Luke Scott

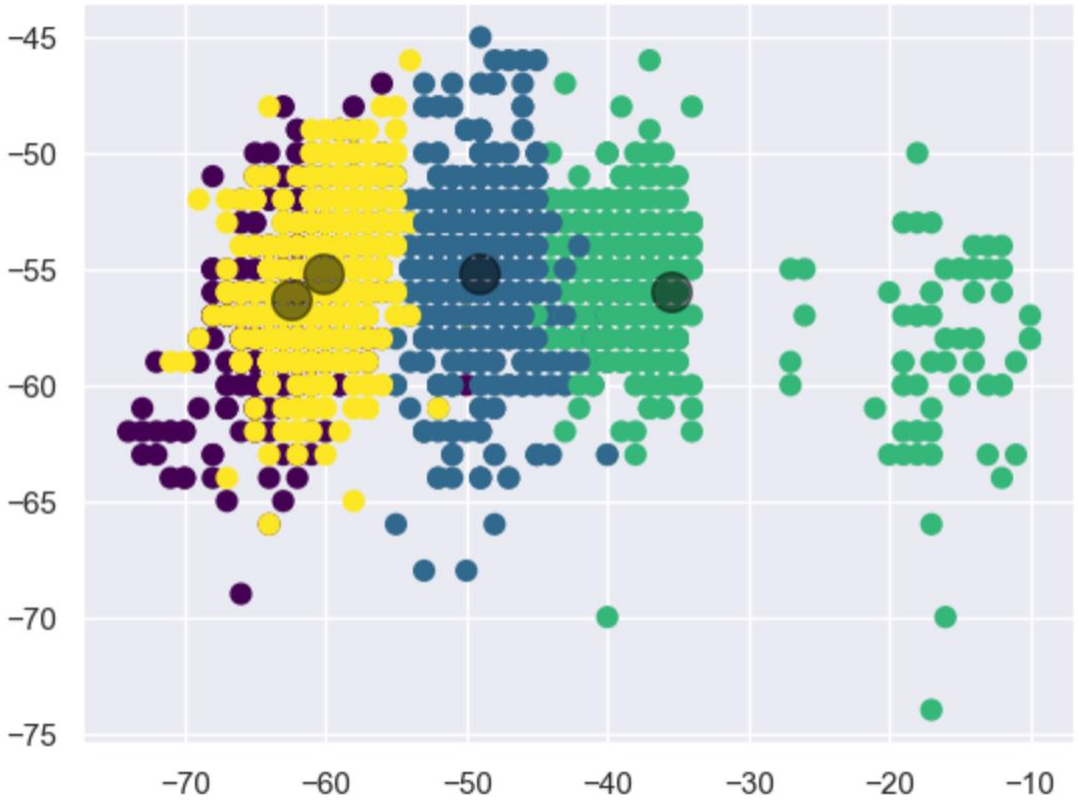
COSC 311

Project 2

Dr. Wang

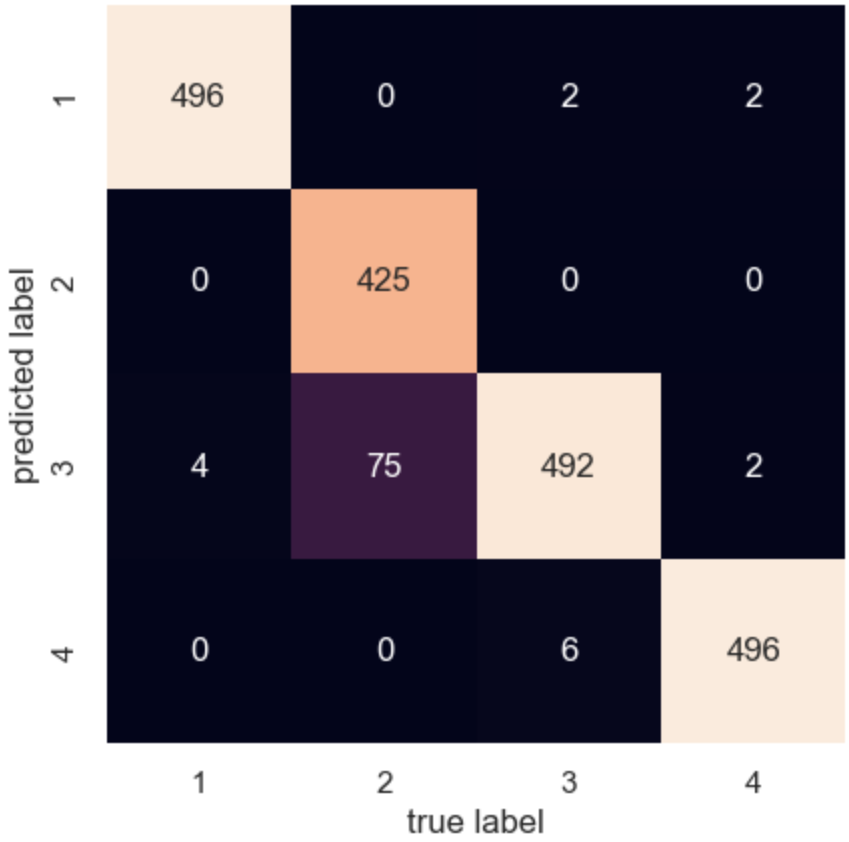
Project Report

------ Part 1 -------



1)

a. I loaded the data, then split it into X and Y values, Y corresponding with the room numbers, and X corresponding with the signals. I then used kmeans to find the centers of the clusters, and plotted them on a scatterplot.

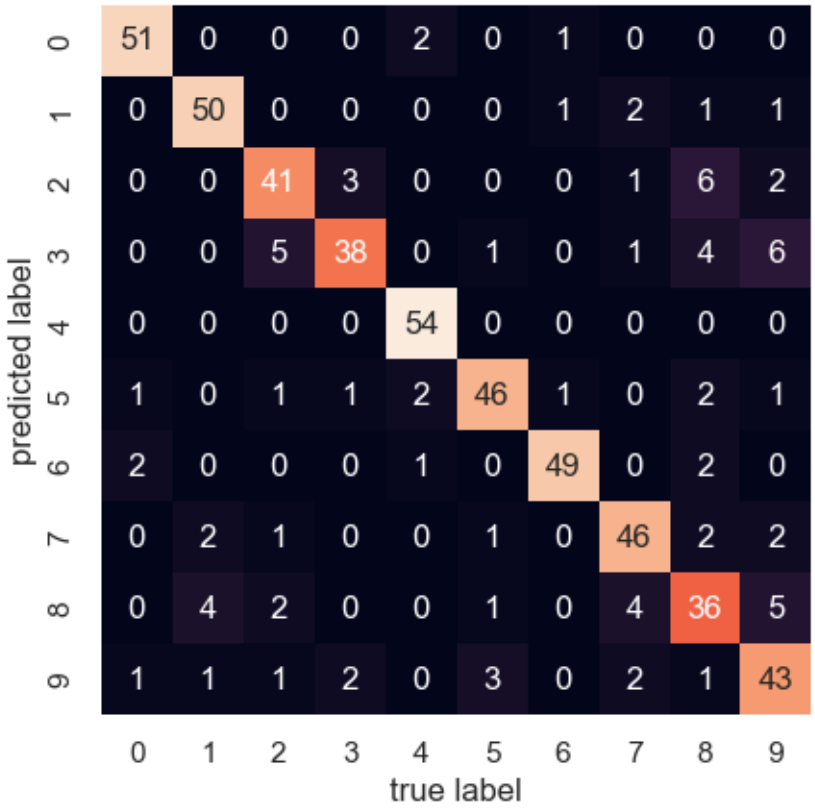


2)

a. The model was highly accurate when predicting rooms one and two, however struggled with predicting rooms 2 and 3. 75 samples were incorrectly predicted into room 3 but actually belonged to room 2.

------ Part 2 -------

1) I conducted a PCA analysis on the digits dataset and found a variance of around 84% when using 3 components. I increased this value to 5, acquiring a 94% accuracy and then transformed the data into 5 dimensions.

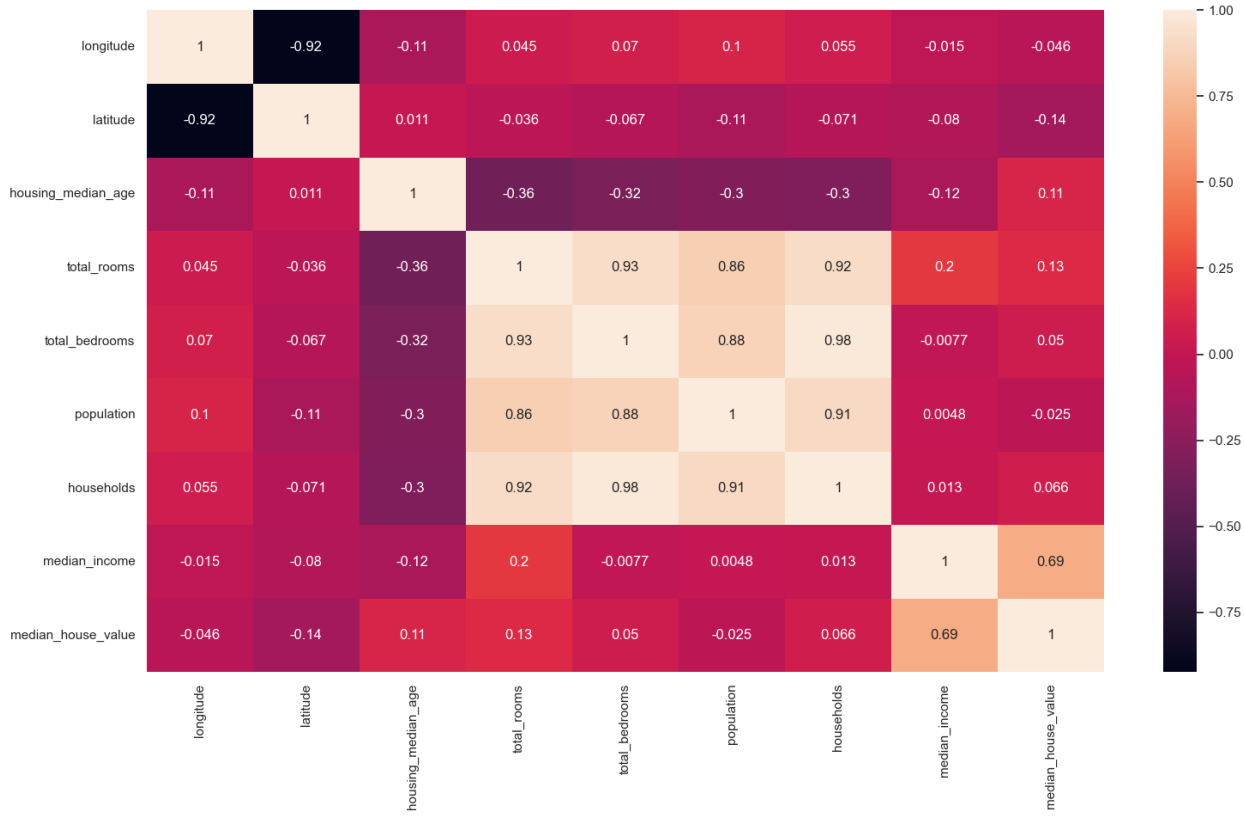


2)

a. To build the classification model, I used a Decision Tree Classifier and used 10 folds for the CVT. The CVT average accuracy was 84% and was successful.

------ Part 3 -------

1)



a. I conducted a correlation matrix on the dataset, and found that 'housing_median_age', 'total rooms', and 'median_income' had the highest correlation with 'median_house_value'. I chose these three attributes to split the data.

2)

```
X = housing_data.drop(labels = ['latitude','longitude',
                                'median_house_value','ocean_proximity',
                               'index','level_0','total_bedrooms',
                               'households','population'], axis = 1)
y = CD['median_house_value']
print(X)
print(y)
```

```
X_train, X_test , y_train, y_test = \
    train_test_split(X, y, test_size = 0.4, random_state = 0 )

X_train.shape, y_train.shape, X_test.shape, y_test.shape

model = LinearRegression()
model
model.fit(X_train, y_train)
predict = model.predict(X_test)
compare1 = pd.DataFrame({"Predicted":predict, "Actual":y_test})
```

a. Split the data.

```
The performance for testing set
--------------------------------
MAE is  60378.62097652892
MSE is  6607470379.44292
RMSE is  81286.34805084381
```

3)

a. Using the training data, I built a Multiple Linear Regression model and tested using the test data. The corresponding performance of the model is shown.