IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Mathematical Modelling of Engineering Problems
Vol. 10, No. 1, February, 2023, pp. 139-148
Journal homepage: http://iieta.org/journals/mmep

Intelligent Modelling Techniques for Predicting Used Cars Prices in Saudi Arabia

Mohammed Gollapalli[1*], Tayma A. Alqahtani[2], Dina H. Alhamed[2], Maryam R. Alnassar[2], Aljawharah M. Alajmi[2], Yasminah H. Alali[2], Mamoun M. Abdulqader[3], Ashraf Saadeldeen[4]

[1] Department of Computer Information Systems, College of Computer Science and Information Technology, Imam Abdulrahman Bin Faisal University, P.O. Box 1982, Dammam 31441, Saudi Arabia
[2] Department of Computer Engineering, College of Computer Science and Information Technology, Imam Abdulrahman Bin Faisal University, P.O. Box 1982, Dammam 31441, Saudi Arabia
[3] Department of Computer Science, College of Computer Science and Information Technology, Imam Abdulrahman Bin Faisal University, P.O. Box 1982, Dammam 31441, Saudi Arabia
[4] Department of Computer Science, Computer Science, Albaha University, P.O. Box 1988, Al Baha 65528, Saudi Arabia

Corresponding Author Email: magollapalli@iau.edu.sa

https://doi.org/10.18280/mmep.100115

Received: 25 May 2022
Accepted: 22 October 2022

Keywords:
car price prediction, machine learning, regression, Linear Regression, XGBoost, Random Forest

ABSTRACT

The production of cars has been decreasing in most countries since the COVID-19 pandemic from 2020 to 2021. Due to this, the used car market has grown to be a booming industry by itself. Recent advances in online portals and platforms have made it possible to get more information about the factors that determine used car values. Hence, car price prediction has become a high-interest field of research. This paper aims to investigate the power of machine learning to build a model that will be able to predict the approximate price of a used car by utilizing the "Saudi Arabia Used Cars" Dataset which is collected from the Syarah platform and available on the Kaggle platform. The model assists both customer and seller to estimate the approximate price of a used car in the market. Three different Machine learning techniques were utilized which are Linear Regression, Random Forest, and XGBoost which score an MSE of 0.15, 0.10, and 0.19 respectively. The Random Forest Regressor algorithm outperformed other algorithms where it achieves the best result on the three evaluated metrics RMSE, MSE, and R-squared.

1. INTRODUCTION

Nowadays there are many different categories of vehicles such as car, coupe, sedan, support car, hatchback, station wagon, convertible, pickup truck, and sport-utility vehicle (SUV). Each of these vehicles has its own characteristic, features, and type of use where the price is specified based on them [1]. Buying a used car is a good alternative for people in many countries because it is affordable and offers the buyers the chance to resell the car again after a while, which could result in some profit [2, 3]. An economic report indicated that during the COVID-19 pandemic from 2020 to 2021 Auto factories around the world have announced plans to temporarily suspend production to disinfect facilities and prevent the spread of coronavirus and that gave an opportunity for the used car market to be a booming industry by itself [4].

In the kingdom of Saudi Arabia, the demand for buying used cars has grown due to several factors one of the main factors was the raised in tax from 5% to 15% which starts recently in the July of 2020 until now [5]. Where this decision affects the trend of the buyers to look at the used cars instead of buying a new car. Selling a used car proves to be challenging since people find it difficult to recognize its fair value coupled with car prices depend on distinctive features and factors, which car owners need to know to determine the value of their vehicles. Accurate car price prediction requires expert knowledge [6].

Naturally, the most significant factors are the car model, age, mileage, brand, horsepower, fuel type used along with consumption per mile are extremely affect the price of a car especially the fuel type due to frequent changes in the price of fuel [7]. In addition, the different features such as type of transmission, exterior color, safety, door number, dimensions, air condition, interior, and whether it has navigation or not will influence the car price [7]. Because of the large number of features and factors that consumes the time and efforts of human in pricing cars, machine learning techniques can be utilized to build a model that predicts the price of the used cars. Thus, many previous studies have addressed the task of predicting used car prices by utilizing machine learning for various kinds of datasets. One of these studies achieved an accuracy of 93% [8].

The current studies have not used real-world Saudi Arabia datasets to predict used car prices in SAR. Therefore, this paper aims to help the Saudi Arabia car industry by working on a real-world Saudi Arabia dataset provided by the Kaggle platform [9] to build a machine learning model capable of predicting used car prices, which is considered critical for car showrooms. Likewise, it could benefit citizens who would like to sell their cars at a fair price. The dataset was collected from the Syarah platform through the year 2021 which allows users to advertise their used cars for sale. The dataset contains 8248 records and 13 features (Make, Type, Year, Origin, Color, Options, Engine Size, Fuel Type, Gear Type, Mileage, Region,

Price, Negotiable). Based on research conducted and the background of the cars industry this dataset contains features that affect the price of a car such as Year, Type, Fuel Type, Mileage, and Color which emphasizes the quality of the dataset. The experiment design framework consists of data collection, data preparation/pre-processing, data analysis, model building, and model evaluation as shown in Figure 1.

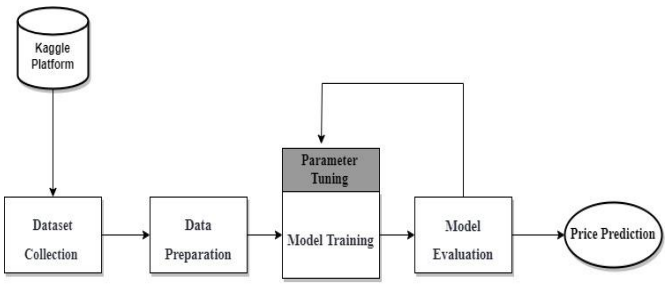

Figure 1. Experimental design

Furthermore, an experimental comparison was conducted to compare the performance of the proposed model in terms of MSE and R-squared metrics. The proposed model works on the regressor problem means it is intended to predict the approximate price of the car. Based on the experiment set up the Random Forest Regressor algorithm achieves the best result for the three evaluated metrics RMSE, MSE, and R-squared with a value of 0.32, 0.10, and 0.82 respectively.

The structure of this research paper is as follows. In section 2, the literature review was described that summarizes similar studies that have been done previously. In section 3, the Data preprocessing where we give a Data overview and Data preparation/pre-processing. In section 4, Materials and Method. We give a description of the proposed Technique, the experimental setup, the performance measure criteria, the optimization strategy, the result and dissection, and the experimental comparison.

2. LITERATURE REVIEW

Mammadov [9] aimed to build a Linear Regression model to predict the prices of cars for the U.S market to aid a new entrant in the U.S automobile industry to understand the important pricing factors as an accurate estimation of automobile prices requires specialized expertise. The dataset used was collected from the web portal fred.stlouisfed.org using a web scraper. The researchers follow the Data Science Methodology (DSM) phases and decided to choose an optimal number of features where model accuracy and generalization both are at a satisfactory level to be used in the model rather than arbitrarily features. The model with optimal features has RMSE 2455,6 and r-sq. 0.9 for the test dataset.

Gajera et al. [8] aims to predict the price of used cars, by using machine learning supervised algorithms such as K-nearest-neighbors, linear-regression, Random Forest, Decision Tree, and XG boost. The data set that they have used to train their model consists of over 92K records. The features that it includes are year of registration, fuel type, car model, kilometers traveled, car brand, and gear type which all contribute to its worth. However, based on the frequency plots that they made they dropped two categories (ethanol and CNG) from their fuel type feature. Furthermore, they removed the records from the cars that act as outliers which are cars that cost more than 400K euros or less than 0 euros. The results concluded that the Random Forest Regressor has the lowest RMSE with a value of 3702.34 as well the highest R-Squared value of 0.93. The limitation of their research is having small data set to make a strong inference, as well as they believed having more features could result in good predictors.

Venkatasubbu and Ganesh [1] aim to compare the performance of three machine learning algorithms in predicting the price of a used car, which are regression trees, lasso regression, and multiple regression. The dataset was collected from Kelly Blue Book Central Edition 2005, which contains 804 records of cars. Data has been randomly divided into training 70%, which represents 563 records and testing 30%, which represents 241 records. Based on the compression result, Regression Tree has the lowest error rate at 3.512% with the highest performance, compared to Lasso Regression at 3.581% and Multiple Regression at 3.468%.

Pal et al. [10] presented a model for predicting prices of used cars using Linear Regression and Random Forest along 500 Decision Trees with the 'Used Car Database' dataset from Kaggle containing the attributes and prices of more than 370,000 used cars and a 70:20:10 split ratio is used for training, testing, and cross-validation. However, Random Forest outperformed Linear Regression as it Solved the overfitting problem resulting in an accuracy of 95.82% for training and 83.62% for testing. Moreover, applying more methods like genetic algorithms and fuzzy logic would be considered in future work.

Monburinon et al. [3] conducted a comparative study for the prediction of prices for used cars. Three models were compared which are Random Forest regression, gradient boosted regression trees and multiple Linear Regression. The dataset was collected from a website called German e-commerce which is available on the Kaggle platform. The final dataset consists of 304,133 rows and 11 features after data preparation. Mean absolute error (MAE) was used to compare the results of these different regression-based models. The results showed that gradient boosted regression trees give the lowest performance with MSE equal to 0.28 then Random Forest regression comes behind which scores MSE equal to 0.35 then finally multiple Linear Regression that scores MSE equal to 0.55. The researchers intended in future work to develop this research by fine-tuning every model parameter and utilizing additional appropriate data engineering for creating better training data. Moreover, one hot encoding could be used in future work along with implementing the model in real-world applications.

Our research is dedicated to Saudi Arabia citizens and residents as the opposite of the previous research. We used a dataset that was collected from Saudi Arabia which is helpful for both the buyer and the seller in our community as the price range of cars differs from other countries. On the other hand, the previous studies were conducted in foreign countries. Thus, if the citizens would like to sell or buy their cars at a fair price the models from previous studies won't provide them with an accurate result since the training dataset differs where the prices are affected by the geometric location. However, if the model was trained on the Dataset of Saudi Arabia's used car market it will give them accurate results. Thus, our model is able to predict the prices in the Saudi Arabia market only. Moreover, the highest accuracy research was conducted for only old used cars with an R2 of 93%. Additionally, our study used a new dataset that has been collected recently in 2021.

3. DATA PREPROCESSING

3.1 Data overview

3.1.1 Description of dataset

In this study, the "Saudi Arabia Used Cars" version 3 dataset from the Kaggle platform [11] was used. The dataset was collected in the year 2021 from the Syarah platform. Additionally, Syarah is a platform for selling new and used cars in Saudi Arabia where buyers and sellers can gather in one place to sell or buy cars. The dataset contains 8,248 records of used cars information as illustrated in Table 1 as the attribute column display the dataset attributes and the description column display the meaning of each attribute.

Table 1. Description of attributes

Attribute	Description
Make	The company name
Type	Type of used car
Year	Manufacturing year
Origin	Origin of used car
Color	Color of used car
Options	Options for a used car
Engine_Size	The engine size of a used car
Fuel_Type	Fuel type of used car
Gear_Type	Gear type of used car
Mileage	Number of miles traveled by used car
Region	The region of the advertised used car for sale
Price	Used car price
Negotiable	True if the price is 0, that means it's negotiable

3.1.2 Statistical analysis of the dataset

The dataset contains four numerical attributes which are "Year", "Engine_Size", "Mileage", and "price". Table 2 displays the count, mean, standard deviation, minimum, Q1, Q2, Q3, and the maximum of the numerical attributes. Furthermore, it appeared that Mileage has a maximum value of 20,000,000 which is not possible as the average person drove 14,000 miles per year. Also, for the Price attribute, it was observed that 25% of the values are zero which is not suitable for the presented model. Moreover, extreme values will be treated in the preparation/pre-processing section.

Table 2. Summary statistics for numerical attributes

	Year	Engine_Size	Mileage	Price
count	8032	8032	8032	8032
mean	2014.097	3.287774	149152.8	53699.23
std	5.758021	1.518001	347512.2	71993.85
min	1963	1	100	0
25%	2012	2	37000	0
50%	2016	3	101785.5	37000
75%	2018	4.4	195000	73625
max	2022	9	20000000	1150000

Table 3. Summary statistics for categorical

	count	unique	top	freq
Make	8032	59	Toyota	2037
Type	8032	381	Land Cruiser	372
Origin	8032	4	Saudi	5961
Color	8032	15	White	3477
Options	8032	3	Full	3191
Fuel_Type	8032	3	Gas	7858
Gear_Type	8032	2	Automatic	6968
Region	8032	27	Riyadh	3236

As for the categorical attributes the dataset contains eight of them which are "Make", "Type", "Origin", "Color", "Options", "Fuel_Type", "Gear_Type", and "Region". Table 3 displays the count, unique, top, and frequent of the categorical attributes. Furthermore, it appeared that Toyota Land Cruiser is the most selling car, and white is the most preferable color. As for car sales, Riyadh is the most popular city as all of them get the highest counted value.

3.2 Data preparation

3.2.1 Feature selection

It appeared that whenever column "Negotiable" is set for zero it means the car does not have a set price and is open for negation, in this case, this column is not needed since the model is considered with predicting the prices only, due to that "Negotiable" feature is dropped. Moreover, the 'Origin' column is dropped as it does not contribute to the model and for the "Fuel Type" column due to it is imbalanced data as shown in Figure 2 where gas has higher values compared to diesel and hybrid, hence it is dropped.

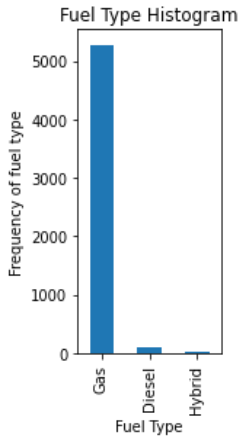

Figure 2. Fuel type

3.2.2 Dealing with extreme values

As shown in Table 2 some values seem not realistic to consider such as the car price and mileage, which will affect the skewness of the data as shown in Figure 3 below. According to Figure 3, the target variable has a skewness that is most pronounced between the values of 0.00 and 200,000 riyals. In order to not skew the results, the rows that contain prices less than 5,000 SAR are dropped. Also, the values that exceed 700,000 km in the Mileage column are dropped.

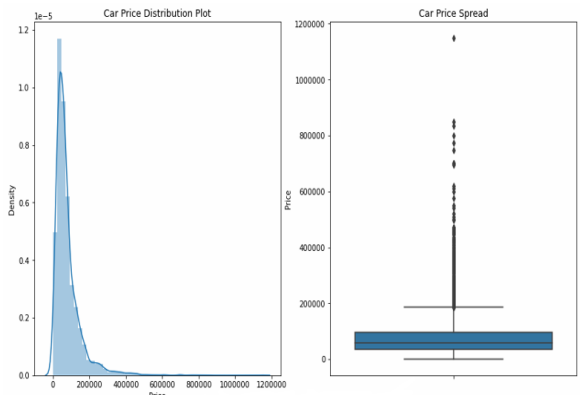

Figure 3. Car price distribution & spread

Table 4. The results of encoding categorical features

Before applying the Encoder				After applying the Encoder			
Make	Type	Region	Color	Make	Type	Region	Color
Chrysler	C300	Riyadh	Black	10.99527	10.99527	10.99527	10.99527
Nissan	Sunny	Riyadh	Silver	10.99527	10.99527	11.31961	10.99527
Hyundai	Elantra	Riyadh	Grey	10.99527	10.99527	10.95372	10.99527
Hyundai	Elantra	Riyadh	Silver	10.83211	10.83211	10.88253	10.6086
Honda	Accord	Riyadh	Navy	10.99527	10.99527	10.90477	10.99527

3.2.3 Data distribution

It is observed from Figure 3 that the data is skewed to the right so in this case, the Log Transformation will be applied, which is a method used to reduce the skewness of the data and distribute it normally by replacing each variable x with a log(x) [12]. Logarithmic transformations give a bell shape that is roughly similar to the original distribution. Furthermore, Figure 4 shows the normal distribution of data by Log Transformation.

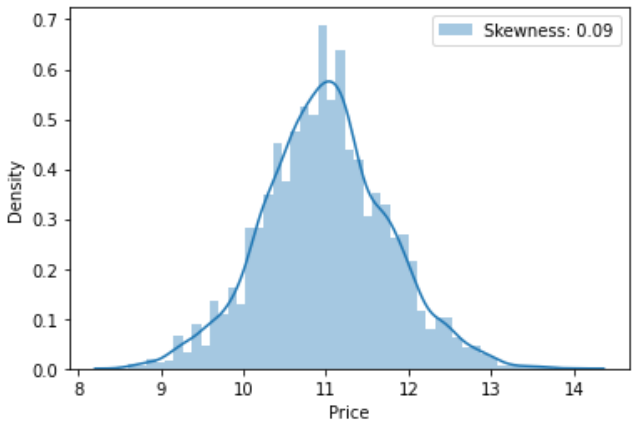

Figure 4. Normal distribution

3.2.4 Features engineering

Most of the algorithm models require numerical input and output as they cannot process any categorical data before converting it into integers to achieve state-of-the-art results. Furthermore, the dataset contains six categorical variables and two of them dropped previously as it is not needed anymore while the rest shown in Table 4 will be converted into integers by two different encoders. The used encoders were Catboost Encoder and Label Encoder. Catboost Encoder is recently introduced by Yandex researchers which gained popularity since it overcome the target leakage problem [13-15]. Moreover, The Label Encoder has a simple approach where labels are converted into a machine-readable form [16]. Table 4 present 'Make', 'Type', 'Region', 'Color' variables after encoding them into numbers.

4. MATERIALS AND METHODS

4.1 Description of the proposed techniques

4.1.1 Linear regression

A linear model that determines a linear connection between the input variables and the output variable is known as Linear Regression. The output variable might be computed by combining the input variables in a linear fashion. Linear regression analysis can be divided into two types of models which are multiple Linear Regression and simple Linear Regression. When there is just one input variable, simple Linear Regression is employed, while multiple Linear Regression is used when there are several input variables. The analysis of this section will be mainly focused on multiple Linear Regression, also known simply as multiple regression which is the technique of studying the relations between the dependent variable and multiple independent variables. There are two variables which are the dependent variable *Y* and the independent variable X_i *(i=1,2, 3, ...)* that will impact the variable *Y*. Simple regression can be expressed as the equation below:

$$Y = a_0 + a_1 X + e \text{ [17]}$$

The independent variable is *X*, and the dependent variable is *Y*. The vertical axis intercept of the regression line is a_0 , while the slope of the regression line is a_1. *e* will be used to show random factors affection on the dependent variable [17]. Multiple regression is a generalization of simple Linear Regression to the case of several independent variables which can be expressed as the equation below [18]:

$$Y = a_0 + a_1 x_1 + .. + a_n x_n + e \quad [18]$$

4.1.2 Random Forest

A Random Forest (RF) classifier is a type of ensemble classifier that is used for both regression and classification problems where the main focus of our work is regression problems. The RF model consists of several trees that work separately for predicting a class label. There are two ways to get the final prediction of the Random Forest. The most popular way is by taking the class with the majority voting to be the final class label prediction. The second way is to calculate the average for all individual trees predictions to get the final prediction [19]. RF is made up of numerous decision trees that are created during the training process and result in class labels. In RF, combining the bootstrap aggregating (Bagging) improves the performance of a single tree classifier [20]. Bagging means training each decision tree with various random samples of rows and features with replacement. The number of rows and features in the original training data should be greater than the number of rows and features for each sample. By using bagging some rows could be used in the training of more than one decision tree. One of the advantages of RF is that it usually achieves high classification accuracy. Also, RF can deal with noise and outliers in the data and it has fewer chances of overfitting.

Figure 5 shows the architecture of the Random Forest model in the proposed system. You can notice that each tree runs separately in parallel with no interaction between them. It works by feeding a Random Forest classifier a pre-processed sample of n samples. RF generates N distinct trees, each of which yields a classification result. The result of the RF is taken by the majority voting of all trees or averaging the votes [21, 22].

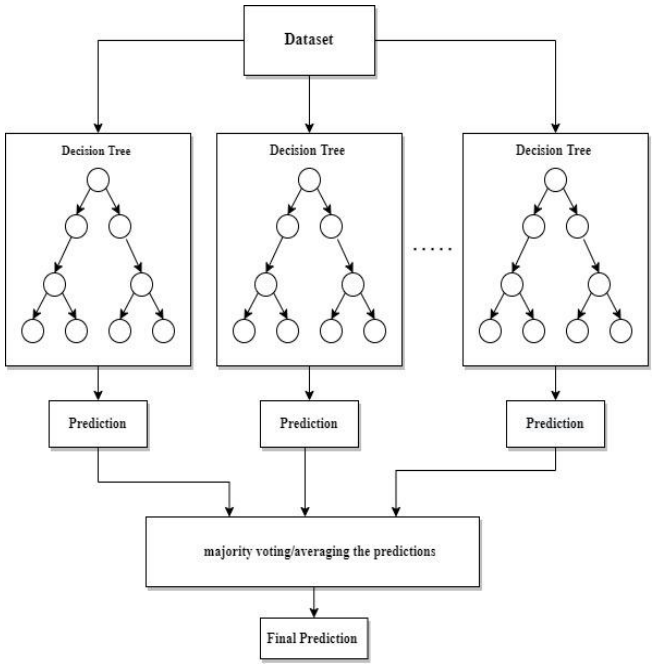

Figure 5. The architecture of the Random Forest

4.1.3 XGBoost

Extreme Gradient Boosting (XGBoost) is a sophisticated machine learning technique that was recently introduced. XGBoost is a supervised learning ensemble technique that is based on the Gradient Boosting concept. XGBoost may be used in both regression and classification problems. The XGBoost algorithm was initially established at the University of Washington as a research project where Tianqi Chen and Carlos Guestrin gave a talk at the SIGKDD Conference in 2016 [23]. XGBoost has not been credited for the several winning Kaggle competitions only but also for being the brains behind a number of cutting-edge industry applications. It combines the predictions of "weak" classifiers to achieve a "strong" classifier. Because XGBoost operates in parallel, the learning process is sped up, resulting in a faster and more accurate modelling process. In the equation below, $y^{(t)}{}_i$ is calculated which is the final tree model:

$$y^{(t)}{}_i = \sum_{k=1}^{t} f_k(x_i) = y^{(t-1)}{}_i + f_t(x_i) \quad [24]$$

Also, $y^{(t-1)}$ represents the previously created tree model, and $f_t(x_i)$ represents the newly created tree model, and it indicates the total number of base tree models [24-26].

4.2 Experimental setup

This experiment was carried out in a Google Research product called the Google Colab environment. Python code can be written and executed in the browser by anyone, making it a great tool for machine learning, data analysis, and education [27]. Therefore, we used it for our regression models. It was used to train the three proposed models Linear Regression, Random Forest, and XGBoost. The Saudi Used Car dataset originally contains 13 features including the target variable which is the price of the used cars and considers a regression type. The dataset contains 8248 records and does not contain any missing values however there were outliers in the dataset that need to be handled. Consequently, we exclude any records with a car that has a price below 5 thousand SAR. Furthermore, we dealt with an extreme value that appears in the dataset that was not within the scope of the experiment. Therefore, we exclude and record that the mileage for the car is above 700 thousand. Moreover, we convert the categorical variables into integers that were in the columns [Make, Type, Color, Region, Gear Type, Options]. Also, a feature selection technique was conducted. Therefore, the features [Negotiable] have been removed as it does not have any meaning for the prediction. As well as the features [Origin, Fuel Type] has been removed as well as their values mostly consist of one specific type and there is not much diversion. Therefore, the input for the given model was [Make, Type, Year, Color, Options, Engine Size, Gear Type, Mileage, Region, and Price]. For comparison with the previous study using the same dataset, we used the same train test splitting to divide the data into 70%, which represents 5,625 records and testing 30%, which represents 2,410 records.

4.3 Performance measure criteria for the proposed models

4.3.1 R Squared (R2)

In a regression model, R2 is a statistical measure that represents the performance of the constructed model and an indicator of how much of a dependent variable's variation is explained by an independent variable (s). The values of R-squared fall in the range between 0 and 1. R- squared is different than MSE and RMSE in that the R2 score is independent of context means in MSE the value we get after determining MSE is a squared unit of the target attribute. for instance, the target attribute is in meter(m) then the calculated MSE we get is in meter squared. With R squared we have a baseline model to compare which none of the other metrics can provide. In a normal case, R2 is when the score is between zero and one for instance 0.8 which indicates that the model is capable to explain 80% of the variance of the data [28, 29].

The formula of R squared is:

$$\text{R2 squared} = 1 - \frac{SSr}{SSm} \text{ [29]}$$

where, the squared sum error of the regression line is SSr, while the squared sum error of the mean line is SSm [29].

4.3.2 Mean squared error

In this work, one of the evaluation methods that are used is mean squared error. It will be considered to evaluate the three proposed methods and analysis the best among them. The reason it has been chosen to rely on is that our problem that predicts the price of the used car in Saudi Arabia is a regression problem. The formula for MSE is:

$$MSE = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n} \text{ [30]}$$

where:

- y_i is the i^{th} actual value
- $\hat{y}_I$ is the predicted value the model has predicated
- n is the number of data points

MSE is calculated similarly to variance. It is calculated by taking the actual value, subtracting it from the predicted value, and squaring the difference. Divide the sum of all the calculated values by the number of observations and repeat the operation for all observed values. The lower the value of MSE the better, 0 means the model is perfect however MSE's basic

value is relying on deciding on one prediction model over the other [30, 31].

4.3.3 Root mean squared error

The third approach of measuring the performance for the proposed method is the root mean squared error. The formula for RMSE is:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} \|y_i - \hat{y}_i\|^2}{n}} \text{ [32]}$$

In order to calculate RMSE, we must first determine the difference between each data point's actual and predicted value. After this, we compute the mean of residuals and take the square root of that. As a result of the fact that it requires actual measurements at each predicated data point, RMSE is considered to be one of the most common methods in supervised learning [32]. The value of RMSE that can relatively predict the data accurately is between 0.2 and 0.5 and this is based on the rule of thumb.

4.4 Optimization strategy

Hyperparameters optimization has a significant impact on the performance of a model [33, 34]. Hence, in order to adjust hyperparameters for optimization of models, many experiments were conducted by utilizing a grid search method to determine the optimal values of hyperparameters for a given model [35]. The GridSearchCV method search through the best combinations of values for each parameter that is given in a set of a grid of parameters using the Cross-Validation method. Table 5 contains the optimization values of the parameters for the Random Forest and XGBoost. Regarding the Linear Regression, the default parameter value was used. While in the Random Forest Grid Search has been used to choose the optimal values between max depth: [11, 12, 13, 14, 15] and N estimators: [30, 50, 100] where it has been found the optimal parameter values are max depth of 11, and N estimators of 50. While for the XGBoost, the best parameter values are a learning rate of 0.1, N estimators of 100, and a max depth of 15 to be the best for our problem.

Table 5. Optimization parameters for the proposed algorithms

Classifiers	Parameter	Optimal value chosen
Random Forest	Max depth	11
	N estimators	50
XGBoost	learning rate	0.1
	N estimators	100
	Max depth	15

4.5 Result and discussion

Summarizing of experimental results for the three proposed models, the Random Forest achieves the best in terms of RMSE, MSE, and R-squared out of the three models as shown in Table 6.

Table 6. Result of the proposed models

Model	RMSE	MSE	R-squared
Linear Regression	0.38	0.15	0.74
Random Forest	**0.32**	**0.10**	**0.82**
XGBoost	0.44	0.19	0.67

The value of Random Forest RMSE is the lowest compared to the other models and a good value for it is usually between 0.5 and 0.2. While the MSE Random Forest has the lowest value and when even the value is closer to zero is considered to perform better [36, 37]. Moreover, the R-squared Random Forest gives the highest value and whenever the value is closer to 1 is better. Random Forest outperforms the other model with all the three-evaluation methods, and that is a good indicator of its powerful performance of Random Forest, and that might be a reason because it is an ensemble model which consists of multiple decision trees. Furthermore, it uses averaging to improve the predictive accuracy as well as control over-fitting.

4.6 Experimental comparison

To compare the performance of the proposed models, the results obtained from the experiment were compared with the benchmark studies. The criterion for the benchmark was the studies that predict the price of the used car where they have used the same models with a similar and different obtain dataset. The dataset we consider is Saudi Arabia Used Cars Dataset which consisted of a total of 8248 records and 13 features. Table 7 contains the comparison of the proposed models with the benchmark studies that used the same models with different obtained datasets. Those studies were used in section 2. The literature review with the R-squared and MSE measurements that compared accordingly. These measurements were previously mentioned in section 4.3.

Table 7. Comparison of the proposed method with the benchmark studies

Study	Dataset	Model	R-squared	MSE
[9]	web portal fred.stlouisfed.org	Linear regression	0.91	-
[8]	Data with over 92K records	▪ Linear Regression ▪ Random Forest Regressor ▪ XG boost	0.76 0.93 0.92	-
[3]	German e-commerce website	Random Forest regression	-	0.35
Proposed model 1	Saudi Arabia Used Cars Dataset	Linear Regression	0.74	0.147
Proposed model 2	Saudi Arabia Used Cars Dataset	Random Forest Regressor	0.82	0.100
Proposed model 3	Saudi Arabia Used Cars Dataset	XGBoost	0.67	0.193

As shown in Table 7 all the proposed models outperform study [3] in the MSE metric. In contrast, our proposed models have lower R-squared scores than [8, 9] due to their larger datasets. Table 8 and Table 9 contains comparison of the proposed models with code found on the Kaggle platform that used the same models which are Linear Regression and Random Forest Regressor with the same obtained dataset. the measurement compared accordingly is R-squared. R-squared measurement was previously mentioned in section 4.3.

Table 8. Comparison of the proposed method with the benchmark studies

	Dataset	Model	R-squared
Benchmark study [35]	Saudi Arabia Used Cars Dataset	Linear Regression	0.55
Proposed model 1	Saudi Arabia Used Cars Dataset	Linear Regression	0.74

Table 9. Comparison of the proposed method with the benchmark studies

	Dataset	Model	R-squared
Benchmark study [35]	Saudi Arabia Used Cars Dataset	Random Forest Regressor	0.80
Proposed model 2	Saudi Arabia Used Cars Dataset	Random Forest Regressor	0.82

Based on Table 8 and Table 9 we can conclude that the two proposed models outperformed the benchmark study as it gave 0.74 for Linear Regression which shows a significant increase in accuracy compared to 0.55 in the benchmark study and 0.82 for Random Forest regression which shows a reasonable increase in accuracy compared to 0.80 in benchmark study in terms of R-squared metric [38-40].

5. DISCUSSIONS

Based on the experimental analysis that has been conducted it conclude that Random Forest Regression model is the best model as it outperforms the Linear Regression and XGBoost models in all the evaluation methods considered in this paper. In order to investigate the rank of the features in terms of their importance to the prediction by studying the effect of the features on the target feature, which is the price of used cars, we utilized a pre-defined function that comes with the Random Forest model to calculate the importance of features then ordered them from the most important to the less important feature [41, 42]. Figure 6 illustrates the ordered features from the most important to the less important feature that significantly affects the price. The top 5 features that made a significant contribution toward the price are Type, Year, Engine size, Make, and Milage.

In Table 10 below, the price feature has been categorized into three classes which are cheap, affordable, and expensive. The cheap category represents the prices that are under 40,000 SAR, and the affordable category represents the prices between 40,000 SAR to 150,000 SAR, while the expensive category represents the prices above 150,000 SAR.

The Attribute Type represents the type of the used car, and it has 381 unique types, However, we took into consideration the top 5 Types of used cars. the most common car among all the types is Camry with 62 occurrences. In the Cheap cars, it was 4.25%, while in the Affordable it was 4.18% and there wasn't any appearance of Camry Car in the expensive category. The second type was Accent with 57 instances and the majority was in the Cheap category with 53 instances translating to 11.86% of the Cheap category while only 0.39% in the affordable. The third most common type was Land Cruiser with 54 occurrences and the highest percentage of the category that have this type is the Expensive category with a percentage of 13.87%. The fourth most common type is Taurus with 53 occurrences, and the affordable category had the highest percentage of it. The fifth one was the Corolla with 46 with 53 occurrences and both the Cheap and the affordable have a similar percentage of 3.13%.

The year of manufacture of the car has been categorized into five classes which are before 2006, 2006-2010, 2010-2014, 2014-2018, and 2018-2022. As demonstrated in the Table above the obtained dataset contains the highest number of cars in the years of manufacture between 2014-2018 in both classes cheap and affordable with the percentage of 28.19%, and 59.77% respectively however in the expensive class it has the higher count of cars fall in the range of 2018-2022 with the percentage of 53.28%. This indicates that cars of modern manufacture tend to be more expensive compared to the rest.

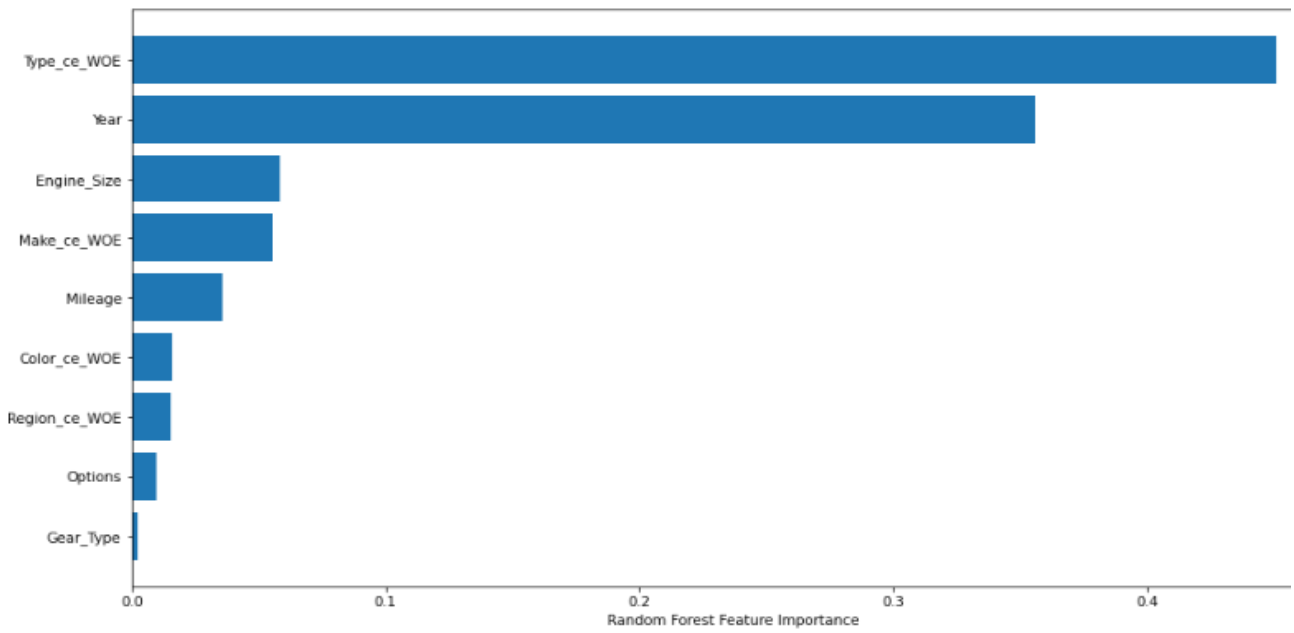

Figure 6. Feature importance

Table 10. Attribute analysis

Attribute	Cheap		Affordable		Expensive	
	count	Percentage	count	Percentage	count	Percentage
Type						
Taurus	4	0.89%	49	4.76%	0	0%
Corolla	14	3.13%	32	3.11%	0	0%
Land Cruiser	5	1.12%	30	2.92%	19	13.87%
Camry	19	4.25%	43	4.18%	0	0%
Accent	53	11.86%	4	0.39%	0	0%
other	352	78.75%	871	84.64%	137	86.13%
Year						
<2006	87	19.46%	12	1.17%	0	0%
2006-2010	86	19.24%	29	2.82%	0	0%
2010-2014	112	25.06%	201	19.53%	0	0%
2014-2018	126	28.19%	615	59.77%	64	46.72%
2018-2022	36	8.05%	172	16.71%	73	53.28%
Engine size						
1.0-2.0	236	52.80%	244	23.71%	25	18.25%
2.0-3.0	77	17.23%	298	28.96%	26	18.98%
3.0-4.0	56	12.53%	259	25.17%	38	27.73%
4.0-5.0	32	7. 16%	50	4.86%	25	18.25%
>5.0	46	10.29%	178	17.30%	23	16.79%
Make						
Nissan	57	12.75%	48	4.66%	4	2.92%
Chevrolet	31	6.94%	77	7.48%	4	2.92%
Toyota	76	17.00%	271	26.34%	22	16.06%
Hyundai	76	17.00%	125	12.15%	0	0%
Ford	54	12.08%	111	10.79%	0	0%
other	57	34.23%	397	38.58%	107	78.10%
Mileage						
<52,000	94	21.03%	246	23.91%	97	70.80%
52,000-104,100	73	16.33%	303	29.45%	29	21.17%
104,100-156,100	76	17.00%	213	20.70%	10	7.30%
156,100-208,100	59	13.20%	112	10.88%	0	0%
208,100-260,100	42	9.40%	67	6.51%	1	0.73%
>260,100	103	23.04%	88	8.55%	0	0

The engine size is a measurement of the overall volume of the engine's cylinders, which is expressed in liters. The engine size has been categorized into five classes which are 1.0-2.0, 2.0-3.0, 3.0-4.0, 4.0-5.0, and > 5. As observed from the table, the engine size has a significant effect on the price where the smaller the engine, the less expensive the car. In the cheap class a percentage of 52.80% for the cars that fall in the category of 1-2 liters however the highest percentage of cars in the affordable class is 28.96% for the 2-3 liters category. Also, the expensive class scored the highest percentage of 27.73% for the cars in the 3-4 liters category. This indicates when the engine size is increased, the price is more likely to increase.

The make attribute represents the company name for the used car. This attribute has 59 unique values, but only the top 5 frequent companies were presented in Table 1. Moreover, the most popular company among all the classes (cheap, affordable, and expensive) was Toyota with a percentage of 17%, 26.34%, and 16.06% respectively. Furthermore, the second most popular company is Hyundai with a percentage of 17%, and 12.15% for cheap and affordable classes as there were not any appearance of it in the expensive class.

The mileage represents the total number of miles the used car has travelled. The mileage has been categorized into six categories which are <52,000, 52,000-104,100, 104,100-156,100, 156,100-208,100. 208,100-260,100 and >260,100. In the cheap class, a percentage of 23.04% for the cars that fall in the category of <52,000, and the highest percentage of cars in the affordable class is 29.45% for the 52,000-104,100 category. Also, the expensive class scored the highest percentage of 70.80% for the cars in the < 52,000 category. This indicates the cars that travel fewer miles tend to be more expensive however the cheap and affordable classes tend to have cars that travel more miles.

6. CONCLUSIONS

This study conducted comparative research on various regression-based model performances. The data set used in this study comes from Saudi Arabia Used Cars Dataset on the Kaggle platform. Several regression machine learning models were applied to the dataset, which are multiple Linear Regression algorithms, Random Forest Regression, and Extreme Gradient Boosting Regression. All these models are tested using the same training data. As a criterion for comparing the results are used R squared, Mean Squared Error, and Root Mean Squared. Therefore, Random Forest provides the best recommendation for evaluating prices among the other machine learning algorithms, with RMSE and MSE values of 0.31 and 0.10, respectively. As for the R-square, Random Forest provides the highest value of 0.82, which is nearly 1. According to the literature review, the best results were achieved by Gajera et al. [8] with an R-squared of 0.93 using the Random Forest Regressor. However, their dataset contained over 92K records, which is much more than the size of our dataset that contained 8248 records and still achieved a compatible result with an R-squared of 0.82. Further, they did not use real-world Saudi Arabia datasets. In our paper, we focused on investigating the real-world Saudi Arabia dataset

as our main contribution. The previous studies have not used real-world Saudi Arabia datasets for predicting used car prices in SAR. Therefore, our goal is to help Saudi Arabia's car industry by building a machine learning model that predicts used car prices. The limitation of this study is regarding the size of the dataset where more training data could provide better results. As part of future work, frequency encoding can be used as a more realistic approach to categorical data interpretation as an alternative to CatBoost encoding. The models could also be applied for public use on a web or mobile-based applications. where all kinds of characteristic variables can be input to the model through an interface and then will directly print out the prediction price on the same interface, thus improving the efficiency and competitiveness of the used car market.

REFERENCES

[1] Venkatasubbu, P., Ganesh, M. (2019). Used cars price prediction using supervised learning techniques. International Journal of Engineering and Advanced Technology, 9(1S3): 216-223. https://doi.org/10.35940/IJEAT.A1042.1291S319

[2] Samruddhi, K., Kumar, R.A. (2020). Used car price prediction using k-nearest neighbor based model. International Journal of Innovative Research in Applied Sciences and Engineering, 4: 629-632. https://doi.org/10.29027/IJIRASE.v4.i3.2020.686-689

[3] Monburinon, N., Chertchom, P., Kaewkiriya, T., Rungpheung, S., Buya, S., Boonpou, P. (2018). Prediction of prices for used car by using regression models. In 2018 5th International Conference on Business and Industrial Research (ICBIR), pp. 115-119. https://doi.org/10.1109/ICBIR.2018.8391177

[4] سكاي نيوز عربية | 'بالشلل'فيروس كورونا يهدد صناعة السيارات. https://www.skynewsarabia.com/business/1330074, accessed on Feb. 19, 2022.

[5] Alsughayer, S.A. (2021). VAT compliance challenges among SMEs: Evidence from Saudi Arabia. Journal of Accounting, Finance and Auditing Studies, 7(3): 34-59. https://doi.org/10.32602/jafas.2021.017

[6] Used Car Price Prediction Using Supervised Machine Learning | by Shubham Jain | Medium. https://shubh17121996.medium.com/used-car-price-prediction-using-supervised-machine-learning-ea9dace76686, accessed on Feb. 17, 2022.

[7] Gegic, E., Isakovic, B., Keco, D., Masetic, Z., Kevric, J. (2019). Car price prediction using machine learning techniques. TEM Journal, 8(1): 113. https://doi.org/10.18421/TEM81-16

[8] Gajera, P., Gondaliya, A., Kavathiya, J. (2021). Old car price prediction with machine learning. International Research Journal Mod. Engineering Technological Science, 3: 284-290.

[9] Mammadov, H. (2021). Car price prediction in the USA by using liner regression. International Journal of Economic Behavior, 11(1): 99-108.

[10] Pal, N., Arora, P., Kohli, P., Sundararaman, D., Palakurthy, S.S. (2019). How much is my car worth? A methodology for predicting used cars' prices using Random Forest. In: Arai, K., Kapoor, S., Bhatia, R. (eds) Advances in Information and Communication Networks. FICC 2018. Advances in Intelligent Systems and Computing, vol 886. Springer, Cham. https://doi.org/10.1007/978-3-030-03402-3_28

[11] Saudi Arabia Used Cars Dataset | Kaggle. https://www.kaggle.com/turkibintalib/saudi-arabia-used-cars-dataset, accessed on Mar. 18, 2022.

[12] Log Transformation: Purpose and Interpretation | by Kyaw Saw Htoon | Medium. https://medium.com/@kyawsawhtoon/log-transformation-purpose-and-interpretation-9444b4b049c9, accessed on Mar. 18, 2022.

[13] Benchmarking Categorical Encoders | by Denis Vorotyntsev | Towards Data Science. https://towardsdatascience.com/benchmarking-categorical-encoders-9c322bd77ee8, accessed on Mar. 18, 2022.

[14] How CatBoost encodes categorical variables? | by Adrien Biarnes | Towards Data Science. https://towardsdatascience.com/how-catboost-encodes-categorical-variables-3866fb2ae640, accessed on Mar. 18, 2022.

[15] Categorical Encoding with CatBoost Encoder - GeeksforGeeks. https://www.geeksforgeeks.org/categorical-encoding-with-catboost-encoder/, accessed on Mar. 18, 2022.

[16] ML | Label Encoding of datasets in Python - GeeksforGeeks. https://www.geeksforgeeks.org/ml-label-encoding-of-datasets-in-python/, accessed on Mar. 18, 2022.

[17] Rong, S., Bao-Wen, Z. (2018). The research of regression model in machine learning field. In MATEC Web of Conferences, p. 01033. https://doi.org/10.1051/matecconf/201817601033

[18] Fedotova, O., Teixeira, L., Alvelos, H. (2013). Software Effort Estimation with Multiple Linear Regression: Review and Practical Application. Journal of Information Science and Engineering, 29(5): 925-945.

[19] Understanding Random Forest. How the Algorithm Works and Why it Is… | by Tony Yiu | Towards Data Science. https://towardsdatascience.com/understanding-random-forest-58381e0602d2, accessed on Mar. 19, 2022.

[20] Madanan, M., Venugopal, A., Velayudhan, N.C. (2020). Applying an optimal feature ranking and selection algorithm and Random Forest classifier algorithm along with k-fold cross validation for classification of blood cancer cells. European Journal of Molecular & Clinical Medicine, 7(11): 774-789.

[21] IEEE Xplore Full-Text PDF. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8369054, accessed on Mar. 18, 2022.

[22] Dewi, C., Chen, R.C. (2019). Random Forest and support vector machine on features selection for regression analysis. International Journal of Innovative Computing, 15(6): 2027-2037. https://doi.org/10.24507/ijicic.15.06.2027

[23] Li, W., Yin, Y., Quan, X., Zhang, H. (2019). Gene expression value prediction based on XGBoost algorithm. Frontiers in Genetics, 10: 1077. https://doi.org/10.3389/fgene.2019.01077

[24] Mo, H., Sun, H., Liu, J., Wei, S. (2019). Developing window behavior models for residential buildings using XGBoost algorithm. Energy and Buildings, 205: 109564. https://doi.org/10.1016/j.enbuild.2019.109564

[25] Gupta, A., Sharma, S., Goyal, S., Rashid, M. (2020).

Novel xgboost tuned machine learning model for software bug prediction. In 2020 International Conference on Intelligent Engineering and Management (ICIEM), pp. 376-380. https://doi.org/10.1109/ICIEM48762.2020.9160152
[26] Osman, A.I.A., Ahmed, A.N., Chow, M.F., Huang, Y.F., El-Shafie, A. (2021). Extreme gradient boosting (Xgboost) model to predict the groundwater levels in Selangor Malaysia. Ain Shams Engineering Journal, 12(2): 1545-1556. https://doi.org/10.1016/j.asej.2020.11.011
[27] Google Colab. https://research.google.com/colaboratory/faq.html, accessed on Dec. 15, 2021.
[28] R-Squared Definition. https://www.investopedia.com/terms/r/r-squared.asp, accessed on Mar. 19, 2022.
[29] Evaluation Metrics for Your Regression Model - Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/05/know-the-best-evaluation-metrics-for-your-regression-model/#h2_9, accessed on Mar. 19, 2022.
[30] Mean Squared Error (MSE) - Statistics by Jim. https://statisticsbyjim.com/regression/mean-squared-error-mse/, accessed on Mar. 18, 2022.
[31] Machine learning: an introduction to mean squared error and regression lines. https://www.freecodecamp.org/news/machine-learning-mean-squared-error-regression-line-c7dde9a26b93/, accessed on Mar. 18, 2022.
[32] Root Mean Square Error (RMSE) - C3 AI. https://c3.ai/glossary/data-science/root-mean-square-error-rmse/, accessed on Mar. 18, 2022.
[33] Gollapalli, M. (2022). Ensemble machine learning model to predict the waterborne syndrome. Algorithms, 15(3): 93. http://doi.org/10.3390/a15030093
[34] Gollapalli, M., Alansari, A., Alkhorasani, H., Alsubaii, M., Sakloua, R., Alzahrani, R., Albaker, W. (2022). A novel stacking ensemble for detecting three types of diabetes mellitus using a Saudi Arabian dataset: pre-diabetes, T1DM, and T2DM. Computers in Biology and Medicine, 147: 105757. http://dx.doi.org/10.1016/j.compbiomed.2022.105757
[35] EDA, Visualization, and Insights | Kaggle. https://www.kaggle.com/code/m0hannad/eda-visualization-and-insights/notebook, accessed on Mar. 19, 2022.
[36] Gollapalli, M., Alabdullatif, L., Alsuwayeh, F., Aljouali, M., Alhunief, A., Batook, Z. (2022). Text mining on hospital stay durations and management of sickle cell disease patients. In 2022 14th International Conference on Computational Intelligence and Communication Networks (CICN), pp. 1-6. http://doi.org/10.1109/CICN56167.2022.10008265
[37] Gollapalli, M., Alamoudi, A., Aldossary, A., Alqarni, A., Alwarthan, S., AlMunsour, Y.Z., Abdulqader, M.M., Mohammad, R.M., Chabani, S. (2022). Modeling algorithms for task scheduling in cloud computing using CloudSim. Mathematical Modelling of Engineering Problems, 9(5): 1201-1209. https://doi.org/10.18280/mmep.090506
[38] Gollapalli, M., Kudos, S.A., Alhamad, M.A., Alshehri, A.A., Alyemni, H.S., Alali, M.O., Mohammad, R.M., Khan, M.A.A., Abdulqader, M.M., Aloup, K.M. (2022). Machine learning models towards prediction of COVID and non-COVID 19 patients in the hospital's intensive care units (ICU). Mathematical Modelling of Engineering Problems, 9(6): 1471-1480. https://doi.org/10.18280/mmep.090605
[39] Gollapalli, M., AlMetrik, M.A., AlNajrani, B.S., AlOmari, A.A., AlDawoud, S.H., AlMunsour, Y.Z., Abdulqader, M.M., Aloup, K.M. (2022). Task failure prediction using machine learning techniques in the google cluster trace cloud computing environment. Mathematical Modelling of Engineering Problems, 9(2): 545-553. https://doi.org/10.18280/mmep.090234
[40] Gollapalli, M. (2015). Literature review of attribute level and structure level data linkage techniques. arXiv preprint arXiv:1510.02395. http://dx.doi.org/10.5121/ijdkp.2015.5501
[41] Gollapalli, M., Li, X., Wood, I. (2013). Automated discovery of multi-faceted ontologies for accurate query answering and future semantic reasoning. Data & Knowledge Engineering, 87: 405-424. http://doi.org/10.1016/j.datak.2013.05.005
[42] Alfaleh, A., Gollapalli, M. (2020). A critical review of data mining techniques used for the management of sickle cell disease. In Proceedings of the 12th International Conference on Computer Modeling and Simulation. http://doi.org/10.1145/3408066.3408105