Exercise 1: Analysing Second Hand Car Sales Data with Supervised andUnsupervised Learning Models

In this exercise you will analyse a mock dataset of second hand car sales in the UK. You can download this dataset as a csv file from Canvas at the following link:

https://canvas.hull.ac.uk/files/5020067/download?download_frd=1

You will see that the dataset contains 50,000 rows, with each row corresponding to the saleof a second hand car. For each car sold, the dataset contains the following information:

- **Manufacturer** – the name of the manufacturer that produced the car.
- **Model** – the name of the model of the car.
- **Engine size** – the size of the engine, in litres.
- **Fuel type** – the type of fuel that the engine uses.
- **Year of manufacture** – the year in which the car was made.
- **Mileage** – the total number of miles that the car has been driven.
- **Price** – the price that the car was sold for, in Pound Sterling (GBP).

NOTE: whilst the names of the car manufacturers and models in this dataset may be familiarto you, be aware that this is a *mock* dataset of *imaginary* car sales data that we generated. In particular, the prices given in this dataset are not intended to be a realistic representationof the actual price of a given car. Furthermore, the years of manufacture contained in this dataset do not necessarily reflect the actual years in which a particular model was in production in the real world.

Goal

Your goal for this exercise is to explore how supervised learning models can be used to predict the price of a second hand car, based on the information contained in this dataset. You will also study how unsupervised learning techniques can be used to identify clusteringpatterns in this dataset.

You will write up the results of your analysis in the style of a scientific report. Your reportshould address the following questions:

a. Compare regression models that predict the price of a car based on a single numerical input feature. Based on your results, which numerical variable in the dataset is the best predictor for a car's price, and why? For each numerical input

feature, is the price better fit by a linear model or by a non-linear (e.g. polynomial)model?

b. Consider regression models that take multiple numerical variables as input featuresto predict the price of a car. Does the inclusion of multiple input features improve

   the accuracy of the model's prediction compared to the single-input feature modelsthat you explored in part (a)?

c. In parts (a) and (b) you only considered models that use the numerical variables fromthe dataset as inputs. However, there are also several *categorical* variables in the dataset that are likely to affect the price of the car. Now train a regression model that uses all relevant input variables (both *categorical* and *numerical*) to predict the price (e.g. a Random Forest Regressor model). Does this improve the accuracy of your results?

d. Develop an Artificial Neural Network (ANN) model to predict the price of a car based on all the available information from the dataset. How does its performance compare to the other supervised learning models that you have considered? Discuss your choices for the architecture of the neural network that you used, and describe how you tuned the hyperparameters in your model to achieve the best performance.

e. Based on the results of your analysis, what is the best model for predicting the priceof a car and why? You should use suitable figures and evaluation metrics to supportyour conclusions.

f. Use the *k*-Means clustering algorithm to identify clusters in the car sales data. Consider different combinations of the numerical variables in the dataset to use as input features for the clustering algorithm. In each case, what is the optimal numberof clusters (*k*) to use and why? Which combination of variables produces the best clustering results? Use appropriate evaluation metrics to support your conclusions.

g. Compare the results of the *k*-Means clustering model from part (f) to at least oneother clustering algorithm. Which algorithm produces the best clustering? Use suitable evaluation metrics to justify your answer.