Estimering av Sykehusopphold

Målet med dette prosjektet er å anvende alt materialet lært i INF161 og fullføre et datavitenskapsprosjekt. Vennligst les beskrivelsen nøye! Dette prosjektet er en obligatorisk del av kurset. Prosjektet utgjør 50% av den endelige karakteren. Karakteren vil baseres på et godt valg av metoder, riktigheten av svarene, klarheten i koden, grundighet og klarhet i rapporteringen.

Krav

Du skal bygge en maskinlæringsmodell for å predikere den forventede lengden på sykehusoppholdet per pasient. Modellen vil bruke pasientopplysninger, inkludert fysiologiske, demografiske og sykdomsalvorlighetsdata på tvers av ni sykdomskategorier, for å predikere forventet oppholdslengde. Målet er å nøyaktig anslå lengden på sykehusopphold for nye pasienter basert på disse variablene. Arbeidet vil bestå av fire deler:

- Datatilberedning (40 poeng):
  - Input: rådata (i `raw_data` mappen). Fire datasett som innholder pasientinformasjon for å predikere pasientenes forventet oppholdslengde på sykehus. Les databeskrivelse i `README`.
  - Output: modellklar data.
  - Funksjonalitet: Dette systemet tar inn rådata og lager en dataframe som kan brukes i maskinlæringsmodellen. Databeskrivelse, data analyse og visualisering, og eventuelle andre datatilberedende steg bør rapporteres og du skal begrunne dine valg du har gjort innenfor dette systemet. Kvaliteten av rapportering er en del av vurderingen.
- Modellering og prediksjon (40 poeng):
  - Input: modellklar data
  - Output: maskinlæringsmodell, forventet generaliseringevne: RMSE
  - Funksjonalitet: Dette systemet tar den forberedte dataframen og bygger en maskinlæringsmodell for å predikere lengden på sykehusoppholdet. Modellvalg, valg av funksjons / forklaringsvariable, håndtering av manglende data (imputeringsteknikk) og variabelutvinning er viktige deler av dette systemet. Du bør evaluere minst 3 forskjellige modelleringsmetoder før du velger den endelige modellen. Ytelsen til systemet er evaluert ved å sammenligne predikert oppholdslengde med faktisk oppholdslengde på et valideringss-/testdatasett. Dette skal gjøres ved å bruke rot-middel-kvadrat-feil (RMSE) av prediksjoner, dvs.

$$\sqrt{\frac{\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}{N}},$$

hvor N er antall prediksjoner, $\hat{y}_i$ er den i-nde prediksjonen og y_i er den tilsvarende sanne verdien. Systemet skal rapportere den forventede generaliserings RMSE. Alle valg, alt fra modellvalg til håndtering av manglende data bør rapporteres og begrunnes. Resultater bør analyseres. Kvaliteten av rapportering er en del av vurderingen.

- Prediksjon (10 poeng):
  - Input: maskinlæringsmodell og nye data
  - Output: prediksjon på nye data
  - Funksjonalitet: Gitt nye datapunkter, skal dette systemet returnere prediksjoner på disse nye datapunktene.
- Nettside (10 poeng):
  - Funksjonalitet: Nettsiden skal la brukerne angi pasientdata og returnere en predikert forventet sykehusopphold. Nettsiden skal komme med brukerdokumentasjon og skal kunne håndtere brukerfeil (f.eks. feil datainput). Merk at dette er et HTML-dokument som finnes på din personlige datamaskin som du åpner med nettleseren din og ikke en nettside som er hostet på internett.

Tidsfrister

Prosjektet består av tre deler med tre forskjellige tidsfrister. I den første delen vil du forberede dataene for analyse. I den andre delen av prosjektet vil du designe en maskinlæringsmodell og predikere lengden sykehusoppholdet. Den siste delen av prosjektet består av å lage et enkelt nettside som kjører ditt prediksjonssystem.

- Tidsfrister:
  - Del 1: Søndag, 22.09, 23.59
  - Del 1 - peer review: Søndag, 29.09, 23.59
  - Del 1 & 2: Søndag, 13.10, 23.59
  - Del 1 & 2 - peer review: Søndag, 20.10, 23.59
  - Endelige prosjekt: Søndag, 3.11, 23.59
- Lever på MittUIB.no/assignments

Innlevering

Alle tidsfrister er obligatoriske. De to første delene vil det ikke gis karakter på, men du skal gi og få peer feedback som er nyttig for å forbedre det endelige prosjektet. For endelig innlevering, vennligst lever inn følgende:

- en PDF rapport som forklare av din tilnærming, dine metodevalg, resultater og analyse, pluss refleksjoner rundt styrker og forbedringspotensialle av ditt arbeid.
- en csv-fil, `predictions.csv`, som inneholder en prediksjon for hver pasient i `sample_data` mappen.

- en zip-fil av koden din. Vennligst inkluder en README.txt-fil i zip-filen din som forklarer hvordan vi skal kjøre koden din.

Vær oppmerksom på at hver ipynb/.py fil og nettside/app må kjøre uavhengig. I tillegg til pakker fra standardbiblioteket kan du bruke følgende Python-pakker: `xlrd`, `numpy`, `pandas`, `polars`, `scipy`, `sklearn`, `matplotlib`, `seaborn`, `requests`, `plotly`, `flask`, `django`, `waitress`. Hvis du bruker andre pakker, vil vi ikke kunne kjøre appen din, og du vil stryke prosjektet.

Koden skal være dokumentert, og triks (f.eks. for å unngå deling på null, for å sikre at den kjører på endelig tid, osv.) skal rapporteres. Begrunnelsen bak alle trinnene i koden skal være tydelig i rapporten.

MERK: Dette prosjektet er en læringsopplevelse. Hvis vi ser at du har kopiert svarene dine fra online ressurser, vil du få 0 poeng.

Modellutvelgelse er en viktig del av oppgaven og vil bli vurdert deretter. Før du anvender maskinlæringsalgoritmer, bør du alltid vurdere (og rapportere) hvilke resultater du forventer. Når du har brukt maskinlæringsalgoritmer, bør du alltid kommentere hvor godt resultatene samsvarer med forventningene dine.