TALLINNA TEHNIKAÜLIKOOL
Infotehnoloogia teaduskond

Elina Kuldkepp 183187IAPM

SOOLISE PALGALÕHE MODELLEERIMINE MASINÕPPE MEETODITEGA

Magistritöö

Juhendaja: Sven Nõmm, Phd
Marge Unt, Phd
Kaja Sõstra, Phd

Tallinn 2020

Autorideklaratsioon

Kinnitan, et olen koostanud antud lõputöö iseseisvalt ning seda ei ole kellegi teise poolt varem kaitsmisele esitatud. Kõik töö koostamisel kasutatud teiste autorite tööd, olulised seisukohad, kirjandusallikatest ja mujalt pärinevad andmed on töös viidatud.

Autor: Elina Kuldkepp

04.08.2020

Annotatsioon

Käesoleva magistritöö põhieesmärgiks on ennustada ja selgitada Eestis olemasolevat palgalõhet, et hiljem, vastavalt saadud tulemustele, seda mõtestatult vähendama hakata. Seni toimunud iga-aastased palgalõhe uuringud on küll täpselt suutnud valitsevat palgalõhe arvutada, kuid tunnuste vähesuse tõttu, ei ole suudetud seda piisavalt hästi selgitada.

Iga aasta oktoobris peavad tööandjad täitma palgalõhe uuringu küsimustiku. Iga neljas aasta asendatakse palgalõhe uuring veidi põhjalikuma töötasu struktuuri uuringuga. Kogutud andmete põhjal arvutatakse palgalõhe.

Autor osales projektis "Soolise palgalõhe kirjeldamine ja seletamine", mille raames täiendati 2014. aasta töötasu struktuuri uuringut eri registritest pärit andmetega, mille näol on tegemist sellisel kujul uudse ja varasemast kindlasti mahukama andmestikuga, mis võimaldab kasutada analüüsis varasemast detailsemaid alagruppe (ametigrupid, haridusgrupid). Uuringu uudsus seisneb ka metoodikate kasutuses. Lisaks tavalistele palga dekomponeerimise meetoditele (Blinder-Oaxaca, kvantiilregressioon), kasutati ka alternatiivsemaid meetodeid meeste ja naiste tunnipalga seletamiseks, rakendades selleks valikut masinõppe ja andmekaeve meetoditest. Kuna autor viis projektis läbi alternatiivsete meetodite osa siis käesolev magistritöö võib paljuski sarnaneda projekti "Soolise palgalõhe kirjeldamine ja seletamine" raportiga.

Lisaks palgalõhe selgitamisele on käesoleva magistritöö eesmärgiks ka välja selgitada, kas ilma palgalõhe uuringuta, kasutades teistest registritest pärit andmeid, on võimalik palgalõhe sama täpselt arvutada. Palgalõhe uuringu küsimustiku täitmine on tööandjatele suureks ja aeganõudvaks kohustuseks.

Lõputöö on kirjutatud eesti keeles ning sisaldab teksti 53 leheküljel, 7 peatükki, 34 joonist, 13 tabelit.

Abstract

Wage Gap Modelling with Machine Learning Methods

The main goal for this thesis is to explain and predict the existing gender pay gap in Estonia in order to reduce it meaningfully. So far, the yearly researches have correctly calculated existing wage gap, but due to the shortage of the available characteristics, researchers have failed to explain it sufficiently.

Every year, in October, employers must participate in the wage gap questionnaire. In every fourth year, pay gap questionnaire is replaced with Structure of Earnings questionnaire which is more thorough. Finally, the pay gap is calculated from the collected data.

Author of this thesis participated in the project „Soolise palgalõhe kirjeldamine ja seletamine“ where the Structure of Earnings questionnaire was improved with new data from many other databases. The improved dataset should help to explain wage gap to successfully reduce it. Novelty of the study is also with the usage of the methods. In addition to the usual wage gap decompositioning methods, such as Blinder-Oaxaca and quantile regression, alternative methods were used to explain female and male hourly pay. Such as a Selection of machine learning and data mining methods.

In addition of explaining the pay gap, one of the goals of this thesis was also explore the possibility to calculate the wage gap by replacing the wage gap questionnaire with synthesized wage information with the same accuracy. This is very important as filling the wage gap questionnaire is bothersome and time consuming for the employers.

The thesis is in Estonian and contains 53 pages of text, 7 chapters, 34 figures, 13 tables.

Lühendite ja mõistete sõnastik

Palgalõhe	Mees- ja naistöötajate palkade erinevus ((meeste keskmine brutopalk - naiste keskmine brutopalk)/meeste keskmine brutopalk).
Karakteristik ehk tegur	Tunnus, mis on kaasatud analüüsi sõltuva tunnuse variatiivsuse seletamiseks.
Selgitamata palgalõhe	Osa palgalõhest, mida ei ole võimalik selgitada erinevustega meeste ja naiste karakteristikutes. Selle aluseks on erinevused meeste ja naiste palgavõrrandite regressioonikordajates.
Selgitatud palgalõhe	Osa palgalõhest, mida on võimalik selgitada erinevustega meeste ja naiste karakteristikutes.
Vaadeldud palgalõhe	Koosneb palgaerinevuste selgitatud ja selgitamata osast, nende kokku liitmisel saadakse vaadeldud palgalõhe.
TSD	Deklaratsioon, kus on kuuliselt ja sissetuleku tüübi kaupa info töötaja sissetulekute ja maksude kohta.
Töötasu struktuur(TSU)	Iga nelja aasta tagant toimuv valikuuring, kus on valitud töötajate kohta muuhulgas info oktoobri bruto tunnitasu ilma lisatasudeta. Asendab ilmumisaastatel palgalõhe uuringut.
Palk ja tööjõud	Firmapõhine deklaratsioon, kus on info muuhulgas firmas töötavate inimeste arvu ning neile makstavate summade kohta.
Palgalõhe uuring	Iga-aastane soopõhine uuring, et arvutada palgalõhet.
EMTAK kood	Eesti majanduse tegevusalade klassifikaator.
Ruutviga(SSE)	Kasutatakse tihti lineaarse regressiooni puhul. Tegeliku andmepunkti väärtusest lahutatakse ennustatud andmepunkti väärtus, mille vahe võetakse ruutu ning liidetakse otsa juba varem arvutatud tegeliku ja ennustatud andmepunktide vahele [1].
F - test	F-testi kasutatakse, et teha kindlaks, kas sõltuva ning kõikide sõltumatute muutujate vahel esineb statistiliselt oluline seos [2].
Gini indeks	Gini indeks mõõdab tunnuse jaotust sõltuva muutuja suhtes. Gini indeks jääb vahemikku 0 - 1. Mida väiksem Gini indeks, seda ühtlasemalt on tunnus jaotatud sõltuva muutuja suhtes [1].
Gradient	Funktsioon, mis kirjeldab suunda mingis ruumipunktis. Veagradient kirjeldab vea suurenemis/vähenemis suunda [3].
Sigmoid	Aktiveerimisfunktsioon, mis surub väljundi 1 ja 0 vahele. Peamiselt kasutatakse mudelitel, mille väljund on tõenäosus [4].

Tanh	Aktiveerimisfunktsioon, mis surub väljundi -1 ja 1 vahele. Kasutatakse enamasti klassifitseerimisel kahe klassi vahel [4].
Ground truth	Etalonväärtus, mida juhendatud masinõppe meetodid üritavad mudelitega ennustada ning mille järgi mudelid kalibreerivad enda parameetreid.
Epoch	*Epoch* on number, mitu korda on närvivõrgus kaalusid uuendatud. Tihti ka number mitu korda on terve andmestik algoritmi poolt läbi töödeldud.
Klaaslagi	Naiste liikumine juhtivatele ametikohtadele on takistatud nii organisatsiooni kultuuri kui ka praktikate tõttu.
Kleepuv põrand	Olukord, kus mehi ja naisi edutatakse küll sarnaselt, kuid naiste samal (juhtival) positsioonil töötamist tasutatakse madalamalt.

Sisukord

Jooniste loetelu

Tabelite loetelu

1. Sissejuhatus

Võrdne palk, võrdse töö eest, on olnud üks Euroopa Liidu alustalasi liidu loomisest saati [5]. Et seda tagada ja kontrollida kogutakse andmeid, iga aasta arvutab iga riik palgalõhe suuruse ning iga nelja aasta tagant antakse välja ka üleeuroopaline palgalõhe uuring "Structure of Earnings Survey", edaspidi SES, kus iga riik on saatnud andmeid ettevõtete kohta, kus on vähemalt 10 töötajat. Kogutud andmed on nii töötajate individuaalsed omadused (sugu, vanus, haridus, töökogemus, amet, töökoormus ja lepingu tüüp) kui ka firmade enda omadused (tegevusala, firma töötajate arv ning sektor). Aastal 2018 anti välja palgalõhe uuring, mis põhines 2014. aasta SES andmetel, kus selgus, et Eesti palgalõhe on Euroopa Liidu üks kõrgemaid (25.6%) [6]. Kuigi Eesti palgalõhe on läinud iga aastaga väiksemaks - 2017 aasta oktoobris oli palgalõhe 20.9% ning aasta hiljem, 2018 aasta oktoobris, teenisid naispalgatöötajad 18.7% väiksemat brutotöötasu, asub Eesti siiski võrreldes teiste Euroopa riikidega, palgaebavõrdsuse tipus [7].

Et palgalõhet märkimisväärselt vähendada, on tähtis seda paremini tundma õppida.

Kuigi üldise soolise palgalõhe arvutamine on üsna lihtne, on märksa keerulisem olemasoleva palgalõhe selgitamine. Klassikaliselt jagatakse sooline palgalõhe selgitatud ja selgitamata osaks. Palgalõhe selgitatud ossa kuuluvad reeglina n-ö objektiivsed tegurid nagu näiteks see, et mehed ja naised kalduvad töötama erinevatel ametialadel, on koondunud erinevatesse sektoritesse, omavad erinevat tööturustaaži (naistel enam pere-eluga seotud töökatkestusi) jne. Palgalõhe selgitamata osa viitab aga sellele, et erinevad tööturutegurid omavad meeste ja naiste jaoks tööturul erinevat tähendust (nt hariduse roll meeste ja naiste palgataseme kujunemisel võib olla erinev). Niisamuti jäävad selgitamata osasse sageli need individuaalsed ja tööturukarakteristikud, mida ei ole mõõdetud või mida ei saa mõõta, aga ka otseselt tööturul aset leidev diskrimineerimine [8].

Käesoleva magistritöö raames on põhieesmärgiks ka laiendada seletavate tunnuste hulka ning analüüsimeetodeid/viise. Kasutatakse lisaandmeid, mida ei ole varem üldiselt palgalõhe selgitamiseks kasutatud: laste arv, emakeel, rahvus, puude olemasolu, puhkusepäevad, haiguspäevad, elukoht. Rakendatakse lisaks nn klassikalistele meetodile nagu seda on Oaxaca-Blinderi dekompositsioon ning kvantiilregressioon ka masinõppe ning andmekaeve meetodeid.

Viimase sammu eesmärgiks on püüd avastada palgaandmetes (uusi) seoseid ja struktuure, mis aitaks ühelt poolt olemasolevaid tulemusi paremini mõista ning teiselt poolt tulevasi analüüse täiendada.

Lisaks palgalõhe selgitamisele on käesoleva magistritöö eesmärgiks ka välja selgitada, kas ilma palgalõhe uuringuta, kasutades ainult teistest registritest pärit andmeid, on võimalik palgalõhe sama täpselt arvutada. Statistikaamet soovib, loodud mudelite töökindluse korral, loobuda palgalõhe uuringust juba 2020 oktoobris.

2. Taust

Euroopa Liit teeb iga nelja aasta tagant palgalõhe uuringu, et mõista erinevate Euroopa riikide palgalõhe olukorda. Siiamaani on nii Eesti kui ka teised Euroopa riigid selleks kasutanud kindlaid andmeid - nii töötajate individuaalsed omadused (sugu, vanus, haridus, töökogemus, amet, töökoormus ja lepingu tüüp) kui ka firmade enda omadused (tegevusala, firma töötajate arv ning sektor). Samuti on üldiselt kasutatud kindlaid klassikaliselt palgalõhe seletamiseks kasutatavaid meetodeid - Oaxaca-Blinderi dekompositsioon (Euroopa sotsiaalse statistika juhid valisid selle peamiseks palgalõhet seletavaks meetodiks) ning kvantiilregressioon [6]. Kuna 2014. aasta Euroopa palgalõhe uuringust (SES), selgus, et Eesti palgalõhe on Euroopa Liidus üks suurimaid (25.6%), otsustati Eestis teha detailsem uuring, et leida nii suure palgalõhe taga täpsemaid põhjuseid ja mustreid. Samuti on Euroopa Komisjon viidanud korduvalt kriitilisele vajadusele Eestis soolise palgalõhe vähendamiseks. Selleks telliti Sotsiaalministeeriumi poolt uurimisprojekt „Soolise palgalõhe vähendamine“ (lühendiga REGE, mis tuleneb inglise keelsest sõnast „REducing GEnder wage gap“) ning pandi kokku töögrupp inimestega Tallinna Ülikoolist, Tallinna Tehnikaülikoolist ning Statistikaametist. Uuringu ülesanne on välja pakkuda lahendusi palgalõhe tõhusamaks jälgimiseks ning hindamiseks, tuues välja sissetulekute ebavõrdsuse analüüsimiseks sobilikke näitajaid ja indikaatoreid [9]. 2014. aasta SES – i, täiendati erinevatest registritest pärit andmetega (nt Rahvastikuregister, Sotsiaalkindlustusameti sotsiaalkaitse süsteemi andmed, tulu- ja sotsiaalmaksu deklaratsiooni andmed, rahvaloenduse andmed). Sellisel kujul on tegemist uudse ning varasemalt mahukama andmestikuga. REGE raames sooviti kasutada ka lisaks klassikalistele palgalõhe meetoditele ka alternatiivseid andmekaeve ja masinõppe meetodeid, mida varem ei ole Eestis sellises kontekstis kasutatud, et avastada andmetes võimalikke uusi struktuure, mis võiks omakorda aidata kaasa nii REGE analüüsi puhul kui ka tulevikus palgalõhe paremale mõistmisele ja seletamisele Eestis. Plaanitud masinõppe ja andmekaeve meetodite puhul keskendutakse (tulenevalt nende loogikast ja rakendusvõimalustest) meeste ja naiste tunnipalga analüüsile – kas ja millised tegurid meeste ja naiste palka Eestis enim seletavad ning kuivõrd need mustrid erinevad või sarnanevad. Lisaks rakendame neid meetodeid detailsemaks analüüsiks ametigruppide või tegevusalade sees. Andmekaeve ja masinõppe meetodite osa viis läbi käesoleva magistritöö autor.

Varasemad Eesti palgalõhe uuringud on toonud välja, et soolist palgalõhet selgitab ja mõjutab Eestis peamiselt tööturu segregatsiooniga seonduv, st erinevad töökoha, töötamise ning ettevõttega seotud tegurid. Lisaks jääb väga suur osa sooliselt palgalõhest selgitamata (selgitatud 15%) [10]. Objektiivsed struktuursed ja individuaalsed tegurid on seni palgalõhet selgitanud üsna tagasihoidlikult ning suurem osa vaadeldud palgalõhest on jäänud selgitamata.

REGE projekti, mille üks eesmärk on kirjeldada ja seletada palgalõhet Eestis, keskendume peamiselt mikrotasandile ehk indiviidi ja tema vahetu keskkonna teguritele. Võimaluste piires kaasame analüüsi ka mesotasandi ehk töökoha faktoreid, et hoomata ettevõtte tasandi tegurite rolli ja olulisust palgalõhe selgitamisel. Selleks rakendatakse klassikalisi analüüsiviise ja -meetodeid, täpsemalt Blinder-Oaxaca dekomponeerimismeetodit ja kvantiilregressiooni, mida on rakendatud ka varasemates palgalõhe analüüsides Eestis [8].

Otsides seotud tööde artikleid, oli peamiseks kriteeriumiks leida, milliseid meetodeid on üldiselt palgalõhe selgitamiseks kasutatud – kas pigem kasutatakse klassikalisi meetodeid või esineb palju juba ka alternatiivseid meetodeid (masinõpe, andmekaeve). Alternatiivsete meetodite olemasolul võrrelda tulemusi klassikaliste meetoditega.

USA – 2017. aastal välja antud artiklis „The Gender Wage Gap: Extent, Trends, and Explanations“, kasutati nii Oaxaca-Blinderi dekompositsioon kui ka kvantiilregressiooni. Oaxaca puhul jaotati testimine kahte rühma – töötaja endaga seotud muutujad ning kõik muutujad koos. Esimese rühma puhul saadi selgitatud osaks 15%, teise rühma puhul 62%. Tunnused ei olnud küll täpselt välja toodud, kuid kuulusid järgmistesse gruppidesse: Hariduse, kogemusega, elukohaga, rassiga, ametiga ja tegevusalaga seotud tunnused. Kasutati ka OLS meetodit, mis on mitme sõltumatu muutujaga lineaarregressioon. Kuid erinevalt käesoleva magistritöö autorist, ei kasutatud OLS regressiooni kõige rohkem palka ja palgalõhet mõjutavate tunnuste leidmiseks, vaid arvutati valem, millel on kõige väiksem viga kasutades kindlaid muutujaid [11].

Serbia – 2017. aastal välja antud artiklis „Application of the Mincer Earning Function in Analyzing Gender Pay Gap in Serbia“ tuuakse välja, et eelnevalt on ka Serbias kasutatud nii Oaxaca-Blinderi dekompositsioon kui ka kvantiilregressiooni. Artiklis endas aga kasutati Mincer teenimise funktsiooni, mis modelleerib haridusaastaid ning ruutfunktsiooni

töökogemustest aastates. Meetodit kasutati selleks, et välja selgitada, kas sugu mõjutab palga kujunemist. Tulemuseks oli, et nii sugu, haridus kui ka töökogemus mõjutavad palka [12].

Paraguay – 2017. aasta Luz Marcela Pineda Mariuci magistritöös kasutatakse OLS – i ning kvantiilregressiooni. Kasutatud tunnused olid vanus, haridus, suhtestaatus, laste arv, elukoht, kooselu vanema põlvkonna esindajaga, firma suurus, amet, tegevusala, lepingu tüüp. Üldiselt uuriti, kas palgalõhe eksisteerib ja milliste tunnuste sees on palgalõhe kõige suurem [13].

Ühendkuningriigid – 2020. aasta artikkel „Understanding Pay Gaps“. Uurimisküsimusteks olid - kas palgalõhe on seletatav erinevate töötajate tunnustega? Ja kui palju palgalõhest suudetakse ära kirjeldada nende tunnustega? Uuringus kasutatakse nii Oaxaca-Blinderi dekompositsioon kui ka kvantiilregressiooni. Kasutatud tunnused olid kogemus, elukoht, vanus, haridus, tööaeg, amet, valdkond, sektor ja laste arv. Tulemustest selgus, et kõige paremini kirjeldavad palgalõhet tööga seotud tunnused ning selgitamata osa on umbes 50% [14].

Vahemeremaad – 2009. aasta aruteludokumendis „Gender Pay Gap and Quantile Regression in European Families“ uuritakse erinevaid palgalõhega seotud küsimusi – ühtedeks põhilisemateks uurimisküsimusteks on kas vahemere maades esineb klaaslae või kleepuva põranda efekti. Selgus, et vahemeremaades kleepuva põranda efekt on üpriski aktuaalne kuigi klaaslae efekt on enamus riikides vähenenud alates 2006. aastast. Samuti tuleb ka sellest uuringust välja, et suurem osa palgalõhest on selgitamata palgalõhe. Kasutati nii Oaxaca-Blinderi dekompositsioon kui ka kvantiilregressiooni, samuti ka OLS – i [15].

Eelnevate näidete põhjal saab väita, et üldiselt kasutatakse palgalõhe selgitamiseks klassikalisi ja palgalõhe selgitamiseks läbi töötatud meetodeid - Oaxaca-Blinderi dekompositsioon kui ka kvantiilregressiooni, palju esines ka OLS regressiooni. Samuti olid ka tulemused väga sarnased. Üldiselt jäi selgitamata osa suremaks kui selgitatud ning põhilised selgitatud osa tunnusteks olid erinevad töökoha, töötamise ning ettevõttega seotud tegurid. Alternatiivsete meetodite kasutamise kohta leidis autor vähe näiteid. Üks põhjalikumaid artikleid selle kohta on 2020. aastal ilmunud „The Gender Pay Gap Revisited: Does Machine Learning offer New Insights?“. Artikkel otsib vastuseid muuhulgas ka küsimusele – Millised tunnused mõjutavad enim palka? Selleks kasutatakse mudeli ehitamist koos „post-double-LASSO“ meetodiga. Lasso regressioon on üks lineaarregressiooni tüüpe, mida kasutatakse mõjutavate tunnuste leidmiseks ning ennustamise parendamiseks. Lasso meetod kasutab kahanemist, milles

vähendatakse väärtuste kaugust soovitud keskpunkti suunas. Kasutatud tunnused jaotati järgmistesse klassidesse: inimkapital, tööturg, demograafia, isikuomadused, firma omadused ning ametiga seotud tunnused. Kasutatud muutujaid oli 5821, mistõttu on keeruline välja tuua ka valituks osutunud muutujaid [16].

Käesoleva magistritöö teiseks osaks on palgalõhe ennustamine. Hetkel peavad tööandjad iga aasta täitma palgalõhe küsitluse. Et tööandjate koormust vähendada, otsustati proovida, kui täpselt saaks arvutada palgalõhe, ilma tööandjate küsitlusvormi täitmiseta.

Otsides artikleid seotud tööde kohta oli peamiseks kriteeriumiks leida, milliseid meetodeid on kasutatud palga ennustamiseks ning, mis on nende tulemused.

Autor ei leidnud ühtegi artiklit palgalõhe ennustamise kohta. Küll aga oli palju artikleid, projekte ja töid, kus ennustati palka. Erinevalt käesoleva magistritöö autorist, ei kasutatud eriti kuupalga või tunnipalga ennustamist vaid ennustati palga vahemikku. Seetõttu ei ole järgnevatel artiklitel ka tulemusi välja toodud, sest need ei ole omavahel võrreldavad.

Artiklis „Analysis of Web Scraped Job Data to Predict Relative Salaries“, sooviti teada, kas töötaja palk on suurem või väiksem mediaanväärtusest. Kasutatud tunnused olid linn, amet. firmanimi, asukoht ja töökoha kokkuvõtte kirjeldus (kasutati kirjelduses esinenud sõnu). Ennustamiseks kasutatakse logistilist regressiooni [17].

Uuringus „Salary Prediction in It Job Market“ jaotati palgad kümnesse klassi, vahemikus 45000-1000000 ning prooviti ennustada, millisesse klassi etteantud palk kuulub. Ennustamiseks kasutati otsustuspuud, otsustusmetsa ja SVM – i [18].

Artiklis „A Comparative Study of Performances of Various Classification Algorithms for Predicting Salary Classes of Employees“ oli ennustamiseks taas vaid kaks klassi – väiksem palk kui 50 000 ning sellest suurem või võrdne. Tunnuseid oli 15, mille seas olid muuhulgas näiteks sugu, haridustase ja firma tüüp. Ennustamiseks kasutati otsustuspuud, Naive-Bayesit ning SVM-i [19].

Uuringus „Salary Prediction in It Job Market with Few High-Dimensional Samples: A Spanish Case Study“ jaotati palgad ennustamiseks nelja klassi – madal, madal-keskmine, kõrge-keskmine ja madal. Kasutatud tunnuste seas olid haridustase, firma suurus, kogemus,

lepingutüüp, töökoormus, amet. Ennustamiseks kasutati muuhulgas otstusmetsa, otsustuspuud, K-lähim naabrit, lineaarregressiooni ja SVM – i [20].

3. Uurimisküsimused

1. Millised tunnused kirjeldavad palga kujunemist, nii soolises lõikes kui ka erinevate EMTAK gruppide siseselt, kasutades lineaarse regressiooni mudeli ehitamist;

2. Kui hästi on võimalik ennustada ametigruppide siseselt, kas vastava suurusega tunnipalk kuulub:

    a. Kõrgepalgalisele mehele;

    b. Kõrgepalgalisele naisele;

    c. Madalapalgalisele mehele;

    d. Madalapalgalisele naisele

    Eeldades, et mida parem on ennustusvõime, seda diskrimineerivam on palga maksmine vastavas ametigrupis. Tulemuste saamiseks kasutatakse otsustuspuud.

3. Kas on võimalik vähendada tööandjate koormust? - Tööandjad peavad iga aasta täitma küsitluse palgalõhe kohta (Palgalõhe uuring). Enamus tööandjate seas tekitab see pahameelt ning lisatööd. Kui täpselt on võimalik ennustada töötajate brutopalka ilma lisatasudeta ning kuu jooksul töötatud tunde, kasutades alternatiivseid andmestikke? Kas on võimalik asendada palgalõhe uuring?

4. Andmed

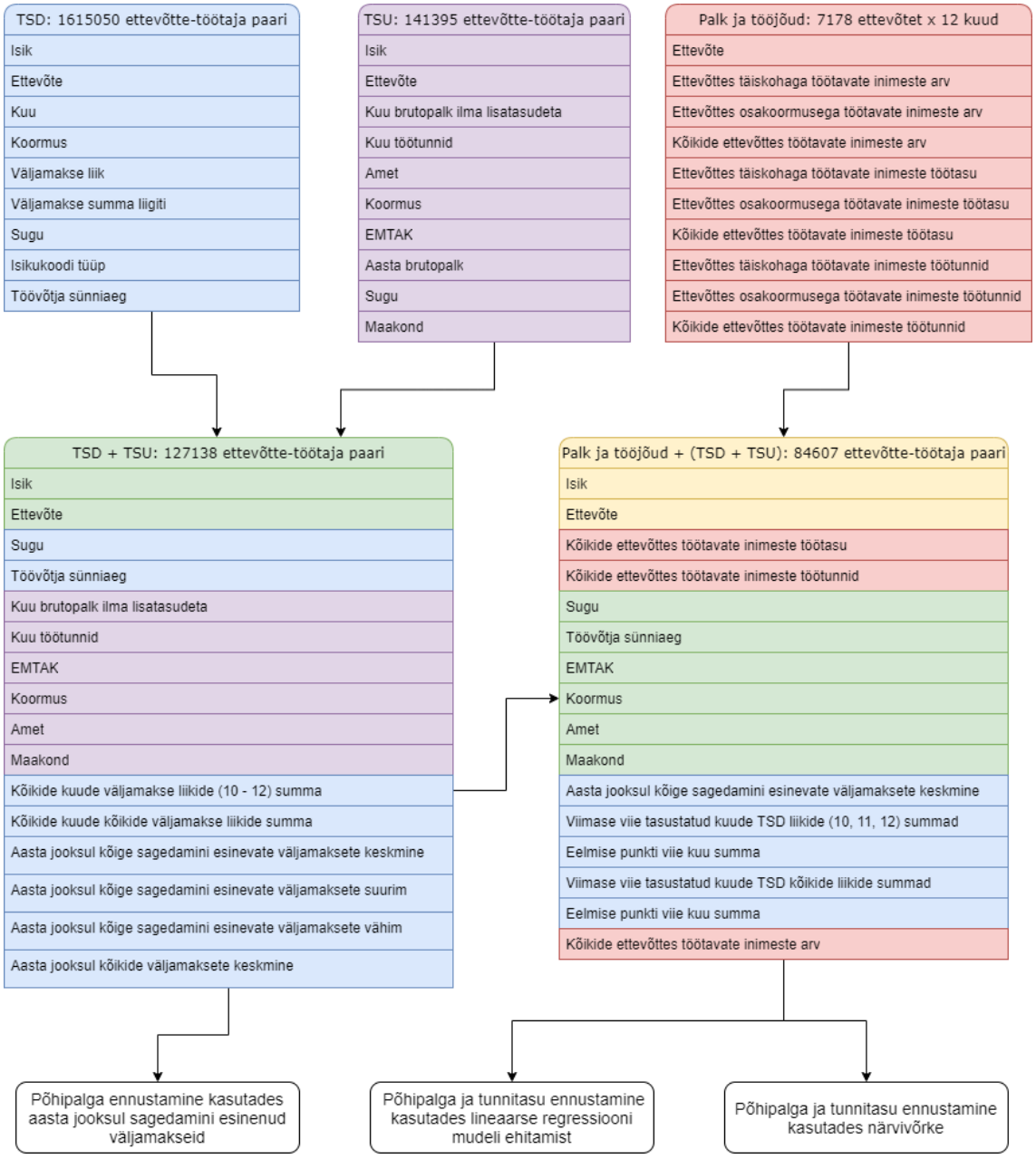
Autor kasutab käesolevas töös järgmiseid tabeleid:

2014. aasta töötasu struktuuri uuring(TSU) - Uuring, mida viiakse läbi iga nelja aasta tagant. Tabelis on info erinevate firmade töötajate kohta - oktoobrikuu brutopalk ilma lisatasudeta, EMTAK kood, oktoobrikuu töötatud tundide arv, amet, maakond, sugu, koormus, aasta brutopalk. Uuringu andmeid laiendati teistest registritest pärit andmetega (nt Rahvastikuregister, Sotsiaalkindlustusameti sotsiaalkaitse infosüsteemi andmed, tulu- ja sotsiaalmaksu deklaratsiooni andmed, rahvaloenduse andmed) - haridustase, sektor, vanus, laste arv, puude olemasolu. Tekkinud tabeli esialgsest 125394-st andmereast, jäi peale ebatäielike (ühe või mitme analüüsiks vajaminevate muutujate kohal puudub väärtus) ridade eemaldamist, alles 111056 andmerida.

2018. Aasta töötasu struktuuri uuring(TSU) - Uuring, mida viiakse läbi iga nelja aasta tagant. Tabelis on info erinevate firmade töötajate kohta - oktoobrikuu brutopalk ilma lisatasudeta, EMTAK kood, oktoobrikuu töötatud tundide arv, amet, maakond, sugu, koormus, aasta brutopalk. Tabelis oli 141395 andmerida, millest jäi peale ebatäielike ridade eemaldamist järgi 127138 andmerida.

2018 TSD - Tabel, kus on erinevate firmade töötajate kohta nende sissetulekud välja toodud väljamakse liigi järgi. Igal real on summa, erineva väljamaksega valitud kuu kohta. Tabelis oli 11216573 rida, 1615050 erineva firma-töötaja paari kohta. Kuid kasutusse läksid ainult nende firma-töötaja paarid, kelle info oli ka 2018 aasta töötasu struktuuri uuringus, et panna kokku kahes erinevast tabelist tulnud info. TSD tabelis ei ole eraldi välja toodud põhi brutopalka ilma lisatasudeta, kuid TSU tabelis on. TSU tabelist tulnud põhi brutopalka ilma lisatasudeta, kasutatakse edaspidi mudelite ennustamisel *ground truth-ina.*

2018 Palk ja tööjõud - Tabel, kus on kuude kaupa välja toodud firmade andmed - töötajate arv, töötajatele makstav töötasu summa ilma lisatasudeta, töötajate poolt töötatud tundide summa. Tabelis on 61018 rida, 7178 firma kohta, millest läks kasutusse 2652 firmat, kelle töötajad on esindatud TSD + TSU koondtabelis.



Joonis 1. Põhipalga ja tundide ennustamiseks kasutatavad andmed

5. Lahendus

5.1 Palka mõjutavate tunnuste analüüs

Selle analüüsi eesmärk on avastada andmetest võimalikke seoseid ja struktuure, mida me klassikalise palgaregressiooni puhul ehk eeldada või oodata ei oskaks. Selles mõttes on tegemist „avastuslikuma“ meetodiga, mis võiks ideaalis pakkuda alternatiivset vaadet (meeste ja naiste) palga kujunemisele (täpsemalt selle seletamisele) Eestis. Fookuses on siin (sarnaselt palgaregressiooniga) indiviidi palk ehk eri palgatasemete kujunemine ja seletamine.

5.1.1 Lineaarse regressiooni mudeli ehitamine

Selle lähenemise esimese sammuna rakendame tunnipalga seletamiseks lineaarset regressioonimudeli loogikat, et ennustada sõltuva muutuja (tunnipalk) väärtust, kasutades ühte või enamat sõltumatut muutujat [21]. Sõltumatute muutujatena kasutati vanust, haridustaset, tööstaaži, koormust, lepingu tüüpi, nädala normi, töötunde, tasustatud päevi, puhkuse päevi, ametikoodi, maakonda, sektorit, emakeelt, rahvust, EMTAK koodi, puuet, sugu, firma suurust ning laste arvu.

Kuna käesolevas töös kasutatakse rohkem kui ühte sõltumatut muutujat nimetatakse seda mitmeseks lineaarseks regressiooniks, mida arvutatakse valemiga:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p, \quad (1)$$

kus $\beta_0,\ \beta_1,\ \beta_2, \ldots, \beta_p$ on kordajad [2]. Kordajaid arvutatakse omakorda valemiga:

$$\hat{\beta} = (X^T X)^{-1} X^T y, \quad (2)$$

kus iga maatriks X rida on sisendite vektor, mille esimene liige on 1 [22].

Fookuses on alustuseks meeste ja naiste palga erinevused, hiljem aga palkade kujunemine tegevusalade (EMTAK peagrupid) lõikes, kuna viimasel (koos ametiga) on varasemate uurimistulemuste (ja etteruttavalt ka siinse palgalõhe dekomponeerimise analüüsi tulemuste) põhjal palgaerinevuste ja -lõhe mõttes oluline seletav jõud. Regressioonimudeli põhine mudeli ehitamine aitab leida, millised sõltumatud muutujad parandavad enim mudeli ennustusvõimet [23]. Kõige paremini mudelit kirjeldavad tunnused leiti, kasutades *forward selection* meetodit,

kus iga iteratsiooniga proovitakse lisada tunnus, mis parandab mudelit kõige paremini. Alustuseks valitakse tunnus, mille põhjal on lineaarse regressiooni ruutviga kõige väiksem. Järgmiseks leitakse tunnus, mis parandab ruutviga koos eelmise tunnusega kõige rohkem, ning mille sarnasus kasutades Pearsoni korrelatsiooni koefitsenti ei ole juba valitud tunnustega üle 0.7. Pearsoni korrelatsiooni koefitsendi leidmine:

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}\ , \tag{3}$$

kus n on kirjete arv ning x ja y on võrdluse all olevad tunnused [22].

Seejärel võrreldakse eelmist ja praegust mudelit. Autor kasutab selleks hüpoteeside statistilist kontrollimist koos F testiga.

Oletatakse, et mingi hüpotees (nimetagem seda nullhüpoteesiks) kehtib. Arutletakse, milliseid hinnangu väärtuseid me nullhüpoteesi kehtides tõenäoliselt võiksime näha. Siis võetakse valim. Kui valimi põhjal arvutatud hinnang tuli ootuspärane, selline nagu ta võiks tulla nullhüpoteesi kehtides, siis võib nullhüpotees õige olla. Kui aga hinnang polnud selline, nagu ta nullhüpoteesi kehtides tõenäoliselt oleks pidanud olema, siis lükatakse nullhüpotees ümber. Nullhüpotees ja alternatiivne hüpotees peavad olema teinetest välistavad [24].

F-skoori arvutatakse järgmise valemiga:

$$F_{stat} = (\frac{SSE_s - SSE_c}{m})(\frac{SSE_c}{n-p-1}), \tag{4}$$

kus SSE_s on SSE väärtus mudelil, millel on vähem muutujaid ja SSE_c on SSE väärtus mudelil, millele proovitakse uut tunnust lisada, m on muutujate arv, mida proovitakse juurde lisada või eemaldada, n on vaatluspunktide arv ning p on tunnuste arv rohkemate tunnustega mudelis [2].

Seejärel tuleb leida F test, mille parameetriteks on number mitu tunnust juurde lisati või maha võeti ning $n - p - 1$, mille tähendus on eespool kirjeldatud.

Kui F – skoor on suurem kui F – test, lükatakse nullhüpotees tagasi, mis tähenda, et tunnuse juurde lisamine parandas mudeli ennustusvõimet.

Kui mudel paraneb oluliselt, lisatakse see muutuja mudelisse ning jätkatakse samasuguse protsessiga, kuni on leitud viis mudelit kõige paremini kirjeldavat tunnust või kuni mudel ei parane enam piisavalt.

Valitud lähenemine osutus valituks, sest lineaarse regressiooni mudeli ehitamine korraga nii ennustab sõltuva muutuja väärtus kui ka saab väga lihtsalt samm-sammu haaval näha, millised tunnused osutusid valituks.

5.1.2 Otsustuspuu

Teise nn alternatiivse lähenemise korral rakendame otsustuspuud, mis kujutab endast struktuurilt vooskeemi, mille tulemusi esitatakse puuna, mille iga sisemine leht on justkui test eri muutujate kohta (nt kas väljas on tuuline või tuulevaikne), iga oks näitab vastava testi tulemust ning välimised lehed esindavad klassi, kuhu vastav andmepunkt ennustuste kohaselt kuulub (otsus, mis tehakse pärast kõikide muutujate testimist) [23]. Teed puu juurest kuni välimise leheni nimetatakse klassifitseerimise reegliteks.

Iga iteratsiooniga proovitakse mudelile lisada uut tunnus, mis osutub, kasutades Gini indeksit, kõige kasulikumaks. Gini indeksi valem:

$$G(v_i) = 1 - \sum_{j=1}^{k} p_j^2, \tag{5}$$

kus $v_1 \ \dots \ v_r$ on r võimalikku väärtust ühes tunnuses, p_j on osa andmepunktidest, millel on väärtus v_i ning, mis kuulub klassi $j \in \{1 \dots k\}$ [23].

Otsustuspuu analüüsi puhul rakendati samu andmestikke ja tunnuseid, aga soost ja tunnipalgast sõltuvana moodustati neli silti/klassi: madalapalgalised naised, kõrgepalgalised naised, madalapalgalised mehed ja kõrgepalgalised mehed. Fookuses on siin ametigrupid ehk analüüs tehakse iga (ISCO peagrupi) ametigrupi sees. Eesmärk on selgitada ametigruppide lõikes välja, kui hästi on võimalik ennustada, millisesse klassi nimetatud andmepunkt kuulub. Klasside leidmiseks arvutati ametikoodi lõikes nii naiste kui ka meeste keskmine palk. Kuigi sisuliselt haakuvad masinõppe analüüsi tulemused palgaregressiooni tulemusega, st keskenduvad meeste ja naiste tunnipalga seletamisele, siis mugavuse mõttes on selle analüüsi tulemused esitatud eraldi sektsioonis.

Valitud lähenemine osutus valituks, sest näitab visuaalselt täpselt ära, kuidas antud tulemuseni jõuti.

5.2 Põhipalga ja töötundide ennustamine palgalõhe arvutamiseks

Valitud Eesti firmad peavad igal aastal osalema palgalõhe uuringus. See tähendab suurt lisatööd - blankettide täitmist. Kuna on selgunud, et firmadele valmistab see palju pahameelt, on Statistikaamet otsustanud proovida palgalõhe uuringust loobuda ning selle asemel kasutada TSD andmeid. Palgalõhe arvutatakse ilma lisatasudeta bruto kuutasu ja kuus töötatud tundide põhjal, kuid TSD andmestikus on välja toodud ainult erinevat liiki väljamaksed koos lisatasudega. Töötasu struktuuri uuringus on küll olemas töötaja brutotöötasu ilma lisatasudeta ning kuus töötatud tunnid, aga kuna see uuring ilmub iga nelja aasta tagant, ei saa seda iga aastaseks palgalõhe arvutamiseks kasutada. Käesolevas magistritöös kasutatakse TSU infot mudelite ehitamiseks (lineaarse regressiooni mudel ehitamiseks ja närvivõrkudes) ja tulemuste valideerimiseks, et hiljem saaks palgalõhe arvutada muudest registritest.

5.2.1 Põhipalga ennustamine kasutades aasta jooksul sagedamini esinenud väljamakseid

TSD andmetes esinevad iga firma töötaja kohta igakuised väljamaksed vastavalt liigile. Kuigi TSD tabelis ei ole eraldi välja toodud töötasu ilma lisatasudeta siis enamasti ei ole lisatasud iga kuu sama suured, mis andis võimaluse eemaldada kuised väljamaksed, mis ei kordunud ja eeldada, et väljamaksed, mille summad kordusid kõige tihedamini võivad osutuda põhipalgaks ilma lisatasudeta. Paljude töötajate puhul esines mitut väljamakset võrdselt - kasutas autor järgmisi lahendusi:

1. Võttes ühe isiku aasta jooksul kõige sagedamini esinevate väljamaksete keskmise
2. Võttes ühe isiku aasta jooksul kõige sagedamini esinevatest väljamaksetest suurima
3. Võttes ühe isiku aasta jooksul kõige sagedamini esinevatest väljamaksetest väikseima
4. Võttes ühe isiku aasta kõikide väljamaksete keskmise

Tulemuste hindamine osutus raskeks, sest TSD pealt ei ole võimalik saada töötatud tunde kindla kuu kohta.

Tulemuste täpsust hindas autor kasutades tabelit Palk ja tööjõud, kus on välja toodud firmas töötavate inimeste arv ning töötajate põhipalga summa. Autor liitis kokku kõikide firma töötajate TSD põhjal ennustatud põhipalgad ning võrdles sama firma originaalse töötajate põhipalga summaga (võrdlusesse läksid ainult firmad, mille kõikide töötajate kohta oli võimalik arvutada ennustatav põhipalk)

Autor arvutas ka palgalõhe eelnevalt ennustatud kuu põhipalkade pealt.

Palgalõhe valem:

$$palgalõhe = \frac{\left(\frac{sum(meeste\ kuupalk)}{sum(meeste\ töötatud\ tunnid)}\right) - \left(\frac{sum(naiste\ kuupalk)}{sum(naiste\ töötatud\ tunnid)}\right)}{\left(\frac{sum(meeste\ kuupalk)}{sum(meeste\ töötatud\ tunnid)}\right)} \quad (6)$$

Et arvutada, kui hästi antud lähenemised põhipalka ennustasid, kasutati TSU tabelist töötatud tunde palgalõhe arvutamiseks ning võrreldi reaalsete põhipalkadega arvutatud palgalõhega.

.

5.2.2 Põhipalga ja töötundide ennustamine kasutades lineaarse regressiooni mudeli ehitamist

Lineaarse regressiooni mudeli ehitamise töötamismehhanismi seletas autor peatükis 4.1.1. Põhipalga ennustamisel oli sõltuv muutuja brutopalk ilma lisatasudeta ning tundide ennustamisel oli sõltuv muutuja töötunnid. Sõltumatute muutujate saamiseks pani autor kokku infot kolmest erinevast tabelist:

1. 2018. aasta Töötasu struktuuri uuring:
    - koormus;
    - EMTAK kood;
    - amet;
    - maakond.

Kuna Töötasu struktuuri uuring toimub iga nelja aasta tagant, aga palgalõhe on vaja arvutada iga aasta, on tabelist valitud tunnused hiljem kätte saadavad ka mujalt registritest, et neid iga aastaselt kasutada saaks.

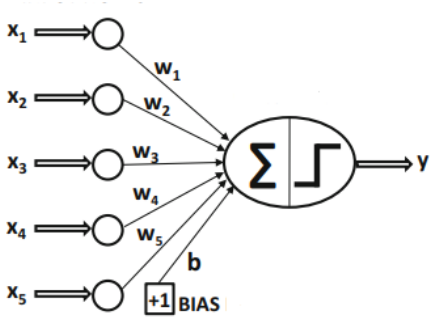
2. Palk ja tööjõud:
    - ettevõtte töötajate arv oktoobrikuu;
    - ettevõtte töötajate brutotasu summa ilma lisatasudeta;
    - ettevõtte töötajate töötatud tundide summa.
3. TSD
    - sünnikuupäevast arvutatud vanus;
    - sugu;
    - ennustatud bruto töötasu kasutades töötaja aasta jooksul kõige sagedamini esinevate palkade keskmist;
    - viie viimase tasustatud kuude TSD liikide (10, 11, 12) summad;
        - eelmise punkti viie kuu summa;
    - viie viimase tasustatud kuude TSD kõikide liikide summa;
        - eelmise punkti viie kuu summa;

Parimal võimalikul juhul oleksid valitud viis kuud august, september, oktoober, november, detsember. Kui mõnel töötajal on puudu näiteks detsembrikuu väljamakse, on tema jaoks valitud kuud juuli, august, september, oktoober, november. Viis viimast tasustatud kuud on valitud asjaolu tõttu, et üldiselt arvutatakse palgalõhe oktoobrikuu andmete põhjal ja autor otsustas võtta viis kuud, mis on oktoobrile kõige lähemal.

5.2.3 Põhipalga ja töötundide ennustamine kasutades närvivõrke

Närvivõrk on algoritmide hulk, mis on tuletatud päris närvivõrkudest ning on disainitud tundma ära erinevaid mustreid. Närvivõrk tõlgendab sissetulevat infot läbi masinliku meeleorgani, sildistades või klusterdades andmeid. Mustrid, mida närvivõrk ära tunneb, on numbrilised vektorid, millesse pärismaailma info (pildid, heli, tekst jne) tuleb tõlkida. Selleks, et närvivõrk suudaks andmeid tõlgendada, on tal vaja juba eelnevalt sildistatud andmeid, mille peal treenida.

Kõige lihtsam närvivõrk on järgmine (neuron):



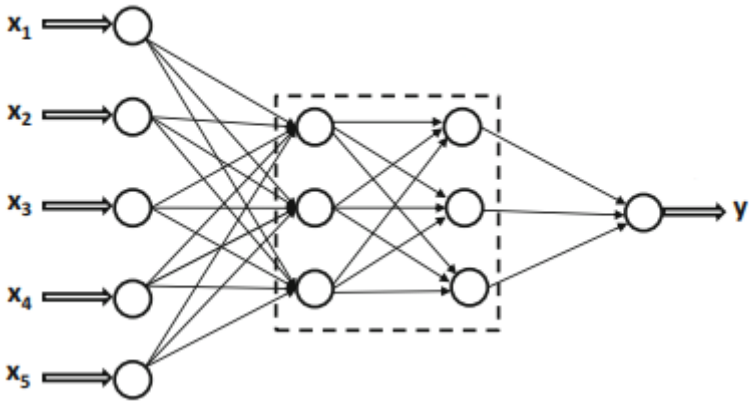
Joonis 2. Neuroni ehitus

,

kus:

- $x_1 - x_5$ on sisendid;
- $w_1 - w_5$, on neile antud kaalud;
- b, on vabaliige;
- $\sum$, on kaalude ja sisendite maatrikskorrutis ning arvutatakse $\bar{W}\ \bar{X} + b = \sum_{j=1}^{d} w_j\, x_j + b$; (7)
- $\sqcap$, on aktiveerimisfunktsioon, mis normaliseerib sisendi väärtused otsuse protsessile vajaliku kujule $\hat{y} = sign\{\bar{W}\ \bar{X} + b\} = sign\{\sum_{j=1}^{d} w_j\, x_j + b\}$; (8)
- y, mis on klassifitseerimise tulemus [1].

Praktikas kasutatakse rohkem kui ühte neuronit rohkem kui ühes kihis, moodustades sügavnärvivõrgu(DNN).



Joonis 3. Sügavnärvivõrgu ehitus

Närvivõrkude treenimine

Treenimine on iteratiivne protsess. Algselt initsialiseeritakse kaalud, siis arvutatakse nende põhjal tulemus, ning seejärel võrreldakse saadud tulemust reaalse tulemusega ning arvutatakse kaofunktsioon (9), mille põhjal uuendatakse kaalusid valitud meetodiga. Protsessi korratakse seni, kuni peatumiskriteerium on täidetud. [25]

$$(\frac{1}{2}(ennustatud - tegelik)^2) \quad (9)$$

Kaalude uuendamine kasutades *gradient descent*-i:

Gradient descent on iteratiivne optimeerimisalgoritm, leidmaks funktsiooni, näiteks närvivõrkude puhul kaofunktsiooni, miinimumi. Lokaalne miinimum leitakse sammudes kaalude veagradiendi negatiivses suunas. Kaalude uuendamise valem on:

$$* W_x = W_x - a(\frac{\delta Error}{\delta W_x}) \ , \quad (10)$$

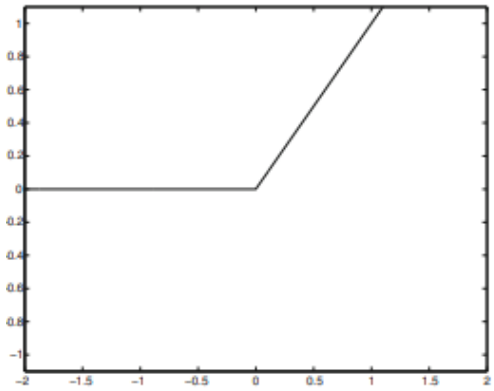
kus:

- $* W_x$, on uuendatud kaal;
- W_x, on eelmine kaal;
- a on õppimiskordaja, mille ülesanne on tagada, et uus uuendatud kaal kindlasti vähendaks kaofunktsiooni;
- $(\frac{\delta Error}{\delta W_x})$ on vea tuletis kaalu suhtes. [26]

Selles töös kasutati Sklearni teeki, kus on funktsioonid närvivõrkudel baseeruvate mudelite treenimiseks, valideerimiseks ja rakendamiseks. Põhifunktsiooniks on MLPRegressor, mille vaikimisi aktiveerimisfunktsiooniks on ReLU ja treenimisalgoritmiks on Adam. Kuna käesolevas magistritöös tuleb ennustada brutopalka ja tunde, mis on positiivsed reaalarvud, ei saa kasutada klassifitseerimispõhist närvivõrku vaid tuleb kasutada regressioonipõhist närvivõrku, mille väljundiks ei ole klassisilt vaid positiivne reaalarv.

Tänapäeval on ReLU suuresti asendanud *sigmoid* ja *tanh* funktsioonid sügavnävivõrkudes. Kasutades *sigmoid* ja *tanh* aktiveerimisfunktsiooni, muutuvad suured arvud kohe 1 ja väiksed arvud lähevad vastavalt 0 või -1, mis tähendab, et edaspidi on funktsioonid sensitiivsed ainult enda keskväärtuse juures. See põhjustab omakorda selle, et mudelil on raske kaalusi muuta ja mudelit parendada. ReLU parandab selle probleemi, samuti on ReLU-l kiirem arvutuskiirus ja paremad õpiomadused. [27]

$$\Phi(v) = max\{v, 0\} \tag{11}$$



Joonis 4. ReLU – Tulemuseks on kas 0 või nullist suurem arv

Gradient descent algoritmile järgnes *stochastic gradient descent*, mis erines selle võrra, et enam ei pea tervet treeningandmestikku läbi käima, et kaalusid uuendada, vaid uuendamist tehakse peale treeningandmete alamhulga läbi töötlemist. Selle eeliseks on, et tulemus ei pruugi enam nii lihtsalt lokaalsesse miinimumi kinni jääda ning kasutamiseks ei lähe vaja nii suurt arvutusvõimsust [1].

Adam algoritm on *Stochastic gradient descent* algoritmi laiendus, mis on arvutusvõimsuselt efektiivne, ei vaja palju mälu ja on üpriski lihtne implementeerida. Seda tõestab ka asjaolu, et Adam on paljudesse närvirõgu pakkidesse on ka sisseehitatud. Adam mõjutab gradient komponenti, kasutades eksponentsiaalselt liikuvat gradientide keskmist ning eksponentsiaalselt liikuvat ruutgradientide keskmisega jagatud õppimiskordajat. [28]

$w_{t+1} = w_t - \frac{a}{\sqrt{\widehat{v_t}+\epsilon}} * \widehat{m_t}$, kus $\widehat{m_t} = \frac{m_t}{1-\beta_1^t}$ ja $\widehat{v_t} = \frac{v_t}{1-\beta_2^t}$ ning

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1)\frac{\delta L}{\delta w_t} \qquad v_t = \beta_2 v_{t-1} + (1 - \beta_2)\left[\frac{\delta L}{\delta w_t}\right]^2 \qquad (12)$$

m ja v algväärtus on 0.

Autorite poolt soovitatud väärtused:

- a = 0.001
- β_1 = 0.9
- β_2 = 0.999
- ε = 10^{-8} [29]

Põhipalga ennustamisel oli sõltuv muutuja brutopalk ilma lisatasudeta ning tundide ennustamisel oli sõltuv muutuja töötunnid. Kasutatud sõltumatud muutujad on kirjeldatud peatükis 4.2.2

5.3 Töövahendid

Palka mõjutavate tunnuste analüüsis kasutati R programmeerimiskeelt, sest kogu töö käis Statistikaameti serveris, ning sinna oli antud hetkel laetud ainult R ja R Studio kuna see valmis mitme inimese koostöös ja nendel oli kasutuskogemus R keelega.

Palgalõhe arvutamise jaoks, laeti autori soovil, Statistikaameti serverisse Pythoni koos masinõppe ja andmekaevega seotud teekidega: Pandas, Sklearn, Numpy ja Scipy.

R ja R Studio

R Studio on arenduskeskkond programmeerimiskeele R jaoks, mis on programmeerimiskeel mõeldud statistiliste arvutuste jaoks. R keel on laialdaselt kasutusel statistikute ja andmekaevega tegelevate inimeste seas.

Python

Python on *de facto* masinõppe programmeerimiskeel. Seda tänu enda lihtsusele ja loetavusele, mis laseb kasutajatel fokuseerida algoritmidele ja tulemustele, selle asemel, et kulutada aega koodi efektiivsele struktureerimisele [30].

Pandas

Pandas on kiire, paindlik ja lihtsalt kasutatav, andmete analüüsiks ning manipuleerimiseks kasutatav avatud lähtekoodiga vahend. See pakub andmestruktuure ning operatsioone, et mugavalt tabuleeritud andmeid analüüsida.

Sklearn

Sklearn (Scikit-learn) on masinõppe teek, mis koosneb erinevatest klassifitseerimise, klusterdamise ning regressiooni algoritmidest. Näiteks tugivektormasinad, otsustusmetsad ja DBSCAN. Sklearn on loodud koostööd tegema Numpy-ga ja Scipy-ga.

Numpy

Numpy on Python-i teek, mis lisab võimekust manipuleerida suuri, mitme-dimensionaalseid vektoreid ja maatrikseid. Lisaks on Numpy-l ka tugi suurele hulgale kõrgetasemelistele matemaatilistele funktsioonidele.

Scipy

Scipy on tasuta ja avatud lähtekoodiga teekide kogum, mida kasutatakse teaduslikeks ja tehnilisteks arvutusteks. Scipy ühendab omavahel mitmeid teeke nagu Numpy, Pandas, Matplotlib ja SymPy.

6. Tulemused

6.1 Palka mõjutavate tunnuste analüüs

6.1.1 Lineaarse regressiooni mudeli ehitamine

Esimese, s.o lineaarse regressiooni mudeli ehitamise analüüsi eesmärk on leida, millised tunnused kirjeldavad palga kujunemist nii soolises lõikes kui ka tegevusalati (mõõdetud EMTAK koodi kaudu). Kõige paremini mudelit kirjeldavad tunnused leiti, kasutades *forward selection* meetodit, kus iga iteratsiooniga lisatakse tunnus, mis parandab siinset mudelit kõige paremini.

6.1.1.1 Sugu

Nagu tabelis Tabel 1 ja joonistel Joonis 14 ja Joonis 15 lisas on näha, mõjutavad naiste ja meeste palku üldiselt samad tunnused samas järjekorras. Lineaarse regressiooni mudeli ehitamisel leiti, et kõige paremini kirjeldavad naiste ja meeste palku amet (ISCO peagrupid), tegevusala (EMTAK koodiga mõõdetud), haridustase, maakond (ehk töökoha asukoht) ja töötaja vanus.

Samas on tabelist Tabel 1 on näha, et need tunnused mõjutavad (seletavad) naiste palku rohkem/paremini, sest leitud tunnuste abil tunnitasu ennustamisel on naiste palga ennustus täpsem kui meeste oma – viga eurodes vastavalt 1,539 ja 2,693 ning vea suhtarv palgaga vastavalt 0,315 ja 0,454. Teisisõnu, võttes arvesse näiteks viit põhinäitajat, suudame naiste palgas olevat variatiivsust edukamalt seletada kui meestel ehk meeste palgas olevat variatiivsust on seletada raskem.

Tabel 1. Regressioonimudeli ehitamise tulemused: meeste ja naiste tunnipalk

	Naised	Mehed
Mõjutavad tunnused	Ametikood EMTAK kood Haridustase Maakond Vanus	Ametikood EMATK kood Haridustase Maakond Vanus
Viga eurodes	1,539	2,693
Vea suhtarv palgaga	0,315	0,454

6.1.1.2 Tegevusala

Tegevusalati (mõõdetud EMTAK peagrupi kaudu) olid palka mõjutavad tegurid samuti üsna sarnased (tabeli aluseks olevad joonised on toodud lisas). Kõige sagedamini, st sõltumata tegevusalast, kerkisid seletavate mõjuteguritena esile ametigrupp, haridustase ja ettevõtte asukoht (maakond), mis mõjutasid kõikide tegevusalade puhul selle sisemist tunnitasu variatiivsust, kuigi mitte samas järjekorras. Amet oli n-ö selgitusjõult esimesel kohal pea kõigis tegevusaladel, v.a A – põllumajandus, metsandus ja kalapüük, kus esimesel kohal oli ettevõtte tegutsemispiirkond; ning Q – tervishoid ja sotsiaalhoolekanne, kus esimesel kohal oli haridus. Valdavalt teisel või kolmandal kohal oli ettevõtte asukoht (maakond) ning haridus, mis loetakse peamiseks inimkapitali tunnuseks, oli üldjuhul kolmandal või neljandal kohal. Haridus osutus olulisemaks kui näiteks piirkond sellistes valdkondandes nagu H – veondus/laondus, I – info ja side, K – finants- ja kindlustustegevus, N – haldus- ja abitegevused, P – haridus ning G – tervishoid ja sotsiaalhoolekanne. Nende tulemuste puhul olulisim on ehk tõsiasi, et kolmest kõige enam seletusvõimet omavast tegurist kaks (ametigrupp, asukoht) on pigem struktuursed kui individuaalsed tegurid, mis annab omakorda alust tugevale tööturu segregatsiooni hüpoteesile. Lisaks esinesid tegevusalati enim seletusvõimet omanud tunnustena sagedamini veel sugu (13 tegevusala puhul 19-st), ettevõtte omandivorm (5 juhul 19-st) ja töötaja vanus (5 juhul 19-st). Tunnused, mis ei mõjutanud oluliselt ühegi tegevusala tunnipalga variatiivust, olid töötatud tunnid, tööturuga seotus (tasustatud päevade arv), emakeel, rahvus ja puue.

Regressioonanalüüsi tulemusi kokku võttes võib öelda, et kesksed ennustajad on nii meeste kui ka naiste palga puhul pigem struktuursed tegurid, s.o amet ja tegevusala ning ka ettevõtte asukoht. Individuaalsetest, s.o inimkapitali teguritest mängis kõige tugevamat rolli ainult

haridus. Mõneti üllatuslikult oli viiendaks enim seletavaks teguriks vanus, mis viitab Eesti kontekstis ilmselt mitte niivõrd karjäärimudelile, vaid tugevale vanusegradiendile karjääri tipphetkel. Ka eri tegevusalade lõikes tegurite mõju hinnates jäid domineerima pigem struktuursed tegurid ning sugu oli endiselt paljudes tegevusalades oluline palgaerinevuste seletaja, kui arvesse oli juba võetud nii ametit, piirkonda kui ka haridust. [8]

Tabel 2. Regressioonimudeli ehitamise tulemused tegevusalade (EMTAK) lõikes

	A Põllumajandus, metsamajandus ja kalapüük	B Mäetööstus	C Töötlev tööstus	D Elektri-energia, gaasi, auru ja konditsioneeritud õhuga varustamine	E Veevarustus; Kanalisatsioon jäätme- ja saastekäsitlus	F Ehitus	G Hulgi- ja jaekaubandus; mootorsõidukite ja mootorrataste remont	H Veondus ja laondus	I Majutus ja toitlustus	J Info ja side
Ridade arv	2278	879	22367	1334	963	5349	12550	7354	2761	4274
Mõjutavad tunnused	Maakond Amet Haridus Vanus Laste arv	Amet Maakond Sugu Haridus Vanus	Amet Sugu Maakond Haridus Vanus	Amet Maakond Leping Sektor Haridus	Amet Maakond Haridus Sugu Vanus	Amet Maakond Haridus Sugu Koormus	Amet Sugu Maakond Haridus Puhkuse-päevad	Amet Sugu Haridus Maakond Leping	Amet Maakond Sugu Haridus Sektor	Amet Sugu Haridus Maakond Laste arv
Viga eurodes	1,739	1,988	1,870	2,200	1,530	1,169	2,180	2, 794	1,395	3,310
Vea suhtarv palgaga	0,394	0,299	0,354	0,306	0,308	0,415	0,440	0,461	0,332	0,466

Tabel 3. Regressioonimudeli ehitamise tulemused, tegevusalade (EMTAK) lõikes jätkub

	K Finants- ja kindlustustegevus	L Kinnisvaraalane tegevus	M Kutse- teadus- ja tehnikaalane tegevus	N Haldus - ja abitegevused	O Avalik haldus ja riigikaitse; Kohustuslik sotsiaalkindlustus	P Haridus	Q Tervishoid ja sotsiaalhoolekanne	R Kunst, meelelahutus ja vaba aeg	S Muud teenindavad tegevused
Ridade arv	2719	1701	2829	4050	7286	19539	8336	3362	1125
Mõjutavad tunnused	Amet Sugu Haridus Maakond Vanus	Amet Firma suurus Maakond Haridus Sugu	Amet Firma suurus Maakond Sektor Haridus	Amet Haridus Maakond Tööstaaž Sugu	Ameti Maakond Haridus Sugu Leping	Amet Haridus Maakond Sugu Puhkuse-päevad	Haridus Amet Sugu Sektor Maakond	Amet Maakond Sektor Sugu Haridus	Amet Maakond Firma suurus Haridus Sektor
Viga eurodes	3,653	2,033	2,741	2,292	1,793	1,155	1,673	1,395	1,761
Vea suhtarv palgaga	0,400	0,533	0,469	0,454	0,258	0,230	0,286	0,322	0,413

[1]

[1] Meenutuseks – EMTAK peagrupid on järgmised: A – põllumajandus, metsamajandus ja kalapüük; B – mäetööstus; C – töötlev tööstus; D – elektrienergia, gaasi, auru ja konditsioneeritud õhuga varustamine; E – veevarustus; kanalisatsioon, jäätme- ja saastekäitlus; F – ehitus; G – hulgi- ja jaekaubandus, mootorsõidukite ja mootorrataste remont; H – veondus ja laondus; I – majutus ja toitlustus; J – info ja side; K – finants- ja kindlustustegevus; L – kinnisvaraalane tegevus; M – kutse-, teadus- ja tehnikaalane tegevus; N – haldus- ja abitegevused; O – avalik haldus ja riigikaitse, kohustuslik sotsiaalkindlustus; P – haridus; Q – tervishoid ja sotsiaalhoolekanne; R – kunst, meelelahutus ja vaba aeg; S – muud teenindavad tegevused; T – kodumajapidamiste kui tööandjate tegevus; kodumajapidamiste oma tarbeks mõeldud eristamata kaupade tootmine ja teenuste osutamine; U – eksterritoriaalsete organisatsioonide ja üksuste tegevus

6.1.2 Otsustuspuu

Nagu öeldud, siis otsustuspuu analüüsi eesmärk oli ameti peagruppide lõikes välja selgitada, kui hästi on võimalik ennustada, millisesse klassi vastav andmepunkt kuulub. Eristati järgmiseid klasse: a) kõrgepalgaline mees; b) kõrgepalgaline naine; c) madalapalgaline mees; d) madalapalgaline naine. Klasside leidmiseks arvutati iga ameti peagrupi (ISCO) sees nii naiste kui ka meeste keskmine palk ning selle põhjal jagati mehed ja naised vastavalt kõrge- ja madalapalgalisteks. Mudeli eeldus on, et mida parem on ennustusvõime, seda diskrimineerivam on palga maksmine vastavas ametigrupis. Otsustuspuude tulemuse koondtabelist näeme (vt Tabel 4), et igas ametigrupis teenivad mehed suuremat tunnipalka kui naised, kuid sellegipoolest on otsustuspuu ennustusvõime üpriski madal, s.o vahemikus 40–60%, ning see ei parane ka gruppides, kus on rohkem ridu. Otsustuspuude ennustusvõime täpsus vahemikus 40–60% tähendab, et ka kõige paremini tunnipalka kirjeldavate tunnustega ei suuda mudel eriti täpselt ennustada, millisesse klassi andmepunkt kuulub, mis võib tähendada, et keskmine tunnipalga suurus ametigrupis ei ole mudeli jaoks piisavalt diskrimineeriv või leidub veel mõni muutuja, mis mõjutab tunnipalga suurust, kuid mille kohta ei ole siinses uuringus andmeid. Olgu selgituseks veel öeldud, et analüüsi puhul lähtuti n-ö etteantud tunnuste hulgast (vt tunnuste kirjeldust eespool alapeatükist „Andmed ja meetodid“).

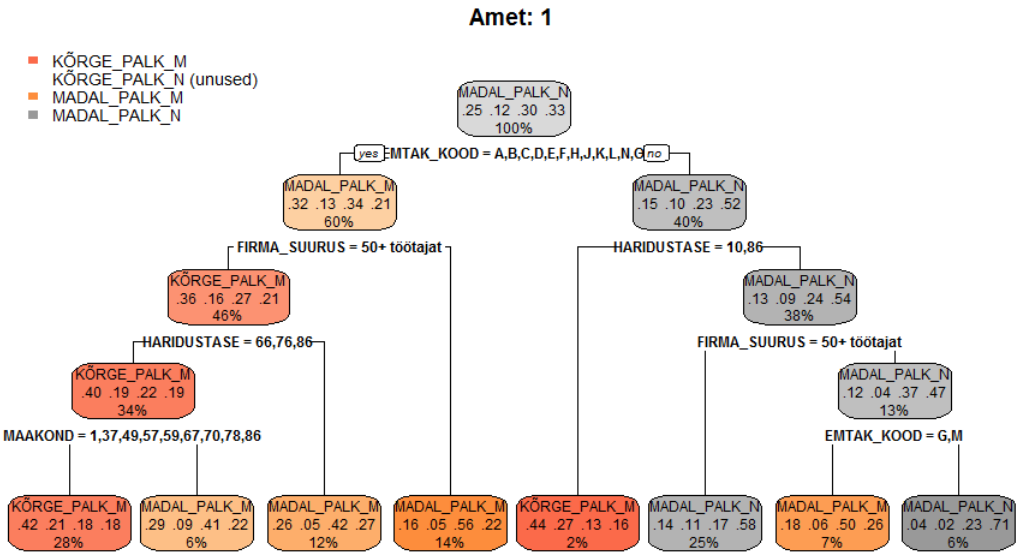
Tabel 4. Otsustuspuu tulemused ameti peagruppide lõikes

	1	2	3	4	5	6	7	8	9
Ridade arv	14274	23249	17397	7498	13290	840	12979	9922	11341
Keskmine palk	11,109	7,072	6,544	5,110	3,670	4,665	5,420	5,065	3,410
Meestel	12,612	8,816	7,859	5,930	4,472	4,877	5,891	5,490	4,219
Naistel	9,295	6,467	5,601	4,820	3,458	4,499	3,861	4,084	2,970
Täpsus	0,500	0,594	0,493	0,468	0,615	0,502	0,534	0,568	0,587

Otsustuspuu lugemist tuleks alustada ülalt. Iga puulehe pealkiri näitab, millisesse klassi satub punkt kõige tõenäolisemalt. Puulehe pealkirja all olevad neli numbrit märgistavad andmepunkti igasse klassi sattumise tõenäosust, teades, kuidas on eelnevad “testid” vastatud. Lehe kõige alumine number näitab, mitu protsenti kõikidest andmepunktidest on käesolevast puulehest läbi käinud.

6.1.2.1 Juhid

Juhtide otsustuspuu põhjal toimub n-ö puu esimese haru jagunemine tegevusala põhjal (andmepunkt kuulub EMTAK kategooriasse A, B, C, D, E, F, H, J, K, L, N või O). Sellest edasi, kui palgasaaja töötab firmas, kus on alla 50 töötaja, kuulub andmepunkt kõige tõenäolisemalt madala palgaga meeste kategooriasse. Otsustuspuu ei paiguta aga ühtegi andmepunkti kõrgepalgaliste naiste kategooriasse, sest ühelt poolt on neid ilmselt võrdlemisi vähem ja teiselt poolt on ilmselt „turvalisem“ diferentseerida kõrgepalgalisi ja madalapalgalisi mehi ning nende kõrval madalapalgalisi naisi, sest sellisel juhul on puu ennustamise täpsus statistiliselt kõige suurem. Ka haridustase kaldub pigem diferentseerima kõrgepalgalisi ja madalapalgalisi mehi, st hariduse mõju determineerib selgemini meeste palgakategooriat. Kui juhipositsioonil palgasaaja ettevõtte suurus on üle 50 töötaja, haridustase on bakalaureusekraad või sellega võrdsustatud haridus ning ta elab Tallinnas, kuulub see palgasaaja kõige tõenäolisemalt kõrge palgaga meeste kategooriasse. Kui haridustase on aga madal (põhiharidus) ja firma suurus üle 50 töötaja, kuulub see juht kõige tõenäolisemalt madalapalgaliste naiste kategooriasse.

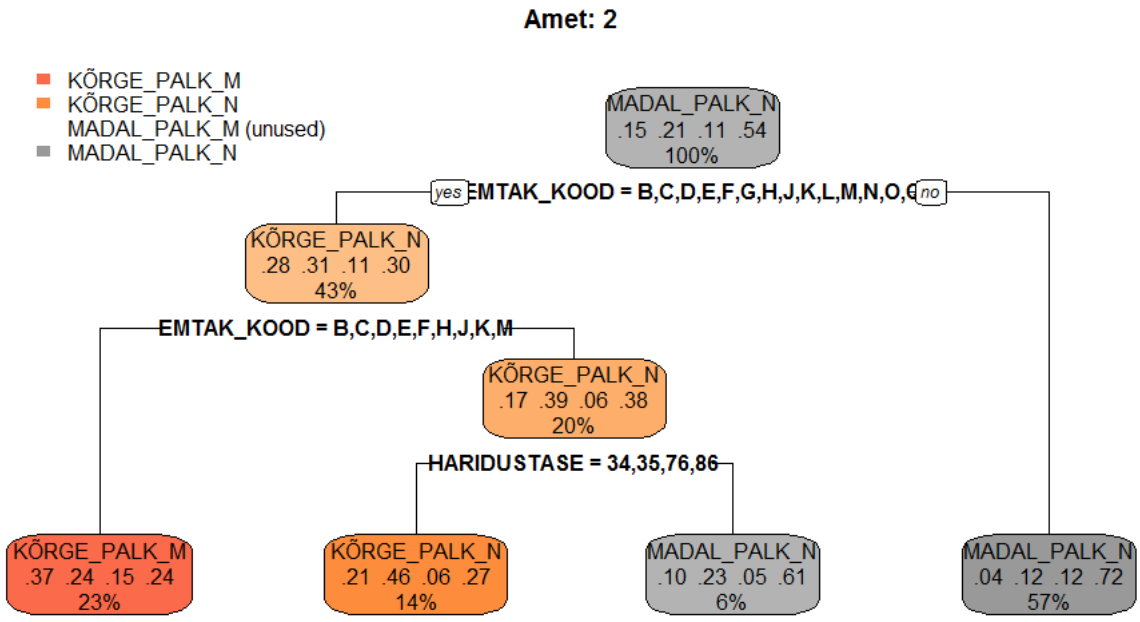


Joonis 5. Juhtide ametigrupi otsustuspuu

6.1.2.2 Tippspetsialistid

Tippspetsialistide otsustuspuu joonise järgi on taas esimene „hargnemine“ tegevusala põhine – kui andmepunkt ei kuulu EMTAK kategooriasse B, C, D, E, F, G, H, J, K, L, M, N, O või Q,

kuulub andmepunkt kõige tõenäolisemalt madalapalgaliste naiste gruppi. Kui tippspetsialist kuulub aga EMTAK gruppi C (töötlev tööstus), kuulub ta kõige tõenäolisemalt kõrge palgaga meeste kategooriasse. Kui tippspetsialist kuulub EMTAK gruppi A (põllumajandus, metsandus, kalapüük), kuulub ta kõige tõenäolisemalt madalapalgaliste naiste kategooriasse. Erinevalt juhtide otsustuspuust on siin osutunud kasutamata kategooriaks madala palgaga mehed ehk siis diferentseeritavateks või kõige lihtsamini ennustatavateks gruppideks on kõrge palgaga mehed ning kõrge või madala palgaga naised. See kõlab mõneti loogilisena ka varasemate uurimistulemuste valguses, kus naistel on üldiselt vähem asja tippjuhtide hulka ning nende karjääri laeks jääb pigem tippspetsialist. Isegi viimase hulgas on vähemalt tunnipalga mõttes meestel endiselt selge positsioon kõrgepalgaliste kategoorias (ehk madalpalgalised mehed ei tule selles ametigrupis üldse esile). Tegevusalade kõrval tuleb arvesse ka haridustase, mis aga diferentseerib naiste kõrgema või madalama palga kategooria tõenäosusi.

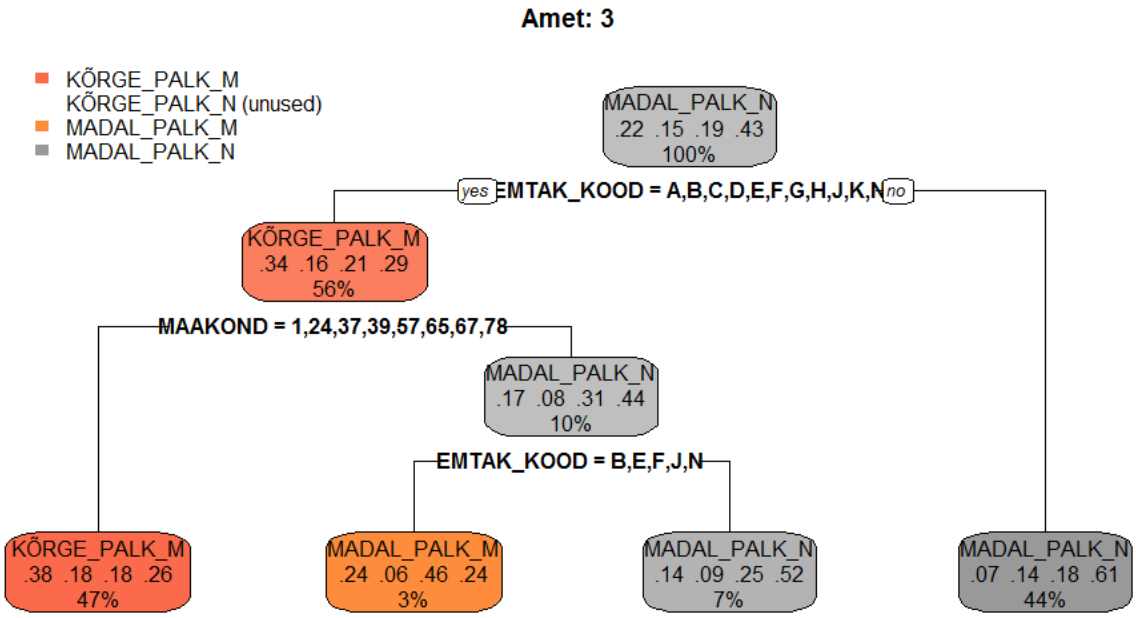


Joonis 6. Tippspetsialistide ametigrupi otsustuspuu

6.1.2.3 Tehnikud ja keskastme spetsialistid

Kolmanda ametigrupi – tehnikud ja keskastme spetsialistid – otsustuspuu järgi, kui andmepunkt kuulub EMTAK kategooriasse A, B, C, D, E, F, G, H, J, K või N, kuulub andmepunkt kõige tõenäolisemalt kõrge palgaga meeste kategooriasse. Kui aga andmepunkt

kuulub ülejäänutesse EMTAK kategooriatesse, kuulub andmepunkt kõige tõenäolisemalt madala palgaga naiste kategooriasse. Otsustuspuu ei paiguta ühtegi andmepunkti kõrgepalgaliste naiste kategooriasse, sest neid andmepunkte on taas võrdlemisi vähem ning puu paigutas need punktid enam-vähem võrdselt madalapalgaliste meeste kategooriasse või kõrgepalgaliste meeste kategooriasse. Kõrgepalgalistel naistel, kes paigutati kõrgepalgaliste meeste kategooriasse, olid paljude tunnuste väärtused suhteliselt sarnased kõrgepalgaliste meeste väärtustega ning vastupidi. Kui tehnik või keskastmespetsialist kuulub EMTAK gruppi A (põllumajandus, metsandus ja kalapüük) ja elab Tartus, siis kuulub ta kõige tõenäolisemalt kõrgepalgaliste meeste kategooriasse.



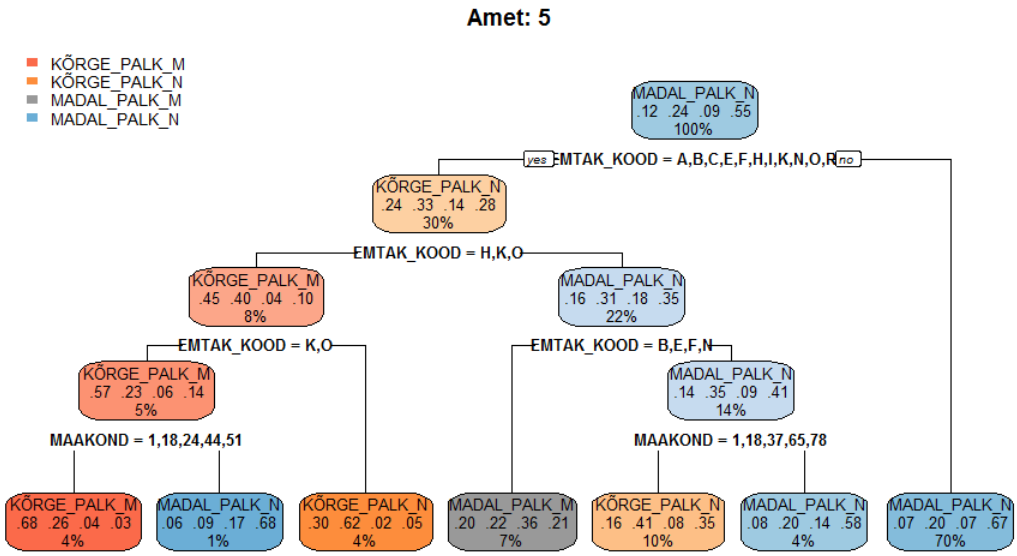
Joonis 7. Tehnikute ja keskastme spetsialistide ametigrupi otsustuspuu

6.1.2.4 Kontoritöötajad ja klienditeenindajad

Kontoritöötajate ja klienditeenindajate ametigrupi puhul arvutas algoritm ainult ühe otsusepuulehe, milleks on madalapalgalised naised. Ükski neist tunnustest ei eraldanud palgagruppe piisava mõjuga ja kuna nimetatud ametigrupis kuulub kõige suurem hulk andmepunkte madalapalgaliste naiste gruppi, ei olnud võimalik otsusepuul märkimisväärselt eristust luua.

6.1.2.5 Teenindus- ja müügitöötajad

Viienda ametite peagrupi, s.o teenindus- ja müügitöötajate põhjal ilmnes, et kui andmepunkt on EMTAK kategoorias kas K või O ning töötaja elab mõnes suurlinnas (Tallinnas, Pärnus, Tartus) või maakondadest kas Ida-Virumaal või Järvamaal, siis kuulub andmepunkt kõige tõenäolisemalt kõrge palgaga meeste kategooriasse. Kui teenindustöötaja kuulub EMTAK gruppi A (põllumajandus, metsamajandus, kalapüük) ning elab Tallinnas, kuulub ta kõrgepalgaliste naiste kategooriasse. Kui teenindustöötaja kuulub EMTAK gruppi K (finants- ja kindlustustegevus) ja elab Tallinnas, kuulub ta kõige tõenäolisemalt kõrgepalgaliste meeste kategooriasse. Teisisõnu, taas hakkab tugevalt mängima ettevõtte tegevusala, kuid selle kõrval tugevalt ka piirkond – andmepunkti võimalus sattuda kõrgepalgaliste naiste kategooriasse on pigem võimalik vaid Tallinnas, samas kui meeste puhul ka teistes piirkondades.



Joonis 8. Teenindus- ja müügitöötajate ametigrupi otsustuspuu

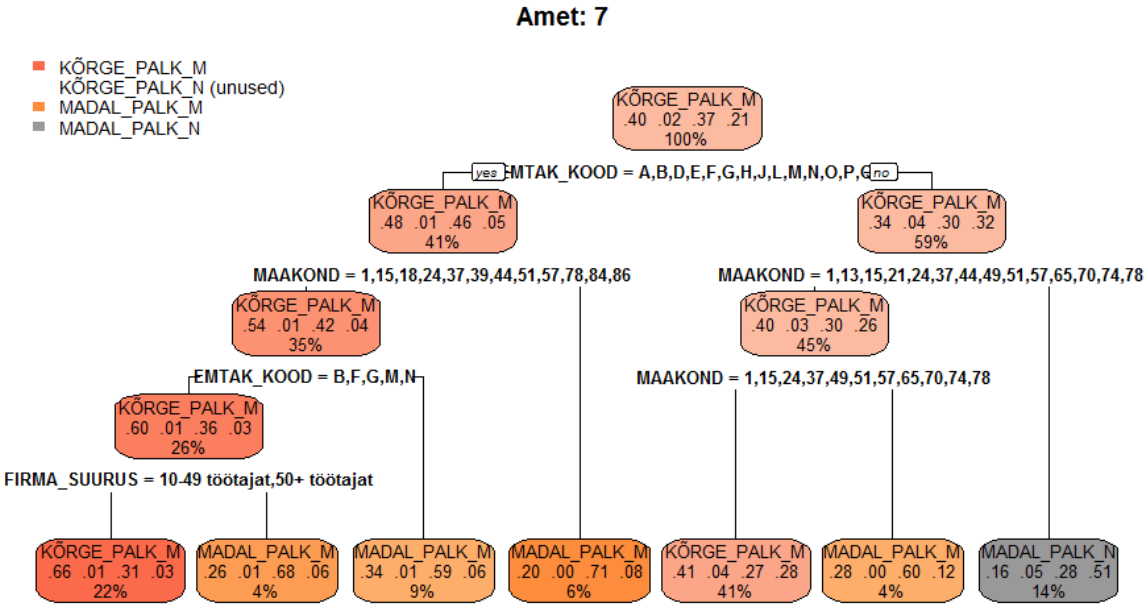
6.1.2.6 Põllumajanduse, metsanduse, kalanduse ja jahinduse oskustöölised

Kui andmepunkt elab maakonnas 1 (Tallinn), 18 (Pärnu), 37 (Harjumaa), 39 (Hiiumaa), 57 (Läänemaa), 82 (Valgamaa), 84 (Viljandimaa), firma suurus on 10 – 49 inimest ja haridustase on 25 (kutseharidus), 35 (kutseharidus põhihariduse baasil), 45 (kutseharidus keskhariduse baasil), kuulub andmepunkt kõige tõenäolisemalt kõrgepalgaliste meeste gruppi.

Kui käesolevasse ametigruppi kuuluv isik elab maakonnas 1 (Tallinn), firma suurus on 10 - 49 töötajat, haridustase on 26 (keskharidus) ja kui puhkusepäevi on rohkem kui 33, kuulub isik kõige tõenäolisemalt kõrgepalgaliste naiste kategooriasse.

6.1.2.7 Oskus- ja käsitöölised

Oskus- ja käsitööliste ametigrupi otsustuspuu järgi toimub esmane „hargnemine“ tegevusalade järgi, eristades kõrgepalgalisi mehi. Kui andmepunkt kuulub EMTAK gruppi B, F, G, M või N ja firma suurus on üle 50 inimese, kuulub andmepunkt kõige tõenäolisemalt madalapalgaliste meeste gruppi. Kui oskus- või käsitööline kuulub EMTAK gruppi A (põllumajandus, metsamajandus, kalapüük) ja elab maakonnas Tallinnas, kuulub oskus- või käsitööline kõige tõenäolisemalt samuti madalapalgaliste meeste kategooriasse. Sarnaselt mõne eelmise ametigrupiga ei paiguta otsustuspuu ka siin ühtegi andmepunkti kõrgepalgaliste naiste kategooriasse. Nagu näha, siis ettevõtte tegevuse kõrval on oluline indikaator ettevõtte asukoht ning lisaks ka ettevõtte suurus.

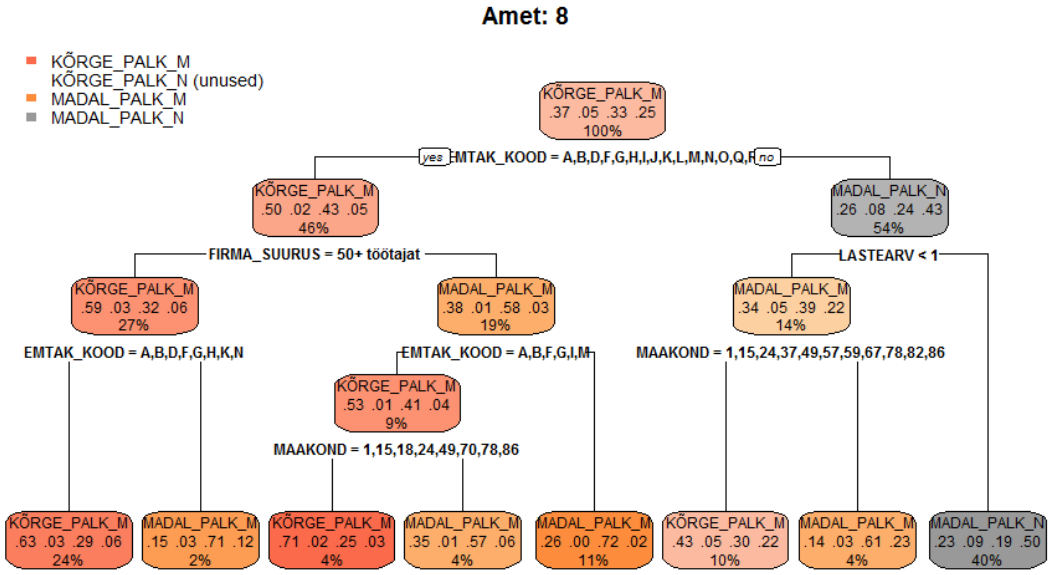


Joonis 9. Oskus ja käsitööliste ametigrupi otsustuspuu

6.1.2.8 Seadme- ja masinaoperaatorid ning koostajad

Seadme- ja masinaoperaatorite otsustuspuu järgi on taas esimene „hargnemine“ ettevõtte tegevusala põhine. Kui andmepunkt ei kuulu EMTAK gruppi A, B, D, F, G, H, I, J, K, L, M, N, O, Q või R ja andmeobjektil ei ole lapsi, kuulub andmepunkt kõige tõenäolisemalt

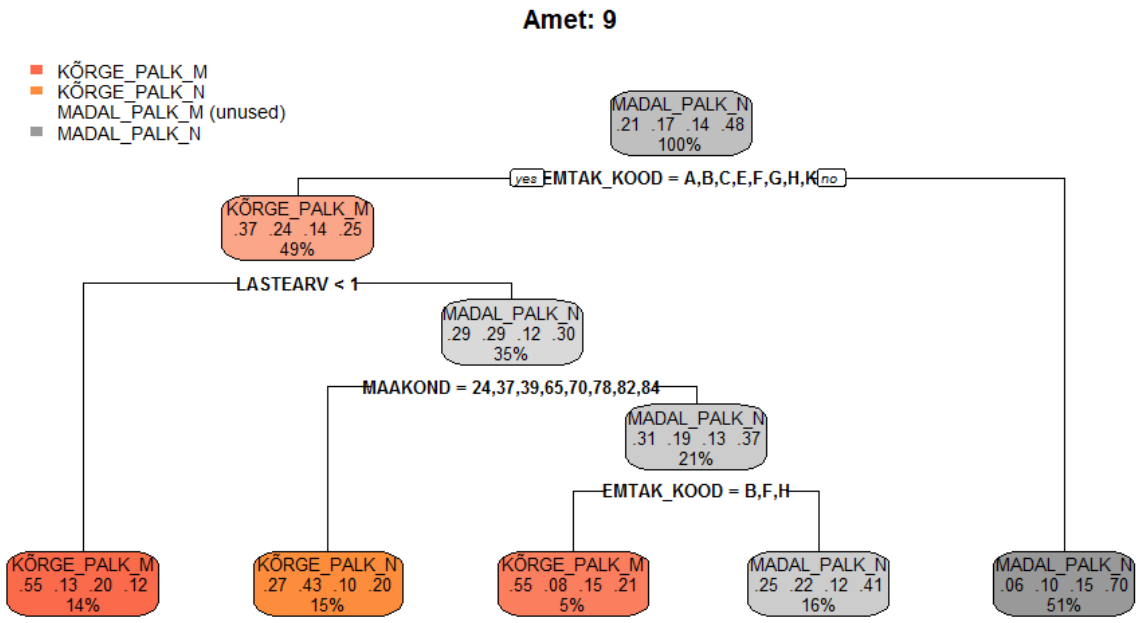
madalapalgaliste meeste gruppi. Kui tal on lapsed, siis kuulub andmepunkt kõige tõenäolisemalt madalapalgaliste naiste gruppi. Kui seadme- või masinaoperaator kuulub EMTAK gruppi A (põllumajandus, metsamajandus ja kalapüük) ja firmas on rohkem kui 50 töötajat, kuulub seadme- või masinaoperaator kõrgepalgaliste meeste kategooriasse. Kui aga seadme- või masinaoperaator kuulub EMTAK gruppi C (töötlev tööstus), ei avalda mõju mitte firma suurus, vaid laste olemasolu – kui lapsi ei ole ning andmepunki töökoha asukoht on Tallinn, kuulub seadme- või masinaoperaator kõrgepalgaliste meeste kategooriasse. Taas ei paigutata ühtegi andmepunki kõrgepalgaliste naiste kategooriasse.



Joonis 10. Seadme- ja masinaoperaatorid ning koostajad

6.1.2.9 Lihttöölised

Lihttööliste ametigrupi otsustuspuu järgi ilmneb, et kui andmepunkt kuulub EMTAK gruppi A, B, C, E, F, G, H või K ja lapsi tal ei ole, kuulub andmepunkt kõige tõenäolisemalt kõrgepalgaliste meeste gruppi, kui lapsed aga on, siis kõige tõenäolisemalt madalapalgaliste naiste gruppi. Kui lihttööline kuulub EMTAK gruppi A (põllumajandus, metsamajandus ja kalapüük) ning tal lapsi ei ole, siis kuulub ta kõige tõenäolisemalt kõrgepalgaliste meeste kategooriasse.



Joonis 11. Lihttööliste ametigrupi otsustuspuu

Kokkuvõtvalt võib otsustuspuude analüüsi puhul välja tuua, et sarnaselt eelneva(te) analüüsi(de)ga on ka sellisel viisil meeste ja naiste palku iseloomustades n-ö domineerival positsioonil struktuursed faktorid, s.o ettevõtte tegevusala. Kuigi ettevõtte tegevusala EMTAK peakoodide tasandil võib olla üsna heterogeenne tunnuse mõttes (selle all on hulk väiksemaid kategooriaid), siis hoolimata sellest viitab selle süstemaatiline esilekerkimine tööturu segregatsiooni olulisusele meeste ja naiste palga seletamisel. Huvitava tähelepanekuna võib välja tuua veel selle, et mitmes peagrupis, nagu näiteks juhid, tehnikud ja keskastme spetsialistid ning oskus- ja käsitöölised, ei ilmnenud (st ei paigutatud sellesse andmepunkte) kõrgepalgaliste naiste kategooriat. Samas ei paigutatud ühtegi andmepunkti madalapalgaliste meeste kategooriasse tippjuhtide puhul. Niinimetatud sinikraede ametigruppides (masinaoperaatorid, lihttöölised) ilmnes olulise eristajana laste olemasolu, kusjuures laste olemasolu suurendas mõlemal juhul andmepunkti paigutamist madalapalgaliste naiste gruppi, samas kui nende puudumine kõrgepalgaliste meeste gruppi. Ilmselt peaks siin otsima ka seoseid hariduse ja vanusega, kuigi selles analüüsis need aspektid järgmiste sammudena esile ei kerkinud. Nende sisuliste tulemuste kõrval ei ole aga väheoluline tõsiasi, et otsustuspuude täpsus jäi siin pigem keskpäraseks või isegi madalaks, s.o 40–60% vahele, mis tähendab, et uuringus kasutatud muutujate põhjal ei ole väga täpselt võimalik ennustada, millisesse gruppi

mingi tunnipalk kuulub. See võib tähendada, et palgaerinevusi on olemasolevate tunnustega siiski pigem raske (ära) seletada. Teiselt poolt võib see aga tähendada, et olenemata kõigist muudest struktuursetest või individuaalsetest tunnustest on tööandjatel ilmselt kalduvus meestele rohkem maksta. Analüüsist ei järeldunud ka, et grupid, kus oli rohkem andmepunkte, oleksid ennustanud väiksema veaga tunnitasu, millest võib järeldada, et gruppe, kus tunnitasu ennustati suurema veaga, mõjutavad leitud tunnused vähem. [8]

6.2 Põhipalga ja töötundide ennustamine palgalõhe arvutamiseks

Palgalõhe arvutamiseks on vaja teada nii töötajate kuus töötatud tunde kui ka kuus teenitud põhipalka ilma lisatasudeta. Hetkel peavad tööandjad täitma iga aasta palgalõhe uuringu küsitluse, mille sisendite hulgas on ka töötatud tunnid ja põhipalk. Iga nelja aasta tagant ilmub TSU (sel aastal palgalõhe uuringut ei toimu), mille peal on ka palgalõhe arvutamiseks vajaminevad tunnused.

Kuna firmade jaoks on väga koormav iga-aastast palgalõhe uuringut täita, soovib Statistikaamet selle kaotada ning arvutada palgalõhet teiste registrite andmete pealt.

Kuna TSU ilmub iga nelja aasta tagant, ei ole võimalik ka sealt iga-aastast palgalõhet arutada. Käesolevas magistritöös kasutatakse TSU-st pärit kuupalka ilma lisatasudeta ning kuus töötatud tunde *ground truth*-ina, et luua mudeleid, mis ennustaksid neid samu tunnuseid võimalikult täpselt, et saaks arvutada võimalikult täpse palgalõhe.

6.2.1 Põhipalga ennustamine kasutades aasta jooksul sagedamini esinenud väljamakseid

Esimese lähenemise juures oli võimalik ennustada ainult töötaja põhipalka, sest tegemist ei ole masinõppe meetodiga ning tundide ennustamiseks ei olnud piisavalt infot.

Palga ennustamiseks kasutati TSD andmeid. Kuigi TSD tabelis ei ole eraldi välja toodud töötasu ilma lisatasudeta siis enamasti ei ole lisatasud iga kuu sama suured, mis andis võimaluse eemaldada kuised väljamaksed, mis ei kordunud ja eeldada, et väljamaksed, mille

summad kordusid kõige tihedamini võivad osutuda põhipalgaks ilma lisatasudeta. Paljude töötajate puhul esines mitut väljamakset võrdselt - kasutas autor järgmisi lahendusi:

1. Võttes ühe isiku aasta jooksul kõige sagedamini esinevate väljamaksete keskmise;
2. Võttes ühe isiku aasta jooksul kõige sagedamini esinevatest väljamaksetest suurima;
3. Võttes ühe isiku aasta jooksul kõige sagedamini esinevatest väljamaksetest väikseima;
4. Võttes ühe isiku aasta kõikide väljamaksete keskmise.

Tulemuste hindamine osutus raskeks, sest palgalõhe arvutamiseks läheb vaja ka kuus töötatud tundide arvu. Selles punktis kasutas autor ennustatud palka ning TSU-st võetud tundide arvu, et palgalõhe arvutada.

Esmalt hindas autor tulemuste täpsust kasutades tabelit Palk ja tööjõud, kus on välja toodud firmas töötavate inimeste arv ning töötajate põhipalga summa. Autor liitis kokku kõikide firma töötajate TSD põhjal ennustatud põhipalgad ning võrdles sama firma originaalse töötajate põhipalga summaga (võrdlusesse läksid ainult firmad, mille kõikide töötajate kohta oli võimalik arvutada ennustatav põhipalk):

Arvutusse läks 201 ettevõtet, sest ainult nende ettevõtete kõikide töötajate kohta oli võimalik ennustada põhipalka.

Tegelik kõikide firmade töötajate palkade summa: 3674236.0

Möödapaneku protsent on arvutatud valemiga: (absoluutväärtuste summa * 100) / 3674236.

1. **Võttes ühe isiku aasta jooksul kõige sagedamini esinevate väljamaksete keskmise**

   Ennustatud kõikide firmade töötajate palkade summa: 3683358.6

   Ennustatud ja tegelike kõikide firmade töötajate palkade vahede

   - absoluutväärtuste summa : 281788.5
   - Möödapaneku protsent: 7.67%

2. **Võttes ühe isiku aasta jooksul kõige sagedamini esinevatest väljamaksetest suurima**

   Ennustatud kõikide firmade töötajate palkade summa: 4109227.3

   Ennustatud ja tegelike kõikide firmade töötajate palkade vahede

   - absoluutväärtuste summa : 522702.4

- Möödapaneku protsent: 14.23%

3. **Võttes ühe isiku aasta jooksul kõige sagedamini esinevatest väljamaksetest väikseima**

   Ennustatud kõikide firmade töötajate palkade summa: 3198970.9

   Ennustatud ja tegelike kõikide firmade töötajate palkade vahede

   - absoluutväärtuste summa : 611627
   - Möödapaneku protsent: 16.65%

4. **Võttes ühe isiku aasta kõikide väljamaksete keskmise**

   Ennustatud kõikide firmade töötajate palkade summa: 3888829.5

   Ennustatud ja tegelike kõikide firmade töötajate palkade vahede

   - absoluutväärtuste summa : 378251.2
   - Möödapaneku protsent: 10.29%

Autor arvutas ka palgalõhe eelnevalt ennustatud kuu põhipalkade pealt kasutades TSU-lt võetud kuu töötundide arvu:

2018. aasta tegelik palgalõhe 141395 rea pealt arvutatuna oli 18.7%. Käesolevas punktis arvutame ja ennustame palgalõhe 127138 andmerea pealt, sest TSD ja TSU linkimisel ning ebatäielike ridade eemaldamisel läks osa ridu kaotsi. 127138 rea pealt on tegelik palgalõhe 18.9%

Hetkese seisuga ei ole palgalõhe uuringu kaotamise järel võimalik kuus töötatud tunde kuskilt mujalt registrist võimalik saada, ent palgalõhe arvutamiseks on töötunde siiski vaja. Autor arvutas ka palgalõhe kasutades töötundide asemel neljakordset koormuse korrutist. Kasutades 127138 andmerida, saadi koormust kasutades palgalõheks 20.4%.

1. **Võttes ühe isiku aasta jooksul kõige sagedamini esinevate väljamaksete keskmise**
   a. Ennustatud palgalõhe töötundidega: 17.7%
   b. Ennustatud palgalõhe koormusega: 18.2%
2. **Võttes ühe isiku aasta jooksul kõige sagedamini esinevatest väljamaksetest suurima**
   a. Ennustatud palgalõhe töötundidega: 19.2%
   b. Ennustatud palgalõhe koormusega: 19.8%

3. **Võttes ühe isiku aasta jooksul kõige sagedamini esinevatest väljamaksetest väikseima**
    a. Ennustatud palgalõhe töötundidega: 15.4%
    b. Ennustatud palgalõhe koormusega: 16.0%
4. **Võttes ühe isiku aasta kõikide väljamaksete keskmise**
    a. Ennustatud palgalõhe töötundidega: 18.1%
    b. Ennustatud palgalõhe koormusega: 18.7%

Kokkuvõtvalt võib öelda, et kõige paremini töötas mudel, kus kasutati isiku aasta kõikidest väljamaksetest (väljamakse liigid 10, 11, 12) võetud keskmist ning koormuse neljakordset korrutist. 127138 andmerea pealt saadud tulemusest (18.9%), erines see kõigest 2% võrra ning täisandmestiku tulemusega (18.7%), oli see võrdne. Kuigi palgatäpsust ennustas paremini mudel, kus kasutati isiku aasta sagedamini esinevatest väljamaksetest (väljamakse liigid 10, 11, 12) võetud keskmist, mis ennustas 7% mööda kõikide töötajate oktoobrikuu palkade summat.

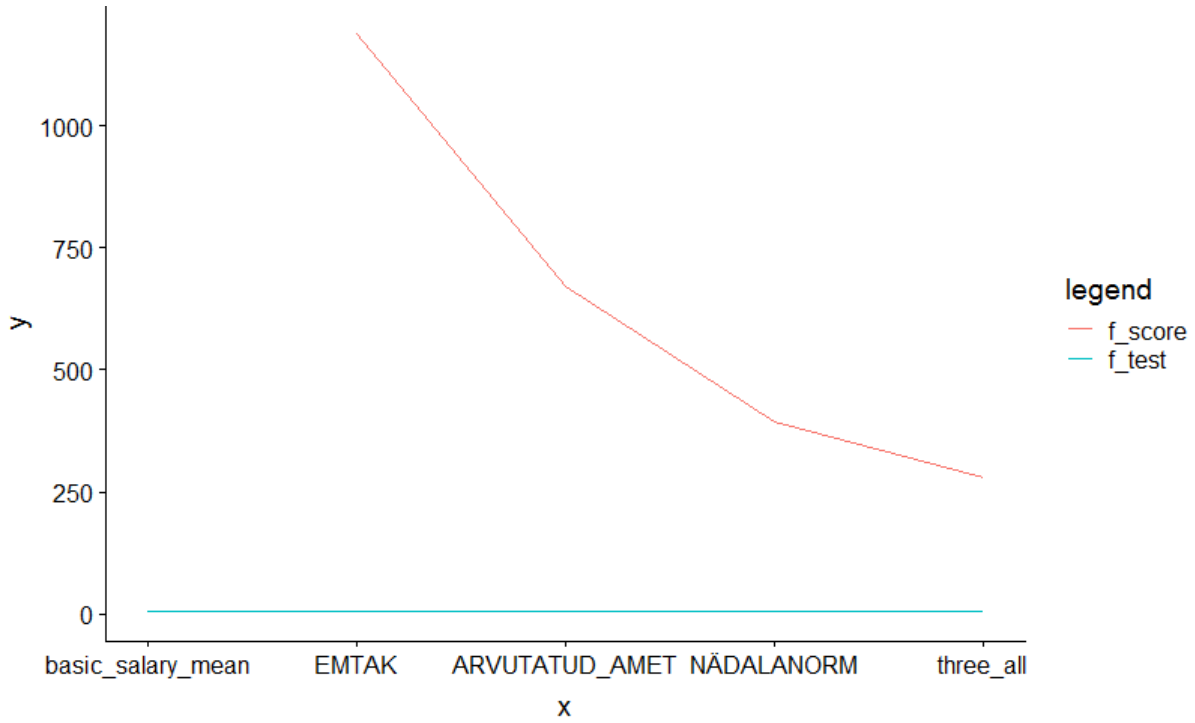
6.2.2 Põhipalga ja töötundide ennustamine kasutades lineaarse regressiooni mudeli ehitamist

Peale kolme tabeli ühendamist jäi alles 84607 rida. Kolme tabeli nimed ning saadud sõltumatud tunnused on kirjeldatud peatükis 4.2.2. Andmed jagati treening-ja testandmeteks. Arvutatud palgalõheks testandmeid kasutades saadi 22.8%. Tulemus ei ole sama, mis oli tegelik 2018. aasta palgalõhe, sest testandmetes oli palju vähem andmeridu.

- Treeningandmeid: 71915
- Testandmeid:12691

6.2.2.1 Ilma lisatasudeta põhipalga ennustamine

Sõltuv tunnus on ilma lisatasudeta oktoobrikuu palk.



Joonis 12. Regressioonimudeli ehitamise tulemus palga ennustamine

F-skoori saab mõõta alates teisest tunnusest, et võrrelda mudeli võimalikku paranemist. Mudel paraneb, kuni punane joon on sinisest kõrgemal.

Kõige paremini ennustasid kuupalka:

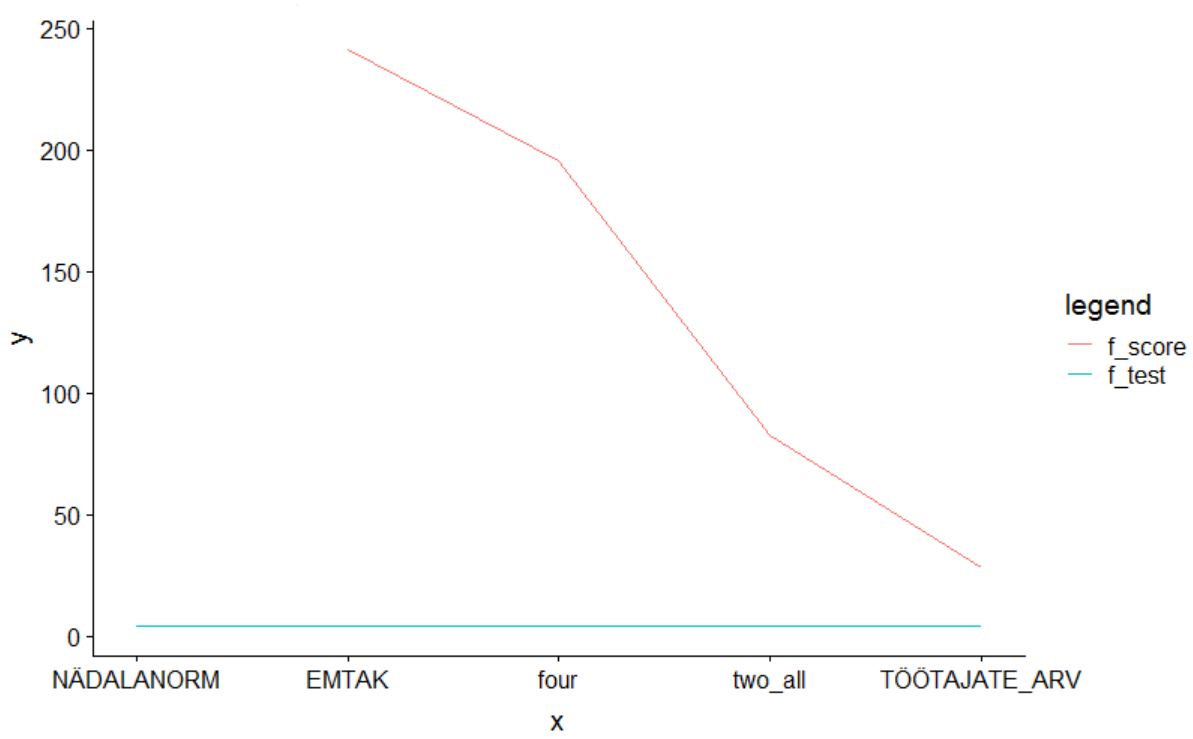
1) TSD pealt ennustatud kõige sagedamini esinevate palkade keskmine (liigid 10, 11, 12);
2) EMTAK kood;
3) Amet;
4) Koormus;
5) TSD pealt kõikide (näiteks oktoobri) väljamaksete tüüpide summa;

Viga eelnevalt nimetatud tunnuste pealt: 197 eurot.

Viga kõiki tunnuseid kasutades: 176 eurot.

6.2.2.2 Töötundide ennustamine

Sõltuv tunnus on oktoobrikuu töötunnid.



Joonis 13. Regressioonimudeli ehitamise tulemus töötundide ennustamine

F-skoori saab mõõta alates teisest tunnusest, et võrrelda mudeli võimalikku paranemist. Mudel paraneb, kuni punane joon on sinisest kõrgemal.

Kõige paremini ennustasid töötunde:

1) Koormus;
2) EMTAK kood;
3) TSD pealt tüüpide 10, 11, 12 (näiteks septembri) väljamaksete summa;
4) TSD pealt kõikide (näiteks novembri) väljamaksete tüüpide summa;
5) Töötajate arv.

Viga eelnevalt nimetatud tunnuste pealt: 18.3 tundi.

Viga kõiki tunnuseid kasutades: 18 tundi.

Autor arvutas palgalõhe kolmel erineval viisil:

a) Ennustatud palkade ja õigete tundidega – 21.6%;
b) Ennustatud tundide ja õigete palkadega - 23.6%;
c) Ennustatud tundide ja ennustatud palkadega – 22.4%.

6.2.3 Põhipalga ja töötundide ennustamine kasutades närvivõrke

Peale kolme tabeli ühendamist jäi alles 84607 rida. Kolme tabeli nimed ning saadud sõltumatud tunnused on kirjeldatud peatükis 4.2.2. Andmed jagati treening-ja testandmeteks.

Kasutatud närvivõrk oli kolme peidetud kihiga (256. 128, 128 neuronit). Treeningalgoritmiks oli *Adam*, sest dokumentatsioon soovitab kasutada *Adami*-t kui tegemist on suure andmehulgaga. Aktiiveerimisfunktsiooniks kasutati ReLU. Algseks õppimiskiiruse väärtuseks oli 0.001 ning õppimiskiiruse väärtuse muutumise algoritm oli *adaptive*, mis tähendab, et õppimiskiirus püsib algväärtuses nii kaua kuni viga läheb iga *epoch* väiksemaks. Kui kahe järjestiku *epoch*-iga ei lähe treeningviga väiksemaks või kui valideerimisskoor ei parane, jagatakse õppimiskiirus 5-ga. Kui 10 *epoch*-i ei parane treenimisviga või valideerimisskoor – *early stopping* lõpetab treeningu. [31]

Arvutatud palgalõheks testandmeid kasutades saadi 21.6%. Tulemus ei ole sama, mis lineaarse regressiooni mudeli ehitamisel, sest testandmetesse kuulusid erinevad read.

- Treeningandmeid: 71915
- Testandmeid:12691

6.2.3.1 Ilma lisatasudeta põhipalga ennustamine

Keskmine ennustatud palga viga oli 143 eurot/kuus.

6.2.3.2 töötundide ennustamine

Keskmine ennustatud tundide viga oli 17 tundi/kuus.

Autor arvutas palgalõhe kolmel erineval viisil:

1. Ennustatud palkade ja õigete tundidega – 22%;
2. Ennustatud tundide ja õigete palkadega – 21.4%;

3. Ennustatud tundide ja ennustatud palkadega – 21.7%.

Arvestades, et kasutati kahe erineva mudeli ennustusi korraga, et arvuta palgalõhe, ei osutunud tulemuste erinevus sugugi suureks. Hinnangute omadused ei pruugi olla ainult mõjutatud asjaolust, et kasutati kahte mudelit, mis põhjendaks hoopis suuremat viga, vaid ka kuna mõlemal mudelil pidid piiratud ressursside tõttu ennustamiseks kasutama kõige tugevamaid mustreid. Nendeks võivad olla stereotüübid, mille alusel on mudel ka ära õppinud diskrimineerima soo alusel palka. Sarnaselt on ka prominentsed süsteemid nagu Amazon-i töölekandideerimismootor ja Apple-i krediidiarvutusmudel uudistes olnud sarnaste diskrimineerimise mustrite kasutamisega [32] [33]. Võib eeldada, et mõlemad mudelid avastasid sama mudeli pärismaailma kirjeldamisele antud andmetest.

See kirjeldab samas ka, miks päris ja ennustatud andmete segunemisel osutusid tulemused ootustest halvemaks. Päris andmed pärinevad pärismaailma mudelist. Pärismaailma süsteemis on igal inimese andmetes individuaalsust, mis väljendub mõlema andmejupi korrelatsioonis. Seega kui segunevad individuaalse seostega andmed üks üheselt, mitte summaarselt, üldistatud andmetega, on mõistetav, et nende suhe ei vasta mõlema maailma mudeli eeldustele.

Antud mudelid on treenitud aasta 2018 andmetega ja võib arvestamata jätta aastate möödumisel toimunud muutusest nagu riigi sekkumine ja hetkel aktuaalne koroonaviirus.

7. Kokkuvõte

Palka mõjutavate tunnuste analüüsi tulemusi kokku võttes võib öelda, et kesksed „ennustajad“ on nii meeste kui ka naiste palga puhul pigem struktuursed tegurid, s.o amet ja tegevusala ning ka ettevõtte asukoht. Individuaalsetest, s.o inimkapitali teguritest mängis kõige tugevamat rolli ainult haridus. Mõneti üllatuslikult oli viiendaks enim seletavaks teguriks vanus, mis viitab Eesti kontekstis ilmselt mitte niivõrd karjäärimudelile, vaid tugevale vanusegradiendile karjääri tipphetkel. Ka eri tegevusalade lõikes tegurite mõju hinnates jäid domineerima pigem struktuursed tegurid ning sugu oli endiselt paljudes tegevusalades oluline palgaerinevuste seletaja, kui arvesse oli juba võetud nii ametit, piirkonda kui ka haridust.

Nende sisuliste tulemuste kõrval ei ole aga väheoluline tõsiasi, et otsustuspuude täpsus jäi siin pigem keskpäraseks või isegi madalaks, s.o 40–60% vahele, mis tähendab, et uuringus kasutatud muutujate põhjal ei ole väga täpselt võimalik ennustada, millisesse gruppi mingi tunnipalk kuulub. See võib tähendada, et palgaerinevusi on olemasolevate tunnustega siiski pigem raske (ära) seletada. Teiselt poolt võib see aga tähendada, et olenemata kõigist muudest struktuursetest või individuaalsetest tunnustest on tööandjatel ilmselt kalduvus meestele rohkem maksta. Analüüsist ei järeldunud ka, et grupid, kus oli rohkem andmepunkte, oleksid ennustanud väiksema veaga tunnitasu, millest võib järeldada, et gruppe, kus tunnitasu ennustati suurema veaga, mõjutavad leitud tunnused vähem.

Kuigi nii regressiooni mudeli ehitamine kui ka närvivõrkude mudel töötasid üpriski hästi, saadi ridade vähesuse tõttu õigest palgalõhes üpriski erinev tulemus. Regressioonimudeli ja närvivõrkude puhul kõikus palgalõhe erinevate testimisandmete valimeid kasutades lausa 5%. Kuigi edaspidi kasutataks juba olemasolevaid treenituid mudeleid, ei ole vaja andmeid test- ja treeningandmeteks teha, vaid saab arvutada kõikide olemasolevate andmeridade peal, jääb alles tunduvalt rohkem andmeridu, kuid praeguse näite põhjal, kaoks siiski erinevate andmestike ühendamisel üle 60 000 rea. Mida vähem andmeridu andmete puudumisel jäi, seda rohkem erines palgalõhe ametlikust 2018. aasta palgalõhest.

Kui oleks võimalik kõikide töötajate kohta saada kõik vajaminevad sisendandmed, ei pruugiks saadud mudelid siiski õigesti valitsevat palgalõhe ennustada. Kuna mudelid on õppinud leidma aasta 2018 palgalõhega andmetest seoseid, siis võib see ka järgnevatel aastatel ennustada

naistele väiksemat tasu kui meestele, arvestamata aastate möödumisel toimunud ühiskondlikest muutusest nagu näiteks riigi probleemi-korrigeeriva sekkumise tõttu.

Põhipalga ennustamine kasutades aasta jooksul sagedamini esinenud väljamakseid töötas väga hästi: kasutades ennustatud palka ning koormust, saadi 127138 rea puhul täpselt sama tulemus, mis oli tegelik palgalõhe arvutatud 141395 reaga – 18.7%. 127138 reaga arvutatud tegeliku palgalõhega, oli erinevus kõigest 0.2%.

Statistikaamet soovib, loodud mudelite töökindluse korral, loobuda palgalõhe uuringust juba 2020 oktoobris. Sellegi poolest, ei julgeks autor veel tuleva aasta palgalõhe ennustada loodud mudelite põhjal, sest 2020 palgalõhe ennustamist võib mõjutada ka hetkel valitsev koroonaviirus, millega mudelid ei ole arvestanud.

Kasutatud kirjandus

[1] C. C. Aggarwal, Neural Networks and Deep Learning, New York: Springer, 2018.

[2] D. Anderson, D. Sweeney, T. Williams, J. Camm ja J. Cochran, Statistics for Business & Economics, Boston: CENAGE Learning, 2017.

[3] kalid, 03 05 2020. [Võrgumaterjal]. Available: https://betterexplained.com/articles/vector-calculus-understanding-the-gradient/.

[4] S. Sharma, „Activation Functions in Neural Networks," 06 09 2017. [Võrgumaterjal]. Available: https://towardsdatascience.com/activation-functions-neural-networks-1cbd9f8d91d6.

[5] European Union, 2020. [Võrgumaterjal]. Available: https://ec.europa.eu/info/policies/justice-and-fundamental-rights/gender-equality/equal-pay/eu-action-against-pay-discrimination_en#establishingequalpayforequalwork.

[6] D. Leythienne ja P. Ronkowski, A decomposition of the unadjusted gender pay gap using Structure of Earnings Survey data, Luxembourg, 2018.

[7] Statistikaamet, 16 10 2019. [Võrgumaterjal]. Available: https://www.stat.ee/pressiteade-2019-122?highlight=palgal%C3%B5he.

[8] K. Täht, M. Unt, E. Kuldkepp, T. Roosalu, T. Lauri, M. Klesment, M. Rokicka ja S. Nõmm, „Soolise palgalõhe kirjeldamine ja seletamine," RASI, Tallinn, 2019.

[9] REGE, „REGE," [Võrgumaterjal]. Available: http://rege.tlu.ee/projekti-tutvustus/. [Kasutatud 27 07 2020].

[10] S. Anspal, H. Biin, E. Kallaste, M. Karu ja L. Kraut, „Sooline palgalõhe. Teoreetilise ja empiirilise kirjanduse ülevaade," Eesti Rakendusuuringute Keskus CENTAR, Poliitikauuringute Keskus PRAXIS, Sotsiaalministeerium, Tallinn, 2009.

[11] F. Blau ja L. Kahn, „The Gender Wage Gap:," *Journal of Economic Literature,* kd. 55, nr 3, pp. 789-865, 2017.

[12] D. Stojanka ja M. Savic, „APPLICATION OF THE MINCER EARNING FUNCTION," *Economics and Organization,* kd. 14, nr 2, pp. 155-162, 2014.

[13] L. M. P. Mariuci, Analyzing the gender pay gap in, 2017.

[14] Z. Amadxarif, M. Angeli, A. Haldane ja G. Zemaityte , „Understanding pay gaps," %1 *Staff Working Paper No. 877*, London, 2020.

[15] C. Nicodemo, „Gender Pay Gap and Quantile," *Discussion Paper No. 3978 ,* Jaanuar 2009.

[16] S. Briel ja M. Töpfer, „The Gender Pay Gap Revisited: Does Machine Learning offer New Insights?," *LASER Discussion Papers,* nr 116, 2020.

[17] M. Salmon, „Analysis of Web Scraped Job Data to Predict Relative Salaries," 17 06 2017. [Võrgumaterjal]. Available: https://medium.com/@msalmon00/analysis-of-web-scraped-job-data-to-predict-relative-salaries-c7237954184a. [Kasutatud 30 07 2020].

[18] M. Navyashree, M. K. Navyashree, M. Neetu, G. R. Pooja ja B. Arun, „Salary Prediction in It Job Market," *International Journal of Computer Sciences and Engineering ,* kd. 7, nr 15, 2019.

[19] S. Chakraborti, „A Comparative Study of Performances of Various," *International Journal of Computer Science and Information Technologies,* kd. 5, nr 2, 2014.

[20] I. Martin, A. Mariello, R. Battiti ja J. A. Hernandez, „Salary Prediction in the IT Job Market with Few High-Dimensional Samples: A Spanish Case Study," *International Journal of Computational Intelligence Systems,* kd. 11, nr 1, pp. 1192-1209, 2018.

[21] D. Freedman, Statistical Models:Theory and Practice, California: Cambridge Univercity Press, 2009.

[22] M. Maathuis, „Regression," Zurich, 2012.

[23] C. C.Aggarwal, Data Mining The Textbook, New York: Springer, 2015.

[24] „Hüpoteeside statistiline kontrollimine," 2010. [Võrgumaterjal]. Available: http://www-1.ms.ut.ee/mart/biomeetria2010/Hypoteesid.pdf?fbclid=IwAR0tKQAVnHeaXlsx7ERs6FBhqHo20tnHxB2NovUo2SdJRkoZCu_OgicuzMc.

[25] M. Nielsen, Neural Networks and Deep Learning, Determination Press, 2015.

[26] HMKCODE, „Backpropagation Step by Step," 03 11 2019. [Võrgumaterjal]. Available: https://hmkcode.com/ai/backpropagation-step-by-step/?fbclid=IwAR1N7x4hHWrOwjOOxnpwV7F10AlSQagv4EgNT3NttU8bhJnLNPGvRe08dl0.

[27] J. Brownlee, „A Gentle Introduction to the Rectified Linear Unit," 03 05 2020. [Võrgumaterjal]. Available: https://machinelearningmastery.com/rectified-linear-activation-function-for-deep-learning-neural-networks/.

[28] D. Kingma ja J. Lei Ba, „ADAM: A Method for Stochastic Optimization," 2015.

[29] R. Karim, „10 Stochastic Gradient Descent Optimisation Algorithms," 22 11 2018. [Võrgumaterjal]. Available: https://towardsdatascience.com/10-gradient-descent-optimisation-algorithms-86989510b5e9.

[30] I. Ghergu, „hackernoon," 02 05 2020. [Võrgumaterjal]. Available: https://hackernoon.com/the-programming-language-for-machine-learning-projects-r9f73ycs.

[31] scikit-learn developers, „sklearn.neural_network.MLPRegressor," scikit-learn, 2007-2019. [Võrgumaterjal]. Available: https://scikit-

learn.org/stable/modules/generated/sklearn.neural_network.MLPRegressor.html. [Kasutatud 12 5 2020].

[32] N. Vigdor, „Apple Card Investigated After Gender Discrimination Complaints,“ The New Your Times, 10 11 2019. [Võrgumaterjal]. Available: https://www.nytimes.com/2019/11/10/business/Apple-credit-card-investigation.html. [Kasutatud 12 05 2020].

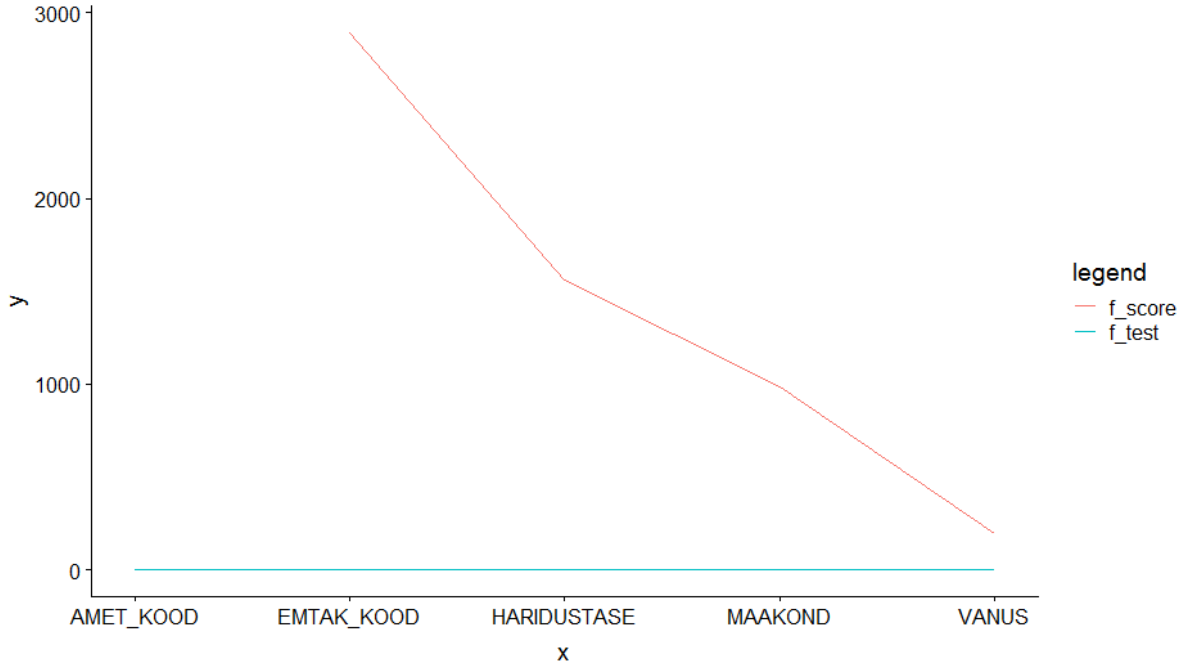
[33] J. Lauret, „Amazon’s sexist AI recruiting tool: how did it go so wrong?,“ Medium, 16 08 2019. [Võrgumaterjal]. Available: https://becominghuman.ai/amazons-sexist-ai-recruiting-tool-how-did-it-go-so-wrong-e3d14816d98e. [Kasutatud 12 05 2020].

[34] Khan Akademy, „Gradient,“ [Võrgumaterjal]. Available: https://www.khanacademy.org/math/multivariable-calculus/multivariable-derivatives/gradient-and-directional-derivatives/v/gradient.

[35] C. Johnson, M. Riggs ja R. Downey, „Fun with numbers: Alternative models for predicting salary levels,“ Manhattan, 1987.

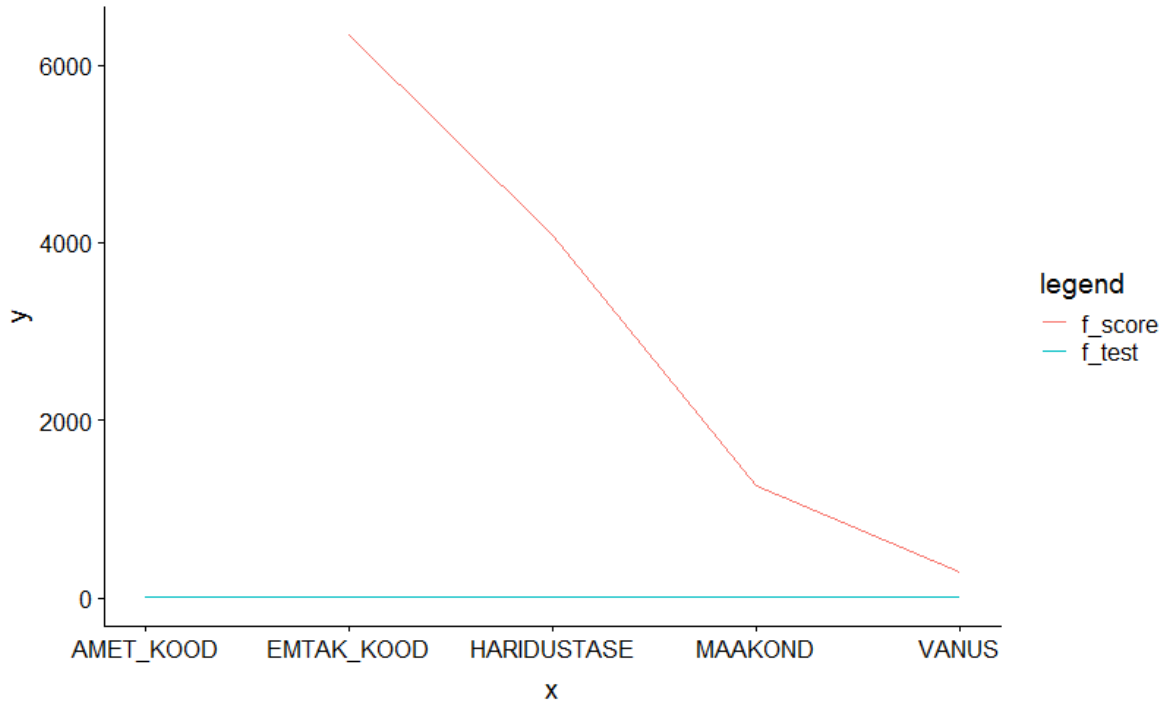
Lisa

Mehed



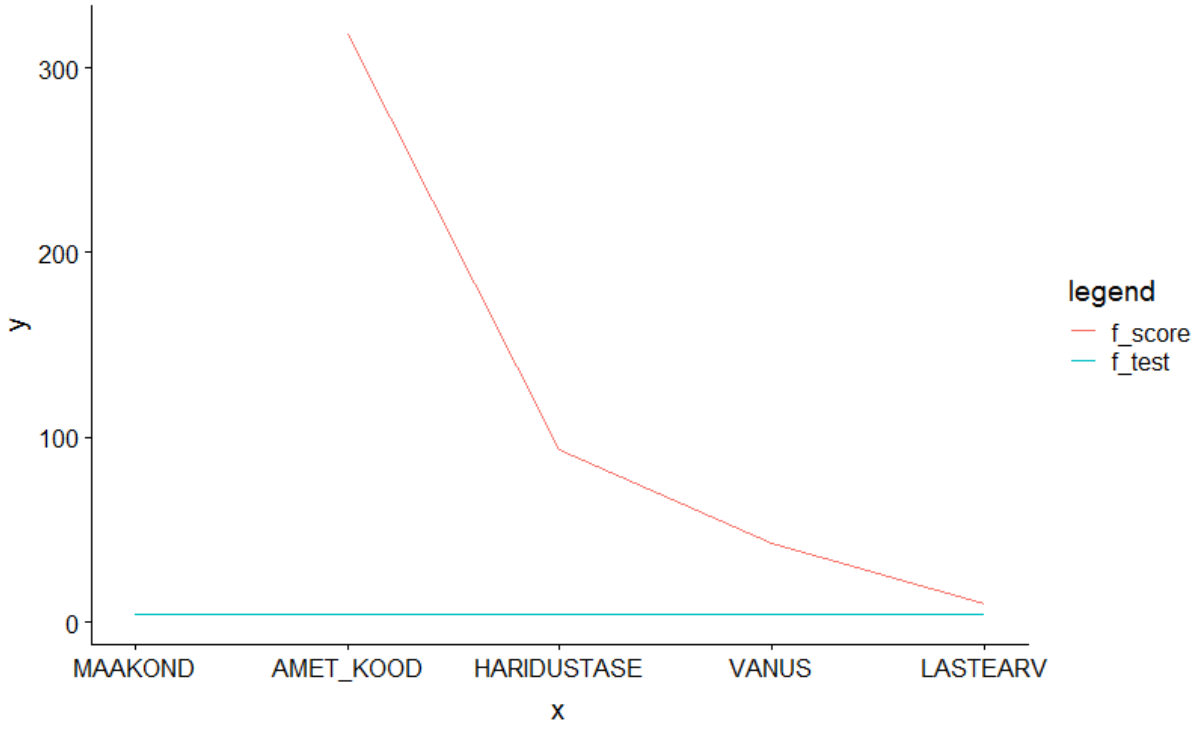
Joonis 14. Regressioonimudeli ehitamise tulemus mehed

Naised



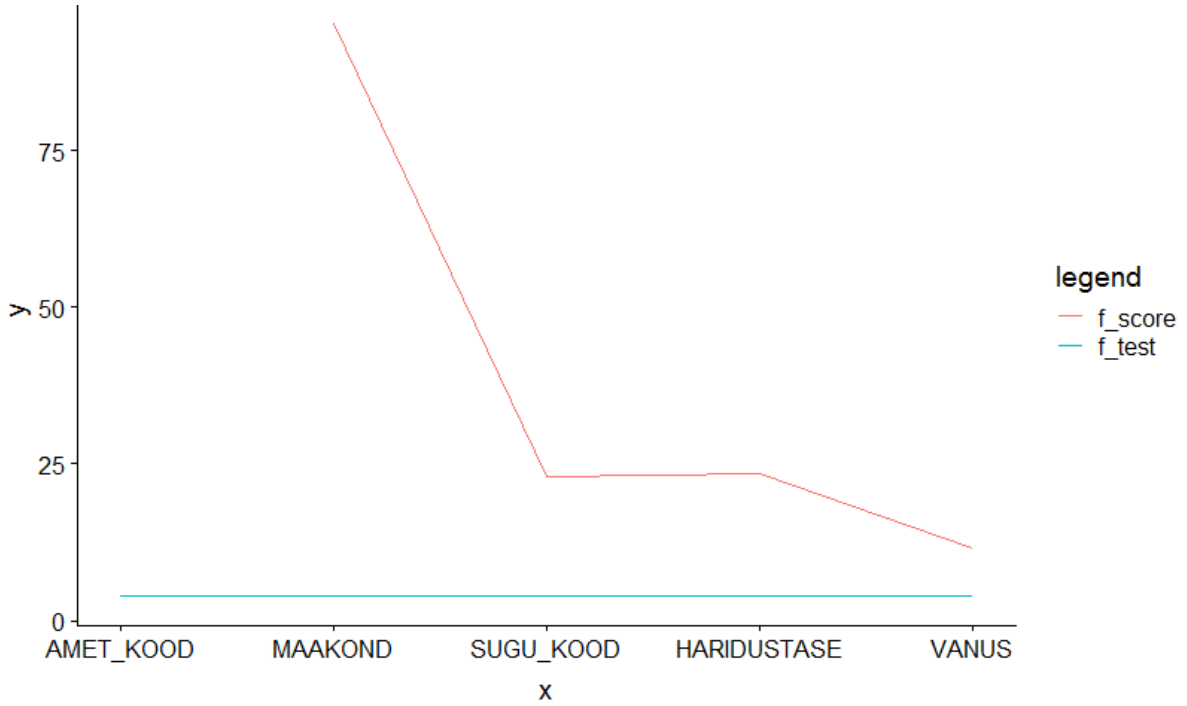
Joonis 15. Regressioonimudeli ehitamise tulemus naised

A – Põllumajandus, metsamajandus ja kalapüük



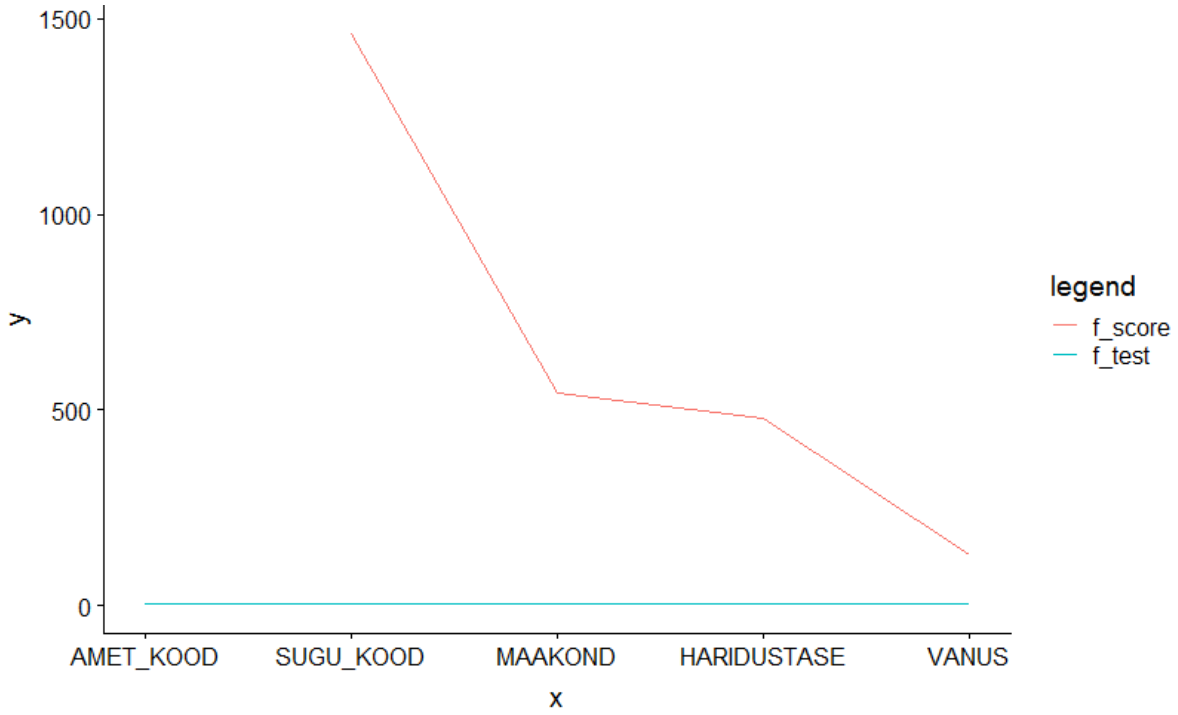
Joonis 16. Regressioonimudeli ehitamise tulemus EMTAK A

B - Mäetööstus



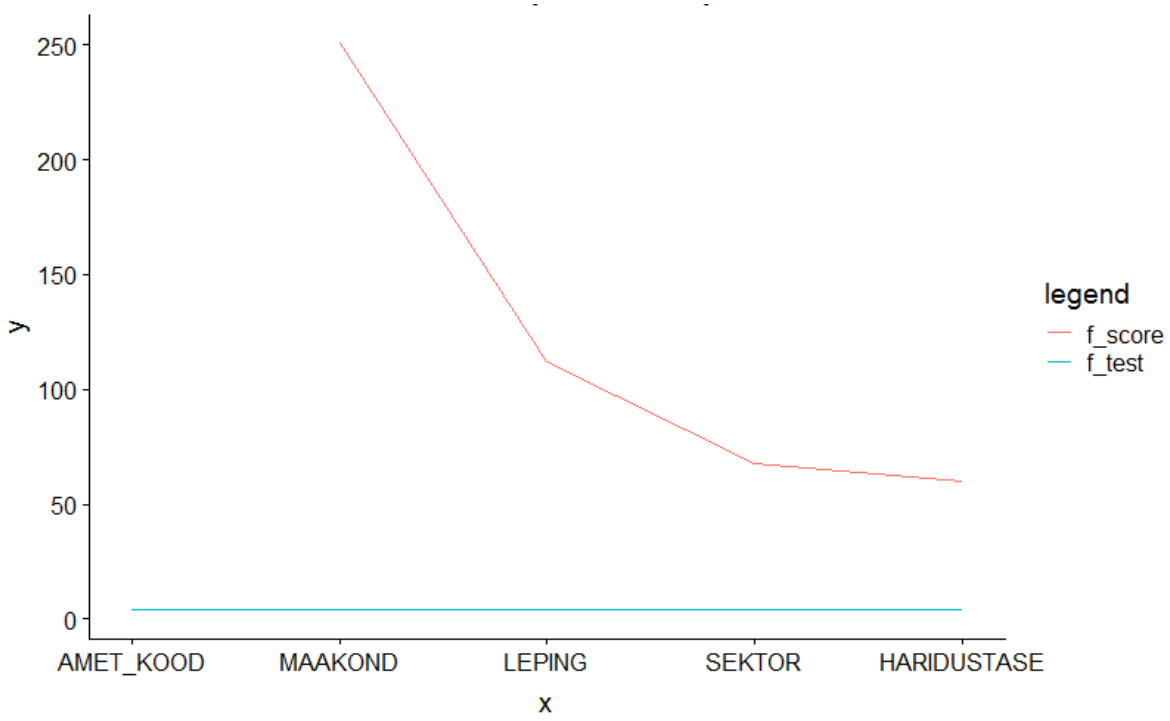
Joonis 17. Regressioonimudeli ehitamise tulemus EMTAK B

C – Töötlev tööstus



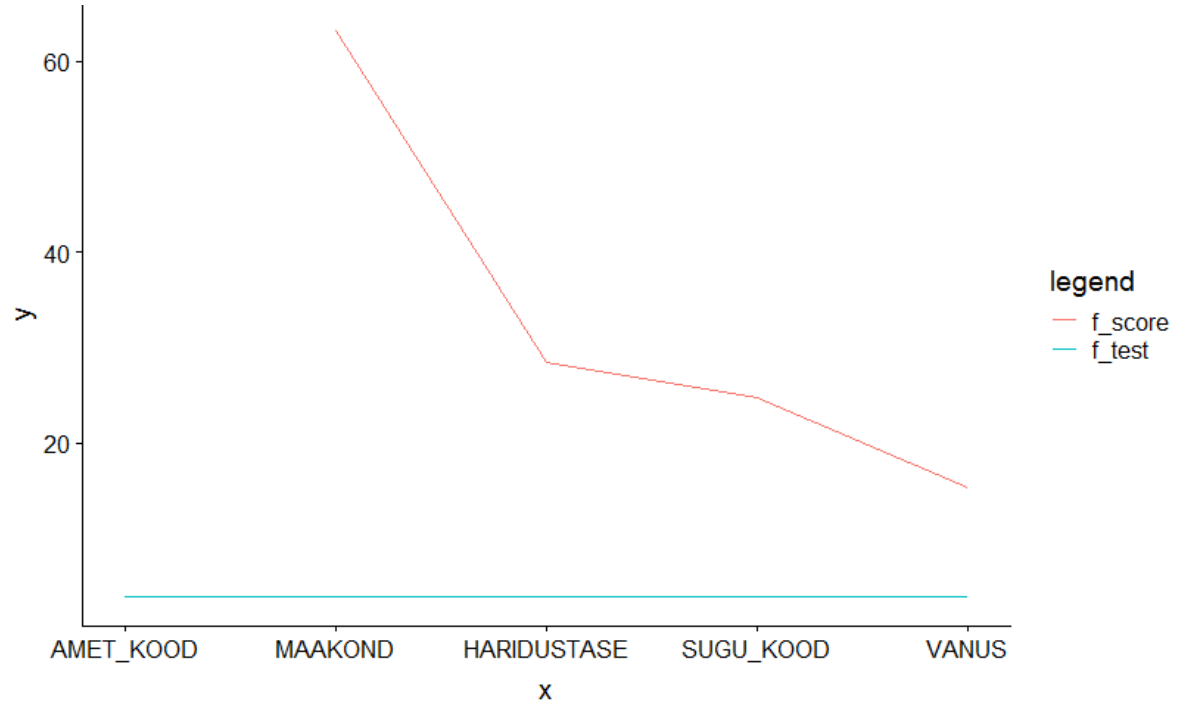
Joonis 18. Regressioonimudeli ehitamise tulemus EMTAK C

D – Elektrienergia, gaasi, auru ja konditsioneeritud õhuga varustamine



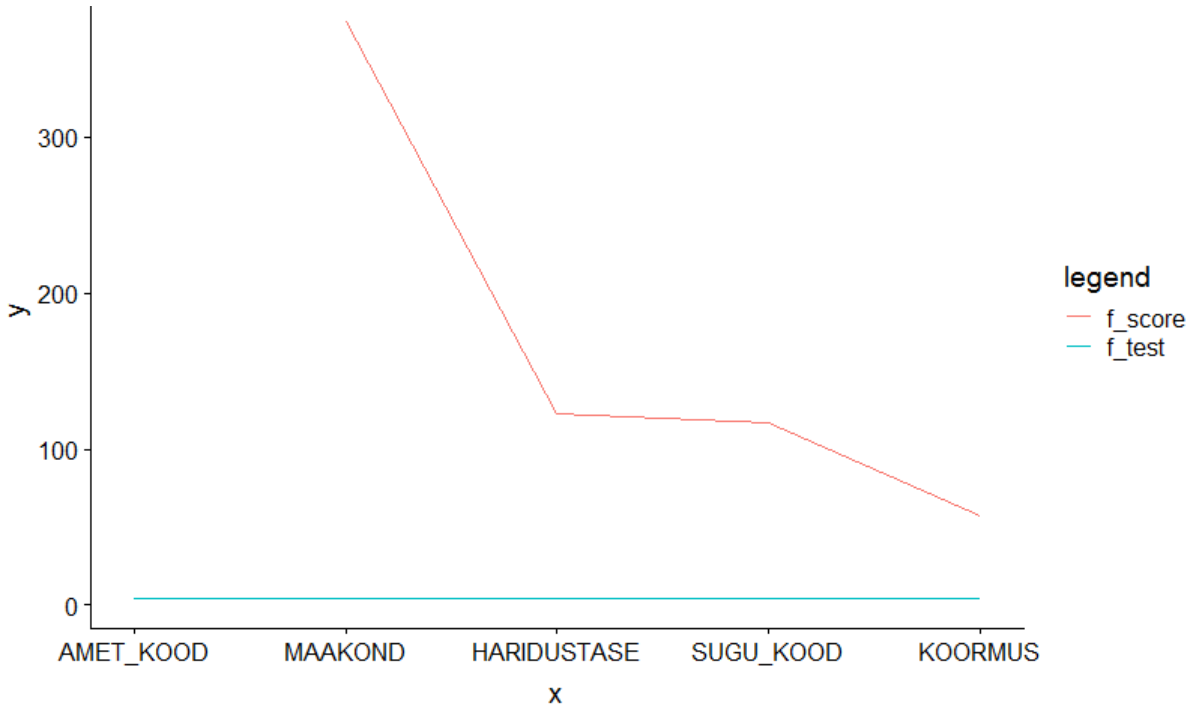
Joonis 19. Regressioonimudeli ehitamise tulemus EMTAK D

E – Veevarustus; Kanalisatsioon, jäätme- ja saastekäsitlus



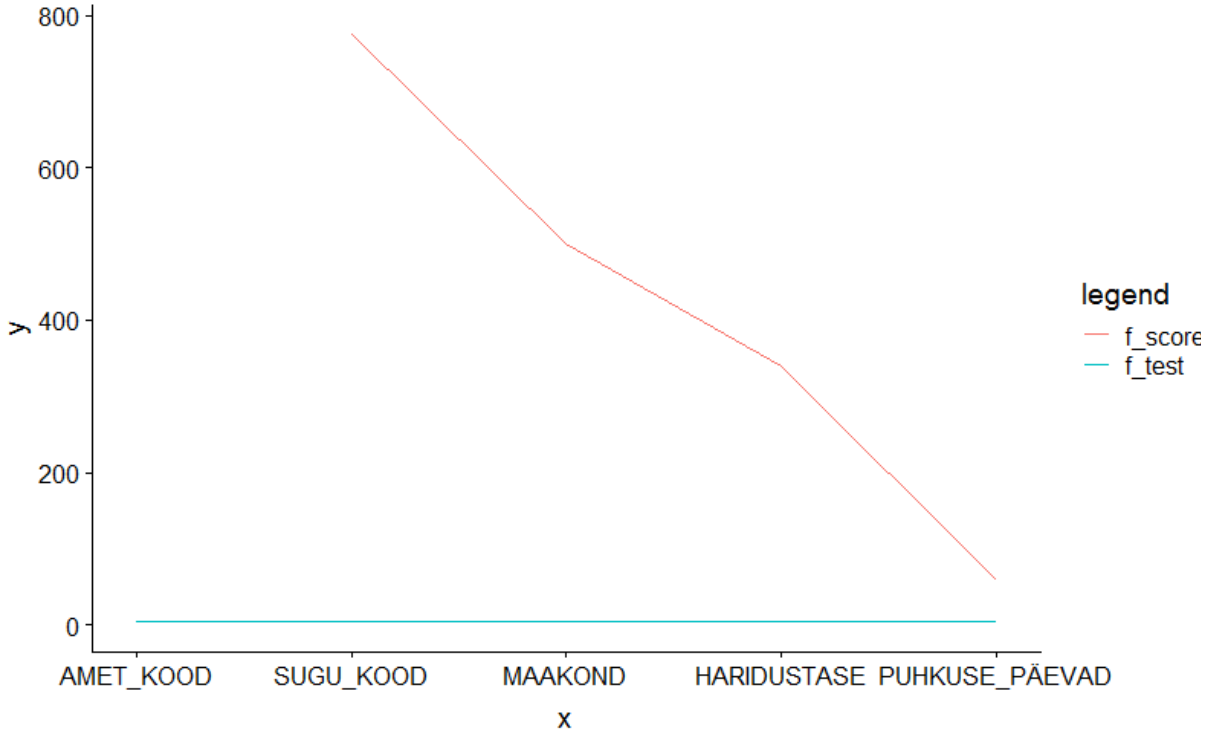
Joonis 20. Regressioonimudeli ehitamise tulemus EMTAK E

F - Ehitus



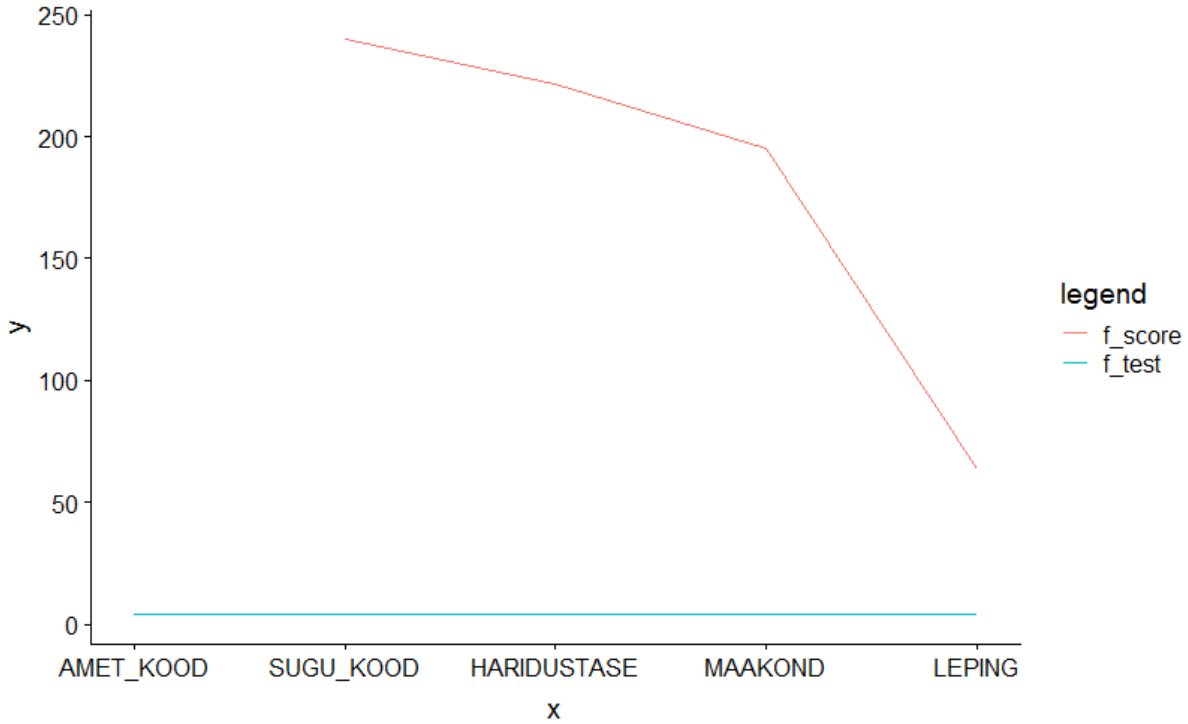
Joonis 21. Regressioonimudeli ehitamise tulemus EMTAK F

G – Hulgi- ja jaekaubandus; mootorsõidukite ja mootorrataste remont



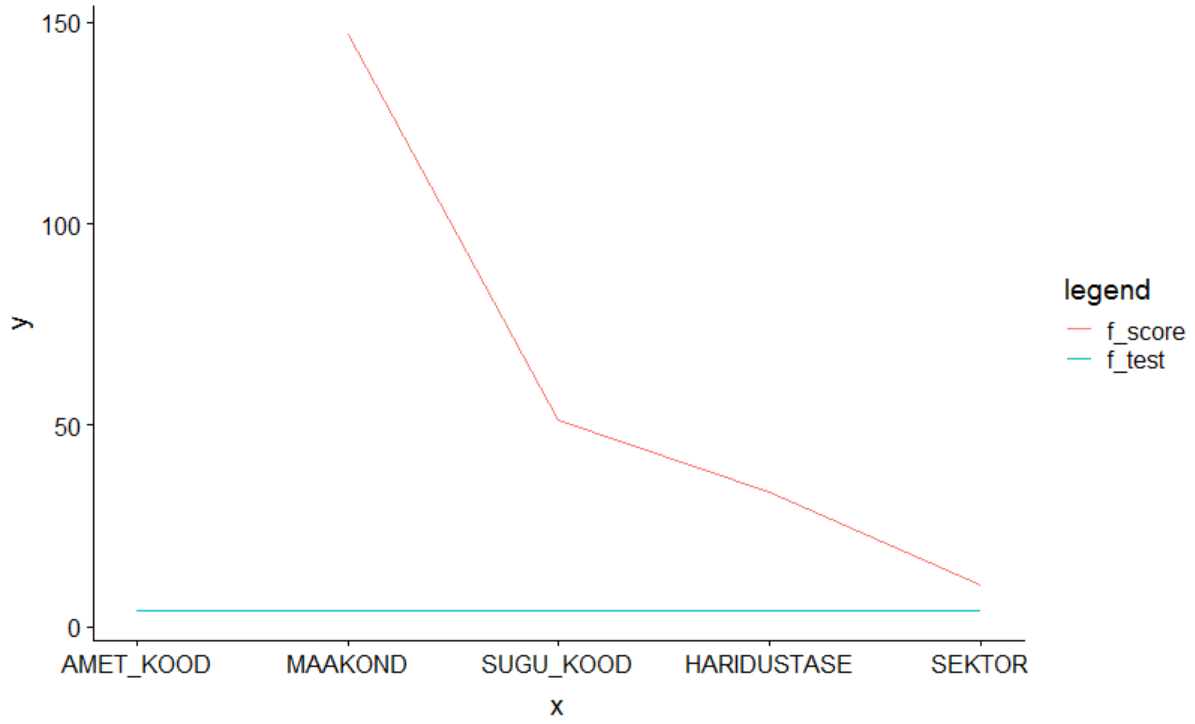
Joonis 22. Regressioonimudeli ehitamise tulemus EMTAK G

H – Veondus ja laondus



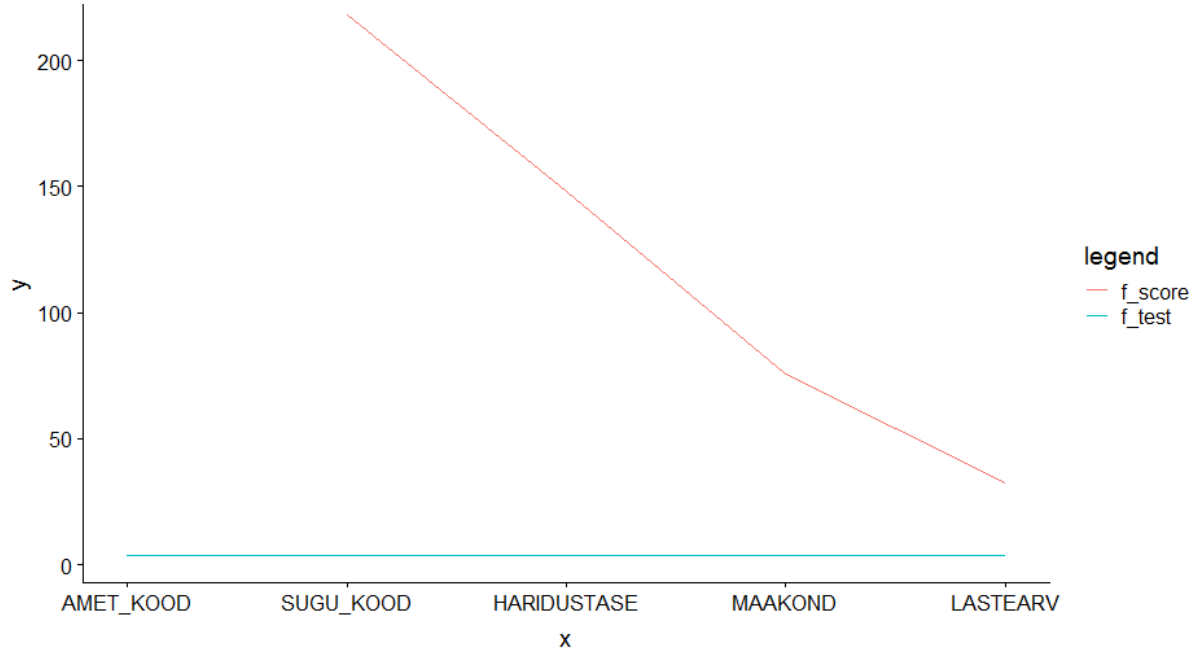
Joonis 23. Regressioonimudeli ehitamise tulemus EMTAK H

I – Majutus ja toitlustus



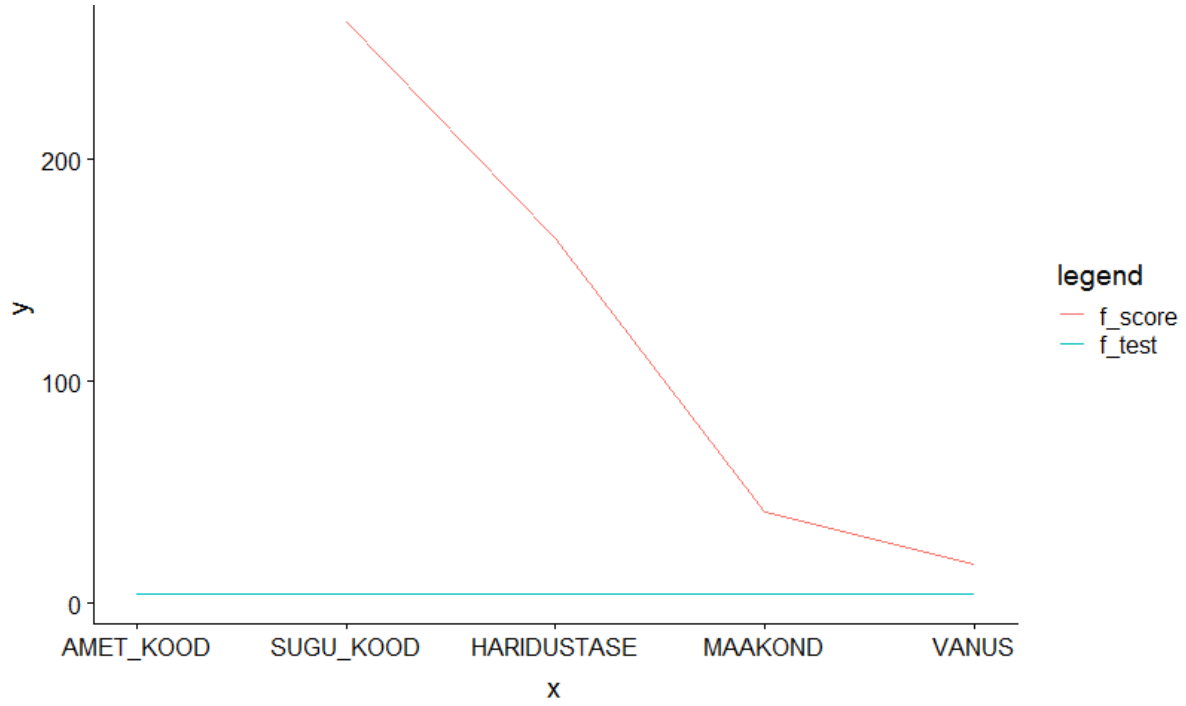
Joonis 24. Regressioonimudeli ehitamise tulemus EMTAK I

J – Info ja side



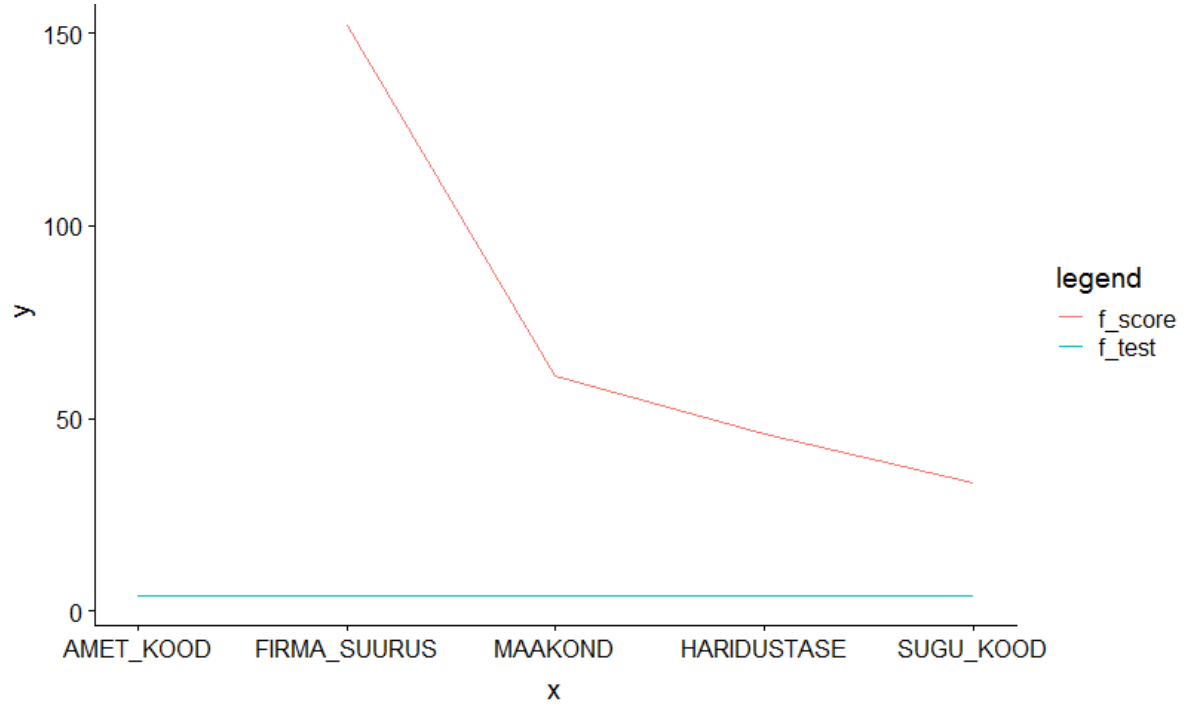
Joonis 25. Regressioonimudeli ehitamise tulemus EMTAK J

K – Finants- ja kindlustustegevus



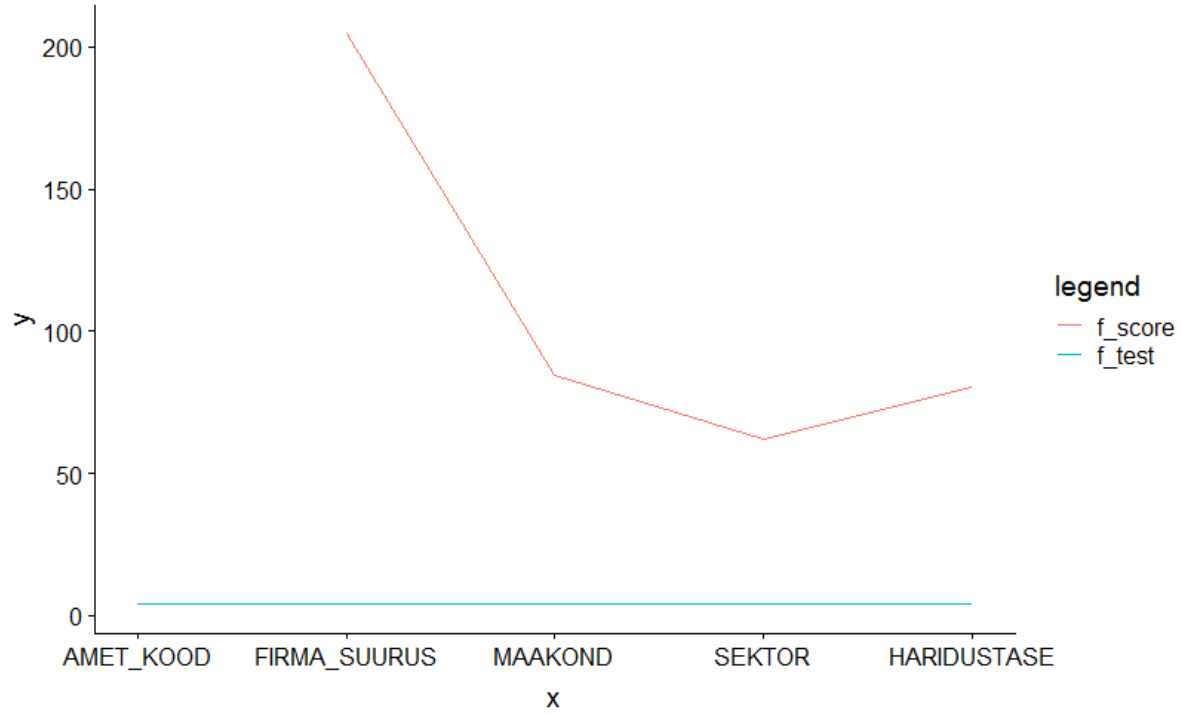
Joonis 26. Regressioonimudeli ehitamise tulemus EMTAK K

L – Kinnisvaraalane tegevus



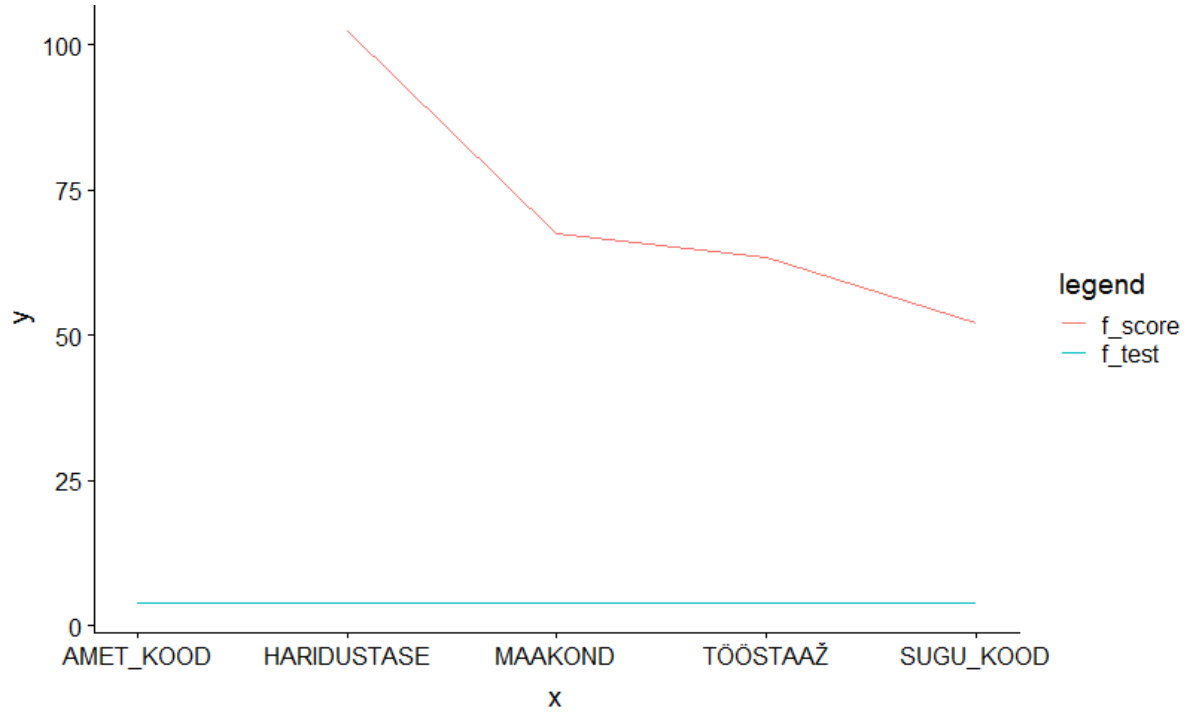
Joonis 27. Regressioonimudeli ehitamise tulemus EMTAK L

M – Kutse-, teadus- ja tehnikaalane tegevus



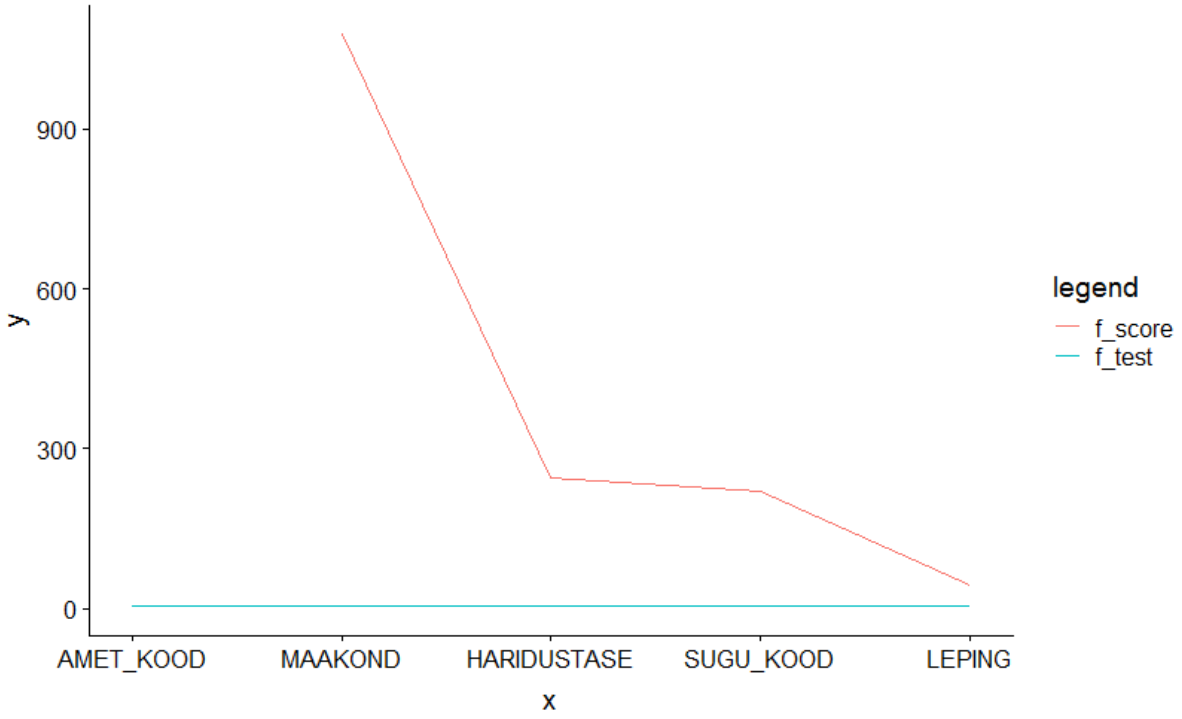
Joonis 28. Regressioonimudeli ehitamise tulemus EMTAK M

N – Haldus- ja abitegevused



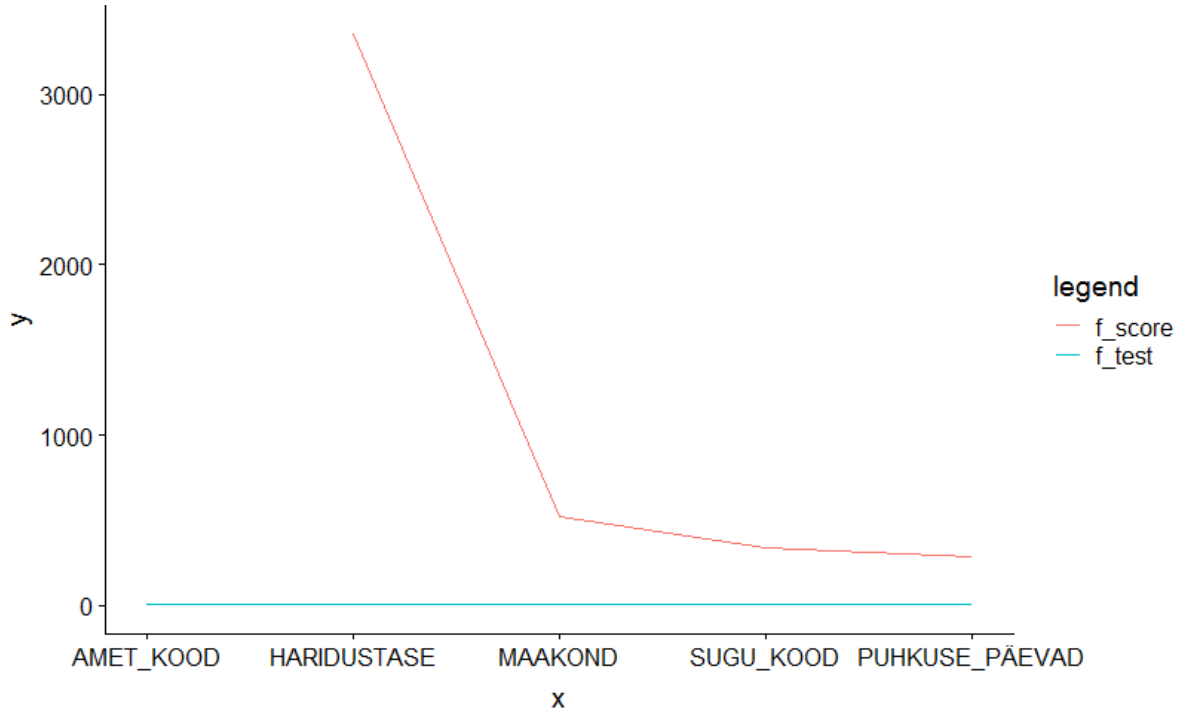
Joonis 29. Regressioonimudeli ehitamise tulemus EMTAK N

O – Avalik haldus ja riigikaitse; Kohustuslik sotsiaalkindlustus



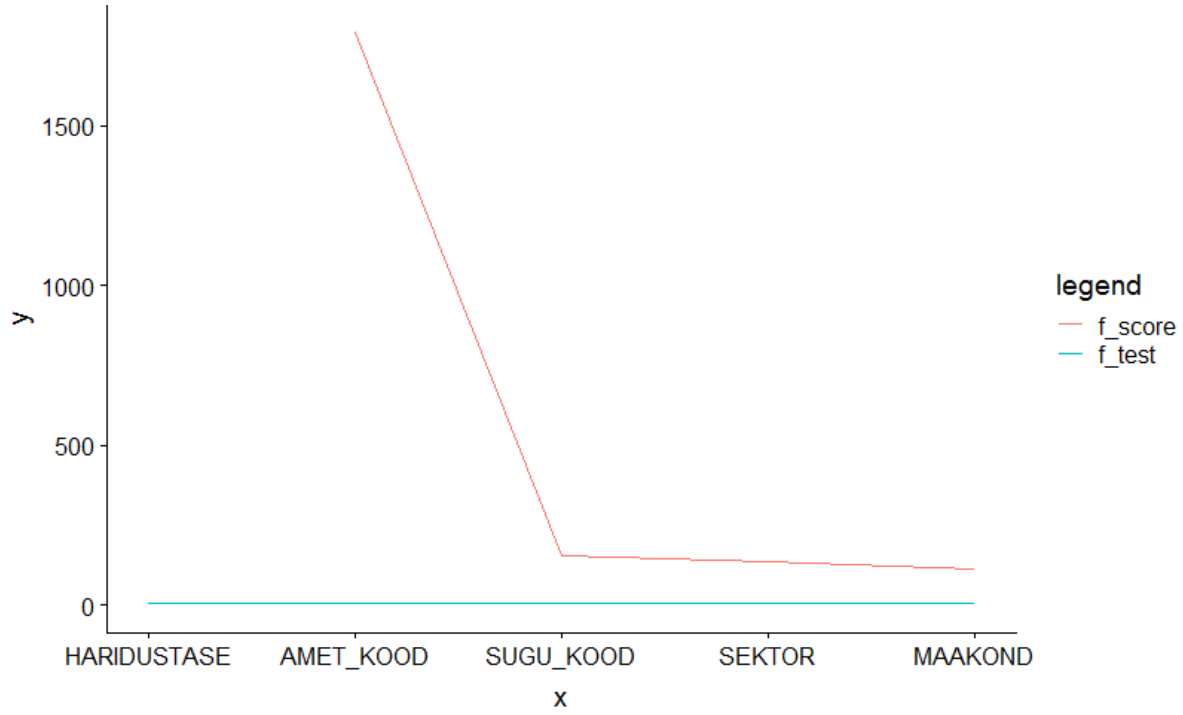
Joonis 30. Regressioonimudeli ehitamise tulemus EMTAK O

P - Haridus



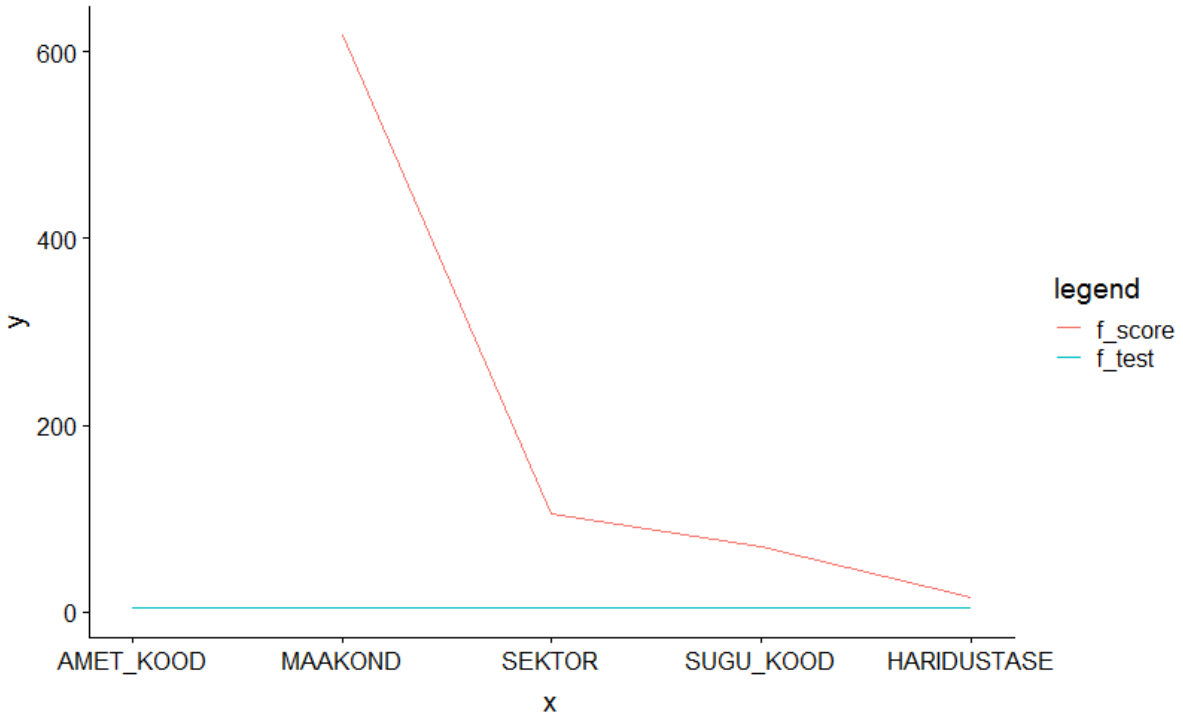
Joonis 31. Regressioonimudeli ehitamise tulemus EMTAK P

Q – Tervishoid ja sotsiaalhoolekanne



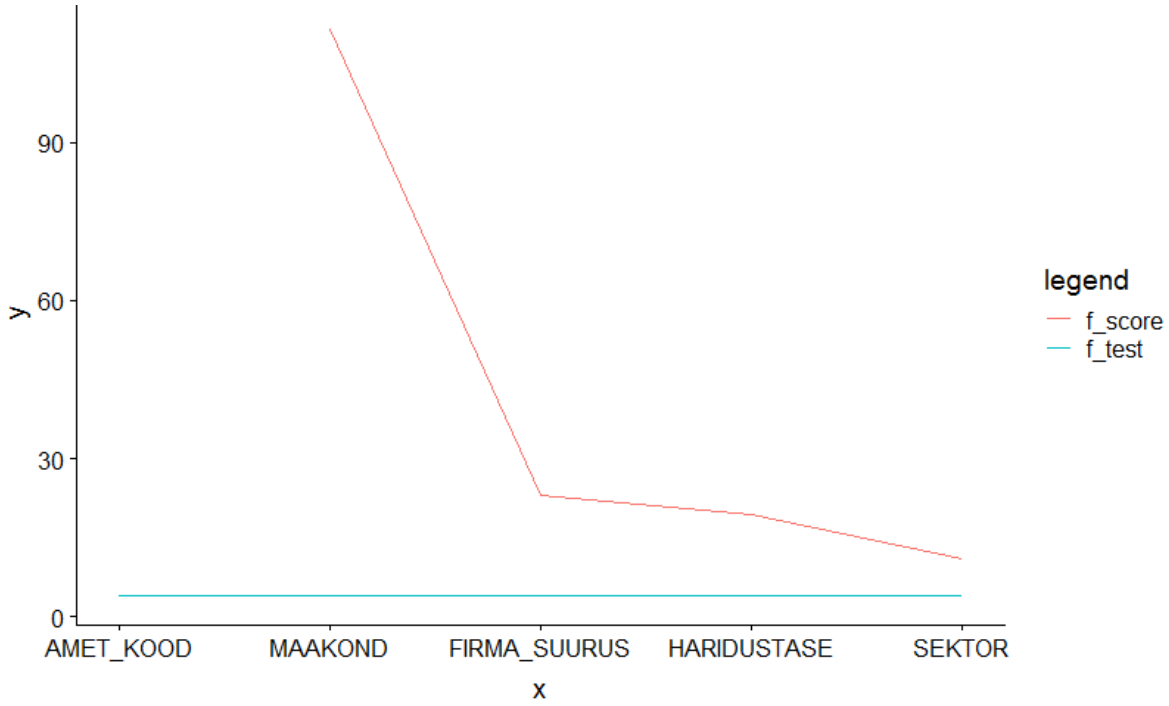
Joonis 32. Regressioonimudeli ehitamise tulemus EMTAK Q

R – Kunst, meelelahutus ja vaba aeg



Joonis 33. Regressioonimudeli ehitamise tulemus EMTAK R

S - Muud teenindavad tegevused



Joonis 34. Regressioonimudeli ehitamise tulemus EMTAK S

1 Juhid

Ennustus	HIGH_M	HIGH_W	LOW_M	LOW_W
HIGH_M	1857	953	755	796
HIGH_W	0	0	0	0
LOW_M	1178	330	2683	1333
LOW_W	530	402	805	2652

Tabel 5. Klassifitseerimisvea tabel Juhid

2 Tippspetsialistid

Ennustus	HIGH_M	HIGH_W	LOW_M	LOW_W
HIGH_M	1963	1309	783	1288
HIGH_W	674	1519	1519	1519
LOW_M	0	0	0	0
LOW_W	743	1948	1622	10317

Tabel 6. Klassifitseerimisvea tabel Tippspetsialistid

3 Tehnikud ja keskastme spetsialistid

Ennustus	HIGH_M	HIGH_W	LOW_M	LOW_W
HIGH_M	3065	1436	1494	2127
HIGH_W	0	0	0	0
LOW_M	108	27	204	105
LOW_W	743	1136	1659	5301

Tabel 7. Klassifitseerimisvea tabel Tehnikud ja keskastme spetsialistid

4 Kontoritöötajad ja klienditeenindajad

Ennustus	HIGH_M	HIGH_W	LOW_M	LOW_W
HIGH_M	0	0	0	0
HIGH_W	0	0	0	0
LOW_M	0	0	0	0
LOW_W	1174	2028	786	3510

Tabel 8. Klassifitseerimisvea tabel Kontoritöötajad ja klienditeenindajad

5 Teenindus- ja müügitöötajad

Ennustus	HIGH_M	HIGH_W	LOW_M	LOW_W
HIGH_M	351	133	19	15
HIGH_W	374	876	115	511
LOW_M	197	213	349	206
LOW_W	666	1949	714	6602

Tabel 9. Klassifitseerimisvea tabel Teenindus- ja müügitöötajad

6 Põllumajanduse, metsanduse, kalastuse ja jahinduse oskustöölised

Ennustus	HIGH_M	HIGH_W	LOW_M	LOW_W
HIGH_M	94	31	45	28
HIGH_W	21	65	12	22
LOW_M	16	6	62	30
LOW_W	46	88	73	201

Tabel 10. Klassifitseerimisvea tabel Põllumajanduse, metsanduse, kalastuse ja jahinduse oskustöölised

7 Oskus- ja käsitöölised

Ennustus	HIGH_M	HIGH_W	LOW_M	LOW_W
HIGH_M	4111	208	2341	1544
HIGH_W	0	0	0	0
LOW_M	827	12	1904	227
LOW_W	283	96	506	920

Tabel 11. Klassifitseerimisvea tabel Oskus- ja käsitöölised

8 Seadme - ja masinaoperaatorid ning koostajad

Ennustus	HIGH_M	HIGH_W	LOW_M	LOW_W
HIGH_M	2235	127	1084	364
HIGH_W	0	0	0	0
LOW_M	515	23	1423	173
LOW_W	916	342	747	1973

Tabel 12. Klassifitseerimisvea tabel Seadme- ja masinaoperaatorid ning koostajad

9 Lihttöölised

Ennustus	HIGH_M	HIGH_W	LOW_M	LOW_W
HIGH_M	1139	240	392	297
HIGH_W	448	714	164	329
LOW_M	0	0	0	0
LOW_W	777	958	1079	4804

Tabel 13. Klassifitseerimisvea tabel Lihttöölised