Mini Project Report

A report submitted in partial fulfilment for the requirements for the award of degree of

BACHELOR OF TECHNOLOGY

In

Information Technology

By

Pranjul Shukla (2007340130040)

Aditya Kumar (2007340130005)

Prashant Sharma (2007340130042)

To

DR. A.P.J. ABDUL KALAM TECHNICAL UNIVERSITY, LUCKNOW

Under the Guidance of

Mr. Shantanu Shukla

Mr. Sandip Kr. Singh

Mr. Abhishek Kr. Yadav

DEPARTMENT OF INFORMATION TECHNOLOGY

RAJKIYA ENGINEERING COLLEGE

ATARRA, BANDA-210201

DEPARTMENT OF INFORMATION TECHNOLOGY

RAJKIYA ENGINEERING COLLEGE

ATARRA, BANDA-210201

This is to certify that the “Mini Project report” entitled “Car Price Prediction System ” submitted by Prashant Sharma, Pranjul Shukla and Aditya Kumar submitted during 2022 – 2023 academic year, in partial fulfilment of the requirement for the award of the degree of Bachelor of Technology in Information Technology in the Department of Information Technology at Rajkiya Engineering College, Banda, under Dr. A.P.J. Abdul Kalam Technical University, Lucknow.

This report is an authentic record of candidates own work carried out under our supervision. The matter embodied in this Mini Project report is original and has not been submitted for the award of any other degree.

College Mini Project Coordinator	Dr. Vibhash Yadav
Mr. Shantanu Shukla	(Associate Professor)
	Head of Department
	Information Technology

ACKNOWLEDGEMENT

It is indeed a great pleasure to express our sincere thanks to our august supervisor **Mr. Shantanu Shukla Sir, Mr Abhishek Kr. Yadav Sir and Mr. Sandip Kr. Singh Sir**, Department of Information Technology of Rajkiya Engineering College, Banda for their continuous support in this Mini Project. They were always there to listen and to give advice. There were always there to meet and talk about our ideas, to proofread and mark up our mini project, and to ask us good questions to help me to think through our problems. Without their encouragement and constant guidance, I could not have finished this project.

I am highly indebted to Director **Prof. Sheo Prasad Shukla Sir** for the facilities provided to accomplish this Mini Project.

I would like to thank my Head of the Department **Dr. Vibhash Yadav Sir** for his constructive criticism throughout my Mini Project.

I am thankful to my family whose unfailing love, affection, sincere prayers and best wishes had been a constant source of strength and encouragement.

I am extremely great full to my department staff members and friends who helped me in successful completion of this Mini Project.

Submitted by

Prashant Sharma (2007340130042)

Pranjul Shukla (2007340130040)

Aditya kumar (2007340130005)

ABSTRACT

Determining whether the listed price of a used car is a challenging task, due to the many factors that drive a used vehicle's price on the market. The focus of this project is developing machine learning models that can accurately predict the price of a used car based on its features, in order to make informed purchases. We implement and evaluate various learning methods on a dataset consisting of the sale prices of different makes and models across cities in the India. Our results show that Random Forest model yield the best results. Conventional linear regression also yielded satisfactory results, with the advantage of a significantly lower training time in comparison to the aforementioned method.

The price of a new car in the industry is fixed by the manufacturer with some additional costs incurred by the Indian Government in the form of taxes. So, customers buying a brand-new vehicle may be confident of the money they make investments to be worth. But, due to the increased prices of new cars and the financial incapability of the customers to buy them, used Car sales are on a global increase. Therefore, to find the car price which would be best suited for the buyer in India, we are going to predict its cost with the help of Machine Learning algorithms which are made available by the Python Environment. Our dataset comprises data related to different car brands with a set of parameters (Name, Year, Fuel Type, Transmission, Owner Type, Price). The primary purpose is to design a model for a given dataset and predict the car price with better accuracy.

CONTENTS

S.NO	Chapter Name	Page No.
1	INTRODUCTION	6
2	DATASET	7
3	PROPOSED METHODS	10
3.1	DATASET COLLECTION	10
3.2	DATA PRE-PROCESSING	10
3.3	MISSING VALUE IDENTIFICATION	10
3.4	FEATURE SELECTION	11
3.5	SCALING AND NORMALIZATION	11
3.6	SPLITTING OF DATA	11
3.7	DESIGN AND IMPLEMENTATION OF CLASSIFICATION MODEL	11
3.8	MACHINE LEARNING CLASSIFIER	11
4	MODELING AND ANALYSIS	12
4.1	LINEAR REGRESSION	12
4.2	LASSO REGRESSOR	12
4.3	DECISION TREE	12
4.4	RANDOM FOREST	12
5	MEASURMENT	13
6	RESULTS AND DISCUSSION	15
7	RESULTS AND ANALYSIS	16
8	CONCLUSION	18
9	REFERENCES	19

CHAPTER 1 - INTRODUCTION

Determining whether the listed price of a used car is a challenging task, due to the many factors that drive a used vehicle’s price on the market. The focus of this project is developing machine learning models that can accurately predict the price of a used car based on its features, in order to make informed purchases. We implement and evaluate various learning methods on a dataset consisting of the sale prices of different makes and models . We will compare the performance of various machine learning algorithms like Linear Regression, Lasso Regression, Random Forest Regressor, Decision Tree Regressor and choose the best out of it. Depending on various parameters we will determine the price of the car. Regression Algorithms are used because they provide us with continuous value as an output and not a categorized value because of which it will be possible to predict the actual price a car rather than the price range of a car. User Interface has also been developed which acquires input from any user and displays the Price of a car according to user’s inputs.

The used car market is an ever-rising industry, which has almost doubled its market value in the last few years. The emergence of online portals such as CarDheko, Quikr, Carwale, Cars24, and many others has facilitated the need for both the customer and the seller to be better informed about the trends and patterns that determine the value of the used car in the market. Machine Learning algorithms can be used to predict the retail value of a car, based on a certain set of features. Different websites have different algorithms to generate the retail price of the used cars, and hence there isn't a unified algorithm for determining the price. By training statistical models for predicting the prices, one can easily get a rough estimate of the price without actually entering the details into the desired website.

CHAPTER 2 - Data Set

For this project, we are using the dataset on used car sales from all over the United States, available on Kaggle. It consists of several car related variables and one target variable. The objective of the dataset is to predict Price of the car. The dataset consists of several independent variables and one dependent variable, i.e., the Selling Price. Independent variables include the Present Price , Kms Driven and so on as Shown in Following Table 1:

Table 1 Dataset description

Serial no	**Attribute Names**	**Description**
1	Car_Name	Name of the car
2	Year	Year in which car Purchase
3	Selling_Price	Selling price of the car
4	Present_Price	Present price of the car
5	Kms_Driven	How many Kilometer car run
6	Fuel_Type	Type of fuel car use
7	Seller_Type	Type of the person buy the car
8	Transmission	Transmission type
9	Owner	Number of owner of the car

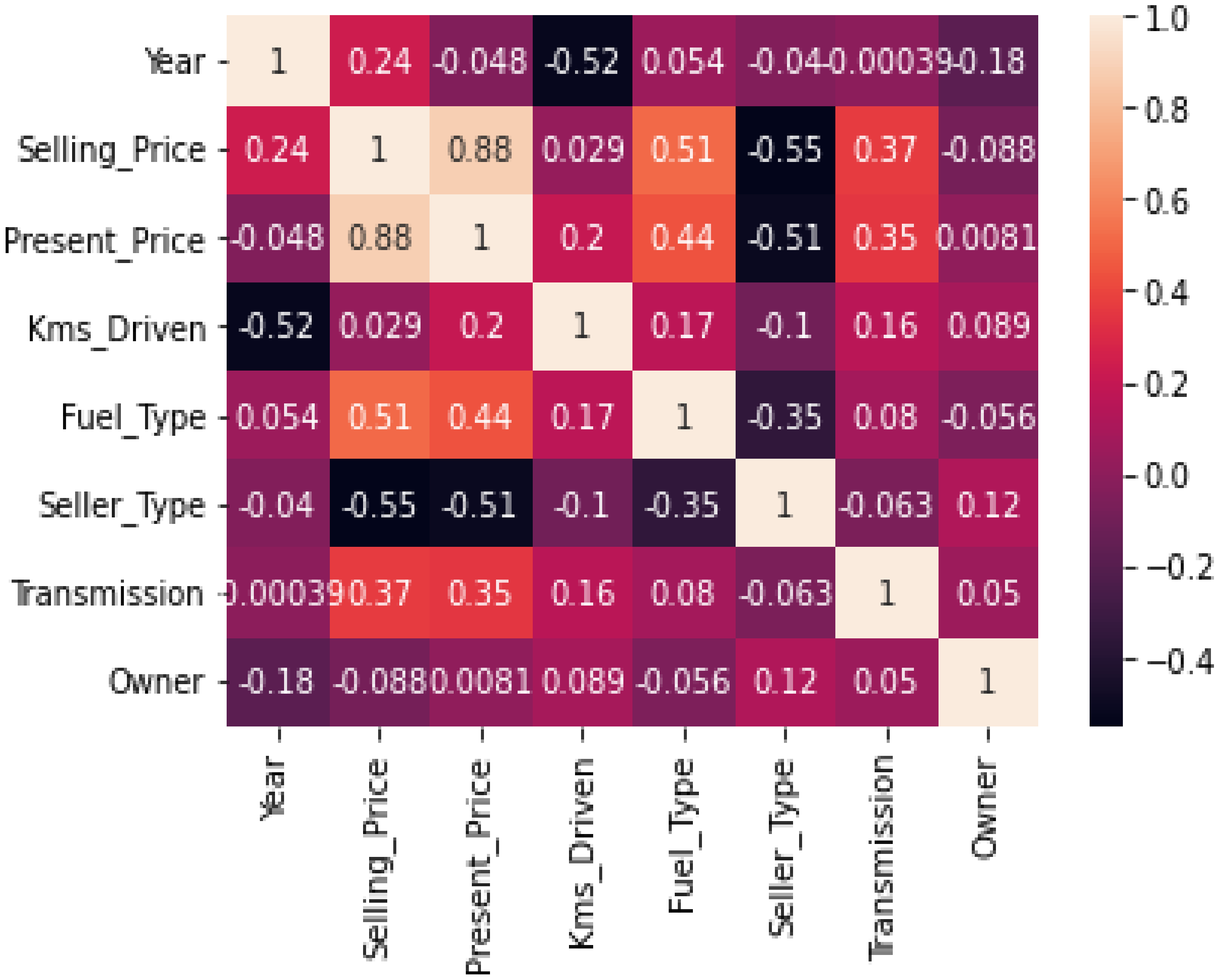
The diabetes data set consists of 310 data points, with 9 features each.

“Selling_Price” is the feature we are going to predict.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 301 entries, 0 to 300
Data columns (total 9 columns):
 #   Column         Non-Null Count  Dtype  
---  ------         --------------  -----  
 0   Car_Name       301 non-null    object 
 1   Year           301 non-null    int64  
 2   Selling_Price  301 non-null    float64
 3   Present_Price  301 non-null    float64
 4   Kms_Driven     301 non-null    int64  
 5   Fuel_Type      301 non-null    object 
 6   Seller_Type    301 non-null    object 
 7   Transmission   301 non-null    object 
 8   Owner          301 non-null    int64  
dtypes: float64(2), int64(3), object(4)
memory usage: 21.3+ KB
```

There is no null values in dataset.

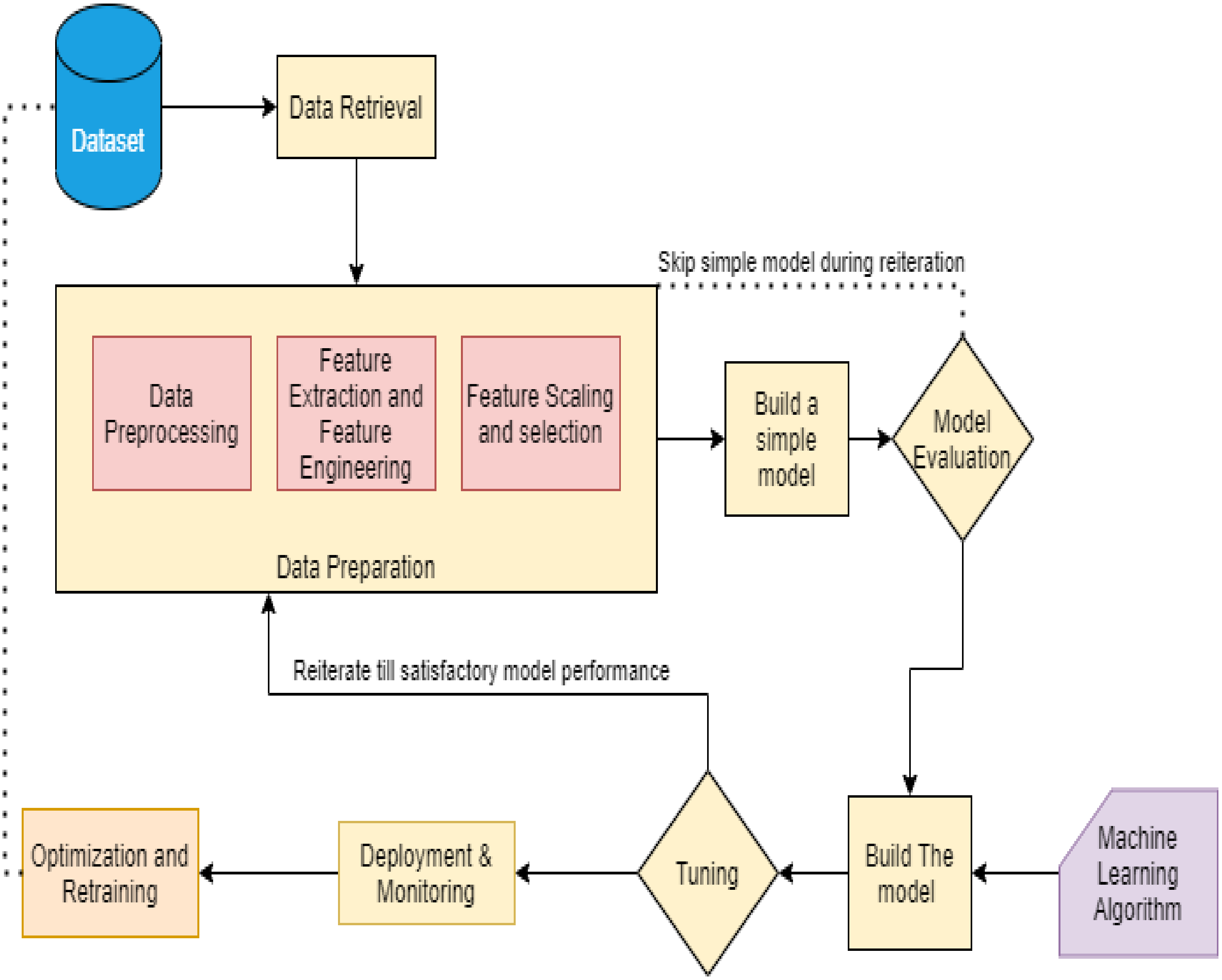
Correlation Matrix:



	Year	Selling_Price	Present_Price	Kms_Driven	Fuel_Type	Seller_Type	Transmission	Owner
Year	1	0.24	-0.048	-0.52	0.054	-0.04	-0.00039	-0.18
Selling_Price	0.24	1	0.88	0.029	0.51	-0.55	0.37	-0.088
Present_Price	-0.048	0.88	1	0.2	0.44	-0.51	0.35	0.0081
Kms_Driven	-0.52	0.029	0.2	1	0.17	-0.1	0.16	0.089
Fuel_Type	0.054	0.51	0.44	0.17	1	-0.35	0.08	-0.056
Seller_Type	-0.04	-0.55	-0.51	-0.1	-0.35	1	-0.063	0.12
Transmission	-0.00039	0.37	0.35	0.16	0.08	-0.063	1	0.05
Owner	-0.18	-0.088	0.0081	0.089	-0.056	0.12	0.05	1

It is easy to see that there is only Present_Price has a very high correlation with our Selling_Price value Whereas Seller_Type has a very low correlation with our Selling_Price value.

Flowchart of the Model


Dataset
Data Retrieval
Skip simple model during reiteration
Data Preprocessing
Feature Extraction and Feature Engineering
Feature Scaling and selection
Data Preparation
Build a simple model
Model Evaluation
Reiterate till satisfactory model performance
Optimization and Retraining
Deployment & Monitoring
Tuning
Build The model
Machine Learning Algorithm

CHAPTER 3 - PROPOSED METHODS

I] **Dataset collection** – It includes data collection and understanding the data to study the hidden patterns and trends which helps to predict and evaluating the results. Dataset carries 310 rows i.e., total number of data and 9 columns i.e., total number of features. Features include Car_Name, Year, Selling_Price, Present_Price, Kms_Driven, Fuel_Type, Seller_Type, Transmission, Owner .

II] **Data Pre-processing:** This phase of model handles inconsistent data in order to get more accurate and precise. This dataset doesn't contain missing values. So, we imputed missing values for few selected attributes Year, Kms_Driven, Owner because these attributes cannot have values zero. Then data was scaled using Standard Scaler. Since there were a smaller number of features and important for prediction so no feature selection was done.

III]**Missing value identification**: Using the Panda library and SK-learn, we check for the missing values in the datasets, shown below. We replaced the missing value with the corresponding mean value if we get.

```
Car_Name        0
Year            0
Selling_Price   0
Present_Price   0
Kms_Driven      0
Fuel_Type       0
Seller_Type     0
Transmission    0
Owner           0
dtype: int64
```

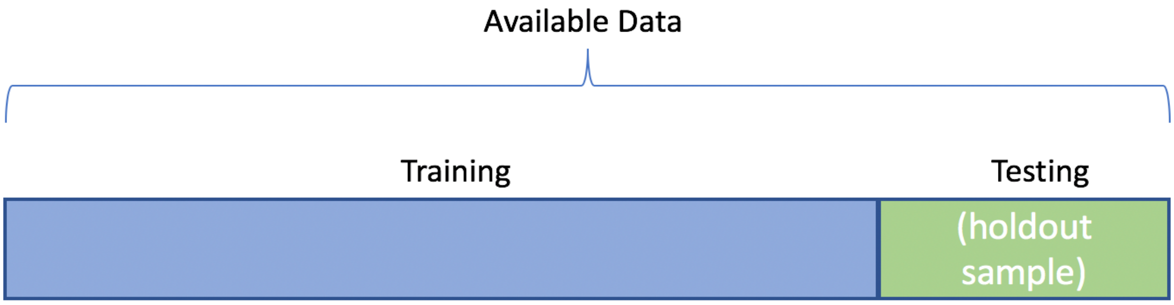
IV] **Feature selection:** Pearson's correlation method is a popular method to find the most relevant attributes/features. The correlation coefficient is calculated in this method, which correlates with the output and input attributes. The coefficient value remains in the range between −1 and 1. The value above 0.5 and below −0.5 indicates a notable correlation, and the zero value means no correlation

Correlation with Selling_Price

```
Present_Price       0.878983
Fuel_Type           0.509467
Transmission        0.367128
Year                0.236141
Kms_Driven          0.029187
Owner              -0.088344
Seller_Type        -0.550724
```

V] Scaling and Normalization: We performed feature scaling by normalizing the data from 0 to 1 range, which boosted the algorithm's calculation speed. Scaling means that you're transforming your data so that it fits within a specific scale, like 0- 100 or 0-1. You want to scale data when you're using methods based on measures of how far apart data points. With these algorithms, a change of "1" in any numeric feature is given the same importance.

VI] Splitting of data: After data cleaning and pre-processing, the dataset becomes ready to train and test. In the train/split method, we split the dataset randomly into the training and testing set. For Training we took 240 sample and for testing we took 70 sample.



VII] Design and implementation of classification model: In this work, comprehensive studies are done by applying different ML Regression techniques like Decision Regressor, Random Forest Regressor, Linear Regression, Lasso Regression.

VIII] Machine learning Regressors: We have developed a model using Machine learning Technique. Used different Regressors techniques to predict car dataset. We have applied Lasso Regressor, Linear Regression, Decision Tree Regressor and Random Forest

Regressor. Machine learning regressor to analyse the performance by finding accuracy of each regressor. All the regressor are implemented using scikit learn libraries in python. The implemented Regression algorithms are described in next section.

CHAPTER 4 - MODELING AND ANALYSIS:

4.1 - Linear Regression: Linear Regression was chosen as the first model due to its simplicity and comparatively small training time. The features, without any feature mapping, were used directly as the feature vectors. No regularization was used since the results clearly showed low variance.

4.2 – Lasso Regressor: Lasso regression is a type of linear regression that uses shrinkage. Shrinkage is where data values are shrunk towards a central point, like the mean. The lasso procedure encourages simple, sparse models (i.e. models with fewer parameters). This particular type of regression is well-suited for models showing high levels of muticollinearity or when you want to automate certain parts of model selection, like variable selection/parameter elimination.

4.3 - Decision Tree: Decision tree regression observes features of an object and trains a model in the structure of a tree to predict data in the future to produce meaningful continuous output. Continuous output means that the output/result is not discrete, i.e., it is not represented just by a discrete, known set of numbers or values.

4.4 - Random Forest: Random Forest is an ensemble learning based regression model. It uses a model called decision tree, specifically as the name suggests, multiple decision trees to generate the ensemble model which collectively produces a prediction. The benefit of this model is that the trees are produced in parallel and are relatively uncorrelated, thus producing good results as each tree is not prone to individual errors of other trees.

CHAPTER 5 – MEASUREMENTS

To find the efficient classifier for diabetes prediction we have applied a performance matrices are confusion matrix and accuracy are discussed as follows:

Confusion matrix: - which provides output matrix with complete description performance of the model.

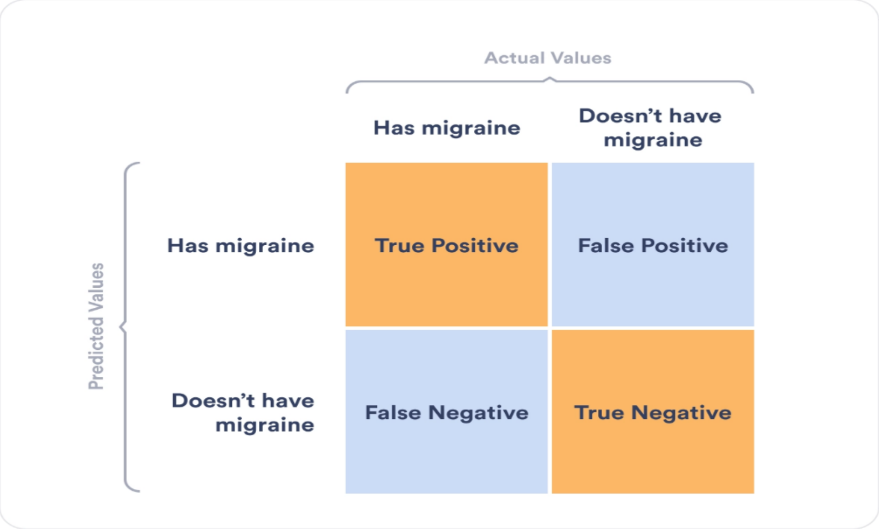
Here,

TP: True positive

FP: False positive

TN: True negative

FN: False negative



Actual Values

Predicted Values	Has migraine	Doesn’t have migraine
Has migraine	True Positive	False Positive
Doesn’t have migraine	False Negative	True Negative

The following performance metrics are used to calculate the presentation of various algorithms.

True positive (TP) – person has disease, and the prediction also has a positive

True negative (TN) – person not having disease and the prediction also has a negative

False positive (FP) – person not having disease but the prediction has a positive

False negative (FN) – person having disease and the prediction also has a positive

TP and TN can be used to calculate accuracy rate and the error rates can be computed using FP and FN values.

True positive rate can be calculated as TP by a total number of persons have disease in reality.

False positive rate can be calculated as FP by a total number of persons do not have disease in reality.

Precision is TP/ total number of person have prediction result is yes.

Accuracy is the total number of correctly classified records.

Accuracy- We have chooses accuracy matrix to measure the performance of all the models. The ratio of number of correct predictions to the total number of predictions Made.

$$\text{Accuracy} = \frac{\text{Number of correct Prediction}}{\text{Total numbers of predictions made.}}$$

CHAPTER 6 - RESULTS AND DISCUSSION

Machine learning regression algorithms developed for prediction of car price in earlier stage. We used 80% of data for training and 20% of data for testing. In this ratio of data splitting Here we found that Random Forest Regressor has the least root mean square error. Comparison of results of all the implemented Regressors are listed in below

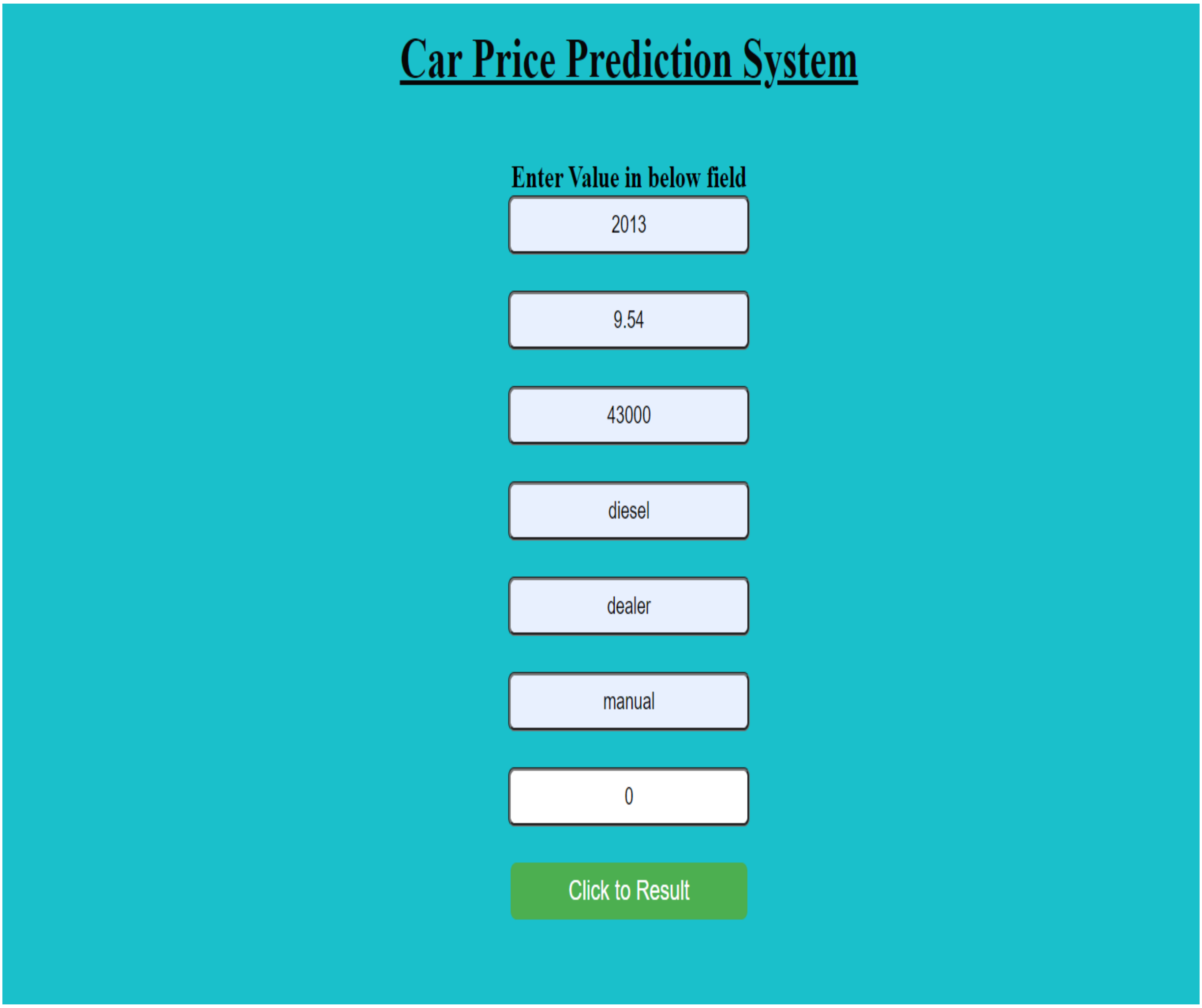
Machine Learning Algorithms	**Error (Root mean square error)**
Linear Regression	1.9558042330415681
Decision Tree	1.654166494205959
Random Forest	1.4500562999484015
Lasso Regressor	2.1219685349371615

Creating a User Interface for Accessibility:

The last part of the project is the creation of a user interface for the model. This user interface is used to enter unseen data for the model to read and then make a prediction. The user interface is created using “Flask” Web app, Hyper Text Markup Language, and Cascading Style Sheets.

CHAPTER 7 – RESULT AND ANALYSIS

The project predicts the Car Price based on the relevant car details collected. When the person enters all the relevant car data required in the online Web portal, this data is then passed on to the trained model for it to make predictions the model then makes the prediction with an accuracy of 85%, which is good and reliable. Following figure shows the basic UI form which requires the user to enter the specific car related data fields. These parameters help to determine the price of the car. Our research has the added benefit of an associated Web app, which makes the model more user friendly and easily understandable for a novice.



Car Price Prediction System

Enter Value in below field

2013

9.54

43000

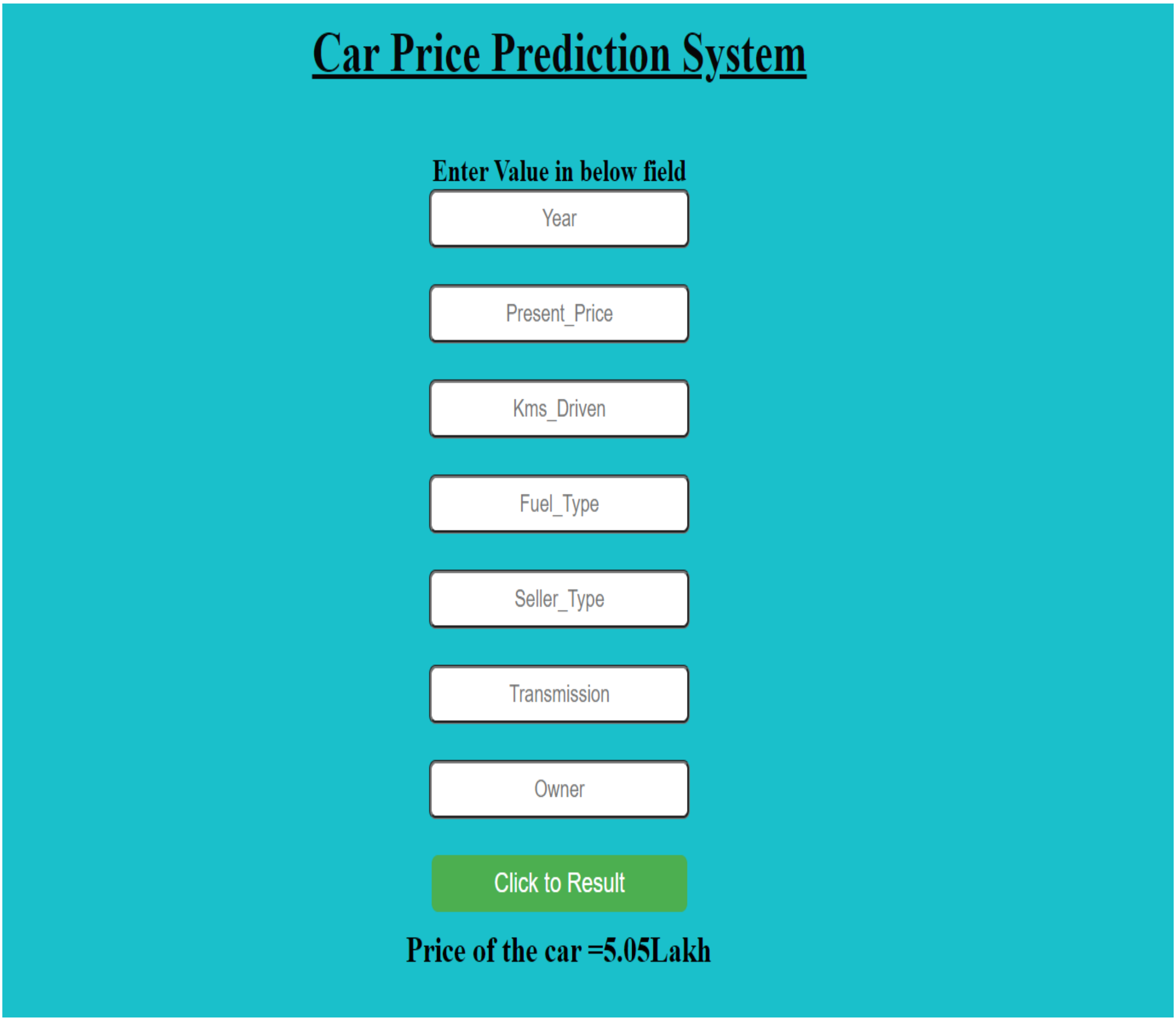
diesel

dealer

manual

0

On submission of this form, data the model gives the result in the form of Table; as shown in following figures;



Car Price Prediction System

Enter Value in below field

Price of the car =5.05Lakh

CHAPTER 8 - CONCLUSION

The increased prices of new cars and the financial incapability of the customers to buy them, Used Car sales are on a global increase. Therefore, there is an urgent need for a Used Car Price Prediction system which effectively determines the worthiness of the car using a variety of features. The proposed system will help to determine the accurate price of used car price prediction. This paper compares 4 different algorithms for machine learning : Linear Regression, Lasso Regression and Decision Tree Regression and Random Forest Regressor.

Using machine learning approaches, this project proposed a scalable framework for India based used cars price prediction. An efficient machine learning model is built by training, testing, and evaluating three machine learning regressors named Random Forest Regressor, Linear Regression, and Bagging Regressor. As a result of pre-processing and transformation, Random Forest Regressor came out on top with 95% accuracy. Each experiment was performed in real-time within the Google Colab environment. In comparison to the system's integrated Jupyter notebook and Anaconda's platform, algorithms took less training time in Google Colab

CHAPTER 9 - REFRENCES

[1] Sameerchand Pudaruth, “Predicting the Price of Used Cars using Machine Learning Techniques”;(IJICT 2014)

[2] Enis gegic, Becir Isakovic, Dino Keco, Zerina Masetic, Jasmin Kevric, ”Car Price Prediction Using Machine Learning”; (TEM Journal 2019)

[3] 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)

[4] Nitis Monburinon, Prajak Chertchom, Thongchai Kaewkiriya, Suwat Rungpheung, Sabir Buya, Pitchayakit Boonpou, “Prediction of Prices for Used Car by using Regression Models” (ICBIR 2018)

[5] Doan Van Thai, Luong Ngoc Son, Pham Vu Tien, Nguyen Nhat Anh, Nguyen Thi Ngoc Anh, “Prediction car prices using qualify qualitative data and knowledge-based system” (Hanoi National University)