ISSN: 2584-0495

Vol. 1, Issue 1, pp.1- 8

IJMIT International Journal of Microsystems and IoT

International Journal of Microsystems and IoT

Indian Society for VLSI Education

ISSN: (Online) Journal homepage: https://www.ijmit.org

Evaluating the Performance of Machine Learning Models for Diabetes Prediction with Feature Selection and Missing Values Handling

Vandana Bhattacharjee, Ankita Priya, Umesh Prasad

Cite as: Vandana Bhattacharjee, Ankita Priya, & Umesh Prasad. (2023). Evaluating the Performance of Machine Learning Models for Diabetes Prediction with Feature Selection and Missing Values Handling. International Journal of Microsystems and IoT, 1(1), 1–8. https://doi.oíg/10.5281/zenodo.8152254

© 2023 The Author(s). Published by Indian Society for VLSI Education, Ranchi, India

Published online: 26 June 2023.

Submit your article to this journal:

Article views:

View related articles:

View Crossmark data:

DOI : https://doi.org/10.5281/zenodo.8152254

Full Terms & Conditions of access and use can be found at https://ijmit.org/mission.php

International Journal of Microsystems and IoT
Vol. 1, Issue 1, pp.1- 8; DOI: https://doi.org/10.5281/zenodo.8152254

Evaluating the Performance of Machine Learning Models for Diabetes Prediction with Feature Selection and Missing Values Handling

Vandana Bhattacharjee[1], Ankita Priya[2] and Umesh Prasad[1]

[1]Birla Institute of Technology, Mesra Off Campus Lalpur, Ranchi, India

[2]Birla Institute of Technology, Mesra, Ranchi, India

ABSTRACT

Diabetes is a growing problem these days and many people are at risk. The effects of modern lifestyle are becoming visible in the spread of several lifestyle diseases. Obesity is one of these diseases which leads to several offshoots, diabetes being one of them. With the growth of Machine Learning (ML) techniques, researchers have been keen to apply them for predicting diabetes in the recent years. It is an important issue that the data set for applying ML techniques needs to be processed in the right way to be able to get informative results. This paper aims to study the effect of different techniques for pre-processing diabetes dataset, and then applying various ML classifiers on them. This research paper evaluates the performance of two algorithms, Logistic Regression and Support Vector Classifier based on the parameters Precision, Accuracy, Recall and F-measure. When the results were compared to other studies conducted by other researchers, it was evident that Logistic Regression performed better than all other classifiers, with the greatest accuracy of 81% utilizing only five features. The experiments have been performed using Python 3.8 and the IDE used was Jupyter Notebook.

KEYWORDS

Driver smoking; feature selection; logistic regression; support vector classifier

1. INTRODUCTION

Diabetes is a chronic health condition affecting a large number of individuals worldwide. Early detection and diagnosis of diabetes are crucial in the management of the disease. Living a stressful life or being overweight and carrying excess weight in the midsection of the body hampers insulin's activity, which eventually results in diabetes. As per the study of the International Diabetes Federation, 451 million people across the world had this disease in 2017, and this number is expected to rise to 693 million people in the next two decades [1]. Diabetes is a result of improper functioning of the pancreas, due to which the level of blood glucose becomes inconsistent, when no insulin is produced, creating type-1 diabetes or insulin resistance in cells, which results in type-2 diabetes [2, 3]. Machine learning as a data analysis tool is becoming very popular and researchers are keen to address the issues of disease prediction using various machine learning algorithms. The usage of Decision Tree, SVM, and Naive Bayes for the diabetes prediction has been suggested by the authors in [4].

This research work focuses on a dataset of diabetic female pregnant patients. The remaining section of the paper is organized as follows: The related work is presented in Section 2. The algorithms and approaches are presented in Section 3. The results are presented in Section 4, and the study is concluded with a discussion of the work's future prospects in Section 5.

2. RELATED WORK

Feature selection and missing value imputation are important steps in pre-processing data for machine learning algorithms. There have been several studies that have evaluated the effectiveness of these techniques for predicting diabetes.

One study conducted by authors used the Pima Indian diabetes dataset to evaluate the effectiveness of feature selection and missing value imputation on prediction accuracy. The authors used several feature selection techniques, including correlation-based feature selection and recursive feature elimination, and compared the performance of different missing value imputation methods, including mean imputation and K-nearest neighbor imputation. The results showed that the combination of recursive feature elimination and K-nearest neighbor imputation resulted in the highest prediction accuracy [5].

Another study conducted by authors to compare the performance of several feature selection techniques, including principal component analysis and genetic algorithms, for predicting diabetes using the Pima Indian diabetes dataset. The authors found that genetic algorithms outperformed other feature selection techniques in terms of prediction accuracy [6].

© 2023 The Author(s). Published by Indian Society for VLSI Education, Ranchi, India

In a different study, researchers used the Pima Indian diabetes dataset to assess the impact of missing value imputation on the accuracy of diabetes prediction. The authors compared the performance of several imputation methods, including mean imputation and expectation-maximization algorithm, and found that expectation-maximization algorithm outperformed other imputation methods in terms of prediction accuracy [7].

Researchers have employed a variety of machine learning methods, such as Decision Tree, Decision Table, etc., to forecast this disease. It has been demonstrated that these learning algorithms are more effective at identifying various diseases [8],[9],[10]. Due to their capacity for managing large amounts of data, their capacity to combine data from various sources to pre-process and handle erroneous data, as well as their capacity to integrate any domain information seamlessly, data mining and machine learning algorithms are advantageous in this regard[11-13][15]. Orabi et al. in [14] studied a group of people of a certain age group and designed a system for diabetes prediction for this group. Decision trees were applied and the obtained results were satisfactory. Researchers in [16] applied classification algorithms Naïve Bayes, Decision Trees and K Nearest neighbour for prediction of diabetes. A detailed review of techniques for Diabetes prediction can be found in [17]. Changsheng et al [18] developed a logistic regression model for predicting diabetes, and in our research we have taken this paper for comparison purpose. Choubey et al in [19] applied naïve bayes with genetic algorithm for feature selection, and then developed a model for classification of Pima Indian diabetes dataset. In [20] Dhomse et al applied principal component analysis for disease prediction. Several other researchers have successfully applied and performed evaluation of classification mining techniques [21-25]. Authors in [26] apply an intelligent approach, using two modules for diabetes prediction. In the first one an artificial neural network model predicts fasting blood sugar, and in the second one studies the relation of fasting blood sugar with the symptoms to predict diabetes.

In terms of recent work, the paper mentions several studies that have explored feature selection and missing value imputation techniques for predicting diabetes. Some notable findings from these studies include:

i) One study evaluated the effectiveness of feature selection and missing value imputation on prediction accuracy using the Pima Indian diabetes dataset. The combination of recursive feature elimination and K-nearest neighbor imputation resulted in the highest prediction accuracy [27].
ii) Another study compared different feature selection techniques and found that genetic algorithms outperformed other methods in terms of prediction accuracy for diabetes using the same dataset.
iii) A separate study assessed the impact of missing value imputation on diabetes prediction accuracy. The expectation-maximization algorithm showed superior performance compared to other imputation methods[28].

Overall, these recent works highlight the importance of feature selection and missing value imputation in improving prediction accuracy for diabetes. The proposed work in the article aligns with these studies by incorporating feature selection and handling missing values in the evaluation of machine learning models.

It is worth noting that the proposed work specifically focuses on the Pima Indian dataset, which includes information from 768 female patients of Pima Indian heritage. The dataset consists of 8 medical predictors and 1 target variable indicating the presence or absence of diabetes. The researchers explore various exploratory statistics of the dataset and replace incorrect values before performing the experiments.

In terms of methodology, the proposed work describes the use of feature selection methods (e.g., correlation-based feature selection, recursive feature elimination, principal component analysis, genetic algorithms), missing value imputation methods (e.g., mean imputation, median imputation, K-nearest neighbor imputation, expectation-maximization algorithm), and machine learning algorithms (e.g., logistic regression, support vector machines) for diabetes prediction. The performance of the classifiers is evaluated using metrics such as precision, recall, F-measure, and accuracy [29].

In conclusion, the proposed work in the article aligns with recent studies by exploring feature selection and missing value handling techniques for diabetes prediction. It focuses on evaluating the performance of Logistic Regression and Support Vector Classifier on the Pima Indian dataset and provides insights into their accuracy and other evaluation metrics.

Overall, these studies demonstrate the importance of feature selection and missing value imputation in predicting diabetes, and suggest that careful selection of these techniques can significantly improve prediction accuracy. However, the optimal selection of these techniques may depend on the specific characteristics of the dataset and the machine learning algorithm used.

3. METHODS AND ALGORITHMS

There are various methods and algorithms that can be used to evaluate the effectiveness of feature selection and missing values in prediction accuracy for diabetes. Here are some commonly used ones:

i) Feature Selection Methods: Several feature selection methods can be used, such as correlation-based feature selection, recursive feature elimination, principal component analysis, and genetic algorithms. These methods can help identify the most relevant features that contribute to prediction accuracy, and reduce the number of irrelevant or redundant features.
ii) Missing Value Imputation Methods: Several missing value imputation methods can be used, such as mean imputation, median imputation, K-nearest neighbor imputation, and expectation-maximization algorithm. These methods can help estimate the missing values in the dataset, and improve the accuracy of the machine learning algorithm.

iii) Machine Learning Algorithms: Various machine learning algorithms can be used to predict diabetes, such as logistic regression, decision trees, support vector machines, random forests, and neural networks. These algorithms can be trained and tested using different combinations of feature selection and missing value imputation methods, and their performance can be compared to identify the optimal combination.

iv) Evaluation Metrics: The performance of the machine learning algorithms can be evaluated using various metrics, such as accuracy, precision, recall, F1 score, area under the receiver operating characteristic curve (AUC-ROC), and mean squared error (MSE). These metrics can help quantify the prediction accuracy and reliability of the machine learning algorithm.

Overall, the selection of methods and algorithms may depend on the specific characteristics of the diabetes dataset, such as the size, dimensionality, and distribution of the data, as well as the research question and objectives of the study.

In this paper the main objective is to perform feature selection and evaluate its effect on prediction accuracy. We also aim to pre-process the data and handle missing values. The datasets have been evaluated by the logistic regression and support vector machine classifier. We now present the algorithm for feature selection and replacement of missing values.

Algorithm1

Input: Dataset D, Classifier set {Ci}, Features subset f // F is the complete set of features Output: Accuracy

```
Start

1. For each classifier Ci in the Classifier set:
   a) For each record in the Dataset D:
      i) Find missing values within the feature subset f.
      ii) Replace the missing values with either the
          Mean or Median.
   b) Apply the classifier Ci on the preprocessed
      dataset.
   c) Record the accuracy of the classifier Ci.
2. End for loop.
3. Return the accuracy values obtained for each classifier.
   End
```

This algorithm takes a dataset D, a set of classifiers {Ci}, and a subset of features f as input, and returns the accuracy of the classifiers on the dataset after performing missing value imputation on the specified subset of features.

The algorithm first loops through each classifier Ci in the provided set. For each classifier, it then loops through each record in the dataset and checks if any values are missing within the feature subset f. If there are missing values, they are replaced with either the mean or median value of the feature, depending on the implementation.

Once all missing values have been imputed, the classifier is applied to the modified dataset, and its accuracy on the data is recorded. This process is repeated for each classifier in the set.

Finally, the algorithm returns a list of accuracy values for each classifier, which can be used to compare the performance of the different classifiers on the dataset with missing values imputed using mean or median imputation.

Note that the performance of the algorithm may vary depending on the choice of feature subset and the specific classifiers used. Additionally, there may be other methods for missing value imputation that could be used instead of mean or median imputation, depending on the characteristics of the dataset and the goals of the analysis.

In this research work we applied two classifiers, namely logistic regression and support vector classifier for computing the classification parameters.

4. EXPERIMENTS AND RESULTS

4.1 Datasets

The Pima Indian dataset is a well-known dataset used for machine learning and data mining research related to diabetes. The dataset contains information from 768 female patients of Pima Indian heritage who were aged 21 or older and living near Phoenix, Arizona, USA. The data was collected by the National Institute of Diabetes and Digestive and Kidney Diseases and is publicly available. The dataset includes 8 medical predictors and 1 target variable, which is a binary variable indicating whether or not the patient has diabetes. The 8 medical predictors are:

i) Pregnancies: the number of times the patient has been pregnant

ii) Glucose: plasma glucose concentration 2 hours in an oral glucose tolerance test

iii) Blood Pressure: diastolic blood pressure (mm Hg)

iv) Skin Thickness: triceps skin fold thickness (mm)

v) Insulin: 2-Hour serum insulin (mu U/ml)

vi) BMI: body mass index (weight in kg/(height in m)^2)

vii) Diabetes Pedigree Function: diabetes pedigree function (a function which scores likelihood of diabetes based on family history)

viii) Age: age of the patient in years

The dataset also includes missing values for some of the predictors, which makes it a suitable dataset for evaluating the effectiveness of missing value imputation methods.

In this study, the Pima Native dataset was employed. Table 1 provides information about the dataset. Though we did not come across any NaN values in the dataset but we did have values in the dataset not in compliance with the values as stated below:

- Glucose level cannot be above 150 or below 70.
- Blood Pressure cannot be below 55.
- Skin thickness cannot be 0.
- BMI index cannot be 0.

Table. 1 PIMA Dataset Description.

S. No.	Attribute
1	Number of times pregnant (P)
2	Plasma glucose concentration (G)
3	Blood pressure (Diastolic) (BP)
4	Triceps skin fold thickness(mm) (ST)
5	2-Hour serum insulin (I)

6	Body mass index(kg/m2) (BM)
7	Diabetes pedigree function (DP)
8	Age (years) (A)
9	Class Variable (True or False)

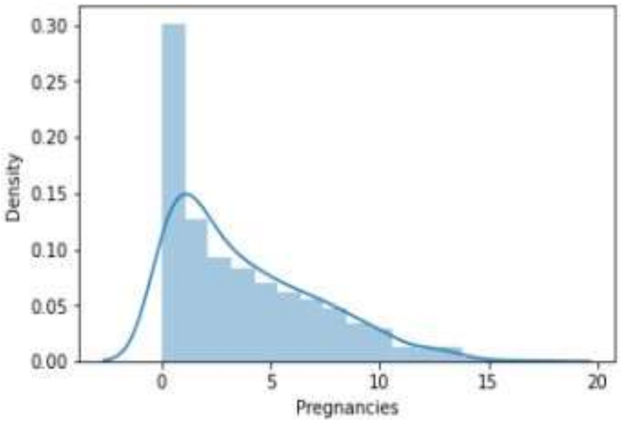


Fig. 1 Exploratory statistics of Pregnancies variable

From Figure 1, we can see that the distribution of pregnancy counts is right-skewed, with a majority of observations falling in the range of 0-5.

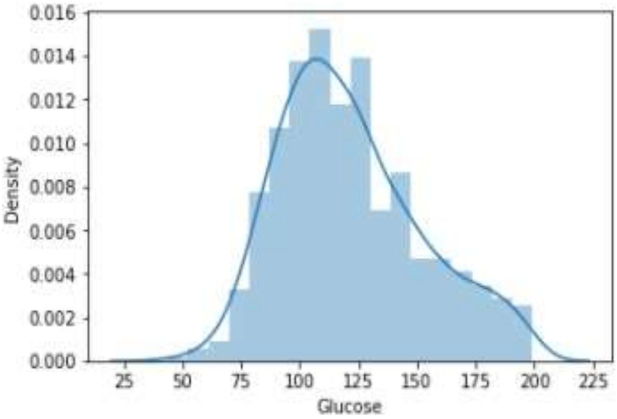


Fig. 2 Exploratory statistics of Glucose variable

From Figure 2, we can see that the distribution of glucose values is roughly normal, with a majority of observations falling in the range of 80-140.

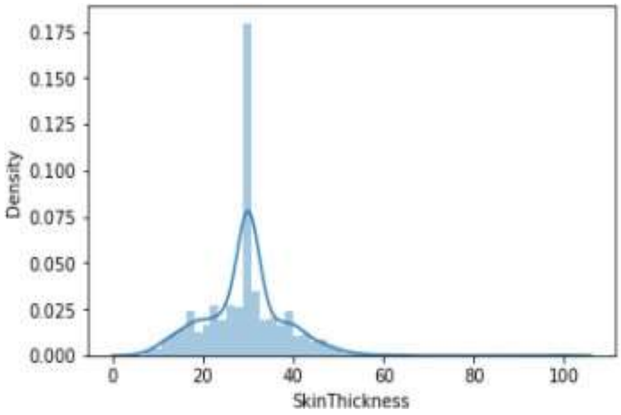


Fig. 3 Exploratory statistics of Skin Thickness variable

From Figure 3, we can see that the distribution of skin thickness values is right-skewed, with a majority of observations falling in the range of 0-30.

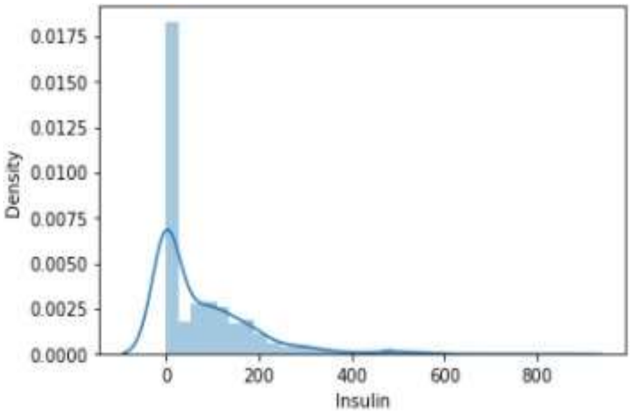


Fig. 4 Exploratory statistics of Insulin variable

From Figure 4, we can see that the distribution of insulin values is right-skewed, with a majority of observations falling in the range of 0-200.

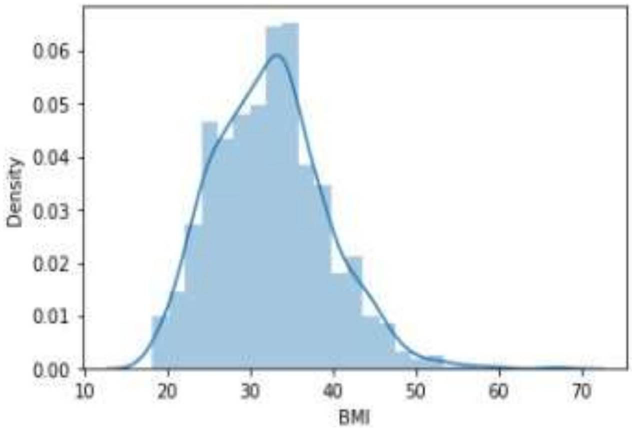


Fig. 5 Exploratory statistics of BMI variable

From Figure 5, we can see that the distribution of BMI values is right-skewed, with a majority of observations falling in the range of 20-40.

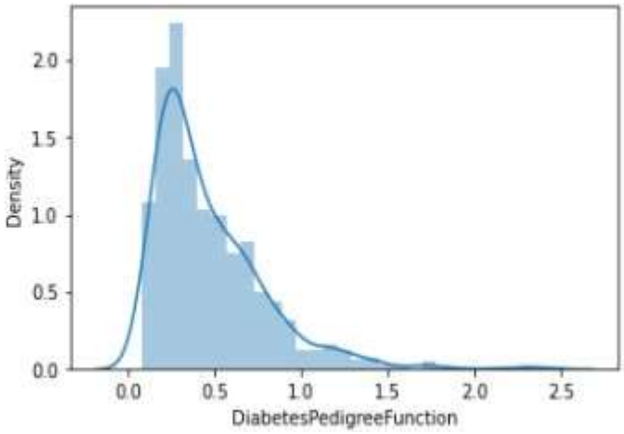


Fig.6 Exploratory statistics of Diabetes Pedigree function

From Figure 6, we can see that the distribution of pedigree function values is right-skewed, with a majority of observations falling in the range of 0-1.

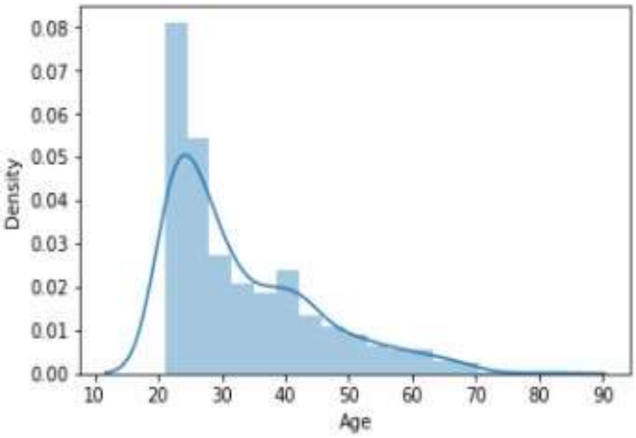


Fig.7 Exploratory statistics of Age variable

From this histogram, we can see that the distribution of age values is right-skewed, with a majority of observations falling in the range of 20-40.

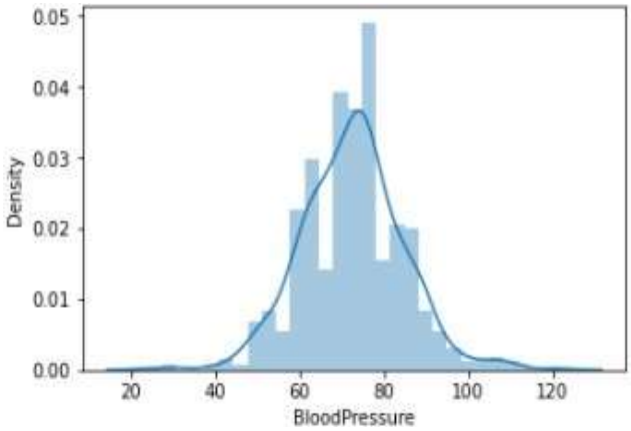


Fig. 8 Exploratory statistics of Blood Pressure variable

From this histogram, we can see that the distribution of blood pressure values is approximately normally distributed, with a majority of observations falling in the range of 60-90.

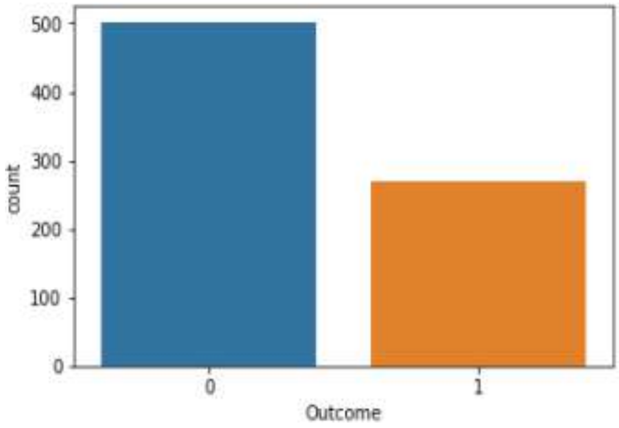


Fig. 9 Exploratory statistics of Outcome

The values which were noted to be incorrect with respect to the parameters were selected for replacement. The exploratory statistics after replacing with mean are given in the Figure 1 to 9.

4.2 Evaluation Parameters

Precision is one of the evaluation metrics commonly used in classification problems such as diabetes prediction. It measures the proportion of true positive cases (patients predicted to have diabetes who actually have it) among all the cases predicted as positive.

Other commonly used evaluation metrics for classification problems include recall, which measures the proportion of true positive cases among all actual positive cases, F-measure, which is a harmonic mean of precision and recall, and accuracy, which measures the proportion of correct predictions among all cases. Therefore, using precision, recall, F-measure, and accuracy as evaluation parameters is appropriate for this research work.

- Precision (P) is calculated as the ratio of the number of true positive cases to the total number of cases predicted as positive, which is the sum of true positive and false positive cases. The formula for precision is:

  $$P = TP(TP + FP) \qquad (1)$$

  Where TP is the number of true positive cases and FP is the number of false positive cases.
- Recall (R) is calculated as the ratio of the number of true positive cases to the total number of actual positive cases, which is the sum of true positive and false negative cases. The formula for recall is:

  $$R = TP(TP + FN) \qquad (2)$$

  Where TP is the number of true positive cases and FN is the number of false negative cases.
- The F-measure (also known as F1 score) is the harmonic mean of precision and recall, and it provides a single metric to evaluate a model's performance. The formula for F-measure is:

  $$FM = 2 * (R * P)/(R + P) \qquad (3)$$

  Where R is recall and P is precision.
- Accuracy is another commonly used evaluation metric in machine learning, and it measures the overall correctness of the model's predictions. The formula for accuracy is:

  $$A = (TP + TN)/\ TE \qquad (4)$$

  Where TP is the number of true positives, TN is the number of true negatives, and TE is the total number of examples in the dataset.

4.3 Results

The experiments were executed for all features, 5 features and 6 features out of the total 8 features. In each case we evaluated on No replacement of zero values, replacement with mean and replacement with median. The results of classification are summarized in Table 2(a) and Table 2(b).

Table. 2(a) Classification results for Support vector classifier

Features	Replace-ment	Logistic Regression Classifier			
		Accuracy (%)	Precision (%)	Recall (%)	F Measure

All	None	75	73	51	60.04
All	Mean	74	69	54	60.58
All	Median	75	73	51	60.04
5 features	None	**78**	**78**	**57**	**65.86**
5 features	Mean	75	73	51	60.04
5 features	Median	**78**	**78**	**57**	**65.86**
6 features	None	75	71	54	61.34
6 features	Mean	76	74	54	62.43
6 features	Median	75	71	54	61.34

5 features: (P,G,BP, ST, A)** 6 features: (P,G,BP, ST,I, A) **

** P – Pregnancies; G – Glucose; BP – Blood Pressure; ST – Skin Thickness; I – Insulin; A- Age

Table 2(a) shows the classification results for the Support Vector Classifier. The experiments were conducted using different feature sets (All, 5 features, and 6 features) and different replacement methods (None, Mean, Median) for zero values. The evaluation metrics include Accuracy, Precision, Recall, and F-Measure.

For the Support Vector Classifier:

- When using all features, regardless of the replacement method, the accuracy ranges from 74% to 75%.
- When using 5 features, the accuracy ranges from 75% to 78%.
- When using 6 features, the accuracy ranges from 75% to 76%.

Table. 2(b) Classification results for Logistic regression

Features	Replace-ment	Logistic Regression Classifier			
		Accuracy (%)	Precision (%)	Recall (%)	F Measure
All	None	78	74	62	67.47
All	Mean	77	73	59	65.25
All	Median	78	74	62	67.47
5 features	None	**81**	**80**	**65**	**71.72**
5 features	Mean	79	79	59	67.55
5 features	Median	**81**	**80**	**65**	**71.72**
6 features	None	80	79	62	69.47
6 features	Mean	79	79	59	67.55
6 features	Median	80	79	62	69.47

5 features: (P,G,BP, ST, A)** 6 features: (P,G,BP, ST,I, A)**

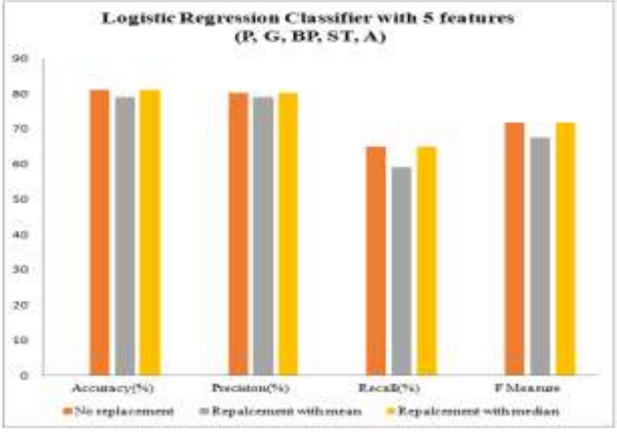
For Logistic Regression Classifier (Table 2(b)):

- When using all features, regardless of the replacement method, the accuracy ranges from 77% to 78%.
- When using 5 features, the accuracy ranges from 79% to 81%.
- When using 6 features, the accuracy ranges from 79% to 80%.

In both classifiers, using 5 features consistently outperforms using all features. Additionally, the replacement method for zero values (None, Mean, Median) does not have a significant impact on the classification results.

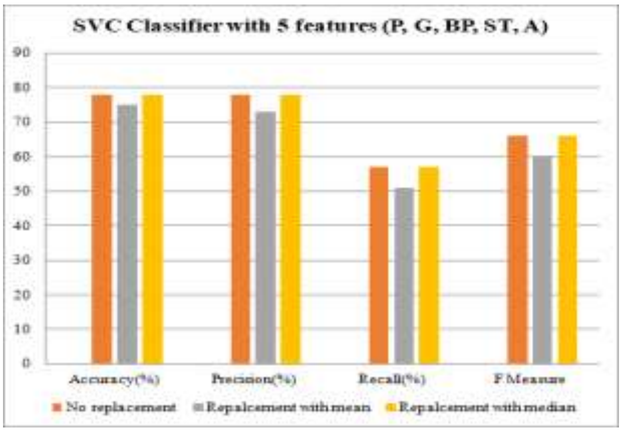
The features used in the experiments are as follows:

- 5 features: P (Pregnancies), G (Glucose), BP (Blood Pressure), ST (Skin Thickness), A (Age)
- 6 features: P (Pregnancies), G (Glucose), BP (Blood Pressure), ST (Skin Thickness), I (Insulin), A (Age)



** P – Pregnancies; G – Glucose; BP – Blood Pressure; ST – Skin Thickness; I – Insulin; A- Age

Fig. 10 Logistic Regression Classifier



** P – Pregnancies; G – Glucose; BP – Blood Pressure; ST – Skin Thickness; I – Insulin; A- Age

Fig. 11 SVC Classifier

The graphical visualization of the best results which were obtained for 5 features are given in Figure 10 and Figure 11.

Table. 3 Comparative results with other researchers

Classifier	Accuracy (%)
Support Vector Machine [Sisodia et al, 2018]	65.10
Naïve Bayes [Sisodia et al, 2018]	76.30
Decision Tree [Sisodia et al, 2018]	73.82
Logistic Regression [Changsheng Zhua et al, 2019]	77

K-Nearest Neighbours [Changsheng Zhua et al, 2019]	75
Support Vector Machine [Changsheng Zhua et al, 2019]	76
Naïve Bayes [Changsheng Zhua et al, 2019]	74
XGBoost [Changsheng Zhua et al, 2019]	76
Logistic Regression Proposed – 5 features	81
Support Vector Classifier Proposed – 5 features	78

5. CONCLUSION

In conclusion, our study demonstrates the importance of feature selection and missing values handling in improving the accuracy of machine learning models for diabetes prediction. Accuracy has always been an intuitive measure of goodness of a model. However, since this dataset is disease prediction and detecting a healthy person as diabetic would be not advisable, which means precision values should also be taken into account. Similarly, If a sick patient goes through the test and predicted as not sick, this would also be an undesirable outcome, hence, recall values should also be taken into account in defining the model goodness. Clearly stated, recall is the classifier's ability to recognize all of the positive samples. All the classification algorithms, namely, logistic regression and support vector classifier were applied on the dataset and results presented in Table 2(a) and Table2 (b). The highest accuracy of 81%, precision of 80% and recall of 65% are achieved with the logistic regression classifier applied on five features. It was further noted that replacement of missing values by mean reduced the accuracy, however replacement by median did not change the accuracy. Further, upon comparing our work with other researchers, we can see from Table 3, that the accuracy of our algorithm is the highest. Our findings have practical implications for the development of accurate and reliable machine learning models for diabetes prediction As part of ongoing work, authors propose to experiment with data augmentation approaches to improve the classification accuracy, since in medical domain collection of large datasets may not be always feasible. We also plan to work on more machine learning algorithms for classification and other feature selection approaches.

ACKNOWLEDGEMENTS

The authors thank the reviewers for their valuable suggestions which helped in preparing the paper in its present form.

REFERENCES

1. Cho, N.; Shaw, J.; Karuranga, S.; Huang, Y.; Fernandes, J.D.R.; Ohlrogge, A.; Malanda, B. IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045. Diabetes Res. Clin. Pr. 2018, 138, 271–281.
2. Sanz, J.A.; Galar, M.; Jurio, A.; Brugos, A.; Pagola, M.; Bustince, H. Medical diagnosis of cardiovascular diseases using an interval-valued fuzzy rule-based classification system. Appl. Soft Comput. 2014, 20, 103–111. [CrossRef]
3. Varma, K.V.; Rao, A.A.; Lakshmi, T.S.M.; Rao, P.N. A computational intelligence approach for a better diagnosis of diabetic patients. Comput. Electr. Eng. 2014, 40, 1758–1765.
4. Deepti Sisodiaa , Dilip Singh Sisodia, Prediction of Diabetes using Classification Algorithms, Procedia Computer Science 132 (2018) 1578–1585
5. Maniruzzaman, M., Rahman, M. J., Al-MehediHasan, M., Suri, H. S., Abedin, M. M., El-Baz, A., & Suri, J. S. (2018). Accurate diabetes risk stratification using machine learning: role of missing value and outliers. Journal of medical systems, *42*, 1-17.
6. Badawy, M., Ramadan, N., & Hefny, H. A. (2022). Healthcare Predictive Analytics Using Machine Learning and Deep Learning Techniques: A Survey.
7. Olisah, C. C., Smith, L., & Smith, M. (2022). Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective. Computer Methods and Programs in Biomedicine, *220*, 106773.
8. Aishwarya, R., Gayathri, P., Jaisankar, N., 2013. A Method for Classification Using Machine Learning Technique for Diabetes. International Journal of Engineering and Technology (IJET) 5, 2903–2908.
9. Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., Chouvarda, I., 2017. Machine Learning and Data Mining Methods in Diabetes Research. Computational and Structural Biotechnology Journal 15, 104–116. doi:10.1016/j.csbj.2016.12.005.
10. Dhomse Kanchan B., M.K.M., 2016. Study of Machine Learning Algorithms for Special Disease Prediction using Principal of Component Analysis, in: 2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication, IEEE. pp. 5–10.
11. Kumar, P.S., Umatejaswi, V., 2017. Diagnosing Diabetes using Data Mining Techniques. International Journal of Scientific and Research Publications 7, 705–709
12. Aljumah, A.A., Ahamad, M.G., Siddiqui, M.K., 2013. Application of data mining: Diabetes health care in young and old patients. Journal of King Saud University - Computer and Information Sciences 25, 127–136. doi:10.1016/j.jksuci.2012.10.003
13. Fatima, M., Pasha, M., 2017. Survey of Machine Learning Algorithms for Disease Diagnostic. Journal of Intelligent Learning Systems and Applications 09, 1–16. doi:10.4236/jilsa.2017.91001.
14. Orabi, K.M., Kamal, Y.M., Rabah, T.M., 2016. Early Predictive System for Diabetes Mellitus Disease, in: Industrial Conference on Data Mining, Springer. Springer. pp. 420–427.
15. Kumar, P.S., Umatejaswi, V., 2017. Diagnosing Diabetes using Data Mining Techniques. International Journal of Scientific and Research Publications 7, 705–709.
16. Uswa Ali Zia, Dr. Naeem Khan, Predicting Diabetes in Medical Datasets Using Machine Learning Techniques, International Journal of Scientific & Engineering Research Volume 8, Issue 5, May-2017

17. Souad Larabi-Marie-Sainte, Linah Aburahmah, Rana Almohaini and Tanzila Saba, Current Techniques for Diabetes Prediction: Review and Case Study
18. Changsheng Zhua, Christian Uwa Idemudiaa and Wenfang Feng, Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques. Elsevier, https://doi.org/10.1016/j.imu.2019.100179, 2019
19. Choubey, D.K., Paul, S., Kumar, S., Kumar, S., 2017. Classification of Pima indian diabetes dataset using naive bayes with genetic algorithm as an attribute selection, in: Communication and Computing Systems: Proceedings of the International Conference on Communication and Computing System (ICCCS 2016), pp. 451–455.
20. Dhomse Kanchan B., M.K.M., 2016. Study of Machine Learning Algorithms for Special Disease Prediction using Principal of Component Analysis, in: 2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication, IEEE. pp. 5–10.
21. Fatima, M., Pasha, M., 2017. Survey of Machine Learning Algorithms for Disease Diagnostic. Journal of Intelligent Learning Systems and Applications 09, 1–16. doi:10.4236/jilsa.2017.91001.
22. Iyer, A., S, J., Sumbaly, R., 2015. Diagnosis of Diabetes Using Classification Mining Techniques. International Journal of Data Mining & Knowledge Management Process 5, 1–14. doi:10.5121/ijdkp.2015.5101, arXiv:1502.03774.
23. Kumar, D.A., Govindasamy, R., 2015. Performance and Evaluation of Classification Data Mining Techniques in Diabetes. International Journal of Computer Science and Information Technologies, 6, 1312–1319.
24. Kumar, P.S., Umatejaswi, V., 2017. Diagnosing Diabetes using Data Mining Techniques. International Journal of Scientific and Research Publications 7, 705–709.
25. Perveen, S., Shahbaz, M., Guergachi, A., Keshavjee, K., 2016. Performance Analysis of Data Mining Classification Techniques to Predict Diabetes. Procedia Computer Science 82, 115–121. doi:10.1016/j.procs.2016.04.016.
26. Tarik A. Rashid, S.M.A., Abdullah, R.M., Abstract, 2016. An Intelligent Approach for Diabetes Classification, Prediction and Description. Advances in Intelligent Systems and Computing 424, 323–335. doi:10.1007/978-3-319-28031-8
27. Christo, V. E., Nehemiah, H. K., Brighty, J., & Kannan, A. (2022). Feature selection and instance selection from clinical datasets using co-operative co-evolution and classification using random forest. IETE Journal of Research, 68(4), 2508-2521.
28. Awawdeh, S., Faris, H., & Hiary, H. (2022). EvoImputer: An evolutionary approach for Missing Data Imputation and feature selection in the context of supervised learning. Knowledge-Based Systems, 236, 107734.
29. Saranya, G., & Pravin, A. (2022). A novel feature selection approach with integrated feature sensitivity and feature correlation for improved prediction of heart disease. Journal of Ambient Intelligence and Humanized Computing, 1-15.

AUTHORS

Vandana Bhattacharjee is working as a Professor in the Department of Computer Science and Engineering, Birla Institute of Technology, Mesra, Ranchi, and presently as Director Incharge of BIT Lalpur Campus. She completed her B. E. (CSE) in 1989 from BIT Mesra and her M. Tech and Ph. D in Computer Science from Jawaharlal Nehru University New Delhi in 1991 and 1995 respectively. She has several National and International publications in Journal and Conference Proceedings. She has authored a book on Data Analysis. She is a Life Member of Computer Society of India. Her research areas include Machine Learning and its application to Software Cost Estimation, Software Fault prediction, Classification of Images, Disease prediction, analysis of Remote sensing images. She is also currently working on representation learning.

E-mail: vbhattacharya@bitmesra.ac.in)

Ankita Priya is a Computer Science and Engineering graduate from Birla Institute of Technology, Mesra, Ranchi, with a passion for coding and research. Currently working at a reputable data solutions company, she utilizes her skills to solve complex problems and contribute to innovative projects. With a strong interest in machine learning and emerging areas of computer science, Ankita is eager to pursue a master's degree in the same field to deepen her knowledge and continue her research efforts. Her enthusiasm for coding, research, and further education sets her on an exciting path towards a future filled with innovation and transformative discoveries.

E-mail: ankitapriya2011@gmail.com

Umesh Prasad is an Assistant Professor in the department of Computer Science and Engg at BIT Mesra, off-campus Lalpur. He was awarded the PhD degree from Jadavpur University in Engineering in the year 2012. His research interests include interdisciplinary studies with a special focus upon energy management, ICT applications, Artificial Intelligence, and Machine Learning and its application. He has published more than 20 research papers and articles in various journals and conferences.

E-mail: umesh@bitmesra.ac.in