2024 IEEE 16th International Conference on Computational Intelligence and Communication Networks (CICN) | 979-8-3315-0526-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/CICN63059.2024.10847398

Comparative Effectiveness of Classification Algorithms in Predicting Diabetes

Fares A. Dael
Management Information Systems
İzmir Bakırçay University
İzmir, Turkey
fares.dael@bakircay.edu.tr

D. Mareyev
Department of Computational and Data Science, Astana IT University
Astana, Kazakhstan
mrvdnl@mail.ru

Ibraheem Shayea
Electronic and Communication Engineering
Istanbul Technical University
Istanbul, Turkey
shayea@itu.edu.tr

Kulniyazova, Korlan S.
Department of Systems Analysis and Control, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan
k_korlan@mail.ru

Gulnara Abitova
Department of Intelligent Systems and Cybersecurity, Astana IT University,
Astana, Kazakhstan
gulnara.abitova@astanait.edu.kz

***Abstract*— Diabetes mellitus poses a significant global health challenge, with increasing prevalence, particularly in low socioeconomic regions. Accurate and early diagnosis is crucial to prevent the severe long-term complications associated with diabetes. This study conducts a comprehensive comparison of six prominent machine learning algorithms—K-Nearest Neighbors (K-NN), Naive Bayes, Support Vector Machine (SVM), Decision Trees, Random Forest, and Logistic Regression—in predicting diabetes using a dataset of 768 individuals with diverse diabetic indicators from Kaggle. Each algorithm is rigorously evaluated based on precision, recall, and F1-score to determine the most effective method for diabetes diagnosis. The results indicate that Logistic Regression outperforms the other algorithms, achieving an accuracy of 81%. This superior performance is attributed to Logistic Regression's ability to effectively delineate linear separations, which is crucial for distinguishing between diabetic and non-diabetic individuals. The study underscores the importance of feature selection and model tuning in enhancing predictive performance. The findings suggest that integrating Logistic Regression into clinical settings can significantly improve the accuracy and timeliness of diabetes diagnosis, potentially leading to better patient outcomes and reduced healthcare costs.**

Keywords*— *Diabetes Diagnosis, Machine Learning, K-Nearest Neighbors, Naive Bayes, Support Vector Machine, Decision Trees, Random Forest, Logistic Regression.

I. INTRODUCTION

The incidence of diabetes has increased significantly in recent decades and has become a major global public health problem. According to the World Health Organization, the number of people with diabetes increased from 108 million in 1980 to 422 million in 2014, with particularly dramatic increases in low- and middle-income countries [1]. Diabetes mellitus is a burgeoning global health crisis characterized by chronic hyperglycemia resulting from defects in insulin secretion, insulin action, or both (American Diabetes Association [ADA], 2022)[2]. The International Diabetes Federation (IDF) has reported a staggering prevalence of approximately 463 million adults with diabetes in 2019, a figure projected to soar to 700 million by 2045 (International Diabetes Federation[3]. The escalating incidence is particularly alarming in regions with lower socioeconomic status, where healthcare access and preventive measures are often scarce[3].

Early and accurate diabetes detection is paramount for effective disease management and averting complications such as cardiovascular disease, neuropathy, nephropathy, and retinopathy [4]. Traditional diagnostic methods, including fasting plasma glucose (FPG) tests, oral glucose tolerance tests (OGTT), and HbA1c measurements, offer reliable results but necessitate multiple patient visits and extensive time commitments [4]. This has ignited interest in developing automated, data-driven diagnostic tools to streamline the diagnostic process and alleviate the burden on healthcare systems [5].

Machine learning, a subset of artificial intelligence, has emerged as a promising avenue for medical diagnostics, capitalizing on vast datasets to uncover patterns and make accurate predictions [5], [6]. In recent years, various machine learning algorithms have been employed to enhance diagnostic precision and efficiency in the medical domain [7]. Among these, classification algorithms have demonstrated significant potential in forecasting diabetes by analyzing a multitude of risk factors and biomarkers [7].

This study aims to comparatively assess the effectiveness of six widely used machine learning classification algorithms—K-Nearest Neighbors (K-NN), Naive Bayes, Support Vector Machine (SVM), Decision Trees, Random Forest, and Logistic Regression—in diagnosing diabetes. By leveraging a well-established dataset from Kaggle encompassing 768 individuals with diverse diabetic indicators, this research seeks to identify the algorithm exhibiting superior performance in terms of precision, recall, and F1-score.

The selection of these algorithms is grounded in their diverse methodological approaches, providing a

979-8-3315-0526-4/24/$31.00 ©2024 IEEE
DOI: 10.1109/CICN.2024.228

Table 1: The first 10 records of the diabetes dataset

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1
5	5	116	74	0	0	25.6	0.201	30	0
6	3	78	50	32	88	31.0	0.248	26	1
7	10	115	0	0	0	35.3	0.134	29	0
8	2	197	70	45	543	30.5	0.158	53	1
9	8	125	96	0	0	0.0	0.232	54	1

comprehensive evaluation of their strengths and limitations in the context of diabetes diagnosis. K-Nearest Neighbors is a straightforward, instance-based learning algorithm that classifies data points based on their similarity to neighboring points [8], [9]. Naive Bayes, a probabilistic classifier, applies Bayes' theorem under the assumption of feature independence [10], [11]. Support Vector Machine constructs optimal hyperplanes to segregate different classes in a high-dimensional space [12], [13], while Decision Trees and Random Forests create predictive models through a series of binary decisions [14], [15]. Logistic Regression, a widely employed linear model, estimates the probability of a binary outcome based on predictor variables [16].

Through rigorous evaluation and comparison, this study aims to provide valuable insights into the most effective machine learning approaches for diabetes diagnosis. The findings will contribute to the ongoing efforts to integrate advanced machine learning techniques into clinical practice, ultimately enhancing diagnostic accuracy, improving patient outcomes, and reducing healthcare costs.

II. RELATED WORKS

The application of machine learning to predict diabetes has gained significant traction in recent years. Numerous studies have explored the potential of various algorithms in identifying individuals at risk of developing the disease[17], [18], [19], [20].

Several studies have focused on the use of traditional classification algorithms such as Naive Bayes, Decision Trees, and Logistic Regression for diabetes prediction [7], [21], [22]. These studies have demonstrated varying degrees of success in terms of accuracy, sensitivity, and specificity. However, the performance of these models can be influenced by factors such as dataset size, feature selection, and model hyperparameter tuning.

More recently, there has been growing interest in employing ensemble methods like Random Forest and Support Vector Machines for diabetes prediction [23]. These algorithms have shown promise in improving predictive performance compared to traditional methods. However, these models can be computationally expensive and may require extensive feature engineering.

While the aforementioned studies have contributed to the field, a comprehensive comparative analysis of multiple algorithms considering performance metrics such as precision, recall, and F1-score is still lacking. Furthermore, the impact of different dataset characteristics on algorithm performance has not been extensively explored.

This study aims to address these gaps by conducting a rigorous comparison of six widely used classification algorithms on a well-established dataset. The findings of this research will provide valuable insights into the relative strengths and weaknesses of these algorithms for diabetes prediction.

III. METHODOLOGY

I. DATA COLLECTION AND PROCESSING

Data collection and processing are critical steps in any analysis or research project, as they determine the quality and reliability of the conclusions that can be drawn from the analysis.

A. EXPLORING THE DATA

For this study, was selected a dataset with information about diabetes, including 768 records and 9 attributes, taken from the website kaggle.com [24]. The first 10 records of the dataset are presented in B, where:

1) Pregnancies - number of pregnancies;
2) Glucose - glucose concentration in blood plasma;
3) BloodPressure - diastolic blood pressure (mm Hg);
4) SkinThickness - thickness of the triceps skin fold (mm);
5) Insulin - 2-hour serum insulin (μU/ml);
6) BMI - body mass index;
7) DiabetesPedigreeFunction - diabetes pedigree function;
8) Age - age in years;
9) Outcome - presence of diabetes (0 - absent, 1 - present).

B shows that the data contains zero values for such indicators as blood pressure, skin thickness and blood insulin levels. Zero values may indicate missing data rather than actual biological measurements. In real conditions, these parameters cannot be equal to zero since this does not correspond to human physiological norms.

The presence of such anomalies can significantly distort the results of the analysis, as they introduce bias and can lead to incorrect statistical conclusions.

It follows from this that in the future, for the algorithms to work correctly, it is necessary to replace the zeros with the average or median value for this characteristic among all observations.

Data types in the dataset range from integer (int64) for most variables to float (float64) for the BMI and Fig fn df Diabetes Pedigree Function variables The ratio of records with and without diabetes (the 'Outcomes' column) is presented in 0.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 768 entries, 0 to 767
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Pregnancies               768 non-null    int64
 1   Glucose                   768 non-null    int64
 2   BloodPressure             768 non-null    int64
 3   SkinThickness             768 non-null    int64
 4   Insulin                   768 non-null    int64
 5   BMI                       768 non-null    float64
 6   DiabetesPedigreeFunction  768 non-null    float64
 7   Age                       768 non-null    int64
 8   Outcome                   768 non-null    int64
dtypes: float64(2), int64(7)
memory usage: 54.1 KB
```

Fig. 1: Data columns and their data types

A. DATA CLEANING

In the dataset in question, the columns 'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin' and 'BMI' contain zero values. In order to fill zero values with the average or median value of the corresponding column, it is necessary to replace zero values with NaN and make a visual assessment of the distribution of their values by constructing the corresponding histograms. Based on the visual assessment of the data, the 'Glucose' and 'BloodPressure' columns are replaced with the corresponding mean values, and the 'SkinThickness', 'Insulin' and 'BMI' columns are replaced with the median values.

B. The first 10 records of the diabetes dataset

The histograms of each column after excluding zero values are depicted in.

After processing the data set, the most optimal classification algorithm is selected.

CLASSIFICATION ALGORITMS AND RESULT

For further analysis of the dataset, it is proposed to consider six classification algorithms, such as: k-nearest neighbors (K-NN), support vector machine (SVM), Naive Bayes, Decision Tree, Random Forest and Logistic Regression.

Each algorithm will be evaluated against a classification report that includes metrics such as: precision, recall, f1-score, support and accuracy.

A. K-NEAREST NEIGHBORS (K-NN)

The k-nearest neighbors (K-NN) algorithm is one of the main methods in the field of machine learning for classification problems. This algorithm is based on a simple idea: an object is classified by the majority vote of its neighbors, with the object assigned the class most frequently found among its k nearest neighbors.

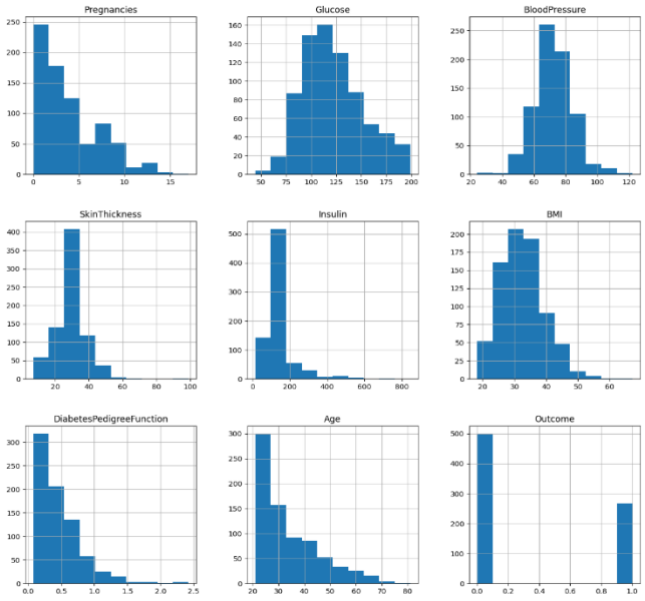

Fig. 2: Data before data cleaning

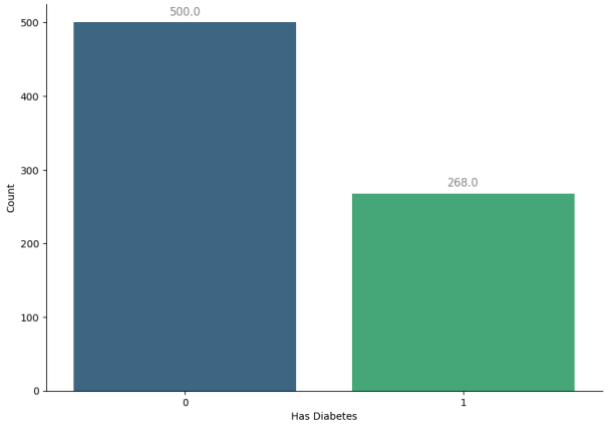

Fig. 3: Data after data cleaning

Traditionally, the K-NN algorithm consists of three main stages:

1) Selecting the number k, which determines the number of neighbors to consider;

2) Defining a distance metric to measure distances between data points;

3) Classify a new data point by identifying the k nearest neighbors and making a decision based on the most frequently occurring class among the neighbors.

However, the traditional K-NN method has disadvantages, especially related to computational efficiency and selection of the optimal value of k. The paper [25] mentions issues such as k selection, nearest neighbor selection, nearest neighbor search, and classification rules that remain relevant for research in this area.

In addition, a new approach [26] has been proposed that simplifies the process by integrating k selection and k nearest

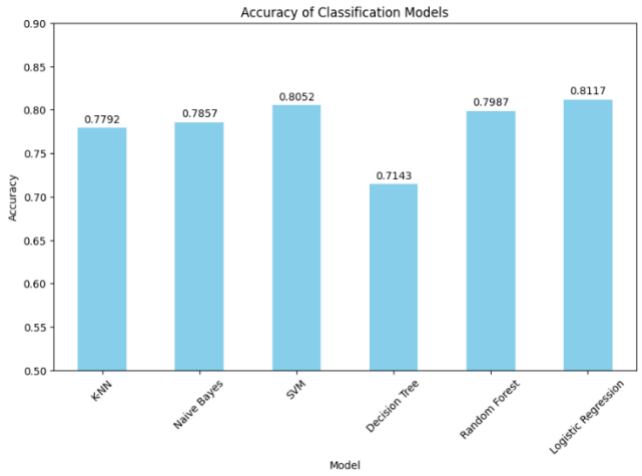

Fig. 4: Proportion of records with and without diabetes

neighbor search into a single matrix operation. This method uses least squares loss function minimization to fit the test data with the training samples, followed by group lasso for efficient sparse learning and neighbor selection.

The new approach not only improves computational efficiency, but also proposes a new classification rule that improves the classification performance of K-NN. Experimental evaluation shows that the proposed method outperforms traditional approaches in terms of both classification performance and computational cost.

To build a data classification model from the diabetes dataset using the K-NN algorithm, the scikit-learn library is used using the KNeighborsClassifier module. The classification report of the K-NN model is shown in 0. From the report, it can be seen that the accuracy of the model is 78%.

Table 2: Classification report of the K-NN model

	precision	recall	f1-score	support
0	0.85	0.83	0.84	107
1	0.63	0.66	0.65	47
accuracy			0.78	154
macro avg	0.74	0.75	0.74	154
weighted avg	0.78	0.78	0.78	154

B. SUPPORT VECTOR MACHINE (SVM)

The traditional support vector machine (SVM) algorithm is a powerful machine learning technique used for classification and regression. SVM searches for a hyperplane in high-dimensional space that best separates different classes of data. The main idea is to maximize the gap between classes, which makes the model robust to noise in the data. However, when dealing with imbalanced data, traditional SVM may be biased towards the larger class.

To improve the performance of SVM with unbalanced data, a modified approach was developed [27]. This approach proposes a hybrid feature selection model (SVM-mRMRe) for cancer classification based on high-dimensional microarray data. This model uses the SVM technique combined with the minimum redundancy and maximum relevance (mRMRe) method to effectively manage the high dimensionality of the data and select the most informative genes.

Additionally, another example of an improved use of SVM can be considered - the cost-sensitive SVM (CMSVM) algorithm for classifying network traffic [28]. This approach includes a multi-class SVM with active learning that can adaptively assign weights to different applications.

To build a data classification model from the diabetes dataset, the SVM algorithm uses the scikit-learn library using the SVC module with a linear optimization kernel kernel='linear' and parameter C=1.0. The classification report of the SVM model is shown in 0. From the report, it can be seen that the accuracy of the model is 81%.

Table 3: Classification report of the SVM model

	precision	recall	f1-score	support
0	0.83	0.91	0.87	107
1	0.73	0.57	0.64	47
accuracy			0.81	154
macro avg	0.78	0.74	0.75	154
weighted avg	0.80	0.81	0.80	154

C. NAIVE BAYES

The Naive Bayes classifier algorithm is a well-known and effective method in the field of machine learning for classification problems. Based on the principles of probability theory, this algorithm uses Bayes' theorem to predict the probability of an object belonging to a certain class based on its features.

The Naive Bayes classifier assumes that all features in the data are independent of each other within each class, which

simplifies calculations and makes the method fast even on large amounts of data. Despite its "naivety", this method often shows high efficiency in many real-world applications, such as spam filtering, medical diagnostics and text classification.

A recent study [29] reports on the use of a naive Bayes classifier to identify malicious applications on the Android operating system. The study used a dataset containing 2854 malicious and 2870 benign programs, each of which is characterized by 116 permission features. The Naive Bayes classifier showed high efficiency, achieving a classification accuracy of 92.4%, which confirms its potential in the task of identifying malignant software.

To build a data classification model from the diabetes dataset using the Naive Bayes algorithm, the scikit-learn library is used using the GaussianNB module. The classification report of the Naive Bayes model is presented in 0. The report shows that the accuracy of the model is 79%.

Table 4: Classification report of the Naive Bayes model

	precision	recall	f1-score	support
0	0.84	0.86	0.85	107
1	0.66	0.62	0.64	47
accuracy			0.79	154
macro avg	0.75	0.74	0.74	154
weighted avg	0.78	0.79	0.78	154

D. DECISION TREE

Decision Tree algorithm is a popular method in machine learning that is used for classification and regression. He constructs a decision-making model in the form of a tree, where the leaves represent decisions or end results and the branches represent the conditions that lead to those decisions. The advantages of decision trees include their interpretability and ability to easily handle both numeric and categorical data.

One way to use decision trees is to diagnose diseases in mobile medical networks, such as smartphones with health sensors [30]. The problem is that devices have limited resources - low battery life and little processing power - and the data must remain confidential.

To solve this problem, the PDTC-LRC protocol is proposed, which reduces power consumption and ensures data protection. The basis of the protocol is the new FCRS encryption scheme, which allows you to perform calculations on encrypted data efficiently. The protocol makes it possible to diagnose diseases while maintaining the confidentiality of user data and medical cloud models [30].

Decision Tree algorithms can effectively improve scheduling problems in cloud environments by considering multitasking and resource utilization [31], [32].

Cloud computing allows users to flexibly use various resources, but the problem of allocating these resources is one of the key challenges.

A new Task Scheduling-Decision Tree (TS-DT) algorithm is proposed that improves performance by minimizing task execution time (makespan), improving resource utilization, and load balancing. To evaluate the effectiveness of TS-DT, a comparison was made with the existing HEFT, TOPSIS-EWM and QL-HEFT algorithms [31].

The results show that TS-DT outperforms these algorithms, reducing task execution time by 5.21%, 2.54% and 3.32% respectively, improving resource utilization by 4.69%, 6.81% and 8.27%, and improving load balancing by 33.36%, 19.69% and 59.06 %. The main disadvantage of the proposed algorithm is the increase in energy consumption [31].

To build a data classification model from the diabetes dataset, the Decision Tree algorithm uses the scikit-learn library using the DecisionTreeClassifier module. The classification report of the Decision Tree model is presented in 0. The report shows that the accuracy of the model is 71%.

Table 5: Classification report of the Decision Tree model

	precision	recall	f1-score	support
0	0.79	0.80	0.80	107
1	0.53	0.51	0.52	47
accuracy			0.71	154
macro avg	0.66	0.66	0.66	154
weighted avg	0.71	0.71	0.71	154

E. RANDOM FOREST

The Random Forest algorithm is an ensemble machine learning technique that uses multiple decision trees to solve classification and regression problems. This method creates a "forest" that is more robust and accurate than individual decision trees. In Random Forest, each tree is trained on a randomly selected subset of the training data, and random subsets of features are used to select the best split at a node. Once all the trees are built, Random Forest aggregates their predictions to produce a final result, which improves accuracy and reduces the risk of overfitting compared to a single decision tree.

The main difference between Random Forest and Decision Tree is that Random Forest uses an ensemble of many trees to reduce overfitting and improve accuracy, whereas Decision Tree uses a single tree, which can result in less stable predictions.

The Random Forest algorithm is an effective tool for detecting credit card fraud due to its ability to process large amounts of data and reduce variations, improving the overall performance of the fraud detection system [33].

Experimental results show [33] that using Random Forest along with other machine learning methods such as logistic regression (LR), artificial neural networks (ANN), and support vector machines (SVMs) can create robust classifiers for fraud detection. The Random Forest ensemble approach helps improve fraud detection accuracy by combining multiple decision trees trained on random subsamples of data.

The Random Forest algorithm can be effectively used to diagnose heart diseases and predict patient survival, overcoming the limitations associated with class imbalance and high data dimensionality [34]. The IWRF algorithm

improves prediction accuracy by using class weighting to account for unbalanced data and using Bayesian optimization to select hyperparameters. The method also includes a feature selection algorithm (Inf-FSs), which reduces data dimensionality and improves model efficiency. Experiments show that IWRF outperforms other machine learning methods in accuracy and F-measure, making it a promising tool for medical diagnosis [34].

To build a data classification model from the diabetes dataset using the Randon Forest algorithm, the scikit-learn library is used using the RandomForestClassifier module with the number of trees within the ensemble n_estimators=100. The classification report of the Random Forest model is presented in 0. The report shows that the accuracy of the model is 80%.

Table 6: Classification report of the Random Forest model

	precision	recall	f1-score	support
0	0.85	0.86	0.86	107
1	0.67	0.66	0.67	47
accuracy			0.80	154
macro avg	0.76	0.76	0.76	154
weighted avg	0.80	0.80	0.80	154

F. LOGISTIC REGRESSION

Logistic regression is a statistical analysis technique that is used to predict the likelihood of belonging to a particular category. It is a type of linear regression specifically adapted for binary classification problems and can be extended to multi-class classification using one-vs-rest or one-vs-one techniques.

Logistic regression is based on a logistic function, or sigmoid, which converts any real number into a range between 0 and 1. This makes it ideal for modeling probability. In the context of logistic regression, the dependent variable Y represents the probability that a given example belongs to one of the classes, usually "1". The model predicts the probability that Y=1, given a set of predictors X.

The logistic regression algorithm can be used to diagnose heart diseases [35]. Logistic regression is used to predict the probability of a patient having a disease based on a variety of input parameters.

Despite its simplicity, logistic regression showed high accuracy in predicting heart disease, especially after hyperparameter optimization using GridSearchCV. Logistic regression is one of the methods used in the study to build predictive models.

The main steps in applying logistic regression include collecting and preprocessing data, selecting optimal features, training the model, and evaluating its performance. A study shows that using GridSearchCV for hyperparameter optimization significantly improves the accuracy of a logistic regression model [35].

Thus, logistic regression can be effectively used for early diagnosis of heart diseases, which contributes to timely treatment and reduction of mortality from cardiovascular diseases.

To build a data classification model from the diabetes dataset using the Logistic Regression algorithm, the scikit-learn library is used using the LogisticRegression module. The qualification report of the LogisticRegression model is presented in 0. The report shows that the accuracy of the model is 81%.

Table 7: Classification report of the Logistic Regression

	precision	recall	f1-score	support
0	0.84	0.91	0.87	107
1	0.74	0.60	0.66	47
accuracy			0.81	154
macro avg	0.79	0.75	0.76	154
weighted avg	0.81	0.81	0.81	154

IV. DISSCUSION

Based on the analysis of algorithms for classifying diabetes data, the Logistic Regression algorithm turned out to be the most effective with an accuracy of 81%. The second and third places in terms of efficiency were taken by the SVM and Random Forest algorithms with an accuracy of 80.5% and 79.9%, respectively. The efficiency of the algorithms is presented in **Error! Reference source not found.**.

The effectiveness of algorithms is determined based on precision, recall, F1-measure and overall accuracy:

– K-Nearest Neighbors (K-NN): Accuracy 0.78 with best performance for class 0 (F1-measure 0.84).

– Support Vector Machine (SVM): Accuracy of 0.805 with strong recall for class 0 (0.91), but lower performance for class 1.

– Naive Bayes: Accuracy of 0.79 with consistent performance across all metrics, slightly below the Class 1 F1 measure (0.64).

– Decision Tree: Accuracy 0.71 with overall inferior performance across all metrics.

– Random Forest: Accuracy 0.80 with good balance across classes, showing high F1 measures (0.86 for class 0 and 0.67 for class 1).

– Logistic Regression: Accuracy of 0.81 with the best overall weighted average metrics among all models, indicating strong performance in both classes.

Logistic Regression and SVM show better overall accuracy and demonstrate strong class classification capabilities. Random Forest also shows high accuracy and balanced class performance, making it suitable for scenarios where equal attention to both classes is important. The Decision Tree model shows the weakest performance, which may be due to overfitting or insufficient ability to capture the complexity of the data.

Logistic regression was found to be the most appropriate, perhaps due to the model's good ability to linearly separate the data, which is often effective when the categories of the target variable are fairly clearly demarcated. While SVM and

random forest, while providing competitive results, may require additional parameter tuning to improve results on similar datasets.

In this work, for the Random Forest algorithm, the parameter that regulates the number of trees within the n_estimators ensemble is set to 100. For the SVM algorithm, the linear optimization kernel kernel='linear' and the parameter C=1.0 were used.

V. CONCLUSION

This study compared six machine learning algorithms to determine the most effective method for classifying diabetes. The statistical analysis performed on the medical data included the following key steps:

– A publicly available dataset of patient medical records was used, containing key indicators related to diabetes;

– To improve data quality, missing values were eliminated and anomalies were corrected, resulting in more accurate and reliable modeling;

– Various machine learning algorithms were used, including Logistic Regression, SVM, Random Forest, Decision Tree, Naive Bayes and K-NN;

– Models were trained on prepared data, with special attention paid to preventing overfitting and improving generalization ability;

– The performance of each model was assessed using indicators such as accuracy, sensitivity and F-measure. These metrics allowed us to objectively compare the performance of different algorithms.

This study examined the problem of determining the most effective algorithm for classifying diabetes, which is key to improving diagnostic methods in the medical field. The study identified logistic regression as the most suitable method, providing a classification accuracy of 81%, making it particularly valuable for practical use in healthcare settings.

The main advantage of logistic regression is its ability to effectively separate categories, which is ideal for the diagnosis of diabetes, where a clear distinction between healthy and sick patients is critical for early detection and prevention of complications. This technique allows not only to improve the accuracy of predictions, but also to speed up the diagnostic process, reducing the burden on medical personnel and increasing the availability of testing for the general population.

REFERENCES

[1] "Diabetes." Accessed: Aug. 06, 2024. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/diabetes

[2] "Powered By Purpose: Making a Difference".

[3] Home *et al.*, "IDF Diabetes Atlas 2021 | IDF Diabetes Atlas." Accessed: Aug. 06, 2024. [Online]. Available: https://diabetesatlas.org/atlas/tenth-edition/

[4] P. Nagaraj and P. Deepalakshmi, "An intelligent fuzzy inference rule-based expert recommendation system for predictive diabetes diagnosis," *International Journal of Imaging Systems and Technology*, vol. 32, no. 4, pp. 1373–1396, 2022, doi: 10.1002/ima.22710.

[5] M. Abadi *et al.*, "TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems," Mar. 16, 2016, *arXiv*: arXiv:1603.04467. doi: 10.48550/arXiv.1603.04467.

[6] N. Tendikov *et al.*, "Security Information Event Management data acquisition and analysis methods with machine learning principles," *Results in Engineering*, vol. 22, p. 102254, Jun. 2024, doi: 10.1016/j.rineng.2024.102254.

[7] F. Lotte *et al.*, "A review of classification algorithms for EEG-based brain-computer interfaces: a 10 year update," *J Neural Eng*, vol. 15, no. 3, p. 031005, Jun. 2018, doi: 10.1088/1741-2552/aab2f2.

[8] A. X. Wang, S. S. Chukova, and B. P. Nguyen, "Ensemble *k*-nearest neighbors based on centroid displacement," *Information Sciences*, vol. 629, pp. 313–323, Jun. 2023, doi: 10.1016/j.ins.2023.02.004.

[9] H. Erkil, İ. Aktı, F. A. Dael, I. Shayea, and A. A. El-Saleh, "Comparative study of K-NN algorithm for transportation mode detection using mobile phone sensor data," *AIP Conference Proceedings*, vol. 3153, no. 1, p. 020006, Jun. 2024, doi: 10.1063/5.0216676.

[10] D. D. Putri, G. F. Nama, and W. E. Sulistiono, "Analisis Sentimen Kinerja Dewan Perwakilan Rakyat (DPR) Pada Twitter Menggunakan Metode Naive Bayes Classifier," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 10, no. 1, Art. no. 1, Jan. 2022, doi: 10.23960/jitet.v10i1.2262.

[11] A. Secilmis, N. Aksu, F. A. Dael, I. Shayea, and A. A. El-Saleh, "Machine Learning-Based Fire Detection: A Comprehensive Review and Evaluation of Classification Models," *JOIV : International Journal on Informatics Visualization*, vol. 7, no. 3–2, pp. 1982–1988, Nov. 2023, doi: 10.30630/joiv.7.3-2.2332.

[12] S. Huang, N. Cai, P. P. Pacheco, S. Narrandes, Y. Wang, and W. Xu, "Applications of Support Vector Machine (SVM) Learning in Cancer Genomics," *Cancer Genomics & Proteomics*, vol. 15, no. 1, pp. 41–51, Jan. 2018.

[13] F. Abdulhafidh Dael, U. Yavuz, and A. A. Almohammedi, "Performance Evaluation of Time Series Forecasting Methods in The Stock Market: A Comparative Study," in *2022 International Conference on Decision Aid Sciences and Applications (DASA)*, Mar. 2022, pp. 1510–1514. doi: 10.1109/DASA54658.2022.9765177.

[14] J. Hu and S. Szymczak, "A review on longitudinal data analysis with random forest," *Briefings in Bioinformatics*, vol. 24, no. 2, p. bbad002, Mar. 2023, doi: 10.1093/bib/bbad002.

[15] Y. Izza, A. Ignatiev, and J. Marques-Silva, "On Tackling Explanation Redundancy in Decision Trees," *Journal of Artificial Intelligence Research*, vol. 75, pp. 261–321, Sep. 2022, doi: 10.1613/jair.1.13575.

[16] Y. Li, F. Lu, and Y. Yin, "Applying logistic LASSO regression for the diagnosis of atypical Crohn's disease," *Sci Rep*, vol. 12, no. 1, p. 11340, Jul. 2022, doi: 10.1038/s41598-022-15609-5.

[17] S. Ghane, N. Bhorade, N. Chitre, B. Poyekar, R. Mote, and P. Topale, "Diabetes Prediction using Feature Extraction and Machine Learning Models," in *2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC)*, Aug. 2021, pp. 1652–1657. doi: 10.1109/ICESC51422.2021.9532818.

[18] A. Pati, M. Parhi, and B. K. Pattanayak, "A review on prediction of diabetes using machine learning and data mining classification techniques," *International Journal of Biomedical Engineering and Technology*, vol. 41, no. 1, pp. 83–109, Jan. 2023, doi: 10.1504/IJBET.2023.128514.

[19] B. Kerimkhan *et al.*, "Automation of flow analysis in scleral vessels based on descriptive-associative algorithms," *Sci Rep*, vol. 13, no. 1, p. 4650, Mar. 2023, doi: 10.1038/s41598-023-31866-4.

[20] A. A. Ahmed *et al.*, "Review on Hybrid Deep Learning Models for Enhancing Encryption Techniques Against Side Channel Attacks," *IEEE Access*, pp. 1–1, 2024, doi: 10.1109/ACCESS.2024.3431218.

[21] M. Hassan, M. A. Butt, and M. Z. Baba, "Logistic Regression Versus Neural Networks: The Best Accuracy in Prediction of Diabetes Disease," *Asian Journal of Computer Science and Technology*, vol. 6, no. 2, Art. no. 2, Sep. 2017, doi: 10.51983/ajcst-2017.6.2.1782.

[22] A. Mahabub, "A robust voting approach for diabetes prediction using traditional machine learning techniques," *SN Appl. Sci.*, vol. 1, no. 12, p. 1667, Nov. 2019, doi: 10.1007/s42452-019-1759-7.
[23] P. Piotrowski, D. Baczyński, M. Kopyt, and T. Gulczyński, "Advanced Ensemble Methods Using Machine Learning and Deep Learning for One-Day-Ahead Forecasts of Electric Energy Production in Wind Farms," *Energies*, vol. 15, no. 4, Art. no. 4, Jan. 2022, doi: 10.3390/en15041252.
[24] "Diabetes Dataset." Accessed: Aug. 06, 2024. [Online]. Available: https://www.kaggle.com/datasets/akshaydattatraykhare/diabetes-dataset
[25] S. Zhang, "Challenges in KNN Classification," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 10, pp. 4663–4675, Oct. 2022, doi: 10.1109/TKDE.2021.3049250.
[26] S. Zhang and J. Li, "KNN Classification With One-Step Computation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 3, pp. 2711–2723, Mar. 2023, doi: 10.1109/TKDE.2021.3119140.
[27] P. E. Kafrawy, H. Fathi, M. Qaraad, A. K. Kelany, and X. Chen, "An Efficient SVM-Based Feature Selection Model for Cancer Classification Using High-Dimensional Microarray Data," *IEEE Access*, vol. 9, pp. 155353–155369, 2021, doi: 10.1109/ACCESS.2021.3123090.
[28] S. Dong, "Multi class SVM algorithm with active learning for network traffic classification," *Expert Systems with Applications*, vol. 176, p. 114885, Aug. 2021, doi: 10.1016/j.eswa.2021.114885.
[29] A. B. Yilmaz, Y. S. Taspinar, and M. Koklu, "Classification of Malicious Android Applications Using Naive Bayes and Support Vector Machine Algorithms," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 10, no. 2, Art. no. 2, May 2022.
[30] S. Alex, K. J. Dhanaraj, and P. P. Deepthi, "Private and Energy-Efficient Decision Tree-Based Disease Detection for Resource-Constrained Medical Users in Mobile Healthcare Network," *IEEE Access*, vol. 10, pp. 17098–17112, 2022, doi: 10.1109/ACCESS.2022.3149771.
[31] H. Mahmoud, M. Thabet, M. H. Khafagy, and F. A. Omara, "Multiobjective Task Scheduling in Cloud Environment Using Decision Tree Algorithm," *IEEE Access*, vol. 10, pp. 36140–36151, 2022, doi: 10.1109/ACCESS.2022.3163273.
[32] L. Rzayeva *et al.*, "Enhancing LAN Failure Predictions with Decision Trees and SVMs: Methodology and Implementation," *Electronics*, vol. 12, no. 18, Art. no. 18, Jan. 2023, doi: 10.3390/electronics12183950.
[33] F. K. Alarfaj, I. Malik, H. U. Khan, N. Almusallam, M. Ramzan, and M. Ahmed, "Credit Card Fraud Detection Using State-of-the-Art Machine Learning and Deep Learning Algorithms," *IEEE Access*, vol. 10, pp. 39700–39715, 2022, doi: 10.1109/ACCESS.2022.3166891.
[34] A. Abdellatif, H. Abdellatef, J. Kanesan, C.-O. Chow, J. H. Chuah, and H. M. Gheni, "Improving the Heart Disease Detection and Patients' Survival Using Supervised Infinite Feature Selection and Improved Weighted Random Forest," *IEEE Access*, vol. 10, pp. 67363–67372, 2022, doi: 10.1109/ACCESS.2022.3185129.
[35] G. N. Ahmad, H. Fatima, S. Ullah, A. Salah Saidi, and Imdadullah, "Efficient Medical Diagnosis of Human Heart Diseases Using Machine Learning Techniques With and Without GridSearchCV," *IEEE Access*, vol. 10, pp. 80151–80173, 2022, doi: 10.1109/ACCESS.2022.3165792.