Prediction of Car Purchase based on User Demands using Supervised Machine Learning

Mohd Samee Uddin [1], Rabab Fatima Hussain [2], Asfiya Samreen [3], Saleha Butool [4]

[1, 2, 3] B.E. Student, [4] Assistant Professor,
Department of IT, Lords Institute of Engineering and Technology,
Hyderabad, Telangana, India.

Published in IJIRMPS (E-ISSN: 2349-7300), Volume 11, Issue 1, (January-February 2023)

Abstract

One of the key sectors of the national economy is the auto industry. Cars are becoming more and more common as a form of private transportation. When a buyer wants to purchase the ideal vehicle, particularly a car, an evaluation is necessary. Because it is an expensive vehicle, there are a lot of conditions and elements to consider before buying a new one, including price, headlamp, cylinder volume, and spare parts. Therefore, it is crucial for the consumer to choose a purchase that can meet all of the criteria before making any other decisions. In our research, we therefore suggest various well-known methods to improve accuracy for a car purchase. These algorithms were used on our dataset, which consists of 50 data. With a prediction accuracy of 86.7%, Support Vector Machine (SVM) produces the best result of the bunch. In this study, we also present comparison findings for all data samples using various methods for precision, recall, and F1 score.

Keywords: Supervised Machine Learning, Naive Bayes, Random Forest Tree, Support Vector Machine, KNN, Cosine Similarity

1. Introduction

People prefer to come up with ideas and take actions in this modern, technologically advanced era that will help them both now and in the future. For instance, a person's life decisions will be influenced by his desire to choose a job, his plans for where he will live, and his desire to take a luxurious trip. Because these must be taken into account when determining how much utility there will be in the future. People are naturally drawn to making choices that enhance utility. Due to the utility's connection to the financial system in our daily lives.

The automobile industry is currently one of the most significant global industries. Even though Bangladesh is a small nation in South Asia, demand for automobiles is growing daily. To get from one place to another, people drive their own cars. The most useful and adaptable of those is the four-wheeled private vehicle. An entire nation's economic growth is largely dependent on its transportation infrastructure. Because an effective transportation infrastructure provides both economic and social opportunities, which benefits markets. Then earlier

Customers would rather have that assurance than making a new car purchase. Because it costs a significant sum of money to purchase a new car from an economic standpoint. Because of this, it's critical to learn about cars that are good or poor based on the experiences of previous purchasers. A modern car's lifespan is dependent on so many different components. Based on the pricing, spare part, customer review, cylinder volume, and resale price, this project forecasts the likelihood that a customer will purchase a car. A fantastic and urgent topic is estimating the possibility or probability of purchase for autos. To compare which algorithm provides greater accuracy, this study utilized four well-known algorithms: Naive Bayes, SVM, Random Forest Tree, and K-nearest Neighbor.

The rest of the paper is assembled as follows. Section 2 of the paper is analysis of related works. Section 3 describes the particulars of the proposed method for finding better accuracy. Section 4 evaluates the experiment and displays the experiment results phase-wise. Section 5 winds up the paper in a nutshell and highlights some future work that can be done.

2. Related Work

Some customers liked high-quality components; others chose inexpensive or expensive parts with all the characteristics they need; yet others were only weak for well-known automakers. Despite the fact that some factors, such as color, comfort, seating capacity, etc. are well recognized, choosing the ideal car remains a challenging undertaking. To determine which algorithm provides greater accuracy in predicting car purchases, we therefore tried to compare various algorithms.

An implementation of Nave Bayes Classification method is proposed by Fitrina et al. [3]. Naive Bayes is known as a simple probabilistic classifier. They applied this method for predicting purchase. They used a dataset on 20 car purchase data and got 75% accuracy. Srivasta et al. [4] applied the powerful learning method, Support Vector Machine (SVM) on different types of data like Diabetes Data, Heart Data, Satellite Data and Shuttle Data. Those datasets have multi classes. They are also proven the analysis of the comparative consequences the use of divers kernel functions on their paper.

A comparative analysis of machine learning algorithm was proposed by Ragupathy et al. [5]. In their paper, they tried to identify and classify sentiment, conveyed in main text. They have collected their data from social media like Twitter, comments, blog posts, news, status updates etc. They also applied Naive Bayes, Decision Tree, K-Nearest Neighbour and Support Vector Machine classifiers for their comparison purpose. Their goal was to find out the most efficient classification technique and SVM came out with 72.7% which was the best accuracy. Another prediction system using supervised machine learning technique was proposed by Noor et al. [6]. They used multiple linear regression method and predicted vehicle price. They got 98% accuracy on their system. Pal et al. [7] proposed a methodology for predicting used cars costs. In their paper, they used Random Forest classifier to predict the costs of used cars. To train the data, they created a Random Forest with 500 Decision trees. Finally, they got 95.82% as training accuracy and 83.63% as testing accuracy. Pudaruth et al. [8] proposed another methodology for predicting used cars prices. In that paper, he applied multiple linear regression analysis, k-nearest neighbours, Naive Bayes and Decision Tree which were used to make the predictions. Osisanwo F.Y. et al. [9] proposed Supervised machine learning technique. They compared seven different Supervised learning algorithms and described those. They also found out the most effective classification algorithm established on dataset.

A new defect classification technique was proposed by Veni et al. [11] to predict the class label of "severity" tuple. Those data tuples were described by various attribute like Phase attribute, Defect, Phase Defected, Impact and Weight. They applied Naive Bayes classifier for prediction purpose. Jayakameswaraiah et al. [2] developed a data mining system to analyze cars. They proposed TkNN clustering algorithm to predict the right car. They also shown the comparison of KNN and their proposed novel TkNN clustering. Another car price prediction technique was proposed by Gegic et al. [12] where they used three machine learning techniques. They got 92.38% accuracy on combination of all ML methods.

Despite these well-known works, there are some harder works as well. We concentrate on comparing four different types of well-known machine learning algorithms in order to determine which algorithm provides the highest accuracy for our dataset.

3. Methodology

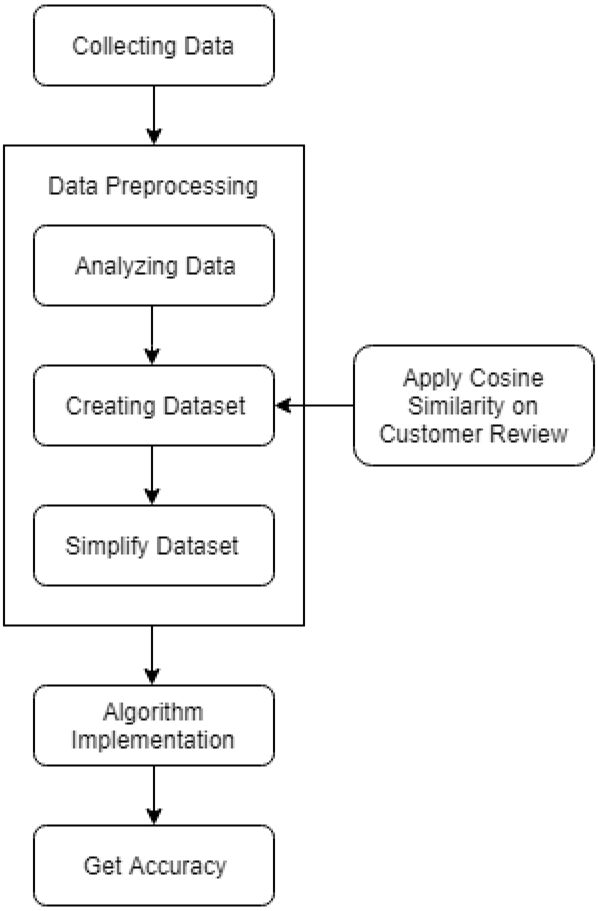
The aim of this research is to analyze the accuracy of different predictive algorithms that can predict the probability of purchasing a car. A comparative analysis of machine learning algorithm was proposed by Ragupathy. In this project, they tried to identify and classify sentiment, conveyed in main text. They have collected their data from social media like Twitter, comments, blog posts, news, status updates etc. They also applied Naive Bayes, K-Nearest Neighbor, Random forest and Support Vector Machine classifiers for their comparison purpose. We have collected our data from social media. After successfully creating our dataset, we evaluate algorithms by splitting dataset. We use 70% of data as training and 30% of data as testing. To evaluate the results, we have used precision recall, execution time, accuracy measurement of the algorithms. The advantages of proposed methodology is as follows:

(1) The most common approach is machine learning, a method that needs a significant data set for training and learning the aspects and sentiments associated, and using this machine learning algorithms for predicting car purchase.

(2) We choose supervised learning algorithms. Those are Nave Bayes, Support Vector Machine (SVM), K-nearest neighbor algorithm (KNN) and Random Forest. we apply above mentioned four algorithms for our dataset. We select one algorithm as our desired algorithm which provides best result for the dataset.

(3) Algorithms: Support Vector Machine and Naive Bayes, Random Forest and KNN.

Figure 1: Depicts the Workflow of the Proposed Methodology



A. Collecting Data

Data collection is in fact the first and most fundamental step in the machine learning pipeline. It's part of the complex data processing phase within an ML lifecycle. From this comes another important point: data collection directly impacts the performance of an ML model and the final results.

B. Data Pre-processing

Data pre-processing is the process of preparing raw data suitable for a machine learning model. This is the first and crucial step in building a machine learning model.

Creating a machine learning project does not always involve clean and formatted data. And if you're going to do anything with the data, cleaning and formatting is a must. So we use the data processing task for this. Real-world data typically contains noise, missing values, and possibly unusable data that cannot be directly used in machine learning models. Data pre-processing is necessary to clean the data and apply it to the machine learning model, which also increases the accuracy and efficiency of the machine learning model. It involves the following steps: Getting the dataset, Importing libraries, Importing datasets, Finding Missing Data, Encoding Categorical Data, Splitting dataset into training and test set, Feature scaling.

C. Simplifying Dataset

Since data is the most important thing in machine learning, the first time that we should do is simplifying the data that we have collected. Because almost all the data we get at the beginning is still messy and a lot of noise. That means we have to filter the large data that we have obtained so that it can be processed later.

Here are the steps you can take to simplify data:

Reduce Noise Data

Suppose we have excel data that contains various column and row data. For example, the column for the date of birth. In that case, the column can be filled with characters, integers, and so on. In this step, we have to homogenize the column date of birth so that it is ready to use.

Reduce Dimensionality

In general, the variables we have may range from thousands or tens of thousands of variables such as customer data, transactions and so on. As much as possible we reduce the data variables that have similarities in them. For example, the dataset contains columns for date of birth and age. We can analyze that the column date of birth and age has the same information. So to simplify it we can reduce the column date of birth and only use the column age.

Find Important Variables and Combinations

Look for important variables or their combinations in the dataset. Suppose we have 10 variables. From these 10 variables, we can analyze which ones are important and which ones are not. We can eliminate these unimportant variables to get a new dataset that is ready to be processed. An example is eliminating the variable name in the dataset. Because the variable name does not provide value in the analysis process that we will do.

In this phase, we again do an update of our dataset. We assume some numeric value for our data. We assume Expensive as 3, Affordable as 2 and Normal as 1 numeric value for our Price attribute. Same as Low (1), Medium (2), High (3) for Spare Part and Cylinder Volume, again same as Expensive (3), Affordable (2), Normal (1) as Resale Price attribute and Yes (1), No (0) for Buy attribute.

Table 1: A Portion of Simplified Dataset

Price	Spare Part	Cylinder Volume	Resale Price	Car’s Review	Buy
3	1	2	3	1	1
2	2	2	2	1	1
1	3	1	2	0	0
3	1	2	3	1	1
2	2	3	1	0	0

We also apply Cosine Similarity for mining the text as customer review in our Cars Review attribute. Applying Cosine Similarity, we found 1 for positive review and 0 for negative review. For finding cosine similarity, we have a dataset which contains customer review, according to the review their is an output which was positive or negative. Then according to the dataset, we measure our collected customer’s review positive or negative. Cosine similarity measure two sentence according to the function,

Cosine similarity is a metric used to measure the similarity of two vectors. Specifically, it measures the similarity in the direction or orientation of the vectors ignoring differences in their magnitude or scale.

Both vectors need to be part of the same inner product space, meaning they must produce a scalar through inner product multiplication. The similarity of two vectors is measured by the cosine of the angle between them.

$$similarity(A, B) = cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|}$$

The Cosine Similarity measure the value between 0 to 1. After measuring two sentences Cosine Similarity, if the value is greater than or equal to 0.5, then the review is positive. If the value is smaller than 0.5, then the review is negative. We have assumed the threshold value, for better performance of the similarity measurement.

Cosine similarity is beneficial for applications that utilize sparse data, such as word documents, transactions in market data, and recommendation systems because cosine similarity ignores 0-0 matches. Counting 0-0 matches in sparse data would inflate similarity scores.

D. Algorithm Implementation

To predict something, first of all, we have to learn our machine. Those machines can learn with the proper algorithm. There are three types of machine learning algorithms. They are supervised learning, Unsupervised learning, Semi-supervised learning. Among those, we choose supervised learning algorithms. Those are Nave Bayes, Support Vector Machine (SVM), K-nearest neighbor algorithm (KNN) and Random Forest tree.

(1) Naive Bayes

The Naive Bayes algorithm is a supervised learning algorithm based on Bayes theorem used to solve classification problems. It is mainly used for text classification that contains high-dimensional training data. The Naive Bayes classifier is one of the simplest and most powerful classification algorithms that helps build fast machine learning models that can make quick predictions. It is a probabilistic classifier, which means it predicts the target based on probability It is regarded as nave because of its assumption shortens calculation1. The overall formula can be written as,

$$P(c/f) = \frac{P(c) * P(f/c)}{P(f)} \qquad (1)$$

Here, c = class, f = features

P(c/f): Posterior Probability
P(c): Class Prior Probability
P(f/c): Likelihood
P(f): Predictor Prior Probability

Applications of Naive Bayes Classifier:
A. It is used for credit evaluation.
B. It is used in the classification of medical data.
C. It can be used for real-time prediction because the Naive Bayes classifier is an eager learner.
D. It is used in text classification such as spam filtering and sentiment analysis.

(2) Support Vector Machine (SVM)

The purpose of the SVM algorithm is to create the best line or decision boundary that can separate the n-dimensional space into classes so that we can easily place a new data point into the correct class in the future. This best decision boundary is called the hyperplane. SVM selects the extreme points/vectors that help create the hyperplane. These extreme cases are called support vectors and therefore the algorithm is called a support vector machine. Consider the diagram below with two different classes classified using a decision boundary or hyperplane. SVM has many applications such as handwriting recognition, classification based on email accounts, facial recognition of people or other animals, etc.

(3) K-Nearest Neighbour Algorithm (KNN)

The K-Nearest Neighbour (KNN) algorithm is a popular machine learning technique used for classification and regression tasks. It relies on the idea that similar data points tend to have similar labels or values. During the training phase, the KNN algorithm stores the entire training dataset as a reference. KNN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.

KNN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using KNN algorithm. KNN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems. KNN is a non-parametric algorithm, which means it does not make any assumption on underlying data. It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.

KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

(4) Random Forest Tree

Random Forest is one of the most popular and commonly used algorithms by Data Scientists. Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

Random forest is a versatile machine learning algorithm developed by Leo Breiman and Adele Cutler. It leverages an ensemble of multiple decision trees to generate predictions or classifications. By combining the outputs of these trees, the random forest algorithm delivers a consolidated and more accurate result.

Its widespread popularity stems from its user-friendly nature and adaptability, enabling it to tackle both classification and regression problems effectively. The algorithm's strength lies in its ability to handle complex datasets and mitigate overfitting, making it a valuable tool for various predictive tasks in machine learning.

Our Random Forest model system deals with ID3 algorithm to train and uses the Gini index to measure it. Gini index is used for calculating the uprightness of split criteria. The Gini impurity measure can be written as:

$$\text{Gini}(Xn) = \sum K_{pnk}(1\text{-}pnk) \qquad (3)$$

Where pnk is the fraction of times. k is an element of class occurs in a split. Xn is an element of set X.

(5) Get Accuracy

Here, we apply the aforementioned four algorithms to our dataset. We choose as the desired algorithm one algorithm that gives the best result for the data set.

4. Experiments and Results

(a) Implementation

For implementing car prediction algorithm, we have used anaconda which is an environment of Python 3.7 with a lot of packages of machine learning. Our processor is Intel Core i3 with clock rate 2.4 GHz, RAM 4 GB. We used Windows 10 (64-bit) as operating system.

(b) Evaluation Dataset

We have collected our data from different shops in Bangladesh and also from social media. After successfully creating our dataset, we evaluate algorithms by splitting dataset. We use 70% of data as training and 30% of data as testing. A simple statistics of dataset given in Table 2.

Table 2: Simple Statistics of Dataset

Attributes	Number of Count
Data Collected	50
Training Data	35
Testing Data	15

C. Evaluation Measurement

To evaluate the results, we have used precision-recall, execution time, accuracy measurement of the algorithms.

(1) Precision, Recall and F1 Score

Precision is a ratio of accurately predicted positive observations. Recall is a ratio of accurately predicted positive observations to all observations in actual class-yes. F1 score is weighted average of the precision and recall. This scores evaluate how a model has performed.

Precision

Precision is defined as the ratio of correctly classified positive samples (True Positive) to a total number of classified positive samples (either correctly or incorrectly).

Precision = True Positive ÷ True Positive + False Positive
Precision = TP ÷ TP + FP
TP: True Positive
FP: False Positive

Recall

The recall is calculated as the ratio between the numbers of Positive samples correctly classified as Positive to the total number of Positive samples. The recall measures the model's ability to detect positive samples. The higher the recall, the more positive samples detected.

Recall = True Positive ÷ True Positive + False Negative

Recall = TP ÷ TP + FN

TP: True Positive

FN: False Negative

F1 Score

The F1 score combines precision and recall using their harmonic mean, and maximizing the F1 score implies simultaneously maximizing both precision and recall. Thus, the F1 score has become the choice of researchers for evaluating their models in conjunction with accuracy.

$$\text{F1 Score} = \frac{TP}{TP + \frac{1}{2}(FP + FN)}$$

Precision, recall and F1 score of several algorithms are given in Table 3:

Table 3: Precision, Recall and F1 Score of Mentioned Algorithms

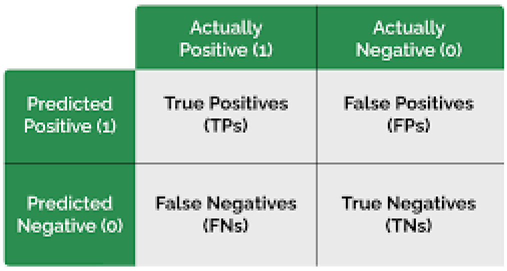
Algorithm	Precision	Recall	F1 Score
Random Forest	0.60	0.60	0.60
KNN	0.75	0.73	0.73
Naive Bayes	0.39	0.43	0.41
SVM	0.89	0.87	0.86

(2) Accuracy of Several Algorithms

Accuracy is a measurement of how a model predicts correctly to the total number of input samples. In our proposed method, all algorithms have split the dataset into 70% training and 30% testing. A simple equation of accuracy calculation is given below,

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

Figure 3: Confusion Matrix



	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

Several algorithms accuracy comparison are given in Figure 2.

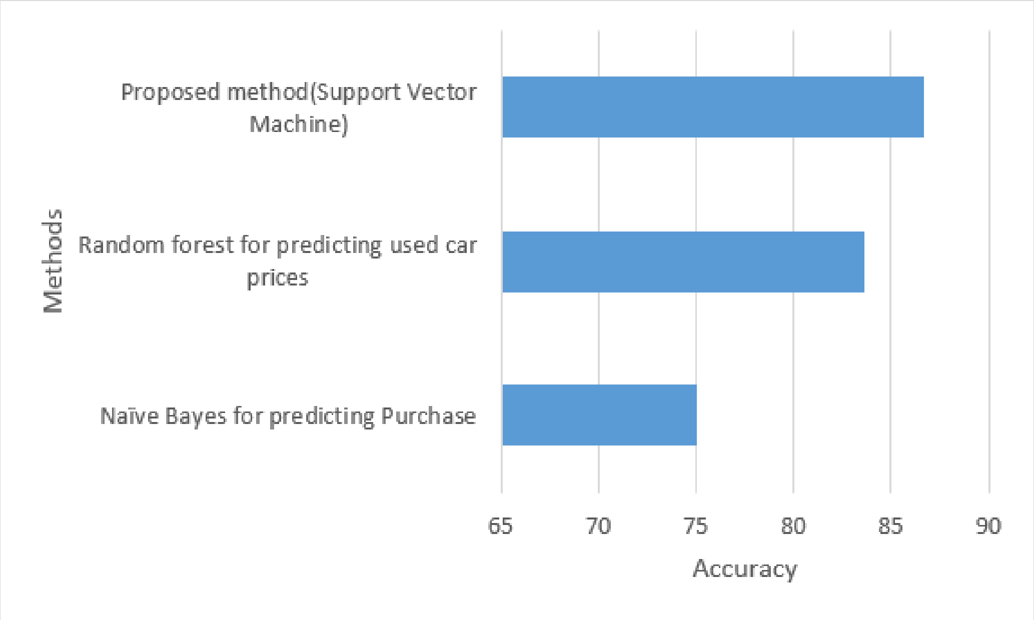
From Figure 2, we can see Support Vector Machine gives highest accuracy (87.6%) than Random Forest, K-nearest Neighbour (KNN) and Naive Bayes. That means Support Vector Machine can classify approximately 44 car purchase data of 50 dataset.

(3) Comparison with other Methods

There are many study about car price prediction from many years. Several study use several machine learning techniques. A methodology for predicting purchase used Naive Bayes algorithm and get 75% accuracy [3]. Another work for predicting used cars prices used Random Forest and get test accuracy 83.63% [7].

On the other hand, our proposed method used Cosine Distance for review analysis and after that using Support Vector Machine got 86.7% accuracy. There are some obstacles determined in different classifiers like Naive Bayes classifier cannot cope with more amounts of data with ease. A simple comparison shown in Figure 4:

Figure 4: Comparison with Other Method



5. Conclusion

Customer purchase prediction aims to help the customer to make the right decision whether or not he will buy a car, and these prediction results are of great importance for conducting future commercial activities. To obtain accurate predictions of customer purchases, machine learning framework is developed based on historical behavioral data. This project mainly focused on Customer's review and used Cosine Similarity to analyze customer's review. Finally, a real-word dataset is adopted to test the feasibility of the proposed prediction framework. The experimental results and comparative analysis verify the validity of the proposed model. Several algorithms were applied on dataset. Those algorithms were then compared according to their accuracy. Support Vector Machine has given the highest accuracy among all the algorithms.

6. Limitations and Future Work

In order to anticipate the outcome, this project used 5 features. More features will be gathered in the future to expand the dataset and aid with prediction. In our future study, we'll employ more effective strategies to obtain improved accuracy as well as more sophisticated machine learning methods including fuzzy logic, decision trees, artificial neural networks, ordinal least squares regression (OLSR), fuzzy logic, etc. By selecting more features for classification, the scope of this job can be increased.

References

[1] M.R. Busse, D.G. Pope, J.C. Pope, J. Silva-Risso, “The psychological effect of weather on car purchases”, The Quarterly Journal of Economics, vol. 1, no. 44, p. 44, 2014.

[2] M. Jayakameswaraiah and S. Ramakrishna, “Development of data mining system to analyze cars using TkNN clustering algorithm”, International Journal of Advanced Research in Computer Engineering Technology, vol. 3, no. 7, 2014.

[3] F. Harahap, A.Y.N. Harahap, E. Ekadiansyah, R.N. Sari, R. Adawiyah, and C.B. Harahap, “Implementation of naive bayes classification method for predicting purchase”, in 2018 6th International Conference on Cyber and IT Service Management (CITSM). IEEE, 2018, pp. 1–5.

[4] K.S. Durgesh and B. Lekha, “Data classification using support vector machine”, Journal of Theoretical and Applied Information Technology, vol. 12, no. 1, pp. 1–7, 2010.

[5] R. Ragupathy and L. Phaneendra Maguluri, “Comparative analysis of machine learning algorithms on social media test”, International Journal of Engineering and Technology (UAE), vol. 7, pp. 284–290, 03 2018.

[6] K. Noor and S. Jan, “Vehicle price prediction system using machine learning techniques”, International Journal of Computer Applications, vol. 167, no. 9, pp. 27–31, 2017.

[7] N. Pal, P. Arora, P. Kohli, D. Sundararaman, and S.S. Palakurthy, “How much is my car worth? A methodology for predicting used cars prices using random forest”, in Future of Information and Communication Conference. Springer, 2018, pp. 413–422.

[8] S. Pudaruth, “Predicting the price of used cars using machine learning techniques”, Int. J. Inf. Comput. Technol, vol. 4, no. 7, pp. 753–764, 2014.

[9] F. Osisanwo, J. Akinsola, O. Awodele, J. Hinmikaiye, O. Olakanmi, and J. Akinjobi, “Supervised machine learning algorithms: Classification and comparison”, International Journal of Computer Trends and Technology (IJCTT), vol. 48, no. 3, pp. 128–138, 2017.

[10] S. Veni and A. Srinivasan, “Defect classification using na¨ıve bayes classification”, International Journal of Applied Engineering Research, vol. 12, no. 22, pp. 12 693–12 700, 2017.

[11] E. Gegic, B. Isakovic, D. Keco, Z. Masetic, and J. Kevric, “Car price prediction using machine learning techniques”, 2019.

[12] M. Jabbar, “Prediction of heart disease using k-nearest neighbor and particle swarm optimization”, Biomed. Res, vol. 28, no. 9, pp. 4154– 4158, 2017.

[13] M.C. Sorkun, “Second-hand car price estimation using artificial neural network”.

[14] Q. Yuan, Y. Liu, G. Peng, and B. Lv, “A prediction study on the car sales based on web search data”, in The International Conference on E-Business and E-Government (Index by EI), 2011, p. 5.