 – 9:00
☐ 8:00 – 8:99
☐ 7:00 – 7:99
☐ 6:00 – 6:99
☐ 5:00 – 5:99

**2. Programi studimor që ju ndiqni:**

☐ Bachelor Informatikë
☐ Bachelor Teknologji Informacioni dhe Komunikimi

**3. Sa provime keni dhënë deri më tani:**

☐ 2 – 4
☐ 5– 7
☐ 7 – 9
☐ mbi 9 Ju lutem specifikoni numrin e provimeve _____

**4. Sa provime keni marrë deri më tani:**

☐ 2 – 4
☐ 5
☐ 6
☐ mbi 6 Ju lutem specifikoni numrin e provimeve _____

**5. Sa kredite keni fituar deri më tani:**

☐ 20
☐ 21 - 30
☐ 31 - 40
☐ mbi 41 Ju lutem specifikoni numrin e krediteve _____

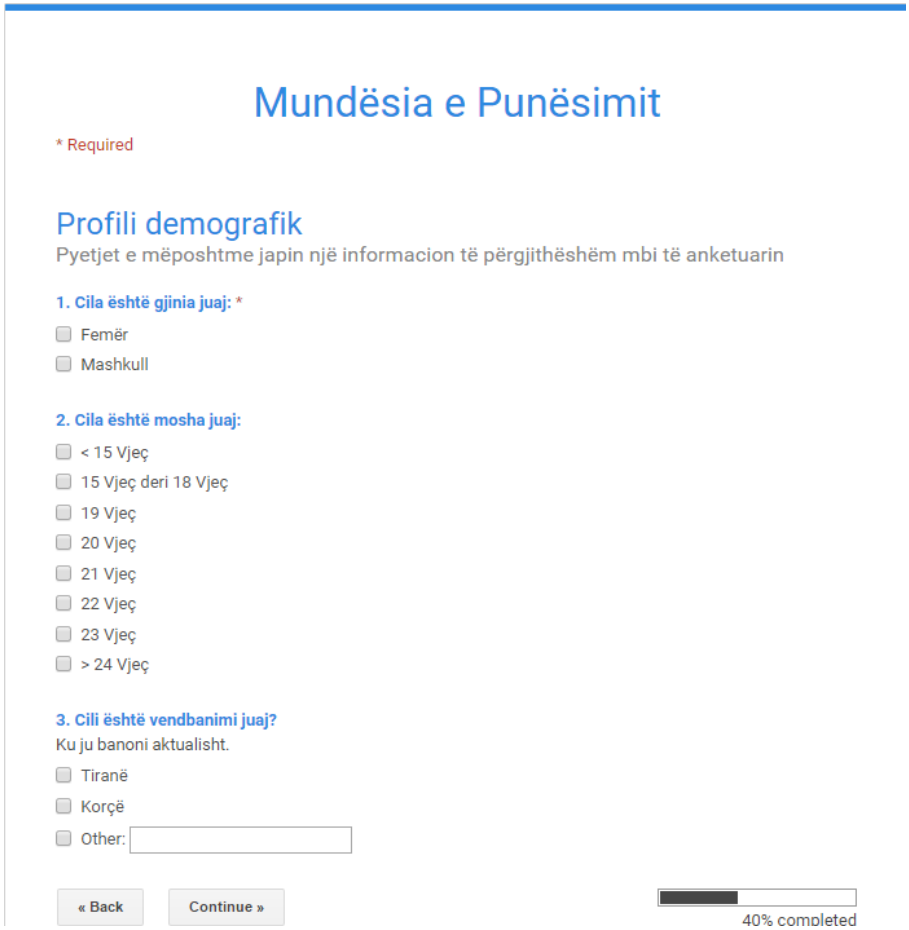
## PYETËSOR ( Shtator 2013)

Për të mbledhur të dhënat në kohë reale në mënyrë sa më të saktë edhe të shpejtë është përdorur mundësia e pafundme që ofron interneti dhe rrjetet sociale në ditët e sotme. Fillimisht është ndërtuar një pyetësor online duke përdorur funksionalitetin **Google Docs**. Nëpërmjet këtij funksioni hapet një dritare ku lejohet të zgjidhet një paraqitje vizuale e këndshme për pyetësorin, të shënohet titulli i pyetësorit që do të krijohet dhe më pas specifikohen të gjitha pyetjet duke u bazuar edhe tek qëllimi që ka studimi. Në përfundim të ndërtimit të pyetësorit jepet mundësia e shpërndarjes (duke u klikuar butoni **Send form)** së pyetësorit në rrjete të ndryshme sociale. Pyetësori u ndërtua online dhe u formulua në një mënyrë të thjeshtë. Jam kujdesur që pyetësori të përmbante pyetjet e nevojshme për të mbledhur të dhënat kryesore që do të duheshin më vonë për të ndërtuar bashkësinë e të dhënave duke u bazuar edhe tek qëllimi i studimit.

Duke qenë se pyetësori është ndërtuar online, më poshtë janë dhënë figurat e katër faqeve që përmbante pyetësori.

***Faqja kryesore e pyetësorit***

***Faqja 1 – Pyetjet rreth profilit demografik të të anketuarit***



Mundësia e Punësimit

* Required

Profili demografik

Pyetjet e mëposhtme japin një informacion të përgjithëshëm mbi të anketuarin

**1. Cila është gjinia juaj: ***

- Femër
- Mashkull

**2. Cila është mosha juaj:**

- < 15 Vjeç
- 15 Vjeç deri 18 Vjeç
- 19 Vjeç
- 20 Vjeç
- 21 Vjeç
- 22 Vjeç
- 23 Vjeç
- > 24 Vjeç

**3. Cili është vendbanimi juaj?**

Ku ju banoni aktualisht.

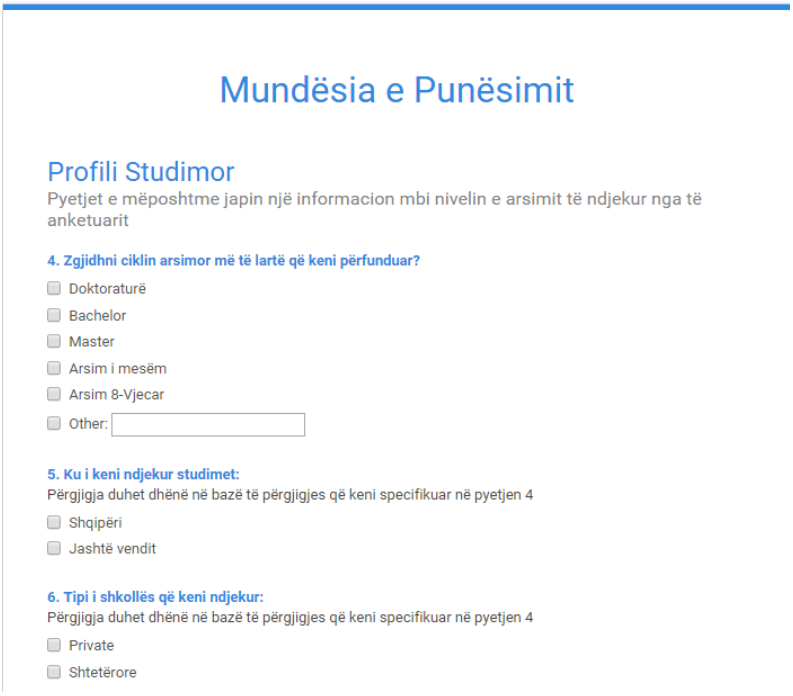
- Tiranë
- Korçë
- Other:

« Back  Continue »

40% completed

***Faqja 2 Pyetjet rreth profilit studimit të të anketuarit***

***Pyetjet 4-6***



Mundësia e Punësimit

Profili Studimor

Pyetjet e mëposhtme japin një informacion mbi nivelin e arsimit të ndjekur nga të anketuarit

**4. Zgjidhni ciklin arsimor më të lartë që keni përfunduar?**

- Doktoraturë
- Bachelor
- Master
- Arsim i mesëm
- Arsim 8-Vjecar
- Other:

**5. Ku i keni ndjekur studimet:**

Përgjigja duhet dhënë në bazë të përgjigjes që keni specifikuar në pyetjen 4

- Shqipëri
- Jashtë vendit

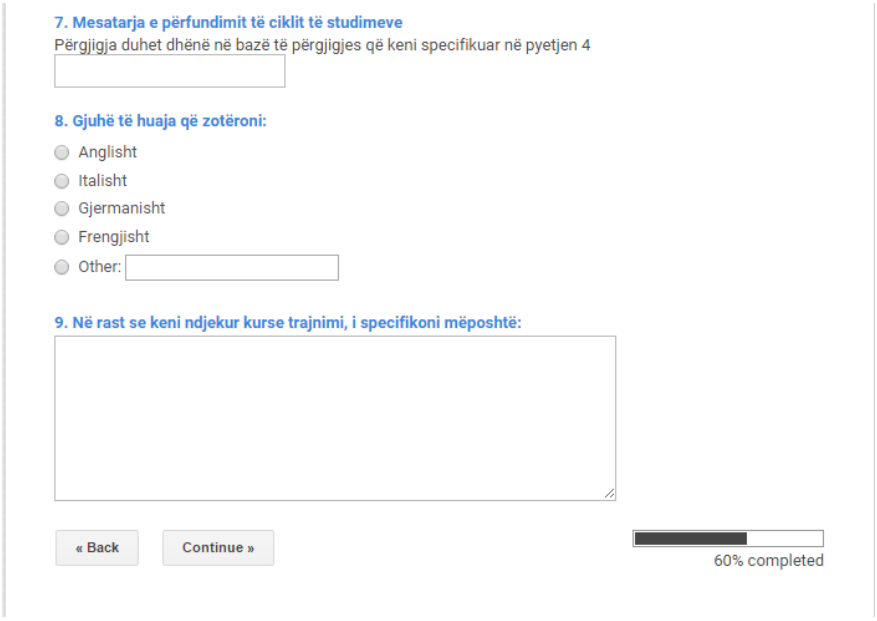
**6. Tipi i shkollës që keni ndjekur:**

Përgjigja duhet dhënë në bazë të përgjigjes që keni specifikuar në pyetjen 4

- Private
- Shtetërore

***Faqja 3 Pyetjet rreth profilit studimit të të anketuarit***

***Pyetjet 7-8***

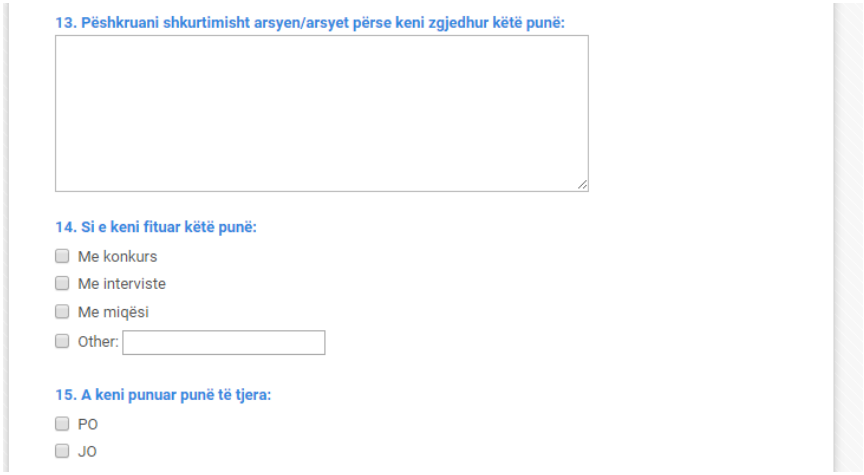



***Faqja 3 Pyetjet rreth profilit të punësimit të të anketuarit***

***Pyetjet 10-12***

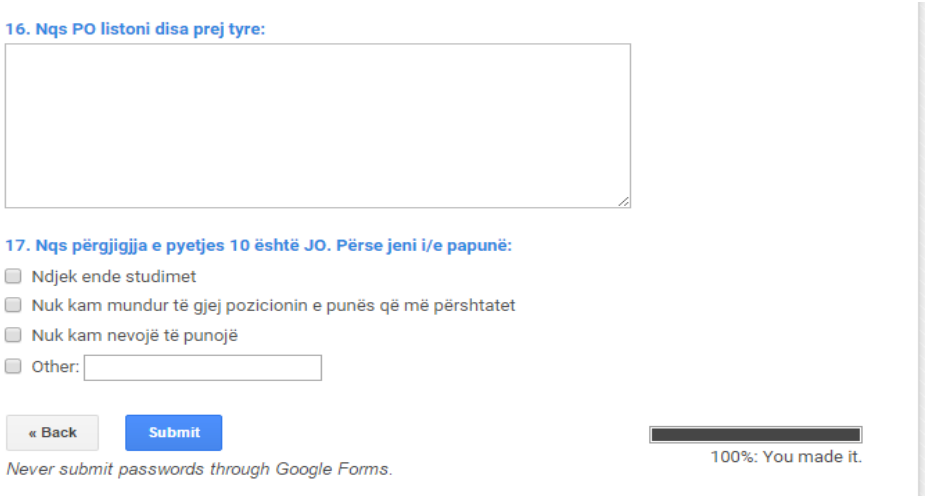
***Faqja 4 Pyetjet rreth profilit të punësimit të të anketuarit***

***Pyetjet 13 - 15***



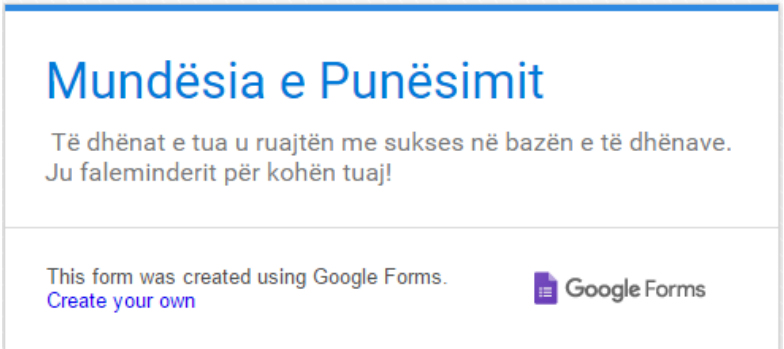

***Faqja 3Pyetjet rreth profilit të punësimit të të anketuarit***

***Pyetjet 16-17***




*Në përfundim të tij i anketuari duhet të klikojë butonin Submit, në mënyrë që të ruhen përgjigjet e tij*

***Faqja e fundit. Ku falënderohet i anketuari për kohën e tij.***



# B. Aneksi 2

Në studimin tonë kemi përdorur aplikacionin WEKA për të zbatuar metodat dhe teknikat Data Mining. Ashtu siç është theksuar dhe më sipër, të dhënat e studimit janë marrë nga pyetësorë të shkruar, online dhe nga sistemi i menaxhimit të studentëve nga një fakultet i Universitetit të Tiranës. Gjatë shpjegimit të studimit janë dhënë shembuj dhe rezultate të krijimit të proceseve të ndryshme. Në këtë pjesë po i listojmë të gjithë proceset e krijuara. Për secilin prej proceseve të krijuar, rezultatet nuk janë tjetër veçse statistika të nxjerra mbi kritere të ndryshme të vendosura nga operatorët e përdorur. Kështu më poshtë po paraqesim të dhënat statistikore të arritura nga ky studim:

***Rasti II: Parashikimin e performancës së studentëve, duke përdorur të dhënat e mbledhura nga sistemi menaxhimit të studentëve, të një fakulteti të Universitetit të Tiranës***

## 1. Rezultati i ekzekutimit të algoritmit J48

Në figurën e mëposhtme jepet një print screen i ndërfaqes së programit WEKA, në të cilin është ekzekutuar algoritmi J48 për testin e Training Set.

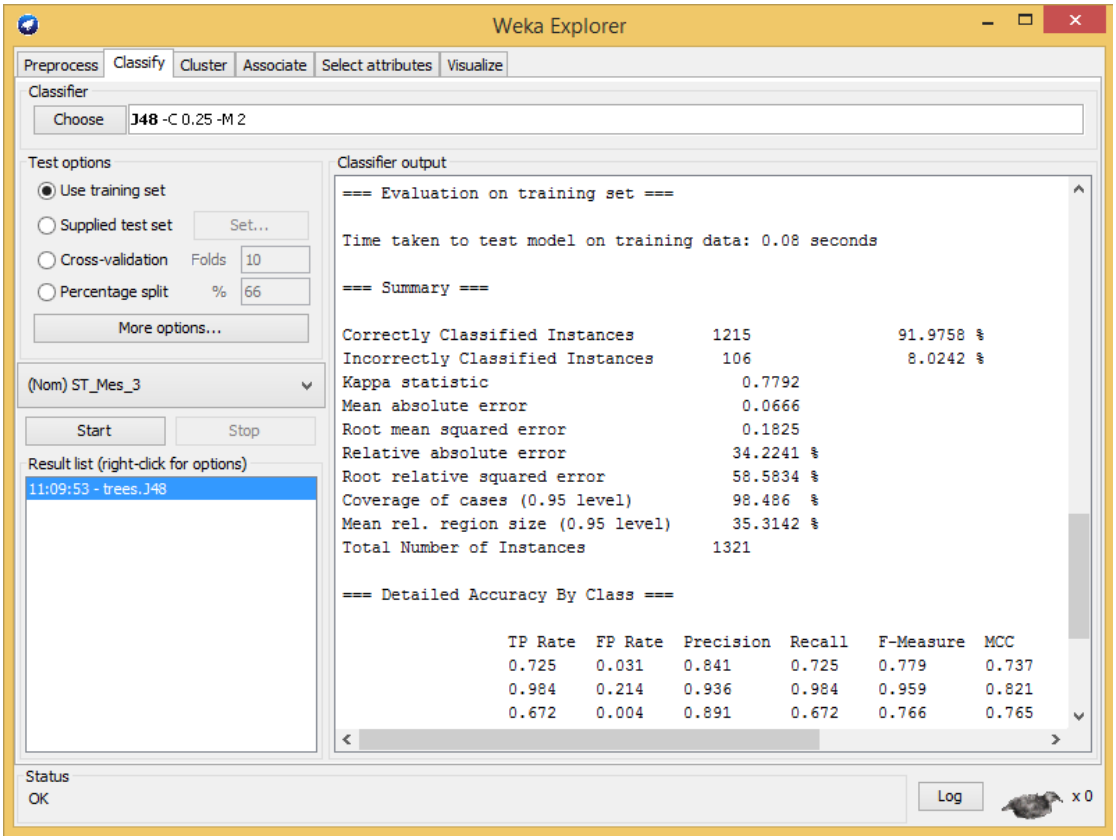



**Figura B.1 Rezultati i algoritmit J48 në programin WEKA për rastin II të studimit**

### Testi 1 – Training Set

Të dhënat e ekzekutimit:

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     TEST-weka.filters.unsupervised.attribute.Remove-R1-
weka.filters.unsupervised.attribute.Remove-R2
Instances:    1321
Attributes:   8
              ST_Gjini
              ST_Mosha
              VendB
              ST_Shtet
              SP
              ST_Mes_1
              ST_Mes_2
              ST_Mes_3
Test mode:    evaluate on training data
```

Pema e vendimit e gjeneruar:

```
=== Classifier model (full training set) ===

J48 pruned tree
------------------
ST_Mes_1 = D
|   ST_Mes_2 = A: C (104.0/21.0)
|   ST_Mes_2 = D: D (775.0/15.0)
|   ST_Mes_2 = C: D (157.0/29.0)
|   ST_Mes_2 = B
|   |   SP = IT: D (68.0/18.0)
|   |   SP = CS
|   |   |   VendB = A: C (5.0/1.0)
|   |   |   VendB = C: D (1.0)
|   |   |   VendB = B
|   |   |   |   ST_Mosha <= 20: D (2.0)
|   |   |   |   ST_Mosha > 20: C (6.0/1.0)
ST_Mes_1 = C
|   ST_Mes_2 = A: B (15.0/1.0)
|   ST_Mes_2 = D: D (49.0/5.0)
|   ST_Mes_2 = C: C (62.0/3.0)
|   ST_Mes_2 = B: C (25.0/6.0)
ST_Mes_1 = B
|   ST_Mes_2 = A
|   |   SP = IT: A (2.0)
|   |   SP = CS: B (7.0/3.0)
|   ST_Mes_2 = D: D (3.0/1.0)
|   ST_Mes_2 = C: C (3.0)
|   ST_Mes_2 = B: B (23.0/1.0)
ST_Mes_1 = A
|   ST_Mes_2 = A: A (10.0)
|   ST_Mes_2 = D: C (2.0/1.0)
|   ST_Mes_2 = C: B (1.0)
|   ST_Mes_2 = B: A (1.0)

Number of Leaves  :             21
```

```
Size of the tree :      30

Time taken to build model: 0.09 seconds
```

Parametrat e vlerësimit të performancës së algoritmit:

```
=== Evaluation on training set ===

Time taken to test model on training data: 0.08 seconds

=== Summary ===

Correctly Classified Instances        1215               91.9758 %
Incorrectly Classified Instances      106                8.0242 %
Kappa statistic                          0.7792
Mean absolute error                      0.0666
Root mean squared error                  0.1825
Relative absolute error                 34.2241 %
Root relative squared error             58.5834 %
Coverage of cases (0.95 level)          98.486  %
Mean rel. region size (0.95 level)      35.3142 %
Total Number of Instances             1321

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.725    0.031    0.841      0.725    0.779      0.737    0.933     0.784     C
               0.984    0.214    0.936      0.984    0.959      0.821    0.949     0.972     D
               0.672    0.004    0.891      0.672    0.766      0.765    0.978     0.782     B
               0.765    0.000    1.000      0.765    0.867      0.873    0.982     0.907     A
Weighted Avg.  0.920    0.168    0.917      0.920    0.916      0.804    0.948     0.928
```

Matrica e çrregullimeve:

```
=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 174  65   1   0 |   a = C
  15 987   1   0 |   b = D
  18   2  41   0 |   c = B
   0   1   3  13 |   d = A
```

Pema e vendimit e gjeneruar për secilin nga testet është e njëjtë si strukturë (ndryshon vetëm në disa vlera të gjetheve për klasifikimin e instancave të gabuara), kështu që për tre testet e mëposhtme do të japim vetëm rezultatin për vlerësimin e performancës dhe matricën e çrregullimeve.

**Testi 2 – Vlerësimi i kryqëzuar i 3-fishtë**

Performanca e algoritmit:

```
Test mode:    3-fold cross-validation
======.....=======

Time taken to build model: 0.02 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1203               91.0674 %
Incorrectly Classified Instances      118                8.9326 %
Kappa statistic                          0.7545
Mean absolute error                      0.0713
Root mean squared error                  0.1923
Relative absolute error                 36.5856 %
Root relative squared error             61.7402 %
Coverage of cases (0.95 level)          98.1832 %
Mean rel. region size (0.95 level)      36.1847 %
Total Number of Instances             1321

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.704     0.033     0.824      0.704     0.760      0.714     0.914     0.780     C
               0.980     0.226     0.932      0.980     0.955      0.803     0.941     0.968     D
               0.607     0.003     0.902      0.607     0.725      0.730     0.957     0.704     B
               0.824     0.005     0.700      0.824     0.757      0.756     0.915     0.597     A
Weighted Avg.  0.911     0.178     0.908      0.911     0.907      0.783     0.936     0.917
```

Matrica e çrregullimeve:

```
=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 169  69   1   1 |   a = C
  19 983   1   0 |   b = D
  17   2  37   5 |   c = B
   0   1   2  14 |   d = A
```

**Testi 3 – Vlerësimi i kryqëzuar i 13-fisht**

Performanca e algoritmit:

```
Test mode:    13-fold cross-validation
======.....=======
Time taken to build model: 0.02 seconds
```

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1204               91.1431 %
Incorrectly Classified Instances       117                8.8569 %
Kappa statistic                          0.7577
Mean absolute error                      0.0715
Root mean squared error                  0.194
Relative absolute error                 36.724  %
Root relative squared error             62.259  %
Coverage of cases (0.95 level)          97.8047 %
Mean rel. region size (0.95 level)      36.6011 %
Total Number of Instances             1321

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.717    0.034    0.823     0.717    0.766      0.721    0.899     0.742     C
                 0.979    0.217    0.934     0.979    0.956      0.808    0.929     0.958     D
                 0.607    0.004    0.881     0.607    0.718      0.721    0.922     0.651     B
                 0.765    0.005    0.684     0.765    0.722      0.720    0.906     0.551     A
Weighted Avg.    0.911    0.171    0.908     0.911    0.908      0.787 0.923      0.899
```

Matrica e çrregullimeve:

```
=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 172  66   1   1 |   a = C
  20 982   1   0 |   b = D
  17   2  37   5 |   c = B
   0   1   3  13 |   d = A
```

### 2. Rezultati i ekzekutimit të algoritmit REPTree

Në figurën e B.2 jepet një printscreen i ndërfaqes së programi WEKA, në të cilin është ekzekutuar algoritmi REPTree për testin e Training Set.

Të dhënat e ekzekutimit:

```
=== Run information ===

Scheme:       weka.classifiers.trees.REPTree -M 2 -V 0.001 -N 3 -S 1 -L -1 -I 0.0
Relation:     TEST-weka.filters.unsupervised.attribute.Remove-R1-
weka.filters.unsupervised.attribute.Remove-R2
Instances:    1321
Attributes:   8
              ST_Gjini
```

*ST_Mosha*
*VendB*
*ST_Shtet*
*SP*
*ST_Mes_1*
*ST_Mes_2*
*ST_Mes_3*

*Test mode: evaluate on training data*

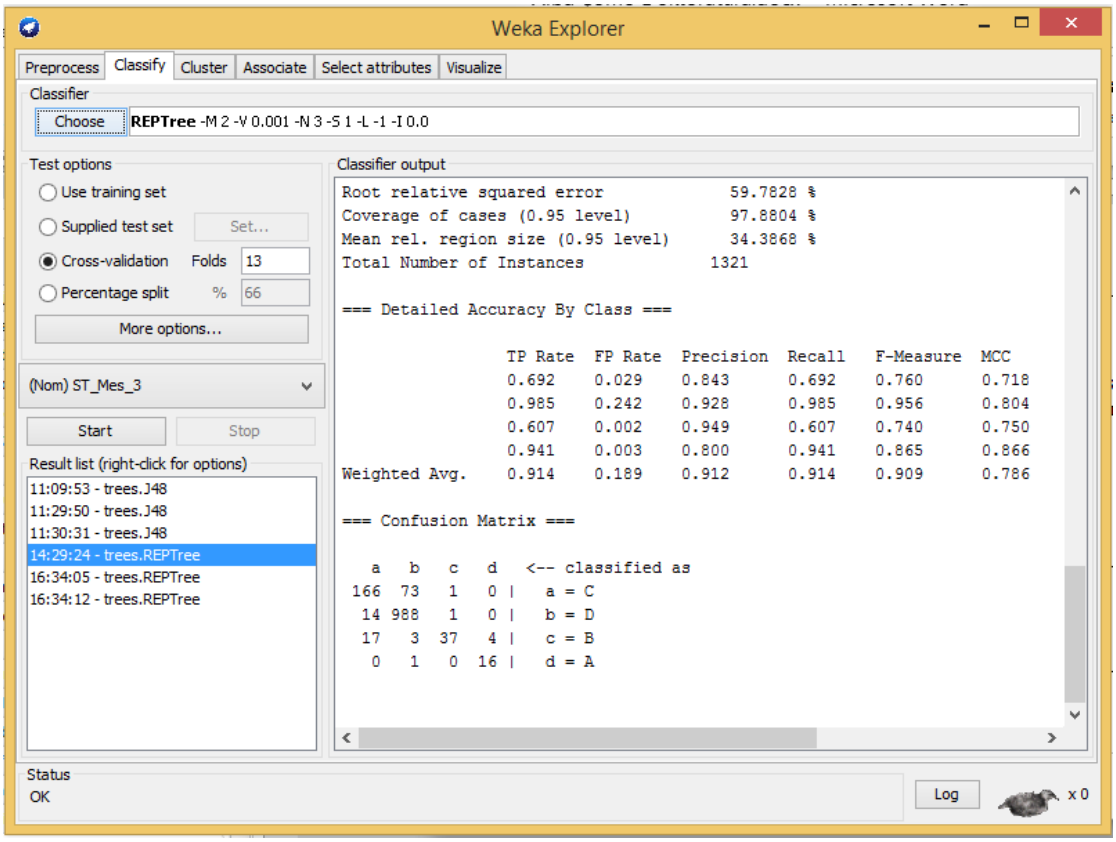



**Figura B.2 Rezultati i algoritmit REPTree në programin WEKA për rastin II të studimit**

Pema e vendimit e gjeneruar:

*=== Classifier model (full training set) ===*

*REPTree*
*==========*

*ST_Mes_2 = A*
*| ST_Mes_1 = D : C (62/13) [42/8]*
*| ST_Mes_1 = C : B (11/1) [4/0]*
*| ST_Mes_1 = B : A (7/3) [2/1]*
*| ST_Mes_1 = A : A (6/0) [4/0]*
*ST_Mes_2 = D : D (569/17) [260/6]*
*ST_Mes_2 = C*
*| ST_Mes_1 = D*

```
|   |   SP = IT : D (40/2) [28/5]
|   |   SP = CS
|   |   |   ST_Gjini = M : D (30/12) [12/0]
|   |   |   ST_Gjini = F
|   |   |   |   VendB = A : D (9/4) [5/1]
|   |   |   |   VendB = C : C (2/1) [1/0]
|   |   |   |   VendB = B : D (21/2) [9/1]
|   ST_Mes_1 = C : C (47/3) [15/0]
|   ST_Mes_1 = B : C (1/0) [2/0]
|   ST_Mes_1 = A : B (1/0) [0/0]
ST_Mes_2 = B
|   ST_Mes_1 = D : D (43/17) [39/10]
|   ST_Mes_1 = C : C (17/4) [8/2]
|   ST_Mes_1 = B : B (13/1) [10/0]
|   ST_Mes_1 = A : A (1/0) [0/0]

Size of the tree : 24

Time taken to build model: 0.06 seconds
```

Parametrat e vlerësimit të performancës së algoritmit:

```
=== Evaluation on training set ===

Time taken to test model on training data: 0.11 seconds

=== Summary ===

Correctly Classified Instances        1207               91.3702 %
Incorrectly Classified Instances      114                8.6298 %
Kappa statistic                          0.7598
Mean absolute error                      0.0694
Root mean squared error                  0.1862
Relative absolute error                 35.6399 %
Root relative squared error             59.7828 %
Coverage of cases (0.95 level)          97.8804 %
Mean rel. region size (0.95 level)      34.3868 %
Total Number of Instances             1321

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.692    0.029    0.843      0.692    0.760      0.718    0.924     0.778     C
               0.985    0.242    0.928      0.985    0.956      0.804    0.941     0.967     D
               0.607    0.002    0.949      0.607    0.740      0.750    0.970     0.764     B
               0.941    0.003    0.800      0.941    0.865      0.866    0.981     0.884     A
Weighted Avg.  0.914    0.189    0.912      0.914    0.909      0.786    0.939     0.922
```

Matrica e çrregullimeve:

```
=== Confusion Matrix ===

   a   b   c   d   <-- classified as
```

```
166  73   1   0 |   a = C
 14 988   1   0 |   b = D
 17   3  37   4 |   c = B
  0   1   0  16 |   d = A
```

Pema e vendimit e gjeneruar për secilin nga testet është e njëjtë si strukturë (ndryshon vetëm në disa vlera të gjetheve për klasifikimin e instancave të gabuara), kështu që për tre testet e mëposhtme do të japim vetëm rezultatin për vlerësimin e performancës dhe matricën e çrregullimeve.

**Testi 2 – Vlerësimi i 3-fold**

Performanca e algoritmit:

```
Test mode:    3-fold cross-validation
=====.....=======

Time taken to build model: 0.11 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1195               90.4618 %
Incorrectly Classified Instances       126                9.5382 %
Kappa statistic                          0.7406
Mean absolute error                      0.0718
Root mean squared error                  0.1935
Relative absolute error                 36.8222 %
Root relative squared error             62.1227 %
Coverage of cases (0.95 level)          97.2748 %
Mean rel. region size (0.95 level)      34.3679 %
Total Number of Instances             1321

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.721    0.043    0.790      0.721    0.754      0.703    0.915     0.780     C
                 0.970    0.226    0.931      0.970    0.950      0.782    0.935     0.965     D
                 0.590    0.003    0.900      0.590    0.713      0.719    0.935     0.717     B
                 0.765    0.003    0.765      0.765    0.765      0.762    0.916     0.822     A
Weighted Avg.    0.905    0.180    0.902      0.905    0.901      0.765    0.931     0.918
```

Matrica e çrregullimeve:

```
=== Confusion Matrix ===

   a   b   c   d   <-- classified as
```

```
 173  66   1   0 |   a = C
  29 973   1   0 |   b = D
  17   4  36   4 |   c = B
   0   2   2  13 |   d = A
```

**Testi 3 – Vlerësimi i 13-fold**

Performanca e algoritmit:

```
Test mode:    13-fold cross-validation
======.....=======

Time taken to build model: 0.03 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1200               90.8403 %
Incorrectly Classified Instances       121                9.1597 %
Kappa statistic                          0.7445
Mean absolute error                      0.0719
Root mean squared error                  0.1937
Relative absolute error                 36.915  %
Root relative squared error             62.1718 %
Coverage of cases (0.95 level)          97.6533 %
Mean rel. region size (0.95 level)      35.2574 %
Total Number of Instances             1321

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.688    0.031    0.833      0.688    0.753      0.710    0.900     0.760     C
                 0.983    0.252    0.925      0.983    0.953      0.792    0.924     0.956     D
                 0.607    0.004    0.881      0.607    0.718      0.721    0.929     0.701     B
                 0.706    0.002    0.800      0.706    0.750      0.748    0.906     0.835     A
Weighted Avg.    0.908    0.197    0.905      0.908    0.903      0.773    0.920     0.907
```

Matrica e çrregullimeve:

```
=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 165  74   1   0 |   a = C
  16 986   1   0 |   b = D
  17   4  37   3 |   c = B
   0   2   3  12 |   d = A
```

**3. Rezultati i ekzekutimit të algoritmit Random Tree**

Në figurën e mëposhtme jepet një printscreen i ndërfaqes së programit WEKA, në të cilin është ekzekutuar algoritmi Random Tree për testin e Training Set.

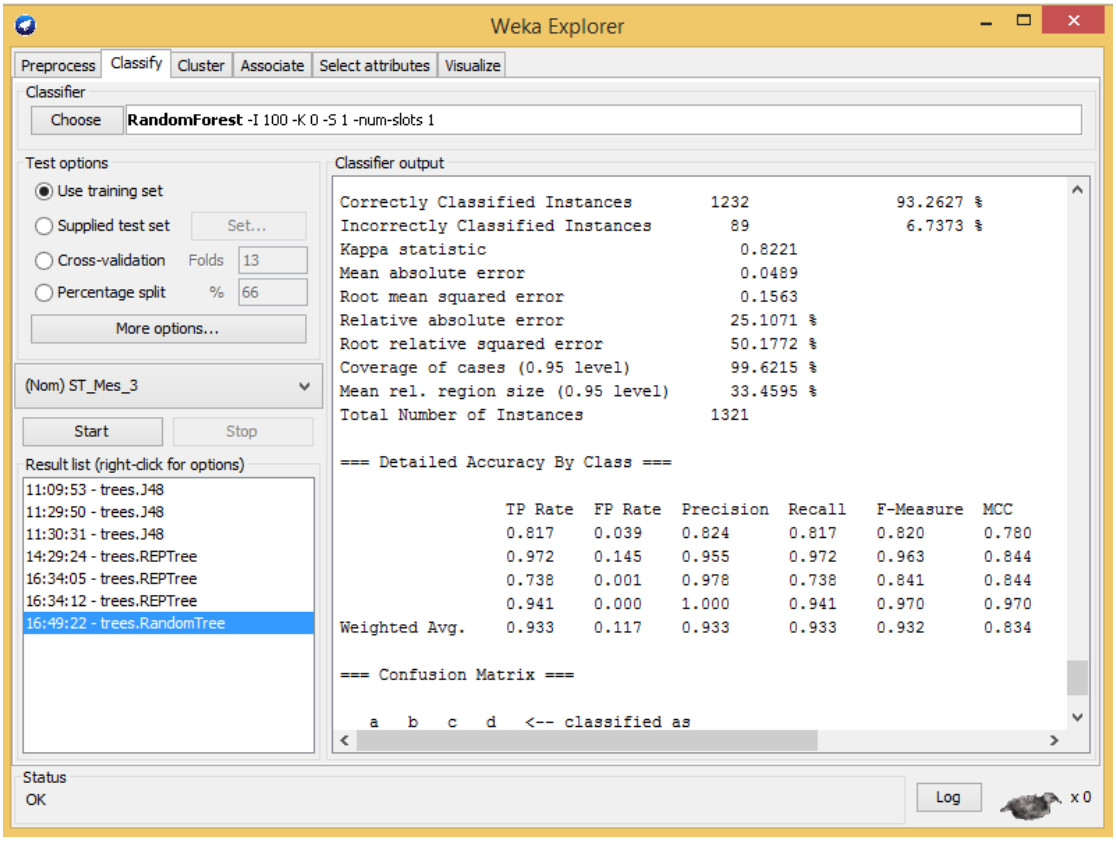



**Figura B.3 Rezultati i algoritmit Random Tree në programin WEKA për rastin II të studimit**

Të dhënat e ekzekutimit:

*=== Run information ===*

*Scheme: weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1*
*Relation: TEST-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.unsupervised.attribute.Remove-R2*
*Instances: 1321*
*Attributes: 8*
*ST_Gjini*
*ST_Mosha*
*VendB*
*ST_Shtet*
*SP*
*ST_Mes_1*
*ST_Mes_2*
*ST_Mes_3*
*Test mode: evaluate on training data*

Pema e vendimit e gjeneruar:

*=== Classifier model (full training set) ===*

*RandomTree*

```
==========

ST_Mosha < 20.5
|  VendB = A
|  |  ST_Mes_1 = D
|  |  |  ST_Mes_2 = A
|  |  |  |  SP = IT
|  |  |  |  |  ST_Gjini = M : C (7/1)
|  |  |  |  |  ST_Gjini = F : C (12/3)
|  |  |  |  SP = CS : B (1/0)
|  |  |  ST_Mes_2 = D : D (62/0)
|  |  |  ST_Mes_2 = C
|  |  |  |  SP = IT
|  |  |  |  |  ST_Gjini = M : D (5/0)
|  |  |  |  |  ST_Gjini = F : C (2/1)
|  |  |  |  SP = CS
|  |  |  |  |  ST_Gjini = M : C (5/2)
|  |  |  |  |  ST_Gjini = F : C (3/1)
|  |  |  ST_Mes_2 = B
|  |  |  |  ST_Gjini = M
|  |  |  |  |  SP = IT : D (1/0)
|  |  |  |  |  SP = CS : C (1/0)
|  |  |  |  ST_Gjini = F : D (7/1)
|  |  ST_Mes_1 = C
|  |  |  ST_Mes_2 = A : B (2/0)
|  |  |  ST_Mes_2 = D : D (4/0)
|  |  |  ST_Mes_2 = C
|  |  |  |  ST_Gjini = M : C (4/0)
|  |  |  |  ST_Gjini = F : C (3/1)
|  |  |  ST_Mes_2 = B : C (5/0)
|  |  ST_Mes_1 = B : B (3/0)
|  |  ST_Mes_1 = A : A (1/0)
|  VendB = C
|  |  ST_Gjini = M
|  |  |  ST_Mes_2 = A : C (1/0)
|  |  |  ST_Mes_2 = D : D (7/0)
|  |  |  ST_Mes_2 = C : D (1/0)
|  |  |  ST_Mes_2 = B : C (1/0)
|  |  ST_Gjini = F
|  |  |  ST_Mes_1 = D
|  |  |  |  ST_Mes_2 = A : D (1/0)
|  |  |  |  ST_Mes_2 = D : D (7/0)
|  |  |  |  ST_Mes_2 = C : C (1/0)
|  |  |  |  ST_Mes_2 = B : D (1/0)
|  |  |  ST_Mes_1 = C : D (1/0)
|  |  |  ST_Mes_1 = B : B (1/0)
|  |  |  ST_Mes_1 = A : C (0/0)
|  VendB = B
|  |  ST_Mes_2 = A
|  |  |  SP = IT
|  |  |  |  ST_Gjini = M : C (8/0)
|  |  |  |  ST_Gjini = F : C (9/3)
|  |  |  SP = CS
|  |  |  |  ST_Mes_1 = D : C (0/0)
|  |  |  |  ST_Mes_1 = C : B (3/1)
```

```
|   |   |   |   ST_Mes_1 = B : A (1/0)
|   |   |   |   ST_Mes_1 = A : A (2/0)
|   |   ST_Mes_2 = D
|   |   |   ST_Mes_1 = D
|   |   |   |   SP = IT
|   |   |   |   |   ST_Gjini = M : D (28/4)
|   |   |   |   |   ST_Gjini = F : D (38/2)
|   |   |   |   SP = CS
|   |   |   |   |   ST_Gjini = M : D (61/1)
|   |   |   |   |   ST_Gjini = F : D (70/1)
|   |   |   ST_Mes_1 = C : D (8/0)
|   |   |   ST_Mes_1 = B : C (0/0)
|   |   |   ST_Mes_1 = A : C (0/0)
|   |   ST_Mes_2 = C
|   |   |   ST_Mes_1 = D
|   |   |   |   SP = IT : D (17/0)
|   |   |   |   SP = CS : D (18/3)
|   |   |   ST_Mes_1 = C : C (6/0)
|   |   |   ST_Mes_1 = B : C (0/0)
|   |   |   ST_Mes_1 = A : C (0/0)
|   |   ST_Mes_2 = B
|   |   |   SP = IT
|   |   |   |   ST_Mes_1 = D
|   |   |   |   |   ST_Gjini = M : D (3/0)
|   |   |   |   |   ST_Gjini = F : D (5/1)
|   |   |   |   ST_Mes_1 = C : C (0/0)
|   |   |   |   ST_Mes_1 = B : B (2/0)
|   |   |   |   ST_Mes_1 = A : C (0/0)
|   |   |   SP = CS
|   |   |   |   ST_Gjini = M
|   |   |   |   |   ST_Mes_1 = D : D (1/0)
|   |   |   |   |   ST_Mes_1 = C : C (2/1)
|   |   |   |   |   ST_Mes_1 = B : B (2/0)
|   |   |   |   |   ST_Mes_1 = A : C (0/0)
|   |   |   |   ST_Gjini = F
|   |   |   |   |   ST_Mes_1 = D : D (1/0)
|   |   |   |   |   ST_Mes_1 = C : B (1/0)
|   |   |   |   |   ST_Mes_1 = B : C (0/0)
|   |   |   |   |   ST_Mes_1 = A : C (0/0)
ST_Mosha >= 20.5
|   ST_Mes_1 = D
|   |   ST_Mes_2 = A
|   |   |   ST_Gjini = M
|   |   |   |   ST_Mosha < 21.5
|   |   |   |   |   VendB = A : C (9/1)
|   |   |   |   |   VendB = C : C (0/0)
|   |   |   |   |   VendB = B : C (1/0)
|   |   |   |   ST_Mosha >= 21.5 : C (11/0)
|   |   |   ST_Gjini = F
|   |   |   |   VendB = A
|   |   |   |   |   ST_Mosha < 21.5 : C (11/4)
|   |   |   |   |   ST_Mosha >= 21.5 : C (12/4)
|   |   |   |   VendB = C : C (1/0)
|   |   |   |   VendB = B
|   |   |   |   |   ST_Mosha < 21.5 : C (8/1)
```

```
|  |  |  |  |  ST_Mosha >= 21.5 : C (12/2)
|  |  ST_Mes_2 = D
|  |  |  SP = IT
|  |  |  |  ST_Gjini = M : D (84/0)
|  |  |  |  ST_Gjini = F
|  |  |  |  |  ST_Shtet = A
|  |  |  |  |  |  VendB = A : D (19/0)
|  |  |  |  |  |  VendB = C : C (0/0)
|  |  |  |  |  |  VendB = B
|  |  |  |  |  |  |  ST_Mosha < 21.5 : D (29/3)
|  |  |  |  |  |  |  ST_Mosha >= 21.5 : D (22/2)
|  |  |  |  |  ST_Shtet = B : D (1/0)
|  |  |  SP = CS
|  |  |  |  VendB = A
|  |  |  |  |  ST_Mosha < 21.5 : D (45/0)
|  |  |  |  |  ST_Mosha >= 21.5
|  |  |  |  |  |  ST_Gjini = M : D (30/0)
|  |  |  |  |  |  ST_Gjini = F : D (31/1)
|  |  |  |  VendB = C : D (32/0)
|  |  |  |  VendB = B
|  |  |  |  |  ST_Mosha < 21.5
|  |  |  |  |  |  ST_Gjini = M : D (66/1)
|  |  |  |  |  |  ST_Gjini = F : D (61/0)
|  |  |  |  |  ST_Mosha >= 21.5 : D (84/0)
|  |  ST_Mes_2 = C
|  |  |  VendB = A
|  |  |  |  ST_Gjini = M
|  |  |  |  |  SP = IT
|  |  |  |  |  |  ST_Mosha < 21.5 : C (4/2)
|  |  |  |  |  |  ST_Mosha >= 21.5 : D (2/0)
|  |  |  |  |  SP = CS : D (7/0)
|  |  |  |  ST_Gjini = F
|  |  |  |  |  ST_Mosha < 21.5
|  |  |  |  |  |  SP = IT : D (1/0)
|  |  |  |  |  |  SP = CS : D (9/3)
|  |  |  |  |  ST_Mosha >= 21.5
|  |  |  |  |  |  SP = IT : C (2/0)
|  |  |  |  |  |  SP = CS : D (2/0)
|  |  |  VendB = C
|  |  |  |  SP = IT : D (1/0)
|  |  |  |  SP = CS
|  |  |  |  |  ST_Gjini = M
|  |  |  |  |  |  ST_Mosha < 21.5 : C (2/1)
|  |  |  |  |  |  ST_Mosha >= 21.5 : D (2/0)
|  |  |  |  |  ST_Gjini = F : C (2/1)
|  |  |  VendB = B
|  |  |  |  SP = IT
|  |  |  |  |  ST_Gjini = M
|  |  |  |  |  |  ST_Mosha < 21.5 : D (12/0)
|  |  |  |  |  |  ST_Mosha >= 21.5 : D (9/1)
|  |  |  |  |  ST_Gjini = F
|  |  |  |  |  |  ST_Mosha < 21.5 : D (7/0)
|  |  |  |  |  |  ST_Mosha >= 21.5 : D (6/1)
|  |  |  |  SP = CS
|  |  |  |  |  ST_Mosha < 21.5
```

```
|  |  |  |  |  |  ST_Gjini = M : D (12/4)
|  |  |  |  |  |  ST_Gjini = F : D (10/0)
|  |  |  |  |  ST_Mosha >= 21.5
|  |  |  |  |  |  ST_Gjini = M : D (7/3)
|  |  |  |  |  |  ST_Gjini = F : D (8/1)
|  |  ST_Mes_2 = B
|  |  |  SP = IT
|  |  |  |  VendB = A
|  |  |  |  |  ST_Gjini = M
|  |  |  |  |  |  ST_Mosha < 21.5 : C (6/3)
|  |  |  |  |  |  ST_Mosha >= 21.5 : D (3/1)
|  |  |  |  |  ST_Gjini = F
|  |  |  |  |  |  ST_Mosha < 21.5 : D (5/2)
|  |  |  |  |  |  ST_Mosha >= 21.5 : D (6/1)
|  |  |  |  VendB = C : C (0/0)
|  |  |  |  VendB = B
|  |  |  |  |  ST_Gjini = M
|  |  |  |  |  |  ST_Mosha < 21.5 : D (2/0)
|  |  |  |  |  |  ST_Mosha >= 21.5 : C (6/3)
|  |  |  |  |  ST_Gjini = F
|  |  |  |  |  |  ST_Mosha < 21.5 : D (12/4)
|  |  |  |  |  |  ST_Mosha >= 21.5 : D (10/1)
|  |  |  SP = CS
|  |  |  |  VendB = A
|  |  |  |  |  ST_Gjini = M : C (2/0)
|  |  |  |  |  ST_Gjini = F : C (2/1)
|  |  |  |  VendB = C : D (1/0)
|  |  |  |  VendB = B
|  |  |  |  |  ST_Mosha < 21.5 : C (4/0)
|  |  |  |  |  ST_Mosha >= 21.5 : C (2/1)
|  ST_Mes_1 = C
|  |  ST_Mes_2 = A : B (10/0)
|  |  ST_Mes_2 = D
|  |  |  VendB = A
|  |  |  |  ST_Shtet = A
|  |  |  |  |  ST_Mosha < 21.5
|  |  |  |  |  |  ST_Gjini = M : D (5/1)
|  |  |  |  |  |  ST_Gjini = F : D (3/1)
|  |  |  |  |  ST_Mosha >= 21.5
|  |  |  |  |  |  ST_Gjini = M : C (1/0)
|  |  |  |  |  |  ST_Gjini = F : D (1/0)
|  |  |  |  ST_Shtet = B : D (1/0)
|  |  |  VendB = C : D (2/0)
|  |  |  VendB = B
|  |  |  |  ST_Mosha < 21.5 : D (9/0)
|  |  |  |  ST_Mosha >= 21.5
|  |  |  |  |  SP = IT : D (1/0)
|  |  |  |  |  SP = CS
|  |  |  |  |  |  ST_Gjini = M : D (5/1)
|  |  |  |  |  |  ST_Gjini = F : D (6/1)
|  |  ST_Mes_2 = C
|  |  |  SP = IT : C (3/0)
|  |  |  SP = CS
|  |  |  |  ST_Mosha < 21.5 : C (26/0)
|  |  |  |  ST_Mosha >= 21.5
```

```
|  |  |  |  |  VendB = A
|  |  |  |  |  |  ST_Gjini = M : C (4/0)
|  |  |  |  |  |  ST_Gjini = F : C (2/1)
|  |  |  |  |  VendB = C : C (2/0)
|  |  |  |  |  VendB = B
|  |  |  |  |  |  ST_Gjini = M : C (6/1)
|  |  |  |  |  |  ST_Gjini = F : C (6/0)
|  |  ST_Mes_2 = B
|  |  |  ST_Shtet = A
|  |  |  |  VendB = A
|  |  |  |  |  ST_Gjini = M
|  |  |  |  |  |  ST_Mosha < 21.5 : C (2/1)
|  |  |  |  |  |  ST_Mosha >= 21.5 : C (3/0)
|  |  |  |  |  ST_Gjini = F : B (1/0)
|  |  |  |  VendB = C : C (1/0)
|  |  |  |  VendB = B
|  |  |  |  |  ST_Gjini = M
|  |  |  |  |  |  ST_Mosha < 21.5
|  |  |  |  |  |  |  SP = IT : C (1/0)
|  |  |  |  |  |  |  SP = CS : C (2/1)
|  |  |  |  |  |  ST_Mosha >= 21.5 : C (4/1)
|  |  |  |  |  ST_Gjini = F : C (2/0)
|  |  |  ST_Shtet = B : C (1/0)
|  ST_Mes_1 = B
|  |  ST_Mes_2 = A
|  |  |  SP = IT : A (2/0)
|  |  |  SP = CS
|  |  |  |  ST_Gjini = M
|  |  |  |  |  ST_Mosha < 21.5
|  |  |  |  |  |  VendB = A : A (1/0)
|  |  |  |  |  |  VendB = C : C (0/0)
|  |  |  |  |  |  VendB = B : B (1/0)
|  |  |  |  |  ST_Mosha >= 21.5 : B (1/0)
|  |  |  |  ST_Gjini = F : A (1/0)
|  |  ST_Mes_2 = D
|  |  |  ST_Gjini = M : C (1/0)
|  |  |  ST_Gjini = F : D (2/0)
|  |  ST_Mes_2 = C : C (3/0)
|  |  ST_Mes_2 = B
|  |  |  ST_Mosha < 21.5
|  |  |  |  VendB = A : B (2/0)
|  |  |  |  VendB = C : D (1/0)
|  |  |  |  VendB = B : B (7/0)
|  |  |  ST_Mosha >= 21.5 : B (7/0)
|  ST_Mes_1 = A
|  |  VendB = A : A (4/0)
|  |  VendB = C : A (1/0)
|  |  VendB = B
|  |  |  SP = IT : C (1/0)
|  |  |  SP = CS
|  |  |  |  ST_Mes_2 = A : A (3/0)
|  |  |  |  ST_Mes_2 = D : B (1/0)
|  |  |  |  ST_Mes_2 = C : B (1/0)
|  |  |  |  ST_Mes_2 = B : C (0/0)
```

```
Size of the tree : 273

Time taken to build model: 0.16 seconds
```

Parametrat e vlerësimit të performancës së algoritmit:

```
=== Evaluation on training set ===

Time taken to test model on training data: 0.05 seconds

=== Summary ===

Correctly Classified Instances        1232               93.2627 %
Incorrectly Classified Instances        89                6.7373 %
Kappa statistic                          0.8221
Mean absolute error                      0.0489
Root mean squared error                  0.1563
Relative absolute error                 25.1071 %
Root relative squared error             50.1772 %
Coverage of cases (0.95 level)          99.6215 %
Mean rel. region size (0.95 level)      33.4595 %
Total Number of Instances             1321

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.817    0.039    0.824      0.817    0.820      0.780    0.974     0.904     C
               0.972    0.145    0.955      0.972    0.963      0.844    0.982     0.993     D
               0.738    0.001    0.978      0.738    0.841      0.844    0.994     0.915     B
               0.941    0.000    1.000      0.941    0.970      0.970    0.999     0.962     A
Weighted Avg.  0.933    0.117    0.933      0.933    0.932      0.834    0.981     0.973
```

Matrica e çrregullimeve:

```
=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 196  43   1   0 |   a = C
  28 975   0   0 |   b = D
  14   2  45   0 |   c = B
   0   1   0  16 |   d = A
```

Pema e vendimit e gjeneruar për secilin nga testet është e njëjtë si strukturë (ndryshon vetëm në disa vlera të gjetheve për klasifikimin e instancave të gabuara), kështu që për tre testet e mëposhtme do të japim vetëm rezultatin për vlerësimin e performancës dhe matricën e çrregullimeve.

**Testi 2 – Vlerësimi i 3-fishtë**

Performanca e algoritmit:

```
Test mode:    3-fold cross-validation
```

```
=====.....======

Time taken to build model: 0.44 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1173               88.7964 %
Incorrectly Classified Instances      148                11.2036 %
Kappa statistic                          0.6974
Mean absolute error                      0.0767
Root mean squared error                  0.2137
Relative absolute error                 39.3364 %
Root relative squared error             68.5966 %
Coverage of cases (0.95 level)          96.5935 %
Mean rel. region size (0.95 level)      35.5602 %
Total Number of Instances             1321

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.654    0.049    0.748      0.654    0.698      0.638    0.902     0.736     C
                 0.963    0.239    0.927      0.963    0.945      0.758    0.939     0.973     D
                 0.607    0.012    0.712      0.607    0.655      0.642    0.940     0.692     B
                 0.765    0.003    0.765      0.765    0.765      0.762    0.967     0.845     A
Weighted Avg.    0.888    0.191    0.882      0.888    0.884      0.731    0.933     0.915
```

Matrica e çrregullimeve:

```
=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 157  71  12   0 |   a = C
  36 966   1   0 |   b = D
  17   3  37   4 |   c = B
   0   2   2  13 |   d = A
```

**Testi 3 – Vlerësimi i 13-fishtë**

Performanca e algoritmit:

```
Test mode:    13-fold cross-validation
=====.....======

Time taken to build model: 0.16 seconds

=== Stratified cross-validation ===
=== Summary ===
```

```
Correctly Classified Instances        1183               89.5534 %
Incorrectly Classified Instances       138               10.4466 %
Kappa statistic                          0.7188
Mean absolute error                      0.0726
Root mean squared error                  0.2036
Relative absolute error                 37.2652 %
Root relative squared error             65.3615 %
Coverage of cases (0.95 level)          97.0477 %
Mean rel. region size (0.95 level)      35.6927 %
Total Number of Instances             1321

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.708    0.050    0.759     0.708    0.733      0.676    0.913     0.755     C
                 0.964    0.220    0.932     0.964    0.948      0.774    0.942     0.973     D
                 0.557    0.007    0.791     0.557    0.654      0.651    0.933     0.691     B
                 0.706    0.004    0.706     0.706    0.706      0.702    0.967     0.858     A
Weighted Avg.    0.896    0.177    0.891     0.896    0.892      0.750    0.936     0.919
```

Matrica e çrregullimeve:

```
=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 170  65   5   0 |   a = C
  35 967   1   0 |   b = D
  19   3  34   5 |   c = B
   0   2   3  12 |   d = A
```

## Literatura


[1] Beck, J. E., and J. Mostow. 2008. "How Who Should Practice: Using Learning Decomposition to Evaluate the Efficacy of Different Types of Practice for Different Types of Students." In *Proceedings of the 9th International Conference on Intelligent Tutoring Systems*. Berlin, Heidelberg: Springer-Verlag, 353–362.

[2] Baker, R.S.J.d. (2007) Is Gaming the System State-or-Trait? Educational Data Mining Through the Multi-Contextual Application of a Validated Behavioral Model. Complete On-Line
Proceedings of the Workshop on Data Mining for User Modeling at the 11th International Conference on User Modeling 2007, 76-80.

[3]Baker, R (2010). Data Mining for Education. To appear in McGaw, B., Peterson, P., Baker, E. (Eds.) International Encyclopedia of Education (3rd edition). Oxford, UK: Elsevier.

[4] Baker, R.S.J.d., Corbett, A.T., Aleven, V. (2008) More Accurate Student Modeling Through Contextual Estimation of Slip and Guess Probabilities in Bayesian Knowledge Tracing.
Proceedings of the 9th International Conference on Intelligent Tutoring Systems, 406-415.

[5] Baker, R.S.J.d., Yacef K., The State of Educational Data Mining in 2009: A RevieW and Future Visions, 2009.

[6] Barnes, T., Bitzer, D., Vouk, M. (2005) Experimental Analysis of the Q-Matrix Method in Knowledge Discovery. Lecture Notes in Computer Science 3488: Foundations of Intelligent Systems, 603-611.

[7] Castro, F., Vellido, A., Nebot, A. Mugica, F. (2007). Applying Data Mining Techniques to e-Learning Problems. In: Jain, L.C., Tedman, R. and Tedman, D. (eds.) Evolution of Teaching and Learning Paradigms in Intelligent Environment. Studies in Computational Intelligence, 62, Springer-Verlag, 183-221.

[8] Chang, K.M., Beck, J.E., Mostow, J., Corbett, A. (2006). A Bayes Net Toolkit for Student Modeling in Intelligent Tutoring Systems. In International Conference on Intelligent Tutoring Systems, Jhongli, Taiwan, 104-113.

[9] Corbett, J. (2001) Supporting Inclusive Education: A Connective Pedagogy. London: Routledge Falmer.

[10] Cortez P., Silva A.. Using data mining to predict secondary school student performance. In the Proceedings of 5th Annual Future Business Technology Conference, Porto, Portugal. 2008

[11] Çomo A., Ninka I., Munguli B. (2015) Using Classification Methods in Higher Education. International Journal of Basic Sciences and Applied Computing (IJBSAC). Impact Factor 0.48. ISSN: 2394-367X (Online), Volume-1 Issue-8, Page No.: 5-8, August 2015.

[12] Dekson.D.E, Jaichandran.R (2015) Neural Network Based Performance Monitoring System For An E-Learning Portal. International Conference On Recent Advancement In Mechanical Engineering &Technology (Icramet' 15) Issn: 0974-2115

[13] Gong, Y., Rai, D., Beck, J. E. & Heffernan, N. T. (2009) Does Self-Discipline Impact Students Knowledge and Learning? . In Barnes, Desmarais, Romero & Ventura (Eds) Proc. of the 2nd International Conference on Educational Data Mining. pp. 61-70. ISBN: 978-84-613-2308-1.

[14] Gopalakrishnan T., Gowthami V. S., Kavya M. (2014) A Survey on Various Learning Styles Used in E-Learning System. International Journal of Modern Trends in Engineering (IJMTER-2014), Impact Factor (SJIF): 1.711 e-ISSN: 2349-9745

[15] Han J. & Kamber M. (2006), Data Mining Concepts and Techniques, 2nd Edition. Morgan Kaufman Publisher. ISBN: 1-55860-901-6

[16]Hanna, M. (2004). Data mining in the e-learning domain. In CampusWide Information Systems, Volume 21, Number 1, 29-34.

[17] Heathcote, E., Dawson, S. (2005). Data Mining for Evaluation, Benchmarking and Reflective Practice in a LMS. In World conference on E-learning in corporate, government, healthcare & higher education, Vancouver, Canada, 326-333.

[18] Ho T.K (2002), Multiple Classfier combination: lessons and next Steps. In: Kandel, Bunken E. (eds) Hybrid Methos in Pattern Recognition, pp171-198. World Scientific, New York.

[19]Ian H. Witten, Eibe Frank, Mark A. Hall, Third Edition (2005); Data Mining, Practical Machine Learning Tools and Techniques;

[20] Jonsson, A., Hasmik, J. Johns, Mehranian, H., Arroyo, I., Woolf, B., Barto, A., Fisher, D., Mahadevan, S. (2005). Evaluating the Feasibility of Learning Student Models from Data, In Educational Data Mining AAAI Workshop, Pittsburgh, 1-6.

[21] Klösgen W., Zytkow J. (2002) Handbook of Data Mining and knowledges discovery, Oxford Univ. Press, Oxford.

[22]Krzysztof J. Cios, Witold Pedrycz, Roman W. Swiniarski, Lukasz Kurgan (2007). Data Mining: A Knowledge Discovery Approach. Springer Science & Business Media. E-ISBN: 978-0-387-36795-8

[23] Kumar S. A. & Vijayalakshmi M. N. (2011), Efficiency of Decision Trees in Predicting Student's Academic Performance, First International Conference on Computer Science, Engineering and Applications, CS and IT 02, Dubai, pp. 335-343.

[24] Madhyastha, T. and Tanimoto, S. 2009. Student Consistency and Implications for Feedback in Online Assessment Systems. In Proceedings of the 2nd International Conference on Educational Data Mining, 81-90.

[25].Machine Learning Group at University of Waikato, Available on: http://www.cs.waikato.ac.nz/ml/weka/

[26] Maimon O., Rokach L., Data Mining and Knowledges Discovery Handbook. Chaper 1 – Introduction to Knowledge Discovery in Databases.

[27] Malhotra, N. K., and Peterson, M. (2006). Basic Research Marketing: A Decision-Making Approach (2 nd Ed.). New Jersey: Pearson Education, Inc

[28]Minaei-Bidgoli, B., Kashy, D.A., Kortemeyer G. & Punch WF (2003), Predicting student performance: an application of data mining methods With the educational Web-based systems LON-CAPA, 33rd ASEE/IEEE Frontiers in Education Conference, Boulder, pp.13-18.

[29] Mitra S. & Archarya T. (2003), Data Mining: Multimedia, Soft Computing and Bioinformatics. Wiley-Interscience; 1 edition. ISBN: 978-0471460541

[30] Mohammed M. Abu Tair, Alaa M. El-Halees, Mining Educational Data to Improve Students' Performance: A Case Study. International Journal of Information and Communication Technology Research, Volume 2 No. 2, February 2012

[31] Murthy S. K. (1998) Automatic construction of decision trees from data: A multidisciplinary survey. Data Mining and Knowledge Discovery f. 345–389

[32] Oprea, C. (2014) Performance Evaluation Of The Data Mining Classification methods. Annals of the "Constantin Brâncuşi" University of Târgu Jiu, EconomySeries,Special Issue/2014- Information society and sustainable development "ACADEMICA BRÂNCUŞI" PUBLISHER, ISSN 2344 – 3685/ISSN-L 1844 - 7007

[33]Pavlik, P., Cen, H., Wu, L., Koedinger, K. (2008) Using Item-Type Performance Covariance to Improve the Skill Model of an Existing Tutor. *Proceedings of the First International Conference on Educational Data Mining*, 77-86.

[34] Perera, D., Kay, J., Koprinska, I., Yacef, K. and Zaiane, O. 2009. Clustering and sequential pattern mining to support team learning. IEEE Transactions on Knowledge and Data Engineering 21, 759-772

[35]Rahkila, M., Karjalainen, M. (1999). Evaluation of learning in computer based education using log systems. In ASEE/IEEE frontiers in education conference, San Juan, Puerto Rico, 16–21.

[36]Ramli, A.A. (2005). Web usage mining using apriori algorithm: UUM learning care portal case. In International conference on knowledge management, Malaysia, 1-19.

[37]Romero, C. & Ventura, S. (2007), Educational Data Mining: a Survey from 1995 to 2005, Expert Systems With Applications, Elsevier, pp. 135-146

[38] Romeo C. & Ventura S. (2010) Educational Data Mining: A review of the state of art. IEEE Transaction Systems Man and Cybernetics Part C (Applications and Reviews). Impact Factor 2.17.

[39] Romero C., Ventura S., Espejo, P.G. & Hervás C. (2008), Data mining algorithms to classify students, I International Conference on Educational Data Mining (EDM). Montreal, pp. 8-17.

[40] Romero, C., Ventura, S., Pechenizkiy, M., & Baker, R.S.J.d. (Eds.). (2010): Handbook of Educational Data Mining. CRC Press

[41] Saunders M., Lewis P. And Thornhill A. (2000), Research Method for Business Students, England, Pearson Education Limited

[42] Scott W. McQuiggan, Jonathan P. Rowe, Sunyoung Lee, and James C. Lester (2008) Story-based Learning: The Impact of Narrative on Learning Experiences and Outcomes.

Intelligent Tutoring Systems Volume 5091 of the series Lecture Notes in Computer Science pp 530-539

[43] Superby J.F., Vandamme J.P. & Meskens N. (2006), Determination of Factors Influencing the Achievement of the First-year University Students using Data Mining Methods, Proceedings of the 8th international conference on intelligent tutoring systems, Educational Data Mining Workshop, (ITS`06), Jhongali, pp. 37-44.

[44]Witten, Ian H. and Frank, Eibe (2005), "Data Mining: Practical machine learning tools and techniques", 2nd Edition, Morgan Kaufmann, San Francisco.

[45] Wu, A., Leung, C. (2002). Evaluating learning behavior of Web-Based Training (WBT) using Web log. In International Conference on Computers in Education, New Zealand, 736-737.

[46]Yu. C.H., Digangi, S., Jannasch-pennell, A.K., Kaprolet, C. (2008). Profiling students who take online courses using data mining methods. In Online Journal of Distance Learning Administration, XI(II), 1-14. [293]Zafra, A., Ventura, S. (2009), Predicting student grades in learnin

[47]Zekić-Sušac M., Frajman-Jakšić A. & Drvenkar N., (2009), Neuron NetWorks and Trees of Decision-making for Prediction of Eficiency in Studies. Ekonomski Vjesnik(2), 314-327

[48] Zheng, S. Xiong, S., Huang, Y., Wu, S. (2008). Using methods of association rules mining optimization in web-based mobile learning system. In International Symposium on Electronic Commerce and Security. Guangzhou, China, 967- 970.

[49] Zikmund WG, (2003). Business research methods. 7th ed. Ohio: Thomson Learning

[50] Zorrilla, M.. E., Menasalvas, E., Marin, D., Mora, E., Segovia, J. (2005). Web usage mining project for improving web-based learning sites. In International Conference on Computer Aided Systems Theory, Las Palmas de Gran Canaria, Spain, 205-210.

[51]Zukhri, Z., Omar, K. (2007). Solving new student allocation problem with genetic algorithms.

# Përmbledhje

*Për institucionet e arsimit të lartë, qëllimi i të cilave është të kontribuojnë në përmirësimin e cilësisë, suksesi i krijimit të kapitalit human është subjekti i një analize të vazhdueshme. Për këtë arsye parashikimi i suksesit të studentëve është thelbësor për institucionet e arsimit të lartë, sepse cilësia e procesit të të mësuarit është aftësia për të kënaqur dhe plotësuar nevojat e studentëve. Në këtë drejtim janë mbledhur të dhëna dhe informacione të rëndësishme, të cilat janë vënë në dispoziocion dhe të autoriteteve përkatëse, në mënyrë që të ruhet cilësia. Cilësia e institucioneve të arsimit të lartë nënkupton sigurimin e shërbimeve, të cilat plotësojnë më së shumti nevojat e studentëve, stafit akademik, dhe pjesëmarrësve të tjerë në sistemin arsimor. Pjesëmarrësit në procesin arsimor, duke përmbushur detyrimet e tyre nëpërmjet veprimtarive të duhura, krijojnë një sasi tepër të madhe të dhënash të cilat kanë nevojë të mblidhen dhe më pas të integrohen dhe të përdoren. Duke shndërruar këto të dhëna në njohuri, arrihet kënaqësia e të gjithë pjesëmarrësve : studentëve, profesorëve, administratës dhe komunitetit shoqëror.*

*Megjithëse për shumë kohë data mining është zbatuar me sukses në botën e biznesit, përdorimi i saj në arsimin e lartë është ende relativisht i ri. Përdorimi i data mining në arsim nënkupton identifikimin dhe nxjerrjen nga të dhënat të njohurive të reja dhe shumë të vlefshme. Qëllimi i përdorimit të data mining ishte zhvillimi i një modeli nga i cili mund të nxirren përfundime mbi suksesin akademik të studentëve. Metoda dhe teknika të ndryshme të data mining janë krahasuar gjatë këtij punimi duke shqyrtuar të gjithë variablat që mund të kenë ndikim në suksesin e studentit, si variablat shoqërore, demografike, etj.*

*Rezultatet e marra nga aplikimi i metodave të data mining, ndihmojnë institucionet e arsimit të lartë për të parashikuar sjelljen e studentëve. Kështu që i gjithë qëllimi i studimit do të përmblidhej në një pyetje të vetme, që do të ishte:*

- *Si mund të përdoren të dhënat e përftuara nga mjedisi arsimor dhe teknikat e data mining, për parashikimin e suksesit të studentit në arsimin e lartë?*

*Për të gjetur zgjidhjen më të mirë për problemin e përmendur, fillimisht mblidhen dhe përpunohen të dhënat nga mjedisi arsimor dhe më pas zbatohen teknika të caktuara të data mining sipas aftësisë që shfaq secila teknikë për të bërë parashikimin më të saktë dhe për të mundësuar krijimin e një modeli i cili do të shërbejë si një bazë për zhvillimin e një sistemi vendimesh mbështetës në arsimin e lartë.*

## Absctract

*For the institutions of higher education, whose focus is the improvement of education quality, the success in creating human capital is subject to a continuous analysis. For this reason, sudent success forecast is crucial per these institutions, because the quality of the learning process is the ability to satisfy and fulfill the students needs. With this in mind, important information and data has been gathered and has been given to the appropriate authorities so that quality could be preserved. The quality of the higher education institutions means offering the services that fulfill as well as possible the needs of students, academic staff and other partecipants in the education system. The actors in the education process, by fulfilling their obligations through everyday actions, create a big quanitity of data which needs to be collected, processed and used. By transforming this data into knowledge we can achieve the happiness of all partecipants in the education process: studens, professors, administration and social comunity.*

*Although datamining has been used for a long time in the business world, its use in the higher education institutions is relatively recent. The use of datamining in education means the identification and exctraction of new and useful knowledge from existing data. The purpose of the data mining usage was the development of a model that could be used to predict the academic success of the students.During this study, different methods and techiques have been compared by observing all the variables that might influence the sutdent succes, such as sociological variables, demografic ones etc.*

*The results from the aplication of data mining methods, help higher education institutions to predict the student behaviour. If the whole study could be concentrated into one question, this question would be:*

- *How can data mining techniques be used on data gathered from education institutions, to predict the success of sudents in higher education?*

*To find the best possible solution to the up-mentioned problem, data was gathered and prepared from the education system, then specified data mining techniques are applied according to the ability of each technique to predict as correctly as possible, and to enable the creation of a new model that will server as the foundation of a decision system to support high education institutions.*