Machine Learning, Spring 2023: Project 2

House Price

KAMAL SUBEDI(G35111003)

Platform: Jupyter Notebook  
Device: MacOS M2  
Programming language: Python3, Python libraries  
Dataset: https://www.kaggle.com/uciml/pima-indians-diabetes-database

**Data Description:**

This dataset contains data on 21613 different housing sale price based on various factor for Kind County in seattle. It has the data for one complete i.e from May 2014 to May 2015. Using various features like bedroom, square living, sqftlot, bathroom, floor and view etc, we have to predict sales price of particular house in that location. The dataset consists of several independent variables and one target (dependent) variable. Independent variables include the number of Bedroom, square living, sqftlot, bathroom etc and independent output variable which is price.

**Algorithm Description**

For the Linear regression model, we will use the various independent variables to predict the sales price of house.

Before building the Linear regression model, we need to split the dataset into training and testing sets. We will use a 80-20 split, where 80% of the data will be used for training and 20% will be used for testing. We will randomly select the observations for each set.

I choose a 80-20 split because we want to have enough data to build an accurate predict, but we also want to have enough data to evaluate its performance. Using a larger training set may result in a more accurate predict, but it may also increase the risk of overfitting to the training data, resulting in poor performance on new, unseen data. Additionally, using a smaller testing set may increase the variability of the evaluation metrics, making it harder to draw conclusions about the classifier's performance. 80-20 split strikes a reasonable balance between these concerns.

Then use, Euclidean and Manhattan distance metric to predict the neighbor and classify.

Below, we can see the scatter plot of feature variable and target variable. From the plot below, we can see that sqft\_living and grad highly influence the target variable. As we are asked to take the single independent variable. So, we are choosing sqft\_living as feature variable and dropping other independent variables.

Chart, treemap chart

Description automatically generated

Fig: Scatter plot

**Algorithm Result:**

We got the confusion matrix and accuracy value as below:

Confusion matrix = [[82 17]

[23 32]]

A picture containing text, clock

Description automatically generatedChart, treemap chart

Description automatically generated  
**Fig: Confusion matrix**



Chart, line chart

Description automatically generated

Fig: Graph between accuracy and K.

Here, we can see that, the accuracy is maximum is when K=5. And the K is increasing the accuracy score got deceasing after k=7.

**Runtime:**

