

**Koostöökokkulepe nr 1.9- 8/21-445-1**

## **Ettevõtjate varase hoiatamise teenus**

### **Prototüüp ja ärianalüüs – praktiline teostus**

**Lõpparuanne  
30. aprill 2022.a.**

**Eksperimentaalstatistika tiim  
Statistikaamet**

**Tallinn 2022**

## Sisukord

Sissejuhatus.....	4
Peatükk I – teoreetilised lähtealused.....	5
1.1. Ärianalüütiline probleemi olemuse üldkirjeldus.....	6
1.2. Kirjandusallikate kasutamine.....	6
1.3. Kaks põhimõtet lähtenurka.....	8
1.4. Aruandlus 3.0 toetus käesolevale projektile.....	9
1.5. Tegevusala määramine.....	10
1.6. Maksejõuetuse terminoloogiline määratlus.....	12
1.7. Majandussuhtarvude kasutamine.....	13
1.8. Hoiatuste tasemed ja teadete kordamine.....	20
1.9. Andmete sobivuse kontroll.....	22
1.10. Sisendparameetrite kvaliteedi tõstmine.....	24
1.11. Täiendavate parameetrite leidmine.....	25
1.12. Parameetrite paljusus.....	25
1.13. Hinnang ettevõtete vaheliste arvelduste põhjal.....	27
1.14. Andmed.....	28
Peatükk II – praktiline teostus.....	34
2.1. Süsteemi ülesehituse põhimõtted.....	35
2.2. Bilansiandmete analüüs.....	36
2.2.1. Bilansiandmete klasterdamine.....	36
2.2.2. Suhtarvude klasterdamine.....	41
2.2.3. Bilansiandmete ja suhtarvude koosklasterdamine.....	43
2.3. Transaktsiooniandmete analüüs.....	46
2.4. Mudelite loomine ja andmete testimine.....	51
2.4.1. Üldised lähte-eeldused mudelite loomisel.....	51
2.4.2. Prognoosimudelite kasutamine.....	54
2.4.3. Prognoosimudeli reaaleluline kontroll.....	56
2.4.4. Euclidean distant andmete analüüsis.....	56
2.4.5. Loss funktsiooni kasutamine mudeli optimeerimisel.....	57
2.4.6. Eksperimentaalstatistiline tehisnärvivõrkude lähtenurk.....	61
Peatükk III – töö täiendamise käigus aset leidnud tegevused.....	66
3.1. Täiendavate majandussektorite mudelite koostamine.....	67
3.2. Reaaleluliste testide ettevalmistamine.....	68
3.3. Prototüübi järk-järguline parandamine.....	68
3.4. Tehnilised lahendused prototüübi kasutamiseks.....	68
3.5. Andmete kasutamine.....	69
3.6. Vigade kõrvaldamine.....	70
KOKKUVÕTE.....	72
Allikad.....	77
Lisad.....	78
Lisa 1.: Tegevusalade defineerimine EMTAK abil.....	78
Lisa 2.: Kasutatavad majandussuhtarvud.....	80
Lisa 3.: Pankrotistunud ettevõtteid klastrite ja aastate lõikes.....	83

Lisa 4.: Täiendavad klasterdamised.....	84
Lisa 5.: Klasterite taandamine kahe parameetri peale.....	85
Lisa 6.: PCA tehnikaga klasterdamine kahedimensionaalseks.....	86
Lisa 7.: Muutujate omavaheline korrelatsioon Transaktsiooniandmete analüüs.....	87
Lisa 8.: Klasteritesse jagunemine tegevusalade lõikes.....	88
Lisa 9.: TwoLayerNet kood.....	89
Lisa 10.: Layers kood.....	94
Lisa 11.: Optim kood.....	98

## Sissejuhatus

Käesolev materjal esitab esmased seisukohad ja põhimõtted ettevõtjate varase hoiatamise teenuse väljatöötamise kontekstis lähtudes peamistest ärianalüütilistest kaalutlustest. Välja on töötatud prototüüp ja kontseptuaalsed lahendused varase hoiatamise teenuse käivitamiseks Eestis. Käesolevas materjalis esitatakse prototüübi ja seonduva materjali seis 31. märtsi 2022. aasta seisuga. Töö teadaolevate probleemide kõrvaldamiseks ja tagasiside arvesse võtmiseks jätkub aprillis 2022.a. Lõplik-täiendatud raport valmib mais 2022.a.

Materjalis esitatakse nii teoreetilised lähtealused, nendel põhinevad analüütilised kaalutlused kui ka esmased tulemused nende analüüsikomponentide lõikes, mida praeguses töö faasis oli võimalik ja loogiline läbi viia (s.t. on kontsentreeritud prototüübi kesksetele momentidele, mitte infotehnoloogilise lõpplahenduse detailidele). Käesolevas materjalis on esitatud ka kommentaarid selle kohta, kui praktilised lahendused ei ole kattunud varasemalt üldiseks arusaamaks olnud teoreetiliste seisukohtadega, analüüsitud on lahknemise põhjuseid ja on tehtud vastavasisulised järeldused.

Käesolev materjal koosneb kahest peatükist:

- Peatükk I – teoreetilised lähtealused;
- Peatükk II – praktiline teostus.

Peatükk I esitab olulisemad teoreetilised lähtealused ja viited teadusartiklitele, millel käesolevas töös kasutatud põhimõtted on tuginenud. Peatükis on toodud ka esmased praktilised näited ja mõned esmased päringud, klasterdamised, järeldused, masinõppe tulemid. Teoreetilisse peatükki on need lisatud eesmärgiga illustreerida teoreetiliste teemade haakumist praktilise tööga ja näidata vahetumalt seoseid, kuidas teoreetiline materjal on kasutust leidnud.

Peatükk II toob välja praktilise arenduse, mis põhineb peatükis I toodud teoreetilistel seisukohtadel ja nende seisukohtade praktilisel tõlgendusel. Peatükk II tugineb eelmisel peatükil ja seal on lahti kirjutatud tehnilised momendid ja lahendused, milleni on käesoleva dokumendi esitamise hetkeks prototüübi loomisel jõutud. Näidatud on nii tööprotsessi kui ka mudelit, milleni on jõutud. Tuuakse ära tulemusi ja tõlgendatakse neid.

Osa prototüübiga seonduvast tehnilisest materjalist paikneb käesoleva dokumendi põhitekstis, osa on paigutatud dokumendi lisadesse. Põhitekstis on toodud eelkõige see materjal, mis tekstiliste selgituste mõistmiseks on vajalik hoida võrdlusena teksti juures, lisades on eelkõige kaudsemalt illustreeriv ning tehniliselt mahukam materjal.

Praktilise teostuse käigus on katsetatud erinevaid lähtenurki ja meetodeid, mõned neist on andnud paremaid, mõned kehvemaid tulemusi. Samas ongi prototüübi väljatöötamise mõte selles, et analüüsida ja saada tulemeid erinevatel viisidel, et siis praktilise tulemi põhjal otsustada, mis töötab ja mis mitte, mida on mõistlik infosüsteemi väljatöötamisel aluseks võtta ja mida mitte.

## **Peatükk I – teoreetilised lähtealused**

I peatükis esitatakse teoreetilised lähtealused ja viited konkreetsetele kirjandusallikatele, millest need teoreetilised lähtealused tulevad. Peatükis tuuakse välja ka praktilisi lähtesamme ja ka esmaseid tulemeid, kuid need on eelkõige mõeldud illustreerimaks teoreetiliste seisukohtade kasutamist praktilises töös. Sisulisemal määral on praktilist mudeli väljatöötamist kirjeldatud II peatükis.

## 1.1. Ärianalüütiline probleemi olemuse üldkirjeldus

Ärianalüütilise probleemipüstituse keskselt on vajalik hinnata, kas konkreetne analüüsitav ettevõtte võib potentsiaalselt sattuda maksejõuetuse olukorda. Nimetatud hinnangu andmine peab olema tagatud automaatselt arvutialgoritmide abil. Hinnang peab võtma arvesse ettevõtte tegevusvaldkonda (erinevate valdkondade ettevõtted võivad sattuda maksejõuetuse olukorda väga erinevate bilansside ja kasumiaruannete seisudega), ettevõtte suurust (erineva suurusega ettevõtted on rahaliselt haavatavad erinevatel viisidel), geograafilist paiknemist (asukoht võib avaldada väga suurt mõju müügile ja ka sisendite hankimise võimele) ja teatud määral ka õiguslikku konteksti (erinevad ettevõtted on reguleeritud erinevalt – näiteks tööjõuintensiivsed ettevõtted on mõjutatud suurel määral tööturu seadusandlusest). Ettevõtetele võimaliku makseraskuse saabumise hinnangu andmisel tuleb lähtuda kõigi ülaltoodud elementide/olemusmääratluste võimalikest kombinatsioonidest.

Käesoleva prototüübi väljatöötamise käigus kasutatakse masinõppe algoritme ja tehisintellektse töö põhimõtteid. Tulenevalt sellest on tehtud tööd teoreetiliste lähtekohtadega, analüüsitud teaduskirjandust ja tutvutud senimaani koostatud lahendustega erinevates teadusinstituutides. Teostatud on analüüs, mida võtta senimaani tehtust üle ning millises suunas prototüübi tööd eelkõige suunata (optimaalse sisendi juures maksimaalselt adekvaatse väljundi saamiseks).

## 1.2. Kirjandusallikate kasutamine

Teoreetiliste lähtenurkade täpsustamiseks ja lähteraamistiku fikseerimiseks on läbi töötatud teatud hulk teadusartikleid. Artiklid on valitud selliselt, kus kirjeldatakse masinõppe ja tehisintellekti lahenduste kasutamist ettevõtete majandustulemuste prognoosimiseks ja võimaliku maksejõuetuse olukorra leidmiseks.

Alljärgnevalt on toodud välja lühikokkuvõtted olulisematest sisenditest, mida kasutatud teadusartiklid on käesoleva töö jaoks andnud ja millele toetatakse käesoleva prototüübi rajamisel lähte-eelduste fikseerimisel ja meetodite valikul.

Töö käigus on läbi vaadatud ja analüüsitud suuremas mahus kirjandust, kui siinkohal viidatud, viidatud on vaid see osa, mida on otsustatud ühel või teisel viisil käesolevas töös kasutada. Ülaltoodu aga tähendab, et teoreetilise kirjanduse loetellu võib lisanduda töö käigus uusi allikaid vastavalt sellele, kuidas töö edeneb ning milliseid momente peetakse vajalikuks edaspidiselt lahendada.

Mitmed alltoodud artiklid on välja töötatud teatud spetsiifilist konteksti arvestades, mis ei ole otseselt kattuv käesoleva töö eesmärgiga. Sellisel juhul on artiklitest võetud see osa sisendist, mis kattub käesoleva töö eesmärgiga ja on otseselt käesoleva töö jaoks rakendatav.

Kou et al. (2021) pakuvad välja pankroti ennustusmudelit väikese ja keskmise suurusega ettevõtetele, mis põhineb nende ettevõtete vaheliste arvelduste andmetel. Tegemist on alternatiivse lahendusega nendele mudelitele, mis põhinevad ettevõtte raamatupisamislike andmete (bilansi, kasumiaruande) majandussuhtarvude analüüsil ja jätab nn. klassikalise suhtarvude analüüsi üldse kõrvale. Metoodika autorid väidavad, et erinevalt raamatupidamislikest aruannetest on ettevõtete vaheline tehinguinfo usaldusväärsem. Lisaks on tehinguinfo näol tegemist dünaamilise infoga, kui raamatupidamise aruanded väljendavad mingit hetkeseisu ja on oma olemuselt staatilised. Tehinguinfo näitab hetkeseisu ja selle muutumist, kui raamatupidamise aruanded näitavad, milline oli ettevõtte seis kunagi minevikus. Tehinguinfo võimaldab välja arvutada ka nüüdispuhasväärtuse – NPV (net present

value), mis on oluliseks ettevõtte „tervise“ indikaatoriks, kuid mida raamatupidamislikud aruanded arvutada ei võimalda.

Perboli ja Arabnezhad (2021) selgitavad, et raamatupidamislike aruannete suhtarvude analüüsi tulemusena on võimalik prognoose teha eelkõige lühiajaliselt ja eelkõige suuremate ettevõtete puhul. Nad pakuvad välja masinõppe meetodi, mille abil on võimalik väikese ja keskmise suurusega ettevõtete pankrotiprognose teha pikemas perspektiivis (kuni 60 kuud) ja oluliselt suurendada prognooside täpsust lühiajalise ajaperioodi lõikes (kuni 12 kuud). Autorid väidavad, et nende poolt välja pakutav meetodika on hea lähtealus majanduspoliitika kujundamiseks, ehk siis annab võimaluse laiemalt tõlgendada suhtarvude analüüsi tulemit.

Perboli ja Arabnezhad (2021) rõhutavad, et masinõppest adekvaatse tulemi saamiseks on oluline eelnevalt valida õiged tunnused, mida analüüsida ning pakuvad selleks välja oma lähtekohad. Põhjalikult käsitlevad nad seda, kuidas rakendada masinõpet prognoosimisel ning millisel viisil tagada sisendid tulemi analüütiliseks toetamiseks. Tuuakse sisse andmete kahekordse treenimise põhimõte. Autorid tagavad ka raamistiku, kuidas tulemusi tõlgendada.

Qu et al. (2019) toovad välja erinevaid masinõppel ja närvivõrkude analüüsil põhinevaid mudeleid, mida nad näevad asjakohased ja perspektiivikad olevat maksejõuetuse olukordade prognoosimisel lähtudes algsena saada olevast majandusolukorda kirjeldavast raamatupidamislikust informatsioonist. Masinõppe meetoditena käsitlevad nad järgnevat: Multivariant Discrimination Analysis (MDA), Logistic Regression (LR), Ensemble method ja Support Vector Machines (SVM). Närvivõrkude analüüsil põhinevatest meetoditest käsitlevad nad järgnevat: Neural Networks (NN), Deep Belief Network (DBN) ja Convolutional Neural Network (CNN).

Shi ja Li (2019) toovad laiapõhjalises kirjandusel põhineval uurimustöös välja millist liiki maksejõuetuse probleeme masinõppe meetoditega (või masinõppe lähedaste meetoditega) lahendatakse ning milliseid meetodeid nende probleemide lahendamiseks kasutatakse. Inglise keeles toodud maksejõuetuse probleemterminid, mida nad käsitlevad, on järgmised: bankruptcy prediction, default prediction, financial failure, financial distress, insolvency, business failure. Lahendusmeetoditena käsitletavat metoodikad/tegevused on järgmised: neural network, support vector machine, decision tree, genetic algorithm, fuzzy, rough set, data mining, case-based reasoning, data envelopment analysis, adaboost, K-nearest neighbors, bayesian network.

Cialone (2020) esitab kaasuse analüüsi, kus on analüüsitud 14 965 itaalia ettevõtet bilansi suhtarvude pinnalt, 13 845 neist ettevõtetest on tegutsevad ja 1 120 pankrotis ning on loodud mudel hindamaks maksejõuetuse tekkeks ühe aastase perioodi jooksul. Mudelis on kasutatud 88 sisendmuutujat. Kasutatud on närvivõrkudel põhinevat analüüsi Deep Fully Connected and Convolutional Neural Networks.

Cialone (2020) nimetab kõige klassikalisemaks (varemalt rohkem kasutatud) ettevõtete võimaliku maksejõuetuse prognoosimise mudelitest Multiple Discriminant Analysis ja Logistic Regression, samas kui viimasel ajal on rohkem hakatud sisse tooma Random Forests, Boosting ja Neural Networks. Vähemstatistilise meetodina, kuid mida üha rohkem kasutatakse, on levinud ka Contingent Claim analysis. Deep Neural Networks omab üha suuremat tähtsust.

Chen (2020) selgitab, et masinõppe meetodid on üha populaarsemad majandusandmete analüüsis, kuid peamiseks pudelikaelaks on tulemi interpreteerimise selgus. Lahendusena pakub ta CART või (boosted) ensemble of decision trees meetodeid. Kui nendest lähtealustest mitte lähtuda, siis võib tulemusi tõlgendamine olla võrrelda nn. „musta kasti“ vaatega.

Ucoglu (2020) annab ülevaate, kuidas masinõppe meetodeid kasutatakse majandusarvestuses ja auditeerimises. Ta annab ülevaate nelja suure audiitorbüroo näitel, kuidas need bürood masinõpet

oma töös kasutavad, analüüsivad bürood on: Ernst and Young, PwC, Deloitte ja KPMG. Ucoglu prognoosib, et aastal 2025 viiakse 30% audititest läbi masinõppe ja tehisintellektseid meetodeid kasutades.

Cao ja You (2021) toovad oma käsitluses välja, et masinõppe mudelid, eelkõige need, mis ei ole lineaarse lähenemisega, tagavad märksa informatiivsema ja täpsema rahavoogude prognoosivõime, kui tavapäraselt kasutatavad rahavoogude prognoosimise meetodid. Oluline on ka see, et masinõppe mudelid võimaldavad leida majanduslikult mõistlikke seoseid, mida tavapäraseid prognoosimudeleid kasutades ei pruugi märgata. Masinõpe võib tuua välja ka täiesti uusi olulisi momente, mida üldse varemalt ei ole mudelisse sisse pandud.

Choudhry (2018) näitab ära, kuidas masinõpet saab kasutada panganduses ühelt poolt bilansside analüüsil ja teiselt poolt ka klienditeeninduse poole haldamisel, ehk siis toob sisse nii analüütilise kui ka inimliku dimensiooni. Käsitluses on oluline ka see, et autor toob välja, milliste meetoditega ja millistel juhtudel rakendada juhendamata ja millisel juhul juhendatud õpet (unsupervised versus supervised learning).

Aliaj et al. (2020) selgitavad, et sarnaste majandusandmete põhjal võib mõni ettevõtte jõuda maksejõuetuse seisuni, kuid samas mõni teine mitte. Sageli langetab otsuse sisuliselt laenu andev pank, kes võib teatud hetkedel laenu tagasi kutsuda või laenugraafiku muutmisele mitte vastu tulla. Autorid vaatavad itaalia laenuandmeid ettevõtetele koosmõjus nende ettevõtete maksevõimetuse tekke momentidega.

Andrés et al. (2004) toovad enda poolt pakutavas finantstulemuste prognoosimise mudelis välja need aspektid ja rõhuasetused, mis on eelkõige mõistlikud kaasata prognooside tegemisse, millest rohkem lähtuda ja mis annavad parema kvaliteediga prognoosi. Masinõpet abstraktselt rakendades ei pruugi tulemid olla väga reaalsuskesksed.

Amel-Zadeh et al. (2020) kasutavad masinõppe random forest mudelit leidmaks, millised majandusarvestuse muutujad on kõige tähtsamad ennustamiseks kõige adekvaatsemalt ettevõtte vaba rahavoogu ja teisi näitajaid, mis viitavad üldiste majandustegevuse näitajate paranemisele.

León et al. (2016) lähtuvad oma käsitluses eeldusest, et ettevõtte (panga) bilanss on unikaalne iga konkreetse ettevõtte osas ja ühtlasi ettevõtet kirjeldav. Nende uurimustöö neural networks masinõppe meetodi abil kinnitab, et nii see on. Kui masin suudab ära tunda ettevõtte bilansse ja neid ettevõtteid nende bilansside alusel kirjeldada, siis võib masin bilansside põhjal aru saada ka nende ettevõtte majandustervisest.

Petropoulos et al. (2019) pakuvad välja dünaamilise bilansside simulatsiooni mootori, mis deep learning meetodil prognoosib bilansi võtmenäitajaid. Autorid võrdlevad saadud tulemusi teiste bilansi prognoosimise meetoditega – staatilise bilansiprognosi meetodiga ja dünaamilise bilansiprognosi meetodiga. Masinõppe abil langeb oluliselt prognoosiviga, eriti üheaastase ette prognoosimise perioodi jooksul.

### **1.3. Kaks põhimõttelist lähtenurka**

Töös on kaks põhimõttelist lähtenurka võimaliku maksejõuetuse prognoosimisel:

- (a) nn. klassikaline majandussuhtarvudel põhinev loogika;
- (b) majandustransaktsioonide analüüsil põhinev loogika.

Majandussuhtarvudel põhinev loogika. Kasutatakse bilansi ja kasumiaruande suhtarve ning nende pinnalt prognoositakse masinõppe meetodeid kasutades võimalikke tulevikutsenaariumeid. Teades,



milliste suhtarvude millised väärtused viitavad maksevõime vähenemisele on võimalik prognoosida võimaliku maksevõimetuse saabumise olukorda. Valdav osa käesolevas materjalis viidatud teaduskirjandust kirjeldab majandussuhtarvudel põhinevaid prognoose.

Majandustransaktsioonide analüüsil põhinev loogika. Kou et al. (2021) pakuvad välja maksejõuetuse ennustumudelit väikese ja keskmise suurusega ettevõtetele, mis põhineb nende ettevõtete vahelistel arvelduste andmetel. Tegemist on alternatiivse lahendusega nendele mudelitele, mis põhinevad ettevõtte raamatupisamislike andmete (bilansi, kasumiaruande) majandussuhtarvude analüüsil ja jätab nn. klassikalise suhtarvude analüüsi üldse kõrvale. Käesoleva töö juures vaadeldakse masinõppe meetoditega ettevõtete vahelisi arveldusi alternatiivina klassikalisele suhtarvudel põhinevale loogikale.

Nii majandussuhtarvudel põhinev kui ka majandustransaktsioonidel põhinev loogika hakkavad teineteist täiendama, mõlemad analüüsid viiakse läbi ja mõlema osas käivitatakse masinõppe meetodid. Kahte analüüsi võrreldakse ja neile leitakse ühisosa. Kui üks analüüs annab väga oluliselt erineva tulemi teisest, siis on põhjust täiendavaks analüüsiks, et selgitada, miks on saadud selline tulemus, kus võib paikneda oluline viga.

#### **1.4. Aruandlus 3.0 toetus käesolevale projektile**

Varase hoiatamise teenuse täpsuse tõstmisele aitab kaasa, kui sisendina on võimalik kasutada Aruandlus 3.0 projekti raames sissetulevat informatsiooni. Töö Statistikaameti poolt selle informatsiooni hanke parandamiseks käib pidevalt ja süsteemselt ning laiemalt koostöös teiste riigiasutustega (nt KeM, JuM, RIK, EMTA, MKM, Eesti Pank) nn. andmepõhise aruandluse suunana reaalarajamajanduse kontseptsiooni laiemaks kasutusele võtmiseks. Esmaseid andmeid juba laekub Statistikaametile, edasise osas on põhjust olla veendumusel, et informatsiooni laekumine järjest paraneb.

Aruandlus 3.0 aitab varase hoiatamise sisendinfo kvaliteeti tõsta kolmest aspektist tulenevalt:

- paraneb andmete detailsuse aste;
- paraneb andmete laekumise kiirus;
- tekib täiendav võimalus majandusinfo modelleerimiseks.

Andmete detailsuse aste. Aruandlus 3.0 kaudu laekuvad ettevõtete tehinguandmed. Tehinguandmed on täpsemad, kui majandusaasta aruannetest saadavad andmed. See asjaolu võimaldab analüüsida ettevõtet täpsemalt, kui koondbilansi ja kasumiaruande põhjal on võimalik seda teha. Varase hoiatuse algoritme on võimalik muuta täpsemaks, arvuti saab masinõppe sisendina kasutada täpsemaid parameetreid ja väljundtulemust on võimalik kasutada rohkema informatsiooni allikana.

Andmete laekumise kiirus. Aruandlus 3.0 on mõeldud reaalarajas toimiva lahendusena, mis tähendab seda, et süsteemi normaalse funktsioneerimise korral on võimalik saada andmeid kuupõhiselt (kui mitte kiiremini). See annab võimaluse hinnata, mis ettevõttega toimub ja kas ettevõtte on endiselt samas/võrreldavas seisus, mis viimase majandusaasta aruande põhjal tehtud analüüs ütleb. Varase hoiatuse süsteem saab täiendava sisendi ettevõtte stabiilsuse hindamiseks.

Majandusinfo modelleerimine. Ettevõtete majandusaasta aruanded on olemas teatud (mineviku) ajalise seisuga, vahest aasta-poolteist vana informatsioon. Kui Aruandlus 3.0 kaudu laekub tehinguinfot reaalarajas (või väiksema viivitusega, kui aruanded endid), siis selle info (ja eelnevalt teada olevate suhete põhjal) on võimalik modelleerida bilansse ja kasumiaruandeid hetkeseisuga (teatud täpsusastemega). Viimane omakorda võimaldab anda varase hoiatuse hinnanguid sisuliselt reaalarajas (või siis vastavalt näiteks kuu-paari ajalise nihkega, mitte aastataguse info põhjal).

Ajalises perspektiivis vaadatuna on hetkeseisuga võimalik olla veendumusel, et Aruandlus 3.0 projekti raames informatsiooni laekumise hulk (rohkem andmeedastajaid) ja paljutahulisus (andmed rohkemate erinevate andmejaotuste lõikes) paraneb ning seda rohkem on võimalik andmeid kasutada sisendina varase hoiatamise teenuse juures. Varase hoiatamise teenuse väljatöötamise juures võetakse sisendina koheselt kasutusele täiendavaid Aruandlus 3.0 projekti raames laekunud andmeid, niipea, kui need laekuvad.

## 1.5. Tegevusala määratlemine

Tegevusala määratlemine on oluline, sest erinevates tegevusalades on potentsiaalne maksejõuetus nii suhtarvude kui ka ettevõtete vaheliste transaktsioonide mõistes täiesti erineval viisil kirjeldatav. Näiteks alljärgnevate ettevõtete võimaliku maksejõuetuse määramisel tuleb bilansi-, kasumiaruande- ja transaktsiooniväärtustele panna hoopis erinevad võimaliku maksejõuetuse ennustamise parameeterväärtused (näited on illustreerivad ja nimetused „rahvalikud“):

- suur hulgifirma;
- väike konsultatsioonifirma;
- teadusarenduskeskus;
- põllumajandusettevõtte;
- üksik taksosõitja;
- metallitööstusettevõtte;
- hambaravikabinet;
- töövaidlusbüroo;
- hamburgeri putka;
- sadam;
- kolme inimese koristusfirma.

Ülaltoodud nimetused on näited, kuid neid vaadates on ilmselge, et need erinevad käiberaha vajaduselt, võimalikult pikaajalise laenuvajaduselt, varude haldamise põhimõtelt, tootmisvahendite vajaduselt, üldiselt bilansimahult, rentaabluselt, käivetelt, käibekiiruselt, sesoonsustundlikkuselt ja paljude muude parameetrite poolest. Seega ei ole võimalik välja töötada piisavalt täpset ja asjalikku masinõppe algoritmi, mis suudaks kõigi ülaltoodud ettevõtete kohta leida parameetrid, mis viitavad võimalikule maksejõuetusele, need parameetrid (eelkõige parameetrite nn. „punased väärtused“) on erinevad. Küll aga on lahendused võimalikud, kui vaadatakse sarnase tegevusala ettevõtteid.

Tegevusalade määratlemise alusel kasutatakse Eesti Majanduse Tegevusalade Klassifikaator EMTAK koodi. Samas ei ole käesoleva töö raames võimalik läheneda selliselt lihtsakoeliselt, et kasutada näiteks EMTAK koodi esimesel tasandil (s.t. võtta aluseks koodi esimene number), teisel tasandil (võtta aluseks kaks esimest numbrit) jne. Selline, küll tehniliselt lihtsam käsitlus, annaks eksitava pildi. Alljärgnevalt toodud näidetega on eksitava pildi tekkimise loogikat põhjendatud.

Näiteks, kui soovida defineerida meditsiinisektori tegevusvaldkonda, siis muuhulgas peavad olema kaasatud järgmised tegevusalad.

- 87101 Hooldusraviasutuste tegevus
- 86211 Üldarstiabi osutamine

Ülaltoodud kahe tegevusala näitel on mõistetav, et meditsiinisektor peaks sisaldama mõlemat ülaltoodud tegevusala. Seega ei ole võimalik minna EMTAK teisele tasemele, tuleb jääda EMTAK esimesele tasandile ja sisse võtta kõik tegevusalad, mille EMTAK algab numbriga „8“. Samas jällegi, siis satuks meditsiinisektori hulka järgmine tegevusala:

- 88911 Lapsehoiuteenus

Selge on see, et lapsehoiuteenus ei ole arstiabi, meditsiinisektor. Samas see on arvestatav tegevusala, mida kasutatakse, mida ettevõtte oma tegevuseks märgivad ja see ei ole ala, mida saaks niisama „mürana“ ignoreerida. Minnes aga EMTAK teisele tasandile ja võttes sisse näiteks koodi „86“ jätame me välja „87“, hooldusraviasutuste tegevuse – tegemist on ikkagi raviasutustega ja kui rääkida meditsiinisektorist, siis ilma nendeta ei saa sellest rääkida.

Soovides defineerida finantssektorit vastandume muuhulgas järgmiste tegevusaladega:

- 64191 krediidasutused (pangad)
- 64301 investeerimisfondid
- 64911 kapitalirent
- 65201 edasikindlustus
- 65301 pensionifondid
- 66111 finantsturgude haldamine
- 66121 väärtpaberite maaklerlus
- 68101 enda kinnisvara ost ja müük

Kui me jääme EMTAK esimese taseme juurde, siis tuleb finantssektorisse sisse ka enda kinnisvara ost ja müük, mis arusaadavalt sinna sektorisse ei sobi. Järelikult esimese tasandi juurde jääd ei saa. Kui minna teise tasandi juurde, siis see ju tähendaks seda, et määratletakse kaks esimest kohta, ehk siis ülaltoodud näite kontekstis, defineerimaks finantssektorit, millised kaks kohta, „64“, „65“ või „66“ peaks määraltema? Tegelikult vajame kõike kolme.

Võttes järgmiseks näiteks kinnisvarasektori, vastandume näiteks järgmise olukorraga:

- 41101 hoonestusprojektide arendus
- 41201 elamute ja mitteeluhoonete ehitus
- 42111 teede ja kiirteede ehitus
- 42121 raudtee- ja metroo ehitus
- 42131 silla- ja tunneliehitus
- 68101 enda kinnisvara ost ja müük
- 68201 kinnisvara üürileandmine
- 68311 kinnisvarabüroode tegevus
- 68321 hoonete ja üürimajade haldus

Kõik ülaltoodud tegevusalad mahuvad kinnisvarategevuse alla. Mis on aga eriti tähtis, sinna alla mahuvad tegevusi, mis on EMTAK esimese koodi järgi „4“ ja samuti, mis on esimese koodi järgi „6“. Lihtsustatud lähenemisnurka kasutades peaksime võtma EMTAKi esimese taseme, koodid „4“ ja „6“. Seda aga jällegi teha ei saa, sest tulenevalt varasemast näitest läheksid sinna alla ka sellised tegevusalad, nagu alltoodud ja mis ei ole kindlasti kinnisvarasektori tegevused:

- 64191 krediidasutused (pangad)
- 64301 investeerimisfondid
- 64911 kapitalirent
- 65201 edasikindlustus
- 65301 pensionifondid
- 66111 finantsturgude haldamine
- 66121 väärtpaberite maaklerlus

Proovides määratleda haridust näeme näiteks järgmisi tegevusalasid:

- 85101 lastesõimede tegevus
- 85102 lasteaedade tegevus

85521 muusika ja kunstikoolitus  
85591 keeleõpe  
85592 arvutiõpe  
85529 Muu huvikoolitus

EMTAK teise koha peale, ehk siis kood „85“ jääda ei saa, sest siis oleks sees ka lastesõimede ja lasteaedade tegevus. Teatud juhtudel võib see sobida, kuid kui haridust kitsamalt analüüsida, siis peaksid need asutused välja jääma. Seega teisele kohale ei saa jääda, läheme kolmandale, kood „855“. Siis võib aga endiselt sees olla muu huvikoolitus (näiteringid, fotoringid), mis ei pruugi jällegi sobida. Minnes neljandale kohale, kood „8559“, saame (ülaltoodud näite kontekstis) sisse küll keeleõppe ja arvutiõppe, kuid muusika ja kunstikoolitus jääksid välja. See on aga täiesti tavapärase hariduse osa ja ei saaks välja jääda.

Ülaltoodud näited on vaid osad, kuid siiski piisavalt näitlikud mõistmaks, et tegevusalade määramisel ei saa kasutada lihtsustatud lähenemist (võtta EMTAKi teatud tasand, mõned kohad).

Lahendusena nähakse siinkohal varianti, et tegevussektorid defineeritakse kombineerides 5 kohalisi EMTAK koodi, kombinatsioonid teostatakse vastavalt püstitatud ülesandele. Käesoleva materjali lisas 1 on toodud neli defineeritud tegevusvaldkonda: (a) meditsiinivaldkond, (b) finantsteenuste valdkond, (c) kinnisvaravaldkond ja (d) haridusvaldkond.

Lisas 1 toodud valdkondade taoline määramine on selles mõttes tinglik, et samad valdkonnad võib määrata teisiti, vajadusel kitsamalt või hoopis laiemalt, kasutades siis vastavalt teisi või teisel määral vajalikke EMTAK koodi. Oluline on aga see, et sektorite/tegevusalade/tootmisharude määramisel selliselt, et nende põhjal tehtavatest järeldustest oleks ka reaalselt kasu varase hoiatuse signaali väljastamisel, selleks on oluline need valdkonnad defineerida vastavalt eesmärgile. Kiiret ja kerget lahendust ei ole.

## 1.6. Maksejõuetuse terminoloogiline määratlus

Käesolev töö on suunatud sellele, et võimalikult varakult leida viiteid sellele, et ettevõtte on sattumas maksejõuetuse olukorda. Loodav mudel ei ole mõeldud pankrottide ennustamiseks või likvideerimismenetluste ettekuulutamiseks. Samas, kui vaadata viidatud teaduskirjandust, siis suur osa maksejõuetuse analüüsile suunatud masinõppe lahendustest on mõeldud ennustamiseks ette pankrotte.

Pankrot on süvenenud maksejõuetuse viimane aste, kuid kui masinõppe meetodid on välja töötatud pankrottide ennustamiseks, siis nende samade meetodite abil on võimalik tuvastada ka varasemaid maksejõuetuse faase (aasta-paar, või rohkem) enne pankrotisündmuse asetleidmist. Need varased faasid on (kus on oht maksejõuetuse tekkimiseks või on esmased märgid selle kergekujulisest esinemisest) kohaks, kus lähtuvalt tootmisharu/sektori spetsiifikast võib olla õige aeg ettevõttele varase hoiatuse signaali väljastamiseks. Seega, kui käesolevas analüüsis võetakse aluseks teaduskirjandust, mis masinõppe meetoditega ennustab pankrotte, siis sealt võetakse vaid see osa, mis on vajalik ja toeks varase maksejõuetuse faasi tuvastamiseks, mitte pankroti, kui maksejõuetuse (üldiselt) viimase faasi tuvastamiseks/ennustamiseks.

Pankrotiseadus (PankrS) § 1 lg 1 sätestab, et pankrot on võlgniku kohtumäärusega välja kuulutatud maksejõuetus. PankrS § 1 lg 2 selgitab, et võlgnik on maksejõuetu, kui ta ei suuda rahuldada võlausaldaja sissenõutavaks muutunud nõuet ja see suutmatus ei ole võlgniku majanduslikust olukorrast tulenevalt ajutine. PankrS § 1 lg 3 toob juurde juriidilisest isikust võlgniku maksejõuetuse määratluse, kus võlgniku vara ei kata võlgniku kohustusi ja tegemist ei ole ajutise olukorraga.

Varase hoiatuse väljastamise hetkel ei ole tegemist kindlasti püsiva maksejõuetuse seisuga, tõenäoliselt ei olda ka jõutud ajutise maksejõuetuse olukorda. Samas annab pankrotiseadus viiteid kahe analüüsi suuna peale:

- milline on potentsiaal, et ei suudeta rahuldada võlausaldajate nõudeid;
- milline on olukord varadega.

Nõuete rahuldamise võimekus. Tegemist on analüüsisuunaga, mis vaatab ettevõtte käibevahendeid, raha käibekiirust, üldist likviidust ja kõike seonduvat. Eesmärgiks on mõista, kas ettevõtte võib sattuda olukorda, kus tema majanduspositsioon muutub sedavõrd mittelikviidseks, et olemata küll majanduslikult raskes seisus ei suudeta ikkagi majandustegevusest tulenevaid kohustusi katta.

Varaline olukord. Tegemist on analüüsisuunaga, mis vaatab kuidas ettevõtte varaline positsioon muutub ajas ja kas see muutumine võib viidata maksejõuetuse tekkele tulevikus. Üldised põhimõtted on omakapitali suuruse jälgimine, milline on võla osakaal (teel raskustesse tavaliselt võla osakaal kasvab, omakapital väheneb nii suhteliselt kui ka nominaalselt).

Kahte ülaltoodud lähtenurka kombineerides, kui näiteks on viiteid sellele, et väheneb nii nõuete rahuldamise võimekus kui teisalt halveneb ka varaline olukord, siis on kindlasti põhjust hoiatussignaali väljastamiseks (kui siis juba hilja ei ole, võimalik, et see tuleb väljastada juba varem).

Maksejõuetuse eelfaas, mille pealt hoiatus väljastatakse, võib kanda väga erinevaid nimetusi (pigem on oluline tehniline defineerimine, millega praeguses töös tegeletakse). Näiteks kasutatud teaduskirjanduses antakse ka masinõppega analüüsitavale maksevõimetusele erinevaid nimetusi. Shi (2019) toob välja järgnevad terminid:

- Bankruptcy prediction;
- Default prediction;
- Financial failure;
- Financial distress;
- Insolvency;
- Business failure.

Seega terminoloogiliselt võib nii maksejõuetuse hetke kui ka sellele eelnevat faasi (millelt varase hoiatuse signaal väljastatakse) nimetata erinevalt, kuid oluline on see, et tehniliselt oleks see täpne ning masinõppe lahendused oleksid treenitud seda leidma.

## 1.7. Majandussuhtarvude kasutamine

Käesolevas töös esmaselt kasutatavad majandussuhtarvud (bilansi ja kasumiaruande pinnalt toodud suhtarvud) ja nende kasutamisoskuse põhimõtted antud töös on esitatud lisa 2. Tegemist on esmase valikuga, mis töö käigus võib muutuda, tõenäoliselt täieneda.

Cialone (2020) toob välja alljärgnevad bilansilised/kasumiaruande näitajad ja neil põhinevad suhtarvud, millel oma närvivõrkudel tehtud analüüsi rajas (järjekord selline, nagu Cialone poolt esitatud):

Current Assets  
Bank Debt to Sales  
Debt/EBITDA ratio  
Total Debt to Equity  
Receivables Average Collection Period  
Payables Average Settlement Period  
EBITDA to Interest Expenses  
Working Capital to Revenues

Current Ratio  
 Total Fixed Tangible Assets to Equity  
 (Equity+Long Term Debts)/Fixed Assets  
 Current Debts to Total Debts  
 Long Term Debts To Total Debts  
 Interest Expenses to Gross Sales  
 Total Equity  
 Net Financial Position  
 Total Assets to Equity  
 Return on Equity  
 Return on Assets  
 Return on Sales  
 Return on Investment  
 Gross Sales  
 Gross Working Capital Turnover  
 Invested Capital Turnover  
 Total Fixed Assets  
 Total Liabilities to Equity  
 Total Current Liabilities  
 Equity to Total Assets  
 (Assets-Inventories)/Debts  
 Long Term debt  
 Total Assets  
 EBITDA  
 EBITDA/Gross Sales  
 Total Credits  
 Total Debt  
 Number of Employees

Üldjoontes kapib ülaltoodu käesoleva materjali lisas 2 toodud näitajate loogikaga. Aluseks on võetud bilansi ja kasumiaruande näitajaid, nende põhjal on arvutatud suhteid ning siis saadud tulem on võetud sisendiks masinõppe meetodite rakendamiseks. Suhteliselt suur osakaal on bilansi ja kasumiaruande nominaal numbritel (võrreldes selle lähtenurgaga, et sageli on analüüsi aluseks üksnes suhtarvud ja nominaalnumbreid ei kasutata).

Alljärgnevas tabelis on toodud näitajad jagatud kategooriate kaupa.

Nominaalnumbrid	Kapitali struktuur	Efektiivus	Likviidsus	Rentaablus
Current Assets	Bank Debt to Sales	Receivables Average Collection Period	Current Ratio	Return on Equity
Total Equity	Debt/EBITDA Ratio	Payables Average Settlement Period	Gross Working Capital Turnover	Return on Assets
Net Financial Position	Total Debt to Equity		Invested Capital Turnover	Return on Sales
Gross Sales	EBITDA to Interest Expenses			Return on Investment
Total Fixed Assets	Working Capital to Revenues			
Total Current Liabilities	Total Fixed Tangible Assets to Equity			
Long Term Debt	(Equity+Long Term Debts)/Fixed Assets			
Total Assets	Current Debts to Total Debts			
EBITDA	Long Term Debts To Total Debts			

Total Credits	Interest Expenses to Gross Sales			
Total Debt	Total Assets to Equity			
Number of Employees	Total Liabilities to Equity			
	Equity to Total Assets			
	(Assets-Inventories)/ Debts			
	EBITDA/Gross Sales			

Tabel 1.: Näitajate jaotus kategooriate kaupa.

Ülaltoodud tabelist nähtub, et Cialone (2020) on eelkõige rõhku pannud ühelt poolt bilansi ja kasumiaruande nominaalnumbrite ja teiselt poolt struktuuri suhtarvude analüüsile. Efektiivsuse poole pealt vaadatakse vaid saada olevate rahade laekumise perioode ja seda, kui millises ajalisel raamistikus makstakse arveid (need kaks nimetatud näitajat jällegi toovad välja, kuidas raha liigub ettevõttesse sisse ja ettevõtetest välja – selles mõttes on tegu võtmenäitajatega).

Likviidsuse poole pealt vaatab Cialone (2020) kolme näitajat. Esiteks klassikalist Current Ratio't, (lühiajaliste võlgnevuste kattekordajat), ehk siis kuidas käibevaraga on võimalik katta lühiajalisi kohustusi – mida rohkem on käibevara lühiajaliste kohustuste vastu, seda parem. Siis ta vaatab käibekapitali käivet ja investeeritud kapitali käivet, kui mitu pööret nimetatud kapitalid aasta jooksul teevad (ehk siis mitu korda kasutab ettevõtte aasta jooksul oma käibekapitali ja mitu korda kasutab investeeritud kapitali – mida rohkem, seda parem).

Rentaablust vaatab Cialone (2020) neljas kategoorias, ehk siis milline on tootlikkus koguvarade, omakapitali, müügi ja investeeringute vastu. Omakapital ja bilansi maht on üldised ettevõtte suuruse (majandusulatus) näitajad, müük ja investeeringud teised kaks võtmenäitajat – müük (käive) näitab, kuidas turul läheb, investeeringud näitavad seda, kuidas omanikud selle müügi vastu on panustanud (millist raha mängu pannud, et sellist käivet saavutada ja selle käibe pealt siis tootlust teenida).

Esmapilgul vaadates tundub Cialone (2020) analüüs nominaalnumbrite ja kapitali struktuuri poole kaldu olevat, kuid kui analüüsida, mida efektiivsuse, likviidsuse ja rentabluse osas on sisse toodud, siis on võetud teatud kõige rohkem võtmenäitajateks loetud sisendid.

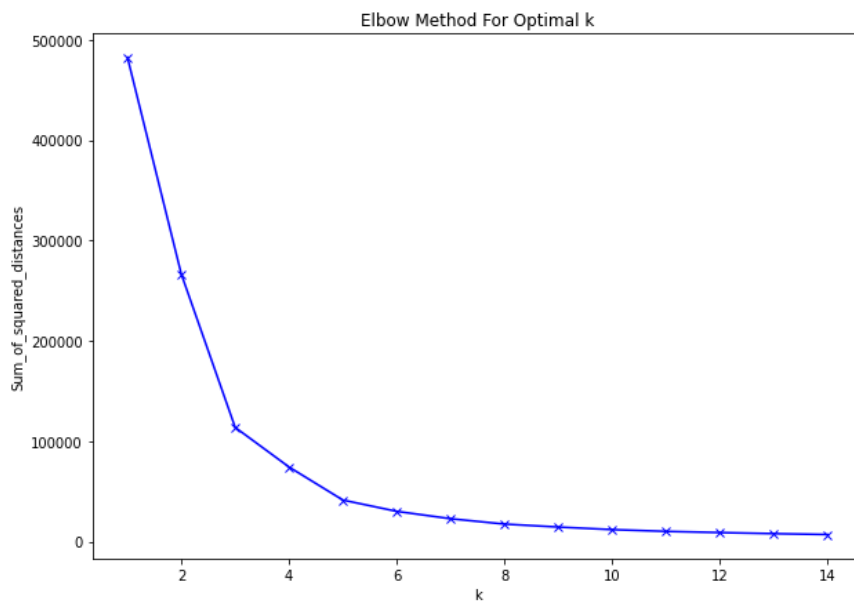
Cialone (2020) valik on ka põhjendatud tema seisukohaga, et väga suur osa tööst (peamine osa tööst) läheb andmete ettevalmistamisele ja sellele, et tagada õiged sisendparameetrid. Ta on kasutanud nii tegutsevate ettevõtete kui ka pankrotistunud ettevõtete andmeid, kusjuures andmed on väga olulisel määral tegutsevate ettevõtete poole kaldu (neid on oluliselt rohkem). Andmete tasakaalustamiseks on kasutatud kahte meetodit: (a) Synthetic Minority Oversampling Technique (SMOTE) ja (b) Adaptive Synthetic Sampling Method (ADASYN).

Käesoleva töö järgse prototüübi ettevalmistamise käigus lähtutakse ühelt poolt teoreetilise teaduskirjanduse analüüsist (nii ärianalüüsi kui ka masinõppe poole pealt), kuid teiselt poolt ka on hakatud praktiliselt läbi katsetama, kuidas üks või teine teoreetiline kontseptsioon reaalandmete peal töötab. Seega ei ole mindud selliselt, et teoreetiliselt on teema väga lõpuni välja töötatud ja siis hakatakse peale praktilise poolega. Praktilist poolt arendatakse koos teoreetiliste kontseptsioonide läbitöötamisega.

Esmases analüüsis on kasutatud majandusaasta aruannete andmeid ja ka pankrotistunud ettevõtete andmebaasi. Pankrotistunud ettevõtete andmebaasi sidumisega kogu majandusaasta aruannete andmebaasiga on saadud Cialone (2020) võrreldav andmestik. Kuna ka reaalandmestik on väga tegutsevate ettevõtete poole kaldu (pankrotistunud ettevõtete andmeid on vähe võrreldes reaalselt tegutsevate ettevõtetega), siis on andmete tasakaalustamiseks kasutatud samuti Synthetic Minority Oversampling Technique (SMOTE) tehnikat. Analüüsi aluseks on võetud iga ettevõtte puhul viie

majandusaasta andmed. Pankrotistunud ettevõtete puhul on andmebaas moodustatud selliselt, et on võetud pankroti hetk ja sellest on tagasi võetud viis aastat, ehk siis 1 aasta enne pankroti väljakuulutamist, 2 aastat enne pankroti väljakuulutamist jne. Ettevõtete pankrotid on välja kuulutatud erinevatel aastatel, seega andmebaas on moodustanud nii, et viis aastat, mis pankrotile eelneb võivad aastanumbrite lõikes olla ettevõtete erinevad, kuid oluline on see, et iga pankrotistunud ettevõtte jaoks on olemas viis eelnevat tegutsemisaastat (sealt saab otsida pankrotile eelnenud mustreid).

Esmaselt on tehtud ära klasterdamisanalüüs (k-means algoritmiga) nägemaks, milline on üldine pilt ettevõtluses. Klasterite arvu määramisel on kasutatud Elbow meetodit – meetod Klasterite maksimumarvu määramiseks, kus täiendava klasteri lisamine ei anna enam oluliselt paremat lisatulemit. Klasterite arvuks analüüsis on määratud 5, määramise aluseks olev Elbow meetodi kõver toodud alljärgneval joonisel.



Joonis 1.: Klasterite arvu määramine.

K-means algoritm jagab n objekti k klasterisse selliselt, et iga objekt kuuluks sellesse klasterisse, mille keskpunktile see kõige lähemal on.

Esmases klasterdamises on kasutatud näitajad on alljärgnevad:

- BI\_100\_1 - Bilanss, käibevarad
- BI\_150\_1 - Bilanss, materiaalne põhivara
- BI\_180\_1 - Bilanss, põhivara kokku
- BI\_190\_1 - Bilanss, varad kokku
- BI\_240\_1 - Bilanss, lühiajalised laenukohustused
- BI\_250\_1 - Bilanss, võlad ja ettemaksud
- BI\_290\_1 - Bilanss, lühiajalised kohustused
- BI\_310\_1 - Bilanss, pikaajalised kohustused
- BI\_370\_1 - Bilanss, kohustused kokku
- BI\_400\_1 - Bilanss, osakapital
- BI\_40\_1 - Bilanss, raha
- BI\_40\_2 - Bilanss, raha eelmise aruandeperioodi lõpus
- s1 - BI\_100\_1/ BI\_190\_1 – käibevara osakaal varades
- s2 - BI\_290\_1/ BI\_190\_1 – lühiajalised kohustused varades



- s3 - BI\_310\_1/ BI\_190\_1 – pikaajalised kohustused varades  
s4 - BI\_370\_1/ BI\_190\_1 – kohustused kokku varades  
s5 - BI\_400\_1/ BI\_190\_1 – osakapital varades  
s6 - BI\_40\_1/ BI\_100\_1 – raha osakaal käibevaras

Nagu ülaltoodust näha on esmases klasterdamises mindud bilansi nominaalnumbrite ja kapitali struktuuri suhtarvude juurde. Võetud on bilansi maht (alusnäitaja), kontsentreeritud on kohustustele (maksejõuetus on suutmatus täita oma kohustusi) ja käibevarale (käibevara näitab rahalist manööverdamisruumi, ka seda, kui palju on võimalik vajadusel kiiresti kohustusi katta – likviidsuse üks kriteeriumitest).

Kontsentreeritud on bilansilistele näitajatele, sest bilanss näitab seisu, kasumiaruanne näitaks aastasisest dünaamikat. Ka aastasisese dünaamika muustrite sissetoomine on oluline, kuid süsteemi loomine toimub etapiviisiliselt, alustatakse staatilisest vundamendist.

Alljärgnevas tabelis on toodud klasterdamise esmased tulemused.

clusters	BI_100_1	BI_150_1	BI_180_1	BI_190_1	BI_240_1	BI_250_1	BI_290_1	BI_310_1	BI_370_1	BI_400_1	BI_40_1	BI_40_2
0	202 479	109 020	319 297	521 742	45 686	67 410	115 668	90 903	217 362	18 326	53 184	45 727
1	323 292	250 233	569 848	836 682	91 308	130 467	215 077	186 630	421 360	17 568	59 752	54 504
2	58 414	44 133	64 355	124 112	3 410	18 066	32 470	18 724	56 763	943	30 585	28 267
3	241 737	125 545	328 041	569 779	52 436	99 852	154 158	101 494	267 362	13 824	53 021	46 453
4	34 822	23 583	43 630	78 344	2 768	9 764	17 945	15 126	38 506	622	23 074	20 430

Tabel 2.: Klasterdamise esmased tulemused.

Alljärgnevas tabelis on toodud ettevõtete kirjete arv klastrite lõikes (ühe ettevõtte kohta on mitu kirjet).

Klastri number	Klastrisse sattunud ettevõtete arv
0	493 985
1	463 494
2	169 205
3	505 960
4	149 264

Tabel 3.: Klastrisse sattunud ettevõtete kirjete arv.

Nagu eelnevalt selgitatud, on andmestikku lisatud pankrotistunud ettevõtete andmed. Alljärgnevas tabelis on toodud pankrotistunud ettevõtete arv klastrite lõikes.

Klastri number	Pankrotistunud ettevõtteid
0	237
1	658
2	12
3	471
4	14

Tabel 4.: Pankrotistunud ettevõtete arv.

Lisas 3 on toodud pankrotistunud ettevõtete arv aastate ja klastrite lõikes.

Üheks põhialuseks esmase testimise juures on võetud bilansi maht. Varade maht näitab muude tingimuste samaksjäädes ettevõtte majandustegevuse ulatust, majandusvõimu, stabiilsust (suurem peaks olema stabiilsem) jms. Alljärgnevas tabelis on toodud välja klastritesse koondunud ettevõtete analüüs suuruse (varade mahu BI\_190\_1) järgi (algoritm on klasterdamisel võtnud aluseks kõiki parameetreid, analüüs on tehtud tulenevalt varade mahust).

clusters	BI_100_1	BI_150_1	BI_180_1	BI_190_1	BI_240_1	BI_250_1	BI_290_1	BI_310_1	BI_370_1	BI_400_1	BI_40_1	BI_40_2
0	202 479	109 020	319 297	521 742	45 686	67 410	115 668	90 903	217 362	18 326	53 184	45 727
1	323 292	250 233	569 848	836 682	91 308	130 467	215 077	186 630	421 360	17 568	59 752	54 504
2	58 414	44 133	64 355	124 112	3 410	18 066	32 470	18 724	56 763	943	30 585	28 267
3	241 737	125 545	328 041	569 779	52 436	99 852	154 158	101 494	267 362	13 824	53 021	46 453
4	34 822	23 583	43 630	78 344	2 768	9 764	17 945	15 126	38 506	622	23 074	20 430

Tabel 5.: Ettevõtete jaotus klastritesse varade mahu alusel.

Kõige suuremad ettevõtted on koondunud klastrisse 1, keskmise varade mahuga suurusjärgus 837 000 eurot. Järgmised ettevõtted, klastritesse 0 ja 3, on n.ö. keskmised (ehk pigem suured-keskmised) keskmise varade mahuga vastavalt 522 000 eurot ja 570 000 eurot. Klastritesse 2 ja 4 on koondunud ettevõtted, mida võib suhteliselt väikesteks pidada.

Võrreldes tabelit 4 ja tabelit 5 on näha, et kõige rohkem pankrotte on kõige suuremate ettevõtete hulgas, klastris 1. Järgnevad nii pankrottide arvu, kui suuruse järgi klastrid 3 ja 0. Kõige vähem pankrotte on väikeste hulgas, sisuliselt mõned loetud pankrotid.

Praktiline loogika on ka seda lähenemist toetav, sest väikesed ettevõtted ei jõua sageli ametliku pankrotini (vähe võlausaldajaid, tihedamad sidemed, väiksemad nõuded jms.), küll aga leidub suurtel alati mõni võlausaldaja, kes nõuab ametlikku menetlust (tunneb, et ta tõrjutakse muidu ebaõiglaselt varade jaotusest kõrvale, mängus on sageli maksuvõlg jne.).

Alljärgnevas tabelis on toodud välja ülalpool esitatud suhtarvud ja bilansi need andmed, mille pealt suhtarvud on arvatud.

clusters	BI_100_1	BI_190_1	BI_290_1	BI_310_1	BI_370_1	BI_400_1	BI_40_1	s1	s2	s3	s4	s5	s6
0	202 479	521 742	115 668	90 903	217 362	18 326	53 184	0,39	0,22	0,17	0,42	0,04	0,26
1	323 292	836 682	215 077	186 630	421 360	17 568	59 752	0,39	0,26	0,22	0,50	0,02	0,18
2	58 414	124 112	32 470	18 724	56 763	943	30 585	0,47	0,26	0,15	0,46	0,01	0,52
3	241 737	569 779	154 158	101 494	267 362	13 824	53 021	0,42	0,27	0,18	0,47	0,02	0,22
4	34 822	78 344	17 945	15 126	38 506	622	23 074	0,44	0,23	0,19	0,49	0,01	0,66

Tabel 6.: Bilansi pealt arvatud suhtarvud klastrites.

Ülaltoodud tabelis toodud suhtarvud on järgmised:

- s1 - BI\_100\_1/ BI\_190\_1 – käibevara osakaal varades
- s2 - BI\_290\_1/ BI\_190\_1 – lühiajalised kohustused varades
- s3 - BI\_310\_1/ BI\_190\_1 – pikaajalised kohustused varades
- s4 - BI\_370\_1/ BI\_190\_1 – kohustused kokku varades
- s5 - BI\_400\_1/ BI\_190\_1 – osakapital varades
- s6 - BI\_40\_1/ BI\_100\_1 – raha osakaal käibevaras

Alljärgnevalt on toodud klastreid kirjeldav koondtabel.

	Klaster 0	Klaster 1	Klaster 2	Klaster 3	Klaster 4
Ettevõtete kirjete arv	493 985	463 494	169 205	505 960	149 264
Pankrotte	237	658	12	471	14
Ettevõtete suurus varade mahult	521 742	836 682	124 112	569 779	78 344

Sõnaline ettevõtte suurus	Keskmi- suured ettevõtted	Suured ettevõtted	Väikesed ettevõtted	Keskmi- suured ettevõtted	Väikesed ettevõtted
s1 – Käibevara osakaal varades	0.39	0.39	0.47	0.42	0.44
s1 kommentaar	Min väärtus	Min väärtus	Max väärtus	Keskmine	Suuruselt 2.
s6 – Raha osakaal käibevaras	0.26	0.18	0.52	0.22	0.66
s6 kommentaar	Suuruselt 3.	Min väärtus	Suuruselt 2.	Suuruselt 4.	Max väärtus
s4 – kohustused kokku varades	0.42	0.50	0.46	0.47	0.49
s4 kommentaar	Min väärtus	Max väärtus	Suuruselt 4.	Suuruselt 3.	Suuruselt 2.
s3 – pikaajalised kohustused varades	0.17	0.22	0.15	0.18	0.19
s3 kommentaar	Suuruselt 4.	Max väärtus	Min väärtus	Suuruselt 3.	Suuruselt 2.
s2 – lühiajalised kohustused varades	0.22	0.26	0.26	0.27	0.23
s2 kommentaar	Min väärtus	Suuruselt 2.-3.	Suuruselt 2.-3.	Max väärtus	Suuruselt 4.
s5 – osakapital varades	0.04	0.02	0.01	0.02	0.01
s5 kommentaar	Max väärtus	Suuruselt 2.-3.	Min väärtus	Suuruselt 2.-3.	Min väärtus
Kirjeldav järelalus	Käibevara vähe, raha keskmiselt, kohustusi vähe, pikaajalisi kohustusi keskmiselt, lühiajalisi kohustusi vähe	Käibevara vähe, raha vähe, kohustusi palju, pikaajalisi kohustusi palju, lühiajalisi kohustusi keskmiselt	Käibevara palju, raha keskmiselt, kohustusi keskmiselt, pikaajalisi kohustusi vähe, lühiajalisi kohustusi keskmiselt	Käibevara keskmiselt, raha keskmiselt, kohustusi keskmiselt, pikaajalisi kohustusi keskmiselt, lühiajalisi kohustusi palju	Käibevara keskmiselt, raha palju, kohustusi keskmiselt, pikaajalisi kohustusi keskmiselt, lühiajalisi kohustusi keskmiselt

Tabel 7.: Klasterdustulemust kirjeldav koondtabel.

Ülaloodud tabeli kontekstis on oluline, et hinnanguvärve (roheline, kollane, punane) on kasutatud selliselt, et punane on „halb“, kollane „enam-vähem“ ja roheline „hea“. Need ei ole korrelatsioonis min ja max väärtustega. Näiteks käibevara min väärtus on halb (seega min on punane) ja kohustuste max väärtus on samuti halb (seega hoopis max on punane).

Ülaloodud tabeli põhjal saab järelalus järgmist: Väiksematel on suurem käibevara osakaal bilansis, seega rohkem suhtelist paindlikkust ja suhtelist võimekust probleemolukordade lahendamiseks. Väiksematel paikneb suhteliselt suurem osa käibevarast rahas, mis jällegi tagab suhteliselt võttes parema manööverdamisruumi. Kõige suuremad omavad ka kõige suuremat kohustuste osakaalu bilansis. Kõige väiksem on kohustuste osakaal bilansis suuruselt kolmandate ettevõtete (keskmiselt suured – Klaster 0) ettevõtetel. Ülejäänutel on kohustusi keskmisel määral.

Pikaajalisi kohustusi on kõige rohkem suurematel ettevõtetel, väikeste nihetega on pikaajaliste kohustuste bilansiproportsiooni suurus otseses korrelatsioonis varade mahuga. Pikaajaliste kohustuste suur osakaal näitab sõltuvust finantseerijatest (pankadest), mis võib majanduskeskkonna muutudes viia olukorrani, kus senisel viisil läbi viidud majandustegevus ei saa jätkuda (kuni selleni välja, et pangad kutsuvad laenu ennetähtaegselt tagasi ja ettevõtte peab lõpetama tegevuse). Pikaajalised kohustused on tavaliselt suures mahus pangalaenu ja nende kohustuste suur osakaal näitab teiselt poolt usaldusväärset pankade silmis (ehk siis teistpidi jälle majandustervist, mille peale laenu on antud).

Lühiajaliste kohustuste vähesuselt paistab silma Klaster 0, mis oli ka minimaalses väärtuses kõigis kohustuste mahus ja oli ka suhteliselt väikese (tagant teine) pikaajaliste kohustuste mahus, varade mahult on suuruselt kolmas. Ehk siis Klaster 0 kirjeldab ühte kahest keskmiselt suurte gruppist, kuid mis selgelt torkab silma konservatiivsusega võla poolest (väiksem laenuvõimendus teiselt poolt toob kaasa ka väiksema käibevara, sest osa laenuvahenditest paikneb tavaliselt paratamatult käibevaras). Teine keskmiselt suurte grupp (Klaster 3) on lühiajaliste kohustuste maksimumnäitajaga, ka suurema käibevara osakaaluga (ehk siis jääb võlgu, võtab „kiiralaenu“ ja sellevõrra on likviidseid vahendeid

rohkem). Kokkuvõttes seega on kaks keskmiselt suurte gruppi, kuid need selgelt erinevad majanduskonservatiivsusest.

Osakapital on oma olemuselt rohkem administratiivne näitaja, kuid teades varade mahtu, kohustuste mahtu ja osakapitali saame leida selle osa omakapitalist, mis ei ole osakapital (ehk siis jaotamata kasumi, reservid jms.). Praegusel esmasel kujul seda analüüsi läbi viidud ei ole, kuid mudeli täpsustamisel tuleb sellega arvestada.

Üldiselt näitab tabel 7, et suured ettevõtted (Klaster 1) on jäigad, suhteliselt kohustustega koormatud (olguigi, et tänu oma suurusele usaldusväärsed ja saavad laenu) ning täiendavalt näitab tabel 4, et suurte hulgas on ajalooliselt olnud ka rohkem pankrotte. Teised turuosaliselised lasevad suurtel endale võlgu jääda, usaldavad järeelmaksule, pangad usaldavad, kuid siiski on neis olemas olemuslik jäikus, mis suurendab maksejõuetuse tõenäosust. Jäik struktuur on potentsiaalse nõrkuse allikaks (nagu ka suur majandusulatus võib turutingimuste muutudes eelise asemel probleemiks osutada).

Klaster 0, mis on konservatiivne keskmine suur, selles sektoris on ajalooliselt pankrotte olnud 237. Teine keskmiselt suur, Klaster 3, mis on laenamise poolest agressiivne, selles sektoris on ajalooliselt pankrotte olnud 471, sisuliselt peaaegu täpselt poole rohkem. Näide hea selles mõttes, et võõrvahendid sisaldavad olemuslikku majandusriski ja see majandusrisk ka reaalselt avaldub.

Töö käigus on läbi viidud erinevaid klasterdamisi, kasutatud selleks erinevaid klasterdamise algoritme, mõned täiendavad näited sellest on toodud käesoleva materjali lisa 4.

Lisas 5 toodud joonisel on esitatud olukord, kus graafiliselt on esitatud 20 parameetri taandamine 2-le parameetrile, et oleks võimalik neid kahedimensionaalsel joonisel kuvada. Klasterid tähistatud erinevate värvidega, klasteri tsentroidid on tähistatud musta värvi x-ga. Tegemist on tehnilise visualisatsiooniga, mitte lõpp(inim)kasutaja jaoks mõeldud tõlgenduspildiga.

**Ülaloodud analüüs ei ole lõplik, vastupidi, väga esmane algus sellest, kus suunas prototüübi väljatöötamine liigub. See on käesolevas materjalis esitatud mitte järelduste tegemiseks, vaid näiteks, kuidas masinat õpetatakse aru saada sellest, milline ettevõtte võib omada nn. majandusnõrkuse tunnuseid, mis omakorda võivad välja viia maksejõuetuse olukorda.**

Ehk siis majandustunnuste, tegevustunnuste jms. järgi jaotatakse ettevõtted klasteritesse (gruppidesse, alagruppidesse, seostesse jne.). Masin saab teada, mis on „hea“, mis on „halb“ (värvid ülal), mis mida mõjutab, millised on ühe või teise klasteri piirväärtused (piirväärtused parameetrite lõikes, nende omavahelised kombinatsioonid), kus võivad hakata tekkima riskid, kuidas hinnata nende riskide materialiseerumise tõenäosust jne.

Ülal on esitatud praegune olukord, töö lähteseis. Sellele järgnevad edasised sammud:

1. Täiendavad klasterdamised;
2. Kasumiaruande poole sissetoomine, rahavoogude hindamine (praegu vaid bilanss);
3. Täiendavate suhtarvude sissetoomine (lisa 2);
4. Analüüsi viimine sektorite tasandile (4 näidet lisa 1);
5. Klasterite outlierite defineerimine (positiivses ja negatiivses suunas);
6. Täiendavate parameetrite leidmine (unsupervised learning).

## 1.8. Hoiatusteate tasemed ja teadete kordamine

Varase hoiatuse signaal väljastatakse kolme taseme korral (käesolevas materjalis ollakse seisukohal, et tasemed on mõistlik seostada maksejõuetusmenetluse analoogiaga):

1. tase – roheline tase, kõige nõrgem signaali tase, kõige väiksem hälbimine normaalsest olukorrast;

2. tase – kollane tase, vahepealse raskusastmega signaali tase, keskmine hälbimine normaalsest olukorrast;
3. tase – punane tase, kõige suurema raskusastmega signaal, kõige kaugem hälbimine normaalsest olukorrast.

### Tase 1

1. taseme hoiatus väljastatakse ettevõttele siis, kui on selgelt näha, et ettevõtte hälbib normaalsest olukorrast (mida tähendab „selgelt näha“, see defineeritakse edasise töö käigus, see on sektoripõhine). 1. taseme hoiatus võib olla ka põhjendamatu, kuna iga ettevõtte on unikaalne, siis ei pruugi leitud hälbimine tegelikult viidata maksejõuetusele või selle tekkele. Samas üldjuhul peaks 1. taseme hoiatus viitama sellele, et on esmaseid märke, mille süvenemisel võib ettevõtte maksevõime hakata vähenema.

### Tase 2

2. taseme hoiatus väljastatakse ettevõttele siis, kui maksevõime on juba vähenenud ja kui on põhjendatud kahtlus (arengudünaamika viitab, et ettevõtte maksevõime on süsteemses vähenemises), et maksevõime võib väheneda sinnamaale, kus ettevõtte võib sattuda pankrotimenetlusse (mida tähendab „põhjendatud kahtlus“, see defineeritakse edasise töö käigus, see on sektoripõhine). On tekkinud ajutised makseraskused ja kui midagi ette ei võeta, siis on oht püsiva maksejõuetuse tekkimiseks. 2. taseme hoiatus viitab sellele, et midagi tuleb väga kohe ja kiirelt ette võtta, tegemist ei ole lihtsalt valehäire või subjektiivse konservatiivsushinnanguga. Üheks võimalikuks lahenduseks olukorra päästmisel oleks näiteks saneerimismenetluse käivitamine.

### Tase 3

3. taseme hoiatus väljastatakse ettevõttele siis, kui on tõenäoliselt tekkinud oht, et ettevõtte peab esitama pankrotiavalduse (võib olla tekkinud ka ÄS § 176 või § 301 viidatud vara vähenemise olukord). 3. taseme hoiatuse eesmärk on juhtida ettevõtte juhatuse tähelepanu asjaolule, et ettevõtte on tõenäoliselt (niivõrd, kuivõrd arvuti seda järeldada suudab) sattunud PankrS § 1 lg 2, või PankrS § 1 lg 3 määratletud olukorda ja juhatuse peab astuma ÄS § 180 lg 5<sup>1</sup> või ÄS § 306 lg 3<sup>1</sup> sätestatud samme.

ÄS § 180 lg 5<sup>1</sup> sätestab, et kui osaühing on maksejõuetu ning maksejõuetus ei ole tema majanduslikust olukorrast tulenevalt ajutine, peab juhatuse viivitamata, kuid mitte hiljem kui 20 päeva möödumisel maksejõuetuse ilmnemisest esitama kohtule osaühingu pankrotiavalduse (ÄS § 306 lg 3<sup>1</sup> sätestab analoogia aktsiaseltsi osas).

PankrS § 1 lg 2 selgitab, et võlgnik on maksejõuetu, kui ta ei suuda rahuldada võlausaldaja sissenõutavaks muutunud nõuet ja see suutmatus ei ole võlgniku majanduslikust olukorrast tulenevalt ajutine. PankrS § 1 lg 3 toob juurde juriidilisest isikust võlgniku maksejõuetuse määratluse, kus võlgniku vara ei kata võlgniku kohustusi ja tegemist ei ole ajutise olukorraga.

### Varase hoiatuse teate kordamine

Süsteem programmeeritakse varase hoiatuse teadet kordama. Kordamine on vajalik, sest eesmärgiks on saavutada olukord, kus ettevõtte võtaks hoiatussignaali saamise peale midagi ette.

Algoritm väljastab uue hoiatusteate kahel juhul:

- (a) kui süsteem näeb, et hoiatuse tase on muutunud halvemuse suunas;
- (b) eelmisest hoiatusest on möödunud kuus kuud ja olukord ei ole muutunud.

## Teate sisu

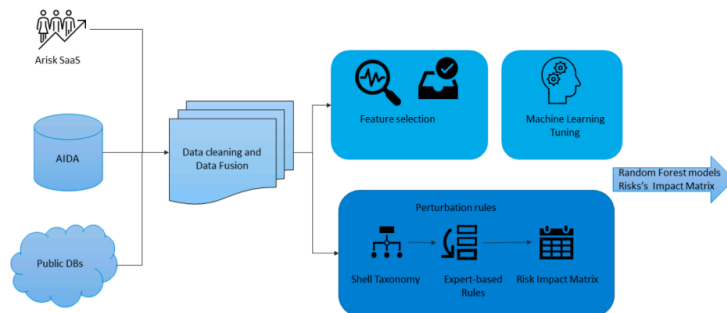
Varase hoiatuse teade sisaldab informatsiooni selle kohta, miks konkreetse taseme teade on väljastatud, mille tõttu süsteem arvab, et maksevõime on kahjustumas või kahjustunud. Teate sisu peab olema ettevõttele edasise tegevuskava aluseks, et oma maksevõimet parandada (s.t. et info maksevõime vähenemise asjaolude kohta peab olema esitatud piisava täpsusega, et sellest oleks kasu).

## 1.9. Andmete sobivuse kontroll

Varase hoiatuse süsteemi väljatöötamisel on oluline lähtuda, et aluseks võetakse kvaliteetsed andmed, võetakse kõige õigemad andmed ja kui andmeid omavahel seostatakse, et siis oleks tegemist mõistliku ja sobiva seostamisega (linkimisega). Praktilise poole pealt tähendab see seda, et tuleb sisendina kasutatavaid andmeid pidevalt üle vaadata ja proovida erinevaid sisendlahendusi (seeläbi hinnata andmete sobivust). Loodav prototüüp peab põhinema kõige ratsionaalsemal ja täpsemal sisendandmete kasutamisel, kus võimalusel on välistatud andmeliiasust ja „andmemüra“ tekitavaid parameetreid.

Perboli ja Arabnezhad (2021) rõhutavad, et masinõppest adekvaatse tulemi saamiseks on oluline eelnevalt valida õiged tunnused, mida analüüsida ning pakuvad selleks välja oma lähtekohad. Üks lähtekoht käesolevas töös andmete sobivuse kontrolliks on nimetatud käsitlus. Tsiteeritud allikas on vaid üks paljudest, mis andmete õige ettevalmistuse ja õigete analüüsitavate tunnuste vajalikkust rõhutavad.

Perboli ja Arabnezhad (2021) toovad oma käsitluses ära alljärgneva joonise, kus on olulist rõhku pandud tunnuste valikule ja andmete esmasele puhastamisele/linkimisele. Olulisel kohal on ka edasine machine learning tuning, ehk siis täiendavate parameetrite leidmine ja nende piirväärtuste paika panemine.



Joonis 2.: Sisendtunnuste valiku olulisus.

Ülaltoodud jooniselt on näha, et andmete linkimisel ja eelneval puhastamisel on väga suur roll, kuid täiendavalt sellele hakatakse veel valida, milliseid parameetreid sisse tuua ja milliseid mitte. Alles peale seda etappi on võimalik minna masinõppe faasi.

Perboli ja Arabnezhad (2021) poolt mudelisse sisse viidavad parameetrid on toodud alljärgnevas tabelis. Nagu tabelist näha on ühelt poolt rõhku pandud rahavoogudele (kasumlikule rahavoole) ja teiselt poolt võlgadele, mis ettevõttel on. Tootmisvõimsus on toodud täiendavalt sisse, kuid pigem lisaparaameeter.

Data set features.

Feature	Feature value	Feature type
ATT10	Absolute value	Revenue/profit
ATT11	Index/percentage (%)	Revenue/profit
ATT12	Absolute value	Revenue/profit
ATT13	Absolute value	Revenue/profit
ATT14	Index/percentage (%)	Revenue/profit
ATT15	Index/percentage (%)	Cost/debt
ATT16	Absolute value	Cost/debt
ATT17	Index/percentage (%)	Cost/debt
ATT18	Absolute value	Cost/debt
ATT19	Index/percentage (%)	Cost/debt
ATT20	Absolute value	Production
ATT21	Absolute value	Production
ATT22	Index/percentage (%)	Revenue/profit
ATT23	Absolute value	Production
ATT24	Index/percentage (%)	Cost/debt

Tabel 8.: Perboli ja Arabnezhad (2021) pakutud parameetrid.

Ülaltoodud lähenemine on andnud Perboli ja Arabnezhad (2021) mudelile võimaluse vaadata pankrotitõenäosust tegevusvaldkondade, käibe, asukoha, COVID-19, toetusmeetmete/laenu suhtes, vt. joonis allpool.

Italian companies bankruptcy with regard to the activity.

Risk of bankruptcy	Activity	Count	Prob < 50%	50% ≤ Prob < 70%	Prob ≥ 70%
Short-term	Industry	69,351	58%	18%	24%
	Commerce	46,524	56%	19%	25%
	Public	3,630	51%	23%	26%
	Service	42,097	47%	21%	32%
Middle-term	Industry	69,351	53%	19%	28%
	Commerce	46,524	51%	21%	28%
	Public	3,630	48%	23%	29%
	Service	42,097	43%	22%	35%
Long-term	Industry	69,351	52%	19%	29%
	Commerce	46,524	50%	22%	29%
	Public	3,630	46%	23%	31%
	Service	42,097	42%	22%	36%

Italian companies bankruptcy with regard to company revenues (millions of euros).

Risk of bankruptcy	Revenue	Count	Prob < 50%	50% ≤ Prob < 70%	Prob ≥ 70%
Short-term	< 5	127,009	51%	20%	29%
	5 ≤ X < 10	17,965	64%	18%	18%
	10 ≤ X < 15	6,519	65%	16%	19%
Middle-term	≥ 15	10,109	65%	16%	19%
	< 5	127,009	47%	21%	32%
	5 ≤ X < 10	17,965	60%	19%	21%
Long-term	10 ≤ X < 15	6,519	61%	17%	22%
	≥ 15	10,109	60%	18%	22%
	< 5	127,009	46%	22%	32%
	5 ≤ X < 10	17,965	60%	18%	22%
	10 ≤ X < 15	6,519	61%	17%	22%
	≥ 15	10,109	59%	18%	23%

Italian companies bankruptcy with regard to company location.

Risk of bankruptcy	Location	Count	Prob < 50%	50% ≤ Prob < 70%	Prob ≥ 70%
Short-term	Northeast	39,775	60%	18%	22%
	Northwest	53,045	57%	18%	25%
	Center	36,724	49%	20%	31%
	South	32,058	48%	22%	30%
Middle-term	Northeast	39,775	56%	19%	25%
	Northwest	53,045	53%	20%	27%
	Center	36,724	44%	22%	34%
	South	32,058	44%	23%	34%
Long-term	Northeast	39,775	55%	19%	26%
	Northwest	53,045	52%	20%	28%
	Center	36,724	43%	22%	35%
	South	32,058	42%	23%	35%

The bankruptcy of Piedmont companies pre- and post-COVID-19, as well as after the financial support policy (at 10%, 20%, or 30% of a company's past-year revenues).

Risk of bankruptcy	Prob. <50%	50% ≤ Prob. < 70%	Prob. ≥70%	Mean risk
Pre-COVID-19	70.7 %	28.7 %	0.6 %	29 %
Post-COVID-19	13.8 %	84.6 %	1.6 %	39 %
Loan 10% of revenues	15.8 %	82.7 %	1.5 %	40 %
Loan 20% of revenues	27.7 %	71.5 %	0.7 %	33 %
Loan 30% of revenues	20.1 %	79.0 %	0.9 %	38 %

Joonis 3.: Pankrotitõenäosused erinevate kontekstide lõikes.

Ülaltoodud tulemid on näiteks, millised võimalikud sisendid võivad andmete meetodite sobivuse kontekstis viia erinevatele tulemitel. Käesolevas töös klasterdatakse erinevaid andmeid erinevate

meetoditega ja otsitakse lõplikku optimaalset tulemit tagamaks maksevõimetuse hinnang erinevates kontekstides (sektorid) ja erinevate parameetrite (ettevõtet iseloomustavad näitajad) lõikes. Teiste riikide analüütilised kogemused on selles kontekstis väga olulised ja käesolevasse materjali lisatud eesmärgiga näidata, millest antud töös lähtutakse.

Andmete sobivuse kontrolli viiakse käesolevas töös läbi süsteemselt lähtuvalt, milliseid parameetreid valitakse ja kuidas need koos tööle hakkavad.

## 1.10. Sisendparameetrite kvaliteedi tõstmine

Vastavalt sellele, kuidas süsteem toodab esialgseid tulemusi, tuleb süsteemi sisse tuua täiendavaid parameetreid, võimalusel nn. täpsuskvaliteedi suurendamisele suunatud parameetreid.

Andrés et al. (2004) toovad välja mõned parameetrid, mis aitavad täpsemalt defineerida äritegevuse kvaliteeti, vt. joonis allpool.

Dimension	Variable	Code
Debt quality	$\frac{\text{Current Liabilities}}{\text{Total Debt}}$	V1
Indebtedness	$\frac{\text{Equity Capital}}{\text{Total Debt}}$	V2
Use of fixed capital	$\frac{\text{Tangible Fixed Assets} + \text{Intangible Fixed Assets}}{\text{Total Employment}}$	V3
Debt cost	$\frac{\text{Financial Expenses}}{\text{Total Debt}}$	V4
Short-term liquidity	$\frac{\text{Current Assets}}{\text{Current Debt}}$	V5
Share of labour costs	$\frac{\text{Labour Cost}}{\text{Added Value}}$	V6
Size	Net Sales (EUR thousands)	V7
Average sales per employee	$\frac{\text{Net Sales (EUR thousands)}}{\text{Total Employment}}$	V8

Joonis 4.: Andrés et al. (2004) analüütilist täpsust suurendavad parameetrid.

Ülaltoodud jooniselt oleks oluline näiteks tuua parameetrit V1, mis näitab võlgade kvaliteeti, ehk siis kui palju ollakse „küla peale niisama võlgu ja kiirlaenu võtnud“ (Current Liabilities) ning kui suure osa moodustab see siis koguvõlast (võla muuks komponentideks on pangast võetud kvaliteetlaenu). See parameeter aitab hinnata, kui juhuslik on laenuvõtmine ja kuivõrd laenu võivad ohtlikuks muutuda ühe või teise stsenaariumi korral. Käesolevas töös on seda juba arvestatud selliselt, et klasterdamises on võetud eraldi arvesse lühiajalised ja pikaajalised kohustused ja vaadatud vastavaid struktuuri suhtarve – seega esmane töö selles suunas, käib, kuid on vaja edasiselt „rafineerida“.

Oluline on siinkohal rõhutada ka parameetreid V1 ja V2, mis toovad sisse inimkapitali dimensiooni, näitavad kui palju on kaasatud töötajaid (V6) ja kuivõrd see kaasatud töötajaskond on tegelikult tootlik (V8).

Aliaj et al. (2020) käsitus on heaks näiteks, kuidas on võimalik sisse tuua täiendavalt täpsustavaid kvaliteetparameetreid. Nad käsitlevad teemat küll panga bilansi keskselt, seega ei saa sealt otseselt midagi „kopeerida“, kuid oluline on põhimõte ja lähtenurk (vt. alltoodud joonis).



ID	Description	ID	Description
L1	Granted amount of loans	B1	Revenues
L2	Used amount of loans	B2	ROE
L3	Bank's classification of firm	B3	ROA
L4	Average amount of loan used	B5	Total turnover
L5	Overdraft	B6	Total assets
L6	Margins	B7	Financial charges/operating margin
L7	Past due (loans not returned after the deadline)	B8	EBITDA
L8	Amount of problematic loans		
L9	Amount of non-performing loans		
L10	Amount of loans protected by a collateral		
L11	Value of the protection		
L12	Amount of forborne credit		

Table 1. Main attributes for the loan (L) and the balance-sheet (B) datasets.

Joonis 5.: Aliaj et al. (2020) kvaliteetparameetrite võimalused.

Näiteks parameetrid L3 (kuidas ettevõtte on klassifitseeritud), L4 (kui palju laenu on tegelikult ära kasutatud), L6 (marginaalid), L7 (varasemad võlgnevused). Tegemist on ühelt poolt tehniliste, kuid teisalt kvaliteetiparameetritega (näiteks varasemad võlgnevused ei ole tänased tehnilised näitajad, kuid on hinnangut loovad näitajad).

### 1.11. Täiendavate parameetrite leidmine

Rafineerimaks käsitlust on vaja sisse tuua täiendavaid parameetreid, mis midagi olulist juurde annaksid. Täiendavate parameetrite leidmiseks on kaks võimalust:

- tehtud töö tulemusi vaadates on võimalik järeltada, kus suunas oleks mõistlik edasi minna ja mida lisada;
- kasutada unsupervised learning (juhendamata) masinõppe meetodit, kus masin ilma täiendavate juhusteta vaatab, milliseid parameetreid võiks lisada.

Choudhry (2018) näitab ära, kuidas juhendatud (supervised) ja juhendamata (unsupervised) masinõppe on kasutatava ja milliseid tehnikaid selle jaoks sobivad on, vt. alltoodud joonis.

Exhibit 2 Machine learning problem types classification

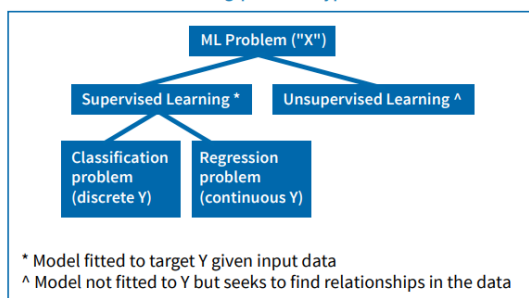


Exhibit 3 Machine learning models

Type	Techniques
Supervised Learning (Target Y given)	Decision trees Artificial neural networks Support vector machines k-nearest neighbours (non-parametric) Random forests (non-parametric) Naïve Bayes
Unsupervised Learning	k-means clustering Hierarchical clustering analysis
Recommender systems	Blend of different methodologies

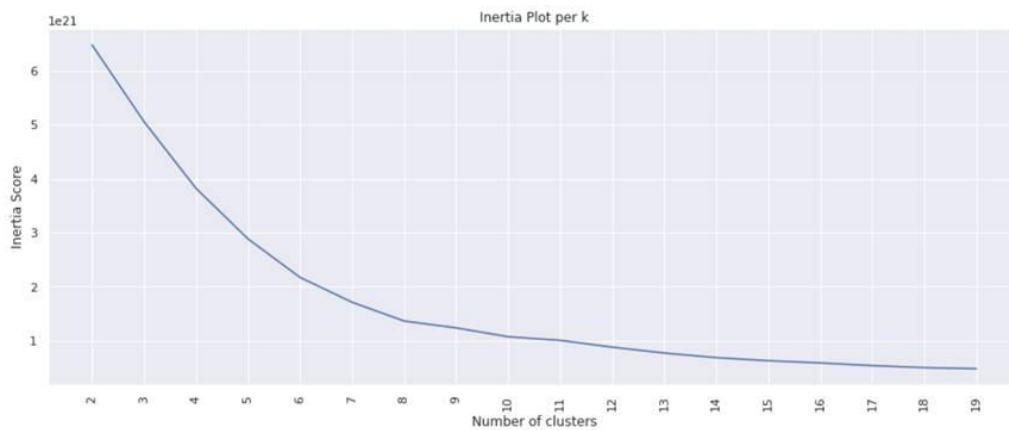
Joonis 6.: Choudhry (2018) põhjal masinõppe rakendamine.

Ülaltoodud joonise vasakpoolisel pildil on näha, milliseid meetodeid soovitatakse kasutada ühe või teise masinõppe puhul. Käesoleva töö juures on kasutatud k-means tehnikat supervised learning kontekstis, mida joonis ei paku. Samas joonis ei ole absoluutne juhtnõr, vaid pigem üks võimalik lahendus.

### 1.12. Parameetrite paljusus

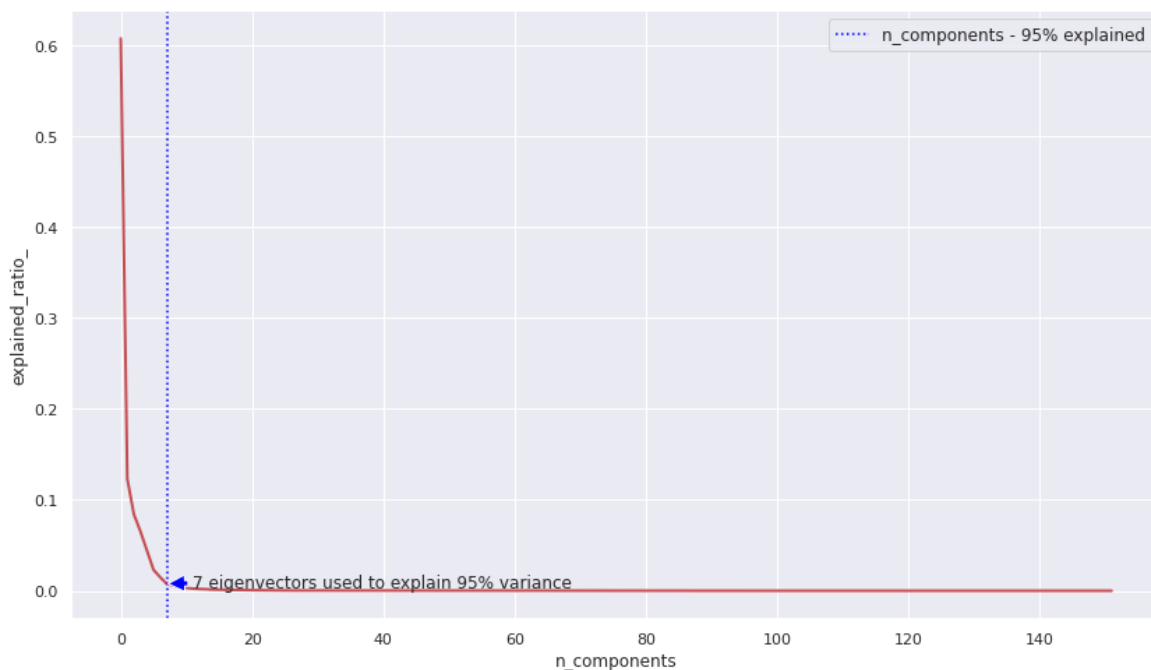
Täiendavate parameetrite lisamisel on oluline õigesti määrata klastrite arv (mille kohta ka ülal on üks selgitus juba antud). Rohkem parameetreid võib tekitada soovi määratleda suurem arv klastreid.

Samas on oluline silmas pidada asjaolu, et klasteritele tuleb anda sisu ja kui klasterid on liiga teineteise sarnased, siis on see raske. Antud juhul on ka suurema arvu parameetrite juures mindud 5 klasteriga



Joonis 7.: Klasterite arvu määramine.

Käesoleva töö juures ollakse jõutud seisu, kus bilansi andmestikus on suurusjärgus 150 erinevat tunnust. Tunnuste suur arv muudab arvutusi keerukaks. Probleemi lahendamiseks on kasutatud PCA (Principal Component Analysis) meetodikat – dimensioonide alandamise tehnika, mis seob korrelatsioonis olevad muutujad väikese arvu mittekorreleeruvate muutujate (peamiste muutujate) külge. Tulemuseks saadakse kompressitud väikesest arvust andmetest kätte maksimaalne hulk informatsiooni. Antud töös on senimaani tehtu põhjal kompressitud muutujate arv 7 muutuja peale, nagu näidatud alltoodud joonisel.



Joonis 8.: Muutujate arvu taandamine seitsme põhiparameetri peale.

PCA tehnikaga saadud seitse parameetrit on muudetud kahedimensionaalseks jooniseks, esitatud lisas 6. Kahedimensionaalseks viimine on vajalik seetõttu, et seeläbi on võimalik seda joonisel kujutada, seitsmemõõtmelist joonist teha ei saa. Telgedel kujutatud parameetrid on matemaatilised,

nende sisu ei ole majanduslikus mõttes otseselt tõlgendatav. Samas saab seeläbi määratleda klastrid ja klastrite sisu analüüsid (igas klastris paiknevad lähedaseks tunnistatud ettevõtete/ettevõttekirjete kogumid) on võimalik anda neile majanduslik kirjeldus, nagu on kirjeldus antud varemalt toodud tabelis 7.

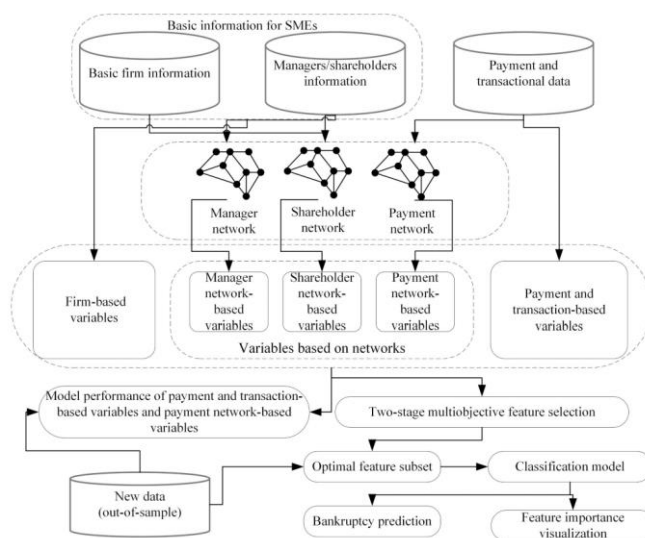
Nagu selgitatud, on PCA tehnika kasutamise juures oluline korrelatsioon. Käesoleva töö käigus koostatud korrelatsioonimaatriks on esitatud lisan 7. Korrelatsioonimaatriksis on toodud vaid olulisemad muutujad, ei ole esitatud kõiki 150 muutujat, mida bilansiandmestikust on võimalik välja võtta. Maatriks on mõeldud illustatsiooniks antud juhul.

### 1.13. Hinnang ettevõtete vaheliste arvelduste põhjal

Ülalkäsitletu on lahenduste kogum, mis põhineb klassikalisel bilansi ja kasumiaruande ning seotud andmete analüüsil. Teine suund (nagu ka varasemalt mainitud) on anda hinnanguid ettevõtte maksevõimele ainuüksi ettevõtete vaheliste arveldusandmete põhjale (tegelikult kombineeritakse antud töös mõlemat suunda, kui lähtetöötlus on erinev). Käesolevas töös arvelduste suunaga alles hakatakse andmepõhiselt tööle, seega praeguse seisuga tehnilisi tulemeid veel ei ole.

Kou et al. (2021) pakuvad välja pankroti ennustusmudelit väikese ja keskmise suurusega ettevõtetele, mis põhineb nende ettevõtete vaheliste arvelduste andmetel. Tegemist on alternatiivse lahendusega nendele mudelitele, mis põhinevad ettevõtte raamatupidamislike andmete (bilansi, kasumiaruande) majandussuhtarvude analüüsil ja jätab nn. klassikalise suhtarvude analüüsi üldse kõrvale. Metoodika autorid väidavad, et erinevalt raamatupidamislikest aruannetest on ettevõtete vaheline tehinguinfo usaldusväärsem.

Kou et al. (2021) illustreerivad oma lähenemist järgmisel algoritmilisel skeemil, vt. alljärgnev joonis. Käesolevas töös võetakse esitatud arveldustel põhinev mudel aluseks, kuid ei kopeerita seda detailselt.



Joonis 9.: Kou et al. (2021) arveldustel põhinev mudel.

Ainuüksi arvelduste analüüsil põhinev maksevõime ennustamise metoodika raskuskohaks on see, et kuidas üksnes rahaülekannete põhjal saada piisaval arvul muutujaid, et sellele saaks masinõpet rajada. Kou et al. (2021) lahendus, nagu alljärgnevalt illustreeritud, on kasutada nii käibeid kui hinnangulisi tehingukasumeid lugedes nii transaktsioonide arvu kui ka väärtust, vaadates seda erinevate päevade intervallide kaupa ning välise muutujaga tuues sisse volatiilsust (riski).

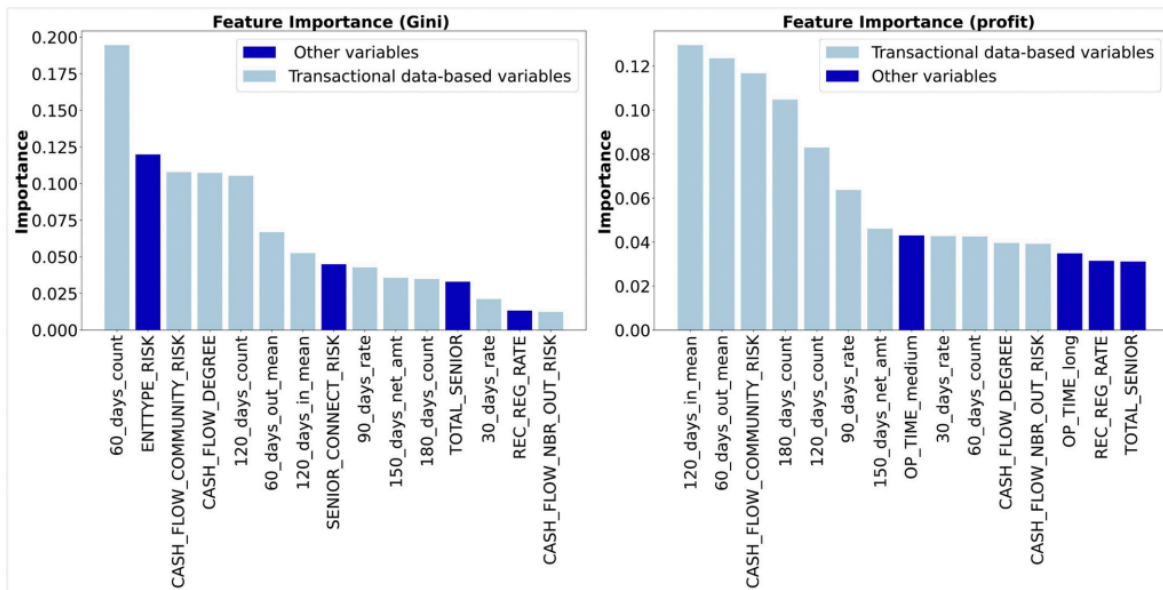


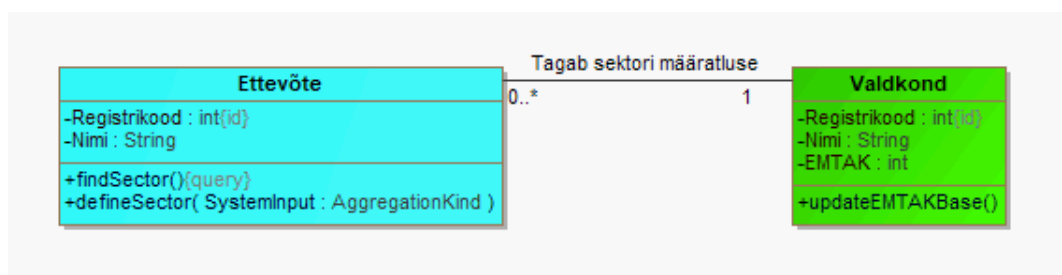
Fig. 6. Feature importance for features in an optimal feature subset.

Joonis 10.: Kou et al. (2021) lahendus transaktsioonimuutujatele.

## 1.14. Andmed

Käesolevas töös kasutatakse peamiselt ettevõtete majandusaasta aruannete andmeid (bilansi ja kasumiaruande analüüsiks, üldiselt võimalusel viie aasta lõikes), neid toetatakse käibedeklaratsioonide KMD ja KMD INF andmetega (transaktsioonandmete saamiseks). Täiendavalt tuuakse sisse tulu- ja sotsiaalmaksu deklaratsioonide TSD andmed, mida toetatakse töötamise registri (TÖR) andmetega (andmekvaliteedi tõstmiseks, vt. joonis 4 ülal). Kasutatakse pankrotistunud ettevõtete majandusaasta aruannete andmeid (viis aastat enne pankroti väljakuulutamist). Võimalusel (töö selle võimaluse loomiseks käib) tuuakse sisse maksuvõlgade andmed. Siinkohal toodud loetalu ei pruugi olla lõplik, töö käigus andmefookus pidevalt täpsustub.

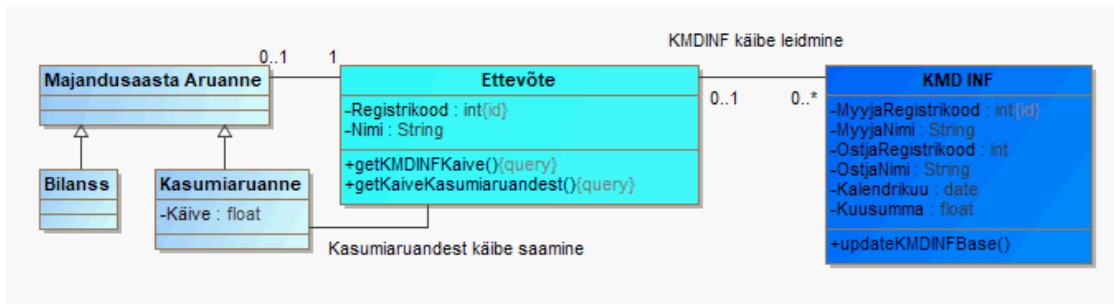
Sektoriaalne paiknemine määratakse EMTAK baasi alusel, nagu toodud alljärgneval klassidiagrammil:



Joonis 11.: Tegevussektorite määramine.

Ülaltoodud jooniselt on näha, kuidas tegevusvaldkond määratletakse EMTAK koodi kasutamise abil. Nagu varemalt käesolevas dokumendis selgitatud ja lisa 1 näidatud, valitakse välja 5-kohalised EMTAK koodid, millest kombineeritakse tootmisharud/sectordi. Seda teeb protseduur +updateEMTAKBase().

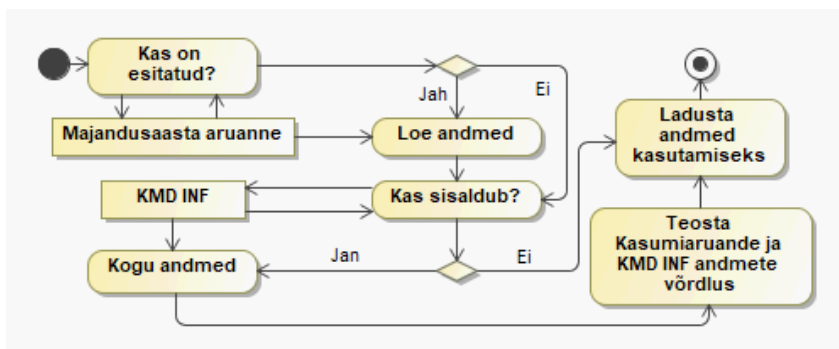
Transaktsioonandmete kasutamisel võetakse KMD INF ja KMD andmetele lisa ka ettevõtte majandusaasta aruannetest, nagu toodud alljärgneval klassidiagrammil. Oluline on võrrelda käivate vastavust, saada siduda pikaajalist vaadet (majandusaasta aruanded) lühiajalise operatiivandmestikuga (KMD ja KMD INF on igakuiselt saadaval).



Joonis 12.: KMD INF andmete sidumine majandusaasta aruannete andmetega.

Ülaltoodud jooniselt on näha, kuidas ühelt poolt kasutatakse käivet KMD INF andmebaasist (päring +getKMDINFKaive()) ja teiselt poolt kasutatakse käivet ettevõtte majandusaasta aruande kasumiaruandest (+getKaiveKasumiaruandest()). Ülaltoodud joonise kontekstis ei ole see oluline, kuid tegelikult toetatakse KMD INF andmeid ka KMD deklaratsioonide andmetega.

Protsessiskeem KMD INF andmete ja majandusaasta aruannete andmete sidumiseks.



Joonis 13.: Protsessiskeem majandusaasta aruannete sidumiseks KMD INF andmetega.

Ülaltoodud protsessiskeem kinnitab eelöeldut, kuid olulise rõhuasetusena tuleks tähelepanu pöörata sellele, et teostatakse KMD INF ja kasumiaruande andmete võrdlus.

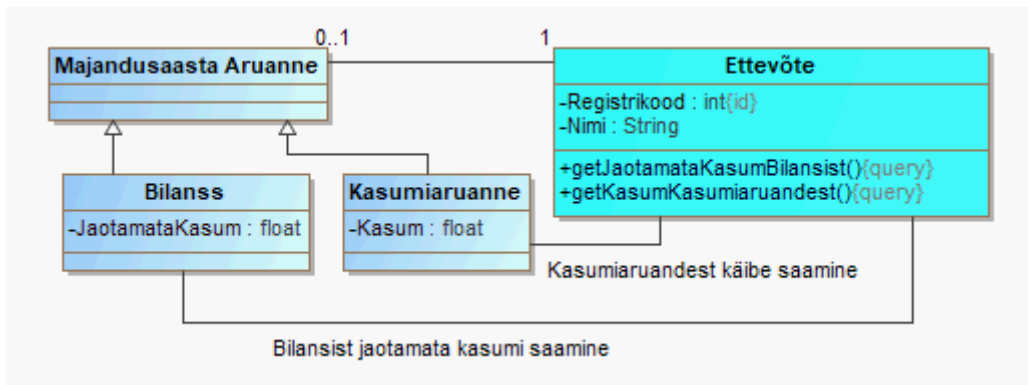
Alljärgnevalt ülaltoodud skeemi algoritmiline kirjeldus.

```

While Ettevõtte.Registrikood ≠ 0 do
  For every Ettevõtte.Registrikood do
    IF Ettevõtte.Registrikood contains in Majandusaasta Aruanne
      then get Kasumiaruanne.Käive
    else
      Kasumiaruanne.Käive = Missing
    end
    IF Ettevõtte.Registrikood ∈ KMD INF
      then getKMDINFKaive()
    else
      getKMDINFKaive() = Missing
    end
    compare getKMDINFKaive() to Kasumiaruanne.Käive
  end
end

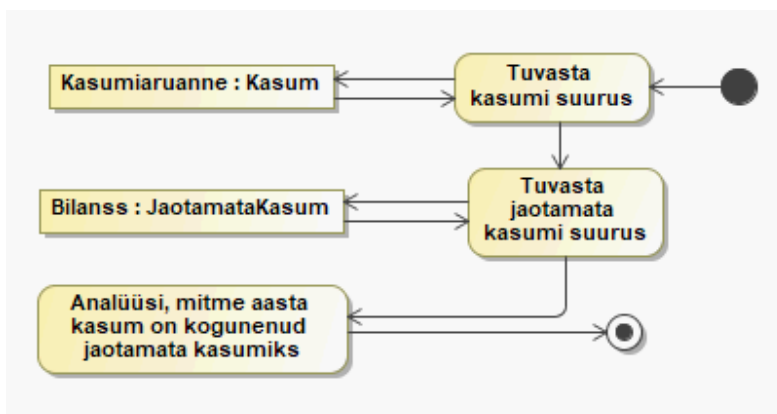
```

Majandusaasta aruande kaudu võetakse arvesse ka ettevõtte kasumit.



Joonis 14.: Kasum majandusaasta aruandest.

Protsessiskeemi põhiselt vaadatakse kasumit jooksvalt majandusaasta aruannetest, kuid tuuakse välja ka täiendavalt jaotamata kasum – see võimaldab saada täiendavat ajalist perspektiivi sellele, kuidas kasum on tekkinud ja milline on kasumi tekitamis dünaamika (arusaadavalt võetakse arvesse ka dividendide maksmist).



Joonis 15.: Kasumi analüüsimise protsessiskeem.

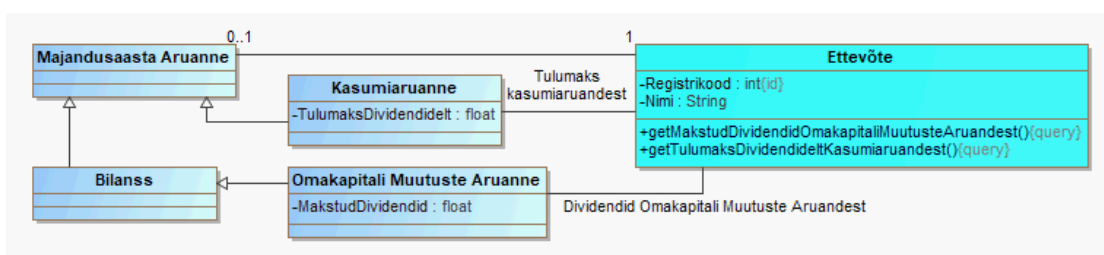
Kasumi analüüsimise algoritmiline skeem:

```

For every Ettevõte.Registrikood do
  get Kasumiaruanne.Kasum
  get Bilanss.JaotamataKasum
  this.KasumlikudAastad=Bilanss.JaotamataKasum/Kasumiaruanne.Kasum
  KasumiAnalüüs = {{KasumlikudAastad},{Bilanss.JaotamataKasum},
                  {Kasumiaruanne.Kasum}}
end

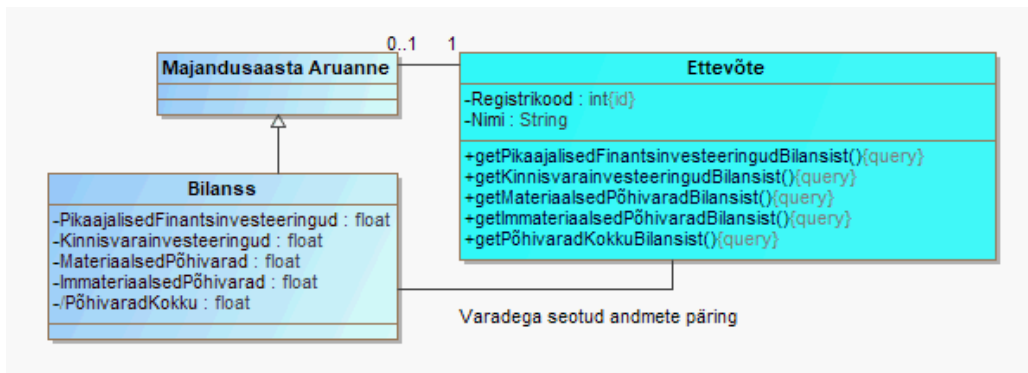
```

Olulise sisendina analüüsitakse ka omakapitali muutust:



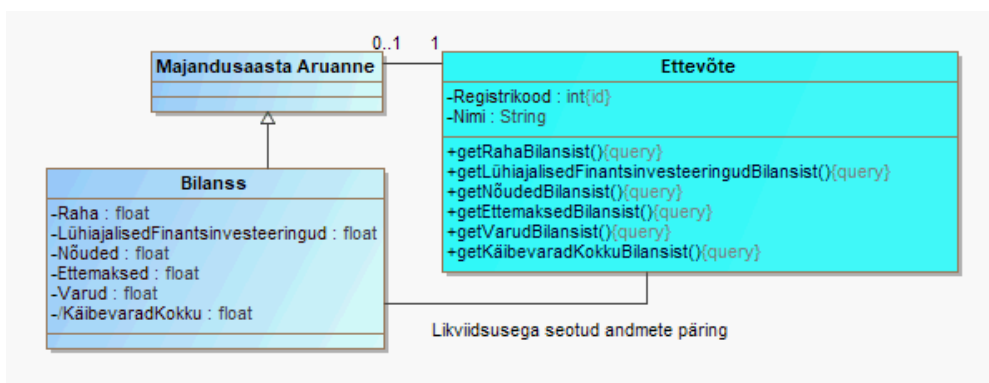
Joonis 16.: Omakapitali muutuse analüüsimine.

Põhjalikuma analüüsi eesmärgil on vajalik anda ka hinnang ettevõtete varade olemusele, mida on kirjeldatud alljärgneval klassidiagrammil. Varade kvaliteet võib avaldada olulist mõju sellele, kas ja kui tugevalt suudab ettevõtte võimaliku maksejõuetuse vastu seista. Näiteks osad kinnisvarainvesteeringud võivad olla väga likviidsed.



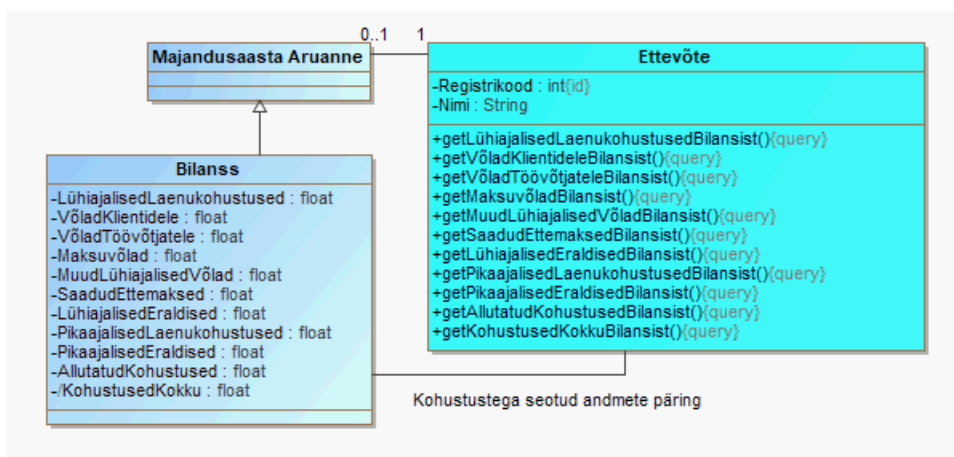
Joonis 17.: Varade kvaliteedi analüüs.

Likviidsus on väga oluline maksevõimekuse ja makseraskuste kontekstis.



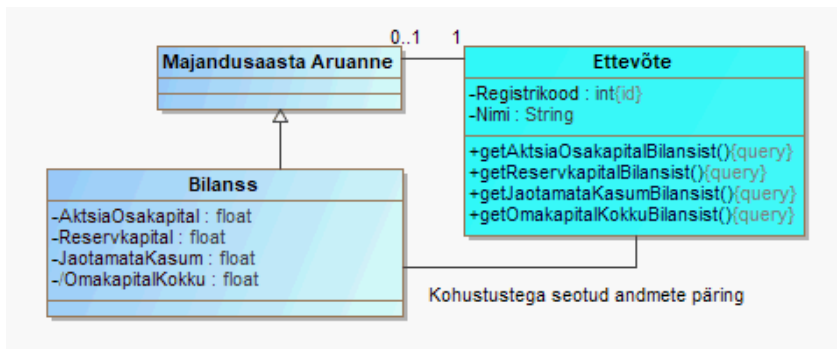
Joonis 18.: Ettevõtte likviidsuse andmed.

Maksevõimekuse seisukohalt on oluline kohustuste analüüs, mis on kirjeldatud alljärgneval klassidiagrammil.



Joonis 19.: Ettevõtte kohustuste analüüs.

Omakapitali analüüs viiakse läbi ka komponentide kaupa (mitte ainult muutustena, nagu ülal toodud). Omakapitali analüüs on vajalik selleks, et saada paremini aru ettevõtte arengudünaamikast, tegevusest nn. ülejääv raha koondub omakapitali alla.

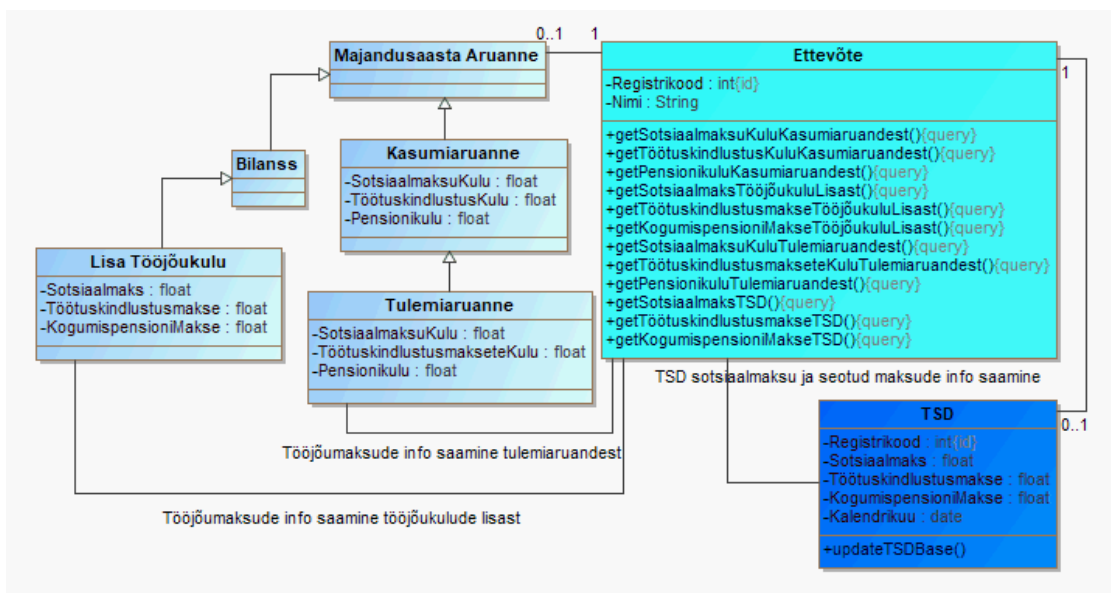


Joonis 20.: Ettevõtte omakapitali analüüs.

Nagu ülal toodud jooniselt nähtub vaadeldakse omakapitali analüüsi läbi aktsia/osakapitali, reservkapitali ja jaotamata kasumi. Jaotamata kasumi reale (dividendide maksmist või mittemaksmist tuleb jälgida) koguneb majandustegevuse tule (positiivne tule) ja selle tulemi kogunemist jälgides on võimalik aru saada tegevusdünaamikast.

Võttes arvesse et majandusaasta aruanded tekivad teatud perioodi tagant, siis operatiivsema ülevaate saamiseks (teades varasemaid dünaamikaid) on jaotamata kasumi teket võimalik modelleerida ka deklaratsioonidest pärit igakuiste andmete pealt.

Tööjõuandmestiku sissetoomine kujutatud alljärgneval klassidiagrammil.

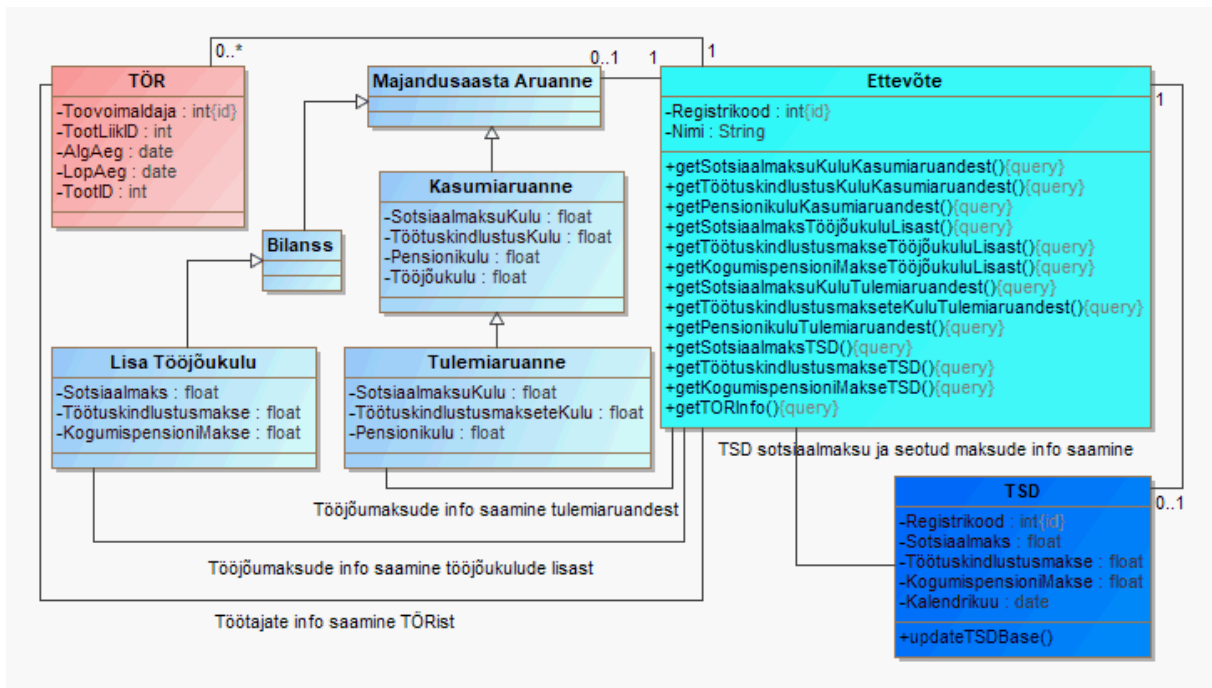


Joonis 21.: Tööjõukulude analüüsi sisend.

Nagu ülal toodud jooniselt nähtub on tööjõukulude peamiseks allikaks majandusaasta aruande bilansi tööjõukulude lisa, kasumiaruande tulemiaruanne ja deklaratsioonide poole pealt TSD deklaratsioon. TSD deklaratsioonide põhjal saab operatiivsemas ajalises mõtmes toetada teatud pikema aja tagant tekkivat majandusaasta aruande informatsiooni.

Alljärgneval klassidiagrammil on välja toodud TÖR andmete sissetoomine andmekvaliteedi tõstmise eesmärgil tööjõuandmete lisamiseks.





Joonis 22.: Töötamise registri andmete kaasamine.

Ülaltoodud klassidiagrammidest on esitatud need, mis praeguses töö faasis on valmis ja millele rajatult on analüüsi läbi viidud või on läbi viimisel. Klassidiagramme (nagu ka protsessiskeeme) lisandub töö käigus juurde ning ühtlasi need täpsustuvad. Ülaltoodu on esitatud eesmärgiga anda võimalikult parimat pilti töö hetkeseisust.

## Peatükk II – praktiline teostus

II peatükk esitab käesoleva töö raames läbi viidud praktilise teostusega seotud asjaolud. Tuuakse välja analüütiline loogika, millel praktiline teostus põhineb, tuuakse näiteid tulemitest ja koodidest/väljunditest. Kõiki läbitud teste, kõiki mudeleid ja klasterdatud klastreid ei esitata, välja tuuakse vaid olulisemat illustratiivset momenti omavad näited. Samas on näited ja esitatud materjal valitud selliselt, et see annaks sisulise ülevaate toimuvast ja töösuundadest.

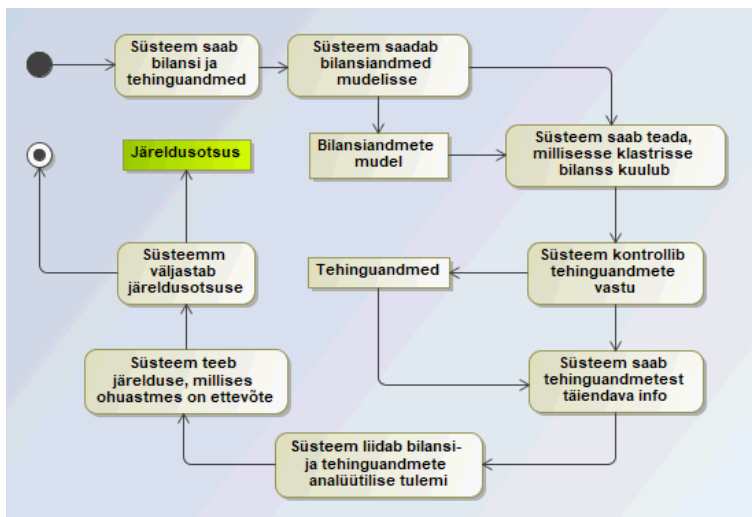
Praktilises teostuses on olulist rõhku pandud klasterdamisele, mis on sisuliselt juhendamata masinõpe (unsupervised learning). Seda on tehtud seetõttu, et lasta masinõppe algoritmidel otsida mustreid, mis võivad viidata maksejõuetuse saabumisele ja üleüldse seostele, mis maksejõuetuse elementidega seonduvad.

Klasterdatud andmete tulemite pealt on loodud mudelid, õpetatud masinat erinevaid mustreid tundma ja selle põhjal järeldusi tegema. Kaasatud on peamiselt bilansinäitajaid, nende näitajate suhtarve ja ettevõtete vaheliste transaktsioonide andmeid. Mudelite edasisel arendamisel plaanitakse kaasata (käesolevas dokumendis veel ei sisaldu) tööjõuandmeid, üldist majandustausta määratlevaid andmestikke.

Peatükki lugedes on oluline silmas pidada, et tegemist on töös oleva produktiga, mis areneb edasi, mille osas on esitatud praeguseni teada olevad lähtekohad.

## 2.1. Süsteemi ülesehituse põhimõtted

Alljärgnevalt on toodud süsteemi ülesehituse põhimõtteline protsessiskeem.



Joonis 2.1.a.: Süsteemi ülesehituse protsessiskeem.

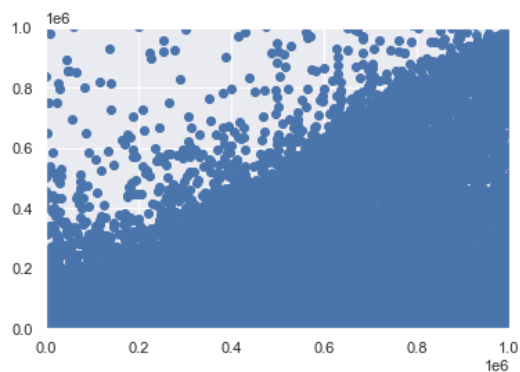
Ülaltoodu näitab, kuidas süsteem peab olema üles ehitatud, põhimõttelise loogika kohaselt. Käesolevas materjalis on lahti selgitatud ja illustreeritud ka süsteemi loomise protsess, sest tegemist on prototüübi loomisega ja prototüüp peab olema võimalikult avatud analüüsiks. Seetõttu on esitatud prototüübi loomise kava, astutud sammud ja nende sammude realiseerimise käigus õpitu (saadud tagasilöögid, tehtud järeldused, sisse viidud parandused).

Süsteem on prototüübi tasandil määratud eksisteerima töötava masinõppe algoritmina, s.t. mitte omama kasutajaliidest välisele kasutajale. Samas töö käigus, võimaldamaks erinevaid testimisi, on hakanud tekkima seisukoht, et veebipõhine kasutajaliides aitaks süsteemi testida ja võimaldada huvitatud isikutel süsteemi paremini tunnetama hakata.

## 2.2. Bilansiandmete analüüs

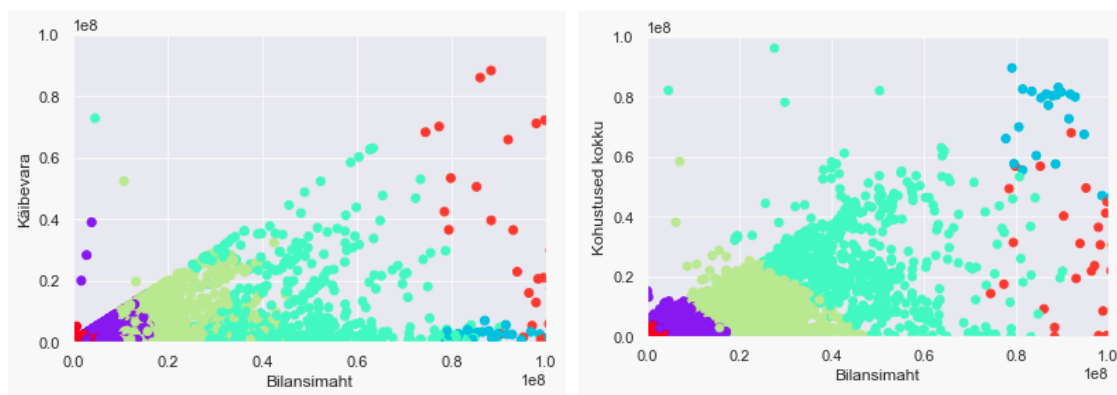
### 2.2.1. Bilansiandmete klasterdamine

Alljärgneval joonisel toodud andmed bilansimahtu (x) ja kohustuste kokku (y) kontekstis. Joonis on hetkel piiritletud 1 miljoniga.

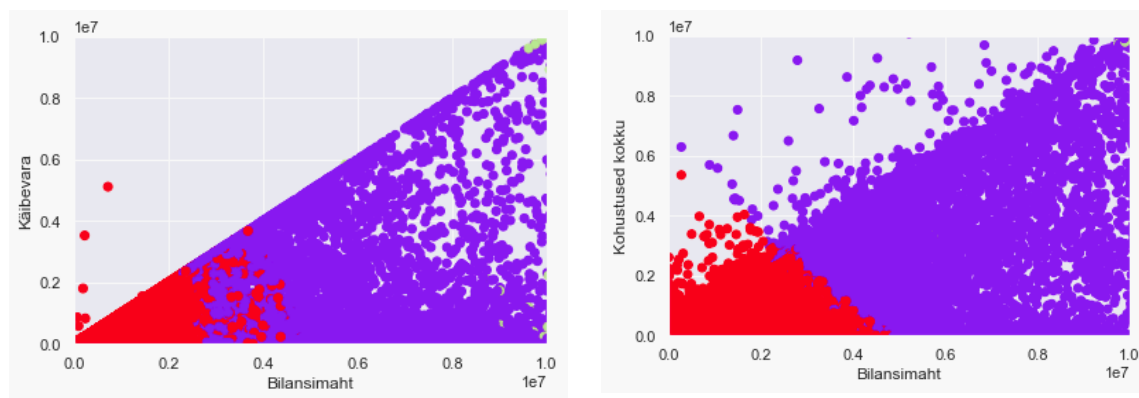


Joonis 2.2.1.a.: Andmed bilanss (x) v. kohustused kokku (y).

Alljärgnevatel joonistel on toodud bilansiandmete klasterdustulemused esitatuna bilansimaht v. kohustused kokku ja bilansimaht v. käibevara. Joonised on piiritletud 100 miljoni ja 10 miljoni. Seega ülaltoodud joonised on suuremad (suuremate summade) vaated alumistest.



Joonis 2.2.1.b.: Bilansiandmete klasterdamine (100 milj. joonise piir).



Joonis 2.2.1.c.: Bilansiandmete klasterdamine (10 milj. joonise piir).

Ülaltoodud jooniselt on näha, et klastrid on moodustunud selgelt suuruse järgi (kõiki parameetreid arvestades, kuid suurus on peamine).

Ülaltoodud analüüsi oli bilansikirjete osas lisatud järgmised:

- käibevara;
- lühiajalised laenud kokku;
- lühiajalised kohustused kokku;
- pikaajalised kohustused kokku;
- kohustused kokku;
- bilanss kokku.

Selline esmane valik tehti seoses sellega, et:

(a) bilansi maht annab esmase ülevaate ettevõtte suurusest, mis mingil määral on tõlgendatav ka turujõuna (sõltub kaasnevatest asjaoludest arusaadavalt); suurus on korrelatsioonis ka teatud inertsusega (suuremat laeva on raskem kõigutada, samas kui suur uppumas on, siis päästa ka ei ole enam midagi);

(b) kohustused näitavad, kui palju on ettevõtte võtnud laene, kui palju see ettevõtte tegelikult üldse omanikele kuulub; laenud võivad olla erineva sisuga ja nn. „ohuastmega“, kuid üldiselt näitab nende osakaal seda, kui võrd elatakse võlgu, laenude dünaamika näitab, kas võlgu elamine on suurenenud või vähenenud (võlgu elamise suurenemine võib näidata toimetulematust omal jõul – kreditorid peavad aidata elus olla);

(c) lühiajaliste ja pikaajaliste kohustuste lahtilöömine näitab, kui võrd on ettevõtte võlgadest tulenevalt lähiperspektiivi ohus (pikaajalised kohustused on pigem stabiilsed, lühiajalised võivad olla kiirelt sissenõutavad, võivad mõjutada ettevõtet lähiperspektiivis); pikaajalised kohustused (eelkõige siis pikaajalised laenud, kuid pikaajaliste kohustuste sisu on peamiselt pikaajalised laenud) näitavad, kui võrd pikaajalise raha laenajad (pangad eelkõige) usaldavad ettevõtet, see usaldus on analüüsi tulemus; lühiajaliste-pikaajaliste kohustuste vahekord näitab muuhulgas stabiilsust pikemas perspektiivis;

(d) lühiajalised laenud näitavad seda osa lühiajalistest kohustustest, mis on võetud finantseerimaks jooksvat tegevust, peamiselt puuduva käibekapitali katteks; kui ettevõtte peab võtma lühiajalisi laene (tavaelus on teatud võrdlusmoment võimalik tuua kiirlaenudega), siis see on viide võimaliku maksejõuetuse suunal; süsteem jälgib lühiajaliste laenude osakaalu lühiajalistes kohustustes nagu ka nende laenude suurust ja muutumist ajas;

(e) käibevara on likviidsed vahendid, mille abil saab ettevõtte ennast vajaduse korral finantseerida (mida ka kasutatakse jooksvaks rahastamiseks); ettevõtte võib olla suur ja tugeval turupositsioonil, kuid kui tal ei ole jooksvaid käibevahendeid, siis võib ta vaatamata kõigele sattuda makseraskustesse.

Ülaltoodud näitajate põhjal tehtud klasterdamine näitas (eelnevalt esitatud illustratsioonid), et masin oskab eristada erinevaid ettevõtteid, jõuda järeldusele, kuhu klastrisse üks või teine ettevõtte kuulub, kuid eelkõige tekkisid klastrid suuruse põhjal. Suurtel ja väikestel on arusaadavalt erinev majandusdünaamika ja tulenevalt sellest on suuruse põhjal eristamine põhjendatud, kuid adekvaatsema hinnangu saamiseks on vaja täpsemat lähtenurka.

Tulenevalt ülaltoodust toodi bilansiandmete juurde sisse järgmised näitajad:

- raha;
- ettemaksed;
- põhivara.

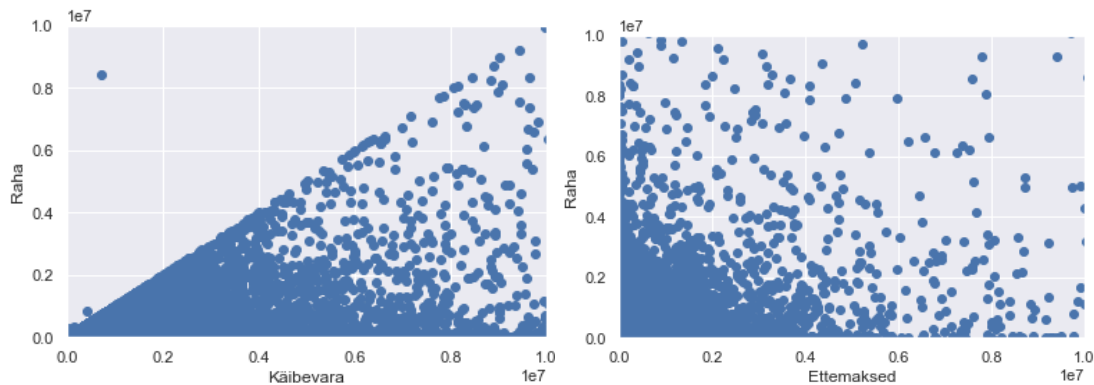
Täiendavalt sissetoodud näitajad aitavad klasterdustulemusi täpsustada tulenevalt alljärgnevast loogikast:

(a) raha näitab sisuliselt seda võimekust, kui millisel määral suudab ettevõtte täna oma kohustusi katta; tegemist on piltlikult nagu sularahaga taskus, mille eest jooksvalt midagi maksta – tegemist on käibevara täiendava täpsustamisega;

(b) ettemakse on ettevõtte vara, ühelt poolt näitab ettemaksete suurus, kui palju on kohustusi ette makstud ja seega on võimekus hinnata kohustuste reaalsust (ettemaksete tõttu on need vahest kunstlikult väiksemad), teiselt poolt näitab see ka ettevõtte finantsvõimekust (nn. „küla peal“ olev raha on ka ettevõtte võimekuse üks osa);

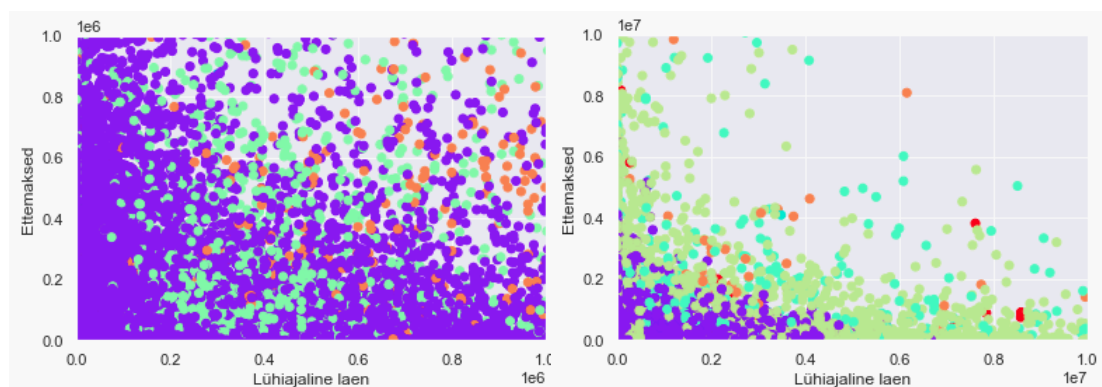
(c) põhivara näitab analüüsis seda, kuidas ettevõtte on oma vahendeid paigutanud, kuidas oma tegevuse organiseerinud; põhivara vähesus näitab teatud oludes ettevõtte ebastabiilsust (kogu kulutatud raha läheb väikevahenditesse, s.t. otse kuludesse, suuri seadmeid ei osteta, tootmisinvesteeringuid ei tehta); oluline on põhivara dünaamika ajas (põhivara kahanemine võib näidata ettevõtte hääbumist või suutmatust oma tehnoloogiat uuendada – hakatakse pooltooteid sisse ostma, ise neid teha ei suudeta).

Alljärgnevalt on toodud täiendavate bilansiparameetritega klasterdatud bilansiandmete tulemuste illustreerimised.



Joonis 2.2.1.d.: Raha v. käibevahendid ning raha v. ettemaksed.

Ülaltoodud joonistelt on näha, et kellel on rohkem käibevara, sellel (väiksema bilansimahu juures) on ka rohkem raha. Suuremate ettevõtete juures tekib seos, et rohkem käibevara ei tähenda enam, et käibevara oleks suurel määral rahas, kuid samas leidub ka neid, kelle jaoks käibevara just raha tähendabki.

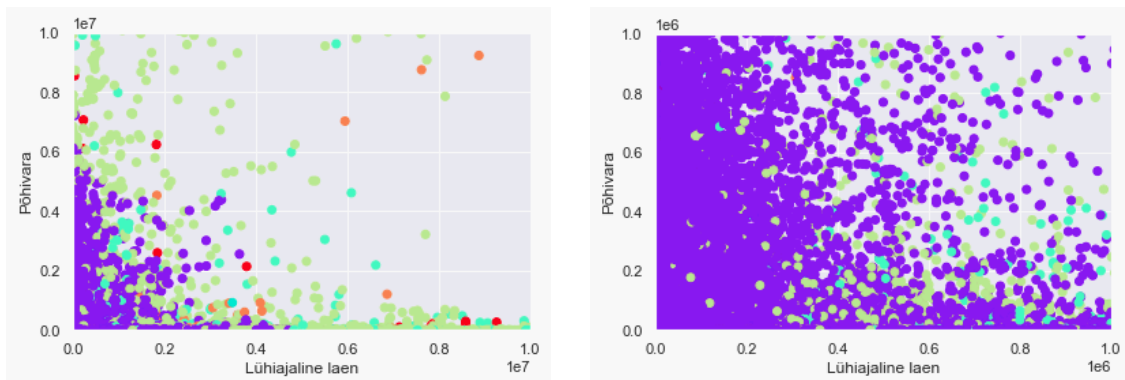


Joonis 2.2.1.e.: Bilansiandmete klasterdamine ettemaksed v. lühiajaline laen.

Ülaltoodud joonisel (klasterdustulemused) on selgelt näha muster, et kui sul on rohkem lühiajalist laenu, siis sul on vähem ettemakseid ja kui sa teed ettemakseid, siis mitte lühiajalisest laenust.

Süsteem toob välja erinevaid klastreid, kuid antud muster on oluline ja paljuütleval (samas väga loogiline) selleks, et masin saaks hinnata lühiajalise laenu ja ettemaksete mõju ettevõtte maksevõimelisusele.

Ülaltoodud kaks joonist on ühe ja sama joonise väljavõtte erinevates skaalade mõõtkavades.



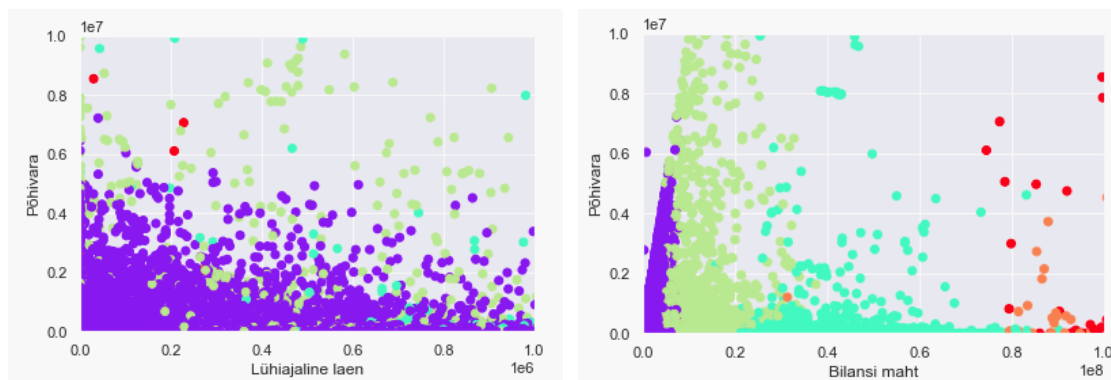
Joonis 2.2.1.f.: Bilansiandmete klasterdamine põhivara v. lühiajaline laen.

Ülaltoodud joonisel on selgelt näha, et ka lühiajalise laenu ja põhivara vahelised seosed, näha on, et kellel on rohkem lühiajalist laenu, sellel on vähem põhivara:

(a) lühiajalise laenu sõltuvuses olevad ettevõtte (a la kiirlaenusõitlased) ei oma üldse majanduslikku võimekust endale arvestataval määral põhivara soetada;

(b) ettevõtte, kes on sattunud lühiajalise laenu sõltuvusse, on oma põhivara realiseerinud; selle variandi kohaselt on eriti oluline asjaolu, et süsteem jälgiks, milline on põhivarade dünaamika olnud aastate lõikes.

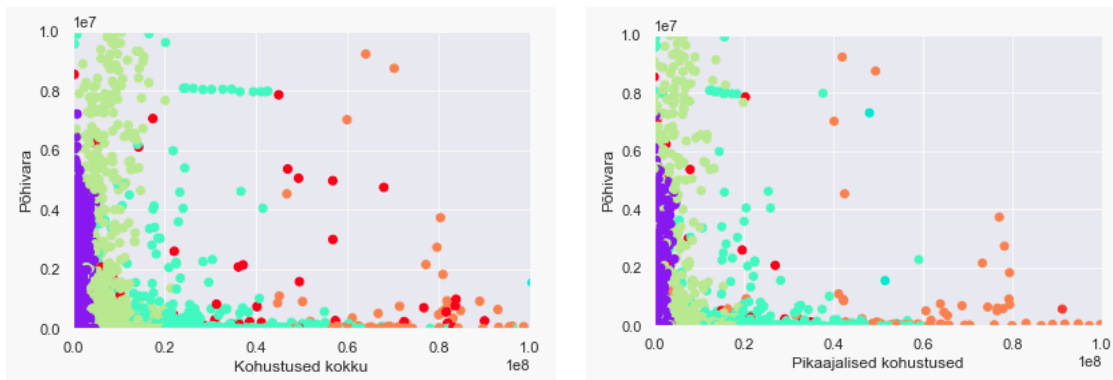
Arvestades, et põhivara olemasolu ja selle soetamise (likvideerimise) dünaamika näitab ettevõtte majanduslikku vundamenti, on põhivara vaadeldud ka teiste näitajate vastu.



Joonis 2.2.1.g.: Bilansiandmete klasterdamine põhivara v. lühiajaline laen, v. bilansi maht.

Ülaltoodud joonisel on veel täiendavas mõõtkavas toodud ära klasterdamise tulemus põhivara v. lühiajaline laen (ehk siis majanduslik mõistlikkus ja jätkusuutlikkus v. kiirlaenu sõltuvus) ning paremal pool põhivara versus bilansi maht.

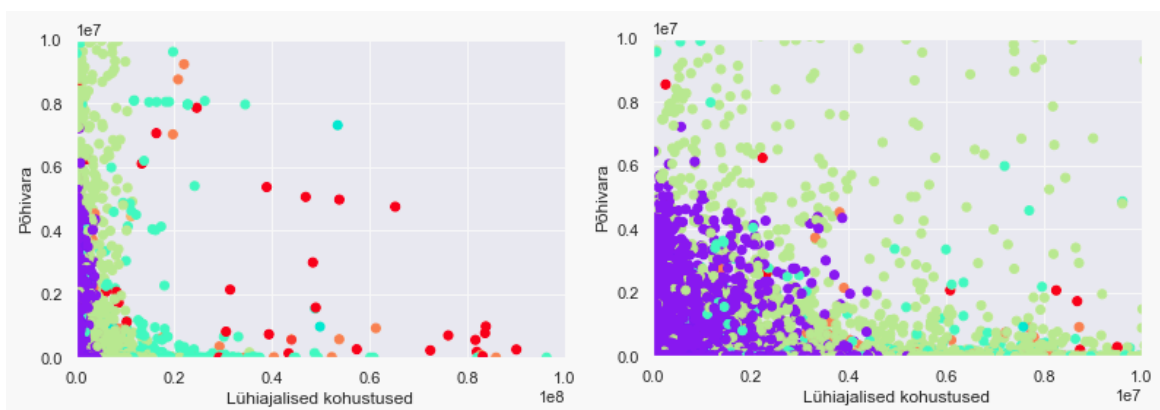
Bilansi mahu ja põhivara seoseid klasterdustulemuste kontekstis vaadates on näha selgelt eristuvaid klastreid nagu ka seoseid põhivara ja bilansi mahu vahel. Bilansi maht näitab üldjuhul põhimõtteliselt ettevõtte suurust, kuid erinevates suurusevahemikes on bilansi struktuurid erinevad (väga suurte bilansimahtude puhul ei ole valdavatel juhtudel tegemist põhivara suure osakaaluga bilansi mahu suhtes).



Joonis 2.2.1.h.: Bilansiandmete klasterdamine põhivara v. pikaajalised kohustused, v. kohustused kokku.

Ülaltoodud jooniselt on näha, et põhivara omab konkreetset mustrit (moodustuvad selged klastrid) kohustused kokku ja pikaajalised kohustused kokku vaadates. Selgelt on dünaamikat, kus suurem põhivara on positiivses seoses suuremate kohustustega (võetud on näiteks pikaajalist laenu, selle arvelt on soetatud põhivara ja arendatud äritegevust). Samas on dünaamikat, kus ollakse hädas väga suurte kohustustega, ja kas tulenevalt sellest või proovides seda olukorda lahendada, on põhivara osakaal minimaalne (selline suhe on kindlasti maksejõulisuse seisukohalt vaadates probleemne).

Alljärgnevatel joonistel on vaadeldud klasterdustulemust kontekstis põhivara v. lühiajalised kohustused. Ülal vaadeldi põhivara v. lühiajalised laenud. Lühiajalised laenud ja lühiajalised kohustused võivad olla suurel määral kattuvad (kuid see olukord on pigem võimalikule maksejõuetusele viitav, sest tegemist on a la kiirraenuvõtjaga oma elutegevuse seeshoidmiseks), kuid võivad ka mitte kattuda (majandustegevuse struktuurist tulenevad lühiajalised kohustused, mis ei viita kuidagi maksejõuetusele). Seega, kuna põhivara on osutunud suhteliselt heaks näitajaks, on oluline seda analüüsida ka lühiajaliste kohustuste vastu.



Joonis 2.2.1.i.: Bilansiandmete klasterdamine põhivara v. lühiajalised kohustused.

Ülaltoodud joonistelt on näha, et lühiajalised kohustused omavad põhivara suhtes täiesti konkreetset eristatavaid klastermustreid. On klastreid, kus vähem lühiajalisi kohustusi tähendavad selgelt rohkem põhivara ja on klastermustreid, kus sisuliselt ettevõtte „valib“ kas hoida raha põivaras või lasta bilansis kasvada lühiajalistel kohustustel (ehk siis OR muster; loogilisem oleks küll AND muster selles küsimuses, kus põhivara on võimalik üles „pumbata“ lühiajaliste kohustuste kasvatamise – kasvada laskmise, s.t. panna lühiajalist raha pikaajalistesse investeringutesse, mis on suur viga tavaliselt – abil).



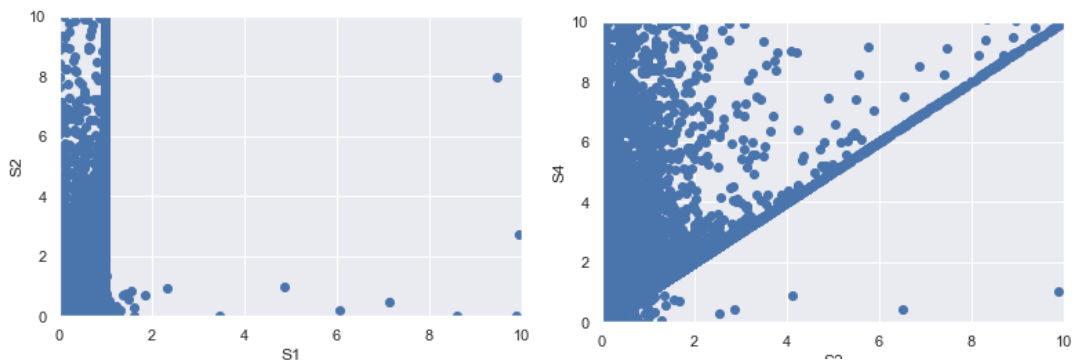
Üldiselt saab väita, et täiendavate parameetrite sissetoomine bilansi analüüsi on põhjendatud. Põhivara, raha ja ettemaksete lisamine mudelisse annab oluliselt juurde – võib välja tuua väga selgeid mustreid, mille alusel järeldusi teha.

### 2.2.2. Suhtarvude klasterdamine

Ettevõtte majanduslikku seisut on võimalik kirjeldada ja analüüsida erinevate bilansi suhtarvude abil. Suhtarvud, mida käesolevas töös on arvesse võetud, on kirjeldatud käesoleva materjali lisa 2. Masinõppe seisukohal on oluline valida esimeseks katsetamiseks olulisemad ja siis vajadusel lisada sinna teisi. Esmaseks klasterdamiseks on aluseks võetu järgmised suhtarvud:

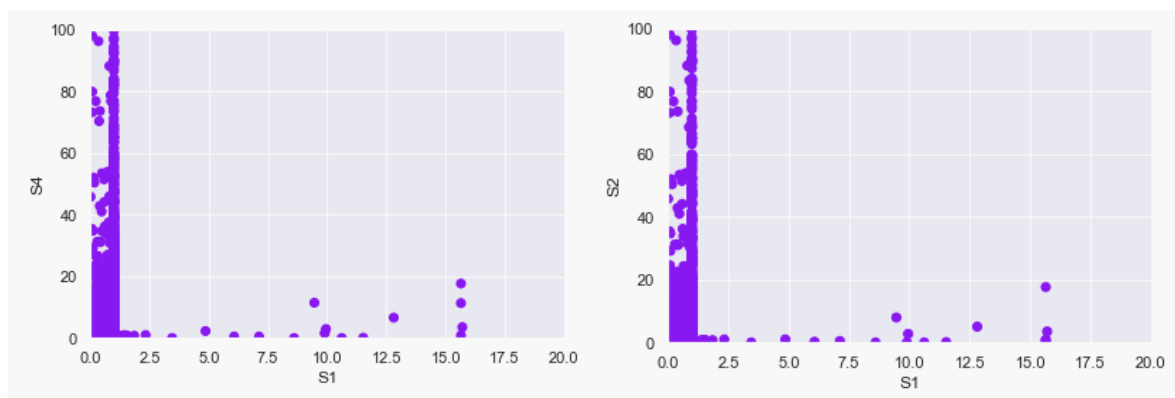
- s1 – käibevara osakaal varades;
- s2 – lühiajalised kohustused varades;
- s3 – pikaajalised kohustused varades;
- s4 – kohustused kokku varades;
- s5 – osakapital varades;
- s6 – raha osakaal käibevaras.

Alljärgneval joonisel on illustratiivselt kujutatud suhtarvude omavahelisi seoseid.



Joonis 2.2.2.a.: Kasutatud suhtarvud, S1 v. S2, S3 v. S4.

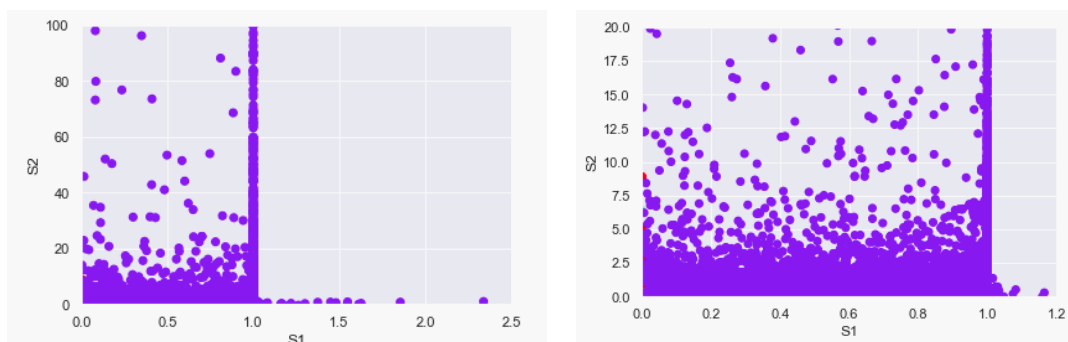
Alljärgneval joonisel on toodud välja suhtarvude klasterdustulemused. Hetkel on kasutatud klasterdamist KMeans(12) põhjal.



Joonis 2.2.2.b.: Suhtarvude klasterdamine S1 v. S4 ja S1 v. S2.

Nagu ülaltoodud jooniselt näha, moodustuvad küll mustrid, kuid sisuliselt on selles vahemikus näha vaid ühte klasterit, mis edasise analüüsi jaoks ei ole väga toetav.

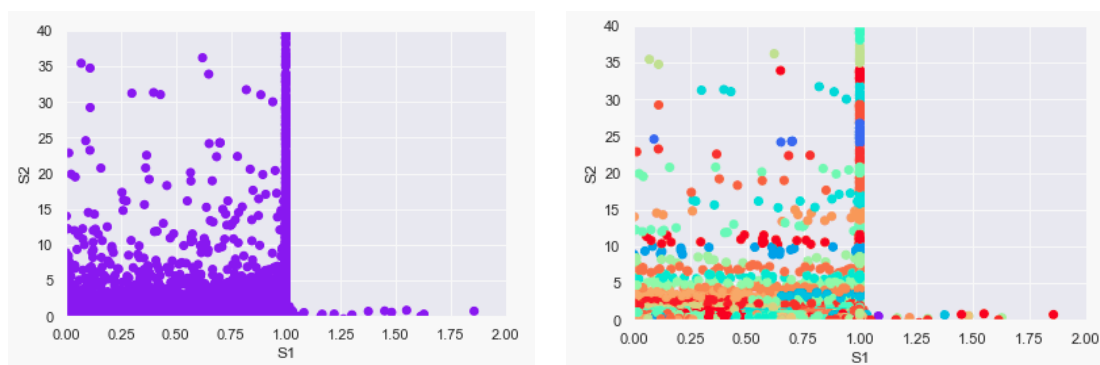
Alljärgneval joonisel on toodud suhtarvude klasterdustulemus S1 v. S2 löikes natuke suuremal määral sissesuurendatuna.



Joonis 2.2.2.c.: Suhtarvude klasterdamine S1 v. S2.

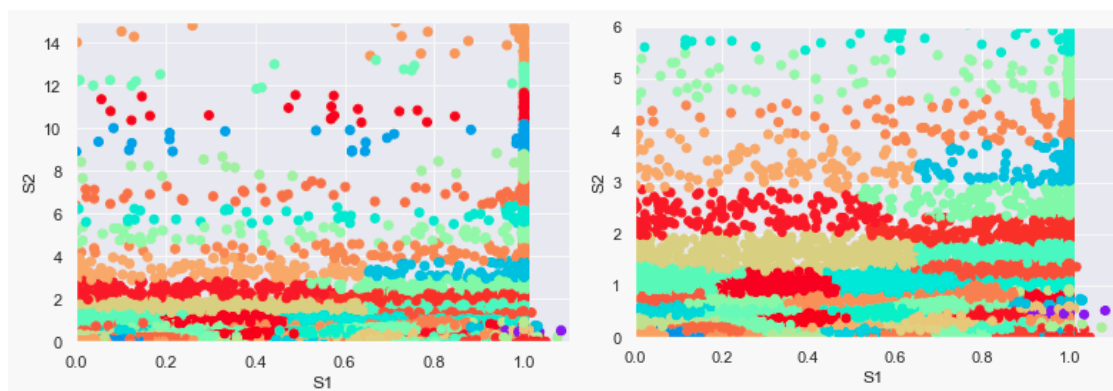
Nagu ülaltoodust nähtub on võimalik leida küll teatud mustrit ja selle põhjal järeldusi teha, kuid KMeans(12) annab siiski selles vahemikus (kus suhtarvud peaks peamiselt olema) vaid ühe klasteri. Seega see ei ole abiks.

Testimise mõttes, viimaks äärmuslikke klastreid nn. eemale ja seega saamaks klasterdust ka sinna piirkonda, kus paikneb peamine huvi suhtarvude seisukohalt vaadates, viisime edasise klasterduse läbi KMeans(200) alusel. Tulemus allpool toodud joonisel, võrdlus KMeans(12)-ga.



Joonis 2.2.2.d.: Suhtarvude klasterdamine S1 v. S2, KMeans(12) v. KMeans(200).

Nagu ülaltoodud jooniselt nähtub (sama mõõtkava, samad suhtarvud) aitab klastrite arvu suurendamine luua olukorda, kus suhtarvude klasterdamisest võib ka analüüsile reaalset kasu tulla.



Joonis 2.2.2.e.: Suhtarvude klasterdamine S1 v. S2, KMeans(200).

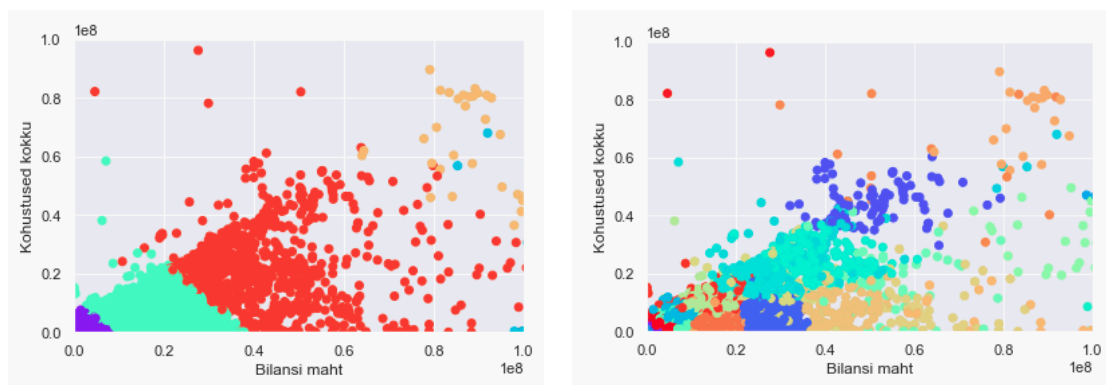
Ülaloodud joonisel on esitatud suhtarvude klasterdamine S1 ja S2 võrdluses KMeans(200)-ga, mõnevõrra rohkem sisse suurendatud kujul. Selgelt eristuvad klastrid ja mustrid, kuid hetkel selliselt, et neid on raske majanduslikult üheselt tõlgendada.

Käesolevas töös on suhtarvude klasterdamist läbi viidud rohkemate variantidega ja rohkemate suhtarvudega, kui ülalpool esitatud, kuid üldjoontes on tulemus sarnane, mis ülallooduga illustreeritud. Bilansisuhtarvud on majandusanalüütikas küll väga olulised näitajad, kuid eraldivõetuna neid masinale õpetada ei ole kerge. Või õigemini suhtarvude pinnalt ei suuda masin juhendamata õppega leida üheselt mustreid (vähemalt käesolevas töös proovitud meetoditega), mille alusel võiks otseselt väita, kas tegemist on võimaliku makseraskusega või mitte.

Samas bilansinäitajata alusel (eelnevalt kirjeldatud) suutis masin juhendamata õppega leida mustreid, mis omavad majanduslikku mõttekust. Seega tehti töö käigus järgimine samm: ühendati bilansiandmed ja bilansi pealt arvatud suhtarvud üheks terviklikuks andmestikuks ja seda kasutati sisendina klasterdamisel (vt. allpool).

### 2.2.3. Bilansiandmete ja suhtarvude koosklasterdamine

Alljärgnevalt on esitatud bilansiandmete klasterdamise tulemused, kui varemalt toodud bilansiandmetele on lisatud varemalt käsitletud suhtarvud.



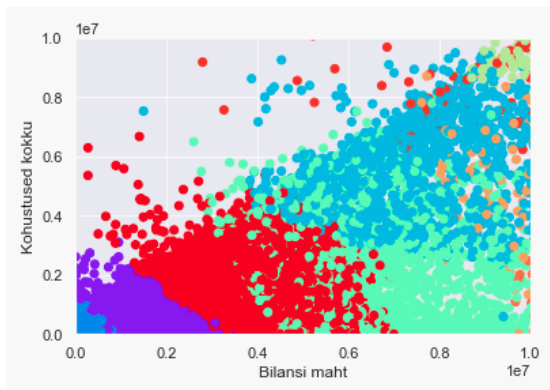
Joonis 2.2.3.a.: Bilansiandmete klasterdamine koos suhtarvudega, KMeans(12) ja KMeans(50).

Ülaloodud joonisel on vasakul joonisel tehtud bilansiandmete ja suhtarvude koosklasterdamine 12 klastrite kontekstis, paremal joonisel on tehtud sedasama 50 klastrite kontekstis. 50 klastrite kontekstis tehtud klasterdamine näitab olulises piirkonnas (kus peamine analüüs tehakse) rohkem klastreid. Lisaks on 50 klastriga joonist vaadates üheselt selge, et sealt on võimalik eristada rohkem olulisi mustreid, kui 12 klastrite puhul. 12 klastrite korral on võimalik eristada peamiselt ettevõtte suurusest tulenevaid erinevaid klastreid, 50 klastrite korral on selgelt eristuvad mustrid (a la suur bilansimaht, vähem kohustusi; kohustusi proportsionaalselt bilansi mahuga jne.).

Ülaloodud klasterdamistulemuse tõlgendamisel (parempoolne joonis, KMeans(50)) on oluline juhtida ka tähelepanu sellele, et osad klastrid on kompaktsemad, selgelt määratletud, teine osa on jällegi hajutatud. Masinale on mõistlik õpetada eelkõige neid klastreid, mis on selgelt määratletud ja nn. kompaktsete majandustunnustega. Kompaktseid klastreid on ka omakorda kahte liiki: (a) osad on määratletud selgete esiletulevate majanduslike joontega (teatud suuruses bilansimahu ja kogukohustuste proportsioon vms.) ja (b) osad on küll kompaktsed ja jooniste mõistes selgelt eristuvad klastrid, kuid neile on raske omistada mingit konkreetset majanduslikku spetsiifikat. Majandusandmete erinev klasterdamine teenib seda eesmärki, et leida need kombinatsioonid ja olukorrad (mustrid), mis ütlevad nii majanduslikult midagi sisuliselt, kui samas annavad ka masinale võimaluse neid eristada.

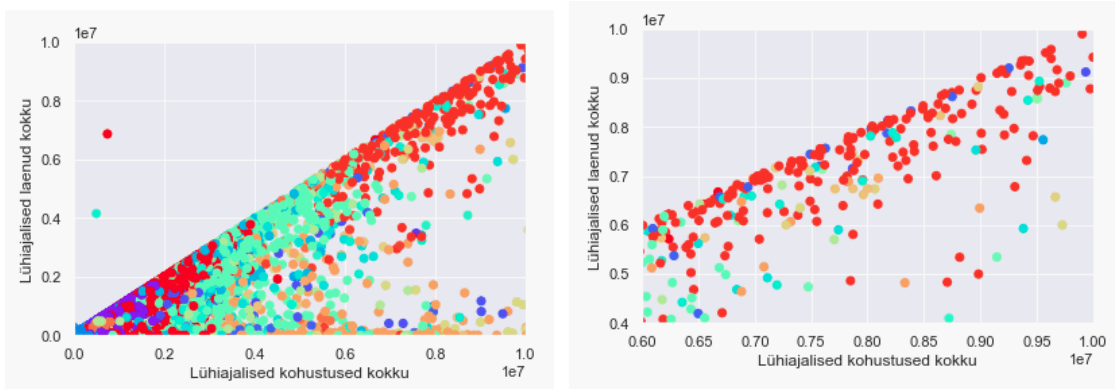
Oluline on ülaltoodud jooniste koha pealt rõhutada ka seda, et ülaltoodud joonisel on skeemid kahedimensionaalsed (kuna joonised on kahedimensionaalsed), kuid kokku on klasterdatud mitmeid erinevaid bilansi näitajaid ja selle pinnalt arvatud suhtarve. Jooniseid saab välja tuua erinevate tunnuste osas (võttes kaks tunnust kõigist võimalikest ja pannes need joonisele). Teistsuguste tunnuste ülesjoonistamisel näeme teistsuguseid klastrite pilte.

Alljärgneval joonisel on bilansiandmete ja suhtarvude klasterdustulemus kohustused kokku v. bilansi maht kontekstis toodud suurendatuna. Jooniselt on näha, et väikestes mahtudes käituvad klastrid bilansi mahu ja kohustused kokku suhtes proportsionaalselt, suuremates mahtudes (joonis 2.2.3.a, parempoolne ülal) käituvad need erinevalt.



Joonis 2.2.3.b.: Bilansiandmete klasterdamine koos suhtarvudega, KMeans(50).

Alljärgneval joonisel on sama klasterduse (bilansi näitajad ja suhtarvud koos, KMeans(50)) välja toodud lühiajalised laenud kokku v. lühiajalised kohustused kokku kahes erinevas mõõtkaavas. Lühiajalisi laene ja lühiajalisi kohustusi omavahel vaadates on võimalik välja joonistada mustrit, kus eristuvad nn. ettevõtetest „kiirlaenuvõtjad“ (olukord, kus lühiajalised laenud on pea võrdsed lühiajaliste kohustuste summaga).



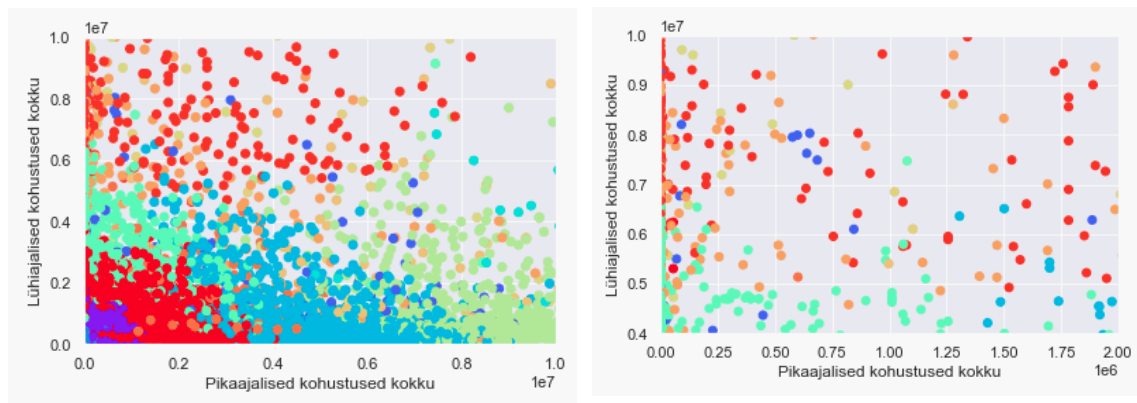
Joonis 2.2.3.c.: Bilansiandmete klasterdamine koos suhtarvudega, lühiajalised laenud v. lühiajalised kohustused, KMeans(50).

Ülaltoodud jooniselt on näha, et selgelt on moodustumas telg, kus paiknevad ettevõtted, kelle lühiajaliste laenude maht on võrdne lühiajaliste kohustuste mahuga (kui mõningatel juhtudel on tegemist olukorraga, kus lühiajalised laenud ületavad lühiajalisi kohustusi, siis on tõenäoliselt tegemist mingit liiki andmeprobleemiga – lühiajalised laenud saavad olla lühiajaliste kohustustega võrdsed või väiksemad, mis ka joonistelt selgelt välja tuleb).

Näiteks suurusjärgus umbes kuue miljoni eurose lühiajaliste kohustuste mahu pealt algaval klastril (punane klaster, parempoolsel joonisel välja suurendatud) on selgelt need tunnused, et ettevõtted finantseerivad ennast „kiiralaenudega“, mis on selgelt viide sellele, et võimalik maksejõuetus võib tekkida.

Teisalt jällegi kolme kuni kuue miljoni eurose lühiajaliste kohustuste mahu juures olev klaster (roheline klaster) sisaldab nii „kiiralaenusõitlasi“ kui ka natuke konservatiivsemaid ettevõtteid. Seega hinnangud ei saa (vähemalt selle klatri kontekstis olla ühesed).

Alljärgneval joonisel on sama klasterdus esitatud pikaajalised kohustused kokku v. lühiajalised kohustused kokku (erinevate mõõtkavades).



Joonis 2.2.3.d.: Bilansiandmete klasterdamine koos suhtarvudega, pikaajalised kohustused v. lühiajalised kohustused, KMeans(50).

Ülaltoodud jooniselt on võimalik leida varemalt käsitletud punane „kiiralaenuvõtjate“ klaster, kuid selles vaates on selge, et osadel seal olevatel ettevõtetel on ka pikaajaliste kohustuste koormus suur (mis antud juhul muudab neid riskantsemaks, sest lisaks „kiiralaenudele“ on peal ka „eluasemelaenu“, kuid ka need vajavad teenindamist).

Huvitav klaster on sinine klaster (pikaajaliste kohustuste mahu kolm kuni kaheksa miljonit eurot juures). Lühiajaliste kohustuste ja lühiajaliste laenude vaates see klaster välja ei joonistunud, mis tähendab, et seda riski klastril ei ole. Pigem võib pikaajaliste kohustuste (tõenäoliselt suures mahus pangalaenu) järgi hinnata, et ettevõtte on pankade silmis usaldusväärne ja läbinud sealsete analüütikute usaldusuuringu positiivselt. Klaster oli ka selgelt näha bilansi mahu ja kohustused kokku võrdluses (joonis 2.2.3.b). See klaster omab selgeid piirjooni ja seda on võimalik majanduslikult tõlgendada.

Ülaltoodud joonised ja tõlgendused on näiteks, mida käesoleva prototüübi väljatöötamise juures otsitakse ja millest lähtuda püütakse. Töös olevaid mustreid on rohkem, järeltõlget, mida teha püütakse, on rohkem. Koondjäreltõlge, mida bilansi andmete ja bilansi suhtarvude koosklasterdamise tulemuste analüüsis võib välja tuua, on see, et sellise klasterdamise tulemusena on süsteem juhendamata masinõppe meetodil suuteline leidma mustreid, mida on võimalik majanduslikult tõlgendada ja mida masin on suuteline eristama.

## 2.3. Transaktsiooniandmete analüüs

Täiendava suunana analüüsis prototüübi väljatöötamiseks on sisse toodud transaktsiooniandmete klasterdamine. Transaktsiooniandmed tähendavad ühe ettevõtte poolt teisele ettevõttele tasutud summased, mis on eristatavad summade suuruse alusel. Ühtlasi on iga konkreetse maksja puhul teada summa saajate arv (on teada konkreetset ettevõtet, mis on summade saajaks), mis võimaldab välja tuua tehingupartnerite arvu. Seega on transaktsiooniandmete sisuks nii tehingupartnerite vahel liikuvad summad kui ka tehingupartnerite arv. Info tuleb käibemaksu andmetest.

Käibedeklaratsiooni lisal KMD\_INF on toodud vähemalt 1000 euroste tehingute andmed. Antud juhul on kasutusel A-osa, kus on toodud müügiarved.

Info siit (<https://www.emta.ee/ariklient/maksud-ja-tasumine/kaibemaks/kaibedeklaratsiooni-ja-aruanete-esitamine/kmd-inf-osa-muugiarved-taitmise-juhised>):

*KMS § 11 lõikes 1 toodud üldreegli kohaselt on käive toimunud, kui kaup on ostjale lähetatud või kättesaadavaks tehtud või teenus on osutatud või kauba/teenuse eest on makse laekunud, olenevalt sellest, milline toiming oli varasem.*

*Arve andmed tuleb nii vormil KMD kui ka KMD INF kajastada sellel kuul, mil nimetatud arve osas on käive osaliselt või täielikult tekkinud vastavalt käibe tekkimise ajale. Seega kajastatakse arve andmed KMD INF-il samal maksustamisperioodil, mil need kuuluvad deklareerimisele KMD-l.*

Käibedeklaratsiooni lisas real 9 kajastub maksustamisperioodil vormi KMD lahtrites 1 + 2 kajastatud maksustatav käive. Seega, kui käibe summana arvestada R1 + R2 + R3, siis võib käibe jagunemist mõtestada järgmiselt:

- 20% määraga maksustatavad toimingud ja tehingud (R1) + 9% määraga maksustatavad toimingud ja tehingud (R2), mis kajastuvad KMD\_INF-il real 9;
- 20% määraga maksustatavad toimingud ja tehingud (R1) + 9% määraga maksustatavad toimingud ja tehingud (R2), mis ei kajastu KMD\_INF-il real 9;
- 0% määraga maksustatavad toimingud ja tehingud.

See tähendab, et käibedeklaratsioonil ridadel 1 ja 2 on toodud KMD\_INFil kajastatud ettevõtete vaheliste tehingute väärtus, KMD\_INFil kajastamata ettevõtete vaheliste tehingute väärtus ning müük eraisikutele.

Varasem analüüs on näidanud, et KMD\_INFil kajastatud tehingute väärtus on moodustanud ligikaudu poole KMD-l ridadel 1 ja 2 toodud väärtusest.

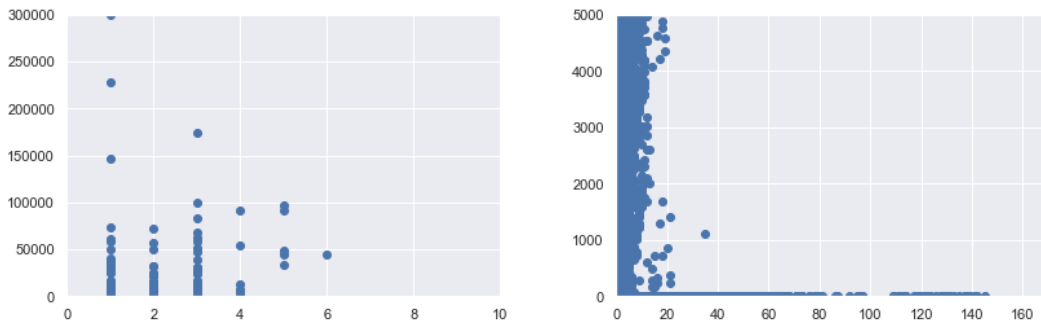
KMD INF andmetel põhineval klasterdussisendiks on arvestatud järgmised näitajad:

- partnerite arv; ettevõtete arv, millele konkreetne ettevõtte on oma teenuseid/toodangut müünud (kellelt ta on raha saanud);
- keskmine makse; keskmine makse, mida ettevõtte on perioodi jooksul partnerite lõikes teinud;
- mediaanmakse; mediaan maksetelt, mida ettevõtte on perioodi jooksul partnerite lõikes teinud;
- summa; koondsumma, mida ettevõtte on perioodi jooksul partneritele välja maksnud, kõigi partnerite lõikes kokku;
- maksimum; maksimaalne makse, mida ettevõtte on perioodi jooksul mõnele partnerile maksnud;
- miinimum; minimaalne makse, mida ettevõtte on perioodi jooksul mõnele partnerile maksnud.

Perioodiks on esmase klasterdamise juures võetud 30 päeva, ehk siis üks kalendrikuu. Täiendavalt võib perioodiks võtta 60 päeva ja 90 päeva, neid võib kõiki eraldi vaadelda, kuid võib ka koondina vaadelda (klasterdada). Eraldiklasterdamine ja siis üheks tervikuks liitmine annab võimaluse hinnata aegrida, ajas toimunud dünaamikat (maksevõime vähenemine/suurenemine, ühest hoiatuskategooriast teise liikumine). Transaktsiooniandmete suureks eeliseks on asjaolu, et neid on

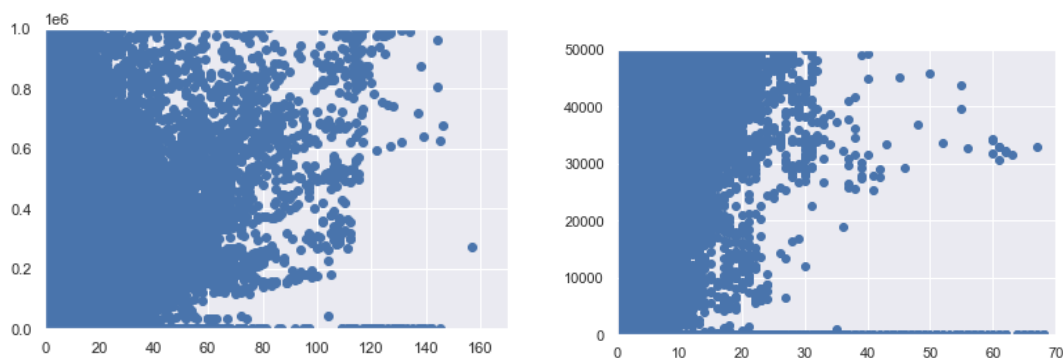
võimalik saada ja analüüsida kuupõhiselt, seega kiiresti on võimalik võtta seisukohta ettevõtte majandusseisu muutumise osas (mitte tagant järele: varane hoiatus v. ajalooline hoiatus).

Alljärgneval joonisel on toodud näiteid transaktsioonide andmete olemusest.



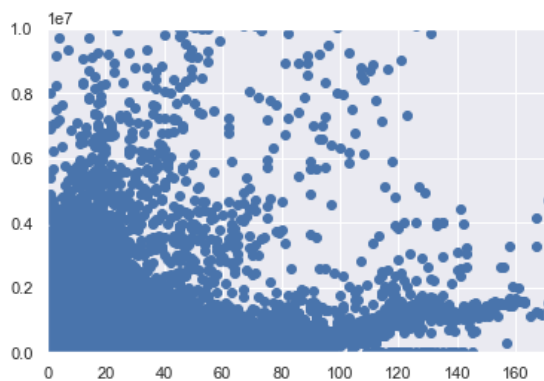
Joonis 2.3.a.: Transaktsiooniandmed, bilansiklaster 4, kogu kinnisvara.

Ülaltoodud joonise vasakpoolisel joonisel on toodud välja transaktsiooniandmed, mis on seotud ühe konkreetse bilansi/suhtarvude klasterduse tulemuse klastriga (erinevate klasterduste tulemklastrite vahel otsitakse seoseid). Parempoolisel joonisel on kujutatud väljavõtte graafikust, kus on kõik kinnisvarasektori transaktsioonandmed. Vertikaaltelgedel on kujutatud kogusumma, horisontaalteljel partnerite arv.



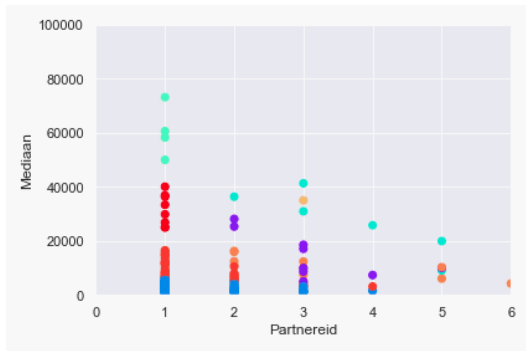
Joonis 2.3.b.: Transaktsiooniandmed, summad v. partnerite arv, kinnisvara.

Ülaltoodud joonistel on veel näidatud transaktsioonide andmeid, teises mõõtkavas, iseloomustamaks nende andmete umbkaudset olemust. Sama alltoodud joonisel.



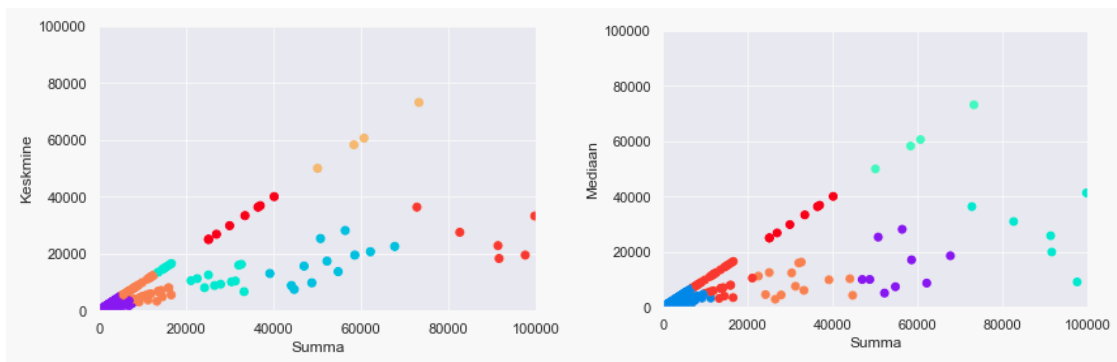
Joonis 2.3.c.: Transaktsiooniandmed, summad v. partnerite arv, kinnisvara.

Üldiseloomustavalt võib transaktsioonandmete kohta öelda, et tegemist on tehnilises mõttes teineteisest vähemeristuvate andmetega, kui seda on bilansiandmed ja nende suhtarvud. See võimaldab ühelt poolt „kergemat klasterdamist“ (s.t. vähem erisusi ja nüansse), kuid teiselt poolt on nendele andmetele raskem anda majanduslikku sisu.



Joonis 2.3.d.: Transaktsiooniandmed klasterdatud, seos bilansiklaster 4-ga.

Ülaltoodud joonisel on näha bilansiklaster 4-ga seotud andmete klasterdustulemus esitatuna partnerite ja mediaansummade lõikes. Alltoodud joonisel on esitatud sama klasterdus keskmine v. summad ja mediaan v. summad lõikes.



Joonis 2.3.e.: Transaktsiooniandmed klasterdatud, seos bilansiklaster 4-ga.

Ülaltoodud jooniselt on näha, et selgelt on eristatavad erinevad klastrid (nagu eelnevalt öeldud, on transaktsioonandmed hästi klasterdatavad andmed).

Jooniselt on näha ka seda, et klastrid on väikesed. Tegemist on andmetega, kus on 270 kirjet ja seetõttu on klastrid väikesed. Ainult 270 kirjet on seetõttu, et need olid kirjed mis vastasid bilansiandmete klasterdustulemusele 4 klastri osas.

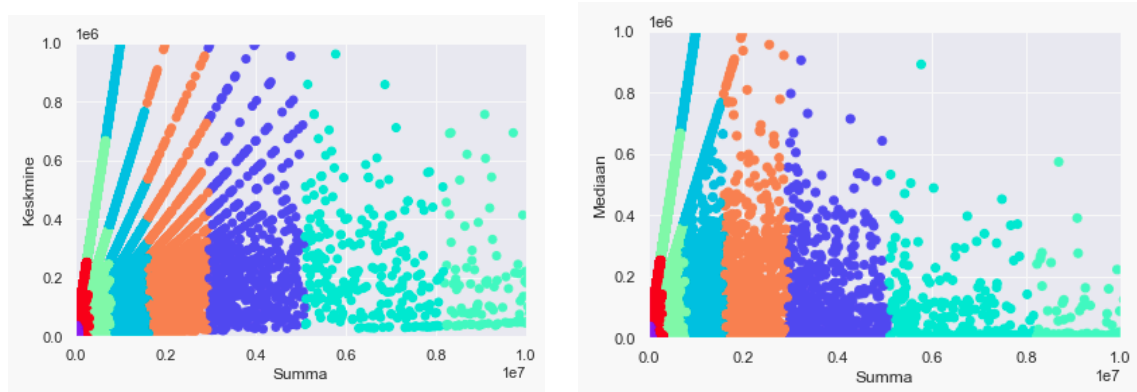
Ülaltoodust tuleb välja ka probleem, mis tekib erinevate klasterdustulemuste sidumisel omavahel. Sektorid, mida analüüsitakse (kinnisvara, hotellindus jne.) on teatud suurusega (mitte väga suured). Nende klasterdamisel tekivad klastrid, mis on veel väiksemad. Nendele klastritele vastavad teistes andmestikus teatud piiratud hulk kirjeid. Kui need on veel klasterdatud, siis on see vastavus veel piiratum.

Mudeleid ehitades tekivad väga erinevad, piiratud suurusega andmekogumid, mis ühelt poolt küll kirjeldavad täpselt majanduslikke tingimusi, kuid mille hulk on suur. Projekti raames proovitakse seda olukorda lahendada, katsetada erinevaid viise klasterdamiseks ja andmete sidumiseks, et klastrite arvu ja majanduslike näitajate sisukust tasakaalus hoida. Majanduslike näitajate sisukust ei saa liiga



madalaks lasta, sest siis ei anna mudel vajalikke tulemusi piisava täpsusega, et selle põhjal midagi ennustada (kedagi hoiatada) saaks.

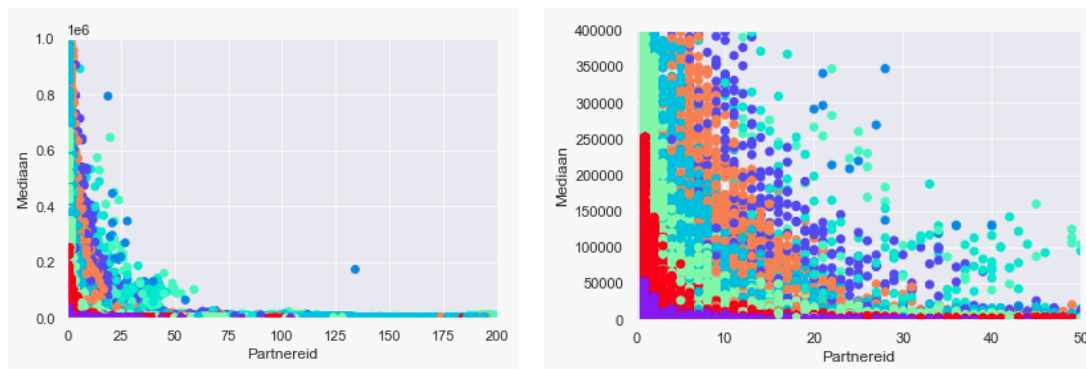
Alljärgneval joonisel on välja toodud kinnisvarasektori klasterdatud andmed, summad mediaanide ja keskmiste vastu.



Joonis 2.3.f.: Transaktsioonandmed klasterdatud, kinnisvarasektor.

Nagu ülaltoodud jooniselt nähtub, on transaktsioonandmete puhul võimalik välja tuua erinevaid klastreid, mis on selgelt eristatavad, võrreldavates suurustes, kompaktsed. Igati ilus klasterdustulemus, parem kui bilansiandmete põhjal, kuid samas on transaktsioonandmete klastrite majandusliku tõlgendamise probleem keerukam (nagu joonistelt näha, on erisused peamiselt liikunud summade suuruse alusel, kuid see ei ole otseseks näitajaks, mille alusel saaks hinnata maksejõuetust).

Alljärgneval joonisel on sama transaktsioonandmete klasterdus esitatud partnerid v. mediaan skaalal erinevates mõõtkaavades.



Joonis 2.3.g.: Transaktsioonandmed klasterdatud, kinnisvarasektor.

Nagu ülaltoodud jooniselt nähtub joonistavad transaktsioonandmete klastrid selgeid mustreid ka partnerid v. mediaan skaaladel. Samas ka need mustrid on eelkõige jaotuvad eelkõige suurusest tulenevast (a la rohkem partnereid ja suurem mediaan versus vähem partnereid ja väiksem mediaan), mis otseselt ei viita maksejõuetuse võimalikkuse asjaoludele.

Töö projektis transaktsioonandmetega jätkub, kuid praeguses faasis on võimalik tuua välja kaks järeldust.

ESIMENE JÄRELDUS. Transaktsioonandmed moodustavad selgelt eristatavaid, võrreldava suurusega kompaktsed klastreid, mida on samas majanduslikult raskem maksejõuetuse tunnuseid sisaldavatena tõlgendada. Olulise lisandväärtuse saab loodetavasti sellest, kui

transaktsiooniklastreid vaadatakse koos bilansiklastrite vastavate seostega (seotakse ära bilansiklastrid teatud kindlate transaktsioonidega transaktsiooniklastrites).

TEINE JÄRELDUS. Tuleb sisse tuua täiendavad perioodid näitajate kohta (30 päeva mediaanile lisada 60 päeva mediaan, 90 päeva mediaan, sama teha keskmiste, miinimumide ja maksimumidega) ning analüüsida, kas klastrid annavad rohkem maksejõuetuse tunnustega järelusi. Ühtlasi kõrvutada erinevate aegade klastreid, vaadata nende ajalist muutumist, ehk ettevõtete kulgemist ühest klastrist teise ja selle majanduslikku tähendust.

Illustreerimaks viimast järelust on allpool toodud joonis. Nimelt on joonisel näha transaktsioonandmete üks vaheklasterdus, kus masin on kasterdussisendina arvesse võtnud ka perioodi. Perioodi arvestamine ei ole sisendina ei ole vale, kuid kui andmed on väga „lakoonilised“, nagu transaktsioonide andmed on, siis võib periood hakata mängima liiga olulist tähtsust ja see ei ole õige. Allpool toodud joonis illustreerib sellist valeklasterdamist ning seega ühtlasi põhjendab ära selle, miks on vajalikud korduvklasterdused ja kõige õigema momendi leidmine.



Joonis 2.3.h.: Transaktsiooniandmed klasterdatud, perioodi liiga suur osakaal.

Lisas 8 toodud kinnisvarasektori jagunemine klastritesse tegevusalade lõikes.

## 2.4. Mudelite loomine ja andmete testimine

### 2.4.1. Üldised lähte-eeldused mudelite loomisel

Klasterdatud andmete pinnalt on käesoleva töö käigus loodud mudeleid, mida masinõppe meetodil süsteemile õpetatakse. Klasterdatud andmed võimaldavad määratleda ettevõtteid, kellel võib tulevikumomendil tekkida maksejõuetuse olukord. Klasterdatud andmeid ja nende tunnuseid kasutame klassifikatsiooni mudeli treenimisel.

Töö käib järgmise loogika alusel:

### **Bilansiandmed -> klassifikatsiooni mudel -> prognoos**

Töö käigus on katsetatud erinevaid mudeleid: naïve bayes (gaussian, bernoulli, categorical, multinomial), svm, gradient boosting jne. Otsustasime RandomForest kasuks, kuna out-of box tulemus oli juba paljulubav.

Alljärgneval joonisel on toodud RandomForest baasmudeli väljund.

## RandomForest baasmudel

```
model = treeni(RandomForestClassifier())
```

```
The Training clf Accuracy is: 0.8687710962870535
```

```
classification_report =
```

	precision	recall	f1-score	support
0	0.08	0.98	0.14	700
0_A_1	1.00	0.84	0.91	58829
0_A_2	0.64	0.87	0.73	1925
0_B	0.83	0.82	0.83	827
1	0.67	1.00	0.80	2
2	1.00	1.00	1.00	2
3	0.77	0.55	0.64	31
4	0.83	0.92	0.87	195
accuracy			0.85	62511
macro avg	0.73	0.87	0.74	62511
weighted avg	0.97	0.85	0.90	62511

```
Aega kulus: 1207.2708015441895 sekundit
```

Joonis 2.4.1.a.: RandomForest mudeli tulemus.

Nagu ülaltoodud jooniselt nähtub, on kohati saavutatud mudeli õpetamisel suhteliselt heal tasemel f1-score, samuti ka head precision, recall ja accuracy näitajad.

Kasutatud RandomForest mudeli efektiivsemaks muutmiseks on toodud sisse hüperparameetrite muutmise aspekt. Alljärgneval joonisel on esitatud hüperparameetrid, mida mudelis on kasutatud.

Parameters currently in use:

```
{'bootstrap': True,
 'ccp_alpha': 0.0,
 'class_weight': None,
 'criterion': 'gini',
 'max_depth': None,
 'max_features': 'auto',
 'max_leaf_nodes': None,
 'max_samples': None,
 'min_impurity_decrease': 0.0,
 'min_impurity_split': None,
 'min_samples_leaf': 1,
 'min_samples_split': 2,
 'min_weight_fraction_leaf': 0.0,
 'n_estimators': 100,
 'n_jobs': None,
 'oob_score': False,
 'random_state': None,
 'verbose': 0,
 'warm_start': False}
```

Joonis 2.4.1.b.: RandomForest mudelisse lisatud hüperparameetrid.

Kõiki parameetreid mudeli tuunimisel kasutama ei hakata, alljärgnevalt on välja toodud parameetrid, mida muudetakse parema tulemuse saamiseks:

- n\_estimators = number of trees in the forest;
- max\_features = max number of features considered for splitting a node;
- max\_depth = max number of levels in each decision tree;
- min\_samples\_split = min number of data points placed in a node before the node is split;
- min\_samples\_leaf = min number of data points allowed in a leaf node;
- bootstrap = method for sampling data points (with or without replacement).

Mõistmaks, mis suurusjärgus need parameetrid olla võiks kasutame RandomSearchCV-d.

Mudelile antakse ette lai parameetrite vahemik ja vastuseks saadakse natuke kitsam parameetrite valik.

Algne lai parameetrite valik, mida läbi skanneeritakse (k-fold CV – 3 foldi) on toodud alljärgneval joonisel.

```

from sklearn.model_selection import RandomizedSearchCV
# Number of trees in random forest
n_estimators = [int(x) for x in np.linspace(start = 200, stop = 2000, num = 10)]
# Number of features to consider at every split
max_features = ['auto', 'sqrt']
# Maximum number of levels in tree
max_depth = [int(x) for x in np.linspace(10, 110, num = 11)]
max_depth.append(None)
# Minimum number of samples required to split a node
min_samples_split = [2, 5, 7, 10]
# Minimum number of samples required at each leaf node
min_samples_leaf = [1, 2, 3, 4]
# Method of selecting samples for training each tree
bootstrap = [True, False]
# Create the random grid
random_grid = {'n_estimators': n_estimators,
               'max_features': max_features,
               'max_depth': max_depth,
               'min_samples_split': min_samples_split,
               'min_samples_leaf': min_samples_leaf,
               'bootstrap': bootstrap}

```

Joonis 2.4.1.c.: Skanneeritavate parameetrite valik.

Alljärgneval joonisel on toodud saadav parameetrite piirkond, kus mudel kõige paremini töötab/prognoosib.

```

rf_random.best_params_

{'n_estimators': 200,
 'min_samples_split': 7,
 'min_samples_leaf': 3,
 'max_features': 'auto',
 'max_depth': 100,
 'bootstrap': False}

```

Joonis 2.4.1.d.: Parima prognoosivõimekuse parameetrite piirkond.

RandomSearchCV tulemus suunatakse GridSearchCV-sse, kus toimub parameetrite „peenähälestamine“. Skanneeritakse eelmise otsingu tulemuste ümbruses (k-fold CV – 3 foldi).

```

from sklearn.model_selection import GridSearchCV
# Create the parameter grid based on the results of random search
param_grid = {
    'bootstrap': [False],
    'max_depth': [80, 90, 100, 110],
    'max_features': ['auto'],
    'min_samples_leaf': [2, 3, 4],
    'min_samples_split': [5, 7, 9],
    'n_estimators': [100, 200, 300, 400]
}

```

Joonis 2.4.1.e.: GridSearchCV kasutamine.

Alljärgneval joonisel on toodud GridSearchCV tulemus.

```
grid_search.best_params_
```

```
{'bootstrap': False,  
 'max_depth': 110,  
 'max_features': 'auto',  
 'min_samples_leaf': 3,  
 'min_samples_split': 7,  
 'n_estimators': 300}
```

Joonis 2.4.1.f.: GridSearchCV tulemus.

GridSearchCV tulemusi kasutatakse edasisel mudeli treenimisel.

## 2.4.2. Prognoosimudelite kasutamine

Alljärgnevalt on toodud materjal, mis kirjeldab käesoleva töö käigus tehtavad andmete prognoosimist, toodud on illustratiivsed näited andmaks ülevaadet, realselt katsetatakse rohkemal viisil ja mitmete erinevate lahendustega.

Alljärgneval joonisel toodud andmete sisselugemine.

```
def data_preparation(x):  
    ...  
    Algandmete töötlus, et viia andmed mudelile sobivale kujule  
    ...  
    df_x = pd.read_csv(x)  
    df_x.replace([np.inf, -np.inf], np.nan, inplace=True)  
    df_x.fillna(0, inplace=True)  
  
    return df_x  
  
# Loeme sisse testandmed  
to_predict = data_preparation('../..//data/raw/forecasting.csv')  
  
to_predict
```

Joonis 2.4.2.a.: Andmete sisselugemine.

Alljärgneval joonisel toodud näidisväljavõtte sisseloetavatest andmetest.

PERIOD_NM	BI_100_1	BI_150_1	BI_180_1	BI_190_1	BI_240_1	BI_250_1	BI_290_1	BI_310_1	BI_370_1	BI_400_1	BI_40_1	BI_40_2
2019	18201.0	2012.0	2012.0	20213.0	0.0	1158.0	1158.0	0.0	1158.0	2556.0	8577.0	8327.0
2017	787467.0	331317.0	342454.0	1129921.0	7121.0	556432.0	563553.0	0.0	563553.0	0.0	93436.0	10271.0
2012	1316520.0	186972.0	186972.0	1503492.0	2206.0	675259.0	692465.0	9368.0	701833.0	51200.0	783465.0	720148.0
2018	1080009.0	2600.0	2600.0	1082609.0	7500.0	160584.0	168084.0	541719.0	709803.0	23966.0	10904.0	27491.0
2016	84700.0	98060.0	98060.0	182760.0	56699.0	0.0	56699.0	20000.0	150049.0	2550.0	38474.0	29164.0
2019	20000.0	56000.0	120000.0	1200000.0	23000.0	3000.0	4000.0	40000.0	44000.0	3000.0	12000.0	11000.0
2019	10000.0	30000.0	40000.0	500000.0	120000.0	2000.0	120000.0	240000.0	360000.0	5000.0	3000.0	3500.0
2018	7000.0	20000.0	20000.0	100000.0	25000.0	4000.0	30000.0	40000.0	70000.0	10000.0	5000.0	4000.0

Joonis 2.4.2.b.: Andmenäidis.

Alljärgneval joonisel on näidatud varasemalt treenitud mudelite sisselugemine.

```
# Loeme varasemalt treenitud RF mudelid
rf_base = joblib.load('../models/rf_base.sav')
rf_tuned = joblib.load('../models/rf_tuned.sav')
```

Joonis 2.4.2.c.: Varasemalt treenitud mudelite sisselugemine.

Alljärgneval joonisel toodud prognoosimine.

```
def predict_data(model, model2, data):
    """
    Mudelite testimine
    """
    X_predict = data.drop(columns=['kood'])
    predictions = model.predict(X_predict)
    predictions2 = model2.predict(X_predict)

    data['pred'] = predictions
    data['pred_2'] = predictions2

    return data
```

```
# Prognoosimine
predictions = predict_data(rf_base, rf_tuned, to_predict)
predictions
```

Joonis 2.4.2.d.: Mudelite kasutamine prognoosimiseks.

Alljärgneval joonisel on toodud prognoosimise tulemus, tulpadesse pred ja pred\_2 on kirjutatud klastrid, kuhu masin arvab ühte või teist testitud andmerida kuuluvat.

PERIOD_NM	BI_100_1	BI_150_1	BI_180_1	BI_190_1	BI_240_1	BI_250_1	BI_290_1	BI_310_1	BI_370_1	BI_400_1	BI_40_1	BI_40_2	pred	pred_2
2019	18201.0	2012.0	2012.0	20213.0	0.0	1158.0	1158.0	0.0	1158.0	2556.0	8577.0	8327.0	0_A_1	0
2017	787467.0	331317.0	342454.0	1129921.0	7121.0	556432.0	563553.0	0.0	563553.0	0.0	93436.0	10271.0	0_A_1	0_A_1
2012	1316520.0	186972.0	186972.0	1503492.0	2206.0	675259.0	692465.0	9368.0	701833.0	51200.0	783465.0	720148.0	0_A_2	0_A_2
2018	1080009.0	2600.0	2600.0	1082609.0	7500.0	160584.0	168084.0	541719.0	709803.0	23966.0	10904.0	27491.0	0_A_1	0_A_1
2016	84700.0	98060.0	98060.0	182760.0	56699.0	0.0	56699.0	20000.0	150049.0	2550.0	38474.0	29164.0	0_A_1	0_A_1
2019	20000.0	56000.0	120000.0	1200000.0	23000.0	3000.0	4000.0	40000.0	44000.0	3000.0	12000.0	11000.0	0_A_1	0_A_1
2019	10000.0	30000.0	40000.0	500000.0	120000.0	2000.0	120000.0	240000.0	360000.0	5000.0	3000.0	3500.0	0_A_1	0_A_1
2018	7000.0	20000.0	20000.0	100000.0	25000.0	4000.0	30000.0	40000.0	70000.0	10000.0	5000.0	4000.0	0_A_1	0_A_1
2020	10.0	6000.0	12000.0	250000.0	150000.0	10.0	180000.0	30000.0	210000.0	2000.0	10.0	300.0	0_A_1	0_A_1
2021	200.0	2000.0	5000.0	120000.0	90000.0	200.0	100000.0	5000.0	105000.0	10.0	200.0	400.0	0_A_1	0_A_1
2020	193097.0	43095.0	43095.0	236192.0	6133.0	150722.0	156855.0	6245.0	163100.0	2556.0	5946.0	37448.0	0_A_1	0_A_1
2021	3341507.0	1552252.0	62212413.0	65553920.0	696896.0	767586.0	1464482.0	28243839.0	29708321.0	10000.0	3262134.0	731815.0	4	4
2021	43694.0	9115.0	9893.0	53587.0	4213.0	531.0	4744.0	860.0	5604.0	2500.0	24050.0	36550.0	0_A_1	0_A_1

Joonis 2.4.2.e.: Prognoositulemus.

Kaks prognoosi: pred on baasmudel ja pred\_2 on hüperparameetrid tuunitud. Mingis osas annavad nad erinevaid tulemusi.

### 2.4.3. Prognoosimudeli reaaleluline kontroll

Töö käigus viiakse läbi prognoosimudelite reaalelulist kontrolli. Võetakse mõni ettevõtte, mille mudel on paigutanud mõnda kindlasse klastrisse, analüüsitakse ettevõtte majanduskäitumist kolmandatest allikatest pärit informatsiooni alusel ja siis tehakse järeldus, kas masin võis midagi õigesti järeldada või mitte.

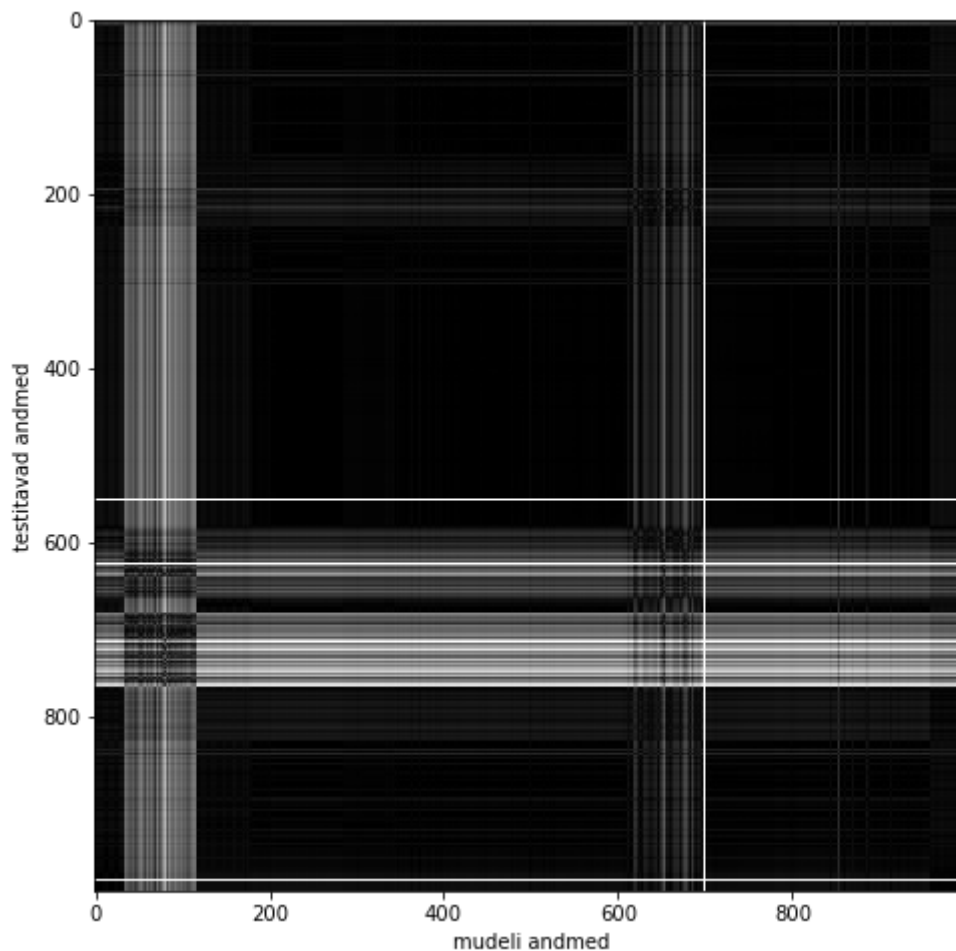
Näiteks pidas süsteem riskantseks sellise ettevõtte käitumist, mille osakapital on 10 000 eurot, mis 2019. aastal kasvatas oma investeringuid kinnisvara ostuks 60 protsenti 12.3 miljoni euroni (s.t. nii palju investeeriti viimasel aastal), mille kinnisvaraportfelli mahuks oli 35 miljonit eurot, laenukoormus tõusis aastaga 9 miljonilt eurolt 15 miljoni eurole (viimasel aastal võeti juurde 6 miljonit laenu).

Süsteem leidis tugeva laenuvõimendusega bilansi, mille kasvukiirus oli väga erinev (suurem) normaalsest ja tegi järelduse, et selline laenul toetuv kiire kasv võib olla riskantne. Järeldus on õige, sest kinnisvarahindade kukkudes võib makseraskuste risk (täiendavad pangatagatise ja sissemaksete nõuded) olla reaalne stsenaarium, samuti ei pruugi kinnisvara tagada laenukulu katmist.

### 2.4.4. Euclidean distant andmete analüüsis

Andmete suhtelise kauguse määramisel (testitavad andmed v. mudeli andmed; õpetamise andmed v. testitavad andmed) kasutatakse Euclidean distantsti:

$$d = \sqrt{[(x_2 - x_1)^2 + (y_2 - y_1)^2]}$$



Joonis 2.4.4.a.: Tulem distantismaatriks.

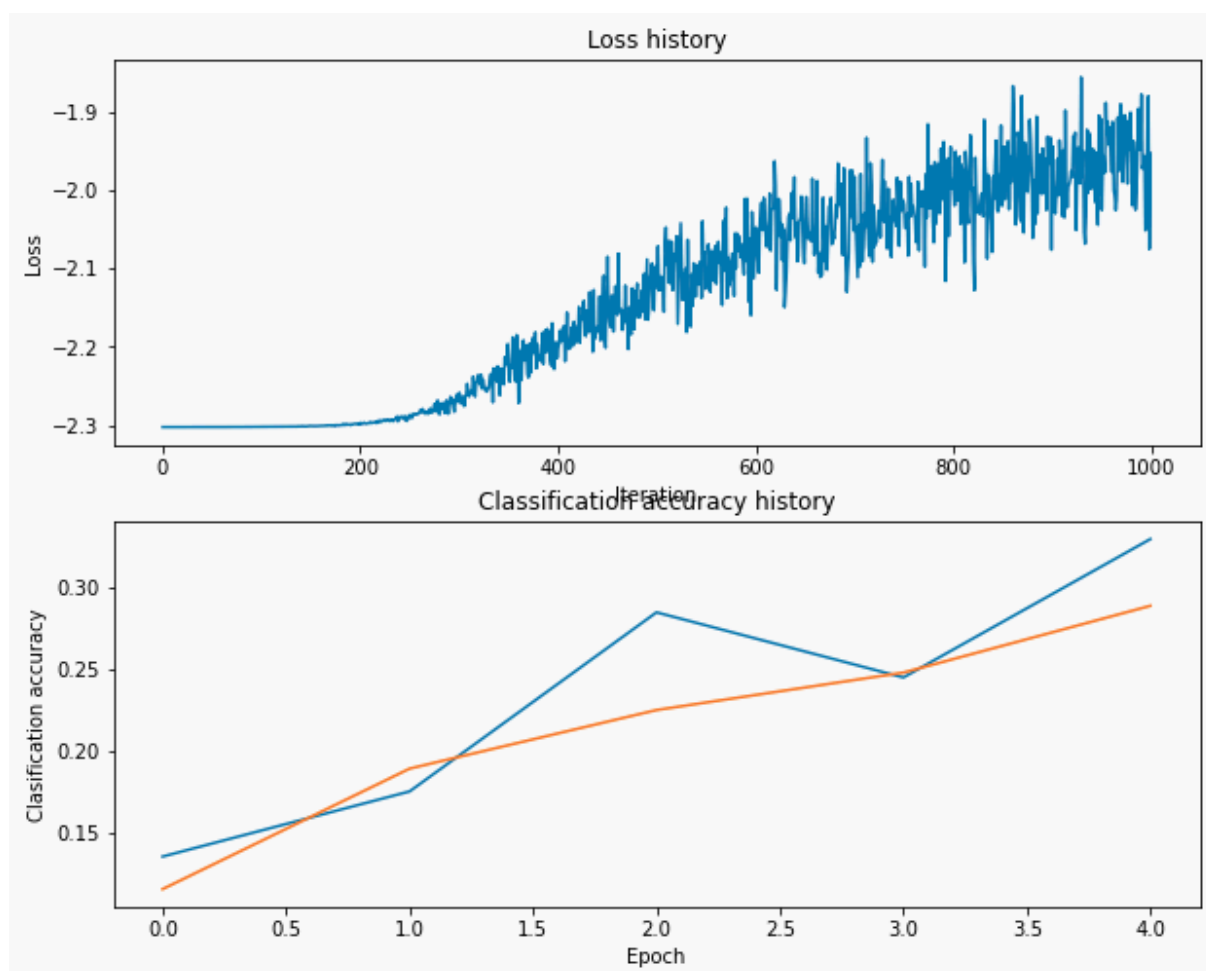


Ülaltoodud joonisel on toodud distantsmaatriks, mis näitab erinevust mudeli andmete ja testitavate andmete (ehk siis prognoosis kasutatavate sisendandmete) vahel. Tegemist on tööriistaga, mida käesolevas töös palju kasutatakse. Sellega saab vaadata treenimisandmete ja testimisandmete sarnasust, mudeliandmete ja prognoosiandmete erinevust, kontrollida prognoosiandmeid treenimisandmete vastu jne.

#### 2.4.5. Loss funktsiooni kasutamine mudeli optimeerimisel

Loss funktsioon ehk kulufunktsioon on kasulik tööriist mudeli optimeerimise juures. Mudeli täpsemaks muutmise (optimeerimise) eesmärgiks kulufunktsiooni minimeerimine. Käesoleva töö käigus on oluline optimeerida mudel (mudelid), mis annavad ettevõtetele hinnanguid.

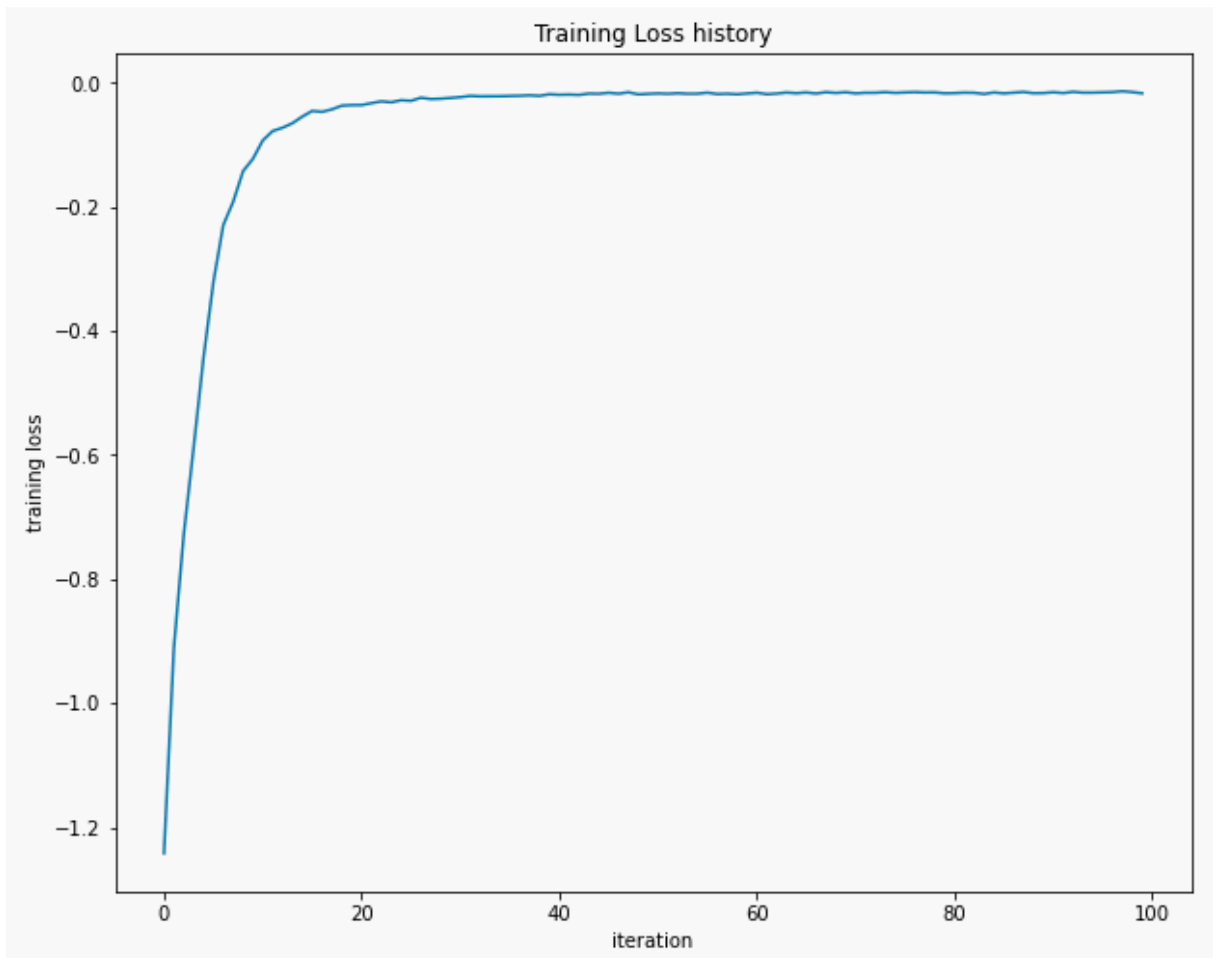
Alljärgneval joonisel on toodud „kuluajalugu“ ja ühtlasi ka funktsioonid, kuidas paralleelselt paraneb mudeli täpsus (Classification accuracy).



Joonis 2.4.5.a.: Loss ajalugu ja Classification accuracy.

Nagu ülaltoodud jooniselt nähtub kulu (loss) väheneb ja mudeli täpsus pareneb. Toodud näitajad ei ole lõppnäitajad, kuna mudelilt eeldatakse paremaid parameetreid. Samas on need illustratsiooniks, kuidas käesoleva töö raames mudelitega tööd tehakse ja üha paremaid lahendusi saada soovitakse.

Alljärgneval joonisel toodud training loss ajalugu teisel treenimisjuhul.



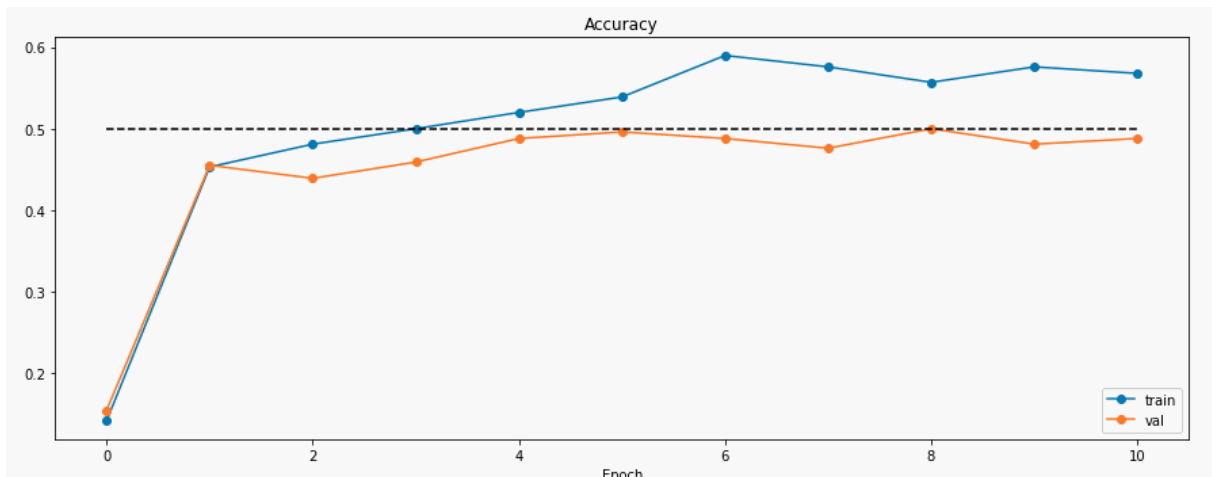
Joonis 2.4.5.b.: Loss ajalugu.

Allpool toodud loss funktsioon rohkemate iteratsioonide korral. Jooniselt on näha, et mida rohkem iteratsioone on, seda väiksemaks loss funktsioon läheb.



Joonis 2.4.5.c.: Loss ajalugu.

Alljärgneval joonisel on näha, kuidas accuracy on viidud juba märksa paremale tasemele (kui näiteks joonisel Joonis 2.4.5.a. – olgugi, et seal on toodud Classification accuracy). Toodud on nii treening kui ka valideerimisajalugu.



Joonis 2.4.5.d.: Treenimis ja valideerimisajalugu.

Alljärgnevalt koodinäide training loss ajaloo plottimiseks.

```
plt.subplot(2, 1, 1)
plt.title('Training loss')
plt.plot(solver.loss_history, 'o')
plt.xlabel('Iteration')

plt.subplot(2, 1, 2)
plt.title('Accuracy')
plt.plot(solver.train_acc_history, '-o', label='train')
plt.plot(solver.val_acc_history, '-o', label='val')
plt.plot([0.5] * len(solver.val_acc_history), 'k--')
plt.xlabel('Epoch')
plt.legend(loc='lower right')
plt.gcf().set_size_inches(15, 12)
plt.show()
```

Joonis 2.4.5.e.: Training loss plottimine.

Alljärgnevalt koodinäide lõpliku treenimislossi arvutamiseks ja plottimiseks.

```
net = init_toy_model()
stats = net.train(X, y, X, y, learning_rate=1e-1, reg=5e-6, num_iters=100, verbose=False)

print('Final training loss: ', stats['loss_history'][-1])

# plot the loss history
plt.plot(stats['loss_history'])
plt.xlabel('iteration')
plt.ylabel('training loss')
plt.title('Training Loss history')
plt.show()
```

Final training loss: -0.01713275892330469

Joonis 2.4.5.f.: Training loss lõplik hinnang ja plot.

Alljärgnevalt näide training loss minimiseerimisest ja accuracy parandamisest.

(Iteration 1 / 4900) loss: 2.300403  
(Epoch 0 / 10) train acc: 0.141000; val\_acc: 0.154000  
(Epoch 1 / 10) train acc: 0.453000; val\_acc: 0.455000  
(Iteration 501 / 4900) loss: 1.442970  
(Epoch 2 / 10) train acc: 0.481000; val\_acc: 0.439000  
(Iteration 1001 / 4900) loss: 1.455718  
(Epoch 3 / 10) train acc: 0.500000; val\_acc: 0.459000  
(Iteration 1501 / 4900) loss: 1.386509  
(Epoch 4 / 10) train acc: 0.520000; val\_acc: 0.488000  
(Iteration 2001 / 4900) loss: 1.374266  
(Epoch 5 / 10) train acc: 0.539000; val\_acc: 0.496000  
(Iteration 2501 / 4900) loss: 1.179609  
(Epoch 6 / 10) train acc: 0.590000; val\_acc: 0.488000  
(Iteration 3001 / 4900) loss: 1.315973  
(Epoch 7 / 10) train acc: 0.576000; val\_acc: 0.476000  
(Iteration 3501 / 4900) loss: 1.294193  
(Epoch 8 / 10) train acc: 0.557000; val\_acc: 0.500000  
(Iteration 4001 / 4900) loss: 1.371872  
(Epoch 9 / 10) train acc: 0.576000; val\_acc: 0.481000  
(Iteration 4501 / 4900) loss: 1.143015  
(Epoch 10 / 10) train acc: 0.568000; val\_acc: 0.488000

Alljärgnevalt näide parima accuracy otsimisest.

Training hidden\_size: 400  
Training learning\_rate: 0.003  
Training reg: 0.025  
Training batch\_size: 500  
iteration 0 / 1200: loss -2.302402  
iteration 100 / 1200: loss -1.649690  
iteration 200 / 1200: loss -1.509460  
iteration 300 / 1200: loss -1.383044  
iteration 400 / 1200: loss -1.513232  
iteration 500 / 1200: loss -1.415503  
iteration 600 / 1200: loss -1.431067  
iteration 700 / 1200: loss -1.428305  
iteration 800 / 1200: loss -1.259119  
iteration 900 / 1200: loss -1.285956  
iteration 1000 / 1200: loss -1.279291  
iteration 1100 / 1200: loss -1.136774  
Current val\_acc: 0.537  
Best Accuracy: 0.537  
Best Hidden Size: 400  
Best Learning Rate: 0.003  
Best reg: 0.025  
Best batch\_size: 500  
Training hidden\_size: 400  
Training learning\_rate: 0.003  
Training reg: 0.03  
Training batch\_size: 500  
iteration 0 / 1200: loss -2.302366  
iteration 100 / 1200: loss -1.741899  
iteration 200 / 1200: loss -1.589219  
iteration 300 / 1200: loss -1.432846  
iteration 400 / 1200: loss -1.482398  
iteration 500 / 1200: loss -1.392184  
iteration 600 / 1200: loss -1.401652

iteration 700 / 1200: loss -1.345791  
iteration 800 / 1200: loss -1.334304  
iteration 900 / 1200: loss -1.374942  
iteration 1000 / 1200: loss -1.112893  
iteration 1100 / 1200: loss -1.142935  
Current val\_acc: 0.508  
Training hidden\_size: 400  
Training learning\_rate: 0.003  
Training reg: 0.035  
Training batch\_size: 500  
iteration 0 / 1200: loss -2.302311  
iteration 100 / 1200: loss -1.713333  
iteration 200 / 1200: loss -1.517223  
iteration 300 / 1200: loss -1.536061  
iteration 400 / 1200: loss -1.389710  
iteration 500 / 1200: loss -1.371919  
iteration 600 / 1200: loss -1.359302  
iteration 700 / 1200: loss -1.442901  
iteration 800 / 1200: loss -1.327040  
iteration 900 / 1200: loss -1.130709  
iteration 1000 / 1200: loss -1.229156  
iteration 1100 / 1200: loss -1.193381  
Current val\_acc: 0.508  
Training hidden\_size: 400  
Training learning\_rate: 0.003  
Training reg: 0.05  
Training batch\_size: 500  
iteration 0 / 1200: loss -2.302266  
iteration 100 / 1200: loss -1.594231  
iteration 200 / 1200: loss -1.592442  
iteration 300 / 1200: loss -1.582597  
iteration 400 / 1200: loss -1.403937  
iteration 500 / 1200: loss -1.298283  
iteration 600 / 1200: loss -1.398085  
iteration 700 / 1200: loss -1.219892  
iteration 800 / 1200: loss -1.228358  
iteration 900 / 1200: loss -1.197194  
iteration 1000 / 1200: loss -1.280821  
iteration 1100 / 1200: loss -1.017234  
Current val\_acc: 0.486  
Final: Best Accuracy: 0.537  
Final: Best Hidden Size: 400  
Final: Best Learning Rate: 0.003  
Final: Best reg: 0.025  
Final: Best batch\_size: 500  
And finally: BE CAREFUL WITH THOSE NEURAL NETWORKS!!!

#### 2.4.6. Eksperimentaalstatistiline tehisnärvivõrkude lähtenurk

Käesoleva töö käigus kasutatakse ka eksperimentaalstatistilist tehisnärvivõrkudel põhinevat lähtenurka. Töö käesolevas faasis on raske öelda, kui võrd tehisnärvivõrkude kasutamine ennast antud töös õigustab, kuid alljärgnevalt on toodud näiteid, mida ja kuidas käesoleva töö raames neid kasutatakse.

Alljärgneval joonisel toodud Affine Forward koodi näidis.

```

# Test the affine_forward function

num_inputs = 2
input_shape = (4, 5, 6)
output_dim = 3

input_size = num_inputs * np.prod(input_shape)
weight_size = output_dim * np.prod(input_shape)

x = np.linspace(-0.1, 0.5, num=input_size).reshape(num_inputs, *input_shape)
w = np.linspace(-0.2, 0.3, num=weight_size).reshape(np.prod(input_shape), output_dim)
b = np.linspace(-0.3, 0.1, num=output_dim)

out, _ = affine_forward(x, w, b)
correct_out = np.array([[ 1.49834967,  1.70660132,  1.91485297],
                        [ 3.25553199,  3.5141327,  3.77273342]])

# Compare your output with ours. The error should be around 1e-9.
print('Testing affine_forward function:')
print('difference: ', rel_error(out, correct_out))

Testing affine_forward function:
difference: 9.7698500479884e-10

```

Joonis 2.4.6.a.: Affine Forward.

Alljärgneval joonisel toodud Affine Backward koodi näidis.

```

# Test the affine_backward function
np.random.seed(231)
x = np.random.randn(10, 2, 3)
w = np.random.randn(6, 5)
b = np.random.randn(5)
dout = np.random.randn(10, 5)

dx_num = eval_numerical_gradient_array(lambda x: affine_forward(x, w, b)[0], x, dout)
dw_num = eval_numerical_gradient_array(lambda w: affine_forward(x, w, b)[0], w, dout)
db_num = eval_numerical_gradient_array(lambda b: affine_forward(x, w, b)[0], b, dout)

_, cache = affine_forward(x, w, b)
dx, dw, db = affine_backward(dout, cache)

# The error should be around 1e-10
print('Testing affine_backward function:')
print('dx error: ', rel_error(dx_num, dx))
print('dw error: ', rel_error(dw_num, dw))
print('db error: ', rel_error(db_num, db))

Testing affine_backward function:
dx error: 1.0908210113205496e-10
dw error: 2.273805557790167e-10
db error: 7.736978834487815e-12

```

Joonis 2.4.6.b.: Affine Backward.

Alljärgneval joonisel toodud Relu Forward Function koodi näidis.

```

# Test the relu_forward function

x = np.linspace(-0.5, 0.5, num=12).reshape(3, 4)

out, _ = relu_forward(x)
correct_out = np.array([[ 0.,          0.,          0.,          0.],
                        [ 0.,          0.,          0.04545455, 0.13636364],
                        [ 0.22727273, 0.31818182, 0.40909091, 0.5]])

# Compare your output with ours. The error should be around 5e-8
print('Testing relu_forward function:')
print('difference: ', rel_error(out, correct_out))

Testing relu_forward function:
difference: 4.999999798022158e-08

```

Joonis 2.4.6.c.: Relu Forward Function.

Alljärgneval joonisel toodud Relu Backward Function koodi näidis.

```

np.random.seed(231)
x = np.random.randn(10, 10)
dout = np.random.randn(*x.shape)

dx_num = eval_numerical_gradient_array(lambda x: relu_forward(x)[0], x, dout)

_, cache = relu_forward(x)
dx = relu_backward(dout, cache)

# The error should be around 3e-12
print('Testing relu_backward function:')
print('dx error: ', rel_error(dx_num, dx))

Testing relu_backward function:
dx error: 3.2756349136310288e-12

```

Joonis 2.4.6.d.: Relu Backward Function.

```

from layer_utils import affine_relu_forward, affine_relu_backward
np.random.seed(231)
x = np.random.randn(2, 3, 4)
w = np.random.randn(12, 10)
b = np.random.randn(10)
dout = np.random.randn(2, 10)

out, cache = affine_relu_forward(x, w, b)
dx, dw, db = affine_relu_backward(dout, cache)

dx_num = eval_numerical_gradient_array(lambda x: affine_relu_forward(x, w, b)[0], x, dout)
dw_num = eval_numerical_gradient_array(lambda w: affine_relu_forward(x, w, b)[0], w, dout)
db_num = eval_numerical_gradient_array(lambda b: affine_relu_forward(x, w, b)[0], b, dout)

print('Testing affine_relu_forward:')
print('dx error: ', rel_error(dx_num, dx))
print('dw error: ', rel_error(dw_num, dw))
print('db error: ', rel_error(db_num, db))

Testing affine_relu_forward:
dx error: 6.750573928879482e-11
dw error: 8.162015570444288e-11
db error: 7.826724021458994e-12

```

Joonis 2.4.6.e.: Affine Relu Forward Function.

```

▶ np.random.seed(231)
num_classes, num_inputs = 10, 50
x = 0.001 * np.random.randn(num_inputs, num_classes)
y = np.random.randint(num_classes, size=num_inputs)

dx_num = eval_numerical_gradient(lambda x: softmax_loss(x, y)[0], x, verbose=False)
loss, dx = softmax_loss(x, y)

# Test softmax_loss function. Loss should be 2.3 and dx error should be 1e-8
print('Testing softmax_loss:')
print('loss: ', loss)
print('dx error: ', rel_error(dx_num, dx))

Testing softmax_loss:
loss: 2.302545844500738
dx error: 9.384673161989355e-09

```

Joonis 2.4.6.f.: Softmax Loss Function testing.

Alljärgnevatel joonistel toodud TwoLayerNet testing koodi näidis.

```

▶ np.random.seed(231)
N, D, H, C = 3, 5, 50, 7
X = np.random.randn(N, D)
y = np.random.randint(C, size=N)

std = 1e-3
model = TwoLayerNet(input_dim=D, hidden_dim=H, num_classes=C, weight_scale=std)

print('Testing initialization ... ')
W1_std = abs(model.params['W1'].std() - std)
b1 = model.params['b1']
W2_std = abs(model.params['W2'].std() - std)
b2 = model.params['b2']
assert W1_std < std / 10, 'First layer weights do not seem right'
assert np.all(b1 == 0), 'First layer biases do not seem right'
assert W2_std < std / 10, 'Second layer weights do not seem right'
assert np.all(b2 == 0), 'Second layer biases do not seem right'

print('Testing test-time forward pass ... ')
model.params['W1'] = np.linspace(-0.7, 0.3, num=D*H).reshape(D, H)
model.params['b1'] = np.linspace(-0.1, 0.9, num=H)
model.params['W2'] = np.linspace(-0.3, 0.4, num=H*C).reshape(H, C)
model.params['b2'] = np.linspace(-0.9, 0.1, num=C)
X = np.linspace(-5.5, 4.5, num=N*D).reshape(D, N).T
scores = model.loss(X)
correct_scores = np.asarray(
    [[11.53165108, 12.2917344, 13.05181771, 13.81190102, 14.57198434, 15.33206765, 16.09215096],
     [12.05769098, 12.74614105, 13.43459113, 14.1230412, 14.81149128, 15.49994135, 16.18839143],
     [12.58373087, 13.20054771, 13.81736455, 14.43418138, 15.05099822, 15.66781506, 16.2846319 ]])
scores_diff = np.abs(scores - correct_scores).sum()
assert scores_diff < 1e-6, 'Problem with test-time forward pass'

print('Testing training loss (no regularization)')
y = np.asarray([0, 5, 1])
loss, grads = model.loss(X, y)
correct_loss = 3.4702243556
assert abs(loss - correct_loss) < 1e-10, 'Problem with training-time loss'

```

Joonis 2.4.6.g.: TwoLayerNet testing 1.



```

model.reg = 1.0
loss, grads = model.loss(X, y)
correct_loss = 26.5948426952
assert abs(loss - correct_loss) < 1e-10, 'Problem with regularization loss'

for reg in [0.0, 0.7]:
    print('Running numeric gradient check with reg = ', reg)
    model.reg = reg
    loss, grads = model.loss(X, y)
    for name in sorted(grads):
        f = lambda _: model.loss(X, y)[0]
        grad_num = eval_numerical_gradient(f, model.params[name], verbose=False)
        print('%s relative error: %.2e' % (name, rel_error(grad_num, grads[name])))
        assert rel_error(grad_num, grads[name]) < 0.6, "Error with gradient for " + name

```

```

Testing initialization ...
Testing test-time forward pass ...
Testing training loss (no regularization)
Running numeric gradient check with reg = 0.0
W1 relative error: 1.52e-08
W2 relative error: 3.30e-10
b1 relative error: 8.37e-09
b2 relative error: 1.34e-10
Running numeric gradient check with reg = 0.7
W1 relative error: 2.53e-07
W2 relative error: 7.98e-08
b1 relative error: 1.56e-08
b2 relative error: 7.76e-10

```

Joonis 2.4.6.h.: TwoLayerNet testing 2.

Täiendav kood on toodud lisades 9-11.

## Peatükk III – töö täiendamise käigus aset leidnud tegevused

Varase hoiatamise teenuse ärianalüüsi ja prototüübi väljatöötamise lõpparuande esimene versioon esitati 31. märtsil 2022. aastal. 12. aprillil 2022. aastal toimus vastavateemaline ümarlaud, kus arutati prototüübi seisu ja edasisi võimalikke arengusuundi. 22. aprilliks oli määratud tagasiside andmise tähtaeg.

Käesoleva materjali peatükk III kirjeldab:

- mida on töö raames tehtud täiendavalt peale 31. märtsi 2022. aasta raporti esitamist, ehk siis kuidas on täiendatud prototüüpi võttes arvesse seda, mis oli teada märts 2022. aasta lõpuks, millised pudelikaelad olid avaldunud ja mis suunas oli plaanitud edasi töötada;
- kuidas on töö raames arvestatud ja millisel määral on arvestamisel 12. aprillil arutatud seisukohad;
- kuidas on töö raames arvestatud ja millisel määral on arvestamisel 22. aprilliks määratud tähtjaks laekunud tagasiside;
- muud olulised tegevused ja seisukohad, mida on läbi viidud ja millega on arvestatud aprillis 2022.a.

Käesolevale materjalile on lisatud ka kokkuvõttev peatükk. Osad seisukohad ja järeldused, mis on tehtud aprillis 2022.a. saadud tagasiside tulemusena, ei ole kajastatud mitte käesolevas III peatükis, vaid kokkuvõtvas peatükis. Materjali jaotus käesoleva III peatüki ja kokkuvõtva peatüki vahel lähtub sellest, et kokkuvõttes on toodud eelkõige kompaktset ülevaadet andvad mõtted ja tulevikku suunatud lähtenurgad. Info terviklikkuse mõttes on III peatükk ja kokkuvõttev peatükk teineteist täiendavateks.

### 3.1. Täiendavate majandussektorite mudelite koostamine

Käesoleva materjali lisa 1 on toodud neli defineeritud tegevusvaldkonda: (a) meditsiinivaldkond, (b) finantsteenuste valdkond, (c) kinnisvaravaldkond ja (d) haridusvaldkond. Peamiselt on senimaani töötatud kinnisvaravaldkonna analüüsiga. Tegemist on mõneti juhusliku valikuga põhimõttel, et „kusagilt peab pihta hakkama“. Samas on kinnisvaravaldkond valitud alustuseks seetõttu, et tegemist on paratamatult (sõltuvalt, kui laialt seda defineerida, vt. lisa 1) Eesti majandusruumi oluliselt mõjutava majandussektoriga. Varase hoiatuse süsteemi korrektseks toimimiseks tuleb prototüüpi katsetada erinevates majandussektorites.

Prototüüp võib hästi toimida mõnes majandussektoris, kuid mitte sama hästi mõnes teises. Oluline on saada korrektseid vastuseid nii väikestes nishisektorites kui ka suurema haardeulatusega majandusharudes, sest vastasel juhul ei ole väljastatavatest hoiatustest kasu ning nad ei täida oma eesmärki. Süsteemi arendamist on alustatud suurtest sektoritest, kuid see ei tähenda, et väiksemaid sektoreid ignoreeritakse või vähemtähtsateks peetakse. Praegusel juhul on sektorid jaotatud, nagu alljärgnevalt näidatud, kui testimise käigus võib tekkida vajadus sektorite teistsuguse defineerimise järele.

Aprillis 2022.a. tehti mudelitega (s.t. erinevate algoritmidega) katsetusi lisa 1 toodud sektorite andmetel:

- meditsiinivaldkond;
- finantsteenuste valdkond;
- haridusvaldkond.

Täiendavalt defineeriti aprillis ja mudelitega prooviti katsetada andmetega järgmistest majandussektoritest:

- majutusteenused;
- toitlustus;
- meelelahutus;
- põllumajandus;
- infotehnoloogia;
- elektritootmine;
- hulgikaubandus;
- jaekaubandus;
- logistika.

Osade ülaltoodud valdkondade defineerimise ja nende valdkondade andmetega saadi selgemaid tulemusi, teistega vähem selgeid tulemusi. Valdkonnad on ka väga erinevad oma erinevate probleemidega. Näiteks meditsiinivaldkond on tugevalt avaliku sektori valdkond. Finantsteenuste puhul on valdkonda kaasatud keskpang, kommerts pangad, kindlustusettevõtjad, finantsnõustajad, valuutavahetajad (kes kõik on väga erineva ettevõtjad oma olemuselt ja keskpang ei ole üldse ettevõtja). Elektritootmisega tegeleb nii Eesti Energia kui ka palju väikeseid tootjaid (jällegi väga erinevad ettevõtted). Ka põllumajandustootjad on tulenevalt oma tegevusvaldkonnast väga erinevad (piimatootjate puhul kehtivad ühed reeglid, teraviljakasvatajate puhul teised).

Ülaltoodu on näited mõningatest tehnilistest asjaoludest, mis on ilmnenu (ja mis on sellise töö juures paratamatu) täiendavate valdkondade defineerimise tulemusena. Prototüübi katsetamise ja edasise arendamise juures on oluline nendele probleemidele tähelepanu pöörata ja neid järk-järgult lahendada (kas siis sektoreid täpsustada või mudeleid vastavalt häälestada või mõlemat). Oluliseks abistavaks

asjaoluks mudeli täpsustamisel on reaaleluliste testide läbiviimine (vt. alapeatükk „Reaaleluliste testide ettevalmistamine“).

### **3.2. Reaaleluliste testide ettevalmistamine**

Aprillis 2022.a. on töö raames aktiivselt tegeletud reaaleluliste testide ettevalmistamisega. Reaaleluliste testide läbiviimine tähendab seda, et võetakse reaalsed ettevõtete andmed ja edastatakse need prototüübile (andmed sisestatakse mudelisse). Prototüüp teeb ettevõtte kohta oma järelduse ja väljastab selle. Prototüüp ütleb, et kas ettevõtte peaks saama mingit liiki hoiatuse (hoiatusel on kolm taset) ning ühtlasi väljastab prototüüp ka täiendava informatsiooni selles osas, miks on vastav hoiatus ettevõttele väljastatud.

Reaaleluliste testide korraldamise osas läbi viidud arutelud on viinu selleni, et teste on mõistlik ja optimaalne korraldada koos Harju Ettevõtlus- ja Arenduskeskus'ega. Harju Ettevõtlus- ja Arenduskeskus on olnud väga abivalmis ja asjalik kaasamõtteleja, kuidas prototüüp saaks elujõuliseks ja kuidas reaalsel ettevõtjatel sellest kõige rohkem kasu oleks. Keskus saab mängu tuua oma praktilise kompetentsuse ja sidemete võrgustiku aitamiseks ühelt poolt jõuda õige testgrupini ning teiselt poolt saada ka täiemahulist tagasisidet, mida arvavad testitavad ettevõtjad ja kuidas nad näevad sellisest hoiatusest endale parimat võimalikku kasu tõusvat.

Käesolevas materjalis ollakse seisukohal, et Harju Ettevõtlus- ja Arenduskeskus on heaks, õigeks ja igas mõttes optimaalseks (parimaks) partneriks prototüübi testimisel reaalsete ettevõtjate (ettevõtete) peal koos vajaliku tagasiside saamisega nendelt ettevõtjatelt.

### **3.3. Prototüübi järk-järguline parandamine**

12. aprillil 2022. aastal toimunud ümarlinal leidis muuhulgas käsitlemist asjaolu, et masinõppe mudelitel põhinev lahendus varase hoiatuse tarvis on oma olemuselt lahendus, mis õpib järk-järgult ning saab targemaks vastavalt sellele, kui palju seda kasutatakse (kui palju mudel sisendit saab ja selle siis endale salvestab – võtab edaspidiste otsuste tegemisel aluseks).

Seda asjaolu on silmas peetud aprillis 2022. aastal aset leidnud mudeli täiendamisel ja parendamisel. Mudelit on treenitud üha asjakohasemate andmete ja mudelile on õpetatud erinevaid majandussektoreid. Rakendatud on lahendusi, mis proovivad välistada n.ö. valepositiivsete tulemuste teket (s.t. selliste tulemuste teket, kus masin annab hoiatuse, kuid hoiatuseks tegelikult ei ole põhjust – ettevõtte võib teatud mustriks eristuda ja seetõttu tähelepanu alla sattuda, kuid see eristumine ei pruugi veel olla mingilgi määral viide võimalikule maksejõuetusele).

Mudeli täiendamine leiab aset ka siis, kui hakatakse ellu viima (ülalmainitud) reaalelulisi teste. Mudel saab tagasisidet, kas ta käitub õigesti ja vajadusel saab mudelit täiendavalt õpetada. Seetõttu, nagu ka 12. aprilli kohtumisel räägitud, võib mudel alguses olla vähemtäpne ja ka testimise käigus oma täpsusastet suurendada. Mudeli testimiseelse ettevalmistamise käigus plaanitakse prototüüp viia sellele tasemele, et testimine annaks juba koheselt positiivseid tulemeid, kuid konservatiivse stsenaariumi kohaselt peab arvestama ka sellega, et mudeli häälestamine jätkub testimise käigus ja testimistulemuste põhjal peale testide läbiviimist.

### **3.4. Tehnilised lahendused prototüübi kasutamiseks**

Aprillis 2022.a. on töötatud tehniliste lahenduste kallal selleks, et prototüüp oleks kasutatav, s.t. eelkõige testitav ja kui testid viivad sobivale tulemusele, siis edaspidiselt kasutatav alusena infosüsteemi loomiseks.

Koostöökokkuleppes ei ole nõutud (ja ollakse oldud seisukohal), et tuleks luua kasutajaliides. Seega ei ole mudel nähtav/kasutatav kolmandatele osapooltele (ka neile, kes tööd jälgivad ja testimisega otseselt seotud on) võrguaadressi vahendusel. Kasutajaliidese, eriti turvalisel kujul, ehitamine konfidentsiaalsusnõuetele alluvate andmete peale on arvestatava mahuga töö ja pidades vajalikuks eelkõige töötunde suunata mudeli enda tegemisele, ei ole kasutajaliidese loomist vajalikuks peetud.

Samas on olnud vajalik (ja sellega on ka tegeletud), et prototüübi mudel oleks kasutatav ja testitav. Esmase ehk siis baaslahendusena pakutakse prototüübi testimiseks lahendust, kus testitavad andmed edastatakse Statistikaametile ja prototüübi arendamisega tegelevad inimesed lasevad prototüübi mudelil anda hinnangu, mida prototüübi mudel antud andmetest arvab. Sellisel kujul ei ole sisendi saamise formaat oluline, prototüübiga tegelevad inimesed saavad viia andmed prototüübile sobivale kujule.

Teise lahendusena (lisaks ülalkirjeldatud baaslahendusele) ikkagi kaalutakse prototüübi arendamise poole pealt seda, et luuakse lahendus, kus kasutajaliidese läbi saavad autoriseeritud (prototüübi arendamist jälgivad inimesed ja testimisega seotud inimesed) kasutajad läbi veebibrowseri sisestada testitavad andmed ja saada mudeli poolt vastus (s.t. kas nende andmete pinnalt väljastatakse mingit liiki hoiatus ja kui väljastatakse, siis millist liiki hoiatus koos vastavate selgitustega). Eelnevalt kirjeldatud põhjustel nimetatud kasutajaliidest ei loeta süsteemi baasversiooniks, kuid arvestades asjaoluga, et kasutajaliides annaks oluliselt juurde efektiivsust süsteemi testimisel (süsteem saab testimise tulemusena ka õppida, mida rohkem testitakse, seda paremaks muutub süsteem), siis kasutajaliidesega tegeletakse ka aktiivselt (kasutajaliidesega tegeleti aktiivselt aprillis 2022. aastal ning jätkatakse tegevust ka mais 2022. aastal).

### **3.5. Andmete kasutamine**

Prototüübi loomist alustati bilansi- ja kasumiaruande andmete pinnalt. Identifitseeriti bilansse ja kasumiaruandeid, mis võiksid viidata probleemidele (või potentsiaalsetele probleemidele) maksevõime vähenemise kontekstis. Masinõppe meetodil analüüsiti bilansi- ja kasumiaruande struktuure, neid omakorda verifitseeriti klassikaliste majandusanalüüsi suhtarvudega (kasutatud suhtarvud toodud käesoleva materjali lisas 2).

Teise suunana prototüübi loomisesse ja arendamisesse toodi sisse ettevõtete vaheliste transaktsioonide andmed. Analüüsimudelid rajati sellele, et kes ja kui palju, millise sagedusega, millistes summades, milliste mustritega kellele maksab, kellel on kui palju tehingupartnereid ja kuidas see tehingupartnerite arv muutub ajas või siis, kui vaadelda teatud osa maksete alajaotusi (näiteks vaid suuri makseid jne.).

Kui prototüübi loomise esimeses faasis oli peamine rõhk bilansi- ja kasumiaruande andmetel, siis aprillis 2022.a. pandi peamist rõhku transaktsiooniandmete mudelite loomisele, mis võimaldaks suhteliselt lakooniliste transaktsiooniandmete pinnalt teha sisulisi järeldusi ettevõtte maksevõime muutumise kohta (transaktsiooniandmed on mõnes mõttes „vähemütlevad“, kui näiteks konkreetset bilansilised näitajad – näiteks suur kohustuste maht, ajas suurenev kohustuste maht, võib viidata väga konkreetset maksevõime vähenemisele, samas tehingupartnerite arvu vähenemine ei pruugi viidata kaduvale turule, ettevõtte lihtsalt kontsentreerub teatud kliendisegmendile, juurde tuleb tuua tehingute mahud klientidega). Saamaks mudeleid „mõistlikku juttu rääkima“ transaktsiooniandmete pinnalt on märksa keerukam, kui bilansiandmete pinnalt. Samas transaktsiooniandmed on sisuliselt igakuiselt saada olevad andmed, kui majandusaasta aruannetest pärineva bilansid ja kasumiaruanded saavad andmeanalüüsiks kättesaadavaks tänasel päeval sageli aasta-poolteise viivitusega (sellises kontekstis on raske rääkida „varasest hoiatamisest“ – andmed on lihtsalt vanad selleks).

Kui Eestis juurutatavate reaalajamajanduse lahenduste tulemusena hakkab tekkima lahendusi, kus majandusaasta aruanded (või nende komponendid) hakkavad andmebaasidesse jõudma reaalajas (või suurusjärgus ühe kalendrikuu pikkuste tähtaegadega), siis on võimalik varase hoiatuse signaale nende pinnalt väljastada. Senikaua kui seda veel ei juhtu, senikaua on oluline teha põhjalikku tööd transaktsiooniandmetega. Käesoleva prototüübi antud kontekstis on oluline rõhutada, et ühelt poolt proovib prototüüp panustada reaalajamajanduse arengusse (transaktsiooniandmete kasutamisega näiteks), teiselt poolt on projekti jaoks väga oluline, kui reaalajamajanduse lahendused areneksid (siis on prototüübi ja edasise infosüsteemi jaoks olemas üha adekvaatsemat sisendit hoiatuste avaldamiseks).

Aprillis 2022.a. hakati mudelis lisaks majandusaasta aruannete andmetele testima ka ekspordi andmeid. Osade sektorite ekspordikäivate osakaal on arvestatavalt suur ja seeläbi ei ole ainult Eestis tehtava käibe alusel võimalik adekvaatseid seisukohti nende ettevõtete maksevõimekuse muutuse kohta anda, samas peab prototüüp (hilisem infosüsteem) suutma väljastada varaseid hoiatusi ka nende ettevõtete osas, mis tegelevad suurel määral oma toodangu/teenuste ekspordiga.

Olulise andmesisendina, mille lisamist prototüübile on alustatud ja aprillis 2022.a. esmaselt testiti, on tööturгу puudutavad andmed. Töötajate arvu suurenemine või vähenemine, ettevõtte suutlikkus maksta konkurentsivõimelist palka ja palgatavate töötajate kvalifikatsioon (varasem kogemus, haridus) on näitajad, mis nii staatiliselt kui ka dünaamiliselt (kui nende näitajate osas on muutuseid) sisaldavad informatsiooni ettevõtte maksevõimekuse kohta.

Aprillis 2022.a. tegeleti ka sellega, et saada prototüübi mudelisse sisendit maksuvõlgade andmete pealt. Samas maksuvõlgade andmetele ei ole töö käigus veel juurdepääsu saadud ning tulenevalt sellest ei ole neid mudelisse lisatud. Samas, kui see võimalus tekib ja mudeli arendamine on paratamatult alles töös, siis lisatakse ka maksuvõlgu puudutavad andmed. Maksuvõlad on üldjuhul näitajaks vähenenud maksevõimest (kui maksevõimega ei ole probleeme, siis tavaliselt maksuvõlgu ei teki, võlgu ollakse makse ikka sellepärast, et raha on vähe, mitte et põhimõtteliselt tahetaks võlgu olla) ja maksuvõlgade muutuse dünaamika võimaldab anda hinnanguid ettevõtte maksevõime muutusele.

Aprillis 2022.a. alustati tööd ka kolmandate andmete lisamise eelduste loomiseks. Kolmandate andmete all peetakse silmas alljärgnevat:

- aadressandmed: võimalus õpetada prototüübile asukohaspetsiifilisi asjaolusid (näiteks turism Tallinnas on midagi muud, kui turism Põlvas, Pärnu ja Kuressaare on spetsiifilised turismilinnad);
- ilmastikuandmed: võimalus õpetada prototüübile ilmastiku spetsiifikast tulenevaid asjaolusid, näiteks põllumajandustootmise kontekstis on oluline, et kas tootmine on Järvamaal, saartel või Kagu Eestis;
- rahvastikuandmed: võimalus õpetada prototüübile seda, kus inimesed elavad, kus rohkem sünnitakse, kus on rohkem surmasid – siit tulenevalt on võimalik näidata turu kujunemist ja tuua sisse sellest tulenevaid näitajaid;
- elektriandmed: tunni ajalise täpsusega olemas olevad elektritarbimiste näidud kõigi mõõtepunktide lõikes võimaldavad reaalajas hinnata majandusaktiivsust (elukondlikku aktiivsust), mis annab reaalajamajanduse kontekstis võimaluse astuda kvalitatiivne samm edasi saamaks sisendit reaalajas; prototüübile on võimalik need sisendid oluliste parameetritena sisse tuua.

### **3.6. Vigade kõrvaldamine**

Aprillis 2022.a. on tegeletud ka vigade kõrvaldamisega. Vigade all peetakse siinkohal silmas eelkõige mudeli poolt antavaid ebaõigeid tõlgendusi või andmesisendi ebaadekvaatseid lahendusi mudelile. Mõlemat liiki vead on paratamatult töö käigus tekkivad ebatäpsused, kuid töötava ja infosüsteemi

loomise aluseks kasutatava prototüübi saamiseks tuleb need vead kõrvaldada (nende tekkimist pidevalt jälgida ja siis nende tekkimise korral need vead kõrvaldada).

Ühte liiki tüüpilise probleemasjaoluna on esinenud olukorda, kus mudel võtab sisendina pakutavast informatsioonist mõnda sisendit ebaproportsionaalselt suure osakaaluga ja tulemused on selles suunas kallutatud (omamata aga selleks mingit objektiivset majanduslikku sisu). Näiteks transaktsioonandmete puhul tekkis olukord, kus aastaarv sai liiga suure osakaalu, tulemusena tekkis näiteks selline järeldus (mis arusaadavalt on ebaõige), et „kui oled käivet teinud aastatel 2017 ja 2018 alguses, siis on sul potentsiaalne kalduvus sattuda 2022. aastal maksejõuetuse olukorda.“

Teist liiki tüüpilise probleemasjaoluna on esinenud olukorda, et kui määratleda majandussektor ja siis sealt omakorda eraldada rohkem ja vähemprobleemseid sektoreid nii bilansianalüüsi kui ka transaktsioonandmete analüüsi alusel, siis sisendandmete hulk masinale vastava mudeli õpetamiseks jääb liiga väikeseks.

## KOKKUVÕTE

Käesoleva töö käigus on välja töötatud prototüüp ja kontseptuaalsed lahendused varase hoiatamise teenuse käivitamiseks Eestis. Käesolev materjal esitab nii esmased seisukohad kui ka põhimõtted, millest varase hoiatamise teenus saaks ja peaks juhinduma. Toodud on kirjeldus, mida on tehtud, milliseid probleeme on töö käigus tulnud kohata ning millisel viisil on neid lahendatud. Ühtlasi on antud selgitused ka selle kohta, miks on mindud täpselt sellist teed prototüübi väljatöötamisel, nagu on mindud.

Alljärgnevalt on käesolevas kokkuvõttes toodud ära mitmed võtmeseisukohad, mida välja töötatud varase hoiatuse teenuse prototüübi juures on oluline arvesse võtta ning millest soovitatakse lähtuda edasise töö osas varase hoiatamise teenuse juurutamisel Eestis.

### ***Prototüübi testimine***

Varase hoiatamise teenuse prototüüp on kontseptuaalselt läbi töötatud ja see on ka esmaselt testitud. Praeguses faasis ollakse jõudnud sinnamaale, et oleks võimalik hakata läbi viima reaalelulisi teste. Reaaleluliste testide all peetakse silmas seda, et ettevõtluskonsultandid (Harju Ettevõtlus- ja Arenduskeskus) edastavad prototüübile reaalselt tegutsevate ettevõtete majandusandmeid ja prototüüp annab vastuseid, et kas tuleks väljastada varane hoiatus, kui jah, siis milline hoiatus tuleb väljastada ning hoiatuse väljastamise korral väljastatakse ka selgitus, miks see hoiatus väljastati. Ettevõtluskonsultandid viivad ettevõttega läbi intervjuu ja kontrollivad seeläbi, kas väljastatud hoiatus (juhul, kui see väljastati) on ettevõtte seisukohast asjakohane. Intervjuu käigus saadakse tagasiside, mida on võimalik kasutada prototüübi edasiseks täiendamiseks.

Siinkohal kirjeldatud testimise viis on formuleerunud prototüübi osas peetud arutelude tulemusena ja käesolevaga ollakse seisukohal, et tegemist on efektiivse ja asjakohase viisiga prototüübi testimisel. Samuti ollakse siinkohal seisukohal, et Harju Ettevõtlus- ja Arenduskeskus on kompetentne ja asjakohane organisatsioon testimiste läbiviimiseks. Samas, kui prototüübi testimisel võib tulla päevakorrade mõni teine meetod või täiendavalt kirjeldatule on otstarbekas kasutada mõnda teist ja täiendavat lahendust, siis ollakse siinkohal valmis alternatiivseid testimise variante päevakorda võtma.

### ***Konsultantide infovaate loomine***

Prototüübi osas peetud arutelude käigus on tekkinud ideed, et kui varase hoiatamise süsteem väljastab hoiatuse, siis väljastab see ka selgituse selles osas, miks hoiatus väljastati ja mida selle hoiatusega peale hakata. On avaldatud arvamust, et ettevõtja ei pruugi olla kompetentne (ettevõtjate taustad on väga erinevad) selle teate sisu ja konteksti lahti mõtestama ja oluline oleks konsultandi kaasamine.

Samas, kui tekib konsultandi kaasamise olukord, siis oleks ehk mõistlik sisse tuua täiendav infoaken konsultandi tarvis (konsultandi vaade, konsultandile suunduv kommenteeritud koond vms.), mis aitab konsultandil ettevõttele selgitada, milles seisneb tema probleem ja kuidas on võimalik seda probleemi lahendada, mida ette võtta probleemi lahendamise suunas jne. Ehk siis ollakse siinkohal seisukohal, et edasi peaks töötama selles suunas, et tagada täiendavat informatsiooni juhaks, kui ettevõtte vajab hoiatuse tõlgendamisel konsultandi abi. Sellisel juhul peaks konsultant nägema kogu ettevõttele suunduvat informatsiooni ja saama ka ehk veel täiendava selgitava sisendi rohkemates elementides (õigusliku koha pealt peab siis ettevõtte selleks nõusoleku andma?).



Oluline on siinkohal rõhutada, et varane hoiatus suunatakse üksnes ettevõtte volitatud isikule. Konsultantide kaasamine saab olla mõeldav peale seda, kui ettevõtte volitatud isik ise avaldab selleks soovi ja annab vastava nõusoleku. Konsultantide kaasamine on võimalus, mitte automaatselt kaasnev protsess. Eelkõige on konsultantide kaasamine mõeldav ja vajalik just keerukamate olukordade puhul – teatud juhtudel on varase hoiatuse olemus väga ilmne ja üheselt mõistetav, teistel juhtudel on selle sisu mõistmine keerukam nagu ka on keerukam mõista seda, et mida täpselt peaks tegema selleks, et varase hoiatuse alusprobleemi lahendada.

### ***Prototüübi edasiarendus***

Prototüüp on välja töötatud eesmärgiga olemaks lähtealuseks varase hoiatamise infosüsteemi väljatöötamiseks. Seda on prototüübi väljatöötamisel silmas peetud ja sellest on lähtutud. Samas on prototüübi väljatöötamise käigus (ja kaasnenud arutelude käigus) ilmnunud mitmed ideed ja lähtenurgad, mida võiks täiendavalt prototüübi edasiarendusse kaasata ning mis võiksid saada välja töötatava infosüsteemi olulisteks komponentideks. Hetkel on tekkinud kaks peamist ideed.

Esiteks võiks prototüüpi ja siis ka hiljem infosüsteemi arendada välja suunas, et kui süsteem väljastab ettevõttele varajase hoiatuse, siis ühtlasi saab süsteem ettevõtjale väljastada ka stsenaariumeid, et juhul, kui ettevõtja käitub näiteks ühel viisil, siis maksejõuetus süveneb ja kuhu siis lõpuks jõutakse, kuid samas, kui ettevõtja käitub teisel viisil, siis hoopis maksevõime paraneb. Varase hoiatuse eesmärk on panna ettevõtjaid mõtlema ja osutada probleemidele tähelepanu ning seeläbi tagada olukord, kus ettevõtlussektori üldine maksevõime tase paraneks. Stsenaariumid edasiseks tegevuseks aitavad ettevõtjal lahti mõtestada ja aru saada, et mida täpselt siiski peaks tegema, et probleeme lahendada, kui hoiatus on saadud (hoiatusel võib olla ju juures selgitus, et miks hoiatus väljastati, kuid sugugi mitte kõik ettevõtjad ei suuda veel selgituse pinnalt ise edasisi arengustsenaariumeid prognoosida). Varajase hoiatuse mudelites on sees palju olulist informatsiooni ja seda tuleb sinna töö käigus järjest juurde, see on väga heaks vundamendiks edasiste stsenaariumite väljapakkumiseks.

Teiseks on võimalik prototüüpi edasi arendada konsultandi vaate arendamise suunas, s.t. mida konsultandi vaatesse lisada ja milliseid tööriistu konsultandile täiendavalt anda, et konsultant saaks ettevõtet parimal võimalikul viisil toetada ja varajases hoiatusest oleks seeläbi kõige rohkem praktilist kasu.

Kahe ülaltoodud idee valguses tuleks võimalusel kaaluda edasist tööd viisil, et paralleelselt infosüsteemi väljaarendamisega tegeletakse ka prototüübi edasise arendamisega ülalmainitud suundades. Need kaks tööd (infosüsteemi loomine ja prototüübi paralleelne edasiarendus) teineteist ei sega ja saavad aset leida samaaegselt. Kõigepealt käivitub infosüsteem n.ö. baasversioonis, hiljem hakatakse sinna lisama täiendavaid komponente.

### ***Varajase hoiatuse lahenduse kasutamine teistes riikides***

Varajase hoiatuse sisseviimise probleem ei ole küsimuseks ainult Eestis, vaid ka teistes riikides. Käesoleva töö raames välja töötatav lahendus kasutab maksimaalsel määral infotehnoloogia viimaseid arenguid (masinõpe, tehisnärvivõrgud jms.). Selline lähenemine varajase hoiatuse teenuse küsimusele on innovatiivne ja sellest omakorda tulenevalt ei pruugi teised riigid olla oma vastavas tegevuses analoogsel viisil lahendatud teenuse käivitamise lävel. Isegi, kui mõnes riigis on kasutatud tehisnärvivõrke ja masinõpet analoogiliste mudelite loomiseks, siis tulemuseks võivad olla teistsugused mudelid (mudeleid treenitakse konkreetsete andmete pealt, määratletakse sisendid jne., selles küsimuses on võimalik minna väga erinevat arenguteed pidi).

Lähtuvalt ülaltoodust ollakse siinkohal seisukohal, et välja töötatav varajase hoiatuse prototüüp ning sellel põhinev infosüsteem võivad olla toodeteks, mis pakuvad huvi ka teistes riikides ning seda informatsiooni saab Eesti pakkuda omapoolse panusena rahvusvahelisse koostöösse.

Treenitud mudelid ja nende tööks vajalik infotehnoloogiline raamistik on sisuliselt arvutitarkvara ja see on nii skaleeritav kui ka jagatav (vajadusel kohandatav). Erinevates riikides on väikesed erinevused, kuid üldjoontes on lahendamist vajavad probleemid sarnased.

Statistikaameti seisukohalt vaadates võib siinkohal väita, et on olemas kogemus erinevate riikide statistiliste andmetega (peamiselt läbi Eurostati koostöö) ja selle kogemuse/teadmise põhjal on võimalik väita, et mudelit on võimalik toita andmetega väga erinevatest riikidest.

Ülaltoodust lähtudes on siinkohal oluline väita, et analoogiliselt Eestile on ka teistes riikides vajadus analoogilise teenuse järgi ja võttes arvesse käesoleva projekti raames tehtut oleme olukorras, kus Eesti saab pakkuda oma lahendust partnerriikidele.

Juhul, kui plaanitakse edasi liikuda ideega, et tutvustada käesoleva projekti raames tehtavat ka teistele riikidele, siis sellel eesmärgil on plaanis arendada välja veebiaadress, kus rahvusvaheliselt mõistetavas keeles kirjutatakse lahti projekti raames tehtu ja lahendused, mida projekti tulemusena loodud tarkvara tagab (nagu ka see, kuidas seda tarkvara oma riigi spetsiifikale on võimalik kohandada).

### ***Reaalajamajanduse arendus***

Käesoleva projekti läbiviimise käigus on jõutud mitmel viisil tulemuseni, et reaalajamajandus ja sellega seotud arengusuunad on väga tähtsad varajase hoiatuse kontseptsiooni elluviimisel ja süsteemseks tööksse rakendamiseks. Varajase hoiatuse lahendus toetab reaalajamajanduse tegevusi, kuid samas on reaalajamajanduse lahendused väga oluliseks sisendiks varajase hoiatuse teema juurutamisel ja arendamisel.

Käesoleva projekti arengut jälginud ja nõustanud reaalajamajanduse suuna eest vastutavad inimesed on olnud väga suureks abiks ja toeks, et varajase hoiatuse lahendus saaks elujõuliseks muutuda. Varajase hoiatuse teema edasisel arendamisel on oluline, et jätkuks sama kompetentne toetus reaalajamajanduse suuna eest vastutavate inimeste poolt, nagu see seni on olnud.

### ***Infosüsteemi loomine***

Prototüübile järgneb infosüsteemi loomine. Käesolevaga ollakse seisukohal, et infosüsteemi loomist võiks kaaluda järgmise lähteloogika alusel. Statistikaamet võiks olla infosüsteemi väljatöötajaks (nagu prototüübi arendustöö näitab, on pidevalt oluline vaadata andmetesse n.ö. sisse, teha pidevaid reaalsusteste saamaks aru, kas masin jõuab ka õigetele järeldustele – andmetele saavad peale vaadata Statistikaameti inimesed). Süsteemi omanikuks võiks olla Justiitsministeerium, kuna varajase hoiatuse tasemed viiakse kokku õiguslikult defineeritud maksejõuetuse astmetega (millal on lihtsalt oht, millal peaks käivitama saneerimise, millal peaks esitama pankrotiavalduse).

Justiitsministeeriumi esindaja on käesoleva projekti käigus esitanud väga kompetentseid ja asjalikke märkusi/juhtnööre, mis on olnud suureks abiks prototüübi kujundamiseks selliseks, et paralleelselt ei hakataks arendama „kahte õiguskeskkonda“ Eestis, vaid viiakse loodav süsteem võimalikult kooskõlla muu õiguspraktikaga. Tulenevalt sellest pakutakse siinkohal välja, et oleks mõistlik kaaluda lahendust, kus loodava varase hoiatamise süsteemi omanikuks oleks Justiitsministeerium.

Siinkohal esitatud lähteloogika on alus aruteludeks, mitte lõplikult fikseeritud seisukoht.

### ***Edasine ajakava***

Alljärgnevalt on välja toodud mõned ajalised orientiirid, kuidas käesoleva projektiga edasi liigutakse:

Prototüübi testimine: mai-juuni 2022.a.

Prototüübi testimise tulemuste analüüs: juuni-juuli 2022.a.

Prototüübi testimise tulemusena prototüübis muudatuste tegemine: juuni-juuli 2022.a.

Prototüübi pinnalt infosüsteemi koostamise plaani väljapakkumine: juuli-august 2022.a.

Infosüsteemi koostamise alustamine: oktoober 2022.a.

Infosüsteemi valmimise tähtaeg: 2023.a. lõpuks.

Süsteemi skaleeritavaks muutmise lahenduse plaani väljapakkumine: juuli-august 2022.a.

Prototüübi edasiarenduse plaani väljapakkumine: juuli-august 2022.a.

Prototüübi edasiarenduse algus: september 2022.a.

Edasiarendatud prototüübi esitus: veebruar 2023.a.

Prototüübi edasiarendus on planeeritud ajaliselt paralleelselt sellega, kui arendatakse infosüsteemi ja lõppema selliselt, et edasiarendatud prototüübi arengutulemusi oleks võimalik kaasata ka juba infosüsteemi väljaarendamisesse.

Ülaltoodud ajakava tähtajad on väljapakutud variandis, mõeldud arutamiseks ja vajadusel muutmiseks.

## Lisad

Käesolevale materjalile on lisatu järgmised lisad:

Lisa 1.: Tegevusalade defineerimine EMTAK abil

Lisa 2.: Kasutatavad majandussuhtarvud

Lisa 3.: Pankrotistunud ettevõtteid klastrite ja aastate lõikes

Lisa 4.: Täiendavad klasterdamised

Lisa 5.: Klastrite taandamine kahe parameetri peale

Lisa 6.: PCA tehnikaga klasterdamine kahedimensionaalseks

Lisa 7.: Muutujate omavaheline korrelatsioon

Lisa 8.: Klastritesse jagunemine tegevusalade lõikes

Lisa 9.: TwoLayerNet kood

Lisa 10.: Layers kood

Lisa 11.: Optim kood

## Allikad

Aliaj, T., Anagnostopoulos, A., Piersanti, S. (2020). Firms Default Prediction with Machine Learning. Bank of Italy (2020).

Amel-Zadeh, A., Calliess, J.-P., Kaiser, D., Roberts, S. (2020). Machine Learning-Based Financial Statement Analysis. Oxford-Man Institute of Quantitative Finance, University of Oxford (2020), pp. 1-55.

Andrés, J., Lorca, P., Bahamonde, A., Coz, J., J. (2004). The Use of Machine Learning Algorithms for the Study of Business Profitability: A New Approach Based on Preferences. *The International Journal of Digital Accounting Research*. Vol 4, N 8 (2004), pp. 99-124.

Cao, K., You, H., (2021). Fundamental Analysis via Machine Learning. Hong Kong University of Science and Technology. (2021), pp. 1-62.

Chen, J., M. (2020). An Introduction to Machine Learning for Panel Data. Michigan State University (2020), pp. 1-67.

Choudhry, M. (2018). Machine Learning in Banking: How To Transform both Balance Sheet Management and Customer Service Provision. *The European Financial Review*. (April-May 2018), pp. 2-8.

Cialone, G., (2020). Bankruptcy Prediction by Deep Learning. Stanford University CS230 Winter 2020.

Kou, G., Xu, Y., Peng, Y., Shen, F., Chen, Y., Chang, K., Kou, S. (2021). Bankruptcy prediction for SMEs using transactional data and two-stage multiobjective feature selection. *Decision Support Systems*. No 140 (2021), pp. 1-14.

León, C., Moreno, J., F., Cely, J. (2016). Whose Balance Sheet is this? Neural Networks for Banks' Pattern Recognition. *Borradores de Economía*. Núm 959 (2016), pp. 1-34.

Perboli, G., Arabnezhad, E. (2021). A Machine Learning based DSS for mid and long-term company crisis prediction. *Expert Systems With Applications*. No 174 (2021), pp. 1-12.

Petropoulos, A., Siakoulis, V., Vlachogiannakis, N., Stavroulakis, E. (2019). Deep-Stress: A deep learning approach for dynamic balance sheet stress testing. Bank of Greece. (2019), pp. 1-22.

Qu, Y., Quan, P., Lei, M., Shi, Y. (2019). Review of bankruptcy prediction using machine learning and deep learning techniques. *Procedia Computer Science*. No 162 (2019), pp. 895-899.

Shi, Y., Li, X. (2019). A bibliometric study on intelligent techniques of bankruptcy prediction for corporate firms. *Heliyon*. No 5 (2019), pp. 1-12.

Ucoglu, D. (2020). Current machine learning applications in accounting and auditing. *PressAcademia Procedia*. Vol 12 (2020), pp. 1-7.

## Lisa 1.: Tegevusalade defineerimine EMTAK abil

### Meditsiini valdkond:

Nimetus	EMTAK
Haiglaraviteenused	EMTAK 86101
Üldarstiabi osutamine	EMTAK 86211
Eriarstiabi osutamine	EMTAK 86221
Hambaravi osutamine	EMTAK 86231
Kiirabi ja parameedikute tegevus	EMTAK 86901
Meditsiinilaborite tegevus	EMTAK 86902
Õendusabi osutamine	EMTAK 86903
Ämmaemandate tegevus	EMTAK 86904
Sanatooriumide ravitegevus	EMTAK 86905
Hooldusraviasutuste tegevus	EMTAK 87101
Diagnostikakabinettide ja -keskuste tegevus	EMTAK 86906
Vaimupuudega ja psüühiliste erivajadustega isikute hoolekandeesutuste tegevus	EMTAK 87201
Ainesõltlaste hoolekandeesutuste tegevus	EMTAK 87202
Vanurite ja puuetega inimeste hoolekandeesutuste tegevus	EMTAK 87301
Asendushooldusteenust osutavate hoolekandeesutuste tegevus	EMTAK 87901
Mujal liigitamata tervishoiualad	EMTAK 86909
Meditsiini- ja ortopeediakaupade jaemüük spetsialiseeritud kauplustes	EMTAK 47741
Muu meditsiini- ja hambaraviinstrumentide ning materjalide tootmine	EMTAK 32509

### Finantsteenuste valdkond:

Nimetus	EMTAK
Keskpanga tegevus	EMTAK 64111
Krediitiasutused (pangad)	EMTAK 64191
Valdusfirmade tegevus	EMTAK 64201
Usaldusfondide, investeerimisfondide ja sarnaste finantsüksuste investeerimine	EMTAK 64301
Kapitalirent	EMTAK 64911
Pandimajad	EMTAK 64921
Muu laenuandmine, v.a pandimajad	EMTAK 64929
Muude mujal liigitamata finantsteenuste osutamine, v.a kindlustus ja pensionifondid	EMTAK 64991
Elukindlustus	EMTAK 65111
Kahjukindlustus	EMTAK 65121
Edasikindlustus	EMTAK 65201
Pensionifondid	EMTAK 65301
Finantsturgude haldamine	EMTAK 66111
Väärtpaberite ja kaubalepingute maaklerlus	EMTAK 66121
Valuutavahetus	EMTAK 66129
Finantsnõustamine	EMTAK 66191
Muud mujal liigitamata finantsteenuste abitegevusalad	EMTAK 66199
Riskide ja kahju hindamine	EMTAK 66211
Kindlustusagentide ja -vahendajate tegevus	EMTAK 66221
Muud kindlustuse ja pensionifondide abitegevusalad	EMTAK 66291
Fondide valitsemine	EMTAK 66301

### Kinnisvara valdkond:

Nimetus	EMTAK
Hoonestusprojektide arendus	EMTAK 41101
Elamute ja mitteeluhoonete ehitus	EMTAK 41201
Teede ja kiirteede ehitus	EMTAK 42111
Raudteede ja metroo ehitus, hooldus ja remont	EMTAK 42121
Silla- ja tunneliehitus	EMTAK 42131
Vee-, gaasi- ja kanalisatsioonitrasside ehitus	EMTAK 42211
Kaevude ja puuraukude rajamine ning likvideerimine	EMTAK 42212
Elektri- ja sidevõrkude ehitus	EMTAK 42221

Vesiehitus	EMTAK 42911
Mujal liigitamata rajatiste ehitus	EMTAK 42991
Lammutamine	EMTAK 43111
Ehitusplatside ettevalmistus ja puhastus	EMTAK 43121
Maakuivendus- ja maaparandustööd, k.a metsa- ja põllumaade kuivendus	EMTAK 43122
Muud kaeve- ja mullatööd	EMTAK 43129
Kontrollpuurimine ja sondimine	EMTAK 43131
Elektrijuhtmete ja -seadmete paigaldus	EMTAK 43211
Tulekahjualarmide, häire- ja valvesignalisatsiooni paigaldus	EMTAK 43212
Telekommunikatsioonikaablite ja antennide paigaldus	EMTAK 43213
Kütte-, ventilatsiooni- ja kliimaseadmete paigaldus	EMTAK 43221
Veetorustiku ja sanitaarseadmete paigaldus	EMTAK 43222
Isolatsioonitööd	EMTAK 43291
Mujal liigitamata ehituspaigaldustööd	EMTAK 43299
Krohvimine	EMTAK 43311
Mistahes materjalist akende, uste ja treppide paigaldus	EMTAK 43321
Muude ehituspuusepatoodete paigaldus	EMTAK 43329
Põranda- ja seinakatete paigaldus	EMTAK 43331
Värvimine ja klaasimine	EMTAK 43341
Muu ehitiste viimistlus ja lõpetamine	EMTAK 43391
Katusetööd	EMTAK 43911
Vundamendi-, betooni- jm müüritööd	EMTAK 43991
Pottsepatööd, ahjude ja kaminade tegemine	EMTAK 43992
Tellingute ja tööplatvormide püstitus ja demonteerimine, rentimine	EMTAK 43993
Muud eriehitustööd	EMTAK 43999
Enda kinnisvara ost ja müük	EMTAK 68101
Enda või renditud kinnisvara üürileandmine ja käitus	EMTAK 68201
Kinnisvarabüroode tegevus	EMTAK 68311
Hoonete ja üürimajade haldus	EMTAK 68321
Aiandus- ja suvilaühistute jms haldus	EMTAK 68322
Muu kinnisvarahaldus või haldusega seotud tegevused	EMTAK 68329

## Haridusvaldkond:

Nimetus	EMTAK
Lastesõimede tegevus	EMTAK 85101
Lasteaedade tegevus	EMTAK 85102
Lasteaed-algkoolide tegevus	EMTAK 85201
Algkoolide tegevus	EMTAK 85202
Lasteaed-põhikoolide tegevus	EMTAK 85311
Põhikoolide tegevus	EMTAK 85312
Gümnaasiumide tegevus	EMTAK 85313
Kutseõppeasutuste tegevus	EMTAK 85321
Rakenduskõrgkoolide tegevus	EMTAK 85411
Ülikoolide tegevus	EMTAK 85421
Spordikoolid	EMTAK 85511
Muu spordi- ja vabaajakoolitus	EMTAK 85519
Muusika- ja kunstikoolitus	EMTAK 85521
Tantsukoolide tegevus	EMTAK 85522
Muu huvikoolitus	EMTAK 85529
Sõiduõpe	EMTAK 85531
Keeleõpe	EMTAK 85591
Arvutiõpe	EMTAK 85592
Muu mujal liigitamata koolitus	EMTAK 85599
Haridust abistavad tegevused	EMTAK 85601
Teadus- ja arendustegevus biotehnoloogia vallas	EMTAK 72111
Teadus- ja arendustegevus muude loodus- ja tehnikateaduste vallas	EMTAK 72191
Teadus- ja arendustegevus sotsiaal- ja humanitaarteaduste vallas	EMTAK 72201

## Lisa 2.: Kasutatavad majandussuhtarvud

Suhtarv / suhtarvu saamise valem	Selgitus suhtarvu funktsiooni kohta	Kuidas suhtarvu tõlgendada / mida näitab
<b>LIKVIIDSUST KIRJELDAVAD SUHTARVUD</b>		
Likviidsussuhtarvud näitavad, kui rahaliselt vaba on ettevõtte positsioon, see annab alust anda esmast hinnangut, kuivõrd suure töenäosusega on ettevõtte sattumas finantsraskustesse. Finantsraskused vormistuvad eelkõige viisil, kuidas ettevõtte suudab oma kohustusi täita väliste partnerite eest ja seetõttu on oluline, kuidas ettevõtte likviidsus selles kontekstis tuge pakub. Tulenevalt ülaltoodust vaadatakse likviidsussuhtarve esmasena.		
<b>Käibekapital</b> Working Capital <i>käibekapital = käibevarad – lühiajalised kohustused</i>	Näitab sisuliselt raha kogust, mida ettevõtte saab kasutada oma igapäevase majandustegevuse läbiviimiseks.	Mida rohkem on ettevõttel käibekapitali, seda parem. Kas käibekapital on piisav või seda on puudu, seda saab hinnata koosmõjus teiste majandusnäitajatega.
<b>Käibekapitali käive</b> Working Capital Turnover <i>käibekapitali käive = müügitulu / käibekapital</i>	Näitab, mitu korda ettevõtte oma käibekapitali aasta jooksul kasutab.	Näitaja võiks jääda vahemikku 2-10, pigem 5-8. Number näitab kordi, mitu korda käibekapital pöörduv aasta jooksul. Mida suurem on number, seda efektiivsemalt käibekapitali kasutab, kuid samas ka seda haavatavam on, kui käibekapital peaks kaduma/väheneb (samamoodi jälegi suhteliselt väiksema rahaga saab ettevõtte välja aidata).
<b>Lühiajaliste võlgnevuste kattekordaja</b> Current Ratio <i>lühiajaliste võlgnevuste kattekordaja = käibevarad / lühiajalised kohustused</i>	Näitab maksevõime taset seeläbi, millisel määral käibevarad ületavad lühiajalisi kohustusi.	Vahemikus 1.0-1.5 tuleks kordajat lugeda nõrgapoolseks (alla selle päris nõrgaks). Vahemik 1.5-2.0 loetakse tugevaks, kus üldjuhul võlgade tasumisega probleeme ei nähta. Üle selle on näitaja veel parem, kui see võib omakorda viidata ülekapitaliseeritusele (kapitali ebaefektiivsele kasutamisele).
<b>Maksevõime kordaja</b> Quick Ratio <i>maksevõime kordaja = (käibevarad - varud - ettemaksud) / lühiajalised kohustused</i>	Näitab ettevõtte lühiajalist likviidsust, ehk seda, kuidas ettevõtte saab rahuldada lühiajalisi kohustusi oma kõige likviidsemate varadega. Näitaja siis olukorraks, kui ettevõttel tekiks vajadus lühiajalised kohustused ära maksta.	Kui maksevõime kordaja on vahemikus 0.9-1, siis loetakse, et ettevõtte suudab teenindada lühiajalisi kohustusi ilma probleemideta. Üle 1 on väga hea näitaja. Samas, kui näitaja on väga suur, siis on tegemist rahaliste vahendite ebaefektiivse kasutamisega.
<b>Maksevalmiduse kordaja</b> Cash Ratio <i>maksevalmiduse kordaja = (raha + lühiajalised finantsinvesteeringud) / lühiajalised kohustused</i>	Näitab seda, kui palju lühiajalisi kohustusi on ettevõtte suuteline katma sisuliselt koheselt. Hea näitaja ettevõtte maksejõuetuse riski hindamiseks olukorras kui juhtub mingi ootamatus ja eelnevalt tuleviku tähtajale lükatud kohustused vajavad kohest tasumist.	Kui maksevalmiduse kordaja on vahemikus 0.5-1, siis loetakse seda normaalseks. Üle selle on näitaja väga hea. Kuid kui näitaja läheb üle 2-3, siis on ettevõtte liigselt ülekapitaliseeritud, rahalisi vahendeid kasutatakse ebaefektiivselt.
<b>KAPITALI STRUKTUURI KIRJELDAVAD SUHTARVUD</b>		
Kapitali struktuuri kirjeldavad suhtarvud näitavad, millise struktuuri järgi paikneb ettevõtte kapital. Suhtarvud võimaldavad vaadata ettevõttesse sisse, mõista selle tegelikku turusuutlikkust ja konkreetse tootmisharu kontekstis anda hinnang, kas ettevõtte võib sattuda finantsilistesse raskustesse. Kapitali struktuuri suhtarvud annavad eelnevalt kirjeldatud likviidsusnäitajatele n.ö. raamistiku.		
<b>Võlakordaja</b> Debt Ratio <i>võlakordaja = kohustused kokku / koguvara</i>	Näitab, kui suures mahus on ettevõtte olemasolevate varade saamiseks kasutatud võõrkapitali. Hea näitaja n.ö. üldise võlausaldaja riski hindamiseks ja sellest tuleneva võimalike makseraskuste riski hindamiseks.	Hinnang näitajale sõltub suuresti sellest, mis sektoris tegutsetakse, kui palju seal võlga üldiselt kasutatakse, kui stabiilne on sektor ja mis liiki võlaga on tegemist. Vahest on ka 30% suur näitaja, teistes oludes võib see olla kuni 80% ilma, et sisulist probleemi oleks.
<b>Kapitaliseerituse tase</b> Debt to Debt plus Equity <i>kapitaliseerituse tase = pikaajalised kohustused / (pikaajalised kohustused + omakapital)</i>	Näitab, kui agressiivselt on ettevõtte laenu võtnud ja millisel määral riske see laenuvõtmine on kaasa toonud. Võlakordajast erineb seeläbi, et on suunatud otseselt laenukohustuste hindamisele (võlakordaja vaatab kohustusi/võõrkapitali laiemalt). Laenuandjad on sageli need, kellel on kõige otsesem võimalus mõjutada ettevõttes toimuvat.	Mida suurem on see näitaja, seda suurem on finantsrisk. Laenuandjad võivad otsustada laenude varasema tagasikutsumise, mis võib olla ka automaatne tulenevalt kolmandate ettevõttega seotud äriolude asjaolude ilmnemisest.
<b>Intresside kattekordaja</b> Interest Coverage Ratio <i>intresside kattekordaja = (ärikasum + finantskulud) / finantskulud</i>	Näitab ettevõtte võimekust katta oma intressikulud ärikasumist, ehk siis mitu korda suudab ettevõtte maksta oma intressikulud jooksvalt teenitavast rahavoost. Näitaja suurus ja dünaamika (mis suunas areneb erinevate perioodide lõikes) näitab, kas ettevõtte on sattumas üle	Mida suurem on näitaja, seda parem. Kui näitaja on väga väike, finantskulude teenindamine võtab enamuse ärikasumist, siis tekib olukord, kus ettevõtte n.ö. „upub võlgade alla“.



	jõu käivate kohustuste ette ja seega maksejõuetusse.	
<b>Omakapitali võlasiduvus</b> <b>Debt to Equity Ratio</b> <i>omakapitali võlasiduvus = kohustused kokku / omakapital</i>	Näitab võõrkapitali kasutamise osatähtsust omakapitali suhtes väljendades seeläbi laenuriski suurust omakapitali vastu. Näitaja abil on võimalik vaadata, kuidas välistest võlausaldajatest tulenev risk väljendub omakapitali suhtes ja seega kuidas see võib viidata potentsiaalsele maksejõuetusele.	Üldine seisukoht on, et 2.0 on halb näitaja ja 1.0 on hea näitaja. Seega, mida suurem on näitaja, seda halvem, mida väiksem, seda parem. Oluline on tegevussektori kontekst, osad sektorid on rohkem võla peal tegutsejad, kui teised.
<b>Omakapitali osatähtsus</b> <b>Equity to Assets Ratio</b> <i>omakapitali osatähtsus = omakapital / koguvara</i>	Näitab, kui palju on ettevõtte varade hulgas n.ö. ettevõtte omanikele kuuluvat vara. Mida väiksem on näitaja, seda vähem on majanduslikus mõttes ettevõtte oma omanike kontrolli all ja seda rohkem mõjutatav välistest asjaoludest. Annab võimaluse hinnata, kuivõrd on maksejõuetuse risk ettevõtte enda inimeste juhtida.	Mida suurem on see näitaja, seda parem. Kuid kui näitaja on 100% või seal ligil, siis põhimõtteliselt tähendab see seda, et seal ei ole üldse kasutatud väliseid vahendeid ja see viitab enamikel juhtudel mitteamurkale majandamisele.
<b>RENTAABLUST KIRJELDAVAD SUHTARVUD</b>		
Rentaabluse suhtarvud näitavad, kuivõrd on ettevõtte tootlik, et edasi areneda, kasvada ja jätkusuutlikult tegutseda. Mida rentabilim on ettevõtte, seda suurema tõenäosusega suudab ta oma rahalisi probleeme ka lahendada, kui on sattunud maksevõime osas probleemide piirile.		
<b>Käiberentaablus</b> <b>Net Profit Margin</b> <i>käiberentaablus = puhaskasum / müügitulu</i>	Näitab, kui suure osa moodustab puhaskasum müügitulus, ehk kui palju saab ettevõtte puhast kasu igast käibe läbi laekunud eurost, mis on ettevõtte osa sellest käibest. Võimaldab analüüsida, kuidas käibe muutus (kasv) vähendab maksevõimetuse riski, millise kõvera alusel.	Mida suurem on see näitaja, seda parem. Maksimumpiirajat ei ole (ehk siis teoreetiliselt 100%). Näitaja „normaalne“ väärtust sõltub tegevussektorist, kuid vahemik 10-15% on üldiselt hinnatav positiivselt.
<b>Varade rentabilus</b> <b>Return on Assets</b> <i>varade rentabilus = EBIT / koguvara</i>	Näitab, millise tootlikkusega on varad, ehk siis millise tootlikkusega on vahendid, mille eest varad on omandatud. Võimaldab hinnata maksejõuetuse riski varade vastu.	Mida suurem on varade rentabilus, seda parem. Sõltuvalt tegevussektorist ja varade olemusest võib varade liigne koormatus tuua kaas n.ö. „katkimineku riski“, mis omakorda peataks osaliselt tegevuse.
<b>Omakapitali rentabilus</b> <b>Return on Equity</b> <i>omakapitali rentabilus = (puhaskasum - eelisaktiade dividendid) / keskmine omakapital</i>	Näitab omakapitali kasutamise efektiivsust, ehk millise efektiivsusega omakapital töötab. Võimaldab lisaks varade tootlikkusele hinnata, millist rolli mängib omakapitalile tinglikult omastatav tootlikkus võimaliku maksejõuetuse välistamisel.	Mida suurem on omakapitali tootlikkus, seda parem. Samas, kui omakapitali osakaal on bilansis väike, siis (muude tingimuste samaksjäädes) rohkem varasid (s.t. laenuvõimendust) võib viia omakapitali tootlikkuse üles, kuid selle ülesviimise negatiivne pool on liigne laenuvõimendus.
<b>ESMAST EFEKTIIVSUST KIRJELDAVAD SUHTARVUD</b>		
Esmast efektiivsust kirjeldavad suhtarvud võimaldavad vaadata rohkem ettevõtte sisse, kas ettevõtte igapäevased operatsioonid on käivitatud efektiivselt, kas igapäevases tegevuses on mingeid olulisi puudujääke, millest tulenevalt on negatiivsed (ebasoodsad) teised (ülal käsitletud) suhtarvud. Esmase efektiivsuse suhtarve vaadeldakse võimalike makseraskuste tekkimise seisukohalt viimasena, kuna üldine lähenemine on üldiselt üksikule, kõigepealt üldine visioon ja arusaam ettevõttest, siis laskumine erinevate detailide tasandile.		
<b>Nõuete käibesagedus</b> <b>Accounts Receivables Turnover</b> <i>nõuete käibesagedus = müügitulu / keskmine nõuete maht</i>	Näitab, mitu korda ületab müügi käibe nõuete maht (nõuded, mis on eelkõige suunatud klientide vastu seoses laekumata rahaga). Näitab ettevõtte tegevuse efektiivsust seoses teostatud müügi eest raha kättesaamisega (mida võivad mõjutada ka vaidlused toote/teenuse kvaliteedi üle). Ehk siis ühtlasi näitaja, mis võib iseloomustada ettevõtte turusuutlikkust sisulises mõttes.	Mida suurem see näitaja on, seda parem, sest siis on raha kättesaamine teenuse/toote eest hea. Sõltub väga sektoris tavaks olevast makserutiinist.
<b>Nõuete käibevälde</b> <b>Receivables Average Collection Period</b> <i>nõuete käibevälde = 360 / nõuete käibesagedus</i>	Tegemist on nõuete käibesageduse tuletisnäitajaga, mis näitab, kui palju aega võtab tavaliselt keskmiselt klientidelt raha kättesaamine. Kui kliendid ei maksa raha ära, siis see võib põhjustada makseraskusi.	Näitaja väljendatakse päevades. Näitaja ei peaks olema väga palju suurem, kui antud maksetähtjad (mõni viibimine alati tekib, päris maksetähtaja pikkusesse näitajat ei saa).
<b>Varade käibekordaja</b> <b>Inventory Turnover</b> <i>varade käibekordaja = müüdü toodangu kulu / keskmised varud</i>	Näitab, mitu korda ettevõtte varusid müüakse, s.t. kui efektiivselt varusid kasutatakse. Varud on lattu toodetud toodang, mis vajab realiseerimist. Kui ettevõtte toodab lattu ja klient toodangut ära ei osta, siis võib ettevõtte sattuda makseraskustesse.	Mida suurem see näitaja on, seda parem, seda vähem seisavad tooted laos ja seda vähem on laoseisu all raha kinni.
<b>Varade käibevälde</b> <i>varade käibevälde = 360 / varade käibekordaja</i>	Tegemist on varade käibekordaja tuletisnäitajaga, mis näitab, mitme päeva jooksul varud turustatakse. Kui ettevõtte varude (s.t. lattu toodetud toodangu turustamine) võtab kaua aega, siis võib tekkida maksejõuetuse risk.	Mida väiksem see näitaja on, seda parem, seda vähem seisab valmis toodang laos ja seda rutem saadakse (muude tingimuste samaksjäädes) toodangu eest raha kätte.

<b>Talitlustersükk</b> <i>talitlustersükk = varude käibevälde + nõuete käibevälde</i>	Näitaja näitab, kui millise perioodi jooksul (kui pika perioodi jooksul) ettevõtte oma toodanguvarudest lahti saab ja ka klientidelt selle eest raha kätte saab. Mida lühem on periood, seda väiksema tõenäosusega satub ettevõtte finantsraskustesse.	Näitaja on päevades, kui see näitaja on natuke suurem, kui tavaliselt sektoris antud maksetähtaeg, siis on väga hea (alla maksetähtaja ei saa see näitaja kukkuda ja seega on olemas mingi optimaalne aluspiir näitajale).
<b>Lühiajaliste kohustuste käibekordaja</b> <b>Accounts Payable Turnover Ratio</b> <i>lühiajaliste kohustuste käibekordaja = müüdüd toodangu kulu / keskmised lühiajalised kohustused</i>	Näitaja näitab, kui efektiivne on lühiajaliste kohustuste teenindamine müüdüd toodangu kulu suhtes vaadatuna. Ehk siis, mitu korda aruandeperioodi jooksul tasutakse lühiajalised kohustused. Mida rohkem suudab ettevõtte lühiajalisi kohustusi tasuda, seda väiksem on maksejõuetuse risk.	Mida suurem on näitaja, seda parem. Optimaalne (keskmine) näitaja on väga sektori spetsiifikast tulenev.
<b>Lühiajaliste kohustuste käibevälde</b> <b>Payables Average Settlement Period</b> <i>lühiajaliste kohustuste käibevälde = 360 / lühiajaliste kohustuste käibekordaja</i>	Tegemist on lühiajaliste kohustuste käibekordaja tuletisnäitajaga, mis näitab, mitme päeva jooksul suudetakse tasuda lühiajalised kohustused. Mida rutem on ettevõttel võimalik lühiajalised kohustused ära maksta, seda väiksem on tõenäosus, et ettevõtte võib sattuda makseraskustesse.	Mida väiksem on näitaja, seda parem, seda kiiremini suudab ettevõtte vajadusel lühiajalised kohustused ära maksta. Liiga väike number viitab jälle sellele, et ettevõttel ei ole sisulisel määral lühiajalisi kohustusi või tegutseb ettevõtte liigse raha tingimustes (liigse raha tingimustes selles mõttes, et sama raha kogusega saaks teha märksa ulatuslikumat äri – konkurendid suudaksid väiksema rahaga teha sama ulatuslikku äri).
<b>Finantseerimistsükk</b> <i>finantseerimistsükk = talitlustersükk – lühiajaliste kohustuste käibevälde</i>	Näitaja näitab päevade arvu, mille osas peab ettevõtte leidma tegevuseks täiendavaid finantsvahendeid. Mida väiksem on see number, seda väiksem on risk maksejõuetuseks.	Suuruse optimaalne tase tuleb tootmisharust, kuid üksik ettevõtte võib selle näitaja tõttu oluliselt erineada ilma, et oleks põhjust muretseda. Samas negatiivne tsükkel näitaks liigset raha selles mõttes, et sama rahaga saaks teha rohkem äri.
<b>Koguvarede käibekordaja</b> <b>Assets Turnover Ratio</b> <i>koguvarede käibekordaja = müügitulu / keskmine koguvare</i>	Näitaja toob välja, kui efektiivselt kasutatakse ettevõtte varasid. Tegemist on esmase efektiivsuse mõiste suhteliselt suure üldistusastmega näitajaga (kogukäive v. kogu varade maht), annab kõige rohkem ülevaatelise pildi varade efektiivsust kasutamisest, nüansid ei hakka segama.	Mida efektiivsemalt kasutatakse ja eeldusel, et nende varadega tehakse õiget asja, siis on suurem efektiivsus parem – ehk mida rohkem varadega raha teenitakse, seda parem. Mida suurem näitaja, seda parem on majandusseis, seda kiiremini võib seis vajadusel pareneada ja seda väiksem on tõenäosus maksejõuetuseks.

**Lisa 3.: Pankrotistunud ettevõtteid klastrite ja aastate lõikes**

	<b>Klaster</b>	<b>Ettevõtteid</b>
2009	1	192
	2	4
2010	1	129
	2	1
2011	1	131
	2	2
2012	1	107
	2	1
	3	1
2013	1	99
	2	2
	3	2
2014	2	1
	3	157
2015	2	1
	3	89
2016	3	102
	4	3
2017	0	1
	3	120
	4	2
2018	0	114
	4	3
2019	0	71
	4	2
2020	0	48
	4	4
2021	0	3

## Lisa 4.: Täiendavad klasterdamised

Tabel L1: Keskmised, K-means: kood, EMA\_BI\_ID-d ei ole klasterdamisel arvestatud. PERIOD\_NM logaritmiti enne klasterdamist (peaks kuupäevad masinale võrreldavaks tegema). BI\_100 - BI\_40 tunnused normaliseeritud MinMaxScaleriga enne klasterdamist. Suhted arvutatud hiljem.

Tabel L2: Mediaanid, K-means: kood, EMA\_BI\_ID-d ei ole klasterdamisel arvestatud. PERIOD\_NM logaritmiti enne klasterdamist (peaks kuupäevad masinale võrreldavaks tegema). BI\_100 - BI\_40 tunnused normaliseeritud MinMaxScaleriga enne klasterdamist. Suhted arvutatud hiljem.

Tabel L3: Keskmised, kõik sama nagu L1, L2 aga lisaks on eemaldatud "outlierid", mis üldiselt klasterdamisalgorütmile ei meeldi. Z-score = 3 ehk 99,7% andmetest jääb alles, 0,3% piirjuhtudest jäävad välja.

Tabel L4: Mediaanid, kõik sama nagu L1, L2 aga lisaks on eemaldatud "outlierid", mis üldiselt klasterdamisalgorütmile ei meeldi. Z-score = 3 ehk 99,7% andmetest jääb alles, 0,3% piirjuhtudest jäävad välja.

### Tabel

#### L1:

clusters	BI_100_1	BI_150_1	BI_180_1	BI_190_1	BI_240_1	BI_250_1	BI_290_1	BI_310_1	BI_370_1	BI_400_1	BI_40_1	BI_40_2
0	174 950	213 481	405 298	436 558	116 311	82 703	126 130	312 141	226 825	14 372	48 832	44 403
1	174 383	266 334	447 299	447 716	129 206	101 252	119 751	360 983	216 132	17 151	71 777	61 763
2	325 517	486 705	914 517	834 326	235 214	168 224	261 025	739 489	504 326	20 947	64 214	60 113
3	247 690	236 418	492 256	570 159	129 936	140 719	197 929	388 452	328 450	14 812	50 645	45 265
4	184 973	257 111	439 333	463 400	127 089	104 896	128 688	366 798	233 655	17 984	63 250	56 165

### Tabel

#### L2:

clusters	BI_100_1	BI_150_1	BI_180_1	BI_190_1	BI_240_1	BI_250_1	BI_290_1	BI_310_1	BI_370_1	BI_400_1	BI_40_1	BI_40_2
0	10181.00	6666.00	8683.00	17696.0	4230.0	2916.00	4272.0	14190.00	6267.00	2556.00	3973.0	3196.00
1	9968.00	8986.00	6679.50	15836.0	4708.0	3778.00	3146.0	16547.00	4240.00	2500.00	5571.0	4457.00
2	9750.55	6353.17	10147.51	17954.0	5603.0	3604.36	6328.0	17707.43	9250.44	2556.47	3025.0	2684.29
3	10062.00	6205.50	9250.00	17938.0	4200.0	2906.00	5039.0	14274.50	7321.00	2556.00	3535.0	2909.00
4	10927.00	9291.00	8333.00	18282.0	5145.0	4274.00	4100.0	17837.00	5819.00	2508.00	5317.0	4103.00

### Tabel

#### L3:

clusters	BI_100_1	BI_150_1	BI_180_1	BI_190_1	BI_240_1	BI_250_1	BI_290_1	BI_310_1	BI_370_1	BI_400_1	BI_40_1	BI_40_2
0	129 446	130 841	275 062	305 567	79 277	63 094	98 388	228 106	176 836	8 644	27 760	26 318
1	132 930	157 411	281 662	311 089	78 576	72 622	85 604	261 983	159 131	7 847	44 924	39 380
2	126 057	129 782	272 419	301 666	70 715	59 797	84 892	226 328	156 247	8 047	34 827	30 835
3	123 261	159 042	270 417	288 120	75 935	67 832	76 424	254 436	143 102	7 005	50 171	43 088
4	130 621	125 326	275 445	310 881	71 027	59 857	91 851	215 041	163 163	8 564	30 264	27 053

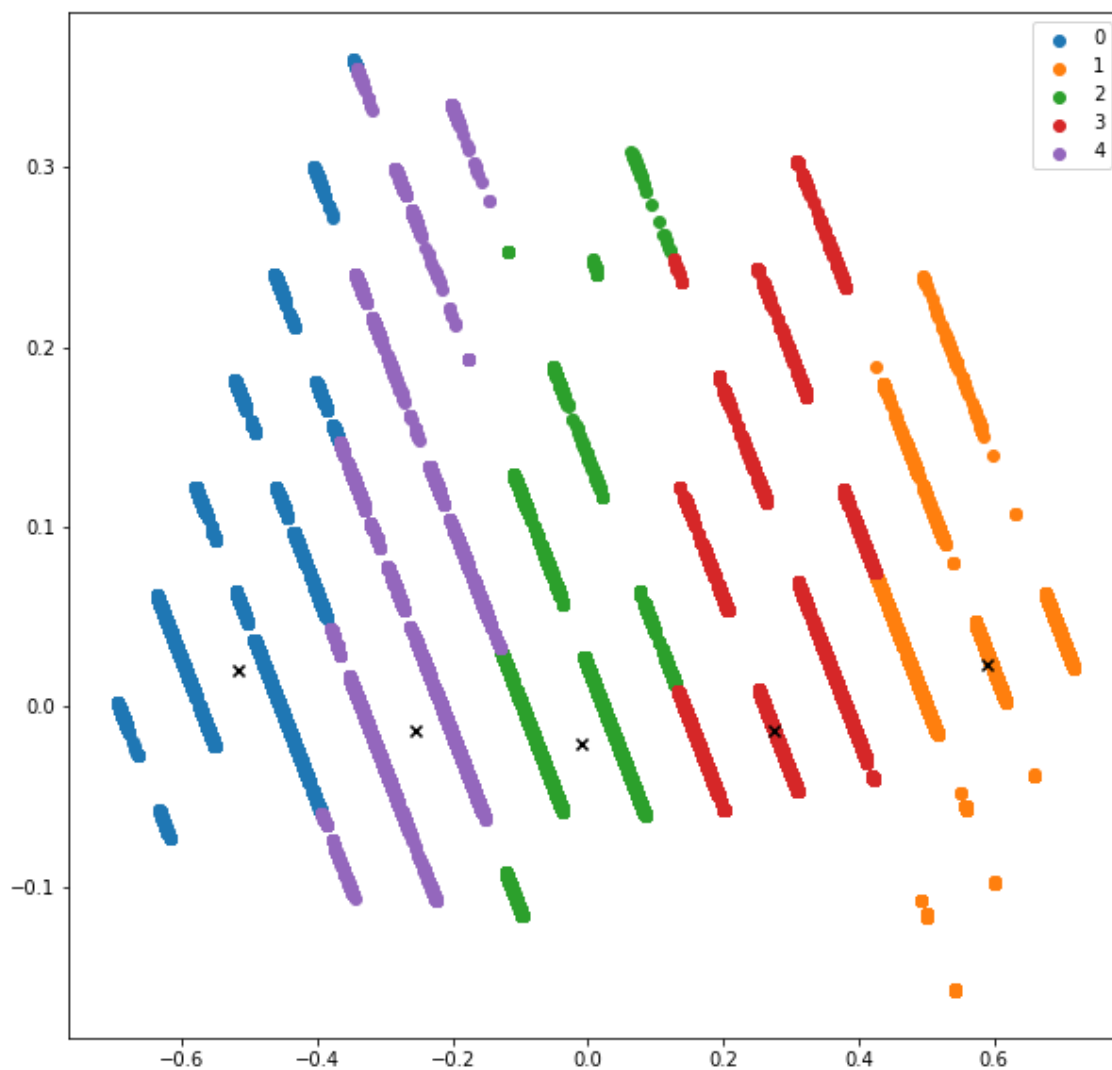
### Tabel

#### L4:

clusters	BI_100_1	BI_150_1	BI_180_1	BI_190_1	BI_240_1	BI_250_1	BI_290_1	BI_310_1	BI_370_1	BI_400_1	BI_40_1	BI_40_2
0	9 671	6 283	10 000	17 785	5 520	3 564	6 262	17 379	9 128	2 556	3 010	2 684
1	10 855	9 170	8 198	18 144	5 094	4 226	4 060	17 591	5 757	2 508	5 271	4 068
2	10 118	6 600	8 571	17 584	4 200	2 890	4 234	14 043	6 204	2 556	3 951	3 179
3	9 902	8 862	6 567	15 718	4 638	3 733	3 116	16 292	4 194	2 500	5 524	4 417
4	9 998	6 148	9 135	17 813	4 150	2 877	5 000	14 065	7 239	2 556	3 516	2 895

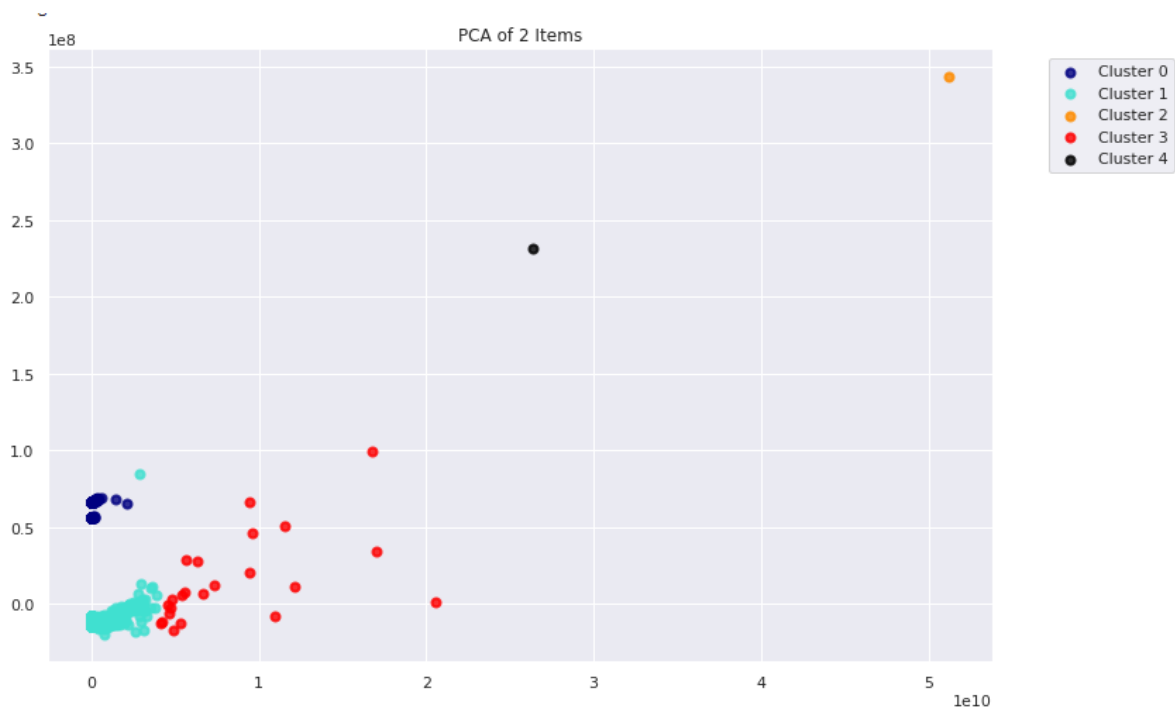
### Lisa 5.: Klasterite taandamine kahe parameetri peale

Joonisel on esitatud olukord, kus graafiliselt on esitatud 20 parameetri taandamine 2-le parameetrile, et oleks võimalik neid kahedimensionaalsel joonisel kuvada. Klasterid tähistatud erinevate värvidega, klasteri tsentroidid on tähistatud musta värvi x-ga.



### Lisa 6.: PCA tehnikaga klasterdamine kahedimensionaalseks

Joonisel on PCA tehnikaga saadud seitse tunnust klasterdatud kahedimensionaalseks jooniseks, saadud on viis klastrit.



### Lisa 7.: Muutujate omavaheline korrelatsioon

	kood	staatus	PERIOD_NM	BI_100_1	BI_150_1	BI_180_1	BI_190_1	BI_240_1	BI_250_1	BI_290_1	BI_310_1	BI_370_1	BI_400_1	BI_40_1	BI_40_2	s1	s2	s3	s4	s5	s6
kood	1.000000	-0.024100	-0.033166	-0.006850	-0.004403	-0.008073	-0.011140	-0.007672	-0.004887	-0.006858	-0.005279	-0.007768	-0.004544	-0.005636	-0.006116	0.015823	-0.000112	0.002195	0.000838	-0.004491	0.025907
staatus	-0.024100	1.000000	-0.042746	0.002152	0.002349	0.002071	0.003241	0.005233	0.002791	0.004622	0.001984	0.004214	0.001559	0.001043	0.000702	-0.003108	-0.000112	0.000230	-0.000066	-0.000281	-0.003138
PERIOD_NM	-0.033166	-0.042746	1.000000	-0.003808	-0.004550	-0.005264	-0.005969	-0.005117	-0.003565	-0.004763	-0.004268	-0.005858	-0.000104	-0.001713	-0.002626	0.004657	-0.002420	-0.000792	-0.002383	-0.002954	-0.014811
BI_100_1	-0.006850	0.002152	-0.003808	1.000000	0.170758	0.542504	0.523649	0.159248	0.757904	0.622311	0.208730	0.505656	0.038898	0.702865	0.599072	0.006890	-0.000057	-0.000061	-0.000074	-0.000176	-0.001057
BI_150_1	-0.004403	0.002349	-0.004550	0.170758	1.000000	0.599501	0.627658	0.263008	0.333489	0.372403	0.667988	0.634145	0.093135	0.251102	0.262894	-0.002711	-0.000037	-0.000023	-0.000034	-0.000116	0.000235
BI_180_1	-0.008073	0.002071	-0.005264	0.542504	0.599501	1.000000	0.799267	0.349011	0.381958	0.404244	0.699406	0.712677	0.166513	0.317255	0.323169	-0.006950	-0.000060	-0.000064	-0.000078	-0.000194	-0.000625
BI_190_1	-0.011140	0.003241	-0.005969	0.523649	0.627658	0.799267	1.000000	0.407315	0.637897	0.742528	0.718525	0.932495	0.169825	0.485813	0.497602	-0.004795	-0.000087	-0.000094	-0.000113	-0.000277	-0.001163
BI_240_1	-0.007672	0.005233	-0.005117	0.159248	0.263008	0.349011	0.407315	1.000000	0.098939	0.480657	0.255094	0.456262	0.045526	0.064382	0.107378	-0.002621	0.000187	0.000013	0.000181	-0.000157	-0.001057
BI_250_1	-0.004887	0.002791	-0.003565	0.757904	0.333489	0.381958	0.637897	0.098939	1.000000	0.881770	0.251571	0.709757	0.019175	0.669306	0.604944	-0.006284	0.000019	-0.000021	0.000020	-0.000111	-0.000706
BI_290_1	-0.006858	0.004622	-0.004763	0.622311	0.372403	0.404244	0.742528	0.480657	0.881770	1.000000	0.293266	0.804730	0.030974	0.577614	0.526052	-0.007076	0.000090	-0.000027	0.000076	-0.000161	-0.001106
BI_310_1	-0.005279	0.001984	-0.004268	0.208730	0.667988	0.699406	0.718525	0.255094	0.251571	0.293266	1.000000	0.754129	0.092596	0.250059	0.238212	-0.002924	-0.000033	0.000072	-0.000015	-0.000127	-0.000431
BI_370_1	-0.007768	0.004214	-0.005858	0.505656	0.634145	0.712677	0.932495	0.456262	0.709757	0.804730	0.754129	1.000000	0.076270	0.476384	0.504159	-0.006739	0.000032	0.000033	0.000046	-0.000188	-0.001029
BI_400_1	-0.004544	0.001559	-0.000104	0.038898	0.093135	0.166513	0.169825	0.045526	0.019175	0.030974	0.092596	0.076270	1.000000	0.026022	0.007200	-0.001988	-0.000025	-0.000030	-0.000033	0.000022	-0.000340
BI_40_1	-0.005636	0.001043	-0.001713	0.702865	0.251102	0.317255	0.485813	0.064382	0.669306	0.577614	0.250059	0.476384	0.026022	1.000000	0.853464	0.000649	-0.000081	-0.000088	-0.000105	-0.000255	0.001467
BI_40_2	-0.006116	0.000702	-0.002626	0.599072	0.262894	0.323169	0.497602	0.107379	0.604944	0.526052	0.238212	0.504159	0.072000	0.853464	1.000000	0.000083	-0.000079	-0.000073	-0.000083	-0.000276	0.001007
s1	0.015823	-0.003108	0.004657	0.006890	-0.002711	-0.006950	-0.004795	-0.002621	-0.006284	-0.007076	-0.002924	-0.006739	-0.019188	0.000649	0.000083	1.000000	0.000369	0.000391	0.000487	0.001437	0.002406
s2	-0.000112	-0.000112	-0.002420	-0.000057	-0.000037	-0.000060	-0.000087	0.000187	0.000019	0.000090	-0.000033	0.000032	-0.000025	-0.000081	-0.000079	0.000369	1.000000	0.023180	0.966477	0.099906	0.000405
s3	0.002195	0.000230	-0.000792	-0.000061	-0.000023	-0.000064	-0.000094	0.000013	-0.000021	-0.000027	0.000072	0.000033	-0.000030	-0.000088	-0.000073	0.000391	0.023180	1.000000	0.260139	0.032362	0.000284
s4	0.000838	-0.000066	-0.002383	-0.000074	-0.000034	-0.000078	-0.000113	0.000181	0.000020	0.000076	-0.000015	0.000046	-0.000033	-0.000105	-0.000083	0.000487	0.966477	0.260139	1.000000	0.103773	0.000479
s5	-0.004491	-0.000281	-0.002954	-0.000176	-0.000116	-0.000194	-0.000277	-0.000157	-0.000111	-0.000161	-0.000127	-0.000188	0.000022	-0.000255	-0.000276	0.001437	0.966477	0.260139	1.000000	0.000407	0.000407
s6	0.025907	-0.003138	-0.014811	-0.001057	0.000235	-0.000625	-0.001163	-0.001057	-0.000706	-0.001106	-0.000431	-0.001029	-0.000340	0.001467	0.001007	0.002406	0.000405	0.000284	0.000479	0.000407	1.000000

## Lisa 8.: Klasteritesse jagunemine tegevusalade lõikes

### Klasteritesse jagunemine tegevusalade lõikes

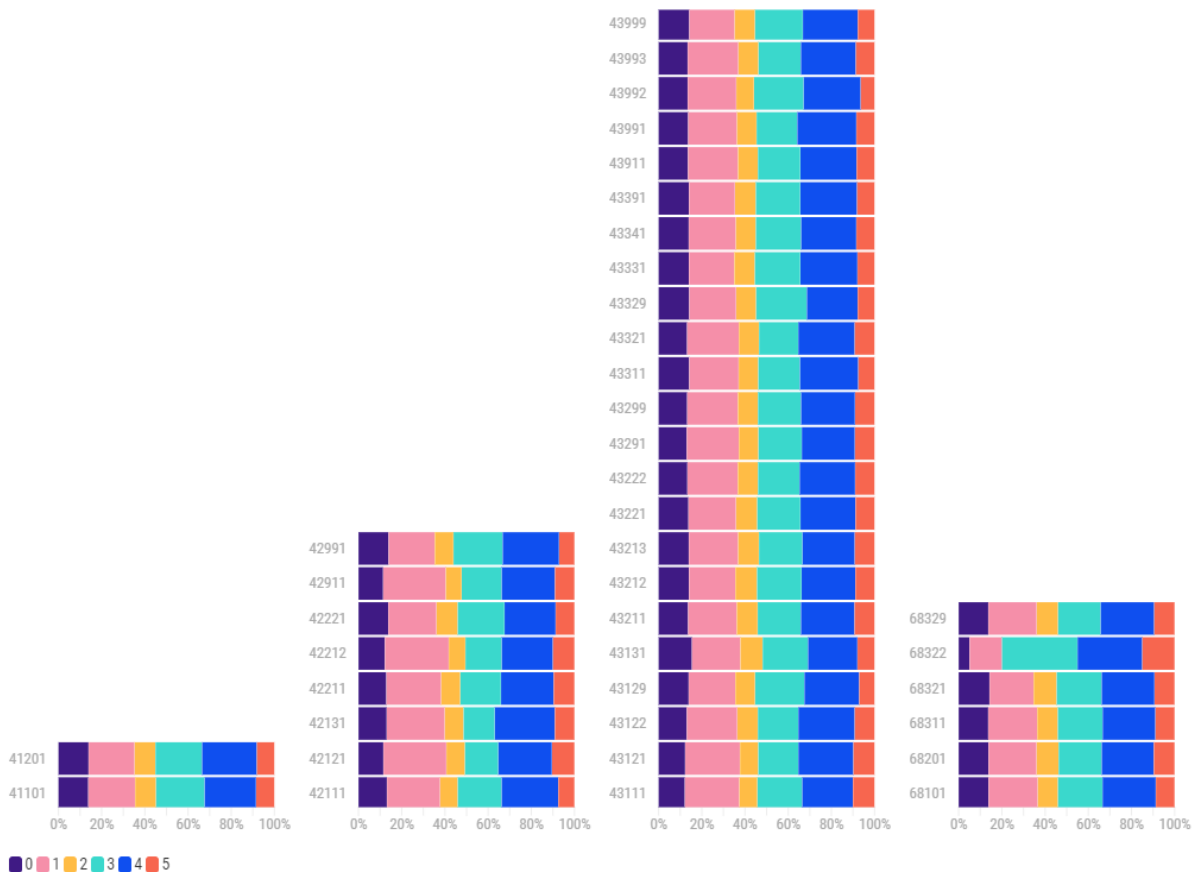
Kõik ▾

Hoonete ehitus

Rajatiste ehitus

Eriehitustööd

Kinnisvaraala tegevus





## Lisa 9.: TwoLayerNet kood

```
from builtins import range
from builtins import object
import numpy as np
```

```
from layers import *
from layer_utils import *
```

```
class TwoLayerNet(object):
```

```
    def __init__(self, input_dim=3*32*32, hidden_dim=100, num_classes=10,
                 weight_scale=1e-3, reg=0.0):
```

```
        """
```

```
        Initialize a new network.
```

```
        Inputs:
```

- input\_dim: An integer giving the size of the input
- hidden\_dim: An integer giving the size of the hidden layer
- num\_classes: An integer giving the number of classes to classify
- weight\_scale: Scalar giving the standard deviation for random initialization of the weights.
- reg: Scalar giving L2 regularization strength.

```
        """
```

```
        self.params = {}
```

```
        self.reg = reg
```

```
        self.cache = {}
```

```
        self.params['W1'] = weight_scale * np.random.randn(input_dim, hidden_dim)
```

```
        self.params['b1'] = np.zeros(hidden_dim)
```

```
        self.params['W2'] = weight_scale * np.random.randn(hidden_dim, num_classes)
```

```
        self.params['b2'] = np.zeros(num_classes)
```

```
    def loss(self, X, y=None):
```

```
        """
```

```
        Compute loss and gradient for a minibatch of data.
```

```
        Inputs:
```

- X: Array of input data of shape (N, d<sub>1</sub>, ..., d<sub>k</sub>)
- y: Array of labels, of shape (N,). y[i] gives the label for X[i].

```
        Returns:
```

```
        If y is None, then run a test-time forward pass of the model and return:
```

- scores: Array of shape (N, C) giving classification scores, where scores[i, c] is the classification score for X[i] and class c.

```
        If y is not None, then run a training-time forward and backward pass and return a tuple of:
```

- loss: Scalar value giving the loss

- grads: Dictionary with the same keys as self.params, mapping parameter names to gradients of the loss with respect to those parameters.

```
"""
```

```
scores = None
```

```
W1, b1 = self.params['W1'], self.params['b1']
```

```
W2, b2 = self.params['W2'], self.params['b2']
```

```
hidden, self.cache['hidden'] = affine_relu_forward(X, W1, b1)
```

```
scores, self.cache['out'] = affine_forward(hidden, W2, b2)
```

```
# If y is None then we are in test mode so just return scores
```

```
if y is None:
```

```
    return scores
```

```
loss, grads = 0, {}
```

```
loss, delta3 = softmax_loss(scores, y)
```

```
loss = loss + 0.5*self.reg*np.sum(W1**2) + 0.5*self.reg*np.sum(W2**2)
```

```
delta2, grads['W2'], grads['b2'] = affine_backward(delta3, self.cache['out'])
```

```
_, grads['W1'], grads['b1'] = affine_relu_backward(delta2, self.cache['hidden'])
```

```
grads['W2'] += self.reg * W2
```

```
grads['W1'] += self.reg * W1
```

```
return loss, grads
```

```
class FullyConnectedNet(object):
```

```
"""
```

A fully-connected neural network with an arbitrary number of hidden layers, ReLU nonlinearities, and a softmax loss function. This will also implement dropout and batch normalization as options. For a network with L layers, the architecture will be

{affine - [batch norm] - relu - [dropout]} x (L - 1) - affine - softmax

where batch normalization and dropout are optional, and the {...} block is repeated L - 1 times.

Similar to the TwoLayerNet above, learnable parameters are stored in the self.params dictionary and will be learned using the Solver class.

```
"""
```

```
def __init__(self, hidden_dims, input_dim=3*32*32, num_classes=10,
```

```
    dropout=0, use_batchnorm=False, reg=0.0,
```

```
    weight_scale=1e-2, dtype=np.float32, seed=None):
```

```
    self.use_batchnorm = use_batchnorm
```

```
    self.use_dropout = dropout > 0
```

```
    self.reg = reg
```

```
    self.num_layers = 1 + len(hidden_dims)
```

```
    self.dtype = dtype
```

```
    self.params = {}
```

```

for i in range(self.num_layers):
    if i == 0 :
        self.params['W' + str(i+1)] = weight_scale * np.random.randn(input_dim, hidden_dims[i])
        self.params['b' + str(i+1)] = np.zeros(hidden_dims[i])
        if self.use_batchnorm:
            self.params['gamma' + str(i+1)] = np.ones(hidden_dims[i])
            self.params['beta' + str(i+1)] = np.zeros(hidden_dims[i])
        elif i < self.num_layers - 1:
            self.params['W' + str(i+1)] = weight_scale * np.random.randn(hidden_dims[i-1], hidden_dims[i])
            self.params['b' + str(i+1)] = np.zeros(hidden_dims[i])
            if self.use_batchnorm:
                self.params['gamma' + str(i+1)] = np.ones(hidden_dims[i])
                self.params['beta' + str(i+1)] = np.zeros(hidden_dims[i])
        else:
            self.params['W' + str(i+1)] = weight_scale * np.random.randn(hidden_dims[i-1], num_classes)
            self.params['b' + str(i+1)] = np.zeros(num_classes)

self.dropout_param = {}
if self.use_dropout:
    self.dropout_param = {'mode': 'train', 'p': dropout}
    if seed is not None:
        self.dropout_param['seed'] = seed

self.bn_params = []
if self.use_batchnorm:
    self.bn_params = [{'mode': 'train'} for i in range(self.num_layers - 1)]

# Cast all parameters to the correct datatype
for k, v in self.params.items():
    self.params[k] = v.astype(dtype)

def loss(self, X, y=None):
    """
    Compute loss and gradient for the fully-connected net.

    Input / output: Same as TwoLayerNet above.
    """
    X = X.astype(self.dtype)
    mode = 'test' if y is None else 'train'

    if self.use_dropout:
        self.dropout_param['mode'] = mode
    if self.use_batchnorm:
        for bn_param in self.bn_params:
            bn_param['mode'] = mode

    scores = None

    a = {'layer0' : X}
    self.cache = {}

```

```

for i in range(self.num_layers):
    W, b = self.params['W'+str(i+1)], self.params['b'+str(i+1)]
    l, l_prev = 'layer'+str(i+1), 'layer'+str(i)
    if mode == 'train':
        bn_params = {'mode': 'train'}
    else:
        bn_params = {'mode': 'test'}
    if i < self.num_layers - 1:
        if self.use_batchnorm and self.use_dropout:
            gamma, beta = self.params['gamma'+str(i+1)], self.params['beta'+str(i+1)]
            a[l], self.cache[l] = affine_batchnorm_relu_dropout_forward(a[l_prev], W, b, gamma, beta,
bn_params, self.dropout_param)
        elif self.use_dropout:
            a[l], self.cache[l] = affine_relu_dropout_forward(a[l_prev], W, b, self.dropout_param)
        elif self.use_batchnorm :
            gamma, beta = self.params['gamma'+str(i+1)], self.params['beta'+str(i+1)]
            a[l], self.cache[l] = affine_batchnorm_relu_forward(a[l_prev], W, b, gamma, beta, bn_params)
        else:
            a[l], self.cache[l] = affine_relu_forward(a[l_prev], W, b)
    else:
        a[l], self.cache[l] = affine_forward(a[l_prev], W, b)

scores = a['layer'+str(self.num_layers)]

if mode == 'test':
    return scores

loss, grads = 0.0, {}

last = self.num_layers
d = {}
loss, dout = softmax_loss(scores, y)
grads = {}
w = 'W' + str(last)
b = 'b' + str(last)
c = 'layer' + str(last)
dh, grads[w], grads[b] = affine_backward(dout, self.cache[c])
loss += 0.5 * self.reg * np.sum(self.params[w]**2)
grads[w] += self.reg * self.params[w]
for i in reversed(range(last - 1)):
    w = 'W' + str(i+1)
    b = 'b' + str(i+1)
    gamma = 'gamma' + str(i+1)
    beta = 'beta' + str(i+1)
    c = 'layer' + str(i+1)
    if self.use_batchnorm and self.use_dropout:
        dh, grads[w], grads[b], grads[gamma], grads[beta] =
affine_batchnorm_relu_dropout_backward(dh, self.cache[c])
    elif self.use_dropout:
        dh, grads[w], grads[b] = affine_relu_dropout_backward(dh, self.cache[c])
    elif self.use_batchnorm:

```

```
        dh, grads[w], grads[b], grads[gamma], grads[beta] = affine_batchnorm_relu_backward(dh,
self.cache[c])
    else:
        dh, grads[w], grads[b] = affine_relu_backward(dh, self.cache[c])

    loss += 0.5 * self.reg * np.sum(self.params[w]**2)
    grads[w] += self.reg * self.params[w]

return loss, grads
```

## Lisa 10.: Layers kood

```
from builtins import range
import numpy as np
```

```
def affine_forward(x, w, b):
```

```
    """
```

```
    Computes the forward pass for an affine (fully-connected) layer.
```

```
    The input x has shape (N, d_1, ..., d_k) and contains a minibatch of N
    examples, where each example x[i] has shape (d_1, ..., d_k). For example,
    batch of 500 RGB CIFAR-10 images would have shape (500, 32, 32, 3). We
    will reshape each input into a vector of dimension D = d_1 * ... * d_k,
    and then transform it to an output vector of dimension M.
```

```
    Inputs:
```

- x: A numpy array containing input data, of shape (N, d\_1, ..., d\_k)
- w: A numpy array of weights, of shape (D, M)
- b: A numpy array of biases, of shape (M,)

```
    Returns a tuple of:
```

- out: output, of shape (N, M)
- cache: (x, w, b)

```
    """
```

```
    out = None
```

```
    out = x.reshape(x.shape[0], -1).dot(w) + b
```

```
    cache = (x, w, b)
```

```
    return out, cache
```

```
def affine_backward(dout, cache):
```

```
    """
```

```
    Computes the backward pass for an affine layer.
```

```
    Inputs:
```

- dout: Upstream derivative, of shape (N, M)
- cache: Tuple of:
  - x: Input data, of shape (N, d\_1, ... d\_k)
  - w: Weights, of shape (D, M)

```
    Returns a tuple of:
```

- dx: Gradient with respect to x, of shape (N, d\_1, ..., d\_k)
- dw: Gradient with respect to w, of shape (D, M)
- db: Gradient with respect to b, of shape (M,)

```
    """
```

```
    x, w, b = cache
```

```
    dx, dw, db = None, None, None
```

```
    dx = dout.dot(w.T).reshape(x.shape)
```

```
    dw = x.reshape(x.shape[0], -1).T.dot(dout)
```

```
db = np.sum(dout, axis=0)
```

```
assert dx.shape == x.shape, "dx.shape != x.shape: " + str(dx.shape) + " != " + str(x.shape)
assert dw.shape == w.shape, "dw.shape != w.shape: " + str(dw.shape) + " != " + str(w.shape)
assert db.shape == b.shape, "db.shape != b.shape: " + str(db.shape) + " != " + str(b.shape)
```

```
return dx, dw, db
```

```
def relu_forward(x):
```

```
    """
```

```
    Computes the forward pass for a layer of rectified linear units (ReLU).
```

```
    Input:
```

```
    - x: Inputs, of any shape
```

```
    Returns a tuple of:
```

```
    - out: Output, of the same shape as x
```

```
    - cache: x
```

```
    """
```

```
    out = None
```

```
    out = np.maximum(x, 0)
```

```
    cache = x
```

```
    return out, cache
```

```
def relu_backward(dout, cache):
```

```
    """
```

```
    Computes the backward pass for a layer of rectified linear units (ReLU).
```

```
    Input:
```

```
    - dout: Upstream derivatives, of any shape
```

```
    - cache: Input x, of same shape as dout
```

```
    Returns:
```

```
    - dx: Gradient with respect to x
```

```
    """
```

```
    dx, x = None, cache
```

```
    mask = x > 0
```

```
    dx = dout * mask
```

```
    return dx
```

```
def dropout_forward(x, dropout_param):
```

```
    """
```

```
    Performs the forward pass for (inverted) dropout.
```

```
    Inputs:
```

```
    - x: Input data, of any shape
```

- dropout\_param: A dictionary with the following keys:
  - p: Dropout parameter. We drop each neuron output with probability p.
  - mode: 'test' or 'train'. If the mode is train, then perform dropout; if the mode is test, then just return the input.
  - seed: Seed for the random number generator. Passing seed makes this function deterministic, which is needed for gradient checking but not in real networks.

Outputs:

- out: Array of the same shape as x.
- cache: tuple (dropout\_param, mask). In training mode, mask is the dropout mask that was used to multiply the input; in test mode, mask is None.

```
"""
p, mode = dropout_param['p'], dropout_param['mode']
if 'seed' in dropout_param:
    np.random.seed(dropout_param['seed'])
```

```
mask = None
out = None
```

```
if mode == 'train':
```

```
elif mode == 'test':
```

```
cache = (dropout_param, mask)
out = out.astype(x.dtype, copy=False)
```

```
return out, cache
```

```
def dropout_backward(dout, cache):
```

```
"""
Perform the backward pass for (inverted) dropout.
```

Inputs:

- dout: Upstream derivatives, of any shape
- cache: (dropout\_param, mask) from dropout\_forward.

```
"""
dropout_param, mask = cache
p, mode = dropout_param['p'], dropout_param['mode']
```

```
dx = None
if mode == 'train':
```

```
elif mode == 'test':
    dx = dout
return dx
```

```
def softmax_loss(x, y):
```

```
"""
Computes the loss and gradient for softmax classification.
```



Inputs:

- x: Input data, of shape (N, C) where  $x[i, j]$  is the score for the  $j$ th class for the  $i$ th input.
- y: Vector of labels, of shape (N,) where  $y[i]$  is the label for  $x[i]$  and  $0 \leq y[i] < C$

Returns a tuple of:

- loss: Scalar giving the loss
- dx: Gradient of the loss with respect to x

```
"""
shifted_logits = x - np.max(x, axis=1, keepdims=True)
Z = np.sum(np.exp(shifted_logits), axis=1, keepdims=True)
log_probs = shifted_logits - np.log(Z)
probs = np.exp(log_probs)
N = x.shape[0]
loss = -np.sum(log_probs[np.arange(N), y]) / N
dx = probs.copy()
dx[np.arange(N), y] -= 1
dx /= N
return loss, dx
```

## Lisa 11.: Optim kood

```
import numpy as np
```

```
"""
```

This file implements various first-order update rules that are commonly used for training neural networks. Each update rule accepts current weights and the gradient of the loss with respect to those weights and produces the next set of weights. Each update rule has the same interface:

```
def update(w, dw, config=None):
```

Inputs:

- w: A numpy array giving the current weights.
- dw: A numpy array of the same shape as w giving the gradient of the loss with respect to w.
- config: A dictionary containing hyperparameter values such as learning rate, momentum, etc. If the update rule requires caching values over many iterations, then config will also hold these cached values.

Returns:

- next\_w: The next point after the update.
- config: The config dictionary to be passed to the next iteration of the update rule.

NOTE: For most update rules, the default learning rate will probably not perform well; however the default values of the other hyperparameters should work well for a variety of different problems.

For efficiency, update rules may perform in-place updates, mutating w and setting next\_w equal to w.

```
"""
```

```
def sgd(w, dw, config=None):
```

```
"""
```

Performs vanilla stochastic gradient descent.

config format:

- learning\_rate: Scalar learning rate.

```
"""
```

```
if config is None: config = {}
```

```
config.setdefault('learning_rate', 1e-2)
```

```
w -= config['learning_rate'] * dw
```

```
return w, config
```

```
def sgd_momentum(w, dw, config=None):
```

```
"""
```

Performs stochastic gradient descent with momentum.

```

config format:
- learning_rate: Scalar learning rate.
- momentum: Scalar between 0 and 1 giving the momentum value.
  Setting momentum = 0 reduces to sgd.
- velocity: A numpy array of the same shape as w and dw used to store a
  moving average of the gradients.
"""

```

```

if config is None: config = {}
config.setdefault('learning_rate', 1e-2)
config.setdefault('momentum', 0.9)
v = config.get('velocity', np.zeros_like(w))

next_w = None

v = config['momentum']*v - config['learning_rate']*dw
next_w = w + v

config['velocity'] = v

return next_w, config

```

```

def rmsprop(x, dx, config=None):

```

```

    """
    Uses the RMSProp update rule, which uses a moving average of squared
    gradient values to set adaptive per-parameter learning rates.

```

```

    config format:
    - learning_rate: Scalar learning rate.
    - decay_rate: Scalar between 0 and 1 giving the decay rate for the squared
      gradient cache.
    - epsilon: Small scalar used for smoothing to avoid dividing by zero.
    - cache: Moving average of second moments of gradients.
    """

```

```

if config is None: config = {}
config.setdefault('learning_rate', 1e-2)
config.setdefault('decay_rate', 0.99)
config.setdefault('epsilon', 1e-8)
config.setdefault('cache', np.zeros_like(x))

next_x = None

config['cache'] = config['decay_rate'] * config['cache'] + (1 - config['decay_rate']) * dx**2
next_x = x - config['learning_rate'] * dx / (np.sqrt(config['cache']) + config['epsilon'])

return next_x, config

```

```

def adam(x, dx, config=None):

```

```
"""
```

Uses the Adam update rule, which incorporates moving averages of both the gradient and its square and a bias correction term.

config format:

- learning\_rate: Scalar learning rate.
- beta1: Decay rate for moving average of first moment of gradient.
- beta2: Decay rate for moving average of second moment of gradient.
- epsilon: Small scalar used for smoothing to avoid dividing by zero.
- m: Moving average of gradient.
- v: Moving average of squared gradient.
- t: Iteration number.

```
"""
```

```
if config is None: config = {}
config.setdefault('learning_rate', 1e-3)
config.setdefault('beta1', 0.9)
config.setdefault('beta2', 0.999)
config.setdefault('epsilon', 1e-8)
config.setdefault('m', np.zeros_like(x))
config.setdefault('v', np.zeros_like(x))
config.setdefault('t', 1)

next_x = None

config['t'] += 1
config['m'] = config['beta1']*config['m'] + (1-config['beta1'])*dx
config['v'] = config['beta2']*config['v'] + (1-config['beta2'])*(dx**2)
mt = config['m'] / (1-config['beta1']**config['t'])
vt = config['v'] / (1-config['beta2']**config['t'])
next_x = x - config['learning_rate'] * mt / (np.sqrt(vt) + config['epsilon'])

return next_x, config
```