COMP47490 Assignment 1

Deadline: Submit no later than **Friday, Nov 4, 2022**. For those with UCD approved accommodations, the deadline is Sunday, Nov 6, 2022.

Instructions

Submit your assignment as one Jupyter notebook file (not a DOC/DOCX/ODT/ZIP/PDF file) via the module Brightspace page.

Exam should be completed individually. ***Any evidence of plagiarism will be reported to the CS plagiarism committee and it can result in a Fail grade***.

Note that there is a bonus question at the end that makes the total marks of the assignment as 110. But you will be graded as if the total is out of 100. So, to get an A+, you only need to score 95/110, to get an A you need 90/110 and so on.

Please keep the whole code in a single Jupyter notebook. In your notebook, please split the code and explanations into many little cells so it is easy to see and read the results of each step of your solution. Please remember to name your variables and methods with self-explanatory names. Please remember to write comments and where needed, justifications, for the decisions you make and code you write.
Your code and analysis is like a story that awaits to be read, make it a nice story please. Aim to keep the notebook clear and concise, with the key code and discussion.

Download the file pulsar_star_<student_number>.csv from Brightspace (My Learning -> Assignment1 Datasets). So, if your student number is 12345678, then download pulsar_star_12345678.csv. When downloading your dataset, please ensure that your student number is correct. **Submissions using an incorrect dataset will receive a 0 grade.**

This dataset contains data for classifying a star as a pulsar or not. Features are based on a number of values extracted from astronomical imaging.

Pulsar stars are a very rare type of Neutron star that produce radio emission detectable on Earth and they are of considerable scientific interest as probes of space-time and states of matter. Their emission spreads across the sky and produces a detectable pattern of broadband radio emission. However in practice almost all detections are caused by radio frequency interference and noise, making legitimate signals hard to find.
The main purpose of this problem is to build a simple classifier in order to predict wether a detected signal comes from a pulsar star or from other sources such as noises, interferences, etc. In other words, the goal in this assignment is to work with the data to build and evaluate prediction models that capture the relationship between the descriptive features and the target feature "target_class".

Each candidate is described by 8 continuous variables. The first four are simple statistics obtained from the integrated pulse profile (folded profile). This is an array of continuous variables that describe a longitude-resolved version of the signal that has been averaged in both time and frequency. The remaining four variables are similarly obtained from the DM-SNR curve. These are summarised below:

1. Mean of the integrated profile.
2. Standard deviation of the integrated profile.
3. Excess kurtosis of the integrated profile.

4. Skewness of the integrated profile.
5. Mean of the DM-SNR curve.
6. Standard deviation of the DM-SNR curve.
7. Excess kurtosis of the DM-SNR curve.
8. Skewness of the DM-SNR curve.

Using your dataset, perform the tasks below using Python and Scikit-learn.

Task 1: Prepare a data quality plan for the dataset. Mark down all the features where there are potential problems or data quality issues. Propose solutions to deal with the problems identified. Explain why did you choose one solution over potentially many other. It is very important to provide justification for your thinking in this part and to list potential solutions, including the solution that will be implemented to clean the data. In particular, pay attention to missing data and carefully address this issue. **[15 marks]**

Task 2: Normalise or Standardise your features as necessary. Carefully decide the normalisation or standardisation technique used. **[5 marks]**

Task 3: Carefully decide the evaluation measure that is best suited to this application and the dataset. Justify your choice -- What characteristics of the application and the dataset made you decide the evaluation measure you chose. **[5 marks]**

Task 4: Compare a decision tree classifier, a kNN classifier and four SVM classifiers (one each with "linear", "poly", "rbf" and "sigmoid" kernel) based on the evaluation measure selected in Task 3. Carefully decide the evaluation methodology for this comparison (e.g., cross validation or a single train/validation/test split or other alternatives). Explore the effect of different parameter settings on these classifiers and find the winner classifier/ parameter setting. Why do you think you got those comparison results? In particular, are you surprised at the relative performance of "linear", "rbf" and "sigmoid" kernels? **[25 marks]**

Task 5: Based on a filter technique, identify the three most discriminative features and the three least discriminative features in this dataset. Run the SVM classifiers with the four kernels on the top three and the bottom three features. How do the results compare? **[10 marks]**

Task 6: Carefully identify the most discriminating features to predict the binary outcome of the dataset using one wrapper feature selection technique. This should be done for each of the decision tree, kNN and four SVM classifiers from part Task 4. Report and discuss the differences between the feature subsets produced by the filter (Task 5) and the wrapper technique. **[15 marks]**

Task 7: Compare the performance of different classifiers using the different feature subsets found in Tasks 5 and 6 and compare it to the results on original dataset that you reported in Task 4. Have the results improved or worsened after feature selection? Is the relative performance of different classifiers and configuration settings in line with your expectation? **[10 marks]**

Task 8: Plot the ROC curves for the "1" class and the different classification models. What do you learn from this ROC curve? Which classifier/configuration is best suited for this task? Are you satisfied with the performance? **[15 marks]**

Task 9: **BONUS Question.** *This part is open-ended* -- Take the exploration and the discussion deeper than what is asked in the above questions and gain further insights into (i) correlation of the various features with the target class, (ii) feature selection and feature

importance, (iii) relative performance of different classifiers (different kernels in case of SVM) and different parameter settings w.r.t different evaluation measures and (iv) effect of different ways of imputing missing values on the final performance of different classifiers. **[10 marks]**

Grading Guideline

A	Careful data-cleaning and normalisation taking all necessary precautions into account. Carefully explored the space of classifier parameters and found a very good setting. The evaluation methodology is correct and extreme care has been taken to avoid test data peeking. The evaluation measures are well-justified and the overall results are impressive. The student has gone well beyond what was asked in the exploration.
B	Careful data-cleaning and normalisation taking most of the necessary precautions into account. Carefully explored the space of classifier parameters and found a good setting. Minor flaws in the evaluation methodology or feature selection. The evaluation measures are well-justified, but not applied consistently across the different classifier/configuration setting. The overall results are correct and reliable.
C	Data-cleaning and normalisation taking most necessary precautions into account. A basic exploration of the space of classifier parameters and classification models, without going in depth. The evaluation methodology and feature selection process has many flaws. Only one evaluation measure was selected or the evaluation measures were not used consistently. The overall results are correct.
D	Limited data-cleaning and normalisation, but fail to take crucial necessary precautions into account. Insufficient exploration of the space of classifier parameters/classifiers. The evaluation measures are not appropriate for the dataset/task or they were not used properly and consistently. The evaluation methodology has serious flaws and/or there is considerable amount of test data peeking. The overall results have limited utility due to flaws in data cleaning, feature selection, evaluation methodology and evaluation measures.
E	Limited data-cleaning and/or data normalisation. Some crucial errors in exploration and the overall results are incorrect and can't be relied on. The evaluation measures are not appropriate for the dataset.
F	No data-cleaning or data normalisation. Major errors in exploration. Incorrect evaluation methodology and wrong evaluation metric used. Overall results fail to convince.