A Project/Dissertation Review-1 Report

on

House Price Prediction

Submitted in partial fulfillment of the
requirement for the award of the degree of

Program Name with Specialization

Under The Supervision of
Name of Supervisor : MD. ANAS.
Designation

Submitted By

Prateek Kumar Singh	21SCSE1010884
Nitesh Kumar	21SCSE1010954
Md.Rizwan Ahmad	21SCSE1010811
Aditya Prajapati	21SCSE1420051

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
GALGOTIAS UNIVERSITY, GREATER NOIDA
INDIA

MONTH, YEAR

Table of Contents

1.Abstract

Data science has grown rapidly over the last decade. In machine learning, many applications and algorithms are evolving day by day. One such application found in trade publications is home price forecasting. House prices are rising every year, making it necessary to model house price projections. These built models help customers purchase homes that meet their needs. The proposed work makes use of the following attributes or characteristics of the house. B. The number of bedrooms available in the house, the age of the house, the travel opportunities of the place, the school facilities available near the house, the shopping malls available near the place of the house. Housing availability and house price projections based on desirable features of housing are modeled in the proposed work and the model is built for a small town located in the West Godavari district of Andhra Pradesh. This work includes decision tree classification, decision tree regression, and multiple linear regression, implemented using Scikit-Learn machine learning tools.

Data Mining is extracting knowledge or useful pattern from large databases. Classification is one of the data mining functionalities, employed for finding a model for class attribute which is a function of other attribute values

Index Terms-

- Decision tree
- House price prediction
- decision tree regression
- multiple linear regression.

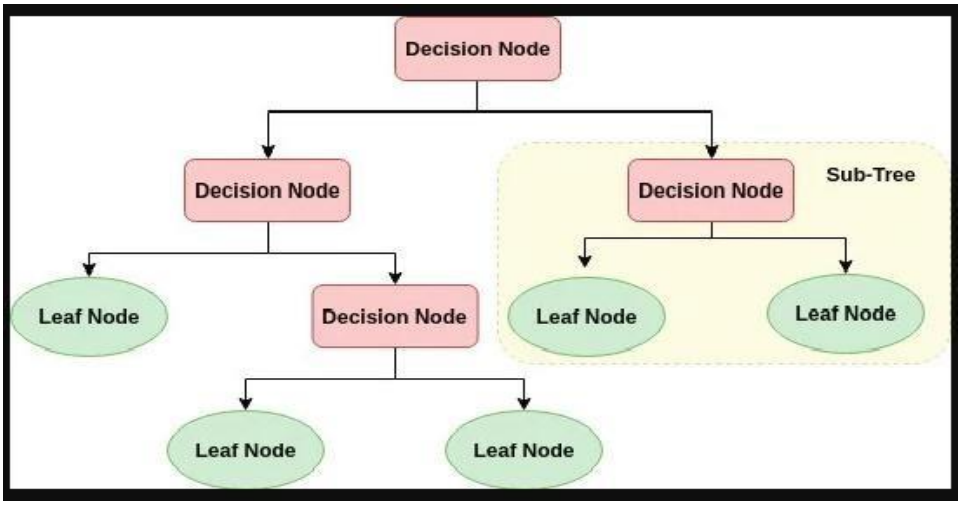
CHAPTER-1
Introduction

Decision Tree is a tool, which can be employed for Classification and Prediction. It has a tree shape structure, where each and every internal node represents test on an attribute and the branches out of the node denotes the test outcomes.
80% of the known dataset can be used as training set and 20% can be used as test data set. Each record in the dataset denotes X and Y values, where X is a set of attribute values and Y is the class of the record which is the last attribute in the dataset. Using the training set Decision Tree Classifier model is constructed and tested

with test data to identify the accuracy level of the classifier.
Decision Tree formation as shown in fig. 1 employs divide and conquer strategy for splitting the training data into subsets by testing an attribute value. This involves attribute selection measures; the attribute which is to be tested first is the one which is having high information gain. Same splitting process is recursively performed on the subsets derived [2]. The splitting process of a subset ends when all the tuples belong to the same attribute value or when no remaining attributes or instances are left with. Decision Tree formation does not need any basic domain knowledge. It can handle data of high dimensions as well. Decision Tree Classifiers have good accuracy in classification.
Once the Decision Tree is formed, new instances can be classified easily by tracing the tree from root to leaf node. Classification through Decision Tree does not require much computation. Decision Trees are capable of handling both continuous and Categorical type of attributes.
To avoid generation of meaningless and unwanted rules in Decision Trees, tree should not be deeper which results in over fitting. Such a tree with over fitting works more accurate with training data and less accurate with test data. Pre pruning and Post pruning are the techniques used in Decision Tree to reduce the size of the trees and avoid over fitting. In Post Pruning the Decision Tree branches and hence the level (depth) of the tree are reduced after completely building the tree. In Pre Pruning, care is taken to avoid over fitting while building the tree itself.
Decision Trees find its major applications in areas such as medicine, weather, finance, entertainment, sports, etc. Decision Tress can also be used for prediction, data manipulation and handling of missing values. As an example in digital mammography it is used for classifying tumor cells and normal cells [3].

This paper discusses about an application of Decision Tree, for purchasing a house in a city based on attribute values such as transport facilities, number of bed rooms, and availability of schools, shopping facilities and

CHAPTER-2
Literature Survey



I. RELATED WORK

Patel and Upadhyay [4] have discussed various pruning methods and their features and hence pruning effectiveness is evaluated. They have also measured the accuracy for glass and diabetes dataset, employing WEKA tool, considering various pruning factors. ID3 algorithm splits attribute based on their entropy. TDIDT algorithm is one which constructs a set of Information gain is partially inclined towards tests with several outcomes. Hence information Gain obtained by splitting on attribute is highest and such a splitting is hopeless for classification. Followed by ID3 its successor C4.5 arrived, which used Gain Ratio as in [8] has utilized decision tree approach for finding the resale prices of houses based on their significant characteristics. In this paper, hedonic based regression method is employed for identifying the relationship between the prices of the houses and their significant characteristics. Ong et al. [9] and Berry et al.

[10] have also used hedonic based regression for house prediction based on significant characteristics Shinde and Gawande [11], predicted the sale price of the houses using various machine learning algorithms like, lasso, SVR, Logistic regression and decision tree and compared the accuracy. Alfiyatin et al. [12] has modeled

a system for house price prediction using Regression and Particle Swarm Optimization (PSO). In this paper, it has been proved that the house price prediction accuracy is improved by combining PSO with regression. problems in Random Forests, Decision Trees, and Categorical Predictors. Using three real data sets, the authors have illustrated how the absent levels affect the performance of the predictors

II. ATTRIBUTE SELECTION MEASURES

Redundant attributes which are considered inappropriate for the data mining task is removed using a process called Attribute selection [14]. Hence a desirable set of attributes results which is the ultimate goal of Attribute selection algorithms. This attribute set produces analogous classification results as that of using all the attributes. Best split attributes selection measures are defined in terms of impurity reduction from parent.

This work includes two parts namely,

(i) Decision Tree Classifier is used to predict the availability of houses as per the users' requirement constraints and it produces responses like yes or no respectively to tell whether a house is available or not.

(ii) Decision tree regression and Multiple Linear Regression methods are used to predict the prices of the houses.

A real time dataset is prepared by analyzing the location named Tadepalligudem of West Godavari District in Andhrapradesh of India. The dataset contains the following features of the houses such as Number of bedrooms, age of the house, transport facility, schools available in the nearby location and shopping facilities.

The proposed method helps to search houses in big cities based on the following attributes.

1. Number of bedrooms (1BHK, 2BHK and 3BHK).
2. Transport facility such as availability of bus facility, train facility and flight facility.

3. School facility such as availability of Government schools, matriculation and CBSE.
4. Shopping facility such as small markets, general stores, shopping malls
5. Prices of the houses from 10 lakhs to 30 lakhs.
6. Age of the house varying from one to five years.

The proposed work is implemented using Scikit Learn, a machine learning tool.

A. Scikit Learn

The Scikit-Learn (SK Learn) is a Python Scientific toolbox for machine learning and is based on SciPy, which is a well-established Python ecosystem for science, engineering and mathematics. Scikit-learn provides an ironic environment with state of the art implementations of many wellknown machine learning algorithms, while sustaining an easy to use interface tightly integrated with the Python language [16],[17]. Scikit-learn features various functionalities like Clustering algorithms, Regression, Classification including random forests, gradient boosting, support vector machines, *k*-means and DBSCAN, and it has been designed to interoperate in conjunction with the Python scientific and numerical libraries SciPy and NumPy.

The step by step implementation using SK Learn is as follows.

Step 1: Import the required libraries. Step 2: Load the dataset.
Step 3: Assign the values of columns 1 to 6 in the Dataset to "X".
Step 4: Assign the values of column 7 which is the class label to "Y".
Step 5: Fit decision tree classifier to the dataset. Step 6: Predict the class label for the test data.

The decision tree classifier shown in fig. 2 is constructed using Scikit Learn and the respective specifications involved are as shown below. It uses Gini index as the measure to select the relevant attributes for testing and splitting the training set.

```
DecisionTreeClassifier(class_weight=None,
criterion='gini',
max_depth=None, max_features=None, max_leaf_nodes=None,
min_impurity_decrease=0.0,
min_impurity_split=None,
min_samples_leaf=1,
```

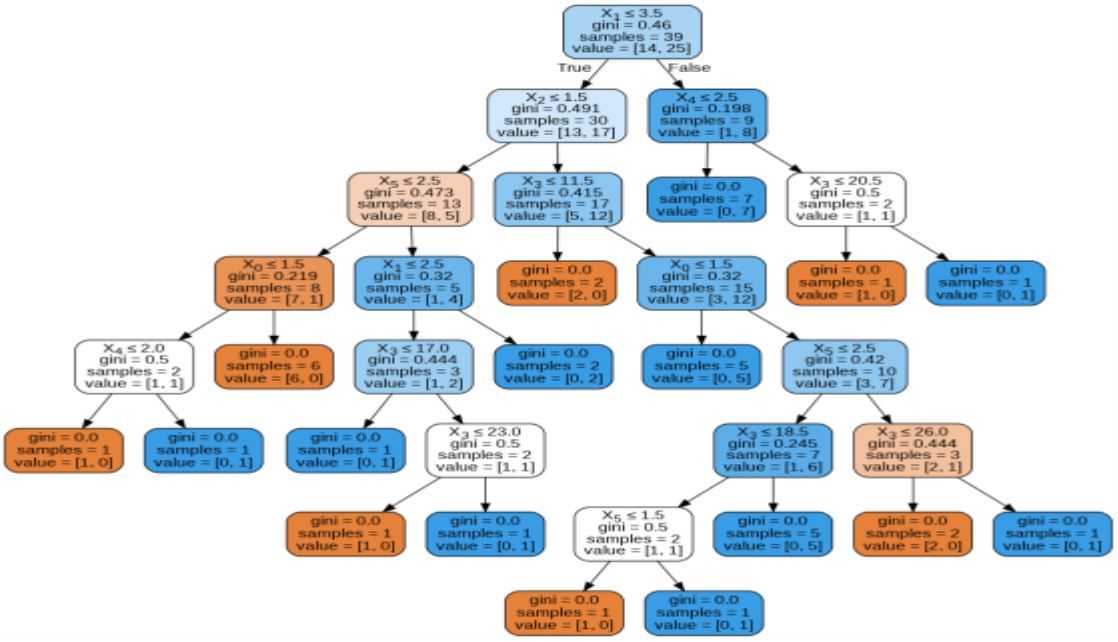
min_samples_split=2,
min_weight_fraction_leaf=0.0,
presort=False, random_state=None, splitter='best')

EXPERIMENTAL EVALUATION

A. *House Availibility Prediction*

The Decision tree output for classifying the availability of houses has discrete binary values like Yes or No. The output of the Decision tree Regression used for house price prediction is a continuous one. The continuous values (Prices) are predicted with the help of a decision tree regression model.

Table 1 shows the sample dataset with ten records considered with some essential features for the area, Tadepalligudem selected in West Godavari District of Andhra Pradesh. But the original dataset consists of 50 records of different combinations. In the table for attribute 3, travelling facility, range is taken from 1 to 3, where 1 denotes Bus facility is available nearby, 2 denotes Bus and train facilities are available nearby and 3 denotes both bus and train facilities are farther to the house location. In shopping facility attribute, 1 denotes less shopping facility with vegetable market and small grocery shops, 2 denotes departmental stores and some small malls and 3 denotes super markets with all facilities. In school attribute 1 denotes that government schools are alone available nearby, 2 denotes that government and private schools are available nearby and 3 denotes that government, private and CBSE schools all are available



I. CONCLUSION AND FUTURE SCOPE

This article uses the most fundamental machine learning algorithms like decision tree classifier, decision tree regression and multiple linear regression. Work is implemented using Scikit-Learn machine learning tool. This work helps the users to predict the availability of houses in the city and also to predict the prices of the houses. Two algorithms like decision tree regression and multiple linear regression were used in predicting the prices of the houses. Comparatively the performance of multiple linear regression is found to be better than the decision tree regression in predicting the house prices. In future the dataset can be prepared with more features and advanced machine learning techniques can be for constructing the house price prediction model.