Tentamenopdracht Programming for data science **(103169)**

Vakcode : ICT.BIM.DSPROG.V19(ICT.BIM.DSPROG.V17) (t2)

Datum : dinsdag 23 juni 2020

Tijd : 09.30 - 12.30 uur

|  |  |  |
| --- | --- | --- |
| Klas: | Lokaal: | Aantal: |
| ICTBIM2a t/m e  Herkansers | online | 152 |
|  |  |  |
|  |  |  |
|  |  |  |

Opgesteld door : Hiemstra, Paul

Docenten : HAP01, RSM03, MLM01, PSS06, CWR01

Gecontroleerd door : Ruben Cijsouw

Rekenmachine : alle rekenmachines toegestaan

Literatuur : alle literatuur die de student relevant vindt

Overige hulpmiddelen : computer, internet

Opgaven inleveren : nee

CONTROLEER VOORAF DE VOLGENDE GEGEVENS:

Deze tentamenopdracht bevat:

5 opgaven

**8** genummerde pagina's

Waarschuw de surveillant als één van deze aantallen niet klopt!

Aanwijzingen:

In totaal zijn voor de vragen 100 punten te halen.

Samenwerken met medestudenten of uitwerkingen uitwisselen is beslist NIET toegestaan.

# Opgave 1 (20 punten)

*Business understanding*

Ons is gevraagd om een aantal analyses te doen over studenten in het hbo.

*Data understanding*

We krijgen uiteindelijk toegang tot de informatie door een koppeling met de database. Voor het ontwikkelen hebben we hiervan een kopie ontvangen met gegevens uit de afgelopen jaren.

Jou is gevraagd om deze data beschikbaar te maken voor de minder technische analisten in de organisatie.

**Opgave 1a** (3 punten)

Gebruik de gegeven functies om een overzicht te geven van de tabellen in de SQlite-database.

**Opgave 1b** (4 punten)

Geef voor de **twee** tabellen met gegevens over de ingeschreven studenten en over de eerstejaarsstudenten in de database een overzicht van de karakteristieken.

Geef voor elke tabel inzicht in

* het aantal rijen en kollomen in de tabel
* de variabelen/kolommen in deze tabel
* de datatypes van de variabelen in deze tabel

**Opgave 1c** (10 punten)

Om de data makkelijk uit de database te halen, willen we een view aanmaken met de data uit de tabellen 'ingeschrevenen\_hbo' en 'eerstejaars\_hbo'. Deze dienen op de juiste wijze te worden samengevoegd. We willen dat alle records uit de 'ingeschrevenen\_hbo' tabel in de view komen.

*Hint*:

GEMEENTENUMMER, BRIN\_NUMMER\_ACTUEEL, OPLEIDINGSCODE\_ACTUEEL, OPLEIDINGSVORM en GESLACHT maken een regel in beide tabellen uniek.

**Maak een SQL-query om de gevraagde view te creëren**. Let er op dat in ieder geval alle regels uit de tabel `ingeschrevenen\_hbo` worden opgenomen in de view.

**Opgave 1d** (3 punten)

Gebruik de door jouw gecreëerde view om **alle rijen** van de analytische dataset in de pythonomgeving te laden.

# Opgave 2 (20 punten)

*Business understanding*

We hebben een dataset over lego. Een online aanbieder gebruikt deze data voor haar webshop. De meeste mensen kopen volledige lego-sets. Hiervoor is het belangrijk om te weten welke lego-sets er zijn er en hoeveel exemplaren er van een set voorradig zijn.

Er komt steeds meer vraag naar losse lego-onderdelen. Om de klanten goed van dienst te kunnen zijn is ons gevraagd om de gegevens over de losse lego-onderdelen te verkennen.

In deze vraag doe jij de noodzakelijke data understanding en data preparation.

*Data understanding*

Het informatiemodel van deze data staat in de afbeelding hieronder. Afbeelding met schermafbeelding

Automatisch gegenereerde beschrijving

Voor deze verkenning hebben we alleen de gegevens uit de onderste vier klassen uit het informatiemodel nodig. Deze gegevens zijn als vier losse bestanden aangeleverd in het mapje 'data vraag 2'.

**Opgave 2a** (9 punten)

Maak één analytische dataset waarbij voor elk lego-onderdeel uit de voorraad alle informatie over kleur, categorie, naam, aantal en reserve-onderdeel op één regel staat.

Plaats de analytische dataset in één DataFrame.

* Voeg de gegevens uit `parts` en `part\_categories` samen. Zorg dat elke variabele één keer voorkomt en zorg voor een 'logische' naamgeving van elke kolom.
* Voeg de gegevens uit `inventory\_parts` toe aan de analytische dataset. Zorg dat elke variabele één keer voorkomt en zorg voor een 'logische' naamgeving van elke kolom.
* Voeg de gegevens uit `colors` toe aan de analytische dataset. Zorg dat elke variabele één keer voorkomt en zorg voor een 'logische' naamgeving van elke kolom.

**Data-exploratie**

**Opgave 2b**. (3 punten)

Exploreer de samengevoegde data. Genereer overzichten die inzicht geven in de volgende kenmerken van de dataset:

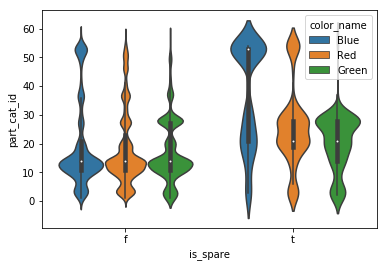
* de laatste regel van de dataset
* het aantal variabelen
* de datatypen van elke variabele
* het totaal aantal records
* het aantal unieke waarden van elke variabele

**Opgave 2c** (4 punten)

Maak een staafgrafiek die inzicht geeft in de spreiding van het aantal onderdelen binnen een `part category`. Zorg dat elke categorie een individuele staaf heeft.

**Opgave 2d** (4 punten)

Een concurrent heeft op hun website een visualisatie staan die wij ook graag willen opnemen in de rapportage omdat deze inzicht geeft in de spreiding van de reserve-onderdelen per categorie en per kleur.

Maak onderstaande visualisatie m.b.v. een *violin plot zo* goed mogelijk na. Je hoeft nog geen aandacht te besteden aan de kleurstelling en de positie van de legenda. 

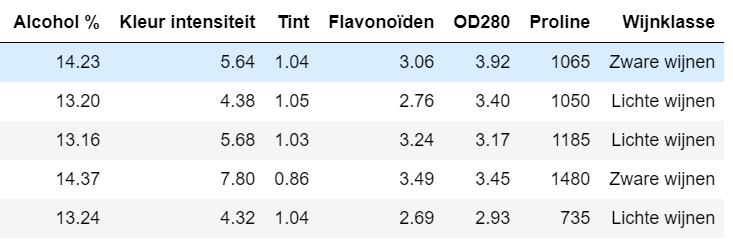
# Opgave 3 (20 punten)

*Business understanding*

In deze opdracht gaan we aan slag met een dataset over wijnen (`vraag3\_Wijnen.xlsx`). De groothandel die deze wijnen verkoopt, onderscheidt de wijnen meestal in twee simpele categorieën: lichte en zware wijnen. Aan jouw de taak om te kijken of hier een betere classificatie te maken is.

*Data understanding*

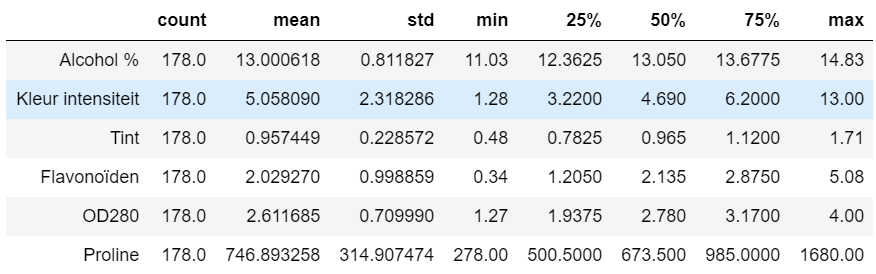
De dataset ziet er als volgt uit



De volgende kolommen zijn beschikbaar:

* Alcohol %, geeft aan hoeveel alcohol er in een wijn zit
* Kleur intensiteit en tint, beschrijven hoe sterk de kleur is en wat die kleur is
* Flavonoïden, OD280 en Proline zijn belangrijke ingrediënten van wijn en geven aan hoeveel er in zit
* Wijnklasse, de oorspronkelijke classificatie van de wijnhandel

Per variabele zijn dit de beschrijvende statistieken:



**Vraag 3a (3 punten)**

We beginnen met de data te verkennen.

* Importeer de dataset
* Maak een pairplot waarbij de wijnklasse als kleur wordt weergegeven
* Leidt aan de hand van de grafiek af op basis van welke variabele(-n) de wijnhandel de oorspronkelijke classificatie had gemaakt

*Data modeling*

**Vraag 3b (3 punten)**

Het is duidelijk dat de kolommen van de data erg sterk van elkaar verschillen. We willen daarom de data gaan normaliseren.

* Maak een kopie van de dataset met daarin alleen de numerieke kolommen
* Normaliseer deze nieuwe dataset door eerst het gemiddelde er af te trekken en in een tweede stap te delen door de standaarddeviatie
* Toon aan dat het gemiddelde in je nieuwe dataset in alle kolommen ongeveer 0 is en de standaarddeviatie ongeveer 1

**Vraag 3c (6 punten)**

Het is nu tijd om een KMeans clustering op de data te gaan toepassen. We willen als eerste weten hoeveel clusters er nodig zijn.

* Maak een Elbow chart door de inertie te berekenen voor een Kmeans clustering. Bereken dit voor het aantal clusters tussen de 1 en 6 door middel van een loop.
* Hoeveel clusters lijkt jou optimaal? Waarom kies je hiervoor?

**Vraag 3d (5 punten)**

Het is nu tijd om de definitieve clustering te gaan draaien

* Draai nog een keer de clustering op de genormaliseerde data op basis van je optimaal aantal clusters
* Voeg vervolgens je gevonden labels toe aan de niet-genormaliseerde dataset
* Produceer van deze niet-genormaliseerde dataset een pairplot, waarbij de kleuring gebaseerd is op de clustering. Zorg ervoor dat je hierbij niet de clusteringsnummers zelf plot als losse figuren.

**Vraag 3e (3 punten)**

Als laatste willen we kijken naar de kenmerken van de verschillende clusters.

* Produceer per cluster de gemiddelde waardes voor elke numerieke kolom
* Geef een interpretatie van elk cluster

# Opgave 4: Lineaire regressie (20 punten)

# *Business Understanding*

Vluchten die te laat vertrekken is een grote bron van irritatie voor passagiers. Een vliegveld heeft je gevraagd om inzicht te geven in de redenen waarom vliegtuigen later vertrekken dan hun geplande vertrektijd. De klant heeft jullie een dataset geleverd (`vraag4\_flights.csv`) waar een grote hoeveelheid vluchtgegevens in staan.

*Data Understanding*

**Vraag 4a (5 punten)**

De klant heeft de hypothese dat de vertraging (`DEPARTURE\_DELAY`) een verband heeft met de afstand van de vlucht (`DISTANCE`) en hoe laat het vliegtuig vertrekt (`DEPARTURE\_TIME`).

* Bereken en visualiseer de correlaties tussen de variabelen die hierboven genoemd worden.
* Interpreteer de correlaties die je gevonden hebt. Ben jij het met de hypothese van de klant eens?

*Modeling*

**Vraag 4b (6 punten)**

Bouw een regressiemodel tussen `DEPARTURE\_DELAY` als afhankelijke variabele en `DISTANCE` en `DEPARTURE\_TIME` als onafhankelijke variabele. Train het model op 70% van de data, en leg 30% achteruit om het model te testen. Gebruik `42` als random state.

* Geef als output in ieder geval de werkelijke en voorspelde waarden van `DEPARTURE\_DELAY`, de berekende coëfficiënten en de intercept.

**Vraag 4c (5 punten)**

De klant heeft nog een aanvulling op het model: zij vermoedt dat de luchtvaarmaatschappij (`AIRLINE`) ook een invloed heeft. Breidt het model verder uit met deze categorische variabele.

* Geef als output in ieder geval de werkelijke en voorspelde waarden van `DEPARTURE\_DELAY`, de berekende coëfficiënten en de intercept.
* Welke luchtvaartmaatschappij heeft de minste vertraging als je kijkt naar de coëfficiënten van het model?

*Evaluation*

**Vraag 4d (4 punten)**

Vergelijk de twee modellen op basis van hun R2. Welke model zou je kiezen en waarom? Wat vindt je van de kwaliteit van het model?

# Opgave 5 (20 punten)

*Business Understanding*

Een klant heeft je gevraagd om een model te bouwen wat op basis van de tekst van een vraag kan voorspellen wat voor categorie vraag het is. Het spel waarbinnen deze vragen en antwoorden een rol spelen is Jeopardy!.

Als basis heb je de dataset `vraag5\_JEOPARDY\_CSV.csv` gekregen waar de vragen en bijbehorende categorie in staan.

*Data preparation*

**Vraag 5a (5 punten)**

Lees de data in, en zet deze data om naar een term-document matrix die geschikt is voor Naive Bayes. Print de data en de term-document matrix op het scherm.

*Modeling*

**Vraag 5b (5 punten)**

Train op basis van de data en het term-document matrix een Naive Bayes model wat voorspelt wat de categorie van een vraag is.

*Evaluation*

**Vraag 5c (5 punten)**

Bereken van het model uit vraag 5b de `f1\_score` en laat de `confusion matrix` zien. Geef daarna ook een interpretatie van de scores: wat vind je van het model?

Onder de aanname dat de categorieën op alfabetische volgorde in de confusion matrix staan, welke van de categorieën wordt het beste en welke het slechtste geclassificeerd? Hoe zie je dit in de confusion matrix?

**Vraag 5d (5 punten)**

Maak nog een Naive Bayes model, maar dan met de 1000 meest voorkomende woorden. Bereken ook hoe goed dit model presteert, vergelijk het met de prestaties bij vraag 5c en trek een conclusie: welk model kies jij? Onderbouw deze keuze vanuit je berekende prestaties.

Geef ook aan welke van de categorieën het meest verbeterd is qua performance. Onderbouw je antwoord.

**Einde Tentamenopdracht**

Vergeet niet je tentamenopdracht in te leveren via de elo

Vergeet niet maandag de reviewopdracht te maken