# **COMP309**

# **Assignment 4**

# **Elliott Rose - 300540768**

**Part 1)**

*Using an initial data analysis (IDA) I was able to find the following information about the diamonds.csv data set:*

* Contains 53,940 instances.
* The X, Y and Z values in the data have minimums of 0. This means there are some  
  missing values, since doesn't make sense for the dimensions to be 0.
* There is an unnamed column which shows the counts the instance, this information isn’t valuable so can be removed.
* 3 categorical features, namely: Cut, color, and clarity.
* 7 numerical features (excluding “Unnamed”), those being: Carat, depth, table, x, y, z, price.

*Due to these findings the following steps in my preprocessing were taken:*

* The first column was removed as it has no relevance to the data/ model prediction.
* Imputed the 0 values in the dimension columns (x, y and z) with the mean.
* Encoded categorical features using an OrdinalEncoder because values in the  
  categorical features have a clear order which would influence the diamond price.
* Split the target feature from the rest of the data to perform a train test split.
* Scaled numerical features using a StandardScaler as the feature values are all using  
  different scales.

*Shown below are the results from the different models:*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *Model* | *MSE* | *RMSE* | *RSE* | *MAE* | *RUNTIME* |
| *LinearRegression()* | *2335265.55* | *1528.16* | *579.83* | *895.14* | *0.004603* |
| *KNeighborsRegressor()* | *2207730.63* | *1485.84* | *548.17* | *831.48* | *0.185915* |
| *Ridge()* | *2335259.54* | *1528.16* | *579.83* | *895.33* | *0.000000* |
| *DecisionTreeRegressor()* | *3613430.34* | *1900.90* | *897.19* | *1050.34* | *0.011345* |
| *RandomForestRegressor()* | *1990624.58* | *1410.89* | *494.26* | *792.20* | *0.517433* |
| *GradientBoostingRegressor()* | *1879375.87* | *1370.90* | *466.64* | *776.10* | *0.015626* |
| *SGDRegressor()* | *2347426.68* | *1532.13* | *582.85* | *883.99* | *0.015625* |
| *SVR()* | *7577653.59* | *2752.75* | *1881.48* | *1349.92* | *63.247901* |
| *LinearSVR()* | *3301263.80* | *1816.94* | *819.68* | *985.26* | *0.000000* |
| *MLPRegressor()* | *2122725.74* | *1456.96* | *527.06* | *823.62* | *0.017392* |