Baigiamųjų projektų temos

Dalykas: Data Science
Projekto trukmė: Birželio 17 - Liepos 10 dienos (56 akademinės valandos)
Projekto rezultatų pristatymas: Liepos 15-17 dienomis vyks projektų rezultatų pristatymas kur kiekvienas studentas pristatys tiek projektą ir rezultatus, tiek pasirinktą darbo metodiką. Po pristatymo vyks aptarimas. Pristatymui paruoškite skaidres, taip pat projekto kodą įdėkite į viešą github repositoriją (nepamirškite pakomentuoti README.md)

Problema #1: paieška pagal nuotrauką (arba reverse image search)

Problema
Įsivaizduokite - gatvėje pamatote taip ilgai ieškotą paltą. Raštas, ilgis, rankovės ir medžiaga - viskas atitinka jūsų norus, tačiau spėjate paltą tik nufotografuoti. Grįžus namo, tikitės pagal nuotrauką rasti šį paltą populiariausiame e-commerce puslapyje, tačiau turite tik nuotrauką ir žinote, kad naudojantis tekstine paieška paltą surasti yra misija (ne)įmanoma.

Užduotis
Užduoties tikslas yra sukurti ML algoritmą kuris atliktų paiešką tik pagal nuotrauką ir grąžintų labiausiai panašius kandidatus į ieškomą paltą (ar kitą rūbą). Kitaip tariant - paduodant vieno palto nuotrauką, algoritmas grąžintų keleto kandidatų nuotraukas.

Galimas sprendimas
Pasiūlyti duomenų rinkiniai:

- Clothing Dataset [*proof-of-concept* rinkinys]
- DeepFashion Dataset [*Consumer-to-shop Clothes Retrieval benchmark* rinkinys]

Daugelis *reverse image search (RIS)* sprendimų remiasi tuo, kad konvoliucinis neuroninis tinklas gali būti panaudotas gauti nuotraukos požymius (vadinamus *embeddings/features*). Vėliau tik reikia įvertinti kurie iš duombazėje turimų *embeddings* yra labiausiai panašūs į ieškomos nuotraukos *embedding* ir tam dažniausiai naudojama *cosine similarity*. Atminkite, kad PoC dalyje nuotraukos požymių generavimui galima naudoti jau išmokytą CNN modelį (pavyzdžiui naudoti modelį jau išmokytą klasifikuoti ImageNet duomenų rinkinį) ir šis modelis leis rasti kandidatus pagal spalvą ar rašto informaciją, tačiau jei norėsite rasti kandidatus pagal tikslesnius požymius kaip rankovės ilgis ar kategorija - nuotraukos požymių gavimo (*feature extractor)* modelį reikės *fine-tunint*.

Uždavinį pradėkite nuo veikimo schemos kūrimo ir vėliau apsvarstykite koks jūsų problemos *scope* - ar kursite RIS įrankį skirtą paltų paieškai tarp visų galimų rūbų, ar bet kokio viršutinio drabužio paieškai, o galbūt jūsų problemos scope bus tik paltai, batai ir maikutės?

Daugiau informacijos apie potencialius RIS sprendimus galima rasti čia (žiūrėkite *Reverse image search for industrial applications*) ir čia.

Problema #2: prognozuoti ar e.recepte išrašytas vaistas bus įsigytas

Problema
Tik dalis išrašytų e.receptų yra sėkmingai įsigyti. Priežasčių tam gali būti ypač daug ir, pavyzdžiui, sprendimą įsigyti vaistą gali lemti faktoriai kaip gyvenamoji vieta, ligos sunkumas ar gydytojo kvalifikacija.

Užduotis
Sukurti ML algoritmą skirtą prognozuoti ar klientas sėkmingai įsigys jam paskirtą vaistą.

Galimas sprendimas
Pasiūlyti duomenų rinkiniai:
- e.recepto atvirieji duomenys [skelbiami kas mėnesį]

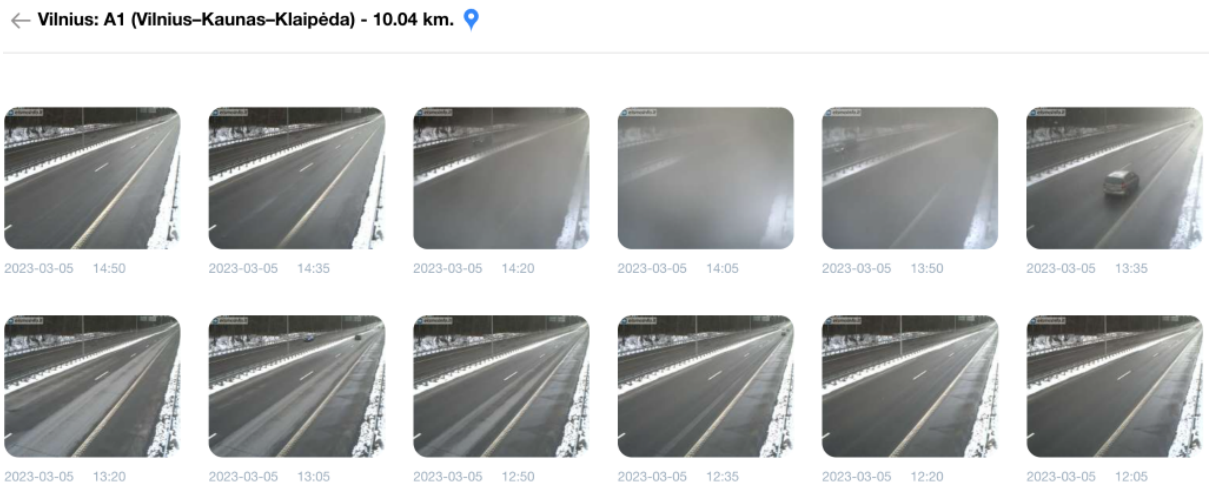
Problema panaši į turėtas paskaitų metu - susipažinti su duomenimis ir išvalyti duomenų rinkinį, apgalvoti faktorius galinčius turėti įtakos *target* kintamąjam ir sukurti *scikit-learn pipeline* prognozavimui atlikti. Sukurtą modelį *padeployinkite*.

Nepamirškite išradingai įvertinti kokie dar faktoriai gali turėti įtakos (pavyzdžiui, galbūt vaisto populiarumas yra geras faktorius prognozuojant ar vaistas bus įsigytas ir šis įvertis gali būti aproksimuotas per parduotų vaistinėms preparatų skaičių), taip pat ko trūksta, kad pasiektumėte puikų modelio tikslumą.

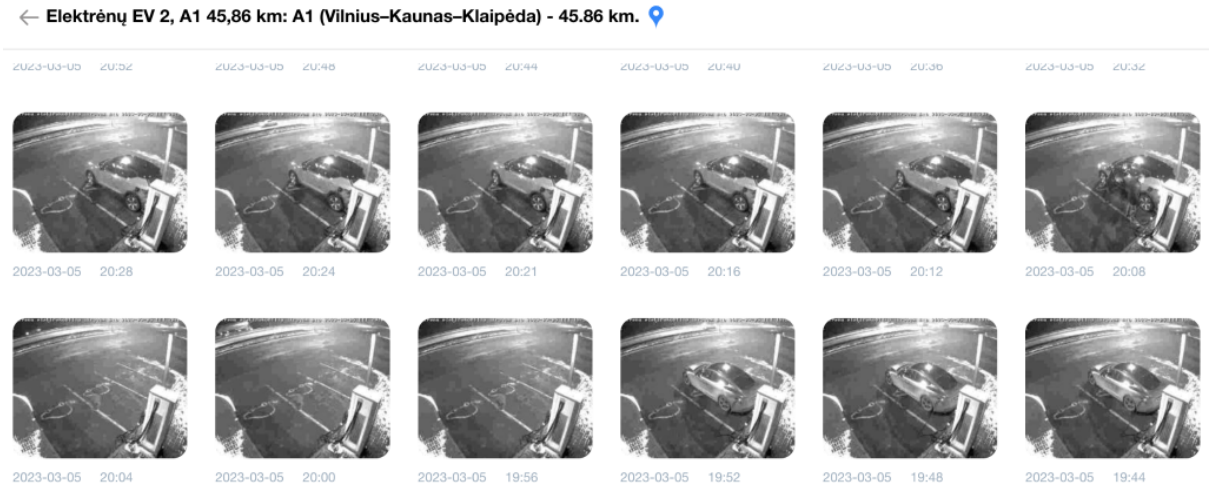
Problema #3: automobilių aptikimas paveikslėliuose

Problema
Lietuvos keliuose yra įrengtos vaizdo kameros, kurios stebi eismą. Ši informacija svetainėje https://eismoinfo.lt/ yra pateikia beveik realiu laiku. Sistemos trūkumas - joje nepateikiama informacija ar nuotraukoje yra fiksuojama transporto priemonė ar ne.

Užduotis
Sukurti modelį, kuris gebėtų iš eismo stebėjimo įrenginių nuotraukų prognozuoti ar juose vaizduojama transporto priemonė ar ne.



Galima užduotį siaurinti ir pasirinkti pavyzdžiui tik elektromobilių įkrovimo vietų stebėjimo fotokameras.



Duomenys

Šiai užduočiai atlikti paruoštų duomenų rinkinio nėra. Jį reikėtų pasiruošti patiems. Visos nuotraukos ir eismo stebėjimo archyvas gali būti pasiekiamas per svetainę https://eismoinfo.lt/ ir pasirinkus “Vaizdo kameros”. Ten pasirinkus konkrečią kamerą galima matyti laike padarytus kadrus.

Galimas sprendimas

Kadangi duomenų rinkinys neegzistuoja, tad visų pirma reiktų susikurti nuotraukų surinkimo įrankį (scraper). Susirinkus nuotraukas, jas susižymėti (labelling), pavyzdžiui nuotraukoje egzistuoja automobilis žymime 1, jei neegzistuoja - 0. Turint pasiruoštą duomenų rinkinį pritaikyti transfer learning ir fine tuning tam, jog sukurtume klasifikavimo modelį.

Duomenų surinkimas

1. Pasiimam kamerų sąrašą iš https://eismoinfo.lt/eismoinfo-backend/camera-info-table. “id” nurodo kameros ID, o “image” - paskutinią padarytą nuotrauką.

2. Norint gauti kiekvienos kameros nuotraukas formuojama sekanti užklausa: https://eismoinfo.lt/eismoinfo-backend/camera-info-table/KAMEROS_ID?pageNumber=0&pageSize=NUOTRAUKU_SKAICIUS, pavyzdžiui: https://eismoinfo.lt/eismoinfo-backend/camera-info-table/1167?pageNumber=0&pageSize=50, paimks 1167 kameros 50 paskutinių nuotraukų. Nuotraukos ID yra “id” laukelis. Laukas “date” užkoduotas kaip UNIX timestamp. Norint gauti nuotrauką reikia paimti “image” lauke nurodytą nuorodą ir iš jos parisiųsti nuotrauką, pavyzdžiui: https://www.eismoinfo.lt/eismoinfo-backend/image-provider/camera/old?id=178750915. Nuotraukos adresas suformuotas taip: https://www.eismoinfo.lt/eismoinfo-backend/image-provider/camera/old?id=NUOTRAUKOS_ID.
3. Pagrindinė biblioteka reikalingi darbui atlikti: https://pypi.org/project/requests/. Jos pagalba galėsit daryti užklausas į eismoinfo.lt kad pasiimtumėt duomenis, taip pat išsaugoti nuotraukas į failus.

Problema #4: automobilių kainos prognozavimas

Problema
Kaip teisingai nustatyti savo automobilio kainą jį parduodant? Kaip nuspręstar perkamo automobilio kaina adekvati?

Užduotis
Sukurti modelį, kurio pagalba būtų galima prognozuoti automobilio kainą pagal tam tikrus jo parametrus. Užduotį galima susiaurinti iki konkretaus gamintojo automobilių.

Duomenys
Duomenys yra saugomi automobilių portaluose, tokiuose kaip www.autoplius.lt, www.autogidas.lt, www.autobilis.lt ir pan.. Norint sukurti duomenų rinkinį - reikia duomenis scrapinti.

Galimas sprendimas
Sukuriamas duomenų surinkimo įrankis (scraper), kurio pagalba sukuriamas apmokymo duomenų rinkinys. Turint duomenis - sprendžiamas regresijos uždavinys. Nustatomi parametrai, kurie bus naudojami kainos nustatymui, atliekamas duomenų išvalymas/ paruošimas, tada atliekamas duomenų modeliavimas.

Problema #5: BVP prognozavimas

Problema
Kaip teisingai nustatyti, kurios šalys labiausiai “paaugo”?

Užduotis

1. Surasti top 10 šalių, kurios labiausiai paaugo “Gross domestic product per capita” atžvilgiu.
2. Nubrėžti grafikus, kurie iliustruotų, kaip keitėsi šalių populiacija iš **The Organization for Economic Cooperation and Development** (OECD).
3. Suskirstykite šalis į 5 klasterius naudodamiesi GDP ir "Volume of exports of goods".
4. Sukurkite modelį, kuris prognozuoja “Gross domestic product per capita”. Būkite atidūs ir nenaudokite laukų, kurie tiesiogiai susiję su GDP.
5. Supaprastinkite 4 punkte sukurtą modelį taip, kad jis būtų kuo tikslesnis ir turėtų ne daugiau 5 kintamųjų (features).

Duomenys
Excel failas
https://www.imf.org/-/media/Files/Publications/WEO/WEO-Database/2022/WEOOct2022all.ashx

Duomenų apibrėžimas
https://www.imf.org/en/Publications/WEO/weo-database/2022/October/download-entire-database

Galimas sprendimas
Susipažinkite su duomenimis. Išvalykite arba užpildykite trūkstamas reikšmes. Nustatykite kokio tipo tai yra uždavinys - klasifikavimo ar regresijos? Kurkite modelį.