Projekt “OPENDATA ESTONIA - Estonia's culture ministry's grant applications”

Sissejuhatus andmeteadusesse (LTAT.02.002)

Anneli Klamas
Anett-Kristin Palmar
Enrih Sinilaid

Tartu
2019

Task 2

Taust

Meie klient, kes tegeleb ürituste planeerimise ja korraldamisega, on juba pikemat aega mõelnud sellele, et millised üritused saavad või ei saa riigilt toetust korraldamiseks. Selleks on nad eelnevalt töötanud erinevate firmadega, kes on proovinud luua algoritme või andmemudeleid, mis oskaksid ennustada kui suure tõenäosusega etteantud üritus toetust saaks. Kahjuks aga ei õnnestunud ükski varasem koostöö projekt. Peamiseks probleemiks oli andmete vähesus erinevat tüüpi üritustest, kui suurt rahastust sooviti ja kui palju ka lõpuks anti, kui üldse. See probleem aga leidis lahenduse eelmise aasta suvel, kui Eesti Kultuuriminister hakkas avalikustama kõiki andmeid selle kohta, mis üritustele rahastust andi ja millistele ei antud.
Kuna Kultuuriminister täiendab seda andmestiku igapäevselt, siis meie klient ootas natuke rohkem kui aasta, et oleks kogunenud piisavalt andmeid, et nendega midagi peale hakata. Siis korraldati konkurss, mille jooksul selgitati välja milline ettevõte nendega koostööd tegema hakkab. Konkursi võitsime meie, Firma T18, ja nüüd on meie ülesandeks välja töötada keskkond või andmemudel, mille põhjal saaksime ennustada, et kui suure tõenäosusega etteantud üritus rahastust saaks. See oleks esimene etapp. Kui esimene etapp õnnestub, siis minnakse kohe järgmisele etapile, milles luuakse tavakasutajale lihtsasti kasutatav keskkond, kuhu saaks etteantud ürituse andmed sisestada ja siis tagastatakse kasutajale ennustus.
Me leiame, et suudame mõlemad etapid läbida edukalt ja koostöö lõpuks esitada kliendile produkt, mis vastaks tema nõuetele ja soovidele.

Eesmärgid

Meie eesmärgiks on võtta andmestik, milles leiduvad erinevad üritused ja kas need on edukalt toetuse saanud või ei. Lisaks veel see, kui palju toetust sooviti ja mis kategooria alla see üritus langes.
Selle andmestiku põhjal loome andmemudeli, mis suudaks ennustada, kas etteantud üritus saaks rahastust või mitte. Kuna andmestik on reaalajas täienev, siis mudel peab ka koos andmestikuga arenema.
Lisaks sellele on meie eesmärgiks veel luua rakendus või kasutajaliides, mis teeks andmete sisestuse ennustamiseks tavakasutajale kergeks.

Loeme projekti õnnestunuks kui:

- Ennustus on korrektne 90% ajast
- Andmemudel on võimeline ise arenema väga vähesel määral spetsialisti sekkumisega
- Kasutajaliides töötab sujuvalt ja on kergesti mõistetav ja kasutatav

Inventuur

- Andmestik toetustest json formaadis
- Jupyter Notebook
- Android Studio
- Andmestiku kontaktisiku email: indrek.reimand@kul.ee
- Õppejõud

Nõuded

- Plakati tähtaeg: 19 detsember.
- Projekti nõuded: Võime ennustada piisava täpsusega toetuse saamist, võime andmeid lisada ja muuta, kasutajaliidese kasutamine

Riskid:

- Võib juhtuda, et ennustamine ei ole piisavalt täpne suurem osa ajast

Terminoloogia:

Toetus	Abistavad ressursid riigilt, mis aitaksid projekti õnnestumisele kaasa.
Menetlus	Protsess, mille jooksul tehakse kindlaks, kas riik toetab projekti või mitte. Ja kui toetab, siis millises koguses.
Ennustatav prognoos	Projekti andmete sisestamisel antakse tõenäosus, mis ütleb seda, kui suure tõenäosusega projekt saaks toetust riigilt
Kasutajaliides	On visuaalne komponent, millesse tavakasutaja saab sisestada kergesti andmeid, mille põhjal tehakse ennustus ja siis tagastatakse tagasi kasutajale

Andmestik	On kogu korduvatest taotlustest, kus on välja toodud projekt nimi, kategooria, taotluse suurus, menetluse staatus, otsus, antud rahastus jne.
Andmemudel	On treenitud mudel andmestikust, mis suudaks näiteks ennustada andmestiku põhjal, et millise tõenäosusega projekt toetust saab.

Maksumus:

2 nädala jooksul töötab iga töötaja selle projekti kallal 30h. See teeb kokku 90h tööd ja tunnimaksvusega 15 eurot/tund läheks projekt maksma 1350 eurot. Hind on hinnakirjas arvestatud, seega firma ei jää selle projektiga miinustesse.

Andmetöötlus

Eesmärgid:

- Töötav ja 90% ajast õigesti ennustav andmemudel.
- Andmemudel on isearenev ja nõuab vähe tähelepanu spetsialistidelt.
- Poster ja andmeliides olemas 19. detsembriks.

Õnnestunud andmetöötlus:

Loeme andmetöötlust õnnestunuks, kui andmemudel suudab ennustada piisava täpsusega ja muutub aja jooksul aina täpsemaks.

Task 3

Andmete kogumine

Nõuded:

Vajame andmeid erinevate ürituste taotlustest, kus oleks välja toodud, millises koguses toetust sooviti, kuidas seda menetleti ja mis otsustati, kui palju rahastust anti, kui kaua otsuse jõustumiseks kulus ja mis kategooria alla üritus langes.

Juurdepääsetavus:

Kuna tegemist on Eesti kultuuriministeeriumi poolt avalikustatud andmetega, siis on ligipääs nendele kõigil.

Valimise kriteerium:

Andmed on avalikustatud leheküljel: https://opendata.riik.ee/andmehulgad/toetuste-menetlemise-infos-steem/, kus andmestik on JSON formaadis. Sellest andmestikust kasutame kõiki välju välja arvatud nime välja, sest meie jaoks on tähtis ürituse kategooria, taotlus, menetluse tulemus, rahastus ja protsessi kestvus.

Limiteerivad tegurid:

Hetkese seisuga ei ole me veel leidnud ühtegi limiteerivat tegurit.

Kokkuvõtvalt kogumisest:

Meie projekti õnnestumiseks vajame hetkese seisuga kultuuriministreeriumi avalikustatud ja pidevas uuendamises olevat andmestiku erinevate ürituste taotluste kohta, mida on võimalik kõigil kasutada ja mille leiab riigi opendata andmebaasist. See andmestik on juba töötlemiseks soodsaks failiformaadis, mis tähendab, et me ei pea kulutama ajalist ressursi andmestiku kohendamisega. Hetkel on andmestikus välja toodud üle 1500 erinevat taotlust, mis on piisav meie projekti jaoks.

Andmete seletus

Andmestikus on järgmised väljad:

Võti	Seletus
id	Taotlust identiteeriv kood, mis on unikaalne
application_code	Dokumendi kood on unikaalne väärtus, mis on igal dokumendil erinev
applicationround_title	Dokumendi tiitel
approved_amount	Kui palju rahastust anti
cost_statement_submission_date	Kuupäev mil anti aruanne maksumustest
cost_statement_submission_deadline	Kuupäev, mille ajaks peab aruanne maksumustest esitatud olema
managing_organization_name	Organisatsioon, kes projektiga tegeleb
name	Organisatsiooni nimi
project_name	Projekti nimi, millele taotlust soovitakse
registration_date	Taotluse registreerimise kuupäev
registry_code	Unikaalne registreerimise kood
requested_amount	Kui suures koguses rahalist toetust sooviti
status	Taotluse staatus
status_txt	Taotluse staatus kirje kujul
submission_date	Taotluse esitamise kuupäev
domain_code	Kood, mis näitab millisesse domeeni taotlus kuulub
domain_name	Kategooria

Andmetest lähemalt:

Võti	Kuju
id	Integer tüüpi unikaalne suurus
application_code	Formaadis x.x.x/y-x, kus x on number ja y on erisuurses arv
applicationround_title	Nimetus String formaadis
approved_amount	Integer tüüpi suurus
cost_statement_submission_date	Date tüüpi suurus kujul: yyyy-mm-dd hh:mm:ss
cost_statement_submission_deadline	Date tüüpi suurus kujul: yyyy-mm-dd hh:mm:ss
managing_organization_name	Nimetus String formaadis
name	Nimetus String formaadis
project_name	Nimetus String formaadis
registration_date	Date tüüpi suurus kujul: yyyy-mm-dd hh:mm:ss
registry_code	Unikaalne integer tüüpi suurus
requested_amount	Integer tüüpi suurus
status	Erinevad väärtused: evaluate_ok, committee, committee_ok, committee_not_ok, decision_ok, decision_not_ok
status_txt	Erinevad väärtused: Menetluses, Otsus jõustunud
submission_date	Date tüüpi suurus kujul: yyyy-mm-dd hh:mm:ss
domain_code	Kood, mis näitab millisesse domeeni taotlus

	kuulub
domain_name	Kategooria

Andmete kvaliteet:

Me hindasime andmete kvaliteedi heaks, sest andmestikust on meil saadaval kõik vajalikud komponendid, mida me vajame oma projekti täitmiseks

Task 4

Planeering:

Ülesanded	Kellele	Ajaline maht
Esmane andmetöötlus JSON formaadist masintöötlemiseks soodsaks formaati	Kõik	10h
Andmete treenimine	Kõik	30h
Kasutajaliidese loomine	Kõik	15h
Kasutajaliidese ühendamine treenitud andmemudeliga	Kõik	20h
Plakati tegemine	Kõik	5h
Testimine ja parandused	Kõik	10h

Kuna kõik panustavad iga ülesande puhul, siis panustatav ajakulu on enamvähem võrdne. Ei välista võimalust, et projekti jooksul võib planeering senisest muutuda kas ülesannete, ülesande lahendajate või ülesande ajalise mahu poolest, aga kõik muudatused jäädvustatakse ja projekti lõpus kajastatakse. Arvestatakse aga sellega, et iga liikme ajaline maht on vähemalt 30h.

Kasutatavad tööriistad:

Projekti jooksul peame kasutama erinevaid tööriistu ja meetodeid, et üht või teist teha. Andmete sisselugemiseks, töötlemiseks, treenimiseks ja muutmiseks kasutame keelt Python, kus kasutame erinevaid mooduleid, mis hõlbustaksid eelnevalt nimetatud tegevuste lahendamist. Keskkond, mida kasutame, on Jupyter Notebook.
Kasutajaliidese loomiseks kasutame aga Java tööriista Android Studio, kus on olemas kõik vajalik, et luua nutiseadmel jooksev kasutajaliides.