

# Big Data Mining

Illir Perolli

Universiteti "Ukshin Hoti"

Fakulteti Shkencave Kompjuterike

Prizren

Email: 200327006.m@uni-prizren.com

**Abstrakti—** Nxjerrja e të dhënave është një proces i nxjerrjes së informacionit të fshehur, të panjohur, por potencialisht të dobishëm nga të dhënat masive. Big Data ka ndikime të mëdha në zbulimet shkencore dhe krijimin e vlerës. Ky punim prezanton metodat në nxjerrjen e të dhënave dhe teknologjitë në Big Data. Diskutohen sfidat e nxjerrjes së të dhënave dhe nxjerrjes së të dhënave nga të dhëna të mëdha.

**Fjalët kyçe—**big data, data mining, Big Data Analytics

## I. HYRJE

Big Data Mining (nxjerrja e të dhënave të mëdha) i referohet teknikave kolektive të nxjerrjes së të dhënave që kryhen në grupe të mëdha / vëllim të dhënash ose të dhëna të mëdha. Nxjerrja e madhe e të dhënave bëhet kryesisht për të nxjerrë dhe për të marrë informacionin ose modelin e dëshiruar nga sasia e madhe e të dhënave. Kjo zakonisht kryhet në sasi të mëdha të të dhënave të pastruara që ruhen me kalimin e kohës nga një organizatë. Në mënyrë tipike, nxjerrja e të dhënave të mëdha punon në kërkimin e të dhënave, përsosjen, nxjerrjen dhe algoritmeve të krahahasimit. Nxjerrja e të dhënave të mëdha kërkon gjithashtu mbështetje nga pajisjet themelore të llogaritjes, posaçërisht procesorët dhe kujtesën e tyre, për kryerjen e operacioneve / pyetjeve (queries) në sasi të madhe të të dhënave. Teknikat dhe proceset në Big data mining përdoren gjithashtu brenda analizave të të dhënave të mëdha dhe inteligjencës së biznesit për të dhënë informacion të përmbledhur të synuar dhe të rëndësishëm, modele (patterns) dhe / ose marrëdhënie midis të dhënave, sistemeve, proceseve dhe më shumë [1].

## II. ÇFARË ËSHTË DATA MINING?

Nxjerrja e të dhënave është një teknikë për zbulimin e modeleve interesante, si dhe modeleve përshkruese dhe të kuptueshme nga të dhënat në shkallë të gjerë. Data mining mund të përdoret për të gjetur korrelacione ose modele midis dhjetra fushave në bazën e të dhënave të mëdha relacionale. Nxjerrja e të dhënave është gjithashtu procesi i zbulimit ose gjetjes së disa formave të reja, të vlefshme, të kuptueshme dhe potencialisht të dobishme të të dhënave [2].

Procesi i gërmimit të të dhënave për të zbuluar lidhjet e fshehura dhe për të parashikuar trendet e ardhshme ka një histori të gjatë. Ndonjëherë referuar si "zbulimi i njohurive

në bazat e të dhënave", termi "miniera e të dhënave" nuk u krijua deri në vitet 1990. Por themeli i tij përfshin tre disiplina të ndërthurura shkencore: statistikën (studimi numerik i marrëdhënieve të të dhënave), inteligjenca artificiale (inteligjenca e ngjashme me njeriun e shfaqur nga softueri dhe / ose makinat) dhe të mësuarit makinerik (algoritme që mund të mësojnë nga të dhënat për të bërë parashikime). Ajo që ishte e vjetër është përsëri e re, pasi teknologjia e nxjerrjes së të dhënave vazhdon të evoluojë për të ecur me hapin me potencialin e pakufizuar të të dhënave të mëdha dhe fuqisë informatike të përballueshme [3].

Gjatë dekadës së fundit, përparimet në fuqinë dhe shpejtësinë e përpunimit na kanë mundësuar që të shkojmë përtej praktikave manuale, të lodhshme dhe që marrin shumë kohë në analizën e shpejtë, të lehtë dhe të automatizuar të të dhënave. Sa më komplekse të dhënat e mbledhura, aq më shumë potencial ka për të zbuluar njohuritë përkatëse. Shitësit me pakicë, bankat, prodhuesit, ofruesit e telekomunikacionit dhe siguruesit, ndër të tjera, po përdorin miniera të të dhënave për të zbuluar marrëdhënie midis gjithçkaje, nga optimizimi i çmimeve, promovimet dhe demografia të mënyra se si ekonomia, rreziku, konkurrenca dhe mediat sociale po ndikojnë në modelet e tyre të biznesit, të ardhurat, operacionet dhe marrëdhëniet me klientët [3].

Nxjerrja e të dhënave është një proces i përdorur nga kompanitë për t'i kthyer të dhënat e papërpunuara në informacion të dobishëm. Duke përdorur softuer për të kërkuar modele në grupe të mëdha të të dhënave, bizneset mund të mësojnë më shumë rreth klientëve të tyre për të zhvilluar strategji më efektive të marketingut, për të rritur shitjet dhe për të ulur kostot. Nxjerrja e të dhënave varet nga mbledhja efektive e të dhënave, deponimi dhe përpunimi kompjuterik [4].

### A. Etimologjia

Në vitet 1960, statisticienët dhe ekonomistët përdorën terma si peshkimi i të dhënave ose gërmimi i të dhënave për t'iu referuar asaj që ata e konsideruan praktikën e keqe të analizimit të të dhënave pa një hipotezë a-priori. Termi "miniera e të dhënave" u përdor në një mënyrë të ngjashme kritike nga ekonomisti Michael Lovell në një artikull të botuar në Revista e Studimeve Ekonomike në 1983. Lovell tregon se praktika "maskohet nën një shumëllojshmëri pseudonimesh, duke filluar nga" eksperimentimi "(pozitiv) te" peshkimi "ose" përgjimi "(negativ) [5][6].

## B. Histori rreth Data Mining

Nxjerrja manuale e modeleve nga të dhënat ka ndodhur me shekuj. Metodatat e hershme të identifikimit të modeleve në të dhëna përfshijnë teoremën e Bayes (vitet 1700) dhe analizën e regresionit (vitet 1800). Përhapja, mbizotërimi dhe fuqia në rritje e teknologjisë kompjuterike kanë rritur në mënyrë dramatike aftësinë e mbledhjes, ruajtjes dhe manipulimit të të dhënave [6].

Ndërsa grupet e të dhënave janë rritur në madhësi dhe kompleksitet, analiza e drejtpërdrejtë e të dhënave "praktike" është shtuar gjithnjë e më shumë me përpunimin indirekt, të automatizuar të të dhënave, të ndihmuar nga zbulime të tjera në shkencën kompjuterike, posaçërisht në fushën e të mësuarit të makinës, të tilla si rrjetet nervore, analiza e grupeve, algoritme gjenetike (vitet 1950), pemë vendimesh dhe rregulla vendimi (vitet 1960) dhe makina vektoriale mbështetëse (vitet 1990) [6].

Nxjerrja e të dhënave është procesi i zbatimit të këtyre metodave me synimin për të zbuluar modele të fshehura në grupe të mëdha të të dhënave. Ajo kapërcen hendekun nga statistikat e aplikuara dhe inteligjenca artificiale (të cilat zakonisht ofrojnë sfondin matematikor) të menaxhimit i bazës së të dhënave duke shfrytëzuar mënyrën e ruajtjes dhe indeksimit të të dhënave në bazat e të dhënave për të ekzekutuar algoritmet aktuale të mësimit dhe zbulimit në mënyrë më efikase, duke lejuar që metoda të tilla të zbatohen në grupe gjithnjë e më të mëdha të të dhënave [6].

## C. Pse është e rëndësishme kërkimi i të dhënave?

### Atëherë pse është e rëndësishme nxjerrja e të dhënave?

Ju keni parë numrat marramendës - vëllimi i të dhënave të prodhuara dyfishohet çdo dy vjet. Vetëm të dhënat e pastruara përbëjnë 90 përqind të universit dixhital. Por më shumë informacion nuk do të thotë domosdoshmërisht më shumë njohuri [3].

Nxjerrja e të dhënave ju lejon të:

- Shikoni të gjithë zhurmën kaotike dhe të përsëritur në të dhënat tuaja.
- Kuptoni se çfarë është relevante dhe më pas shfrytëzoheni mirë atë informacion për të vlerësuar rezultatet e mundshme.
- Përsheptoni ritmin e marrjes së vendimeve të informuara.

## D. Miniera e të Dhënave dhe Statistikë

Ekziston një mbivendosje e madhe midis kërkimit të të dhënave dhe statistikave. Në fakt, shumica e teknikave të përdorura në kërkimin e të dhënave mund të vendosen në një kornizë statistikore. Sidoqoftë, teknikat e kërkimit të të dhënave nuk janë të njëjta me teknikat tradicionale statistikore [7].

Metodat statistikore tradicionale, në përgjithësi, kërkojnë shumë ndërveprime të përdoruesve për të vërtetuar korrektësinë e një modeli. Si rezultat, metodat statistikore mund të jenë të vështira për t'u automatizuar. Për më tepër, metodat statistikore zakonisht nuk shkojnë mirë në grupe

shumë të mëdha të të dhënave. Metodatat statistikore mbështeten në testimin e hipotezave ose gjetjen e korrelacioneve bazuar në mostra më të vogla, përfaqësuese të një popullsie më të madhe [7].

Metodat e nxjerrjes së të dhënave janë të përshtatshme për grupe të mëdha të të dhënave dhe mund të automatizohen më lehtë. Në fakt, algoritmet e minierave të të dhënave shpesh kërkojnë grupe të mëdha të dhënash për krijimin e modeleve të cilësisë [7].

## E. Si funksionon Data mining?

Nxjerrja e të dhënave përfshin eksplorimin dhe analizimin e blloqeve të mëdha të informacionit për të mbledhur modele dhe trende kuptimplota. Mund të përdoret në mënyra të ndryshme, të tilla si marketing i bazës së të dhënave, menaxhim i rrezikut të kredisë, zbulim i mashtrimit, filtrim i postës elektronike të bezdisshme, apo edhe për të dalluar ndjenjën ose mendimin e përdoruesve.

Procesi i nxjerrjes së të dhënave ndahet në pesë hapa. Së pari, organizatat mbledhin të dhëna dhe i ngarkojnë ato në magazinat e tyre të të dhënave. Tjetra, ata ruajnë dhe menaxhojnë të dhënat, ose në serverat e brendshëm ose në cloud. Analistët e biznesit, ekipet e menaxhimit dhe profesionistët e teknologjisë së informacionit hyjnë në të dhëna dhe përcaktojnë se si ata duan t'i organizojnë ato. Pastaj, softueri i aplikimit rendit të dhënat bazuar në rezultatet e përdoruesit, dhe së fundmi, përdoruesi përfundimtar paraqet të dhënat në një format të lehtë për t'u ndarë, të tilla si një grafik ose tabelë [8].

## F. Shembull i nxjerrjes së të dhënave

Dyqanet ushqimore janë përdorues të njohur të teknikave të data mining. Shumë supermarkete ofrojnë karta besnikërie falas për klientët që u japin atyre mundësi për të ulur çmimet që nuk janë në dispozicion për jo-anëtarët. Kartat e bëjnë më të lehtë për dyqanet të gjurmojnë se kush po blen çfarë, kur e blejnë atë dhe me çfarë çmimi. Pas analizimit të të dhënave, dyqanet më pas mund t'i përdorin këto të dhëna për t'u ofruar klientëve kupona të synuara për zakonet e tyre të blerjes dhe të vendosin kur t'i vendosin artikujt në shitje ose kur t'i shesin me çmim të plotë [8].

Nxjerrja e të dhënave mund të jetë një shqetësim kur një kompani përdor vetëm informacione të zgjedhura, të cilat nuk janë përfaqësuese të grupit të përgjithshëm të mostrës, për të provuar një hipotezë të caktuar [8].

### III. ÇFARË ËSHTË BIG DATA?

Të dhënat e mëdha është një term që përshkruan vëllimin e madh të të dhënave - të strukturuar dhe të pastrukturuar - që përmbajt një biznes në baza ditore. Por nuk është sasia e të dhënave që është e rëndësishme. Është ajo që bëjnë organizatat me të dhënat që kanë rëndësi. Të dhënat e mëdha mund të analizohen për njohuri që çojnë në vendime më të mira dhe lëvizje strategjike të biznesit [9].

Të dhënat e mëdha është një fushë që trajton mënyra për të analizuar, nxjerrë në mënyrë sistematike informacione, ose përndryshe merren me grupe të dhënash që janë shumë të mëdha ose komplekse për t'u trajtuar nga programi tradicional i aplikimit për përpunimin e të dhënave. Të dhënat me shumë raste (rreshta) ofrojnë fuqi më të madhe statistikore, ndërsa të dhënat me kompleksitet më të lartë (më shumë attribute ose kolona) mund të çojnë në një normë më të lartë të zbulimit të rremë. Sfidat e të dhënave të mëdha përfshijnë kapjen e të dhënave, ruajtjen e të dhënave, analizën e të dhënave, kërkimin, ndarjen, transferimin, vizualizimin, pyetjen, azhurnimin, privatësinë e informacionit dhe burimin e të dhënave. Të dhënat e mëdha fillimisht u shoqëruan me tre koncepte kryesore: vëllimi, shumëllojshmëria dhe shpejtësia. Kur trajtojmë të dhëna të mëdha, mund të mos marrim shembull por thjesht vëzhgojmë dhe gjurmojmë se çfarë ndodh. Prandaj, të dhënat e mëdha shpesh përfshijnë të dhëna me madhësi që tejkalojnë kapacitetin e softverit tradicional për tu përpunuar brenda një kohe dhe vlere të pranueshme [10].

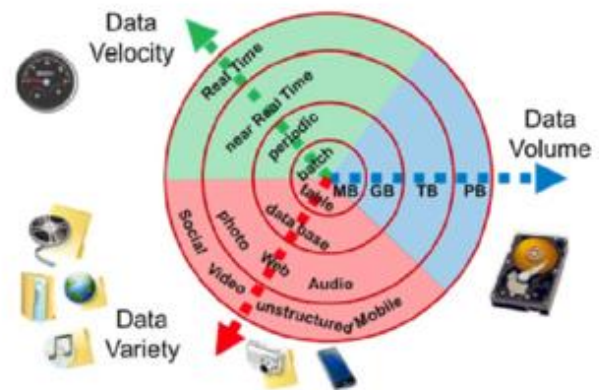
Përdorimi aktual i termit të dhëna të mëdha ka të bëjë me përdorimin e analizave parashikuese, analizave të sjelljes së përdoruesit, ose disa metodave të tjera të avancuara të analizave të të dhënave që nxjerrin vlerë nga të dhënat, dhe rrallë në një madhësi të veçantë të të dhënave. "Ka pak dyshim se sasi të të dhënave tani në dispozicion janë me të vërtetë të mëdha, por kjo nuk është karakteristikë më e rëndësishme e këtij ekosistemi të ri të të dhënave. Shkencëtarët, drejtuesit e biznesit, praktikuesit e mjekësisë, reklamantët dhe qeveritë në mënyrë të rregullt takojnë vështirësi me grupe të mëdha të dhënash në zona duke përfshirë kërkimet në internet, fintech, informatikën urbane dhe informatikën e biznesit [11].

Termi "të dhëna të mëdha" i referohet të dhënave që janë aq të mëdha, të shpejta ose komplekse sa është e vështirë ose e pamundur të përpunohen duke përdorur metoda tradicionale. Akti i hyrjes dhe ruajtjes së sasive të mëdha të informacionit për analizat ka qenë prej kohësh. Por koncepti i të dhënave të mëdha fitoi vull në fillimin e viteve 2000 kur analisti i industrisë Doug Laney artikuloi përkufizimin aktual të të dhënave të mëdha:

**Vëllimi:** Organizatat mbledhin të dhëna nga një larmi burimesh, duke përfshirë transaksione biznesi, pajisje inteligjente (IoT), pajisje industriale, video, media sociale dhe më shumë. Në të kaluarën, ruajtja e tij do të kishte qenë një problem - por ruajtja më e lirë në platforma si data lakes dhe Hadoop e kanë lehtësuar barren [9].

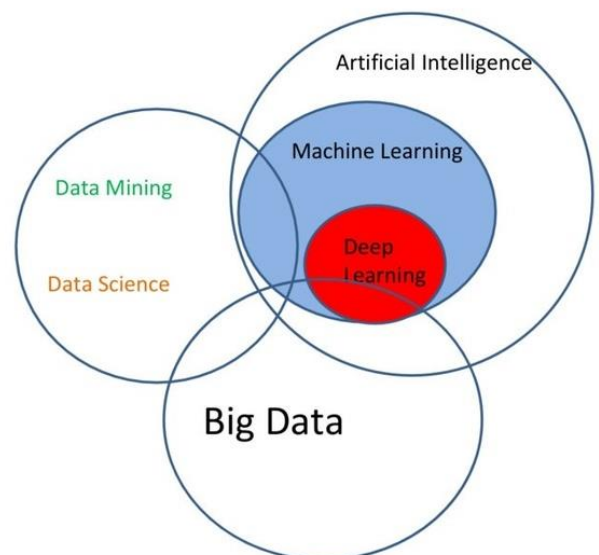
**Shpejtësia:** Me rritjen e IoT, të dhënat kalojnë tek bizneset me një shpejtësi të paparë dhe duhet të trajtohen në kohën e duhur. Etiketat, sensorët dhe matësit inteligjentë të RFID po nxisin nevojën për t'u marrë me këto të dhëna në kohë gati reale [9].

**Shumëllojshmëria:** Të dhënat vijnë në të gjitha llojet e formateve - nga të dhënat numerike të strukturuar në bazat e të dhënave tradicionale deri te dokumentet e pa strukturuar të tekstit, postat elektronike, videot, audiot, të dhënat e aksioneve dhe transaksionet financiare [9].



### IV. CILI ËSHTË NDRYSHIMI MIDIS BIG DATA DHE DATA MINING?

Data Mining dhe Big data janë dy gjëra të ndryshme, ndërsa të dyja kanë të bëjnë me përdorimin e grupeve të mëdha të të dhënave për të trajtuar të dhënat që do t'i shërbejnë qëllimit tonë, ato janë dy terma të ndryshëm në aspektin e funksionimit për të cilin përdoren. Big Data i referohet një koleksioni të grupeve të të dhënave të mëdha (p.sh.- të dhëna në fletët Excel të cilat janë shumë të mëdha për tu trajtuar me lehtësi). Mining Data nga ana tjetër i referohet aktivitetit të kalimit nëpër një pjesë të madhe të të dhënave për të kërkuar informacione përkatëse ose përkatëse.



## V. PSE ËSHTË E RËNDËSISHME BIG DATA?

Shumica e gjërave në skenarin e sotëm drejtohen nga përfitimi që japin në drejtim të përfitimeve monetare, këto mjete ndihmojnë në sigurimin e informacionit kuptimplotë për marrjen e vendimeve më të mira të biznesit dhe gjithashtu mund të përdoren për të studiuar gjëra të ndryshme të tjera që mund të përfitojnë nga njerëzimi.

Rëndësia e të dhënave të mëdha nuk sillet rreth asaj se sa të dhëna keni, por asaj që bëni me to. Ju mund të merrni të dhëna nga çdo burim dhe t'i analizoni ato për të gjetur përgjigje që mundësojnë:

- 1) zvogëlimin e kostos,
- 2) uljen e kohës,
- 3) zhvillimin e produktit të ri dhe ofertat e optimizuara,
- 4) vendimmarrjen inteligjente.

Kur kombinoni të dhëna të mëdha me analitikë me fuqi të lartë, mund të përmbushni detyra të lidhura me biznesin si:

- Përcaktimi i shkaqeve rrënjësore të dështimeve, çështjeve dhe defekteve në kohë gati reale.
- Gjenerimi i kuponëve në pikën e shitjes bazuar në zakonet e blerjes së klientit.
- Rillogaritja e plotë e portofoleve të rrezikut në minuta.
- Zbulimi i sjelljes mashtruese para se të ndikojë në organizatën tuaj [9].

## VI. PSE ËSHTË E RËNDËSISHME DATA MINING?

Minifikimi i të dhënave është i rëndësishëm për shkak të arsyeve të ndryshme, më e rëndësishmja dhe e dobishme prej tyre është të kuptosh se çfarë është e rëndësishme dhe ta përdorësh mirë për të vlerësuar gjërat ndërsa të dhënat e reja vijnë në skenë, kjo nga ana tjetër degëzohet në raste të ndryshme përdorimi në vende si industria e kujdesit shëndetësor, analiza e tregut financiar etj [12].

## VII. KRAHASIMI MES BIG DATA DHE DATA MINING

Duke i kuptuar mjaft mirë të dy konceptet, mund të themi se ato janë 2 koncepte shumë të ndryshme. Koncepti kryesor nëse shikojmë në Data Mining është të gërmojmë në të dhëna dhe të analizojmë modelin dhe marrëdhënien që mund të jetë më e dobishme në algoritmet e parashikimit si Linear Regresioni në Inteligjencën Artificiale. Koncepti kryesor në Big Data nga ana tjetër është shpejtësia, burimi, siguria e sasisë së madhe të të dhënave që kemi në dispozicion.

Mund të thuhet se Data Mining nuk varet nga Big Data, pasi mund të bëhet nga çdo sasi e të dhënave (preferenciale e madhe, pasi jep më shumë raste testimi dhe rrjedhimisht rezultate të sakta) qofshin ato të mëdha apo të vogla. Nga ana tjetër Big Data është shumë e varur nga Data mining pasi kemi nevojë për të gjetur përdorimin e vëllimit të madh të të dhënave që kemi, nuk ka dobi pa analizën e tyre.

## VIII. SA REALISHT JANË TË MËDHA BIG DATA?

Ne kemi hyrë në epokën e të dhënave për të mirë. Gjithçka që bën në internet, madje edhe jashtë linjës, lë gjurmë në të njëjtën gjë - nga cookies në profilet e mediave sociale. Pra, sa të keni njohur të vërtetë? Sa për të bërë në baza ditore? Mirësevini në epokën Zettabyte [13].

Të dhënat maten në bit dhe bajte. Një bit përmban një vlerë prej 0 ose 1. Tetë bit bëjnë një bajt. Pastaj kemi kilobajt (1000 bajt), megabajt (1000<sup>2</sup> bajt), gigabajt (1000 (bajt), terabajt (1000<sup>4</sup> bajt), petabajt (1000<sup>5</sup> bajt), ekzabajt (1000<sup>6</sup> bajt) dhe zettabajt (1000<sup>7</sup> bajt) [13].

Trafiku në internet është vetëm një pjesë e ruajtjes totale të të dhënave, e cila përfshin gjithashtu të gjitha pajisjet personale dhe të biznesit. Vlerësimet për kapacitetin total të ruajtjes së të dhënave që kemi tani, në vitin 2019, ndryshojnë, por tashmë janë në diapazonin 10-50 zettabyte. Deri në vitin 2025 kjo vlerësohet të rritet në intervalin prej 150-200 zettabytes [13].

Padyshim që krijimi i të dhënave do të fiksohet vetëm në vitet e ardhshme, kështu që mund të pyesni veten: a ka ndonjë kufizim të ruajtjes së të dhënave? Jo në të vërtetë, ose më saktë, ka kufij, por janë aq larg sa nuk do të arrijmë askund afër tyre në çdo kohë të shpejtë. Për shembull, vetëm një gram ADN mund të ruajë 700 terabajt të dhëna, që do të thotë se ne mund të ruajmë të gjitha të dhënat tona që kemi tani në 1500 kg ADN - të paketuara dendur, do të futeshin në një dhomë të zakonshme. Megjithatë, kjo është shumë larg nga ajo që ne jemi në gjendje të prodhojmë aktualisht [13].

## IX. BIG DATA MINING

Nxjerrja e të dhënave përfshin eksplorimin dhe analizimin e sasive të mëdha të të dhënave për të gjetur modele për të dhëna të mëdha. Teknikat dolën nga fushat e statistikave dhe inteligjencës artificiale (AI), me pak menaxhim të bazës së të dhënave të hedhura në përzierje.

Në përgjithësi, qëllimi i nxjerrjes së të dhënave është ose klasifikimi ose parashikimi. Në klasifikim, ideja është që të renditen të dhënat në grupe. Për shembull, një tregtar mund të jetë i interesuar në karakteristikat e atyre që u përgjigjen përkundrejt atyre që nuk iu përgjigjen një promovimi [14].

Këto janë dy klasa. Në parashikim, ideja është të parashikojmë vlerën e një ndryshoreje të vazhdueshme. Për shembull, një tregtar mund të jetë i interesuar të parashikojë ata që do t'i përgjigjen një promovimi [14].

Algoritmet tipike të përdorura në kërkimin e të dhënave përfshijnë sa vijon:

**Asocimi:** Asocimi bën një korrelacion midis dy ose më shumë artikujve për të identifikuar një model. Për shembull, një supermarket mund të përcaktojë që klientët shpesh blejnë krem pana kur blejnë luleshtrydhe dhe anasjelltas. Shoqërimi përdoret shpesh në sistemet e pikave të shitjes për të përcaktuar tendencat e zakonshme midis produkteve [14].

**Klasifikimi:** Atribute të shumëfishta mund të përdoren për të identifikuar një klasë të veçantë të artikujve. Klasifikimi cakton artikujt në kategori ose klasa të synuara për të parashikuar me saktësi se çfarë do të ndodhë brenda klasës. Disa industri përdorin klasifikimin me klientët. Për shembull, një kompani bankare mund të përdorë një model klasifikimi për të identifikuar aplikuesit e kredisë si rreziqe kredie të ulëta, të mesme ose të larta. Organizata të tjera klasifikojnë audiencat aktuale dhe ato të synuara në grupe moshe dhe shoqërore për fushatat e marketingut [14].

**Grumbullimi:** Grumbullimi është metoda me të cilën rekordet (të dhënat) grupohen së bashku. Zakonisht kjo është bërë për t'i dhënë përdoruesit një pamje të nivelit të lartë të asaj që po ndodh në bazën e të dhënave [14].

**Pemët e Vendimeve (Decision trees):** Pemët e Vendimeve përdoren për kategorizimin ose parashikimin e të dhënave. Një pemë vendimi fillon me një pyetje të thjeshtë që ka dy ose më shumë përgjigje. Çdo përgjigje çon në një pyetje të mëtejshme që përdoret për të klasifikuar ose identifikuar të dhëna që mund të kategorizohen, ose kështu që një parashikim mund të bëhet bazuar në secilën përgjigje. Grafiku i një peme vendimi përfaqëson se si një ofrues i telefonit celular mund të klasifikojë klientët që kombinojnë, ose ata që nuk rinovojnë kontratat e tyre telefonike [14].

**Modele sekuenciale:** Modelet sekuenciale identifikojnë trendet ose dukuritë e rregullta të ngjarjeve të ngjashme. Kjo teknikë e nxjerrjes së të dhënave përdoret shpesh për të kuptuar sjelljet e blerjes së përdoruesve. Shumë shitës me pakicë përdorin të dhëna dhe modele vijuese për të vendosur për produktet që ata shfaqin. Me të dhënat e klientëve ju mund të identifikoni që klientët të blejnë një koleksion të veçantë të produkteve së bashku në periudha të ndryshme të vitit, sipas IBM. Në një aplikacion për shporta, ju mund ta përdorni këtë informacion për të sugjeruar automatikisht që artikuj të caktuar të shtohen në një shportë bazuar në frekuencën e tyre dhe historikun e kaluar të blerjeve [14].

**Këtu është një shembull i pemës së klasifikimit.** Merrni parasysh situatën kur një kompani telefonike dëshiron të përcaktojë se cilët klientë rezidencialë ka të ngjarë të shpërbimin shërbimin e tyre [14].

Kompania telefonike ka informacion që përbëhet nga atributet e mëposhtme: sa kohë personi ka pasur shërbimin, sa shpenzon në shërbim, nëse shërbimi ka qenë problematik, nëse ai ka planin më të mirë të thirrjes që i nevojitet, ku jeton, si ai është i vjetër, nëse ai ka shërbime të tjera të bashkuara së bashku, informacione konkurruese në lidhje me planet e transportuesve të tjerë dhe nëse ai ende e ka shërbimin [14].

Sigurisht, ju mund të gjeni shumë më shumë attribute se kjo. Atributi i fundit është ndryshorja e rezultatit; kjo është ajo që softueri do të përdorë për të klasifikuar klientët në një nga dy grupet - ndoshta të quajtura mbajtës dhe rreziqe të fluturimit [14].

Grupi i të dhënave është i ndarë në të dhëna trajnimi dhe një grup të dhënash testesh. Të dhënat e trajnimit përbëhen nga vëzhgime (të quajtura attribute) dhe një ndryshore e rezultatit (binare në rastin e një modeli klasifikimi) - në këtë rast, qëndruesit ose rreziqet e fluturimit [14].

Algoritmi drejtohet nga të dhënat e trajnimit dhe vjen me një pemë që mund të lexohet si një seri rregullash. Për shembull, nëse klientët kanë qenë me kompaninë për më shumë se dhjetë vjet dhe ata janë mbi 55 vjeç, ka të ngjarë të qëndrojnë si klientë besnikë [14].

Këto rregulla drejtohen më pas mbi grupin e të dhënave të provës për të përcaktuar se sa i mirë është ky model në "të dhëna të reja". Masat e saktësisë janë dhënë për modelin. Për shembull, një teknikë e njohur është matrica e konfuzionit. Kjo matricë është një tabelë që ofron informacion në lidhje me atë se sa raste ishin saktë kundrejt klasifikuar gabimisht [14].

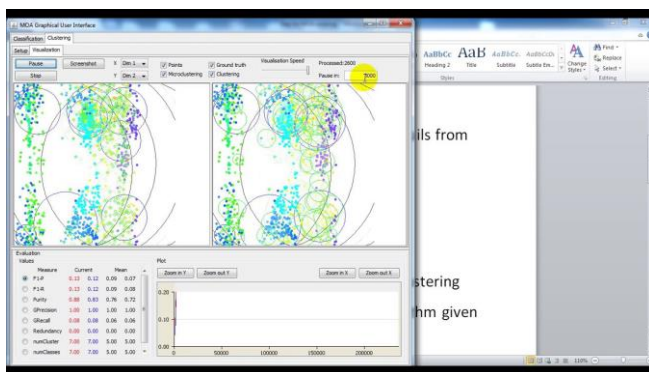
Nëse modeli duket mirë, ai mund të vendoset në të dhëna të tjera, pasi është në dispozicion (domethënë, duke e përdorur atë për të parashikuar raste të reja të rrezikut të fluturimit). Bazuar në modelin, kompania mund të vendosë, për shembull, të dërgojë oferta speciale për ata klientë të cilët mendon se janë rreziqe fluturimi [14].

## X. DISA PROGRAME QË PËRDOREN PËR KËRKIMIN NË BIG DATA

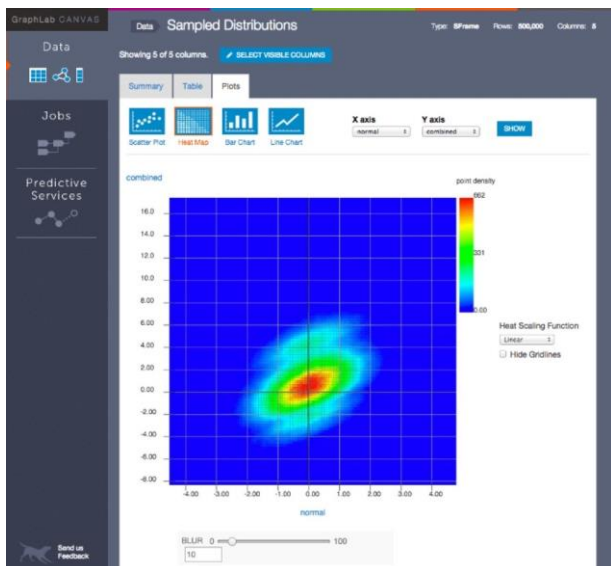
Në kërkimet dhe analizat e të dhënave të mëdha, disa mjete dhe burimeve të hapura janë si më poshtë:

Apache Mahout: Softuer i shkallëzuar për të mësuar makinerinë dhe kërkimet e të dhënave bazuar kryesisht në Hadoop. Ka zbatime të grumbullimit, klasifikimit, filtrimit bashkëpunues dhe kërkimeve të shpeshta të modelit [15][16].

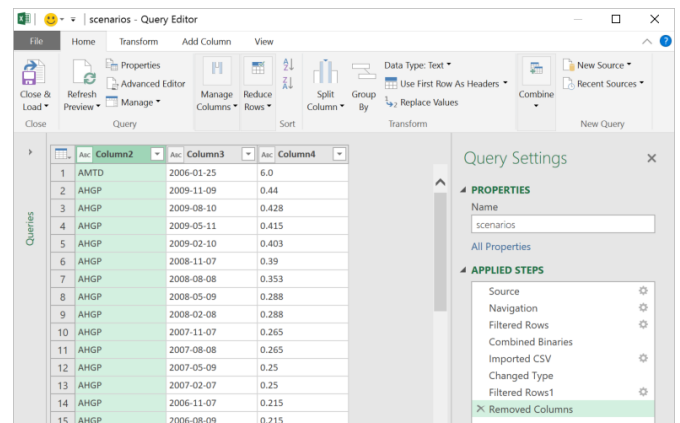
MOA: Transmetoni softuerin e kërkimit të të dhënave për të kryer minimin e të dhënave në kohë reale. Ka zbatime të grumbullimit, klasifikimit, regresionit, nxjerrjeve të shpeshta të artikujve dhe nxjerrjeve të shpeshta të grafikëve. R: gjuhë programimi me burim të hapur dhe mjedis softuer i krijuar për llogaritjen statistikore, kërkimet / analizat e të dhënave dhe vizualizimin [15][16].



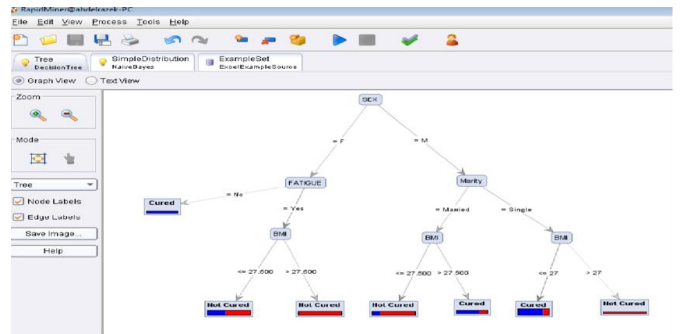
GraphLab: një sistem paralel grafik i nivelit të lartë i ndërtuar pa përdorur MapReduce [15][16].



Excel: Ofron aftësi të fuqishme të përpunimit të të dhënave dhe analizave statistikore [15][16].

A screenshot of the Microsoft Excel Query Editor. The main window displays a table with 15 rows and 4 columns. The columns are labeled 'Column2', 'Column3', 'Column4', and 'Column5'. The rows contain data points, including dates and numerical values. On the right side, there is a 'Query Settings' pane with a 'PROPERTIES' section and an 'APPLIED STEPS' section. The 'APPLIED STEPS' section lists various operations such as 'Source', 'Navigation', 'Filtered Rows', 'Combined Binaries', 'Imported CSV', 'Changed Type', 'Filtered Rows1', and 'Removed Columns'.

Rapid-I Rapidminer: Rapidminer është softuer me burim të hapur që përdoret për kërkim të të dhënave, mësim makine (Machine Learning) dhe analiza parashikuese [15][16].



Duke treguar Algoritmin e Pemëve të Vendimeve (Decision trees), krijuar nga (Rapid I, Rapidminer Ver.4.6, Berlin, Gjermani)

## XI. SFIDAT E DATA MINING DHE BIG DATA MINING

Big data mining është më sfiduese krahasuar me algoritmet tradicionale të data mining. Të dhënat e mëdha shpesh ruhen në vende të ndryshme. Ndërsa algoritmet tipike të data mining kërkojnë që të gjitha të dhënat të ngarkohen në memorien kryesore, lëvizja e të dhënave nëpër vende të ndryshme është e kushtueshme. Sfidat e mëdha të nxjerrjes së të dhënave dhe vështirësitë në hartimin e algoritmeve ngrihen nga vëllimet e mëdha të të dhënave, nga karakteristikat komplekse dhe dinamike të të dhënave. Sfidat e algoritmeve të mëdha të nxjerrjes së të dhënave renditen si më poshtë [17]:

- Një sistem i madh i kërkimit të të dhënave duhet të mundësojë një mekanizëm shkëmbimi informacioni dhe shkrirjeje për të siguruar që të gjitha faqet e shpërndara (ose burimet e informacionit) të mund të punojnë së bashku për të arritur një qëllim global të optimizimit [17].



• **Kërkimi nga të dhëna të rralla, të pasigurta dhe jo të plota:** Të dhënat e rralla nuk mund të përdoren për të nxjerrë përfundime të besueshme. Qasjet e zakonshme janë të përdorin zvogëlimin e dimensionit ose zgjedhjen e tipareve për të zvogëluar dimensionet e të dhënave ose për të përfshirë me kujdes mostra shtesë, të tilla si metodat gjenerike të mësimi (generic unsupervised learning methods) në data mining. Për të dhëna të pasigurta, secili artikull i të dhënave paraqitet si disa shpërndarje të mostrave [17].

Zgjidhjet e zakonshme janë marrja në konsideratë e shpërndarjeve të të dhënave për të vlerësuar parametrat e modelit. Shumica e algoritmeve të data mining mund të trajtojnë të dhëna jo të plota ose që mungojnë. Futja e vlerave që mungojnë është një metodë për të prodhuar modele të përmirësuara [17].

• **Kërkimi i të dhënave komplekse dhe dinamike:** Aktualisht, nuk ka asnjë model të njohur efektiv dhe efikas të të dhënave për të trajtuar kompleksitetin e të dhënave të mëdha (të strukturuar, të pa strukturuar dhe gjysmë të strukturuar) [17].

• **Trajtimi i një sasive të madhe të dhënash:** Disponueshmëria e madhe e të dhënave e bën të vështirë marrjen e vendimeve. Të dhënat që mund të kenë ndërmarrjet janë rritur në mënyrë eksponenciale nga disa vitet e fundit [18].

Ata kanë të dhëna për gjithçka, që nga ajo që i pëlqen një konsumatori, tek mënyra se si reagojnë, tek një aromë e veçantë, tek restoranti mahnitës që u hap në Itali fundjavën e kaluar.

Këto të dhëna tejkalojnë sasinë e të dhënave që mund të ruhen dhe llogariten, si dhe të merren. Sfida nuk është aq e disponueshme, por menaxhimi i këtyre të dhënave. Së bashku me rritjen në të dhëna të pastruara, disponueshmëria e të dhënave është në formate të shumta si video, audio, media sociale, të dhëna të pajisjeve inteligjente etj. Disa nga mënyrat më të reja të zhvilluara për të menaxhuar këto të dhëna janë një hibrid i bazave të të dhënave relacionale të kombinuara me bazat e të dhënave NoSQL. Një shembull i kësaj është MongoDB [18].

#### • Siguria e të dhënave:

Në rritjen e të dhënave, çështja kryesore është sigurimi i të dhënave. Shumë organizata pretendojnë se përballen me probleme me Sigurinë e të Dhënave. Kjo ndodh të jetë një sfidë më e madhe për ta sesa shumë probleme të tjera që lidhen me të dhënat. Të dhënat që vijnë në ndërmarrje vihen në dispozicion nga një gamë e gjerë burimesh, disa prej të cilave nuk mund të besohet të jenë të sigurt dhe të pajtueshme brenda standardeve organizative. Ata kanë nevojë të përdorin një shumëllojshmëri të strategjive të mbledhjes së të dhënave për të ndjekur nevojat e të dhënave. Këto të dhëna bëhen të disponueshme nga burime të shumta, dhe për këtë arsye kanë probleme të mundshme të sigurisë [18].

Ju kurrë nuk mund ta dini se cili kanal i të dhënave është i kompromentuar, duke kompromentuar kështu sigurinë e të dhënave të disponueshme në organizatë dhe duke u dhënë

hakerëve një shans për të hyrë brenda. Tani është thelbësore të prezantoni praktikën më të mira të Sigurisë së të Dhënave për mbledhjen, ruajtjen dhe rikthimin e sigurt të të dhënave [18].

#### • Mungesa e burimeve të afta

Ka një mungesë të profesionistëve të aftë të të dhënave të mëdha në dispozicion në këtë kohë. Kjo është përmendur nga shumë ndërmarrje që kërkojnë të përdorin më mirë Big Data dhe të ndërtojnë sisteme më efektive të Analizës së të Dhënave.

Ekzistojnë mungesa e njerëzve me përvojë dhe të Shkencëtarëve të Çertifikuar të të Dhënave ose Analistëve të të Dhënave në dispozicion për momentin, gjë që e bën grumbullimin e numrit të vështirë dhe ndërtimin e depërtimit të ngadaltë. Përsëri, trajnimi i njerëzve në nivelin e hyrjes mund të jetë i kushtueshëm për një kompani që merret me teknologji të reja. Në vend të kësaj shumë po punojnë për zgjidhje automatizimi që përfshijnë Mësimin e Makinës (Machine Learning) dhe Inteligjencën Artificiale për të krijuar njohuri, por kjo gjithashtu kërkon staf të trajnuar mirë ose kontraktimin e jashtëm të zhvilluesve të aftë [18].

## XII. 14 FUSHA KU NXJERRJAE TË DHËNAVE (DATA MINING) PËRDORET GJERËSISHT

### • Shëndetësia e ardhshme

Nxjerrja e të dhënave mban potencial të madh për të përmirësuar sistemet shëndetësore. Përdor të dhëna dhe analiza për të identifikuar praktikën më të mira që përmirësojnë kujdesin dhe ulin kostot. Studiuesit përdorin qasje të data mining si baza të të dhënave shumë-dimensionale, të mësuarit në makinë, informatikë e butë, vizualizimi i të dhënave dhe statistikën. Kërkimet mund të përdoren për të parashikuar vëllimin e pacientëve në çdo kategori. Janë zhvilluar procese që sigurojnë që pacientët të marrin kujdesin e duhur në vendin e duhur dhe në kohën e duhur. Nxjerrja e të dhënave mund të ndihmojë gjithashtu siguruesit e kujdesit shëndetësor të zbulojnë mashtrimet dhe abuzimet [19].



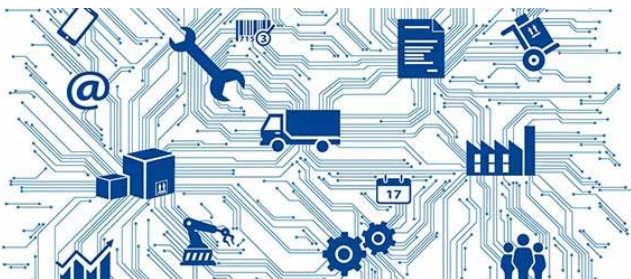
- **Analiza e Shportës së Tregut**

Analiza e shportës së tregut është një teknikë modelimi e bazuar në një teori që nëse blini një grup të caktuar artikujsh ka më shumë të ngjarë të blini një grup tjetër artikujsh. Kjo teknikë mund t'i lejojë shitësit me pakicë të kuptojë sjelljen e blerjes së një blerësi. Ky informacion mund të ndihmojë shitësin me pakicë të njohë nevojat e blerësit dhe të ndryshojë paraqitjen e dyqanit në përputhje me rrethanat. Përdorimi i analizës diferenciale mund të bëhet krahasimi i rezultateve midis dyqaneve të ndryshme, midis klientëve në grupe të ndryshme demografike [19].



- **Inxhinieri Prodhuese**

Dituria është pasuria më e mirë që një ndërmarrje prodhuese do të zotërojë. Mjetet e kërkimit të të dhënave mund të jenë shumë të dobishme për të zbuluar modele në procesin kompleks të prodhimit. Nxjerrja e të dhënave mund të përdoret në modelimin e nivelit të sistemit për të nxjerrë marrëdhëniet midis arkitekturës së produktit, portofolit të produktit dhe të dhënave për nevojat e klientit. Mund të përdoret gjithashtu për të parashikuar zhvillimin e produktit, kohën, koston dhe varësinë midis detyrave të tjera.



- **Arsim**

Ekziston një fushë e re në zhvillim, e quajtur Mbledhja e të Dhënave Arsimore, shqetësimet për zhvillimin e metodave që zbulojnë njohuri nga të dhënat që vijnë nga Mjedise Arsimore. Qëllimet e Mbledhjes së të dhënave identifikohen si parashikimi i sjelljes së të nxënimit në të ardhmen, duke studiuar efektet e mbështetjes arsimore dhe avancimin e njohurive shkencore në lidhje me të mësuarit. Nxjerrja e të dhënave mund të përdoret nga një institucion për të marrë vendime të sakta dhe gjithashtu për të parashikuar rezultatet e studentit. Me rezultatet institucioni mund të përqendrohet në atë që të mësojë dhe si të mësojë [19].

Modeli i të nxënimit të studentëve mund të kapet dhe të përdoret për të zhvilluar teknika për t'i mësuar ata.

- **Menaxhimi i Marrëdhënieve me Konsumatorët**

Menaxhimi i Marrëdhënieve me Konsumatorët ka të bëjë me blerjen dhe mbajtjen e klientëve, gjithashtu përmirësimin e besnikërisë së klientëve dhe zbatimin e strategjive të fokusuara tek klientët. Për të mbajtur një marrëdhënie të duhur me një klient, një biznes duhet të mbledhë të dhëna dhe të analizojë informacionin. Kjo është ajo ku nxjerrja e të dhënave luan rolin e saj. Me teknologjitë e data mining të dhënat e mbledhura mund të përdoren për analiza. Në vend që të hutohen se ku të përqendrohen për të mbajtur klientin, kërkuesit e zgjidhjes marrin rezultate të filtruara [19].

- **Zbulimi i Mashtrimit**

Miliarda dollarë janë humbur nga veprimi i mashtrimeve. Metodat tradicionale të zbulimit të mashtrimit janë shumë kohë dhe komplekse. Nxjerrja e të dhënave ndihmon në sigurimin e modeleve kuptimplota dhe kthimin e të dhënave në informacion. Çdo informacion që është i vlefshëm dhe i dobishëm është njohuri. Një sistem i përsosur i zbulimit të mashtrimit duhet të mbrojë informacionin e të gjithë përdoruesve. Një metodë e mbikëqyrur përfshin mbledhjen e të dhënave të mostrave. Këto regjistra klasifikohen mashtrues ose jo-mashtrues. Një model është ndërtuar duke përdorur këto të dhëna dhe algoritmi është bërë për të identifikuar nëse rekordi është mashtrues apo jo [19].



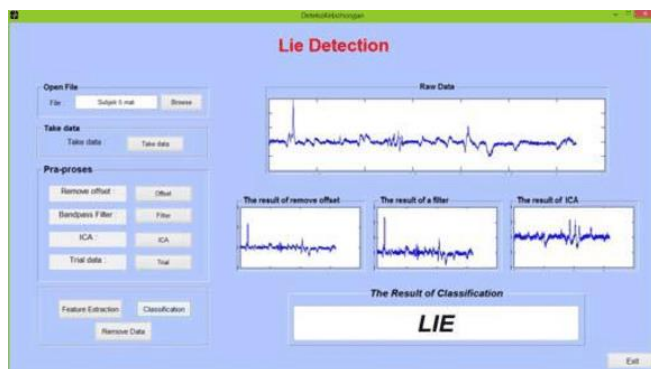


- **Zbulimi i ndërhyrjeve**

Çdo veprim që do të kompromentojë integritetin dhe konfidencialitetin e një burimi është një ndërhyrje. Masat mbrojtëse për të shmangur ndërhyrjen përfshijnë vërtetimin e përdoruesit, shmangien e gabimeve të programimit dhe mbrojtjen e informacionit. Nxjerrja e të dhënave mund të ndihmojë në përmirësimin e zbulimit të ndërhyrjeve duke shtuar një nivel të përqendrimit në zbulimin e anomalisë. Ndihejmon një analist të dallojë një aktivitet nga aktiviteti i zakonshëm i rrjetit të përditshëm. Nxjerrja e të dhënave gjithashtu ndihmon në nxjerrjen e të dhënave që janë më të rëndësishme për problemin [19].

- **Zbulimi i gënjeshtreve**

Të kapësh një kriminel është e lehtë ndërsa nxjerrja e së vërtetës prej tij është e vështirë. Zbatimi i ligjit mund të përdorë teknikat e kërkimeve për të hetuar krimet, për të monitoruar komunikimin e terroristëve të dyshuar. Kjo e regjistruar përfshin gjithashtu nxjerrjen e tekstit. Ky proces kërkon të gjejë modele kuptimplota në të dhëna që zakonisht janë tekst i pastrukturuar. Kampioni i të dhënave i mbledhur nga hetimet e mëparshme krahasohet dhe krijohet një model për zbulimin e gënjeshtres. Me këtë model mund të krijohen procese sipas nevojës [19].



- **Segmentimi i klientit**

Studimi tradicional i tregut mund të na ndihmojë në segmentimin e klientëve, por nxjerrja e të dhënave shkon thellë dhe rrit efektivitetin e tregut. Kërkimet e të dhënave ndihmojnë në radhitjen e klientëve në një segment të veçantë dhe mund të përshtatin nevojat sipas klientëve. Tregu ka të bëjë gjithmonë me mbajtjen e klientëve. Nxjerrja e të dhënave ju lejon të gjeni një segment të klientëve bazuar në cënueshmërinë dhe biznesi mund t'i ofrojë ata me oferta speciale dhe të rrisë kënaqësinë [19].

- **Bankë Financiare**

Me banka të kompjuterizuara kudo, një sasi e madhe e të dhënave supozohet të gjenerohet me transaksione të reja. Nxjerrja e të dhënave mund të kontribuojë në zgjidhjen e problemeve të biznesit në bankë dhe financa duke gjetur modele, shkaqe, dhe korrelacione në informacionin e biznesit dhe çmimet e tregut që nuk janë menjëherë të dukshme për menaxherët, sepse vëllimi i të dhënave është shumë i madh ose gjenerohet shumë shpejt për t'u parë nga ekspertët. Menaxherët mund t'i gjejnë këto informacione për segmentimin, shënjestrimin, blerjen, mbajtjen dhe mirëmbajtjen më të mirë të një klienti fitimprurës [19].

- **Mbikëqyrja e korporatave**

Mbikëqyrja e korporatave është monitorimi i sjelljes së një personi ose grupi nga një korporatë. Të dhënat e mbledhura përdoren më shpesh për qëllime të marketingut ose shiten korporatave të tjera, por gjithashtu ndahen rregullisht me agjencitë qeveritare. Mund të përdoret nga biznesi për të përshtatur produktet e tyre të dëshirueshme nga klientët e tyre. Të dhënat mund të përdoren për qëllime të marketingut të drejtpërdrejtë, siç janë reklamata e synuara në Google dhe Yahoo, ku reklamata i drejtohen përdoruesit të motorit të kërkimit duke analizuar historikun e tyre të kërkimit dhe postat elektronike [19].

- **Analiza e hulumtimit**

Historia tregon se ne kemi qenë dëshmitarë të ndryshimeve revolucionare në kërkime. Nxjerrja e të dhënave është e dobishme në pastrimin e të dhënave, para-përpunimin e të dhënave dhe integrimin e bazave të të dhënave. Studiuesit mund të gjejnë ndonjë të dhënë të ngjashme nga baza e të dhënave që mund të sjellë ndonjë ndryshim në hulumtim. Mund të dihet identifikimi i sekuencave të përbashkëta dhe korrelacioni midis aktiviteteve. Vizualizimi i të dhënave dhe kërkimet vizuale të të dhënave na ofrojnë një pamje të qartë të të dhënave [19].

- **Hetimi penal**

Kriminologjia është një proces që synon të identifikojë karakteristikat e krimit. Në të vërtetë, analiza e krimit përfshin eksplorimin dhe zbulimin e krimeve dhe marrëdhëniet e tyre me kriminelët. Vëllimi i lartë i të dhënave të krimit dhe gjithashtu ndërlikueshmëria e marrëdhënieve midis këtyre llojeve të të dhënave e kanë bërë kriminologjinë një fushë të përshtatshme për zbatimin e teknikave të nxjerrjes së të dhënave. Raportet e krimeve të bazuara në tekst mund të shndërrohen në skedarë të përpunimit të tekstit. Këto informacione mund të përdoren për të kryer procesin e përputhjes së krimit [19].

- **Bioinformatika**

Qasjet e Data Mining duken idealisht të përshtatshme për Bioinformatikë, pasi ajo është e pasur me të dhëna. Minimi i të dhënave biologjike ndihmon në nxjerrjen e njohurive të dobishme nga grupe të dhënash masive të mbledhura në biologji dhe në fusha të tjera të lidhura me shkencat e jetës, siç janë mjekësia dhe neuroshkenca. Zbatimet e nxjerrjes së të dhënave në bioinformatikën përfshijnë gjetjen e gjeneve, konkluzionin e funksionit të proteinave, diagnostikimin e sëmundjes, prognozën e sëmundjes, optimizimin e trajtimit të sëmundjes, rindërtimin e rrjetit të ndërveprimit të proteinave dhe gjeneve, pastrimin e të dhënave dhe parashikimin e vendndodhjes së proteinave nën qelizore [19].

### XIII. KONKLUSIONE

Ne jemi duke jetuar në epokën e të dhënave të mëdha ku sasi të mëdha të të dhënave heterogjene, gjysmë të strukturuar dhe të pastrukturuar gjenerohen vazhdimisht në shkallë të paparë. Të dhënat e mëdha zbulojnë kufizimet e teknikave ekzistuese të minierave, duke rezultuar në një seri sfidash të reja në lidhje me minierat e mëdha të të dhënave. Miniera e të dhënave të mëdha është një fushë premtuese kërkimesh, ende në fillimet e saj. Përkundër punës së kufizuar të bërë në minierat e të dhënave të mëdha deri më tani, ne besojmë se kërkohet shumë punë për të kapërcyer sfidat e saj në lidhje me heterogjenitetin, shkallëzueshmërinë, shpejtësinë, saktësinë, besimin, origjinën, privatësinë dhe interaktivitetin.

Nxjerrja e të dhënave mund të përdoret për të zbuluar njohuri të fshehura, të panjohura, por të dobishme nga të dhëna masive, të paqarta, të zhurmshme, jo të plota dhe të rastit. Analiza e të dhënave të mëdha kërkon që kërkimet e shpërndara të rrjedhave të të dhënave duhet të kryhen në kohë reale. Nevojiten shumë kërkime në analizat praktike dhe teorike për të siguruar metoda të reja për kërkime të shpërndara të të dhënave me rrjedha të mëdha të të dhënave. Sfidat e algoritmeve të mëdha të minierave të të dhënave janë: të mësuarit lokal dhe bashkimi i modelit për burime të shumta informacioni; kërkimi nga të dhëna të rralla, të pasigurta dhe jo të plota; dhe kërkimet e të dhënave komplekse dhe dinamike.

### XIV. BIBLIOGRAPHY

- [1] "Techopedia," 24 December 2020. [Online]. Available: <https://www.techopedia.com/definition/30215/big-data-mining>.
- [2] G. W. Lidong Wang, "Data Mining Applications in Big Data," 3 September 2015.
- [3] "Sas," [Online]. Available: [https://www.sas.com/en\\_gb/insights/analytics/data-mining.html](https://www.sas.com/en_gb/insights/analytics/data-mining.html). [Accessed 24 December 2020].
- [4] M. C. Lovell, "Data Mining". The Review of Economics and Statistics, 2015.
- [5] W. W. Charemza and D. F. Deadman, "Data Mining". New Directions in Econometric Practice, 2014.
- [6] M. Kantardzic, Data Mining: Concepts, Models, Methods, and Algorithms., 2014.
- [7] "Oracle," [Online]. Available: [https://docs.oracle.com/cd/B28359\\_01/datamine.111/b28129/process.htm#CHDJCFAG](https://docs.oracle.com/cd/B28359_01/datamine.111/b28129/process.htm#CHDJCFAG). [Accessed 25 December 2020].
- [8] A. Twin, "Data Mining," 20 September 2020.
- [9] M. M. G. K.-U. S. D.-H. Tran, "Change detection in streaming data in the era of big data: models and issues," 2014.
- [10] T. Breur, "Statistical Power Analysis and the contemporary "crisis" in social sciences," July 2016.
- [11] J. M. S. M. Reichman OJ, "Challenges and opportunities of open data in ecology," February 2015.
- [12] 27 December 2020. [Online]. Available: [https://www.sas.com/en\\_us/insights/analytics/data-mining.html](https://www.sas.com/en_us/insights/analytics/data-mining.html).
- [13] L. Gavin, "Is Big Data Really That Big?," 13 October 2020.
- [14] J. S. Hurwitz, Big Data For Dummies, 2015.
- [15] A. Bifet, "Mining Big Data in Real Time," 2014.
- [16] S.-W. M. Y.-H. L. M. Chen, "Big data: A survey," Mobile Netw Appl," 2014.
- [17] X. Z. G.-Q. W. W. D. X. Wu, "Data Mining with Big Data," Knowledge and Data Engineering," 2014.
- [18] Y. Vaghela, "Four Common Big Data Challenges," 28 June 2018.
- [19] Rajkumar, "Top 14 useful applications for data mining," 20 August 2014.