TALLINNA TEHNIKAÜLIKOOL
Infotehnoloogia teaduskond

Airika Andruse 211930IABM

Arvutikasutajate rahvuse ennustamine klaviatuurikasutuse iseärasuste järgi masinõppe meetodite abil

Magistritöö

Juhendaja: Avar Pentel
MSc

Tallinn 2023

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: Airika Andruse

10.05.2023

Annotatsioon

Klahvivajutuste järgi on võimalik ennustada arvutikasutaja isikuomadusi nagu näiteks tema vanust, sugu, haridustaset või ka emakeelt.

Käesoleva magistritöö eesmärgiks on masinõppe algoritmide abil ennustada klaviatuurikasutuse iseärasuste järgi arvutikasutajate rahvust, peamiselt Ida-Virumaal elavate eesti ja vene emakeelt kõnelejate näitel. Lisaks püütakse välja selgitada kui suure täpsusega on vastavalt masinõppe algoritmidele parimal juhul ennustamine võimalik ja missuguseid klaviatuuri kasutusdünaamikal põhinevaid omapärasid leidub eesti ja vene rahvusest inimeste seas.

Andmete kogumiseks loodi viieosaline veebirakendus, mis salvestab trükkimisel klaviatuuri klahvivajutusi ja nende aegu. Vormist saadud andmeid eeltöödeldi, analüüsiti ja nende põhjal loodi masinõppe algoritmide järgi ennustusmudelid, millega tuvastati arvutikasutajate rahvust. Ennustamiseks kasutati Weka programmi, kus loodi kuue masinõppe algoritmi mudelid.

Magistritöö parima ennustustulemustega meetoditeks olid *Support Vector Machine* (SVM(SMO)) ja *Multilayer Perceptron* (MLP), mille kümnekordse ristvalideerimise kaalutud keskmine F1-skoorid olid 0.957. Head tulemused saadi ka *Random Forest*-i ja *Simple Logistic*-u masinõppe algoritmide treenimisel (0.929 ja 0.953). Antud magistritööst saab järeldada, et klaviatuurikasutuse järgi on võimalik ennustada, kas inimene on eesti või vene rahvusest.

Lõputöö on kirjutatud eesti keeles ning sisaldab teksti 36 leheküljel, 7 peatükki, 14 joonist, 15 tabelit.

Abstract

Predicting the nationality of computer users by keyboard usage specifics using machine learning methods

It is possible to predict the personal characteristics of a computer user, such as their age, gender, education level or even their mother tongue, based on their keystrokes.

The aim of this master´s thesis is to use machine learning algorithms to predict the nationality of computer users, mainly Estonian and Russian native speakers living in Ida-Virumaa. In addition, the aim is to find out with what accuracy the best case prediction is possible according to the machine learning algorithms and what kind of differences based on keystroke dynamics are found among people of Estonian and Russian nationality.

To collect the data, a five-part web application was created, which records keyboard keystrokes and their timings when typing. The data from the form was pre-processed, analysed and based on these predictive models using machine learning algorithms were created to identify the nationality of computer users. The Weka program was used for the prediction, wherein models of six machine learning algorithms were created.

The method with the best prediction performance in the master's thesis were the Support Vector Machine (SVM(SMO)) and Multilayer Perceptron (MLP) with a weighted average F1-score of ten-fold cross-validation 0.957. Good results were also obtained by Random Forest and Simple Logistic machine learning algorithms' training (0.929 ja 0.953). It can be concluded from this master's thesis that it is possible to predict whether a person is of Estonian or Russian nationality by keyboard usage.

The thesis is in Estonian and contains 36 pages of text, 7 chapters, 14 figures, 15 tables.

Lühendite ja mõistete sõnastik

AdaBoost	*Adaptive Boosting*, adaptiivne võimendamine, masinõppe algoritm
Biomeetrik	*Biometric characteristic*, isiku bioloogiline ja käitumislik karakteristik
CSV	*Comma-Separated Values*, komaeraldusega väärtused, tekstifaili tüüp, mis sisaldab vorminguta andmeid: kirjeid eraldab reavahetuskood, kirje välju eraldavad komad. Failinime laiend: .csv
CV	*Cross-validation*, ristvalideerimine
C4.5 DT	*C4.5 decision tree*, C4.5 otsustuspuu, masinõppe algoritm
Eksimismaatriks	*Confusion matrix*, maatriks, milles registreeritakse katselistele näidetele mingi reeglistiku rakendamisel saadavate õigete ja väärade liigitusjuhtude arv.
EM	*Expectation-Maximization*, eelduste maksimeerimine, masinõppe algoritm
F-skoor	*F-score / F1-score / F-measure*, f-mõõdik või f-väärtus täpsemalt ka F1-skoor on veamõõdik.
Klassifitseerima	Mingite tunnuste alusel liikideks või rühmadeks jaotama, teatud liiki või rühma paigutama
Klasterdamine	Objektide rühmitamine nende omavahelise seose, nt omavahelise kauguse, läheduse või sarnasuse alusel
Klahvirütm	*Keystroke dynamics*, klahvivajutuste dünaamika näitajate kogum
Linear Regression	Lineaarne regressioon
Logistic Regression	Logistiline regressioon
ML	*Machine learning*, masinõpe
MLP	*Multilayer Perceptron*, mitmekihiline tunnetus, masinõppe algoritm
SL	*Simple Logistic regression*, lihtne logistiline regressioon, masinõppe algoritm
RF	*Random forest*, juhuslik mets, masinõppe algoritm
Regressioonanalüüs	*Regression, regression analys*, sõltuvate ja sõltumatute muutujate vaheliste seoste leidmine
SMO	*Sequential Minimal Optimization* (klassifitseerimisel)
SMOreg	*Sequential Minimal Optimization regressor*(regressioonil)
SVM	*Support vector machines*, tugivektormasinad, masinõppe algoritm

Sisukord

Jooniste loetelu

Tabelite loetelu

1 Sissejuhatus

Iga inimene, kes arvutit või muud nutiseadet kasutab jätab veebis enda tegevustega maha digitaalse jalajälje. Näiteks kirjutatakse otsingu termineid, kirju või muid tekste, laaditakse üles fotosid, videosid, sotsiaalmeedia postitusi jne [1]. Arvuti sisendseadmetega nagu hiir ja klaviatuur tehakse arvutis peamisi tegevusi, millede kasutusega saab kasutajate kohta erinevaid andmeid koguda.

Klahvivajutuse dünaamika või teisisõnu ka klahvirütm kuulub biomeetriliste tegurite alla, mis tähendab et andmed ehk näitajate kogum on mõõdetavad tema käitumise kaudu [2]. Biomeetriline omadus ehk biomeetrik on „isiku bioloogiline ja käitumuslik karakteristik, mida saab avastada ja millest saab välja eraldada eristatavad korratavad biomeetrilised erisused isikute automaatseks tuvastuseks, näiteks sõrme papillaarkurru struktuur, näo topograafia, näonaha tekstuur, käe topograafia, sõrme topograafia, iirise struktuur, käe veenistruktuur, peopesa kurrustiku struktuur, reetina muster jt“ [3].

Masinõpe (ing k *Machine Learning*) on Sõnaveebi (2022) definitsiooni järgi „arvuti treenimine etteantud andmete, reeglite põhjal otsuseid ja ennustusi tegema, vastavate algoritmide koostamine (nt tehislike närvivõrkude, klasterdamise meetodil)“ [4]. Ühendades klaviatuurivajutuse dünaamika ja masinõppe algoritmide kasutuse on võimalik ennustada arvutikasutaja isikuomadusi nagu näiteks tema vanust, sugu, haridustaset või ka emakeelt.

1.1 Taust ja probleem

Biomeetriliste omaduste välja selgitamine on oluline küberturvalisuse valdkonnas, õigusorganitele või ametiasutustele, kes uurivad ja koguvad andmeid kurjategijate või ka kannatanute kohta ning tuvastavad isikuid. Riigi Infosüsteemiameti (RIA) küberturvalisuse teenistuse juhi Gert Auväärt tõi välja RIA küberturvalisuse aastaraaamatu 2022 [5] artiklis, et „Küberkuritegevus on maailmas üks tulutoovamatest ning seeläbi ka kõige suurema kasvu teinud kuriteoliikidest.“ Klaviatuuri klahvivajutuse dünaamika uurimine on aktuaalne küberkuritegevus vähendamise või ennetamise

seisukohalt, millega on võimalik inimesi või inimgruppe identifitseerida või isikuid profileerida.

Võrdluses kasutaja sisestatud isikuandmeid ja klaviatuurkasutust analüüsides võib nende eristudes kahtlustada näiteks valeandmete esitamist, petuskeeme või ka identiteedivargust, mille tulemusel püütakse esitada ennast kellegina, kes ei olda. Lisaks võib rahvuse ennustamise järgi võimalik teada saada, kust võib küberkurjategijate poolt suunatud rünnak aset leida. Või teisest küljest võib kannatanute klaviatuurikasutuse järgi ennustada, kellele rünnakud on suunatud. Äriline kasu võibolla Eesti Riigi Infosüsteemiameti ennetustegevuses või küberturvalisust pakkuvatele ettevõtetele, kes saavad klaviatuurikasutuse kaudu ennustades uurida juhtumeid või pakkuda sellist tarkvara, mis analüüsib, kontrollib kasutajate klaviatuurikasutuse põhjal autentimist.

Veebikasutajate tegevusi jälgides on võimalik soovitada teatud arvutikasutajate gruppidele kindlat sisu. Näiteks kui pangapettused on aktuaalsed vene rahvusest või venekeelt kõnelevate inimeste seas siis saab neile ohutusteavet saata. Veel saavad näiteks e-poed teada, kes on nende suuremad kliendigrupid ning luua neile emakeelse sisuga reklaame või veebilehti.

Klaviatuuri klahvikasutuse ennustustulemusi saab kasutada rahvuste ja keelte põhiste gruppide klassifitseerimisel näiteks keeleõppes või keeletestide kontrollimisel. Olümpiaadidel, tasemetöödel või muudel sarnastel testidel saab tuvastada või klassifitseerida sooritajat rahvuse või kaudu. Näiteks vene emakeelega õpilane sooritab arvutis venekeele olümpiaadi, mis on mõeldud ainult eesti keel emakeelt kõnelejatele, siis võib petturi sooritust mitte arvestada või eemaldada.

1.2 Eesmärk

Käesoleva magistritöö eesmärgiks on masinõppe meetodite abil ennustada klaviatuurikasutuse iseärasuste järgi arvutikasutajate rahvust, peamiselt Ida-Virumaal elavate eesti ja vene emakeelt kõnelejate näitel. Lisaks püütakse peamisele eesmärgile toetudes leida ka eesti ja vene rahvusest inimeste klaviatuurikasutusel enim eristuvaid klahvivajutusi või tähekombinatsioone. Eeldati, et inimesed on harjunud kasutama enda kõnekeelele vastavaid klaviatuure rohkem ja ühed tähekombinatsioonid on klaviatuuril trükkides iseloomulikumad ühele rahvusele kui teisele.

Magistritöös on püstitatud kolm uurimisteema küsimust:

1. Mis on klaviatuuri klahvivajutuse dünaamika järgi kasutajate rahvus?
2. Kui suure täpsusega on võimalik ennustada klahvivajutuse dünaamika järgi kasutajate rahvust?
3. Kas eri rahvustel esineb iseloomulikke klaviatuurikasutuse tähekombinatsioone, kui jah siis missuguseid klaviatuurikasutuse tähekombinatsioone esineb?

Klaviatuurikasutuse andmete kogumiseks tehakse veebirakendusena vorm, millega kogutakse klahvivajutustest saadud informatsiooni. Vormist saadud andmeid analüüsitakse ja nende põhjal luuakse masinõppe algoritmide järgi ennustusmudelid, millega tuvastada arvutikasutajate rahvust ja neile iseloomulikke klaviatuurikasutuse klahvi- või tähekombinatsioone.

Magistritööga leitakse masinõppe abil parima ennustustäpsusega meetod, millega tuvastada veebirakenduse vormi täitjate rahvust. Eesmärgist tulenevalt loodetakse oodatavaks ennustustäpsuseks saada vähemalt 80% ja parimal juhul 90% lähedased ennustustulemused.

1.3 Ülevaade tööst

Magistritöö on jaotatud viide peamisesse etappi. Esiteks antakse ülevaade eelnevatest uuringutest, kus ennustatakse klaviatuurikasutusest tulenevalt arvutikasutajate isikuomadusi. Teiseks kirjeldatakse andmeid, nende kogumist, eeltöötlust ja analüüsimist. Kolmandaks tutvustatakse masinõpet ja töös kasutatavaid masinõppe meetodeid. Neljandaks keskendutakse masinõppe algoritmide kaudu genereeritud tulemustele vormiosade kaupa. Viiendas ehk viimases osas esitatakse parimate ja olulisemate ennustustulemuste kokkuvote, arutelu ja järelduste ülevaade.

2 Varasemad uuringud

Selles peatükis tutvustatakse klaviatuuri dünaamika uurimise kujunemist, uuringu teema kasvutrendi algust, varasemate uuringute aktuaalsust, vajalikkust ja leitud ennustustulemusi.

1860. aastal mõistsid telegraafi-operaatorid, et klahvirütmi järgi on inimesi on võimalik ära tunda. Võrdlusena toodi välja, et tippimine on sarnaselt kõnekeelele hästi äratuntav. Teise maailmasõja ajal kasutas sõjaväeluure morsekoodi saatja tuvastamiseks punktide ja kriipsude unikaalseid vajutamisviise. Asjatundlikel operaatoritel võimaldas sõnumi edastuskiiruse ja stiili kaudu järeldada, kas teises asukohas olev sõnumisaatja on oma või vaenlane [6].

Alates eelmise sajandi lõpust on klaviatuurivajutuse dünaamikat käsitletud juba üle 40 aasta [7, 8, 9, 10, 11, 12]. Kui inimene trükib, saab järjestikuste klahvivajutuste vaheliste viivituste, klahvivajutuste kestuse, sõrmede paigutuse ja klahvidele avaldatava surve põhjal konstrueerida unikaalse profiili. Korrapäraselt trükitud sümbolite, tähtede või sõnade puhul võivad käekirja tunnused olla isikupärased just konkreetsele inimesele. Klahvivajutuse dünaamika on protsess, mille käigus analüüsitakse kuidas kasutaja trükib, jälgides klahvivajutusi, klaviatuuri sisestusi näiteks tuhandeid kordi sekundis ja püüdes isikut tuvastada harjumuspärase rütmi alusel ning mustrite järgi. [13].

Sellel sajandil on teema vastu järjest enam huvituma hakatud. Paulo Henrique Pisani ja Ana Carolina Lorena 2013. aastal avaldatud uuringus „*A systematic review on keystroke dynamics*“ käsitleti väljaandeid klaviatuurikasutuse dünaamikast, mille abil püütakse kasutajaid tuvastada nende trükirütmi järgi. Uuringu tulemused näitavad, et 2005. aastast 2012. aastani on igal aastal klaviatuurikasutust käsitlevate väljaannete arv olnud 15-st kuni 25 artiklini, mis viitab valdkonna kasvutrendile [14].

Klahvivajutuse dünaamika tüüpiline uuring koosneb trükkimise andmete kogumisest või eelnevalt kogutud andmetest, millest tavalised on klahvivajutuste ajad nagu klahvide otsimis- ja hoidmisajad. Need on arvutatud vajutus- ja vabastamisaegadest [15].

Uuringutes rakendatakse peamiselt klassifitseerimisalgoritme või vähemal määral ka regressioonanalüüsi. Leitakse parima ennustustäpsusega klassifikaatoreid ning selgitatakse välja ennustustulemusi. Täpsustulemusi väljendatakse lisaks ennustustulemustele protsentides või osa kaalule sageli ka vigade ja eksimuste määrana. Vigade määr on õiguspärase kasutaja andmete protsent, mida peetakse ekslikult mitte-eeldatava klassi kuuluva isiku vastuseks. Eksimusmäär on teiste kirjutaja kirjutusproovide protsent, mida peetakse ekslikult õiguspärase kasutaja vastuseks. Igas uuringus tuleb arvestada, et klaviatuuril kogutud andmete ennustamisel võib esineda veamäärasid [12].

Varasemalt on klaviatuuri- ja arvutihiirekasutuse iseärasusi uuritud konkreetsemalt autentimise tehnoloogiate välja töötamiseks, peamiselt soo või vanuse ennustamiseks aga emakeele või rahvuse tuvastamiseks pigem vähem. Järgnevate tutvustavate uuringute läbiviimisega on püütud välja selgitada valeandmetega kasutajate identifitseerimist. Sotsiaalmeedia kontode loomisega valetatavad inimesed end vastupidiselt meestest naisteks ja vanuselt nooremaks, et näiteks pedofiilina suhelda teismeliste tütarlastega. Või noored märgivad enda vanuse vanemaks, et saada ligipääsu täiskasvanutele mõeldud veebikeskkondade sisule. Mõlema eelnevalt kirjeldatud näite ohtude vältimiseks on oluline valeandmete esitamist teadvustada, uurida ja analüüsida.

2012. aasta alguses avaldatud teadusartiklis *„A New Soft Biometric Approach For Keystroke Dynamics Based On Gender Recognition”*, mille autorid Giot ja Rosenberger, püüdsid välja selgitada, et uue pehme biomeetriline lähenemisega klahvivajutuse dünaamika jaoks on võimalik ära tunda inimese sugu. Uuringu ennustustäpsus katsesooritajate soo määramisel jäid tulemused 87% ja 92% (0.87 ja 0.92) vahele, mis näitab, et väga edukalt on klaviatuuri trükkimise kaudu võimalik teada saada, kas isik on mees või naine [16].

Eestis on Avar Pentel enda kahes töös, mis koostati 2017 aastal — „*Predicting Age and Gender by Keystroke Dynamics and Mouse Patterns*“ teadusartiklis ja 2019 aastal — „*Predicting User Age by Keystroke Dynamics*“ teadusartiklis, ennustanud klahvivajutuse dünaamika järgi kasutajate vanust ja sugu. Esimeses Penteli artiklis olid kõik ennustustulemused üle baastaseme ehk üle 0.5 ning parimaks ennustustäpsuseks saadi kuni 73% ehk 0.73. Teise artikli parimad ennustustäpsused olid juba paremad 0.88, 0.9 ja 0.92 [17, 18].

Tallinna Ülikooli Henri Vajaku (2019) bakalaureuse töös "Põlvkondade interaktsioonide erinevused arvutite sisendseadmete kasutamisel", kus on kasutatud nii klaviatuuri klahvivajutusi kui ka arvutihiire koordinaatide andmeid, on teada saadud parim ennustustäpsus erinevatel vanusegruppide kuuluvusest kuni 0.88-ni. [19].

Eelnevad uuringud näitavad, et klaviatuuri dünaamika kaudu isikugruppide ennustamine on aktuaalne ja tulemused ulatuvad ka üle 90% täpsuseni. Antud magistritöös püütakse klaviatuurikasutuse kaudu ennustada just rahvust aga keelte või rahvusega käsitlevaid uuringuid klaviatuuri dünaamikaga on loodud ja läbiviidud vähem kui soo või vanusegruppidega. Järgmised teadustööd on sellised, kus on klaviatuurikasutuse ennustamisel välja toodud erinevad tekstilised, keelelised või isikute päritolu aspektid.

Gunetti, Picardi ja Ruffo 2005. aasta septembris ilmunud teadusartiklis, kus viidi läbi erinevate keelte klahvivajutuste eksperiment ja analüüs juhtumiuuringuna, nimetusega *„Keystroke Analysis of Different Languages: A Case Study"* on leitud, et eri keeltes vabas vormis teksti kirjutusdünaamika kaudu saab kasutajaid tuvastada. Aga tuuakse välja, et nende tulemuste põhjal saadakse paremaid ennustustäpsusi pikemate vabas vormis tekstide kui lühemate analüüsist [20].

2012. aasta Bergsma, Post ja Yarowsky tehtud teadusartiklite siilimeetria või stülomeetria analüüsis („*Stylometric Analysis of Scientific Articles*") on ennustatud teadusartikli autorite sugu, kas autor on inglise keelt emakeelena kõneleja või mitte ja kas artikkel on avaldatud konverentsi või seminari materjalina. Uuringus leiti, et inglise keelt emakeelena kõnelejate ennustustäpsus F1-skoori tulemus oli 91,6% [21].

Teadusartiklis keeleliselt täiustatud klahvivajutuse dünaamika kasutamine, et ennustada masinakirjutaja kognitiivseid ja demograafilisi näitajaid (*„Utilizing linguistically enhanced keystroke dynamics to predict typist cognition and demographics"*) on sarnaselt eelmisele artiklile ennustatud inglise keelt emakeelena või mitte emakeelena kõnelejaid. Antud artiklis toodi välja, et inglise keelt mitte emakeelt kõnelejatel oli aeglasem trükkimiskiirus, sümbolite ja sõnade vahel pikemad pausid ning rohkem parandusklahvide kasutust [22].

Keelest sõltuva väljakutsepõhise klahvivajutuse dünaamika (*„Language dependent challenge-based keystroke dynamics"*) uuringus viidi läbi katse 60 osalejaga, kuue erineva rahvusest inimese seas. Nad pidid kirjutama sõnu inglise keeles, prantsuse keeles,

mis ei olnud ühegi osaleja emakeel ning ka enda emakeeles ja keelegrupi kaudu olevas suguluskeeles. Selles eksperimendis ei kasutatud juhuslikult kirjutatud sõnades keelespetsiifilisi tähemärke. Leiti, et emakeelseid sõnu trükiti kiiremini kui võõrkeelseid. Võrreldes prantsuse keelega kirjutati ka emakeele suguluskeelseid sõnu kiiremini ning kohati inglise keelsete ja emakeelsete sõnade trükkimisel ei erinetud oluliselt. Järeldati, et rahvuste eristamine on võimalik [23].

Klaviatuuri dünaamika keelelisi ja õpilaste päritolu aspekte on kajastatud veel 2020. aasta artiklis klahvivajutuse andmete uurimine kahes kontekstis: kirjakeele ja programmeerimiskeele mõjutavus õpiväljundite ennustatavuses teadusartiklis (*„A Study of Keystroke Data in Two Contexts: Written Language and Programming Language Influence Predictability of Learning Outcomes"*). Uuringus võrreldi Ameerika ja Soome programmeerimise kursuste õpilaste klaviatuurivajutusi Java ja Pythoni programmeerimiskeeltes. Töös leiti õpilaste kirjakeeles ja eelnevalt mainitud programmeerimiskeeltes esinevad olulised eristuvad sümbolid ja tähekombinatsioonid. Analüüsitud on Ameerika ja Soome õpilaste trükkimiskiirusi, digammide esinemisi, latentsusi ning võrreldud neid kursuse lõpptulemusega. Parim ennustustulemus saadi juhusliku metsa klassifikaatori kaudu, mille Pythoni programmeerimiskeele puhul oli 62% ja Javal 72% [24]. Kuigi eelnev uuring ei uuri otseselt kõnelejate rahvuse ennustust siis näitab see, et nii kirjakeele kui ka programeerimiskeeltele on klaviatuuritrükkimisel omapärased tähekombinatsioonid, antud kontekstis siis digraafid ja nende esinemissagedused.

Tallinna Ülikoolis on Aleksandr Rõbakov (2022) enda bakalaureuse töös „Arvutikasutajate emakeele määramine tema klaviatuuri kasutusdünaamika järgi" püüdnud välja selgitada emakeele ennustust klaviatuurikasutuse järgi. Autor analüüsis eesti ja vene keele rääkijate klaviatuurikasutuse harjumusi. Kokkuvõttena on kirjutatud, et tulemuste põhjal on võimalik kasutajaid keeleliselt grupeerida. Parimad ennustus täpsused juhusliku metsa algoritmi kasutades suudeti klassifitseerides saada üle 80%. Töös toodi välja ka eesti ja vene keele kõnelejatel esinevad tähekombinatsioonid n-grammidena nagu a_n_d, n_d, space_b, a_space_t jt, millest esimesed kaks on keeleklassidel eristavad tunnused. Aga selles töös esines ka puudusi, millest mõjuvam oli vähene osalejate arv ning veel keelegruppide suuruste ja vanuselised erinevused [25].

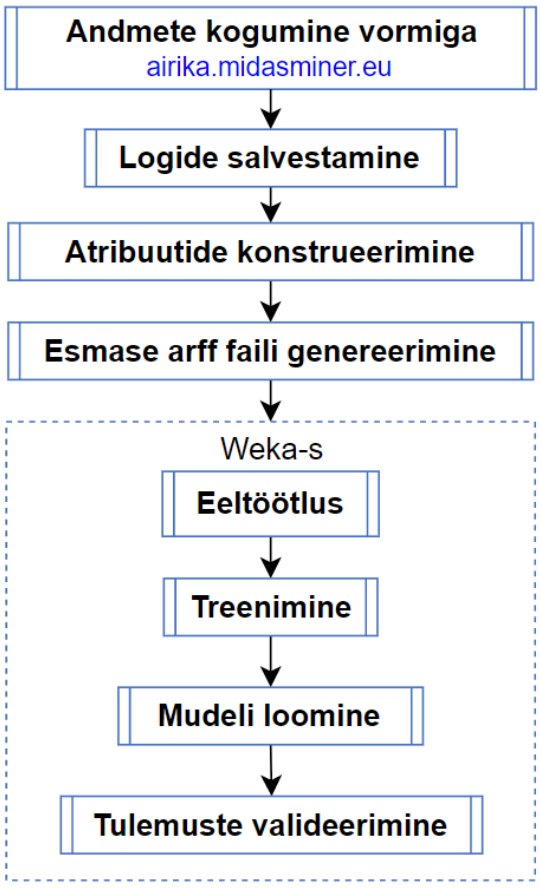
Viimane töö on peamine, mille rakendamist on soovitud käesoleva magistritööga edasi arendada. Rahvuse ennustamiseks eesti ja vene keelt kõnelevate inimeste klaviatuuri kasutusdünaamikat kasutades on eesmärgiks kaasata suuremat hulka osalejaid, kus on valimis võrdsemalt eesti ja vene keelt kõnelejad. Demograafiliste andmetena on eesmärgiks küsida lisaks soole ja vanusele ka rahvust. Kuigi sugu, vanus ja rahvus jäävad anonüümsuse ja rahvus otsese seose tõttu antud töös peamisest ja algsest analüüsist välja siis nende teadasaamine on hiljem kasulik kasutada õppematerjalina. Näiteks ennustades klaviatuurikasutuse andmete järgi sooritajate sugu, vanust või vanusegruppi ja rahvust. Rahvuse küsimisega saab kontrollida, kas klaviatuurikasutuse järgsed ennustustulemused vastaksid kasutaja enda poolt määratud rahvusele.

Lisaks soovitakse koguda etteantud tekstide kui ka vabatekstide trükkimisest saadud klahvivajutuste andmeid. Fikseeritud tekstidel tuleks arvestada selliseid bi- ja trigramme ehk järjestikuseid tähekombinatsioone, kus esineb kaks ja kolm tähte, mis on keeleliselt sama kuju ja kõlaga aga erikeelelistel klaviatuuridel asuvad teistes kohtades, näiteks a, o, t, k või selliseid tähti, mis on kõlalt erinevad, aga kujult samad, näiteks c, p.

3 Andmed

Antud peatükis antakse ülevaade andmeanalüüsi protsessist, kirjeldatakse lahendust, mis loodi andmete kogumiseks ja klahvivajutuste andmete logide loomist. Teiseks tuuakse välja valim ning missugused andmed jäid edaspidiseks meetodite rakendamiseks ehk missuguseks kujunes lõplik valim. Kolmandaks tutvustakse andmete eeltöötlust.

Andmeanalüüsi protsessi on etapiti näha joonisel 1. Etappide täpsemad kirjeldused ja selgitused on välja toodud 3.1 - 4.3 alapeatükkides.



Joonis 1. Andmeanalüüsi protsess

3.1 Andmete kogumine

Magistritööks vajalike andmete kogumiseks loodi veebirakendusena viieosaline vorm, mille täitmisel salvestatakse kasutaja trükitud sisestused.

Vorm on kättesaadav veebilehena aadressil airika.midasminer.eu. Veebirakenduse koostamisel on arvestatud varasemates klahvivajutuse dünaamikaga seotud töödes kasutatavaid testide või katsete lahendusi, kus on küsitud demograafilisi andmeid, etteantud tekstide kui ka vabas vormis tekstide sisestusi.

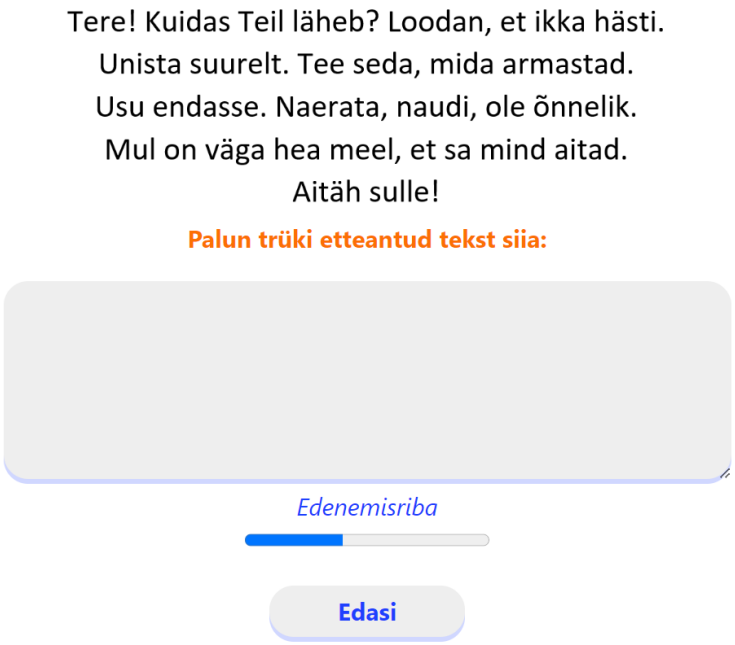
Veebilahenduse tagarakendus (ingl k *backend* või *back-end*) ehk serveripoolne osa on loodud PHP (ingl k *Hypertext Processor*) skriptimiskeeles. Eesrakendus (ingl k *fontend* või *front-end*) ehk kliendipoolne osa on loodud kasutades HTML (ingl k *Hyper Text Markup Language*) markeerimiskeelt, CSS (ingl k *Cascading Style Sheets*) stiilikujunduskeelt ja JavaScript-i, mis on veebiarenduses interpreteeritav objektorienteeritud programmeerimiskeel.

Vormi esimeses osas (Joonis 2.) peab kasutaja enda kohta ära märkima vastavalt sisestuse järjekorrale järgnevad andmed:

1. sugu,
2. vanus,
3. rahvus,
4. eesti keeles arvutis kirjutamise oskuse hindamine 1-9 punktiskaalal,
5. inglise keeles arvutis kirjutamise oskuse hindamine 1-9 punktiskaalal,
6. vene keeles arvutis kirjutamise oskuse hindamine 1-9 punktiskaalal.

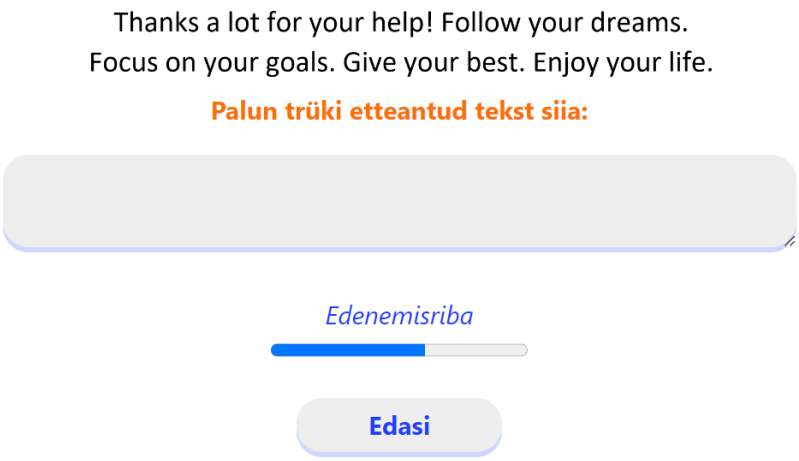
Joonis 2. Soo, vanuse, rahvuse määramine ja arvutis kirjutamiste oskuste hindamine eesti, inglise ja vene keeles 1-9 punktiskaalal

Vormi teises osas (Joonis 3.) tuleb kasutajal maha kirjutada etteantud eesti keelne tekst. Eesti keelne tekst koosneb 31 sõnast, mille pikkus on tühikutega 183 sümbolit ja ilma tühikuteta 153 sümbolit.



Joonis 3. Etteantud eesti keelse teksti mahakirjutamine

Vormi kolmandas osas (Joonis 4.) tuleb kasutajal maha kirjutada samuti etteantud tekst, mis on inglise keeles. Inglise keelne tekst koosneb 19 sõnast, mille pikkus on tühikutega 101 sümbolit ja ilma tühikuteta 83 sümbolit.

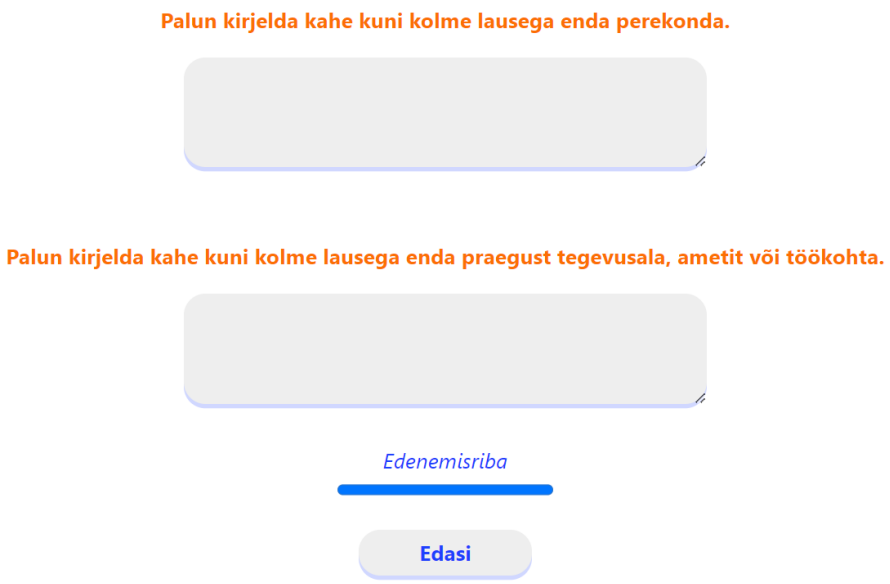


Joonis 4. Etteantud inglise keelse teksti mahakirjutamine

Eelviimases ehk vormi neljandas osas (Joonis 5.) kirjutab kasutaja vabas vormis teksti, milles kirjeldab etteantud pildil toimuvat tegevust.

Joonis 5. Vabas vormis etteantud pildil toimuva tegevuse kirjeldamine

Viimases ehk viiendas vormiosas (Joonis 6.) tuleb vastajal vabas vormis kirjeldada esimeseks enda perekonda ning teiseks enda praegust tegevusala, ametit või töökohta. Mõlema kirjelduse vastuseks oodatakse kahte kuni kolme lauset.



Joonis 6. Vabas vormis enda perekonna ja praeguse tegevusala, ameti või töökoha kirjeldamine

Kui test on täidetud ja kõigi viie osa sisestused salvestatud ning ära saadetud, tänatakse lõpetuseks vormi täitjat (Joonis 7.).

Joonis 7. Lõppvaade vormi täitnule

Kokku sisaldab vorm ühteteistkümmet andmesisestusvälja. Esimene ja kolmas on etteantud valikuvariant, milles saab valida ainult ühe eelnevalt määratletud üksteist välistavate valikute hulgast. Teiseks on numbrilise sisestuse väli, mille minimaalseks väärtuseks on viis ja maksimaalseks väärtuseks sada. Neljandast kuuenda sisestusena on 1-9 hinnanguskaalariba. Seitsmendas ja kaheksandas väljas tuleb tekstilahtrisse trükkida etteantud tekst. Viimastesse üheksandast kuni üheteistkümnendasse tekstilahtritesse tuleb kirjutada kahe- kuni kolmelauselised vabas vormis tekstid.

Vormi täitmise ajavahemiku arvestamisel on lähtutud eeldatavast kasutajate erinevatest trükkimiskiirustest ja -oskustest. Kogu vormi lugemiseks ja lahtrite sisestuste tegemiseks kulub vastajal umbes 5-10 minutit.

3.2 Logi

Iga kasutaja trükkimise andmetest salvestatati klahvivajutuste logi (Joonis 8.), mis koosneb vajutatud ja lahtilastud klaviatuuri klahvide koodidest, tähistest, suunast (alla *kup* on vajutamine ja üles *kdn* on vabastamine) ja ajast.

```
69,e,kdn,1681302481333
69,e,kup,1681302481420
32, ,kdn,1681302482101
32, ,kup,1681302482248
76,l,kdn,1681302482815
76,l,kup,1681302482906
65,a,kdn,1681302482924
65,a,kup,1681302483011
80,p,kdn,1681302483082
80,p,kup,1681302483157
```

Joonis 8. Klaviatuurivajutuse logi näide osaliste andmetega

Iga vormiosa klahvivajutuste logi salvestati kasutaja kohta eraldi, sellepärast et eelnevalt eeldati kui kasutaja ei täida kogu vormi täielikult siis on analüüsiks vähemalt osaliselt täidetud vormi andmed. Kõigi vormiosade klahvivajutuste logidest moodustati iga vormitäitja kohta klahvivajutuste objekt.

3.3 Valim

Veebirakenduse aadressi jagati pöördumise ja e-kirjana (Lisa 2.) Ida-Virumaa gümnaasiumite õppejuhtidele, TalTech Virumaa kolledži töötajatele, TalTech Nõustamiskeskuse (ingl k *TalTech Student Counselling Office*) Facebooki lehel, autori enda Facebooki lehel ja personaalselt tuttavate seas ajavahemikul 17.03-25.04.2023. Täpsustavalt oli pöördumises eelnevalt palutud vormi sisestus läbi viia arvutiga, kuna paremaks analüüsiks sobivad just klaviatuuril sisestatud andmed. Mobiilseadmetega tehtud katsed ei sobi edasiseks analüüsiks ja neid ei arvestatud. Veel paluti kirjutatavaid tekste mitte kopeerida (*copy*) ja kleepida (*paste*), vaid klaviatuuril trükkida.

Kirjutamine Ida-Virumaa koolidele seisnes soovis saada vastajaid nii eesti kui ka vene keelt kõnelevate inimeste seast võimalikult võrdselt ehk rahvused või keelegrupid oleksid võimalikult tasakaalus. Tasakaalustamine on oluline masinõppes, sellepärast et mõned algoritmid võivad vähemusosi ignoreerida ning klassifitseerivad kõik juhtumid enamusklassi ja seeläbi saadakse kõrgem klassifitseerimise täpsus [26].

Kokku koguti 544 vastaja sisestatud andmed. Paraku kõik vastajad täpsustavat kirja ei lugenud või jäi neile märkamata, et paluti vormi mitte täita telefoni või tahvelarvutiga. Osad vastanutest kirjutasid suvalisi vastuseid. Lisaks oli muust rahvustest vastajaid ainult 10. Mobiilseadme või tahvelarvutiga täidetud andmeid, muust rahvusest ja suvaliselt täidetud vormi vastajaid oli kokku 208. Need vastused jäeti edaspidisest meetodite rakendamisest välja.

Edasist analüüsi hakati teostama 336 vastaja andmetest. Eesti rahvusest oli 175 vastanut, kellest mehi 51, naisi 124 ja nende keskmine vanus oli 29,8 eluaastat. Vene rahvusest oli 161 vastanut, kellest mehi 54, naisi 107 ja nende keskmine vanus oli 23,2 eluaastat. Noorim vastanu oli 8 aastane ja vanimad kaks 68 aastased. Eesti ja vene rahvustest vastanute keskmine vanus oli 27 eluaastat.

Eestlased hindasid vene keeles arvutis kirjutamise oskust keskmiselt 2,8 punktilise hinnanguga, mis näitab et eestlased pigem ei ole kokku puutunud vene klaviatuuril trükkimisega ja ei oska vene keeles arvutis kirjutada. Venelased hindasid eesti keeles arvutis kirjutamise oskust keskmiselt 5,9 punktilise hinnanguga, mis näitab et Eestis elavad vene rahvusest inimesed kirjutavad ka eesti keelsel klaviatuuril.

Vormiosade kohta valiti erisuurused vastanute andmed, sest mõned vastanud vastasid ainult osadele vormiosadele või lahtrisse oodatud mõnelauselise vastuse asemel kirjutati üks või kaks sõna. Täheldati ka, et vormiosadele, kus oli oodatud kirjeldusi vastati inglise keeles. Järgevalt on selgitatud, kuidas vorm osade andmed valiti ja kui palju neid jäi.

1) I osa ehk demograafiliste andmete ja arvutil kirjutamise keeleoskuste hinnangute sisestamine jäeti suurema osana analüüsist välja, kuna edasi oli vaja anonüümseid vastuseid. Ainult eesti ja vene keeleoskuste hinnangutest loodi uus atribuut, millest täpsem kirjeldus on järgnevas 3.3 peatükis. Uue atribuudi andmeid oli 336.

2) II osa ehk eesti keelse etteantud teksti trükkimiselt valiti vastused, kus oli rohkem kui 150 klahvivajutust. II osa analüüsiks sobisid kõigi 336 vastanu andmed.

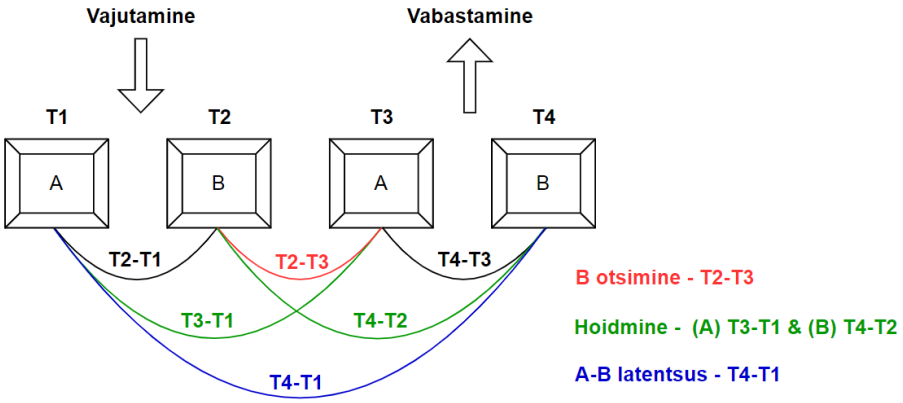
3) III osa ehk inglise keelse etteantud teksti trükkimiselt valiti vastused, kus oli rohkem kui 90 klahvivajutust. III osa analüüsiks jäi 331 vastanu andmed.

4) IV osa ehk vabas vormis pildil toimuva tegevuse kirjeldamisel valiti vastused, kus oli rohkem kui 50 klahvivajutust. IV osa analüüsiks jäi 310 vastanu andmed.

5) V osas ehk vabas vormis enda pere ja praegust tegevusala, ametit või töökoha kirjeldamisel valiti vastused, kus oli rohkem kui 50 klahvivajutust. V osa analüüsiks jäi 310 vastanu andmed.

3.4 Andmete eeltöötlus

Andmeanalüüsiks vajalike andmete eeltöötlusega alustati veebirakenduse koodis, kus kasutati klaviatuuri klahvivajutuste ja -vabastamiste aegu ning nendega arvutati edaspidi kirjeldatud keskmiste aegade atribuudid. Joonisel 9. on esitatud visuaalselt, kuidas klahvivajutuste ja -vabastamise aegadest arvutatakse trükkimisest saadud atribuudid nagu klahvi allhoidmisaeg, klahvide vaheline aeg, kahe klahvi vajutuse aeg ehk latentsus [17].



Joonis 9. Esmaste klahvivajutuse atribuutide arvutamine

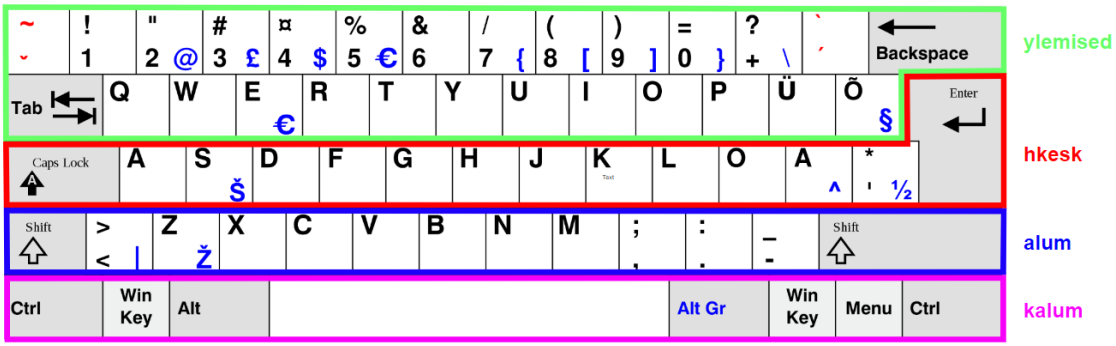
Klahvivajutuste logidest konstrueeriti ja arvutati välja erinevate klahvide otsimise ja allhoidmise ning n-grammide keskmised ajad. Näiteks kui tekstis esines mitu korda täht „a“, mille klahvikood on 65, siis tähe a otsimiseks kulunud keskmine aeg on 65_seek ja allhoidmiseks kulunud keskmine aeg 65_hold. N-gramm on mitme järjestikuse klahvi, tähe või sümboli vajutamiseks kulunud keskmist aega. Antud töös kasutatakse bigramme, mis on kahe klahviline n-gramm ja trigramme, mis on kolme klahviline n-gramm. Bigramm a_n on näiteks järjestikuste tähtede a ja n vajutamiseks kulunud keskmine aeg, klahvikoodidena 65_78 ja trigramm a_e_r on näiteks tähtede a, e ja r vajutamiseks kulunud keskmine aeg, klahvikoodidena 65_69_82 (Joonis 10.).

```
@attribute 65_hold numeric
@attribute 65_seek numeric
@attribute 65_78 numeric
@attribute 65_69_82 numeric
```

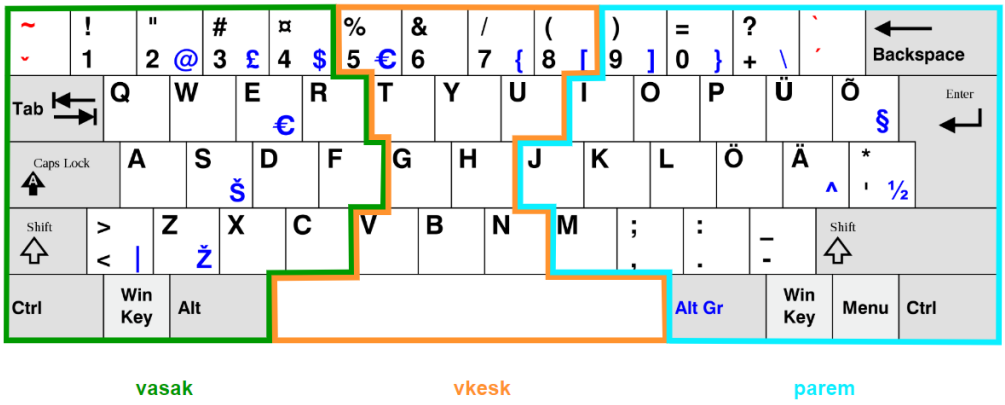
86: 65_hold Numeric	87: 65_seek Numeric	101: 65_78 Numeric	98: 65_69_82 Numeric
115.25	125.45	520.0	816.0
108.5	11.428571428571	585.0	333.0
105.03571428571	67.0	134.0	354.0

Joonis 10. Klahvivajutuste logidest konstrueeritud atribuutide näited

Veel jagati klaviatuur osadeks horisontaalselt ülemised, keskmised, alumised ja kõige alumised klahvid (Joonis 11.) ning vertikaalselt vasakud, keskmised ja paremad klahvid (Joonis 12.). Jagatud osade kohta arvutati klahvivajutusteks kulunud keskmised ajad määratletud klahvide asukohti arvestades.



Joonis 11. Klaviatuuri osad horisontaalselt



Joonis 12. Klaviatuuri osad vertikaalselt

Parandusklahvide nagu *backspace* (klahvikood 8), *delete* (klahvikood 46) ja ka nootleklahvide kohta arvutati nende esinemise suhteline sagedus testitäitja kohta. Sõnavahelised ajad ehk pausid leiti arvutades keskmised ajad nende n-grammide kohta, mis sisaldasid tühiku klahvivajutusi klahvikoodiga 32.

Lisaks on iga vormitäitja kohta standardiseeritud tema vormiosade atribuutide andmed. See tähendab, et ühe vormiosa erinevate atribuutide keskmisest ajast arvutati standardhälve ning seejärel lahutati iga üksiku atribuudi väärtusest keskmine ja jagati tulemus standardhälbega. Standardiseerimine on vajalik selleks, et tuua esile suhtelised erinevused erinevate klahvikombinatsioonide sisestamiste kiirustes.

Põhilises andmeanalüüsis kasutati teise kuni viienda vormiosi. Vormi esimeses osas pidi kasutaja sisestama demograafilised andmed ehk enda kohta soo, vanuse ja rahvuse. Need

atribuudid eemaldati. Lisaks tuli inimesel hinnata 1-9 punktiskaalal enda arvutis kirjutamise oskus eesti, inglise ja vene keeles. Eesti ja vene keeleoskuste hinnangutest genereeriti pideva tunnuse ennustamiseks atribuut „uus klass“: eesti keele oskuse hinnang - vene keele oskuse hinnang, mille väärtused jäid vahemikku -8 kuni 8 (17 väärtust). Seejärel eemaldati ka keelehinnangute atribuudid. Esimese vormiosa atribuudid on eemaldamise tõttu tabelis 1. märkimata jäetud. Atribuuti “uus klass” klassifitseerimisel ei kasutatud.

Iga vormiosa kohta loodi arff (*Attribute-Relation File Format*) fail, mis sisaldasid atribuute, klasse ja csv (*comma-separated values*) formaadis andmeid (Joonis 13.).

```
@attribute 219_hold numeric
@attribute 219_seek numeric
@attribute 219_78 numeric
@attribute 219_78_78 numeric
@attribute class {ee,ru}

@data
511.392857,727.154762,0.21121,0.438753,0.263446,
7,0.570418,0.44344,-1.64984,0.20536,0.656335,0.1
-0.337775,-0.23333,0.509833,1.356268,0.165588,1.
.866227,0.05624,1.028224,0.283036,1.25502,0.1210
631,0.813577,0.853376,-0.267754,0.995824,-0.0409
```

Joonis 13. II vormiosa arff faili näide.

Magistritöö andmete eeltöötluse filtreerimisel ja hiljem ka andmeanalüüsi läbiviimiseks kasutati Weka-t [27], mis on avatud lähtekoodiga tarkvara. Weka-s avati vormiosade arff failid ja eemaldati *RemoveUseless* filtriga need atribuudid, mis ei varieerunud üldse või varieerusid liiga palju. Filter kustutab kõik läbivalt konstantsed atribuudid automaatselt ja ka need, mis ületavad maksimaalset maksimaalset variatsiooniparameetrit. Maksimaalse dispersiooni analüüsi kasutatakse ainult nominaalsete atribuutide puhul [28]. Seejärel asendati *EMImputation* filtriga puuduvad arvväärtused *Expectation Maximization* (EM) algoritmi abil [29]. EM on latentse muutuja mudel, mida kasutatakse maksimaalse tõenäosusliku meetodi leidmiseks [30]. *Normalize* ehk atribuutide normaliseerimist viidi ka läbi aga see üldisi ennustustulemusi ei mõjutanud kuid võimaldas paremini hinnata iga üksiku atribuudi mõju klasside ennustamisele. Kuna eesti klass oli suurem ja vene klass väiksem siis tehti eesti klassile *undersampling* tasaklaalustamine, jättes välja need instantsid, milles oli liiga palju puuduvaid väärtusi.

Tasakaalustatud klassides oli võrdselt nii eesti kui vene andmeid esimeses kahes testiosas 161 ja kahes viimases 155.

Kokkuvõtlikult on andmete eeltöötluses arvutatud ja alles jäänud klahvivajutuste atribuudid välja toodud tabelis 1. Täpsemalt on juurde lisatud atribuutide kirjeldused ja arvud vormiosade kaupa.

Tabel 1. Klahvivajutuste atribuutide nimetused, kirjeldused ja arvud vormiosade kaupa

Atribuudi nimetus	Atribuudi tähendus	Vormiosad			
		II	III	IV	V
otsimine (*seek*)	klahvi vabastamise keskmine aeg, eelmise klahvi vabastamisest kuni uue klahvi vajutamiseni	24	26	15	15
hoidmine (*hold*)	klahvi allhoidmise keskmine aeg, klahvi vajutamisest vabastamiseni	24	26	15	15
n-grammi latentsus (*n-gram latency*)	kahe, kolme (tähe/sümboli) järjestikuse klahvivajutuse keskmine aeg	277	148	69	69
osa otsimine	klaviatuuri teatud piirkonna klahvide leidmisele kulunud keskmine aeg	9	9	9	9
osa hoidmine	Klaviatuuri teatud piirkonna klahvide allhoidmisele kulunud keskmine aeg	9	9	9	9
parandused (*backspace, delete*)	parandavate klahvide esinemise suhteline sagedus	1	1	1	1
paus (*space2, space3*)	sõna viimase tähe klahvi vabastamise, tühiku allavajutamise, vabastamise keskmine aeg kuni vajutatakse alla uue sõna esimene täht	2	2	2	2
KOKKU		**346**	**211**	**120**	**120**

4 Meetodid

Käesolevas peatükis selgitatakse üldiselt masinõpet, töös kasutatavaid masinõppe algoritme ja andmeanalüüsi meetodeid, millega ennustati lõplikusse valimisse jäänud sooritajate rahvust klaviatuurikasutuse iseärasuste järgi.

4.1 Masinõpe

Sissejuhatuses välja toodud masinõppe definitsioonile lisades on masinõpe (ingl k *machine learning*, ML) andmepõhine tehnoloogia, mis kasutab eelnevalt kogutud andmete põhjal erinevaid algoritme matemaatiliste mudelite koostamiseks ja prognooside tegemiseks. Prognoositava väljundi täpsus sõltub andmemahust. Suurem andmehulk aitab luua paremaid mudeleid, mille ennustus on täpsem. Mõiste masinõpe võttis esmakordselt kasutusele Arthur Samuel 1959. aastal. Masinõpet peetakse tehisintellekti (ingl k *artificial intelligence*, AI) alamhulgaks, mis on seotud algoritmide arendamisega. Masinõppe algoritmid võimaldavad arvutil andmetest ja varasematest meetodite rakendamistest iseseisvalt õppida. Tänapäeval kasutatakse masinõpet näiteks pildi- või kõnetuvastuse, emaili filtreerimise, Facebooki ja Instagrami sildistamise, sisusoovituse süsteemides ning isesõitvate autodes, virtuaalsete assistentides või tehisintellektide veebirakendustes [30].

Masinõppes eristatakse põhiliselt kahte tüüpi tunnuseid: diskreetsed (ingl k *discrete*), mis on tekstilised ja kuuluvad lõplikku hulka ja pidevad (ingl k *continuous*), mis on arvulised. Tunnuseid saab vaadelda ka binaarsetena, kas need on esidatud või mitte, näiteks õige või vale (0 või 1), kuulub klassi või mitte [31].

Masinõppe mudelid jagunevad peamiselt kolmeks: juhendatud õpe (ingl k *supervised learning*), juhendamata õpe (ingl k *unsupervised learning*) ja pooleldijuhendatud õpe (ingl k *semi-supervised learning*) [32]. Veel arvestatakse masinõppe liigina ka stiimulõpet (ingl k *reinforcement learning*) [30, 33].

Juhendatud masinõpe on etteantud ehk näidisandmete põhjal soovitava tulemuse ennustamine, näiteks rämpsposti liigitamine epostis eraldi kausta, müügitulu

prognoosimine andmete põhjal või objektide tuvastamine piltide kaudu. Juhendatud masinõpet kasutatakse klassifitseerimisel (ingl k *classification*) ja regressioonanalüüsis (ingl k *regression, regression analysis*), millede meetodid on näiteks Naïve Bayes, lineaarne ja logistiline regressioon, juhuslik mets, tugivektormasinad jt [30, 32].

Juhendamata masinõpe on etteantud andmete põhjal mudeli iseseisvalt mustrite leidmine. Andmestik ei ole eelnevalt klassifitseeritud ega kategoriseeritud ning algoritm peab ilma juhendamiseta ise leidma sarnaseid mustreid või grupeeringuid. Juhendamata masinõpet kasutatakse klasterdamiseks (ingl k *clustering*) ja seoste leidmiseks (ingl k *association*), mille peamisteks meetodideks on peakomponentanalüüs (ingl k *principal component analysis*, PCA) ja üksikväärtuste eraldamine (ingl k *singular value decomposition*, SVD). Veel kuuluvad juhendamata masinõppe alla närvivõrkude mudelite ehk neuronvõrgustike (ingl k *neural networks*), k-keskmise (ingl k *k-means*) ja tõenäosuslike (ingl k *probabilistic*) klasterdamise meetodid. Näiteks on juhendamata masinõppe uuriva andmeanalüüsi tegevusteks klientide andmete põhjal klientide segmenteerimine, müügistrateegiate leidmine või pildi ja mustrite tuvastamine [30, 32, 33].

Pooleldijuhendatud masinõpe on üks osa juhendatud ja teine osa juhendamata masinõppe algoritmide kasutamist, millega leitakse sarnasusi kui ka ennustustulemusi. Kasutatav meetod on otstarbekas siis kui analüüsiks ei ole suurt hulka andmeid, mida on vaja juhendatud masinõppes ja kui andmete grupeerimine on liiga aeganõudev või töömahukas, mis on eelduseks juhendamata masinõppes [32].

Lisaks on stiimulõpe, mis leiab mudeli tegevusplaani ning viib soovitava tulemuseni. Mudel sarnaneb juhendatud masinõppega ainult sellepoolest, et algoritmi ei treenita näidisandmete abil. Stiimulõpe toimib läbi katse-eksitus meetodi, mille õigete tegevuste ja vigade eest saadakse tagasisidet ning läbi selle õpitakse [30, 32, 33].

Antud töö liigitub edasise praktilises osas läbiviidud meetodite poolest põhiliselt juhendatud masinõppe alla, milles kasutatakse klassifitseerimisalgoritme ja vähesel määral viiakse läbi regresioonanalüüs. Mõlemaid algoritme kasutatakse masinõppes ennustamiseks ja need töötavad sildistatud andmekogumitega. Peamine erinevus regressiooni- ja klassifikatsioonialgoritmide vahel on see, et regressioonalgoritme kasutatakse pidevate väärtuste, näiteks hinna, palga, vanuse jne, ennustamiseks ja klassifikatsioonialgoritme kasutatakse diskreetsete väärtuste, näiteks mehe või naise, tõe

või vale, rämpsposti või mitte rämpsposti jne, ennustamiseks või klassifitseerimiseks [30].

Klassifitseerimine on protsess, mille käigus leitakse funktsioon, mis aitab jagada andmekogumi erinevatel parameetritel põhinevatesse klassidesse. Klassifitseerimise puhul treenitakse arvutiprogrammi treeningandmestikul ja selle põhjal liigitatakse andmed erinevatesse klassidesse. Regressioon on statistiline meetod, mille abil modelleeritakse sõltuvate (siht-) ja sõltumatute (ennustavate) muutujate vahelist seost ühe või mitme sõltumatu muutuja abil [30].

4.2 Kasutatud masinõppe algoritmid

Järgnevalt on tutvustatud neid masinõppe algoritme, mida antud töö analüüsiks kasutati. Algoritmide valik põhineb varasemate uuringute statistilise ennustuse meetodite valikust, autori õpi- ja katsetuskogemusest ning eelnevalt läbiviidud analüüsi parimatest ennustustulemustest.

Magistritöö andmeanalüüsi läbiviimiseks kasutati Weka-t [27]. Programm sisaldab masinõppe algoritmide kogumikku, mis on välja töötatud andmekaevandamise tegevuste jaoks [34]. Wekas käivitati klassifitseerimiseks juhusliku metsa, tugivektormasina, C4.5 otsutuspuu, lihtsa logistilise regresiooni, AdaBoost ja mitmekihilise tunnetuse algoritme ning eesti keelse etteantud teksti regresioonanalüüsiks rakendati juhusliku metsa, tugivektormasina ja lineaar regressiooni algoritme.

Random Forest (RF) on eesti keeles juhuslik mets ja see on masinõppe algoritm, mis ühendab mitme otsustuspuu väljundit, et jõuda ühe tulemuseni. Juhusliku metsa algoritm koosneb otsustuspuude kogumikust ja iga puu koosneb treeningkogumist võetud asendusvalimist, mida nimetatakse algvalimiks. Sellest treeningvalimist üks kolmandik on testandmed. Juhusliku funktsiooni tegevusega lisatakse andmekogumisse rohkem mitmekesisust ja vähendatakse otsustuspuude vahelist korrelatsiooni. Sõltuvalt probleemi tüübist varieerub ennustuse määramine. Juhuslikku metsa kasutatakse nii klassifitseerimis- kui ka regressiooniprobleemide lahendamisel. Klassifitseerimise puhul saadakse ennustatud klass enamushääletuse, mis tähendab kõige sagedasema kategoorilise muutuja alusel. Juhusliku metsa eelisteks on üleõppimise (ingl k *overfitting*)

ohu vähendamine, paindlikkus ja funktsiooni olulisuse lihtsasti määramine. Puuduseks on aeganõudev protsess, see vajab suuremaid andmekogusid ja on keeruline [32].

Tugivektormasinad, ing k. ***Support Vector Machines (SVM)*** on masinõppe algoritm, mida kasutatakse nii klassifitseerimise kui ka regressiooniprobleemide lahendamiseks. Tugivektormasina algoritmi eesmärk on luua parim joon või otsustuspiir, mis suudab andmestiku klassideks eraldada, et hiljem saaks hõlpsasti uue andmepunkti õigesse kategooriasse paigutada. Seda parimat otsustuspiiri nimetatakse hüpertasandiks ja see on klasse eraldav sirgjoon. Tugivektormasin valib äärmuslikud punktid/vektorid, mis aitavad hüpertasandit luua. Neid äärmuslikke juhtumeid nimetatakse tugivektoriteks ja seetõttu nimetatakse algoritmi tugivektormasinateks. Tugivektormasinate tugevusteks on hea arvutus- või ennustusvõime, kui klasside vahel on selge eralduvus ja see on ka suhteliselt mälusäästlik. SVM ei sobi suurte andmetega töötamiseks ja kui on ebamäärased andmed siis on jõudlus aeglasem ja klassifitseerimisel võivad esineda vead [30]. Tugivektormasinate treenimiseks on järjestikuse minimaalse optimeerimise algoritm ***Sequential Minimal Optimization*** (***SMO***), mida kasutatakse ruutprogrammeerimise (ingl k *quadratic programming (QP)*) probleemi lahendamiseks klassifitseerimisel. Algoritm jaotab andmed osadeks ja optimeerib iga osa eraldi. See asendab puuduvad väärtused, muudab nominaalsed atribuudid binaarseteks ja vaikimisi normaliseerib atribuudid. Võrreldes tavalise tugivektormasinate algoritmiga on järjestikuse minimaalse optimeerimise algoritm kiirem ja töötab paremini ebamääraste andmete jaoks [35]. Teisipidi aga regressioonil kasutatakse järjestikuse minimaalse optimeerimise regressiooni algoritm ***Sequential Minimal Optimization regression*** (***SMOreg***), mis loob regressioonimudeli, kasutades vaikimisi RegOptimizer optimeerimisalgoritmi [36].

C4.5 decesion tree ehk ***J48*** on otsustuspuul põhinev meetod, mida kasutatakse statistilise klassifikaatorina. Otsustuspuu on puustruktuurne klassifikaator, mille sisemised sõlmed esindavad andmekogumi tunnuseid, harud esindavad otsustusreegleid ja iga leht esindab tulemust. Otsustuspuu on visuaalselt lihtsasti arusaadav ja mõistetav [30]. Võrreldes teiste otsustuspuudega kasutab C4.5 algoritm ühekordset kärpimise protsessi (ingl k *Single Pass Pruning Process*), millega ennetatakse ülepaigutust. Veel on C4.5 algoritmil võimalik töötada nii diskreetsete kui ka pidevate andmetega ning see on kasulik osaliselt puudulike andmete analüüsimisel [37].

Logistiline regressioon ehk ***Logistic Regression*** ennustab kategoorilise sõltuva muutuja väljundit. Seetõttu peab tulemus olema kategooriline või diskreetne väärtus. See võib olla kas jah või ei, 0 või 1, tõene või väär jne, kuid selle asemel, et anda täpset väärtust 0 ja 1, annab see tõenäosuslikud väärtused, mis jäävad 0 ja 1 vahele. Muutujal ei tohiks olla multikollineaarsust. Logistiline regresioon on populaarne aga seda kasutatakse erinevalt lineaarsest regressioonist eelkõige klassifitseerimisel [27]. Lihtsat logistilist regresiooni, ingl k ***Simple Logistic (SL)***, kasutatakse ühe binaarse muutuja ennustamiseks ühe teise muutuja abil ning ka kahe sellise muutuja vahelise arvulise seose määramiseks. *Simple Logistic* on kergesti tõlgendatav mudel [38].

AdaBoost (AB) ehk adaptiivse võimenduse nominaalklassifikaator on algoritm, mida kasutatakse ansamblimeetodina. Kõige tavalisem *AdaBoost*-iga kasutatav hindaja on ühe tasemega otsustuspuud, mis tähendab ainult 1 jaotusega otsustuspuud. Neid puid nimetatakse ka otsustuskändudeks. *AdaBoost* loob mudeli ja annab andmeridadele võrdsema mõjutuskaalu. Võimendamise põhimõte seisneb selles, et kõigepealt koostatakse treeningandmete põhjal mudel ja seejärel teine mudel, et parandada esimeses mudelis esinevad vead [39].

Lineaarne regresioon ehk ***Linear Regression*** näitab lineaarset seost sõltuva ja ühe või mitme sõltumatu muutuja vahel. Lineaarset regressiooni saab jagada kahte tüüpi algoritmideks: lihtsaks lineaarseks regressiooniks (ingl k *Simple Linear regression*), mis on numbrilise sõltuva muutuja väärtuse ennustamine kasutades ühte sõltumatut muutujat; mitmeks lineaarseks regressiooniks (ingl k *Multiple Linear regression*), mis on arvuliselt sõltuva muutuja ennustamine rohkem kui ühe sõltumatu muutujaga. Lineaarne regressioon on masinõppes üks lihtsamaid ja populaarsemaid algoritme, mida kasutatakse regressioonanalüüsis [30].

Mitmekihiline tunnetus ehk ***Multilayer Perceptron (MLP)*** on neuronvõrgustik või närvivõrgu mudel, milles soovitud mudel ehitatakse eksperimentaalsete sisend- ja väljundpaaride korrelatsiooni abil. Algoritm põhineb õppimisprotsessidel, on universaalne ja väga paindlik funktsioonide ühtlustamisel. MLP-d kasutatakse peamiselt klassifitseerimisel ja ennustamisel [40].

4.3 Tulemuste valideerimine

Mudelite headuse hindamiseks kasutati 10-kordset ristvalideerimist, mis jagab andmed juhuslikult kümne võrdse suurusega osadeks. Üks osa on valideerimiseks ja üheksa osa treeningandmeteks. Tegevust korratakse kümmekorda, mille valideerimistulemustest leitakse keskmised väärtused.

Klassifitseerimise hindamiseks arvutatakse erinevaid mõõdikuid nagu täpsus (ingl k *precision*), saagis (ingl k *recall*) ja f-skoor. Täpsus (1) on tegelik osakaal kõigist positiivsetest ennustustest. Saagis (2) on õigesti tuvastatud osakaal tegelikest õige positiivsetest ennustustest. F-skoor (3) või f-mõõdik või f-väärtus (ingl k *f-score or f-measure*), täpsemalt ka F1-skoor on veamõõdik, mis mõõdab mudeli tulemuslikkust. F1-skoor on täpsuse ja saagise harmooniline keskmine. F-skoori arvutamisel arvestatakse õigeid positiivseid (TP) kordi kui positiivne tulemus ennustati õigeks, vale positiivseid (FP) kordi kui positiivne tulemus ennustati valeks ja vale negatiivseid (FN) kordi kui negatiivne tulemus ennustati valeks [41].

$$Precision = \frac{TP}{TP+FP} \quad (1)$$

$$Recall = \frac{TP}{TP+FN} \quad (2)$$

$$F\text{-}score = 2 * \frac{Precesion*Recall}{Precesion+Recall} = \frac{2TP}{2TP+FP+FN} \quad (3)$$

Täpsus - *Precision*
Saagis - *Recall*
TP - õige positiivne, *true positive*
FP - vale positiivne, *false positive*
FN – vale negatiivne, *false negative*
F-skoor – *F-score* (F1-skoor)

Mõõdikute põhjal luuakse üldiselt klassifitseerimisel kasutatav eksimismaatriks (ingl k *confusion matrix*), mis on õigete ja valede ennustustulemuste kokkuvõte (Tabel 2.).

Tabel 2. Eksimismaatriks

		Tegelik klass	
		Positiivne	Negatiivne
Ennustatud klass	Positiivne	TP	FP
	Negatiivne	FN	TN

5 Tulemused masinõppe meetodite rakendamisest

Siin peatükis antakse ülevaade klaviatuurikasutuse iseärasuste ennustusttulemustest, mis on esitatud iga vormiosa kaupa eraldi. Tulemused on saadud kogutud andmete põhjal ja eelmises meetodite peatükis kirjeldatud masinõppe algoritmide rakendamiste abil.

5.1 Ennustamise täpsused

Kasutatud masinõppe algoritmid, mis on edasiste tabelite üleval lahtrites, on esitatud järgnevate lühenditena:

- **RF** - *Random Forest*
- **SVM (SMO)**- *Support Vector Machine (Sequential Minimal Optimization)*
- **C4.5** - *C4.5 decesion tree või J48*
- **SL** - *Simple Logistic*
- **AB** - *AdaBoostM1*
- **LR** - *Linear Regression*
- **SVM (SMOreg)** - *Support Vector Machine (Sequential Minimal Optimization regression)*
- **MLP** - *Multilayer Perceptron*

Keeleklassi kuuluvuse ehk klassifitseerimiste ennustustulemused, mis on rakendatud loetelus viie esimese ja viimase masinõppe algoritmiga, on tabelites esitatud F1-skooridena ja nende kaalutud keskmistena. Eksimismaatriksitel (ingl k *confusion matrix*) on keeleklassidena eesti = EE ja vene = RU ning arvudena õigesti ja valesti ennustatud instantside jaotus keeleklasside vahel. Eksimismaatriks näitab klassifitseerimismudeli ennustusvõimet.

Lisaks on esitatud klaviatuurikasutuse järgi klassi määramisel enim eristumist mõjutavat atribuuti, mis on leitud *Simple Logistic* mudeli järgi ja järjestatud suuremast väiksemani koefitsiendi järgi. Mida suurem on koefitsent seda kauem aega võtab nende klahvide otsimine ja vajutamine.

Ainult teise vormiosa ehk etteantud eestikeelse teksi kohta tehti regresioonanalüüs.

5.1.1 Etteantud eesti keelse teksti tulemused

Etteantud eesti keelse teksti ehk teise vormiosa klassifitseerimise ennustustulemused, ennustusvõime eksimismaatriks ja parima eristusvõimega atribuudid on välja toodud tabelites 3-6 .

Teise vormiosa parimateks klassifitseerimise ennustusmudeliteks osutusid *Support Vector Machine* (SVM(SMO)) ja *Multilayer Perceptron* (MLP), mille kaalutud keskmised F1-skoorid saadi 0.957 ehk 95,7%. Üle 90%-lised ennustustulemused saavutati ka juhusliku metsa (RF) ja lihtsa logistilise regressiooni (SL) algoritmidega.

Tabel 3. II vormiosa klassifitseerimise ennustustulemused F1-skoori järgi

Keeleklassid	RF	SVM (SMO)	C4.5	SL	AB	MLP
Eesti	0.929	0.957	0.896	0.953	0.824	0.957
Vene	0.928	0.956	0.892	0.954	0.828	0.957
Kaalutud keskmine	0.929	0.957	0.894	0.953	0.826	0.957

Tabel 4. II vormiosa eksimismaatriks (*confusion matrix*) ennustustulemustest

Keele-klassid	RF		SVM (SMO)		C4.5		SL		AB		MLP	
	EE	**RU**	**EE**	**RU**	**EE**	**RU**	**EE**	**RU**	**EE**	**RU**	**EE**	**RU**
Eesti	151	10	155	6	147	14	153	8	131	30	154	7
Vene	13	148	8	153	20	141	7	154	26	135	7	154

Eestlaste kolm enim eristumist mõjutanud atribuuti olid N_A_U, U_hold ja Õ_seek. Järgnesid K_U_I, R_M_A ja N_I. Venelaste kolm enim eristumist mõjutanud atribuuti olid M_I_D, Space2 ja V_hold. Järgnesid Space_seek, U_D ja K_seek.

Tabel 5. II vormiosa parima eristusvõimega atribuudid Simple Logistic mudeli järgi

Eesti				Vene			
1.	N_A_U	**6.**	N_I	**1.**	M_I_D	**6.**	K_seek
2.	U_hold	**7.**	U_S_U	**2.**	Space2	**7.**	T_E_I
3.	Õ_seek	**8.**	M_U_L	**3.**	V_hold	**8.**	T_A_space
4.	K_U_I	**9.**	D_I	**4.**	Space_seek	**9.**	U_U_R
5.	R_M_A	**10.**	Corrective*	**5.**	U_D	**10.**	O_D

Corrective* - on kogum erinevaid klahve - delete, backspace ja neli nooleklahvi, mille eristumise mõju erinevalt teistest atribuutidest tuleneb vajutatud korda arvust.

Teise vormiosa regressioonanalüüsil leiti *Support Vector Machine* (SVM(SMOreg)) algortimiga kõige suurem korrelatsioonikoefitsent 0.97. Korrelatsioonikoefitsent ehk -kordaja näitab lineaarset korrellatsiooni ennustatud väärtuse ja tegeliku väärtuse vahel. Väikseimad vead leiti samuti tugivektormasinate mudelil - keskmine absoluut viga 0.89 ja ruutkeskmise viga 1.24.

Tabel 6. II vormiosa regressioonanalüüsi tulemused

	RF	**SVM (SMOreg)**	**LR**
Korrelatsioonikoefitsient	0.86	0.97	0.95
Keskmine absoluut viga	2.32	0.89	1.26
Ruutkeskmise viga	2.85	1.24	1.76

5.1.2 Etteantud inglise keelse teksti tulemused

Etteantud inglise keelse teksti ehk kolmanda vormiosa klassifitseerimise ennustustulemused, ennustusvõime eksimismaatriks ja parima eristusvõimega atribuudid on välja toodud tabelites 7-9.

Kolmanda vormiosa parimateks klassifitseerimise ennustusmudeliteks osutusid samuti *Support Vector Machine* (SVM(SMO)) ja *Multilayer Perceptron* (MLP), mille kaalutud keskmised F1-skoorid saadi 0.953 ehk 95,3%. Üle 90%-lised ennustustulemused saavutati ka juhusliku metsa (RF) ja lihtsa logistilise regressiooni (SL) algoritmidega.

Tabel 7. III vormiosa klassifitseerimise ennustustulemused F1-skoori järgi

Keeleklassid	**RF**	**SVM (SMO)**	**C4.5**	**SL**	**AB**	**MLP**
Eesti	0.908	0.953	0.846	0.944	0.725	0.953
Vene	0.912	0.954	0.844	0.944	0.728	0.954
Kaalutud keskmine	0.910	0.953	0.845	0.944	0.727	0.953

Tabel 8. III vormiosa eksimismaatriks (*confusion matrix*) ennustustulemustest

Keele-klassid	**RF**		**SVM (SMO)**		**C4.5**		**SL**		**AB**		**MLP**	
	EE	**RU**	**EE**	**RU**	**EE**	**RU**	**EE**	**RU**	**EE**	**RU**	**EE**	**RU**
Eesti	143	18	151	10	137	24	151	10	116	45	153	8
Vene	11	150	5	156	26	135	8	153	43	118	7	154

Eestlaste kolm enim eristumist mõjutanud atribuuti olid S_T, A_L ja A_space_L. Järgnesid Y_O, R_E ja R_E_A. Venelaste kolm enim eristumist mõjutanud atribuuti olid M_seek, O_N ja B_hold. Järgnesid L_seek, N_seek ja E_S.

Tabel 9. III vormiosa parima eristusvõimega atribuudid Simple Logistic mudeli järgi

Eesti				**Vene**			
1.	S_T	**6.**	R_E_A	**1.**	M_seek	**6.**	E_S
2.	A_L	**7.**	U_S	**2.**	O_N	**7.**	N_K_S
3.	A_space_L	**8.**	C_U_S	**3.**	B_hold	**8.**	K_S_space
4.	Y_O	**9.**	J_O_Y	**4.**	L_seek	**9.**	Space_B_E
5.	R_E	**10.**	R_space_G	**5.**	N_seek	**10.**	Space_H

5.1.3 Vabas vormis pildil toimuva tegevuse kirjeldamise tulemused

Vabas vormis pildil toimuva tegevuse kirjeldamise ehk neljanda vormiosa klassifitseerimise ennustustulemused, ennustusvõime eksimismaatriks ja parima eristusvõimega atribuudid on välja toodud tabelites 10-12.

Neljanda vormiosa parimaks klassifitseerimise ennustusmudeliks osutus *Random Forest* (RF), mille kaalutud keskmine F1-skoor saadi 0.796 ehk 79,6%.

Tabel 10. IV vormiosa klassifitseerimise ennustustulemused F1-skoori järgi

Keeleklassid	**RF**	**SVM (SMO)**	**C4.5**	**SL**	**AB**	**MLP**
Eesti	0.805	0.737	0.665	0.703	0.707	0.762
Vene	0.788	0.734	0.623	0.703	0.693	0.754
Kaalutud keskmine	0.796	0.735	0.644	0.703	0.700	0.758

Tabel 11. IV vormiosa eksimismaatriks (*confusion matrix*) ennustustulemustest

Keele-klassid	**RF**		**SVM (SMO)**		**C4.5**		**SL**		**AB**		**MLP**	
	EE	**RU**	**EE**	**RU**	**EE**	**RU**	**EE**	**RU**	**EE**	**RU**	**EE**	**RU**
Eesti	130	25	115	40	109	46	109	46	112	43	120	35
Vene	38	117	42	113	64	91	46	109	50	105	40	115

Eestlaste kolm enim eristumist mõjutanud atribuuti olid K_K_I, A_hold ja Õ_seek. Järgnesid V_seek, A_D ja I_L. Venelaste kolm eristuvat atribuuti olid Space_S_A, M_seek ja L_seek.

Tabel 12. IV vormiosa parima eristusvõimega atribuudid Simple Logistic mudeli järgi

Eesti				**Vene**	
1.	K_K_I	**6.**	I_L	**1.**	Space_S_A
2.	A_hold	**7.**	Backspace_seek	**2.**	M_seek
3.	Õ_seek	**8.**	Alum_seek	**3.**	L_seek
4.	V_seek	**9.**	Space3		
5.	A_D				

5.1.4 Vabas vormis enda pere ja praeguse tegevusala, ameti või töökoha kirjeldamise tulemused

Vabas vormis enda pere ja praeguse tegevusala, ameti või töökoha kirjeldamise ehk viienda vormiosa klassifitseerimise ennustustulemused, ennustusvõime eksimismaatriks ja parima eristusvõimega atribuudid on välja toodud tabelites 13-15.

Viienda vormiosa parimaks klassifitseerimise ennustusmudeliks osutus samuti *Random Forest* (RF), mille kaalutud keskmine F1-skoor saadi 0.793 ehk 79,3%.

Tabel 13. V vormiosa klassifitseerimise ennustustulemused F1-skoori järgi

Keeleklassid	**RF**	**SVM (SMO)**	**C4.5**	**SL**	**AB**	**MLP**
Eesti	0.800	0.748	0.671	0.699	0.656	0.770
Vene	0.787	0.756	0.644	0.695	0.660	0.759
Kaalutud keskmine	0.793	0.752	0.658	0.697	0.658	0.764

Tabel 14. V vormiosa eksimismaatriks (*confusion matrix*) ennustustulemustest

Keele-klassid	**RF**		**SVM (SMO)**		**C4.5**		**SL**		**AB**		**MLP**	
	EE	**RU**	**EE**	**RU**	**EE**	**RU**	**EE**	**RU**	**EE**	**RU**	**EE**	**RU**
Eesti	128	27	114	41	108	47	109	46	101	54	122	33
Vene	37	118	36	119	59	96	48	107	52	103	40	115

Eestlaste kolm enim eristumist mõjutanud atribuuti olid I_hold, Ä_hold ja Backspace_hold. Järgnesid S_Õ_Õ, Õ_seek ja V_A. Venelaste kolm enim eristumist mõjutanud atribuuti olid A_L_O, M_seek ja Õ_backspace. Järgnesid L_O_K, I_N ja Õ_Õ_backspace.

Tabel 15. IV vormiosa parima eristusvõimega atribuudid Simple Logistic mudeli järgi

Eesti				**Vene**			
1.	I_hold	**6.**	V_A	**1.**	A_L_O	**6.**	Õ_Õ_backspace
2.	Ä_hold	**7.**	Space_S_Õ	**2.**	M_seek	**7.**	Backspace_Ö_Ö
3.	Backspace_hold	**8.**	N_I	**3.**	Õ_backspace	**8.**	Space_seek
4.	S_Õ_Õ	**9.**	N_space_I	**4.**	L_O_K	**9.**	S_A
5.	Õ_seek	**10.**	V_seek	**5.**	I_N	**10.**	O_N

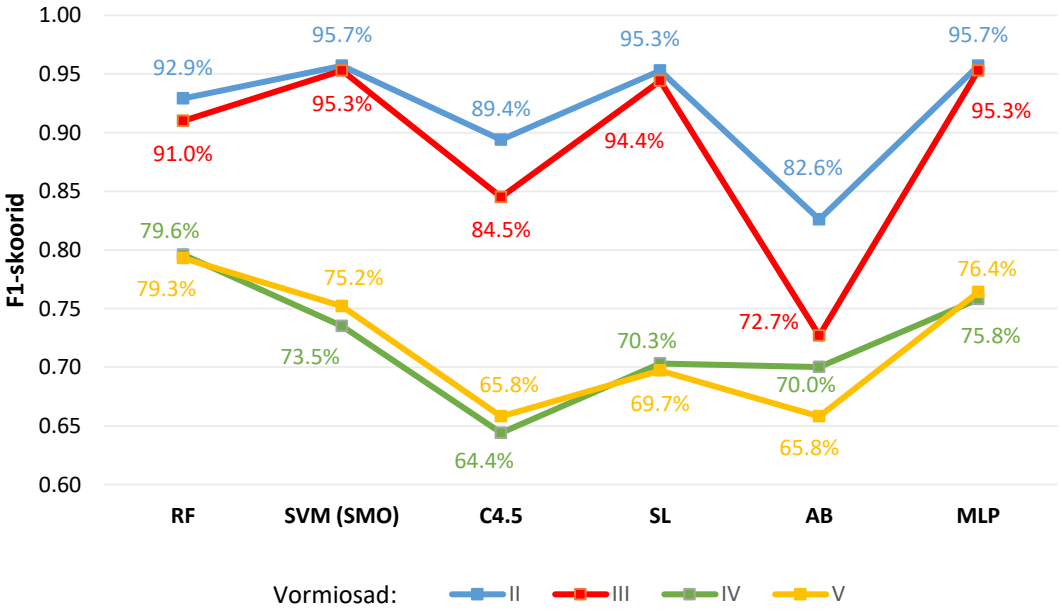
6 Olulisemate tulemuste analüüs ja järeldused

Viimases sisupeatükis tuuakse välja parimad ja olulisemad ennustustulemused. Kokkuvõtlikult esitletakse masinõppe algoritmide klassifitseerimise ennustustäpsused. Tehakse tulemuste põhjal järeldused, mis annavad vastused sissejuhatuses püstitatud uurimisteema küsimustele.

Klaviatuuri klahvivajutuse dünaamika järgi on võimalik ennustada lisaks varasemalt uuritud arvutikasutaja isikuomadustele nagu vanus, sugu, haridustase, emakeel ka rahvust. Käesolev magistritöö kinnitab, et rahvust saab ennustada. Andmed põhinevad Eestis elavate eesti ja vene keelt kõnelejate vormitäitmise trükkimiselt saadud klahvivajutustest, mille andmeanalüüsi tulemusena selgusid, kas vormitäitja oli rahvuselt eestlane või venelane. Kuigi eeldati, et inimesed on harjunud kasutama enda kõnekeelele vastavaid klaviatuure rohkem siis rahvus, kõnekeel ja klaviatuuril kirjutamise keel ei ole tingimata omavahel seotud. Näiteks kui Eestis sündinud, kogu elu Eestis elanud ja Eesti koolis õppinud vene kodukeelt kõnelev noor, kes peab ennast rahvuselt venelaseks, ei ole kokku puutunud vene klaviatuuril trükkimisega siis ei tea ta vene klaviatuuril olevate tähtede asukohti ja ei pruugi osata vene keeles arvutil trükkida. Sellisel juhul võib klassifitseerimisel teda eestlaseks pidada. Või näiteks Eestis elavad poolvene-pooleesti rahvusest inimesed (poluvernikud) võisid märkida rahvuseks eestlase aga klaviatuurikasutuse järgi klassifitseeriti ta venelaseks, kuna tema trükkimise muster oli rohkem vene klassile sarnane. Seesugused eeldused selles töös analüüsis kinnitust ei saanud. Oli 17 neid vastanuid, kes identifitseerisid ennast eestlasena aga kelle vene keeles arvutis kirjutamise oskuse hinnang oli parem kui eesti keeles ja neid oli 14, kes identifitseerisid ennast venelasena aga kelle eesti keeles arvutis kirjutamise oskuse hinnang oli parem kui vene keeles.

Parimad ennustustäpsused, mis antud töös leiti, olid teise ja kolmanda vormiosade klassifitseerimise tulemused F1-skooride järgi ulatudes 95,7%-ni. Joonisel 14. on graafiliselt esitatud kuue valitud masinõppe algoritmi nelja erineva vormiosa klassifitseerimiste F1-skooride kaalutud keskmiste tulemused.

II vormiosa ja III vormiosa olid etteantud eesti ja inglise keelsete tekstide kirjutamine. IV vormiosa ja V vormiosa olid vabas vormis pildi, perekonna ja tegevusala või töökoha kirjeldamine. Kuna etteantud teksti trükiti paremini kui vabas vormis oodatud vastuseid siis joonisel 14. on näha ka ennustustulemuste erinevus. Etteantud tekstide tulemused jäävad 72,7 - 95,7% vahele ja vabavormi tekstide omad 64,4 - 79,6% vahele. Vabas vormis oodati kahe- kuni kolmelauselisi vastuseid aga sisestatud vastused olid lühemad kui etteantud tekstidel ning vastajate omavahelises võrdluses olid vabatekstide vastused pikkuste pooles erinevad. Mõni vastas kirjelduste osades üksikute sõnadega või mõne sõnaga aga teised kirjutasid pikemaid vastuseid, mis koosnesid paarist kuni kolmest lihtlausest. Kuna vabavormide vastuste erinevuse tõttu on klahvivajutuste ühisosa enamike vormitäitjate suhtes väike ja esines ka puuduvaid väärtusi siis nendest vormiosadest sai andmeanalüüsiks kasutada väiksemat hulka atribuute. Antud juhul võib järeldada, et madalam ennustustulemus vabatekstide puhul on seotud vähesemate andmetega ja madalama andmete kvaliteediga.



Joonis 14. Vormiosade klassifitseerimiste F1-skooride kaalutud keskmiste tulemused

Veel eeldati, et ühed tähekombinatsioonid on klaviatuuril trükkides iseloomulikumad ühele rahvusele kui teisele, mis osutus õigeks. Näiteks eesti rahvusest vastajad kirjutasid klaviatuuril aeglasemalt bigramme nagu S_T, A_L, Y_O, R_E, A_D, N_I, D_I, U_S, I_L, V_A ja trigramme N_A_U, K_U_I, R_M_A, K_K_I, S_Õ_Õ, U_S_U, M_U_L, R_E_A, C_U_S, J_O_Y. Vene rahvusest vastajad kirjutasid kauem U_D, O_N, I_N, O_D, E_S,

S_A bigramme ja M_I_D, A_L_O, L_O_K, T_E_I, U_U_R, N_K_S trigramme. Eestlastel võttis rohkem aega ka U, A, I ja Ä tähtede ning parandusklahvide vajutamine ning Õ ja V tähtede, tagasivõtu, alumise klaviatuuri osa klahvide otsimine. Venelastel võttis rohkem aega sõnede vahelised pausid, V ja B tähtede vajutamine ning tühiku klahvi, M, L, N, K tähtede otsimine.

Kui vaadelda lisaks eesti ja vene klaviatuuridel erinevate asukohtadega tähti nagu A, E, K, M, L, N, O, T siis on näha, et venelased otsisid M, L, N, K klahve kauem ja trükkisid kauem ka M_I_D, tühik_S_A, A_L_O, T_E_I, E_S, tühik_B_E, S_A, O_D, O_N, N_K_S, K_S_tühik, L_O_K, I_N klahvi- või tähekombinatsioone, mis sisaldasid eelnevalt väljatoodud erineva asukohaga tähti. Sellest saab järeldada, et klahvidele, mis asuvad erikeelelistel klaviatuuridel teistes kohtades kulub leidmisele ja trükkimiseks rohkem aega.

Arvati ka, et eestikeelsete täpitähtede otsimine on vene rahvusest inimestel keerulisem leida, aga eristava tunnusena ei tulnud venelastel näiteks Ä täht välja, mis või olla tingitud ka sellest, et mõned vastanud kirjutatid Ä tähe A tähena. Õ ja Ö tähed esinesid oluliste atribuutidena vene rahvusest inimeste trükkimisel Õ_Õ_tagasivõtu ja tagasivõtu_Ö_Ö klahvikombinatsioonidena. Ä täht esines eestlaste ameti või tegevusala vormiosa kirjeldamises teise tähtsuselt eristava atribuudina. Eestlastel oli ka Õ tähe vajutamine oluline eristatav atribuut S_Õ_Õ, tühiku_S_Õ klahvikombinatsioonide trükkimisel ja Õ tähe otsimisel. Täielikult ei saa järeldada, et venelastel võttis kõigi eesti täpitähtede trükkimine kauem aega aga Õ ja Ö trigrammide esinemine näitab, et nende kirjutamine oli aeglasem.

Kuigi rahvuse ennustustulemused saadi etteantud tekstide puhul väga head siis vabatekstide uuringuosa võiks paremaks muuta. Vabavormi tekstilahtritele võiks määrata näiteks sümbolite või sõnade arvulise piirangu, et vastajad ei jätaks vastamata või ei kirjutaks ainult üksiksõnalisi vastuseid. Sel juhul oleks ka andmete sisestused eeldatavasti korrektsemad.

Üldiselt saab magistritöö tulemustega rahule jääda, aga teemat võiks põhjalikumalt ka edasi uurida ja teisi masinõppe algoritme rakendada.

7 Kokkuvõte

Käesoleva magistritöö eesmärgiks oli masinõppe meetodite abil ennustada klaviatuurikasutuse iseärasuste järgi arvutikasutajate rahvust, peamiselt Ida-Virumaal elavate eesti ja vene emakeelt kõnelejate näitel. Klaviatuuri klahvivajutuse dünaamika uurimine on aktuaalne küberkuritegevuse vähendamise või ennetamise seisukohalt. Näiteks identifitseerides ja profileerides isikuid või inimgruppe on võimalik ennustada, mis keelt võivad kurjategijad kõneleda ja kus nad võivad tegutseda. Kannatanute rahvuste kohta teades saab neile saata ennetatavat, hoiatavat või instruktiivset informatsiooni nende emakeeles.

Töö jaoks vajalike andmete kogumiseks loodi veebirakendusena viieosaline vorm, mis asub aadressil airika.midasminer.eu. Vorm salvestab trükkimisel klaviatuuri klahvivajutusi ja nende aegu. Kokku täitis vormi 544 inimest, millest lõppvalimisse jäi 336 sooritaja andmed. Masinõppe ja andmeanalüüsi programmi Weka-t kasutati andmete eeltöötluses filtreerimiseks ja andmeanalüüsi läbiviimiseks. Põhiliselt kuue valitud masinõppe algortimiga, milleks olid *Random Forest, Support Vector Machine (SMO), C4.5, Simple Logistic, AdaBoost ja Multilayer Perceptron* leiti mudelite ennustustäpsused.

Magistritöö eesmärk sai täidetud. Läbiviidud andmeanalüüsi ennustustulemused jäid etteantud tekstidel 72,7% - 95,7% vahele ja vabavormi tekstidel 64,4% - 79,6% vahele. Parimad ennustustulemused olid 95,7%, mis leiti *Support Vector Machine* (SVM(SMO)) ja *Multilayer Perceptron* (MLP) masinõppe algoritme kasutades. Veel leiti, et klaviatuurikasutusele olulisema eristumisvõimega rahvusele iseloomulikud tähekombinatsioonid olid N_A_U, S_T, K_K_I eesti rahvusele ja M_I_D, tühik_S_A, A_L_O vene rahvusele .

Töö tulemustest saab järeldada, et eri rahvustest inimesed kirjutavad klaviatuuril erinevalt ja neil on rahvusele omaseid klaviatuurikasutuse tähekombinatsioone. Antud töö ennustustulemusi ületas 64% piiri ja üldiselt võib nendega rahule jääda. Teisest küljest on klahvivajutuse kaudu rahvuse ennustamisel veel uurimis- ja arenguvõimalusi.

Paremate tulemuste saamiseks võiks uurida veel Venemaal elavate venelaste klaviatuuri klahvivajutuste dünaamikat. Võiks kaasata ka viie või enam erineva rahvusega inimesi. Lisaks di- ja trigrammidele oleks huvitav teada saada nelja või rohkema n-grammi esinemisi eri rahvustel.

Magistritöö tulemusi saab rakendada automaatsetes kontroll-, soovitussüsteemides või küberturvalisuses, kus anonüümsete isikuomaduste väljaselgitamine on oluline. Näiteks elektrooniliste keeletestide kontollimiseks, e-poodides rahvusele vastava keeleliste reklaamide või sisulehtede loomiseks, küberjuhtumite uurimisteks või analüüsi- ja autentimistarkvara välja töötamiseks.

Antud magistritööga kogutud andmeid saab kasutada õppetöö raames, näiteks masinõppe või andmekaeve loengutes ja praktikumides.

Summary

The aim of this master's thesis was to use machine learning methods to predict the nationality of computer users based on the characteristics of keyboard usage, mainly using Estonian and Russian native speakers living in Ida-Virumaa. The study of keyboard's keystroke dynamics is topical from the point of view of cybercrime reduction or prevention. For example, by identifying and profiling individuals or groups of people, it is possible to predict what language criminals might speak and where they might operate. Knowing the nationalities of the victims, preventive, deterrent or instructive information can be sent to them in their mother tongue.

To collect the data needed for the thesis, a five-part form was created as a web application, available at airika.midasminer.eu. The form records keyboard keystrokes and their timings as they are typed. A total of 544 people filled in the form, of which 336 respondents were in the final sample. Weka, a machine learning and data analysis program, was used to pre-process the data for filtering and data analysis. The prediction accuracies of the models were found using mainly six selected machine learning algorithm, which were Random Forest, Support Vector Machine (SMO), C4.5, Simple Logistic, AdaBoost and Multilayer Perceptron.

The objective of the master's thesis was met. The prediction results of the data analysis conducted ranged from 72.7% to 95.7% for the pre-written texts and from 64.4% to 79.6% for the free-form texts. The best prediction results were 95.7%, which were found using Support Vector Machine (SVM(SMO)) and Multilayer Perceptron (MLP) machine learning algorithms. Furthermore, it was found that the letter combinations with the most significant differentiation power for keyboard usage were N_A_U, S_T, K_K_I for the Estonian nationality and M_I_D, space_S_A, A_L_O for the Russian nationality.

From the results of the study, it can be concluded that people of different nationalities type differently on the keyboard and have nationality-specific keyboard letter combinations. The majority of the prediction results of this work exceeded the the 64% limit and one can generally be satisfied. On the other hand, there is still room for further research and development in the prediction of nationality through keystrokes. The

dynamics of keyboard keystrokes of Russians living in Russia could be further investigated for better results. Five or more different nationalities could also be included. In addition to di- and trigrams, it would be interesting to find out the occurrences of four or more n-grams in different nationalities.

The results of this thesis can be applied to automated verification, recommendation systems or cyber security, where the identification of anonymous personal characteristics is important. For example, for controlling electronic language tests, for creating nationality-appropriate linguistic advertisements or content pages in e-shops, for investigating cyber incidents, or for developing analysis and authentication software.

The data collected in this master's thesis can be used for learning purposes, for example in lectures and practical courses on machine learning or data mining.

Kasutatud allikad

[1] S. Paulus, "100 sekundi video: Digitaalne jalajälg ja privaatsus", Tartu Ülikool, ERR Novaator, 12.02.2016. [Online]. Available: https://novaator.err.ee/258569/100-sekundi-video-digitaalne-jalajalg-ja-privaatsus. [Accessed 20.01.2023].

[2] Cybernetica, "klahvirütm," Andmekaitse ja infoturbe leksikon, 2011-2023. [Online]. Available: https://akit.cyber.ee/term/7087-keystroke-dynamics. [Accessed 15.01.2023].

[3] Cybernetica, "biomeetrik," Andmekaitse ja infoturbe leksikon, 2011-2023. [Online]. Available: https://akit.cyber.ee/term/12. [Accessed 15.01.2023].

[4] Eesti Keele Instituut, "masinõpe," Sõnaveeb, 11.08.2022. [Online]. Available: https://sonaveeb.ee/search/unif/dlall/dsall/masin%C3%B5pe/1. [Accessed 15.01.2023].

[5] G. Auväärt, "Küberturvalisuse aastaraamat 2022," Riigi Infosüsteemi Amet, 15.02.2022. [Online]. Available: https://www.ria.ee/kuberturvalisus/kuberruumi-analuus-ja-ennetus/olukord-kuberruumis?view_instance=0&current_page=1#aastaraamatud. [Accessed 16.01.2023].

[6] P. Axbom, "Your unique typing rhythm can reveal your identity," 30.11.2020. [Online]. Available: https://axbom.com/keystroke-dynamics/. [Accessed 13.02.2023].

[7] G. Forsen, M. Nelson and J. R. Staron, "Personal attributes authentication techniques". Technical Report RADC-TR-77-333, Rome Air Development Center, 1977.

[8] R. S. Gaines, W. Lisowski, S. J. Press and N. Shapiro, "Authentication by keystroke timing: some preliminary results," The Rand Corporation, 1980. [Online]. Available: https://apps.dtic.mil/sti/pdfs/ADA484022.pdf. [Accessed 13.02.2023].

[9] D. Umphress and G. Williams, "Identity verification through keyboard characteristics," 25.04.1985. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0020737385800365#preview-section-abstract. [Accessed 30.01.2023.].

[10] J. Garcia, "Personal identification apparatus". USA Patent 4621334A, 4.11.1986.

[11] R. Joyce and G. Gupta, "Identity Authentication Based on Keystroke Latencies," Communications of the ACM, 1990. [Online]. Available: https://dl.acm.org/doi/pdf/10.1145/75577.75582. [Accessed 14.02.2023].

[12] K. S. Killourhy, "A Scientific Understanding of Keystroke Dynamics," 2012. [Online]. Available: http://reports-archive.adm.cs.cmu.edu/anon/2012/CMU-CS-12-100.pdf. [Accessed 14.02.2023].

[13] F. Monrose and A. D. Rubin, "Keystroke dynamics as a biometric for authentication," 24.01.2000. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0167739X9900059X. [Accessed 30.01.2023.].

[14] P. H. Pisani and A. C. Lorena, "A systematic review on keystroke dynamics," J Braz Comput Soc, 10.07.2013. [Online]. Available: https://journal-bcs.springeropen.com/articles/10.1007/s13173-013-0117-7. [Accessed 13.02.2023].

[15] A. Goodkind and A. Rosenberg, "Muddying The Multiword Expression Waters: How Cognitive Demand Affects Multiword Expression Production," 05.06.2015. [Online]. Available: https://aclanthology.org/W15-0914.pdf. [Accessed 02.03.2023].

[16] R. Giot and C. Rosenberger, "A New Soft Biometric Approach For Keystroke Dynamics Based On Gender Recognition," 2012. [Online]. Available: https://www.researchgate.net/publication/220490378_A_New_Soft_Biometric_Approach_For_Keystroke_Dynamics_Based_On_Gender_Recognition. [Accessed 15.02.2023].

[17] A. Pentel, "Predicting Age and Gender by Keystroke Dynamics and Mouse Patterns," 2017. [Online]. Available: https://www.researchgate.net/publication/317400387_Predicting_Age_and_Gender_by_Keystroke_Dynamics_and_Mouse_Patterns. [Accessed 30.12.2022].

[18] A. Pentel, "Predicting User Age by Keystroke Dynamics," 2019. [Online]. Available: https://www.researchgate.net/publication/325392442_Predicting_User_Age_by_Keystroke_Dynamics. [Accessed 30.12.2022].

[19] H. Vajak, "Põlvkondade interaktsioonide erinevused arvutite sisendseadmete kasutamisel," Tallinna Ülikool, 2019. [Online]. Available: https://www.etera.ee/zoom/59748/view? [Accessed 18.12.2022].

[20] D. Gunetti, C. Picardi and G. Ruffo, "Keystroke Analysis of Different Languages: A Case Study," 2005. [Online]. Available: https://www.researchgate.net/publication/221460678_Keystroke_Analysis_of_Different_Languages_A_Case_Study. [Accessed 15.02.2023].

[21] S. Bergsma, M. Post and D. Yarowsky, "Stylometric Analysis of Scientific Articles," 3-8. 2012. [Online]. Available: https://aclanthology.org/N12-1033.pdf. [Accessed 24.03.2023].

[22] D. G. Brizan, A. Goodkind, P. Koch, K. Balagani, V. V. Phoha and A. Rosenberg, "Utilizing linguistically enhanced keystroke dynamics to predict typist cognition and demographics," 2015. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S107158191500097X?via%3Dihub [Accessed 24.03.2023].

[23] P. Bours and S. Brahmanpally, "Language Dependent Challenge-based Keystroke," 23-26.10.2017. [Online]. Available: https://ieeexplore.ieee.org/document/8167838. [Accessed 25.03.2023].

[24] J. Edward, J. Leinonen and A. Hellas, "A Study of Keystroke Data in Two Contexts: Written Language and Programming Language Influence Predictability of Learning Outcomes". 2020. [Online]. Available: https://dl.acm.org/doi/10.1145/3328778.3366863. [Accessed 27.03.2023].

[25] A. Rõbakov, "Arvutikasutaja emakeele määramine tema klaviatuuri kasutusdünaamika järgi," Tallinna Ülikool. 2022. [Online]. Available: https://www.etera.ee/zoom/198757/view?. [Accessed 18.12.2022].

[26] E. Frank, "Oversampling and Undersampling," 30.01.2019. [Online]. Available: https://waikato.github.io/weka-blog/posts/2019-01-30-sampling/. [Accessed 30.12.2022].

[27] E. Frank, M. A. Hall and I. H.Witten, "The WEKA Workbench," Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques" Morgan Kaufmann, Fourth Edition, 2016. [Online]. Available: https://www.cs.waikato.ac.nz/ml/weka/Witten_et_al_2016_appendix.pdf [Accessed 30.12.2022].

[28] R. Kirkby, "Class RemoveUseless," [Online]. Available: https://weka.sourceforge.io/doc.dev/weka/filters/unsupervised/attribute/RemoveUseless.html. [Accessed 30.04.2023].

[29] J. Schafer, "Analysis of Incomplete Multivariate Data," New York: Chapman and Hall, 1997. [Online]. Available: https://weka.sourceforge.io/doc.packages/EMImputation/weka/filters/unsupervised/attribute/EMImputation.html. [Accessed 30.04.2023].

[30] javaTpoint, "Machine Learning Tutorial," 2011-2021. [Online]. Available: https://www.javatpoint.com/machine-learning. [Accessed 13. 02. - 30.04. 2023].

[31] R. Sirel, "Sissejuhatus masinõppesse," 06.10.2015. [Online]. Available: https://slideplayer.com/slide/14581461/. [Accessed 20.04.2023].

[32] International Business Machines Corporation (IBM), "What is machine learning?," 2023. [Online]. Available: https://www.ibm.com/topics/machine-learning. [Accessed 20.04.2023].

[33] K. Eljand, "Andmeteadus, Masinõpe ja tehisintellekt," 04.11.2021. [Online]. Available: https://abi.ria.ee/download/attachments/32407555/1.%20Andmeteadus%2C%20masin%C3%B5pe%20ja%20tehisintellekt%20%284%29.pdf?api=v2. [Accessed 20.04.2023].

[34] Waikato University, "Weka 3: Machine Learning Software in Java," [Online]. Available: Weka 3: Machine Learning Software in Java. [Accessed 20.12.2022].

[35] J. C. Platt, "Fast Training of Support Vector Machines using Sequential Minimal Optimization," Microsoft Research, 21.04.1998. [Online]. Available: https://www.microsoft.com/en-us/research/uploads/prod/1998/04/sequential-minimal-optimization.pdf. [Accessed 25.04.2023].

[36] S. Shevade, S. Keerthi, C. Bhattacharyya and K. Murthy, "Improvements to the SMO Algorithm for SVM Regression," 1999. [Online]. Available: https://www.researchgate.net/publication/3302831_Improvements_to_SMO_algorithm_for_SVM_regression. [Accessed 25.04.2023].

[37] F. R. Shamil, "C4.5 Algorithm in Data Mining," T4Tutorials, 2023. [Online]. Available: https://t4tutorials.com/c4-5-algorithm-in-data-mining/. [Accessed 24.04.2023].

[38] StatsTest.com, "Simple Logistic Regression," 2023. [Online]. Available: https://www.statstest.com/simple-logistic-regression/. [Accessed 26.04.2023].

[39] A. Saini, "Master the AdaBoost Algorithm: Guide to Implementing & Understanding AdaBoost," 26.04.2023. [Online]. Available:

https://www.analyticsvidhya.com/blog/2021/09/adaboost-algorithm-a-complete-guide-for-beginners/. [Accessed 30.04.2023].

[40] S. Lek and Y. Park, "Multilayer Perceptron," 06.08.2008. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/B9780080454054001622. [Accessed 30.04.2023].

[41] Y. Sasaki, "The truth of the F-measure," 26.10.2007. [Online]. Available: https://www.researchgate.net/publication/268185911_The_truth_of_the_F-measure. [Accessed 29.04.2023].

Lisa 1 – Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks[1]

Mina, Airika Andruse

1. Annan Tallinna Tehnikaülikoolile tasuta loa (lihtlitsentsi) enda loodud teose „Arvutikasutajate rahvuse ennustamine klaviatuurikasutuse iseärasuste järgi masinõppe meetodite abil“, mille juhendaja on Avar Pentel
   1.1. reprodutseerimiseks lõputöö säilitamise ja elektroonse avaldamise eesmärgil, sh Tallinna Tehnikaülikooli raamatukogu digikogusse lisamise eesmärgil kuni autoriõiguse kehtivuse tähtaja lõppemiseni;
   1.2. üldsusele kättesaadavaks tegemiseks Tallinna Tehnikaülikooli veebikeskkonna kaudu, sealhulgas Tallinna Tehnikaülikooli raamatukogu digikogu kaudu kuni autoriõiguse kehtivuse tähtaja lõppemiseni.
2. Olen teadlik, et käesoleva lihtlitsentsi punktis 1 nimetatud õigused jäävad alles ka autorile.
3. Kinnitan, et lihtlitsentsi andmisega ei rikuta teiste isikute intellektuaalomandi ega isikuandmete kaitse seadusest ning muudest õigusaktidest tulenevaid õigusi.

10.05.2023

1 Lihtlitsents ei kehti juurdepääsupiirangu kehtivuse ajal vastavalt üliõpilase taotlusele lõputööle juurdepääsupiirangu kehtestamiseks, mis on allkirjastatud teaduskonna dekaani poolt, välja arvatud ülikooli õigus lõputööd reprodutseerida üksnes säilitamise eesmärgil. Kui lõputöö on loonud kaks või enam isikut oma ühise loomingulise tegevusega ning lõputöö kaas- või ühisautor(id) ei ole andnud lõputööd kaitsvale üliõpilasele kindlaksmääratud tähtajaks nõusolekut lõputöö reprodutseerimiseks ja avalikustamiseks vastavalt lihtlitsentsi punktidele 1.1. ja 1.2, siis lihtlitsents nimetatud tähtaja jooksul ei kehti.

Lisa 2 – Pöördumine / e-kiri vormi jagamisel

Tere!

Minu nimi on Airika Andruse, olen Talllinna Tehnikaülikooli (TalTech) äriinfotehnoloogia eriala magistrant ning kirjutan magistritööd teemal „Arvutikasutajate rahvuse ennustamine klaviatuurikasutuse iseärasuste järgi“.

Kogun magistritöö jaoks vajalikud andmed testi kaudu, kus inimene täidab klaviatuuril testis etteantud väljad. Selleks palun teil täita allolev test, mis on anonüümne. Test koosneb viiest osast, selles tuleb määrata enda sugu, vanus ja rahvus, hinnata enda arvutis kirjutamise oskusi, trükkida etteantud tekstid, kirjeldada pildil olevat tegevust ning vastata kahele küsimusele.

NB! Palun testi täita vastavalt ettekirjutatud lausele, kindlasti arvuti või sülearvutiga (mitte telefoni või tahvelarvutiga), sest edasiseks analüüsiks on vajalik kasutada arvuti või sülearvuti klaviatuuril sisestatud andmeid.

Veel palun testis kirjutatavaid tekste mitte kopeerida (copy) ja kleepida (paste), vaid klaviatuuril trükkida.

Testi täitmine võtab umbes 5-10 minutit.
Test on aadressil: airika.midasminer.eu

Tänan ette koostöö eest
Airika Andruse
Äriinfotehnoloogia
TalTech