USER CAR PREDICTION

(EGR598 COURSE PROJECT)

ABSTRACT

This project's primary goal is to forecast used car prices based on each model's unique attributes. It enables you to decide whether a used car is worth the asking price provided by different online used car sites. Additionally, it is helpful to those who intend to buy or sell an automobile. Manufacturers set the prices for new cars on the market, with the government incurring additional expenses in the form of taxes. As a result, customers can feel confident in purchasing new cars because their investment is reasonable. However, as new car prices rise and consumers are unable to buy new cars due to a lack of funds, used car sales are increasing globally.Our project effectively determines the value of a car based on various characteristics. There are websites that offer this service, but their forecasting methods may not be optimal. Various models and systems also allow us to predict the actual market value performance of used vehicles. When buying or selling, it is important to know the actual market value.

INDEX

INTRODUCTION

Because there are so many variables that affect how much a used automobile costs on the market, figuring out the list price of one is a challenging undertaking. The goal of this project is to create a machine learning model that can precisely forecast a used car's pricing based on its features so that consumers can make an informed decision. The prices you see when searching his website for used cars to buy or sell are not precise enough. Sometimes the asking price is too high, and other times it's too low. We are left unsure of whether to buy or sell a car at that price as a result. The goal of the used car industry is to profit from customers.This covers commissions and extra revenue generated from customers. It can be challenging to determine whether a used car is worth the asking price when perusing online classifieds. The mileage, fuel type, number of owners, year, and other elements can all have an impact on a car's genuine value. A seller's difficulty is how to appropriately price a secondhand car. Create a model to forecast used car prices using machine learning techniques based on the data you already have. There are millions of used cars flooding the market. You may produce precise used car prices using the data on your own platform.

1. MOTIVATION

During lockdown lot of people have faced the transportation problem due to lack of safety (COVID-19) in public transportation. Due to this reason, they thought of buying own cars. As, some people can't afford to buy new cars. So, they are shifting to second hand cars. To get the estimation price of the second-hand car, we thought to develop a project using Machine Learning Algorithms.

2. PURPOSE

It can be challenging to determine whether a used car is worth the asking price when perusing online classifieds. The actual value of an automobile can vary depending on a number of factors, including mileage, make, model, and year of production. From a seller's perspective, it might be difficult to determine the right price for secondhand cars [2–3]. Our "Used Automobile Price Prediction System" is a model that predicts used car prices based on a variety of criteria using machine learning algorithms and existing data.

3. SCOPE

We'll demonstrate how to develop a model that can forecast automobile prices. Use the machine learning workflow you previously learnt to anticipate a car's market value based on its characteristics. Information on various autos is included in the dataset we use. This project's primary goal is to forecast used automobile prices based on a variety of factors. is to predict used car prices based on various characteristics.

SYSTEM ANALYSIS

1. LITERATURE SURVEY

The project is to use machine learning to predict the cost of used cars. In this research, we look into the use of supervised machine learning methods to forecast used automobile prices in Mauritius. The forecasts are supported by historical information gathered from daily publications. The predictions were made using a variety of methodologies, including multiple linear regression analysis, k-nearest neighbors, naive bayes, and decision trees.

MACHINE LEARNING:

The research that develops computer algorithms that transform data into intelligent actions is called "machine learning." It is made with interfaces for statistics, data science, and computer science. It is considered part of artificial intelligence. Machine learning algorithms build models based on sample data, called training data, to make predictions and decisions without being explicitly programmed to do so. machine learning algorithms are used in various applications such as: Credit card, fraud detection, sentiment detection, sentiment analysis, email filtering, etc. where it is arduous to make traditional algorithms in order to perform the desired tasks assigned.

METHODOLGY

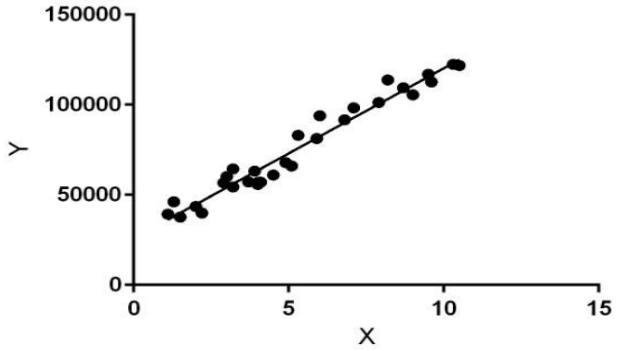
There are two main stages to the system:

Training Phase: Using the dataset's data, the system is taught to fit a model (a line or a curve) based on the proper algorithm.

Phase of testing: The system is given inputs and its functionality is checked. Precision is examined. Consequently, the data utilized to develop or validate the model should be sound. The system must employ the appropriate algorithms to complete two distinct jobs because it is intended to identify and forecast used automobile prices. Before deciding on a particular method for further use, we compared the accuracy of many algorithms. I picked the best candidate for the job.

2. EXISTING SYSTEM

- **LINEAR REGRESSION:** In the context of machine learning, linear regression is essentially an algorithm that relies heavily on supervised learning. Activate a regression task. Based on independent variables, regression models seek to predictors. Finding connections between variables and predictions is its core application. Regression models vary according to the number of independent variables used and the kind of relationship between dependent and independent variables that is taken into account.



Predicting the value of a dependent variable (y) based on a designated independent variable is the task carried out by linear regression (x). By using this method, it is possible to determine the linear relationship between the input (x) and the result (y). Therefore, linear regression is so named. In the graphic above, X represents an individual's job history and Y represents their wage. The regression line offers the model's best fit.

Hypothesis function for Linear Regression:

$$y = \theta_1 + \theta_2.x$$

While training the model we are given:

x: input training data (univariate – one input variable(parameter))

y: labels to data (supervised learning)

When training the model – it fits the best line to predict the value of y for a given value of x. The model gets the best regression fit line by finding the best θ_1 and θ_2 values.

θ_1: intercept

θ_2: coefficient of x

Once we find the best θ_1 and θ_2 values, we get the best fit line. So, when we are finally using our model for prediction, it will predict the value of y for the input value of x.

COST FUNCTION (J):

The model seeks to forecast the y-values such that the error difference between the predicted and true values is as small as possible by obtaining a best-fit regression line. To obtain the optimal values that minimize the difference between the predicted y-value (pred) and the true y-value, it is crucial to change the 1 and 2 values (y).

The Root Mean Squared Error (RMSE) between the predicted y value (pred) and the true y value is the cost function(J) of linear regression (y).

$$minimize \frac{1}{n} \sum_{i=1}^{n} (pred_i - y_i)^2$$

$$J = \frac{1}{n} \sum_{i=1}^{n} (pred_i - y_i)^2$$

DECISION TREE: Although they can be used to solve classification and regression problems, decision trees are a supervised learning technique that work best for classification problems. It is a tree-structured classifier with internal nodes denoting data set characteristics, branches denoting decision rules, and each leaf node denoting a classification outcome..

1. A decision node and a leaf node are the two nodes in a decision tree. While leaf nodes are the outcomes of decisions and have no more branches, decision nodes are used to create decisions and have numerous branches.

2. Based on the features of a specific data collection, a choice or test is made.

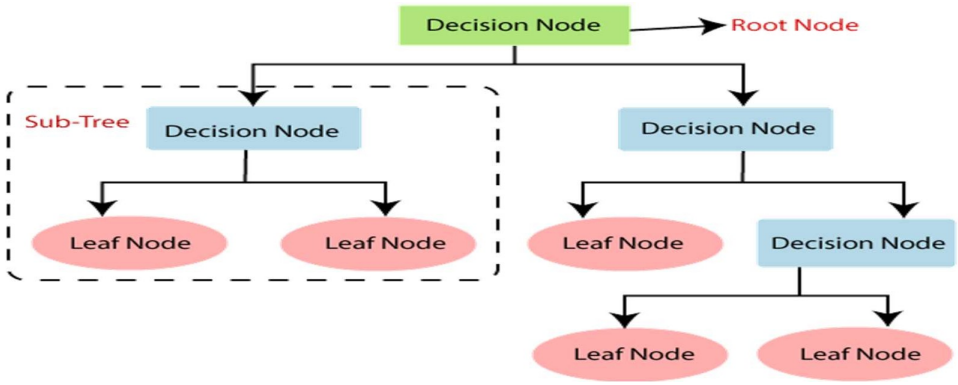
3. A visual representation of all potential answers to a choice or problem based on the conditions provided.

4. Because it grows from a root node to additional branches to create a tree-like structure, it is known as a decision tree. The CART algorithm, which stands for Classification and Regression Tree Algorithm, is used to construct the tree.

A decision tree merely poses a question and divides the tree into subtrees according to the response (Yes/No).

Decision trees are typically designed to mimic how people think when making decisions, making them simple to comprehend.

Because the decision tree displays a tree-like structure, the logic behind it is simple to comprehend.



DECISION TREE TERMINOLOGIES:

1. **Root node:** The decision tree's root node serves as its origin. In addition to being split into two or more homogeneous sets, it represents the complete data collection.
2. **Leaf node:** After receiving a leaf node, the tree cannot be further divided because a leaf node is the ultimate output node.
3. **Split:** A decision node or root node is split into sub-nodes in accordance with predetermined criteria.
4. A branch or subtree is a tree created by slicing another tree.
6. **Pruning**: Removing undesirable limbs from a tree is known as pruning.
7. **Parent/child nodes:** In a tree, the root node is referred to as the parent node, and all other nodes are referred to as children.

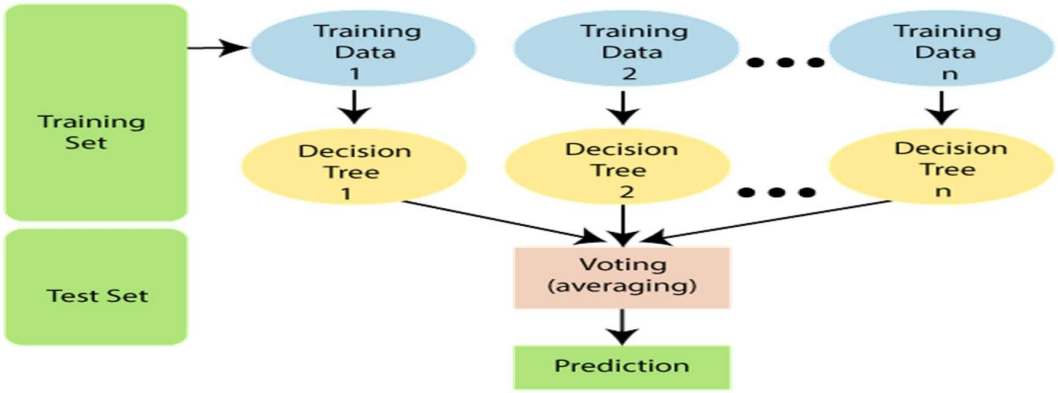
Step 1: First, begin the tree with the root node, which should include the entire dataset.
Step 2: To identify the top attributes in the data set, use the Attribute Selection Measure. Step 3: Divide S into groups of elements that have the best values for each property.
Create a decision tree node with the best attributes in step four.
Step 5: Recursively construct a fresh decision tree using the segment of the Step 3 dataset. Continue doing this until the nodes cannot be further categorized and the last node can only be referred to as a leaf node.

DISADVANTA GES

1. The model is time-consuming.
2. More difficult to compute than other algorithms.
3. The Random Forest Algorithm might be able to address the overfitting issue.

3. PROPOSED SYSTEM

RANDOM FOREST MODEL: A random forest is a type of classifier that takes the average of a group of decision trees over various subsets of a given dataset to increase the dataset's predictive accuracy.



ADVANTAGES:

1. It lessens decision trees' overfitting and enhances accuracy.
2. Adaptable to situations involving regression and classification.
3. is effective with both continuous and categorical values.
Automate data missing values.
5. To find a solution, it is necessary to compile various decision trees.

4. HARDWARE REQUIREMENTS

- **Operating system :** The software required to manage and run computing devices, such as smartphones, tablets, laptops, supercomputers, web servers, vehicles, network towers, and smart watches, is known as an operating system. This graphical user interface (GUI) serves as a conduit between the user and the technical components of the computer.

- Processor (Dual core 2.4 GHz): The integrated electronic circuit known as a processor is what executes the calculations necessary for a computer to function. The operating system passes fundamental instructions to the processor, which carries out logic, input/output (I/O), and other operations (OS). On processor operations, the majority of other processes rely.

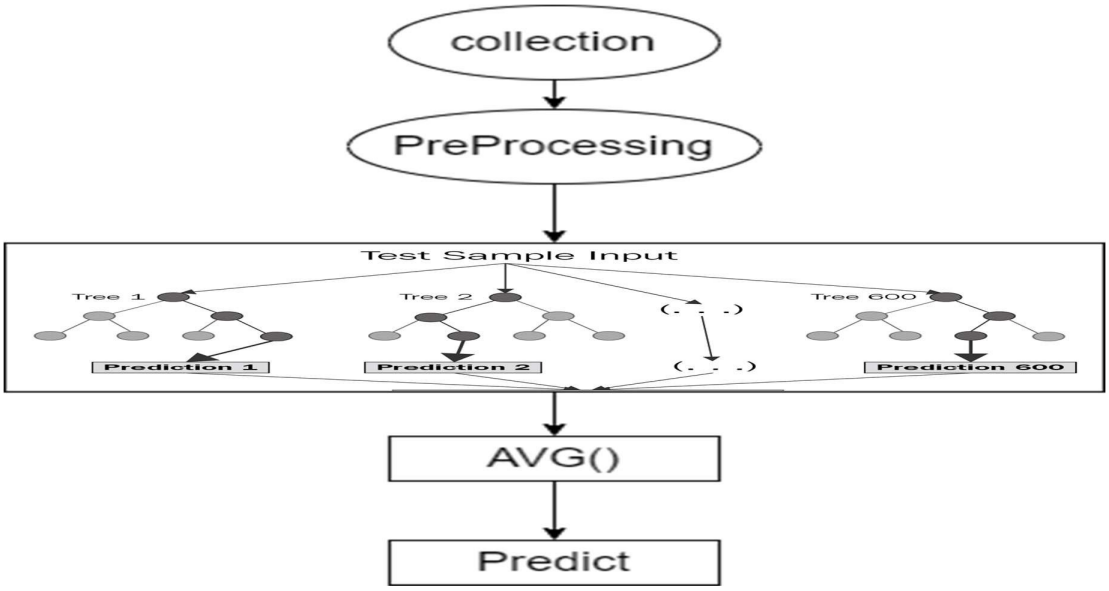
- **RAM:** Your computer's short-term memory, known as Random Access Memory (RAM), is utilized to manage all the open programs and tasks. Without RAM, no application, file, game, or stream will function.

5. SOFTWARE REQUIREMENTS

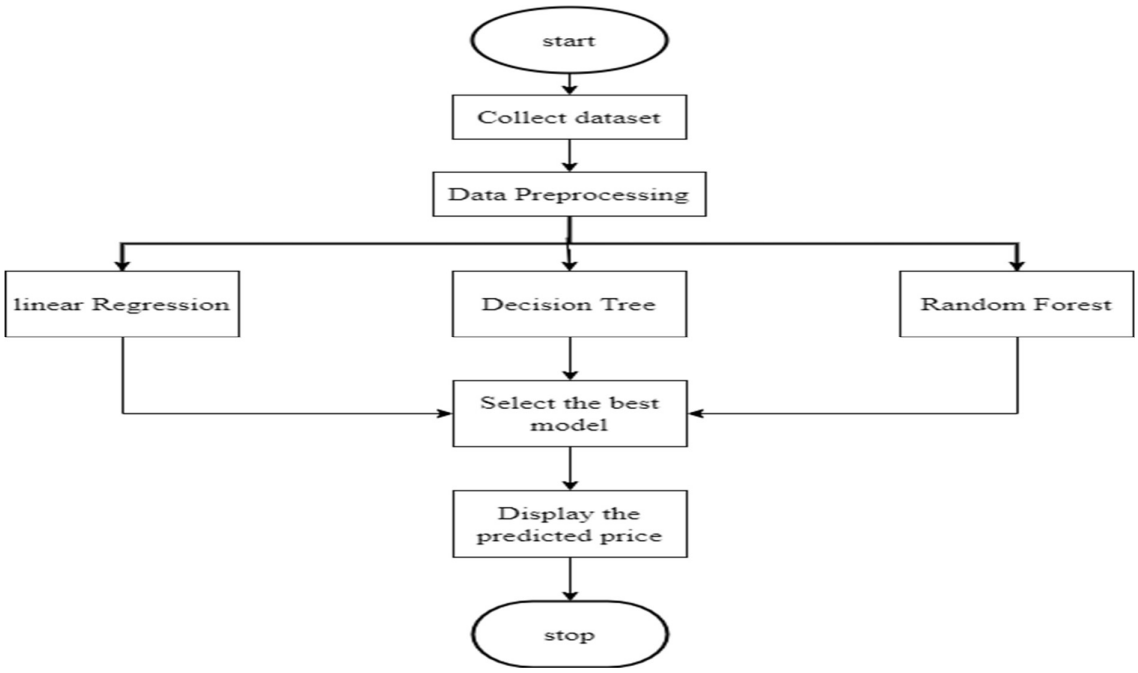
- **Google colab**: Colaboratory, also known as Colab, is a creation of Google Research. Colab excels at machine learning, data analysis, and education because it lets anyone write and execute arbitrary Python code from the browser.
- **Python:** Outstanding library ecosystem One of the key causes for Python's popularity as the most used programming language for AI is the variety of libraries available. Python libraries make it possible to access, modify, and convert data, which is necessary for machine learning (ML) applications.

SYSTEM DESIGN

1. ARCHITECTURE



DATAFLOW DIAGRAM



2. RANDOM FOREST ALGORITHM:

One of the most potent machine learning algorithms, the Random Forest Algorithm is built on the idea of decision tree algorithms. The random forest algorithm builds a structure of decision trees in the form of a forest. The accuracy of recognition increases as the number of trees increases. The Bootstrap approach is used to create trees. Bootstrapping builds a single tree with permutations by randomly selecting components and samples from the data source. The random forest method chooses the most effective decision tree technique for classification from among the randomly chosen pieces. The Gini index and information mining techniques are also used by the random forest algorithm to determine the optimum partitioner. This procedure is repeated until n trees are produced using Random Forest.

Random Forest Algorithm Features

1. Compared to decision tree algorithms, it is more accurate.
2. Offers a practical means of handling missing data.
3. is able to forecast outcomes sensibly without over-adjusting the settings.
4. Fixes the overfitting issue with decision trees.
5. A subset of the elements at each node split point in the random forest tree is chosen at random.

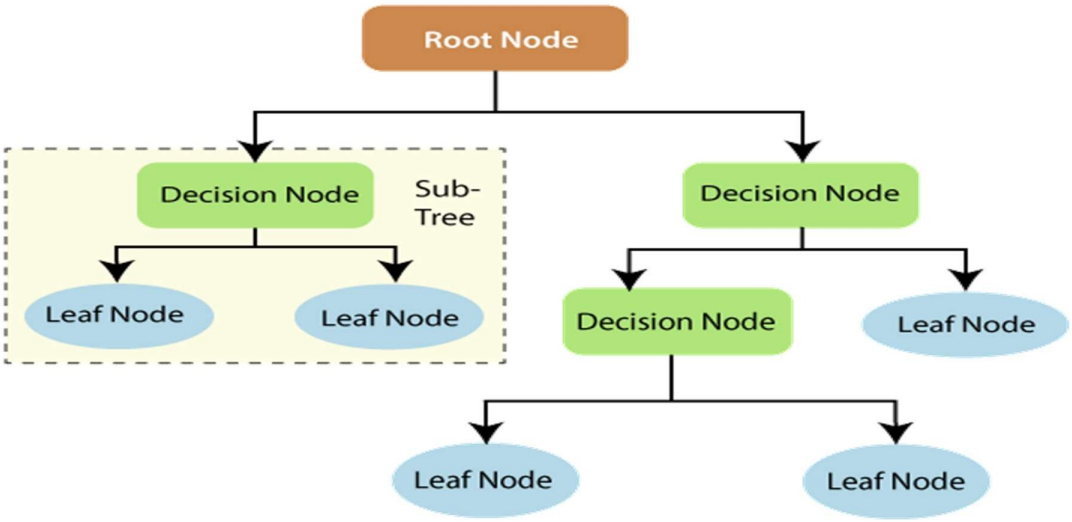
How random forest algorithm works

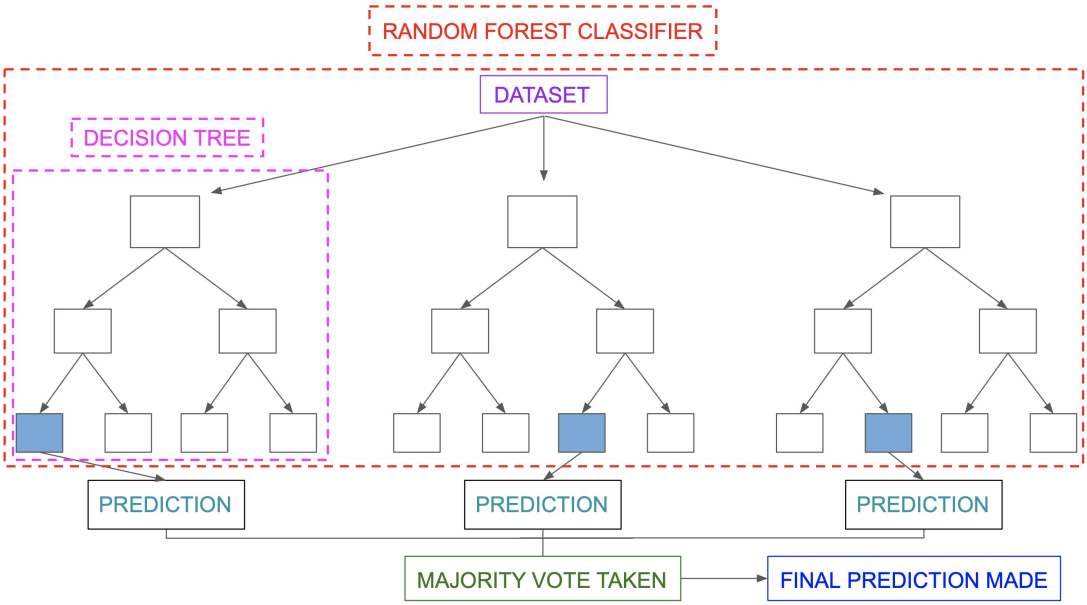
Understanding decision trees

Random forest algorithms' fundamental building pieces are decision trees. Using a tree structure, a decision tree is a decision assistance method. You can better comprehend the operation of a random forest algorithm by reviewing decision trees.

Decision nodes, leaf nodes, and root nodes are the three parts of a decision tree. The training data set is divided into branches via the decision tree algorithm, and those branches are further divided into branches. This pattern keeps going until it encounters a leaf node. Leaf nodes cannot be separated from one another.

The attributes that are utilized to predict outcomes are represented by nodes in a decision tree. Links to the leaves are provided by decision nodes. The three different sorts of nodes in a decision tree are depicted in the diagram below.





3. PSEUDO CODE

Steps

1.Pick "k" elements at random from the total of "m" elements. When and where

2.Use the ideal split point to calculate node "d" between elements "k."

3.Use the optimal split to divide the node into child nodes.

4.Repeat steps 1 through 3 until there are "l" nodes.

5.To generate "n" trees, repeat steps 1 through 4 "n" times to build the forest.

Algorithm 1 Random Forest

Precondition: A training set $S := (x_1, y_1), \ldots, (x_n, y_n)$, features F, and number of trees in forest B.

```
function RandomForest(S, F)
    H ← ∅
    for i ∈ 1, ..., B do
        S^(i) ← A bootstrap sample from S
        h_i ← RandomizedTreeLearn(S^(i), F)
        H ← H ∪ {h_i}
    end for
    return H
end function
function RandomizedTreeLearn(S, F)
    At each node:
        f ← very small subset of F
        Split on best feature in f
    return The learned tree
end function
```

SYSTEM IMPLEMENTATION

1. PACKAGES

In our project we are using the following libraries in python.

- Numpy
- Pandas
- Seaborn
- Matplotlib
- Sklearn

2. Modules of the project:

DATASET OVERVIEW:- The used automobiles offered on Cardekho.com are described in this listing. Each of its nine columns, such as Car Name, which provides details about automobile manufacturers, contains information about a certain attribute. the year of the brand-new automobile you bought. Sale price The cost of buying the car. Future pricing projections will use this label as their primary reference. miles of a vehicle Fuel This method retrieves the vehicle's fuel type (CNG, petrol, diesel, etc.). Whether the seller is an individual or a seller is indicated by seller type. Communication is discloses if the vehicle is automatic or manual. the vehicle's former owner's owner number. The car's current catalog price is given by the term current price.

- ❑ Step 1: Acquiring dataset and importing all the essential library

- ❑ Step 2: Data Pre Processing

• Data preprocessing is a necessary task to clean the data and fit it for the machine learning model, which also increases the accuracy and efficiency of the machine learning model.It involves below steps:

- • Reading the dataset using pd.read_csv
- • datasets should be imported
- • Libraries should be imported
- • Feature scaling

- ❑ Step 3: Data Visualization

• This is the most important step of the data science life cycle, here we understand the behavior of the data and try to make some meaningful insight out of it.

- ❑ Step 4: Pre Modeling
  - • Split the dataset into dependent and independent variables.
  - • split the dataset into training and testing sets.

- ❑ Step 6: Choosing best fit model for our dataset
  - • Use the concept of cross validation score and make use of it to choose the best fit model for the coding.

DATA PRE-PROCESSING

We made a histogram of the data in order to better comprehend it. Due to the high price sensitivity of used cars, we discovered that the data set contains a significant number of outliers. The most recent model years with low mileage typically sell for a premium, although the data was rather erratic. The accident's circumstances and conditions may have a substantial impact on the car's value. We trimmed the data set to 3 standard deviations from the mean to weed out outliers since we lacked information on the history and condition of the vehicles. manufacture, model, and state were combined into a single hot vector.

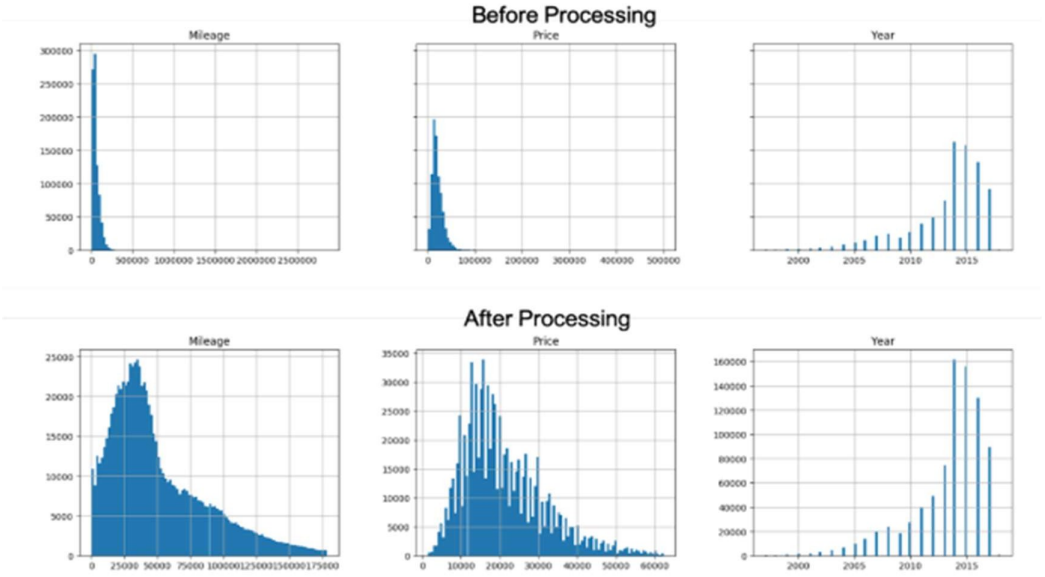


Fig 1: (Top) Histogram for raw data. (Bottom) Histogram after pruning.

Because they were specific to each vehicle and did not improve the training process, some characteristics, such the VIN number, were lost during training. Additionally, we discovered that using specific machine learning frameworks, certain categorical features share similar values. When employing frameworks like XGBoost that demand distinctive function names, this creates issues (across functions). For instance, the string "Genesis" cannot appear in both the "Make" and "Model" fields at the same time. In order to interact with these frameworks, certain common function names have been eliminated and renamed.

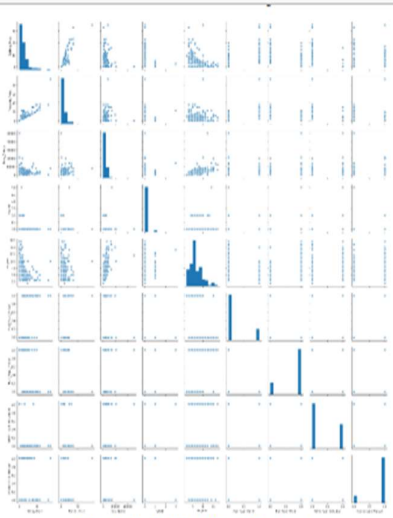


Fig-1: Pair Plot Final Dataset

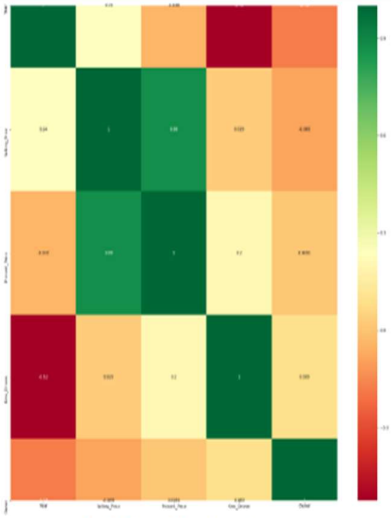


Fig-2: Heat Map Final Dataset

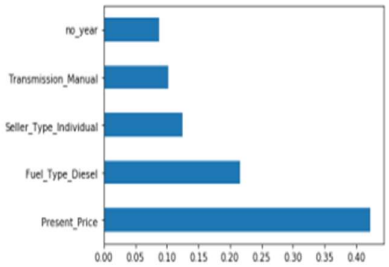


Fig-3: Feature in Bar Graph

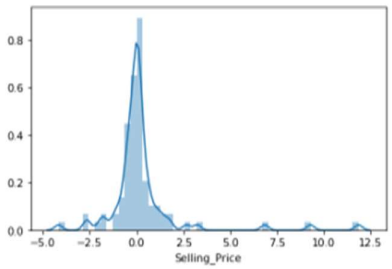


Fig-4: Prediction of the Dataset

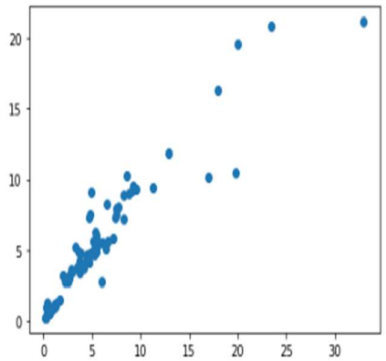


Fig-5: Evaluation of the Model

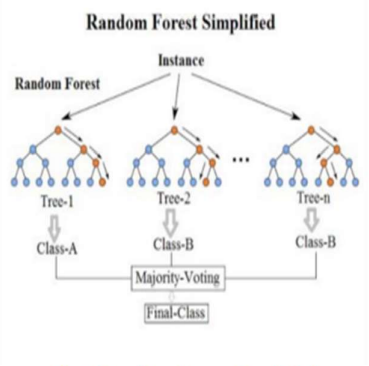


Fig-6: Random Forest Simplified

SYSTEM EVALUATION

1. EVALUATION METRICS

Mean Absolute Error (MAE): The absolute average difference between the observed and predicted data is measured, but significant prediction errors are not penalized.

$$\mathbf{MAE} = \frac{1}{n}\sum_{i=1}^{n} |\boldsymbol{x}_i - \boldsymbol{x}|$$

Mean Square Error (MSE): The average squared distance between the actual data and the

predicted data is measured here.

MSE formula = $(1/n) * \Sigma(\text{actual} - \text{forecast})^2$

Root mean square error : One of the most used metrics for measuring forecast quality is the root mean square error or root mean square deviation. Due to the fact that RMSE requires and utilizes actual measurements at each anticipated data point, it is frequently utilized in supervised educational applications.
The formula is:

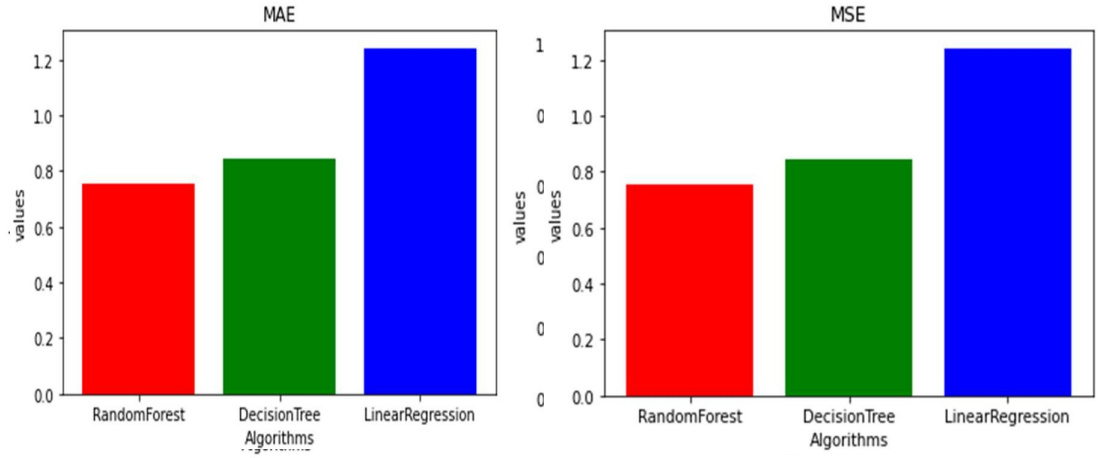
$$RMSE = \sqrt{\overline{(f-o)^2}}$$

- f = forecasts
- o = observed values

A crucial indicator for assessing the effectiveness of a regression-based machine learning model is the R2 score. The coefficient of determination, commonly known as R squared, is pronounced as R sq. It measures how much variance in the forecasts that the data set can account for.

The formula is:

$$R^2_{adj} = 1 - \left[\frac{\left(1 - R^2\right)(n-1)}{n-k-1} \right]$$



2. PLOTS

We want our R2 score to be maximum and other errors to be minimum for betterresults.

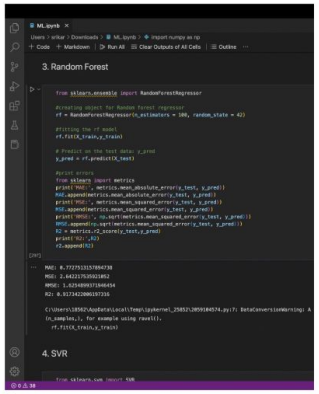
Random forest regressor is giving better results. Therefore, we will hypertune thismodel and then fit, predict.

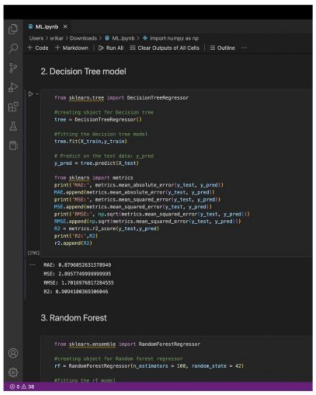
3. DISCUSSION:

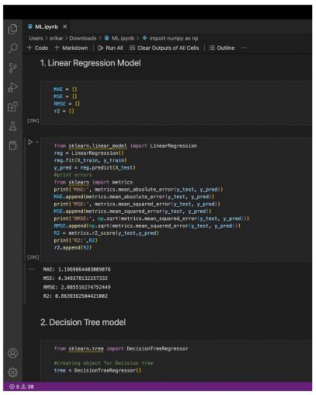
Three algorithms were utilized in our study to choose the best model. The best-fitting model was determined using a number of criteria, including mean absolute error, mean squared error, mean squared error, and R2 score. We must increase the R2 score and reduce the other mistakes among these four metrics in order to achieve better results. As a result, the Random Forest Regressor outperforms the Linear Regression and Decision Tree algorithms in terms of performance.

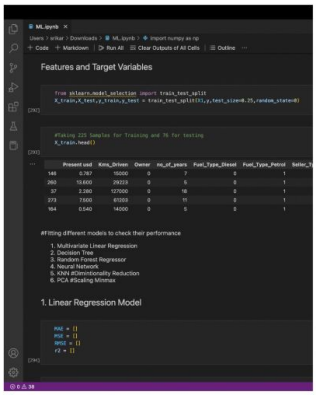
7. SCREENSHOTS

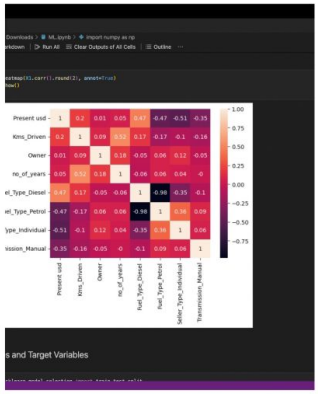
7.1 Code Screenshots

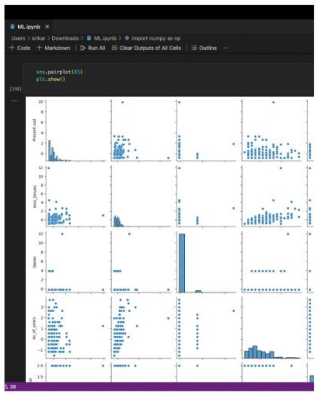


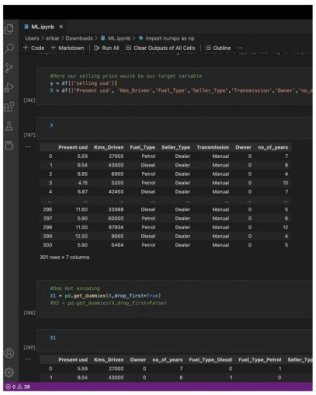


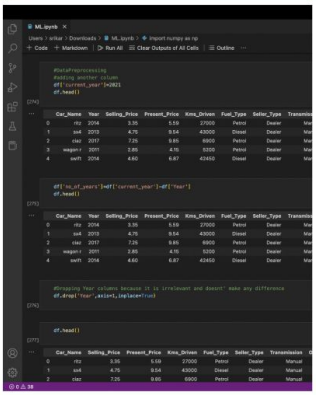


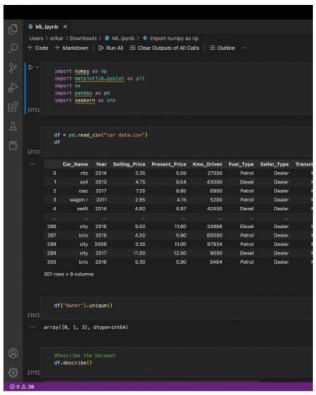


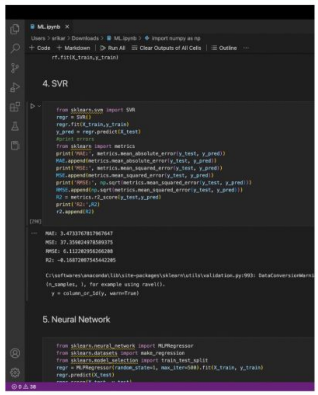












CONCLUSION

As new car prices have soared and consumers can no longer afford them, used car sales are rising globally. Because of this, there is a critical need for a system that can accurately estimate the worth of a used car from various attributes. With the proposed (random forest) approach, it is possible to anticipate used car prices with greater accuracy.

FUTURE SCOPE

Future connections between this machine learning model and several websites that offer real-time data for price prediction are possible. Large historical car price data can be added as well, which will assist the machine learning model's accuracy. As a user interface to communicate with the user, we can construct an Android application.

REFERENCES

1. https://www.kaggle.com/jpayne/852k-used-car-listings
2. N. Monburinon, P. Chertchom, T. Kaewkiriya, S. Rungpheung, S. Buya and P. Boonpou, "Prediction of prices for used car by using different models," 2018 5th International Conference on Business and Industrial Research (ICBIR), Bangkok, 2018, pp. 115-119.
3. https://scikit-learn.org/stable/modules/classes.html: Scikit-learn: Machine Learning in Python,
   Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.
4. https://github.com/topics/car-price-prediction
5. DATASET: https://www.kaggle.com/nehalbirla/vehicle-dataset-from-cardekho?select=car+data.csv
6. Sameerchand Pudaruth, “Predicting the Price of Used Cars using Machine LearningTechniques” ;(IJICT 2014).
7. Ning sun, Hongxi Bai, Yuxia Geng, Huizhu Shi, “Price Evaluation Model in Second Hand Car
   System Based on BP Neural Network Theory”; (Hohai University Changzhou, China).