Shri Vile Parle Kelavani Mandal's
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING
(Autonomous College Affiliated to the University of Mumbai)
NAAC Accredited with "A" Grade (CGPA : 3.18)

Department of Computer Science and Engineering (Data Science)

Subject: Machine Learning – I (DJ19DSC402)

AY: 2022-23

Experiment 10

(Mini Project)

Aim: Design a classifier to solve a specific problem in the given domain.

Tasks to be completed by the students:
Select a specific problem from any of the given domain areas, such as: Banking, Education, Insurance, Government, Media, Entertainment, Retail, Supply chain, Transportation, Logistics, Energy and Utility.
Task 1: Select appropriate dataset, describe the problem and justify the suitability of your dataset.
Task 2: Perform exploratory data analysis and pre-processing (if required).
Task 3: Apply appropriate machine learning algorithm to build a classify. Perform appropriate testing of your model.
Task 4: Submit a report in the given format.

- Introduction
- Data Description
- Data Analysis
- Reason to select machine learning model
- Algorithm
- Result Analysis
- Conclusion and Future Scope.
- Python notebook

Task5: Presentation

Shri Vile Parle Kelavani Mandal's

DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING

(Autonomous College Affiliated to the University of Mumbai)

NAAC Accredited with "A" Grade (CGPA : 3.18)

Department of Computer Science and Engineering (Data Science)

Report on Mini Project

Machine Learning -I (DJ19DSC402)

AY: 2022-23

TITLE OF THE PROJECT: HOUSE PRICE PREDICTION

NAME:DIVYESH KHUNT **SAP ID:60009210116**

Guided By

Dr. Surekha Janrao

CHAPTER 1: INTRODUCTION

The dataset provides information on the properties available for sale in various cities of India. The data includes features such as the location of the property, its size in square feet, number of bedrooms, whether it is under construction or ready to move, and other related features. The target variable is the price of the property in lakhs (one lakh is equivalent to 100,000 Indian Rupees). This dataset can be used to develop models for predicting the prices of properties based on their features, which could be useful for real estate agents and buyers. Additionally, the dataset could be analyzed to gain insights into the Indian real estate market and the factors that influence property prices.

Problem statement:

Today with the increasing population there is also an increase of demand on land and area. For this many brokers and real estate agents are available which help people to get the desired houses. With the help of this model, we can replace the agents with and app or web consisting of this model. The goal of this project is to build a model that can predict the price of apartment in different locations based on various attributes such as location, square feet, resale value, etc. This model can be used by real estate agents or buyers to make informed decisions about the price of the apartment.

CHAPTER 2: DATA DESCRIPTION

Dataset Attributes:

POSTED_BY: Whether the listing was posted by an owner or a dealer.

UNDER_CONSTRUCTION: Whether the apartment is under construction or ready-to-move. RERA: Whether the apartment is registered under

RERA or not. BHK_NO: Number of bedrooms in the apartment.

BHK_OR_RK: Whether the apartment is a BHK or a studio (RK).

SQUARE_FT: Total square footage of the apartment.

READY_TO_MOVE: Whether the apartment is ready to move.

RESALE: Whether the apartment is a resale or a new property.

ADDRESS: The location of the apartment.

LONGITUDE: The longitude of the apartment's location.

LATITUDE: The latitude of the apartment's location.

TARGET(PRICE_IN_LACS): The target variable or the price of the apartment in lacs

CHAPTER 3: DATA ANALYSIS

Missing values: The dataset does not contain any missing values.

Data distribution: We can plot histograms for the numerical features to understand their distributions. For example, the feature "SQUARE_FT" seems to have a right-skewed distribution.

Correlation: We can also calculate the correlation matrix to see the relationships between the features. It seems that the feature "BHK_NO." has a high positive correlation with the target variable, while the features "READY_TO_MOVE" and "RESALE" have a negative correlation.

- In the given dataset, **SVM** is chosen as a regression algorithm to predict the target variable TARGET(PRICE_IN_LACS) which represents the price of the house in Indian Rupees. Here are some reasons why SVM is a good choice for this problem:
- SVM is a powerful algorithm for both classification and regression problems. It is known to have good performance on small-to-medium sized datasets.
- SVM can handle non-linear relationships between features and the target variable. It does this by mapping the features to a higher-dimensional space where the relationship can be linear, using a kernel function.
- In the given dataset, we have a relatively small number of features (11) and a moderate number of samples (30,000). This is a good scenario for SVM, as it can find an optimal decision boundary that maximizes the margin between the support vectors while avoiding overfitting.
- SVM is a robust algorithm that is less sensitive to outliers in the data. This is important in the context of real estate prices, as there may be outliers in the data due to unique features of certain properties.
- SVM can handle both continuous and categorical features, making it flexible for datasets with mixed data types like this one.
- After training the SVM model on the given dataset, we obtained a testing R-squared score of 0.39 and a mean squared error of 6911.

These scores indicate that the SVM model has a moderate level of accuracy in predicting house prices. The line graph of training and testing accuracy shows that the model's training and testing accuracy increase as the number of training samples increases. However, there is a significant gap between the training and testing accuracy, which suggests that the model may be overfitting the training data.

- Overall, SVM is a good choice for predicting house prices in this dataset due to its flexibility, robustness, and ability to handle non-linear relationships. However, further tuning of hyperparameters and feature engineering may be necessary to improve the model's accuracy and reduce overfitting.