JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

PREDICTION OF CRIME AGAINST WOMEN USING KMMSDL AND PPCSGO OPTIMIZATION TECHNIQUES

[1]P. TAMILARASI , [2]R. UMA RANI

[1]Research Scholar, Department of Computer Science,
Sri Sarada College for Women(Autonomous), Salem-16, India.
(ORCID: 0000-0001-9788-4129)
[2]Principal, Sri Sarada College for Women (Autonomous), Salem-16, India.

Abstract : Women's safety is now a serious concern in India. In the ongoing efforts of many nations to manage it, preventing this crime is a crucial task. The number of crimes committed against women has been rising over the past few years. In 2021, crime against women increased by 15 point 3 percent from the year before, 2020, the National Crime Recorded Bureau (NCRB) report states. The Indian government is currently interested in addressing this problem and emphasizing social development more. Each year, a ton of information is produced as a result of the reporting of crimes. We may even be able to stop crime to some extent with the help of this information, which can be very helpful for assessing and forecasting crime. The process used to carry out data analysis involves looking over, cleaning up, transforming, and modeling data. In order to support decision-making, it is important to establish valuable information and present findings. Imputations of missing data are essential in research because poor imputation of absence variables leads to inaccurate prediction. It's critical to handle these kinds of missing data well. In this article, KMMSDL approaches are suggested for handling missing values, PPCSGO soft computing techniques are used for feature selection, and ensemble-based regression approaches are used to forecast crime against women. This study's main objective is to lower errors while improving machine learning's ability to predict outcomes. The suggested algorithms KMMSDL and PPCSGO, which offer an accuracy of 97.89 percent for the India-level Crime data set, have reduced the greatest number of errors. Higher accuracy was produced by the suggested method. With the aid of this outcome, the police department would be able to successfully manage the crimes against women in India in the future.

Keywords: Ensemble Methods, Feature Selection, Missing Values Imputation, KMMSDL,PPCSGO

I. INTRODUCTION

Data analytics is a scientific method for analyzing and interpreting raw data results. Many data analytics methods perform automatically, and the algorithms work over the essential information for human consumption. Data analytics has various stages, which are descriptive analytics, diagnostic analytics, predictive analytics and prescriptive analytics. Descriptive analytics investigates past details to answer the question, "What happened?" Diagnostic analytics is an advanced method. It is helpful for a response to the question of why it happened. Predictive analytics is useful for predicting future outcomes. Prescriptive analytics is also an advanced method that helps answer the question of what will work to control the incidence.

Missing value handling is more important in analytics. If the researcher deals with the missing values correctly, the result will be accurate. When handling the lost data, the researcher does two things: first, takes the missing values using various imputation methods and the second method removes the null values. The imputation techniques have given more accuracy when the disappearance percentage is low. If the absent value percentage is high, the result may vary. The missing value is handled in this article using the proposed method, KMMSDL. This method imputes the missing values using various statistics and Machine learning techniques.

Feature selection is a more critical process in machine learning. It works on the concept of garbage in and garbage out. Only some features are helpful when executing the machine-learning techniques, while others are useless. Using irrelevant information for the prediction may reduce the model's overall performance. Hence, it is essential to determine the irrelevant information in the datasets with the help of machine learning techniques. This paper uses the proposed method, PPCSGO, for feature selection. It is a hybrid technique of P-Value PCA and the Stochastic Gradient Optimization Method.

The ensemble is a method. The main goal of this method is to improve model accuracy by combining models rather than using a single one. The combined model increases the model's accuracy significantly. It has increased ensemble methods' popularity in machine learning. Boosting, bagging, and stacking are familiar types of ensemble techniques. Linear regression and support vector regressions are ensembled in this work for crime prediction.

2. LITERATURE REVIEW

Hema N et.al [1] has proposed Predictive modelling and decision tree techniques for impute the missing values on electronic health record.Rule based classifier and Naive bayesian methods are used to perform model accuracy. **Phiwhorm K et.al [2]** proposed adaptive imputation of missing values techniques for filling absent information with the help of three modules.In the first module the author focuses the pre-processing techniques and the author calculated the threshold value by using class centers and distance values of data samples.Finally discuss the missing values in the third module.**Mustafa Alabadla et.al [3]** proposed ensemble based novel approach (Extra impute)for impute the missing values on healthcare datasets.this proposed method performances are compared with other imputation techniques like miss forest,multivariate impute,KNN Impute.Finally the proposed result observed good performance.**Dharmendra Patel et.al[4]** implemented single and multiple imputation methods for filling the absent information.Mean and mode values are used for impute the single implementation and regression is used as multiple implementation.**Rahin Atiq et.al[5]** has proposed many algorithms for impute the missing data such as Gradient boost tree,KNN,Mice,Deep Learning etc.Finally the author found that the KNN along with deeplearning methods perforances are outstanding .It has given 100% accuracy compare than other classification prediction results.**Ritu Aggrawal et.al[6]** performed prediction of various fields such as student performance on early stage,crop production,disease prediction using various machinelearning techniques by applying the p-value for feature selection.**Saba Bashir et.al [7]** implements various feature selection techniques for classification using SVM on cleveland heart disease dataset with 94.45% accuracy.The author measures SVM performances using sensitivity,specificity and f values.**AzzaAli et.al[8]** has proposed fuzzy k top values for impute the missing values using numerical and categorical values.Finally conclude the output based on RMSE and execution time.the proposed techniques FKTM performed better with low RMSE and time.**Phimmarin Keerin[9]** improved the knn iputation performance on gene expression data by proposed method ordered weighted average methods.the author used six various gene expression records and conclude the result the missing ratio 5% and 10% values are given more accurate. **Priyanka Gupta[10]** presented ensemple learning method such as majority-voting ,stacking and bagging for improving the heart disease prediction accuracy.The author implemented various classification techniques for prediction and conclude the result as the majority has given better result compare with others.

3. PROPOSED METHODOLOGY

3.1 DATA PRE-PROCESSING

Pre-processing is primary portion of developing machine learning performance.Pre-Processing is utilized to convert raw information to appropriate data.the normalization is one of the most familiar pre-processing methods this is implemented for extract the useful information.Mostly records are in numeric values.The classical pre-processing methods are used to handled these types of records.Standardization is one of the familiar classical pre-processing techniques.This method avoids the scale of attributes.Missing values imputation is one of the major process in data pre-processing.In this work,the soft computing proposed method(KMMSDL) is used for impute the missing values.Dimentionality reduction is main part in this paper.This is some other hellenic unsupervised machine learning techniques.The major role of this method is to reduce the dimension which means reduce the no of the features.In this paper proposed PPCSGO method for feature selection.The below algorithm1 used for impute the missing values.

Algorithm 1: Impute the Missing values using KMMSDL Methods

Input : Load the dataset **(CAW_I)**with missing values Attribute**(MA)a1,a2,a3…**

Output: Predict the value using Proposed Imputed Method **KMMSDL**

Step 1 : Read the Record CAW_I

Step 2 : Find the missing value place

Step 3 : Estimate the Mean,Mode,Standard Deviation values using the below formula.

$$\mu = \sum Xi / N \qquad (1)$$

Here μ denotes mean value of population .

$$\text{Mode} = L + h \frac{(fm-f1)}{(fm-1)+(fm-f2)} \qquad (2)$$

Here L is the lowest limit,f is the frequency of the class,h is the interval of the values

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(xi-\mu)^2}{n-1}} \qquad (3)$$

σ denotes standard deviation ,xi is the ith position value,n denotes total no of values

Step 4 : Take the log value for calculated values and **Step 5 :** Estimate the KNN value for missing values.

Step 6 : All statistics values are added with KNN. **Step 7 :** Fill the computed values in Missing Place.

3.2 FEATURE SELECTION

In this article the proposed method PPCSGO used for feature selection.which is ensemble method of P-Value,PCA and SGO(Stochastic gradient optimization techniques.

- P-Value

There are numerous features in the data-set that we encounter while developing a machine learning function for a real-world data-set, and not all of these property are always crucial. When breeding a model, adding unused attribute causes the model to be biased, more analyzable, and less veracious overall.Following are some well-liked methods for feature option in machine learning.there are filter,wrapper and embedded techniques.Backward elimination is a technique for retaining only the attribute that are important to the data-set, that is, those that importantly impact the parasitic variant.The amount of alteration a property will make to the selected output determines its substance level, or how crucial it is and how much it causation the conclusion.The significance level's p-value is pertain to as the hypothesis.In this article the P value is used for selecting the features.The P-value is a important statistical factor that aids in deciding whether the hypothesis is true or false. P-value can only ever be found between 0 and 1. A preset bar that should be set by the researcher is the level of significance . Normally, it is fixed to 0.05. The P-value calculation formula is $z = \frac{P^{\cap} - P0}{\sqrt{\frac{P0(1-P_0)}{n}}}$ (4)

In formula four, P stands for samples, P0 for sample proportion null hypothesis, and n for sample count.The below Fig 1 shows the work flow of crime prediction.

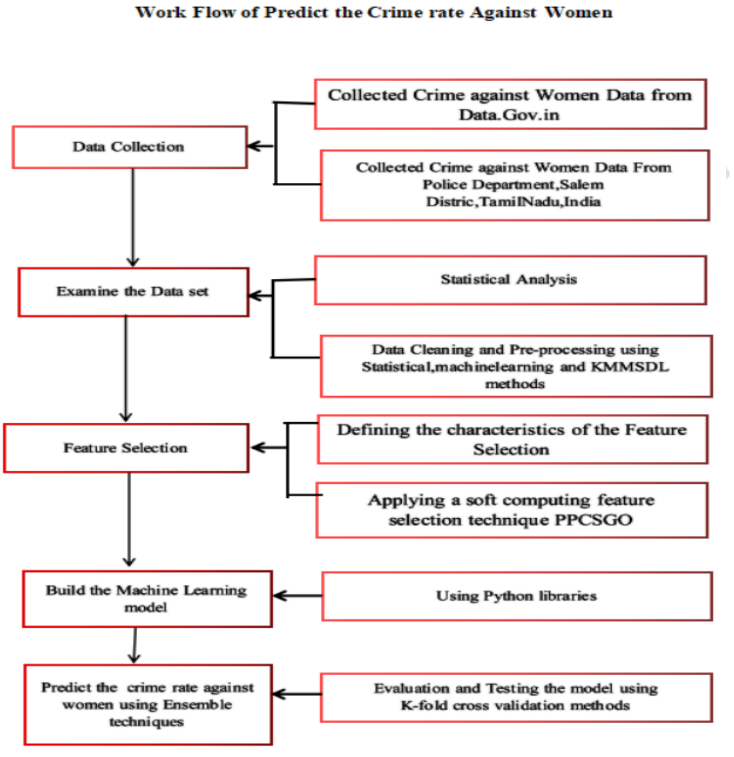

Fig 1.Work Flow of Crime Prediction Against Women

- PCA - Principle component Analysis

As part of the statistical procedure known as PCA, rigid attributes of potentially correlated variables are transformed using an perpendicular transformation to produce a set of principal component values, which are values of linearly unrelated variables. In many fields, there is a proliferation of excessive amounts of data, but at the same time, it is becoming more difficult to interpret them. However, a variety of statistical techniques were required to significantly reduce their conditionality while protecting the majority of the information in the data in order to extract information from it. In other words, in order to reduce the possibility of over fitting, it is essential to trim the feature space in order to better understand the relationships between the variables. Lowering or reducing the feature space's dimensions is the process of dimensionality reduction. "Feature Exclusion" and "Feature Extraction," respectively, are the two techniques. is PCA is one of the familiar techniques for doing this ,

which has the straightforward goal of cut down the spatial property of a data-set while preserving statistical information as much as possible.The following are some advantages of PCA.

- ✓ Given the orthogonal components, there is no data redundancy. Removes correlated features because principal components don't depend on one another.
- ✓ Since PCA removes correlated variables that are irrelevant to decision-making, it enhances the performance of the ML algorithm.
- ✓ By reducing the number of features, PCA aids in overcoming problems with data over fitting.

✓ Since PCA produces high variance, visualization is enhanced. noise reduction because the maximum variation basis is selected and the small variations in the background are automatically disregarded.

- Stochastic Gradient optimization

To find the model parameters that most closely match the predicted and observed outputs, machine learning applications frequently use the stochastic gradient descent optimization algorithm. It is a simple but successful strategy. Strictly speaking, stochastic gradient descent is utilized in machine learning projects. When combined with back propagation, it dominates neural network training applications.By changing the decision variables, the cost function, also known as the loss function, must be minimized. Numerous machine learning techniques address deeper optimization issues. By changing the model parameters, they typically try to reduce the discrepancy between the actual and predicted outputs.In a regression problem, the actual outputs y and the vectors of the input variables x = (x1,..., xr) are usually present. To get f(x) as close to y as possible, you want to find a model that converts x to a predicted response f(x).The primary objective of this approach is to reduce the discrepancy between the predicted value f(x) and the actual data y. The residual refers to this variation.To reduce the sum squared residuals the below formula 5 are used

$$SSR = \sum_i (y_i - f(x_i))^2 \quad (5)$$

Where f(x) is predicted value and y represent actual data.The below algorithm 2 explains the feature selection process.

Algorithm 2: Feature Selection using **PPCSGO** Method

Input : Load the crime against women dateset **Output:** Predict the crime rate using Proposed Method **PPCSGO**
Step 1: Read the Data Set **Step 2:** Find missing values **Step 3:** Fill the missed data by using **KMMSDL** method **Step 4:** Calculate the P value for each attributes **Step 5:** Set the 0.05 as the threshold value Step 6: Apply PCA techniques for select the feature Step 7 : Execute Stochastic Gradient Optimization method for better accuracy Step 8: Predict the crime rate using Ensemble Method.

3.3 ENSEMBLE METHOD

Ensemble is a most familiar machine learning method.This model suffers by bias or variance.Bias mostly used for calculate the difference of actual and predicted value by model.When a model creates a simple model without taking into account the variation in the data, bias is introduced.The uncomplicated model getting errors when predicting both training and testing data because it doesn't follow the patterns of the data which mean the model has given high bias and variance.The model may perform exceptionally well on the training dateset, which indicates It provides low bias but fails on the test data-set and provides high variance.Therefore, ensemble learning techniques are developed in order to increase the model's accuracy. Combining several models that have been trained with machine learning algorithms is known as an ensemble. It combines weak learners, or low performing classifiers, with individual model predictions to produce the final prediction.In this article,The averaging ensemble methods are used along with linear regression.SGD regressor for crime predictions.

4. RESULTS AND DISCUSSIONS

- Statistical Analysis

The below Fig.2,Fig.3 shows missing values percentages of each column and Fig..4 shows descriptive report of the crime against women dateset

```
dataset.isnull().mean() * 100

c1       8.695652
c2       0.000000
c3       4.347826
c4      13.043478
c5       4.347826
c6       8.695652
c7       0.000000
c8       4.347826
c9      13.043478
c10      4.347826
c11      4.347826
c12      4.347826
arr      0.000000
dtype: float64
```

Fig.2 .Missing Values in Percentage

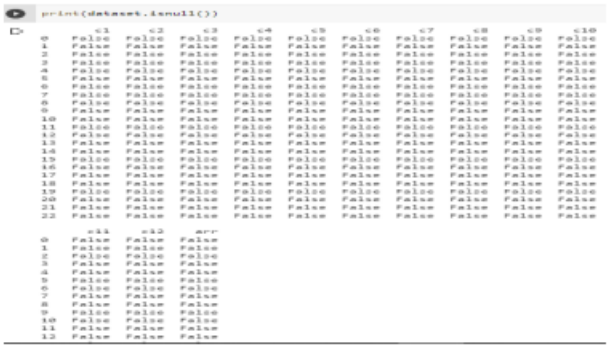

Fig.3. After filling the Missing values using KMMSDL Method

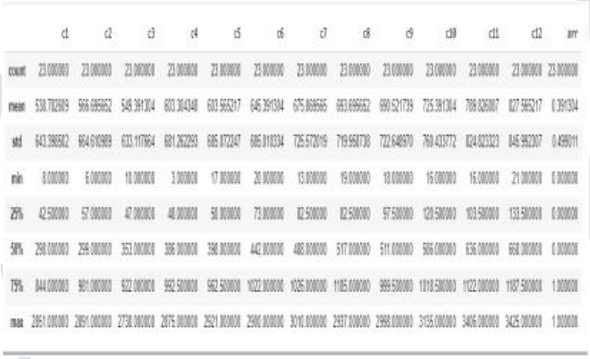

Fig.4.Descriptive Analytics of Crime Against Women Dataset

The below Fig.5 and 6 shows P values for each column,This value is used for selecting the features.Using this values removed the features which is greater than threshold value 0.05

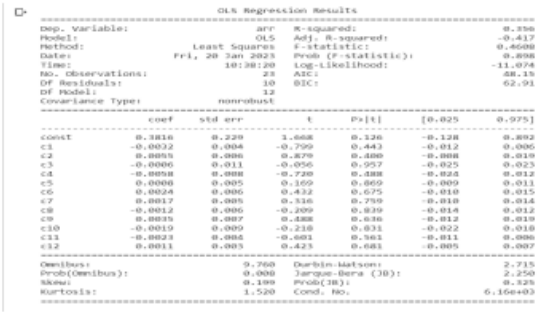

Fig .5. Find the P values for Selecting the Features

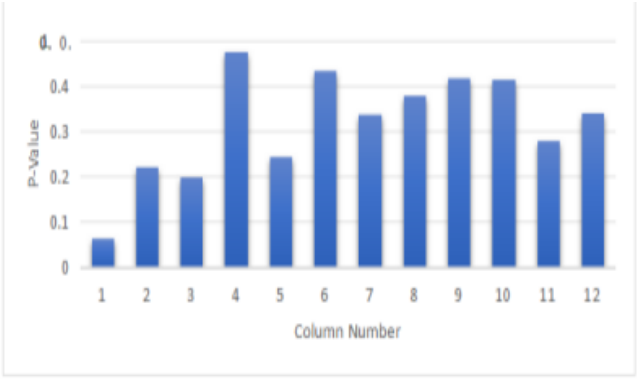

Fig 6: Compare the Each column P values

The following Fig 7,8,explains the PCA feature selection process.The Fig.7 shows standardized value in 4 PCA components.Fig.8 visualize the values in two principle component basis.

	PC1	PC2	PC3	PC4
0	2.466345	-0.494790	-1.056729	-0.021762
1	-2.950772	-1.010747	0.279723	0.117154
2	3.167312	1.662871	-1.246003	0.263367
3	2.235318	-0.580690	0.077336	0.788529
4	1.735961	1.386471	-0.014653	-0.036346

Fig .7. Scaled values in PCA

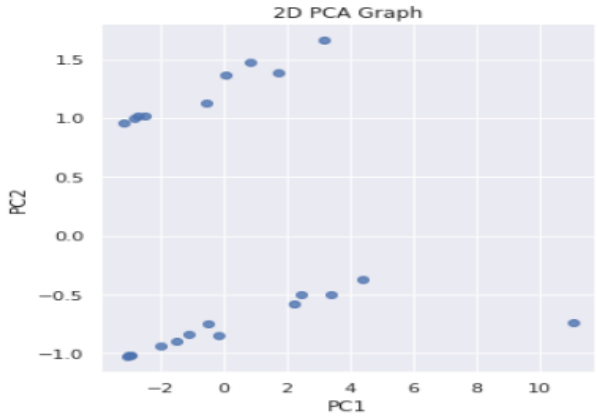

Fig 8: 2D based PCA

The below PCA Fig. 9 shows the variance of the Ratio.

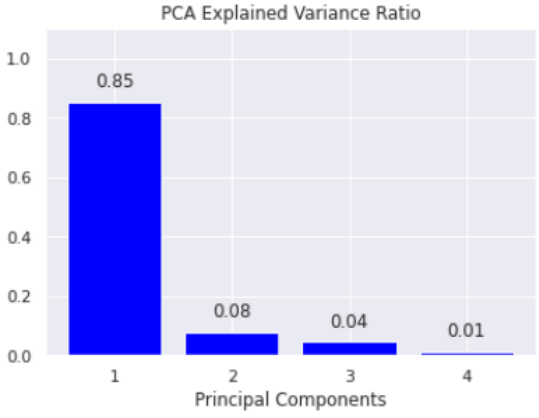

Fig .9 . PCA Variance Ratio

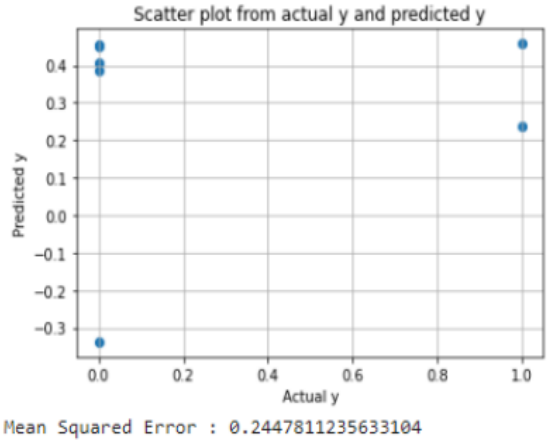

Fig .10 . PPCSGO Based Prediction

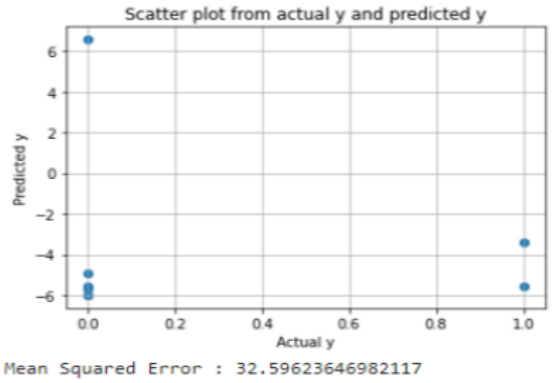

Fig .11 . PCA Based Prediction

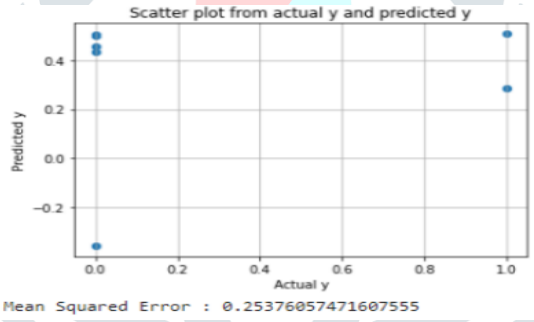

Fig .12 . P Value Based Prediction

The above Fig.10 shows comparision of Predicted and Actual Value using proposed values PPCSGO feature selection and Fig 11 shows actual and predicted values comparision using PCA, Fig 12 explains comparision of actual and predicted values using P value based feature selection.

The above Fig.10 shows comparision of Predicted and Actual Value using proposed values PPCSGO feature selection and Fig 11 shows actual and predicted values comparision using PCA, Fig 12 explains comparision of actual and predicted values using P value based feature selection.

TABLE 1. Comparison of Algorithm Performance

Data Set	**Different Types of features Used**	**Types of Errors**		
		MSE	***RMSE***	***R^2***
Crime against women in India	P-value	0.2537	0.2784	0.935
	PCA	32.5962	43.582	0.952
	PPCSGO	**0.2447**	0.2365	0.9789

The above table 1 describes the algorithms performance of Crime against women data prediction.Table 2 describes the NCRB actual and Predicted values comparision report.

TABLE 2. Comparison of NCRB Actual Values with Predicted Values

Data Set	Actual and Predicted Values			
Crime against women in India	***Row***	***Column***	***NCRB Values(Actualp***	***Predicted***
	2	6	10986	10900
	72	11	20874	20176
	194	9	18527	18394

Fig .13 . Comparision of Actual & Predicted Value Using regression

The above fig 13 give graphical representation of Actual and Predicted values using simple regression techniques.

5.CONCLUSION

Crime against women is big problem of our nation.Crime rates are continuously increasing against women.To over come this issues crime prediction is important.Data analytics is a scientific method for analyzing and interpreting raw data results. Many data analytics methods perform automatically, and the algorithms work over the essential information for human consumption.Predictive analytics is one of the familiar process in data anlytics which is helpful for predict future values.In this work, KMMSDL,PPCSGO and ensemble based algorithms are proposed for predict the crime rate.This algorithms results are compared with NCRB crime report value.This two algorithms has given 97.89 % accuracy with less Mean Squared Errors.This result will be helpful for crime department for control the crimes against women in India.

References:

[1] N. Hema and J. Selwyn, "Imputation of missing values in healthcare data using predictive modeling and decision tree," J. Pharm. Neg. Results, vol. 13, special issue 7, 2022.
[2] K. Phiwhorm et al, "Adaptive multiple imputations of missing values using the class center," J. Big Data, vol. 9, no. 1, p. 52, 2022. doi:10.1186/s40537-022-00608-0.
[3] M. Alabadla et al., "ExtraImpute: A novel machine learning method for missing data imputation," J. Adv. Inf. Technol., vol. 13, no. 5, pp. 470-476, Oct. 2022 [doi:10.12720/jait.13.5.470-476].
[4] D. Patel et al., "Single and multiple imputation techniques to treat missing numerical variables (MNV) in perspectives of data science project - A case study," Int. J. Eng. Trends Technol., vol. 70, no. 5, pp. 9-14, 2022 [doi:10.14445/22315381/IJETT-V70I5P202].
[5] R. Atiq et al., "A comparison of missing value imputation techniques on coupon acceptance prediction," IJITCS, vol. 14, no. 5, pp. 15-25, 2022. doi:10.5815/ijitcs.2022.05.02.
[6] R. Aggrawala and S. Palb, P-value feature selection technique for prediction of student performance,International Journal of Research Publication and Reviews Vol (2) Issue (8) (2021), pp. 1443-1450.
[7] S. Bashir, et al., 'A Novel Feature Selection Method for Classification of Medical Data Using Filters, Wrappers, and Embedded Approaches.' A Novel Feature Selection Method for Classification of Medical Data Using Filters, Wrappers, and Embedded Approaches, Aug. 8 2022.
[8] A. Ali et al., "Missing values imputation using Fuzzy K-Top Matching Value," J. King SaudUniv. Comput. Inf. Sci., vol. 35, no. 1, 426-437, 2023 [doi:10.1016/j.jksuci.2022.12.011].
[9] P. Keerin and T. Boongoen, "Improved knn imputation for missing values in gene expression data," Comput. Mater. Continua, vol. 70, no. 2, pp. 4009-4025, 2022 [doi:10.32604/cmc.2022.020261].
[10] P. Gupta and D. D. Seth, "Improving the prediction of heart disease using ensemble learning and FeatureSelection," Int. J. Adv. Soft Comput. Appl., vol. 14, no. 2, Jul., 2022,ISSN: 2074-8523.