SCIENTIFIC OASIS

Journal of Operations Intelligence

Journal homepage: www.jopi-journal.org
eISSN: 3009-4267

A Comparative Analysis of the Machine Learning Methods for Predicting Diabetes

Mohammad Maydanchi[1,*], Mehrbod Ziaei[2], Mehrdad Mohammadi[1], Armin Ziaei[3], Mina Basiri[4], Fatemeh Haji[5], Kazhal Gharibi[1]

[1] Department of Industrial and Systems Engineering, Auburn University, AL, USA
[2] Department of Electrical Engineering, University of Science and Culture, Tehran, Iran
[3] Department of Engineering and Computer Science, The University of Texas at Dallas, TX, USA
[4] Department of Industrial and Systems Engineering, Tarbiat Modares University, Tehran, Iran
[5] Department of Computer Engineering, Ferdowsi University of Mashhad, Mashhad, Iran

ARTICLE INFO

Article history:
Received 27 February 2024
Received in revised form 5 April 2024
Accepted 15 May 2024
Available online 18 May 2024

Keywords:
Diabetes Prediction; Machine Learning; Classification Models; Healthcare; Comparative Analysis.

ABSTRACT

Diabetes can lead to various health problems and complications, such as cardiovascular disease, kidney damage (nephropathy), eye issues, neuropathy, and foot ailments. Therefore, early diagnosis of diabetes can be immensely beneficial in preventing the development of these conditions. Utilizing machine learning is one method to detect diabetes in individuals at an early stage. In this study, we compare the performance of nine machine learning classification models in predicting diabetes. These models include XGBoost, gradient boosting, AdaBoost, logistic regression, decision tree, KNN, perceptron, random forest, and naïve Bayes. We utilize several evaluation metrics, focusing on the F1-score, area under the curve (AUC), and computational runtime. Our comparison reveals that complex tree-based models exhibit the highest F1-score and AUC, albeit with longer execution times.

1. Introduction

Diabetes is a persistent health condition characterized by its chronic nature, impacting the process through which the body converts food into energy, giving rise to life-threatening, debilitating, and expensive complications while diminishing overall life expectancy. In individuals with diabetes, the body either produces insufficient insulin or struggles to utilize it effectively. When insulin is deficient or cells fail to respond appropriately, excess blood sugar remains in the bloodstream. This prolonged elevation of blood sugar levels can lead to severe health complications, including but not limited to heart disease, vision impairment, and kidney disease [1].

Diabetes manifests in two primary types: Type 1 diabetes manifests through an autoimmune reaction, where the body targets itself, disrupting insulin production. This form of diabetes,

* *Corresponding author.*
E-mail address: mzm0181@auburn.edu

https://doi.org/10.31181/jopi21202421

comprising 5-10% of cases, is believed to result from this self-directed immune response. In contrast, Type 2 diabetes emerges primarily from insulin resistance or inadequate insulin secretion. Lifestyle factors such as dietary choices and physical activity levels heavily influence its development. Type 2 diabetes, the most prevalent form, accounts for approximately 90-95% of all diabetes diagnoses [2].

Approximately 537 million adults aged 20-79 grapple with diabetes worldwide, equating to 1 in 10 individuals. Projections indicate a rise to 643 million by 2030 and a further increase to 783 million by 2045. Over three-quarters of adults with diabetes reside in low-income and middle-income countries. In 2021, diabetes caused 6.7 million deaths, equating to one every 5 seconds. Additionally, diabetes has a significant economic impact, with health expenditures totaling at least USD 966 billion, representing a 316% increase over the past 15 years [3].

According to data from the American Diabetes Association [4], in 2021, diabetes affected 38.4 million individuals in the United States, comprising 11.6% of the total population. Among these, 29.7 million received a formal diagnosis. At the same time, 8.7 million remained undiagnosed, as shown in Figure 1, where the X-axis is the time period. The Y-axis is an age-adjusted percentage of total diabetes; diagnosed diabetes has a positive slope from 2001 to 2020, and undiagnosed diabetes has a constant slope throughout these years [4]. Approximately 1.2 million Americans are annually diagnosed with diabetes. Furthermore, in 2021, a staggering 97.6 million individuals aged 18 and older in the United States exhibited signs of prediabetes.

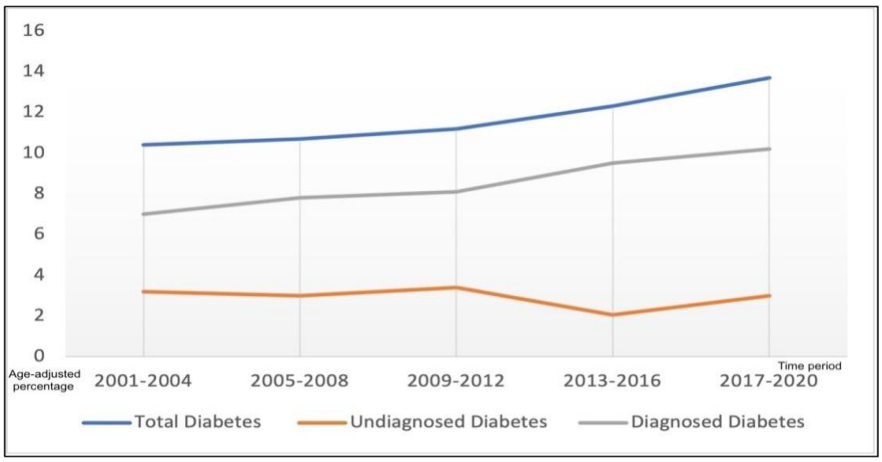

Fig. 1. Diabetes diagnoses statistics

Early diagnosis of diabetes, along with analysis of its causes and its relationship with health factors, is of paramount importance. One of the methods to achieve this goal is using artificial intelligence techniques to predict diabetes. Sun *et al.,* [5] employed machine learning (ML) models and encryption methods to jointly analyze vertically partitioned data. They aimed to explore the relationship between type 2 diabetes mellitus (T2DM) status and annual healthcare costs across different healthcare categories. They proposed an extension of the personal health train (PHT) architecture. This design empowers researchers to deploy data processing algorithms (application trains) directly to individual data sources, eliminating the need to centralize the data required for analysis.

The global burden of diabetes (GBD) 2021 conducted estimates of diabetes prevalence and impact across 204 countries and territories, spanning 25 age groups for both males and females individually and collectively. These estimates encompassed the burden of lost years in terms of healthy life, quantified as disability-adjusted life-years (DALYs, comprising years of life lost (YLLs) and years lived with disability (YLDs)). The death ensemble model approach calculates diabetes-related

deaths [6]. They employed a methodical approach, conducting systematic reviews and opportunistic searches. Data contributed by country collaborators and the world health organization (WHO) are also utilized in this process. In addition, Jamali *et al.,* [7] present a dynamic approach, where their examination of limitations underscores the significance of offloading computation within the confines of time constraints.

In conducting this cohort analysis, Schön *et al.,* [8] utilized a sample of 927 participants, aged 18–69 years, from the German diabetes study (GDS) who had recently developed Type 2 diabetes. These participants were positioned within a previously established two-dimensional tree based on nine straightforward clinical and laboratory variables, adjusted for age and sex. The assessment of insulin sensitivity was performed using a hyper-insulinaemic-euglycaemic clamp.

The growing accessibility of data and ML tools has led to more researchers developing datasets and models to aid in analyzing and alleviating the crisis, which is evaluated by implementing different ML methods for the problem of opioid disorder [9]. Gaudelet *et al.,* [10] developed graph machine learning (GML), which introduces a new classification of ML techniques that exploit the structural characteristics inherent in graphs and other irregular datasets, enabling the acquisition of effective feature representations.

Mbunge *et al.,* [11] implemented various sophisticated ML algorithms such as logistic regression, artificial neural networks (ANN), support vector machines (SVM), random forest, decision trees, Adaboost, bagging, and XGBoost have been employed to address diverse diseases.

In another usage, Ibrahim *et al.,* [12] utilized COVID-19 datasets from Morocco, Sudan, Uganda, Rwanda, Cameroon, Gabon, South Africa, Namibia, Nigeria, and Senegal; predictions for the COVID19 pandemic were made. In Burkina Faso, random forest regressors and gaussian processes were applied to forecast malaria epidemics.

Soltaninejad *et al.,* [13] developed a holistic framework for sales prediction by combining technical analysis, time series modeling, ML, neural networks, and random forest techniques into an integrated approach.

Artificial intelligence systems have the potential to aid healthcare professionals and public health institutions in enhancing healthcare service delivery by facilitating early disease detection, prediction, diagnosis, identification of diseases, and mapping hotspot areas, as suggested by Mbunge *et al.,* [11] and by Shill et al., [14] in the process of developing new learning methods.

In addition, Abubakar *et al.,* [15] implemented a range of deep learning models, including VGG16, VGG19, ResNet50, ResNet101, ResNet152, DenseNet121, DenseNet201, AlexNet, and Xception, the researchers sought to detect malaria in images of blood smears. Their investigation involved extracting discriminative features from the dataset images and employing machine-learning algorithms to categorize each sample, discerning whether a patient was infected based on a given blood sample.

Maydanchi *et al.,* [16] conducted screening for risk factors, which expedites the identification of cardiovascular diseases (CVD), leading to a swifter and more efficient response, thereby mitigating the risk of mortality. This paper evaluates six classification models such as AdaBoost, random forest, decision tree, KNN, naïve bayes, and perceptron, in forecasting CVD symptoms.

Haseli *et al.,* [17,18] employed a multi-faceted approach in their decision-making process, emphasizing the need to consider numerous factors and analyze results from diverse viewpoints. This underscores the complexity inherent in tackling multi-criteria decision-making problems.

Hennebelle *et al.,* [19] proposed an ML-based smart healthcare framework within an integrated IoT-edge-cloud computing system. This system analyzes diabetes risk factors using medical sensors/devices and predicts the incidence of Type 2 diabetes in individuals.

Toscano-Pulido *et al.,* [20] Introduce a tailored hybrid evolutionary multi-objective optimization (EMO) algorithm to improve quality within the Chesapeake Bay Watershed. This algorithm selects cost-effective best management practices (BMPs) to mitigate pollution levels, tackling the complexities of large-scale optimization.

MacKay *et al.,* [21] highlight that ML is frequently employed to identify patterns and trends within extensive sets of predictor variables, which would be challenging for researchers to discern otherwise. The authors emphasize that restricting data sets to only a handful of predetermined variables may impede the full realization of the potential of ML tools in healthcare.

Considering the mentioned usage of ML in health care, we can narrow it down to find the literature on Diabetes.

Joshi *et al.,* [22] introduced a study on diabetes prediction using ML techniques to forecast diabetes using three distinct supervised ML approaches: ANN, logistic regression, and SVM. In another paper, the methodology of Modak *et al.,* [23] undergoes thorough scrutiny via performance evaluation metrics such as the confusion matrix, sensitivity, and accuracy measurements. CatBoost is the most efficacious among the ensemble techniques assessed, demonstrating a notable accuracy rate.

An approach combining semi-supervised learning with extreme gradient boosting was deployed by Tasin *et al.,* [24] to forecast insulin characteristics within a confidential dataset. Methodologies like SMOTE and ADASYN were employed to address the issue of class imbalance. The researchers explored various ML classification techniques, including decision trees, SVM, random forest, logistic regression, k-nearest neighbors (KNN), and assorted ensemble strategies to identify the optimal algorithm for prediction. Following extensive training and testing of all classification models, the proposed system demonstrated its superior performance with the XGBoost classifier utilizing the ADASYN methodology, achieving an accuracy rate of 81%, an f1-score of 0.81, and an AUC of 0.84. Additionally, a domain adaptation approach was implemented to underscore the adaptability and versatility of the proposed system.

Moreover, Mujumdar *et al.,* [25] found that their previous method's classification and prediction accuracy were unsatisfactory. In their paper, they propose a new diabetes prediction model aiming to enhance the classification of diabetes. This model integrates additional external factors associated with diabetes and standard factors like Glucose, BMI, Age, and Insulin. Including these new factors substantially improves classification accuracy compared to their previous dataset. Moreover, they introduce a pipeline model for diabetes prediction to enhance classification accuracy further.

Lai *et al.,* [26] constructed predictive models employing gradient boosting machine (GBM) and logistic regression techniques. The effectiveness of these models in discrimination was evaluated using the AUC of the receiver operating characteristics (ROC). Sensitivity, indicating the accuracy in identifying Diabetes Mellitus patients, was refined using the adjusted threshold method and the class weight method. Furthermore, comparisons were drawn with alternative learning machine methods, including decision trees and random forests.

The study by Sarwar *et al.,* [27] explores predictive analytics in healthcare, employing six distinct ML algorithms. A dataset comprising patient medical records is acquired for experimental purposes, with each algorithm applied to the dataset. The performance and accuracy of these algorithms are thoroughly examined and juxtaposed. By comparing various ML techniques utilized in this investigation, insights are gained into which algorithm proves most effective for diabetes prediction. The primary objective of this paper is to assist healthcare professionals and practitioners in the early detection of diabetes through the utilization of ML techniques.

Saru and Subashree [28] utilized data mining techniques to explore and identify suitable approaches for the efficient classification of a diabetes dataset and the extraction of valuable patterns. This study conducted medical bioinformatics analyses to predict diabetes. The WEKA software served as the mining tool for diabetes diagnosis. The Pima Indian diabetes database obtained from the UCI repository was employed for analysis. The dataset was thoroughly examined to develop an effective model for predicting and diagnosing diabetes.

Our study is unique in the recent diabetes dataset as the sole comprehensive comparison of nine ML models for predicting diabetes. This aspect makes our work significant by filling a gap in existing literature and offering readers a broad perspective on these models' performance. The objective of our study is to compare the performance of these models, including XGBoost, gradient boosting, AdaBoost, logistic regression, decision tree, KNN, perceptron, random forest, and naïve bayes, in predicting diabetes. In Section 2, we discuss the dataset, data preparation, and the classification methods utilized in this study. Section 3 presents the results of evaluation metrics. Finally, Section 4 outlines the conclusions drawn from this study.

2. Methodology

The research draws upon the diabetes prediction dataset obtained from Kaggle. This dataset represents a comprehensive compilation of medical and demographic attributes of patients, meticulously curated alongside their corresponding diabetes status, classified as either positive or negative. Through its rich repository of patient data, encompassing vital health indicators and personal information, the dataset serves as a valuable resource for investigating predictive models and advancing our understanding of diabetes diagnosis.

The dataset contains data from 100,000 individuals designed for the prediction of diabetes. The dataset consists of both categorical and continuous variables. Table 1 presents these variables. The categorical variables include gender, hypertension, heart disease, and smoking history. The target variable, diabetes, is also a categorical variable. Patients are categorized into two classes: class 0 for those without diabetes and class 1 for those with diabetes. In addition, age, BMI (Body Mass Index, indicating the level of body fat), HbA1c (meaning the average level of blood sugar during the past two to three months), and blood glucose are variables that are continuous.

Table 1
Categorical and continuous variables

Feature	Type	Range of Value
Gender	Categorical	Male/ Female
Age	Continuous	Float [0.08, 80]
Hypertension	Categorical	0/ 1
Heart Disease	Categorical	0/ 1
Smoking History	Categorical	never/ No info/ current/former/ ever/ not current
BMI	Continuous	Float [10, 95.7]
HbA1c level	Continuous	Float [3.5, 9]
Blood Glucose Level	Continuous	Integer [80, 300]
Diabetes	Categorical	0/ 1

Exploratory data analysis (EDA) analyzes datasets through statistical and visual methods to understand their structure, patterns, and potential relationships between variables. This article has two types of analysis: Univariate analysis and bivariate analysis. In the following, we will discuss them.

Figure 2 presents a count chart showing the distribution of a population by gender, revealing a predominant proportion of females at 58.41%, compared to 41.57% of males, with a marginal representation of non-binary or other identified genders at 0.02%.

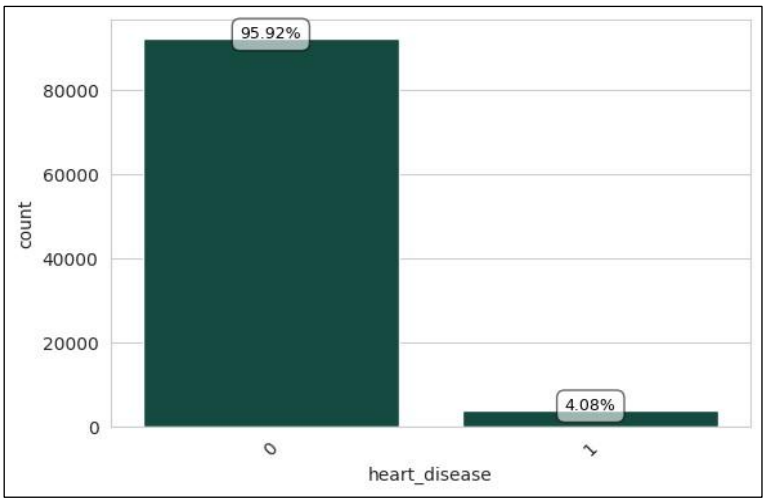

Fig.2. Distribution of a population by gender

Figure 3 provides a stark visual representation of the prevalence of hypertension within the surveyed cohort. A commanding majority of the population, constituting 92.24%, is denoted as not having hypertension, while a minority of 7.76% identified as hypertensive.

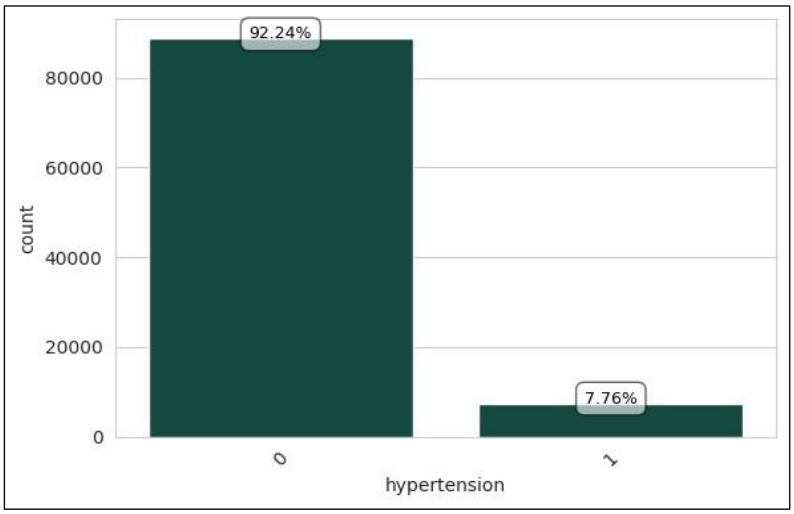

Fig.3. Prevalence of hypertension

Figure 4 illustrates the distribution of heart disease within the study's demographic, where an overwhelming 95.92% of participants do not have heart disease, differing sharply from the 4.08% who do.

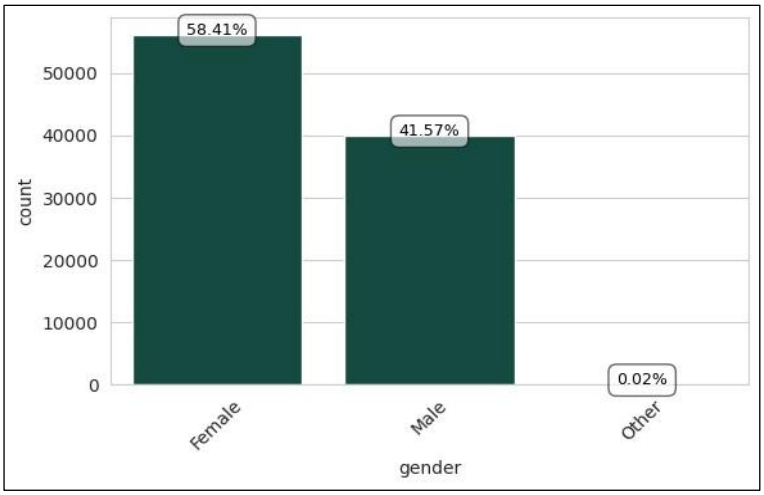

Fig.4. Distribution of heart disease

Figure 5 delineates the smoking history profiles within the examined sample, providing a multifaceted perspective on smoking behaviors. The largest group, constituting 35.78%, reports never having smoked, closely followed by 34.21%, for which there is no information regarding their smoking habits. Active smokers account for 9.57% of the population. In comparison, former smokers comprise a comparable 9.67%, indicating a balanced distribution between current and past smokers within the cohort. Notably, 4.16% are identified as 'ever' smokers, a category that likely encapsulates individuals who have smoked infrequently or sporadically. Lastly, the category 'not current' smokers, representing 6.62%, may include individuals who have ceased smoking recently and have not yet classified themselves as 'former' smokers.

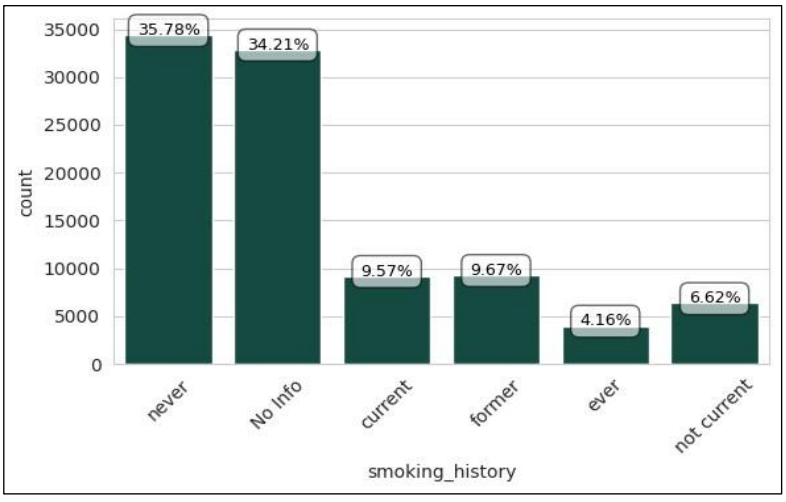

Fig.5. Distribution of smoking history

Figure 6, central to our investigation, delineates the distribution of diabetes within the studied cohort, serving as the target feature. The graphical representation illustrates a pronounced

imbalance, with 91.18% of the sample population not being affected by diabetes. In comparison, a minority of 8.82% have been diagnosed with the condition.

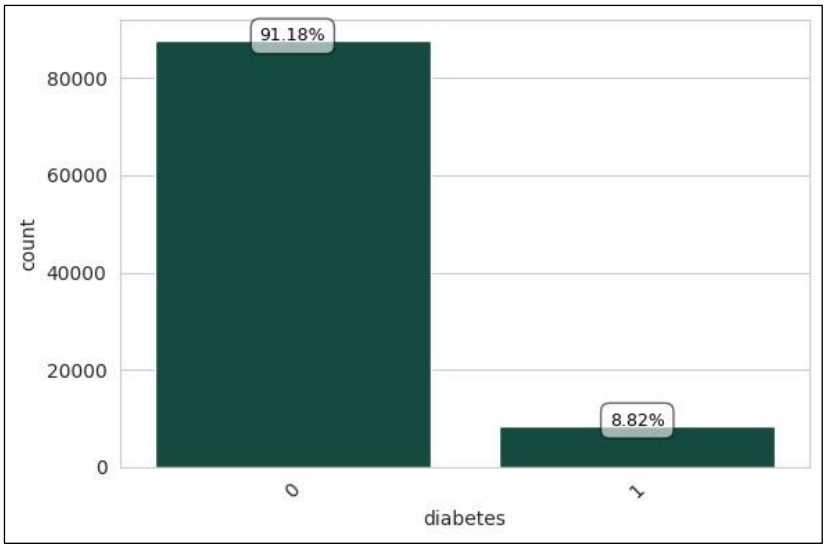

Fig.6. Distribution of diabetes

Figure 7 presents a dual-plot graphical analysis of the age distribution within the dataset. The left plot showcases a histogram overlaid with a kernel density estimate, providing a visual summary of the age distribution among the study participants. The histogram's bars represent the count of individuals within specific age ranges, revealing the data's skewness or symmetry. The mean age is marked by a dashed red line and annotated at 41.79 years, with a standard deviation of 22.46 years, implying a wide age range among the subjects. The kernel density curve accentuates the overall distribution shape, indicating variability and potential outliers in age. The right plot features a box plot, a robust representation of age dispersion through quartiles. The box encloses the interquartile range (IQR), which spans 35 years, reflecting the middle 50% of the data, centered around a median age of 43.

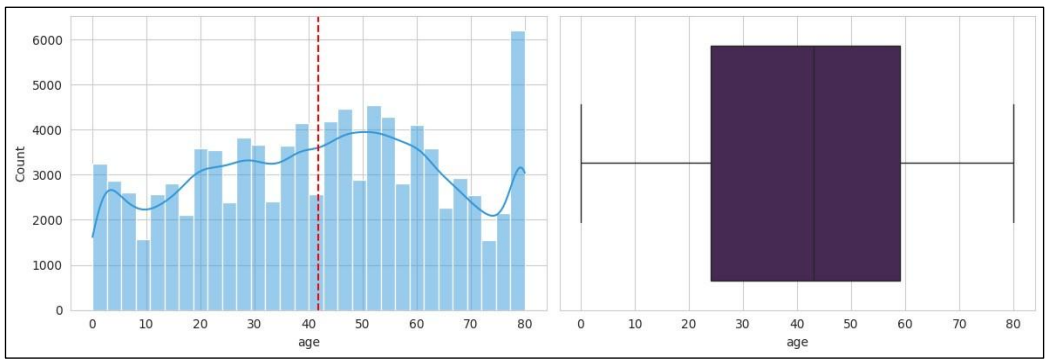

Fig.7. Histogram and boxplot for age

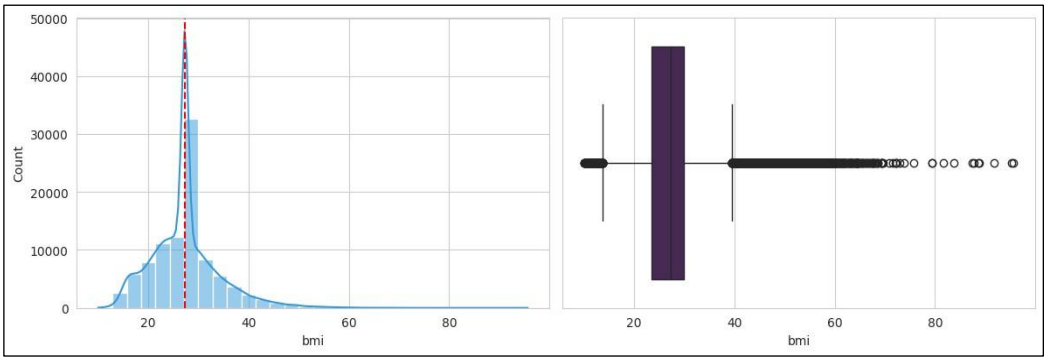

Fig.8. Histogram and boxplot of BM

In Figure 8, we compare BMI distribution among participants using a histogram and a box plot. This comparison is essential for predicting diabetes because BMI is known to be a risk factor for the condition. On the left, the histogram reveals the frequency of various BMI scores, shaded in blue, with a corresponding density curve that outlines the distribution's shape. The mean BMI is pinpointed at 27.32 with a standard deviation of 6.77, highlighted by a dashed red line, suggesting a moderate average BMI but with considerable spread across the population. This spread indicates the population's weight status variability, from underweight to obese categories. To the right, th box plot provides a more detailed view of the BMI distribution's quartiles. The median BMI, identical to the mean, is 27.32, indicating a symmetric distribution of BMI values around the central tendency. An IQR of 6.46 suggests that the middle 50% of BMI values are relatively clustered.

Figure 9 offers a composite view of the distribution of Hemoglobin A1c (HbA1c) levels in the dataset, a critical biomarker for diabetes management and diagnosis. The histogram on the left panel illustrates the frequency distribution of HbA1c levels across the study population, overlayed with a kernel density curve that emphasizes the data's distribution pattern. The mean HbA1c level is at 5.53%, as the dashed red line indicates, with a standard deviation of 1.07%. The right panel features a box plot that delineates the quartile distribution of HbA1c levels, where the median is slightly higher than the mean at 5.80%, hinting at a slight right skew in the data. The IQR is relatively narrow at 1.40%, signifying that the central 50% of HbA1c values are clustered closely around the median. However, outliers are present, as shown by the points beyond the whiskers.

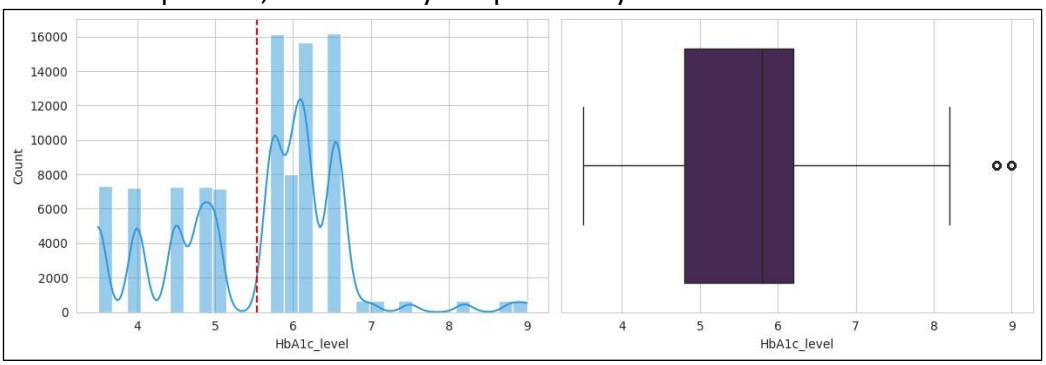

Fig.9. Histogram and boxplot of Hemoglobin A1c

Figure 10 provides a close look at the distribution of blood glucose levels, a vital indicator for diagnosing and monitoring diabetes. With its overlaying kernel density estimation, the histogram on the left side visualizes the distribution of blood glucose levels among participants. The graph exhibits several peaks, indicating possible groupings or ranges of values within the data. The mean value is 138.22 mg/dL, delineated by a dashed red line and a 40.91 mg/dL standard deviation. This relatively

high mean, in conjunction with the observed variability, might suggest a skew towards higher glucose levels within this particular cohort, which could be indicative of a prevalence of prediabetes or undiagnosed diabetes. In contrast, the box plot on the right offers a succinct summary of the data's dispersion via quartiles. The median is slightly higher than the mean at 140.00 mg/dL, reinforcing the skew observed in the histogram. An IQR of 59.00 mg/dL indicates a broad middle range of glucose values, and the presence of outliers, as evidenced by the points beyond the whiskers, suggests significant deviations from typical glucose levels in some participants.

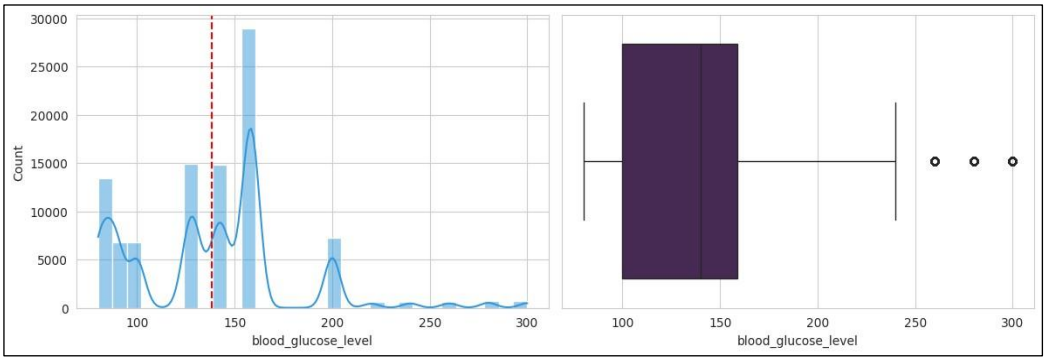

Fig.10. Histogram and boxplot of blood glucose level

Figure 11 displays a bar chart that quantifies the incidence of diabetes across gender divisions within a dataset. It depicts 51,714 females without diabetes and 4,447 with the condition. The male population shows 35,932 individuals without diabetes and 4,035 with it. The 'Other' gender category is represented by a nominal count of 18 individuals with diabetes.

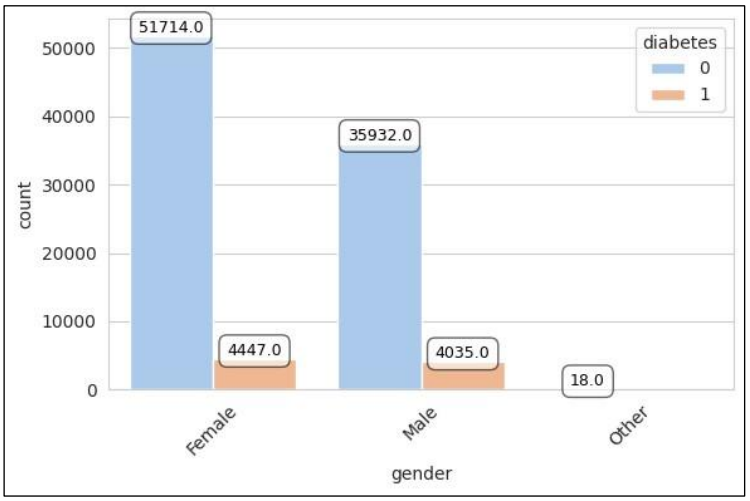

Fig.11. Diabetes incidence across genders: bar chart analysis

Figure 12 presents a bar chart that explores the relationship between hypertension and diabetes within a given population. The chart shows that among individuals without hypertension, 82,289 do not have diabetes, and 6,396 do. Conversely, of those with hypertension, 5,375 are not diabetic, and a significant number, 2,086, have diabetes. This notable difference in the diabetic count between those with and without hypertension may suggest a correlation between hypertension and the prevalence of diabetes.

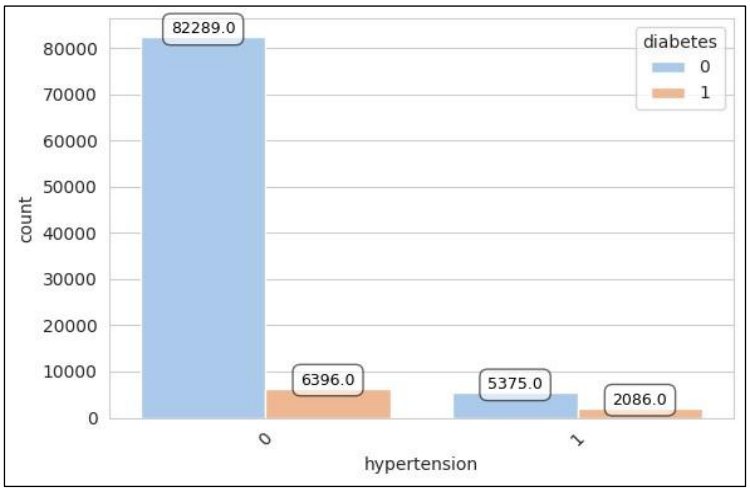

Fig.12. Hypertension and diabetes: bar chart analysis

Figure 13 shows a bar chart that correlates the occurrence of heart disease with the prevalence of diabetes. Among individuals without heart disease, 85,008 do not have diabetes, whereas 7,215 do, indicating a lower relative prevalence of diabetes within this group. In stark contrast, the group with heart disease has a smaller non-diabetic count of 2,656 but a notably higher diabetic count of 1,267, suggesting a substantial prevalence of diabetes among those with heart disease.

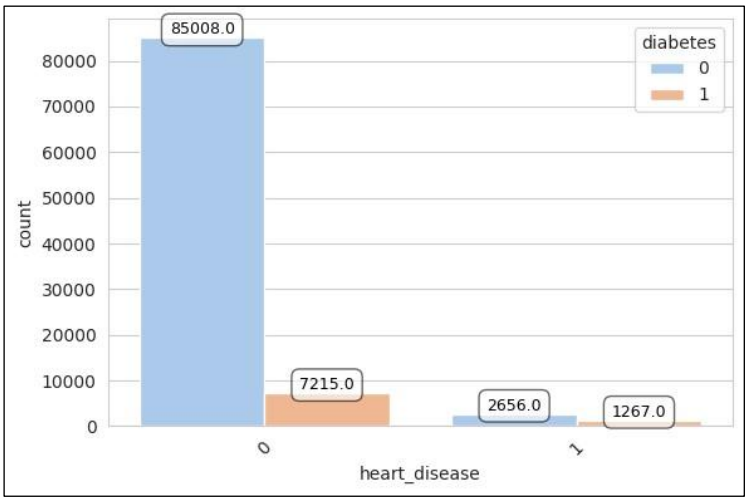

Fig.13. Heart disease and diabetes: bar chart analysis

Figure 14 conveys a detailed bar chart that delineates the relationship between individuals' smoking history and diabetes status. The 'Never' smokers group is the largest among those without diabetes, totaling 31,061, and includes 3,337 individuals with diabetes. Similarly, for the 'No info' category, there are 31,442 without diabetes and a smaller fraction of 1,445 with it. The 'Current' smokers are fewer, with 8,249 not having diabetes and 948 with the condition. 'Former' smokers show a count of 7,709 without diabetes and a higher proportion of 1,590 with diabetes compared to 'Current' smokers. Those categorized as 'Ever' smokers have 3,526 without diabetes and 472 with it. Lastly, the 'Not current' smokers, possibly indicating those who have quit recently, include 5,677 without diabetes and 690 with the condition.

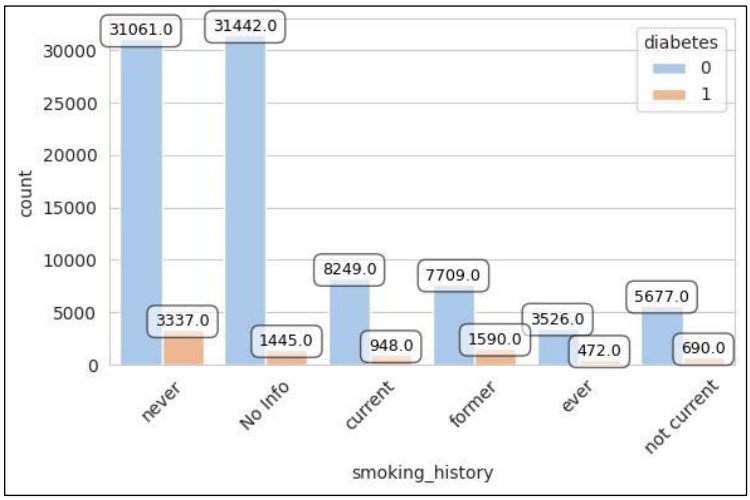

Fig.14. Smoking history and diabetes: bar chart analysis

Figure 15 pairs a histogram with a box plot to compare the age distribution of a population segmented by diabetes status. The histogram on the left indicates a bimodal age distribution with two peaks, one for the younger demographic and another for an older age group. It shows that the mean age of individuals without diabetes class 0 is 39.94 years, with a standard deviation of 22.23 years. In contrast, the mean age for those with diabetes class 1 is significantly higher at 60.93 years, with a narrower standard deviation of 14.55 years, suggesting less variability in age among diabetic individuals. The box plot on the right further explores this age disparity. For non-diabetic individuals, the median age is 40 with an IQR of 35 years, indicating a wide spread of ages. For those with diabetes, the median age jumps to 62 years with a tighter IQR of 20 years, reaffirming the histogram's indication of a more concentrated age range within the diabetic segment of the population.

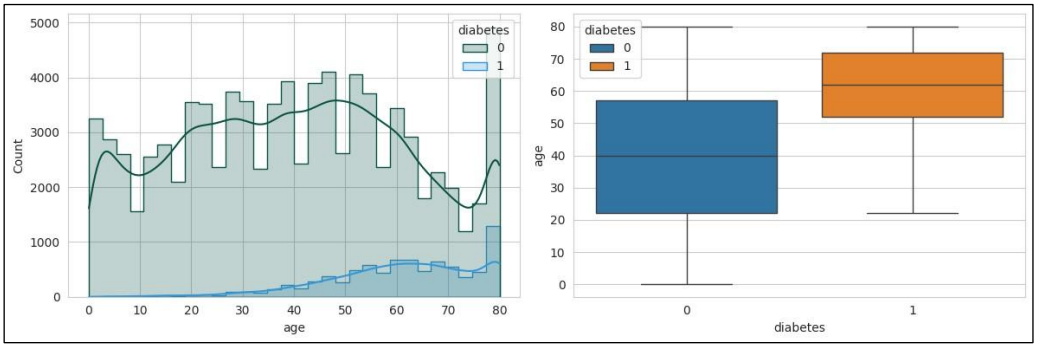

Fig.15. Comparing age distribution by diabetes status: histogram and box plot analysis

Figure 16 combines a histogram and a box plot to examine the BMI distributions relative to diabetes status within a population. The histogram on the left displays a sharp peak for the non-diabetic group (class 0) with a mean BMI of 26.87 and a standard deviation of 6.51, indicating a moderate average BMI with some variability. For the diabetic group (class 1), the mean BMI is higher at 32.00, with a standard deviation of 7.56, suggesting that individuals with diabetes tend to have an average BMI.

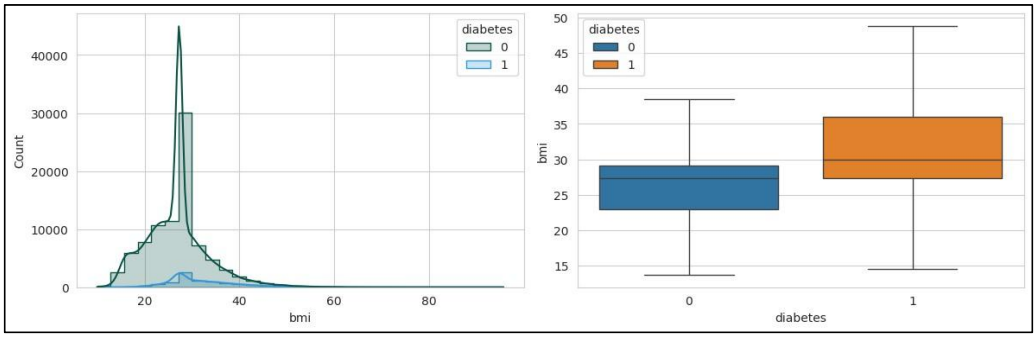

Fig.16. Examining BMI distributions by diabetes status: histogram and box plot analysis

The box plot on the right summarizes the BMI data, segmented by diabetes status. For the nondiabetic group, the median BMI is 27.32, with an IQR of 6.18, showing a relatively normal distribution of BMI values. In contrast, the diabetic group's median BMI increases to 29.98 with a broader IQR of 8.62, indicating a heavier distribution towards higher BMI values.

Figure 17 provides an insightful look at the HbA1c levels stratified by diabetes status within a study population. The histogram to the left highlights a stark contrast in HbA1c levels between individuals without diabetes (class 0) and those with the condition (class 1). For the non-diabetic group, the mean HbA1c level is relatively low at 5.40, with a standard deviation of 0.97, portraying a concentration of lower HbA1c values. On the other hand, the diabetic group exhibits a higher mean HbA1c level of 6.93 with a standard deviation of 1.08, indicating a shift towards higher HbA1c values typically associated with diabetes.

The box plot on the right echoes these findings, showing a median HbA1c level of 5.80 for nondiabetics with an IQR of 1.40, compared to a significantly higher median of 6.60 for diabetics with an IQR of 1.40.

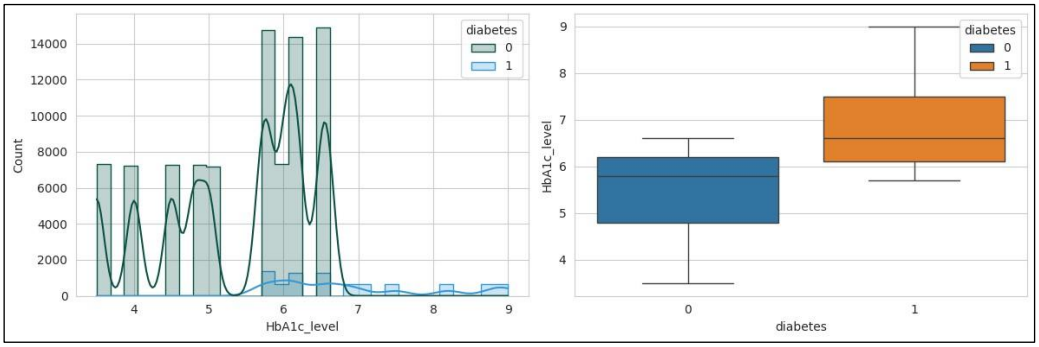

Fig.17. HbA1c levels stratified by diabetes status: histogram and box plot analysis

Figure 18 shows a histogram and a box plot to analyze the distribution of blood glucose levels concerning diabetes status. The histogram reveals a clear distinction: individuals without diabetes (class 0) have a mean blood glucose level of 132.82 mg/dL with a standard deviation of 34.24 mg/dL, suggesting moderate glucose levels with some spread. For those diagnosed with diabetes (class 1), the mean level escalates to 194.03 mg/dL and exhibits more significant variability, as indicated by a 58.63 mg/dL standard deviation.

The box plot on the right accentuates this differentiation, showing a median blood glucose level of 140.00 mg/dL for non-diabetics, with an IQR of 58.00 mg/dL. The diabetic group's median is significantly higher, at 160.00 mg/dL, and the IQR widens to 95.00 mg/dL, signifying a broader dispersion of values, which is typical given the nature of the disease.

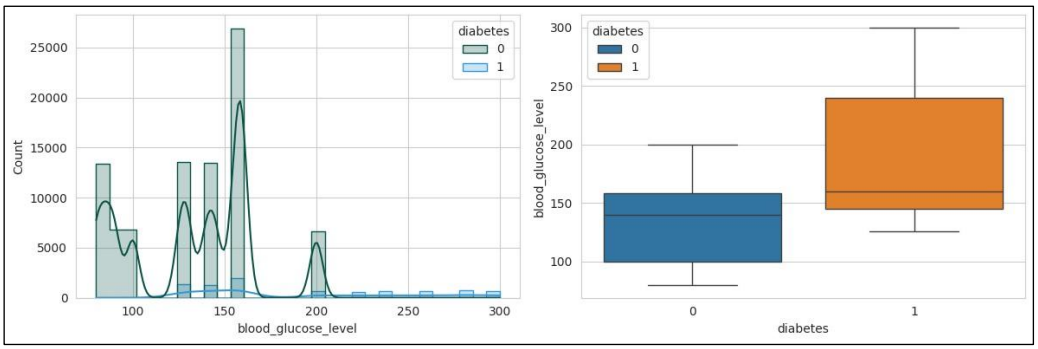

Fig.18. Blood glucose levels distribution by diabetes status: histogram and box plot analysis

Figure 19 shows the correlations among features. The highest positive correlations identified are between *diabetes* and *blood glucose level* at 0.42 and *diabetes* and *HbA1c level* at 0.4, which are considered moderate. A positive correlation of 0.34 is noted between *age* and *BMI*. In contrast, 0.26, 0.25, 0.23, and 0.22 correlations are observed between *age* and *diabetes*, *age* and *hypertension*, *age* and *heart disease*, and *age* and former *smoking history*, respectively. Furthermore, *BMI* and *diabetes* show a positive correlation of 0.21, and *hypertension* and *diabetes* correlate at 0.2, all classified as weak positive correlations. Moderate positive correlations are observed within the dataset, notably between *diabetes* and *blood glucose level* at 0.42 and *diabetes* and *HbA1c level* at 0.4. Weak positive correlations are noted, including *age* and *BMI* at 0.34, *age* and *diabetes* at 0.26, *age* and *hypertension* at 0.25, *age* and *heart disease* at 0.23, *age* and *former smoking history* at 0.22, *BMI* and *diabetes* at 0.21, and *hypertension* and *diabetes* at 0.2.

Negligible correlations are observed for *gender other* with various health metrics: a slight positive correlation with *blood glucose level* at 0.00046 and *BMI* at 0.00012; and minor negative correlations with *smoking history No Info* at -0.00069, *smoking history ever* at -0.0011, *HbA1c level* at -0.0015, *heart disease* at -0.0027, *hypertension* at -0.0038, *diabetes* at -0.0041, and both former and *current smoking history* at -0.0043, all indicating a lack of significant linear relationship. Weak negative correlations include *smoking history No Info* and *age* at -0.28, *smoking history No Info* with *smoking history former* and *smoking history current*, both at -0.24, as well as between *smoking history never* and *smoking history former*, and *smoking history current* with *smoking history never*, each at -0.24. Additionally, Figure 19 shows that smoking history *No Info* and *BMI* have a correlation of -0.22, with a similar pattern observed between smoking history *No Info* and current smoking history at -0.2. A moderate negative correlation is noted between smoking history *No Info* and never having smoked at -0.55.

In preprocessing the dataset, several steps were taken to ensure its quality and suitability for analysis. Firstly, it was observed that there were no null samples present, indicating that the dataset was relatively clean in this regard. However, a substantial number of rows were found to be duplicated, totaling 3854 duplicates. These duplicate rows were removed to avoid skewing the analysis and to ensure the dataset's integrity.

Next, the dataset was split using a 20-80 split, with 20% reserved for testing and 80% for training. Additionally, standard scaling was applied to the data to ensure that all features were on a similar scale, which can aid in the performance of specific ML algorithms.

The dataset was also examined for class imbalance, revealing that class 0 had 87664 samples before balancing. In comparison, class 1 had 8482 samples, indicating a moderate imbalance. This paper used the Synthetic Minority Over-sampling Technique (SMOTE) to deal with the data imbalance.

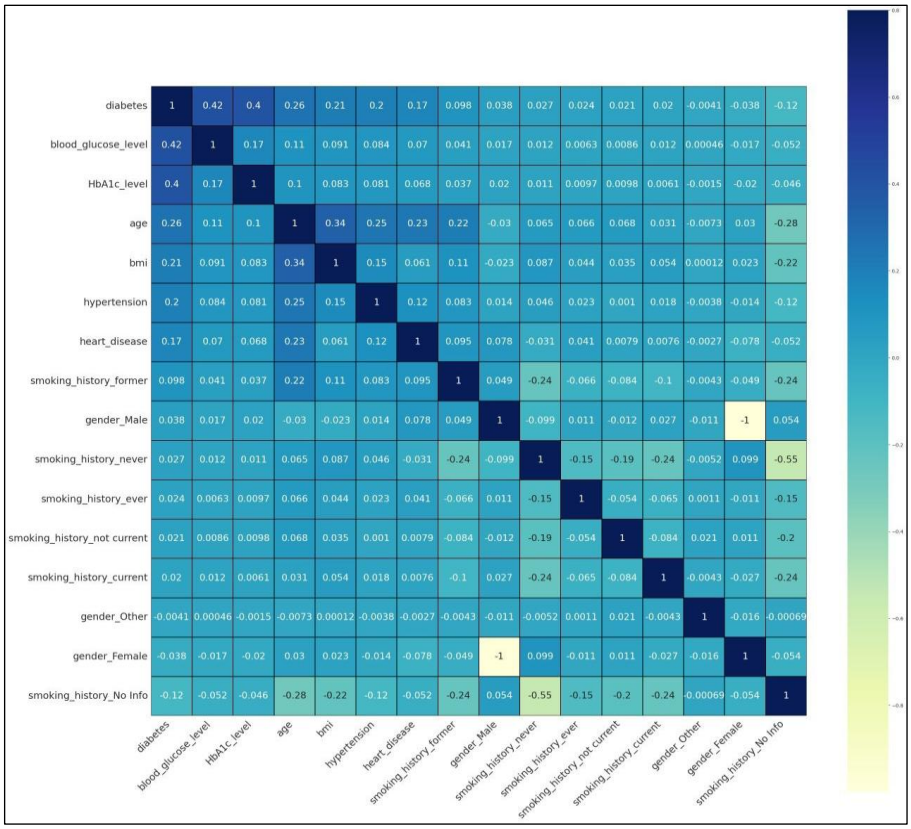

Fig.19. Correlation table of the variables

Furthermore, certain features underwent encoding for compatibility with ML algorithms. Hypertension and heart disease features were encoded as binary variables, while gender and smoking history were subjected to one-hot encoding. These encoding techniques help convert categorical variables into a format that ML models can interpret effectively. On the other hand, age, BMI, HbA1c, and blood glucose level features did not require encoding, likely due to their numerical nature and direct usability in analysis without transformation.

The classification ML methods compared in this paper are perceptron, KNN, naïve bayes, XGBoost, gradient boosting, AdaBoost, logistic regression, decision tree, KNN, perceptron, random forest, and naïve bayes.

The perceptron is a supervised learning algorithm specifically designed for binary classification tasks. It allows neurons to process and learn from elements within the training dataset [29]. The KNN algorithm, a versatile supervised ML technique, is well-known for addressing classification and regression issues [30]. Naïve Bayes works by estimating the probability of a post-test outcome using the values of predictive variables. It then assigns new samples to classes by comparing these probabilities [31]. Boosting algorithms seek to build robust predictive models by leveraging multiple weak learners in combination. This ensemble technique includes methods like AdaBoost, XGBoost, and gradient boost [32]. Random forest and decision tree are both supervised learning methods. Random forest applies a collection of decision trees to improve prediction accuracy [33].

On the other side, a decision tree model functions by iteratively dividing the dataset into subsets based on feature values, aiming to create more uniform subsets regarding the target variable [34]. Logistic regression, a cornerstone in predictive modeling, leverages the logistic function's mathematical principles to classify data into binary categories. By mirroring networks of interconnected nodes, it adeptly handles complex analyses, making informed predictions on binary target variables across diverse fields [35,36].

Each ML method has a specific set of parameters. The parameters were fine-tuned with *GridSearchCV* to optimize performance. In the XGBoost model, the *learning rate* is set to 0.2, which is critical in controlling the step size at each iteration to minimize the loss function. The *max depth* parameter is chosen as 7, determining the maximum depth of a tree, which is essential for capturing the complexity of data without overfitting. Additionally, *n estimators* is set to 300, indicating the number of trees in the model, where a higher number can improve performance but increases the risk of overfitting and computational cost.

For gradient boosting, the *learning rate* of 0.2 is crucial as it influences the rate at which the model learns, balancing speed and accuracy. The *max depth* is set to 5. The *n estimators* parameter, set at 300, is significant as it dictates the number of sequential trees built, affecting both model accuracy and complexity.

In AdaBoost, the *learning rate*, set at 1.0, is essential for adjusting the weights of the weak learners, determining how quickly the algorithm adapts. The *n estimators* parameter, set to 200, defines the number of weak learners to be used, impacting the model's ability to generalize.

For logistic regression, the C parameter, valued at 0.1, controls the degree of regularization, influencing the complexity and generalizability of the model. In order to further improve the model's robustness, the penalty type, L2, is critical for controlling overfitting. A linear solver, *liblinear*, is used to solve the problem. This solver is ideal for small datasets as well as binary classification problems.

In the decision tree model, the criterion of entropy is significant for determining the quality of a split, directly impacting the model's ability to capture relevant patterns in the data. The *min samples leaf* parameter, set at 2, is key for preventing the tree from growing too complex and overfitting. *Min samples split*, set at 5, helps decide when a node will be split into further sub-nodes, balancing detail and overfitting.

KNN relies on the Manhattan metric to define the distance measure between nearest neighbors, which significantly influences the model's performance. The *n neighbors* parameter, set to 3, determines the number of neighbors to consider, which is crucial for the classification decision. The weights parameter, set to distance, ensures that closer neighbors significantly influence the outcome, enhancing the model's accuracy.

In the perceptron model, the alpha value of 0.0001 is important as it regulates the learning rate, affecting the speed and quality of convergence. The *max iter* parameter, set at 1000, defines the maximum number of passes over the training data, balancing computation time and model accuracy. The L2 penalty helps prevent overfitting by adding a regularization term to the loss function.

For the random forest model, the *n estimator* parameter, set to 500, is crucial as it indicates the number of trees in the forest. The criterion of entropy is used to measure the quality of a split, influencing the decision-making of the individual trees. The *min samples split*, set at 2, is essential for defining the minimum number of samples required to split an internal node, affecting the depth and complexity of the trees.

In the naïve bayes model, *var smoothing*, set at 1e-6, is a crucial parameter as it adds a value to the variance of the distribution, helping to smooth categorical data and improve model performance.

Each parameter plays a crucial role in the respective models, affecting their ability to learn from the data and make accurate predictions. These parameters are shown in Table 2.

Table 2
Parameters of classification models

Model	Parameters
XGBoost	learning rate: 0.2, max depth: 7, n estimators: 300
Gradient Boosting	learning rate: 0.2, max depth: 5, n estimators: 300
AdaBoost	learning rate: 1, n estimators: 200
Logistic Regression	C: 0.1, penalty: l2, solver: liblinear
Decision Tree	criterion: entropy, min samples leaf: 2, min samples split: 5
KNN	metric: Manhattan, n neighbors: 3, weights: distance
Perceptron	alpha: 0.0001, max iter: 1000, penalty: l2
Random Forest	criterion: entropy, min samples split: 2, n estimators: 500
Naïve Bayes	var smoothing: 1e-6

3. Result

This section aims to comprehensively evaluate the performance of nine distinct supervised ML classification models. The models are A Comparison of Methods for Predicting Diabetes: XGBoost, gradient boosting, AdaBoost, logistic regression, decision tree, KNN, perceptron, random forest, naïve bayes. These models are specifically evaluated in their ability to predict the occurrence of diabetes classification. This analysis provides valuable insights into the application of ML methods in diabetes prediction.

In evaluating the performance of ML models, accuracy stands out as one of the most important metrics. This metric is a fundamental yardstick for assessing the model's overall ability to make correct predictions, providing a quantitative measure of its performance across various tasks and datasets. In this study, accuracy represents the number of samples that are correctly classified as either a person with diabetes or a person without diabetes overall.

Table 3 presents the accuracy of each model. This table shows that XGBoost and gradient boosting exhibit the highest accuracy at 97%, while perceptron has the lowest accuracy, standing at 85%.

In most literature comparing the prediction performance of classification models, accuracy is the primary metric. However, in a dataset with an imbalance issue, this metric is unreliable because the number of samples in one of the classes significantly exceeds the number in another, leading accuracy to primarily reflect the model's performance on predicting the majority class. As mentioned in the previous section, the majority group is class 0, with 87,664 samples, and the minority group is class 1, with 8,482 samples. This moderately imbalanced dataset necessitates using other metrics to compare the models.

In detecting individuals with heart disease, prioritizing classifying samples within class 1 over class 0 is essential. This is because giving precedence to predicting diabetes in individuals at risk holds greater importance than predicting its absence. Consequently, we assess the model's performance on class 1 using metrics such as precision, recall, and f1-score.

The precision for class 1 indicates the proportion of samples truly belonging to class 1, and the model correctly classifies them within class 1. The recall score for class 1 denotes the percentage of samples in class 1 that are accurately classified. As depicted in Table 3, gradient boosting exhibits the highest precision for class 1 at 93%, while perceptron demonstrates the lowest precision at 35%.

Additionally, logistic regression achieves the highest recall for class 1 at 89%, whereas naïve bayes records the lowest recall at 63%. Since each model may have the highest rank in either precision or recall metrics, making direct comparisons becomes challenging. Therefore, we use the f1-score, the harmonic mean of precision and recall.

Table 3 shows that gradient boosting, XGBoost, and AdaBoost exhibit the highest f1-scores at 81%, 80.5%, and 79%, respectively. These models demonstrate closely competitive performances, and notably, all belong to the family of tree-based models. The lowest F1-score rank is attributed to the perceptron at 49%.

Table 3
A comprehensive evaluation of accuracy, f1-score, precision, and recall metrics for class1 and class 0

Model	Accuracy	Class 0 (No)			Class 1 (Yes)		
		Precision	Recall	F1- score	Precision	Recall	F1-score
Gradient Boosting	97%	97%	100%	98%	93%	71%	81%
XGBoost	97%	97%	99%	98%	90%	73%	80%
AdaBoost	96%	98%	98%	98%	80%	78%	79%
Random Forest	96%	98%	98%	98%	77%	76%	77%
Decision Tree	95%	98%	97%	98%	74%	75%	75%
KNN	93%	98%	94%	96%	57%	78%	66%
Naïve Bayes	94%	96%	97%	97%	68%	63%	65%
Logistic Regression	89%	99%	89%	93%	43%	89%	58%
Perceptron	85%	98%	85%	91%	35%	83%	49%

The ROC curve is another metric to evaluate and compare ML prediction models. The x-axis of the curve represents the false positive rate. This rate corresponds to the ratio of samples incorrectly classified as class 1 to the total number of samples that do not belong to class 1. The yaxis of the curve illustrates the true positive rate, also known as recall. We can construct ROC curves by systematically computing these two rates for each classifier applied to class 1 within the dataset. Figure 20 shows the ROC curve of the nine models. The model with a larger AUC is regarded as the superior predictor for identifying the presence of diabetes. adaBoost, gradient boosting, and XGBoost achieve the highest ranks with AUC values of 0.977, 0.976, and 0.974, respectively. Their performances are very close to each other in this metric, just as they are in the f1-score. On the other hand, the decision tree model secures the lowest rank with an AUC of 0.870.

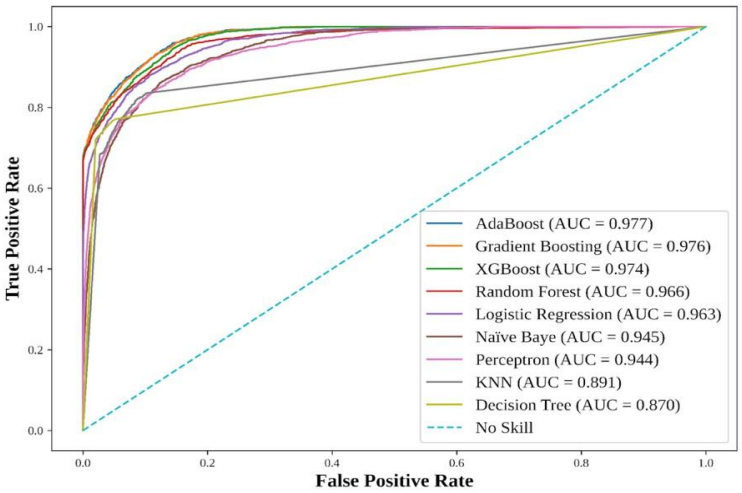

Fig.20. ROC curve for prediction models

In this paper, we treat the runtime of the model dataset as the final metric for comparing prediction models. Nevertheless, in specific instances, the runtime of the prediction model itself might take precedence. Table 4 presents the sorted performances of the nine classification models, organized based on their respective runtimes. Among these models, naïve bayes shows the quickest runtime of 0.11 seconds, while gradient boosting is the slowest with 144.42 seconds. This disparity in runtimes sheds light on the distinct computational efficiencies of these models, providing valuable insights for model selection based on time considerations.

Table 4
Run time of the prediction models

Model	Second
Naïve Bayes	0.11
KNN	0.47
Perceptron	0.56
Logistic Regression	0.89
Decision Tree	0.95
XGBoost	4.1
AdaBoost	33.87
Random Forest	85.1
Gradient Boosting	144.42

4. Conclusion

In this research, a comprehensive performance analysis is carried out on nine classification models, specifically XGBoost, logistic regression, gradient boosting, AdaBoost, random forest, decision tree, KNN, perceptron, and naïve bayes, with the aim of predicting diabetes. The research showcases the performance of these ML methods across six evaluation metrics, including accuracy, precision, recall, f1-score, AUC of ROC curve, and runtime.

The dataset exhibits a moderate imbalance; thus, we addressed this imbalance in the training dataset using the SMOTE method. However, the test dataset remains imbalanced. It made accuracy an inadequate measure for comparing supervised learning models.

Moreover, as the f1-score serves as a harmonic mean of precision and recall, it encapsulates both aspects in a single metric. Consequently, prioritizing the ranking of models based on the f1-score becomes a more comprehensive approach, considering both precision and recall simultaneously rather than examining them separately. The f1-score rankings from highest to lowest for the first five positions are as follows: gradient boosting, XGBoost, AdaBoost, random forest, and decision tree. These models exhibit f1-scores that significantly outperform those of other models.

Moreover, the rankings for AUC, listed from highest to lowest for the first five positions, are as follows: AdaBoost, gradient boosting, XGBoost, random forest, and logistic regression. These models demonstrate AUC values that notably exceed those of other models.

Conversely, the rankings for runtime, arranged from quickest to slowest for the first five positions, are as follows: naïve bayes, KNN, perceptron, logistic regression, and decision tree. The training time of these models is much shorter than that of other models.

The metrics indicate that tree-based models, especially those employing complex structures like bagging and boosting, demonstrate superior predictive capabilities for diabetes; however, their computational speed is notably slower than that of other models. This suggests that efforts to optimize the efficiency of boosting and bagging methods could significantly improve diabetes prediction. Additionally, for future research, we recommend exploring dimensionality reduction

techniques such as principal component analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE), linear discriminant analysis (LDA), independent component analysis (ICA), and autoencoders to reduce data size and enhance runtime efficiency.

Furthermore, this study emphasizes the significance of leveraging ML, particularly classification models, for disease diagnosis, explicitly focusing on conditions like diabetes. Healthcare institutions and hospitals are encouraged to implement the recommended models from this study to enable early disease detection before it progresses within patients' bodies. The methods, code, and results presented in this paper can be utilized to develop software or applications, allowing individuals to input their health information and receive awareness regarding their susceptibility to diabetes. Such a tool would empower many to proactively prevent diabetes by predicting and seeking advice from specialists for preventive measures or diabetes management. Moreover, this application could be integrated into smartwatches or smartphones, enabling individuals to monitor relevant health criteria continuously.

In our current study, we employed traditional classification ML methods. However, the capability of neural networks and their broad application, which is from decision-making [37] through the health industry, convince us to consider them for future research endeavors. These techniques include Feedforward Neural Networks (FNN), Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM) Networks, Gated Recurrent Unit (GRU) Networks, Siamese Networks, and Capsule Networks. Integrating these sophisticated deep learning models holds the potential for substantial enhancement in f1-score, AUC, and runtime, enabling more effective analysis in subsequent investigation.

Author Contributions

Conceptualization, methodology, M.M., M.Z., M.M., A.Z., M.B., F.H., and K.G.; software, F.H., and K.G.; validation, formal analysis, resources, investigation, M.M., M.Z., F.H., and K.G.; writing—original draft preparation, writing—review and editing, visualization, M.M., M.Z., M.M., A.Z., M.B., F.H., and K.G.; supervision, M.M., and M.Z. All authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgment

This research was not funded by any grant.

References

[1] Heald, A. H., Stedman, M., Davies, M., Livingston, M., Alshames, R., Lunt, M., ... & Gadsby, R. (2020). Estimating life years lost to diabetes: outcomes from analysis of National Diabetes Audit and Office of National Statistics data. Cardiovascular endocrinology & metabolism, 9(4), 183-185. https://doi.org/10.1097/XCE.0000000000000210

[2] Centers for Disease Control and Prevention. Diabetes. (2023). https://www.cdc.gov/diabetes/basics/diabetes.html/Accessed September 5, 2023.

[3] IDF Diabetes Atlas. Diabetes around the world in 2021. (2021). https://www.diabetesatlas.org/ Accessed March 1, 2023.

[4] American Diabetes Association. Statistics about diabetes. (2023). https://diabetes.org/about-diabetes/statistics/about-diabetes/ Accessed November 2, 2023.

[5] Sun, C., van Soest, J., Koster, A., Eussen, S. J., Schram, M. T., Stehouwer, C. D., ... & Dumontier, M. (2022). Studying the association of diabetes and healthcare cost on distributed data from the Maastricht Study and Statistics Netherlands using a privacy-preserving federated learning infrastructure. Journal of Biomedical Informatics, 134, 104194. https://doi.org/10.1016/j.jbi.2022.104194

[6] Ong, K. L., Stafford, L. K., McLaughlin, S. A., Boyko, E. J., Vollset, S. E., Smith, A. E., ... & Brauer, M. (2023). Global, regional, and national burden of diabetes from 1990 to 2021, with projections of prevalence to 2050: a systematic analysis for the Global Burden of Disease Study 2021. The Lancet, 402(10397), 203-234. https://doi.org/10.1016/S0140-6736(23)01301-6

[7] Jamali, H., Karimi, A., & Haghighizadeh, M. (2018, February). A new method of cloud-based computation model for mobile devices: energy consumption optimization in mobile-to-mobile computation offloading. In Proceedings of the 6th International Conference on Communications and Broadband Networking (pp. 32-37). https://doi.org/10.1145/3193092.3193103

[8] Schön, M., Prystupa, K., Mori, T., Zaharia, O. P., Bódis, K., Bombrich, M., ... & Schrauwen-Hinderling, V. (2024). Analysis of type 2 diabetes heterogeneity with a tree-like representation: insights from the prospective German Diabetes Study and the LURIC cohort. The Lancet Diabetes & Endocrinology, 12(2), 119-131. https://doi.org/10.1016/S2213-8587(23)00329-7

[9] Garbin, C., Marques, N., & Marques, O. (2023). Machine learning for predicting opioid use disorder from healthcare data: a systematic review. Computer Methods and Programs in Biomedicine, 107573. https://doi.org/10.1016/j.cmpb.2023.107573

[10] Gaudelet, T., Day, B., Jamasb, A. R., Soman, J., Regep, C., Liu, G., ... & Taylor-King, J. P. (2021). Utilizing graph machine learning within drug discovery and development. Briefings in bioinformatics, 22(6), bbab159. https://doi.org/10.1093/bib/bbab159

[11] Mbunge, E., & Batani, J. (2023). Application of deep learning and machine learning models to improve healthcare in sub-Saharan Africa: Emerging opportunities, trends and implications. Telematics and Informatics Reports, 100097. https://doi.org/10.1016/j.teler.2023.100097

[12] Ibrahim, Z., Tulay, P., & Abdullahi, J. (2023). Multi-region machine learning-based novel ensemble approaches for predicting COVID-19 pandemic in Africa. Environmental Science and Pollution Research, 30(2), 3621-3643. https://doi.org/10.1007/s11356-022-22373-6

[13] Soltaninejad, M., Aghazadeh, R., Shaghaghi, S., & Zarei, M. (2024). Using Machine Learning Techniques to Forecast Mehram Company's Sales: A Case Study. Journal of Business and Management Studies, 6(2), 42-53. https://doi.org/10.32996/jbms.2024.6.2.4

[14] Shill, P. C., Wu, R., Jamali, H., Hutchins, B., Dascalu, S., Harris, F. C., & Feil-Seifer, D. (2023). WIP: Development of a Student-Centered Personalized Learning Framework to Advance Undergraduate Robotics Education. In 2023 IEEE Frontiers in Education Conference (FIE) (pp. 1-5). IEEE. https://doi.org/10.1109/FIE58773.2023.10343234

[15] Abubakar, A., Ajuji, M., & Yahya, I. U. (2021). DeepFMD: computational analysis for malaria detection in blood-smear images using deep-learning features. Applied System Innovation, 4(4), 82. https://doi.org/10.3390/asi4040082

[16] Maydanchi, M., Ziaei, A., Basiri, M., Azad, A. N., Pouya, S., Ziaei, M., ... & Sargolzaei, S. (2023). Comparative Study of decision tree, adaboost, random forest, Naïve Bayes, KNN, and perceptron for heart disease prediction. In SoutheastCon 2023 (pp. 204-208). IEEE. https://doi.org/10.1109/SoutheastCon51012.2023.10115189

[17] Haseli, G., Sheikh, R., & Sana, S. S. (2020). Base-criterion on multi-criteria decision-making method and its applications. International journal of management science and engineering management, 15(2), 79-88. https://doi.org/10.1080/17509653.2019.1633964

[18] Haseli, G., & Sheikh, R. (2022). Base criterion method (BCM). In Multiple criteria decision making: Techniques, Analysis and Applications (pp. 17-38). Singapore: Springer Nature Singapore. https://doi.org/10.1007/978-981-16-7414-3_2

[19] Hennebelle, A., Materwala, H., & Ismail, L. (2023). HealthEdge: a machine learning-based smart healthcare framework for prediction of type 2 diabetes in an integrated IoT, edge, and cloud computing system. Procedia Computer Science, 220, 331-338. https://doi.org/10.1016/j.procs.2023.03.043

[20] Toscano-Pulido, G., Razavi, H., Nejadhashemi, A. P., Deb, K., & Linker, L. (2024). Large-Scale Multiobjective Optimization for Watershed Planning and Assessment. IEEE Transactions on Systems, Man, and Cybernetics: Systems. https://doi.org/10.1109/TSMC.2024.3361679

[21] MacKay, C., Klement, W., Vanberkel, P., Lamond, N., Urquhart, R., & Rigby, M. (2023). A framework for implementing machine learning in healthcare based on the concepts of preconditions and postconditions. Healthcare Analytics, 3, 100155. https://doi.org/10.1016/j.health.2023.100155

[22] Jangir, S. K., Joshi, N., Kumar, M., Choubey, D. K., Singh, S., & Verma, M. (2021). Functional link convolutional neural network for the classification of diabetes mellitus. International Journal for Numerical Methods in Biomedical Engineering, 37(8), e3496. https://doi.org/10.1002/cnm.3496
[23] Modak, S. K. S., & Jha, V. K. (2023). Diabetes prediction model using machine learning techniques. Multimedia Tools and Applications, 1-27. https://doi.org/10.1007/s11042-023-16745-4
[24] Tasin, I., Nabil, T. U., Islam, S., & Khan, R. (2023). Diabetes prediction using machine learning and explainable AI techniques. Healthcare technology letters, 10(1-2), 1-10. https://doi.org/10.1049/htl2.12039
[25] Mujumdar, A., & Vaidehi, V. (2019). Diabetes prediction using machine learning algorithms. Procedia Computer Science, 165, 292-299. https://doi.org/10.1016/j.procs.2020.01.047
[26] Lai, H., Huang, H., Keshavjee, K., Guergachi, A., & Gao, X. (2019). Predictive models for diabetes mellitus using machine learning techniques. BMC endocrine disorders, 19, 1-9. https://doi.org/10.1186/s12902-019-0436-6
[27] Sarwar, M. A., Kamal, N., Hamid, W., & Shah, M. A. (2018). Prediction of diabetes using machine learning algorithms in healthcare. In 2018 24th international conference on automation and computing (ICAC) (pp. 1-6). IEEE. https://doi.org/10.23919/IConAC.2018.8748992
[28] Saru, S., & Subashree, S. (2019). Analysis and prediction of diabetes using machine learning. International journal of emerging technology and innovative engineering, 5(4).
[29] Minsky, M., & Papert, S. (1969). An introduction to computational geometry. Cambridge tiass., HIT, 479(480), 104.
[30] Zhang, S., Li, X., Zong, M., Zhu, X., & Cheng, D. (2017). Learning k for knn classification. ACM Transactions on Intelligent Systems and Technology (TIST), 8(3), 1-19. https://doi.org/10.1145/2990508
[31] Langarizadeh, M., & Moghbeli, F. (2016). Applying naive bayesian networks to disease prediction: a systematic review. *Acta Informatica Medica*, *24*(5), 364. https://doi.org/10.5455/aim.2016.24.364-369
[32] Demir, S., & Sahin, E. K. (2023). An investigation of feature selection methods for soil liquefaction prediction based on tree-based ensemble algorithms using AdaBoost, gradient boosting, and XGBoost. *Neural Computing and Applications*, *35*(4), 3173-3190. https://doi.org/10.1007/s00521-022-07856-4
[33] Speiser, J. L., Miller, M. E., Tooze, J., & Ip, E. (2019). A comparison of random forest variable selection methods for classification prediction modeling. *Expert systems with applications*, *134*, 93-101. https://doi.org/10.1016/j.eswa.2019.05.028
[34] Song, Y. Y., & Ying, L. U. (2015). Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry*, *27*(2), 130. https://doi.org/ 10.11919/j.issn.1002-0829.215044
[35] Rymarczyk, T., Kozłowski, E., Kłosowski, G., & Niderla, K. (2019). Logistic regression for machine learning in process tomography. Sensors, 19(15), 3400. https://doi.org/10.3390/s19153400
[36] LaValley, M. P. (2008). Logistic regression. Circulation, 117(18), 2395-2399. https://doi.org/10.1161/CIRCULATIONAHA.106.682658
[37] Haseli, G., Ranjbarzadeh, R., Hajiaghaei-Keshteli, M., Ghoushchi, S. J., Hasani, A., Deveci, M., & Ding, W. (2023). HECON: Weight assessment of the product loyalty criteria considering the customer decision's halo effect using the convolutional neural networks. *Information Sciences*, *623*, 184-205. https://doi.org/10.1016/j.ins.2022.12.027