**ENSEMBLE MACHINE LEARNING MODEL FOR PREDICTING USED CAR PRICES IN KENYA.**

**MOSES ONSERIO JAMES**

**MDATC01/6051/2022**

**A RESEARCH PROPOSAL SUBMITTED TO THE DIRECTORATE OF POSTGRADUATE STUDIES FOR PARTIAL FULFILLMENT FOR THE REQUIREMENTS FOR THE AWARD OF MASTERS OF SCIENCE IN DATA SCIENCE OF THE CO-OPERATIVE UNIVERSITY OF KENYA.**

# **DECLARATION**

This proposal is my original work and has not been presented for any other degree or award in any university.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Moses Onserio James Date

I confirm that the candidate did the work presented in this proposal under my supervision.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Dr. Ronald Ojino Date

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Dr. Fidelis Mukudi Date

The Co-operative university of Kenya.

# **DEDICATION**

I dedicate this work to my family, my loving parents Mr. James Onserio Onserio and Mrs. Perpetual Onserio, and my loving sisters Esther Onserio and Macvivian Onserio who have unconditionally supported me.

# **ACKNOWLEDGEMENT**

First, I take this opportunity to thank the management of the University for the Financial Support and conducive environment for learning. Special thanks to Dr. Shem Mbandu, Dean School of Computing and Mathematics, and Dr. Charles Katila, Chairperson Department of Computer Science and Information Technology who shared insights into the degree programme and its relevance that led to my enrolment in the University.

Special thanks to my supervisors, Dr. Ronald Ojino and Dr. Fidelis Mukudi for advice, guidance, and encouragement that generated insights that shaped the research.

To my father and mother, I say thank you for both financial and emotional support. You will always remain in my heart. To my siblings Esther and Macvivian who give hope in this World. Let this work inspire you.

Thanks to all Master of Science in Data Science students with whom I shared knowledge, staff, and lecturers at The Co-operative University of Kenya for your support.

# **ABSTRACT**

This study aims to develop an ensemble machine-learning model to predict used car prices in Kenya. Given the challenges of inconsistent and subjective pricing in the Kenyan used car market, this research seeks to provide a data-driven solution by incorporating key vehicle attributes such as brand, mileage, year of production, engine size, fuel type, torque, condition, transmission, drive, and horsepower. The study will apply Design Science Research methodology and the model will integrate machine learning techniques; Random Forest, Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Gradient Boosting, and SHAP for feature explanation to enhance accuracy and transparency. The goal is to improve used car pricing accuracy, and market efficiency, foster trust, and support better decision-making for buyers, sellers, and other stakeholders. This study will contribute to the application of machine learning in emerging markets, providing a standardized tool for predicting used car prices in Kenya.

Contents

[**DECLARATION** i](#_Toc189082388)

[**DEDICATION** ii](#_Toc189082389)

[**ACKNOWLEDGEMENT** iii](#_Toc189082390)

[**ABSTRACT** iv](#_Toc189082391)

[**LIST OF TABLES** vi](#_Toc189082392)

[**LIST OF FIGURES** vii](#_Toc189082393)

[**LIST OF SYMBOLS AND ABBREVIATIONS** viii](#_Toc189082394)

[**CHAPTER 1: INTRODUCTION** 1](#_Toc189082395)

[**1.1 Background study** 1](#_Toc189082396)

[**1.2 Problem Statement** 3](#_Toc189082397)

[**1.3 Objectives** 4](#_Toc189082398)

[**1.4 Significance of the Study** 5](#_Toc189082399)

[**1.5 Scope of study** 5](#_Toc189082400)

[**1.6 Justification of the study** 6](#_Toc189082401)

[**1.7 Limitations of study** 6](#_Toc189082402)

[**CHAPTER 2: LITERATURE REVIEW** 7](#_Toc189082403)

[**2.1 Introduction** 7](#_Toc189082404)

[**2.2 Research Gap** 15](#_Toc189082405)

[**2.3 Assumptions of the Study** 17](#_Toc189082406)

[**2.4 Expected Outcomes of the Study** 17](#_Toc189082407)

[**2.5 Theoretical frameworks** 18](#_Toc189082408)

[**2.6 Conceptual framework** 19](#_Toc189082409)

[**CHAPTER 3: RESEARCH METHODOLOGY** 21](#_Toc189082410)

[**3.1 Introduction** 21](#_Toc189082411)

[**3.2 Design Science Research Paradigm and Pragmatism as a Paradigm in DSR** 21](#_Toc189082412)

[**3.3 Ethical Considerations** 33](#_Toc189082413)

[**REFERENCES** 34](#_Toc189082414)

[**APPENDICES** 38](#_Toc189082415)

# 

# **LIST OF TABLES**

[**1. Summary of key research gaps** 16](#_Toc189083885)

# **LIST OF FIGURES**

**Figure 1. Conceptual framework**………………………………………………………………**20**

# **LIST OF SYMBOLS AND ABBREVIATIONS**

Mathematical and Statistical Symbols and abbreviations

ANN - Artificial Neural Network

CC - Engine capacity in cubic centimeters

IQR - Interquartile Range

K-Fold - A cross-validation method where K represents the number of splits in the dataset

KNN - K-Nearest Neighbors

Ksh - Kenyan (currency)

LSTM - Long Short-Term Memory

MAE - Mean Absolute Error

ML - Machine Learning

OLS - Ordinary Least Squares

R² - R-squared (coefficient of determination)

RMSE - Root Mean Square Error

SHAP (SHapley Additive exPlanations)

SLA - Service Level Agreement

SVM - Support Vector Machine

VIF - Variance Inflation Factor

# **CHAPTER 1: INTRODUCTION**

## **1.1 Background study**

The number of vehicles in Kenya has been growing at a rapid rate of 12% annually, with the national registered fleet standing at 4 million as of 2018 (Kenya National Bureau of Statistics, 2023). This significant growth highlights vehicles' critical role in the Kenyan economy. Regular pricing of vehicles is essential for a variety of purposes, including insurance, resale, leasing, and accounting among others. However, the existing systems for automobile pricing exhibit considerable variability, as values often differ significantly even for identical vehicles. This inconsistency points to the inefficiencies in traditional pricing methods (Lessmann & Voss, 2017).

Currently, obtaining a car price requires contacting licensed experts, such as evaluation firms, insurance agents, or an individual who knows the field. These methods rely heavily on expert opinions and standardized depreciation formulae, which consider factors like mileage, age, transmission, fuel type, engine capacity, horsepower, and condition(accident or accident-free) among others. While these traditional approaches are widely used, they often lack uniformity and consistency. The advent of machine learning has shown promise in automating vehicle pricing, offering more reliable and accessible solutions.

Predicting the prices of used cars is particularly relevant in today’s automotive market, as used vehicles have become a priority for many buyers due to their affordability. For instance, in the United States alone, approximately 40 million used-cars are sold annually. On average, used cars are 50% cheaper than new ones, allowing buyers to save on financing costs and avoid the sharp depreciation new vehicles face, which can be as high as 11% within the first drive from the lot (National Automobile Dealers Association, 2023).

The Kenyan used-car market primarily relies on imports, with Japanese brands of vehicles like Toyota dominating due to their affordability (Doctor, 2024). However, the market faces challenges, such as fluctuating import duties and currency depreciation, which significantly affect car prices (Mobility Foresights, 2024). Recent reports indicate a sharp rise in the cost of used cars in Kenya. For example, between September 2023 and early 2024, the price of a used Mercedes C-Class 2017 model increased from KSh 3.8 million to at least KSh 4.4 million, while the Toyota Harrier’s price rose from KSh 3.8 million to KSh 4 million (Munda & Mutua, 2024). Smaller models, such as the Mazda Demio and Honda Fit, have also experienced substantial price hikes. The primary drivers of these price increases include higher import duties—which rose from 25% to 35% in July 2023—excise taxes, VAT, rising credit costs, and the depreciation of the Kenyan shilling against the dollar. These factors have resulted in reduced vehicle imports, with spending dropping by $162 million in 2023 (National Treasury Tax Expenditure Report, 2024). In response, households and dealers have increasingly turned to older, locally available cars due to reduced costs. The Kenya Auto Bazaar Association attributes this trend to declining household purchasing power and limited access to loans for small traders (Africa, 2024)

At the same time, challenges such as varying pricing methodologies complicate the market further. Pricing a car for resale often involves engaging an expert or referencing market trends, but these methods are inconsistent and highly subjective. Sellers may price their vehicles too high and face long delays in selling, or they may set prices too low and incur losses. Buyers, on the other hand, struggle with unreliable price estimates, leading to inefficiencies in market transactions (Peach Cars Report, 2024).

A robust and standardized model for predicting the prices of used-cars in Kenya is therefore essential. Such a model could incorporate key factors to provide more accurate and uniform price. This would benefit both buyers and sellers by reducing market inefficiencies, enhancing transparency and fostering trust in the used-car market. With the increasing integration of digital technologies in the automotive sector, machine learning provides an opportunity to revolutionize car pricing in Kenya. By leveraging data-driven approaches, a predictive model can help bridge the gap between subjective pricing and objective market realities. This initiative not only addresses the growing need for consistent and reliable pricing tools but also aligns with global trends of adopting technology to solve real-world challenges in the used-car market. This study aims to help customers estimate the price value of a used-car they are willing to buy or sell. To achieve this, this study aims to build a machine learning model that predicts car prices based on various features such as make, mileage, year of production, fuel type, engine capacity, torque, acceleration, horse power, transmission and drive. The study will explore how machine learning and big data analytics can be applied to predict the price of used-cars using a local dataset based in Kenya. It will employ various statistical models and machine learning algorithms to lay the foundation for developing and understanding the need for accurate price prediction techniques.

## **1.2 Problem Statement**

Pricing used cars in Kenya is a significant challenge due to the reliance on inconsistent and subjective pricing methods. Sellers often overestimate vehicle prices to maximize profits, leading to long delays in sales, while buyers struggle with unreliable pricing information, which breeds mistrust. Most existing studies for instance a research by Pillai (2020) used Artificial Neural Networks (ANN), for used car price prediction in the U.S. market, Samruddhi and Kumar (2020) proposed a K-Nearest Neighbor (KNN) based model to predict the prices of used cars in India and

Mustapha et al. (2022) who proposed a regression model to predict the resale value of used cars in Morocco with exception of a few , have primarily focused on single machine learning algorithms, partial variables or markets in Europe, the United States, India, Morocco and Asia. These models are not directly applicable to the Kenyan context. There is currently no data-driven, standardized tool to address these challenges in Kenya’s used-car market. To solve this problem, there is a need for a tailored ensemble machine learning model that incorporates key vehicle attributes for instance make, year of manufacture, drive, engine size, mileage, fuel type, horse power, torque and acceleration. Such a model will provide accurate price predictions, reduce inefficiencies and improve market transparency.

## **1.3 Objectives**

**1.3.1 Main Objective**

To develop an ensemble machine learning model that accurately predicts the prices of used cars in the Kenyan market.

**1.3.2 Specific Objectives**

1. To determine the key features to consider when predicting the prices of used cars in Kenya.
2. To develop an ensemble machine learning model for predicting used car prices in Kenya by integrating Random Forest, Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Gradient Boosting, while leveraging SHAP for interpretability.
3. To validate the developed ensemble model using the real-world dataset.
4. To evaluate the performance of the validated model using appropriate metrics.

**1.3.3 Research Questions**

1. What are the key features influencing the price of used cars in Kenya?
2. How can an ensemble machine learning model integrating Random Forest, SVM, KNN, and Gradient Boosting improve the accuracy and interpretability of used car price prediction in Kenya?
3. How well does the developed ensemble model perform when applied to a real-world dataset of used car prices in Kenya?
4. Which performance metrics provide the most reliable assessment of the validated ensemble model’s accuracy in predicting used car prices in Kenya?

## **1.4 Significance of the Study**

The purpose of this research is to develop an advanced ensemble machine learning model to predict used car prices in Kenya more accurately by combining Random Forest, Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Gradient Boosting, and SHAP. The study seeks to address the existing pricing challenges by creating a data-driven solution that integrates vehicle-specific attributes. By providing reliable and standardized price estimates, the model will enhance trust, improve market efficiency and support better decision-making for buyers, sellers other stakeholders and contribute to the growing body of knowledge on applying machine learning in emerging markets.

## **1.5 Scope of study**

The scope of this study is to develop an ensemble machine learning model to accurately predict the prices of used cars in Kenya. This encompasses both vehicles that were initially used within Kenya and those imported from other countries but have a history of prior use. The model will use real-world datasets with key attributes such as name, model, make, year of manufacture, mileage, engine size, fuel type, horsepower, transmission, condition, car name, body type, source (Kenyan/foreign used) and acceleration. Ensemble techniques: Random Forest, Support Vector Machines (SVM), Gradient Boosting, and K-Nearest Neighbors (KNN) will be applied to enhance accuracy. SHAP (SHapley Additive exPlanations) will also be used to explain feature contributions, ensuring transparency.

## **1.6 Justification of the study**

The Kenyan used car market plays a vital role in transportation, offering affordable mobility options to a significant portion of the population. While characterized by a preference for durable Japanese imports and locally used cars, the market faces challenges such as price uncertainty, limited transparency, and the proliferation of counterfeit parts. (Chepkwony, 2022).These issues hinder consumer confidence, lead to pricing inefficiencies, and stifle market growth. Developing a machine learning-based price prediction model specifically tailored to the Kenyan context is crucial. This model will enhance market transparency by providing objective price estimates, empowering buyers with informed decision-making, and reducing the risk of overpricing. By incorporating key factors like brand, engine size, age, drive, transmission, fuel type, mileage, acceleration, and horsepower, the model will offer more accurate predictions than existing global models. This research will contribute significantly to improving market efficiency, reducing fraud, and fostering greater trust between buyers and sellers in the Kenyan used-car market.

## **1.7 Limitations of study**

The study is limited to the constraint of financial resources. To ensure the feasibility of data collection within budget constraints, the research will primarily rely on an online data collection method which is web scraping.

# **CHAPTER 2: LITERATURE REVIEW**

## **2.1 Introduction**

The literature review explores studies conducted by various researchers on car price prediction using machine learning techniques. It examines the key variables employed in these studies, such as mileage, year of production, brand, model, fuel type, and other features influencing car prices. The review also discusses the machine learning models applied, including linear regression, Random Forest, Gradient Boosting, and Neural Networks, highlighting their respective outcomes and performance metrics. Furthermore, the chapter identifies the future recommendations proposed by these studies, such as incorporating advanced algorithms, expanding datasets, or improving model explainability. By synthesizing these findings, the review highlights how these studies address existing challenges in car price prediction while aligning with the research gaps this study aims to address. This forms the foundation for developing a more accurate, scalable, and interpretable model tailored to the research context.

Some studies have modeled used car pricing. However, most rely on single modeling techniques, partial variables, or have a narrow scope. For example, Gegic et al. (2019) developed a car price prediction model for the Bosnia and Herzegovina market using ensemble machine learning techniques. The study employed **Artificial Neural Networks (ANN), Support Vector Machines (SVM),** and **Random Forest** algorithms, combining them to enhance prediction accuracy. The data, scraped from autopijaca.ba using PHP, was tested in a Java-based application. The model achieved an accuracy of **87.38%,** demonstrating promising results. However, the authors suggested future research involving diverse datasets and scope for improved generalization.

Chandak et al. (2019) explored the correlation between car attributes for instance kilometers driven, year of registration, fuel type, and fiscal power of used-car prices. The study applied **K-Nearest Neighbors (KNN)** and **Classification and Regression Trees (CART)** on a dataset containing **300,000 entries.** CART outperformed KNN with a lower root mean square error (RMSE) of **4961.64**, compared to KNN's RMSE of **5581.96.** The study emphasized the importance of incorporating additional features, such as horsepower and torque, in future models for improved accuracy.

Artificial Neural Networks (ANN) were utilized by Pillai (2020) for used-car price prediction in the U.S. market. The model, trained on 140,000 car records across 30 popular brands and tested on 35,000 cars, achieved a Mean Absolute Percentage Error (MAPE) of 11% and an R² value of 0.96 by employing embedding techniques for categorical variables. This outperformed simpler models like linear regression and random forests, demonstrating ANN's potential for highly accurate price predictions. However, this study is limited to a single model and scope. Further research is needed to compare ANN's performance with other techniques, such as Random Forest and Support Vector Machines (SVM), to identify the most accurate model for used-car price prediction.

Samruddhi and Kumar (2020) proposed a K-Nearest Neighbor (KNN) based model to predict the prices of used cars in India. They used data collected from the Kaggle website to train the model. The data was tested with different training and testing ratios, and the model achieved an accuracy of approximately 85%. This demonstrated that the KNN regression algorithm can effectively predicting used car prices and for buyers and sellers in the used-car market. However, for future research, they recommended the use of other advanced machine learning techniques and different scopes to validate the model with different methods to enhance the optimization of the model with improved accuracy.

Bukvić et al. (2020) presents an overview of data-driven models for predicting the price of used vehicles in the Croatian market. The authors focused on key factors like the production year and kilometers driven. They gathered data from the online marketplace "Njuškalo" and cleaned it by removing redundant and missing values. The study used linear regression to predict car prices. They also compared the accuracy of this method with classification algorithms. The aim was to analyze the vehicle market and predict price trends based on available data. They concluded that the predicted model has the highest accuracy with linear regression where main features (price and model) are available. This study is limited to India and paid attention to only key factors (Production year and Kilometers driven) with linear regression as the specific machine learning technique. This calls for a need to do research that brings many features together and combines various machine learning techniques to identify which predicts the price better.

Bharambe et al. (2022) proposed three regression algorithms to predict the price of used cars. They considered various factors to ensure reliable and accurate predictions. The study used three supervised machine learning techniques: linear regression, lasso regression and ridge regression. Python libraries like Numpy, Pandas, and Sklearn were used to build the model and design the project’s graphical user interface (GUI). The accuracy of the models was compared, with linear regression achieving 83.65%, lasso regression 87.09%, and ridge regression 84.00%. The final price prediction was made using lasso regression, as it provided the highest accuracy. For future work, the authors plan to collect more data and explore advanced techniques like Random Forest, Artificial Neural Networks (ANN), and Convolutional Neural Networks (CNN) to improve the user experience and prediction accuracy.

Budiono et al. (2024) proposed using the K-Nearest Neighbors (KNN) regression model to predict used-car prices. KNN is a machine learning algorithm that works well with multi-dimensional data and is resistant to noisy data, making it ideal for predicting used-car prices based on various factors. The model analyzes data such as the year of production, car type, condition, and other relevant details. The study used 504 used car data points collected through web scraping. The model achieved an error rate of just 8.3% and an R2R^2 value of 98.8%. These results can help both buyers and sellers better understand and estimate used car prices in the market. This research used only a single model to predict the price of used cars. It calls for a need to combine the various advanced machine learning techniques to better predict the price.

Amik et al. (2021) proposed a model to help customers in Bangladesh predict the price of pre-owned cars, as there are no existing online services for this purpose. Their goal was to create a system that helps potential buyers make informed decisions when purchasing used cars. They collected and processed a dataset and performed exploratory data analysis. Several machine learning regression algorithms, such as linear regression, LASSO regression, decision tree, random forest, and extreme gradient boosting (XGBoost), were applied to the data. After comparing the performance of each method, XGBoost was chosen as the best model, accurately predicting prices over 91% of the time. The most important features for predicting car prices were found to be the brand, model, year, transmission type, body type, fuel type, engine capacity, and kilometers driven. The dataset was collected from Bikroy.com website, and the analysis revealed new market trends. The XGBoost model, with an R Squared score of 91.32%, was deployed as a web application for future use. The researchers also proposed a plan for future to collect a larger dataset with more relevant features for improved predictions.

Peerun et al. (2022) conducted study aimed to determine if artificial neural networks (ANNs) could accurately predict second-hand car prices in Mauritius. The researchers collected data on 200 used cars from various sources, including car websites and newspaper advertisements, over a one-month period in August 2014. Features like the car's manufacturing year, brand, engine capacity (in cubic centimeters), paint type (normal or metallic), transmission type (manual or automatic), mileage (kilometers driven), and price (in Mauritian rupees) were included in the dataset. Four machine learning algorithms were tested, and support vector machine (SVM) regression performed slightly better than ANNs and linear regression. However, for higher-priced cars, some predictions were significantly off from actual prices. They made use of Mean Absolute Error as a performance metric. This indicates the need for further research with a larger dataset and additional experiments using different neural network structures and different performance metrics to improve accuracy.

Varshitha et al.,(2022) conducted research in India to develop supervised learning models, including Artificial Neural Networks (ANN) and Random Forest, for predicting used-car prices based on vehicle datasets. The study aimed to create an accurate and reliable prediction system by analyzing various car attributes. The ANN model was implemented using the Keras Regressor algorithm, while other machine learning algorithms, including Random Forest, Lasso, Ridge, and Linear Regression, were tested on the dataset. Among these, the Random Forest model performed the best, achieving the lowest error with a Mean Absolute Error (MAE) of 1.097 and an R² value of 0.772. This research demonstrated that the Random Forest model is more effective than simpler models, providing better predictions for used car prices. The findings also suggest that such models could potentially address fraud in the used-car market with further refinement and future advancements. This also necessitates the utilization of diverse advanced machine learning techniques to enhance the accuracy of used-car price prediction.

Mustapha et al. (2022) proposed a regression model to predict the resale value of used cars in Morocco, aiming to assist buyers and sellers in estimating accurate car prices. The study utilized a dataset collected from Avito, a local online e-commerce platform, where a Python script, leveraging the BeautifulSoup library, was used to crawl car listings. Important features such as mileage, fuel type, production year, car make, model, and fiscal power were extracted from these listings. After data collection, the dataset was cleaned by removing null values and outliers. Missing values in columns like mileage and fiscal power were imputed using their mean values to avoid losing valuable data. The final dataset comprised 8,000 records.In terms of model performance, the Gradient Boosting Regressor (GBR) outperformed all other models tested, achieving an R² score of approximately 0.80. This result indicated that the model explained 80% of the variation in car prices, though the R² score was not high enough to consider the model a perfect fit. Additionally, the GBR model showed relatively weak performance in minimizing root mean squared errors (RMSE). Other models, such as KNN and Artificial Neural Networks (ANN), were also tested, but the GBR consistently demonstrated superior performance. They recommended additional of more features in future studies for more accuracy, which could improve both the R² score and reduce RMSE, leading to more accurate predictions of used car prices.

Liang Han et al. (2020) presented a price suggestion system for online second-hand items using both text descriptions and images. The system aims to assist sellers in setting effective and reasonable prices for their second-hand items uploaded to online platforms. They designed a multi-modal system that extracts visual and textual features, alongside statistical item features, to assess whether the uploaded item is suitable for a reasonable price suggestion. This assessment is done using a binary classification model, while a regression model is used to generate price suggestions for items deemed qualified. The system includes two constraints in the joint training of the classification and regression models and utilizes a customized loss function to optimize the price suggestions, maximizing the seller's gain and facilitating online transactions. The authors also propose a set of evaluation metrics to assess the performance of their system, with experiments on a large real-world dataset demonstrating its effectiveness. While the system may be effective in making price suggestions, it may lack transparency in explaining why a particular price is suggested for an item. Providing insight into how the different features contribute to the final price suggestion could help sellers trust and better utilize the system.

Kanwal et al. (2017) introduced a vehicle price prediction system that uses supervised machine learning, specifically multiple linear regression, to predict vehicle prices with 98% accuracy. The model relies on the following factors: vehicle's model, make, city, version, color, mileage, alloy rims, and power steering, to predict its price as the dependent variable. The system simplifies model complexity by selecting only the most relevant features from the dataset, eliminating unnecessary variables during the processing phase. This results in a more efficient and accurate prediction model. They, however, recommend future work to use advanced methods like fuzzy logic , K-nearest neighbors and genetic algorithms to get better results.

Pal et al. (2018) used Random Forest, a supervised learning method, to predict the prices of used cars. The model was selected after conducting thorough exploratory data analysis to evaluate the impact of each feature on price. A Random Forest with 500 decision trees was constructed to train the dataset. The experimental results demonstrated a training accuracy of 95.82% and a testing accuracy of 83.63%. The model effectively predicts car prices by selecting the most correlated features, ensuring reliable and accurate outcomes. The researchers used only one technique to predict the price. Better accuracy can be achieved by combining Random Forest with other models like Gradient Boosting or ensemble them using techniques like stacking or blending to leverage the strengths of different algorithms.

Ahtesham and Zulfiqar (2022) aimed to develop machine learning models, including linear regression, gradient boosting, and decision tree algorithms, to predict the prices of used cars based on their features. The implementation was carried out using Apache Spark, a big data processing tool, alongside several machine learning libraries such as pandas. The dataset used for this study was sourced from PakWheels, consisting of 56,186 instances and 16 attributes. While significant research has been conducted in this domain, the authors highlight that limited work has utilized PySpark for such predictions. Their proposed models achieved varying levels of accuracy, with linear regression producing approximately 50% accuracy, gradient boosting achieving 89%, and decision tree regression obtaining 86%. For future work, the authors suggested exploring models like K-nearest neighbors (KNN) and other algorithms to achieve improved prediction results.

Sachin Kumar and Damandeep Kaur (2022) proposed a car price prediction system for an Asian market using machine learning techniques. The dataset for this research was obtained from Kaggle website and underwent preprocessing steps such as normalization, standardization, and cleaning to eliminate noise and improve the performance of machine learning algorithms. Various algorithms were evaluated to identify the one best suited for the dataset. The authors implemented the Random Forest algorithm to develop the prediction model, which demonstrated strong predictive capabilities. The final model was integrated into a Python-based application using Flask, enabling user interaction. Additionally, the system was deployed on Heroku for accessibility and ease of use. The study achieved impressive results in tackling the car price prediction problem. For future research, the researchers recommended to extend the system to work effectively with multiple datasets, enhancing its versatility and robustness.

## **2.2 Research Gap**

From the reviewed literature, it is evident that significant progress has been made in predicting used car prices using various machine learning techniques such as **K-Nearest Neighbors (KNN), Artificial Neural Networks (ANN), Support Vector Machines (SVM), Random Forest** and **XGBoost.** However, the following research gaps can be identified:

1. **Limited Use of Ensemble Methods:** While studies like Gegic et al. (2019) and Bharambe et al. (2022) touched on ensemble techniques, many of the models rely on single algorithms, such as linear regression, KNN, or ANN. Combining multiple models (ensemble methods) has shown a potential to improve prediction accuracy, but this has not been widely applied.
2. **Local Context and Limited Scope**: Most studies have been conducted in markets such as the **U.S., Europe, India and Mauritius**, with **limited research in the African context**, particularly Kenya. The Kenyan used car market has unique challenges, including **economic factors** for instance high import duties, **aging car inventory** and **market inefficiencies**. This requires models tailored to the Kenyan context.
3. **Feature Engineering Gaps:** Some studies (Chandak et al., 2019) emphasized the importance of incorporating additional features like **horsepower, torque or maintenance costs** to improve accuracy, but this has not been widely adopted.
4. **Small and Outdated Datasets:** Many studies used small or geographically limited datasets. For instance, Peerun et al. (2022) only collected 200 car records. To improve generalizability, there is a need for **larger, more diverse, and up-to-date datasets** in Kenya.

**1. Summary of key research gaps**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Study** | **Techniques** | **Strengths** | **Weaknesses** | **Focus** |
| Gegic et al.(2019) | Ensemble (ANN, SVM,Random Forest) | High accuracy(87.38%) | Limited data source (Bosnia and Herzegovina), needs diverse datasets for  improved generalization | Used car market in Bosnia and Herzegovina |
| Chandak et al. (2019) | KNN, CART | Identified key features (mileage, year, fuel type) | Lower accuracy compared to other studies (RMSE: 4961.64), recommends including more features (horsepower, torque) | Not specified |
| Samruddhi & Kumar (2020) | KNN | Achieved good accuracy (85%) | Limited data source (Kaggle website), recommends using advanced techniques for improved accuracy | Used car market in India |
| Bukvić et al. (2020) | Linear Regression | Focuses on key factors (year, mileage), good accuracy with linear regression | Limited to linear regression, needs exploration of advanced techniques and using more features | Used car market in Croatia |
| Bharambe et al. (2022) | Linear Regression, Lasso Regression, Ridge Regression | Compare multiple algorithms (highest accuracy with Lasso Regression: 87.09%) | Limited data, proposes using advanced techniques (Random Forest, ANN, CNN) for improvement | Not specified |
| Mustapha et al. (2022) | Gradient Boosting Regressor | Achieved good accuracy (R²: 0.80) | Lower accuracy in minimizing RMSE, recommends adding more features for improvement | Used car market in Morocco |
| Pal et al. (2018) | Random Forest | Achieved good accuracy (training: 95.82%, testing: 83.63%) | Used single technique (Random Forest), proposes combining with other models (Gradient Boosting) for improvement | Not specified |

## **2.3 Assumptions of the Study**

1. The available data is assumed to be of reasonable quality, with minimal missing values and inconsistencies. Data cleaning and preprocessing steps will be necessary to ensure data accuracy and reliability.
2. It is assumed that the developed ensemble machine learning model will apply to various pricing scenarios and different categories of used cars within the Kenyan market.
3. The model is assumed to be robust to variations in market conditions and changes in vehicle attributes over time.
4. It is assumed that the data to be collected on car prices accurately reflects all local market trends, including fluctuations in demand, supply, depreciation, and applicable costs such as value-added tax (VAT) and other duties.

## **2.4 Expected Outcomes of the Study**

The study is expected to produce a reliable and more accurate **ensemble machine learning model** capable of predicting the prices of used cars in Kenya based on key attributes such as brand, car age, fuel type, transmission and engine size. The model is anticipated to outperform traditional pricing methods in terms of accuracy, consistency, and reliability. Additionally, the study is expected to identify and highlight the most significant factors influencing car prices, providing valuable insights into pricing trends in the Kenyan market. The developed model will serve as a practical tool for car buyers, sellers, and other stakeholders, enabling them to make informed decisions and promoting transparency in the used car market. Furthermore, the study is expected to contribute to the field of machine learning by demonstrating the applicability of ensemble techniques in solving real-world problems, particularly in emerging markets like Kenya.

## **2.5 Theoretical frameworks**

**2.5.1 Hedonic Pricing Theory**

The **Hedonic Pricing Theory (HPT)** is the most suitable theory for this research on predicting the price of used cars in Kenya. The theory posits that the price of a good or service is determined by its individual attributes or characteristics (Hargrave, 2020). Hedonic Pricing Theory (HPT) is grounded in the characteristics theory of value, which posits that consumers derive utility not from goods themselves, but from the attributes they embody (Lancaster, 1966). Rosen (1974) extended this by arguing that goods are bundles of characteristics and consumers have varying preferences for these attributes. HPT, therefore, emphasizes that a good's price is determined by the value consumers place on its individual characteristics, allowing for the estimation of "implicit prices" for specific attributes. This framework provides a robust foundation for understanding how the diverse characteristics of a used car, such as brand, age, mileage, drive, engine capacity, fuel type, transmission, acceleration and aspiration collectively influence the car's market price. Each of these attributes contributes to the overall value of the car, and buyers and sellers assign varying values to them based on their preferences. This theory aligns perfectly with the research as it provides a foundation for analyzing how specific car features impact pricing. By identifying and quantifying these key attributes, the machine learning model can offer accurate and fair price predictions, making the **Hedonic Pricing Theory** central to understanding and developing the model for predicting used car prices in Kenya.

## **2.6 Conceptual framework**

In this section, the independent and dependent variables that will be used in the course of implementation of this project will be analyzed. Cumulatively, these factors will contribute to the end result of the algorithms that will be used in the ensemble framework.

**Output layer.**

**Dependent variable.**

Predicted used car price.

**Input layer.**

**Independent variables.**

Year

Make

Mileage

Transmission

Engine Capacity

Fuel type

Acceleration

Horsepower

Car name

Source (Kenyan/foreign used)

Annual insurance

Condition (Accident/Accident-free)

Body type

**Mediating variables.**

Year

***Figure 1. Conceptual framework.***

**Explanation of the labels**

Independent Variables: These are the input features for instance model, age and mileage that directly influence the price of the car.

Mediating variables: Represents the model year of the car or the year the car was sold. Year mediates the impact of age and mileage on the price because newer models generally depreciate differently compared to older models, even if they have the same mileage.

Dependent Variable: This is the output of the model (the predicted price of the used car.)

# **CHAPTER 3: RESEARCH METHODOLOGY**

## **3.1 Introduction**

This chapter outlines the research methodology to be employed in this study, which aims to develop an ensemble machine learning model for accurately predicting used car prices in the Kenyan market. Given the complex nature of the problem and the need for a practical, data-driven solution, this research will adopt the Design Science Research (DSR) paradigm. DSR emphasizes the creation and evaluation of artifacts (Ahmed, D., & Sundaram, 2011). In this case, the ensemble machine learning model is designed to solve identified problems in a real-world context. By iteratively designing, developing, and evaluating the model, this research addresses the challenges of inconsistent and subjective pricing prevalent in Kenya's used car market.

## **3.2 Design Science Research Paradigm and Pragmatism as a Paradigm in DSR**

DSR will provide a structured framework for conducting this research, focusing on developing and evaluating practical solutions to real-world problems. This study will align with the core principles of DSR, which include:

**Problem Identification and Motivation**

**Objectives for a Solution**

**Artifact Design and Development**

**Demonstration**

**Evaluation**

**Communication**

**figure 2. DSR principles**

**Pragmatism as a paradigm in DSR**

Pragmatism, with its emphasis on practical consequences, aligns perfectly with the core principles of Design Science Research (DSR) (Hevner et al., 2004). This philosophical framework encourages DSR researchers to prioritize the creation of artifacts that solve real-world problems and have a tangible impact (March & Smith, 1995). In the context of developing an ensemble model for predicting the price of used cars, a pragmatic approach would prioritize understanding the needs and perspectives of both buyers and sellers (user-centric design). This involves gathering desired features of a price prediction tool. Furthermore, a pragmatic approach emphasizes iterative development and experimentation, involving building and refining the ensemble model through multiple iterations and evaluating the model's performance against real-world data. This iterative process would involve experimenting with different combinations of machine learning algorithms, feature engineering techniques, and model parameters to optimize prediction accuracy and user satisfaction. Ultimately, a pragmatic approach ensures that the final model is not only accurate but also usable and valuable to its intended users by focusing on practical outcomes like providing clear and concise explanations of the predictions and integrating the model into existing platforms or applications used by buyers and sellers.

**3.2.1 Problem Identification and Motivation and Objectives of the Solution**

The problem identification and motivation, along with the objectives of the solution, have already been built upon the discussions presented in Chapter One under Section 1.2 (Problem Statement) and Section 1.3 (Objectives of the Study).

**3.2.2 Artifact Design and Development**

**3.2.2.1 Data Collection**

**Target population/Data source**

The target population for this study comprises used cars listed on Kenya's online motor bazaars: Kai and Karo, Motor Hub, Gigi, Maridady, Auto Box, Kenjap, and Khushi Motors and SBT (Japanese used cars). This population encompasses vehicles with diverse features and pricing, offering a broad representation of the used car market that can be used well in this study. Their structured format also reduces the risk of data inconsistencies, making preprocessing easier and more efficient.

**Data Features**

Key features to be collected include: Mileage, Year of production, Horsepower, Engine size, Seats, Price, Fuel type, Transmission, Car name, Source(Kenyan used or foreign used), Body Type, Annual insurance, and Condition(Accident or Accident-free)

**Data Collection Tools**

Python library, BeautifulSoup will be used to mine data: BeautifulSoup is a Python library widely used for web scraping and data mining, enabling developers to extract and process data from HTML or XML documents. It provides an easy-to-use interface for navigating, searching, and modifying webpage elements. Data Miner (software for **data mining)** will also be used to scrape the data.

**Programming Language**

Python will be used. Python is a versatile, high-level programming language widely used in data science and machine learning due to its simplicity, readability, and extensive ecosystem of libraries (Sohel, 2024). Libraries such as NumPy and pandas enable efficient data manipulation and analysis, while scikit-learn, TensorFlow, and PyTorch provide robust tools for building machine learning models. Python's clean syntax makes it beginner-friendly, while its active community ensures ample resources and support for solving problems. Additionally, Python is open-source, platform-independent, and seamlessly integrates with other tools, making it ideal for the entire data science workflow, from data collection to model deployment..

**Libraries**

* Data Collection**:** BeautifulSoup and Data Miner.
* Data Preprocessing**:** Pandas, NumPy and Scikit-learn.
* Modeling: Scikit-learn, XGBoost and SHAP.
* Evaluation and Visualization**:** Matplotlib and sklearn.

BeautifulSoup and Data Miner (Data Collection): These are essential for web scraping, allowing you to collect data from online platforms or sources where used car information is listed.

Pandas (Data Preprocessing): Pandas is used for data manipulation and analysis, helping you clean, structure, and handle datasets in a way that is easy to work with for modeling.

NumPy (Data Preprocessing): NumPy enables efficient numerical operations, helping with mathematical transformations and operations on your datasets, especially when dealing with large arrays or matrices.

Scikit-learn (Data Preprocessing & Modeling): Scikit-learn provides a wide range of tools for preprocessing, feature engineering, and implementing machine learning algorithms, making it a go-to library for building predictive models.

XGBoost (Modeling): XGBoost is a powerful machine learning algorithm for regression tasks, known for its high performance, speed, and accuracy, which is essential for used car price prediction model.

SHAP (Modeling & Evaluation): SHAP helps in explaining the model’s predictions by providing insights into feature importance and how each input affects the output, ensuring model transparency and trustworthiness.

Matplotlib (Evaluation & Visualization): Matplotlib is used to create visualizations like graphs and plots, which help in understanding model performance, feature relationships, and trends in your data.

**3.2.2.2 Data Preprocessing**

**Data Cleaning:**

Duplicate Removal: Records with identical feature values will be flagged and removed.

Handling Missing Values: Numerical data will be imputed using mean or median values where depending on the distribution of the data that will be collected. Categorical data will be handled appropriately based on their type: Binary variables: Will be encoded using binary encoding (0/1).Nominal variables: Will be encoded using one-hot encoding to avoid introducing an artificial order.

Outlier Detection**:** Boxplots, scatterplots, and interquartile ranges (IQR) will be used to identify and handle extreme values in numerical features.

**Feature Engineering**

**Derived Features:**

"Car Age" will be calculated as the difference between the current year and the year of production and "Mileage per Year" to normalize mileage based on car age.

**Categorical Encoding:**

One-hot encoding is a method used to convert nominal (categorical) variables into a numerical format suitable for machine learning algorithms. For variables like **fuel type** in a used car price prediction model, one-hot encoding will be used to ensure that the categorical information is represented numerically without introducing unintended relationships.

**Scaling and Normalization:**

StandardScaler will be applied to normalize numerical features, ensuring uniform feature scaling for distance-based models like KNN.

**Feature Selection**

1. Recursive Feature Elimination (RFE): To iteratively remove the least important features based on model performance.
2. Feature Importance from Random Forest and XGBoost: Features with low importance scores will be excluded. SHAP values will provide additional insights into feature relevance.

**3.2.2.3 Model Development**

**Base Models:**

**Random Forest**: In the Kenyan used car market, where a wide variety of brands, models, and conditions exist, Random Forest plays a critical role. This ensemble learning method constructs multiple decision trees and averages their predictions, making it effective at capturing complex interactions among features like make, model, year, mileage, and condition. Its ability to handle numerous variables and reduce the risk of overfitting compared to single decision trees makes Random Forest a reliable tool for predicting car prices in a diverse and dynamic market.

**Support Vector Machines (SVM)**: The Kenyan used car market likely exhibits non-linear relationships between factors and car prices such as depreciation, which may not follow a simple linear pattern. Support Vector Machines (SVM) excel in modeling such non-linear relationships by identifying the optimal hyperplane or decision boundary that separates data points. This ability to handle complex patterns makes SVM a powerful tool for generating more accurate car price predictions.

**K-Nearest Neighbors (KNN)**: K-Nearest Neighbors (KNN) is particularly useful in the Kenyan context because it effectively captures localized price trends. In a market where regional demand and supply heavily influence car prices, KNN predicts a car's price by examining the prices of its closest neighbors based on features like mileage, year, and make. By focusing on local patterns, KNN ensures that regional price variations are reflected in its predictions.

**Gradient Boosting**: Gradient Boosting stands out in the Kenyan used car market by effectively capturing the intricate interplay of factors that influence prices. This ensemble method builds a sequence of weak learners typically decision trees where each learner corrects the errors of its predecessor. The result is a highly accurate model capable of understanding dynamic and diverse pricing trends, empowering both buyers and sellers to make well-informed decisions.

**Experimentation Process**

The development of the used car price prediction model will follow a structured experimentation process:

**Data Extraction**

Data will be extracted from various Kenyan online car marketplaces: Kai and Karo, Motor Hub, Gigi, Maridady, Auto Box, SBT, Kenjap, and Khushi Motors using BeautifulSoup and Data Miner. The data will include key features: Mileage, Year of production, Horsepower, Engine size, Seats, Price, Fuel type, Transmission, Car name, Source(Kenyan used or foreign used) Body Type, Annual insurance, Condition(Accident or Accident-free).

**Data Conversion**

The collected raw data will be cleaned and preprocessed using Pandas and NumPy. Duplicate records will be identified and removed to prevent data redundancy. Missing values, which can significantly impact model performance, will be addressed. We will first identify missing values using Pandas functions like “isnull()” and “sum()”. Imputation strategies will then be applied: for numerical features, mean or median imputation will be used, while mode imputation will be employed for categorical variables. If a feature exhibits a high percentage of missing values for instance 30%, it will be considered for removal. Similarly, rows with excessive missing values may also be dropped. For categorical variables, we will use one-hot encoding. This technique creates separate binary columns for each category, where a value of 1 indicates the presence of that category and 0 indicates its absence. For example, if the “Fuel Type” feature has categories such as “Petrol,” “Diesel,” and “Electric,” one-hot encoding will generate three new columns one for each fuel type. A car with a petrol engine would have a 1 in the “Petrol” column and 0 in the others, while a diesel car would have a 1 in the “Diesel” column and 0 elsewhere. This transformation ensures that categorical data is structured in a format suitable for machine learning models without introducing any unintended ordinal relationships

**Feature Selection**

Feature selection techniques, including correlation analysis and importance ranking (via Random Forest), will be applied to identify and retain the most relevant features. This step ensures that only significant factors are used for model training, enhancing performance, and reducing computational complexity.

**Model Selection and Application**

After feature selection, various machine learning algorithms will be tested to identify the best-performing models:

1. Random Forest: This ensemble learning method captures intricate interactions among multiple features, effectively predicting prices in the complex used car market.

2. Support Vector Machines (SVM): SVM will model non-linear relationships by finding optimal decision boundaries, thereby improving prediction accuracy.

3. K-Nearest Neighbors (KNN): KNN will predict car prices based on similarities with nearby listings, effectively capturing regional price differences.

4. Gradient Boosting: This method will sequentially correct the errors of weak learners, refining predictions and adapting to dynamic pricing trends.

**Ensemble Strategy**

To make the predictions even more accurate, the research will use the Weighted Averaging technique to combine the price predictions from each of the individual models above, giving more weight to the models that have performed best in the past.

**Model Interpretability**

To make the model understandable and trustworthy, we will use a technique called SHAP (SHapley Additive exPlanations). This will help in understanding how each feature for instance mileage, year or make influences the price. For example, one can see how much a higher mileage affects the price. Secondly, to identify the most important factors that drive used car prices in the Kenyan market. Lastly, to make the model more transparent and easier for users to understand and trust.

**3.2.3 Demonstration**

The developed ensemble model will be applied to a real-world dataset of used car listings from Kenyan online marketplaces. This demonstration will assess the model's ability to predict prices accurately based on the input features and evaluate its practical applicability in the Kenyan market.

**3.2.4 Evaluation**

**Evaluation Metrics**

The following metrics will be used to assess model performance: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and R-squared (R²):

RMSE (Root Mean Squared Error) is a widely used metric for assessing the accuracy of predictions in regression machine learning tasks. It measures the deviation of predicted values from the actual values using the Euclidean distance. Since RMSE involves squaring the errors, it is sensitive to large prediction errors, which can disproportionately increase its value. A lower RMSE indicates greater accuracy in the model's predictions. The formula for calculating RMSE is provided below, where “ŷ” represents the predicted value and “y” represents the actual value.

R-squared(R2) is another commonly used metric in regression machine learning, often described as a normalized version of MSE (the square of RMSE). It remains unaffected by the scaling of values in the dataset, focusing instead on the proportion of errors relative to the average value. This means that the absolute magnitude of errors does not influence R-squared, only their relative size. Like RMSE, R-squared can also be affected by outliers or unusually large errors. Its values range from 0 to 1, with 1 indicating perfect prediction accuracy, where the model's predictions completely align with the actual data. Higher R-squared values suggest better model performance.

MAPE (Mean Absolute Percentage Error) is another metric used to evaluate prediction accuracy. It calculates the absolute percentage error between predicted and actual values for each instance, then averages these errors to produce the MAPE value. Unlike RMSE and R-squared, MAPE does not square the errors, and the individual errors depend on the relative magnitude of the actual and predicted values. A MAPE of zero indicates perfect predictions, while lower values signify better model performance. (GeeksforGeeks, 2024).

**Cross-Validation**

K-Fold Cross-Validation will be used: It ensures robust performance evaluation by splitting the data into training and testing subsets multiple times based on the size of data that will be collected. 80% of the data will be used for training and 20% of the data will be used for testing.

**Benchmarking**

The ensemble model’s performance will be benchmarked against: Individual base models and Simplistic baselines such as mean or median predictions.

**Expected Outcomes**

The primary objective is for the ensemble model to demonstrate superior performance compared to individual machine learning algorithms; KNN, SVM, Random Forest and XGBoost, in terms of both accuracy and robustness. By combining the strengths of these diverse models, the ensemble approach aims to achieve more reliable and consistent price predictions. Furthermore, the integration of SHAP (SHapley Additive exPlanations) will provide valuable insights into the key factors that most significantly influence used car prices. This knowledge will empower stakeholders, such as car buyers, sellers, and online marketplaces, to make more informed pricing decisions and develop data-driven pricing tools.

**3.2.5 Communication of Results**

The research findings will be disseminated through: Academic Publicationswhich is submission of papers to relevant journals, conference presentationsandstakeholders who are the car dealers, industry associations, and policymakers.

## **3.3 Ethical Considerations**

Ethical practices are crucial throughout the research process. Data scraping will be conducted responsibly, adhering to the terms of service of the online platforms used. Any sensitive or proprietary information encountered will be anonymized to protect user privacy and intellectual property rights. Furthermore, the research findings will be presented transparently, acknowledging any potential biases or limitations of the model. This transparency fosters trust and ensures responsible use of the developed pricing model.

# **REFERENCES**

Africa, O. (2024). *Surging Prices and Shifting Trends: The Challenges Facing Kenya’s Second-Hand Car Market*. OfficePhase Africa. https://officephase.com/2024/02/12/surging-prices-and-shifting-trends-the-challenges-facing-kenyas-second-hand-car-market/#:~:text=The%20Kenya%20Auto%20Bazaar%20Association,diminished%20purchasing%20power%20of%20households.

Africa, O. (2024). *Surging Prices and Shifting Trends: The Challenges Facing Kenya’s Second-Hand Car Market*. OfficePhase Africa. https://officephase.com/2024/02/12/surging-prices-and-shifting-trends-the-challenges-facing-kenyas-second-hand-car-market/

Ahtesham, M., & Zulfiqar, J. (2022). Used Car Price Prediction with Pyspark. In *Lecture notes in networks and systems* (pp. 169–179). https://doi.org/10.1007/978-3-031-01942-5\_17

Amik, F. R., Lanard, A., Ismat, A., & Momen, S. (2021). Application of machine learning techniques to predict the price of Pre-Owned cars in Bangladesh. *Information*, *12*(12), 514. https://doi.org/10.3390/info12120514

AnalytixLabs. (2023, December 26). Random Forest Regression — How it Helps in Predictive Analytics? *Medium*. https://medium.com/@byanalytixlabs/random-forest-regression-how-it-helps-in-predictive-analytics-01c31897c1d4

Bhatnagar, P., Lokesh, G. H., Shreyas, J., Flammini, F., & Gautam, S. (2024). An Analysis of Car Price Prediction using Machine Learning. *An Analysis of Car Price Prediction Using Machine Learning*, *4*, 11–15. https://doi.org/10.1145/3674029.3674032

*Blog Post | Peach Cars*. (n.d.). https://peachcars.co.ke/blog/challenges-of-selling-your-car-in-kenya-and-how-to-overcome-them

Budiono, D. A., Utomo, K. S., Wibowo, K. J., & Wiradinata, M. J. (2024). Used Car Price Prediction Model: A Machine Learning approach. *Budiono | International Journal of Computer and Information System (IJCIS)*. https://doi.org/10.29040/ijcis.v5i1.147

Bukvić, L., Škrinjar, J. P., Fratrović, T., & Abramović, B. (2022). Price prediction and classification of Used-Vehicles using supervised Machine Learning. *Sustainability*, *14*(24), 17034. https://doi.org/10.3390/su142417034

*Climate Change 2022: Impacts, adaptation and Vulnerability*. (n.d.). IPCC. https://www.ipcc.ch/report/ar6/wg2/

GeeksforGeeks. (2024a, February 26). *Regression in machine learning*. GeeksforGeeks. https://www.geeksforgeeks.org/regression-in-machine-learning/

GeeksforGeeks. (2024b, July 29). *Regression metrics*. GeeksforGeeks. https://www.geeksforgeeks.org/regression-metrics/

Han, L., Yin, Z., Xia, Z., Tang, M., & Jin, R. (2020). Price Suggestion for Online Second-hand Items with Texts and Images. *Proceedings of the 30th ACM International Conference on Multimedia*, *4*, 2784–2792. https://doi.org/10.1145/3394171.3413759

Hao, L., Umar, M., Khan, Z., & Ali, W. (2020). Green growth and low carbon emission in G7 countries: How critical the network of environmental taxes, renewable energy and human capital is? *The Science of the Total Environment*, *752*, 141853. https://doi.org/10.1016/j.scitotenv.2020.141853

Hargrave, M. (2020, November 18). *Hedonic pricing: definition, how the model is used, and example*. Investopedia. https://www.investopedia.com/terms/h/hedonicpricing.asp

*Kenya used car market size | Mordor Intelligence*. (n.d.). https://www.mordorintelligence.com/industry-reports/kenya-used-car-market

Mobility Foresights. (2024a, January 22). *Kenya Used Car market 2024-2030*. https://mobilityforesights.com/product/kenya-used-car-market/

Mobility Foresights. (2024b, January 22). *Kenya Used Car market 2024-2030*. https://mobilityforesights.com/product/kenya-used-car-market/

News Desk. (2022, January 5). Used Car imports up in Kenya. *Trendsnafrica | 24/7 Africa News*. https://trendsnafrica.com/used-car-imports-up-in-kenya/

Pal, N., Arora, P., Kohli, P., Sundararaman, D., & Palakurthy, S. S. (2018). How much is my car worth? A methodology for predicting used cars’ prices using Random Forest. In *Advances in intelligent systems and computing* (pp. 413–422). https://doi.org/10.1007/978-3-030-03402-3\_28

Pillai, A. S. (2022, June 16). *A deep learning approach for used car price prediction*. Journal of Science & Technology. https://www.thesciencebrigade.com/jst/article/view/140

*Prediction of Prices Car Price Prediction with Machne Learning*. (2022, October 6). IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/abstract/document/9995772?casa\_token=25vLVXNinokAAAAA:n7PAYqIyFcNtMSN7\_Gwr8ELH0UZUrozxa3qWrBXZZ23q7b8pUJQ9NoeU9oEfZ0LdB6HzWeQQzD93oSrc

*Prediction of used car prices using artificial neural networks and machine learning*. (2022a, January 25). IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/abstract/document/9740817

*Prediction of used car prices using artificial neural networks and machine learning*. (2022b, January 25). IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/abstract/document/9740817/references#references

*Price prediction of used cars using machine learning*. (2021, November 22). IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/abstract/document/9696839?casa\_token=9f4-JFiGMmQAAAAA:t-SChgZjsuVlUeR0MgBqPNQa8i4FSYoO40Mj1wGFxyUwvn1xYsGnWRpuZLCQ008eD1C1Ia6gcBtaIVHr

Tang, J., Gong, R., Wang, H., & Liu, Y. (2023). Scenario analysis of transportation carbon emissions in China based on machine learning and deep neural network models. *Environmental Research Letters*, *18*(6), 064018. https://doi.org/10.1088/1748-9326/acd468

*Used Car Price Prediction using Machine Learning: A Case Study*. (2022, May 18). IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/abstract/document/9800719?casa\_token=HHOMbnUT60kAAAAA:LFcinEBfGYrN2NtKl\_VLFyVPpOhYglEfaNbASBS4534mm2--0YoT1vgo0d4nFUxeCmBsH\_Uboj4400Kr

Wikipedia contributors. (2024, October 2). *Random forest*. Wikipedia. https://en.wikipedia.org/wiki/Random\_forest

Yu, W., Xia, L., & Cao, Q. (2024). A machine learning algorithm to explore the drivers of carbon emissions in Chinese cities. *Scientific Reports*, *14*(1). <https://doi.org/10.1038/s41598-024-75753-y>

Doctor, A. (2024, October 28). The evolution of car technology in Kenya: Past, present, and future. Garage in Nairobi | Automotive Doctor Motor Garage Limited. <https://automotivedoctor.co.ke/the-evolution-of-car-technology-in-kenya-past-present-and-future/>

# **APPENDICES**

**Appendix A: Budget**

|  |  |
| --- | --- |
| **Item** | **Budget Amount** |
| Operating costs | KES 20,000 |
| Travel costs | KES 10,000 |
| Conference costs | KES 20,000 |
| Publication | KES 35,000 |
| Contingency | KES 20,000 |
| Total | KES 105,000 |

**Appendix B: Gantt chart**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **TASK** | **January** | **February** | **march** | **April** | **May** | **June** | **July** |
| **1.Initial planning** |  |  |  |  |  |  |  |
| Objectives &Lit review |  |  |  |  |  |  |  |
| **2.Proposal development** |  |  |  |  |  |  |  |
| Methodology  design |  |  |  |  |  |  |  |
| Proposal writing |  |  |  |  |  |  |  |
| **3.Data collection** |  |  |  |  |  |  |  |
| Dataset collection |  |  |  |  |  |  |  |
| Feature selection and engineering |  |  |  |  |  |  |  |
| Data preprocessing |  |  |  |  |  |  |  |
| **4.Algorithm development** |  |  |  |  |  |  |  |
| Model selection |  |  |  |  |  |  |  |
| Ensemble algorithm design |  |  |  |  |  |  |  |
| Model training and testing |  |  |  |  |  |  |  |
| Performance evaluation |  |  |  |  |  |  |  |
| **5.Writing and dissemination** |  |  |  |  |  |  |  |
| Draft report |  |  |  |  |  |  |  |
| Final report |  |  |  |  |  |  |  |
| Presentation preparation |  |  |  |  |  |  |  |
| Dissemination |  |  |  |  |  |  |  |