COMP47490 Assignment 1

Deadline: Submit no later than **Sunday 31st Oct, 2021**.

Instructions

Answer both questions. 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.

Question 1

This assignment focuses on building and evaluating prediction models for a particular problem and dataset. The problem and data come from an animal shelter concerned with animal welfare and reducing the risk of animal death. The shelter wants to use the data collected about their animals to build a data analytics solution for death risk prediction to help them in their planning towards improving the welfare of the animals they shelter. The shelter collects some data for each animal they take in (columns in the dataset with keyword "intake") and also records the status of each animal when it left the shelter (columns in the dataset with keyword "outcome"). The target variable to predict is a variable called outcome. For this variable, the value “1” indicates that after intake, the animal outcome was negative, e.g., "death", while the value "0" indicates that the animal outcome was a positive one, e.g., was adopted or returned to the owner. The dataset we work with is a sample of the data released by this shelter: https://data.austintexas.gov/Health-and-Community-Services/Austin-Animal-Center-Outcomes/9t4d-g238

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 outcome.

Download the file animal_shelter_<student_number>.csv from Brightspace (My Learning -> Datasets -> Assignment 1). So, if your student number is 12345678, then download animal_shelter_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 is an ***open-ended assignment*** -- You do not have to restrict yourself to what is asked. You can take the exploration and the discussion deeper than what is asked.

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. Always start with an introduction about the problem and your understanding of the problem domain and the data analytics solution. Then describe the steps you take and your findings from each step. Aim to keep the notebook clear and concise, with the key code and discussion.

(a) Convert the features to their appropriate data types (e.g., decide which features are more appropriate as continuous and which ones as categorical types). - Drop duplicate rows and columns, if any. - Drop constant columns, if any. - Save your updated/cleaned data frame to a new csv file.

Prepare a data quality plan for the cleaned CSV file. - 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. - Apply your solutions to obtain a new CSV file where the identified data quality issues were addressed.

Normalise your features as necessary. [10 marks]

[Note that the performance of your classification models in the questions ahead may depend on the data cleaning you do in this part, so please make sure that you clean the data to a decent degree]

(b) Carefully identify the most discriminating features to predict the target category using the filter and wrapper feature selection techniques. Report the feature subsets that these techniques select. In the case of a filter, carefully decide the strategy to choose a subset of the ranked features and justify your choice. In the case of wrapper techniques, carefully select features for at least one Decision Tree, one Naïve Bayes , one SVM and one k-NN classifier. Report and discuss the differences between the feature subsets produced by the filter and wrapper techniques.

What insights can you provide to the animal shelter management based on the feature importance? Are there any actions that they can take to improve the adoption/ return rate of the animals.

For example, an insight can look like this "Based on the information gain/gini coefficient criteria, the filter technique found the animal breed as the most discriminatory feature for this task. On careful examination, I found that a particular breed "Pit Bull Mix" was the top euthanised breed at the shelter. Searching on the web, I found that this could reflect the common stigma toward the Pit Bull breed. Pit Bulls are known to be trained fighting dogs and commonly labelled a “Dangerous Breed”. In the state of Texas Breed Specific Laws can target such “Dangerous Breeds”, with some cities even ruling against the breeds adoption. [15 marks]

(c) Carefully consider the evaluation measure(s) that you use for this exercise and evaluate the performance of at least one Decision Tree, one Naïve Bayes, one k-NN classifier and one SVM classifier on your dataset using the different feature subset(s) identified in (b). Explore the effect of different parameter settings on these classifiers. Which combination of classifier, parameter settings, feature subset gives you the best result, based on your chosen evaluation measure(s).

Describe the evaluation procedure that you used in good detail and justify why you used it. For example, you could have decided to keep 30% data as hold-out test set or used cross-validation with 10 folds or something else. [15 marks]

(d) 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? [10 marks]

(e) Carefully discuss the results obtained in part (c) and (d). To what extent are these results in line with or different from what you learnt about these classifiers in your lectures? For example, is your accuracy higher or lower on the dataset with reduced number of features as compared to the original dataset? Is the relative performance of different classifiers and configuration settings in line with your expectation? Why do you think a particular classifier performed very poorly -- Do you suspect an underlying assumption failing on this dataset? [10 marks]

Question 2

Answers all parts below. **Please provide answers in your own words.** Note that I am not looking for a mere reproduction of lecture slides/recording here, I want to see what you understood from it and what you read further on the topic.

(a) Consider the nightmare situation in which you struggled hard to obtain a very high accuracy (>95%) on your training data for a binary classification task, but when your client ran it on their test data, the accuracy was very low (<50%). This is despite the fact that your dataset is reasonably balanced (majority class < 65%) and you are using a fairly complex learning algorithms with many parameters to fit your dataset. How do you explain this situation? What are the possible causes for this? How can you improve the testing accuracy in this situation? What precautions should you take in your evaluation procedure to avoid this situation? [10 marks]

(b) Explain (in **own words)** what is the kernel trick in SVM? What is the mathematical reasoning that makes it work? Why is the kernel trick important? Give some examples of the kernels together with the situation when you should use it. [10 marks]

[You can include what you explored further on the topic. For example, I ran SVM with different kernels on the face dataset and I found that ....]

(c) Explain, **in your own words**, what is the curse of dimensionality? Give a mathematical intuition as to why a k-nearest neighbour classifier may struggle on a very high-dimensional data. Give examples of various situations in which you would want to reduce the number of features you consider for classification. [10 marks]

(d) You are working on a classification task to separate fraudulent transactions from the normal financial transactions. What evaluation measure(s) will you use for this application and why? What measure will you not use for this application and why? [10 marks]

Grading Guideline

- Q1: 60 marks (10 + 15 + 15 + 10 + 10)
- Q2: 40 marks (10 + 10 + 10 + 10)

	Quality of Exploration in Q1 a,b,c	Quality of Discussion in Q1 d,e	Question 2
A	Careful data-cleaning and normalization taking all necessary precautions into account. Carefully explored the space of classifier parameters and found a very good setting. The evaluation measures are well-justified and the overall results are impressive. The student has gone well beyond what was asked in the exploration.	The quality of discussion reflects an excellent understanding of the underlying machine learning concepts.	The answers reflects an excellent understanding of the underlying machine learning concepts.
B	Careful data-cleaning and normalization taking all necessary precautions into account. Carefully explored the space of classifier parameters and found a good setting. The evaluation measures are well-justified, but not applied consistently across the different classifier/configuration setting. The overall results are correct and reliable.	The quality of discussion reflects very good understanding of the underlying machine learning concepts.	The answers reflects a very good understanding of the underlying machine learning concepts.
C	Data-cleaning and normalization taking most necessary precautions into account. Carefully explored the space of classifier parameters/classifiers and found a good setting. Only one evaluation measure was selected or the evaluation measures were not used consistently. The overall results are correct.	The quality of discussion reflects good understanding of the underlying machine learning concepts.	The answers reflects a good understanding of the underlying machine learning concepts. No major mistake in understanding.
D	Limited data-cleaning and normalization, but fail to take crucial necessary precautions into account. A basic exploration of the space of classifier parameters and classification models. The evaluation measures are not appropriate for the dataset/task or they were not used properly and consistently. The overall results have limited utility.	The quality of discussion reflects a basic understanding of the underlying machine learning concepts. The discussion is largely based on copy/pasting material from lecture slides.	The answers reflects a basic understanding of the underlying machine learning concepts. Some gaps/errors in the reflection provided.
E	Limited data-cleaning and/or data normalization. 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.	The quality of discussion reflects poor understanding of the underlying machine learning concepts.	There are far too many mistakes in the answer -- Reflects poor understanding.
F	No data-cleaning or data normalization. Major errors in exploration. Overall results fail to convince.	No discussion.	Answer provided is incorrect