TALLINNA TEHNIKAÜLIKOOL

Infotehnoloogia teaduskond

Andra Rajaste 222570IAIB

Caroly Märtson 223356IAIB

Haridusandmete analüüs ja integreerimine koolijuhtimise toetamiseks: prototüübi arendus kahe kooli näitel

Bakalaureusetöö

Juhendaja: Ago Luberg

PhD

Tallinn 2025

Autorideklaratsioon

Kinnitame, et oleme 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.

Autorid: Andra Rajaste ja Caroly Märtson

04.06.2025

Annotatsioon

Käesoleva bakalaureusetöö eesmärk on kaardistada üldhariduskoolide vajadused ning töötada välja algeline töölaua prototüüp, mis aitaks koolijuhtidel teha andmepõhised otsuseid, parandada õppeedukust ja õpilaste heaolu. Töö käigus tehakse koostööd Saku Gümnaasiumi ja Mustamäe Riigigümnaasiumiga, et välja selgitada, millised andmed neil olemas on, kuidas neid praegu kogutakse ja millised on takistused nende kasutamiseks. Kuna automatiseeritud süsteemi polnud võimalik luua, koguti andmed koolilt käsitsi. Andmekogumeid analüüsitakse, puhastatakse ja töödeldakse, et luua visualiseeringuid ja seoseid, mis aitaksid kooli juhtkonnal õigeaegselt märgata mustreid, näiteks puudumiste ja õppeedukuse vahelist seost.

Töö tulemusel valmib algeline prototüüp ühe kooli andmete põhjal, kus visualiseeritakse seoseid erinevate andmetüüpide vahel, näiteks puudumised, hinded ja riiklikud rahulolu-küsitluste tulemused. Töö ei paku lõplikku lahendust, kuid loob tugeva aluse edasiseks arendustööks.

Lõputöö on kirjutatud eesti keeles ning sisaldab teksti 47 leheküljel, 7 peatükki, 4 joonist.

Abstract

Analysis and Integration of Educational Data to Support School Management: Prototype Development Based on Two Case Schools

Schools often lack a comprehensive and systematic overview of student absences, grades, and their connections to other factors that influence student well-being and academic performance. The data is usually spread across multiple platforms, and in order to analyze it and make informed decisions, it must be manually processed, which takes a lot of time and effort. Additionally, one major obstacle to building an automated system is identified. The information systems used in schools currently do not allow for the automatic export of student-related data like grades and absences.

The aim of this bachelor thesis is to map the needs of schools and develop a basic dashboard prototype that would help school leaders make data-driven decisions and improve both academic outcomes and student well-being. As part of the work, two schools, Saku Gümnaasium and Mustamäe Riigigümnaasium, were involved to better understand what kind of data they currently have, how it is collected, and what the barriers are to using it effectively. Since it was not possible to create an automated system at this stage, the data was manually collected from one school. These datasets were then analyzed, cleaned, and processed to create visualizations and reveal connections that could help the school leadership detect patterns in time, such as links between absences and academic performance.

As a result of the work, a basic prototype was created using one school's data, visualizing relationships between different types of information, such as absences, grades, and national satisfaction survey results. While the solution presented in this thesis is not final, it lays a solid foundation for further development.

The thesis is in Estonian and contains 47 pages of text, 7 chapters, 4 figures.

Lühendite ja mõistete sõnastik

API	Rakendusliides (*Application Programming Interface*)
CSV	Komaga eraldatud väärtused (*Comma-Separated Values*)
DPA	Andmekaitse järelevalveasutus (*Data Protection Authority*)
DPIA	Andmekaitsealane mõjuhinnang (*Data Protection Impact Assessment*)
DPO	Andmekaitseinspektor (*Data Protection Officer*)
EHIS	Eesti hariduse infosüsteem (*Estonian Education Information System*)
EL	Euroopa Liit (*European Union*)
GDPR	Üldine andmekaitse määrus (*General Data Protection Regulation*)
Harno	Haridus- ja Noorteamet (*Education and Youth Board*)
IKS	Isikuandmete kaitse seadus (*Personal Data Protection Act*)
IT	Infotehnoloogia (*Information Technology*)
JSON	Lihtsustatud andmevahetusvorming (*JavaScript Object Notation*)
KiVa	Kiusamisvastane programm (*Antibullying Program*)
LDAP	Kergprotokoll kataloogipöördusteks (*Lightweight Directory Access Protocol*)
MURG	Mustamäe riigigümnaasium (*Mustamäe State Gymnasium*)
NDJSON	Reajoondusega JSON (*Newline Delimited JSON*)
PCA	Põhikomponentide analüüs (*Principal Component Analysis*)
PDF	Kantav dokumendiformaat (*Portable Document Format*)
PISA	Rahvusvaheline õpilaste hindamise programm (*Programme for International Student Assessment*)
REST	Esitatava oleku edastus (*Representational State Transfer*)
ROU	Rahulolu-uuring (*Satisfaction Survey*)
SAML	Turvaline kinnitamise märgistuskeel (*Security Assertion Markup Language*)
SQL	Struktureeritud päringukeel (*Structured Query Language*)
TLS	Transpordikihi turvalisus (*Transport Layer Security*)
TLÜ	Tallinna Ülikool (*Tallinn University*)
USA	Ameerika Ühendriigid (*United States of America*)
XLS	Exceli tabelifail (*Excel Spreadsheet*)

Sisukord

Jooniste loetelu

1 Sissejuhatus

Hea õppeedukus ja õpilaste heaolu kooliga käivad käsikäes ning nende tasakaalus hoidmine on kooli oluline ülesanne. Selleks koguvad koolid mitmesuguseid andmeid - osaletakse riiklikes rahuloluküsitlustes, viiakse läbi koolisiseseid uuringuid ning kogutakse igapäevast infot hindamise ja puudumiste kohta. Andmeid koguneb palju ning need on sageli hajutatud erinevatesse infosüsteemidesse. Puudub terviklik ja süsteemne ülevaade, mis koondaks need erinevad andmeallikad ühte keskkonda ja võimaldaks analüüsida ja visualiseerida nende omavahelisi seoseid.

Sellise lahenduse puudumine piirab koolil õigeaegselt märgata mustreid, mis võivad viidata probleemidele, ning läbi selle pakkuda õpilastele rohkem tuge ja parandada õppeedukust. Lisaks sellele koostatakse statistikat käsitsi, näiteks Exceli abil. Kuna andmed on kõik laiali, siis nende manuaalselt kokku kogumine, puhastamine ning analüüsimine võtab palju aega. Koondanalüüse koostatakse peamiselt kord aastas, mis vähendab võimalust reaalajas jälgida õpilaste arengut ja tuvastada varakult probleeme.

Lisaks on andmete automaatne kogumine piiratud, kuna olemasolevad infosüsteemid, nagu Stuudium ja eKool, ei paku avatud liidest, mille kaudu saaks korraga kõiki andmeid automaatselt alla laadida ja analüüsiks kasutada. Sageli on andmete eksport võimalik vaid õpilase või klassi kaupa.

Käesoleva bakalaureusetöö eesmärk on kaardistada kahe üldhariduskooli andmekasutuse hetkeseis ja vajadused ning töötada välja algeline prototüüp visuaalsest töölauast, mis toetaks koolijuhte otsuste tegemisel. Eesmärk on luua ka põhjalik teoreetiline alus, et seda saaks tulevikus edasi arendada, kuna selle lõputöö raames valmib algeline prototüüp vaid ühe kooli näitel. Töö keskendub peamiselt sellele, milliseid andmeid koolid koguvad, kuidas neid kasutatakse ning milliseid seoseid oleks võimalik nende vahel luua. Uuritakse, kuidas andmete analüüs ja visualiseerimine võiksid toetada kooli juhtimist ja ennetavat sekkumist.

Töö raames tehakse koostööd kahe üldhariduskooliga, Saku Gümnaasiumi ja Mustamäe Riigigümnaasiumiga. Kogutakse ja uuritakse olemasolevaid andmeid ja nende struktuure. Andmete põhjal viiakse läbi analüüs ning luuakse seoseid, mida visualiseeritakse avatud lähtekoodiga tarkvara OpenSearch abil.

Bakalaureuse töö teises peatükis antakse ülevaade töö taustast, sealhulgas haridusandmete liikidest, Eesti koolides kasutatavatest infosüsteemidest ning kuidas toimub andmeanalüüs kahes üldhariduskoolis. Kolmandas peatükis antakse ülevaade vajaduste kaardistamise protsessist, mis hõlmab koolide aasta kokkuvõtete analüüsi ning kohtumisi koolide ja õppeinfosüsteemi pakkujatega. Neljandas peatükis kirjeldatakse andmeanalüüsi protsessi ning saadud tulemusi. Viiendas peatükis analüüsitakse alternatiivseid lahendusi, tutvustatakse valminud algelist töölaua prototüüpi ning antakse ülevaade valideerimise tulemustest. Kokkuvõttes analüüsitakse saavutatud tulemusi ja tuuakse välja töö piirangud ning edasised arendusvõimalused.

2 Taust

Antud peatükis tutvustatakse töös kasutatavaid haridusandmete liike ja nende kasutusvõimalusi. Seejärel antakse ülevaate Eesti üldhariduskoolides kasutatavatest infosüsteemidest ning kuidas andmeanalüüs kahe üldhariduskooli näitel hariduses toimub. Eesmärgiks on anda ülevaade taustast, et töö sisu oleks selgemini mõistetav.

2.1 Haridusandmete liigid ja kasutusvõimalused

Koolides kogutakse mitmekesiseid andmeid, millel on suur potentsiaal hariduse juhtimisel, enesehindamisel ning õpilaste arengu toetamisel. Haridusandmete alla kuuluvad muuhulgas õpitulemused, puudumised ning rahulolu-uuringute tulemused. Järgnevalt käsitletakse olulisemaid andmetüüpe ja nende kasutusvõimalusi.

2.1.1 Õpitulemused

Õpitulemused on koolides igapäevaselt kogutavad andmed, mis kajastavad õpilaste akadeemilist arengut ja käitumist. Neid kogutakse peamiselt koolide infosüsteemide kaudu, Eestis on levinumad platvormid Stuudium ja eKool. Peamised õpitulemusi kajastavad andmed on:

- Hinded, mis antakse viiepallisüsteemis või koolipõhiste alternatiivsete hindamissüsteemide alusel;
- Puudumiste info, mis hõlmab nii põhjendatud kui ka põhjendamata puudumisi;
- Käitumise info: märkused ja tähelepanekud õpilaste käitumise kohta.

Eestis rakendatakse enamasti kahte tüüpi hindamist: kujundav ja kokkuvõttev. Kujundav hindamine keskendub õpilase arengu toetamisele, andes suulist ja kirjalikku tagasisidet õpitulemuste ja õppeprotsessi kohta. See aitab kaasa õpilase enesehindamise ja eesmärgistamise oskuse arengule. Kokkuvõttev hindamine hõlmab hinnete koondamist trimestri-, poolaasta- või aasta hinneteks ning kasutatakse sageli näiteks järgmiseks klassiks üleviimisel või kooli lõpetamisel. Hinnete andmisel arvestatakse õpilase tulemuste vastavust

õppekavas toodud õpitulemustele [1].

Õpitulemuste andmete põhjal saavad koolijuhid ja õpetajad:

- jälgida õpilaste arengut ja tuvastada abivajajaid;
- hinnata, kas puudumised mõjutavad negatiivselt õpitulemusi;
- planeerida õppe sisu ja tugiteenuseid;
- anda tagasisidet õpilastele ja lapsevanematele;
- luua sisendeid kooli enesehindamiseks ja arengukava koostamiseks;

2.1.2 Rahulolu kooliga

Koolikeskkond on üks olulistest teguritest, mis mõjutab otseselt õpilaste õpitulemusi, heaolu ja motivatsiooni. Koolid koguvad koolikeskkonna ja rahulolu kohta andmeid, et mõista paremini, millised tegurid toetavad või takistavad õppimist. Selleks korraldab Harno igal kevadel riiklike rahuloluküsitlusi, mis võimaldavad koguda tagasisidet õpilastelt, õpetajatelt ja lapsevanematelt [2]. Need andmed on koolidele ja omavalitsustele väärtuslikud tööriistad, et hinnata hariduse kvaliteeti ning kavandada arendustegevusi. Vastavalt koolile korraldatakse ka erinevaid koolisiseseid rahuloluküsitlusi.

Euroopa Komisjoni 2024. aasta haridusvaldkonna ülevaates tuuakse välja, et kooli kuuluvustunne ja turvaline õpikeskkond on seotud paremate õpitulemustega, samas kui koolikiusamine mõjub õpitulemustele negatiivselt. Eestis esineb küll koolikiusamist võrdlemisi sageli, kuid samas on paljudel õpilastel suur kuuluvustunne, mis tasakaalustab negatiivseid mõjusid. Samuti toetavad Eesti koolid üldiselt ennastjuhtivat õppimist, mis toetab loovat mõtlemist, milles Eesti õpilased on saavutanud häid tulemusi Euroopaga võrreldes [3].

Rahulolu- ja koolikeskkonna andmed võimaldavad seega hinnata koolikeskkonda, sotsiaalseid suhteid, kiusamise esinemissagedust ning õppimistingimusi. See on oluline sisend kooli juhtimisele, õpetamise kvaliteedi parendamisele ning ennetustöö kavandamisele.

2.2 Infosüsteemid Eesti koolides

Olenevalt koolist on Eesti üldhariduskoolides kasutusel igapäevaselt peamiselt kaks õppeinfosüsteemi: Stuudium ja eKool [4], [5]. Nende peamine eesmärk on toetada koolide igapäevast õppetööd. Mõlemad platvormid võimaldavad õpetajatel hallata õppetööd - sisestada hindeid, kodutöid, puudumisi, suhelda lapsevanemate ja õpilastega ning dokumenteerida õppimisega seotud protsesse. Samuti võimaldavad need süsteemid juhtkonnal jälgida kooli toimimist ja teha juhtimisotsuseid olemasolevale andmestikule tuginedes. Antud peatükis antakse ülevaade Stuudiumist ja eKoolist, nende kasutusvõimalustest ja piirangutest, sealhulgas andmekasutuse seisukohalt.

2.2.1 Stuudium

Stuudium on toetab õpetajate, juhtkonna, õpilaste ja lapsevanemate igapäevast suhtlust ja infohaldust. Õpetajad saavad ise koostada päevikuid vastavalt ainele või klassile ning sisestada sinna hindeid, puudumisi, kodutöid ja tunnikirjeldusi. Kontrolltööd koondatakse ühisesse kalendrisse, mis võimaldab koolisiseselt töid paremini planeerida.

Süsteemist leiab õpilaste ja lastevanemate kontaktid ning nende aktiivsuse jälgimise tööriistad. Õpilaste hinnetest ja hinnangutest koostatakse automaatselt tunnistused ja õpilasraamatud, mille saab edastada EHISesse. Suhtlus toimub Stuudiumi enda keskkonnas, kus saab saata sõnumeid ja jagada sündmusi vastavalt sihtgrupile. Õpilastel ja vanematel on mugav ülevaade kõigist kooliga seotud tegevustest – tundidest ja hinnetest kuni puudumiste ja arenguvestlusteni.

Koolid saavad Stuudiumis koostada ja hallata aine- ja töökavasid ning õppematerjale, mida saab jagada õpilaste või kolleegidega. Süsteemis on tööriistad erinevate dokumentide (nt individuaalse õppekava, arenguvestluste) salvestamiseks, avalduste ja registreerimiste kogumiseks ning rahuloluküsitluste läbiviimiseks. Samuti pakub Stuudium ülevaatlikku statistikat õppeedukuse ja puudumiste kohta.

Lisafunktsioonidena saab Stuudiumis ehitada veebilehti, koostada ülesannete nimekirju ning siduda kasutajakontod Google Workspace'i teenustega.

2.2.2 eKool

Sarnaselt Stuudiumile pakub ühendab ka eKool erinevaid funktsioone – hinnete, puudumiste, kodutööde, õppematerjalide ning tunniplaanide haldus, pakkudes kõigile osapooltele ajakohast ülevaadet ja võimaldades kiiret infovahetust. Õpilased ja vanemad saavad süsteemi kaudu reaalajas teavitusi oluliste muudatuste ja tulemuste kohta.

Lisaks õppeinfo haldamisele pakub eKool tuge ka sotsiaalse toe ja õpilaste heaolu jälgimisel. Platvormi kaudu saab jälgida koolikohustuse täitmist ning tuvastada murelapsi – näiteks õpilasi, kelle puudumiste hulk viitab võimalikule riskile. Neile määratakse süsteemis vastutav tugispetsialist, kes saab jälgida sekkumismeetmeid ning teha koostööd koolipersonaliga. Samuti võimaldab eKool integreerida API liidestuse abil mitmeid haridusrakendusi.

2.2.3 Infosüsteemide andmekasutus

Andmekasutuse seisukohalt koondavad Stuudium ja eKool suures mahus andmeid õpilaste õpitulemuste, kohaloleku, suhtluse ja tagasiside kohta. Siiski on andmete kättesaadavus ja kasutatavus suuremas analüütilises kontekstis piiratud, kuna puuduvad paindlikud ekspordivõimalused või avatud liidestused (API-d) andmete süsteemseks kasutamiseks väljaspool platvormi. eKooli puhul on küll võimalus integreerida API abil mitmeid haridusrakendusi, kuid puudub võimalus automaatselt andmeid kätte saada ja kasutada. See muudab keeruliseks koolide ja kohalike omavalitsuste jaoks andmepõhiste otsuste tegemise või omaenda andmete põhjal süstemaatilise analüüsi läbiviimise.

Kokkuvõtlikult täidavad Stuudium ja eKool hästi oma esmaseid funktsioone igapäevase õppetöö toetamisel, kuid andmepõhise juhtimise ja laiapõhjalise analüüsi võimalused on nende platvormide puhul piiratud.

2.3 Andmeanalüüs hariduses kahe kooli näitel

Koostöös kahe üldhariduskooliga, Saku Gümnaasiumi ja MURG-iga, viidi läbi olukorra kaardistamise, et mõista, kuidas haridusandmeid praegu kogutakse, hallatakse ja analüüsitakse. Eesmärk oli saada aimu sellest, millised on koolide andmekogumise ja -analüüsi praktikad ning millised vajadused on automatiseeritud tööriistade järele. Järgnev peatükk annab ülevaate Saku Gümnaasiumi ja Mustamäe Riigigümnaasiumi kogutavatest andmetest

ja nende analüüsimise praktikatest ning võrreldakse kahte kooli omavahel.

2.3.1 Saku Gümnaasium

Saku Gümnaasiumis on andmete kogumine mitmekesine, kuid killustatud. Erinevaid mõõdikuid ja näitajaid talletatakse Exceli tabelites, kus igal kooli töötajal võib olla vastutus konkreetse valdkonna andmete kogumise eest. Kuigi andmeid kogutakse regulaarselt ja süstemaatiliselt, puudub nende koondamisel ja analüüsimisel automatiseeritud lahendus. Kogu andmetöötlus toimub käsitsi: koostatakse koondfaile, filtreeritakse andmeid, arvutatakse keskmisi ja protsente ning luuakse graafikuid ja järeldusi.

Kogutav andmestik hõlmab:

- Tasemetööd ja testitulemused: erinevates ainetes, näiteks 4. ja 7. klassi loodusõpetus, matemaatika, eesti keel teise keelena, HARNO üldpädevustestid.
- Küsitlused:
  - Kooli ROU - õpilased, õpetajad, lapsevanemad;
  - KiVa küsitlused - õpilased, õpetajad, klassijuhatajad;
  - Ennastjuhtiva õppija enesehindamised - gümnaasiumiõpilased;
  - EU Kids Online - seosed digikasutuse ja heaolu vahel;
  - Koolilõuna uuring - söömisharjumused ja rahulolu toitlustamisega.
- Rahvusvahelised uuringud ja testid: PISA, ROU, EU Kids Online jms.
- Õpetajate ja juhtkonna uuringud: näiteks TLÜ õpetajauurimus ja juhtide uuring.
- luua sisendeid kooli enesehindamiseks ja arengukava koostamiseks;

Lisaks kogutakse andmeid Haridussilmast – Haridus- ja Teadusministeeriumi hallatavast andmeportaalist, mis koondab statistikat Eesti haridussüsteemi kohta, näiteks eksamitulemused, lõpetajate arv, õpetajate kvalifikatsioon jms [1]. Sealt saab andmeid eksportida CSV-vormingus failidena, näiteks eksamitulemused ja lõpetajate statistika. Koolil on olemas ka tulemusnäitajate koondfail, kus jälgitakse arengukava täitmist läbi konkreetsete mõõdikute.

Kõik see info koondub sageli kokkuvõttefailidesse iga kooliastme kohta, mis on avalikud ja jagatakse tihti juhtkonna ning õpetajate vahel. Kuid kogu see protsess, alates andmete kogumisest kuni statistika ja graafikute loomiseni, toimub manuaalselt, mis on väga

aeganõudev.

2.3.2 Mustamäe Riigigümnaasium

MURG kasutab igapäevaselt eKooli kui peamist õppeinfosüsteemi, mille kaudu jälgitakse jooksvalt õpilaste akadeemilist edukust, eelkõige kursuse mitteläbimisi. Süsteemist saadakse igapäevaselt infot hinnete ja puudumiste kohta, kuid selle analüütiline kasutamine on piiratud. Aasta lõpus koostatakse õppenõukogu protokoll, kuhu kantakse kõik läbikukkunud õpilased. See nõuab täpset andmete läbivaatamist ja käsitsi koondamist. Iga juhtum analüüsitakse individuaalselt, et teha otsuseid ja pakkuda vajadusel tuge.

Kooli üldtööplaan ja sellega seotud andmeanalüüs koostatakse kord aastas käsitsi Excelis. Selleks kasutatakse erinevaid andmeallikaid, näiteks HARNO rahuloluküsitlusi ja Haridussilm.ee portaali andmeid. Aastas korra koostatav ülevaade on kooli juhtimisprotsessi oluline osa, kuid kuna andmeid saadakse välisallikatest vaid kord aastas, ei võimalda see järjepidevat seiret.

eKooli igapäevane andmestik on kooli jaoks väärtuslik, kuid selle kasutamine nõuab manuaalset tööd. Statistika tuleb eraldi alla laadida klasside kaupa .xls failidena, puudub võimalus näha ülekoolilisi koondvaateid. Näiteks ei ole võimalik kiiresti vaadata kooliüleseid trende õppeedukuse või puudumiste kohta.

Kuna eKool ei paku sisulisi analüütilisi tööriistu, tuleb igat õpilast vaadelda eraldi, sageli koostöös õppenõustajaga. Analüüs ei ole pelgalt tehniline, see on igapäevane arendustöö, mille kaudu püütakse mõista, miks tekivad raskused, ning kuidas toetada õpilaste arengut.

2.3.3 Koolide analüüsipraktikate võrdlus

Kuigi Saku Gümnaasium ja MURG kasutavad erinevaid õppeinfosüsteeme, vastavalt Stuudiumit ja eKooli, on nende andmetöötluse ja -analüüsi praktikad sisuliselt väga sarnased. Mõlemad koolid koguvad andmeid erinevatest allikatest ning viivad läbi analüüse, mille tulemusena koostatakse kokkuvõtteid ja järeldusi õppeprotsessi juhtimiseks.

Andmed koondatakse mõlemas koolis Exceli tabelitesse, kus toimub käsitsi töötlus: andmete filtreerimine, arvutuste tegemine ning järelduste sõnastamine. Kuigi andmestikku kogutakse palju ja selle põhjal tehakse ka olulisi juhtimisotsuseid (nt õpilaste toetamine, õppetöö

kvaliteedi hindamine, arengukava seire), puudub automatiseeritud andmeanalüüsi süsteem, mis võimaldaks andmetega tõhusalt ja regulaarselt töötada.

Andmete regulaarne analüüs kooli tasandil toimub mõlemas koolis põhiliselt kord aastas, mis on seotud kooli üldtööplaani koostamisega ja sõltub välisandmete (nt HARNO, Haridussilm) kättesaadavusest. Igapäevaselt kasutatavate infosüsteemide (eKool, Stuudium) andmed (hinded, puudumised) on küll olemas, kuid puuduvad koondväljavõtted ning sageli tuleb andmed alla laadida klasside või õpilaste kaupa eraldi failidena.

Seega võib järeldada, et kuigi andmetel põhinev otsustamine on mõlemas koolis tähtsal kohal, on tehnilised vahendid selleks piiratud ning vajadus automatiseeritud, kasutajasõbraliku ja koolikohase analüüsisüsteemi järele on selgelt olemas.

3 Vajaduste kaardistamine

Järgnev peatükk annab ülevaate sammudest, mida tehti, et kaardistada koolide vajadused. Tegevustena analüüsiti koolide aastakokkuvõtteid ja arengukavasid, kohtuti koolide esindajatega ning infosüsteemi pakkujatega.

3.1 Koolide aastakokkuvõtete ja arengukavade analüüs

Selle alapeatüki eesmärk on uurida, milliseid haridusandmeid koguvad koostöökoolid Saku Gümnaasium ja Mustamäe Riigigümnaasium ning kuidas nende dokumenteeritud tulemusi ja eesmärke saab rakendada andmeanalüütilise süsteemi loomisel. Analüüsiti MURGi arengukava ja Saku Gümnaasiumi õppeaastate kokkuvõtteid, et kaardistada olemasolevad mõõdikud, arenguvaldkonnad ja järeldused, mida koolid ise oma dokumentides teevad.

3.1.1 Mustamäe Riigigümnaasiumi arengukava analüüs

MURGi arengukava (2024–2027) [6] sisaldab mitmeid selgelt määratletud mõõdikuid, mille põhjal hinnatakse kooli arengu edukust. Andmepõhine juhtimine on arengukavas küll olemas, kuid selle rakendamine toimub pigem kord aastas, kui tehakse kokkuvõtteid riiklikest ja koolisisestest uuringutest.

Olulised mõõdikud ja näitajad arengukavas:

- **Õpitulemused ja edasijõudmine**: jälgitakse kursuse mitteläbimisi ning püütakse vähendada täiendavale õppele jäävate õpilaste arvu. See eeldab andmepõhist õpilaste kaardistamist ja toetusmeetmete rakendamist;
- **Puudumiste arv**: kolmandas kooliastmes on seatud eesmärgiks, et põhjuseta puudujaid oleks alla 10%. Arengukavas tuuakse välja, et 2022. a kevadel oli selliseid 31% õpilastest. See viitab puudumiste ja õpitulemuste seose olulisusele;
- **Rahuloluküsitluste näitajad**: arengukavas tuuakse välja nii kooliga rahulolu, enesetõhusus, seotus, eneseregulatsioon kui ka negatiivsed indikaatorid nagu küünilisus ja

kurnatus. Neid mõõdetakse Likerti skaalal ning jälgitakse trende kooliastmete lõikes;

- **Ennastjuhtiv õppimine ja eneseregulatsioon**: on keskne strateegiline eesmärk, mida mõõdetakse mitme eri meetodiga, sh riiklike üldpädevustestide ja koolisiseste testidega;
- **Tulemuslikkus gümnaasiumi lõpus**: üheks võtmenäitajaks on gümnasistide edasine haridustee – sihiks on, et vähemalt 70% lõpetanutest jätkaks õpinguid.

Nendest eesmärkidest saab välja tuua rea andmepunkte, mida tulevane infosüsteem võiks pidevalt jälgida ja visualiseerida: puudumiste trendid, kursuse läbivus, eneseregulatsiooni testide tulemused, rahulolu erinevates kategooriates ning lõpetajate edasine haridustee.

3.1.2 Saku Gümnaasiumi aastakokkuvõtete ja arengukava analüüs

Saku Gümnaasiumis koostatakse igal aastal mahukad õppeaasta kokkuvõtted [7], mille alusel hinnatakse õppeedukust, osalemist üritustel ja arenguvajadusi. Lisaks on olemas arengukava (2023–2027) [8], mis seab eesmärgid ja mõõdikud õppija iseseisvuse, eneseregulatsiooni ja õpimotivatsiooni arendamiseks.

Aastakokkuvõtete sisu:

- **Õppeedukuse näitajad**: tuuakse välja klassigruppide kaupa hinnete jaotus, näiteks mitu õpilast sai hindeks 5. Lisaks õppeedukuse protsent, arvestuslike tööde tulemused jne;
- **Riiklike ja koolisiseste testide tulemused**: tasemetööde ja riigieksamite tulemused on esitatud tabelitena, vahel koos võrdlusega eelmiste aastatega ja üleriigilise keskmisega;
- **Arenguvestluste andmed**: mitu vestlust toimus, millist tagasisidet anti ja millised olid üldised tähelepanekud;
- **Üritused ja saavutused**: mitmed leheküljed on pühendatud spordi- ja ainealaste võistluste tulemustele ning kooli esindamisele erinevatel üritustel;
- **Järgmiseks aastaks seatud eesmärgid**: sh konkreetsete meetodite või teemade fookusesse võtmine (nt kujundav hindamine, individuaalne õpirada).

Arengukava mõõdikud:

- **Õpilaste eneseregulatsioon**: mõõdetakse Harno testidega ja TLÜ abil loodud koolisisese vahendiga;
- **Rahulolu ja koolikeskkond**: igal aastal viiakse läbi riiklikud rahuloluküsitlused (4., 8., 11. klassid), mille tulemused on arengukavas aluseks tegevuste kavandamisel;
- **Puudumised ja järelõpe**: samuti jälgitakse, kui suur on põhjendamata puudumiste arv ja täiendavale õppele jäänute osakaal.

Sarnaselt MURGile saab ka Saku näidete põhjal tuletada, et koolid juba koguvad hulgaliselt andmeid, mille analüütiline kasutus võiks oluliselt kasvada, kui need oleks automatiseeritult kättesaadavad ja visuaalselt esitatud.

3.1.3 Kokkuvõte: ühisjooned ja rakendusvõimalused

Analüüsitud dokumentidest joonistuvad välja mitmed ühised sisulised fookused, mis on otseselt rakendatavad meie loodava süsteemi arenduses. Mõlema kooli puhul korduvad järgmised põhinäitajad ja teemad:

- **Akadeemilised tulemused** – hinded, tasemetööd, eksamitulemused, läbikukkumised;
- **Õppeedukuse ja osalemise üldnäitajad** – õppeedukuse protsent, tasemetööde läbivus, kiitused ja medalid;
- **Arenguvestlused ja isiklik areng** – arenguvestluste maht ja sisu, eneseregulatsiooni ning ennastjuhtiva õppimise oskused;
- **Puudumised ja osalus** – põhjendamata puudumiste statistika, seosed madala akadeemilise eduga;
- **Rahulolu ja koolikeskkond** – küsitlused õpilaste, õpetajate ja lastevanemate rahulolust, heaolunäitajad, üldpädevused.

Ühisjoonena saab välja tuua, et andmeid kogutakse erinevatest allikatest ning need jäävad sageli isoleerituks. Seosed, näiteks hinnete ja puudumiste või rahulolu ja motivatsiooni vahel, jäävad dokumentides kirjeldamata või subjektiivseks.

Sellest lähtuvalt võiks arendatav andmeanalüüsi tööriist pakkuda võimalusi:

- võrrelda ainepõhiseid tulemusi klasside ja aastate lõikes, sealhulgas riiklike testidega;
- jälgida läbikukkumiste trende ning seostada neid puudumiste ja tagasisidega;

- tuua kokku rahuloluküsitluste tulemused eri osapoolte - õpilased, õpetajad, vanemad, ja aastate kaupa;
- sõnastada visuaalselt eesmärkide ja tulemuste vastavus, tuginedes arengukavade mõõdikutele;
- automatiseerida aastapõhiste ülevaadete koostamine, toetudes juba kogutud andmetele.

Selline lähenemine ei aitaks ainult juhtkonnal ja õpetajatel paremini andmeid tõlgendada, vaid toetaks ka strateegilist juhtimist, enesehindamist ja eesmärgipärast arengukava täitmist reaalajas.

3.2 Kohtumised koolidega

Antud peatükk annab ülevaate kohtumistest kahe kooliga, Mustamäe Riigigümnaasiumi ja Saku gümnaasiumiga. Kohtumiste eesmärgiks oli uurida, milliseid andmeid ja mis kujul koolid neid koguvad, kuidas teostatakse andmeanalüüse, kas on huvi ja vajadus käesoleva töö käigus alustatava lahenduse vastu.

3.2.1 Kohtumine Mustamäe Riigigümnaasiumiga

Mustamäe Riigigümnaasiumiga on hetkeseisuga toimunud üks kohtumine. Kohtumisel osalesid käesoleva töö autorid, juhendaja, Piret Zahkna ning MURGi õppejuht ja haridustehnoloog. Kohtumise eesmärk oli saada ülevaade kooli andmekasutuse praktikatest, probleemidest ja ootustest võimaliku uue andmeanalüütilise tööriista suhtes. Alustuseks tutvustati kooli esindajatele käesoleva töö ideed, mis probleemi see lahendab ning kuidas saab kool abiks olla.

Arutelu käigus ilmnes, et kool kasutab igapäevaselt eKooli andmestikku, mille abil jälgitakse eeskätt kursuse läbikukkujate arvu ja õppeedukust. Aasta lõpus koostatakse õppenõukogu protokoll, kuhu kantakse kõik läbikukkunud õpilased, mistõttu on andmete täpne tõlgendamine ja individuaalne kaardistamine juhtkonna jaoks oluline. Kooli esindajate sõnul ei toeta olemasolevad vahendid seda protsessi piisavalt. eKool võimaldab küll hinnete ja puudumiste jälgimist, kuid üksnes klasside ja õpilaste kaupa. Koolikaupa koondülevaated puuduvad, samuti tuleb vajalik statistika igast klassist eraldi käsitsi alla laadida.

Kooli üldine andmeanalüüs toimub peamiselt kord aastas, kui koostatakse üldtööplaan, mille sisendiks on HARNO uuringud, rahuloluküsitlused ja Haridussilm.ee andmestik. Kool märkis, et andmetega tegeletakse pidevalt, et mõista paremini, miks õpilased hätta jäävad või millistes klassides võiks õpetamise kvaliteet olla ebajärjekindel.

Lahenduse ideed võeti vastu ettevaatliku huviga. Kool ei olnud kindel, millisel kujul võiks valmiv tööriist välja näha, ning tõstatati küsimusi selle kohta, milliseid seoseid üldse oleks võimalik hinnete ja puudumiste vahel mõtestatult analüüsida. Samas toodi välja, et kui lahendus aitaks koondada killustunud andmed ning muudaks kooli juhtimise jaoks olulised mustrid ja riskikohad selgemaks, oleks see koolile kindlasti väärtuslik.

Kohtumise lõpus lepiti kokku, et järgmise sammuna uuritakse eKooli poolelt, kas nende platvormil on olemas API-lahendus, mille abil oleks võimalik automaatselt andmeid süsteemi tõmmata. Kui selline liides puudub, selgitatakse välja, kas nad oleksid valmis selle loomiseks koostööd tegema.

3.2.2 Kohtumised Saku Gümnaasiumiga

Saku Gümnaasiumiga kohtuti kokku kolm korda. Kohtumiste eesmärk oli tutvustada lahenduse ideed, uurida kooli andmekasutust ja vajadusi ning koguda sisendandmeid, mille põhjal luua esmaseid analüüse.

Esimesel kohtumisel osalesid käesoleva töö autorid ja juhendaja, Piret Zahkna ning Saku Gümnaasiumi koolijuht Keit Fomotškin. Kohtumisel tutvustati plaani luua ühtne andmepõhine juhtimislaud, mis koondaks erinevatest allikatest pärit haridusandmeid ja muudaks need lihtsalt ligipääsetavaks kooli erinevates rollides töötajatele. Kohtumise käigus tõstatus rida olulisi küsimusi, millele süsteemi loomisel vastuseid otsitakse:

- Millised andmed koolil õppetöö kvaliteedi hindamiseks üldse olemas on?
- Millised andmed on tehniliselt kättesaadavad?
- Kuidas erinevaid andmestikke omavahel ühildada?
- Kas kogutava info põhjal oleks võimalik luua ühtne andmeplatvorm?

Kuna kohtumine oli suunatud eelkõige visiooni tutvustamisele, ei saadud veel detailseid vastuseid konkreetsete andmestike või andmevajaduste kohta. Samas kinnitas kool valmisolekut

teemadega süvitsi minna. Edasiste sammudena lepiti kokku, et vaadatakse põhjalikumalt Stuudiumi võimalusi ja andmeallikaid, sealhulgas mitte-isikustatud andmeid rahulolu- ja üldpädevusuuringutest (nt ROU, HARNO e-testid, TLÜ õpetajauurimus jt). Koolipoolseteks kontaktisikuteks määrati arendusjuht Maris ja haridustehnoloog Kelly. Koostöö hõlbustamiseks loodi ühine pilvekaust, kuhu kogutakse olemasolevaid andmestikke.

Teises kohtumises arendusjuhi Marisega keskenduti mitte-isikustatud andmete kogumise ja kasutamise praktikatele. Kool kogub regulaarselt erinevaid küsitlusandmeid, sealhulgas riiklikke rahuloluuuringuid, üldpädevuste teste ja koolisiseseid tagasisideküsitlusi. Andmeid hoitakse ja töödeldakse käsitsi Google Sheets tabelites ning iga kooliastme kohta koostatakse kokkuvõttefailid. Lisaks kasutatakse Haridussilma andmeid ning on loodud tulemusnäitajate koondfail arengukava mõõdikute jälgimiseks.

Kohtumise tulemusena ilmnes, et kuigi andmeid kogutakse palju, puudub nende vahel terviklik vaade. Andmed on hajali ning vajavad manuaalset töötlust, mis on Marise sõnul väga ajamahukas. Arendusjuht oli valmis jagama mitte-isikustatud küsitluste tulemusi edasise analüüsi ja arendustöö sisendina.

Kolmas kohtumine toimus kooli haridustehnoloogi Kellyga ning keskendus Stuudiumi võimaluste ja andmevaadete tutvustamisele. Arutelu käigus anti ülevaade sellest, millist statistikat ja andmestikku on võimalik Stuudiumi kaudu näha kooli tasandil, sh hinnete, puudumiste ja õppeedukuse kohta.

Selgus, et ka Stuudium ei võimalda andmeid lihtsalt ja koondatult alla laadida, vaid failid tuleb käsitsi eraldi vormistada ning puudub võimalus kogu andmestikku korraga struktureeritult eksportida. Kohtumise lõpus oli haridustehnoloog valmis jagama hinnete ja puudumistega seotud andmeid, nagu ta need süsteemist kätte saab, ning nõustus need eelnevalt anonüümseks tegema.

3.3 Suhtlus infosüsteemide pakkujatega ja andmete ligipääs

Käesoleva töö raames võeti ühendust kahe enimkasutatava õppeinfosüsteemi Stuudiumi ja eKooliga, et selgitada välja, millised on võimalused haridusandmete ligipääsuks, automatiseeritud andmeside loomiseks ning koostööks arendatava prototüübi kontekstis.

Stuudiumi esindajatega ei õnnestunud kontakti saada. Seetõttu puudub hetkel info selle kohta, kas neil oleks huvi ja tehniline valmidus arendada koolidele või kolmandatele osapooltele suunatud andmepäringu API liideseid.

eKooliga toimus aga üks sisuline kohtumine, mille käigus tutvustati neile töö eesmärki ning arendatava andmeanalüütilise lahenduse visiooni. Arutelu käigus toodi eKooli poolt esile mitmeid andmekaitse ja privaatsusega seotud küsimusi: eelkõige, kuidas tagatakse isikuandmete töötlemise turvalisus, millised andmekogud on sobivad ning kes omab ligipääsu millisele andmele. Üldine suhtumine oli siiski konstruktiivne ja koostööaldis. eKool avaldas huvi mõista täpsemalt, milliseid andmeid ja millises vormingus oleks süsteemi toimimiseks vaja.

Kohtumise tulemusel lepiti kokku, et töö autorid edastavad eKoolile täpsustatud kirjaliku ülevaate andmetest, mida oleks vaja ligipääsuks, sealhulgas andmetüübid ja väljad. Selle põhjal lubas eKool hinnata, kas ja millises mahus oleks võimalik arendada eraldi API-liides, mille kaudu saaks andmeid struktureeritult ja turvaliselt edastada. Käesoleva töö kirjutamise hetkel ei ole lõplikku vastust veel saadud.

Kuna mõlemad platvormid ei paku hetkel koolidele võimalust eksportida kogu vajaminevat andmestikku automatiseeritud kujul ega avalikke API-liideseid, tuleb andmed koguda käsitsi koolidest endist. Praktikas tähendab see, et automatiseeritud süsteemi potentsiaal jääb suurel määral kasutamata ning koolid peavad jätkuvalt andmeid käsitsi ette valmistama. Olgu selleks Exceli tabelid, anonüümseks muudetud andmefailid või kokkuvõtvad failid PDF-formaadis. Seetõttu on praegu käesoleva töö käigus teostada analüüse mitteisikustatud failidega.

Tulevikus võiks Stuudiumi ja eKooli avatud andmepäringu liidesed võimaldada senisest palju paremat andmepõhist juhtimist koolides, sealhulgas jooksva õppeedukuse ja puudumiste jälgimist, rahulolu-uuringute sidumist õpitulemustega ning arengukava täitmise automaatset seiret. Selle eelduseks on aga läbimõeldud andmekaitse probleemid, paindlik ja turvaline tehniline lahendus, mille valmidus sõltub infosüsteemide arendajate otsusest.

3.4 Kokkuvõte

Kokkuvõtvalt võib öelda, et nii Mustamäe Riigigümnaasium kui ka Saku Gümnaasium koguvad juba täna ulatuslikult haridusandmeid. Nende dokumentidest ja kohtumistest joonistuvad välja ühised väljakutsed: andmestik on killustunud, analüüs toimub valdavalt käsitsi ning seosed eri andmetüüpide vahel jäävad sageli kasutamata.

Õppeinfosüsteemid ei võimalda hetkel andmeid struktureeritult ega automaatselt eksportida. Kuigi eKool on väljendanud valmisolekut kaaluda andmevahetusliidese loomist, ei ole lõplikke kokkuleppeid hetkel saavutatud. Seetõttu saab arendatav prototüüp hetkel toetuda üksnes koolide poolt jagatud mitte-isikustatud andmestikule, näiteks anonüümseks muudetud hinded ja puudumised ning rahuloluküsitluste tulemused. Koolide valmisolek sellist andmestikku jagada on olemas.

Lahenduse loomisel on seega oluline keskenduda sellele, kuidas koondada erinevad olemasolevad andmed ühtsesse vaatesse, visualiseerida kriitilised näitajad ning toetada strateegilist juhtimist. Süsteem võiks pakkuda järgmisi funktsionaalsusi:

- Aine- ja klassipõhine õpitulemuste võrdlus erinevate aastate lõikes;
- Puudumiste ja hinnete vaheliste seoste tuvastamine;
- Rahulolu- ja üldpädevuste andmete dünaamiline võrdlemine, seda nii klasside, rollide kui ka aastate lõikes;
- Arengukava mõõdikute sidumine tegelike andmetega ning visuaalne jälgimine eesmärkide täitmisest.

Selline tööriist toetaks koolide vajadust andmepõhise juhtimise järele, võimaldaks automatiseerida seni käsitsi tehtavaid analüüse ning looks parema aluse andmetel põhinevatele otsustele. Täieliku potentsiaali saavutamiseks on tulevikus hädavajalik koostöö infosüsteemide arendajatega ning ligipääs reaalajas uuenevatele andmetele.

4 Isikuandmete kaitse ja turvalisus haridusandmete töötlemisel

Haridusandmete analüüsi ja visualiseerimise eesmärgil tuleb arvestada, et sellised andmed nagu õpilaste hinded ja puudumised kvalifitseeruvad isikuandmeteks. Kui tulevikus soovitakse integreerida lahendust õppeinfosüsteemidega nagu eKool või Stuudium, muutub andmekaitse tagamine keskseks eelduseks igasugusele süsteemi laiemale kasutusele.

4.1 Isikuandmete õiguslik käsitlus

Euroopa Parlamendi ja nõukogu määruse (EL) 2016/679 ehk isikuandmete kaitse üldmääruse (GDPR) kohaselt loetakse isikuandmeteks igasugust teavet tuvastatava isiku kohta. Haridusandmete puhul võib see hõlmata nii hindeid, puudumisi kui ka emotsionaalse heaolu mõõdikuid, kui neid saab seostada konkreetse isikuga [9].

GDPR artikkel 6 sätestab, et andmete töötlemine on õiguspärane üksnes teatud juhtudel. Haridusandmete kontekstis võivad sobilikud õiguslikud alused olla järgmised:

- **Avaliku ülesande täitmine** – näiteks kooli seadusest tulenev kohustus tagada kvaliteetne haridus ja jälgida õpilaste arengut;
- **Nõusolek** – kui lapsevanem või õpilane on andnud selgesõnalise loa andmete töötlemiseks konkreetsetel eesmärkidel (nt uuringutes osalemine);
- **Õigustatud huvi** – kui see ei kahjusta andmesubjekti õigusi (nt statistiline analüüs paremate õppemeetodite leidmiseks) [10].

Eestis täpsustab GDPR-i rakendamist isikuandmete kaitse seadus (IKS), mille § 14 sätestab, et andmete töötlemine peab olema lisaks muule ka eesmärgikohane, seaduslik ja turvaline [11].

4.2 Tehnilised meetmed OpenSearch platvormil

OpenSearchi turvamoodul (OpenSearch Security) võimaldab rakendada mitmeid meetmeid, et tagada andmete turvalisus nii hoiustamisel kui kasutamisel [12]. Peamised funktsioonid hõlmavad järgmist:

- **Rollipõhine ligipääs**: igale kasutajale saab määrata õigused konkreetsete indeksite või andmeväljade lõikes;
- **Autentimine ja autoriseerimine**: toetatud on mitmed standardiseeritud mehhanismid, sealhulgas LDAP, SAML või OpenID Connect;
- **Auditilogimine**: süsteem talletab kõik andmetega tehtud tegevused, võimaldades hiljem jälgida ja analüüsida ligipääse või muudatusi;
- **Krüpteeritud ühendused (TLS)**: andmete edastamine ja salvestamine toimub krüpteeritult, mis aitab vältida andmeleket või kõrvaliste isikute juurdepääsu;
- **Dokumendi- ja andmevälja tasemel turvalisus**: võimaldab piirata kasutaja ligipääsu üksikutele dokumentidele või andmeväljadele isegi siis, kui need asuvad samas indeksis.

Tänu neile funktsionaalsustele saab süsteemis defineerida näiteks rollid õpetajatele, kes näevad ainult oma klasside andmeid, või juhtkonnale, kellel on ligipääs koondvaadetele.

4.3 Andmete anonüümistamine ja pseudonüümistamine

Käesolevas töös kasutati mitte-isikustatud andmeid: kõik isikut tuvastavad atribuudid eemaldati. See lähenemine on kooskõlas GDPR artikli 5 punktiga 1(c), mille järgi tuleb töödeldavad andmed piirata minimaalselt vajalikega [9, 13].

Kui tulevikus peaks tekkima vajadus töödelda isikustatud andmeid, tuleb sõlmida koolide ja süsteemiarendaja vahel andmetöötluslepingud ning vajadusel koostada mõjuhinnang (DPIA), nagu nõuab GDPR artikkel 35 [9].

4.4 Isikustatud andmete töötlemine ja lepingulised kohustused

Kui tulevikus hakatakse töötlema isikustatud andmeid (nt otse eKooli või Stuudiumi API kaudu saadud hinded või puudumised), tuleb kooli ja süsteemi arendaja vahel sõlmida

andmetöötlusleping (DPA), mis vastab GDPR artikli 28 nõuetele [9].

Andmetöötlusleping peab sisaldama vähemalt järgmisi elemente:

- **Andmetöötluse eesmärk ja kestus**: näiteks kooli statistika ja juhtimisotsuste toetamine, andmete säilitamine maksimaalselt 5 aastat;
- **Töödeldavate andmete liigid ja andmesubjektide kategooriad**: nt õpilaste hinded, puudumised, küsitluste vastused;
- **Volitatud töötleja kohustused ja õigused**: sh turvameetmed, töötajate konfidentsiaalsuskohustus, auditivalmidus;
- **Tegevused andmelekkega või rikkumisega toimetulekuks**: kuidas teavitatakse kooli ja vajadusel Andmekaitse Inspektsiooni;
- **Andmete kustutamine või tagastamine lepingu lõppemisel**: kas andmed hävitatakse, anonüümistatakse või tagastatakse vastutavale töötlejale;
- **Allhangete korral täiendavad kohustused**: kui arendaja kasutab kolmanda osapoole pilveteenuseid või partnerettevõtteid, peab see olema eraldi lepingus lubatud.

Kõik volitatud töötlejad peavad järgima andmesubjekti õigusi (nt õigus andmete parandamisele, kustutamisele), mille täitmise eest vastutab kool kui vastutav töötleja.

4.5 Mõjuhinnang isikuandmete kaitsele (DPIA)

Kui andmetöötlus toob kaasa suure tõenäosusega kõrge riski füüsiliste isikute õigustele ja vabadustele, tuleb enne töötlemise alustamist koostada andmekaitsealane mõjuhinnang (DPIA) vastavalt GDPR artiklile 35 [9]. Mõjuhinnang on eelkõige vajalik siis, kui:

- töödeldakse suurt hulka isikuandmeid,
- tehakse andmete süstemaatilist jälgimist või hindamist (nt hindetrendid ajas, rahulolu areng),
- töödeldakse eriliigilisi andmeid (nt õpilaste tervise või emotsionaalse seisundi hinnangud).

DPIA peab sisaldama:

- kirjeldust töötlemistoimingutest ja eesmärkidest;
- hinnangut riskidele, mis võivad andmesubjekte mõjutada;

- kirjeldust leevendusmeetmetest, mis aitavad neid riske maandada (nt krüpteerimine, ligipääsupiirangud, pseudonüümistamine) [13].

4.6 Eeltingimused koostööks infosüsteemidega

Kui süsteem integreeritakse koolide ametlike infosüsteemidega (nt eKool, Stuudium), tuleb arvesse võtta järgmisi eeltingimusi:

- Kool jääb andmete vastutavaks töötlejaks – tema määratleb, milleks ja milliseid andmeid kogutakse ja kasutatakse.
- Süsteemi arendaja tegutseb volitatud töötlejana, kellel on lubatud andmeid töödelda ainult vastutava töötleja dokumenteeritud juhiste alusel.
- Kõik andmetöötlustoimingud, kasutajate õigused ja tehnilised kaitsemeetmed tuleb kirjalikult fikseerida.
- Vajadusel tuleb kaasata kooli andmekaitsespetsialist (DPO), kes hindab riske ja tagab, et töötlemine vastaks seadustele [10].
- Süsteemi arhitektuur ja tööprotsessid peavad olema auditeeritavad – see tähendab, et on võimalik logide kaudu tuvastada, kes, millal ja miks andmetele ligi pääses või neid muutis.

Need põhimõtted aitavad tagada, et õpilaste isikuandmeid töödeldakse seaduslikult, turvaliselt ja läbipaistvalt.

5 Andmeanalüüs

Käesolevas peatükis kirjeldatakse Saku Gümnaasiumi andmete põhjal läbi viidud analüütilisi etappe, mille eesmärk oli tuvastada olulisi mustreid ja seoseid nii küsitluspõhistes kui ka hinnetel põhinevates andmestikes. Analüüsi eesmärk oli toetada kooli arenguprotsesse, pakkudes faktipõhist sisendit juhtimisotsusteks ning õpetamise tõhustamiseks.

Andmeanalüüs jagunes kaheks põhisuunaks: (1) küsitlusandmete analüüs, mille raames rakendati klasterdamist ja masinõppelist modelleerimist vastajate profiilide mõistmiseks, ning (2) hinnetepõhine analüüs, kus keskenduti ainetevahelistele seostele ja teadmiste latentsetele struktuuridele, kasutades korrelatsioonanalüüsi ja põhikomponentide analüüsi (PCA). Mõlema suuna puhul pöörati tähelepanu andmete eeltöötlusele, normaliseerimisele ja puuduvate väärtuste käsitlemisele, et tagada tulemuste usaldusväärsus.

Lisaks rõhutatakse peatükis töövoogude automatiseerimist ning failipõhist lähenemist, mis võimaldas teostada korduvaid ja skaleeritavaid analüüse mitme õppeaasta lõikes. Tulemused salvestati masinloetavas vormingus, et neid oleks lihtne edasi töödelda ja visualiseerida.

Järgnevad alajaotused käsitlevad üksikasjalikult kasutatud meetodeid, analüüsitud tunnuseid ning saadud järeldusi, pakkudes süsteemset ülevaadet andmetest tuletatud teadmustest.

5.1 Andmete eeltöötlus

Käesolevas töös kasutati Saku Gümnaasiumi küsitluspõhiseid andmeid, mis hõlmasid erinevate aastate tagasisideküsitlusi õpilastelt, õpetajatelt, lapsevanematelt ja kooli juhtkonnalt. Küsitlused puudutasid mitmeid koolielu aspekte nagu koolikeskkond, õpetajakvaliteet, infovahetus, tugiteenused ja üldine rahulolu.

5.1.1 Küsitluspõhiste andmete töötlusprotsess

Küsitlused olid esitatud Exceli vormingus, mis teisendati CSV-failideks. Seejärel viidi andmed läbi struktureeritud eeltöötlustsükli, mis realiseeriti spetsiaalselt kirjutatud Python-

skripti abil. Skript töötles andmeid järgmiselt:

- **Kaustastruktuuri kasutamine:** Küsitlusandmed organiseeriti kooli nime ja aastate kaupa alamkaustadesse, kust neid automaatselt loeti ja töödeldi.
- **Veerupäiste ja tekstiväljade puhastamine:** Veerupäistest ja tekstiväljadest eemaldati diakriitikud, need teisendati kasutades ASCII-tähestikku ning normaliseeriti väiketähtedeks ja eemaldati tühikud.
- **Ajatempli standardiseerimine:** Ajatempli veerg teisendati ISO 8601 formaati, kui kuupäevavorming oli tuvastatav (nt `2023-03-15T13:45:00`).
- **Likerti-skaala vastuste kaardistamine:** Valikvastustega küsimustes kasutatud vastused nagu *"nõustun täiesti"* või *"pigem ei nõustu"* kaardistati numbrilisele skaalale vahemikus 1 kuni 5. Eraldi käsitleti vastust *"ei oska vastata"*, mis kodeeriti väärtusega `-1`. Kõik vastused normaliseeriti (nt eemaldati täpitähed) enne vastavustabeliga võrdlemist.
- **Tundmatute väärtuste logimine:** Kui vastus ei kattunud ühegi teadaoleva vormiga, lisati see eraldi nimekirja, et võimaldada käsitsi ülevaatamist ja vajadusel parandamist.
- **Tühjade või ebatäielike ridade eemaldamine:** Andmestikust eemaldati tühjad või täielikult puuduolevad kirjed, et tagada andmete kvaliteet.

Töödeldud failid salvestati uude kaustastruktuuri `data/processed_data`, kusjuures iga faili nimi sai juurde tähise `_processed.csv`, viidates, et see on läbinud eeldefineeritud töötlustsükli.

Kokkuvõttes võimaldas see automatiseeritud eeltöötlusprotsess erinevate aastate ja küsitluste ühtset ning korratavat käsitlust, mis on oluline usaldusväärse kvantitatiivse analüüsi teostamiseks.

Hinnetepõhiste andmete töötlusprotsess

Lisaks küsitluspõhistele andmetele töödeldi käesolevas töös ka hinnetepõhiseid andmeid, mis pärinesid Saku Gümnaasiumi erinevate aastate õpitulemustest. Andmed olid esitatud CSV-formaadis, kusjuures igas failis sisaldus veerg hinnete või hinnete keskmisega. Andmete ühtlase ja korratava töötluse tagamiseks rakendati spetsiaalset Python-skripti, mille abil viidi läbi järgmine eeltöötlustsükkel:

- **Kaustastruktuuri kasutamine:** Hinnete failid organiseeriti kooli nime ja aastate

alusel vastavatesse alamkaustadesse, kust need automaatselt loeti ja töödeldi.

- **Hinnete parsimine:** Kui andmefail sisaldas veergu `grades`, kus hinded olid loetletud komadega eraldatud stringina (nt `"5, 4, 3"`), teisendati see veerg, et iga kirje muutuks vastavaks täisarvude loeteluks.
- **Keskmise hinde arvutamine:** Iga kirje jaoks arvutati keskmine hinne olemasolevate hinnete põhjal. Kui hinnete veerg puudus, kuid oli olemas juba arvutatud keskmisega veerg, kasutati seda.
- **Tundmatute vormingute käsitlemine:** Kui hinnete või keskmise hinde veerg puudus, katkestati töötlus antud failiga ning väljastati veateade, et probleem oleks tuvastatav ja parandatav.
- **Töödeldud andmete salvestamine:** Kõik töötluse läbinud andmefailid salvestati uude kaustastruktuuri, kus failide nimedele lisati sufiks `_preprocessed.csv`, tähistamaks, et need on eeltöödeldud.

Selline töötlusprotsess võimaldas hinnetepõhiste andmete ühtset käsitlemist sõltumata algandmete esitusviisist (täishinnete loetelu või eelnevalt arvutatud keskmine). See lihtsustas edasist analüüsi, näiteks korrelatsioonide või keskmiste võrdlemist erinevate aastate ja sihtrühmade lõikes.

5.2 Analüüsi metoodika: küsitlusandmed

Küsitlusandmete analüüsimisel on oluline mõista, millised mustrid ja seosed esinevad erinevate vastajate hinnangutes. Selleks rakendati klasterdamise ja masinõppe meetodeid, et tuvastada sarnaste vastustega rühmi ning selgitada, millised tunnused neid rühmi iseloomustavad. Käesolevas jaotises tutvustatakse kasutatud analüüsimeetodeid ja nende rakendamise loogikat.

5.2.1 Küsitluspõhiste andmete klasterdamine

Klasterdamine võimaldab jagada vastajad gruppidesse lähtuvalt nende sarnasustest, ilma et oleks vaja eelnevalt teada, millised grupid peaksid eksisteerima. See meetod on eriti kasulik haridusandmete puhul, kus võivad esineda selged, kuid varjatud mustrid. Järgnevates alajaotustes kirjeldatakse, kuidas klasterdamiseks andmed ette valmistati ja millist algoritmi kasutati.

Andmete imputeerimine

Puuduvate andmete täitmiseks (*missing data imputation*) kasutati keskmisega imputeerimist, kus iga veeru puuduvad väärtused asendati selle veeru aritmeetilise keskmisega. See lähenemine sobib haridusandmete puhul, kus väärtused on sageli pidevad (nt hinded, hinnangud) ja keskmine aitab vältida andmestiku liigset moonutamist.

Andmete normaliseerimine

Enne klasterdamist skaleeriti kõik tunnused samale skaalale vahemikus $[0, 1]$. Kuna K-means algoritm põhineb Eukleidese kaugusel, on see tundlik erinevatele mõõtkavadele. Skaleerimine aitab vähendada olukordi, kus suurema väärtusega tunnused mõjutavad tulemusi ebaproportsionaalselt.

K-means klasterdamine

Klasterdamiseks kasutati K-means algoritmi ning klastrite arvuks määrati $k = 3$. See meetod võimaldab vastajad grupeerida nende sarnasuste põhjal, toetudes nende vastustele.

- Arvutuslikult efektiivne ka suuremate andmestike puhul;
- Tulemusi on lihtne visualiseerida ja tõlgendada;
- Sobib pidevate tunnuste korral.

Grupipõhine lähenemine

Kuna andmestik jagunes erinevateks rühmadeks (nt klassid või küsimustike tüübid), viidi klasterdamine läbi iga grupi kohta eraldi. See võimaldas paremini arvestada konkreetse grupi eripäradega ning vältida moonutusi, mis võivad tekkida erinevate rühmade segundamisel.

5.2.2 Klastrikuuluvuse selgitamine RandomForesti ja SHAP abil

Et mõista, miks vastaja kuulub teatud klastrisse, kasutati masinõppelist klassifikatsioonimudelit ning SHAP-väärtusi, mis aitavad tulemusi selgitada inimloetaval kujul.

RandomForesti mudel ja tunnuste tähtsuse hindamine

Iga rühma ja aasta kohta treeniti eraldi mudel, mille abil püüti ennustada vastaja kuuluvust kindlasse klastrisse. Mudeli täpsust hinnati segadusmaatriksi ja klassifitseerimisaruande alusel. Samuti määrati tunnuste olulisus, mis näitas, millised vastaja omadused mõjutasid enim klastri määramist.

SHAP-analüüs: individuaalsete prognooside selgitamine

SHAP-väärtused võimaldasid hinnata iga üksiku vastaja kohta, millised tunnused mõjutasid tema klastri määramist. Näiteks võis teatud väide rahulolu kohta oluliselt suurendada tõenäosust kuuluda kõrge rahuloluga klastrisse.

Automatiseeritud tulemuste genereerimine

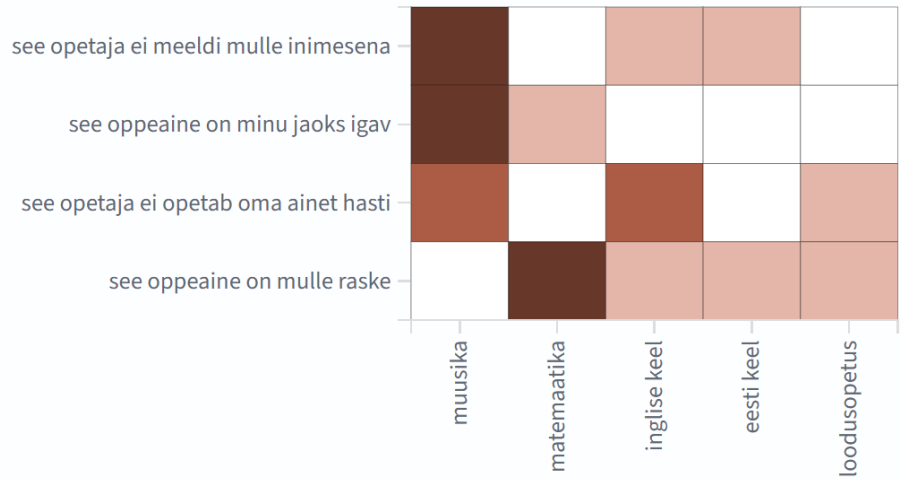
Iga rühma kohta koostati automaatselt inimloetavad kokkuvõtted, mis sisaldasid:

- mudeli täpsuse mõõdikuid (nt klassifitseerimisaruanne);
- olulisemaid tunnuseid ja nende mõju suunda;
- näidisselgitust ühe vastaja kohta;
- tulemuste tõlgendusi ja soovitusi.

See võimaldas õpetajatel ja koolijuhtidel kiiresti mõista, millised teemad mõjutavad enim vastajate hinnanguid või rahulolu taset.

5.2.3 Vaba teksti analüüs ja teemade modelleerimine

Lisaks kvantitatiivsele klasteranalüüsile viidi läbi ka vabade tekstivastuste sisuline analüüs, mille eesmärk oli leida korduvaid teemasid ja mõttemustreid vabades vastustes. Näiteks võimaldas analüüs välja tuua probleemsemad õppeained ning mis on selle põhjuseks. Tulemusi kujutati korrelatsioonimaatriksina. Mida tumedam värv, seda rohkem esines antud õppeainet ja põhjust vabade vastustega küsimustes (Joonis 3).



Joonis 1. 3.-6.klassi probleemsed õppeained ja põhjused 2023

Tekstide eeltöötlus ja tokeniseerimine

Tekstid normaliseeriti, teisendati väiketähtedeks ning eemaldati mittesobivad märgid. Seejärel rakendati lemmatiseerimist ja osa-sõna märgendamist, keskendudes nimisõnadele, tegusõnadele ja omadussõnadele. Kõrvaldatud said sagedased eesti keele stoppsõnad, nagu *"ja", "see", "et"*, mis ei kandnud sisulist tähendust.

Sõnastike ja korpuse koostamine

Koostati sõnastik, millest eemaldati väga harva või liiga sagedasti esinevad sõnad. Alles jäid need, mis esinesid vähemalt kolmes dokumendis, kuid mitte rohkem kui pooltes tekstides. See aitas keskenduda teemadele, mis on sisuliselt informatiivsed ja mitte müra tekitavad.

LDA (Latent Dirichlet Allocation) mudeli treenimine

Teemade avastamiseks kasutati LDA-mudelit, mille eesmärk oli leida kuni kaheksa peamist teemat tekstidest. Mudelit treeniti korduvalt kogu andmestiku peal, et saavutada usaldusväärsed ja arusaadavad tulemused.

Tulemuste ekspordimine ja kasutamine

Lõplik teema-analüüs eksporditi NDJSON-formaadis, mis võimaldas tulemusi lihtsalt edasise analüüsi või visualiseerimise jaoks kasutada. Nii said koolijuhid või õpetajad kiire ülevaate sellest, millised teemad domineerisid vastajate vastustes ning millised valdkonnad vääriksid süvitsi tähelepanu.

5.2.4 Tulemuste automaatne kokkuvõtlik esitamine

Lisaks analüüsile loodi töövoog, mis võimaldas automaatselt koostada kokkuvõtteid klastrite omaduste kohta. Eesmärk oli muuta andmepõhiste järelduste tegemine õpetajatele ja koolijuhtidele võimalikult kiireks ja arusaadavaks.

Andmetöötlus ja tähenduse lisamine

Andmetabelite veerunimed muudeti arusaadavamaks: eemaldati tehnilised prefiksid, numbrid ja sõnastus kohandati loetavaks ning tähenduslikuks. Vajadusel lisati veergudele lühike selgitus, mis andis konteksti.

Klastrianalüüsi kokkuvõtted

Iga aasta ja klastri kohta arvutati keskmised väärtused ning tuvastati need küsimused, mille keskmised hinnangud olid kõige madalamad. Selline lähenemine võimaldas esile tõsta parenduskohti ning anda sisukas kirjeldus iga klastri võimalikust tähendusest.

Trendianalüüs ajas

Lisaks hinnati, kuidas erinevate klastrite esinemissagedus muutus ajas. See võimaldas välja tuua trende, näiteks kas mõni teatud hoiakutega rühm on muutunud aja jooksul arvukamaks või väiksemaks.

Grupi määramine failinimede põhjal

Andmefailide nimesid kasutati selleks, et tuvastada, millisele sihtrühmale (nt vanuserühm või klassitase) andmestik kuulus. See võimaldas automaatselt eristada analüüse rühmiti.

Kokkuvõtete salvestamine

Iga grupi analüüs ja kokkuvõte salvestati eraldi tekstifailina, et neid oleks võimalik jagada ja kasutada konkreetsete sihtrühmade lõikes.

Automaatne töövoog ja eksport

Analüüsiprotsess viidi läbi kõigi andmekataloogide ja aastate lõikes, kus sobivad failid loeti sisse, kombineeriti, ning seejärel koostati iga grupi kohta automaatne kokkuvõte. Tulemused salvestati eraldi alamkataloogi, mis võimaldas neid kiiresti edasi töödelda või jagada.

5.3 Analüüsi metoodika: hinnete andmed

Käesolevas peatükis keskendutakse õpilaste hinneteandmete analüüsimiseks kasutatud metoodikale. Hinnete kujunemine ja nendevahelised seosed kannavad olulist infot nii individuaalse õppija tugevuste ja nõrkuste kui ka laiemate õpimustrite kohta koolis tervikuna. Metoodiline läbimõeldus andmetöötluses on hädavajalik, et saadud tulemused oleksid usaldusväärsed, tõlgendatavad ja rakendatavad haridusotsuste tegemisel.

Analüüs lähtub vajadusest mõista, kuidas erinevad õppeained omavahel korreleeruvad, millised peidetud oskuste dimensioonid hinnetest välja joonistuvad ning kuidas need seosed ajas muutuvad. Selleks rakendati kombinatsiooni klassikalistest statistilistest meetoditest (nt Pearsoni korrelatsioon) ja mitmemõõtmelise andmeanalüüsi võtetest (nt põhikomponentide analüüs). Tähelepanu pöörati ka puuduvate hinnete käsitlemisele ning andmete standardiseerimisele, et vältida eelarvamusi ja võimaldada usaldusväärseid järeldusi.

Peatükk annab ülevaate sellest, kuidas andmed struktureeriti, milliseid eeltöötlusetappe rakendati ning miks valiti just need analüüsimeetodid. Samuti kirjeldatakse, kuidas tulemused salvestati ja struktureeriti, et neid oleks võimalik hiljem tõhusalt kasutada

otsustustoe või visualiseerimise tarbeks.

5.3.1 Hinnete korrelatsioonide ja komponentide analüüs

Hinnete andmete analüüs jagunes kaheks põhietapiks: esiteks hinnete korrelatsioonide tuvastamine ning teiseks põhiliste komponentide analüüs (PCA), et mõista, millised teadmiste valdkonnad (õppeained) kaldusid koos varieeruma ning millised peamised suunad andmestikus esinesid.

Andmete eeltöötlus

Kõikide analüüside eelduseks oli andmete ühtlustamine ja täitmine. Andmestik pivot'iti kujule, kus iga rida vastas ühele õpilasele ning iga veerg ühele õppeainele. Kui mõne aine hinne puudus, asendati see selle aine keskmise hindega, et säilitada korrelatsioonianalüüsi terviklikkus. Selline keskmisega imputeerimine aitab hoida andmestiku statistilist struktuuri moonutamata.

5.3.2 Õppeainete omavahelised seosed: korrelatsioonanalüüs

Õppeainete vaheliste hinnete seoste uurimine võimaldab välja selgitada, kas ja millisel määral ühes aines hästi esinevad õpilased saavutavad häid tulemusi ka teistes ainetes. Selline korrelatsioonanalüüs võib viidata üldistele kognitiivsetele oskustele, metoodilisele kattuvusele või ainespetsiifilisele õpetamisviisile. Järgnevates alajaotustes kirjeldatakse kasutatud metoodikat ja saadud tulemuste tõlgendamist.

Metoodika

Iga aasta ja andmefaili kohta arvutati Pearsoni korrelatsioonimaatriks, mis näitab, kui tugevalt kahe aine hinnete varieerumine on omavahel seotud. Analüüs viidi läbi ainult erinevate ainete paaride kohta (välistati identsete veergude võrdlus) ning eemaldati need seosed, mille korrelatsioon oli väiksem kui 0.1.

- Tugevad positiivsed korrelatsioonid viitavad sellele, et õpilased, kes saavad ühes aines häid hindeid, kalduvad hästi esinema ka teises aines.
- Seoseid arvutati ainult juhul, kui mõlemad ained olid piisavalt täidetud.
- Tulemused salvestati `NDJSON`-formaadis, et võimaldada lihtsat importi visualiseerimis- või otsingumootoritesse (nt Elasticsearch).

Tulemuste tõlgendamine

Analüüsi eesmärk oli tuua esile sellised ainesuhted, mis võiksid viidata üldistele oskustele (nt loogiline mõtlemine või keeleline võimekus) ning anda koolijuhile või õpetajatele sisendit, millistes valdkondades õppimist saab toetada mitmekülgsemalt. Näiteks võib tugev korrelatsioon matemaatika ja füüsika vahel viidata vajadusele toetada matemaatikaoskusi juba varasemas etapis, et tagada edasine edu loodusainetes.

5.3.3 Põhikomponentide analüüs (PCA)

Kuna hinnete andmestik võib sisaldada palju omavahel seotud tunnuseid, on otstarbekas kasutada mõõtmete vähendamise meetodeid, et tuua esile kõige olulisemad teadmiste dimensioonid. Põhikomponentide analüüs (PCA) on statistiline meetod, mis aitab keerukaid andmestikke lihtsustada ilma oluliselt kaotamata informatsiooni. Järgnevas osas selgitatakse PCA rakendamise eesmärki ja selle käigus saadud tulemuste tõlgendust.

Eesmärk ja lähenemine

Et mõista, millised peamised teadmiste "suunad"või latentstruktuurid hinnete andmetes esinevad, rakendati PCA-mudelit. See võimaldas taandada mitmemõõtmelise hinnete andmestiku kolmele põhikomponendile, säilitades võimalikult palju andmete variatiivsust.

- Enne analüüsi normaliseeriti andmed standardiseerimise teel (nullkeskmine, ühikvariants).
- PCA komponendid näitasid, kuidas erinevad õppeained moodustavad ühiseid mõõtmeid, mille kaudu saab õppurite erinevusi kirjeldada.
- Lisaks hinnetele säilitati iga õpilase kohta ka võimalik klassikuuluvus, et võimaldada rühmapõhist analüüsi.

Tulemuste struktuur ja kasutamine

Tulemused salvestati NDJSON-formaadis, kus iga dokument vastab ühele õpilasele ja sisaldab:

- hindepõhiseid PCA-komponente (`pca1`, `pca2`, `pca3`);
- iga aine täidetud (või täidetud keskmisega) hindeid;
- võimalusel ka õpilase klassiinfo;
- ühtset andmeindeksit hilisemaks otsimiseks või visualiseerimiseks.

Rühmapõhine analüüs ja visualiseerimine

Tänu salvestatud klassiinfole oli võimalik komponentide alusel joonistada nt punktdiagramme, mis näitavad, kuidas erinevad klassid paiknevad peamistes teadmistel põhinevates dimensioonides. See võimaldab tuvastada klassidevahelisi erinevusi või mustreid (nt kas mõni klass paistab silma tugeva reaalainete kompetentsiga).

5.3.4 Automatiseeritud töövoog ja failipõhine grupeerimine

Tõhus andmetöötlus eeldab hästi struktureeritud töövoogu, mis võimaldab andmeid süstemaatiliselt töödelda, jälgida ja uuendada. Käesolevas töös kasutati automatiseeritud skripte, mis suutsid iseseisvalt klassifitseerida ja salvestada andmed vastavalt koolile, aastale ja õpilasrühmale. See võimaldas mitte ainult lihtsamat andmete haldamist, vaid ka järjepidevat ja korratavat analüüsiprotsessi.

Failide organiseerimine ja töötlemine

Iga andmefail töödeldi automaatselt ning salvestati vastavasse väljundkausta, mille struktuur vastas kooli ja aasta nimetusele. Failinimedes sisalduvad märksõnad (nt klassitase) võimaldasid automaatselt tuvastada, millisele õpilasrühmale andmestik kuulus.

Mitmeaastane analüüs ja muutuste jälgimine

Andmeid töödeldi ka mitme õppeaasta lõikes, mis võimaldas jälgida, kuidas komponentide ja ainekorrelatsioonide struktuur aja jooksul muutub. Näiteks sai hinnata, kas mõni oskuste mõõde on teatud aastatel rohkem varieeruv või kas mõni aine on hakanud rohkem korreleeruma teistega.

Tulemuste salvestamine ja jagamine

Iga töötlemise tulemus salvestati eraldi `.ndjson` failina, mida on võimalik kasutada edasiseks visualiseerimiseks (nt Kibana, Tableau) või õpetajate ja koolijuhtide otsustusprotsessi toetamiseks.

5.4 Valitud meetodite põhjendus ja teoreetiline taust

Käesolevas töös kasutati mitmeid andmeteaduse ja masinõppe meetodeid, mille valik põhines nii haridusandmete spetsiifikal kui ka analüüsi eesmärkidel. Järgnevalt kirjeldatakse põhjalikult kasutatud meetodite valimise põhjuseid.

5.4.1 Küsitluspõhiste andmete analüüsi meetodid

Küsitluspõhiste andmete töötlus ja analüüs nõuab tähelepanelikkust puuduvate väärtuste, erinevate mõõtkavade ja andmekvaliteedi osas. Käesolevas töös rakendati meetodeid, mis on sobilikud just haridusliku tagasiside analüüsimiseks, kus andmestik võib olla struktuurselt mitmekesine ja osaliselt täitmata.

Keskmisega imputeerimine

Puuduvate väärtuste täitmiseks kasutati lihtsat, kuid tõhusat keskmisega imputeerimise meetodit. See on üks levinumaid lähenemisi pidevate tunnuste puhul, eriti haridusandmetes, mis sageli kannatavad harvaesinemise ja mittetäielikkuse all [14].

- Meetod säilitab andmestiku suuruse, kuna ridu ei eemaldata,
- on arvutuslikult kiire ja efektiivne,
- ei nõua keerukaid eeldusi andmete jaotuse kohta [15, 14].

Keskmisega asendamine sobib eriti hästi hariduslike küsitlusandmete puhul, kus puuduvad hinnangud võivad olla tingitud juhuslikest vastajate puudumistest või tehnilistest vigadest.

Andmete skaleerimine (Min-Max normaliseerimine)

Enne klasterdamist viidi kõik tunnused skaalale [0,1] kasutades Min-Max normaliseerimist. Meetodi valik põhines järgmistel kaalutlustel:

- K-means algoritm tugineb Eukleidese kaugusele, mis on tundlik eri mõõtkavade suhtes [16].
- Skaleerimine tagab, et ükski tunnus ei domineeri kauguse arvutamisel pelgalt suure väärtuse tõttu.
- Min-Max meetod on interpreteeritav ja sobib hästi, kui andmed on piiratud teadaolevas vahemikus [17].

K-means klasterdamine

Klasterdamiseks valiti K-means, mis on:

- arvutuslikult efektiivne ka suuremate andmehulkade puhul;
- hästi tõlgendatav ning laialdaselt kasutatav sarnaste andmestike analüüsimisel;
- sobiv pidevate ja skaleeritud tunnustega andmestike puhul [18, 19].

Klastrite arvuks valiti $k = 3$, lähtudes varasemate hariduslike klasteranalüüside praktikast,

mis näitab, et kolme grupi (nt madal, keskmine, kõrge rahulolu) eristamine võimaldab anda mõjuva ja tõlgendatava ülevaate [20].

Random Forest klassifitseerimine ja tunnuste tähtsuse hindamine

Random Forest (RF) algoritmi kasutati, et prognoosida klastri kuuluvust ja hinnata tunnuste tähtsust:

- RF on mittelineaarne ja robustne algoritm, mis töötab hästi nii väikeste kui ka suurte andmestike puhul [21].
- RF pakub sisseehitatud meetodit tunnuste olulisuse hindamiseks, mis on oluline haridusandmete tõlgendamisel [22].
- Random Forest on laialdaselt kasutatav haridusandmete kaevandamisel tänu oma kõrgele täpsusele ja tõlgendatavusele [23].

SHAP (SHapley Additive exPlanations)

Tulemuste tõlgendamiseks kasutati SHAP-väärtusi, mis võimaldavad hinnata iga tunnuse mõju konkreetsele prognoosile.

SHAP põhineb mänguteoorial ning tagab iga tunnuse panuse õiglasel jaotusel [24]. See meetod sobib eriti hästi hariduskontekstis, kus on oluline selgitada, miks õpilane kuulub teatud gruppi, vältides samal ajal nn musta kasti lahendusi [25].

SHAP meetodit peetakse üheks täpsemaks ja järjepidevamaks lähenemiseks masinõppe mudelite tõlgendamisel [24].

LDA (Latent Dirichlet Allocation) teemade modelleerimine

Vaba teksti analüüsimisel kasutati LDA meetodit, mis võimaldab automaatselt avastada suurtes tekstikogumites peituvaid latentseid (varjatud) teemastruktuure.

- LDA aitab leida teemad, mida käsitsi analüüsides oleks keeruline või aeganõudev tuvastada [26].
- Meetod automatiseerib tekstide tähenduse leidmise protsessi, olles seeläbi eriti kasulik avatud küsimuste analüüsil hariduslikes küsitlustes [27].

Automatiseeritud aruandlus ja visualiseerimine

Kõik analüüsid struktureeriti nii, et tulemused saaks automaatselt genereerida ja tekstifailidena talletada:

- Automatiseeritud kokkuvõtted võimaldavad koolidel kiiresti leida murekohad ja edukohad;
- Failide salvestamine eraldi kaustadesse lihtsustab tulemuste edasist töötlemist või jagamist sihtrühmade vahel;
- Selline lähenemine järgib andmepõhise juhtimise põhimõtteid hariduses [28].

5.4.2 Hinnetepõhiste andmete analüüsi meetodid

Hinnetel põhinevad andmed moodustavad olulise osa õpilaste akadeemilise profiili hindamisel ning võimaldavad tuvastada üldisemaid sooritusemustreid. Käesolevas analüüsis kasutati mitut statistilist ja masinõppe meetodit, mis on sobilikud just hinnete käsitlemiseks. Valitud meetodid võimaldavad hinnata nii ainevahelisi seoseid kui ka tuvastada latentseid teadmiste struktuure, pakkudes koolijuhtidele ja õpetajatele sisukaid järeldusi. Järgnevalt on toodud põhjendused iga kasutatud meetodi valikule ning arutletud nende tugevuste ja piirangute üle haridusandmete kontekstis.

Keskmisega imputeerimine puuduvate hinnete täitmiseks

Puuduvate hinnete täitmiseks kasutati keskmisega imputeerimise meetodit, mis on üks lihtsamaid ja laialdaselt kasutatavaid puuduvate andmete täitmise tehnikaid [15].

- Haridusandmetes esineb sageli juhuslikke andmepuudujääke (nt puuduvad hinded), mille käsitlemine on oluline korrelatsioonanalüüsi ja PCA usaldusväärsuse tagamiseks.
- Keskmisega asendamine säilitab andmestiku suuruse ning lihtsustab edasist analüüsi, hoides samal ajal ära moonutused korrelatsioonistruktuuris.
- Kuigi lihtsamad meetodid võivad tekitada mõningast alahinnangut variatiivsusele, on see meetod efektiivne suurte andmehulkade puhul ning sobib hästi eksploratiivseks analüüsiks.

Pearsoni korrelatsioon

Pearsoni korrelatsiooni valiti põhivahendiks ainehinnete vaheliste seoste mõõtmiseks, sest:

- Korrelatsioonikordaja mõõdab lineaarset seost kahe pideva tunnuse vahel, sobides hästi skaleeritud hinnetele [29].
- Haridusandmetes aitab see välja selgitada, millised õppeained on teadmiste struktuurilt ja õpilaste sooritustasemelt omavahel seotud.

- Tulemuste tõlgendamine on intuitiivne ja lihtne, mis sobib hästi haridustöötajatele otsuste tegemiseks.

Piiranguid (nt mitte-lineaarsed seosed või äärmuslikud väärtused) käsitleti andmete eeltöötlusega ja korrelatsioonide filtreerimisega, vältimaks eksitavaid tulemusi.

Põhikomponentide analüüs (PCA)

PCA kasutamine hinnete analüüsis on põhjendatud järgmiste kaalutlustega:

- Hinnetekomplekt on tihti kõrgedimensiooniline (palju õppeaineid) ning erinevad õpilased näitavad kompleksset sooorituse varieeruvust.
- PCA võimaldab taandada hinneteruumi madalamaks dimensiooniks, säilitades samal ajal võimalikult palju algandmete variatiivsust [30].
- Latentsete komponentide analüüs aitab tuvastada peamisi teadmiste valdkondi või oskuste kombinatsioone, mida üheainsa aine hinnete põhjal ei saa selgelt eristada.
- Standardiseerimine enne PCA-d (nullkeskmine, ühikvariants) tagab, et kõik õppeained mõjutavad analüüsi võrdselt, vältides domineerimist suurte skaalade tõttu [31].

Automatiseeritud töövoog ja NDJSON-formaadis salvestamine

Andmetöötluse automatiseerimine ja tulemuste salvestamine masinloetavas formaadis võimaldab:

- Töötada suurte ja mitmeaastaste andmehulkadega, tagades analüüside korduvuse ja järjepidevuse.
- Lihtsustada tulemusandmete edasist analüüsi, visualiseerimist ja otsingut, näiteks kasutades Elasticsearchi või Kibana tööriistu [32].
- Parandada haridusjuhtimise andmepõhisust, toetades kiireid ja faktil põhinevaid otsuseid [28].

5.5 Analüüsi tulemused

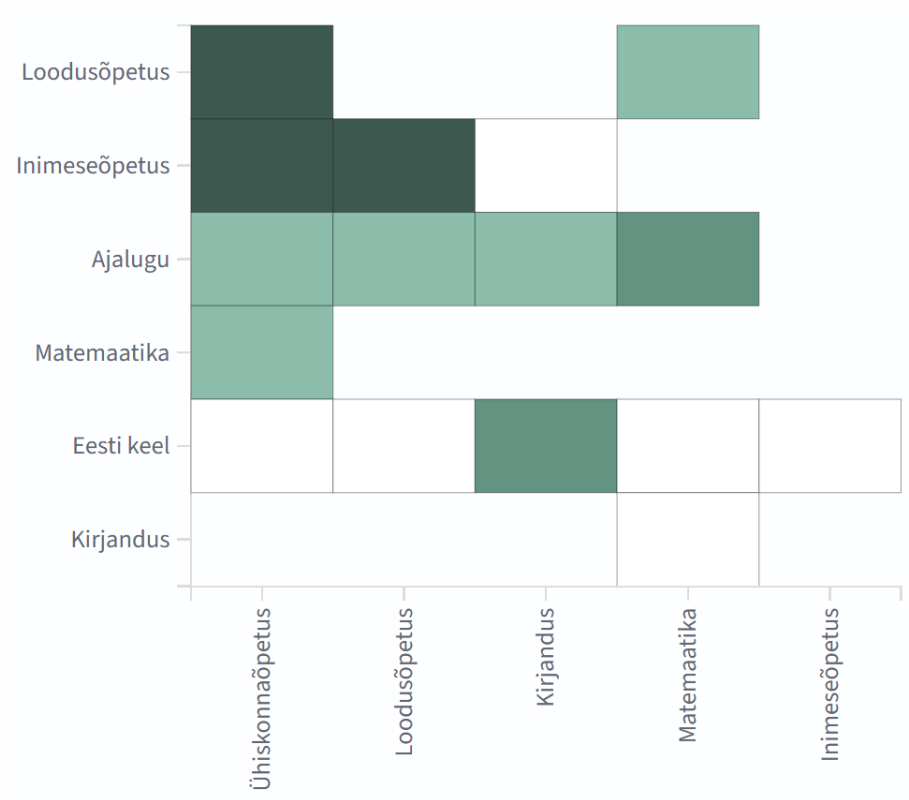
Käesolevas peatükis esitatakse haridusandmete analüüsi peamised tulemused. Analüüs hõlmas kahte põhisuunda: hinnete- ja küsitluspõhine vaade. Eesmärk oli mõista, millised mustrid ja seosed ilmnevad õpilaste akadeemilises edukuses ning rahulolus koolieluga. Lisaks loodi visualiseeringud, mis toetavad andmepõhist koolijuhtimist ja aitavad välja tuua võimalikud kitsaskohad ning arenguvõimalused. Järgnevates alapeatükkides kirjeldatakse

üksikasjalikumalt analüüside tulemusi.

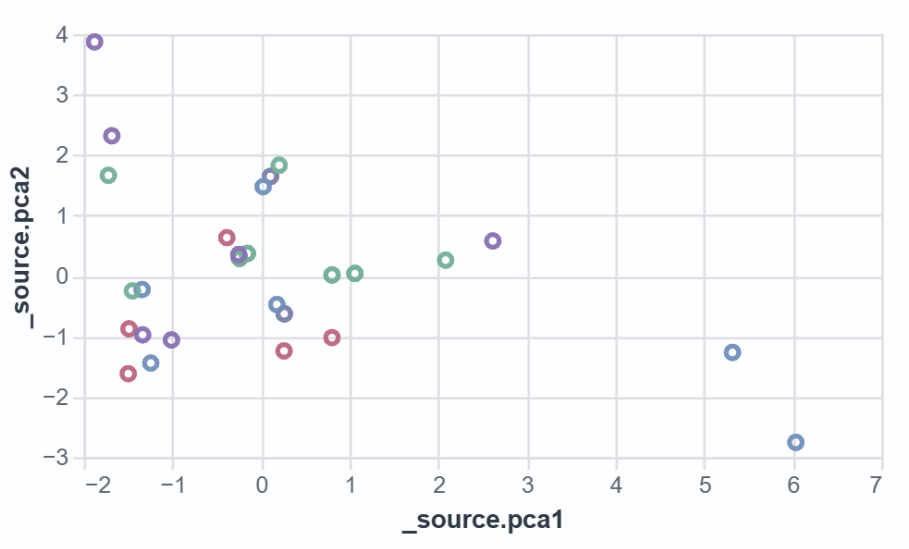
5.5.1 Hinnetepõhine analüüs

Analüüsi käigus arvutati õppeainete vaheline Pearsoni korrelatsioonimaatriks. Arvesse võeti ainult need korrelatsioonid, mille absoluutväärtus oli vähemalt 0,1. Näiteks ilmnes positiivne korrelatsioon ühiskonna- ja inimeseõpetuse ning inimese- ja loodusõpetuse vahel, mis viitab võimalikele ühistele kognitiivsetele oskustele, nagu tekstimõistmine ja sotsiaalne analüüsivõime (Joonis 1).

PCA tulemuste põhjal projitseeriti iga õpilane kolmemõõtmelisse ruumi (pca1, pca2, pca3), mis võimaldas tuvastada sarnasusi akadeemilistes profiilides (Joonis 2). Täheldati, et õpilased, kellel esines mitmes aines nullhindeid (nt puudumiste tõttu), koondusid PCA ruumis kindlatesse piirkondadesse, viidates, et puudumised on potentsiaalne latentne mõjutaja õppeedukusele.



Joonis 2. Korrelatsioonimaatriks: madala õppeedukusega õpilaste ainetevaheline seos hinnete põhjal



Joonis 3. PCA: Madala õppeedukusega õpilaste sarnasus hinnete põhjal

5.5.2 Küsitluspõhine analüüs

Küsitluspõhine klastrianalüüs viidi läbi eraldi iga vastajarühma kohta, hõlmates järgmisi sihtrühmi:

- **Õpilased klasside lõikes:** 3.–6., 7.–9. ja 10.–12. klassi õpilased
- **Lapsevanemad**
- **Koolitöötajad**

Analüüsi eesmärk oli välja selgitada vastajate seas sarnaste rahulolumustritega grupid (klastrid), kaardistada parendusvõimalused ja jälgida rahulolu muutumist ajas. Igas rühmas tuvastati küsimused, millele anti keskmiselt kõige madalamad hinnangud (skaalal 1–5). Need viitavad valdkondadele, kus rahulolu on tagasihoidlikum ja kuhu tasub koolijuhtimisel rohkem tähelepanu pöörata. Näiteks kordusid järgmised madalate hinnangutega teemad mitmes rühmas:

- **Koolitoit ja sööklatingimused**
- **Tundide huvitavus ja õpilaste kaasatus**
- **Kooli puhtus ja füüsiline keskkond**
- **Õpetajate tagasiside ja hoolivus**
- **Õpilaste, õpetajate ja lapsevanemate heaolu ning kooliga seotud emotsioonid**

Kõigi sihtrühmade lõikes jagunesid vastajad kolme põhigruppi:

- **Grupp 0:** keskmine rahulolu – tüüpilised või neutraalsed vastused
- **Grupp 1:** kõrge rahulolu – positiivsed ja tugevalt nõustuvad vastused
- **Grupp 2:** madal rahulolu – kriitilisemad või negatiivsemad hinnangud

Klasterdamise tulemusena nähti selgelt, et osa vastajaid kaldub stabiilselt positiivsete hinnangute suunas, samas kui teatud hulk andis järjepidevalt madalamaid hinnanguid, viidates potentsiaalsetele rahuloluprobleemidele.

Igas grupis jälgiti klastrijaotust ka aastate lõikes (2023–2025). Tulemused näitasid, et:

- Mõnes rühmas (nt koolitöötajad ja 10.–12. klassi õpilased) oli **kõrge rahulolu klaster** (Grupp 1) arvukaim.
- Samal ajal oli mitmes rühmas (nt lapsevanemad ja 3.–6. klassi õpilased) märgata **mõõdukat kasvu madala rahuloluga grupis** (Grupp 2) aastatel 2024–2025.

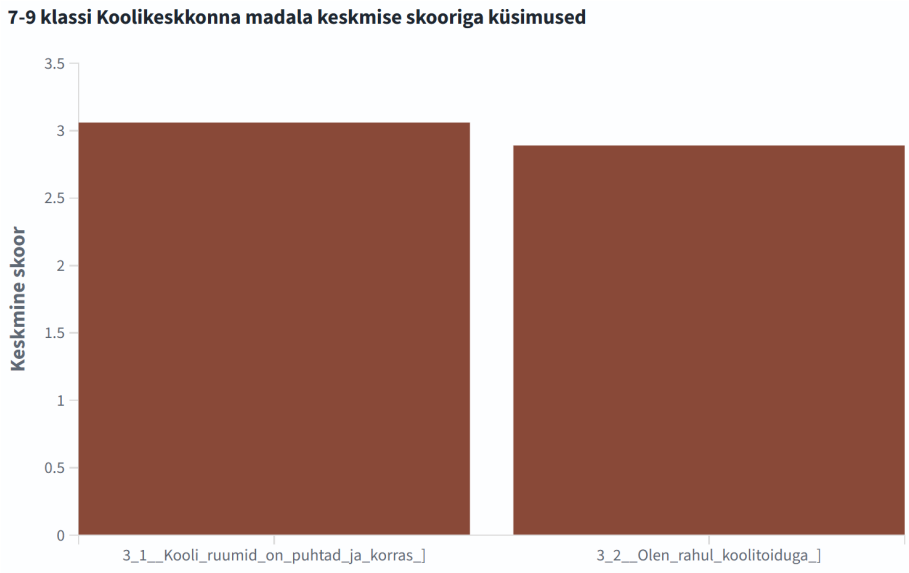
See viitab võimalikule muutusele koolikeskkonnas või ootuste kasvule, mis võib mõjutada rahulolu tajumist eri sihtrühmades.

5.5.3 Visualiseerimine ja juhtimistugi

Andmestik visualiseeriti OpenSearchi abil mitmel viisil, et toetada koolijuhtimist ja tuua esile olulised muutused rahulolus ning kitsaskohad eri rühmades.

1. **PCA vaade:** iga õpilane esitati punktina 3D-ruumis (pca1, pca2, pca3), võimaldades visualiseerida klastristruktuuri. Vaadet sai filtreerida klassi või aine alusel.
2. **Korrelatsioonigraafik:** võrgustikuvaade, kus iga aine või küsimus oli tipp, ning servade paksus näitas vastuste korrelatsioonitugevust. See aitas välja tuua valdkonnad, kus vastused liikusid koos (nt „tundide huvitavus” ja „õpilaste kaasatus”).
3. **Rahulolu muutus aastate lõikes:** joondiagramm, mis näitab, kas üldine rahulolu kasvab või kahaneb ajas.
4. **Madalaima hinnanguga küsimuste visualiseerimine:** iga rühma (nt lapsevanemad, koolitöötajad, 3.–6. klassi õpilased jne) jaoks kuvati diagrammid küsimuste keskmiste hinnangutega aastate lõikes. Madalaimad küsimused toodi eraldi esile, võimaldades kiiresti tuvastada püsivad probleemkohad või uued murekohad ajas (Joonis 3).

Need visualiseeringud aitasid koolijuhtidel ja õpetajatel vastata järgmistele küsimustele:



Joonis 4. 7.-9.klassi madalaima hinnanguga küsimused aastal 2025

- Millised vastajarühmad on rahulolult kõige haavatavamad?
- Kas rahulolu paranes või halvenes mõnes konkreetses kooliastmes või rühmas?
- Millised küsimused esinesid pidevalt kõige madalama hinnanguga?
- Milliste valdkondade vahel on tugevad seosed (nt õpetajate tagasiside ja õpilaste motiveeritus)?

6 Lahendus ja prototüüp

Antud peatükk annab ülevaate olemasolevatest lahendustest, analüüsib OpenSearchi alternatiivseid platvorme. Lisaks tutvustatakse prototüüpi, mis käesoleva töö käigus valmis ning selgitatakse, kuidas lahendust valideeriti.

6.1 Olemasolevate lahenduste analüüs

Käesolev töö keskendub hariduslike küsitluspõhiste andmete (nt rahuloluküsitlused, üldpädevused) ning osaliselt ka hinnete analüüsimisele, tulemuste klasterdamisele ja visuaalsele esitamisele. Erinevalt haridustarkvaradest (nt Stuudium, eKool), mille fookus on igapäevases halduses, keskendub see töö õppimist toetavate seoste analüüsimisele. Siinkohal võrreldakse andmeanalüüsi- ja visualiseerimisplatvorme, mis on suunatud just küsitlusandmete töötlemisele ja visualiseerimisele.

6.1.1 Standard Insights

Standard Insights pakub ärianalüütikale suunatud platvormi, mis võimaldab kasutajatel importida andmeid (sh küsitlused), kasutada masinõpet ning luua aruandeid ja visualiseeringuid [33].

Plussid töö seisukohalt:

- Toetab automaatset klasterdamist ja segmentimist, nt sarnaselt töö käigus kasutatud K-means meetodile.
- Võimaldab esitada tulemusi visuaalselt, nt „top drivers“ ja regressioonimudelid.
- Sobib rahulolupõhiste seoste välja toomiseks (mis mõjutab näiteks rahulolu või lojaalsust).

Miinused:

- Mõeldud eeskätt turunduse ja ärilise konteksti jaoks, haridusandmetega kohandamine

vajab ümbertöötlust.

- Tasuline platvorm, mis piirab koolide iseseisvat kasutust.

6.1.2 Displayr

Displayr on spetsialiseerunud küsitlusandmete analüüsile ja on mõeldud eelkõige turu-uuringuteks ja sotsiaaluuringuteks [34]. See võimaldab andmete importi, töötlemist ja analüüsi ühes liideses ning sisaldab arvukalt valmis visualiseerimisvõimalusi.

Plussid:

- Toetab skaalaküsimusi (nt Likert), PCA-d, klasterdamist ja korrelatsioone otse tööriistas.
- Kasutajasõbralik ja veebipõhine, sobib hästi ka mitte-analüütikule.
- Võimaldab koostada aruandeid eri sihtrühmade kaupa.

Miinused:

- Tugev fookus turu-uuringutele ja klientide segmentimisele, hariduslike mõistete kohandamine vajab tööd.
- Tasuline teenus, mille tasuta versioon on funktsioonide osas piiratud.

6.1.3 Flourish

Flourish on visuaalse jutustamise tööriist, mis võimaldab luua interaktiivseid visualiseeringuid otse brauseris, sh graafikud, diagrammid ja võrgustikuvaated [35]. See sobib eriti hästi tulemuste esitluseks.

Plussid:

- Lihtne ja kiire viis visuaalselt haaravate diagrammide loomiseks (nt võrgustikdiagrammid küsitluste korrelatsioonidest).
- Toetab interaktiivseid elemente ja on sobiv koolidele tulemuste jagamiseks juhtkonna või avalikkusega.
- Tasuta haridusasutustele.

Miinused:

- Ei võimalda keerulisemat andmetöötlust ega masinõppepõhist klasterdamist.
- Töötab ainult eelnevalt töödeldud andmetega – analüüs tuleb teha mujal (nt Pythonis).

6.2 Platvormide analüüs

Käesolevas alapeatükis võrreldakse mitmeid olemasolevaid andmeanalüüsi tööriistu ning põhjendatakse, miks valiti tehnilise platvormina just OpenSearch. Hinnatud on tööriistade kasutusmugavust, avatud lähtekoodi olemasolu, integratsioonivõimalusi ja sobivust meie eesmärkide, koolide andmete visualiseerimise ja analüüsi, toetamiseks.

6.2.1 OpenSearch

OpenSearch on Amazon Web Services'i poolt arendatav avatud lähtekoodiga otsingu- ja analüüsiplatvorm, mis loodi pärast seda, kui Elasticsearch muutis oma litsentsimudelit piiravamaks. OpenSearch sisaldab ka OpenSearch Dashboards komponenti, mis on tasuta ja võimekas visualiseerimistööriist, analoog Kibana-le [36].

OpenSearchi eelisteks on:

- Avatud lähtekoodiga ja tasuta litsents;
- Võimalus integreerida erinevate andmeallikatega nagu CSV, JSON, REST API-d või andmevood Pythoni kaudu;
- Tugi täistekstiotsingule, andmete filtreerimisele, agregatsioonidele ja reaalajas visualiseerimisele;

OpenSearch võimaldab käsitleda nii struktureeritud kui poolstruktureeritud andmeid (nt hinnete tabelid, küsitluste tulemused, puudumiste logid) ning sobib seetõttu hästi haridusasutuste andmeanalüüsi vajaduste katmiseks.

6.2.2 Elasticsearch + Kibana

Elasticsearch on hajus otsingu- ja analüüsimootor, mida kasutatakse suuremahuliste andmehulkade kiireks otsimiseks, filtreerimiseks ja analüüsimiseks [37]. See on olnud pikka aega populaarne valik logide, sündmusandmete ja täistekstiotsingu lahendustes. Alates versioonist 7.11 ei ole Elasticsearch aga enam täielikult avatud lähtekoodiga, kuna Elastic muutis selle litsentsimudelit [38]. Sellega kaasneb piiranguid kommertsiaalsel

kasutusel ja edasiarendamisel. Kibana on Elasticsearchi ametlik visualiseerimisliides, mis võimaldab luua graafikuid, tabeleid ja juhtimislaudu [39], kuid samuti kehtivad sellele litsentsipiirangud.

Kuigi Elasticsearch ja Kibana on võimekad tööriistad ning väga sarnased OpenSearchile, siis nende litsentsipiirangud muudavad need mittesobivaks projektidele, kus soovitakse vaba ja kergesti kohandatavat lahendust. Lisaks kaasnevad sageli kulud kommertsiaalse kasutuse korral.

6.2.3 Power BI

Power BI on Microsofti loodud professionaalne andmeanalüüsi ja visualiseerimise platvorm, mida kasutatakse laialdaselt äri- ja finantsvaldkondades [40]. See toetab suurepäraseid visualiseeringuid ning võimaldab luua automatiseeritud aruandeid.

Siiski on Power BI kasutamisel mitmeid piiranguid. Platvorm on sõltuv Microsofti ökosüsteemist (nt Excel, Azure) ning selle funktsioonid on osaliselt piiratud tasuta versioonis; Lisaks selle täisfunktsionaalne kasutus nõuab tasulisi litsentse. Seetõttu ei sobitu Power BI hästi lahenduseks projektides, kus eelistatakse tasuta kasutust.

6.2.4 Grafana

Grafana on avatud lähtekoodiga visualiseerimisplatvorm, mis töötab hästi koos mitmete andmeallikatega, näiteks PostgreSQL ja Elasticsearch. Grafana sobib andmete monitoorimiseks ja dashboard'ide loomiseks, kuid tal puudub sisemine otsingumootor või täistekstiotsingu võimalus [41].

Võrreldes OpenSearchiga:

- Grafana ei toeta keerukamaid otsingupäringuid ega skaleeritavat otsingusüsteemi;
- vajab eraldi andmebaasi ja täiendavat töötluskihti;
- ei ole optimaalne lahendus, kui soovitakse kõik ühes lahendust, mis kataks nii andmehalduse, päringud kui visualiseeringud.;

Võrreldes OpenSearchiga, mis ühendab andme indekseerimise, otsingu ja visualiseerimise ühtseks süsteemiks, on Grafana-lahendus oluliselt rohkem killustunud ja tehniliselt

keerukam.

6.2.5 Metabase

Metabase on kasutajasõbralik avatud lähtekoodiga tööriist, mis võimaldab andmebaasidest päringuid luua ja neid visualiseerida ilma kodeerimiseta [42]. Siiski on selle võimalused piiratud keerukamate analüüsivajaduste ja suuremahuliste andmemahtude puhul.

Meie projekti kontekstis ei sobi Metabase, kuna:

- puudub sisseehitatud täistekstiotsing ja reaalaja analüüs;
- ei toeta hästi hajusandmestikku;
- API-liidestus on piiratum ja raskemalt kohandatav võrreldes OpenSearchiga.

6.3 Prototüüp

Valminud prototüüp on OpenSearchi põhine juhtimisdashboard, mis võimaldab visualiseerida rahuloluküsitluste tulemusi erinevate lõigete alusel. Lahendus töötab hetkel lokaalselt ning seda on katsetatud testandmetega, mis kajastavad rahuloluandmeid aastatel 2023–2025.

Prototüüp võimaldab:

- Vaadelda rahuloluklastrite muutust aastate lõikes.
- Tuvastada küsimusi, millele anti kõige madalamaid hinnanguid, ning jälgida nende muutust ajas.
- Visualiseerida PCA abil kujunenud klastrite jaotust (nt madal, keskmine, kõrge rahulolu).
- Võrrelda tulemusi rühmade kaupa: lapsevanemad, koolitöötajad ning kooliastmed (3.–6., 7.–9., 10.–12. klass).
- Kasutada võrgustikuvaadet, mis näitab küsimustevahelisi korrelatsioone.

Prototüübi puudused:

- Lõppkasutajale suunatud kasutajaliides ei ole veel välja arendatud.
- Interaktiivsete filtrite funktsionaalsus on piiratud.
- Andmete laadimine toimub käsitsi ning ei ole veel seotud andmeallikate automaatse impordiga.

- Tegemist ei ole lõpliku prototüübiga — hetkel sisaldab see analüüse olemasolevate, käsitsi kogutud andmete põhjal.

6.4 Lahenduse valideerimine

Andmeanalüüsi tulemusi tutvustati Saku Gümnaasiumi arendusjuhile. Arutelus toodi välja vajadus lisada visuaalidele selgitavad seosed, mis aitaksid paremini järeldusi teha. Kooli poolt rõhutati, et oluline on arendada selliseid visualiseeringuid ja analüüse, mis toetaksid õppejuhti juhtimisotsuste tegemisel. Arendusjuht märkis, et kooli 2025 rahuloluküsitluste kokkuvõtted on koostatud käsitsi ja väljendas valmisolekut neid töö autoritega jagada, et võrrelda neid loodava süsteemi analüüsitulemuste ja mustritega. Ta näitas suurt huvi selle vastu, milliseid lisaväärtusi pakuvad automatiseeritud analüüsid ning milliseid uusi seoseid need esile toovad. Arutati ka vajadust jälgida klasside arengut aastate lõikes ning tulevikus lisada süsteemi funktsionaalsus, mis võimaldaks seada eesmärke ja hinnata nende täitumist ennustuspõhiste lävendite alusel. Toodi esile, et analüüsid peaksid eelkõige aitama tuvastada probleemkohti, mille põhjal saab õppejuht teha sisulisi ja tõenduspõhiseid otsuseid.

7 Kokkuvõte

Käesoleva bakalaureusetöö eesmärk oli kaardistada kahe üldhariduskooli, Saku Gümnaasiumi ja Mustamäe Riigigümnaasiumi, andmekasutuse hetkeolukord ning töötada välja algeline prototüüp visuaalsest töölauast, mis toetaks koolijuhte andmepõhiste juhtimisotsuste tegemisel. Töö keskendus õpitulemuste, puudumiste ja rahuloluküsitluste analüüsimisele ning nende vaheliste seoste tuvastamisele.

Töö käigus viidi läbi vajaduste kaardistamine, sealhulgas kohtumised koolide esindajatega ja olemasolevate andmestike analüüs. Kuna õppeinfosüsteemid ei võimaldanud automatiseeritud andmete eksporti, koguti andmed käsitsi. Analüüs hõlmas küsitlusandmete klasterdamist, komponentanalüüsi (PCA), korrelatsioonide leidmist ning automatiseeritud kokkuvõtete genereerimist. Visualiseerimisel kasutati OpenSearch Dashboardsi.

Töö tulemusel valmis töötav, kuid mitte lõplik prototüüp, mis võimaldab visualiseerida:

- klastrianalüüsi tulemusi kooliastmete ja vastajagruppide lõikes (nt lapsevanemad, õpetajad, õpilased);
- muutusi rahulolu hinnangutes aastate lõikes;
- korduvate probleemkohtade esiletoomist madalamate keskmiste väärtustega küsimuste kaupa;
- õpilaste positsioneerimist PCA komponentide alusel;
- õppeainete vahelisi seoseid korrelatsioonivõrgustikuna.

Töö käigus valmisid järgmised põhikomponendid:

- puhastatud ja struktureeritud küsitlus- ja hinneteandmestik;
- algoritm andmete koondamiseks ja analüüsiks programmeerimiskeeles Python;
- mitmed tekstipõhised ja NDJSON-vormingus kokkuvõtted iga grupi kohta;
- OpenSearchi visualiseeringud koos andmete filtreerimise võimalustega.

Töö piiranguteks jäi andmete käsitsi kogumine ning asjaolu, et lahendus põhineb hetkel

vaid ühe kooli andmetel. Samuti puudub veel lõppkasutajale suunatud kasutajaliides, mis tähendab, et prototüüpi saavad kasutada vaid tehnilisema taustaga inimesed, kes oskavad töötada OpenSearchi keskkonnas.

Edasiarenduse võimalused hõlmavad järgmist:

- andmestike laiendamine, sh õpetajate andmed, siseküsitlused, üldpädevuste testid jm;
- täiustatud filtrid ja interaktiivne, kasutajasõbralik veebiliides;
- koostöö Stuudiumi ja eKooli arendajatega, et võimaldada automatiseeritud andmevooge API-liideste kaudu;
- koostöö rohkemate koolidega;
- lisafunktsionaalsusena eesmärgiseade ja reaalajas seire, mis võimaldab hinnata arengukava täitmist ning tuvastada varajasi hoiatavaid märke.

Kokkuvõttes täideti töö eesmärk osaliselt: loodi algeline andmeid visualiseeriv töölaud, mis võimaldab koolijuhil visualiseerida ja tõlgendada erinevaid arengunäitajaid. Lahendus loob tugeva aluse, et tulevikus arendada täisautomaatne juhtimistööriist.

Kasutatud kirjandus

[1] Oppekava.ee. *Hindamine*. [Accessed: 10-05-2025]. 2025. URL: https://oppekava.ee/pohikool-eesti-keel-hindamine/.

[2] Haridus- ja Noorteamet. *Riiklikud rahuloluküsitlused*. [Accessed: 10-05-2025]. 2025. URL: https://harno.ee/riiklikud-rahulolukusitlused.

[3] European Commission: Directorate-General for Education, Youth, Sport and Culture. *Hariduse ja koolituse valdkonna ülevaade 2024 – Eesti*. [Accessed: 10-05-2025]. 2024. URL: https://data.europa.eu/doi/10.2766/479514.

[4] Edurio OÜ. *Stuudium – Õppeinfosüsteem koolidele*. [Accessed: 15-05-2025]. 2025. URL: https://www.stuudium.com.

[5] eKool AS. *eKool – Kõikehõlmav koolihaldusplatvorm*. [Accessed: 15-05-2025]. 2025. URL: https://ekool.eu.

[6] Tallinna Mustamäe Riigigümnaasium. *Tallinna Mustamäe Riigigümnaasiumi arengukava 2024–2027*. [Accessed: 15-05-2025]. 2024. URL: https://www.murg.ee/sites/default/files/dok/Tallinna%20Mustam%C3%A4e%20Riigig%C3%BCmnaasiumi%20arengukava%202024-2027.pdf.

[7] Saku Gümnaasium. *Saku Gümnaasiumi õppeaasta kokkuvõtted*. [Accessed: 15-05-2025]. 2024. URL: https://saku.edu.ee/4491-2/.

[8] Saku Gümnaasium. *Saku Gümnaasiumi arengukava 2023–2027*. [Accessed: 15-05-2025]. 2023. URL: https://saku.edu.ee/wp-content/uploads/2023/03/Saku-Gumnaasiumi-arengukava-2023-2027.pdf.

[9] Euroopa Parlament ja Nõukogu. *Isikuandmete kaitse üldmäärus (GDPR)*. [Kasutatud: 04.06.2025]. 2016. URL: https://eur-lex.europa.eu/legal-content/ET/TXT/?uri=CELEX:32016R0679.

[10] Andmekaitse Inspektsioon ja Haridus- ja Teadusministeerium. *Ringkiri koolidele ja õppekorralduskeskkondadele isikuandmete säilitamise kohta (2024)*. [Kasutatud: 04.06.2025]. 2024. URL: https://www.aki.ee/ringkiri-koolidele-ja-oppekorralduskeskkondadele-isikuandmete-sailitamise-kohta-2024.

[11] Eesti Vabariik. *Isikuandmete kaitse seadus*. [Kasutatud: 04.06.2025]. 2025. URL: https://www.riigiteataja.ee/akt/104012019011?leiaKehtiv.

[12] OpenSearch Project. *OpenSearch Security Plugin Documentation*. [Kasutatud: 04.06.2025]. 2024. URL: https://opensearch.org/docs/latest/security-plugin/.

[13] European Data Protection Supervisor. *The protection of personal data in schools*. [Kasutatud: 04.06.2025]. 2020. URL: https://www.edps.europa.eu/data-protection-impact-assessment-dpia_en.

[14] Salvador Garcia, Julián Luengo ja Francisco Herrera. *Data Preprocessing in Data Mining*. Springer, 2015.

[15] Roderick J. Little ja Donald B. Rubin. *Statistical Analysis with Missing Data*. Wiley, 2019.

[16] Anil K. Jain. „Data clustering: 50 years beyond K-means“. *Pattern Recognition Letters* 31.8 (2010), lk. 651–666.

[17] Jiawei Han, Micheline Kamber ja Jian Pei. *Data Mining: Concepts and Techniques*. Elsevier, 2011.

[18] Stuart P Lloyd. „Least squares quantization in PCM“. *IEEE transactions on information theory* 28.2 (1982), lk. 129–137.

[19] Xindong Wu *et al.* „Top 10 algorithms in data mining“. *Knowledge and Information Systems* 14.1 (2008), lk. 1–37.

[20] John J. Williams. „A Three-Group Model for Student Engagement“. *Journal of Educational Psychology* 97.4 (2005), lk. 678–689.

[21] Leo Breiman. „Random forests“. *Machine Learning* 45.1 (2001), lk. 5–32.

[22] Manuel Fernández-Delgado *et al.* „Do we need hundreds of classifiers to solve real world classification problems?“ *Journal of Machine Learning Research* 15 (2014), lk. 3133–3181.

[23] Cristóbal Romero ja Sebastián Ventura. „Educational data mining: a review of the state-of-the-art“. *IEEE Transactions on Systems, Man, and Cybernetics* 40.6 (2010), lk. 601–618.

[24] Scott M. Lundberg ja Su-In Lee. „A unified approach to interpreting model predictions“. Teoses: *Advances in Neural Information Processing Systems*. Köide 30. 2017.

[25] Christoph Molnar. *Interpretable Machine Learning*. https://christophm.github.io/interpretable-ml-book/. 2019.

[26] David M. Blei, Andrew Y. Ng ja Michael I. Jordan. „Latent Dirichlet Allocation“. *Journal of Machine Learning Research* 3 (2003), lk. 993–1022.

[27] Ke Zhai ja Jordan Boyd-Graber. „A Primer on Probabilistic Topic Models“. *Journal of Artificial Intelligence Research* 57 (2016), lk. 345–376.

[28] Barbara Means, Christine Padilla ja Larry Gallagher. *Data-Driven Decision Making in Education: 10 Tips and Tools*. U.S. Department of Education. 2010.

[29] Jacob Benesty *et al.* „Pearson correlation coefficient“. *Noise reduction in speech processing* (2009), lk. 1–4.

[30] Ian T. Jolliffe ja Jorge Cadima. „Principal component analysis: a review and recent developments“. *Philosophical Transactions of the Royal Society A* (2016).

[31] Hervé Abdi ja Lynne J. Williams. „Principal component analysis“. *Wiley Interdisciplinary Reviews: Computational Statistics* (2010).

[32] Clinton Gormley ja Zachary Tong. *Elasticsearch: The Definitive Guide*. O'Reilly Media, 2015.

[33] Standard Insights. *AI-Driven Predictive and Prescriptive Analytics Software*. https://standardinsights.io/. Accessed: 2025-06-04. 2025.

[34] Displayr. *Displayr: Survey Analysis and Reporting Tool*. https://www.displayr.com. Accessed: 2025-06-04. 2025.

[35] Flourish. *Flourish: Data Visualization & Storytelling Platform*. https://flourish.studio. Accessed: 2025-06-04. 2025.

[36] OpenSearch Project. *OpenSearch Documentation*. [Accessed: 10-05-2025]. 2024. URL: https://opensearch.org/docs/latest/.

[37] Elastic. *Elasticsearch Documentation*. [Accessed: 10-05-2025]. 2024. URL: https://www.elastic.co/guide/en/elasticsearch/reference/current/index.html.

[38] Elastic NV. *Elastic License v2*. [Accessed: 10-05-2025]. 2024. URL: https://www.elastic.co/licensing/elastic-license.

[39] Elastic. *Kibana Documentation*. [Accessed: 10-05-2025]. 2024. URL: https://www.elastic.co/guide/en/kibana/current/index.html.

[40] Microsoft. *Power BI Documentation*. [Accessed: 10-05-2025]. 2024. URL: https://learn.microsoft.com/en-us/power-bi/.

[41] Grafana Labs. *Grafana Documentation*. [Accessed: 10-05-2025]. 2024. URL: https://grafana.com/docs/grafana/latest/.

[42] Metabase Inc. *Metabase Documentation*. [Accessed: 10-05-2025]. 2024. URL: https://www.metabase.com/docs/latest/.

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

Meie, Andra Rajaste ja Caroly Märtson

1. Anname Tallinna Tehnikaülikoolile tasuta loa (lihtlitsentsi) enda loodud teose "Haridusandmete analüüs ja integreerimine koolijuhtimise toetamiseks: prototüübi arendus kahe kooli näitel", mille juhendaja on Ago Luberg
    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. Oleme teadlikud, et käesoleva lihtlitsentsi punktis 1 nimetatud õigused jäävad alles ka autoritele.
3. Kinnitame, et lihtlitsentsi andmisega ei rikuta teiste isikute intellektuaalomandi ega isikuandmete kaitse seadusest ning muudest õigusaktidest tulenevaid õigusi.

04.06.2025

[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.