A Comparative Analysis of Machine Learning Algorithms for the Prediction and Classification of Prices of Used Vehicles in Australia.

Man Chung Chan
Faculty of Business
Humber College
Toronto, Canada
manbentenchan@gmail.com

Yan Pui Siu
Faculty of Business
Humber College
Toronto, Canada
halliesiu@gmail.com

Anjali Patel
Faculty of Business
Humber College
Toronto, Canada
anjali.patel199899@gmail.com

Mathini Kanagaratnam
Faculty of Business
Humber College
Toronto, Canada
mathini.kanex@gmail.com

Sairohit Chowdhary
Faculty of Business
Humber College
Toronto, Canada
sairohitchowdhary@gmail.com

***Abstract*—The pandemic and post-pandemic era have caused a significant fluctuation in the price of used vehicles, leaving prospective buyers and sellers in Australia perplexed about their next steps. The aim of this study is to use a machine learning-based approach to predict the price of a used vehicle and classify the prices as fair or not. Multiple Linear Regression, K-Nearest Neighbour (K-NN), Naive Bayes, Classification and Regression Trees, and Logistic Regression models are being used in this study and evaluated against each to determine the best accuracy of price prediction. Among all the models, K-NN has achieved the highest accuracy of 91.7% in distinguishing fair vehicle prices. The model can be useful not only for potential buyers and sellers but also for manufacturers and dealers, who can adjust their plans and be ahead of their competitors based on the predictions.**

Keywords—Used vehicles, Multiple Linear Regression, K-Nearest Neighbour, Naive Bayes, Classification and Regression Trees, Logistic Regression

I. INTRODUCTION

During the pandemic from around March 2020 to early 2021, the prices of used cars decreased initially due to economic uncertainty and restricted movement reducing demand. However, the demand for used cars increased as people avoided public transportation and started commuting in their vehicles, leading to a shortage of inventory and higher prices. Moreover, disruptions in the supply chain due to border closures and reduced international trade caused a shortage of new cars, further driving up the prices of used cars. Dealers faced a shortage of vehicles and had to offer competitive buy-backs to former customers to fill their lots [1].

As Australia began to recover from the pandemic in mid-2021, the prices of used cars stabilized and even decreased slightly in some regions as more new cars became available, and consumer demand for used cars shifted towards new ones. A report from financial intelligence company Moody's Analytics revealed that used car prices are now below their level a year ago, although they are still significantly higher than before the pandemic. The constant fluctuations in prices between the pandemic and post-pandemic period have a significant impact on customers' affordability, availability, and trade-in values, causing uncertainties. Rising interest rates, shorter waiting times for new cars, and increased acceptance of delivery delays have contributed to the decline in used-car prices [2].

The objective of this study is to develop a machine-learning model that can predict the price of pre-owned vehicles, which can help customers to determine a fair price in a market where prices are volatile. By using this model, buyers can avoid paying an excessively high price for a used vehicle, and sellers can set a suitable price based on their urgency to sell. As a result, both parties can save time and effort while selling or searching for second-hand vehicles.

Additionally, the model can also demonstrate how the value of used vehicles decreases over time, enabling customers to make decisions about which model to purchase if they plan to sell it in the future. This can also help car manufacturers such as Hyundai, Toyota, and Honda to determine which models they should focus on producing in the used car market to stay competitive.

The paper is structured in the following manner: Section II contains the literature survey related to the field of used car price prediction. Section III covers the description of the dataset. In Section IV, the methodology of the study was proposed. Section V covers the techniques used for data pre-processing. Section VI describes each model and Section VII elaborates on the examination of the performance of each model for price prediction of the used cars. Section VIII makes graphic comparisons between algorithms while section IX states the limitation of this paper. Finally, section X specifies the conclusion.

II. Literature Review

Many attempts have been made by researchers in the past to predict used cars across countries using Machine Learning. However, there was limited published research on price prediction in Australia.

In a study titled "Housing Price Prediction Using Machine Learning Algorithms: The Case of Melbourne City, Australia", Phan [3] used machine learning techniques to analyze historical data of house transactions in Australia, resulting in the development of a model capable of predicting property prices. The study found a significant price gap between expensive and affordable suburbs in Melbourne. Although the study focused on property prices which is different from the project, the regression models and methodology could be used as a reference for other topics. The author examined various property-related factors, such as location and condition, to construct the prediction model based on these features. Some features or indicators used in the study could also be applied to used cars.

Numerous studies have been conducted by authors using datasets from various countries to forecast prices for pre-owned vehicles. Samruddhi and Kumar [4] collected historical data from the internet and applied a K-nearest neighbor-based model for analysis in their research entitled "Used Car Price Prediction using K-Nearest Neighbor Based". They experimented with different ratios of train and validation data to obtain various outcomes, with their best outcome achieving an 85% prediction accuracy. To validate their model, they used the K Fold Method with 5 and 10 folds and also applied linear regression, which resulted in a prediction accuracy of 71%. Although this accuracy is not necessarily low, other models may be more appropriate. The authors suggest that in addition to KNN and linear regression, other models should also be considered.

Venkatasubbu and Ganesh [5] proposed Used Car price prediction using supervised learning techniques such as ANOVA, Lasso Regression, Regression Tree, and Tukey test. According to the authors, the prediction rate for all the models was well under the accepted 5% error. The mean error of the regression tree model was found to be more than the mean error rate of the multiple regression and lasso regression models.

Çelik and Osmanoğlu [6] proposed a Prediction of The Prices of Second-Hand Cars in 2019. They used car data from the internet to establish a model. They mainly employed linear regression and split the data at different ratios (70-30% and 80-20%). Their best R-square outcome was 89.1 %, which is relatively high. They applied an R-square to indicate the model's performance, which seemed better than prediction accuracy since it could use more information about the model. However, their study focused only on three factors: price, model, and year of production. In this case, there is some room for improvement; other aspects like mileage and transmission can also be considered.

Hankar, et al. [7] used data gathered from an e-commerce website in Morocco to conduct their study. They employed a multiple linear regression model as a baseline and compared its results with four other models, including K-nearest neighbors regressor (KNN), random forest regressor (RFR), gradient boosting regressor (GBR), and artificial neural network (ANN) based regressor. The models were trained on 80% of the entire dataset, while the remaining 20% was allocated for testing purposes. The findings revealed that the gradient boosting regressor performed the best, achieving the highest R2 score and the lowest root mean squared error among all models tested.

III. Source Of The Data and Description Of The Dataset

The Australian automobile market dataset was obtained from Kaggle [8], and originally collected from Autotrader Australia [9]. This dataset is classified as a secondary source and contains 16 variables and 17,048 records.

Table I: Description Of Dataset

Variable Name	Data Type	Description
ID	Numerical	Vehicle ID
Name	Nominal	Vehicle Name
Price	Numerical	Price of Vehicle in AUD
Brand	Nominal	Brand of Vehicle
Model	Nominal	Model of Vehicle
Variant	Nominal	Variant of Vehicle
Series	Nominal	Series of Vehicle
Year	Numerical	Manufacturing Year
Kilometers	Numerical	Total Distance Driven
Type	Nominal	Type of Vehicle
Gearbox	Nominal	Type of Gearbox
Fuel	Nominal	Type of Fuel Used
Status	Nominal	New, Used, or Demo
CC	Numerical	Size of Engine in Cubic Centimeters
Color	Nominal	Color of Vehicle
Seating Capacity	Numerical	Number of Seats

IV. Objectives and Methodology

This section describes the process we followed in conducting a comparative analysis of machine learning algorithms for the prediction and classification of prices of used vehicles in Australia. Our methodology involved data collection, pre-processing, data splitting, model building, and performance evaluation. All the data pre-processing, model building, and analysis steps were performed in Python. By following a systematic methodology, our primary objective is to provide insights into the most effective machine learning algorithm for price prediction and classification in the used car market, which can aid customers in determining a reasonable or fair price.

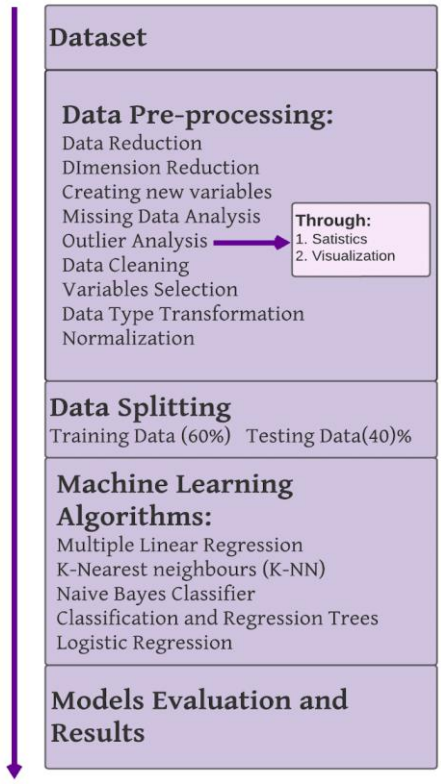

Fig. 1. Methodology.

V. Data Preprocessing

Data pre-processing is a critical step in machine learning that involves preparing and cleaning data before it can be used for analysis or modeling. The methods and techniques employed in this process aim to ensure the consistency, accuracy, and suitability of data for analysis, thereby enhancing the accuracy and performance of models.

A. Data Reduction

Data reduction refers to the process of decreasing the size of the dataset while retaining important information. For our analysis, which focuses solely on price prediction and classification in the used car market, we removed the records of new and demo cars in the "Status" variable.

TABLE II: NUMBER OF ROWS BEFORE AND AFTER DATA REDUCTION

Number of Rows Before Data Reduction	Number of Rows After Data Reduction
17,048	16,304

B. Dimension Reduction

Following the data reduction process, all remaining records in the "Status" variable pertained to used vehicles. Consequently, we removed the "Status" variable to reduce the dimensionality of the dataset.

C. Creating New Variables

We created a new variable called "Age," which subtracts the manufacturing year "Year" from the current year of 2023 to determine the age of the vehicles in the dataset. The age of a vehicle can be a crucial factor with the more relevant and informative feature in determining its market value, as it provides a measure of how long the vehicle has been in use. While the manufacturing year "Year" does not consider the current year or the duration of the vehicle's use accurately in the machine learning models.

In addition to the original variables in the Australian automobile market dataset, we created a new variable called "Price_Classification" to classify the prices of used vehicles as either fair or unfair based on industry standards. Our classification threshold was based on data released by Cox Automotive Australia's data solutions division in June 2022, which indicated that the average listed price of used vehicles on dealer websites was $39,279 [10]. Specifically, prices above $39,279 were classified as "Not Fair," while prices at or below $39,279 were classified as "Fair." We used the "Price_Classification" variable as an output for classification machine learning algorithms, such as K-Nearest Neighbors (K-NN), Naïve Bayes Classifier, Classification Trees, and Logistic Regression.

TABLE III: DESCRIPTION OF NEW VARIABLES CREATED

New Variable Name	Data Type	Description
Age	Numerical	Age of Vehicle
Price_Classification	Nominal	Fair or Not Fair
Price_Classificatio_ Numerical	Numerical	0 = Fair, 1 = Not Fair

D. Missing Data Analysis

After examining the dataset for the Australian automobile market with the newly introduced variables, it was determined that no missing values were present in the dataset.

E. Exploring Variables With Statistics

TABLE IV: DESCRIPTIVE STATISTICS

	Mean	SD.	Min.	Max.	Median
Price	35,638.71	29,747.38	1,000	999,000	29,888
Year	2015.2	4.62	1989	2022	2016
Kilometers	107,796.9	79,181.73	5	2,700,000	92,488.5
CC	2,507.48	889.22	875	7,300	2,359
Seating_ Capacity	5.11	1.13	2	14	5
Age	7.81	4.62	1	34	7

Based on the descriptive statistics, we observed that the "Price" variable has a wide standard deviation of $29,747.38, indicating a large variation in prices. Moreover, the median price of $29,888 is lower than the mean, suggesting that the price

distribution is skewed to the right. There are more cars sold at lower prices than at higher prices.

As for the "Year" and "Age" variables, their standard deviation of 4.62 years suggests some variation in the age of the cars. The median and mean values are similar, implying that the distribution of years and age is relatively symmetrical.

The "Kilometers" variable has a large standard deviation of 79,181.73, implying that there is a wide variation in the mileage of the cars. Moreover, the median value of 92,488.5 kilometers is lower than the mean, indicating a right-skewed distribution. Even after removing the record of new vehicles, the presence of a car with only 5 kilometers (minimum value) indicates the possibility of some new vehicles in the dataset. On the other hand, the maximum value of 2,700,000 kilometers indicates an extreme case.

The standard deviation of 889.22 for the "CC" variable indicates that there is variability in the engine size of the vehicles in the dataset.

F. Exploring Variables with Graphs

After our initial findings that outliers and extreme values are present in the variables "Price", "Kilometers", and "CC", we construct histograms to delve deeper into the skewness and extreme values within their distributions.

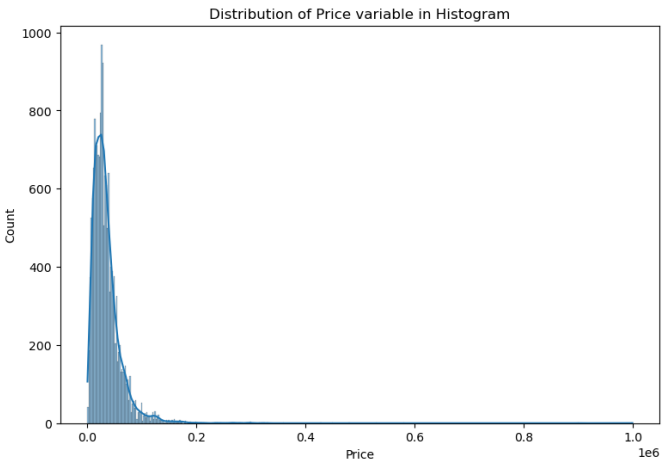

Fig. 2. Distribution of price variable.

We can see that the variable "Price" exhibits positive skewness. Most of the prices are distributed on the left side of the mean while there are outliers distributed on the right.

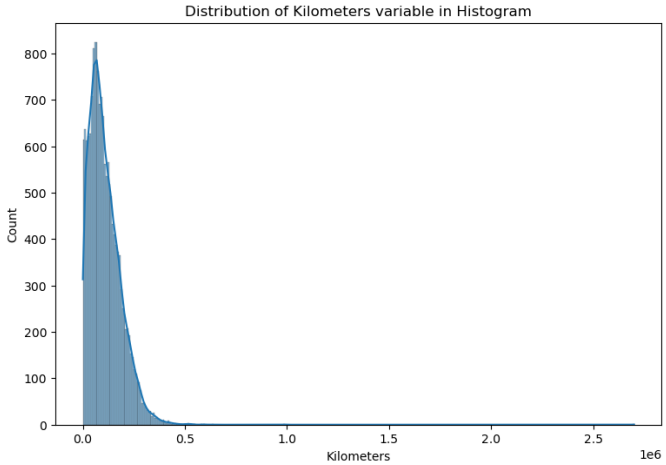

Fig. 3. Distribution of kilometers variable.

We can see that the variable “Kilometers” exhibits a positive skewness again. Most of the kilometers are distributed on the left side of the mean while there are outliers distributed on the right.

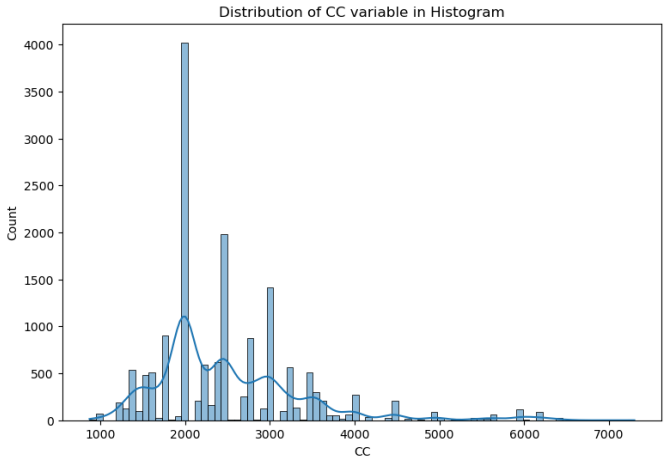

Fig. 4. Distribution of CC variable.

We can see that the variable “CC” is slightly skewed to the right with a long tail on the right-hand side.

G. Outlier Analysis

We used boxplots to identify outliers and extreme values that are significantly different from the rest of the data in the dataset.

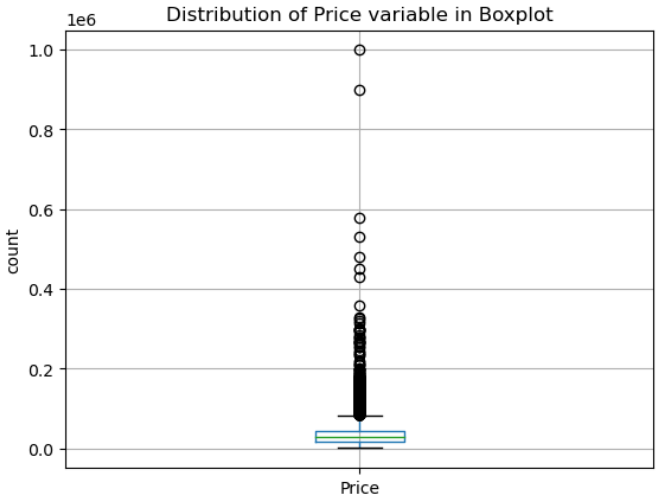

Fig. 5. Boxplot of price variable.

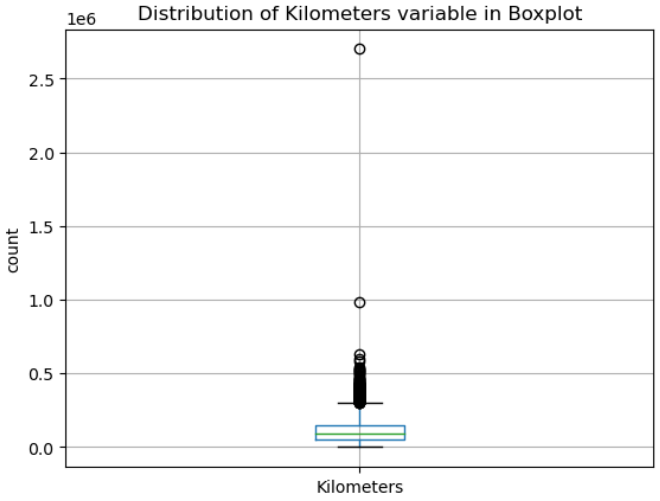

Fig. 6. Boxplot of kilometers variable.

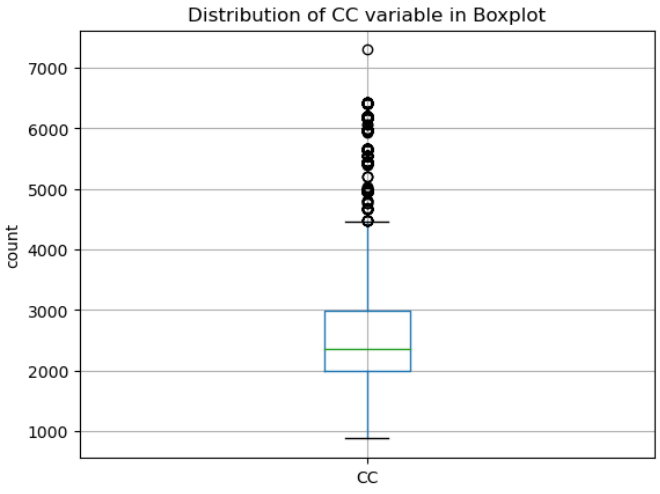

Fig. 7. Boxplot of CC variable.

Once we have gained an understanding of how variables are distributed in histograms and boxplots, we can spot outliers and extreme values in our data. To define outliers precisely, we employ the interquartile range.

$$Interquartile\ Range\ (IQR) = Q3 - Q1 \quad (1)$$

$$Lower\ Bound = Q1 - (1.5 \times IQR) \quad (2)$$

$$Upper\ Bound = Q3 + (1.5 \times IQR) \quad (3)$$

TABLE V: INTERQUARTILE RANGE

	Price	Kilometers	CC
Interquartile Range	26,415	101,414.75	991
Lower Bound	-21,632.5	-101,536.875	504.5
Upper Bound	84,027.5	304,122.125	4,468.5

H. Data Cleaning

Values exceeding the upper bound or falling below the lower bound of the "Price", "Kilometers", and "CC" variables were considered outliers, and we eliminated the rows containing these outliers from our dataset by omission. With this omission of outliers, our dataset is now cleaned and ready for further analysis.

TABLE VI: NUMBER OF ROWS BEFORE AND AFTER OMISSION

Number of Rows Before Omission	Number of Rows After Omission
16,304	14,894

I. Variable Selection

The process of variable selection aims to identify the most important features in a dataset, reduce the dataset's dimensionality and enhance the performance of the model.

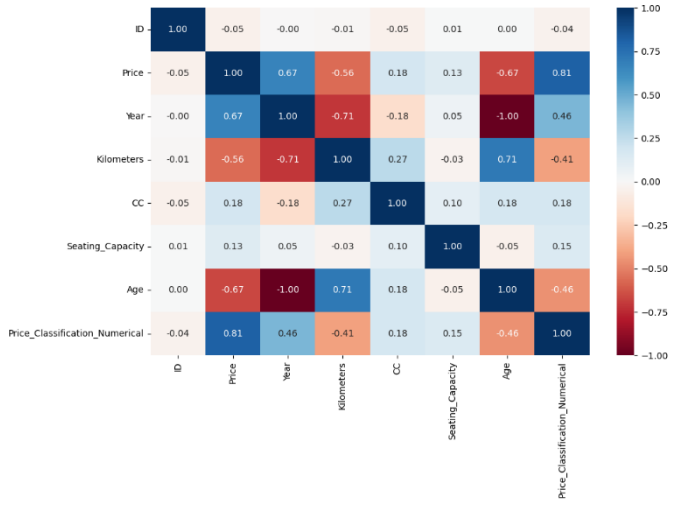

Fig. 8. Heatmap for variable selection.

A moderate positive correlation (+0.67) exists between Price and Year, whereas Price has a moderate negative correlation (-0.56) with Kilometers and a moderate negative correlation (-0.67) with Age. Since the Year variable does not consider the current year or the vehicle's usage duration accurately in machine learning models, we included Age instead of Year in our models.

After reviewing the heatmap and relevant literature, we have selected the most important variables from the dataset for our model. These variables include "Brand", "Kilometers", "Type", "Gearbox", "Fuel", "CC", "Seating_Capacity", and "Age". We used an exhaustive search approach to identify the best subset of predictors by fitting regression models. We evaluated all subsets of predictors using both the adjusted R square and Akaike Information Criterion (AIC).

$$R^2_{adj} = 1 - \frac{n-1}{n-p-1}(1-R^2) \quad (4)$$

$$ACI = n \ln\left(\frac{SSE}{n}\right) + n(1+\ln(2\pi)) + 2(p+1) \quad (5)$$

	n	r2adj	AIC	Age	Brand	CC	Fuel	Gearbox	Kilometers	Seating_Capacity	Type
0	1	0.448747	194492.597124	True	False	False	False	False	False	False	False
1	2	0.543815	192802.048654	True	False	False	True	False	False	False	False
2	3	0.601011	191605.950796	True	False	True	True	False	False	False	False
3	4	0.663837	190075.874072	True	False	True	True	False	True	False	False
4	5	0.667053	189990.972842	True	False	True	True	False	True	True	False
5	6	0.669152	189935.450827	True	True	True	True	False	True	True	False
6	7	0.669285	189932.857835	True	True	True	True	False	True	True	True
7	8	0.669315	189933.046185	True	True	True	True	True	True	True	True

Fig. 9. Exhaustive search.

We found that the model's adjusted R square was the highest when eight predictors were used. Additionally, the AIC suggested that a model with eight predictors was a good fit.

Therefore, in this paper, we employed "Brand", "Kilometers", "Type", "Gearbox", "Fuel", "CC", "Seating_Capacity", and "Age" as predictors for all the machine learning algorithms.

J. *Data Type Transformation*

In this paper, several techniques were used to transform data types, including encoding, normalization, and creating dummy variables. Additionally, new columns were created to categorize numerical variables into distinct groups based on the second-hand car market for Naïve Bayes Classifier.

TABLE VII: DATA TYPE TRANSFORMATION

	Input Variables Requirement	Techniques Used
Multiple Linear Regression Algorithm	Numerical	Encoding, Normalization
K-Nearest Neighbours (K-NN) Algorithm	Numerical	Encoding, Normalization
Naïve Bayes Classifier Algorithm	Categorical	Creating Dummies and New Class
Classification and Regression Trees Algorithm	Categorical / Numerical	Encoding
Logistic Regression	Numerical	Encoding, Normalization

The above data pre-processing activities will improve the accuracy and speed of machine learning algorithms. By removing outliers and inconsistencies from the data, data preprocessing helps to lower the complexity of machine learning models.

VI. MODELING

For the Australian automobile market dataset, we split the data into two sets: 60% for training and 40% for testing. The training data is used to assess the model's ability to detect overfitting, while the testing data is used to evaluate the model's predictive accuracy. This paper employs predictive analytics and supervised machine learning techniques to predict a single variable, namely "Price", through both classification and prediction. Specifically, we consider a predictive model based on Linear Regression, while our classification models include K-Nearest Neighbours (K-NN), Naïve Bayes Classifier, Classification Trees, and Logistic Regression.

A. *Multiple Linear Regression Algorithm*

The Multiple Linear Regression Algorithm is a supervised machine-learning technique used for prediction. The objective of the algorithm is to optimize predictive accuracy and predict the outcome value of new records based on the input values. In our case, Y is the Price of used vehicles while predictors include Age, Brand, CC, Fuel, Gearbox, Kilometers, Seating Capacity, and Type. The output of multiple linear regression analysis is a set of coefficients that represent the relationships between the independent variables and the dependent variable with the following formula:

$$Y = \beta 0 + \beta 1X1 + \beta 2X2 + \cdots + \beta pXp + \varepsilon \quad (6)$$

Where β0,β1, ... βp denote the coefficients while ε represents the noise or unexplained part.

In the equation, each coefficient represents how much the Price variable changes when the corresponding independent variable changes by one unit, while all other independent variables remain constant. The size of the coefficient shows how strong the relationship is between the independent variable and Price. If the coefficient is high, it suggests that the independent variable has a significant impact on the dependent variable, whereas a small coefficient indicates a weaker impact.

TABLE VIII: REGRESSION MODEL

Predictors	Coefficients
Brand	54.619033
Kilometers	-0.095292
Type	15.81328
Gearbox	-223.055852
Fuel	-2471.210165

CC	7.541284
Seating_Capacit	787.023024
Age	-1618.412516

In our case, the coefficient for Brand is 54.619, which means that for every unit change in the Brand variable, the predicted Price of the car increases by $54.62, holding all other independent variables constant. Similarly, the coefficient for Kilometers is -0.095, which means that for every unit increase in the Kilometers variable, the predicted Price of the car decreases by $0.10, holding all other independent variables constant.

The provided table indicates that there is a noteworthy positive correlation between a car's price and its brand and seating capacity. On the other hand, the variables of Fuel and Age demonstrate a strong negative correlation with the car's price.

B. K-Nearest Neighbours (K-NN) Algorithm

The k-Nearest Neighbours Algorithm is a supervised machine-learning technique used for classification. The algorithm is used to classify new records based on the most similar k data points located nearby the focal point. By using the Euclidean Distance method, we can calculate and rank the distance between the focal point and the nearby points. We can then compare the accuracy of different k values by comparing the predicted class to the actual class and choosing the k with the highest accuracy.

$$Euclidean\ Distance = \sqrt{(x_1 - u_1)^2 + (x_2 - u_2)^2 + \cdots + (x_p - u_p)^2} \quad (7)$$

Where x1, x2, and xp denote the focal point while u1, u2, and up denote some other points.

In our analysis of the Australian automobile market dataset, we experimented with twenty different values of k to determine the optimal value. Specifically, we measured the accuracy of our model across different k values on a validation set. Based on our evaluation of the accuracy rates, we found that the best-performing value of k was 5, as it achieved the highest accuracy rate among the twenty values that were tested. Once we selected k=5, we re-ran the algorithm on the combined training and validation sets to generate the classification of new records.

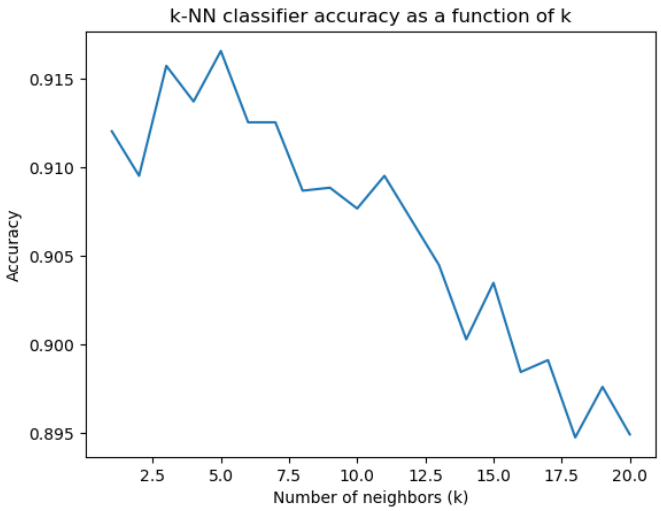

Fig. 10. Finding the optimal k value in terms of accuracy.

C. Naïve Bayes Classifier Algorithm

Before executing the Naïve Bayes Classifier, the input variables are transformed into categorical variables via dummy creation. Additionally, two new variables namely "CC_Classification" and "Kilometers_Category" are created to group the numerical variables "CC" and "Kilometers" into distinct categories. For instance, automobiles with cylinder volumes below 1,201 are classified as small CC vehicles; those with cylinder volumes ranging from 1,201 to 1,401 are classified as medium CC vehicles, while those with cylinder volumes above 1,401 are classified as large CC vehicles [11]. As per Surex's recommendation, a vehicle with a mileage of more than 160,000 kilometers is categorized as high mileage [12]. Conversely, according to Motor Biscuit, a vehicle with mileage below 30,000 miles (48,280 kilometers) is classified as low mileage [13].

After performing data transformation, the Naïve Bayes Classifier is employed as a supervised machine learning algorithm to classify the used vehicle prices into two classes, i.e., fair, or not fair. The classification is based on the following features: car brand, kilometers, car type, gearbox, fuel, CC, seating capacity, and age of the vehicles. The algorithm then calculates the probability of each class for a new record and assigns the record to the class with the highest probability, assuming that the predictor variables are independent. The posterior probability is calculated using a formula that accounts for the probability of each feature for each class [14].

$$P(c \mid x) = \frac{P(x \mid c)P(c)}{P(x)} \quad (8)$$

Where c denotes for data class while x denotes for predictor variable [14].

D. Classification and Regression Trees Algorithm

The Classification and Regression Trees Algorithm is a supervised machine-learning technique that uses a decision tree-based approach to classify and predict car prices by creating a set of rules represented by tree diagrams. In our case, the algorithm's goal is to classify a car price record as fair or unfair.

CART employs a recursive partitioning process to split the data into homogenous groups based on feature values and assigns a prediction value to each group. To construct the decision tree, the algorithm is trained on a labelled dataset of car features and prices, and the Gini impurity measure is used at each splitting node to maximize the information gain. The formula used to calculate the Gini Index is as follows:

$$I(A) = 1 - \sum_{k=1}^{m} P_k^2 \quad (9)$$

Where p is the proportion of cases in rectangle A that belong to class k.

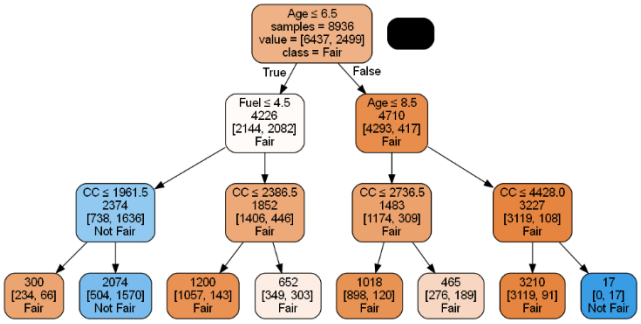

Fig. 11. Decision tree.

In our analysis, we observed that the decision tree's root node divides the car characteristics into two groups based on the "Age" feature, using a threshold of 6.5. If a vehicle's age is less than or equal to 6.5, it goes to the left branch, while any vehicle with an age greater than 6.5 goes to the right branch. The leaf node on the left branch is labeled "Fair," indicating that any car with an age less than or equal to 6.5 is categorized as having a "Fair" price. The tree continues to split based on the highest Gini Index.

E. Logistic Regression Algorithm

Logistic Regression is a machine learning algorithm that builds upon the concept of Linear Regression, but with a categorical output. The goal of logistic regression is to classify whether a car's price is fair or not based on a set of predictor variables. Logistic regression uses a logit function to transform the input variables into a predicted probability value between 0 and 1. The logit function is a transformation of the odds ratio, which is the ratio of the probability of a fair car price occurring to the probability of a fair car price not occurring. The logistic response function and the standard logistic model are defined with the following formulas:

$$p = \frac{1}{1+e^{(\beta 0+ \beta 1x1+ \beta 2x2+\cdots+ \beta qxq)}} \quad (10)$$

$$log(Odds) = \beta 0 + \beta 1x1 + \beta 2x2 + \cdots + \beta qxq \quad (11)$$

Where p is the probability of the positive outcome, that is Fair car price and q is the number of predictors.

In our study, we trained the logistic regression model on a dataset of car prices that were labeled as either fair or not fair. The model learned to associate the input variables with the corresponding fair or not fair labels and classify a new car price based on the input variables.

In logistic regression, the model coefficients correspond to the effect of each predictor variable on the probability of a positive outcome, which in this case is a fair car price. When examining the coefficients, a positive value implies a positive relationship between the predictor variable and the probability of a fair car price. This means that as the predictor variable value increases, the likelihood of a fair price being assigned to the car also increases while keeping other variables constant. Conversely, a negative coefficient indicates a negative relationship, where the likelihood of a fair car price decreases as the predictor variable value increases, holding all other variables constant.

Additionally, the intercept represents the estimated probability of a car price being labeled as "fair" when all predictor variables are set to 0.

TABLE IX: LOGISTIC REGRESSION

Predictors	Coefficients
Intercept	-1.077727671998732
Brand	0.004781
Kilometers	-1.467431
Type	0.005732
Gearbox	-0.050226
Fuel	-0.598894
CC	1.117838
Seating_Capacit	0.319352
Age	-2.09214

Based on our analysis, we found that the intercept value of -1.078 indicates a low baseline probability of the car price is fair. The coefficient for Kilometers is -1.467, which implies that higher Kilometers are linked with lower odds of the car price being fair. On the other hand, the coefficient for CC is 1.118, which suggests that larger engine sizes (CC) are associated with higher odds of the car price being fair. Lastly, the coefficient for Age is -2.092 implies that older cars are associated with lower odds of the car price being fair.

VII. MODEL EVALUATION

The evaluation methods for machine learning algorithms vary depending on their specific type. In this paper, we have categorized the algorithms used into two groups: predictive models and classification models.

TABLE X: PREDICTIVE MODEL AND CLASSIFICATION MODEL

Predictive Model	Classification Model
Multiple Linear Regression	K-Nearest Neighbours (K-NN)
	Naïve Bayes Classifier
	Classification and Regression Trees
	Logistic Regression

The regression statistics and validation errors were employed to assess the performance of the predictive model, whereas the classification model was evaluated using the confusion matrix.

Regression statistics include Mean Error (ME), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Percentage Error (MPE), and Mean Absolute Percentage Error (MAPE) were calculated from the validation data. These measures provide us with accuracy and precision in the prediction.

Validation Errors (also known as residuals) are calculated as the difference between the actual value and the predicted value in the validation data to assess the performance of a linear regression model.

The confusion matrix summarizes the correct and incorrect classifications of the model with the number of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) on the dataset. The rows represent the actual class of price while the columns represent the predicted class of price produced by the model.

A. Multiple Linear Regression Algorithm

TABLE XI: MULTIPLE LINEAR REGRESSION EVALUATION

Mean Error (ME)	-70.2167
Root Mean Squared Error (RMSE)	9885.4414
Mean Absolute Error (MAE)	7383.556
Mean Percentage Error (MPE)	-5.295
Mean Absolute Percentage Error (MAPE)	32.0642

The Root Mean Squared Error (RMSE) of 9885.441 indicates that there is a relatively large average difference between the predicted and actual values, which suggests that the model's accuracy may not be sufficient for practical applications. Additionally, the Mean Absolute Percentage Error (MAPE) of 32.064 is a measure of the average absolute percentage deviation of the predictions from the actual values, and it suggests that the model's predictions have a relatively high percentage error.

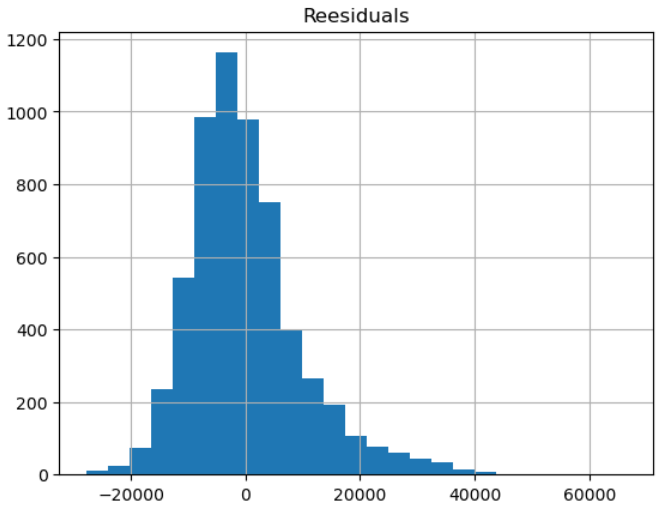

Fig. 12. Residual error of linear regression.

The histogram of residuals shows the distribution of errors around the predicted values, and in this case, the histogram exhibits positive skewness. A positively skewed histogram of residuals suggests that most of the errors are concentrated towards the lower end of the range, with a few larger errors occurring towards the higher end. One potential reason for this skewness could be that the relationship between the price of vehicles (the target variable) and the predictor variables (such as Age, Brand, CC, Fuel, Gearbox, Kilometers, Seating Capacity, and Type) is nonlinear. As a result, the model could make larger errors for some observations when this relationship was nonlinear.

B. K-Nearest Neighbours (K-NN) Algorithm

TABLE XII: CONFUSION MATRIX OF K-NEAREST NEIGHBOURS (K-NN) – VALIDATION DATA

	Predicted Positive	**Predicted Negative**
Actual Positive	4,024(TP)	236(FN)
Actual Negative	261(FP)	1,437(TN)

TABLE XIII: ERROR AND ACCURACY RATE OF K-NEAREST NEIGHBOURS (K-NN)

	Error Rate	**Accuracy**
Training Data	5.54%	94.46%
Validation Data	8.34%	91.66%

Based on the evaluation, it can be concluded that the K-Nearest Neighbours (K-NN) Algorithm performs relatively well in forecasting the classification of new records. The algorithm achieved an accuracy of 91.66% on the validation data and 94.46% on the training data. However, the higher accuracy of the training data suggests that the algorithm may overfit the training data, which could limit its ability to generalize to new and unseen data.

C. Naïve Bayes Classifier

TABLE XIV: CONFUSION MATRIX OF NAÏVE BAYES CLASSIFIER – VALIDATION DATA

	Predicted Positive	**Predicted Negative**
Actual Positive	3,935(TP)	325(FN)
Actual Negative	377(FP)	1,321(TN)

TABLE XV: ERROR AND ACCURACY RATE OF NAÏVE BAYES CLASSIFIER

	Error Rate	**Accuracy**
Training Data	11.66%	88.34%
Validation Data	11.78%	88.22%

Based on the evaluation, the Naïve Bayes Classifier demonstrated an accuracy of 88.34% on the training data and 88.22% on the validation data. The results indicate that the Naïve Bayes Classifier has a lower accuracy rate compared to the K-NN algorithm.

D. Classification and Regression Trees

TABLE XVI: CONFUSION MATRIX OF CLASSIFICATION AND REGRESSION TREES – VALIDATION DATA

	Predicted Positive	**Predicted Negative**
Actual Positive	3,925(TP)	335(FN)
Actual Negative	615(FP)	1,083(TN)

TABLE XVII: ERROR AND ACCURACY RATE OF CLASSIFICATION AND REGRESSION TREES

	Error Rate	**Accuracy**
Training Data	15.85%	84.15%
Validation Data	15.94%	84.06%

After evaluating the Classification and Regression Trees (CART) algorithm using the given data, it was found that the algorithm achieved 84.15% accuracy on the training data and 84.06% accuracy on the validation data. These findings indicate that the CART algorithm can predict the class of new records reasonably well, but with a lower accuracy rate compared to the K-NN and Naïve Bayes Classifier algorithms. The CART algorithm creates a decision tree for classifying records based on a set of features. However, it may not perform as effectively as other machine learning algorithms in some circumstances, particularly when dealing with complex data.

E. Logistic Regression

TABLE XVIII: CONFUSION MATRIX OF LOGISTIC REGRESSION – VALIDATION DATA

	Predicted Positive	**Predicted Negative**
Actual Positive	3,947(TP)	313(FN)
Actual Negative	451(FP)	1,247(TN)

TABLE XIX: ERROR AND ACCURACY RATE OF LOGISTIC REGRESSION

	Error Rate	**Accuracy**
Training Data	13.61%	86.39%
Validation Data	12.82%	87.18%

Based on the evaluation, the Logistic Regression demonstrated an 86.39% accuracy on the training data and an 87.18% accuracy on the validation data. These results suggest that the Logistic Regression algorithm performs moderately well in predicting the class of new records, with a lower accuracy rate compared to the K-NN and Naïve Bayes Classifier algorithms, but a slightly higher accuracy rate than the CART algorithm.

VIII. RESULTS COMPARISONS

Accuracy computed from the validation data is used to evaluate the model's ability to predict new data and estimate future classification errors.

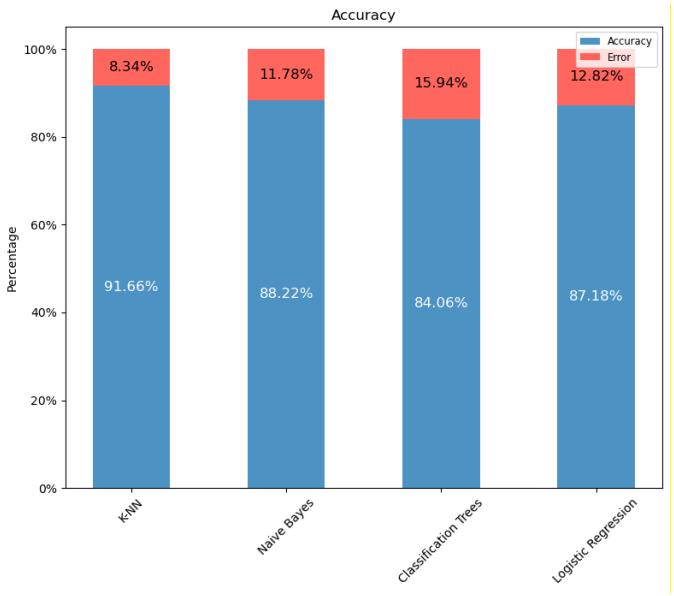

Fig. 13. Comparison of predictive accuracy between classification models.

Among all the algorithms, K-Nearest Neighbors (K-NN) exhibits the highest level of predictive accuracy, followed by Naïve Bayes Classifier and Logistic Regression. Classification trees, on the other hand, demonstrate the lowest accuracy.

The cumulative gains chart is another tool to assess a model's classification performance and aid in making deployment decisions. It measures the model's ability to correctly identify the positive class, with the X-axis indicating the dataset size and the Y-axis representing the proportion of correctly identified positive class records. The cumulative gains are computed by ranking the model's predictions from highest to lowest, and adding up the proportions of the positive class that are correctly identified at each rank. A higher curve on the cumulative gains chart indicates a more effective model in identifying the positive class. The chart can also assist in determining the optimal cut-off point, which is the point that maximizes the identification of the positive class while minimizing false positives.

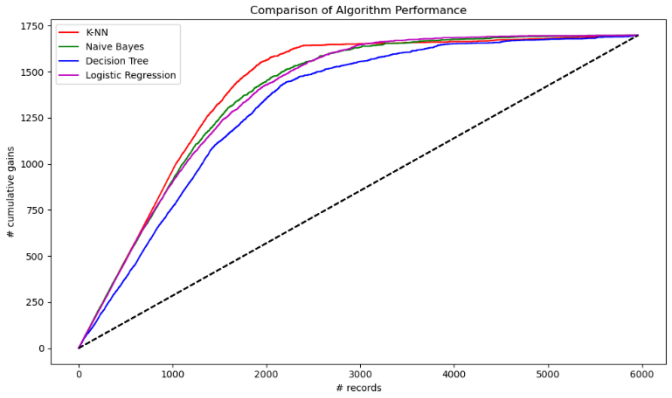

Fig. 14. Cumulative gains chart.

Based on the Cumulative Gains Chart above, it can be observed that K-NN has the steepest curve compared to the other three algorithms. This indicates that K-NN can identify a larger proportion of the positive class with fewer records. This suggests that K-NN is more effective in the early detection of positive cases, providing a strong advantage over the other algorithms.

Moreover, K-NN has a larger area under the curve compared to the other algorithms, indicating better overall performance in distinguishing between the positive and negative classes. Therefore, it can be concluded that K-NN is the most effective algorithm for classifying used vehicle prices, followed by Naïve Bayes Classifier, Logistic Regression, and Classification trees, respectively.

IX. Limitations

Although this study aims to provide a comprehensive approach to predicting the prices of used vehicles, there are several limitations to consider. Firstly, the dataset used in this study is limited to Australia, and therefore the findings may not be generalizable to other regions. Secondly, the study focuses on only eight features of the car, and other external factors such as the supply chain of new and used vehicles during COVID and post-COVID, inflation rate, condition of the car, the level of maintenance, and the previous owner's driving behavior may affect the prices of used vehicles and are not included in this study. Thirdly, despite pre-processing the dataset and removing outliers, the study's results may be influenced by the quality of the data used, and any inconsistencies in the dataset could impact the accuracy of the findings. Finally, it is important to acknowledge that each algorithm used in this study has its assumptions and limitations, which may affect the accuracy of the model's predictions. For instance, Multiple Linear Regression assumes a linear relationship, Naive Bayes assumes independence among predictors, Classification and Regression Trees assume a hierarchical structure, and Logistic Regression assumes a linear relationship on the logit scale.

X. Conclusion

The prices of pre-owned vehicles have significantly risen in recent years, creating a perplexing situation where a new vehicle may be more affordable than a used one, which goes against the traditional notion of buying a second-hand car for cost savings. Out of the five machine learning algorithms we evaluated, the k-NN algorithm exhibited the highest accuracy of 91.7% when validated with the data, outperforming the other algorithms which showed accuracy rates ranging from 84.1% to 88.2%. The k-NN algorithm also displayed the steepest curve with the largest area under the curve in the Cumulative Gains Chart, further confirming its superior performance in identifying the positive class and distinguishing between the positive and negative classes. Therefore, we highly recommend the use of the k-NN algorithm with k set to 5 for predicting pre-owned vehicle prices, as it has demonstrated the highest accuracy.

According to the research conducted by Samruddhi and Kumar [4], the K-Nearest Neighbor algorithm achieved an accuracy rate of 85% in predicting used car prices. In comparison, our research achieved a higher accuracy rate of 91.7%, indicating the better performance of the algorithm in predicting used car prices.

The K-NN algorithm stands out as the most feasible option, given that it is a non-parametric approach that does not rely on any assumptions about the data. In contrast, linear regression and logistic regression may not be well-suited for predicting second-hand vehicle prices, as it may be challenging to assume a linear relationship between the input features, such as seating capacity, and the non-linear target variable. Moreover, the Naive Bayes algorithm may not be the optimal choice for interpreting prices, as the dataset exhibit strong dependencies among the input variables. Finally, the CART algorithm does not perform as effectively as other machine learning algorithms when dealing with complex data and creating rules using the eight predictors in this study.

In this paper, the K-NN algorithm is proposed as a means to effectively interpret the market dynamics of pre-owned vehicle prices. By providing potential buyers with guidance on whether a vehicle's price is reasonable, the algorithm can aid in negotiations with dealers for a more favorable price. Similarly, dealerships can leverage the algorithm to determine optimal prices that maximize their profits, especially as the COVID-related supply constraints normalize. Ultimately, this approach is expected to create a mutually beneficial scenario for both buyers and dealerships, resulting in a win-win outcome.

Acknowledgment

We would like to express our deepest gratitude to our professor, Dr. Salam Ismaeel, for his invaluable guidance and support throughout the research process. His extensive knowledge, constructive feedback, and unwavering encouragement have been instrumental in shaping our research.

References

[1] R. Kurmelovs, "Australia's used car market is in overdrive as dealers chase customers to buy back vehicles," The Guardian, Jul. 7, 2022. [Online]. Available: https://www.theguardian.com/australia-news/2022/jun/07/australias-used-car-market-is-in-overdrive-as-dealers-chase-customers-to-buy-back-vehicles.

[2] W. Stofford, "Australian used car prices are falling (slowly) back to Earth," CarExpert, Mar. 6, 2023. [Online]. Available: https://www.carexpert.com.au/car-news/australian-used-car-prices-are-falling-slowly-back-to-earth.

[3] T. D. Phan, "Housing Price Prediction Using Machine Learning Algorithms: The Case of Melbourne City Australia", 2018 International

Conference on Machine Learning and Data Engineering (iCMLDE), pp. 35-42, 2018.

[4] K. Samruddhi and R. A. Kumar, "Used Car Price Prediction using K-Nearest Neighbor Based Model," Int. J. Innov. Res. Appl. Sci. Eng. (IJIRASE), vol. 4, pp. 629-632, 2020.

[5] P. Venkatasubbu and M. Ganesh, "Used Cars Price Prediction using Supervised Learning Techniques," Int. J. Eng. Adv. Technol. (IJEAT), vol. 9, no. 1S3, pp. 66-71, Dec. 2019.

[6] Ö. Çelik and U. Ö. Osmanoğlu, "Prediction of the prices of second-hand cars," Avrupa Bilim ve Teknoloji Dergisi, no. 16, pp. 77-83, 2019.

[7] M. Hankar, M. Birjali, and A. Beni-Hssane, "Used Car Price Prediction using Machine Learning: A Case Study," in 2022 11th International Symposium on Signal, Image, Video and Communications (ISIVC), May 2022, pp. 1-4.

[8] N. T. C. Lai, "Australian AutoMobile Market," Kaggle, [Online]. Available: https://www.kaggle.com/datasets/nguyenthicamlai/cars-sold-in-australia.

[9] Autotrader, "Cars for Sale," Autotrader, [Online]. Available: https://www.autotrader.com.au/for-sale.

[10] J. Mellor, "Used car prices remain high," GoAuto News Premium, Cox Automotive, [Online]. Available: https://www.coxautoretail.com.au/news/used-car-prices-remain-high/.

[11] S. Pudaruth, "Predicting the price of used cars using machine learning techniques," International Journal of Information and Computer Technology, vol. 4, no. 7, pp. 753-764, 2014.

[12] Surex, "What Is Good Mileage for a Used Car?," [Online]. Available: https://www.surex.com/blog/what-is-good-mileage-for-used-car.

[13] G. DeSantis, "What Does High or Low Mileage Really Mean for Your Car?," MotorBiscuit, [Online]. Available: https://www.motorbiscuit.com/high-low-mileage/.

[14] S. Sayad, "An Introduction to Data Science," Naive Bayesian, [Online]. Available: https://www.saedsayad.com/naive_bayesian.htm.