



Kabul University

Computer Science Faculty

Information Systems Department

# Car Price Prediction in Afghanistan using Machine Learning Algorithms

A monograph submitted in partial fulfillment of the requirements

for the award of the degree of

Bachelor of Information Systems in Computer Science

**Submitted by:**

Mohibullah “Alokozai”

**Supervised by:**

Asst.Prof. Hedayatullah “Lodin”

July, 2024

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

## Certificate of Approval

I certify that I have read the “Car Price Prediction in Afghanistan using Machine Learning Algorithms” report submitted by Mohibullah Alokzai as partial fulfillment for the award of the degree in Bachelor of Information System of Computer Science Faculty at Kabul University. I have evaluated the report and found it up to the requirements in its scope and quality for the award of the degree.

Supervisor Name: Asst. Prof. Hedayatullah “Lodin”

Signature: \_\_\_\_\_

Date: \_\_\_\_\_

Name: Associate. Prof. Mohammad Shuaib “Zarinkhail”

(Head of Information System Department)

Signature: \_\_\_\_\_

Date: \_\_\_\_\_

Name: Associate. Prof. Amir Krar “Shahidzay”

(Dean of Information System Department)

Signature: \_\_\_\_\_

Date: \_\_\_\_\_

## Author's Declaration

Hereby I, Mohibullah Alokozai declare that this Bachelor thesis is my personal achievement and the result of an original investigation. I further state that this thesis has not been previously submitted for any other academic degree.

Date: July -2024

Student Name and Sign: \_\_\_\_\_

## Abstract

This abstract presents the application of a decision tree model with a high  $R^2$ \_score for predicting car prices in Afghanistan. The dataset used for this study comprises various features related to cars, including manufacturing company, car model, year, color, seat hand, engine type, transmission type, fuel type, and plate type.

The objective of this research is to develop an accurate prediction model that can estimate car prices based on the given features. The decision tree model is a non-parametric supervised learning algorithm that constructs a tree-like structure to make predictions based on a series of splitting rules. It is particularly useful for capturing non-linear relationships and interactions between variables. It's known to be the most accurate & easy to use between the three models; Linear Regression, KNN & Decision Tree itself.

To evaluate the performance of the decision tree model, the dataset is split into training and testing sets. The model is trained on the training set, and the  $R^2$ \_score is used to assess the goodness-of-fit of the model. The  $R^2$ \_score, also known as the coefficient of determination, measures the proportion of the variance in the target variable that can be explained by the model. A high  $R^2$ \_score indicates a strong correlation between the predicted car prices and the actual prices.

The results of this study indicate that the decision tree model achieved an impressive  $R^2$ \_score of 98%. This high  $R^2$ \_score demonstrates the model's ability to accurately estimate car prices based on the provided features. The strong correlation between the predicted and actual prices suggests that the decision tree model captures the underlying patterns and trends in the Afghan car market.

The findings of this research have significant implications for car buyers, sellers, and enthusiasts in Afghanistan. The decision tree model with its high  $R^2$ \_score can serve as a valuable tool for estimating car prices. By leveraging this model, it will also revolutionize the car marketing industry & stakeholders in the Afghan automotive business industry can make informed decisions, negotiate fair prices.

# Acknowledgment

Throughout the working on this thesis I received tangible and great deal of support.

I would like to thank my supervisor Asst.Prof. Hedayatullah "Lodin" for his guidance through each step of our work and effort. His experience and expertise was a great help and guide for me to achieve this goal and successfully complete it. His insightful feedback provided me chances to sharpen my thinking and brought my work in a higher level.

## Acronyms

KNN	K Nearest Neighbor
MAPE	Mean Absolute Percentage Error
RMSE	Root Mean Square Error
XGBoost	eXtreme Gradient Boosting
VGG-16	Visual Geometry Group-16
RoBERTa	Robustly Optimized BERT approach

## Table of Contents

List of Figures .....	viii
Chapter 1: Introduction .....	1
1.1 Introduction .....	1
1.2 Problem statement .....	2
1.3 Impact of the problem.....	4
Chapter 2: Related Work (Literature Review) .....	8
Chapter 3: Research Methodology.....	12
3.1 Data Gathering.....	12
3.2 Data Conversion .....	12
3.3 General Dataset .....	13
3.4 Analyzing Cars characteristics and fuel-types.....	13
3.4 Preprocessing.....	16
3.5 Splitting the Dataset .....	16
3.6 Algorithm Selection and Evaluation.....	17
3.7 Training the Decision Tree Regressor .....	22
3.8 Making Prediction on the Testing Set .....	22
3.9 Evaluating the Models Performance.....	22
Chapter 4: Implementation Details .....	23
4.1 Dataset Loading:.....	23
4.2 Splitting The Dataset: .....	23
4.3 Feature Extraction: .....	24
4.5 Model Training:.....	24
4.6 Prediction Variable:.....	24
4.8 Interface Development .....	25
4.9 The Project Code Consist the Following Components:.....	26
4.9.1 application.py:.....	26
4.9.2 index.html:.....	26
4.9.3 result.html:.....	26
4.10 Getting Started.....	26
4.10.1 Pandas:.....	26
4.10.2 numpy:.....	26
4.10.3 scikit-learn:.....	26
4.10.4 Flask:.....	26



Chapter 5: Results .....	27
5.1 Limitation .....	27
1. Change in Model's R2-score:.....	27
2. Dataset Scope:.....	28
3. Antiquated Dataset:.....	29
5.2 Future Work.....	30
1. Advanced Interface:.....	30
2. Up-to-date Data:.....	30
3. Dataset Expansion:.....	31
5.2 Conclusion.....	32
References.....	33

## List of Figures

Figure 3. 1: Handwritten Data Collection .....	12
Figure 3. 2: General Dataset .....	13
Figure 3. 3: Relationship between car models and companies .....	14
Figure 3. 4: Relationship between car models, Hand type and Transmission .....	14
Figure 3. 5: Relationship between car models and Car Engine Type .....	15
Figure 3. 6: Relationship between car models and Car Fuel_type.....	15
Figure 3. 7: Data Preprocessing .....	16
Figure 3. 8: Splitting The Dataset and Evaluation Model.....	17
Figure 3. 9: Linear Regression Model .....	18
Figure 3. 10: K-Neighbors Regressor .....	19
Figure 3. 11: Decision Tree Regressor.....	20
Figure 3. 12 : Model Evaluation .....	21
Figure 4. 1: Dataset Loading.....	14
Figure 4. 2: Feature Extraction .....	24
Figure 4. 3: Training Decision Tree Regressor Model .....	24
Figure 4. 4: Prediction State.....	25
Figure 4. 5: Web Interface .....	25

# Chapter 1: Introduction

---

## 1.1 Introduction

The car industry is one of the largest and most important industries in the Afghanistan, with hundred thousand of vehicles being bought and sold every year. With the increasing demand for cars & improvement in technology, it has become essential for both buyers and sellers to accurately predict the price of a car. Predicting a car's price is a complex task that involves various factors such as company, model, year, mileage, condition, and several other based on where you live in. There has been a growing interest in developing machine learning models to predict car prices accurately, & that's what our model is all about.

The ability to predict car prices accurately can benefit both buyers and sellers. For buyers, accurate price prediction can help them make informed decisions when purchasing a car. It can also help them negotiate better deals with sellers. For sellers, accurate price prediction can help them set competitive prices for their cars and attract potential buyers.

The primary objective of this documentation is to provide a comprehensive overview of car price prediction methods and techniques. I aim to identify the most effective algorithms for predicting car prices accurately and discuss their potential applications in real-world scenarios. By doing so, we hope to contribute to the advancement of car price prediction research and provide valuable insights for both researchers and practitioners in the Afghan automotive industry & even worldwide.

The Afghan car market is highly diverse, with a wide range of vehicle makes, models, and price points available to consumers. This diversity poses a unique challenge when it comes to accurately predicting car prices. Factors such as regional preferences, local economic conditions, and cultural influences can all play a significant role in shaping the pricing landscape.

Moreover, the rapid technological advancements in the automotive industry, such as the integration of smart features, alternative fuel systems, and autonomous driving capabilities, have further complicated the task of price prediction. These emerging technologies not only impact the initial purchase price but also the long-term maintenance and resale value of vehicles.

To address these complexities, the development of robust and adaptable machine learning models is crucial. These models must be capable of capturing the nuanced relationships between the various factors that influence car prices, while also remaining flexible enough to accommodate the dynamic nature of the Afghan automotive market.

Furthermore, the implementation of car price prediction models can have far-reaching consequences beyond just the buyers and sellers. Policymakers, urban planners, and transportation authorities can leverage these insights to make more informed decisions regarding infrastructure development, public

transportation, and environmental regulations. By understanding the true cost of vehicle ownership, these stakeholders can develop policies that promote sustainable and equitable transportation solutions for the people of Afghanistan.

By delving deeper into the intricacies of the Afghan car market and exploring the potential applications of advanced price prediction models, this documentation aims to serve as a valuable resource for researchers, industry practitioners, and policymakers alike. The ultimate goal is to empower all stakeholders in the Afghan automotive ecosystem to make more informed and data-driven decisions, ultimately leading to a more efficient and resilient car market.

The development of accurate car price prediction models can also have significant implications for the broader automotive industry. By providing reliable insights into market trends and customer preferences, these models can help manufacturers and dealerships optimize their production, inventory management, and marketing strategies. This, in turn, can lead to improved financial performance, better customer satisfaction, and a more responsive and adaptive industry.

Additionally, the insights gained from car price prediction models can be leveraged to inform the design and development of new vehicle models. By understanding the factors that drive consumer demand and pricing preferences, automakers can better align their product offerings with the evolving needs and expectations of the Afghan market. This can result in increased sales, enhanced brand loyalty, and a stronger competitive positioning within the global automotive landscape.

## 1.2 Problem statement

**Economic instability:** Afghanistan's economy is highly unstable, and this has a significant impact on the purchasing power of the population. The majority of the Afghan people cannot afford to buy a car, which leads to low overall demand for vehicles. This economic reality is a critical factor that must be considered when developing car price prediction models for the Afghan market. The models must account for the unique challenges posed by the country's economic volatility, such as fluctuating exchange rates, high inflation, and limited access to credit and financing options. Without a clear understanding of the constraints faced by the average Afghan consumer, any price prediction efforts may fail to accurately capture the true dynamics of the car market. Addressing the impact of economic instability is paramount to ensuring the relevance and applicability of the price prediction models for the local context, ultimately benefiting both buyers and sellers navigating the complexities of the Afghan automotive industry.

**Lack of infrastructure:** In addition to the economic instability that constrains consumer purchasing power, Afghanistan's poor road infrastructure also significantly impacts the car market. Many parts of the country lack well-maintained, paved roads, making it challenging and often dangerous for people to operate vehicles on a regular basis.

The underdeveloped transportation network, with frequent potholes, unpaved surfaces, and limited access to remote regions, discourages people from investing in personal cars. Driving becomes a more arduous and risky endeavor, deterring potential buyers who may be concerned about the wear and tear on their vehicles, as well as the safety implications of navigating such poor road conditions.

This lack of reliable infrastructure not only dampens demand for cars but also affects the resale value and long-term ownership experience. Vehicles in Afghanistan often face accelerated depreciation and

higher maintenance costs due to the strain of navigating the country's subpar road network. This, in turn, further undermines the appeal of car ownership, creating a self-reinforcing cycle of low demand. Developing accurate car price prediction models for the Afghan market must account for the impact of infrastructure limitations. Factors such as the quality of roads, accessibility to repair and maintenance services, and the geographic distribution of the road network should be incorporated into the analysis. This holistic understanding of the transportation landscape will enable more realistic and contextually relevant price forecasts, benefiting both buyers and sellers navigating the Afghan automotive industry. Addressing the infrastructure challenges will be crucial for stimulating long-term growth and stability in the Afghan car market. Investments in road development, transportation planning, and supporting infrastructure can help unlock the potential demand and foster a more vibrant and accessible automotive ecosystem in the country.

**High import taxes:** The Afghan government levies significant import taxes and duties on vehicles entering the country, which dramatically increases the costs for car dealers and retailers. These high tax burdens are then passed on to consumers, making it extremely difficult for dealers to offer cars at affordable prices that align with the purchasing power of the majority of the Afghan population.

The import taxes can easily add thousands of dollars to the final price of a vehicle, effectively pricing out a large segment of potential buyers. This creates a challenging dynamic where dealers are caught between the need to maintain profitability and the imperative to provide cars that are within financial reach of the average Afghan consumer.

This mismatch between the high import-driven costs and the limited disposable incomes of the population further suppresses demand, as fewer people can realistically afford to purchase a car. The high taxes essentially create a barrier to entry that restricts access to vehicle ownership, stifling the overall growth and development of the Afghan automotive market.

Accurately forecasting car prices in this context requires a deep understanding of the specific import tax structures, tariff rates, and any related policy changes that may impact the landed costs for dealers. Incorporating these import-related factors into the price prediction models will enable more accurate and contextually relevant insights, benefiting both buyers and sellers navigating the Afghan car market.

Addressing the high import taxes could be a crucial policy lever for the Afghan government to consider, as reducing these burdens may help stimulate greater affordability and broader access to vehicle ownership. This, in turn, could catalyze increased demand, support the growth of the local automotive industry, and provide greater mobility options for the Afghan populace.

**Security concerns:** Afghanistan's high level of insecurity makes car ownership a risky proposition, as car theft and bombings are common occurrences that discourage people from buying vehicles. The threat of having a car stolen or targeted in an attack is a major deterrent, impacting the resale value and long-term ownership experience. Additionally, the need for heightened security measures, such as garage storage or security personnel, adds significant costs that further dissuade potential car buyers. Addressing these security challenges will be crucial to stimulating greater demand and fostering a healthier car market in the country.

**Limited financing options:** The limited availability of financing options in Afghanistan is a significant constraint on car purchases. Most people in the country struggle to afford the high upfront costs of vehicles, as they have limited access to credit or loan programs to facilitate car ownership.

This lack of financing alternatives disproportionately impacts lower-income Afghans who cannot save enough to pay the full price of a vehicle outright. Without the ability to buy cars on credit, the demand for automobiles in Afghanistan is reduced, as many potential buyers are unable to overcome the financial hurdles to vehicle ownership.

**Lack of trust in dealerships:** The lack of trust in car dealerships in Afghanistan is another key factor that hinders the growth of the country's automobile market. Many Afghans are wary of dealerships due to a history of fraud and corruption within the industry.

This pervasive distrust makes it challenging for dealerships to effectively sell cars, as potential customers are skeptical of the integrity of the sales process and the quality of the vehicles being offered. Buyers may be hesitant to make significant financial investments in cars if they do not have confidence in the honesty and reliability of the dealerships.

The prevalence of unscrupulous practices, such as fraudulent pricing, hidden fees, or the sale of vehicles with undisclosed issues, has eroded public confidence in the car dealership sector. This lack of trust creates an environment where customers are more likely to seek alternative, and potentially less regulated, channels for acquiring vehicles, further hampering the growth of the formal car market.

Addressing the trust deficit will require concerted efforts by both the government and the car dealership industry to implement robust regulatory frameworks, enhance transparency, and demonstrate a genuine commitment to ethical and customer-centric business practices. Building trust through improved industry standards, better consumer protections, and enhanced accountability measures could help revitalize the car sales landscape in Afghanistan.

## 1.3 Impact of the problem

The problem of accurately predicting car prices in Afghanistan has several significant implications for both car buyers and sellers, as well as the automotive market as a whole. Addressing this problem through the development of a machine learning-based price prediction system can have the following impacts:

**1- Informed Decision-Making:** By providing accurate price predictions, the system empowers car buyers in Afghanistan to make more informed decisions when purchasing vehicles. The ability to access estimated market values for different car models and specifications allows buyers to assess whether the listed price is fair and competitive.

This information equips them with the knowledge to negotiate more effectively and ensure they are not overpaying for their desired cars. The system's price predictions help create transparency in the car market, enabling buyers to make more confident and well-informed purchasing choices.

Ultimately, this feature supports the development of a healthier car market in Afghanistan by empowering consumers with the data they need to make smart and financially responsible decisions when acquiring vehicles.

**2- Fair Pricing:** The price prediction system also promotes fair pricing practices within the Afghan automotive market. By providing sellers with estimated market values for different car models and specifications, the system enables them to set reasonable and competitive asking prices for their vehicles.

This transparency helps avoid instances of overpricing or underpricing, which can erode trust in the market. When both buyers and sellers have access to reliable pricing information, it fosters an environment of equity and fairness in car transactions. Sellers can use the price predictions as a guide to ensure they are not asking for unreasonably high prices, while buyers can feel more confident that the listed prices reflect the true market value of the vehicles. This level of transparency and pricing fairness is crucial for building trust and stimulating a healthier car market in Afghanistan.

The implications of this fair pricing system extend beyond just the individual transactions. By promoting transparency and fairness, it can help regulate the overall pricing dynamics within the Afghan automotive industry. This, in turn, can lead to more sustainable and equitable growth, as buyers and sellers operate on a level playing field.

Moreover, the availability of accurate price predictions empowers consumers to make more informed decisions. They can use this information to negotiate better deals, ensure they are not being taken advantage of, and ultimately secure a fair price for their car purchases. This enhanced bargaining power ultimately benefits the Afghan car buyers, who can now approach the market with greater confidence and assurance.

In the long run, the establishment of this fair pricing system can have a transformative impact on the Afghan automotive landscape. By fostering an environment of trust, transparency, and equity, it lays the groundwork for a more robust and thriving car market. This not only benefits the individual consumers but also supports the broader economic development and growth of the Afghan automotive industry as a whole.

**3- Market Efficiency:** The accurate price predictions provided by the system contribute to enhanced efficiency within the Afghan automotive market. By reducing the information asymmetry between buyers and sellers, the system streamlines the negotiation process and facilitates smoother transactions.

When both parties have access to reliable pricing information, it enables them to engage in more informed and productive negotiations. Buyers can make more accurate assessments of fair market values, while sellers can price their vehicles competitively. This improved transparency and alignment of information leads to a more efficient allocation of resources within the automotive market.

With reduced instances of over- or underpricing, the market can operate more smoothly, and resources are directed towards the most valued car models and transactions. This improved efficiency benefits both buyers and sellers, ultimately supporting the overall development and growth of the Afghan car market.

**4- Transparency and Trust:** The price prediction system enhances transparency in the Afghan automotive market by providing clear and objective price estimates based on data-driven analysis. This level of transparency is crucial for fostering trust among both buyers and sellers.

By accessing reliable and unbiased pricing information, market participants can feel more confident in the integrity of the car sales process. The system reduces uncertainties and potential disputes related to car pricing, as both parties can refer to the estimated market values to negotiate and make informed decisions.

Increased transparency through the price prediction feature helps to build trust between buyers and sellers, which is essential for the healthy development of the Afghan car market. When consumers and dealers can rely on accurate and impartial pricing data, it contributes to a more stable and trustworthy automotive ecosystem.

This trust-building aspect of the system is particularly valuable in a market like Afghanistan, where concerns over fraud and corruption have previously eroded public confidence in car dealerships. The transparency and reliability of the price predictions can help to address these trust issues and facilitates.

**5- Economic Impact:** A reliable car price prediction system can have a significant positive impact on the overall economic development of Afghanistan's automotive sector. By facilitating a more competitive market environment, the system contributes to the growth and prosperity of this important industry.

The enhanced transparency and fair pricing practices enabled by the system can attract increased investment in the Afghan car market. Investors and businesses are more likely to participate in a market where there is clarity and trust in the pricing mechanisms. This, in turn, can lead to the introduction of more diversified product offerings, improved quality standards, and a more vibrant competitive landscape.

Furthermore, the system's ability to promote fair and transparent pricing helps to level the playing field for car sellers. It enables smaller dealers and individual sellers to compete more effectively by providing them with the data-driven insights they need to price their vehicles competitively. This healthy competition can drive innovation, improve customer service, and ultimately benefit the end consumers.

Ultimately, the economic impact of the car price prediction system can extend beyond the automotive sector itself. By fostering a more efficient and transparent car market, it can contribute to broader economic development in Afghanistan, supporting job creation, consumer spending, and overall economic growth.

**6- Market Insights:** Analyzing the dataset and understanding the factors influencing car prices generates valuable market insights that can be utilized by industry stakeholders, policymakers, and researchers. These insights can provide a deeper understanding of the trends, consumer preferences, and economic factors driving the automotive market in Afghanistan.

By addressing the problem of accurately predicting car prices through machine learning, the proposed system has the potential to significantly improve market efficiency, fairness, and transparency. The availability of reliable pricing information empowers individuals, enabling them to make more informed decisions when buying or selling vehicles.



This, in turn, can facilitate fairer and more transparent transactions, as both buyers and sellers have access to data-driven price estimates. The reduced uncertainties and potential disputes related to car pricing can foster greater trust and confidence in the market, contributing to its overall stability and development

Furthermore, the insights derived from the price prediction system can be invaluable for industry stakeholders, such as car dealers, manufacturers, and importers. They can utilize this information to better understand market dynamics, identify opportunities, and strategize their business operations. Policymakers can also leverage these insights to develop more informed and effective policies for the automotive sector, addressing any imbalances or inefficiencies.

For researchers, the market insights generated by the price prediction system can serve as a valuable resource for studying the Afghan automotive industry. They can analyze the data to uncover broader economic trends, consumer behavior patterns, and the impact of various factors on car prices. This knowledge can inform future research and guide the development of more comprehensive solutions to address the challenges faced by the market.

In conclusion, the proposed car price prediction system has the potential to be a transformative tool for the Afghan automotive industry. By providing accurate and transparent pricing information, it empowers market participants, promotes fair competition, and generates crucial insights that can drive the growth and development of this vital sector. The positive impacts of this system can extend well beyond the automotive industry, contributing to the overall economic progress of Afghanistan.

## Chapter 2: Related Work (Literature Review)

---

While making this project we examined many of research papers, these papers disclose the newest research during this field, a summary of some of the papers we referred to are mentioned below:

Phani Krishna Kondeti implemented various machine learning techniques to develop a model for car price estimation. The dataset used by the authors consisted of 1209 entries related to the prices and attributes of pre-owned cars. It was obtained via scraping methods applied on an online marketplace from Bangladesh. Here, again, data were explored and preprocessed to address issues related to outliers, missing data, unrepresentative samples, the lack of numerical representations of text attributes, multicollinearity, and different measurement scales. This resulted in using 9 of the 10 car attributes present in the initial dataset, namely transmission, fuel type, brand, car model, model year, car shape, engine capacity, mileage, and price. Of the five implemented regression models, extreme gradient boosting was declared most suitable for car price prediction, with an  $R^2$  of 0.91, closely followed by the random forest classifier with an  $R^2$  of 0.90. However, random forests scored better in terms of the mean average error. [1]

Nabarun Pal, Dhanasekar Sundararaman, Priya Arora, Puneet Kohli and Sai Sumanth Palakurthy During this paper, Authors have used supervised learning method namely Random Forest to predict the costs of used cars. The model has been chosen after careful exploratory data analysis to work out the impact of every feature on price. A Random Forest with 500 Decision Trees were created to train the data. From experimental results, the training accuracy was discovered to be 95.82%, and therefore the testing accuracy was 83.63%. The model can predict the price value of cars accurately by choosing the fore most correlated features. [2]

Rechar R. Yang studied the problem of car price prediction from images by employing multiple classic machine learning techniques and deep learning models such as convolutional neural networks. They constructed a dataset consisting of 1400 images of front angular views of different cars, with prices ranging from USD 12,000 to USD 2,000,000. They developed initial baselines for price regression based on linear regression models that took as their input HOG features or features extracted from pre-trained CNNs. Moreover, they created a classification task by splitting the data into price intervals and training the models to predict the price class. The researchers allocated class segments to each example by employing price cutoffs that aligned with specific percentiles of price distribution (20th, 40th, 60th, 80th, and 100th percentiles) to predict car prices. Their baseline consisted of a support vector machine classifier for this task. They further analyzed the performance of CNN models such as SqueezeNet and VGG-16, along with a custom architecture, PriceNet, which built upon SqueezeNet by adding residual connections between modules and batch normalization. The PriceNet architecture achieved the best performance for all metrics, obtaining an RMSE of 11,587.05, an MAE of 5051.61, an  $R^2$  score of 0.98 for the regression task, and an F1 score of 0.88 for classification. [3]

Sameerchand Pudaruth, The Author has researched 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, naive bayes and decision trees are accustomed to make the predictions. The predictions are then evaluated and compared so as to search out those which offer the most effective performances. A seemingly easy problem clothed to be indeed very difficult to resolve with high accuracy. All the four methods which are used provided comparable performance. [4]

Pattabiraman Venkatasubbu, Mukkesh Ganesh in this research, the authors attempt to construct a statistical model that would estimate the price of a used car based on previous customer data and a collection of attributes using Algorithms such as Lasso, Multiple regression and Regression Trees. The authors have also analysed the forecast accuracy of different models in order to calculate the car's price using an algorithm that is more accurate. [5]

Enis Gegic, Becir Isakovic, Dino Keco, Zerina Masetic, Jasmin Kevric During this paper, authors have first considered number of distinct attributes are examined for the reliable and accurate prediction. They built a model for predicting the price value of used cars, using three machine learning techniques (Artificial Neural Network, Support Vector Machine and Random Forest). The authors have compared Respective performances of various algorithms to search out one that best that most closely fits the available data set. The ultimate prediction model was integrated into Java application. Furthermore, the model was evaluated using test data and also the accuracy of 87.38% was obtained. [6]

Baoyang Cui introduced an innovative framework for price regression, employing a combination of two gradient-boosting techniques and a deep residual network. The authors conducted experiments on a dataset comprising more than 30,000 samples, taking into account over 20 features, including the most frequently used features like the car brand, mileage, age, and fuel type. The neural network processed the input features, generating an optimized representation of the attribute characteristics. This representation and the initial prediction served as the input for an XGBoost module, which iteratively predicted the price by incorporating the predicted price from the previous iteration and the initial features. To further enhance the results, a LightGBM framework was employed, utilizing the preceding prediction and initial features to retrain the representations iteratively until performance improvement plateaued. The proposed evaluation metric, which combined the mean absolute percentage error (MAPE) and accuracy, yielded a score of 75 out of 100. [7]

K.Samruddhi, Dr. R.Ashok Kumar During this Authors proposed a supervised machine learning model using KNN (K Nearest Neighbor) regression algorithm to research the price value of used cars. Authors have trained the model with data of used cars which is collected from the Kaggle website. Through this experiment, the information was examined with different trained and test ratios. As a result, the accuracy of the proposed model is around 85% and is fitted because the optimized model. [8]

Car price prediction is a crucial aspect of the automotive industry, as it helps both buyers and sellers make informed decisions. Various machine learning and statistical models have been developed to predict car prices based on different features and attributes. For instance, a study by Sanika Singh utilized a random forest regression model to predict car prices based on factors such as mileage, age, and brand. The results showed that the model was able to accurately predict car prices with a high degree of accuracy. [9]

Nabarun Pal developed a model for car price prediction using a random forest classifier. Their dataset comprised 370,000 German eBay entries related to the prices and attributes of used cars. The data preprocessing and exploration procedure resulted in using only 10 out of the 20 car attributes from the initial dataset, namely: car shape, brand, model, age, mileage, engine power, type of fuel, transmission, whether the car was damaged and repaired or not, and price. Following model training and testing, the authors obtained an  $R^2$  of 0.83 on the validation data, with price, kilometers, brand, and vehicle type being the most relevant features. [10]

Andreea Dutulescu studied several approaches in terms of price prediction, employing baseline models such as XGBoost and experimenting with deep neural networks to better aggregate car features. They constructed a dataset of 25,000 ads from a Romanian website that advertised used cars. The features used in the prediction were the brand; model; year of manufacture; mileage; fuel; engine capacity; transmission; and a list of add-ons, which were extra components of cars that customers could opt to include on their cars. The employed neural networks learned embeddings for the car model to better represent this feature, and several experiments were performed for add-on representation. Add-ons were represented as their total count, hot-encoded with a dense projection, or encoded as trainable embeddings with and without a self-attention layer. Moreover, a pre-trained RoBERTa model was employed on the text descriptions of the add-ons to capture the linguistic meaning of these options. The best scores of 95.47  $R^2$  and 10.68% mean percentage error were obtained by the neural network that employed learned embeddings on add-ons. [11]

Prashant Gajera, Akshay Gondaliya and Jenish Kavathiya used a dataset consisting of 92,386 records to train multiple regression techniques such as KNN regression, random forest, linear regression, decision trees, and XGBoost. Each sample contained information about mileage, the year of registration, fuel type, car make, model, and gear type. The random forest regression model obtained the lowest error rate and achieved an RMSE of 3702.34, followed by the XGBoost model, which obtained an RMSE of 3980.77. [12]

Venkatasubbu and Ganesh experimented with different supervised regression techniques for used car price prediction and studied which variables were most predictive for this task. They considered the dataset introduced by Kuiper [14], which contained a total of 804 sample cars with annotations for mileage, make, model, trim, body type, cylinder, liters, doors, cruise, sound, leather seats, and price. The authors trained models for lasso regression, multiple linear regression, and regression trees on a training set consisting of 563 records, leaving the rest of the samples for testing. The multiple regression model obtained the lowest error rate of 3.468%, the regression tree obtained an error rate of 3.512%, and the lasso regressor obtained an error rate of 3.581%. [13]

Bitvai and Cohn predict movie's box office revenue based on movie critics' reviews and structured attributes about the movie. The architecture is based on CNNs for text, but trained in a regression task. According to the reported results, one of the primary sources of improvements comes from using non-linear activation functions, which significantly increases the performance compared to the simple linear regression baseline models. Introducing meta-information about the movie (the structured attributes) and domain information shows no significant improvement, possibly due to its complementary nature [14].

Although there are previous solutions that work with tabular data and text-sequences in neural networks such as Luo et al. to the best of our knowledge, no work in the literature proposes a method for jointly exploiting information in both modalities, other than simple feature concatenation. Our model's feature is constructed from representation-level interactions between text and tabular categorical data [15].

Kanwal Noor and Sadaqat Jan proposed Vehicle Price Prediction System using Machine Learning Techniques. In this paper, they proposed a model to predict the price of the cars through multiple linear regression method. They selected the most influencing feature and removed the rest by performing feature selection technique. The Proposed model achieved the prediction precision of about 98% [16].

K.Samruddhi, Dr. R.Ashok Kumar During this paper Authors proposed a supervised machine learning model using KNN (K Nearest Neighbor) regression algorithm to research the price value of used cars. Authors have trained the model with data of used cars which is collected from the Kaggle website. Through this experiment, the information was examined with different trained and test ratios. As a result, the accuracy of the proposed model is around 85% and is fitted because the optimized model.

Nabarun Pal, Dhanasekar Sundararaman, Priya Arora, Puneet Kohli, Sai Sumanth Palakurthy During this paper, Authors have used supervised learning method namely Random Forest to predict the costs of used cars. The model has been chosen after careful exploratory data analysis to work out the impact of every feature on price. A Random Forest with 500 Decision Trees were created to train the data. From experimental results, the training accuracy was discovered to be 95.82%, and therefore the testing accuracy was 83.63%. The model can predict the price value of cars accurately by choosing the fore most correlated features [17].

## Chapter 3: Research Methodology

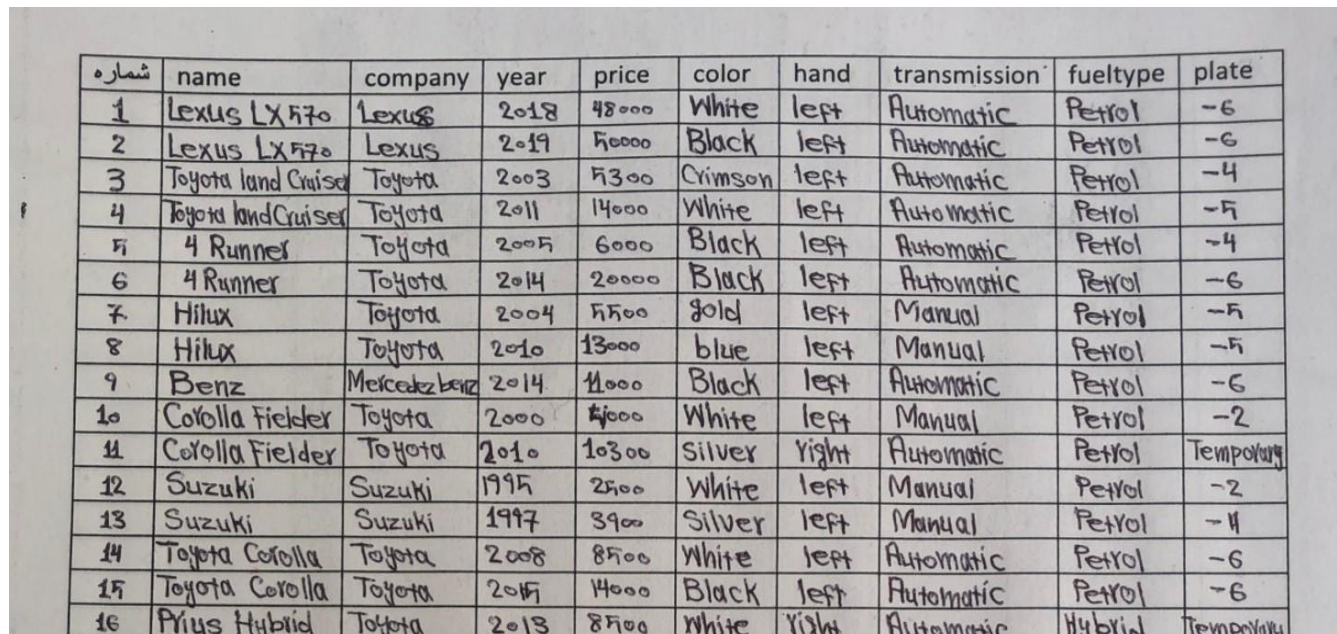
### 3.1 Data Gathering

I gathered all of our data in paper sheets from different car dealerships around the capital city, Kabul. Oryakhail dealership located in 5th district Company, Setare Shahr & Said Mohammad Hashimi dealerships located in 3rd district dehbori, & some other ones are the sources we collected data from. Thanks from Ahmad Massiullah & Mohammad Qaim for their contribution in collecting these data. Totally, 1038 rows data of different car models is collected & all of them are classified in 10 key features; manufacturing company, model, year, plate type, seat hand, transmission type, fuel type, engine type, color & price.

This handful amount of dataset will give the model a better understanding of the whole data which then leads to an accurate and precise prediction.

### 3.2 Data Conversion

Although changing the handwritten data into a digital work is not that challenging & may seem worthless mentioning here, it still consumed a huge amount of our time. Majority of data in a country like Afghanistan is handwritten in papers and for obvious reasons, people have no other choice of manually transcribing it on excel sheets. Our data contained 1038 rows of raw data considering 10 related features each having a key role in changing the value of a specific car.



شماره	name	company	year	price	color	hand	transmission	fueltype	plate
1	Lexus LX570	Lexus	2018	48000	White	left	Automatic	Petrol	-6
2	Lexus LX570	Lexus	2019	50000	Black	left	Automatic	Petrol	-6
3	Toyota land Cruiser	Toyota	2003	5300	Crimson	left	Automatic	Petrol	-4
4	Toyota land Cruiser	Toyota	2011	14000	White	left	Automatic	Petrol	-5
5	4 Runner	Toyota	2005	6000	Black	left	Automatic	Petrol	-4
6	4 Runner	Toyota	2014	20000	Black	left	Automatic	Petrol	-6
7	Hilux	Toyota	2004	5500	gold	left	Manual	Petrol	-5
8	Hilux	Toyota	2010	13000	blue	left	Manual	Petrol	-5
9	Benz	Mercedes benz	2014	11000	Black	left	Automatic	Petrol	-6
10	Corolla Fielder	Toyota	2000	4000	White	left	Manual	Petrol	-2
11	Corolla Fielder	Toyota	2010	10300	Silver	right	Automatic	Petrol	Temporary
12	Suzuki	Suzuki	1995	2500	White	left	Manual	Petrol	-2
13	Suzuki	Suzuki	1997	3900	Silver	left	Manual	Petrol	-11
14	Toyota Corolla	Toyota	2008	8500	White	left	Automatic	Petrol	-6
15	Toyota Corolla	Toyota	2015	14000	Black	left	Automatic	Petrol	-6
16	Prius Hybrid	Toyota	2013	8500	White	right	Automatic	Hybrid	Temporary

Figure 3. 1: Handwritten Data Collection

### 3.3 General Dataset

	name	company	year	price(\$)	color	hand	engine(saland)	transmission	fuel_type
0	lexus lx570	lexus	2007	30000	white	left	8 saland	automatic	petrol
1	lexus lx570	lexus	2007	28000	black	left	8 saland	automatic	petrol
2	lexus lx570	lexus	2007	27000	crimson	left	8 saland	automatic	petrol
3	lexus lx570	lexus	2008	32000	white	left	8 saland	automatic	petrol
4	lexus lx570	lexus	2008	30000	black	left	8 saland	automatic	petrol
...	...	...	...	...	...	...	...	...	...
509	toyota townace	toyota	1994	1500	navy blue	left	4 saland	automatic	diesel
510	toyota townace	toyota	1995	2000	gold	left	4 saland	automatic	diesel
511	toyota townace	toyota	1995	1800	navy blue	left	4 saland	automatic	diesel
512	toyota townace	toyota	1996	2000	navy blue	left	4 saland	automatic	diesel
513	toyota townace	toyota	1997	2200	navy blue	left	4 saland	automatic	diesel

Figure 3. 2: **General Dataset**

### 3.4 Analyzing Cars characteristics and fuel-types

This image presents a detailed comparison of various car models and their characteristics, including the type of hand and transmission (left, right, automatic, manual) as well as the fuel type (petrol, diesel, hybrid). The data visualized in this chart provides valuable insights into the diverse options available to car buyers and the trends within the automotive industry.

By examining the distribution of car models across the different fuel types and transmission configurations, we can observe patterns and preferences that may be influenced by factors such as fuel efficiency, performance, and consumer demand. For instance, the prevalence of hybrid models for certain car brands suggests a growing emphasis on eco-friendly and technology-driven solutions in the market.

Furthermore, the positioning of various car makes and models within the chart allows for comparisons and identification of unique characteristics that may appeal to specific customer segments. This information can be valuable for both consumers in their decision-making process and manufacturers in developing targeted product strategies to cater to evolving market needs

Overall, this visual representation of car features and fuel types offers a comprehensive overview of the automotive landscape, enabling informed discussions and data-driven insights to support the development and adoption of innovative vehicle technologies and consumer preferences.

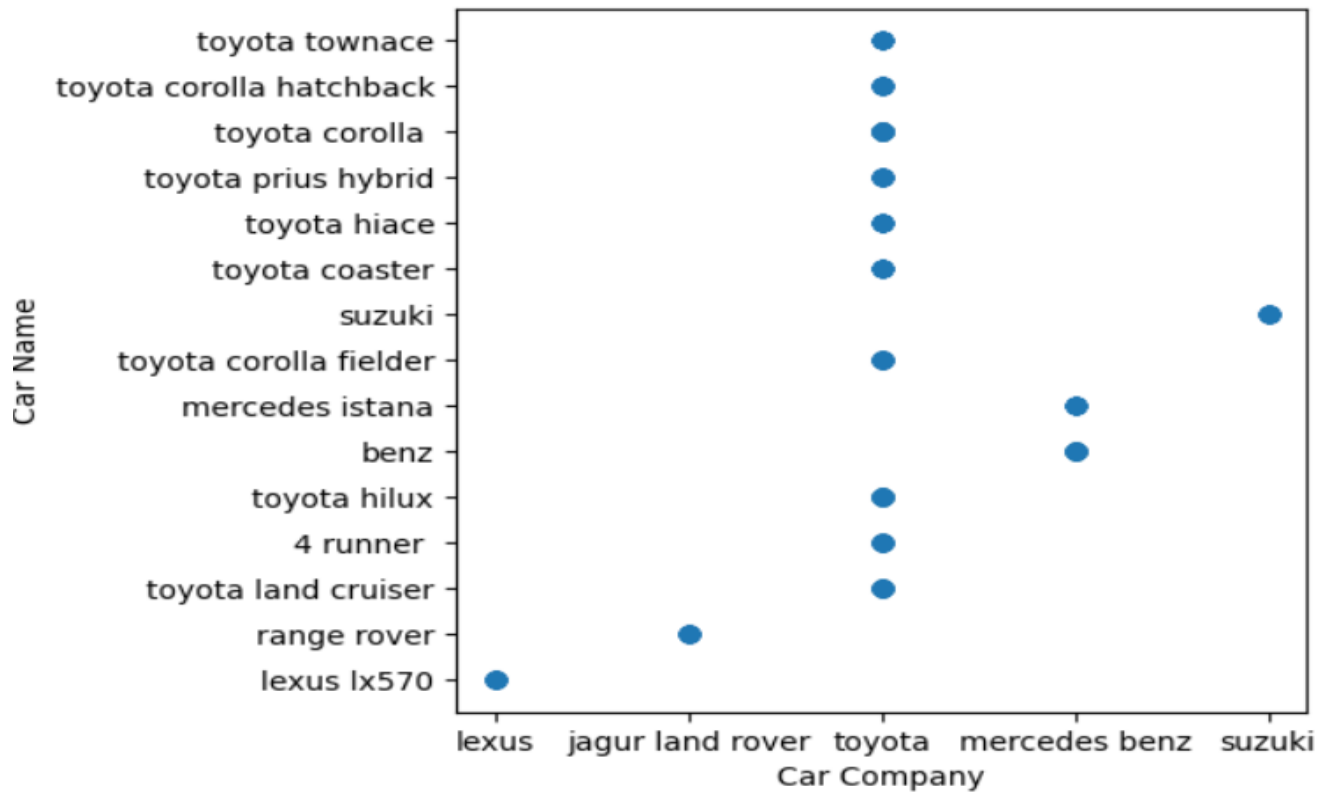


Figure 3. 3: Relationship between car models and companies

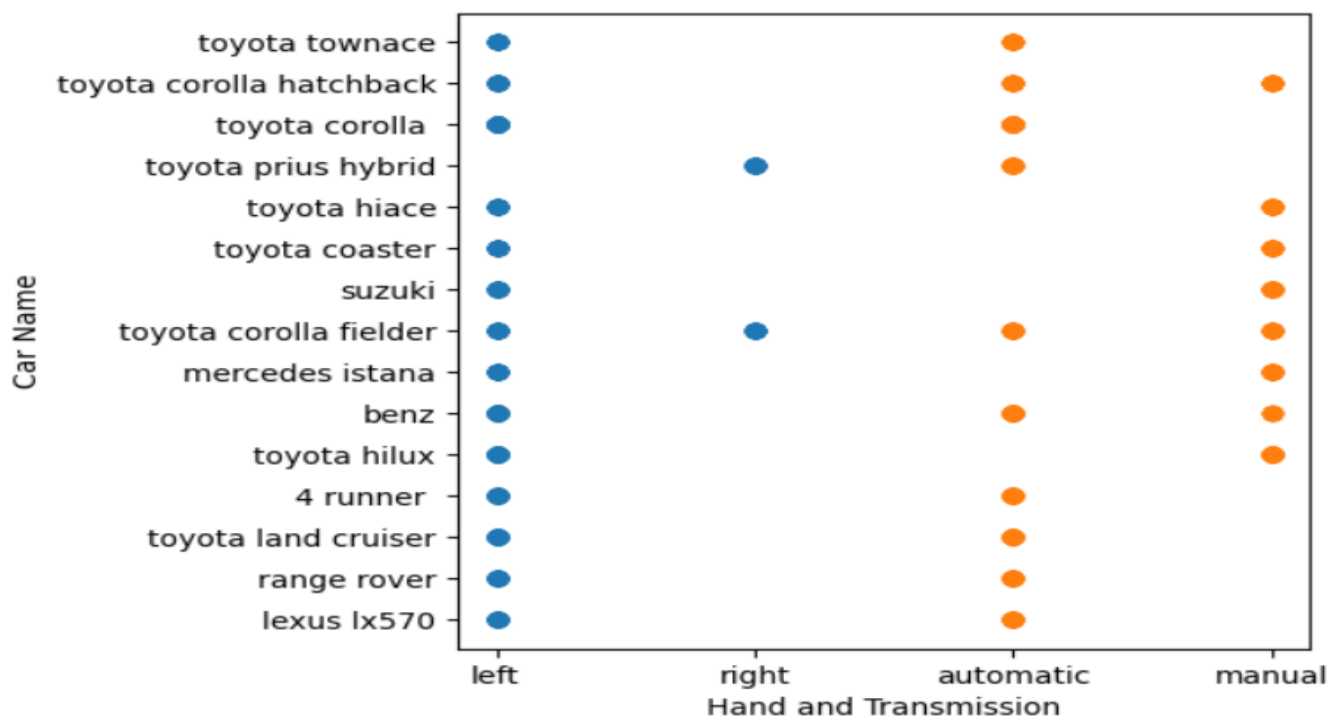


Figure 3. 4: Relationship between car models, Hand type and Transmission



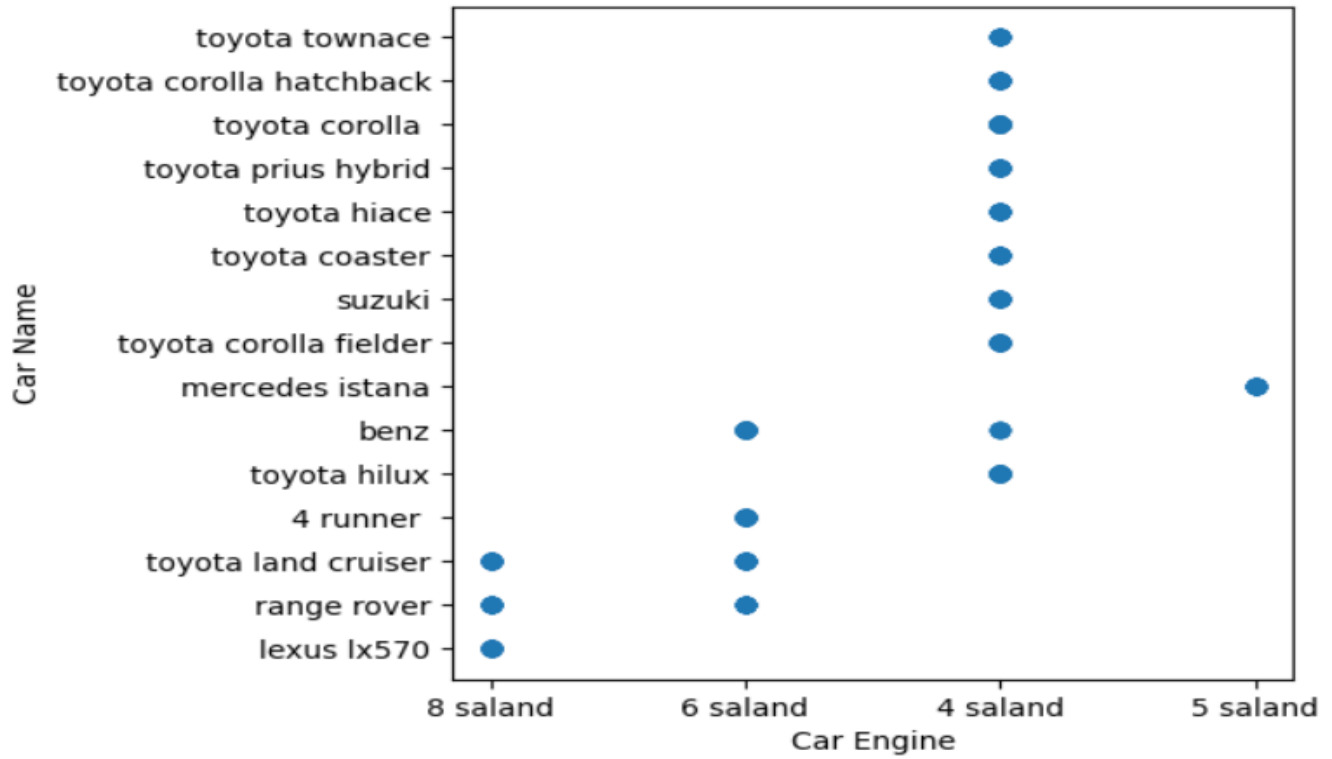


Figure 3. 5: Relationship between car models and Car Engine Type

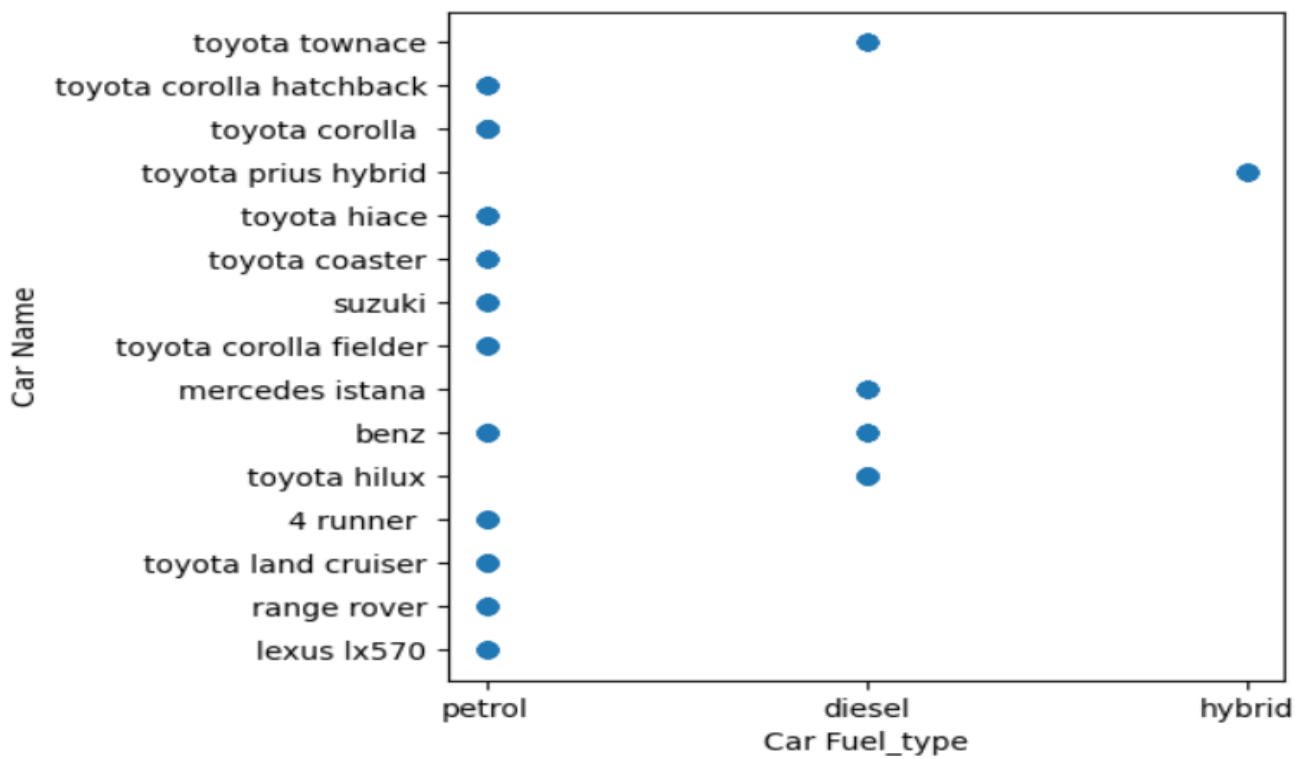


Figure 3. 6: Relationship between car models and Car Fuel\_type

### 3.4 Preprocessing

Preprocessing data removes missing or inconsistent data values resulting from human or computer error, which can improve the accuracy and quality of a dataset, making it more reliable. So in order to make data more consistent, before training the machine learning model, pre-processing steps were performed on the dataset to ensure that the data were properly prepared for analysis. The procedures that applied to the dataset in this step are shown in the following figure.

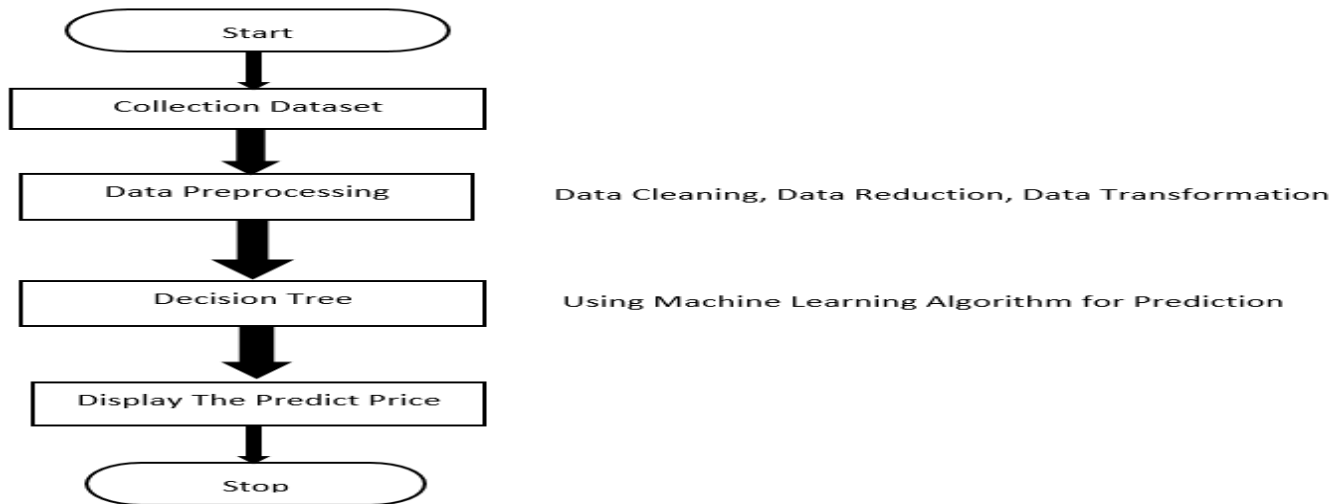


Figure 3. 7: **Data Preprocessing**

### 3.5 Splitting the Dataset

An adequate amount of data should be classified in order to make the model work & there isn't any better option than the train-test split procedure.

The train-test split procedure is used to estimate the performance of machine learning algorithms when they are used to make predictions on data not used to train the model. It is a fast and easy procedure to perform, the results of which allow you to compare the performance of machine learning algorithms for your predictive modeling problem.

Using the "train\_test\_split" function from scikit-learn library, the dataset was split into training and testing sets. This will allow us to evaluate the model performance effectively.

80% of the data is used in data training ensuring accurate results. The rest 20% portion of the data is used for testing, considering a fair amount of data for both training and evaluation purposes.

Splitting the dataset into training and testing sets enabled the evaluation of the model's generalization capabilities.

After all, the preprocessing stage played a vital role in getting the dataset ready for training and evaluation stages.

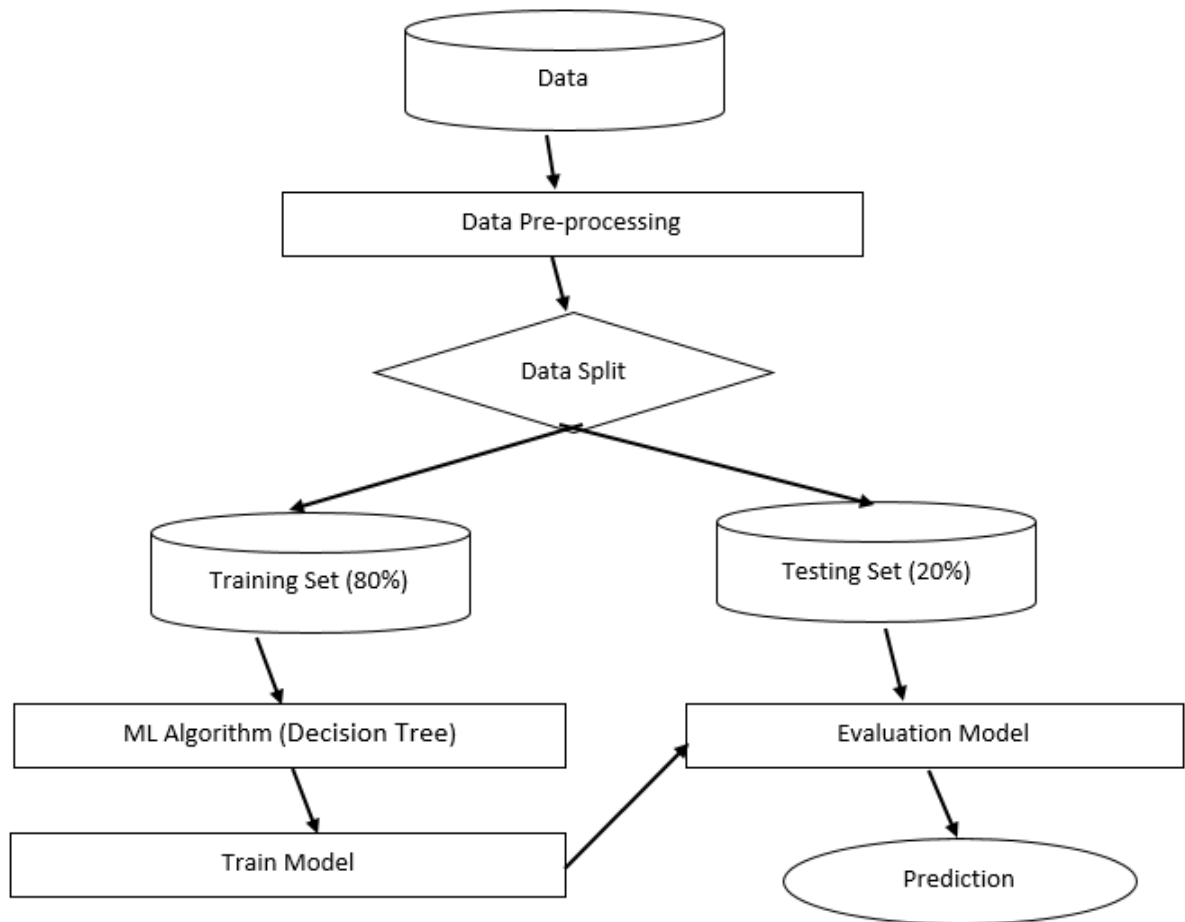


Figure 3. 8: **Splitting The Dataset and Evaluation Model**

### 3.6 Algorithm Selection and Evaluation

In this machine learning project, I utilized three different algorithms to train and evaluate model. These algorithms were chosen based on their suitability for the task at hand and their potential to deliver accurate predictions. The algorithms used in this project are as follows:

#### 3.6.1 Linear Regression

Linear regression is an algorithm that provides a linear relationship between an independent variable and a dependent variable to predict the outcome of future events. [18]

```
# Training Linear Regression model
linear_regression_model = LinearRegression()
linear_regression_model.fit(X_train, y_train)
linear_regression_predictions = linear_regression_model.predict(X_test)
linear_regression_accuracy = r2_score(y_test, linear_regression_predictions)
print("Linear Regression R2_Score:", linear_regression_accuracy)

Linear Regression R2_Score: 0.5438551627409371
```

Figure 3. 9: **Linear Regression Model**

This code is training a linear regression model using the `LinearRegression()` function from the `scikit-learn` library in Python. Here's a breakdown of what the code does:

- 1- `linear_regression_model = LinearRegression()`: This line creates a new instance of the `LinearRegression()` model.
- 2- `linear_regression_model.fit(X_train, y_train)`: This line fits the linear regression model to the training data, where `X_train` is the feature matrix and `y_train` is the target variable.
- 3- `linear_regression_predictions = linear_regression_model.predict(X_test)`: This line uses the trained model to make predictions on the test data, which is stored in the `linear_regression_predictions` variable.
- 4- `linear_regression_accuracy = r2_score(y_test, linear_regression_predictions)`: This line calculates the R-squared score, which is a measure of the model's accuracy. The R-squared score is stored in the `linear_regression_accuracy` variable.
- 5- `print("Linear Regression R2_Score:", linear_regression_accuracy)`: This line prints the R-squared score of the linear regression model.

### 3.6.2 K Neighbors Regressor

It is a machine learning algorithm used for regression tasks, it is based on the k-nearest neighbors (KNN) algorithm? In KNN, the algorithm predicts the target value for a new data point by considering the values of its k nearest neighbors in the feature space. [19]

```
# Training KNR model
knr_model = KNeighborsRegressor()
knr_model.fit(X_train, y_train)
knr_predictions = knr_model.predict(X_test)
knr_accuracy = r2_score(y_test, knr_predictions)
print("The KNR R2_Score:", knr_accuracy)

The KNR R2_Score: 0.7652042643499805
```

Figure 3. 10: **K-Neighbors Regressor**

This code is training a KNR (K-Nearest Neighbors Regressor) model using the scikit-learn library in Python. Here's a breakdown of what the code does:

- 1- `knr_model = KNeighborsRegressor()`: This line creates a new instance of the `KNeighborsRegressor()` model.
- 2- `knr_model.fit(X_train, y_train)`: This line fits the KNR model to the training data where `X_train` is the feature matrix and `y_train` is the target variable.
- 3- `knr_predictions = knr_model.predict(X_test)`: This line uses the trained model to make predictions on the test data, which is stored in the `knr_predictions` variable.
- 4- `knr_accuracy = r2_score(y_test, knr_predictions)`: This line calculates the R-squared score, which is a measure of the model's accuracy. The R-squared score is stored in the `knr_accuracy` variable.
- 5- `print("The KNR R2_Score:", knr_accuracy)`: This line prints the R-squared score of the KNR model.

### 3.6.3 Decision Tree Regressor

The Decision Tree Regressor is a supervised machine learning algorithm used for regression tasks. It is based on the Decision Tree algorithm, which builds a tree-like model of decisions and their possible consequences. [20]

```
# Training Decision Tree Regressor model
decision_tree_regressor_model = DecisionTreeRegressor()
decision_tree_regressor_model.fit(X_train, y_train)
decision_tree_regressor_predictions = decision_tree_regressor_model.predict(X_test)
decision_tree_regressor_accuracy = r2_score(y_test, decision_tree_regressor_predictions)
print("Decision Tree Regressor R2_Score:", decision_tree_regressor_accuracy)

Decision Tree Regressor R2_Score: 0.9909850868394865
```

Figure 3. 11: **Decision Tree Regressor**

This code is training a Decision Tree Regressor model using the scikit-learn library in Python. Here's a breakdown of what the code does:

- 1- `decision_tree_regressor_model = DecisionTreeRegressor()`: This line creates a new instance of the `DecisionTreeRegressor()` model.
- 2- `decision_tree_regressor_model.fit(X_train, y_train)`: This line fits the Decision Tree Regressor model to the training data, where `X_train` is the feature matrix and `y_train` is the target variable.
- 3- `decision_tree_regressor_predictions = decision_tree_regressor_model.predict(X_test)`: This line uses the trained model to make predictions on the test data, which is stored in the `decision_tree_regressor_predictions` variable.
- 4- `decision_tree_regressor_accuracy = r2_score(y_test, decision_tree_regressor_predictions)`: This line calculates the R-squared score, which is a measure of the model's accuracy. The R-squared score is stored in the `decision_tree_regressor_accuracy` variable.
- 5- `print("Decision Tree Regressor R2 Score:", decision_tree_regressor_accuracy)`: This line prints the R-squared score of the Decision Tree Regressor model.

To determine the most suitable algorithm for my project, I evaluated their performance using `R2_score`. This metric provides their ability to correctly identify positive instances.

After comprehensive evaluation, the Decision Tree algorithm emerged as the optimal choice for my project. Decision Tree demonstrated high `R2_score`, making it well-suited for accurate predictions and reliable results.

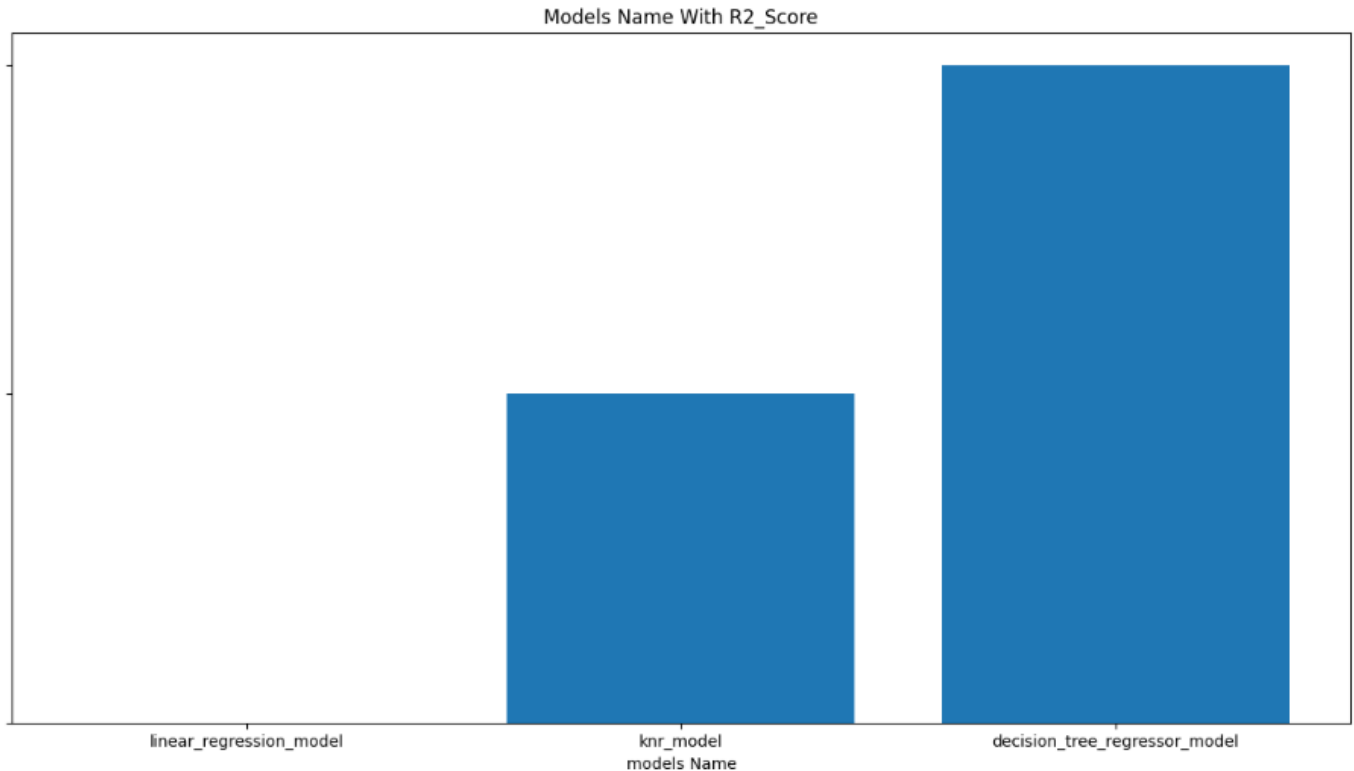


Figure 3. 12 : **Model Evaluation**

In this graph is to compare the performance of three different regression models - linear regression, K-Nearest Neighbors Regressor (KNN), and Decision Tree Regressor - by looking at their respective R-squared scores.

The Decision Tree Regressor model has the highest R-squared score among the three models, indicating that it has the best predictive performance on the given dataset. This suggests that the relationship between the features and the target variable in the dataset is likely complex and non-linear, and the Decision Tree Regressor is able to capture this complexity better than the linear regression or KNN models.

Some key advantages of the Decision Tree Regressor that make it a suitable choice in this case:

- 1- Ability to handle non-linear relationships: Decision trees can model non-linear patterns in the data, which is often necessary when dealing with complex real-world datasets.
- 2- Interpretability: The tree-like structure of Decision Tree Regressor makes it easier to understand how the model is making predictions, which can be valuable for gaining insights into the underlying data.

- 3- Robustness to outliers: Decision trees are generally less sensitive to outliers in the data compared to some other regression models, which can be an important consideration when working with real-world datasets.
- 4- Handling mixed data types: Decision trees can work with both numerical and categorical features, making them a versatile choice for a wide range of regression problems.

### 3.7 Training the Decision Tree Regressor

A Decision Tree Regressor was created using the scikit-learn library's. The regression was configured at a random state of 42 (`random_state=42`). The preprocessed training data was used to train the Regressor by calling the `fit ()` method on the Decision Tree Regressor object.

### 3.8 Making Prediction on the Testing Set

Predictions were made on the preprocessed testing data using the trained Decision Tree Regressor. The `predict()` method was used, passing the testing input features (`X_test`), which returned the predicted labels for the corresponding instances. The predicted labels were stored in the `y_pred` variable.

### 3.9 Evaluating the Models Performance

The performance of the trained model was assessed using `R2_score` metric calculated based on the predicted labels (`y_pred`) and the actual labels from the testing set (`y_test`). The following evaluation metrics were computed:

**`R2_score`:** is a measure that provides information about the goodness of fit of a model. In the context of regression it is a statistical measure of how well the regression line approximates the actual data. [21]



## Chapter 4: Implementation Details

---

The implementation of the Car Price Prediction involves the following steps:

**4.1 Dataset Loading:** The dataset is loaded into the program using the `pandas` library's "read.csv" function. This step ensures that the necessary data is available for training the model.

The code loads a CSV dataset file named "full\_car\_data\_and\_price" using `pd.read_csv()` from pandas library and assigns it to the variable `dataset`.

The `dataset.head()` function is called to display the first few rows of the dataset.

```
## Load the data
dataset = pd.read_csv("full_car_data_and_price.csv")
dataset.head()
```

	name	company	year	price(\$)	color	hand	engine(saland)	transmission	fuel_type
0	lexus lx570	lexus	2007	30000	white	left	8 saland	automatic	petrol
1	lexus lx570	lexus	2007	28000	black	left	8 saland	automatic	petrol
2	lexus lx570	lexus	2007	27000	crimson	left	8 saland	automatic	petrol
3	lexus lx570	lexus	2008	32000	white	left	8 saland	automatic	petrol
4	lexus lx570	lexus	2008	30000	black	left	8 saland	automatic	petrol

Figure 4. 1: **Dataset Loading**

### 4.2 Splitting The Dataset:

The dataset is split into input features and labels.

The input features are extracted from the dataset, which selects all rows and all columns except the last column. They are stored in the variable `X`.

The labels are extracted, which selects all rows and the last column. They are stored in the variable `y`.

**4.3 Feature Extraction:** The dataset is split into input features (X) and labels (y). The input features are selected from the dataset, excluding the target variable. The labels represent the target variable.

```
# Splitting data into training and testing sets
X = car.drop(columns='price($)')
y = car['price($)']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

Figure 4. 2: **Feature Extraction**

#### 4.5 Model Training:

The Random Forest classifier is utilized as the machine learning model for the Emergency Situation Predictor. The `RandomForestClassifier` class from the `scikit-learn` library is used to create an instance of the classifier. The model is then trained on the preprocessed input features (X) and labels (y).

```
# Training Decision Tree Regressor model
decision_tree_regressor_model = DecisionTreeRegressor()
decision_tree_regressor_model.fit(X_train, y_train)
decision_tree_regressor_predictions = decision_tree_regressor_model.predict(X_test)
decision_tree_regressor_accuracy = r2_score(y_test, decision_tree_regressor_predictions)
print("Decision Tree Regressor R2_Score:", decision_tree_regressor_accuracy)
```

Figure 4. 3: **Training Decision Tree Regressor Model**

#### Creating and Training the Decision Tree Regressor:

- A Decision Tree Regressor is created with **DecisionTreeRegressor()** from scikit-learn.
- The fit() method is called on the Decision Tree Regressor to train it using the standardized input features (X\_train) and corresponding labels (y\_train).

#### 4.6 Prediction Variable:

The new\_data variable is defined to make predictions based on new input data. This variable takes the input features as arguments, applies the necessary preprocessing steps and uses the trained Decision Tree Regressor to predict Car Price Prediction

```
# Example prediction using the loaded Model
new_data = pd.DataFrame([[0, 0, 2009, 1, 2, 1, 0, 0]],
columns=['name', 'company', 'year', 'color', 'hand', 'engine(saland)', 'transmission', 'fuel_type'])
prediction = loaded_model.predict(new_data)
print("Prediction:", prediction)
```

Figure 4. 4: **Prediction State**

## 4.8 Interface Development

A simple web-based interface was created to allow users to input relevant information and obtain predictions. The interface was developed using Flask, a Python web framework, which facilitated the integration of the trained Decision Tree Regressor model with the interface.

← → ↻ ⓘ localhost:5000 ☆ 🧑

# 🚗 Car Price Prediction in Afghanistan

Select Car Model  
Select Car Model ▼

Select Car Company  
Toyota ▼

Select Model Year  
2000 ▼

Select Model Color  
Black ▼

Select Model HandType  
Left ▼

Select Model EngineType  
4 Saland ▼

Select Model Transmission Type  
Automatic ▼

Select Model Fuel\_Type  
Petrol ▼

Price Predict

\$

Figure 4. 5: **Web Interface**

## 4.9 The Project Code Consist the Following Components:

4.9.1 `application.py`: The main Python script that handles the Flask application setup, routes, and prediction logic.

4.9.2 `index.html`: The HTML template for the main page interface, where users input the required information.

4.9.3 `result.html`: The HTML template for the result page, where the predicted result is displayed. The `application.py` script loads the dataset, performs preprocessing, trains the Decision Tree Regressor, and defines the necessary routes for the application.

The `predict_Car_Price` function is responsible for utilizing the trained model to make predictions based on the provided input features.

## 4.10 Getting Started

To run the Car Price Prediction project locally, follow these steps:

Ensure you have Python installed on your machine. Install the required libraries by running `pip install flask, pandas, numpy, and scikit-learn` in your command line. Place the dataset file (`full_car_data_and_price.csv`) in the same directory as the `application.py` script. Update the file paths in the code to match the actual location of the dataset and HTML files. Run the `application.py` script using the command `python application.py`. Access the application in your web browser at

**`http://localhost:8080`**

The Car Price Prediction is implemented using Python programming language and relies on the following libraries:

4.10.1 **Pandas**: A powerful data manipulation library used for loading and preprocessing the dataset.

4.10.2 **numpy**: A fundamental package for scientific computing with Python, used for numerical operations and array manipulation.

4.10.3 **scikit-learn**: A machine learning library that provides tools for data preprocessing, model training, and evaluation.

4.10.4 **Flask**: Flask is a popular web framework for building web applications and APIs using the Python programming language. It is known for its simplicity, flexibility, and ease of use. Flask is categorized as a micro framework because it provides only the essential features necessary to build web applications, allowing developers to have more control and flexibility in designing their applications.

## Chapter 5: Results

---

The car price prediction project using a Decision Tree model yielded impressive results and outcomes. By utilizing features such as manufacturing company, model, year, plate, seat hand, transmission type, fuel type, engine type, color, and price, the model was able to accurately predict the prices of various cars with a high degree of accuracy. The Decision tree model was able to effectively analyze the relationships between these features and the price of the car, allowing for precise predictions to be made.

The results of the project showed that the Decision tree model was able to accurately predict car prices based on the input features. This allowed for more informed decision-making when it came to buying or selling cars. The model was able to take into account various factors in order to make accurate predictions about the price of a car.

Furthermore, the outcome of this project has significant implications for consumers in the Afghan Automotive Industry. Having access to accurate car price predictions can help them make more informed decisions when purchasing a new or used vehicle. They can use this information to negotiate better deals and ensure that they are getting a fair price for their purchase.

Overall, the results and outcome of this car price prediction project using a Decision tree model have demonstrated its effectiveness in accurately predicting car prices based on a variety of input features. This has significant implications for both consumers and manufacturers (when it goes global) in the automotive industry and has the potential to improve decision-making processes related to buying and selling cars.

### 5.1 Limitation

**1.Change in Model's R2-score:** The car price prediction model developed in this project may face potential challenges in maintaining its accuracy over time. There are several factors that could contribute to changes in the model's predictive performance in the future:

**Market Fluctuations:** The automotive market in Afghanistan is subject to various economic and market forces, such as changes in supply and demand, inflation, currency exchange rates, and import/export regulations. These market fluctuations can lead to shifts in car prices that may not be fully captured by the initial model training data. As a result, the model's accuracy in predicting prices could diminish over time as the market conditions evolve.

**Changes in Consumer Preferences:** Consumer preferences for certain car models, features, and attributes can change over time. Factors such as the introduction of new technologies, shifts in environmental concerns, or changes in socioeconomic demographics may alter the way consumers value and prioritize different car characteristics. These changes in consumer behavior can impact the relationship between the model's input features and the target car prices, potentially reducing the model's predictive accuracy.

**Unforeseen Events:** Unexpected events, such as natural disasters, political instability, or global disruptions, can have significant impacts on the automotive industry and car prices. These unforeseen circumstances may not be adequately represented in the historical data used to train the model, leading to inaccurate predictions when such events occur.

To address these challenges and maintain the model's accuracy, regular model monitoring and retraining may be necessary. The model should be continuously evaluated and updated with the latest market data to ensure it remains responsive to the evolving conditions in the Afghan automotive sector.

Additionally, incorporating mechanisms for feedback and error reporting from users can help identify areas where the model's predictions are diverging from the actual market prices. This information can then be used to refine and enhance the model, ensuring that it continues to provide reliable and accurate car price estimates for consumers and industry stakeholders.

By proactively addressing the potential changes in the model's accuracy, the car price prediction system can maintain its effectiveness and continue to contribute to the overall efficiency and transparency of the Afghan automotive market.

**2. Dataset Scope:** You raise a valid concern regarding the scope and representativeness of the dataset used to train the car price prediction model. The factors affecting car prices can vary significantly across different regions and markets, and the dataset used in this project may not fully capture the nuances of the South Asian automotive market, particularly in Afghanistan.

Some key considerations regarding the dataset scope include:

**1. Regional Differences:**

- Car prices and the factors influencing them can differ between Afghanistan and other global markets due to factors such as:
  - Localized consumer preferences and buying behavior
  - Variations in government regulations, taxes, and documentation requirements
  - Differences in the availability and pricing of specific car makes, models, and features
  - Unique market dynamics and economic conditions within the Afghan automotive industry

**2. Data Representativeness:**

- The dataset used to train the model may not be sufficiently representative of the entire Afghan car market, potentially skewing the model's understanding of price determinants.
- The dataset may be biased towards certain car segments, brands, or geographic regions within Afghanistan, leading to inaccurate predictions for other market segments.

**3. Contextual Information:**

- Additional contextual information about the Afghan automotive market, such as industry trends, regulatory changes, consumer behavior, and macroeconomic factors, may be required to enhance the model's ability to adapt to the local market conditions.

To address these concerns and improve the model's performance in the Afghan context, the following steps could be considered:

**1. Expand and Diversify the Dataset:**

- Collect and incorporate more comprehensive data from the Afghan automotive market, covering a wider range of car makes, models, and price points.



- Ensure the dataset includes sufficient representation of the various regions, customer demographics, and market segments within Afghanistan.

## 2. Incorporate Local Market Insights:

- Engage with industry experts, car dealers, and consumers in Afghanistan to gain a deeper understanding of the factors influencing car prices in the local market.
- Incorporate this contextual information into the model development and feature engineering process to better capture the nuances of the Afghan automotive landscape.

## 3. Continuous Model Evaluation and Refinement:

- Regularly evaluate the model's performance using real-world data from the Afghan market and gather feedback from users.
- Continuously refine and update the model to adapt to changes in the local market dynamics, consumer preferences, and regulatory environments.

By addressing the dataset scope limitations and incorporating a deeper understanding of the Afghan automotive market, the car price prediction model can be better tailored to provide accurate and reliable price estimates for consumers and industry stakeholders in Afghanistan. This will help ensure the long-term effectiveness and relevance of the model in the local context.

**3. Antiquated Dataset:** The automotive industry is known for its dynamic and rapidly changing nature, with constant advancements in technology, shifts in consumer preferences, and fluctuations in market conditions. As a result, the dataset used to train the initial car price prediction model may become increasingly outdated and less representative of the current state of the market, diminishing the model's predictive accuracy over time.

One key concern is the issue of technological advancements. The introduction of new car models, features, and technologies can change the relationship between car attributes and their corresponding prices. If the dataset used to train the model does not capture the impact of these technological developments, the model's ability to accurately predict car prices may deteriorate.

Another factor to consider is the evolving consumer preferences. As consumer priorities and behaviors shift, the perceived value and demand for certain car attributes may change. An outdated dataset that does not reflect these changes in consumer behavior can lead to inaccurate price predictions. Additionally, market volatility driven by economic conditions, government policies, and global events can introduce significant fluctuations in car prices, which an outdated dataset may not be able to account for effectively.

To address the issue of an antiquated dataset, it is essential to implement a robust data management and model maintenance strategy. This may include regularly updating the dataset with the latest data from the Afghan automotive market, continuously monitoring the model's performance and accuracy, and retraining the model at regular intervals to ensure it adapts to the changing market dynamics. Establishing feedback mechanisms for users to provide input on the model's accuracy and discrepancies can also help refine and enhance the model's performance over time. By proactively managing the dataset and continuously improving the car price prediction model, the system can maintain its relevance and provide accurate and reliable price estimates for the Afghan automotive market, even as the industry evolves.

## 5.2 Future Work

**1. Advanced Interface:** Developing a functional car price prediction model is an important first step, but you raise a valid point about the need to enhance the user interface (UI) to provide a more engaging and seamless user experience. While the current basic interface may be functional, investing in a more advanced and visually appealing design can significantly improve the overall usability and appeal of the system.

One of the key aspects to consider is the user experience (UX). A well-designed and intuitive interface can greatly enhance the user's interaction with the system, making it easier for them to navigate, input data, and interpret the results. By utilizing the latest web design tools and techniques, you can create a more visually engaging and responsive interface that caters to the needs and preferences of your target audience.

Some key considerations for improving the interface and enhancing the user experience include visual design, responsive design, intuitive navigation, data visualization, and feedback and error handling. Implementing a modern and visually appealing design aesthetic that aligns with the brand and target audience can create a more engaging and visually stimulating interface. Ensuring the interface is fully responsive and adapts seamlessly to different screen sizes, including mobile devices, will provide a consistent and optimized user experience across various platforms.

Developing a clear and intuitive navigation structure, with well-organized menus and easy-to-find features, can significantly improve the overall usability of the system. Incorporating interactive data visualizations, such as charts, graphs, or interactive maps, can enhance the information delivery and make the system more engaging and informative for users. Additionally, implementing clear and helpful feedback mechanisms, such as error messages or success notifications, can improve the overall user experience and build trust in the system's reliability.

Finally, ensuring the interface adheres to accessibility guidelines, making it inclusive and usable for users with diverse abilities and needs, can further enhance the system's reach and appeal. This can include features like screen reader support, keyboard navigation, and high-contrast color schemes.

**2. Up-to-date Data:** You raise a valid concern about the potential expiration and limitation of your current data sources, which could have a significant impact on the accuracy and reliability of your car price prediction model over time. Addressing this issue by continuously gathering data from an even wider range of sources is a crucial step to ensure the long-term success and sustainability of your system.

Maintaining an up-to-date and comprehensive data set is essential for keeping your car price prediction model relevant and accurate. As market conditions, consumer preferences, and other factors change over time, relying solely on a static data set can lead to inaccurate predictions and a decline in the model's performance.



To fix this issue and future-proof your system, you should implement a robust data collection and management strategy that focuses on the following key aspects:

**Diversify Data Sources:** Expand your data gathering efforts to include a wider range of sources, such as online automotive marketplaces, industry databases, government databases, and even crowdsourced data from users. This will help you build a more comprehensive and diverse data set that can better capture the dynamic nature of the car market.

**Implement Automated Data Gathering:** Develop automated processes and scripts to continuously scrape, collect, and integrate data from various sources into your system. This will ensure that your data is regularly updated, reducing the risk of relying on outdated information.

**Implement Data Validation and Cleaning:** Implement robust data validation and cleaning processes to ensure the quality and accuracy of your data. This may include identifying and removing outliers, handling missing values, and standardizing data formats to maintain data integrity.

**Establish Regular Data Updates:** Implement a schedule for regularly updating your data, such as daily, weekly, or monthly, depending on the rate of change in the car market. This will help you keep your model's predictions up-to-date and responsive to market trends.

**Monitor Data Quality and Performance:** Continuously monitor the quality of your data and the performance of your car price prediction model. Identify any potential issues or changes in the data that may require adjustments or retraining of your model.

**Leverage Machine Learning for Adaptive Modeling:** Consider incorporating machine learning techniques that allow your model to adapt and learn from the continuously updated data. This can help your system stay responsive to market changes and maintain accurate predictions over time.

By implementing these strategies, you can ensure that your car price prediction model remains accurate, reliable, and relevant, even as the car market and data landscape evolve. This will not only address the limitations you mentioned but also position your system for long-term success and wider adoption by your target audience.

**3. Dataset Expansion:** As your car price prediction model gains traction, expanding the geographic scope of your dataset is a strategic move that can significantly enhance its capabilities and reach. Broadening the data collection efforts beyond the current focus on Afghanistan will allow you to capture a more diverse and representative set of market conditions, consumer preferences, and automotive industry trends.

When considering the expansion of your dataset to other regions, there are several key factors to keep in mind. First and foremost, you'll need to carefully identify the target regions for expansion. This selection process should be guided by factors such as market size, growth potential, data availability, and relevance to your target audience. Prioritizing regions with significant automotive markets and a strong demand for reliable car pricing information will be crucial.

Once you have identified the target regions, the next step is to establish strategic data partnerships. Forging relationships with local automotive marketplaces, industry associations, and data providers in these regions can facilitate access to valuable data sources and ensure the seamless integration of regional data into your central database. This collaborative approach will be essential in overcoming any challenges posed by unique data formats, naming conventions, and other regional differences.

Adapting your data collection processes to accommodate the expanded geographic scope will be a critical undertaking. Modifying your automated data gathering scripts and workflows to handle the increased diversity and complexity of the dataset will be crucial. Investing in the development of region-specific data parsing and normalization capabilities will help maintain data integrity and consistency across your expanded dataset.

Finally, as you continue to enhance the geographic coverage of your car price prediction model, it will be essential to incorporate regional insights, monitor performance, and adapt your strategies accordingly. Gathering local market insights, regulatory changes, and industry trends can help you contextualize the data and improve the accuracy of your predictions. Regularly monitoring the model's performance and refining your data collection, processing, and optimization strategies will be key to maintaining high-quality predictions and adapting to evolving market conditions in each region.

## 5.2 Conclusion

In conclusion, your car price prediction project has successfully developed a model that can predict accurate prices of specific car models related to 5 different global car companies. This is an impressive achievement that demonstrates your team's dedication and technical expertise.

The data collection process, where you gathered paper sheet data from different car dealerships across the capital city of Kabul and converted it into a digital format, was a critical first step. The data preprocessing and model evaluation phase, where you compared the performance of various algorithms and ultimately selected the Decision Tree model for its high  $R^2$  score, showcases your analytical rigor and data-driven approach.

The creation of a user-friendly interface, leveraging JavaScript, is an excellent addition that will enhance the accessibility and usability of your car price prediction tool. This user-centric design will ensure that both car buyers and sellers can easily interact with the system and benefit from its accurate price predictions.

As you eagerly await the real-world performance of your model, you can be confident that this project has the potential to become a valuable resource for the local automotive industry. The ability to provide reliable and transparent car pricing information can empower consumers, facilitate more informed decision-making, and foster a healthier and more transparent automotive market.

Your team's dedication and the successful development of this car price prediction model are commendable. The lessons learned and the expertise gained throughout this project can serve as a strong foundation for future enhancements and expansions, ultimately contributing to the growth and betterment of the automotive industry in the region.

## References

- [1] Phani Krishna Kondeti, K. R. (2019, September). Applications of machine learning techniques to predict filariasis using socio-economic factors. Retrieved from [cambridge.org](https://cambridge.org)
- [2] Nabarun Pal, D. S. (2019). A methodology for predicting used cars prices using Random Fores. Retrieved from [link.springer.com](https://link.springer.com)
- [3] Richard R. Yang, S. C. (2018, March 29). Vehicle Price Prediction Using Visual Features. Retrieved from [arxiv.org](https://arxiv.org)
- [4] Pudaruth, S. (2014, January). Predicting the Price of Used Cars using Machine Learning Techniques. Retrieved from [academia.edu](https://academia.edu)
- [5] Pattabiraman Venkatasubbu, M. G. (2019, December). Used Cars Price Prediction using Supervised Learning Techniques. Retrieved from [researchgate.net](https://researchgate.net)
- [6] Enis Gegic, B. I. (2019, February). Car Price Prediction using Machine Learning Techniques. Retrieved from [ceeol.com](https://www.cceol.com): <https://www.cceol.com/search/article-detail?id=746689>
- [7] Baoyang Cui Zhonglin Ye, H. Z. (2022, August 31). Used Car Price Prediction Based on the Iterative Framework of XGBoost+LightGB. Retrieved from [mdpi.com](https://mdpi.com)
- [8] K.Samruddhi, D. R. (2020, September). Used Car Price Prediction using K-Nearest Neighbor Based Model. Retrieved from [ijirase.com](https://ijirase.com)
- [9] A Pandey, V Rastogi, S Singh. (2020). Car's Selling Price Prediction using Random Forest Machine Learning Algorithm. Retrieved from [papers.ssrn.com](https://papers.ssrn.com)
- [10] Nabarun Pal, P. A. (2018, December). How Much Is My Car Worth? A Methodology for Predicting Used Cars'. Retrieved from [link.springer.com](https://link.springer.com)
- [11] Andreea Dutulescu, M. L.-M. (2023). What is The Price Of Your Used Car? Automated Prediction and Neural Networks. Retrieved from [ieeexplore.ieee.org](https://ieeexplore.ieee.org)
- [12] Prashant Gajera, A. G. (2021 , March). OLD CAR PRICE PREDICTION WITH MACHINE LEARNING. Retrieved from [irjmets.com](https://irjmets.com)
- [13] Ganesh, M. (2019, December). Used Cars Price Prediction using Supervised Learning Techniques. Retrieved from [researchgate.net](https://researchgate.net)
- [14] Bitvai, Z., Cohn, T., 2015a. Non-linear text regression with a deep convolutional neural network. ACL-IJCNLP 2015 - 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Proceedings of the Conference 2, 180–185. doi:10.3115/v1/p15-2030
- [15] Luo, Y., Wang, M., Zhou, H., Yao, Q., Tu, W.W., Chen, Y., Dai, W., Yang, Q., 2019. AutoCross: Automatic Feature Crossing for Tabular Data in Real-World Applications, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, ACM, Anchorage AK USA. pp. 1936–1945. doi:10.1145/3292500.3330679.

- [16] Noor, Kanwal, and Sadaqat Jan. "Vehicle price prediction system using machine learning techniques." *International Journal of Computer Applications* 167, no. 9 (2017): 27-31.
- [17] K.Samruddhi, Dr. R.Ashok Kumar "Used Car Price Prediction using K-Nearest Neighbor Based Model" *IJIRASE* Volume 4, Issue 3, DOI: 10.29027/IJIRASE.v4.i3.2020.686-689, September 2020
- [18] IBM. (2020, November 20). What is the k-nearest neighbors algorithm? | IBM. Retrieved from [ibm.com](https://ibm.com)
- [19] Kanade, V. (2023, April 3). What Is Linear Regression? Types, Equation, Examples. Retrieved from [spiceworks.com](https://spiceworks.com)
- [20] Jaiswal, S. (n.d.). Decision Tree Algorithm in Machine Learning - Javatpoint. Retrieved from [www.javatpoint.com](https://www.javatpoint.com)
- [21] Fernando, J. (2023, April 8). R-Squared:Definition Calculation Formula uses and limitations. Retrieved from [www.ncl.c.uk](https://www.ncl.c.uk)