Artificial Intelligence Approaches for Predicting Diabetes in Egypt

Ayah H. Elsheikh[1], Hossam A. Ghazi[2], Nancy Awadallah Awad [3]

[1]Faculty of Information systems and computers, Sadat academy for management sciences, Cairo, Egypt

[2]Assistant Professor of Internal Medicine, Mansoura Faculty of Medicine, Mansoura, Egypt

[3]Department of Computer and Information Systems, Sadat Academy for Management Sciences, Cairo, Egypt

Abstract

One major public health concern in Egypt is the increasing incidence of diabetes mellitus. It is essential to recognize problems early and treat them effectively [1]. This work applies several machine learning methods to predict diabetes risk using a dataset from Egyptian diabetes and endocrinology clinics. Features including age, BMI, medical history, and other health markers are included in the dataset. Using performance criteria such as confusion matrix, F1-score, recall, accuracy, and precision, we assessed various models including K-Neighbors, Gaussian Naive Bayes, Bernoulli Naive Bayes, Extra Trees, SVC, and Logistic Regression. The findings indicate that diabetes can be accurately predicted using machine learning. Logistic Regression, with a cross-validated accuracy of 0.965, test accuracy of 0.957, precision of 0.94, recall of 0.90, and an F1-score of 0.92, proved to be the most effective model for this dataset.

Keywords

Machine learning, Logistic Regression, Diabetes, Support Vector Classifier, Egypt

1. Introduction

The goal of the large discipline of computer science known as artificial intelligence (AI) is to build machines that are able to carry out activities that normally demand intellect similar to that of humans. Among these tasks include pattern recognition, reasoning, problem solving, and comprehension of natural language[2].It includes a range of methods and tools designed to allow machines to simulate certain parts of human thought processes. Machine Learning (ML) is a crucial subfield of AI that focuses on creating algorithms that enable computers to make judgments or predictions based on the data they are given. In contrast to conventional AI sys-

DOI:10.5121/ijaia.2024.15501

tems, which depend on well defined rules, machine learning algorithms enhance their functionality via experience. In order to do this, big datasets must be trained with models in order to find trends and make data-driven judgments without the need to manually program each unique activity[3].

In general, ML approaches fall into two categories: unsupervised learning, which finds hidden patterns in unlabeled data, and supervised learning, which builds models based on labeled data. These methods are allowing sophisticated data analysis and automation, which is driving substantial innovation and increasing efficiency across a range of sectors[4]. With its high prevalence and rising incidence rates, diabetes mellitus poses a serious and growing public health threat in Egypt. This long-term condition, which is characterized by high blood sugar levels because of insufficient insulin synthesis or usage, can lead to serious side effects such as renal failure, neuropathy, retinopathy, and cardiovascular disease. These issues seriously lower people's quality of life and put a heavy strain on the healthcare system[5].A multitude of variables, including changes in lifestyle, bad eating habits, insufficient physical exercise, and genetic susceptibility, have been connected to the increased incidence of diabetes cases in Egypt. To limit the burden of the disease, better solutions for early identification and care are needed to address this expanding health concern. Using machine learning to anticipate diabetes is a potential way to address this growing health problem in Egypt[6].ML algorithms are capable of precisely identifying risk variables and predicting the probability of acquiring diabetes through the analysis of large datasets. By using early intervention measures, healthcare workers might possibly avoid or postpone the beginning of the disease thanks to this predictive capabilities. This paper suggests applying ML techniques to learn from data trends in order to identify diabetes early on. The algorithm analyzes a number of health markers, such as age, blood pressure, body mass index (BMI), and medical history, to identify those who are at high risk for diabetes. Its goal is to attain high prediction accuracy. By facilitating prompt and individualized medical care, this proactive strategy enhances patient outcomes and lessens the overall burden on healthcare systems.For early identification and control of diabetes, AI and ML applications are critical, especially in places like Egypt where the condition is becoming more common. Public health outcomes may be improved by utilizing these cutting-edge technologies to support successful treatments, increase prediction accuracy, and improve overall health.

2. RELATED WORK

Several recent studies have focused on Artificial intelligence technique for Diabetic Prediction. Maniruzzaman et al.[7] used classification techniques like LR-RF combination for feature selection, NB, DT, RF, AdaBoost considering the evaluative measures such as accuracy and Area under the ROC Curve on National Health and Nutrition Examination Survey dataset, and concluded accuracy 94.25%. K. Hasan and et al.[8]there purpose was To put forward a robust framework for predicting diabetes ,the classefire was used is SVM, KNN, DT, MLP, NB, AdaBoost, XGBoost on PIDD and the final result ACC achieved was 78.9% by using AdaBoost. S. Kumari et al.[9]depended on Improve the accuracy of prediction of diabetes mellitus using a combination of machine learning techniques by using NB, RF and LR algorithms on

PIDD dataset and the evaluative measures was ACC, Precision, Recall, F1-score, AUC, in the end 79.08% accurate results. P. Rajendra et al. [10] the Purpose is to Create a prediction model and investigate many methods to make performance better and accuracy by Linear regression(LR)algorithm on two datasets PIDD and Vanderbilt the evaluative measures Precision, Recall, F1-score after using the algorithm on them the result become 78% accuracy for Dataset 1, 93% accuracy for Dataset 2. Raja Krishnamoorthi and et al.[11] the main object is Unique intelligent diabetes mellitus prediction framework (IDMPF) is developed using machine learning Algorithms, LR, RF, SVM, and KNN on PIDD dataset there Validation Parameter Accuracy and LR high Accuracy 86Raghavendran et al .[12]Analyze a patient dataset to determine the probability of type 2 diabetes by LR, KNN, RF, SVM, NB, AdaBoost Algorithms on PIDD dataset this result conclude AdaBoost performs well 95Salliah Shafi Bhat and et al.[13] compares alot of classification models based on machine learning algorithms for predicting a patients' diabetic condition at the earliest feasible stage using RF, MLP, SVM, DT, GBC, and LR al gorithms on dataset gathered from a doctor in the Indian district of Bandipora in the years April 2021–Feb2022 .the result was RF has the highest accuracy of 98%. Aishwariya Dutta and et al.[14].Employing ML-based ensemble model, in which preprocessing plays a critical role in ensuring robust and accurate prediction, enabled this research to achieve its goal of making an early prediction of diabetes using NB, RF,DT ,XGB and LGB Algorithms on DDC dataset that was introduced from the South Asian country of Bangladesh (2011 and 2017–2018) . Validation Parameters is Auc,Acc the Results is Accuracy 0.735%and AUC0.832%. Jashwanth Reddy et al .[15] in 2022 there purpose is To design an accurate mode for predicting human diabetes using machine learning algorithms like SVM, KNN, LR, NB, GB and RF on also PIDD Dataset there Validation Parameters ACC, ROC, Precision, Recall and FM. The Result Was ACC 80% using RF . Chatrati et al.[16] used classification techniques like SVM, KNN, DTand LR, considering the evaluative measures such as accuracy on PID Ddatase,and concluded that the accuracy for SVM achieve 75% as higher accuracy. Muhammad Exell Febrian and et al.[17] Making an artificial intelligent model that can predict diabetes diseas by k-nn and native bayes Algorithms on PIDD data set and the Accuracy for naive bayes was 76.07Chun-Yang Chou and et al.[18]this study used Microsoft Machine Learning Studio to train the models of various kinds of neural networks, and the prediction results were used to compare the predictive ability of the various parameters for diabetes. There use two-Class Logistic Regression, Two-Class Neural Network, TwoClass Decision Jungle, Two-Class Boosted Decision Tree on the collected data from tests on the patients in the past two years were used as predictors of the models. Validation Parameters are True Positive ,False Positive ,False Negative True Negative, Accuracy ,Precision Recall, F1 Score, AUC. Result was 95,3% Acc for two-class boosted decision tree .

Table 1. Summarized comparison of related work, including Researchers, Year of research, Dataset, Validation Parameters, and Results.

Researchers/Year	**Datasets**	**Validation Parameters**	**Results**
Maniruzzaman and et al. 2020[7]	National Health and Nutrition Examination Survey	ACC, AUC	ACC 94.25%
K. Hasan and et al.[8]	PIMA dataset	ACC	ACC achieved was 78.9% by using AdaBoost
Md. Mehedi Hassan and et al., 2021[19]	Collected from Shaheed Sheikh Abu Naser Specialized Hospital, Khulna	ACC	ACC for Random Forest 97.5%
S. Kumari and et al., 2021[9]	PIDD	ACC, Precision, Recall, F1-score, AUC	79.08% accurate results on PIMA dataset
P. Rajendra and et al., 2021[10]	PIDD and Vanderbilt	Precision, Recall, F1-score	78% accuracy for Dataset 1, 93% accuracy for Dataset 2
C. Yadav and et al., 2021[20]	UCI repository	ACC, Recall, Precision, F1-score	ACC for Bagging ensemble methods was 98%
Muhammad Exell Febrian and et al., 2022[17]	PIDD	ACC	ACC for naive bayes was 76.07%
Raja Krishnamoorthi and et al., 2022[11]	PIDD	ACC	LR high Acc 86%
Jashwanth Reddy and et al., 2022[15]	PIDD	ACC, ROC, Precision, Recall, FM	ACC 80% using RF
Raghavendran and et al., 2022[12]	PIDD	ACC, Precision, Recall, F1-Score, CM	AdaBoost performs well 95%
Salliah Shafi Bhat and et al., 2022[13]	Dataset gathered from an Indian doctor lives in Bandipora in the years April 2021–Feb 2022	ROC Area, Recall, Precision, F-measure, and MCC. K-fold	RF has the highest accuracy of 98%
Aishwariya Dutta and et al., 2022[14]	DDC dataset from the South Asian country of Bangladesh (2011 and 2017–2018)	AUC, ACC	Accuracy 73.5% and AUC 0.832
Chun-Yang Chou and et al.,2023[18]	Collected data from tests on patients over the past two years	True Positive, False Positive, False Negative, True Negative, Accuracy, Precision, Recall, F1 Score, AUC	ACC for two-class boosted decision tree was 95.3%

3. The Proposed Framework

A number of crucial processes are included in the framework that has been developed for the purpose of predicting diabetes using machine learning algorithms: data collection, preprocessing, exploratory data analysis (EDA), dividing the dataset, re-

sampling, model selection, and model assessment. Every step in the framework is explained in depth in the sections that follow.

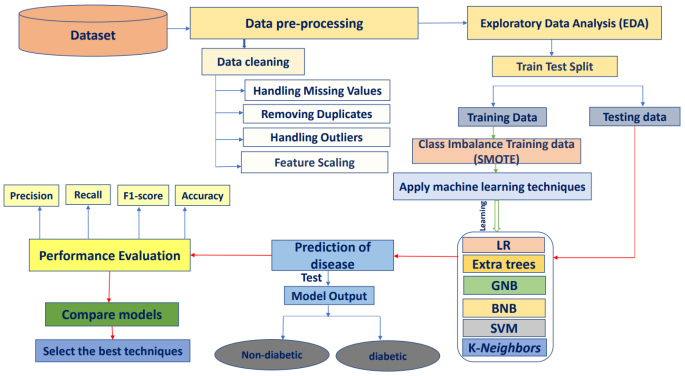

Figure 1: Framework for predicting diabetes using machine learning algorithms.

3.1. Dataset

The dataset included in this paper is inverted from the clinics of Dr. Hossam Arafa, an endocrinologist and diabetic specialist in Egypt. The dataset contains a number of variables that are useful in predicting diabetes, including clinical measures, medical history, and particular symptoms. The dataset included 10,000 patient records in it at first. Following extensive preprocessing, which involved procedures for cleaning and preparation, the dataset was narrowed down to contain 5790 patients. The variables in this dataset include previous surgical history, COVID-19 status, hypertension, obesity, tiredness, dyspnea, thyroid disorders (Primary Hypothyroidism), fatigue, BMI, blood pressure (both systolic and diastolic), and glycated hemoglobin levels. If a patient has diabetes is indicated by the outcome variable. This improved dataset offers a thorough foundation for comprehending the variables linked to diabetes, making it easier to use and contrast different machine learning algorithms to forecast diabetes outcomes.

Feature	Description	Values/Range
Age	Age of the patient	3 to 87
Past Surgical	History of past surgeries	0 = no, 1 = yes
Covid	Whether the patient had COVID-19	0 = no, 1 = yes
HTN (Hypertension)	Presence of hypertension	0 = no, 1 = yes
Primary Hypothyroidism	Presence of primary hypothyroidism	0 = no, 1 = yes
Obesity	Whether the patient is obese	0 = no, 1 = yes
Sense of Lump	Whether the patient has a sense of lump in the body	0 = no, 1 = yes
Dyspnea	Difficulty or labored breathing	0 = no, 1 = yes
Fatigue and Dizziness	Presence of both fatigue and dizziness	0 = no, 1 = yes
BMI (Body Mass Index)	Body mass index of the patient	1 to 97
Blood pressure up	Systolic blood pressure	70 to 220
Blood pressure down	Diastolic blood pressure	40 to 120
Glycated hemoglobin	Level of glycated hemoglobin	3.6 to 17.2
Outcome	Target variable indicating diabetes presence	0 = non-diabetic, 1 = diabetic

Table 1: The dataset features are listed along with their description and values to provide a comprehensive understanding of each feature.

3.2. Data Pre-processing

The data preprocessing steps are crucial in preparing the dataset for machine learning model training and evaluation.The preprocessing steps performed are detailed below:

3.2.1. Data Cleaning

- **Handling Missing Values:** The dataset's missing values were eliminated in order to preserve the analysis's correctness and integrity. This methodology guarantees the utilization of only complete data points, hence mitigating the risk of biases or mistakes that may result from incomplete data.

- **Correcting Errors:**Errors in data input and discrepancies in category vari-

ables were found and fixed. In order to ensure uniformity in data representation and check for inaccurate labels—both of which are essential for successful analysis and modeling—this approach includes.

- **Removing Duplicates:** In order to avoid redundancy and guarantee that every data point in the dataset is distinct, duplicate records were eliminated. This process aids in preserving the dataset's dependability and quality.
- **Handling Outliers:**The Interquartile Range (IQR) approach was used to find and eliminate outliers in important characteristics including BMI and glycated hemoglobin. The IQR is computed as follows:

$$\text{IQR} = Q3 - Q1$$

where $Q1$ and $Q3$ are the first and third quartiles, respectively. Outliers are defined as data points outside the range:

$$[Q1 - 1.5 \times \text{IQR}, Q3 + 1.5 \times \text{IQR}]$$

- **Feature Scaling:**The 'StandardScaler' was utilized for feature scaling in order to standardize the dataset. This method uses the following formula to modify each feature, giving it a mean of (0) and a standard deviation of 1. using the formula:

$$X_{\text{scaled}} = \frac{X - \mu}{\sigma}$$

where X represents the original value, μ is the mean of the feature, and σ is the standard deviation. This process ensures that all features contribute equally to the model's learning process and improves the efficiency and convergence of gradient-based optimization algorithms during model training.

3.3. Exploratory Data Analysis (EDA)

We examine the dataset's feature distribution and examine any correlations between the characteristics in this part. Understanding the links and patterns in the data is crucial for influencing the predictive modeling process, and this study sheds light on those linkages and patterns.

3.3.1. Distribution of Features and Correlation Analysis

Age:

- **Distribution:** The dataset is right-skewed, with older patients showing higher rates of adverse health outcomes.
- **Correlation:** Age strongly correlates with hypertension ($r = 0.55$) and moderately with adverse health outcomes ($r = 0.46$).

BMI (Body Mass Index):

- **Distribution:** Higher BMI is linked to adverse health outcomes.
- **Correlation:** BMI shows moderate correlations with systolic ($r = 0.29$) and diastolic blood pressure ($r = 0.31$), and HbA1c ($r = 0.18$).

Blood Pressure (Systolic and Diastolic):

- **Distribution:** Patients with adverse outcomes have higher blood pressure.
- **Correlation:** Systolic and diastolic pressures are strongly correlated ($r = 0.94$) and moderately with adverse outcomes ($r = 0.35$ each).

Glycated Hemoglobin (HbA1c):

- **Distribution:** Elevated HbA1c indicates poor glucose control.
- **Correlation:** HbA1c strongly correlates with adverse outcomes ($r = 0.85$).

COVID-19: The study found that a history of COVID-19 among participants had a negligible direct impact on the adverse health outcome being studied.

- **Distribution:** The distribution of COVID-19 history across various demographics and health statuses did not reveal significant patterns.
- **Correlation:** Correlation analysis showed very weak associations between COVID-19 and the outcome ($r = 0.096$), as well as minimal correlations with age ($r = 0.071$), hypertension ($r = 0.05$), and glycated hemoglobin ($r = 0.085$). This suggests that COVID-19 history has minimal influence on these health variables.

Hypertension (HTN):

- **Distribution:** Hypertension is common in adverse outcomes.
- **Correlation:** Moderate correlation with adverse outcomes ($r = 0.40$).

Primary Hypothyroidism:

- **Distribution:** Low prevalence in adverse outcomes.
- **Correlation:** Weak negative correlation with adverse outcomes ($r = -0.17$).

Sense of Lump:

- **Distribution:** Rare but more frequent in adverse outcomes.
- **Correlation:** Weak negative correlation with adverse outcomes ($r = -0.16$).

Fatigue and Dizziness:

- **Distribution:** More common in adverse outcomes.
- **Correlation:** Weak negative correlation with adverse outcomes ($r = -0.095$).

Dyspnea:

- **Distribution:** More frequent in adverse outcomes.
- **Correlation:** Weak negative correlation with adverse outcomes ($r = -0.048$).

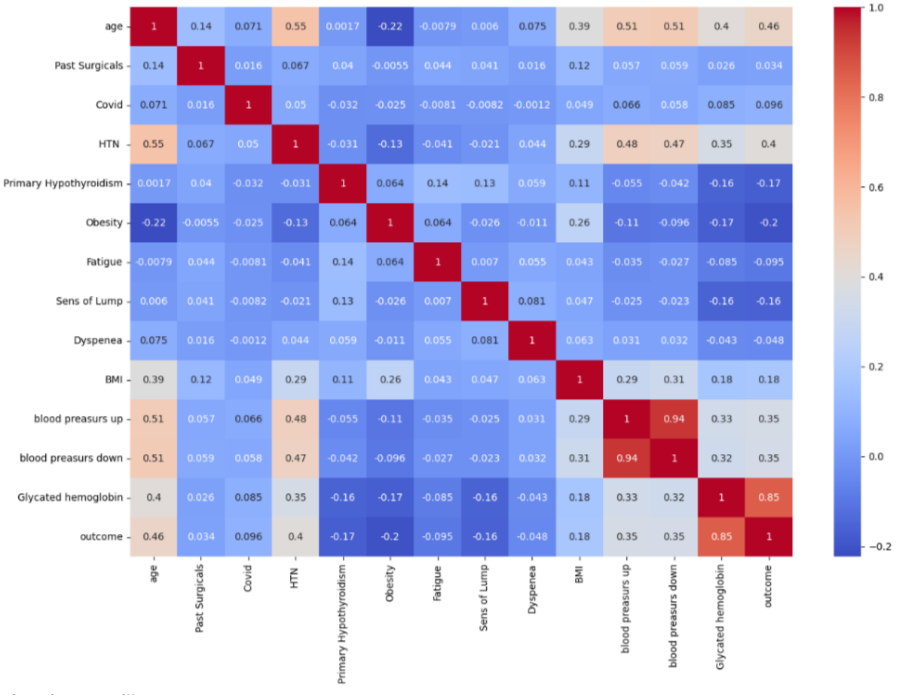

Figure 2: Correlation Matrix of Features

3.4. Splitting Data into Training and Testing

To assess the performance of the model, the dataset was divided into training (80%) and testing (20%) groups.In order to offer an objective assessment of the model's accuracy, this stage makes sure that it is evaluated on untested data.

3.5. Resample Training Dataset

The outcome feature's original distribution showed a significant imbalance, with 4216 cases of non-diabetes and 1574 cases of diabetes (1). As a result of this imbalance, the model may function biasedly, favoring the majority class prediction. The training dataset was resampled using the Synthetic Minority Over-sampling Technique (SMOTE) in order to correct the class imbalance. In order to guarantee that the model is trained on a balanced dataset, this approach creates fake examples for the minority class. The distribution of the outcome feature became balanced after SMOTE was applied to the training dataset, yielding equal numbers of examples of both classes (3379 for each class). The mathematical representation of the SMOTE algorithm is as follows:

$$X_{\text{new}} = X_i + (X_j - X_i) \cdot \delta \quad \text{where} \quad X_i, X_j \in \text{Minority Class}, \quad \delta \sim U(0,1)$$

where:

- X_{new}: The generated synthetic sample.
- X_i: A randomly selected minority class sample.
- X_j: Another randomly selected minority class sample from the k-nearest neighbors of X_i.
- δ: A random number between 0 and 1, drawn from a uniform distribution $U(0,1)$.

This ensures that the model is trained on a balanced dataset, which can improve its ability to generalize and perform well on both classes.[21]

3.6. Apply Machine Learning Techniques

The pre-processed and resampled dataset is subjected to many machine learning methods. This framework takes into account the following algorithms: Support Vector Machine (SVM), K-Neighbors, Gaussian Naive Bayes (GNB), Extra Trees, and Gaussian Naive Bayes (GNB). Binary classification issues are often expressed using the logistic regression approach. It makes an evenulity estimate for a binary result by using one or more predictor factors. The logistic function, which converts a linear feature combination into a probability value between 0 and 1, is the fundamental component of logistic regression. Because of this, it works especially well in situations where the result is a probability or a binary categorization. This makes the algorithm easy to comprehend and analyze, which helps with the data analysis.[22]

Table 2: Logistic Regression Parameters

Parameter	**Description**
C	Regularization strength; a smaller value indicates stronger regularization to prevent overfitting. This helps in managing model complexity and avoids overfitting by penalizing large coefficients.
solver	Algorithm used for optimization, such as 'liblinear' for smaller datasets or 'lbfgs' for larger ones. The choice of solver affects the speed and stability of convergence.
max_iter	Maximum number of iterations for the solver to converge, ensuring that the algorithm has sufficient iterations to reach the optimal solution.

Logistic Regression offers simplicity and ease of interpretation, making it suitable for initial modeling. Its computational efficiency and effectiveness in binary classification problems are additional benefits.[23]

3.6.1. Extra Trees

Extra Trees, or Extremely Randomized Trees, is an ensemble learning technique that builds a large number of decision trees using random subsets of features and data points. Unlike traditional decision trees, Extra Trees introduces greater randomness

in the splitting of nodes, which helps in reducing variance and improving generalization. Each tree is built with a random subset of features and data points, and the predictions are aggregated by averaging (for regression) or majority voting (for classification). This method is known for its robustness and efficiency in handling high-dimensional data[24].

Table 3: Extra Trees Parameters

Parameter	**Description**
n_estimators	Number of trees in the forest; more trees generally improve performance but increase computation time. A larger number of trees enhances model robustness and accuracy.
max_depth	Maximum depth of each tree; controls the complexity and size of the trees. Limiting the depth helps to prevent overfitting and improves model generalization.
min_samples_split	Minimum number of samples needed to divide an internal node; helps in controlling the growth of the trees and reducing overfitting.
min_samples_leaf	Minimum number of samples required to be at a leaf node; prevents creating leaves with very few samples, which enhances the generalization ability of the model.

Extra Trees is known for its high predictive accuracy and fast training times. Its robustness to overfitting and ability to handle complex datasets effectively are key benefits[25].

3.6.2. Gaussian Naive Bayes (GNB)

Gaussian Naive Bayes is a probabilistic classifier built on Bayes' theorem, which presumes that the features follow a Gaussian (normal) allocation. This model is particularly effective when the features are constant and normally distributed. The algorithm studies the probability of each dignity given the features, using the Gaussian distribution to estimate these probabilities. It is known for its simplicity and efficiency, especially in scenarios with high-dimensional data where features are assumed to be independent[26].

Table 4: Gaussian Naive Bayes Parameters

Parameter	**Description**
var_smoothing	A small value added to the variances to prevent numerical instability during computation. This parameter ensures stability and robustness in the model's predictions.

Gaussian Naive Bayes is simple and fast to train, making it ideal for high-dimensional datasets. Its efficiency and good performance with normally distributed data are significant advantages.

3.6.3. Bernoulli Naive Bayes (BNB)

Bernoulli Naive Bayes is a variant of the Naive Bayes algorithm tailored for binary or boolean features. It is built on the presumption that the features are binary and uses a Bernoulli distribution to show the existance or absence of features. This approach is often used in text classification tasks where features represent the presence or absence of words. The model calculates probabilities based on feature occurrence and class labels[27].

Table 5: Bernoulli Naive Bayes Parameters

Parameter	Description
alpha	Additive smoothing parameter; a small constant added to feature counts to handle zero probabilities. This parameter helps avoid zero probabilities and improves model robustness.
binarize	Threshold for binarizing the input features; values above this threshold are considered as 1, and others as 0. This parameter is useful for transforming continuous features into binary format.

Bernoulli Naive Bayes is well-suited for binary feature data and text classification. Its simplicity and efficiency make it effective for high-dimensional and sparse datasets.

3.6.4. Support Vector Classifier (SVC)

Support Vector Classifier (SVC) is a strong classification technique that finds the best hyperplane which maximizes the margin between various classes. It can handle both linear and non-linear classification issue through the use of kernel functions, which map input features into higher-dimensional spaces. SVC is known for its effectiveness in high-dimensional spaces and its ability to work well with a wide range of data distributions[28].

Table 6: Support Vector Classifier Parameters

Parameter	Description
C	Organization parameter; dominates the trade-off between fulfilling a low error on the training data and minimizing model intricacy. This balance affects bias and variance in the model.
kernel	Designates the kernel type to be used (e.g., 'linear', 'rbf' for radial basis function, 'poly' for polynomial). The kernel choice influences the model's ability to handle non-linear data.
gamma	Kernel degree for 'rbf', 'poly', and 'sigmoid' kernels; controls the influence of a single training example. This parameter helps define the shape of the decision boundary.

SVC handles complex decision boundaries well and performs effectively in high-dimensional spaces. The various kernel options and its ability to provide high classification accuracy are key benefits[29].

3.6.5. K-Neighbors (KNN)

K-Neighbors is a straightforward classification technique that uses the majority class among a data point's k-nearest neighbors to give a class to it. The distance between data points is typically computed using metrics such as Euclidean distance. KNN does not require a training phase, making it fast to implement. The choice of k and distance metrics are crucial as they significantly impact the model's performance and accuracy[30].

Table 7: K-Neighbors Parameters

Parameter	Description
n_neighbors	Number of neighbors to consider for classification; a smaller value can lead to overfitting, while a larger value can smooth out predictions. This affects the model's sensitivity to local data.
weights	Weight function used in prediction; 'uniform' applies equal weight to all neighbors, while 'distance' weights neighbors by their distance. This adjusts the influence of neighbors based on proximity.
metric	Distance metric used to compute the distance between points; common metrics include 'minkowski' and 'euclidean'. The choice of metric affects the accuracy of predictions.

K-Neighbors is simple and does not require training, which makes it effective for straightforward classification tasks.

3.7. Models Evaluation

In this Paper, we assessed six machine learning algorithms' performance on a dataset that was divided into training and testing sets. The dataset contained Logistic Regression (LR), Extra Trees, Gaussian Naive Bayes (GNB), Bernoulli Naive Bayes (BNB), Support Vector Classifier (SVC), and K-Neighbors. It was resampled to solve class imbalance. The cross-validated accuracy, test accuracy, precision, recall, and F1-score were used to evaluate each method. The goal of this Paper is to identify, given the dataset, the best model for diabetes prediction.

3.7.1. Cross-Validated Accuracy

Cross-validated accuracy evaluates the generalizability of a model by dividing the dataset into k subsets (folds). The model is trained on $k-1$ folds and tested on the remaining fold, with this process repeated k times. The average accuracy across all folds is the cross-validated accuracy.

$$\text{CV-Acc} = \frac{1}{k}\sum_{i=1}^{k} A_i \tag{1}$$

where:

- k = Number of folds
- A_i = Accuracy on the i-th fold

3.7.2. Test Accuracy

Test accuracy measures the proportion of correctly classified instances in the test set.

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \tag{2}$$

Or, in terms of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN):

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{3}$$

3.7.3. Precision

Precision, also known as Positive Predictive Value, measures the proportion of positive predictions that are correct.

$$\text{Precision} = \frac{TP}{TP+FP} \tag{4}$$

where:

- TP = True Positives (correctly predicted positive instances)
- FP = False Positives (incorrectly predicted positive instances)

3.7.4. Recall

Recall, also known as Sensitivity or True Positive Rate, measures the proportion of actual positive instances that are correctly identified.

$$\text{Recall} = \frac{TP}{TP+FN} \tag{5}$$

where:

- TP = True Positives
- FN = False Negatives (actual positive instances that were incorrectly predicted as negative)

3.7.5. F1-Score

The F1-Score is the harmonic mean of Precision and Recall.

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6}$$

In terms of TP, FP, and FN:

$$F1 = 2 \times \frac{TP}{2TP + FP + FN} \tag{7}$$

These assessment metrics give a thorough grasp of the classifier's capabilities and reveal how effectively the model balances the trade-offs between false positives and false negatives while making class distinctions. Through the assessment of the model's performance on many dimensions, including recall, precision, and overall accuracy, the metrics facilitate an understanding of the classifier's capabilities and shortcomings in terms of accurate result prediction. The performance and confusion measures for each of the six machine learning models evaluated in this article are summarized in the table below.

Table 8: Performance Evaluation

Model	**Cross-Val Accuracy**	**Test Accuracy**	**Precision**	**Recall**	**F1-Score**
Logistic Regression	0.9652	0.9568	0.94	0.90	0.92
Extra Trees	0.9701	0.9560	0.94	0.89	0.92
Gaussian NB	0.8989	0.8774	0.95	0.87	0.91
Bernoulli NB	0.9123	0.8912	0.95	0.88	0.92
Support Vector Classifier (SVC)	0.9664	0.9551	0.93	0.90	0.92
K-Neighbors	0.9449	0.9266	0.89	0.83	0.86

3.8. Compare Models and Select the Best Technique

- **Logistic Regression** with an F1-score of 0.92, recall of 0.90, precision of 0.94, and test accuracy of 0.9568, showed excellent performance. As one of the best-performing models in this research, these results highlight its accuracy in predicting diabetes.

- **Extra Trees** exhibited the highest cross-validated accuracy (0.9701) among the models, with a similar test accuracy (0.9560) to Logistic Regression. How-

Table 9: Confusion Matrix

Model	**True Negative**	**False Positive**	**False Negative**	**True Positive**
Logistic Regression	818	19	31	290
Extra Trees	820	17	34	287
Gaussian NB	731	106	36	285
Bernoulli NB	748	89	37	284
Support Vector Classifier (SVC)	816	21	31	290
K-Neighbors	805	32	53	268

ever, its precision (0.94) and recall (0.89) were slightly lower, suggesting a marginally reduced reliability in identifying true positive cases on new data.

- **Support Vector Classifier (SVC)** attained a 0.9551 test accuracy, along with 0.93 and 0.90 precision and recall ratings. SVC is less appealing as the best model option even if these measures are similar to those of logistic regression and do not significantly differ from it.

- **Gaussian Naive Bayes** showed good precision (0.95), but had a lower test accuracy of 0.8774. Its recall (0.87) was lower, though, suggesting inconsistent results when it came to detecting real positive instances. Similar limits were displayed by **Bernoulli Naive Bayes**, which had a test accuracy of 0.8912 and balanced but lower recall (0.88) and precision (0.95).

- **K-Neighbors** revealed a test accuracy of 0.9266 along with 0.89 and 0.83 for precision and recall, respectively. Although it performs rather well, it is not as well as SVC, Extra Trees, and Logistic Regression, which makes it less ideal for this dataset.

Overall,**Logistic Regression** Since it balances precision, recall, and F1-score performance, is chosen as the most dependable model overall because of its high test accuracy. Logistic regression is the favored model for this article despite the good results also shown by **Extra Trees** and **SVC**. This is because of its consistent predictive power and somewhat superior balance in performance metrics.

Conclusions

The goal of our work is to show how machine learning (ML) algorithms may dramatically improve diabetes prediction and early detection based on input data. Through the use of an extensive dataset obtained from the Egyptian clinic of Dr. Hossam Arafa, the research demonstrates the impressive potential of machine learning models to enhance predictive healthcare, especially in areas such as Egypt where diabetes is a major public health concern.

We compared various machine learning models—Extra Trees, Gaussian Naive Bayes, Bernoulli Naive Bayes, Support Vector Classifier (SVC), and K-Neighbors. Logistic Regression emerged as the most effective for predicting diabetes, with a test

accuracy of 0.9568 and balanced metrics: precision of 0.94, recall of 0.90, and F1-score of 0.92. Extra Trees, while slightly lower in test accuracy (0.9560), excelled in cross-validated accuracy (0.9701). SVC also performed well (0.9551 accuracy) but offered no major advantage. Gaussian and Bernoulli Naive Bayes were less consistent, and K-Neighbors lagged behind the top models.

The importance of cutting-edge machine learning approaches in improving healthcare outcomes is highlighted in this research. In order to increase predicted accuracy, it promotes ongoing improvements in the clinical use of these models, particularly in tackling important issues like data quality and model interpretability. Machine learning models can greatly improve early intervention tactics in healthcare systems, which might result in more individualized and efficient diabetic care. It is recommended that future research concentrate on improving the forecasting accuracy of current models and investigating novel data sources in order to bolster public health activities.

References

[1] W. F. Abd El Aal, T. M. Farahat, S. M. Morsy, N. M. Ebrahim, Prevalence of prediabetes among adolescents and youth in sohag governorate, egypt, BMC Public Health 20 (1) (2020) 1372. doi:10.1186/s12889-020-09502-x.
URL https://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-020-09502-x

[2] IBM, Artificial intelligence, accessed: 2024-07-30 (2024).
URL https://www.ibm.com/topics/artificial-intelligence

[3] B. In, Artificial intelligence, accessed: 2024-07-30 (2024).
URL https://builtin.com/artificial-intelligence

[4] GeeksforGeeks, Machine learning models, https://shorturl.at/FBy0s/, accessed: 2024-07-30 (2024).

[5] N. S. Rani, S. K. M. Yadav, A. S. Nair, Diabetes mellitus: A public health challenge, Journal of Diabetes Research 2022 (2022) 1–10, accessed: 2024-07-30.
URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9390800/

[6] D. A. Shih, E. G. Riley, J. M. Brown, Prevalence and impact of diabetes mellitus in the middle east and north africa, Journal of Diabetes Research 2021 (2021) 1–12, accessed: 2024-07-30.
URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8472500/

[7] M. Maniruzzaman, M. J. Rahman, B. Ahammed, M. M. Abedin, Classification and prediction of diabetes disease using machine learning paradigm, Health information science and systems 8 (2020) 1–14.

[8] M. K. Hasan, M. A. Alam, D. Das, E. Hossain, M. Hasan, Diabetes prediction using ensembling of different machine learning classifiers, IEEE Access 8 (2020) 76516–76531.

[9] S. Kumari, D. Kumar, M. Mittal, An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier, International Journal of Cognitive Computing in Engineering 2 (2021) 40–46.

[10] P. Rajendra, S. Latifi, Prediction of diabetes using logistic regression and ensemble techniques, Computer Methods and Programs in Biomedicine Update 1 (2021) 100032.

[11] J. o. Healthcare Engineering, Retracted: A novel diabetes healthcare disease prediction framework using machine learning techniques (2023).

[12] C. V. Raghavendran, G. Naga Satish, N. Kumar Kurumeti, S. M. Basha, An analysis on classification models to predict possibility for type 2 diabetes of a patient, in: Innovative Data Communication Technologies and Application: Proceedings of ICIDCA 2021, Springer, 2022, pp. 181–196.

[13] S. S. Bhat, V. Selvam, G. A. Ansari, M. D. Ansari, M. H. Rahman, Prevalence and early prediction of diabetes using machine learning in north kashmir: a case study of district bandipora, Computational Intelligence and Neuroscience 2022 (1) (2022) 2789760.

[14] A. Dutta, M. K. Hasan, M. Ahmad, M. A. Awal, M. A. Islam, M. Masud, H. Meshref, Early prediction of diabetes using an ensemble of machine learning models, International Journal of Environmental Research and Public Health 19 (19) (2022) 12378.

[15] D. J. Reddy, B. Mounika, S. Sindhu, T. P. Reddy, N. S. Reddy, G. J. Sri, K. Swaraja, K. Meenakshi, P. Kora, Withdrawn: Predictive machine learning model for early detection and analysis of diabetes (2020).

[16] S. P. Chatrati, G. Hossain, A. Goyal, A. Bhan, S. Bhattacharya, D. Gaurav, S. M. Tiwari, Smart home health monitoring system for predicting type 2 diabetes and hypertension, Journal of King Saud University-Computer and Information Sciences 34 (3) (2022) 862–870.

[17] M. E. Febrian, F. X. Ferdinan, G. P. Sendani, K. M. Suryanigrum, R. Yunanda, Diabetes prediction using supervised machine learning, Procedia Computer Science 216 (2023) 21–30.

[18] C.-Y. Chou, D.-Y. Hsu, C.-H. Chou, Predicting the onset of diabetes with machine learning methods, Journal of Personalized Medicine 13 (3) (2023) 406.

[19] M. M. Hassan, M. A. M. Billah, M. M. Rahman, S. Zaman, M. M. H. Shakil, J. H. Angon, Early predictive analytics in healthcare for diabetes prediction using machine learning approach, in: 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), IEEE, 2021, pp. 01–05.

[20] D. C. Yadav, S. Pal, An experimental study of diversity of diabetes disease features by bagging and boosting ensemble method with rule based machine learning classifier algorithms, SN Computer Science 2 (1) (2021) 50.

[21] A. Vidhya, Overcoming class imbalance using smote techniques, accessed: 2024-08-01 (2020).
URL https://shorturl.at/vZfGW/

[22] J. Brownlee, Logistic regression for machine learning, Machine Learning MasteryAccessed: 2024-07-30 (2024).
URL https://shorturl.at/Svssr/

[23] A. Kumar, Logistic regression explained: From scratch, visually, mathematically, and programmatically, Towards Data ScienceAccessed: 2024-07-30 (2024).
URL https://https://shorturl.at/MQVAU

[24] P. Geurts, D. Ernst, Extremely randomized trees, in: Proceedings of the 11th International Conference on Machine Learning (ICML), 2006, pp. 489–496, accessed: 2024-08-11.
URL https://www.semanticscholar.org/paper/Extremely-randomized-trees-Geurts-Ernst/336a165c17c9c56160d332b9f4a2b403fccbdbfb

[25] H. K, What, when, how: Extratrees classifier, Towards Data ScienceAccessed: 2024-07-30 (2024).
URL https://shorturl.at/QvNuO

[26] M. L. Plus, How naive bayes algorithm works with example and full code, Machine Learning PlusAccessed: 2024-07-30 (2024).
URL https://www.machinelearningplus.com/predictive-modeling/how naive-bayes-algorithm-works-with-example-and-full-code/

[27] G. AI, Part 2: Dive into bernoulli naive bayes, accessed: 2024-08-11 (2023).
URL https://medium.com/@gridflowai/part-2-dive-into-bernoulli-naive-bayes-d0cbcbabb775

[28] GeeksforGeeks, Support vector machine algorithm, accessed: 2024-07-29 (2023).
URL https://www.geeksforgeeks.org/support-vector-machine-algorithm/

[29] Scikit-learn, sklearn.svm.SVC — Scikit-learn 1.3.0 documentation, accessed: 2024-07-30 (2024).
URL https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

[30] A. M. Simple, K-nearest neighbors (knn): A comprehensive guide, MediumAccessed: 2024-07-30 (2024).
URL https://medium.com/ai-made-simple/k-nearest-neighbors-knn-a-comprehensive-guide-7add717806ad