Stochastic Modelling and Computational Sciences

PERFORMANCE ANALYSIS OF MACHINE LEARNING ALGORITHMS ON DIABETES DATASET

Kamaljeet Kaur[1], Er. Amrit Kaur[2] and Dr. Navjot Kaur[3]
[1,2,3]Punjabi University, Patiala, Punjab
[1]kaurkamaljeetk4265@gmail.com, [2]amrit.tiet@gmail.com and [3]Navjot_antal@yahoo.co.in

ABSTRACT

Diabetes is an ongoing illness that has been affecting more and more people over time. It causes an enormous number of fatalities each year. Since quite a few of people have serious health problems because of late being diagnosed, it is crucial to create techniques for early disease identification given the number of deaths caused by it each year. Early detection is therefore essential. Nevertheless, when there are missing values in the information set, ML models do not perform well. . In this research, using the PIMA Indian datasets, we provided a solid machine learning framework to enhance the performance of diabetes prediction. K-NN, SVM, logistic regression, Random Forest, and Naïve Bayes are the five methods are compared on diabetic dataset and two parameters are used to analyze the result which is accuracy and error rate. Missing values are frequently explained by errors made by people while handling data, machine error because of faulty machinery, respondents' unwillingness to answer specific questions, study dropout, and merging unrelated data. However, it's crucial to deal with missing numbers before performing data analysis because doing so could lead to biased or incorrect conclusions. To address this problem, we used different classifiers for imputing the missing values. This research makes a significant contribution to the study and emphasises the significance of dealing with missing values using machine learning Techniques.

Keywords: Dataset, missing significant values. Machine learning models, imputation

1. INTRODUCTION

Diabetes is a chronic health condition that affects how your body turns food into energy. Diabetes is a condition that develops when the blood sugar or blood glucose levels in the body are unusually high. Over time, that can cause serious health problems, such as heart disease, vision loss, and kidney disease. Type 1 diabetes: This type of diabetes is caused by an autoimmune reaction that destroys the cells in your pancreas that make insulin. Type 1 diabetes usually starts in childhood or adolescence, but it can develop at any age. Type 2 diabetes: This type of diabetes is caused by a combination of genetic and environmental factors. Type 2 diabetes is often preventable, but it is also the most common type of diabetes. The goal of analytics for prediction is to enhance medical results, care for patients, resource optimisation, and disease identification accuracy. Numerous researchers are testing different classification algorithms from ML techniques to diagnose diseases. The goal of analytics for prediction is to enhance medical results, care for patients, resource optimisation, and disease identification accuracy. Numerous researchers are testing different classification algorithms from ML techniques to diagnose diseases. Missing values can be caused by several things, including entirely random missing, or not at all random missing. All of these could be the outcome of a system failure during data collecting or a mistake made by a human during data pre-processing. Missing values cause the machine learning models to learn less during the training phase, which has a detrimental impact on classification accuracy [1]. The missing values problem is typically prevalent in all data-related disciplines and results in a variety of challenges, including performance deterioration, issues with data processing, and biased results. Additionally, the amount of missing data, the pattern of missing data, and the process underlying the missingness of the data all have a role in how important missing values are certain methods, such as deleting instances and substituting prospective or approximated values, can be used to address missing values.

Values that are not present can be addressed by a variety of methods, such as deleting instances and substituting prospective or approximated values [2]. Machine learning has recently undergone a lot of research to effectively find missing data. To understand the correlation between input data and the target class, the machine learning model needs a whole set of features, but [3-4]. When there are more missing values in the dataset, machine learning models frequently have trouble producing good performance during the training stage. The procedure of identifying and dealing with the missing values for each input feature during the pre-processing stage is crucial. In recent years, numerous studies have substituted values for missing values using more sophisticated methods, such as K-Nearest Neighbour (KNN), Naive Bayes (NB), Support Vector Machine (SVM) expectation maximisation, and so on. Despite requiring lengthy computation times, these methods outperform the other conventional approaches in terms of classification performance [5].

In the projected research, we will use different machine learning classifier to fill the missing values and compare their accuracies for the diabetic patient.

1.2 Contribution of the Research

The contribution of machine learning models to finding missing values in numeric data sets depends on several factors, such as the size and complexity of the data set, the type of missing values, and the accuracy requirements. However, in general, machine learning models can be a valuable tool for dealing with missing values in numeric data sets. Machine learning models can contribute to finding missing values in numeric data sets in several ways.

- **Prediction:** Machine learning models can be used to predict the missing values based on the other values in the data set. This can be done using a variety of algorithms, such as regression, classification, or clustering.
- **Imputation:** Machine learning models can also be used to impute the missing values. This means that the model will fill in the missing values with estimated values. This can be done using a variety of methods, such as mean imputation, median imputation, or regression imputation.
- **Feature selection:** Machine learning models can be used to select features that are most relevant to predicting the missing values. This can help to improve the accuracy of the prediction or imputation process.

1.3 Related Works

The impact of eating behaviours on drug-treated and untreated mice was examined using a dataset in Rubin et al.'s [6] study on handling missing data. The expectation maximisation algorithm was applied and contrasted with other approaches such the mean substitute regression, the Bayesian technique, and list-wise deletion, which was the least effective option. The authors concluded that the EM algorithm was the most effective technique for the data type they used. By adopting the Generative Adversarial Nets (GAN) design, Yoon et al. [7] created a novel approach for determining missing values. They trained two models—a generative model and a discriminative model—and employed a two-player minimax game. It is important to note that while we are unable to analyse deep learning techniques in the ABS processing platform now, they are still a possibility in the future.

By altering the missingness ratio, Khan, et al. [8] conducted a thorough review of ensemble techniques on 8 datasets. Their findings demonstrated that multiple imputation using EM was second to bootstrapping as the most reliable technique. Bayesian spatial generalised linear models were suggested by Bakar and Jin [9] to infill values for all of Australia's statistical areas.

2. MISSING DATA METHODS

There are two main methods for handling missing data: imputation and deletion. Imputation involves filling in the missing values with estimates. This can be done using a variety of methods, such as [10]:

- **Mean substitution:** This is the simplest imputation method, and it involves replacing all missing values with the mean of the observed values.
- **Last observation carried forward:** This method replaces missing values with the last observed value for the same variable.
- **Multiple imputations:** This is a more sophisticated imputation method that uses multiple estimates to fill in the missing values. This can help to reduce bias in the results.

Deletion involves removing the observations with missing values from the dataset. This can be done either by **list wise deletion** or **pair wise deletion**. The best method for handling missing data depends on the specific dataset and the type of analysis that is being performed. In general, imputation is a better option than deletion, as it can help to preserve the information in the dataset. However, imputation can also introduce bias into the results, so it is important to choose the right imputation method for the specific dataset. Table 1. Summarizing the details of both techniques.

Table 1. Summarization of missing data handling techniques

Technique	Explanation	Advantages	Disadvantages
Imputation	Filling in the missing values with estimates	- Preserves information in the dataset - Can be used with a variety of imputation methods	Can be more computationally expensive than deletion
Deletion	Removing the observations with missing values	- Simple to implement - Can be used with any type of analysis	Can introduce bias into the results

2.1 Machine Learning Models

Machine learning models are algorithms that are trained on data to learn how to make predictions or decisions. There are many different types of machine learning models, but they can be broadly categorized into three types: supervised learning, unsupervised learning, and reinforcement learning [11-12]. Supervised learning models are trained on data that has been labelled, meaning that each data point has a known output. For example, if you are training a model to predict whether an email is spam or not, you would need to have a set of labelled emails, where each email is either labelled as spam or not spam. Once the model is trained, it can be used to predict the output for new data points.

Unsupervised learning models are trained on data that does not have labelled outputs. This means that the model must figure out how to group the data points together on its own. For example, if you are training a model to cluster customer data, you would not have any information about which customers are like each other. The model would have to learn how to group the customers together based on their similarities. Reinforcement learning models are trained on data that consists of rewards and punishments. The model learns to make decisions that maximize the rewards and minimize the punishments. For example, if you are training a model to play a game, the model would learn to make moves that lead to winning the game and avoid moves that lead to losing the game. Here are some of the most common machine learning models:

- **Linear regression:** This is a simple model that can be used to predict a continuous output. For example, you could use linear regression to predict the price of a house based on its features.
- **Logistic regression:** This is a model that can be used to predict a binary output, such as whether an email is spam or not.
- **Decision trees:** These are models that use a tree-like structure to make predictions. Decision trees can be used to predict both continuous and binary outputs.
- **Random forests:** These are ensembles of decision trees. Ensembles are models that are created by combining multiple models. Random forests are often more accurate than individual decision trees.
- **Support vector machines:** These are models that can be used to classify data or to make regression predictions. Support vector machines are often used for high-dimensional data.

3. METHODOLOGY

Figure 1 summarises the steps involved in the materials and processes, and the sections that follow go into further depth on each phase.

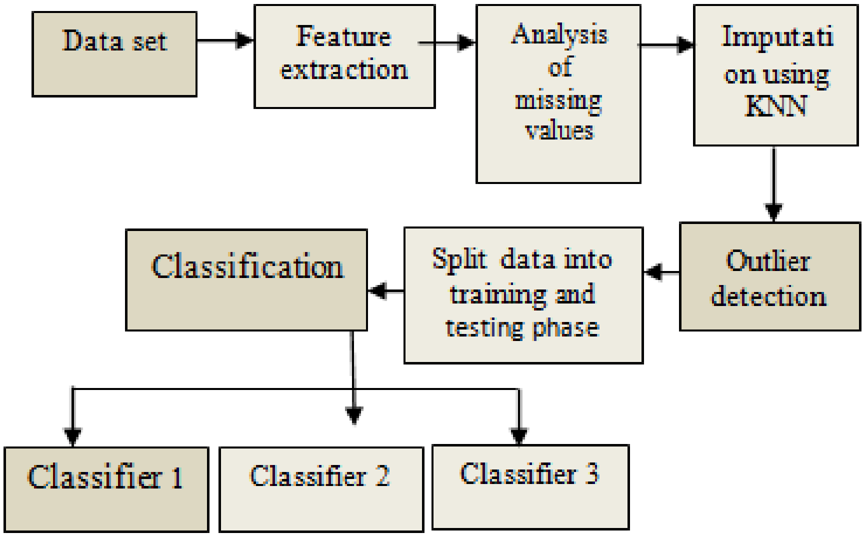

Figure 1: Proposed methodology

3.1 Data Set

The National Institute of Diabetes and Digestive and Kidney Diseases is the initial source of this dataset. Utilising specific diagnostic metrics present in the information set, the data set's goal consists of. In specifically, all patients in this facility are Pima Indian women who are at least 21 years old.

3.2 Feature Extraction

The main extracted features are *pregnancies, glucose, blood pressure, skin thickness, insulin, BMI, diabetes pedigree function, age, and outcome. The data set exhibit 768 rows and 9 columns, where outcome column plays a role of predicted output. If the output will 0 means the person is not diabetic and if its 1 means the person is diabetic. From the dataset of 768, the ration of diabetic person is 268 and rest 500 are non-diabetic. The visualization of data is presented in figure 2.*

Pregnancies Glucose Blood Pressure Skin thickness Insulin

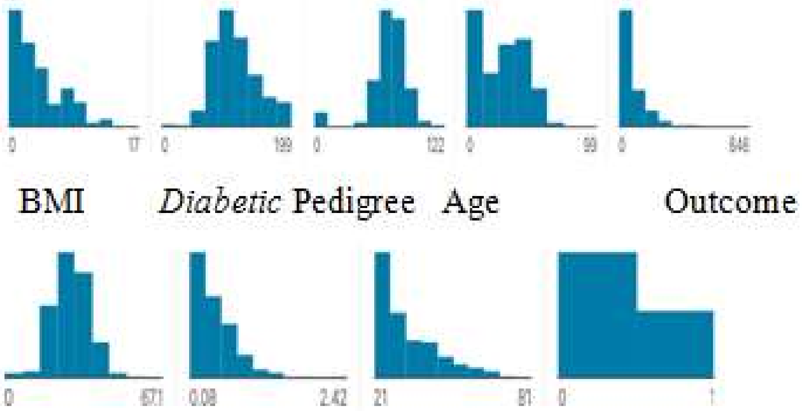

Figure 2. Structure of data

3.3 Analysis of Missing Values

For better accuracy of the model, we need to eliminate the missing values. In the data set, pregnancy column indicates no missing values or a feature that we cannot considered as missing value. We are unable to drop the column because there are about 50% and 30% missing records for insulin and skin thickness, respectively.

3.4 KNN Imputation

KNN imputation uses the k-nearest neighbours' algorithm to impute missing values from a dataset [13]. The step-by-step approach to using K-NN for imputing missing values explained in figure 3.

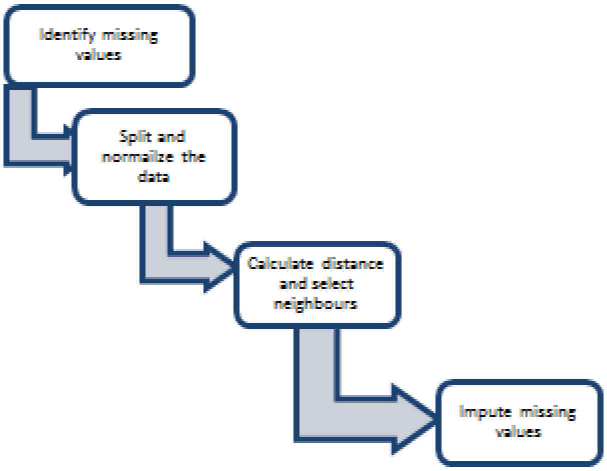
Identify missing values
Split and normailze the data
Calculate distance and select neighbours
Impute missing values

Figure 3. KNN imputation

To impute the missing value, one must first determine the k observations that are the most comparable to the observation with the missing value. K-NN method replaced all missing values in the data set for further better results.

3.5 Outliers Information

Information about the factors that were taken into consideration in relation to the values of the outliers [14] is shown in Figure 4. We can analyse from the figure that glucose, age, and BMI represents stronger association with outcome.

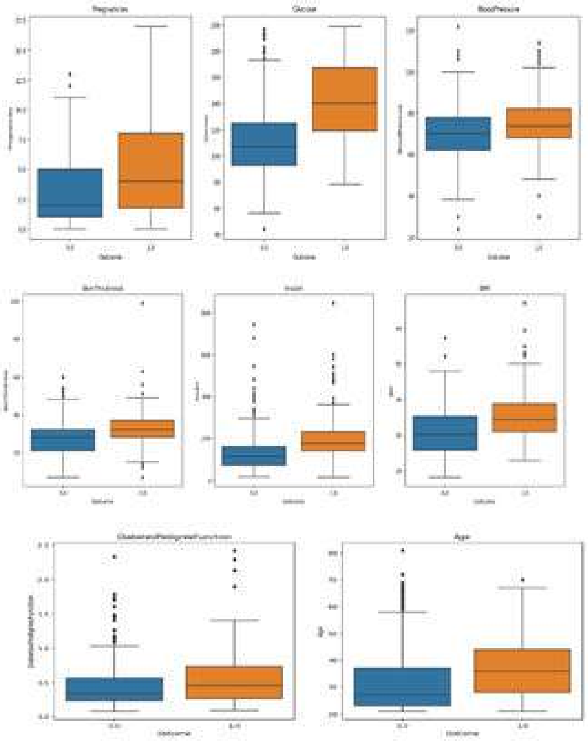

Figure 4: Outliers information

3.7 Splitting of Data

Using this technique, the data set was divided into a training set and a testing set. The model is trained using the training set. Additionally, the system is tested, and its correctness is assessed using the testing set.

3.8 Classification Models

We will use 5 classifiers namely K-NN, SVM, logistic regression, Random Forest, and Naïve Bayes for classification. We will compare their accuracies and choose the best one. We will set the classifier models' default settings when we initialise them.

4. EVALUATION PARAMETERS AND RESULT ANALYSIS

The most reliable ML classifiers, such as KNN, SVM, Naive-Bayes, logistic regression, discriminant models, and ensemble models, are used to assess the classification performance in overall accuracy, and error rate. The performance of a classifier is assessed using a variety of evaluation metrics. These all measures are dependent on four parameters as true negatives, true positives, false positives, and false negatives as shown in Table 2.

Table 2: Confusion Matrix

Class category	Predicted class 1	Predicted class 2
Class 1	TN	FP
Class 2	FN	TP

The following are some of the most typical parameters:

1) **Accuracy:** The proportion of cases that are classified properly.

$$\text{Accuracy}=\frac{TP+TN}{TP+TN+FP+FN} \qquad (1)$$

2) **Error rate:** The number of all incorrect interpretations divided by the sum of all accurate and incorrect interpretations, or the dataset, is used to calculate error rate.

Error rate: $\frac{FP+FN}{TP+TN+FP+FN}$ (2)

3) **Precision:** This is the percentage of favourable situations that are genuinely positive.

Precision: $\frac{TP}{TP+FP}$ (3)

4) **Recall:** This is the proportion of positive events that are genuinely positive and are labelled as such.

Recall: $\frac{TP}{TP+FN}$ (4)

Fig.5 and 6 represents the graphical visualization of performance of classifiers.

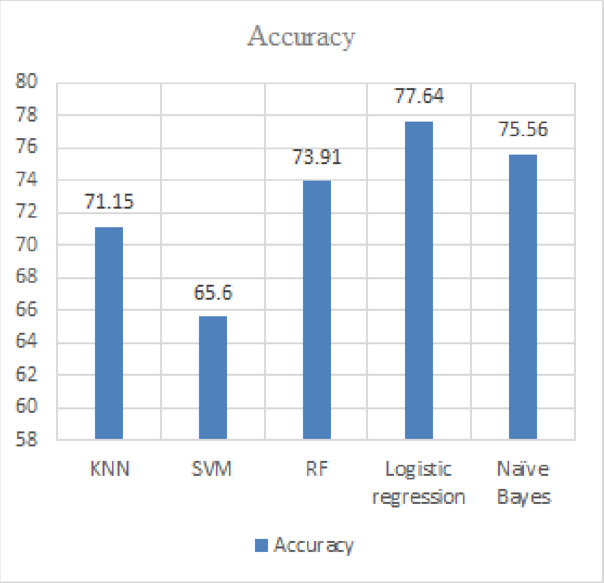

Figure 5: Performance analysis using accuracy.

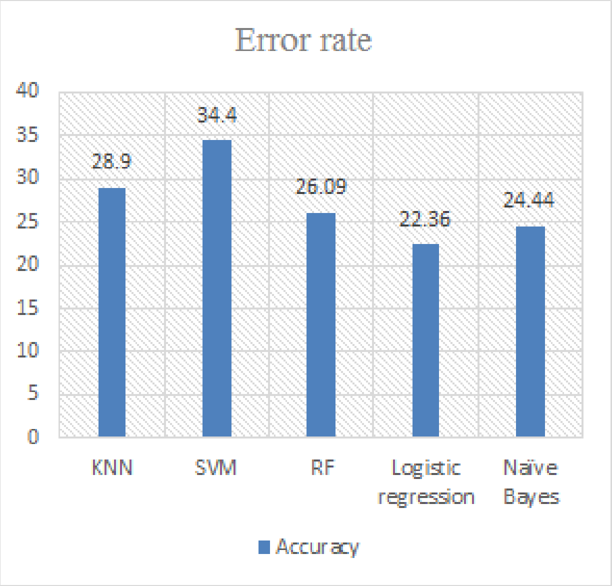

Figure 6: Performance analysis using error rate.

We can analyse from the figures that logistic regression performs better as compares to other models.

5. DISCUSSION AND CONCLUSION

During this research, the researcher described a technique which is used to deal with missing information in a machine learning model for diabetes categorization. This study's major goal was to increase the final result's accuracy while overcoming the drawbacks of traditional missing data imputation techniques.

In this research, using the PIMA Indian datasets, we provided a solid machine learning framework to enhance the performance of diabetes prediction. K-NN, SVM, logistic regression, Random Forest, and Naïve Bayes are the five methods are compared on diabetic dataset and two parameters are used to analyse the results which is accuracy and error rate.

A KNN is used in the proposed method for imputation. The first benefit of using the K-NN is that missing values are imputed more accurately. After imputation, we have used different classifiers which are K-NN, SVM, logistic regression, Random Forest, and Naïve Bayes for evaluating accuracies, where logistic regression is performed well as compared to others. In terms of error rate, SVM has maximum error rate which is 34.4% as compared to other methods. We intend to collect data on the history of diabetes in families and further hone the suggested strategy in the future.

REFERENCES

1. Kang H. The prevention and handling of the missing data. Korean J Anesthesiol. 2013;64(5):402.
2. Ludbrook J. Outlying observations and missing values: how should they be handled? Clin Exp Pharmacol Physiol. 2008;35(5–6):670–8.
3. Graham JW. Missing data analysis: making it work in the real world. Annu Rev Psychol. 2009; 60:549–76.
4. Choudhury A, Kosorok MR. Missing data imputation for classification problems. arXiv preprint arXiv:2002.10709. 2020.
5. Qiu J, Wu Q, Ding G, Xu Y, Feng S. A survey of machine learning for big data processing. EURASIP J Adv Signal Process. 2016;2016(1):1–16.
6. Rubin DB. Inference and missing data. Biometrika. 1976;63(3):581–92.
7. Yoon, J., Jordon, J., & van der Schaar, M. (2018). Missing data imputation using generative adversarial nets
8. Australian Bureau of Statistics. (2019). The Business Longitudinal Analysis Data Environment (BLADE).
9. Bakar, K., & Jin, H. (2019). A real prediction of survey data using Bayesian spatial gen-eralised linear models. Communications in Statistics-Simulation and Computation, 1-16.
10. Makaba, T., & Dogo, E. (2019, November). A comparison of strategies for missing values in data on machine learning classification algorithms. In *2019 International Multidisciplinary Information Technology and Engineering Conference (IMITEC)* (pp. 1-7). IEEE.
11. Mahesh, B. (2020). Machine learning algorithms-a review. *International Journal of Science and Research (IJSR).[Internet]*, *9*(1), 381-386.
12. Mohammed, M., Khan, M. B., & Bashier, E. B. M. (2016). *Machine learning: algorithms and applications*. Crc Press.
13. Malarvizhi, R., & Thanamani, A. S. (2012). K-nearest neighbor in missing data imputation. *Int. J. Eng. Res. Dev*, *5*(1), 5-7.
14. Hodge, V., & Austin, J. (2004). A survey of outlier detection methodologies. *Artificial intelligence review*, *22*, 85-126.