IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 12, Special Issue 2, April 2023

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

|| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | SJIF: 8.423|| Volume 12, Special Issue 2, April 2023 ||

8th National Conference on Advanced Computing Technologies (NCACT'23)

Organized by

Department of CSE, Velammal Engineering College, Chennnai, India

Prediction of House Price Using SMLT

Mrs Sruthi Nath[1], Nithi Sneha[2] , Vandana S[3]

Assistant Professor, Department of Computer Science Engineering, Velammal Engineering College, Chennai, Tamil Nadu, India[1]

U.G. Students, Department of Computer Science and Engineering, Velammal Engineering College, Chennai, Tamil Nadu, India[2-4]

ABSTRACT : Generally, predicting how the house price will perform is one of the most difficult things to do. It can be described as one of the most critical process to predict that. This is a very complex task and has uncertainties. To prevent this problem in One of the most interesting (or perhaps most profitable) time series data using machine learning techniques. Hence, house price prediction has become an important research area. The aim is to predict machine learning based techniques for house price prediction results in error based calculation. The analysis of dataset by supervised machine learning technique(SMLT) to capture several information's like, variable identification, uni-variate analysis, bi-variant and multi-variant analysis, missing value treatments and analyze the data validation, data cleaning/preparing and data visualization will be done on the entire given dataset. To propose a machine learning-based method to accurately predict the house price Index value by prediction results in the form of house price increase or stable state best regression from comparing supervise machine learning algorithms. Additionally, to compare and discuss the performance of various machine learning algorithms. dataset with evaluation classification report, identify the confusion matrix and to categorizing data from priority and the result shows that the effectiveness of the proposed machine learning algorithm technique can be compared with best accuracy MAE,MSE,R^2 .

I. INTRODUCTION

I.

Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data, and apply knowledge and actionable insights from data across a broad range of application domains. Data science can be defined as a blend of mathematics, business acumen, tools, algorithms and machine learning techniques, all of which help us in finding out the hidden insights or patterns from raw data which can be of major use in the formation of big business decisions.

II. RELATED WORK

Urban housing price is widely accepted as an economic indicator which is of both business and research interest in urban computing. However, due to the complex nature of influencing factors and the sparse property of transaction records, to implement such a model is still challenging. To address these challenges, in this work, we study an effective and fine-grained model for urban subregion housing price predictions. Compared to existing works, our proposal improves the forecasting granularity from city-level to mile-level, with only publicly released transaction data. We employ a feature selection mechanism to select more relevant features. a fine-grained forecasting model, JGC MMN, for subregion spatiotemporal housing price prediction. In particular, we modify the structure of DenseNet and adopt the method of bagging by fusing the KF-based method to improve the accuracy.

III. METHODOLOGY

Data Pre-processing:

Validation techniques in machine learning are used to get the error rate of the Machine Learning (ML) model, which can be considered as close to the true error rate of the dataset. If the data volume is large enough to be representative of the population, you may not need the validation techniques. However, in real-world scenarios, to work with samples of data that may not be a true representative of the population of given dataset. To finding the missing value, duplicate value and description of data type whether it is float variable or integer. The sample of data used to provide an unbiased evaluation of a model fit Data Processing Test dataset Training dataset Regression ML Algorithm Model on the training dataset while tuning model hyper parameters. The evaluation becomes more biased as skill on the validation

|| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | SJIF: 8.423|| Volume 12, Special Issue 2, April 2023 ||

8th National Conference on Advanced Computing Technologies (NCACT'23)

Organized by

Department of CSE, Velammal Engineering College, Chennnai, India

dataset is incorporated into the model configuration. The validation set is used to evaluate a given model, but this is for frequent evaluation. It as machine learning engineers use this data to fine-tune the model hyper parameters. Data collection, data analysis, and the process of addressing data content, quality, and structure can add up to a time-consuming to-do list. During the process of data identification, it helps to understand your data and its properties; this knowledge will help you choose which algorithm to use to build your model.
A number of different data cleaning tasks using Python's Pandas library and specifically, it focus on probably the biggest data cleaning task, missing values and it able to more quickly clean data. It wants to spend less time cleaning data, and more time exploring and modeling.

Some of these sources are just simple random mistakes. Other times, there can be a deeper reason why data is missing. It's important to understand these different types of missing data from a statistics point of view. The type of missing data will influence how to deal with filling in the missing values and to detect missing values, and do some basic imputation and detailed statistical approach for dealing with missing data. Before, joint into code, it's important to understand the sources of missing data. Here are some typical reasons why data is missing:

- User forgot to fill in a field.
- Data was lost while transferring manually from a legacy database.
- There was a programming error.
- Users chose not to fill out a field tied to their beliefs about how the results would be used or interpreted.

Variable identification with Uni-variate, Bi-variate and Multi-variate analysis:

- import libraries for access and functional purpose and read the given dataset
- General Properties of Analyzing the given dataset
- Display the given dataset in the form of data frame
- show columns
- shape of the data frame
- To describe the data frame
- Checking data type and information about dataset
- Checking for duplicate data
- Checking Missing values of data frame
- Checking unique values of data frame
- Checking count values of data frame
- Rename and drop the given data frame
- To specify the type of values
- To create extra columns

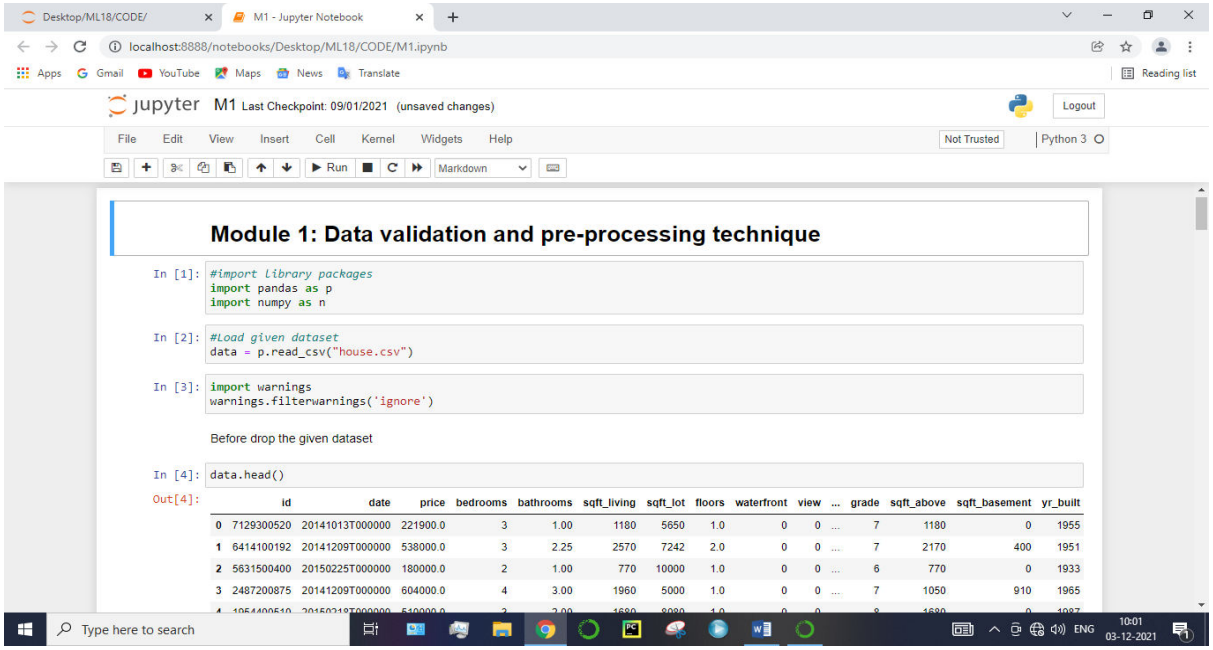
Data Validation/ Cleaning/Preparing Process
Importing the library packages with loading given dataset. To analyzing the variable identification by data shape, data type and evaluating the missing values, duplicate values. A validation dataset is a sample of data held back from training your model that is used to give an estimate of model skill while tuning models and procedures that you can use to make the best use of validation and test datasets when evaluating your models. Data cleaning / preparing by rename the given dataset and drop the column etc. to analyze the uni-variate, bi-variate and multi-variate process. The steps and techniques for data cleaning will vary from dataset to dataset. The primary goal of data cleaning is to detect and remove errors and anomalies to increase the value of data in analytics and decision making.

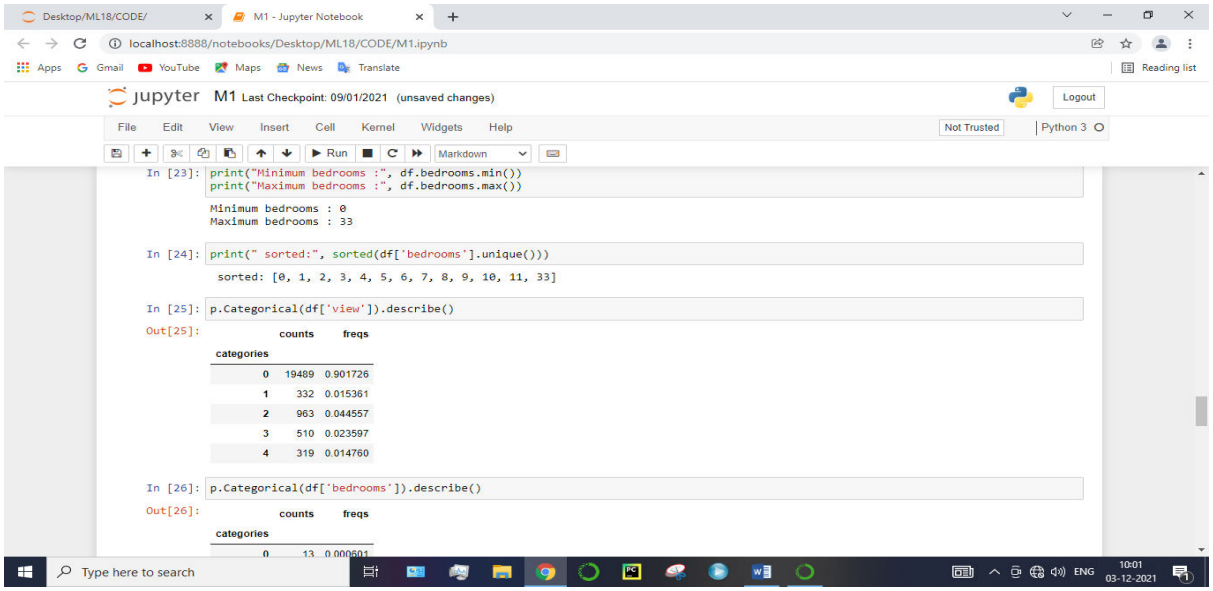
|| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | SJIF: 8.423|| Volume 12, Special Issue 2, April 2023 ||

8th National Conference on Advanced Computing Technologies (NCACT'23)

Organized by

Department of CSE, Velammal Engineering College, Chennnai, India





Preparing the Dataset :

This dataset contains 15129 training 6213 testing records extracted, which were regression pattern:

- Predicting to the House price using regression pattern.

Exploratory Data Analysis of house Prediction

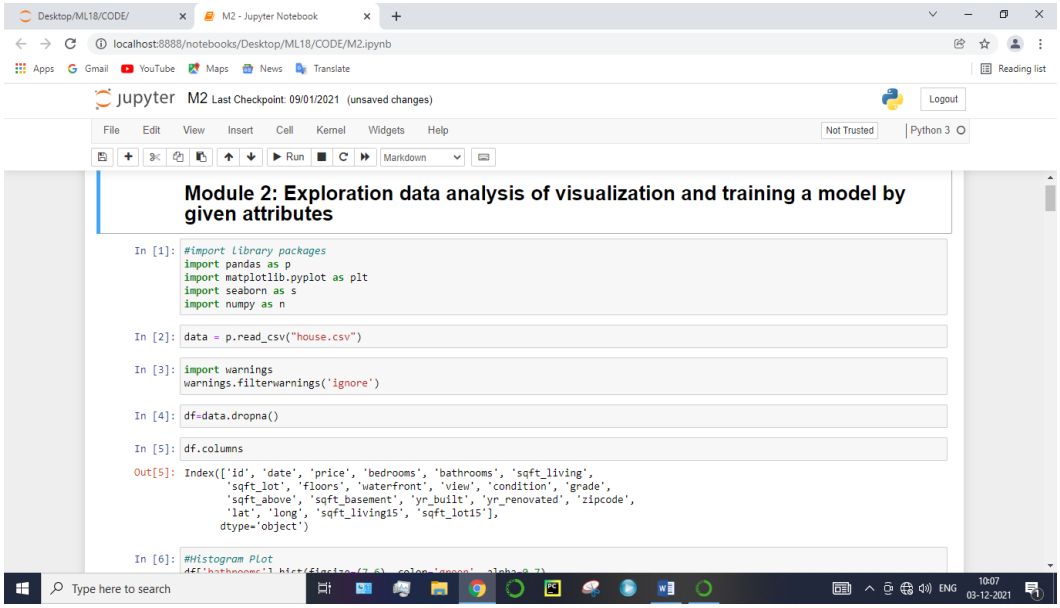
Multiple datasets from different sources would be combined to form a generalized dataset, and then different machine learning algorithms would be applied to extract patterns and to obtain results with MAE.

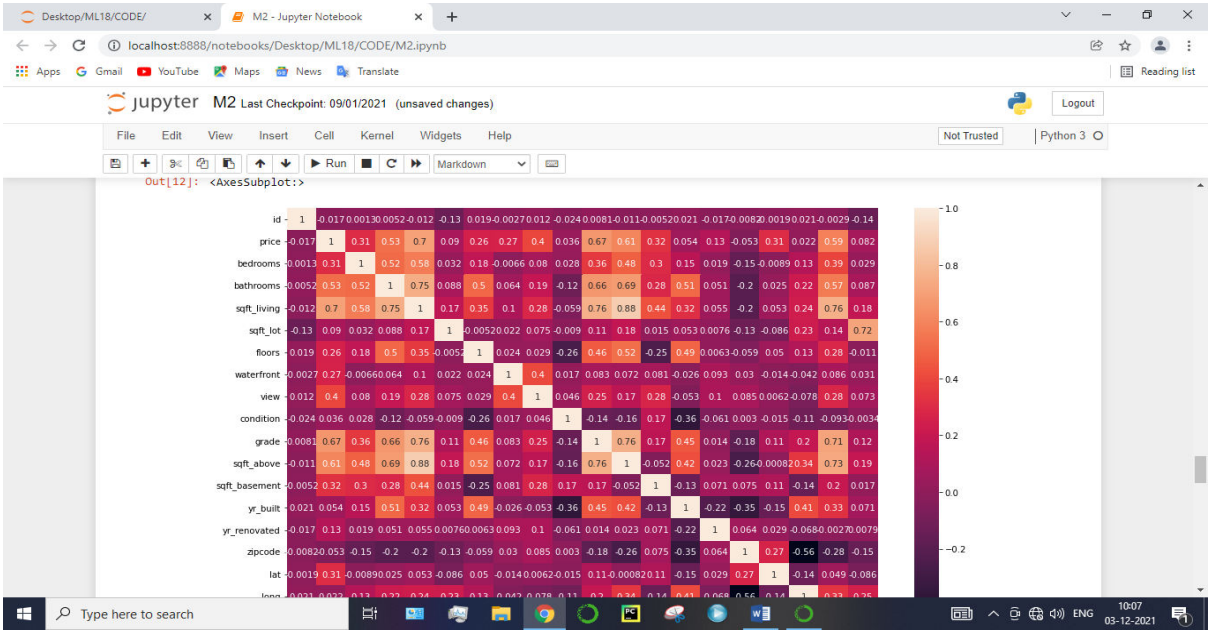
|| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | SJIF: 8.423|| Volume 12, Special Issue 2, April 2023 ||

8th National Conference on Advanced Computing Technologies (NCACT'23)

Organized by

Department of CSE, Velammal Engineering College, Chennnai, India





Comparing Algorithm with prediction in the form of best accuracy result

It is important to compare the performance of multiple different machine learning algorithms consistently and it will discover to create a test harness to compare multiple different machine learning algorithms in Python with scikit-learn. It can use this test harness as a template on your own machine learning problems and add more and different algorithms to compare. Each model will have different performance characteristics. Using resampling methods like cross validation, you can get an estimate for how accurate each model may be on unseen data. It needs to be able to use these estimates to choose one or two best models from the suite of models that you have created. When have a new dataset, it is a good idea to visualize the data using different techniques in order to look at the data from different perspectives. The same idea applies to model selection. You should use a number of different ways of looking at the estimated accuracy of your machine learning algorithms in order to choose the one or two to finalize. A way to do this is to use different visualization methods to show the average accuracy, variance and other properties of the distribution of model accuracies.

|| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | SJIF: 8.423|| Volume 12, Special Issue 2, April 2023 ||

8th National Conference on Advanced Computing Technologies (NCACT'23)

Organized by

Department of CSE, Velammal Engineering College, Chennnai, India

Data Wrangling

In this section of the report will load in the data, check for cleanliness, and then trim and clean given dataset for analysis. Make sure that the document steps carefully and justify for cleaning decisions.

Data collection

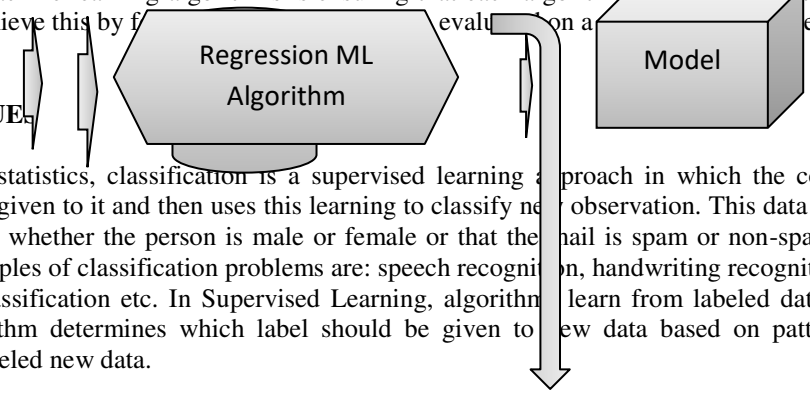
The data set collected for predicting given data is split into Training set and Test set. Generally, 7:3 ratios are applied to split the Training set and Test set. The Data Model which was created using machine learning algorithms are applied on the Training set and based on the test result MAE, Test set process is done.

Building the regression model

The predicting the house price problem, decision tree regression model is effective because of the following reasons: It provides better results in regression problem.

- It is strong in preprocessing outliers, irrelevant variables, and a mix of continuous, regression and discrete variables.
- It produces out of bag estimate error which has proven to be unbiased in many tests and it is relatively easy to tune with.

The key to a fair comparison of machine learning algorithms is ensuring that each algorithm is evaluated in the same way on the same data and it can achieve this by f evalu on a ess.



ALGORITHM AND TECHNIQUES

Algorithm Explanation

In machine learning and statistics, classification is a supervised learning a proach in which the computer program learns from the data input given to it and then uses this learning to classify ne observation. This data set may simply be bi-class (like identifying whether the person is male or female or that the ail is spam or non-spam) or it may be multi-class too. Some examples of classification problems are: speech recogni n, handwriting recognition, bio metric identification, document classification etc. In Supervised Learning, algorithm learn from labeled data. After understanding the data, the algorithm determines which label should be given to ew data based on pattern and associating the patterns to the unlabeled new data.

Random Forest Regression:

Random Forest Regression is a supervised learning algorithm that uses **ensemble learning** method for regression.

Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model.

we have a basic understanding of how the Random Forest Regression model works, we can assess its performance on a real-world dataset. Similar to my previous posts, I will be using data on House Sales in King County, USA.

After importing the libraries, importing the dataset, addressing null values, and dropping any necessary columns, we are ready to create our Random Forest Regression model

From the sklearn package containing ensemble learning, we import the class **Random Forest Regressor**, create an instance of it, and assign it to a variable.

The parameter *n_estimators* creates *n* number of trees in your random forest, where *n* is the number you pass in. We passed in 10.

The *.fit()* function allows us to train the model, adjusting weights according to the data values in order to achieve better accuracy. After training, our model is ready to make predictions, which is called by the *.predict()* method.

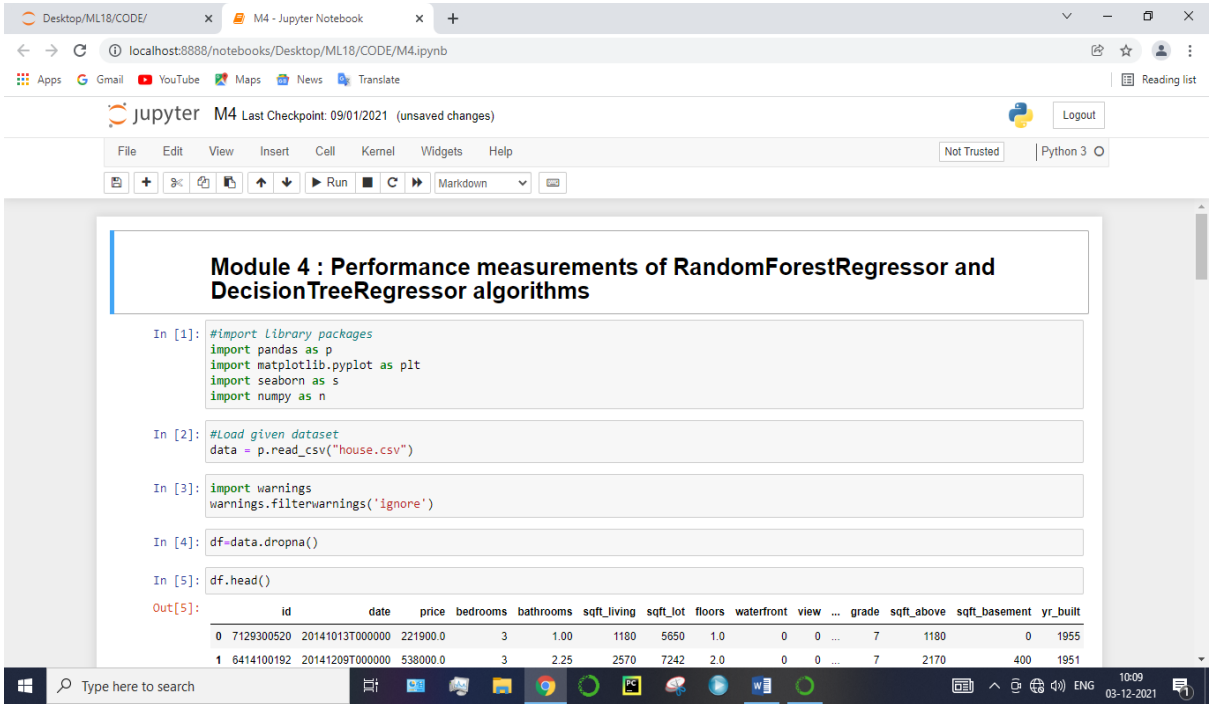
A random forest is a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. The sub-sample size is controlled with the max_samples parameter if Bootstrap = True (default), otherwise the whole dataset is used to build each tree.

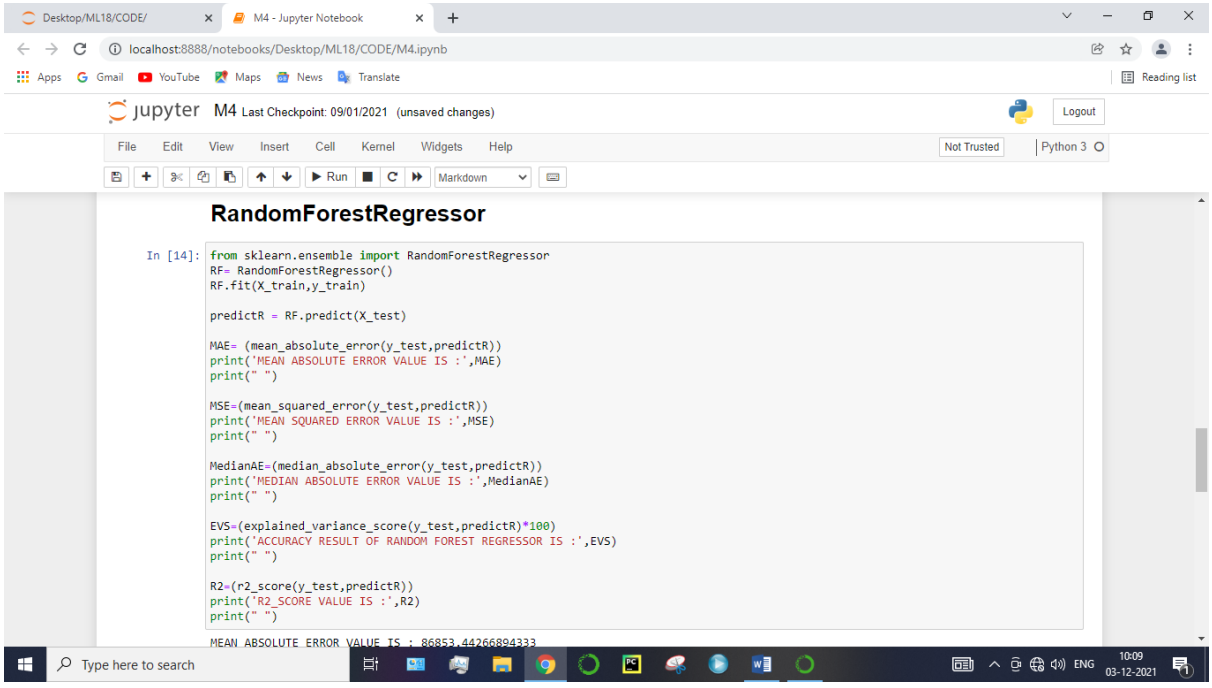
|| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | SJIF: 8.423|| Volume 12, Special Issue 2, April 2023 ||

8th National Conference on Advanced Computing Technologies (NCACT'23)

Organized by

Department of CSE, Velammal Engineering College, Chennnai, India





Decision Tree Regression

Decision Tree - Regression. Decision tree builds regression or classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes.

Decision tree builds regression or classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed.

|| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | SJIF: 8.423|| Volume 12, Special Issue 2, April 2023 ||

8th National Conference on Advanced Computing Technologies (NCACT'23)

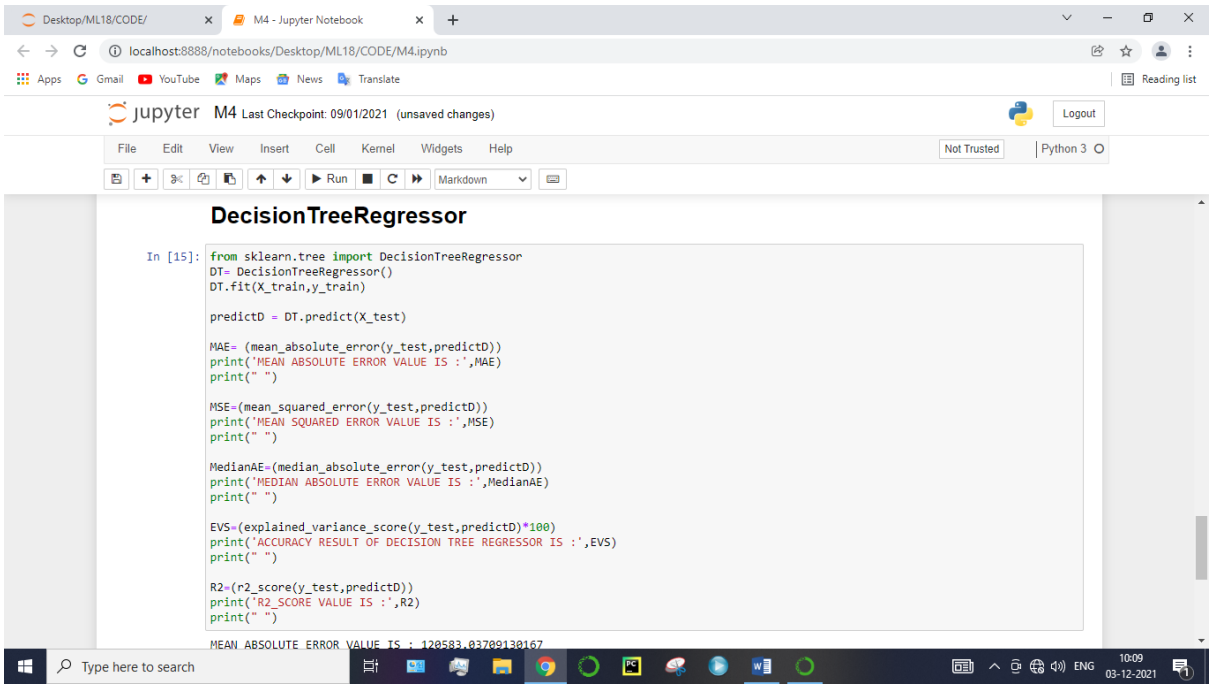
Organized by

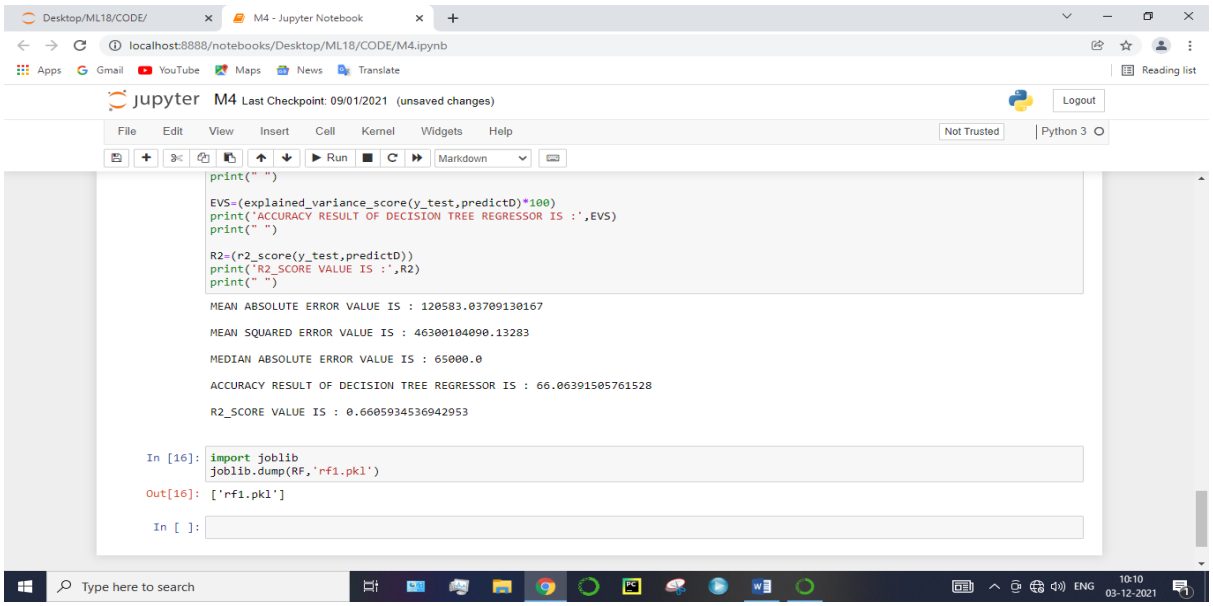
Department of CSE, Velammal Engineering College, Chennnai, India

The final result is a tree with decision nodes and leaf nodes. A decision node (e.g., Outlook) has two or more branches (e.g., Sunny, Overcast and Rainy), each representing values for the attribute tested. Leaf node (e.g., Hours Played) represents a decision on the numerical target. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data.

Decision Tree is one of the most commonly used, practical approaches for supervised learning. It can be used to solve both Regression and Classification tasks with the latter being put more into practical application. It is a tree-structured classifier with three types of nodes.

Decision trees where the target variable can take continuous values (typically real numbers) are called regression trees. Decision trees are among the most popular machine learning algorithms given their intelligibility and simplicity.





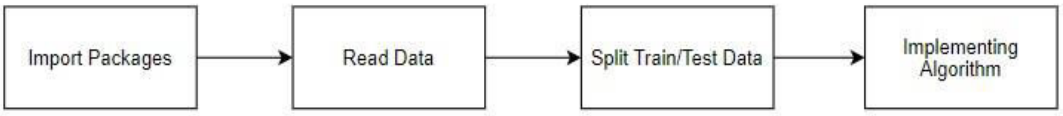
|| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | SJIF: 8.423|| Volume 12, Special Issue 2, April 2023 ||

8th National Conference on Advanced Computing Technologies (NCACT'23)

Organized by

Department of CSE, Velammal Engineering College, Chennnai, India

MODULE DIAGRAM:



GIVEN INPUT EXPECTED OUTPUT

Input : Data

Output : Getting Accuracy

II. IV. EXPERIMENTAL RESULTS

|| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | SJIF: 8.423|| Volume 12, Special Issue 2, April 2023 ||

8th National Conference on Advanced Computing Technologies (NCACT'23)

Organized by

Department of CSE, Velammal Engineering College, Chennnai, India

III. IV. V. CONCLUSION

The analytical process started from data cleaning and processing, missing value, exploratory analysis and finally model building and evaluation. The best accuracy on public test set is higher accuracy score is will be find out. This application can help to find the House Price. We have overseen out how to set up a model that gives clients for an original best methodology with look at future dwelling esteem expectations. Straight previous suggest works bring been used inside our model, something appreciate that that future worth expectations will have an inclination towards even more reasonable qualities. We composed an approach with use in basically the same manner as significantly data as time licenses for our expectation framework, by embracing those thoughts from guaranteeing inclination supporting. These consolidate redesigns we didn't choose as a result of compelled length of the time. A genuine concern for the expectation structure may be the stacking time frame. Additionally, our informational collection takes more than one day ought to get ready. As gone against playing out the calculations consecutively, we may use different processors and equal the calculations in question, which may potentially diminish the planning time Furthermore expectation period. Incorporate even more functionalities under the model, we can give decisions for customer with select a region on the other hand district should deliver rather than entering in the rundown.

REFERENCES

1. John M. Clapp and Carmelo Giaccotto .” Evaluating House Price Forecasts” - 2002
2. Lynn Wu, Erik Brynjolfsson .” How Google Searches Foreshadow Housing Prices and Sales”- 2009
3. Vasilios Plakandaras, Rangan Gupta, Periklis Gogas and Theophilos Papadimitriou , “Forecasting the U.S. Real House Price Index” - 2013
4. Koen de Koning, Tatiana Filatova, Okmyung Bin,” Improved Methods for Predicting Property Prices in Hazard Prone Dynamic Markets” - 2016
5. Gaikwad Purva Chandrakant, Ganjave Pratiksha Namdev, Gorade Pooja Subhash, S. S. Gore, “Implementation of House Price Prediction Model Using Image Processing and Machine Learning” -2019

INNO SPACE
SJIF Scientific Journal Impact Factor

SJIF: 8.423

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INTERNATIONAL CENTRE

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details