Data-mining Opdracht 21 Januari 2020

Stel je een casus voor. Een fietsenmakerbedrijf heeft moeite met de verkoop van fietsen aan zijn of haar klanten en doet wanhopige pogingen om klanten te lokken bij zijn bedrijf. Er zijn gewoon te veel factoren voor de fietsenmaker om te voorspellen wie je moet bereiken. Hij besluit een datascientist in te huren die in één dag tijd uit zijn data bruikbare informatie probeert te halen.
Dit kan op enkele manieren, namelijk, door (i) inzicht te geven in de relaties tussen attributen in de dataset, (ii) deze relaties te visualiseren, (iii) te laten zien welke attributen (features) van belang zijn voor het voorspellen van een potentiele klant en (iv) hoe goed een model kan voorspellen wie van een nieuwe test-set klanten wel of niet een fiets gaat kopen.

Presenteer je bevindingen en proces in een verslag. Dit mag in een jupyter notebook bestand of in een word-bestand. Je mag bij deze opdracht gebruik maken van **WEKA** en van **PYTHON** (en natuurlijk **jupyter notebook**). Hou de volgende structuur aan in je verslag:

Weka leent zich goed voor om de data te importeren en snel in te zien. Verschillende relaties te herkennen en algoritmes snel op toe te passen.

Python leent zich goed voor het verkennen en aanpassen van de data. Er missen hier en daar waardes die opgevuld dienen te worden.

<u>Beoordeling</u>

Je wordt bij deze opdracht beoordeeld op de volgende criteria:

1. De navolgbaarheid van je proces. Als opdrachtgever wil je het proces kunnen begrijpen van je datascientist. Waarom heeft hij/zij deze en deze keuze gemaakt? Wat voegt het toe aan de wens van de opdrachtgever?
2. Is de data voldoende geïnterpreteerd en uitgelegd? Leg bijvoorbeeld per attribuut uit of het een (grote) rol speelt voor het voorspellen van het wel of niet kopen van een fiets. Kijk of je een plot kan maken van de data (wij de docenten zullen hier morgen ondersteuning voor geven).
3. Heb je kunnen voorspellen wie er van de test-set een fiets heeft gekocht of niet. We zullen hier niet streng op zijn sinds het concept hiervan nog vrij lastig is. De voorspelling mag zowel met **WEKA** als met **PYTHON** gemaakt worden. Nog beter als je het met beide doet.

Jullie krijgen vanavond de **training-set**. De **test-set** krijgen jullie morgen in de middag.

Verder krijgen jullie vanavond het **jupyter notebook** bestand die behandeld is in de les vanmiddag. Op de volgende pagina is een cheatsheet met belangrijke formules nog een maal vermeld.

Veel voorkomende functies binnen Python

Dependencies / packages

Import pandas as pd
Import numpy as np
from sklearn.model_selection import train_test_split

Read files

pd.read_csv('*file'*)

Datamanipulatie of **inzage**

dataframe.info()
len(dataframe)
dataframe.keys()
dataframe.describe()
dataframe.count()
dataframe.head()
dataframe.tail()
dataframe.nunique()
dataframe.describe()
dataframe.groupby()
dataframe.mean()
dataframe.isnull()
dataframe.append()
dataframe.fillna()
dataframe.drop()

Computaties

- Dit is voornamelijk om een Boolean functie uit te voeren (*True or False*)

3 < 6 = TRUE
5 == 9 = FALSE
4 => 4 = TRUE
6 =< 5 = FALSE
2 != 6 = TRUE