CS4055 Project

For our dataset we chose California Housing Data (1990) from Kaggle the dataset has 20000 rows and 10 columns. We aimed to predict median house value. The dataset contains details on housing blocks it contains information such as the amount of people who lived in that area, the total bedrooms in the area and the median income of the residents.

Our final model was a random forest it was our best performing model giving us a R squared score of .61. our other models were linear regression and decision tree both of which didn’t achieve about a 37 R squared value. We didn’t achieve a very high accuracy when predicting our target variable, but we believe this is due to the data not providing enough info needed for an accurate prediction. To improve our data we created 3 new features rooms per household, bedrooms per room and people per house these features made a decent improvement upon our accuracy score. We also removed any data which was on an island as all the island data were outliers and skewed our data. When we fit our models we scaled our data too to negate the effect of the large variance in our prediction values.

When we compared the heatmap of median house values plotted using longitude and latitude we seen the higher values were to the left of the graph. When you look at the heatmap.html where population is plotted over a google maps still of California we see that the left of the graph is the coast. This indicated that the median house value increased the closer to the coast you went. Some quick analysis of the data showed this to be true.