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 1

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

FINAL SUBMISSION DATE: 14 Dec 2020 **MARKS:** 85

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 [85 Marks]

Please log into Blackboard and go to CSIS3764 -> Assessments -> Project -> Question 1 and download the data file called “game_stats.csv”. The data contains soccer statistics for world cup matches. Construct machine learning classifiers that is able to predict from which team the Man of the Match will come from, depending on the feature values for that specified soccer match. Create a Jupyter notebook called “CSIS3764_Exam_YourStudentNumber.ipynb”.

The Jupyter notebook should have the following functionality:

- Read the data file "game_stats.csv" into a dataframe called "stats". Keep all the columns. You will later have to determine which columns are important and which can be removed.
- Summarize the data by:
  - Taking a peek at the data.
  - Getting the dimensions of the dataset.
  - Providing a statistical summary for the all the numerical columns.
  - Determining the data types for all the columns.

- Clean the data, check the data for any faulty values and handle all the data errors (display the data after each cleaning step):
  - Handle all missing values. If the number of missing values entail a high percentage of the total samples, the values can be removed otherwise they need to be filled with appropriate values.
  - Convert columns “Man of the Match” and “PSO” from text to numeric values [0, 1].
  - Convert column “Round” from text to numeric values [0, 1, 2, 3, 4, 5].
  - Convert the text values in columns “Team” and “Opponent” into a vector space [0, 1] and merge it with the existing columns forming a new dataframe called “stats_clean”.
  - Drop all the irrelevant columns that are still of type *object*. Confirm that all columns of type *object* have been removed.

- Determine the correlation between Man of the Match and the other data columns, excluding the Team and Opponent columns.
  - Create a heatmap that depicts the correlation.
  - Use 0.11 as the correlation coefficient to determine correlation.
  - Look at positive and negative correlation.
  - List the columns that shows a positive or negative correlation of 0.11 and more with Man of the Match. Do this programmatically.
  - Create box plots between Man of the Match and each of the columns that showed correlation (positive or negative).
  - Remove outliers from these columns that are 2.5 times the variance or more outside the 1st and 3rd quantiles.

- Define X and y:
  - X = Data features
  - y = Man of the Match
  - Determine the dimensions of X and y.

- Train the following classifiers with a training dataset of 90%.
  - Logistic regression
  - K nearest neighbor
  - Decision tree
  - Support vector machines

- Determine the best model by using k-fold cross-validation.
  - Set K = 20 for the k-fold cross-validation
  - Report the accuracy and F1 scores
  - Create box plots to compare the classifiers' F1 scores to determine which classifier performed the best.

- Select the model that produced the highest F1 score to do the following:
  - Make predictions using X_test.
  - Provide the accuracy score, confusion matrix and classification report of the model.

- Explain the confusion matrix and classification report results with regard to the following (Type the answers in the notebook):
  - Interpret the precision score for this model?
  - Interpret the recall score for this model?
  - Interpret the F1 score for this model?
  - Are the FN and FP values of the confusion matrix acceptable?

- Use a dummy classifier as a baseline to compare your selected classifier's accuracy score, confusion matrix and classification report with.
  - https://scikit-learn.org/stable/modules/generated/sklearn.dummy.DummyClassifier.html?highlight=dummyclassifier

[85]

Question 2

Available on Blackboard under CSIS3764 -> Assessments -> Project -> Question 2.

End of Project