E-ISSN 2581 – 7957
P-ISSN 2277 – 3916

CVR Journal of Science and Technology, Volume 26, June 2024
DOI: 10.32377/cvrjst2618

Predictive Modeling of Diabetes Mellitus Utilizing Machine Learning Techniques

N Nagarjuna[1] and Dr. Lakshmi HN[2]
[1]Senior Assistant Professor, CVR College of Engineering/ CSIT Dept., Hyderabad, India
Email: n.nagarjuna@cvr.ac.in
[2]Professor&Head, CVR College of Engineering/ Emerging Technologies Dept., Hyderabad, India
Email: hn.lakshmi@cvr.ac.in

Abstract: **Diabetes mellitus represents a persistent metabolic condition distinguished by elevated levels of blood sugar, which results from the inadequacy of the body to secrete and respond to insulin, leading to health risks and frequent hospitalizations. Accurate predictive models are vital for targeted interventions to reduce readmissions and improve healthcare quality and cost. Early prediction can mitigate its impact, aid in control, and potentially save lives. Machine learning algorithms show promise in medical applications, including diabetes prediction and diagnosis. Limited data quality hinders accurate diabetes prediction due to missing values and inconsistencies. This paper investigates machine learning's potential for predicting and diagnosing diabetes, aiming to enhance accuracy and efficiency in disease management. Feature engineering techniques are applied to preprocess the data and extract relevant features for model development. To address class imbalance, SMOTE (Synthetic Minority Oversampling Technique) is employed. Various machine learning algorithms, including logistic regression, Naïve Bayes, random forests, support vector machines (SVM), K-Nearest Neighbors (KNN), and eXtreme Gradient Boosting (XGBoost), are utilized to build predictive models. The performance evaluation employs standard metrics such as accuracy, recall, precision, and F1-Score. Notably, Random Forest achieves an accuracy of 82% followed by XGBoost(80%) , surpassing other ML algorithms utilized.**

Index Terms: **Diabetes mellitus, Machine learning, Prediction, SVM, logistic regression, Accuracy.**

I. Introduction

Diabetes is a long-term health condition caused by either inadequate production of insulin or the body's ineffective use of insulin. Insulin, a hormone crucial for regulating blood sugar levels, plays a key role in this mechanism. Common symptoms comprise increased thirst, frequent urination, blurred vision, fatigue, and unintentional weight loss. Prolonged diabetes can harm the heart, eyes, kidneys, and nerves, increasing the risk of heart attack, stroke, kidney failure, and vision loss, along with potential foot ulcers and amputation [1].

Currently, 537 million adults worldwide have diabetes, and by 2030, an estimated 643 million people will be living with diabetes, rising to a staggering 783 million by 2045. Over three-quarters reside in low- to middle-income nations. In 2021, diabetes claimed 6.7 million lives, with a health expenditure toll of at least USD 966 billion—a 316% surge over 15 years. Additionally, 541 million adults face Impaired Glucose Tolerance (IGT), heightening their risk of developing type 2 diabetes [2].

There are three major forms of diabetes: type-1, type-2, and gestational diabetes. Type-1, affecting 5-10% of cases, is an autoimmune condition where the body mistakenly attacks insulin-producing cells. This necessitates lifelong insulin therapy to manage blood sugar levels. Type 2, comprising 90-95% of cases, occurs when insulin is ineffective in regulating blood sugar, often diagnosed in adults but increasingly in younger individuals. Regular blood sugar testing is crucial due to subtle symptoms, and prevention through lifestyle changes is feasible. Gestational diabetes affects pregnant women, raising health risks for both mother and baby, with implications for future health conditions. [3].

While a cure for diabetes remains elusive, early, and accurate prediction offers promising avenues for control and prevention. However, predicting diabetes poses a challenge due to the non-linear nature of data. Recent research, employing machine learning's ability to learn without explicit programming, has shown promising results in forecasting diabetes risk. While machine learning holds promise in medicine, ensuring consistent accuracy across different algorithms remains a challenge. Identifying the algorithm with the highest performance is crucial for building better classifiers. The pervasive reach of machine learning across industries extends to medicine, where its potential to revolutionize healthcare is significant [4]. Machine learning and statistics are used in predictive modeling to find patterns in data and estimate the likelihood that certain events will occur [5]. This work aims to create a model predicting diabetes in patients. Afterward, different methods are investigated to increase the model's accuracy.

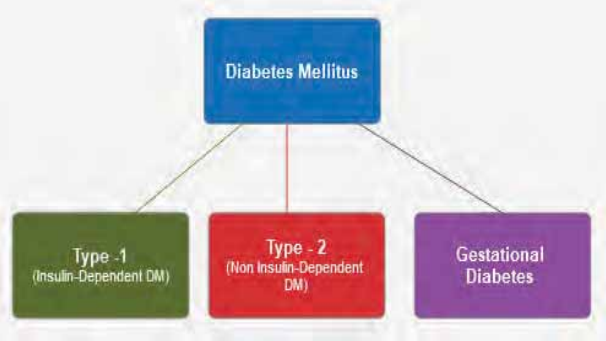


Figure1. Diabetes Mellitus types

Received on 09-03-2024, Revised on 08-04-2024, Accepted on 08-05-2024.

CVR Journal of Science and Technology, Volume 26, June 2024
DOI: 10.32377/cvrjst2618

II. Related Work

In a study by Iyer A. [6], a diabetic dataset was analyzed to identify hidden patterns using classification algorithms. Naive Bayes and Decision Trees were compared, demonstrating the effectiveness of both methods.

In a study by Mercaldo et al. [7], researchers investigated machine learning algorithms for diagnosing diabetes. They compared six algorithms, including J48, multilayer perceptron, JRip, Hoeffding Tree, Bayes Net, and random forest, using the WEKA software and the PIDD dataset. Notably, the Hoeffding Tree algorithm showed promising results for diabetes prediction.

Sisodia et al. [8] compared three classifiers (Naïve Bayes, SVM, Decision Trees) for diabetes prediction using the Pima Indian Diabetes Database. Naïve Bayes achieved the highest accuracy (76.30%) among F-Score, Precision, Recall, and Accuracy metrics.

Maniruzzaman et al. [9] conducted in-depth research on filling in missing variables and rejecting outliers to improve the performance of the machine learning model.

Sneha et al. [10] used feature selection to identify the most informative attributes from the PIMA diabetes dataset for their prediction model. They evaluated several machine learning algorithms, including Support Vector Machine, k-Nearest Neighbors, Naive Bayes, Decision Tree, and Random Forest. Naive Bayes achieved the best accuracy of 82.2%.

Lukmanto et al. [11] utilized feature selection on PIMA Indian dataset using Fuzzy SVM for their prediction with a promising accuracy of 89.02%.

In a study by Ahuja et al. [12], researchers evaluated 15 classification algorithms, including Multilayer Perceptron, using Python. They addressed missing data by imputing missing values with the median and replacing outliers. The performance of the algorithms was assessed using five different dataset selection methods and 2, 4, 5, and 10-fold cross-validation. Their findings revealed that Multilayer Perceptron achieved the highest accuracy (78.7%) when combined with feature selection using Linear Discriminant Analysis.

Morgan-Benita et al. proposed Hard Voting Ensemble Approach (HVEA) for diabetes prediction [13]. Compared to individual models like Logistic Regression (88.01%), Support Vector Machine (89.82%), and Artificial Neural Network (88.46%), HVEA achieved significantly higher accuracy (90.05%) using non-glucose data from Mexico and 10-fold cross-validation.

III. Methodology

The key elements of the methodology for predicting diabetes are visualized in figure 2. The input data is preprocessed by handling missing values and outliers. Feature extraction and selection are performed to reduce complexity and focus on relevant information. Data is split into training and testing sets. The model is trained on the training data and evaluated on the test data. Results are presented as confusion matrices. Using the trained model and selected features, predictions for new diabetic patients are made.

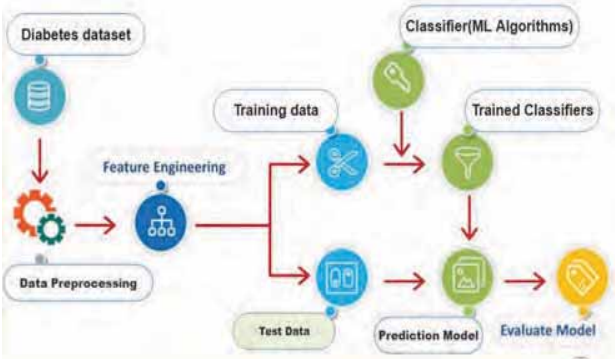


Figure 2. Proposed methodology

A. Dataset

This study utilized the PIDD (Pima Indians Diabetes Database), available from the UCI (University of California), Irvine repository. The dataset contains information on 768 females aged 21 or older. PIDD offers ample recorded instances and requires minimal pre-processing for integration into various learning models. As a result, it is widely used in machine learning and deep learning models for detecting diabetes.

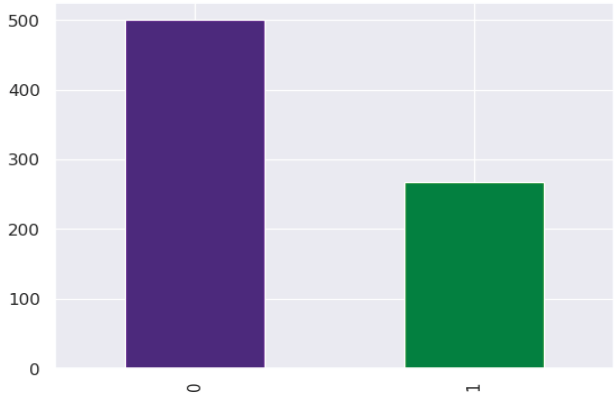


Figure 3. number of individuals with and without diabetes

In Figure 3, a bar chart illustrates the distribution of patients with and without diabetes. X-axis displays the outcome and y-axis displays the count. Of these, 268 have been diagnosed with diabetes (designated as 1), and 500 do not have diabetes (designated as 0). Figure 3 reveals a significant class imbalance in the dataset (268 diabetes vs. 500 non-diabetic). This imbalance can hinder machine learning model performance. To address this, SMOTE (Synthetic Minority Oversampling Technique) is employed. SMOTE generates synthetic samples for the minority class, balancing the class distribution.

Table 1 provides description of dataset. It contains nine feature columns, including the month of pregnancy, glucose, plasma, blood pressure, triceps, insulin, BMI, age, and Diabetes Pedigree Function (DPF). The dataset's missing

CVR Journal of Science and Technology, Volume 26, June 2024
DOI: 10.32377/cvrjst2618

values, which were at first thought to be complete, were eventually determined to be zeros. Zero values, however, were thought to be physiologically impractical for some characteristics, such as age or blood pressure. In a similar vein, readings for plasma glucose, 2-hour serum insulin, and body mass index were implausible, with almost 50% showing zero. Crucially, every attribute present in the database is either a real number or an integer.

TABLE I.
DATA SET DESCRIPTION

Attribute	Description	Data type	Range
Pregnancy	Number of pregnancies	Number	0 –17
Plasma glucose	Blood sugar level	Number	0 –199
Triceps	Subcutaneous fat thickness(mm)	Number	0–99
Blood pressure	blood pressure, expressed in mm Hg	Number	0–122
BMI	Body mass index, kg/m²	float	0 – 67.1
Serum insulin	Two hour serum insulin (μU/mL)	Number	0 – 846
Age	Age in Years	Number	21 – 81
DPF	diabetes risk based on family history	float	0.078 – 2.42
Outcome	Value indicating diabetes diagnosis (Positive/Negative)	Boolean	0,1

B. *Pre-processing*

Machine learning algorithms rely heavily on data for effective model training. Initially, datasets collected from various sources are often in a raw format, prone to inconsistencies that models may struggle with. Pre-processing is crucial to clean the data, involving tasks such as handling missing values, creating new features, and splitting the data into train-test sets. To address the issue of certain features with zero values, such as blood pressure, skin thickness, BMI, mean, and median imputations are employed to replace missing or zero values. Consequently, missing values for select attributes, including Blood Pressure, Glucose level, BMI, Skin Thickness, and Age, were imputed, given that these attributes cannot logically hold zero values. Following the imputation process, the dataset was scaled to normalize all values.

A carefully chosen data set is characterized by features that exhibit substantial correlations with the target class along with substantial discordances with one another. A filter-based feature selection is used for selecting features to identify the uncorrelated features. Figure 4 illustrates the relationship between all features. The darker colors represent less association, while lighter colors indicate greater correlation.

Feature importance identifies the most influential features in a machine learning model, aiding in understanding data relationships, feature selection, model interpretation, and debugging. Figure 5 demonstrates that the Glucose feature has the most significant influence on predicting diabetes.

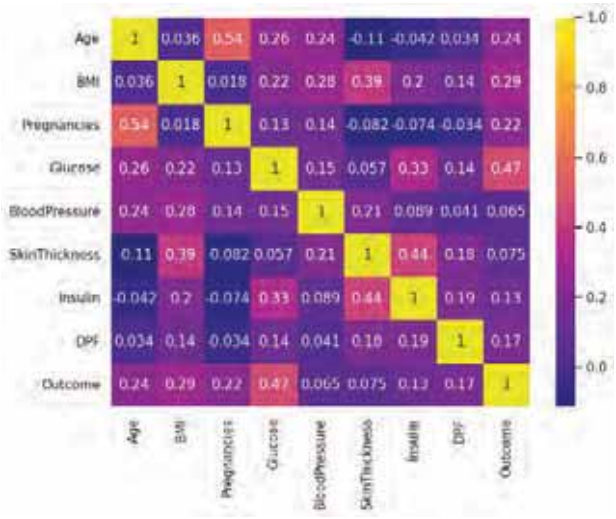


Figure 4. Heatmap showing Correlation between features.

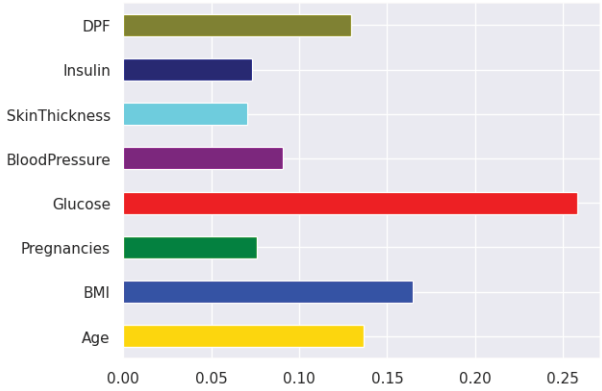


Figure 5. Feature importance of diabetes data set

A. *Algorithms used:*

i. Logistic Regression

For binary classification, logistic regression estimates outcome probabilities from predictor variables. It is widely used in binary classification tasks. Logistic regression estimates the probabilities using the logistic function, which maps any real-valued input into the range 0 and 1 [14].

ii. Support Vector Machines

SVM stands out as an algorithm adept at both classification and regression tasks. It achieves this by identifying, within a high-dimensional space, the hyperplane that best segregates classes. This is accomplished by maximizing the margin, or distance, between these classes. Even in high-dimensional spaces, SVMs remain effective due to their versatility in handling non-linear data through kernel functions [15].

iii. K-Nearest Neighbors (KNN)

For classification tasks, the KNN (K-Nearest Neighbors) algorithm identifies the k closest data points in the feature space and assigns the majority class label. It is simple yet effective for both classification and regression tasks. However, it can be computationally expensive for large datasets [16].

CVR Journal of Science and Technology, Volume 26, June 2024
DOI: 10.32377/cvrjst2618

iv. Naïve Bayes:

The Naive Bayes algorithm is a probabilistic classification method that calculates the probability of each class for a given set of input features. It leverages Bayes' theorem and assumes independence between features to make these classifications. Naïve Bayes is particularly popular in text classification tasks due to its simplicity and efficiency [17].

v. Random Forest

Random Forest, a powerful machine learning algorithm for classification and regression, achieves its strength by combining a multitude of diverse decision trees. Each tree, built on random data subsets and features, casts a vote (classification) or contributes to an average (regression). This democratic approach improves accuracy, handles diverse data, and reduces overfitting, making it powerful and versatile. It shines in various domains but might not always outperform complex algorithms on specific tasks.

vi. Decision Tree

Decision trees, resembling upside-down trees, excel at both classification and regression, splitting data based on features for clear predictions. These models shine in interpretability, handling diverse data, and revealing feature importance. However, they can overfit the data, are sensitive to noise, and favor features with many categories. Despite these limitations, decision trees remain valuable for tasks where understanding the "why" behind predictions is essential[19].

vii. XGBoost

XGBoost, or eXtreme Gradient Boosting, is a highly optimized implementation of the gradient boosting algorithm, widely recognized for its efficiency and scalability. XGBoost boasts built-in features like automated missing value handling, tree pruning, and hyperparameter tuning, streamlining the process. Additionally, it unlocks insights into important features through feature importance scores, enhancing interpretability and model understanding [20].

B. Training and Testing

The PIDD dataset is used to train and evaluate machine learning models in Python. For this, the data is split 70/30 for training and testing.

IV. Results

The final step of evaluating a prediction model involves assessing its performance using various metrics like accuracy, confusion matrix, precision, recall, and F1-score. These metrics rely on ground truth labels, also known as classification labels.

A. Confusion Matrix:

In classification tasks like diabetes prediction, a confusion matrix clearly shows how the model performed.

Table II.
Confusion Matrix for Diabetes Prediction

Predicted	Actual Positive (Diabetes)	Actual Negative (No Diabetes)
Positive (Predicted Diabetic)	True Positives (TPD)	False Positives (FPD)
Negative (Predicted Non-Diabetic)	False Negatives (FND)	True Negatives (TND)

Table 2 depicts the confusion matrix for diabetes prediction. True Positives are the cases where the SVM model correctly predicted patients who actually have diabetes. False Positives are the cases where the model incorrectly classified patients as diabetic when they actually don't have diabetes (Type I Error). False Negatives are the cases where the model incorrectly classified patients as non-diabetic when they actually have diabetes (Type II Error). True Negatives are the cases where the model correctly predicted patients who do not have diabetes.

Table III.
Performance Comparison of ML Algorithms for Diabetes Prediction

Algorithms	Accuracy	Precision	Recall	F1-Score
SVM	0.74	0.74	0.74	0.74
Logistic Regression	0.74	0.71	0.80	0.75
Naïve Bayes	0.77	0.78	0.75	0.77
KNN	0.72	0.71	0.75	0.73
Random Forest	0.82	0.83	0.82	0.82
Decision Tree	0.75	0.74	0.78	0.76
XGBoost	0.80	0.81	0.81	0.80

B. Accuracy:

This metric measures how well the classifier can identify cases of diabetes. Overall, the performance of the algorithms on this metric is good, with all models achieving an accuracy of over 0.72. .Random Forest and XGBoost achieve the highest accuracy (0.82,,0.80), indicating good ability to distinguish diabetic and non-diabetic individuals.

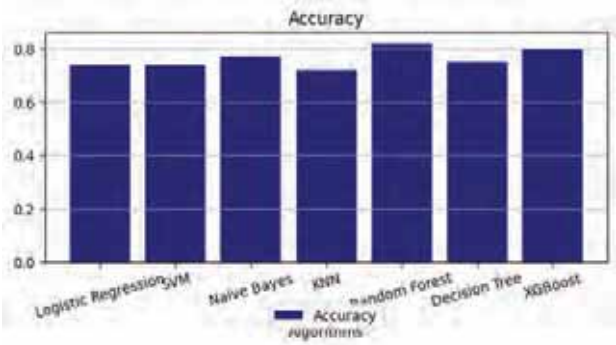


Figure 6. Accuracy of classifiers for diabetes prediction

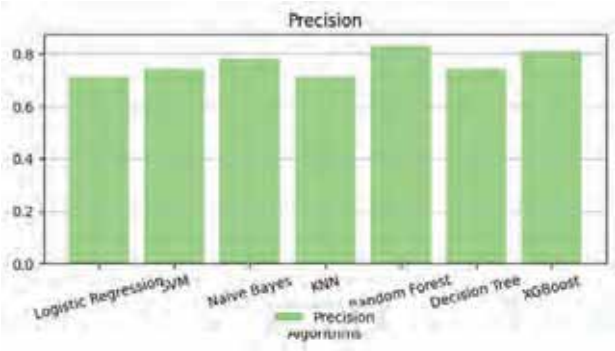
CVR Journal of Science and Technology, Volume 26, June 2024
DOI: 10.32377/cvrjst2618

C. Precision:

$$Precision = (TPD) / (TPD + (FPD). \quad (1)$$

It indicates the proportion of patients predicted to have diabetes who do have it. A high precision signifies the model's effectiveness in identifying individuals with diabetes and minimizing false positives. Here, again Random Forest(0.83) and XGBoost (0.81) outperform other models.

Figure 7. Precision of classifiers for diabetes prediction



D. Recall (Sensitivity):

$$Recall = TPD / (TPD + FND) \quad (2)$$

Recall indicates the proportion of actual diabetic cases that the model correctly identifies. A high recall means that the model is good at identifying all cases of diabetes and avoiding false negatives. Random Forest, followed by XGBoost, performs the best in terms of recall.

Figure 8. Recall of classifiers for diabetes prediction

E. F1-Score:

$$f1 - Score = 2 * \frac{(precision * Recall)}{(Precision + Recall)} \quad (3)$$

This metric balances the importance of precision and recall in a single score, and it takes both these factors into account. It is a good overall measure of a model's performance. Random Forest (0.82) and XGBoost have the best F1-score among all the models, which indicates that it achieves a good balance between precision and recall.

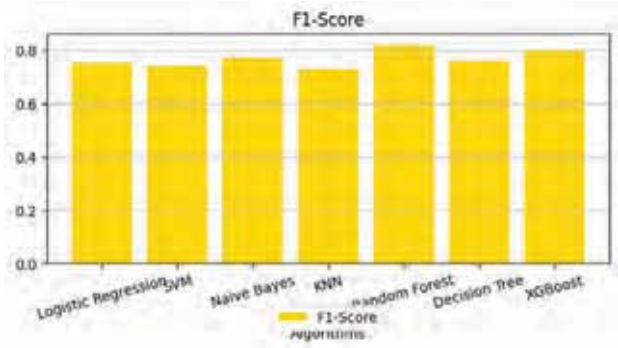


Figure 9. Precision of classifiers for diabetes prediction

Figure 10. Performance metrics of ML classifiers in Line graph

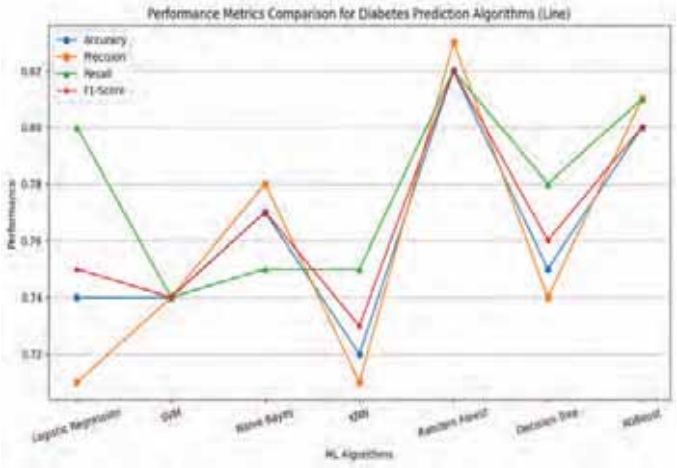


Figure 10 presents a line graph comparing evaluation metrics for various machine learning algorithms.

V. Conclusions

Early detection of diabetes is crucial for many who remain unaware. This paper explores machine learning techniques for high-accuracy diabetes risk prediction. Random Forest and XGBoost algorithms show promise, but success hinges on data preprocessing (cleaning, normalization, feature selection). A key challenge in diabetes prediction is class imbalance. SMOTE, applied during preprocessing, mitigates this issue by generating synthetic data for the underrepresented class, improving model performance. The choice between precision and recall depends on the cost of errors. For minimizing unnecessary tests, prioritize high-precision models like Random Forest. Conversely, for identifying all diabetic cases, focus on high-recall models like XGBoost. Our study achieved 82.5% accuracy using a Random Forest classifier, demonstrating a good balance between identifying healthy individuals and catching diabetic cases. This system's adaptability to other diseases paves the way for broader advancements in automated disease analysis and user-friendly web apps for risk prediction.

CVR Journal of Science and Technology, Volume 26, June 2024
DOI: 10.32377/cvrjst2618

REFERENCES

[1] Diabetes. (2023, April 5). World Health Organization (WHO).
[2] IDF diabetes atlas 2021. (n.d.). IDF Diabetes Atlas | Tenth Edition.
[3] What is Diabetes? (2023, September 5). Centers for Disease Control and Prevention.
[4] Wagai, G., Firdous, S., & Sharma, K. (2022). A survey on diabetes risk prediction using machine learning approaches. Journal of Family Medicine and Primary Care, 11(11), 6929.
[5] Rajendra, P., & Latifi, S. (2021). Prediction of diabetes using logistic regression and ensemble techniques. Computer Methods and Programs in Biomedicine Update, 1, 100032.
[6] Iyer, A., S, J., & Sumbaly, R. (2015). Diagnosis of diabetes using classification mining techniques. International Journal of Data Mining & Knowledge Management Process, 5(1), 01-14.
[7] Mercaldo, F., Nardone, V., & Santone, A. (2017). Diabetes mellitus affected patients classification and diagnosis through machine learning techniques. Procedia Computer Science, 112, 2519-2528.
[8] Sisodia, D., & Sisodia, D. S. (2018). Prediction of diabetes using classification algorithms. Procedia Computer Science, 132, 1578-1585.
[9] Maniruzzaman, M., Rahman, M. J., Al-MehediHasan, M., Suri, H. S., Abedin, M. M., El-Baz, A., & Suri, J. S. (2018). Accurate diabetes risk stratification using machine learning: Role of missing value and outliers. Journal of Medical Systems, 42(5).
[10] Sneha, N., & Gangil, T. (2019). Analysis of diabetes mellitus for early prediction using optimal features selection. Journal of Big Data, 6(1).
[11] Lukmanto, R. B., Suharjito, Nugroho, A., & Akbar, H. (2019). Early detection of diabetes mellitus using feature selection and fuzzy support vector machine. Procedia Computer Science, 157, 46-54.
[12] Ahuja, R., Dixit, P., Banga, A., & Sharma, S. C. (2021). Classification algorithms for predicting diabetes mellitus: A comparative analysis. Pervasive Healthcare, 233-253.
[13] Morgan-Benita, J. A., Galván-Tejada, C. E., Cruz, M., Galván-Tejada, J. I., Gamboa-Rosales, H., Arceo-Olague, J. G., Luna-García, H., & Celaya-Padilla, J. M. (2022). Hard voting ensemble approach for the detection of type 2 diabetes in Mexican population with non-glucose related features. Healthcare, 10(8), 1362.
[14] David W. Hosmer, J., & Lemeshow, S. (2004). Applied logistic regression. John Wiley & Sons.
[15] Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273-297.
[16] Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 13(1), 21-27.
[17] Rish, I. (2001). An empirical study of the naive Bayes classifier. IJCAI 2001 workshop on empirical methods in artificial intelligence (p./pp. 41--46).
[18] Ensembles: Gradient boosting, random forests, bagging, voting, stacking. (n.d.). scikit-learn. Retrieved February 22, 2024, from
[19] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning: With applications in R. Springer Science & Business Media.
[20] XGBoost documentation — xgboost 2.1.0-dev documentation. (2022). XGBoost Documentation — xgboost 2.0.0 documentation.
[21] Pima Indians Diabetes Database. (2023, July 29). Kaggle.