Laboration Maskininlärning

Disease prediction

I det här momentet kommer vi jobba med ett dataset med data för hjärt-kärlsjukdom. Börja med att ladda ned datasetet från Kaggle:
https://www.kaggle.com/sulianova/cardiovascular-disease-dataset.

Läs på vad de olika kolumnerna betyder. Notera att detta dataset innehåller många felaktigheter, exempelvis finns negativa blodtryck och blodtryck som är omöjligt höga.

EDA

Använd pandas, matplotlib och seaborn för att besvara på följande frågor för datasetet:

- Hur många är positiva för hjärt-kärlsjukdom och hur många är negativa?
- Hur stor andel har normala, över normala och långt över normala kolesterolvärden? Rita ett tårtdiagram.
- Hur ser åldersfördelningen ut? Rita ett histogram.
- Hur stor andel röker?
- Hur ser viktfördelningen ut? Rita lämpligt diagram.
- Hur ser längdfördelningen ut? Rita lämpligt diagram.
- Hur stor andel av kvinnor respektive män har hjärt-kärlsjukdom? Rita lämpligt diagram

Feature engineering BMI

Skapa en feature för BMI (Body Mass Index), läs på om formeln på wikipedia.

- Släng de samples med orimliga BMIer och outliers. Notera att detta kan vara svårt att avgöra i vilket range av BMIer som vi ska spara. Beskriv hur du kommer fram till gränserna.
- Skapa en kategorisk BMI-feature med kategorierna: normal range, overweight, obese (class I), obese (class II), obese (class III).

Feature engineering blodtryck

Släng bort samples med orimliga blodtryck och outliers. Likt förra uppgiften är det inte trivialt att sätta gränserna. Skapa en feature för blodtryckskategorier enligt tabellen i denna artikel: https://www.healthline.com/health/high-blood-pressure-hypertensiondefinition. Beskriv hur du kommer fram till gränserna.

Visualiseringar andel sjukdomar

Skapa barplots med en feature mot andelen positiva för hjärt-kärl sjukdom. Exempelvis blodtryckskategorier mot andel positiva, BMI kategori mot andel positiva mm. Gör dessa plots i en figur med flera subplots.

Visualiseringar korrelation

Skapa en heatmap av korrelationer och se om du hittar features som är starkt korrelerade, dvs nära 1 eller features som är starkt negativt korrelerade, dvs nära -1. Kan du förklara varför de kan vara korrelerade?

Skapa två dataset

Skapa en kopia av ditt dataframe. På ena dataframet:

- Ta bort följande features: ap_hi, ap_lo, height, weight, BMI
- Gör one-hot encoding på BMI-kategori, blodtryckskategori och kön

På andra dataframet:

- Ta bort följande features: BMI-kategori, blodtryckskategori, height, weight
- Gör one-hot encoding på kön

Välj modeller

Välj 3-5 maskininlärningsmodeller, gärna så olika som möjligt. För *vardera* dataset som vi skapade i tidigare uppgift gör följande:

- train—validation—test split
- skala datasetet med feature standardization och normalization (de görs inte samtidigt, utan i olika omgångar)
- Definiera hyperparametrar (param_grids) att testa för varje modell
- Använd GridSearchCV() och välj lämpliga mätvärden (scoring)
- Gör prediction på valideringsdata
- Beräkna och spara evaluation score för ditt valda mätvärde
- Kontrollera bästa parametrarna för respektive modell

Ensemble

Använd VotingClassifier() på datasetet som du valt och lägg in de bästa parametrarna för respektive modell.

Vilket dataset väljer du och vilken modell väljer du? Använd den modellen du valt och träna på all data förutom testdatan.

Evalueringar

Gör confusion matrices och classification reports för din valda modell.

Bedömning

Om du har fått någon kodsnutt från någon annan eller hittat i någon sida är det viktigt att du källhänvisar, annars räknas det som plagiat. Skriv en kommentar bredvid koden som du har tagit.

Godkänt

- Löst uppgiften på ett korrekt sätt
- Koden är kommenterad med relevanta kommentarer
- Beskrivit och motiverat val av parametrar, modeller mm
- Dataanalysen och datavisualiseringen är gjord på korrekt sätt

Väl Godkänt

Uppfyllt allt för godkänt samt:

- Koden är tydlig och enkel att följa
- Koden är välstrukturerad med funktioner och/eller OOP i importerade moduler
- Motiverar väl dina val av parametrar, modeller mm
- Dataanalysen och datavisualiseringen är väl genomtänkt och beskriven