**A Project Review ETE Report**

On

# CAR PRICE PREDICTION

# USING MACHINE LEARNING

***Submitted in partial fulfilment of the requirement for the award ofthe degreeof***

BTECH CSE

GALGOTIAS UNIVERSITY

(Established under Galgotias University Uttar Pradesh Act No. 14 of 2011)

**Under The Supervision of Ms.**

**SONIA KUKREJA**


Submitted By

**REYANSH**

(19SCSE1010136)

**PARVEZ AKHTAR**

(19SCSE1010156)

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

# SCHOOL OF COMPUTING SCIENCE AND ENGINEERING
# GALGOTIAS UNIVERSITY, GREATER NOIDA

## CANDIDATE'S DECLARATION

I/We hereby certify that the work which is being presented in the thesis/project/dissertation, entitled **"CAR PRICE PREDICTION USING MACHINE LEARNING"** in partial fulfillment of the requirements for the award of the ~~B.~~TECH sub~~m~~itted in the School of Computing Science and Engineering of Galgotias University, Greater Noida, is an original work carried out during the period of month, Year to Month and Year, under the supervision of Sonia Kukreja… Ass Profs, Department of Computer Science and Engineering/Computer Application and Information and Science, of School of Computing Science and Engineering , Galgotias University, Greater Noida

The matter presented in the thesis/project/dissertation has not been submitted by me/us for the award of any other degree of this or any other places.

REYANSH 19SCSCE1010136

PARVEZ AKHTAR 19SCSCE1O1O156

This is to certify that the above statement made by the candidates is correct to the best of my knowledge.

Supervisor Name

Designation

## CERTIFICATE

The Final Thesis/Project/ Dissertation Viva-Voce examination of REYANSH – 19SCSCE101036 & PARVEZ AKHTAR 19SCSCE1010156 has been held on_____and his/her work is recommended for the award of B.TECH

**Signature of Examiner(s)** **Signature of Supervisor(s)**

**Signature of Project Coordinator** **Signature of Dean**

Date:

Place: Greater Noida

## Abstract

The price of a new car in the industry is fixed by the manufacturer with some additional costs incurred by the Government in the form of taxes. So, customers buying a new car can be assured of the money they invest to be worthy. 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, there is an urgent need for a Used Car Price Prediction system that effectively determines the worthiness of the car using a variety of features.

The existing system includes a process where a seller decides a price randomly and the buyer has no idea about the car and its value in the present day scenario. In fact, the seller also has no idea about the car's existing value or the price he should be selling the car at.

To overcome this problem we have developed a model which will be highly effective. Regression Algorithms are used because they provide us with continuous value as output and not a categorized value. Because of which it will be possible to predict the actual price of a car rather than the price range of a car.

This paper presents a vehicle price prediction system by using the supervised machine learning technique. The research uses multiple linear regression as the machine learning prediction method which offered 98% prediction precision. Using multiple linear regression, there are multiple independent variables but one and only one dependent variable whose actual and predicted values are compared to find precision of results. This paper proposes a system where price is dependent variable which is predicted, and this price is derived from factors like vehicle's model, make, city, version, color, mileage, alloy rims and power steering.

In this paper, we investigate the application of supervised machine learning techniques to predict the price of used cars in Mauritius. The predictions are based on historical data collected from daily newspapers. Different techniques like multiple linear regression analysis, k-nearest neighbours, naïve bayes and decision trees have been used to make the predictions. The predictions are then evaluated and compared in order to find those which provide the best performances. A seemingly easy problem turned out to be indeed very difficult to resolve with high accuracy. All the four methods provided comparable performance. In the future, we intend to use more sophisticated algorithms to make the predictions.

# Table of Contents

# List of Table

# List of Figures

# Chapter 1: Introduction

## Introduction

Vehicle price prediction especially when the vehicle is used and not coming directly from the factory, is both a critical and important task. With the increase in demand for used cars and an up to 8 percent decrease in demand for the new cars in 2013, more and more vehicle buyers are finding alternatives of buying new cars outright. People prefer to buy cars through lease which is a legal contract between buyer and seller. The seller category includes direct seller or third party, business entity, or insurance company. Under a lease contract, the buyers pay regular installments of the item purchased for a predefined period of time. These lease installments are dependent upon the estimated price of the vehicle and thus, sellers are interested to know about the fair estimated price of their vehicles. It is found through studies that finding a fair estimated price of a used car is important as well as challenging. So, there is a need for an accurate price prediction mechanism for used cars. Prediction techniques of machine learning can be helpful in this regard. Machine learning uses two techniques, i.e., inductive and deductive. Deductive/Supervised learning is based on the usage of existing facts and knowledge to deduce new knowledge and facts while in inductive machine learning new computer programs are created by finding patterns and rules in the new data sets which were never explored before. We use the deductive approach of multiple linear regression since it creates new values based on existing values. In this technique, there is a single Try Editor (beta) in Google Docs Get advanced suggestions while you write. Turn it on Dismissdependent variable Y and there can be multiple independent variables X. The relationship among variables is direct or linear. This paper has the following goals: Design: The research includes the design of a system explaining the linear relationship between X and Y which are price and other factors like a model and make of the car. ii. Predict: The research predicts the price of the vehicle using a linear regression model which identifies different patterns and projects and predicts the value of the vehicle. iii. Confirm: The research finds out which variable associated with the vehicle is the best predictor of its price. There are many types of linear regressions and this research uses multiple linear regression where there is more than one independent variable. The data associated with the investigation was very large because there are thousands of used cars and each car's data comprises of values of many features. Both data gathering and analysis are complex. In the beginning, two thousand records of used cars were recorded and the data was obtained from Quikr which is a well-known online company for reselling used and new cars in India. The research used only those cars that contained price details so that the results could be verified. Features like car's model, make,

version, city, color, mileage, engine capacity, alloy rims, power steering, engine type, and price were included.

**Tools Used:**

**i) Anaconda Navigator**: Anaconda Navigator is a graphical user interface which includes Anaconda distribution that helps in the launching of application and in the management of conda packages, environment and channels without using the command line system. It is a package manager, an environment manager, a Python/R data science distribution, and a collection of over 7,500+ open-source packages. It is available for operating systems like Windows, macOS and Linux.

**ii) Jupyter Notebook**: The Jupyter Notebook is an open -source web application that you can use to create and share documents that contain live code, equations, visualizations, and text. Jupyter Notebook is maintained by the people at Project Jupyter. It can be used for data science, statistical modeling, machine learning and many such things

**Technologies Used:**

**i)Machine Learning**: Machine Learning System automatically learn program from data. It used in web Search, spam filter, recommendation system, and placement, credit scoring, recommendation of stock trading, fraud detection etc. There are lot of types of Machine learning but the one that used in our system.

**ii) Python**: It is a high-level general-purpose programming language. It is dynamically typed and also is garbage-collected. Multiple Programming and object-oriented and functional programming are one of the main features of python programming. Flexibility in Python allows it to be a great option for Machine Learning. Developing Scripts in Python is much easier because of all the standard libraires that are present in Python

# Chapter 2: Literature Review

Surprisingly, work on estimated the price of used cars is very recent but also very sparse. In her MSc thesis [3], Listiani showed that the regression model build using Try Editor (beta) in Google Docs Get advanced suggestions while you write. Turn it on Dismisssupport vector machines (SVM) can estimate the residual price of leased cars with higheraccuracy than simple multiple regression or multivariate regression. SVM is Predicting the Price of Used Cars using Machine Learning Techniques 755 better able to deal with very high dimensional data (number of features used to predict the price) and can avoid both over-fitting and underfitting. In particular, she used a genetic algorithm to find the optimal parameters for SVM in less time. The only drawback of this study is that the improvement of SVM regression over simple regression was not expressed in simple measures like mean deviation or variance.

In another university thesis [4], Richardson working on the hypothesis that car manufacturers are more willing to produce vehicles that do not depreciate rapidly. In particular, by using multiple regression analysis, he showed that hybrid cars (cars which use two different power sources to propel the car, i.e. they have both an internal combustion engine and an electric motor) are more able to keep their value than traditional vehicles. This is likely due to more environmental concerns about the climate and because of its higher fuel efficiency. The importance of other factors like age, mileage, make, and MPG (miles per gallon) were also considered in this study. He collected all his data from various websites. Wu et al. [5] used a neuro-fuzzy knowledge-based system to predict the price of used cars. Only three factors namely: the make of the car, the year in which it was manufactured, and the engine style were considered in this study.

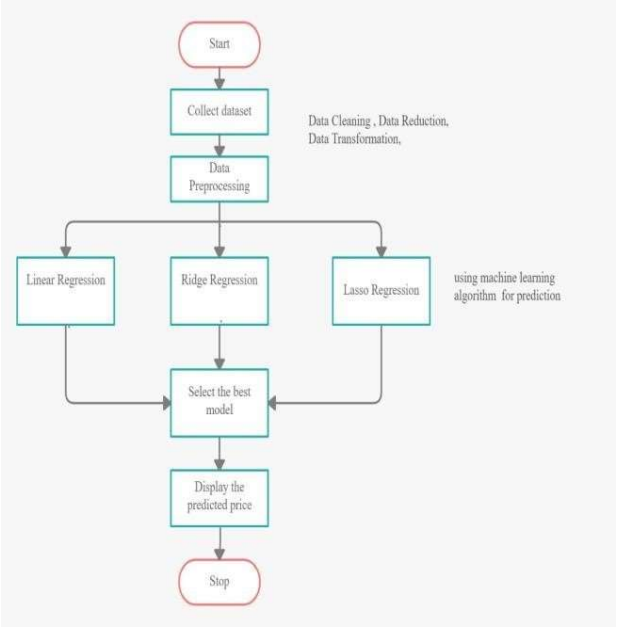
The proposed system produced similar results as compared to simple regression methods. Car dealers in the USA sell hundreds of thousands of cars every year through leasing Try Editor (beta) in Google Docs Get advanced suggestions while you write. Turn it on Dismiss[6]. Most of these cars are returned at the end of the leasing period and must be resold. Selling these cars at the right price has a major economic connotation for their success. In response to this, the ODAV (Optimal Distribution of Auction Vehicles) system was developed by Du et al. [6].

This system not only estimates the best price for reselling the cars but also provides advice on where to sell the car. Since the United States is a huge country, the location where the car is sold also has a nontrivial impact on the selling price of used cars. A k-

nearest neighbor regression model was used for forecasting the price. Since this system was started in 2003, more than two million vehicles have been distributed via this system [6].

## Chapter 3: Project Design

3.1 Proposed System:




Project Prerequisites
The experiment has been conducted by a decent powered Intel® core ™ i7-7500U CPU @ 2.70GHz (4 CPU's) with a memory size of 8GB.. Python has been used as the development language with Development environment being provided by Windows. Anaconda Tools have been used for providing integrated development environment.

3.1 Database:

**https://github.com/knowledgeshelfit/carpr/blob/main/Car_Price.ipynb**

METHODOLOGY

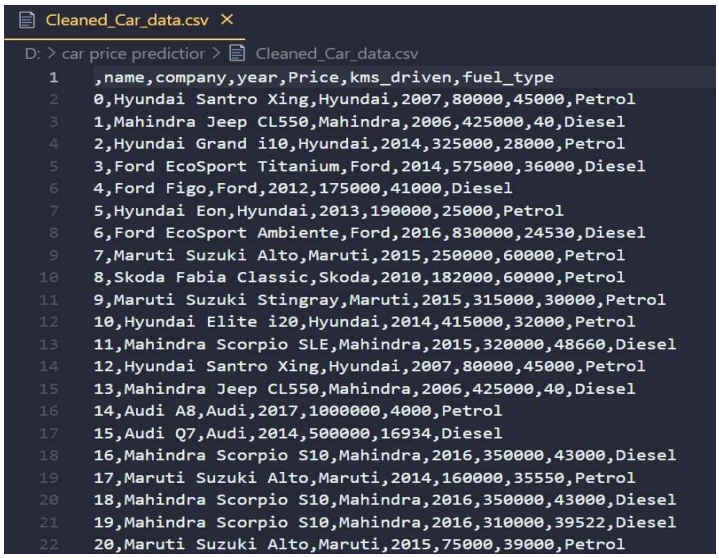
Data is collected from Kaggle > quikr_car.csv . Data is collected and processed.The following attributes were captured for each car: name, company ,year ,Price, kms_driven, fuel_type After raw data has been collected and stored to local database, data preprocessing step was applied. Many of the attributes were sparse and they do not contain useful information for prediction. Hence, it is decided to remove them from the dataset. The collected raw data contains 893 samples .

> car price predictior > quikr_car.csv

```
name,company,year,Price,kms_driven,fuel_type
Hyundai Santro Xing XO eRLX Euro III,Hyundai,2007,"80,000","45,00
Mahindra Jeep CL550 MDI,Mahindra,2006,"4,25,000",40 kms,Diesel
Maruti Suzuki Alto 800 Vxi,Maruti,2018,Ask For Price,"22,000 kms"
Hyundai Grand i10 Magna 1.2 Kappa VTVT,Hyundai,2014,"3,25,000","2
Ford EcoSport Titanium 1.5L TDCi,Ford,2014,"5,75,000","36,000 kms
Ford EcoSport Titanium 1.5L TDCi,Ford,2015,Ask For Price,"59,000
Ford Figo,Ford,2012,"1,75,000","41,000 kms",Diesel
Hyundai Eon,Hyundai,2013,"1,90,000","25,000 kms",Petrol
Ford EcoSport Ambiente 1.5L TDCi,Ford,2016,"8,30,000","24,530 kms
Maruti Suzuki Alto K10 VXi AMT,Maruti,2015,"2,50,000","60,000 kms
Skoda Fabia Classic 1.2 MPI,Skoda,2010,"1,82,000","60,000 kms",Pe
Maruti Suzuki Stingray VXi,Maruti,2015,"3,15,000","30,000 kms",Pe
Hyundai Elite i20 Magna 1.2,Hyundai,2014,"4,15,000","32,000 kms",
Mahindra Scorpio SLE BS IV,Mahindra,2015,"3,20,000","48,660 kms",
Hyundai Santro Xing XO eRLX Euro III,Hyundai,2007,"80,000","45,00
Mahindra Jeep CL550 MDI,Mahindra,2006,"4,25,000",40 kms,Diesel
Audi A8,Audi,2017,"10,00,000","4,000 kms",Petrol
Audi Q7,Audi,2014,"5,00,000","16,934 kms",Diesel
Mahindra Scorpio S10,Mahindra,2016,"3,50,000","43,000 kms",Diesel
Maruti Suzuki Alto 800,Maruti,2014,"1,60,000","35,550 kms",Petrol
Mahindra Scorpio S10,Mahindra,2016,"3,50,000","43,000 kms",Diesel
Mahindra Scorpio S10,Mahindra,2016,"3,10,000","39,522 kms",Diesel
```

**RAW DATA**

since data is collected and we need to clean these samples ,perform cleaning and saves the cleaned samples in csv file.The csv file is later used to load dat into software for building machine learning models.After cleanup process, the data set has been reduced to 815 samples .



Cleaned_Car_data.csv

D: > car price predictior > Cleaned_Car_data.csv

```
,name,company,year,Price,kms_driven,fuel_type
0,Hyundai Santro Xing,Hyundai,2007,80000,45000,Petrol
1,Mahindra Jeep CL550,Mahindra,2006,425000,40,Diesel
2,Hyundai Grand i10,Hyundai,2014,325000,28000,Petrol
3,Ford EcoSport Titanium,Ford,2014,575000,36000,Diesel
4,Ford Figo,Ford,2012,175000,41000,Diesel
5,Hyundai Eon,Hyundai,2013,190000,25000,Petrol
6,Ford EcoSport Ambiente,Ford,2016,830000,24530,Diesel
7,Maruti Suzuki Alto,Maruti,2015,250000,60000,Petrol
8,Skoda Fabia Classic,Skoda,2010,182000,60000,Petrol
9,Maruti Suzuki Stingray,Maruti,2015,315000,30000,Petrol
10,Hyundai Elite i20,Hyundai,2014,415000,32000,Petrol
11,Mahindra Scorpio SLE,Mahindra,2015,320000,48660,Diesel
12,Hyundai Santro Xing,Hyundai,2007,80000,45000,Petrol
13,Mahindra Jeep CL550,Mahindra,2006,425000,40,Diesel
14,Audi A8,Audi,2017,1000000,4000,Petrol
15,Audi Q7,Audi,2014,500000,16934,Diesel
16,Mahindra Scorpio S10,Mahindra,2016,350000,43000,Diesel
17,Maruti Suzuki Alto,Maruti,2014,160000,35550,Petrol
18,Mahindra Scorpio S10,Mahindra,2016,350000,43000,Diesel
19,Mahindra Scorpio S10,Mahindra,2016,310000,39522,Diesel
20,Maruti Suzuki Alto,Maruti,2015,75000,39000,Petrol
```

**CLEADED DATA**

# 4. Module Description

In python have use several libraries, and information about few of them that are used to develop this system is below:

- Numpy:

It offers powerful N-Dimensional Array that are fast and versatile and also helps vectorization and indexing. It also offers itself as a numerical computing tool which can solve many mathematical functions, linear algebra routines or fourier transformations. The core of Numpy is a well-optimized C code so it provides flexibility with the speed.

- Panda

Panda is the fast easy to use tool which is use for data analysis that have been built over the python programming language. It is a powerful tool which can be used for data manipulation. It is one of the most important libraries in the field of Data Analysis and Data Science.

- Sklearn:

It is a simple and efficient tool which is used for data prediction. It is open source so it is available to everyone and is reusable to various context. It is built on the basis of NumPy, SciPy and matplotlib. It can be used for classification, Regression and Clustering

**CODE -:**

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns
```

## 3 Data Connectivity:

In the option 1, we use the concept of Data Visualization . Below are the code snippets of the program that was used during that process:

```
data = pd.read_csv('car_data.csv')
data.head()
```

Out[2]:

| | Car_Name | Year | Selling_Price | Present_Price | Kms_Driven | Fuel_Type | Seller_Type | Transmission | Owner |
|---|---|---|---|---|---|---|---|---|---|
| 0 | ritz | 2014 | 3.35 | 5.59 | 27000 | Petrol | Dealer | Manual | 0 |
| 1 | sx4 | 2013 | 4.75 | 9.54 | 43000 | Diesel | Dealer | Manual | 0 |
| 2 | ciaz | 2017 | 7.25 | 9.85 | 6900 | Petrol | Dealer | Manual | 0 |
| 3 | wagon r | 2011 | 2.85 | 4.15 | 5200 | Petrol | Dealer | Manual | 0 |
| 4 | swift | 2014 | 4.60 | 6.87 | 42450 | Diesel | Dealer | Manual | 0 |

Importing all the required libraires and establishing connections

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns
```

In [3]:
```
data.shape
```

Out[3]:
```
(301, 9)
```

In [4]:
```
data.info()
```

```
<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
```

In [5]:
```
data.describe().T
```

Out[5]:

| | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| Year | 301.0 | 2013.627907 | 2.891554 | 2003.00 | 2012.0 | 2014.0 | 2016.0 | 2018.0 |
| Selling_Price | 301.0 | 4.661296 | 5.082812 | 0.10 | 0.9 | 3.6 | 6.0 | 35.0 |
| Present_Price | 301.0 | 7.628472 | 8.644115 | 0.32 | 1.2 | 6.4 | 9.9 | 92.6 |
| Kms_Driven | 301.0 | 36947.205980 | 38886.883882 | 500.00 | 15000.0 | 32000.0 | 48767.0 | 500000.0 |
| Owner | 301.0 | 0.043189 | 0.247915 | 0.00 | 0.0 | 0.0 | 0.0 | 3.0 |

In [6]:
```
data = data.drop('Car_Name', axis=1)
data.head()
```

Out[6]:

| | Year | Selling_Price | Present_Price | Kms_Driven | Fuel_Type | Seller_Type | Transmission | Owner |
|---|---|---|---|---|---|---|---|---|
| 0 | 2014 | 3.35 | 5.59 | 27000 | Petrol | Dealer | Manual | 0 |
| 1 | 2013 | 4.75 | 9.54 | 43000 | Diesel | Dealer | Manual | 0 |
| 2 | 2017 | 7.25 | 9.85 | 6900 | Petrol | Dealer | Manual | 0 |
| 3 | 2011 | 2.85 | 4.15 | 5200 | Petrol | Dealer | Manual | 0 |
| 4 | 2014 | 4.60 | 6.87 | 42450 | Diesel | Dealer | Manual | 0 |

In [7]:
```
data['Years_old'] = 2021 - data.Year
data.head()
```

Out[7]:

| | Year | Selling_Price | Present_Price | Kms_Driven | Fuel_Type | Seller_Type | Transmission | Owner | Years_old |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 2014 | 3.35 | 5.59 | 27000 | Petrol | Dealer | Manual | 0 | 7 |
| 1 | 2013 | 4.75 | 9.54 | 43000 | Diesel | Dealer | Manual | 0 | 8 |
| 2 | 2017 | 7.25 | 9.85 | 6900 | Petrol | Dealer | Manual | 0 | 4 |
| 3 | 2011 | 2.85 | 4.15 | 5200 | Petrol | Dealer | Manual | 0 | 10 |
| 4 | 2014 | 4.60 | 6.87 | 42450 | Diesel | Dealer | Manual | 0 | 7 |

In [8]:
```
data.drop('Year', axis=1, inplace=True)
data.head()
```

Out[8]:

| | Selling_Price | Present_Price | Kms_Driven | Fuel_Type | Seller_Type | Transmission | Owner | Years_old |
|---|---|---|---|---|---|---|---|---|
| 0 | 3.35 | 5.59 | 27000 | Petrol | Dealer | Manual | 0 | 7 |
| 1 | 4.75 | 9.54 | 43000 | Diesel | Dealer | Manual | 0 | 8 |
| 2 | 7.25 | 9.85 | 6900 | Petrol | Dealer | Manual | 0 | 4 |
| 3 | 2.85 | 4.15 | 5200 | Petrol | Dealer | Manual | 0 | 10 |
| 4 | 4.60 | 6.87 | 42450 | Diesel | Dealer | Manual | 0 | 7 |

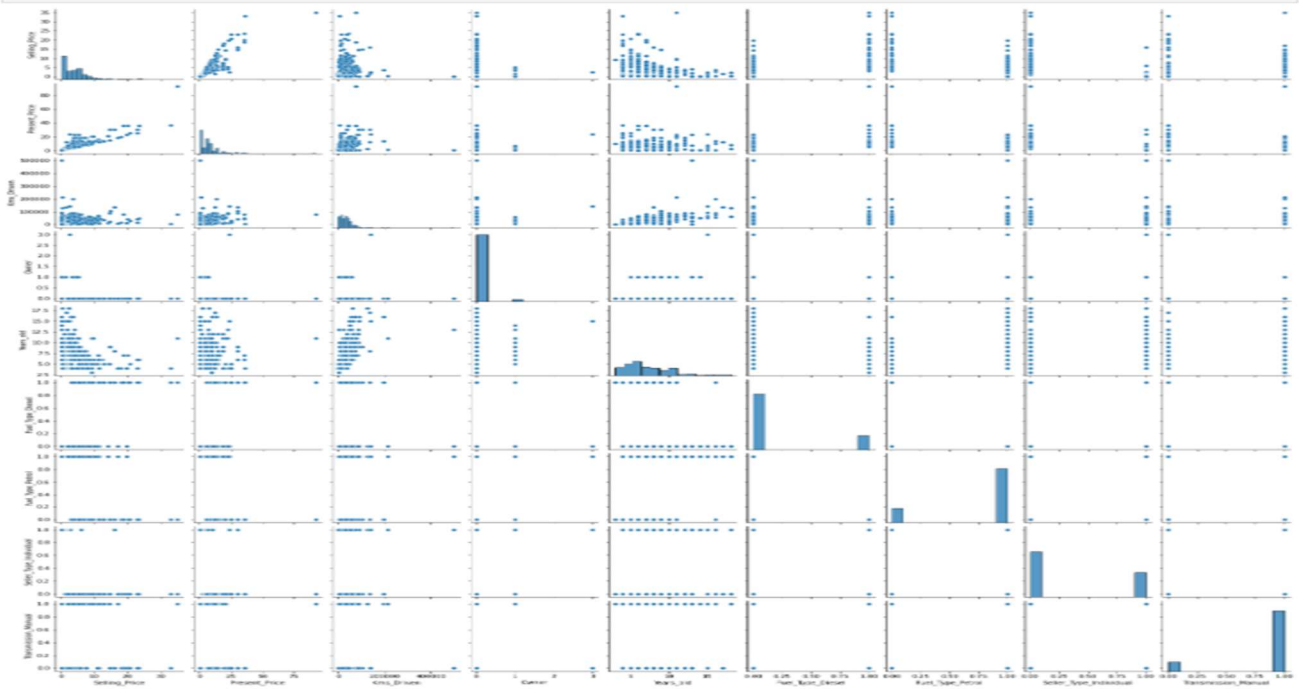
In [9]:
```
data = pd.get_dummies(data,drop_first=True)
```

In [10]:
```
data.head()
```

Out[10]:

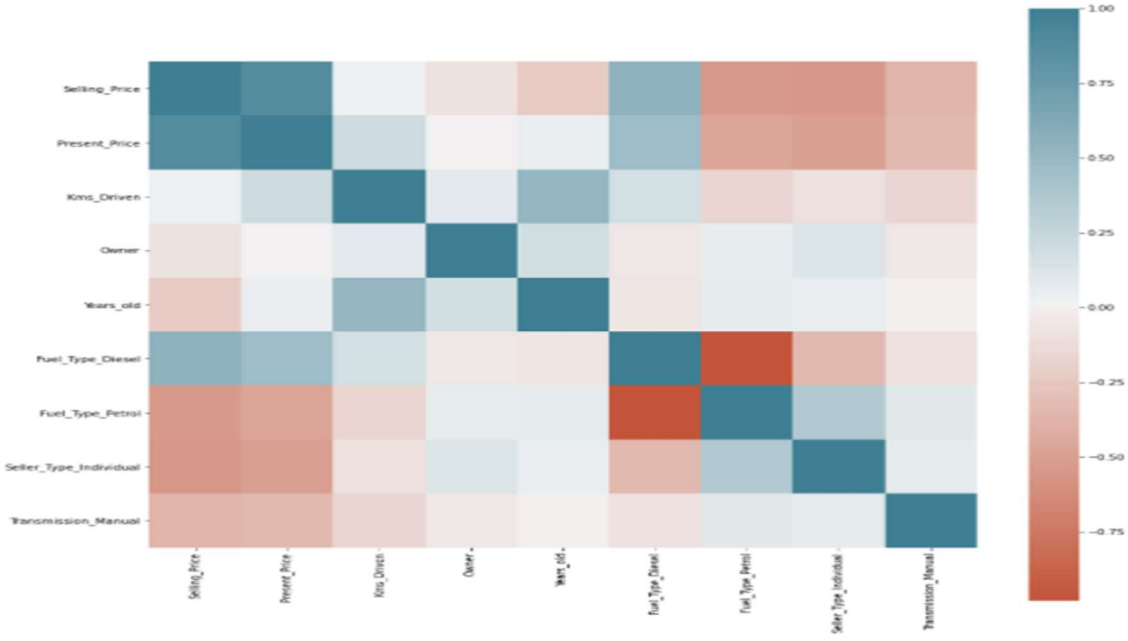
| | Selling_Price | Present_Price | Kms_Driven | Owner | Years_old | Fuel_Type_Diesel | Fuel_Type_Petrol | Seller_Type_Individual | Transmission_Manual |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 3.35 | 5.59 | 27000 | 0 | 7 | 0 | 1 | 0 | 1 |
| 1 | 4.75 | 9.54 | 43000 | 0 | 8 | 1 | 0 | 0 | 1 |
| 2 | 7.25 | 9.85 | 6900 | 0 | 4 | 0 | 1 | 0 | 1 |
| 3 | 2.85 | 4.15 | 5200 | 0 | 10 | 0 | 1 | 0 | 1 |
| 4 | 4.60 | 6.87 | 42450 | 0 | 7 | 1 | 0 | 0 | 1 |

In [11]:
```
sns.pairplot(data);
```



In [12]:
```
plt.figure(figsize=(15,15))
sns.heatmap(
    data.corr(),
    cmap=sns.diverging_palette(20, 220, n=200),
    square=True
);
```

In [12]:
```
plt.figure(figsize=(15,15))
sns.heatmap(
    data.corr(),
    cmap=sns.diverging_palette(20, 220, n=200),
    square=True
);
```



We trained a Machine Learning Model and used that to predict the best fertilizers. Below is the prediction test result that came after when we checked our model for the test data with the code snippet too.

# Chapter 5: Conclusion and Future Scope

## 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 3 different algorithms for machine learning : Linear Regression, Lasso Regression and Ridge Regression.
The data set used in this paper can be very valuable in conducting similar research using different prediction techniques. The prices of vehicles can be predicted using this data set on same or different prediction software as well. The data obtained under this research facilitated in prediction of prices of used cars through linear regression method. Many assumptions were made on the basis of the data set. The proposed system evaluated variables and selected the most relevant variables out of the dataset and reduced the complexity of model by eliminating unrelated variables during processing and analysis phase.

## Future Enhancements

In future this machine learning model may bind with various website which can provide real time data for price prediction. Also we may add large historical data of car price which can help to improve accuracy of the machine learning model. We can build an android app as user interface for interacting with user. For better performance, we plan to judiciously design deep learning network structures, use adaptive learning rates and train on clusters of data rather than the whole dataset.

The future price prediction of used cars with the help of same data set will comprise of using fuzzy logic, KNN and genetic algorithm.

## REFERENCES

[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)