Unraveling Diabetes Detection: A Comparative Study of Machine Learning Approaches

***Abstract*—Looking at healthcare by utilizing machine learning will be at the heart of our job, with the main aim of recognizing early diabetes cases. Giving an early diagnosis of diabetes to prevent the growth of serious complications is clearly the main task to achieve for a thorough treatment of the problem. Diabetes is a chronic disease that is a global problem because it is a permanent disease. Therefore, diabetes is the leading cause of early death. We used machine learning and statistical methodologies to conduct a comparative evaluation of these modeling forecasts as an essential tool for diabetes diagnosis. While keeping the experimentation and evaluation rigorous, we study how successful the models are and also what their advantages and disadvantages are within the field of health care. Our purpose is to distribute the outcomes that model the implementations in a real-life healthcare system. Further highlighting this, it serves as the basis for making blood sugar diagnoses and therapy on an individual basis, which then results in the improvement of mental health. In completing these steps, therefore, personalized medicine may well provide a solution for diabetes treatment in the future.**

***keywords*—Diabetes, Machine learning, Comparative evaluation.**

I. Introduction

A. Background

Diabetes is a long-lasting metabolic disorder that manifests itself with hyperglycemia (excessive sugar) in the blood resulting from inadequate production or utilization of insulin or both. This disease that occurs in many people in the world threatens health by increasing the risk of diseases such as cardiovascular disease, renal damage, nerve damage, and eye impairment in particular if not treated properly. Type 1 diabetes usually emerges during childhood or the teen period. It does not get cured and is to be dealt with for a lifetime by using insulin therapy, but type 2 diabetes found more among adult people is associated with obesity and other lifestyle factors, and it can be cured either by using diet, exercise, or, in cases of need, medicine, and insulin therapy. Diabetes management requires monitoring blood sugar levels, strictly following a diet, regularly exercising, frequently injecting insulin or taking oral pills, hoping everything goes well, and not letting the disease worsen.

It is present that in the modern era, diabetes mellitus has become a rapidly rising kind of health problem that is mainly growing among the demographic communities. To allow for the proper intervention and to keep complications at bay, early and timely detection of diabetes is of great significance. Conventional diagnostic methods can be significantly imprecise in the assessment of individuals who could perhaps develop a disease, to the extent that new techniques should be found. Machine learning is currently gaining in popularity as a Machine learning approach that enables data-driven predictive patterns to generate diabetes diagnostics.

Based on the latest figures from the WHO, it is estimated that more than 422 million people in the world live with diabetes. Experts predict that the numbers will rise even more. One of the most essential tools for diabetes prevention and control is early diagnosis. This allows us to lower the effects of diabetes on both people and their healthcare systems. Nevertheless, the case is that some techniques of conventional diagnostics may not be precise enough to enable us to have a word about people who are already at risk. Machine learning provides a solution that enables, through an advanced process of data mining, the reduction of diagnostic time and the delivery of quality medical services. Algorithms of machine learning take into account a broad variety of factors, which include medical history, genetic propensity, style of life, and lab work, to determine who is likely to suffer from diabetes. With these algorithms at hand, they can detect multitudes of patterns and interconnections among huge resources of data, based on which they predict who may need proactive screening and intervention. What is more, machine learning empowers personalized medicine by providing doctors with patient-specific prognoses and a decision-making process that fits the specifics of the patients. Even though it may be seen as having multiple positive sides to it, machine learning, when it comes to the automated diagnosis of diabetes too, has its own set of challenges that come with it. Gaps in datasets, linkage of features, and interpretability of outcomes are among the main areas where the machine learning model should be improved in building robust models for diabetes detection. Researchers are dedicated to investigating different machine learning approaches, such as deep learning, which could be useful for successful diabetes detection problem-solving and the enhancement of efficiency and accuracy. Hence, it is time to move forward with a comparative study of machine learning models for the detection of diabetes in light of the above facts. This kind of study has been aimed at assessing the accuracy rates of different machine learning algorithms in effectively identifying diabetes-at-risk individuals. Sensitivity, specificity, and predictive value are the current evaluation criteria. This study endeavors to gain a comprehensive understanding of the advantages and drawbacks of the existing machine learning methods that aim at the detection of diabetes disease, on the grounds of which the latest techniques in disease detection are to be developed or pre-existing approaches are to be modified for the betterment of patient outcomes.

B. Motivation

Undertaking this research is motivated by the urgent need to overcome the barriers preventing diabetes diagnosis. Progress in health technology notwithstanding, the perennial concern of missed diagnosis and delayed detection continues to cost in terms of health effects and healthcare costs above the expected. Against this canvas, our research journey starts a mission to analyze and compare a variety of machine learning algorithms with the aim of discovering which one is the most successful in the search for diabetes. We aim to create a way to execute treatments that not only give higher accuracy but also are given at the proper time. In essence, our study looks to fill the gap between technology and health. Through the use of machine learning technology, our mission is to ensure that physicians are well-equipped with cutting-edge tools that make early disease identification and personalized care possible. This involves information hoarding and seeking correlations that might otherwise be overlooked, which is the strength of algorithms. In this regard, our initiative aims at delivering impactful knowledge, which will help the providers make appropriate choices. All in all, our journey is one of empowerment empowerment for clinicians to take the initiative, empowerment for patients to explore the path to wellness with vigor, and empowerment for humanity as a whole in its battle for victory over diabetes.

C. Contributions

This signature report makes its contribution to the healthcare analytics realm by providing a comprehensive evaluation of machine learning models for diabetes diagnosis. From a theoretical standpoint, the study not only explores the basics of the predictive algorithms that have been used but also validates their implementation using empirical data, thus capturing the performance of diverse models such as Naive Bayes, Support Vector Machines, Decision Trees, K-Nearest Neighbours, and Random Forest. We bring together findings obtained by analyzing real-world data and provide practice-centered advice, which may additionally be used to create evidence-informed guidelines. Moreover, in addition to the current study, our research is a basement for yet-to-be-investigated areas in predictive analytics, as this will lead to the establishment of innovations and advancements in diabetes management strategies. Along with the concreteness of certain types of our machine learning algorithms, it is important to highlight that they are both fully applicable and effective in the medical domain. This approach, therefore, ensures that the findings we get are both credible and sensible to make regarding the current developments in healthcare analytics.

II. Literature review

Sisodia *et al.* [**?**] deployed three machine learning algorithms to forecast the outbreak of diabetes. The authors have achieved 76.30% accuracy 6 by using the III-B1 Naive Bayes classifier with the Indian diabetes type II dataset.

Rastogi *et al.* [**?**] utilized four different machine-learning algorithms for the forecasting of diabetes. The scholars came across the fact that logistic regression was a more reliable tool than the other Machine Learning techniques, the accuracy 6 being 82.46% with the Kaggle dataset.

Daanouni *et al.* [**?**] implemented a machine learning algorithm on the dataset of the Pima Indian Tribe for the prediction of diabetes. This was realised by experimenting with neural networks of various kinds, like artificial neural networks, k-nearest neighbours III-B4, deep neural networks, and decision trees III-B3, on the data set. They have concluded that Deep Neural Network is the best algorithm, which is associated with the highest accuracy rate 6 of 90%.

Xue *et al.* [**?**] chooses Support Vector Machine III-B2, Naive Bayes III-B1, and LightGBM machine learning algorithms. The issue that researchers must bear in mind when they want to distinguish successful strategies of prediction is multiple types of exploring or using each others work, according to the authors. The Sylhet Diabetes Hospital from Bangladesh gave this dataset, and it was mentioned that this method is superior to other methods.

Dutta *et al.* [**?**]point out that the classifiers used constitute Naive Bayes III-B1, Random Forest, Decision Tree III-B3, XGBoost, and LightGBM. The most unmistakable way of presenting the finding is by employing logistic regression, which has a 96% record of accuracy 6 acquired from a just-released dataset in Bangladesh.

Zou *et al.* [**?**] believed that early detection of the disease could help physicians manage the patient's condition and improve clinical decision-making. In this case, he provided diabetes prediction in the model through RF, decision trees III-B3, and deep learning networks. The information furnished was gathered from the physical examination records of patients in a Chinese hospital called Luzhou. The two authors' studies in combination point to an accuracy rate 6 of Random Forest III-B5 standing at 80.84%

Soni *et al.* [**?**] research involved six machine learning models: k-Nearest Neighbours III-B4, Support Vector Machine III-B2, Decision Tree III-B3, Logistic Regression, Random Forest III-B5, and Gradient Boosting, which were used to predict diabetes in the case of the Pima Indian Diabetes dataset. The best-performing method is Random Forest III-B5, whose accuracy function 6was 77%. This can be noticed when the results are tested by the machine learning systems.

Muhammad *et al.* [**?**] presented in the year 2020 explore classifiers such as random forest III-B5, gradient boosting, k-Nearest Neighbours III-B4, Support Vector Machine III-B2, logistic regression, and Naive Bayes III-B1. Based on the Mohamed Specialist Hospitals data from Nigeria, the writer coined this name. Unlike algorithms such as Gradient Boosting (82%), k-Nearest Neighbours(75%) III-B4, Naive Bayes (86%) III-B1, Logistic Regression (90%), and Random Forest (90%) III-B5, we conclude that random forest's III-B5 still the one with the highest percent of accuracy 6(87%).

K. VijiyaKumar *et al.* [**?**] used models of ML to make predictions on diabetes. The Random Forest Classifier III-B5 was developed, and its accuracy 6 was tested using type 2 diabetes. High accuracy 6 is within 90% (an approximation).

The fact that our approach showed better performance than other machine learning algorithms for diabetes prediction is especially worth mentioning.

L. V. R. Kumari *et al.* [**?**] used machine learning methods to improve early diabetes prediction accuracy. They have incorporated algorithms like Naive Bayes III-B1, K-Nearest Neighbor III-B4, Logistic Regression and Random Forest III-B5. Among these algorithms, k-Nearest Neighbours III-B4algorithm gives a high accuracy 6of 78.57%.

N. Fazakis *et al.* [**?**] developed a model that used Machine Learning technologies for the prediction of type 2 diabetes. With an Area Under the ROC Curve (AUC) of 0.884, the ensemble WeightedVotingLRRFs machine learning model is suggested as a way to enhance diabetes prediction. With respect to the weighted voting. Additionally, a comparison of various machine learning models that use both inductive and transductive learning, as well as the Leicester and Finnish Diabetes Risk Score systems, is offered.The research analyzed data from the English Longitudinal Study of Ageing (ELSA) database.

M. A. R. Refat *et al.* [**?**], in this analysis, they explore different machine learning and deep learning methods for early prediction of diabetic illness. A diabetic dataset with 17 features, including the class label is used, from the UCI repository. They evaluated the performance of various classification techniques using different metrics. The results indicate that while most algorithms achieved around 90% accuracy 6, the XGBoost classifier demonstrated significantly better performance, reaching almost 100% accuracy 6.

III. Proposed Work

The problem is the existence of diabetes, which has been demonstrated through the deaths of many people, yet it is a disease that has not been awarded a perfect cure. Sometimes diabetes is discovered too late after many years, and the patient may live with its symptoms without even knowing about it. A lot of progress has been made in building the system to diagnose diabetes, and indeed, more efforts have become relevant with the ever-evolving artificial intelligence. Specifically, our study focuses on the implementation of five mainstream Machine-learning algorithms on three repositories of datasets and the inspection of accuracy performance indicators for all the Machine Learning algorithms on all three datasets.

A. Preprocessing

1) Data Collection: Data collection is the essential part number one and the most important step towards building any machine learning model. The accuracy score 6 of the algorithms used are derived from the sample used. The data provided by the three datasets on diabetes at Kaggle has been collected, and we trained our model on this data to predict whether someone has diabetes or not.

2) Data Standardization: From the sklearn collection, we employed the StandardScaler object from the preprocessing module. For this, it cuts down on standardization1by taking the mean out and dividing it through to get a unit variance.

$$z = \frac{x - \mu}{\sigma} \tag{1}$$

If x denotes the data point, μ represents the mean of the dataset, and σ signify the Standard deviation of the dataset.

A fit method of the sklearn library is used to obtain the mean and standard deviation of the data. The process then involves the use of the transformation method, with the mean and standard deviation from previously calculated values deducted from it.

3) Data Splitting: The generic data has been split into train and test data where the split uses the 80-20 rule therefore, 20% of the data is the testing and 80% of the data for training.

B. Models Used

1) Naive Bayes: In the present study, the Gaussian-Naive Bayes variation of the Naive Bayes classifier is utilized, as it fits the continuous data well and uses the idea that all features have a normal distribution. In light of the strict conditionality of independent features, the probabilistic classification classifier Naive Bayes applies the Bayes theorem. The algorithm hypothesizes the presence of one feature of a certain group as unbiased and independent of the other features; that is why it is named Naive Bayes. Spreadsheets may be considered a simple practice to calculate and forecast in due time. When we want an immediate, responsive action, this is when the algorithm tends to be used more often. The model's performance is considered by determining its accuracy 6and drawing a confusion matrix I. The mathematical formula for Bayes Theorem is:

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(B)}{P(A)} \tag{2}$$

Where:
$P(A \mid B)$ is the posterior probability of class (A) given predictor (B).
$P(A)$ and $P(B)$ are the probabilities of class and predictor respectively.
$P(B \mid A)$ is the likelihood which is the probability of the predictor given class.

2) Support Vector Machine: It can be used for linear and non-linear classification, regression, and also for informing about the outliers. The primary goal of this strategy is to define a hyperplane that separates N- dimensional space into distinct classes. The algorithm tries to find such a hyperplane that shows the most succeeding separation margin between two classes. It decides on a hyperplane whose distance from it to the nearest data point on both sides is maximum. And if such a hyperplane is present it can be called the maximum margin hyperplane. SVMs are, in general, designed to perform a lot of different tasks such as image detection, spam detection, face detection, etc.

3) Decision Trees: This algorithm decides choices based on questions and a criteria framework. The approach is just like a flowchart, where a node means a feature, a branch means decision, and a leaf points to the result.
Components of a Decision Tree:-

a) Root Node: This node extends from the top or starting point of the decision tree, where all the features are divided into different branches based on various conditions.

b) Decision Nodes: These nodes, which are also called decision nodes, are constructed by splitting the root node. They represent intermediate decisions.

c) Leaf Nodes: The last nodes of the tree from where even further splitting does not happen. Furthermore, they give concrete results.

The algorithm is composed of recursion in such a way that it rives the data set into minute subsets. Each node is evaluated, and the algorithm determines which feature to split the data. In our algorithm, the maximum depth of 2 means we will use a maximum of 2 levels in our tree. Moreover, this reduces the model's tendency to pick up only specific data patterns, known as overfitting, ensuring that the resultant tree is not overly complex. The criterion used is 'entropy'3 and it is calculated as

$$E(S) = -\sum_{i=1}^{n} p_{C_i} \log_2(p_{C_i}) \tag{3}$$

Where $E(S)$ is the entropy3 of dataset S, p_{C_i} is the proportion of instances in S that belong to class C_i and n is the number of classes. It is used for the information gain4. Information gain is the reduction in entropy or surprise by splitting a feature and it is calculated as

$$Gain(S, A) = E(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} E(S_v) \tag{4}$$

where S is the total sample space, A is the attribute, $Values(A)$ represent the set of all possible values of attribute A, and S_v represents the subset of S for which attribute A has value v.

4) k-Nearest Neighbours (kNN): The approach is fairly easy, straightforward, and accurate based on the samples that are as close or the closest one to the reference as possible, known as the neighbors, where the closeness is calculated using Euclidean Distance 5, Manhattan Distance, etc. The specific model will assess the five nearest neighbors for its predictions (k = 5) and will use Euclidean distance 5 to do so.

$$d(p, q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \ldots + (p_n - q_n)^2} \tag{5}$$

Where p,q are the two points in the data collection. p_i and q_i are the i^{th} coordinates respectively, and n is the number of dimensions or features in the dataset.
It does not require any assumptions about the data, this helps the algorithms to adapt to different patterns and make predictions based on the local structure of data. The choice of 'k' significantly influences the performance of the algorithm. A small value of k can make the model prone to overfitting. On the other hand, a large value of k can increase the bias resulting in high test error. Moreover, the value of k should also be an odd number as it ensures that there will always be a majority class hence preventing a tie.

5) Random Forest: It is a powerful machine-learning algorithm that works by creating several decision trees III-B3during the training phase. While predicting, the algorithm totals the results of all trees; for classification tasks, the preferred method of decision-making is through voting, while for regression tasks, the preferred method is through averaging. It predicts the output with high accuracy even for large datasets. Random Forests are widely used for classification, regression, and reduced overfitting. In our model we designed a Random Forest classifier with estimators=100 to create a maximum of 100 trees, max_depth=100 which makes sure that the maximum depth of any tree generated is not more than 100, min_samples_split=20 which specifies the minimum number of samples required to split an internal node which here is set to 20, the min_samples_leaf which defines the minimum number of samples required for a node to be considered a leaf is one, max_features is set to the square root of the number of features required to consider when looking for the best split. The random state is set to 5, which makes the output deterministic.

IV. Experiment

A. Dataset

A large dataset is needed to undertake a comparative study of various features and models. Here, we have used three datasets from different sources and with different numbers of features in the data set. We didn't create the dataset we used for our research; we have sourced the data sets from Kaggle, making them more resilient and versatile.

1) Dataset_1 (Diabetes Dataset): This Dataset was sourced from Kaggle and has 769 entries and 9 features namely Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, and Outcome.

2) Dataset_2 (Diabetes Prediction Dataset): This Dataset was sourced from Kaggle and has 100001 entries and 9 features namely Gender, Age, hypertension, heart_disease, smoking_history, bmi, HbA1c_level, blood_glucose_level, and diabetes.

3) Dataset_3 (Diabetes Health Prediction Dataset): This Dataset was sourced from Kaggle and has 253681 entries and 22 features namely Diabetes_012, HighBP, HighChol, CholCheck, BMI, Smoker, Stroke, HeartDiseaseorAttack, PhysActivity, Fruits, Veggies, HvyAlcoholConsump, AnyHealthcare, NoDocbcCost, GenHlth, MentHlth, PhysHlth, DiffWalk, Sex, Age, Education and Income.

B. Experiment setting

In this study, we are comparing machine-learning algorithms on various datasets to predict diabetes.
The initial step involves the Standardisation and Label encoding of the datasets, this is achieved using StandardScaler() and LabelEncoder() function. Following this, the datasets are split using an 80-20 split i.e. 80% of the data is used to train the model, and the rest is used to test the model to calculate the accuracy.
Subsequently, the existing machine-learning models are applied to the datasets, and their performance is evaluated using the accuracy score 6 and confusion matrix I that are formulated using

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \tag{6}$$

and

Table I
CONFUSION MATRIX

	Predicted	**Predicted**
Actual	True Positive (TP)	False Negative (FN)
Actual	False Positive (FP)	True Negative (TN)

Where
True Positives (TP): When we predicted that an event happened and it did occur, those are considered True Positives (TP).
True Negatives (TN): We predicted that there would be no event, and indeed, no event occurred.
False Positives (FP): We predicted yes, but the event didnt happen. Also known as Type I error.
False Negatives (FN): We predicted no, but the event did happen. Also known as Type II error.
Subsequently, the accuracy scores derived from these models are represented in a barplot, providing a visual interpretation of the model's performance. This graphical representation facilitates a more intuitive understanding of the distribution and variance of accuracy scores across different models.

V. RESULTS AND DISCUSSION

A. Analysis of Algorithms on the Datasets

In this section, we compared the accuracy of each algorithm using five models: Naive Bayes III-B1, Support Vector Machine III-B2, Decision Tree III-B3, k-Nearest Neighbours III-B4, and Random Forest III-B5 on three different datasets.

1) Dataset_1(Diabetes Dataset): For this dataset, Random Forest Classifier III-B5 was the best fit for predicting Diabetes with its accuracy score 6of 81.2% and confusion matrix I as

Table II
CONFUSION MATRIX FOR DATASET_1

	Predicted Diabetes	Predicted Not Diabetes
Actually Diabetes	86	13
Actually Not Diabetes	16	39

2) Dataset_2 (Diabetes Prediction Dataset): For this dataset, Decision Tree III-B3 was the best fit for predicting Diabetes with its accuracy score 6 of 97.2% and confusion matrix I as

Table III
CONFUSION MATRIX FOR DATASET_2

	Predicted Diabetes	Predicted Not Diabetes
Actually Diabetes	18299	0
Actually Not Diabetes	551	1150

3) Dataset_3 (Diabetes Health Prediction Dataset): For this dataset, Decision Tree III-B3 and Support Vector Machine III-B2 are the best fits for predicting Diabetes with their accuracy score 6 of 86.1% and their confusion matrix I as

Table IV
CONFUSION MATRIX FOR DATASET_3

	Predicted Diabetes	Predicted Not Diabetes
Actually Diabetes	43671	0
Actually Not Diabetes	7065	0

Table V
ACCURACY SCORES ON ALL THE DATASETS

	KNN	**SVM**	**NB**	**RFC**	**DC**
Dataset_1	0.805195	0.779221	0.772727	0.811688	0.733766
Dataset_2	0.96110	0.96130	0.90380	0.97055	0.97245
Dataset_3	0.847406	0.860750	0.774539	0.859094	0.860750

4) Analysis of implemented algorithms on the datasets: We can observe the performance of five different machine learning models k-Nearest Neighbors (kNN) III-B4, Support Vector Machine III-B2, Naive Bayes III-B1, Random Forest Classifier III-B5, and Decision Tree III-B3 on three different datasets (Dataset_1 IV-A1, Dataset_2 IV-A2, Dataset_3 IV-A3. kNN III-B4 performs consistently across all datasets with the highest performance on Dataset_2 IV-A2.
Support Vector Machines III-B2 also shows consistent performance across all datasets, slightly outperforming kNN III-B4 on Dataset_2 IV-A2.
Naive Bayes III-B1 has the lowest performance on Dataset_2 IV-A2 and Dataset_3 IV-A3 compared to the other models. However, its performance on Dataset_1 IV-A1 is comparable to that of Support Vector Machine III-B2. Random Forest Classifier III-B5 performs well on all datasets, with its best performance on Dataset_2 IV-A2. Decision Tree III-B3 has the highest performance on Dataset_2 IV-A2 and Dataset_3 IV-A3, but its accuracy drops on Dataset_1 IV-A1. These observations can help in choosing the right model based on the dataset at hand. However, its important to consider other

factors such as the nature of the data, the interpretability of the model, and the computational resources available.

VI. Graphical Comparison of Classification Models on the Datasets

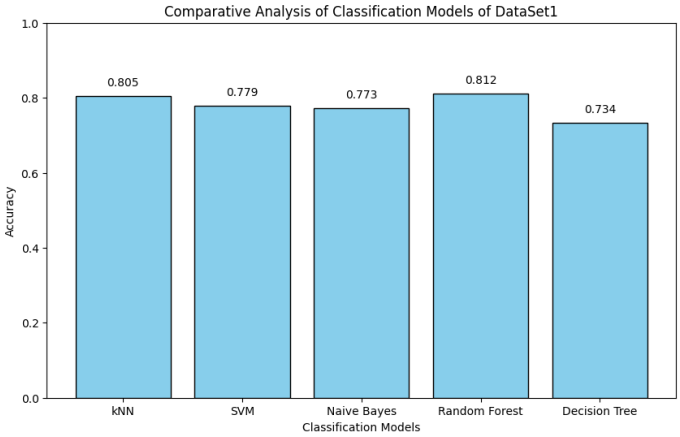

Figure 1. Analysis of all algorithms on Dataset_1

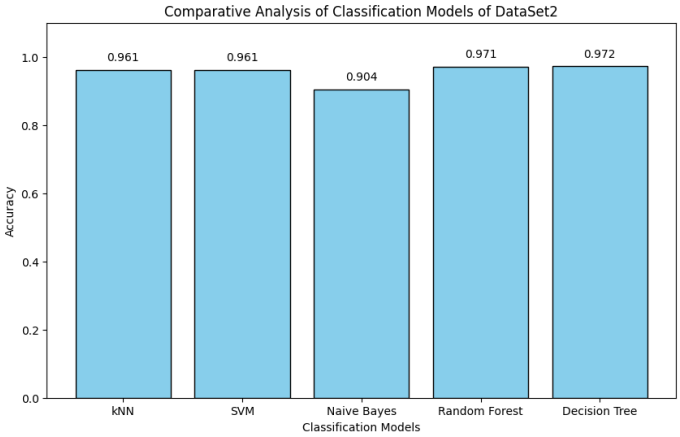

Figure 2. Analysis of all algorithms on Dataset_2

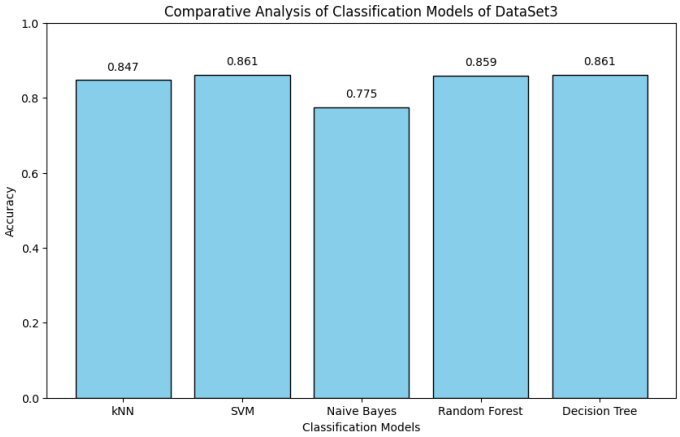

Figure 3. Analysis of all algorithms on Dataset_3

VII. Conclusion and Future Work

It has been identified that this study critically analyzes five machine learning techniques used in predicting diabetes: K-Nearest Neighbors III-B4, Decision Tree III-B3, Random Forest III-B5, Support Vector Machine III-B2, and Naive Bayes III-B1. Hence, we can deduce the effectiveness as well as the performance of these techniques through the application of them to the diagnosis of diabetes. Lastly, our analysis brings out the Random Forest Classifier III-B1 as the most probable model for diabetes prediction since it showed better predictive accuracy and robustness compared to the other models. However, the outcomes of each model were different when the overall scores were measured, which means that diabetic predictors should be chosen based on various parameters.

In addition to this, this research also contributes to the whole knowledge base regarding diabetes prediction by showing how the Machine Learning approaches could be applied to improve early detection and management of this disease. Similarly, these models will assist healthcare workers with better choices and interventions to reduce the likelihood of diabetes complications.

In the near future, the use of machine learning algorithms will enable us to envision an era of diabetes prediction. To boost forecast capability as well as resilience, one can start with the ensemble learning approach, which involves a manifold of models with the best features. Boosting model performances and generalizations through ensemble methods such as stacking, boosting, and bagging has been shown to be very successful in several instances. By including and combining development-specific attributes and data sources like environmental factors, lifestyle habits, or genetic biomarkers, the pathophysiology of diabetes can be elucidated. Machine learning algorithms can process complex relationships and patterns thanks to finding meaning and interconnecting different data sets. It enables advanced forecasting and tailored measures to be used.

References