UNIVERSITY OF THE FREE STATE

BLOEMFONTEIN CAMPUS

CSIS3764

DEPARTMENT: COMPUTER SCIENCE AND INFORMATICS

CONTACT NUMBER: 051 401 2929

SEMESTER TEST 1
30 September 2022

ASSESSOR: 1. Mnr. W.S.J. Marais
MODERATOR: 1. Mev. S.E.S. Campher

TIME: 120 min **MARKS:** 80

INSTRUCTIONS:

- Log into Blackboard and go to *CSIS3764 -> Assessments -> Semester Test 1*.

Question 1 [20 Marks - 20 Minutes]

1.1 Give a detailed definition of Data Science. (6)

1.2 Explain the difference between oversampling and undersampling as it pertains to the data exploratory analysis and pre-processing phase. (2)

1.3 Draw a diagram of the CRISP-DM model and discuss each stage briefly as a model for conducting Data Science projects. (10)

1.4 Briefly explain what data leakage is and how it can be prevented. (2)

[20]

Question 2 [60 Marks - 100 Minutes]

- Please log into Blackboard and go to *CSIS3764 -> Assessments -> Semester Test 1*.
- Download the *airplane_crashes.csv* data file.
- Create a Jupyter Notebook that contains the Python code to import, explore, analyse, clean and pre-process the data for numerical data analysis. Call the notebook *CSIS3764_YourStudentNumber_ST1_2022*.

- The Jupyter Notebook should do the following (NB!!! - Show the changes in the data after each operation):

2.1 Import the data file into a dataframe, named *airplane_crashes* with the column headings as contained in the data file. (1)

2.2 To make the column headings less confusing, change the following dataframe column headings as indicated below: (2)

- Change *Location* to *Crash_Location*.
- Change *Operator* to *Airline_Operator*.
- Change *Type* to *Aircraft_Type*.
- Change *Registration to Registration_Number*.
- Change *Aboard* to *People_Aboard*.
- Change *Fatalities* to *Fatalities_Aboard*.
- Change *Ground* to *Ground_Fatalities*.
- Change *Summary* to *Event_Info*.

2.3 Determine and display the row count, column count and total number of elements in the dataframe. (2)

2.4 Inspect the data by: (3)

- Displaying the first 10 records of the dataframe.
- Generating a statistical summary of <u>all</u> the features.
- Displaying the data type for each column.

2.5 Change the data type of the *Date* column to *DateTime*. (2)

2.6 Split the value of the Date column into three parts by inserting three additional columns between columns Date and Time as indicated below: (4)

- Column *Year*: Populate this column with the Year as specified in the *Date* column.
- Column *Month*: Populate this column with the Month as specified in the *Date* column.
- Column *Day*: Populate this column with the Day as specified in the *Date* column.

All the relevant column values must be integers.

2.7 Add a column, named *Total_Fatalities*, to the end of the dataframe that combines the number of Fatalities Aboard and Ground Fatalities. (2)

2.7.1 Determine and display the maximum number of Total Fatalities that was caused by one of the airplane crashes contained in the dataset. (1)

2.7.2 On which date did the airplane crash occur, regarding question 2.7.1? (3)

2.7.3 From the dataset, find and print the information about the event that occurred which caused the airplane crash regarding question 2.7.1? (3)

2.8 List, in ranking order from highest to lowest number of crashes, the top 20 Airline Operators that had the most recorded crashes along with their respective number of crashes. (2)

2.9 Determine and display the percentage of all airplane crashes that involved South African Airways. Round the percentage to 2 decimal values. (4)

2.10 Plot a pie chart to illustrate the percentage of Fatalities Aboard versus the Ground Fatalities. Round the percentages to zero decimal values. (4)

2.11 By only using the information provided in the pie chart, what deduction can be made from the information depicted in the pie chart? (2)

2.12 Using the Seaborn library, plot two bar charts overlaying each other: (5)

- Plot a bar chart in the colour red that depicts the number of Total Fatalities per year.
- Plot a bar chart in the colour blue that depicts the number of Fatalities Aboard per year.

The blue bar chart should overlay the red bar chart.

2.13 Discuss how the information depicted in the bar chart changes/influences your deduction about the information depicted in the pie chart. (2)

2.14 Determine and display the Aircraft Type that had the highest sum of Fatalities Aboard. (4)

2.15 Discard the *Date*, *Time*, *Flight #*, *Route*, *Registration_Number*, *cn/In* and *Event_Info* columns. (1)

2.16 Handle missing values by doing the following: (3)

- Check for missing values.
- Discard all rows that contain missing values.
- Check if all the records that contained missing values were indeed discarded.

2.17 Convert the text values in the remaining dataframe to numeric values as follows: (6)

- Convert the columns *Airline_Operator* and *Aircraft_Type* to numeric values [0, 1, 2, …], with the highest numeric value depending on the number of unique text values.
- Convert the column *Crash_Location* to a vector space.

2.18 Ensure that all the values in the dataframe are on the same scale. (3)

2.19 Write the cleaned and pre-processed dataframe to a file named *clean_airplane_crashes.csv*. (1)

- After you have completed Question 2, please submit your completed Jupyter Notebook on Blackboard.

[60]

End of Paper