Journal of Industrial and Systems Engineering
Vol. 16, No. 3, pp. 30-62
Summer (July) 2024

JISE

Investigating the impact of missing value imputation methods on the prediction of diabetes using machine learning

Hooman Pourrostami[1], Seyed Amirreza Alavi[1], Ahar Hosseeini[2], Mobina Mousapour Mamoudan[1], Fariborz Jolai [1]*, Amir Aghsami[3]

[1] *School of Industrial Engineering, College of Engineering, University of Tehran, Tehran, Iran*

[2] *Center national health insurance, Tehran, Iran*

[3] *School of Industrial Engineering, K. N. Toosi University of Technology (KNTU), Tehran, Iran*

Abstract

Diabetes poses significant challenges due to its prevalence and the potential consequences of inaccurate or delayed diagnosis. This study focuses on enhancing prediction reliability to mitigate such risks. Initially, it identifies diabetes-related factors through correlation analysis with the target variable and implements models to address missing data. Subsequently, various imputation methods including CART, GMM, and RFR are employed to evaluate these factors. Results from each imputation scenario inform the selection of the most effective method. The study then employs ensemble algorithms like AdaBoost, Bagging, Gradient Boosting, and RF to enhance classification model accuracy. Further refinement is achieved by optimizing hyper-parameters through grid search. Evaluation involves comparing model predictions with those of medical professionals to assess accuracy. The findings reveal superior performance of optimized machine learning models over human predictions, indicating potential for improved diagnosis accuracy and reduced medical errors. This research contributes to advancing predictive modeling in diabetes diagnosis, offering prospects for enhanced community health and reduced socioeconomic burdens.

Keywords: Diabetes, Prediction, Machine learning, Ensemble learning, Gaussian Mixture Models, Imputation methods

1. Introduction

According to the World Health Organization (WHO), the number of people with diabetes increased by 31 million between 1980 and 2014, and, also, in 2014, 8.5% of adults had diabetes. In 2019, diabetes was the direct cause of death for 1.5 million people, and 8% of people with

* Corresponding Author

diabetes died before the age of 70. In most cases, this disease occurs when glucose/blood sugar levels in the human body are very high. However, several factors usually increase a person's diabetes, such as poor diet, age, ethnicity, obesity, a sedentary lifestyle, and a family history of diabetes. Gender, Body Mass Index (BMI), pregnancy, and metabolic status are other factors that contribute to diabetes. It can be said that diabetes is a multifactorial disease, and doctors are required to compare the test results of the patient with the test results of healthy people before making a diagnosis (Khanam & Foo, 2021). Such comparisons can make the diagnosis very misleading. In China, one of the most populous countries in the world, for example, 65.2 million people with diabetes have not been diagnosed (Li et al., 2021; Nozari, 2024).

On the other hand, not diagnosing diabetes on time affects the quality of life of people and causes many problems for the patient. This disease has many side effects and can damage the body’s organs. The most important of these effects are kidney, vision, cardiovascular, and nerve problems and the increase in treatment costs. If we can predict this disease at the right time, its impact on one's life will be less. As a result, its costs are reduced, and communities and health systems are improved.

With increasing advances in technology in the field of medicine, it is possible to diagnose diseases, including diabetes, quickly and accurately. Many diseases can be predicted using machine learning algorithms and data mining techniques. Data mining is a practical technique to find important information from large volumes of data (Mamoudan et al., 2021; Nozari, 2023). Machine learning algorithms are a data science technique used for prediction in various fields, including healthcare (Behdinian et al., 2022). Machine learning is a branch of artificial intelligence that enables systems to learn and improve automatically. Its main purpose is to develop computer programs that can access data and use it to learn.

Another important goal of the machine learning algorithm is to reduce errors and time and increase prediction accuracy. When the patient is referred to the doctor by having laboratory records, diagnosing the disease from the records is sometimes erroneous and causes serious problems. To overcome this problem, machine learning algorithms can be used to reduce the error rate and improve doctors' performance. The performance of these algorithms, especially in classification, is such that by finding the relationship between the features of the existing dataset and ranking these features, discover patterns that the data in these patterns are accurately identified (Werner de Vargas et al., 2022; Nozari et al., 2022).

When the practical works on the data is discussed, one of the most critical topics is the feature selection operation that extracts the required features from the data. Feature selection can be defined as the process of identifying relevant features and removing irrelevant and repetitive features with the aim of observing a subset of features that describe the problem well and with minimal loss of efficiency. Improving the efficiency of machine learning algorithms is one of the most important advantages of using feature selection.

Deep learning and machine learning algorithms have been able to make great improvements in predictions. Also, by using ensemble models, the accuracy of prediction models can be optimized

and increased. By correctly setting the parameters of ensemble algorithms, repeating the algorithm with a wide range of different parameters is avoided. Finally, the desired algorithm reaches the best performance in the shortest time. Considering that the accurate prediction of diabetes can lead to an increase in the health of society, many studies have been conducted in this direction. However, some questions have not yet been fully answered and this study tries to answer them:

- Which of the machine learning models can increase the accuracy in prediction?
- How effective is the use of missing value filling methods in increasing the accuracy of the model?
- Is the performance and accuracy of machine learning algorithms more than doctors?

Although accurate prediction of diabetic patients, directly and indirectly, affects people's quality of life and society's health, no study has provided the most accurate algorithm for predicting diabetes. In general, this article contributes to the literature in several areas. In the first part, using the data we obtained from Kaggle, in addition to examining the features effective on diabetes, we used various imputation missing value algorithms such as statistical methods, Cart, GMM and RFR algorithms and dimensionality reduction methods to select the most effective features on diabetes. In the second part, the features assigned in part one was used in machine learning algorithms to predict the diabetes. In order to identify the best model for predicting diabetes, in the third part, all the classification algorithms obtained from different feature imputations are optimized using ensemble learning model. By comparing the results of these algorithms, we seek to identify the best algorithm for predicting diabetes and improve its performance. To evaluate the proposed algorithms, the accuracy of each of these models is measured with validation tools. Finally, the results will be compared with the doctor’s opinion.

With all these interpretations, research to predict diabetic patients should be done in such a way that the accuracy and validity of the proposed models increase. Therefore, to achieve these goals, we asked a doctor to use the data to determine whether a person has diabetes or not. Then, the doctor's accuracy in predicting diabetic patients is calculated so that they can be used to compare the accuracy of prediction algorithms. Doing this can allow doctors to make accurate predictions to identify diabetic patients in a very short time.

This study makes several notable contributions to the field of diabetes diagnosis and predictive modeling. Firstly, it provides a comprehensive examination and comparison of three distinct imputation methods—Cart, Gaussian Mixture Models, and Random Forest Regressor—for filling in null values, thereby offering insights into the effectiveness of each approach in improving prediction accuracy. Secondly, by employing a diverse array of deep learning and machine learning algorithms, the study enhances the understanding of the predictive capabilities of these models in identifying diabetic patients. Furthermore, the optimization and selection of the most accurate algorithm for predicting diabetes through the utilization of various classification and ensemble algorithms represent a significant advancement in refining predictive modeling techniques for disease diagnosis. Ultimately, by meticulously analyzing the accuracy of the generated models and comparing them with medical professionals' opinions, this study provides

valuable insights into the potential utility of machine learning algorithms in improving diabetes diagnosis accuracy and reducing diagnostic errors.

Generally, in the second part of this article, the literature on the prediction of diabetes patients is discussed. The third part is devoted to methodology. In the fourth part, we analyze the results. In the fifth part, the discussion is presented. In the end, in the sixth part, there is a conclusion.

2. Literature review

One of the most significant health problems in today's advanced and developing societies is the prevalence and prevalence of diabetes. Not only is the economic loss to the sick person great, but the psychological damage to the patient and family is also negligible. Previous studies have shown that almost every aspect of a patient's life can be affected by diabetes and reduced patient satisfaction and quality of life. There are a large number of tools designed to measure a patient's quality of life and identify the most important aspects of life affected by diabetes. Identifying the most important aspects provides the information needed to plan for improving the patient's quality of life. For this reason, many researchers from different societies have predicted diabetes.

Park and Edington (2001) used Sequential Multi-Layer back-diffusion Perceptron (SMLP) to predict diabetes. They found that the results outperformed the regression models. Heikes, Eddy, Arondekar, and Schlessinger (2008) studied a predictor of diabetes risk in the United States from undiagnosed and prediabetes data. Karatsiolis and Schizas (2012) used the SVM algorithm and both the Radial Basis Function (RBF) and Polynomial approaches. These approaches were used to separate the division boundaries between the two groups of patients, potentially increasing the model models accuracy from 81% to 82.2%. Kumari and Chitra (2013) Using the SVM technique on Indian diabetic patients, achieved an accuracy of 65% for the education data and an accuracy of 85% for the model test data. In addition, Sudharsan, Peeples, and Shomali (2014) Used a machine learning model to predict hypoglycemic events occurring within 24 hours in patients with diabetes, yielding a sensitivity of 92% and specificity of the mark is 70%. In the same year Vijayan and Ravikumar (2014), used KNN, K-means, Amalgam KNN, and Adaptive Neuro-Fuzzy Inference System (ANFIS) algorithms on Indian diabetics. They found that the combination of the KNN and K-means performed better than each of these algorithms. Furthermore, these algorithms cover more than 80 precisions.

Santhanam and Padmavathi (2015) Used the k-means method for preprocessing. The SVM method uses genetic algorithms to perform the model and improve the model's accuracy, which increases the model's accuracy from 96% to 98%. Heydari, Teimouri, Heshmati, and Alavinia (2016) Used SVM, Artificial Neural Networks (ANN), DT, KNN, and Bayesian Network (BN) methods to predict type 2 diabetes in Tabriz patients. The ANN algorithm with an accuracy of 0.97 is the best model among all the models used. In the same year, Barale & Shrike, using the KNN method, first corrected the incomplete dataset of Indian diabetic patients and then combined the NN and LR algorithms to obtain a model with 0.99 accuracies. Mansourypoor and Asadi (2017) presented a model that can be considered a suitable alternative for diabetes diagnosis using the reinforcement learning-based evolutionary fuzzy rule-based system (RLEFRBS).

Zou et al. (2018) Used DT, Random Forest (RF), and Nearest Neighbor (NN) methods to predict diabetes mellitus. The dataset includes physical examination data from a hospital in Luzhou, China. It contains 14 properties. They used the Five-fold cross-validation method to validate the models. Principal Components Analysis (PCA) and Minimum Relevance Maximum Relevance (MRMR) methods were used for dimensionality reduction modeling, and RF modeling with 0.808 accuracies gave the best results. Nguyen et al. (2019) Predicted the onset of diabetes using deep learning algorithms, suggesting that sophisticated methods can improve the model's performance. The severe gradient elevation model (XGBoost) was developed to predict diabetes risk in 368 middle-aged and elderly people and ultimately to map the specific probabilities of each potential patient. This study indicates that even though the amount of data used for training is low, the XGBoost model still shows stable and excellent performance compared to traditional machine learning models (Wang, Wang, Chen, Jin, & Che, 2020).

Wang et al. (2020) defined a multi-label classification problem by considering three labels: macrovascular, microvascular, and neuropathy. Due to the imbalance of the classification, a small spherical multi-label and large margin machine (WML-SSLM) has been developed to diagnose the complications of diabetes. In the same year, Haq et al. (2020) developed an intelligent decision-making system based on machine learning algorithms to successfully diagnose diabetes and early treatment. DT machine learning classifier is used to classify the dimensions of the model. A filter-based DT algorithm (ID3) is proposed to select suitable features. The ID3 performance is better than other feature selection techniques, such as the DT Ada Boost group, random forest, and coverage-based feature selection method. DM Type 2 diagnostic datasets were collected from Mortala Nigeria Hospital. These datasets were used to develop a supervised ML model based on logistic regression, SVM, KNN, gradient setup algorithms, and a random forest (Muhammad, Algehyne, & Usman, 2020).

Also, in 2021, the article used the Pima Indian Patient Database to predict whether a person is at risk for diabetes based on specific diagnostic factors (Joshi & Dhakal, 2021). LR compares several predictive models to predict diabetes. This paper uses various selection criteria that are common in machine learning algorithms, such as Akaike's Information Criterion (AIC), Bayesian Information Criterion (BIC), Cp Mallows, Adjusted R^2, and forward and backward selection to identify significant predictors. A classification tree, which is a widely used machine learning technique with remarkable classification power, has also been used to predict the prevalence of diabetes (Fazakis et al., 2021). In another paper, a decision support system with an improved SVM radial discriminant underlying technique for disease prediction was deployed and compared with different machine learning techniques (Harimoorthy & Thangavelu, 2021). Deberneh and Kim (2021) Compared the performance of LR, SVM, RF, and XGBoost algorithms. They also used set techniques such as the Confusion Matrix-based Classification (CIM) integration approach, soft voting, and classification methods and compared their performance with the primary models.

Ghosh et al. (2021) Used SVM, AdaBoost (AB), RF, and Gradient Boosting (GB) to evaluate Indian diabetic patients and then extracted the important features using the MRMR method. RF model with 99.35% accuracy gives the best results. Diagnosis is a branch of Artificial Intelligence (AI) that focuses on the development of algorithms. These techniques can determine the correct

behavior of a system. Khanam and Foo (2021) have applied many machine learning methods to Pima Indian Diabetes (PID) patients, including RF, Naïve Bayes (NB), SVM, KNN, NN, DT, and LR, of which NN has the highest accuracy of 88.6%. Many algorithms have recently been developed to improve machine learning models. One of these algorithms is the Grasshopper Optimization Algorithm (GOA). This technique has been used in 2021 to improve feature selection in Indian patients and to use the support vector machine method for modeling and the 10-fold validation method for model validation with 97% accuracy (Kamel & Yaghoubzadeh, 2021).

Therefore, in 2022, many researchers and physicians have promised to diagnose Machine Learning-based Disease Diagnosis (MLBDD) as cheap and effective. Healthcare data such as X-rays and MRIs, along with tabular data including patients' condition, age, and gender are utilized in the creation of MLBDD systems, which is highly suitable. (Manjurul Ahsan & Siddique, 2021).Lu, Uddin, Hajati, Moni, and Khushi (2022) patient used patients' socio-behavioral characteristics to predict diabetes. They implemented several machine learning models, including LR, KNN, RF, SVM, and ANN, to predict diabetes through behavioral characteristics. These algorithms' Area under Curves (AUC) is between 0.79 and 0.91. Among these models, the RF model has shown the best accuracy. Khademi et al. (2022) studied the effect of environmental conditions in diabetes according to medical data mining technique. Fereidouni et al. (2022) used Decision Trees, Naive Bayes, and Rule Induction models to predict death rate fluctuations concerning various foods in Covid-19 outbreak. Qi, Song, Liu, Zhang, and Wong (2023) used KF_NN models to feature selection. The accuracy of the models are enhanced by Kfpredict algorithms.Doğru, Buyrukoğlu, and Arı (2023) implemented super learner model for cross-validation. Three different datasets have been used for model execution. The important feature are also selected by Chi-square. Abdali et al. (2024) compared several machine learning algorithms to classify patients with mental conditions. to categorize them into different classes according to their disorder types.

Kee et al. (2023) examined ML-based prediction models for CVD risk in T2DM patients. Neural networks show promise, but adherence to PROBAST and TRIPOD standards is crucial for reducing bias and improving clinical applicability. Another study delves into diabetic prediction, emphasizing early classification for effective management. It highlights deep learning's pivotal role in automated retinopathy characterization, offering insights into binary classification and severity assessment, and discusses future directions and challenges in the field (Pan et al., 2023). Other study underscores the significance of diabetes prediction for enhancing treatment outcomes. It proposes a model employing data mining techniques like Random Forest, SVM, Logistic Regression, and Naive Bayes, with logistic regression achieving the highest accuracy at 82.46% (Rastogi and Bansal, 2023).

This study distinguishes itself through a comprehensive exploration of three distinct imputation methods—Cart, Gaussian Mixture Models, and Random Forest Regressor—for managing missing data, providing valuable insights into their effectiveness in improving prediction accuracy in diabetes diagnosis. Additionally, it employs a diverse array of deep learning and machine learning algorithms to predict diabetic patients, enhancing understanding of their predictive capabilities. Furthermore, the study optimizes predictive models using various classification and ensemble

algorithms, aiming to identify the most accurate algorithm for diabetes prediction. Notably, it evaluates model performance against expert opinions, offering practical insights into the potential utility of machine learning algorithms in clinical settings, thus contributing significantly to refining predictive modeling techniques and improving diabetes diagnosis accuracy.

As it is known, a lot of research has been done in the field of predicting diabetes. However, none of these studies have chosen the best imputation algorithm for predicting diabetic patients. However, some researches have focused on the optimization of classification algorithms using ensemble algorithms, but the role of feature missing values completion has not been seen in the articles. In addition, none of the previous articles have compared the performance of these algorithms with the doctor's diagnosis. In this article, in addition to predicting diabetic patients and optimizing them using imputation methods, we also consider the role of different ensemble learning algorithms.

3. Methodology

This study encompasses four primary objectives. Firstly, it focuses on imputing null values utilizing three distinct methods: Cart, Gaussian Mixture Models, and Random Forest Regressor. These methods are integrated into prediction models and subsequently compared for their efficacy. Secondly, the study employs a range of deep learning and machine learning algorithms to predict diabetic patients. Thirdly, various classification and ensemble algorithms are utilized to optimize predictive models and identify the most accurate algorithm for predicting diabetes patients. Finally, the accuracy of the results generated by these algorithms is meticulously analyzed.

The procedural steps of this study are delineated in Figure 1. Initially, different imputation methods are applied to fill in missing data values. Subsequently, classification models are developed using various imputation techniques. Following this, the accuracy of these classification models is compared utilizing tuned ensemble algorithms. Finally, the performance of the developed models is juxtaposed with the opinions of medical professionals to assess their efficacy in diagnosing diabetes.

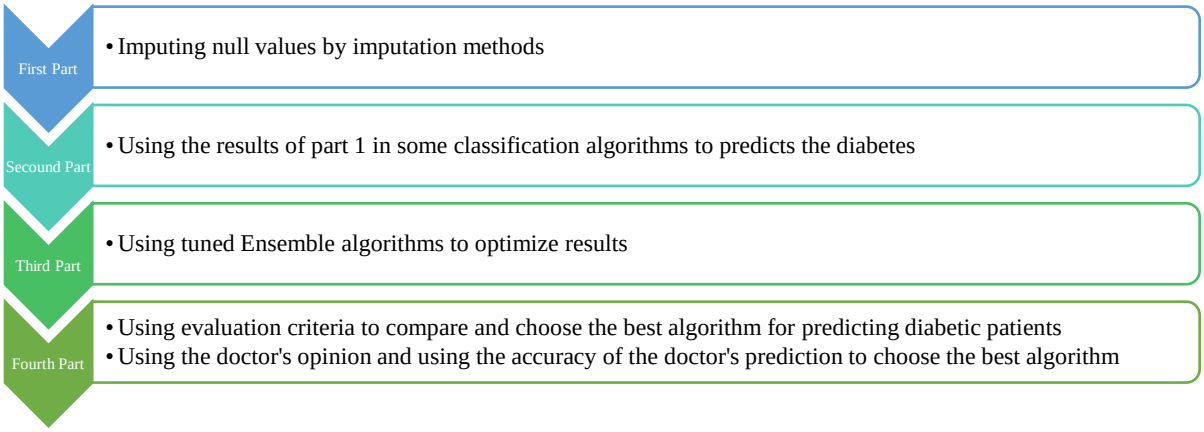

Figure 1:The Steps of This Study

Table 1 :Description of each feature

Features	Description
Pregnancies	Number of times pregnant
Glucose	2-Hour Glucose tolerance test
Blood Pressure	Diastolic blood pressure (mm Hg)
Skin Thickness	triceps skin fold thickness (mm)
Insulin	2-Hour serum insulin (mg/dL)
BMI	Body mass index (BMI) (Weight/Height^2, unit in kg/m2)
DiabetesPedigreeFunction	Diabetes pedigree function
Age	Age (years)
Outcome	Response Variable

Data Description and Data Preprocessing

Diabetes is a disease that can increase the risk of many diseases and can cause many irreparable costs for individuals and society. This article utilizes various factors including pregnancies, blood glucose levels, blood pressure readings, skin thickness, insulin levels, BMI, diabetes family history, and age to diagnose diabetes. The description of each feature is shown in Table 1. In data science projects, the preprocessing section is the most important part of modeling and therefore has a significant impact on the model's accuracy. The data preparation and preprocessing phases typically include data structuring and integration, Noise data management, identifying missing data, managing missing data, converting qualitative data to quantitative data, and normalization and standardization.

Any dataset has some understanding of noise data, depending on the data type and the rows entered. The primary dataset includes 768 instances, but three samples are Anomaly and should remove them. The reason for the abnormality of these three observations is the high levels of insulin in the blood as well as the Glucose, which shows a healthy person, so they should be removed. Three abnormal data are shown in Table 2. The existing dataset contains 765 diabetics and 8 features, along with binary response variables. Table 3 shows the description of this dataset.

Table 2 :Anomaly sample of Dataset

Instance	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age	Outcome
153	1	153	82	42	485	40.6	0.687	23	0
228	4	197	70	39	744	36.7	2.329	31	0
247	0	165	90	33	680	52.3	0.427	23	0

Table 3: Dataset Description

	count	mean	std	min	25%	50%	75%	max
Pregnancies	765	3.845752	3.346249	0	1	3	6	13.5
Glucose	765	121.4895	30.31143	44	99	117	140	199
Blood Pressure	765	72.40392	11.93921	40	64	72	80	122
Skin Thickness	538	29.00929	10.05913	7	22	29	36	57
Insulin	370	132.6108	74.28539	14	75	120	177.5	360
BMI	765	32.35444	6.640657	18.2	27.5	32.35444	36.5	50.25
DiabetesPedigreeFunction	765	0.460799	0.293725	0.078	0.243	0.37	0.624	1.476
Age	765	33.27059	11.77133	21	24	29	41	81
Outcome	765	0.350327	0.477384	0	0	0	1	1

Noise data is specified depending on the characteristics and values of the data. The number of pregnancies in individuals can start from zero and continue to its permissible level, i.e., zero in this column will not be meaningless. But Glucose cannot be zero, and usually, if it is higher than 126, the person is more likely to develop diabetes. The same is true for blood pressure. If the value of blood pressure is zero, it is noise data should be checked. For skin thickness, zero data is meaningless and should be replaced with logical data. One of the most effective features in the dataset is the amount of insulin in the blood. The importance of filling insulin values becomes more important when there is a significant relationship between the response variable and this feature in the correlation matrix. The correlation matrix was shown in Figure 2. Therefore, it is very important to fill in the missing values of this feature. The number of missing data recorded in each column is shown in Figure 3. Different solutions can be used to replace this data depending on the distribution of each feature and the amount of noise data.

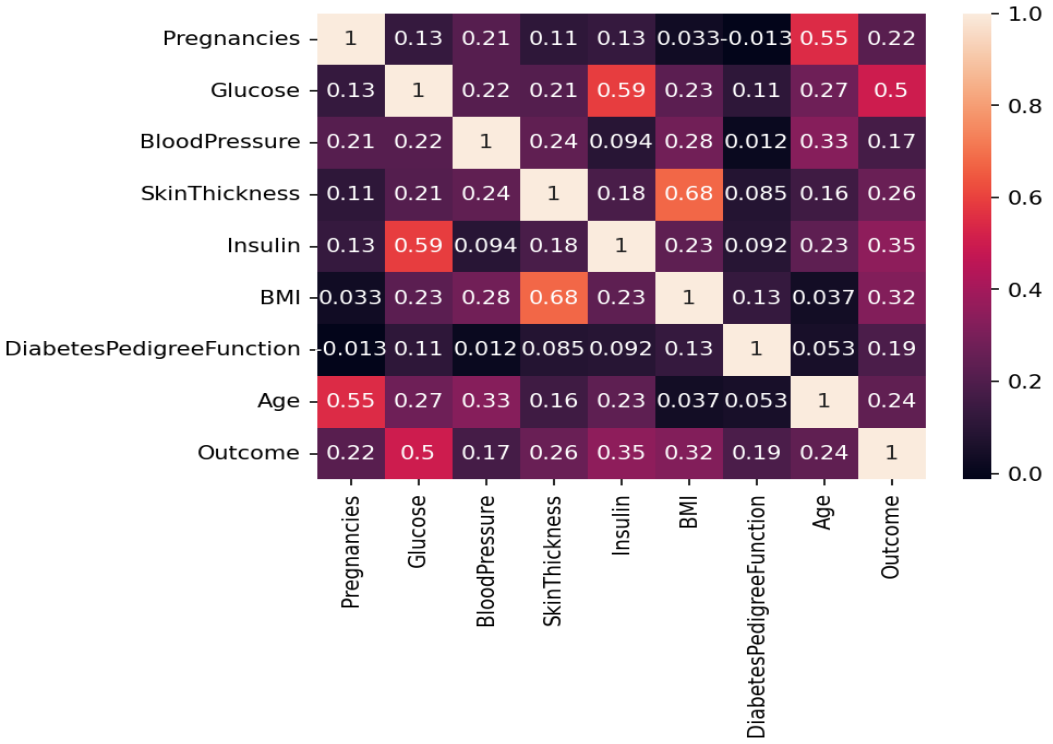

Figure2: Correlation Matrix

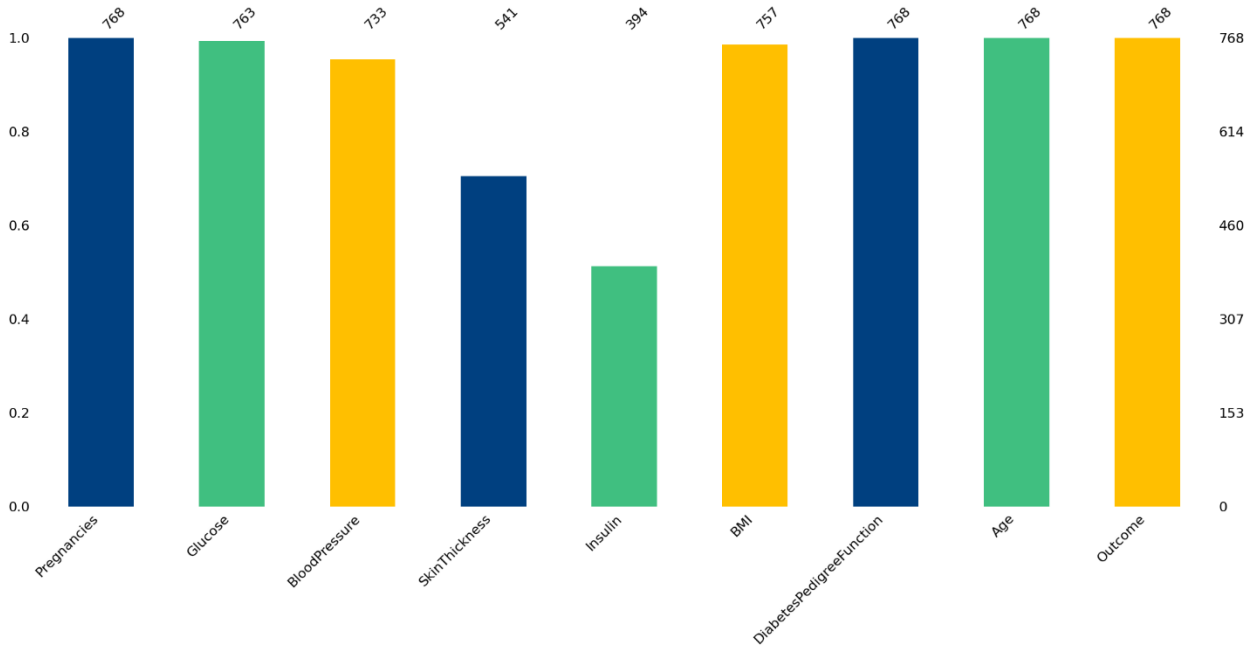

Figure3: Missing values in each feature

3.1 Imputation Algorithms

3.1.1 Decision Tree Regressor (DTR)

Cart imputation, which is also referred to as Classification and Regression Tree (CART) imputation, is a technique utilized to complete missing values in a dataset. It involves creating a decision tree algorithm using the present data points in the dataset, and subsequently employing the model to estimate the missing data points. In fact, this model is used to fill in missing values of insulin and skinthickness. These variables are numerical, so a regression tree was used to predict them. The training set is used to construct the decision tree model, while the test set is used to evaluate the performance of the model. The training data includes the specified values of insulin and skin thickness, while the test data includes the missing values of these two features. According to the number of features of the data set and the type of problem, the hyperparameters of the decision tree model, including the maximum depth of the tree, the maximum feature, and the minimum number of samples for splitting and branching conditions, will be selected. The hyperparameters of the decision tree model are represented in Table 5.

Poisson distribution is used for decision tree segmentation conditions. The Poisson criterion in decision tree regression uses the Poisson deviance to determine the best split in the tree, choosing the split that results in the greatest reduction in the deviance. Intuitively, the Poisson deviance measures how well the model fits the data by comparing the actual counts to the counts predicted by the model. A lower deviance indicates a better fit, with a perfect fit having a deviance of zero. Equation 1 shows whether the predicted values are well fitted on the Poisson distribution; in other words, the smaller the deviation, the more accurate our prediction will be. In this equation, n represents the number of observations, $y(i)$ is the observed count for the i-th observation, and $exp\{X_i\hat{\beta}\}$ is the predicted count for the i-th statement based on the model (Aryai & Goldsworthy, 2023).

$$D = 2 \times \sum_{i=1}^{n} \left[y(i) * \log\left(\frac{y(i)}{exp\{X_i\hat{\beta}\}}\right) - \left(y(i) - exp\{X_i\hat{\beta}\}\right) \right] \quad (1)$$

3.1.2 Gaussian Mixture Models (GMM)

Another method that has been used to fill the missing values of insulin and skin thickness features is Gaussian Mixture Models (GMM). Gaussian mixture models are statistical models that assume that a collection of data points is composed of a combination of Gaussian distributions, each with its own set of unknown parameters. A mixture model can be viewed as a generalization of the k-means clustering procedure to incorporate latent Gaussian centers and covariance structures. The basic idea behind a GMM is to estimate the parameters of the underlying Gaussian distributions that generate the observed data. These parameters include each Gaussian component's mean, covariance, and weights. In practice, researchers often use GMMs for clustering tasks, where they aim to partition the data into distinct groups based on their similarity.

To perform clustering with a GMM, the model parameters are typically estimated using an iterative algorithm such as the Expectation-Maximization (EM) algorithm. The algorithm works by first assigning each data point to a cluster based on the likelihood of it being generated by each Gaussian distribution. Then, the model parameters are updated based on these assignments, and the process is repeated until convergence. One advantage of GMMs is that they can capture complex shapes of data distributions, unlike other clustering algorithms like k-means which assume spherical clusters. They also provide a probabilistic framework for clustering, allowing for uncertainty in the assignment of data points to clusters. The hyperparameters of this model include the number of clusters and covariance type (Segar et al., 2021) . In this model, the covariance type is assigning to full because it can be more flexible in data clustering. The hyperparameters of the GMM are shown in Table 5.

3.1.3 Random Forest Regressor (RFR)

The last method that has been used for missing value imputation is the Random Forest Regressor (RFR). It is based on the random forest algorithm, which is a decision tree-based ensemble method. The concept behind using a random forest for imputation is to build multiple decision trees on different subsets of the data and then combine their predictions to make a final prediction. Each tree in the forest is built by randomly selecting a subset of the features and a subset of the samples from the original dataset. To impute missing values using a random forest regressor, we first split our data into two sets: the set with missing values and the set without missing values. We use the set without missing values to train our random forest model, where the target variable is the feature that has missing values (Özdemir et al., 2023).The hyperparameters of the decision tree model are described in Table 5.

Using a random forest regressor for imputation delivers several benefits. It can effectively handle categorical and continuous variables while also detecting non-linear relationships between features. Additionally, the random forest algorithm is robust to outliers and noisy data.For imputation missing values features, first the value of the skin thickness variable is predicted using other available variables, and then the results obtained from this prediction are used for prediction the insulin variable. In Equation 2, the dependent variables are written as a function of the independent variables (Pourkhodabakhsh, Mamoudan, & Bozorgi-Amiri, 2023).

$$y_1 = f(Glucose, Pregnancies, Blood\ pressure, Age, BMI, DPF) + \in$$
$$y_2 = f(Glucose, Pregnancies, Blood\ pressure, Age, BMI, DPF, Skinthickness) + \in \quad (2)$$

Table 5 :Hyper parameters for Imputation Methods

	Imputation Methods		
	Cart	Gaussian mixture	Random Forest Regressor
Hyper parameters Tuning	random_state=30 criterion='poisson' min_samples_split=45	n_components=7 random_state=30 covariance_type='full'	n_estimators=135 random_state=30 min_samples_split=40

	max_depth=6 max_features='auto'		max_depth=7 max_features='auto'

3.2 Prediction Models

3.2.1 Support Vector Machine

The Support Vector Machine (SVM) is one of the machine learning algorithms. This algorithm can be used for both regression and classification problems. This algorithm can be a great option when you need to predict very accurately. On the other hand, the complex formulations that exist for this model make it very difficult to visualize these algorithms. In the SVM algorithm, each data sample is plotted as a point in the n-dimensional space on the data scatter diagram (n) is the number of properties that a data sample has) and the value of each data attribute is one of the coordinate points of the point on the graph. The SVM utilizes a straight line to classify distinct and diverse data, assuming that the classes can be separated by a linear boundary.

In the training phase of this algorithm, the decision boundary is so that the "minimum" distance, boundary, decision-making, or separator has been selected to have a more reliable margin of the model. When y shows the list of targeted applications or classes, the SVM algorithm will build a hyperplane H that is shown in Equation 3. In Equation 3, x shows the input from netflow data. a is the vector of the weights for each flow feature, b is the bias term, and $\emptyset(0)$ is the fixed feature-space transformation (Tsai, Hu, & Systems, 2022) .

$$H: a^T \emptyset\,(x) + b = 0 \tag{3}$$

The problem of the SVM algorithm is to find the hyperplanes H that separate different classes.

3.2.2 Decision Tree

The Decision Tree (DT) is a map of the possible outcomes of a series of related choices or options that allow an individual or organization to weigh possible actions in terms of costs, opportunities, and benefits. The DT can either advance personal and informal goals and plans or draw an algorithm that predicts the best option based on mathematics. The DT includes some nodes and branches where it classifies the samples so that it grows from the root downwards and finally gaines the leaf nodes (Fereidouni et al., 2022).

A DT typically starts with an initial node, after which potential outcomes branch off. Each of those outcomes leads to other nodes, which creates branches of other probabilities. This branching structure eventually becomes a tree-like diagram. Also, the number of nodes, leaves, characteristics to take into consideration, and the depth of roots determine the size of each DT, resulting in a more understandable DT due to its smaller volume. This is done by pruning. Equation 4 shows how entropies are related. If the sample is completely homogeneous, the entropy is zero, and if it is the same, the entropy is one. In other words, the most entropy is achieved when all categories have equal probabilities of belonging. Each DT must be constructed using a frequency

table for one property and a frequency table for two properties to calculate entropy. Gini Index is also shown in Equation 5. By measuring the degree or likelihood of misclassification of a particular variable by random selection, the Gini Index can be used as a measure of distribution (Haq et al., 2020).

$$Entropy(p_1, p_2, \dots, p_k) = -\sum p_1 \log_2(p_1) \tag{4}$$

$$I_G = 1 - \sum_{J=1}^{C} P_J^2 \tag{5}$$

3.2.3 Naive Bayes

Naive Bayes (NB) is the simplest machine learning algorithm we can apply to our data. As the name implies, this algorithm assumes that all variables in the data set are naïve, that is, they are not related to each other.

The Bayesian classification technique is often used as a simple way to classify and determine a way to tag objects or points. There is no single algorithm for using the simple Bayesian classifier; instead, there is a family of algorithms that operate on the assumption that attributes or variables are independent of each other. For example, the size of the fruit and the color of the variables, which are considered independent variables, can be effective in determining its type. So, if the fruit is red and about 10 cm in size, it is most likely an apple (Blanquero et al., 2021).

Most simple Bayesian techniques and models use the likelihood function maximization method. Although the simple Bayesian classification technique has limited and accessible assumptions, it can well solve real-world problems and be considered a competitor to the "random forest" methods. One of the significant advantages of the simple Bayesian category is the ability to estimate model parameters with a small sample size as a "Training Data" set (Srivastava, Singh, & Singh, 2022).

3.2.4 *k*-nearest neighbor algorithm

k-nearest neighbor algorithm (KNN) is a non-parametric statistical method used for statistical classification and regression. In both cases, KNN contains the closest instructional example in the data space, and its output varies depending on the type used in the classification and regression. In the classification mode, according to the value specified for KNN, it calculates the distance of the point that we want to label with the nearest points. This algorithm decides on the label of the desired point according to the maximum vote of neighboring points. Various methods can be used to calculate this distance, one of the most important of which is the Euclidean distance. The average of the values obtained in the regression model is its output. Since the calculations of this algorithm are based on distance, data normalization can help to improve its performance (Wu et al., 2018).

KNN stands for K-Nearest Neighbor, and the most common use of this algorithm is in machine learning as well as data mining. This non-parametric algorithm makes it simpler and more useful than other algorithms.

3.2.5 Ensemble Learning

Ensemble machine models are actually a collection of machine learning models in which weak models or basic models are combined with each other to obtain better results. When weak models are combined with each other correctly, more accurate and stable models are obtained. In most cases, each of the models alone has high variance and bias, and to solve this problem, the models are combined with each other.

3.2.5.1 Adaptive Boosting

One of the most successful amplification algorithms for binary classification is Adaptive Boosting (AdaBoost). AdaBoost is used with small DTs, and the first tree is created, and its function on each training sample is used to measure the attention of the next tree to the samples. Therefore, the tree should pay attention to all training samples and give more weight to training data that is difficult to predict while giving less weight to data that is easy to predict. Each input training Dataset is presented and labeled as in Equation 6. In this equation, n is the training data set size, $x_i \in T$ and $y_i \in \{-1,1\}$. In this particular method, every individual example in the training data can be represented as a coordinate in a space that has multiple dimensions (Feng et al., 2020).

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n) \tag{6}$$

AdaBoost has the ability to create a probability distribution that covers all the training examples. This distribution can be altered through multiple iterations by introducing a new weak classifier to the data.$D_t(x_i)$ shows the probability distribution and the successive iterations are shown by T. In this algorithm h_t is the weak classifier chosen for the iteration t. $h_t(x_i)$ denoted the class label assigned by x_i. After comparing $h_t(x_i)$ with y_i for $i = 1,2, \dots, n$ we will have an error that we noted it ϵ_i. This error shows the classification error rate for the classifier h_t. Each classifier candidate is trained over iterations using a subsample of all the training data provided by the $D_t(x)$ probability distribution. The higher the probability of $D_t(x)$ for a given training sample x, the greater the chance of selecting it for the instruction of candidate $h(t)$ classifier. The h_t must be selected in a way to minimize the amount of misclassification rate ϵ. Z_t is a normalization factor that ensures that the sum of the updated probability distribution $D_{t+1}(x_i)$ over all samples equals 1.

Equation 7 shows the probability distribution over all the training samples of data at iteration $t + 1$. In each iteration, the model attempts to fit better distribution related to the last step. The core of the AdaBoost algorithm is the trust level α_t assigned to the chosen weak classifier. The bigger the value of error, ϵ_t for a classifier, the lower the trust must be. Equation 8 shows the relation between α_t and ϵ_t (Kumari & Chitra, 2013).

$$D_{t+1}(x_i) = \frac{D_t(x_i) \times exp\,(-\alpha_t \times y_i \times h_{t(x_i)})}{Z_t} \tag{7}$$

$$\alpha_t = \frac{1}{2} \ln \frac{1 - \epsilon_t}{\epsilon_t} \tag{8}$$

The trust level of the candidate classifier $h(t)$ will get a value in the scale of $-\infty$ and ∞, if the epsilon rate gets closer to 1 starting from 0. If the ϵ gets closer to 1, the weak classifier almost completely fails on the overall training Dataset. If the ϵ gets closer to 0, the weak classifier will be powerful as also stated. According to Al-Hadeethi, Abdulla, Diykh, Deo, and Green (2020), the classifier H will be obtained. After n iterations, H classifier in the AdaBoost will be evaluated as shown in Equation 9. Sign is a function that returns +1 if the argument is positive and -1 otherwise.

$$H(x) = sign\left[\sum_{t=1}^{n} \alpha(t) h_t(x)\right] \quad (9)$$

3.2.5.2 Bagging

A Bagging classifier is a type of ensemble learning algorithm that combines multiple base classifiers by training each one on different random subsets of the original dataset. The individual predictions of each classifier are then combined, either through voting or averaging, to generate the final prediction. The idea behind bagging is to reduce variance in the model by averaging out the effects of small variations in the training data. By training multiple models on different subsets of the data, we introduce diversity into the ensemble, which can improve the overall accuracy and robustness of the model. In this research, we implemented the DT algorithm as a base model because it has the highest accuracy than other models. In bagging, all models are run in parallel and the target variable is predicted by the majority vote of the algorithms.

3.3.6.3 Gradient_Boosting

Gradient Boosting Classifier is an algorithm that iteratively chooses a function that moves towards a weak hypothesis or negative gradient, with the goal of minimizing a loss function. By combining multiple weak learning models, it creates a strong predictive model that can handle non-linear data and interactions effectively. In Gradient Boosting, the training of subsequent trees depends on the errors made by the previous trees in the sequence. Each new tree is trained on the residual error of the previous tree's predictions. Gradient Boosting Classifier has a significant advantage over other ensemble methods in that it is adept at effectively dealing with non-linear data and interactions, thereby making it a highly effective tool for classification tasks. In this algorithm, between 50 and 150 decision tree models have been implemented according to different methods of filling the missing values. The learning rate has varied between 0.01 and 0.1. According to Bentéjac, Csörgő, and Martínez-Muñoz (2021), in equation 10, m is the number of implemented algorithms and gamma is the learning rate as the other (Sai et al., 2023).

$$F_M(x) = \sum_{m-1}^{M} \gamma_m h_m(x) \quad (10)$$

3.3.6.4 Random Forest

A random forest is a type of predictive model that uses multiple decision trees on smaller portions of the dataset and combines their results through averaging to enhance accuracy and prevent over-

fitting. To make a prediction with a random forest classifier, first, all the individual decision trees in the forest are run on the new input instances, and each tree provides a class prediction. The class that receives the most votes from the individual decision trees is then chosen as the final prediction of the random forest model. This algorithm also can perform well when the dataset contains many missing value data. Also, the feature selection from each estimator is random and can prevent over-fitting.

A subset of data is injected into each tree in the RF algorithm. For example, if your data set had 1000 rows (i.e., 1000 samples) and 50 columns (i.e., 50 properties) (read the Properties and Dimensions lesson), the RF algorithm for each tree is 100 rows and 20 columns, which Gives a randomly selected form and is a subset of the data set. With these subset datasets, these trees can make decisions and build their classification model (Hernández-Pereira et al., 2022).

Figure 4 shows the four main processes in predicting diabetes. These steps include pre-processing, data separation into training and testing contain validation, model implementation, and evaluation and comparison of implemented models.

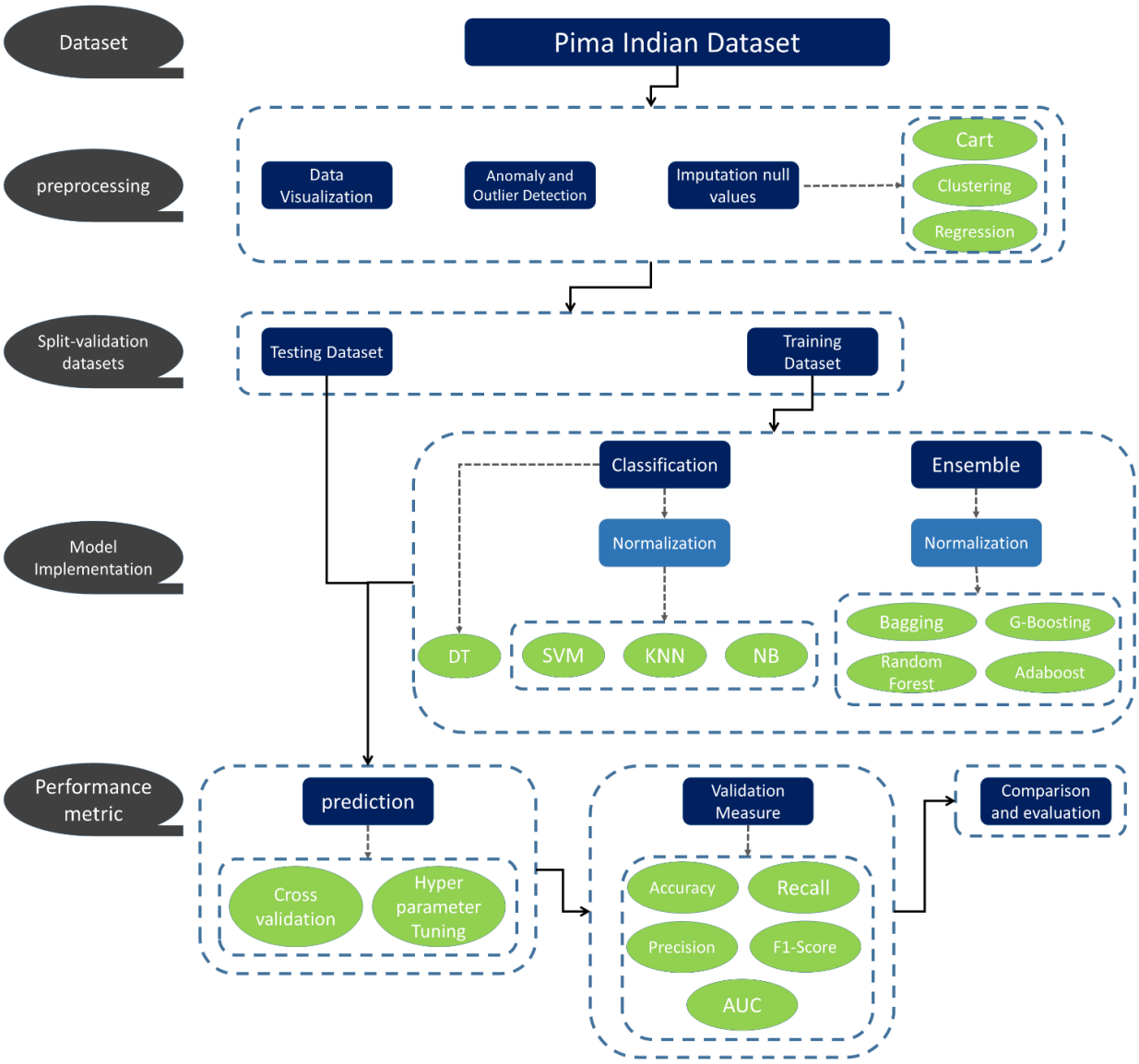

Figure4 :the flow diagram of the diabetes prediction approach

3.4 Experiments and Evaluation

The accuracy of classification algorithms depends on that algorithm's number of correct predictions. In other words, the ratio of correct predictions made by the algorithm to the total number of predictions is used to evaluate the algorithm's accuracy, which Equation 11 shows well (Mamoudan, Jafari, Mohammadnazari, Nasiri, & Yazdani, 2023). The higher the accuracy rate, the higher the number of correctly predicted cases. Accuracy is useful when false positives error is more worrisome than false negatives error. An inaccurate rating system may lead to customer churn and damage to the business, which is why it is so important (Jain, Quamer, & Pamula, 2021). Precision formula is shown in Equation 12. Recall, which is shown in Equation 13, shows how accurate the model was in predicting true positives. This is a useful measure whenever a false

negative is more worrisome than a false positive. Combined with the Recall metric, it creates a sense of Precision. When Precision equals Recall, the metric is at its maximum. The formula for measuring the F1 score is shown in Equation 14. The Area Under Curve (AUC) represented the area under ROC plot as shown in Equation 15.

$$Accuracy = \frac{T_p + T_n}{T_n + T_p + F_p + F_n} \tag{11}$$

$$Precision = \frac{T_p}{T_p + F_p} \tag{12}$$

$$Recall = \frac{T_p}{T_p + F_n} \tag{13}$$

$$F1\ score = \frac{2 \times precision \times recall}{precision + recall} \tag{14}$$

$$AUC = \int_0^1 Roc\ Curve(x)dx \tag{15}$$

3.4.1 Model implemented: Imputation Models

This research implements three assignment methods in the first step to replace missing values with appropriate data. Then the results of these algorithms are used in the prediction of diabetic patients. By using algorithms SVM, DT, RF, LR, NB, KNN and Ensemble learning models diabetic patients have been predicted. To check the accuracy and performance of these algorithms, an expert (doctor) has been requested to diagnose diabetic patients from their data. Finally, the results are analyzed. The results of the model using eight classification and ensemble algorithms by Cart imputation models are shown in Table 6. Also, the model's results using GMM and RFR models are shown in Tables 7 and 8, respectively.

Table 6: The results of 8 classification algorithms with Cart imputation models

Methods	Accuracy	Recall	Precision	F1-score	AUC
DT	82.35%	0.82	0.79	0.80	0.85
SVM	78..43%	0.79	0.73	0.75	0.80
KNN	77.12%	0.76	0.73	0.74	0.82
NB	77.78%	0.76	0.75	0.76	0.80
AdaBoost	80.39%	0.80	0.77	0.78	0.84
Bagging	81.7%	0.82	0.78	0.79	0.84
G_Boosting	83.45%	0.83	0.82	0.82	0.87
RF	81.7%	0.81	0.79	0.80	0.83

Table 7: The results of 8 classification algorithms with GMM imputation models

Methods	Accuracy	Recall	Precision	F1-score	AUC
DT	86.93%	0.86	0.86	0.86	0.91
SVM	80.39%	0.81	0.78	0.78	0.83
KNN	81.05%	0.81	0.78	0.79	0.82
NB	81.7%	0.80	0.78	0.79	0.87
AdaBoost	84.31%	0.83	0.83	0.83	0.92
Bagging	87.58%	0.87	0.86	0.87	0.90
G_Boosting	90.2%	0.89	0.89	0.89	0.95
RF	90.8%	0.89	0.89	0.89	0.95

Table 8 :The results of 8 classification algorithms with RFR imputation models

Methods	Accuracy	Recall	Precision	F1-score	AUC
DT	81.56%	0.82	0.80	0.80	0.83
SVM	79.23%	0.79	0.77	0.77	0.81
KNN	79.08%	0.79	0.75	0.76	0.81
NB	81.05%	0.80	0.78	0.79	0.87
AdaBoost	83.35%	0.83	0.81	0.82	0.91
Bagging	86.27%	0.87	0.84	0.85	0.94
G_Boosting	82.75%	0.82	0.80	0.81	0.84
RF	85.65%	0.85	0.83	0.83	0.93

3.4.2 Hyper Parameters Tuning by Grid search

Hyperparameter tuning is an essential stage in building machine-learning models. It involves fine-tuning the hyperparameters of a model to find the optimal set of values that lead to the best performance. Grid search is a popular method for hyperparameter tuning, especially when the number of hyperparameters is small. In grid search, we specify a range of values for each hyperparameter and then create a grid of all possible combinations of these values. First, we train and test the model using each combination of hyperparameters in the grid. After that, we choose the combination that performs the best. In all the models implemented in this research, the grid search method has optimized the hyperparameters. For example, suppose we have a SVM model with three hyperparameters include C, Kernel and gamma. We can specify a range of values for each hyperparameter, such as [0.1, 1, 10] for C, [rbf,linear] for kernel and [0.5, 1, 2] for gamma. This gives us a grid of eighteen possible combinations of hyperparameters. We then train and evaluate the SVM model for each combination of hyperparameters using cross-validation and select the hyperparameters that give us the best performance on the validation set. The optimal

hyperparameters of each algorithm using CART, GMM and RFR methods are shown in Table 9,10,11 respectively.

Table 9: The tuning parameters of 8 classification algorithms with Cart imputation models

Methods	Hyper Parameters By Grid search Tuning (Cart Model)
DT	{Rastogi, 2023, 'max_depth': 8, 'max_features': 'log2', 'min_samples_split': 40}
SVM	{'C': 1, 'gamma': 1.2, 'kernel': 'rbf'}
KNN	{'leaf_size': 5, 'n_neighbors': 21}
NB	{'var_smoothing': 0.1}
AdaBoost	{n_estimators=185, learning_rate=0.1}
Bagging	{max_depth=8,min_samples_split=40,criterion='gini',max_features= 'log2',ccp_alpha= 0.001 ,'n_estimators': 165}
G_Boosting	{'learning_rate': 0.01, 'max_depth': 6, 'max_features': 'log2', 'min_samples_split': 45, ccp_alpha= 0.01,'n_estimators': 155, 'subsample': 0.8}
RF	{'criterion': 'entropy', 'max_depth': 6, 'max_features': 'log2', 'min_samples_split': 45, 'n_estimators': 160}

Table 10 :The tuning parameters of 8 classification algorithms with GMM imputation models

Methods	Hyper Parameters By Grid search Tuning (GMM Model)
DT	{'ccp_alpha': 0.001, 'criterion': 'gini', 'max_depth': 7, 'max_features': 'log2', 'min_samples_split': 25}
SVM	{'C': 1, 'gamma': 0.8, 'kernel': 'linear'}
KNN	{'leaf_size': 5, 'n_neighbors': 21}
NB	{'var_smoothing': 0.1}
AdaBoost	{n_estimators=45, learning_rate=0.1}
Bagging	{max_depth=8,min_samples_split=40,criterion='gini',max_features= 'log2' ,ccp_alpha= 0.01,{'n_estimators': 20}
G_Boosting	{max_depth=7,criterion='gini',max_features= 'log2', ,min_samples_split=40 ccp_alpha= 0.001,'n_estimators': 40, 'subsample': 0.8}
RF	{'criterion': 'entropy', 'max_depth': 6, 'max_features': 'auto', 'min_samples_split': 30, 'n_estimators': 15}

Table 11: The tuning parameters of 8 classification algorithms with RFR imputation models

Methods	Hyper Parameters By Grid search Tuning (RFR Model)
DT	{'ccp_alpha': 0.01, 'criterion': 'entropy', 'max_depth': 8, 'max_features': 'log2', 'min_samples_split': 45}
SVM	{'C': 1, 'gamma': 0.8, 'kernel': 'linear'}
KNN	{'leaf_size': 5, 'n_neighbors': 23}
NB	{'var_smoothing': 1e-07}
AdaBoost	{n_estimators=70, learning_rate=0.1}

Bagging	{max_depth=8,min_samples_split=40,criterion='gini',max_features= 'log2' ,ccp_alpha= 0.01,{'n_estimators': 30}
G_Boosting	{'learning_rate': 0.1, 'max_depth': 6, 'max_features': 'log2', 'min_samples_split': 35, ccp_alpha= 0.01,'n_estimators': 70, 'subsample': 0.8}
RF	{'criterion': 'gini', 'max_depth': 6, 'max_features': 'auto', 'min_samples_split': 40, 'n_estimators': 110}

3.4.3 Validation of Ensemble Models

Choosing a sufficient number of estimators has a direct impact on the accuracy of the model. Also, this argument will indirectly affect the overfitting of the model. So in this section, the accuracy of each ensemble model is evaluated by the number of estimators. The ensemble models' hyperparameters have been adjusted using grid search. For bagging and gradient boosting models, the accuracy of the model has been calculated for 20 to 75 estimators with an interval limit of 5. The results are shown in Figure 5. Based on the result, the best estimators from Bagging are 20. Also, Gradient Boosting is between 40 and 55. The AdaBoost and Random Forest results also are shown in Figure 6. According to the obtained results, the best estimator for the AdaBoost model is 45. The optimal number of estimators for the random forest model is 15.

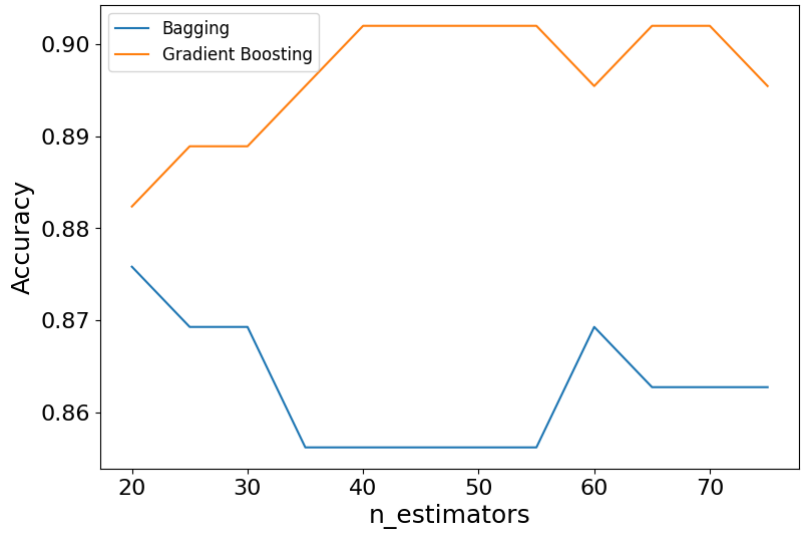

Figure5: Accuracy of the Bagging and Gradient Boosting model according to the estimator

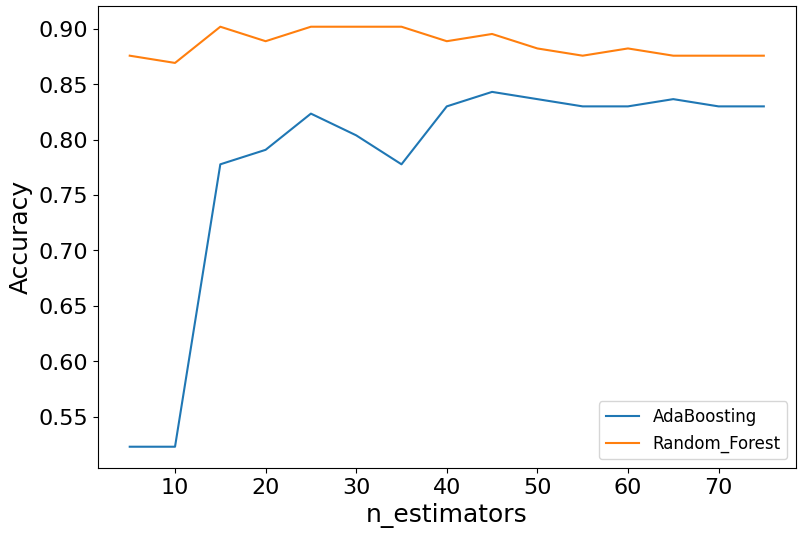

Figure6: Accuracy of the AdaBoost and Random Forest model according to the estimator

4 Comparing models by various Imputation models

In this article, after different methods have been used for imputing missing values to check the performance of each of these algorithms, the resulting models are used in predicting diabetic patients. These algorithms include DT, SVM, KNN and NB. Ensemble learning algorithm has been used to improve the performance of basic algorithms such as AdaBoost,Bagging,Gradient Boosting and RF.

A comparison of the performance of each model is used to select the best model from the tested models and to select the best model for disease prediction. Moreover, the data is divided into two parts based on the 20-80% percent range. Evaluation criteria have been used to evaluate the new data based on their evaluation and accuracy. In addition, Accuracy, Precision, Recall, F1-score and AUC are available. A forecast's accuracy is determined by Accuracy. It is desirable to have higher values for these tests in order to achieve better performance from the model. Essentially, when the measurement is closer to 100%, it indicates a better fit and better overall performance.

For the best algorithm in disease prediction, it has been compared the classification algorithms that include different missing values imputation models. Applying ensemble learning algorithm has improved the accuracy of all basic models from 1 to 10 percent. The hyperparameters of these algorithms are optimized by grid search to the greatest extent possible. Consequently, their error rates are as low as possible.

In Figure 7, the accuracy of classification models is compared by considering three models of filling in missing values. According to the results of the graph, it can be explained that the DT model performed better than other models in all three imputation methods.

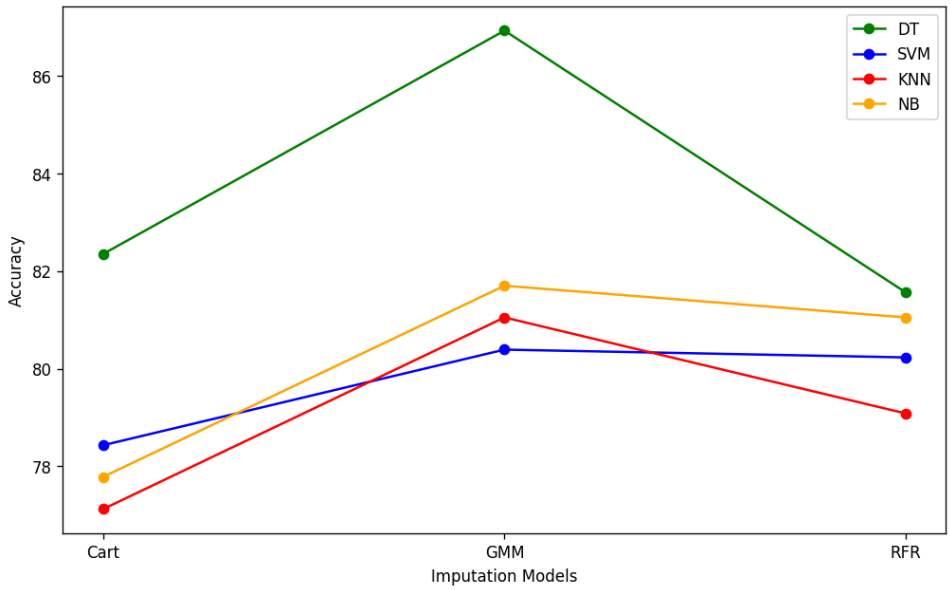

Figure 7: Comparing the Accuracy of 4 classification models

The implementation of ensemble learning models can significantly increase the accuracy of the model. Ensemble models, in reality, merge and enhance the outcomes of multiple less potent models. The results of Adaboost, Bagging, Gradient Boosting and Random Forest models considering missing values filling models are shown in Figure 8.

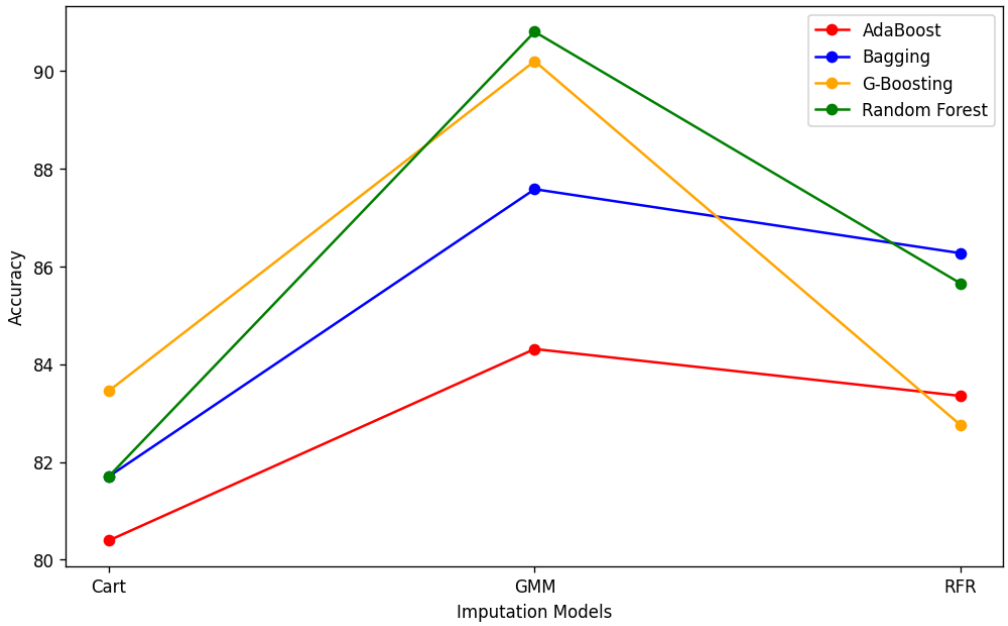

Figure 8: Comparing the Accuracy of 4 ensemble models

4.1 Results of Doctor Opinion

In addition to the use of classification algorithms, the doctor's point of view has also been used in this research. 765 data in this research were provided to the doctor without considering the response variable. Regarding each patient, based on the features in the dataset, the doctor gave his opinion about whether or not the patient has diabetes. Finally, according to the doctor's results, it is determined that the accuracy of the model will be 65%. As it is clear from Table 12, the results of all the models, including the classification and ensemble models considering the lowest accuracy, are at least 12% better than the results obtained by the doctor's opinion. Accuracy, recall, precision, F1 score and AUC of doctor's opinions compared to other models are shown in Table 12.

Table12: Comparison of doctor's opinion with other models

Methods	Accuracy	Recall	Precision	F1-score	AUC
Doctor result	65.00%	0.67	0.64	0.65	0.70
DT_RFR	81.56%	0.82	0.80	0.80	0.83
SVM_Cart	78.43%	0.79	0.73	0.75	0.80
KNN_Cart	77.12%	0.76	0.73	0.74	0.82
NB_Cart	77.78%	0.76	0.75	0.76	0.80
AdaBoost_Cart	80.39%	0.80	0.77	0.78	0.84
Bagging_Cart	81.7%	0.82	0.78	0.79	0.84
G_Boosting_RFR	82.75%	0.82	0.80	0.81	0.84
RF_Cart	81.7%	0.81	0.79	0.80	0.83

According to Figure 9, it can be concluded that KNN_Cart, which is the weakest model among all the classification models, has performed 12% better than the doctor's results. Also, the results of the Gradient Boosting model with an accuracy of 82.75% have improved nearly 18% of the doctor's results. In general, the use of classification and enhanced algorithms have improved the performance of doctors in the diagnosis of diabetes.

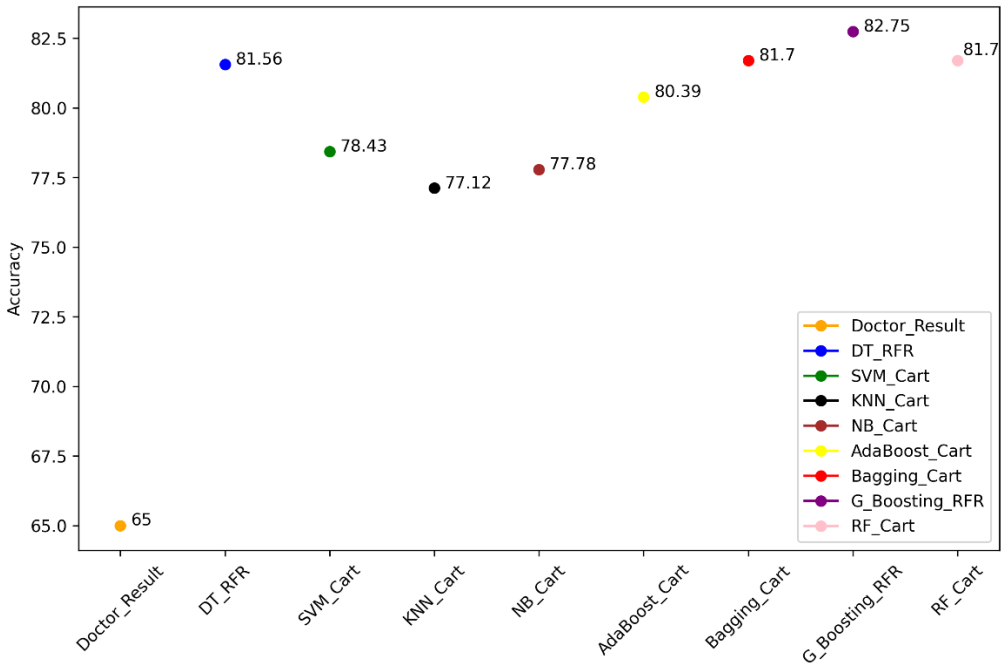

Figure 9 :Comparison of accuracy of all models with doctor's results

5. Discussions

In addition to the importance of correctly imputing features like Insulin and Skinthickness for disease diagnosis, particularly diabetes, this research delves into the utilization of various data completion models, such as Cart, GMM, and RFR. Prior to this study, the literature lacked comprehensive discussions on the optimal techniques for handling missing values, making this research particularly impactful in addressing this gap. Furthermore, the concurrent validation of ensemble methods enhances the robustness of the findings, providing a more reliable basis for decision-making in clinical settings. Moreover, the integration of medical expertise, alongside a thorough comparison with implemented models, adds depth to the analysis. By incorporating insights from healthcare professionals, the study ensures that the developed models are not only statistically sound but also clinically relevant. This collaborative approach strengthens the validity and applicability of the research findings.

Expanding on the machine learning methodologies employed, the research encompasses eight distinct algorithms, namely DT, SVM, KNN, NB, AB, Bagging, Gradient Boosting, and Random Forest. Notably, the results underscore the superiority of ensemble algorithms over traditional classification models, showcasing the efficacy of leveraging collective intelligence for predictive modeling in medical diagnostics. The comparative analysis reveals nuanced insights into the performance of different imputation techniques. While the GMM model emerges as the most accurate method for handling missing values, variations in algorithmic performance are observed across different classifiers. For instance, the DT method exhibits superior accuracy in segregating

diabetic and healthy patients, showcasing its potential as a foundational model within ensemble frameworks.

Additionally, the exceptional accuracy of the Random Forest algorithm within the GMM context highlights the importance of algorithm-selection synergy in optimizing predictive performance. This nuanced understanding not only enhances the interpretability of the results but also provides actionable insights for future research and clinical practice.

The investigation presented herein delves into the pivotal significance of precise feature imputation within the realm of medical diagnostics, with a particular emphasis on diabetes diagnosis. Through an exploration of diverse data completion methodologies such as Cart, GMM, and RFR, the study addresses a notable lacuna within extant literature pertaining to optimal strategies for handling missing values. This discourse is especially pertinent given the burgeoning reliance on machine learning algorithms within healthcare decision-making paradigms.

The concurrent validation of ensemble methods serves to augment the depth and integrity of the inquiry, thereby bolstering the reliability and resilience of the resultant findings. By harnessing the collective intelligence inherent in multiple algorithms, the research not only engenders heightened predictive accuracy but also underscores the imperative of model heterogeneity in mitigating both bias and variance.

Furthermore, the integration of clinical expertise in the development and validation of predictive models constitutes a commendable aspect of this study. Collaboration with healthcare professionals ensures that the devised models not only adhere to statistical rigor but also remain tethered to clinical exigencies. This interdisciplinary synergy not only amplifies the translational potential of research outcomes but also facilitates more efficacious disease diagnosis and patient care protocols.

The comparative evaluation of machine learning algorithms unveils nuanced insights into their respective efficacies and limitations. While ensemble methodologies exhibit superior performance compared to conventional classification models, discernible differentials in algorithmic efficacy emerge contingent upon the employed imputation techniques. For instance, the GMM model evinces superior accuracy in managing missing values relative to Cart and RFR methodologies, thereby underscoring the pivotal importance of tailored data imputation strategies predicated upon dataset intricacies.

Moreover, the identification of the DT method as a cornerstone within ensemble frameworks accentuates its intrinsic potential in ameliorating predictive performance, particularly in discerning between diabetic and non-diabetic cohorts. This discernment not only enriches our comprehension of algorithmic comportment but also furnishes actionable insights conducive to optimizing model selection and configuration in prospective research endeavors and clinical deployments.

6. Conclusion

Diabetes mellitus, or simply diabetes, is a group of diseases that make the body's use of blood sugar (glucose) difficult. Diabetic metabolism is caused either by an insufficient amount of insulin produced by the pancreas or by a lack of insulin that is used efficiently by the body. Blood sugar is released by the hormone insulin. The condition of hyperglycemia, or increased blood sugar, is a common complication of uncontrolled diabetes and over time causes serious damage to many body systems, including the blood vessels and nerves. Diabetes is generally divided into type 1 and type 2, which is predicted to be type 2 diabetes in this research. It is possible for long-term complications of diabetes to develop gradually if diabetes is not diagnosed correctly. The long-term effects of diabetes can be devastating to the heart, blood vessels, kidneys, nerves, eyes, and blood vessels. Therefore, predicting diabetes is very important for preventing its complications.

In this research, classification algorithms have been used to predict diabetes. One practical method to increase the model's accuracy is to use various missing data imputation models methods. Because the case study is a research related to diabetic patients, it is of great importance to fill in features such as glucose and insulin. Therefore, it is not logical to fill these values using the average and mode methods. In fact, the purpose of using this method is to identify the best data completion models. According to the results obtained from the doctor's point of view, it can be concluded that the doctor's accuracy is lower compared to the accuracy of the implementation models.

Based on the conducted research, it is suggested for future research to use effective features based on the doctor's opinion in the field of diabetes diagnosis. Increasing the number of effective features is another thing that is suggested to future researchers. It is also suggested that if they have access to hospital data, they should test the predictive models that have shown high accuracy and present the results. Diagnosing diabetes with X-rays, especially in the pancreas and abdominal areas, can help improve the prognosis of diabetic patients. The results show that, generally, diabetic patients suffer from other complications caused by this disease. Therefore, to prevent these types of complications, a decision-making framework can be designed to help these patients. This framework can provide the patient with appropriate solutions to prevent complications.

References

Abdali, N., Vaezi, M.A., Rabani, M. and Aghsami, A., 2024. A new data-driven decision-making method for therapist patient allocation and scheduling. Journal of Industrial and Systems Engineering.

Al-Hadeethi, H., Abdulla, S., Diykh, M., Deo, R. C., & Green, J. H. (2020). Adaptive boost LS-SVM classification approach for time-series signal classification in epileptic seizure diagnosis applications. Expert Systems with Applications, 161, 113676.

Andrade, L., Rapp, T., & Sevilla-Dedieu, C. (2018). Quality of diabetes follow-up care and hospital admissions. International Journal of Health Economics and Management, 18, 153-167.

Aryai, V., & Goldsworthy, M. J. E. A. o. A. I. (2023). Day ahead carbon emission forecasting of the regional National Electricity Market using machine learning methods. 123, 106314.

Behdinian, A., Amani, M.A., Aghsami, A. and Jolai, F., 2022. An Integrating Machine Learning Algorithm and Simulation Method for Improving Software Project Management: A Case Study. Journal of Quality Engineering and Production Optimization, 7(1), pp.54-74.

Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2021). A comparative analysis of gradient boosting algorithms. Artificial Intelligence Review, 54, 1937-1967.

Blanquero, R., Carrizosa, E., Ramírez-Cobo, P. and Sillero-Denamiel, M.R., 2021. Variable selection for Naïve Bayes classification. Computers & Operations Research, 135, p.105456.

Deberneh, H. M., & Kim, I. (2021). Prediction of Type 2 diabetes based on machine learning algorithm. International Journal of Environmental Research and Public Health, 18(6), 3317.

Doğru, A., Buyrukoğlu, S. and Arı, M., 2023. A hybrid super ensemble learning model for the early-stage prediction of diabetes risk. Medical & Biological Engineering & Computing, 61(3), pp.785-797.

Fazakis, N., Kocsis, O., Dritsas, E., Alexiou, S., Fakotakis, N., & Moustakas, K. (2021). Machine learning tools for long-term type 2 diabetes risk prediction. IEEE Access, 9, 103737-103757.

Feng, D.-C., Liu, Z.-T., Wang, X.-D., Chen, Y., Chang, J.-Q., Wei, D.-F., & Jiang, Z.-M. (2020). Machine learning-based compressive strength prediction for concrete: An adaptive boosting approach. Construction and Building Materials, 230, 117000.

Fereidouni, Z., Mehdizadeh Somarin, Z., Mohammadnazari, Z., Aghsami, A. and Jolai, F., 2022. Analysis of correlation between food consumption habits and COVID-19 outbreak. Journal of Industrial and Systems Engineering, 14(2), pp.86-118.

Ghosh, P., Azam, S., Karim, A., Hassan, M., Roy, K., & Jonkman, M. (2021). A comparative study of different machine learning tools in detecting diabetes. Procedia Computer Science, 192, 467-477.

Harimoorthy, K., & Thangavelu, M. (2021). Multi-disease prediction model using improved SVM-radial bias technique in healthcare monitoring system. Journal of Ambient Intelligence and Humanized Computing, 12(3), 3715-3723.

Heikes, K. E., Eddy, D. M., Arondekar, B., & Schlessinger, L. (2008). Diabetes Risk Calculator: a simple tool for detecting undiagnosed diabetes and pre-diabetes. Diabetes care, 31(5), 1040-1045.

Hernández-Pereira, E., Fontenla-Romero, O., Bolón-Canedo, V., Cancela-Barizo, B., Guijarro-Berdiñas, B., & Alonso-Betanzos, A. J. A. I. (2022). Machine learning techniques to predict different levels of hospital care of CoVid-19. 52(6), 6413-6431.

Heydari, M., Teimouri, M., Heshmati, Z., & Alavinia, S. M. (2016). Comparison of various classification algorithms in the diagnosis of type 2 diabetes in Iran. International Journal of Diabetes in Developing Countries, 36(2), 167-173.

Jain, P. K., Quamer, W., & Pamula, R. J. O. (2021). Sports result prediction using data mining techniques in comparison with base line model. 58(1), 54-70.

Joshi, R. D., & Dhakal, C. K. (2021). Predicting type 2 diabetes using logistic regression and machine learning approaches. International Journal of Environmental Research and Public Health, 18(14), 7346.

Kamel, S. R., & Yaghoubzadeh, R. (2021). Feature selection using grasshopper optimization algorithm in diagnosis of diabetes disease. Informatics in Medicine Unlocked, 26, 100707.

Karatsiolis, S., & Schizas, C. N. (2012). Region based Support Vector Machine algorithm for medical diagnosis on Pima Indian Diabetes dataset. Paper presented at the 2012 IEEE 12th International Conference on Bioinformatics & Bioengineering (BIBE).

Kee, O.T., Harun, H., Mustafa, N., Abdul Murad, N.A., Chin, S.F., Jaafar, R. and Abdullah, N., 2023. Cardiovascular complications in a diabetes prediction model using machine learning: a systematic review. Cardiovascular Diabetology, 22(1), p.13.

Khanam, J. J., & Foo, S. Y. (2021). A comparison of machine learning algorithms for diabetes prediction. ICT Express, 7(4), 432-439.

Kumari, V. A., & Chitra, R. (2013). Classification of diabetes disease using support vector machine. International Journal of Engineering Research and Applications, 3(2), 1797-1801.

Li, J., Yuan, P., Hu, X., Huang, J., Cui, L., Cui, J., . . . Li, J. (2021). A tongue features fusion approach to predicting prediabetes and diabetes with machine learning. Journal of biomedical informatics, 115, 103693.

Lu, H., Uddin, S., Hajati, F., Moni, M. A., & Khushi, M. (2022). A patient network-based machine learning model for disease prediction: The case of type 2 diabetes mellitus. Applied Intelligence, 52(3), 2411-2422.

Mamoudan, M. M., Jafari, A., Mohammadnazari, Z., Nasiri, M. M., & Yazdani, M. (2023). Hybrid machine learning-metaheuristic model for sustainable agri-food production and supply chain planning under water scarcity. Resources, Environment and Sustainability, 14, 100133. doi:https://doi.org/10.1016/j.resenv.2023.100133

Mamoudan, M.M., Forouzanfar, D., Mohammadnazari, Z., Aghsami, A. and Jolai, F., 2023. Factor identification for insurance pricing mechanism using data mining and multi criteria decision making. Journal of Ambient Intelligence and Humanized Computing, 14(7), pp.8153-8172.

Manjurul Ahsan, M., & Siddique, Z. (2021). Machine Learning-Based Heart Disease Diagnosis: A Systematic Literature Review. arXiv e-prints, arXiv: 2112.06459.

Mansourypoor, F., & Asadi, S. (2017). Development of a Reinforcement Learning-based Evolutionary Fuzzy Rule-Based System for diabetes diagnosis. Computers in Biology and Medicine, 91, 337-352. doi:https://doi.org/10.1016/j.compbiomed.2017.10.024

Muhammad, L., Algehyne, E. A., & Usman, S. S. (2020). Predictive supervised machine learning models for diabetes mellitus. SN Computer Science, 1(5), 1-10.

Nguyen, B. P., Pham, H. N., Tran, H., Nghiem, N., Nguyen, Q. H., Do, T. T., . . . Simpson, C. R. (2019). Predicting the onset of type 2 diabetes using wide and deep learning with electronic health records. Computer methods and programs in biomedicine, 182, 105055.

Nozari, H. (2024). Green Supply Chain Management based on Artificial Intelligence of Everything. Journal of Economics and Management, 46, 171-188.

Nozari, H. (Ed.). (2023). Building Smart and Sustainable Businesses with Transformative Technologies. IGI Global.

Nozari, H., Ghahremani-Nahr, J., Fallah, M., & Szmelter-Jarosz, A. (2022). Assessment of cyber risks in an IoT-based supply chain using a fuzzy decision-making method. International Journal of Innovation in Management, Economics and Social Sciences, 2(1).

Özdemir, M. A., Özdemir, G. D., Gül, M., Güren, O., Ercan, U. K. J. M. L. S., & Technology. (2023). Machine learning to predict the antimicrobial activity of cold atmospheric plasma-activated liquids. 4(1), 015030.

Pan, L., Sun, W., Wan, W., Zeng, Q. and Xu, J., 2023. Research Progress of Diabetic Disease Prediction Model in Deep Learning. Journal of Theory and Practice of Engineering Science, 3(12), pp.15-21.

Park, J., & Edington, D. W. (2001). A sequential neural network model for diabetes prediction. Artificial intelligence in medicine, 23(3), 277-293.

Pourkhodabakhsh, N., Mamoudan, M. M., & Bozorgi-Amiri, A. (2023). Effective machine learning, Meta-heuristic algorithms and multi-criteria decision making to minimizing human resource turnover. Applied Intelligence, 53(12), 16309-16331.

Qi, H., Song, X., Liu, S., Zhang, Y., & Wong, K. K. L. (2023). KFPredict: An ensemble learning prediction framework for diabetes based on fusion of key features. Comput Methods Programs Biomed, 231, 107378. doi:10.1016/j.cmpb.2023.107378

Rastogi, R. and Bansal, M., 2023. Diabetes prediction model using data mining techniques. Measurement: Sensors, 25, p.100605.

Sai, M. J., Chettri, P., Panigrahi, R., Garg, A., Bhoi, A. K., & Barsocchi, P. J. I. J. o. C. I. S. (2023). An Ensemble of Light Gradient Boosting Machine and Adaptive Boosting for Prediction of Type-2 Diabetes. 16(1), 14.

Santhanam, T., & Padmavathi, M. (2015). Application of K-means and genetic algorithms for dimension reduction by integrating SVM for diabetes diagnosis. Procedia Computer Science, 47, 76-83.

Segar, M. W., Patel, K. V., Vaduganathan, M., Caughey, M. C., Jaeger, B. C., Basit, M., . . . Wang, T. J. J. D. (2021). Development and validation of optimal phenomapping methods to estimate long-term atherosclerotic cardiovascular disease risk in patients with type 2 diabetes. 64, 1583-1594.

Srivastava, N. K., Singh, S. K., & Singh, U. J. O. (2022). Analysis and prediction of Covid-19 spreading through Bayesian modelling with a case study of Uttar Pradesh, India. OPSEARCH, 1-16.

Sudharsan, B., Peeples, M., & Shomali, M. (2014). Hypoglycemia prediction using machine learning models for patients with type 2 diabetes. Journal of diabetes science and technology, 9(1), 86-90.

Tsai, C.-F., Hu, Y.-H. J. K., & Systems, I. (2022). Empirical comparison of supervised learning techniques for missing value imputation. 64(4), 1047-1075.

Vijayan, V., & Ravikumar, A. (2014). Study of data mining algorithms for prediction and diagnosis of diabetes mellitus. International journal of computer applications, 95(17).

Wang, L., Wang, X., Chen, A., Jin, X., & Che, H. (2020). Prediction of type 2 diabetes risk and its effect evaluation based on the XGBoost model. Paper presented at the Healthcare.

Werner de Vargas, V., Schneider Aranda, J. A., dos Santos Costa, R., da Silva Pereira, P. R., Victória Barbosa, J. L. J. K., & Systems, I. (2022). Imbalanced data preprocessing techniques for machine learning: a systematic mapping study. 65(1), 31-57.

Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., & Tang, H. (2018). Predicting diabetes mellitus with machine learning techniques. Frontiers in genetics, 515.