UNIVERSITY OF THE FREE STATE

BLOEMFONTEIN CAMPUS

CSIS3764

DEPARTMENT: COMPUTER SCIENCE AND INFORMATICS

CONTACT NUMBER: 051 401 2929

PROJECT:
Main End-of-year 2020
PAPER 2

ASSESSOR: 1. Mnr. W.S.J. Marais
MODERATOR: 1. Dr. J.E. Kotzé

FINAL SUBMISSION DATE: 14 Dec 2020 **MARKS:** 35

INSTRUCTIONS:

- Create two Jupyter notebooks for this exam project. One for each question below.
- Rename the notebooks as specified in each question.
- Use comments to indicate the purpose of the code you write.
- To complete the exam project, you need to submit the two notebooks to the following Blackboard links respectively:
  - CSIS3764 -> Assessments -> Project -> Question 1.
  - CSIS3764 -> Assessments -> Project -> Question 2.

Question 1

Available on Blackboard under CSIS3764 -> Assessments -> Project -> Question 1.

Question 2 [35 Marks]

Please log into Blackboard and go to CSIS3764 -> Assessments -> Project -> Question 2 and download the data file called “cluster_data.csv”. The data contains measurements of different parts of three flower species. The relevant parts that are:

- The sepal length.
- The sepal width.
- The petal length.
- The petal width.

Sepals typically function as protection for the flower in bud, and often as support for the petals when in bloom. You are expected to create an unsupervised machine learning model that will be able to arrange flowers into different groups when given the measurements for the sepals and petals.

Create a Jupyter notebook called “CSIS3764_Exam2_YourStudentNumber.ipynb”.

The Jupyter notebook should have the following functionality:

- Read the data file "cluster_data.csv" into a dataframe called "cluster". The columns contain the following information:
  - Column 1: sepal length.
  - Column 2: sepal width.
  - Column 3: petal length.
  - Column 4: petal width.

  Name the columns accordingly.

- Summarize the data by:
  - Taking a peek at the data.
  - Getting the dimensions of the dataset.
  - Providing a statistical summary for all the numerical columns.
  - Determining the data types for all the columns.
  - Create box plots for each of the flower features.

- Clean the data, check the data for any faulty values and handle all the data errors.
- After the data has been cleaned:
  - Get the updated dimensions of the dataset.
  - Create updated box plots for each of the flower features.

- Visualize the data:
  - Create scatter plots between the different pairs of flower features. This should give you 6 separate scatter plots.

- Use k-Means clustering to train a machine learning model that will have the capability to distinguish between three different flower species.
  - Use all the features from the “clusters” dataframe.
  - Print the labels of the trained model.

- Re-create the earlier scatter plots between the different pairs of flower features.
  - Show the clusters that was identified by the model within the data by means of different colours.

- You pick a flower with features (Sepal length = 7, Sepal width = 3, Petal length = 6, Petal width = 2).
  - Use the model that you trained to predict the label (species) of the flower you picked.
  - Print the label that the model predicted.
  - Re-create the earlier scatter plots between the different pairs of flower features and add the data points of the flower that you picked to the scatter plots.

- Has the model predicted the species of the flower that you picked correctly (Type the answer in the Jupyter notebook)?

[35]

End of Project