Projekti üldinfo

Projekti nimi: Vaimse tervise riskide ennustamine

Tiimi liikmed: Kadi Tulver, Rauno Tali, Tõnn Sikk

Github repo: https://github.com/tondza/PAT2023_Menta/

Projekti eesmärk lühidalt on kasutada andmeteaduse vahendeid, et luua ennustusmudel vaimse tervise probleemide esinemise ohuks teatud valimi puhul erinevaid parameetreid arvestades.

Äriline arusaam

Taustainfo

Vaimne tervis on tõusnud tähelepanu huviorbiiti eriti pärast COVID-19 pandeemiat ja sellega kaasnenud isolatsioonimeetmeid. Pandeemia käigus täheldati vaimse tervise halvenemist ja depressiooni süvenemist paljudes erinevates sotsiaalsetes gruppides (alates koolinoortest kuni vanurite ja kontoritöötajateni). TAI ja Tartu Ülikooli poolt läbi viidud rahva vaimse tervise uuringu põhjal on ligi veerandil täiskasvanutest vähemalt üks psüühikahäire diagnoos (Kokkuvõttev pressiteade, sisaldab ka kogu lõpparuande linki: https://www.tai.ee/et/uudised/uuring-levinuimad-vaimse-tervise-probleemid-depressioon-ja-arevushaired). Eriti levinud probleemideks on muuhulgas depressioon ja läbipõlemine, millele antud projekti raames keskendume. Kuna vaimse tervise abi kättesaadavus Eestis on üsna piiratud, on parim viis probleemiga tegelemiseks püüda neid ennetada. Selleks on oluline mõista, millised tegurid võivad olla vaimse tervise häirete riskiteguriteks ja milliste tunnuste kaudu on võimalik neid kõige paremini ennustada.

Meie projekti tiimi liikmetest kaks (Tõnn ja Rauno) on tegevad IT valdkonnas, kus on viimasel ajal suurenenud tähelepanu töötajate vaimse tervise hoidmisele nii koolituste kui kompenseeritavate teenuste näol. Kuna IT valdkonnas on kodukontorist töötamine keskmisest lihtsam ning pärast pandeemiat endiselt väga levinud, on paraku sagedased ka juhtumid, kus inimesed jäävad oma vaimse tervise muredega üksi. Seega võiks projekti raames tehtud analüüsidest olla praktiline kasu nii tiimi liikmetele kui ka kõigile neile, kellega seda infot pärast jagatakse, parandades seeläbi inimeste teadlikkust vaimse tervise probleemide riskiteguritest Tiimi kolmas liige (Kadi) on hariduselt psühholoog ning oma töö kontekstis ka varem töökollektiividele vaimse tervise teemal infot jaganud. Otsene kokkupuude selliste andmetega ning selgem arusaam võimalikest riskiteguritest just Eesti rahvastiku kontekstis aitaks seda infot edaspidi veelgi paremini näitlikustada ja levitada.

Eesmärgid

- Prognoosida üksikisikute tõenäosust vaimse tervise probleemide esinemiseks/tekkimiseks erinevate parameetrite põhjal, nagu vanus, sugu, elukutse, tervisekäitumine jne.
- Juhtida tähelepanu vaimse tervise teemale kursusel osalejate ning oma töökollektiivi hulgas

Edukuse mõõdikud

- Loodud mudeli prognoosimise täpsus on (precision) vähemalt 85% testandmestiku puhul

- Tuvastada vähemalt kolm olulist vaimse tervise probleeme ennustavat tegurit, mis on tavainimesele arusaadavad (lihtsalt inimese poolt interpreteeritav mudel)
- Tähendusrikas võrdlus erinevate inimgruppide vahel (Eesti andmestiku ning mõne muu riigi analoogse andmestiku vahel või erinevate vanusegruppide vahel)
- Visualiseerida olulisemad tulemused kaasahaaraval ja hoomataval moel

Olukorra hinnang

Ressursside ülevaade:

- Kursuse “Praktiline andmeteadus” põhjal omandatud teadmised ja oskused andmestiku puhastamiseks, andmeanalüüsiks, andmete visualiseerimiseks ja ennustusmudelite loomiseks
- Orienteeruvalt 22 tundi tööaega tiimiliikme kohta
- Iga tiimiliikme enda tööjaam/arvuti ja selle arvutusvõimsus
- Vabalt saadavad andmeteaduse tööriistad nagu Jupyter Notebooks ning Pythoni teegid nagu Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn jne.
- ChatGPT Plus subscription
- Ligipääs avalikele anonümiseeritud andmebaasidele

Nõuded, eeldused ja piirangud:

- Nõue: Projekti lõpparuanne peab olema esitatud hiljemalt 17. novembriks
- Nõue: Kuna vaimne tervis on delikaatne teema, peab projektitiim arvestama kõigi isikuandme kaitse ja andmekogude kasutamise tingimustega.
- Eeldus: Projektitiim saab ligipääsu andmestikule, mis sisaldab andmeid nii üksikisikute vaimse tervise probleemide (või nende puudumise) kohta kui ka muude isiku omaduste kohta, mis võivad üksi või kombineeritult mõjutada vaimse tervise ohuteguritena (nt vanus, sugu, perekonnaseis, amet jms).
- Eeldus: Andmestik esindab (vähemalt osaliselt) ka Eesti elanikkonda.
- Piirang: Projektitiim peab leppima saadaolevate andmekogude piirangutega (uue andmekogu loomine scrape’imise vms tehnikaga ei tundu antud teema puhul reaalne).

Riskid ja alternatiivplaanid:

- Risk: Avalikult kättesaadavad andmekogud ei vasta projekti eeldustele
    - Alternatiivplaan: Pöörduda otse võimalike andmekogude omanike poole (nt kontaktid Geenivaramus, Tervise Arengu Instituudis või Peaasi lehel); projekti läbiviimine fiktiivsete andmetega (nt Kaggle);
- Risk: Eraviisilised kokkulepped võimalike andmekogude omanikega vajalike ligipääsude saamiseks (nt Geenivaramu, Peaasi.ee) on liiga aeganõudvad ning lükkavad edasi projekti alustamise aega.
    - Alternatiivplaan: Kui on veendumus vajaliku andmekogu olemasolus ning põhimõtteline nõusolek selle kasutamiseks omaniku poolt, siis võib projekti alustades kasutada ka näidisandmeid / fiktiivseid andmeid, mille struktuur on sarnane päris andmetega (ja hiljem andmed visualiseerida / mudelid treenida päris andmetega)

- Risk: Projektiliikmete muud kohustused ja prioriteedid võivad mõjutada võimekust projekti panustada (panuse vähenemisest kuni loobumiseni)
    - Alternatiivplaan: võimaluse korral probleeme ennetada jaotades tiimiliikmete panust nende ajaressursse arvestades; 1 tiimiliikme loobumise puhul saavad 2 liiget jätkata, vähendades vajadusel skoopi. 2 tiimiliikme loobumise puhul või koostöö raskuste korral saab nõu pidada aine korraldajatega ning valida kitsam uurimisküsimus, mida lahendada individuaalselt.

Terminoloogia:

Vaimne tervis - WHO definitsiooni järgi: heaoluseisund, milles inimene realiseerib oma võimeid, tuleb toime igapäevaelu pingetega, suudab töötada tootlikult ja tulemusrikkalt ning on võimeline andma oma panuse ühiskonna heaks
Psüühikahäire - kliiniliselt äratuntav kogum vaimse tervisega seotud sümptomitest või käitumisviisidest, millega enamasti kaasneb oluline stress ja mis häirib isiku funktsioneerimist
Risk - statistiline hinnang häire või sündmuse tõenäosusele
Läbipõlemine - Maailma terviseorganisatsioon WHO kirjeldab läbipõlemist, kui sündroomi, mis tekib kroonilise tööga seotud stressi tagajärjel, mida ei ole edukalt maandatud. Läbipõlemist kirjeldatakse kolme dimensiooni kaudu: kurnatus, võõrdumine, Vähenenud töö- ja tegutsemisvõime.
Depressioon - püsivalt väljendunud meeleolu alanemine, millega kaasneb elurõõmu kadumine, energia vähenemine ning mille tulemusena langeb toimetulekuvõime ja elukvaliteet.

Allikad:
Eesti rahvastiku vaimse tervise uuring. Lühikokkuvõte, 2023. (https://tai.ee/sites/default/files/2023-03/RVTU_lyhikokkuvote_2023.pdf)
Peaasi koduleht (peaasi.ee)

Kulud ja tulud:

Kulud: Projekti läbiviimisel pole ette näha rahalisi kulusid. Ressursikulud hõlmavad peamiselt aega, mis kulub andmetest arusaamise, puhastamise, analüüside läbiviimise ja raporti kirjutamise peale.

Tulud: Väärtus tiimi liikmetele õpetliku kogemusena päris andmetega töös, tähendusrikaste järelduste tegemisel andmete põhjal. Potentsiaalne väärtus ka paremas arusaamas vaimse tervise probleemide ja nende ennetamise kohta ning nende teadmiste levitamine kursuse teiste liikmete ning oma tuttavate ja töökollektiivi seas.

Andmekaevandamise eesmärgid

Eesmärgiks on ehitada ennustavad mudelid, mis suudavad piisava täpsusega hinnata indiviidi riski vaimse tervise häire tekkimiseks.

Andmekaevandamise edu hindamisel arvestada andmestiku kvaliteeti ja asjakohasust ning mudeli täpsust (mudeli headuse näitajate põhjal, lisaks arvestades kuivõrd on tulemused kooskõlas varasemate uuringutulemustega võimalike riskifaktorite kohta).

Andmete mõistmine

Andmete kogumine

Andmevajaduste kirjeldamine

Andmestik peaks sisaldama demograafilisi näitajaid nagu vanus ja sugu, töökäitumisega seotud tegureid nagu tööroll, töötunnid, tööstaaž, stressitase ning üldiseid hinnanguid tervisekäitumisele. Andmestik peaks kindlasti sisaldama ka andmeid vaimse tervise seisundi kohta, võimaluse korral ka hinnanguid sümptomitele või häire diagnoose, mida ennustada.

Andmete kättesaadavus

Esialgne uurimine näitab, et on saadaval mõningad vaimse tervisega seotud avalikud andmestikud. Siiski on töökohaga seotud depressiooni ja läbipõlemise andmestikud piiratud. Lihtsamini leitavad on Euroopa või Ameerika kontekstis kogutud andmed, kuid kuna oleme huvitatud ka Eestis tehtud uuringutest, on vajalik edasine uurimine ja kontakt potentsiaalsete andmete valdajatega.

Valikukriteeriumide määratlemine

Valitud andmestik peab olema aktuaalne, põhjalik ja esindama sihtrühma. Puuduvad väärtused peaksid olema minimaalsed ja andmestik peaks sisaldama muutujaid, mis on otseselt või kaudselt seotud töökohaga ning mis võimaldaksid ennustada vaimse tervisega seotud näitusid.

Andmete uurimine ja kirjeldamine

Projekti eesmärkide kohaselt soovime leida andmestikku, mis oleks kogutud Eestis viimaste aastate jooksul ning oleks võimalikult suure valimiga. Kuni sellist andmestikku otsime, valisime ülesande nõuete täitmiseks Kaggle'ist andmestiku, mis kajastab 2016. aastal erinevate riikide IT töötajate seas läbi viidud vaimse tervise küsimustikku (https://www.kaggle.com/datasets/osmi/mental-health-in-tech-2016). Selles andmestikus on 1433 vastajat ja 63 erinevat küsimust, kus on talletatud muuhulgas olulisemad demograafilised näitajad (sugu, vanus), erinevad hinnangud töökeskkonnale, vaimsele tervisele ja lisaks muud taustainfot. Mõned huvipakkuvad küsimused, mis võiksid osutuda relevantseks, kui hakkame valima muutujaid, mida ennustusmudelisse kaasata on näiteks: pereliikmete seas leiduvad vaimse tervise probleemid, tööandja hoiakud vaimse tervise küsimuste suhtes, ligipääs vaimse tervise ressurssidele ja negatiivsed kogemused töökeskkonnas. Küsimus on ka selle kohta, kas inimene tegeleb kaugtööga, kuna kaugtöö kogemus pandeemia kontekstis tundus olevat üheks riskiteguriks tööstressi kogemisel.

Oleks huvitav vaadata, kas ka pandeemia eelselt võib see negatiivseid tagajärgi ennustada. Valdavalt on selles andmestikus raporteeritud vaimse tervise diagnoosidena ärevushäireid ning meeleoluhäireid (nt depressioon), mõnel juhul ka stressiga seotud probleeme ning sõltuvushäiret.

Andmete kvaliteet

Andmestikus on palju puuduvaid andmepunkte, kuna kõik inimesed pole igale küsimusele vastanud. Juhul kui puuduvaid andmeid on mõne olulise muutuja lõikes liiga suur osa, siis võib see ennustusmudeli kvaliteedi langetada. Samuti on mitmed vastused kodeeritud ebaühtlaselt, seega tuleb need viia uutele skaaladele, et teha tähendusrikkaid üldistusi ja analüüse. Valimi suurus 1433 peaks olema piisav, et ehitada lihtsamaid mudelid, kuid võib jääda liiga väikseks, kui tahaksime kasutada keerulisemaid masinõppe mudeleid näiteks tehisnärvivõrke.

Projekti planeerimine

1. Eesmärkidele sobiva andmestiku leidmine - andmestike otsimine, andmete sobivuse ja kvaliteedi hindamine, kontakteerumine võimalike andmehalduritega
    - Hinnanguline ajakulu liikme kohta: 2h
    - Meetodid ja töövahendid: Google, valdkonnasisesed kontaktid
2. Uurimistöö teema sisu ja tausta kohta
    - Hinnanguline ajakulu liikme kohta: 2h
    - Meetodid ja töövahendid: Google scholar, Tervise Arengu Instituudi koduleht (tai.ee), Peaasi koduleht (Peaasi.ee), ChatGPT
3. Andmetega tutvumine, andmete puhastamine, eeltöötlus, kirjeldav statistika, meetodite valik
    - Hinnanguline ajakulu liikme kohta: 4h
    - Meetodid ja töövahendid: Python, R, Praktilise andmeteaduse kursuse materjalid, ChatGPT
4. Võrdlev andmeanalüüs, andmetabelite ja tulemusi illustreerivate jooniste loomine
    - Hinnanguline ajakulu liikme kohta: 4h
    - Meetodid ja töövahendid: Python, R, Praktilise andmeteaduse kursuse materjalid, ChatGPT
5. Mudelite treenimine ja testimine
    - Hinnanguline ajakulu liikme kohta: 4h
    - Meetodid ja töövahendid: Python, Praktilise andmeteaduse kursuse materjalid, ChatGPT
6. Tulemuste kokkukirjutamine ja visualiseerimine
    - Hinnanguline ajakulu liikme kohta: 3h
    - Meetodid ja töövahendid: Python, R, ChatGPT
7. Lõppraporti koostamine
    - Hinnanguline ajakulu liikme kohta: 2h
    - Meetodid ja töövahendid: Microsoft Word, Google Scholar

8. Projekti tutvustamiseks ettekande koostamine ja selle presenteerimine
    Hinnanguline ajakulu liikme kohta: 1h
    Meetodid ja töövahendid: PowerPoint, Microsoft Word

Hinnanguline ajakulu liikme kohta kokku: 22h