International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

A Study on Machine Learning Based Predictive Analytics for Early Detection of Diabetes in Senior Citizens

Mr. Kamalesh A[1], Mrs. B. Lakshmi[2]

[1]Student, [2]Assistant Professor

Department of Management Studies, Panimalar Engineering College, Chennai.

ABSTRACT

This study investigates the use of machine learning (ML) models for the early detection of diabetes among senior citizens, a group particularly vulnerable to this chronic condition. Utilizing K-Nearest Neighbors, Decision Trees, and Gradient Boosting, we developed predictive analytics models that analyze health data to identify early signs of diabetes. The models were trained using a dataset encompassing a range of health indicators pertinent to the elderly population. Our evaluation focused on the accuracy, sensitivity, and specificity of each model, revealing their potential to enhance diagnostic processes and preventive healthcare. By enabling earlier intervention strategies, this approach aims to improve the management and outcomes of diabetes care in the aging population.

Keywords: Diabetes, Machine learning, predictive analytics.

INTRODUCTION

The rising global prevalence of diabetes, particularly among senior citizens, presents a critical public health challenge. As age advances, the risk and complications associated with diabetes increase, underscoring the need for early detection and intervention. This study employs advanced machine learning techniques—K-Nearest Neighbors, Decision Trees, and Gradient Boosting—to address this need by developing predictive models that can identify early signs of diabetes in the elderly population. These models analyze comprehensive health data, including clinical, lifestyle, and biometric information, to discern patterns that precede the onset of diabetes. By integrating these predictive models into routine healthcare screenings, we aim not only to enhance early diagnosis but also to facilitate personalized healthcare strategies. This proactive approach could significantly improve the management of diabetes, reducing the incidence of severe complications and improving the overall quality of life for senior patients. The success of this initiative has the potential to set a new standard in preventive healthcare, demonstrating the pivotal role of machine learning in transforming patient care.

OBJECTIVES

Primary Objective:

To develop and validate machine learning models that can predict early onset of diabetes in senior citizens using health data.

Secondary Objectives:

1. To compare the effectiveness and accuracy of K-Nearest Neighbors, Decision Trees, and Gradient Boosting in predicting diabetes.
2. To identify key health indicators and features that are most predictive of diabetes risk among the elderly.
3. To enhance the predictive power of the models through optimization techniques and hyperparameter tuning.

REVIEW OF LITERATURE

Aneesha Chacko, March 2021, Diabetes Prediction Using Machine Learning classification Techniques

Our project aims to predict diabetes using machine learning algorithms on the Pima Indians Diabetes Database. We utilize various classifiers and ensemble methods, finding Random Forest to be the most effective. Glucose levels show a positive correlation with diabetes. Our approach can revolutionize diabetes prediction, aiding early detection. Future enhancements include incorporating more parameters and improving data quality for increased accuracy.

KM Jyoti Rani, August 2020, Diabetes Prediction Using Machine Learning

Diabetes is a global health crisis, affecting millions worldwide. Early prediction is crucial to mitigate its impact on individuals and healthcare systems. Our project employs machine learning algorithms like KNN, Logistic Regression, Random Forest, SVM, and Decision Tree to predict diabetes with high accuracy. We achieve 99% accuracy using the Decision Tree algorithm on the john Diabetes Database. This system holds promise for early disease detection and could be expanded to predict other medical conditions, enhancing healthcare automation and efficiency.

Mallula Venkatesh, July 2023, Multiple Disease Prediction using Machine Learning, Deep Learning and Stream-LIT

The "Multiple Disease Prediction using Machine Learning, Deep Learning and Streamlit" project aims to predict diseases like diabetes, heart disease, kidney disease, Parkinson's disease, and breast cancer using machine learning algorithms such as TensorFlow with Keras, Support Vector Machine (SVM), and Logistic Regression. Deployed via Streamlit Cloud, the user-friendly interface prompts users to input relevant parameters for disease prediction, with high accuracies achieved across various diseases. Future enhancements include expanding disease coverage, integrating more ML algorithms, advanced feature engineering, real-time monitoring, deployment in healthcare settings, and continuous model improvement through collaboration with healthcare professionals.

Arnab Das, Allamsetty Udit Venkata Nagopa Sai, March 2022, Disease Prediction Application Using Machine Learning

The healthcare system relies on machine learning and data processing to predict diseases like breast cancer, heart disease, and diabetes, simplifying patient care decisions. By inputting medical data, the system accurately predicts disease occurrences and recommends suitable hospitals and doctors for treatment. This research aims to predict common diseases efficiently, reducing delays and inaccuracies in medical reporting. By focusing on heart disease, breast cancer, and diabetes, the system improves accuracy and provides recommendations for nearby hospitals with quality care. Future implementations aim to recommend hospitals based on user reviews using the Collaborative Filtering algorithm. This algorithm considers user preferences to provide personalized recommendations, enhancing patient satisfaction and healthcare services.

Dr C K Gomathy, Mr A Rohith Naidu, Dec 2021, The Prediction of Disease using Machine Learning

The Disease Prediction using Machine Learning system utilizes symptoms provided by users to predict diseases, employing the Naïve Bayes classifier for disease prediction. By leveraging machine learning algorithms like linear regression and decision tree, the system predicts diseases such as Diabetes, Malaria, Jaundice, Dengue, and Tuberculosis with an average prediction accuracy probability of 100%. Implemented using the Grails framework, this system offers a user-friendly interface accessible via web application from any location. Additionally, the accuracy of disease risk prediction depends on the diversity of features in hospital data. This systematic review aims to assess the performance, limitations, and future prospects of Disease Predictability Software, potentially informing developers and promoting personalized patient care. The program predicts diseases based on user input symbols, utilizing decision tree, Random Forest, and Naïve Bayes algorithms for data processing, achieving a system accuracy of 98.3%. Overall, these machine learning techniques are designed to effectively predict disease outbreaks.

RESEARCH METHODOLOGY

This study presents a holistic approach to predicting various health conditions prevalent among senior citizens, aiming to enhance healthcare management and early intervention strategies. The methodology is structured into three distinct phases: pre-processing, training, and classification. Each phase is meticulously designed to ensure the robustness and effectiveness of the predictive models.

1. PRE-PROCESSING PHASE

The pre-processing phase plays a crucial role in preparing the raw data for model training and evaluation. It involves several key steps:

Data Collection: The data is sourced from the Open-source Senior care dataset, comprising diverse health-related attributes, medical history, lifestyle factors, and demographic information of senior citizens aged above 60 years. The data collection process ensures compliance with ethical standards and privacy regulations, with informed consent obtained from all participants.

Data Cleaning: Raw data often contains inconsistencies, missing values, and outliers that can adversely affect model performance. In this phase, rigorous data cleaning techniques are applied to address these issues. Missing values are imputed using appropriate strategies such as mean, median, or mode imputation. Outliers are identified and either removed or treated based on domain knowledge.

Feature Engineering: Feature engineering involves transforming raw data into informative features that capture relevant patterns and relationships. This may include creating new features, encoding categorical variables, and scaling numerical features to a common range. Techniques such as one-hot encoding, label encoding, and standardization are applied to ensure compatibility with different modelling algorithms.

Data Partitioning: The pre-processed dataset is divided into training and test sets using stratified sampling to preserve the class distribution. Approximately 80% of the data is allocated for training the models, while the remaining 20% is reserved for independent evaluation. This partitioning strategy helps assess the generalization performance of the models on unseen data.

2. TRAINING PHASE:

In the training phase, a diverse ensemble of machine learning and deep learning models is employed to build predictive algorithms for the targeted health conditions. The models selected for training include:

Logistic Regression: A classical linear model used for binary classification tasks. It models the probability of a binary outcome using a logistic function, making it suitable for predicting disease onset based on input features.

$$\mathrm{P(Y=1|X)} = \frac{1}{1+e-(\beta 0+\beta 1X1+\beta 2X2+\ldots+\beta nXn)}$$

Where:

P(Y=1|X) is the probability of the positive class given the input features X.

β0,β1,...,βn are the coefficients of the logistic regression model.

X1,X2,...,Xn are the input features.

Gradient Boosting: Ensemble learning techniques that combine multiple weak learners (decision trees) to create a strong predictive model. Gradient Boosting algorithm iteratively improve the performance of the model by minimizing a predefined loss function, resulting in highly accurate predictions.

K-Nearest Neighbors Classifier (KNN): A non-parametric method used for classification tasks based on the similarity of data points in the feature space. KNN assigns a class label to a new data point based on the majority class of its nearest neighbors. It is simple to implement and does not require training, making it suitable for lazy learning.

Decision Tree: A tree-like model consisting of nodes representing features, branches representing decisions, and leaf nodes representing class labels. Decision trees are intuitive to interpret and can handle both numerical and categorical data. However, they are prone to overfitting, especially on noisy datasets.

Each model is trained on the pre-processed training data using specific algorithms tailored to its characteristics. Hyperparameters such as learning rate, regularization strength, tree depth, and number of neurons in hidden layers are fine-tuned through grid search or random search to optimize performance and prevent overfitting.

3. CLASSIFICATION PHASE:

The classification phase involves evaluating the trained models on the independent test dataset to assess their predictive performance. The following performance metrics are calculated for each model:

Accuracy: The proportion of correctly classified instances out of the total instances.

Precision: The proportion of true positive predictions out of all positive predictions made by the model.

Recall: The proportion of true positive predictions out of all actual positive instances in the dataset.

F1-score: The harmonic mean of precision and recall, providing a balanced measure of the model's accuracy.

The outputs of each model are analyzed comprehensively to identify strengths, weaknesses, and areas for improvement. Model interpretability and feature importance are also assessed to gain insights into the underlying factors contributing to disease prediction.

4. PERFORMANCE EVALUATION:

When evaluating the performance of classification models, particularly in scenarios with imbalanced datasets, relying solely on metrics like accuracy can be insufficient. This is because accuracy does not consider the distribution of classes and may not adequately capture the model's true effectiveness. One widely used tool for assessing classification results is the confusion matrix, which provides a detailed breakdown of the model's predictions. The matrix contains four quadrants representing True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) predictions. From this matrix, various performance metrics can be derived to assess different aspects of the model's performance.

Some of the key performance metrics include precision, recall (sensitivity), F1-score, and the Receiver Operating Characteristic (ROC) curve. Precision measures the proportion of correctly predicted positive cases among all predicted positive cases, while recall calculates the proportion of correctly predicted positive cases among all actual positive cases. The F1-score is the harmonic mean of precision and recall, providing a balanced measure of a model's performance. The ROC curve plots the true positive rate against the false positive rate at various threshold settings, offering insights into the model's discrimination ability across different thresholds.

These metrics collectively offer a more nuanced understanding of a classifier's performance, allowing researchers and practitioners to make informed decisions about model selection, parameter tuning, and deployment strategies.

$$\mathrm{Accuracy} = \frac{TP+TN}{TP+FP+FN+TN}$$

$$\mathrm{F1-score} = 2\,\frac{precision * recall}{precision + recall}$$

Recall $= \frac{TP}{TP + FP}$

Precision $= \frac{TP}{TP + FN}$

EXPERIMENTAL EVALUATION:

The experimental evaluation of the predictive models is conducted in a Jupyter notebook environment, leveraging cloud-based computing resources for scalability and efficiency. Data science libraries such as TensorFlow, Scikit-learn, Pandas, and NumPy are utilized for data manipulation, model development, evaluation, and visualization. The experiments are designed to validate the efficacy of the proposed methodology in accurately predicting health conditions among senior citizens. Various performance metrics, visualizations, and comparative analyses are employed to assess the robustness and generalization capabilities of the predictive models.

DATA COLLECTION:

The data utilized in this project is sourced from Open-Source Senior care dataset, encompassing a diverse range of health-related attributes and demographic information. The dataset is collected with the consent of the participants and adheres to strict privacy and ethical guidelines. It includes information on medical history, lifestyle factors, physiological parameters, diagnostic tests, and medication usage, providing a comprehensive overview of the health status of senior citizens.

TABLE 1

A SAMPLE OF DIABETES DISEASE PREDICTION DATASET

Note: Here the Target variables are Label Encoded

PAT_ID	PREGNANCIES	GLUCOSE	BLOOD PRESSURE	SKIN THICKNESS	INSULIN	BMI	DIABETES PEDIGREE FUNCTION	AGE	TARGET
1	6	148	72	35	0	33.6	0.627	81	1
2	1	85	66	29	0	26.6	0.351	67	0
3	8	183	64	0	0	23.3	0.672	74	1
4	1	89	66	23	94	28.1	0.167	70	0
5	0	137	40	35	168	43.1	2.288	64	1

The Diabetes disease dataset comprises a total of 2760 observations and 10 Columns, making it a substantial resource for diabetes disease prediction.

TABLE 2

Decoded Target Variable

VARIABLE	INTERPRETATION
OUTCOME	Indicating 0: Absence of diabetes 1: Presence of diabetes

Machine learning algorithms typically work better with numerical data. By converting categories (e.g., "Positive" and "Negative") into numerical values (0 and 1), the model can more easily understand the relationship between the outcome and other features in the data. So, in this case, the Table 2 shows that 0 and 1 have clear meanings (0 signifying absence and 1 signifying presence of diabetes) which makes the data easier for humans to interpret as well. There are other encoding schemes for categorical variables, but binary encoding (0, 1) is a simple and effective approach for many machine learning tasks.

TABLE 3. DATASET SHAPES

Dataset	Shape
Training	(1545, 9)

Validation	(663, 9)
Test	(552, 9)

The dataset used to predict Diabetes disease consists of 2760 entries, with each entry containing information on 9 attributes related to diabetes prediction. Upon splitting the dataset for training, validation, and testing, it was found that the training set comprises 1545 entries, while the validation set contains 663 entries, and the test set contains 552 entries. This division ensures that a significant portion of the data is allocated for training the machine learning models (approximately 56% of the total dataset), allowing them to learn patterns and relationships within the data. The validation set, comprising approximately 24% of the data, serves to tune model hyperparameters and prevent overfitting. Finally, the test set, comprising the remaining 20% of the data, is used to evaluate the performance of the trained models on unseen data, providing an unbiased assessment of their predictive capabilities. This balanced partitioning of the dataset into training, validation, and test sets facilitates robust model development and evaluation for effective diabetes prediction.

TABLE 4. PREVALENCE OF DIABETES

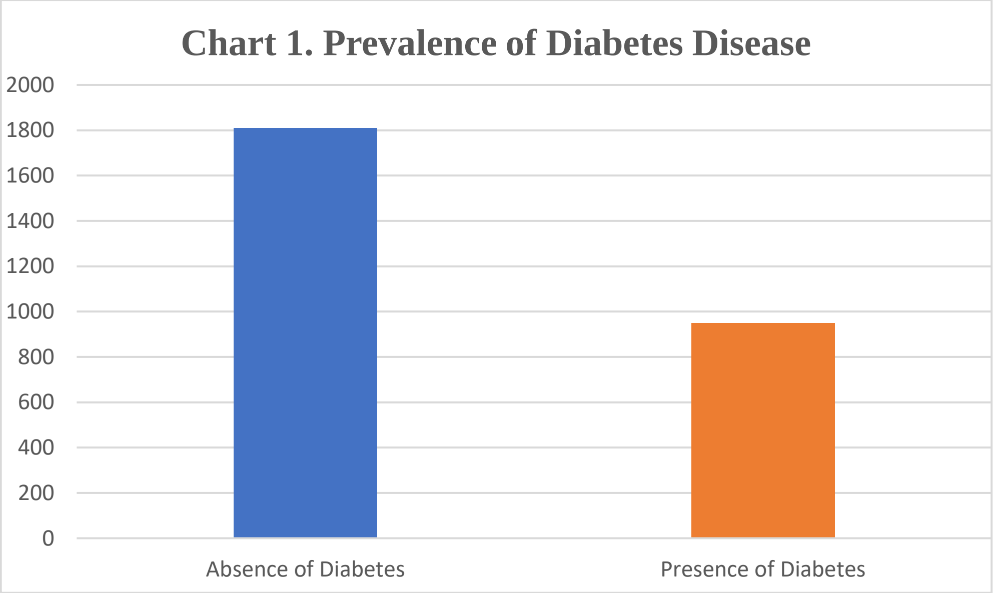
VARIABLE	NO OF PATIENT	PERCENTAGE (%)
Absence of Diabetes	1811	**66**
Presence of Diabetes	949	**34**
TOTAL	2760	100

FINDINGS:

The chart and accompanying data table display the prevalence of diabetes among a sample of 2,760 patients Absence of Diabetes 1,811 patients, constituting 66% of the sample and Presence of Diabetes 949 patients, making up 34% of the sample. A significant majority of the sampled population does not have diabetes. Approximately one-third of the sampled individuals are diagnosed with diabetes, indicating a substantial prevalence within the population studied.

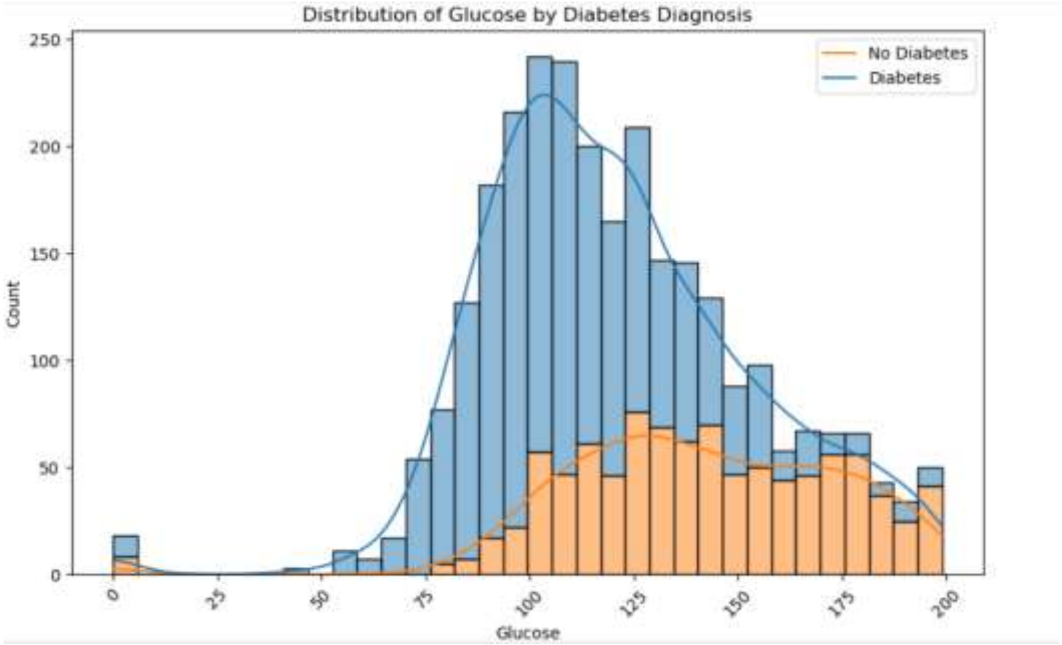
INFERENCE:

The data suggests a high awareness and diagnosis rate of diabetes, as a considerable portion of the sample has been diagnosed. The distribution also highlights the importance of diabetes as a public health concern given its prevalence in the sample. This could have implications for healthcare resource allocation and preventive health measures.



DIABETES DISEASE DISTRIBUTION

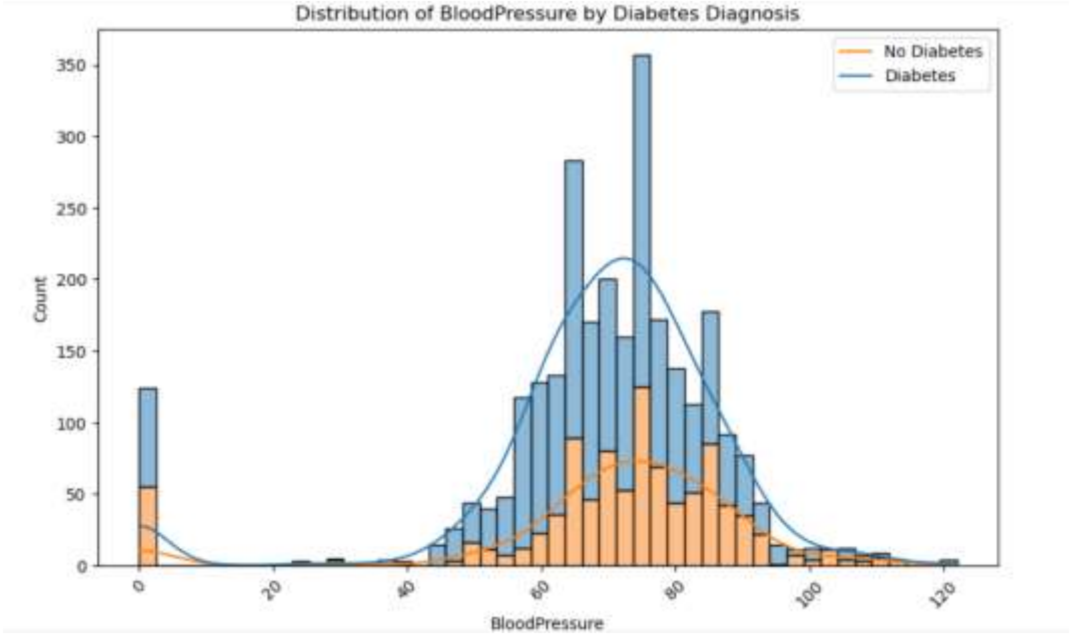
CHART 2. DISTRIBUTION OF GLUCOSE BY DIABETES DIAGNOSIS



INTERPRETATION:

The histogram shows glucose levels in individuals, differentiated by diabetes diagnosis where No Diabetes (Blue)Predominantly normal distribution, with most values around 100 mg/dL, primarily ranging from 70 to 130 mg/dL and Diabetes (Orange) Skewed distribution with a peak around 150 mg/dL, indicating higher glucose levels generally from 120 mg/dL upwards. Individuals with diabetes typically show higher glucose levels compared to those without. There is some overlap between 90 and 130 mg/dL, suggesting this range could be critical for identifying pre-diabetic conditions.

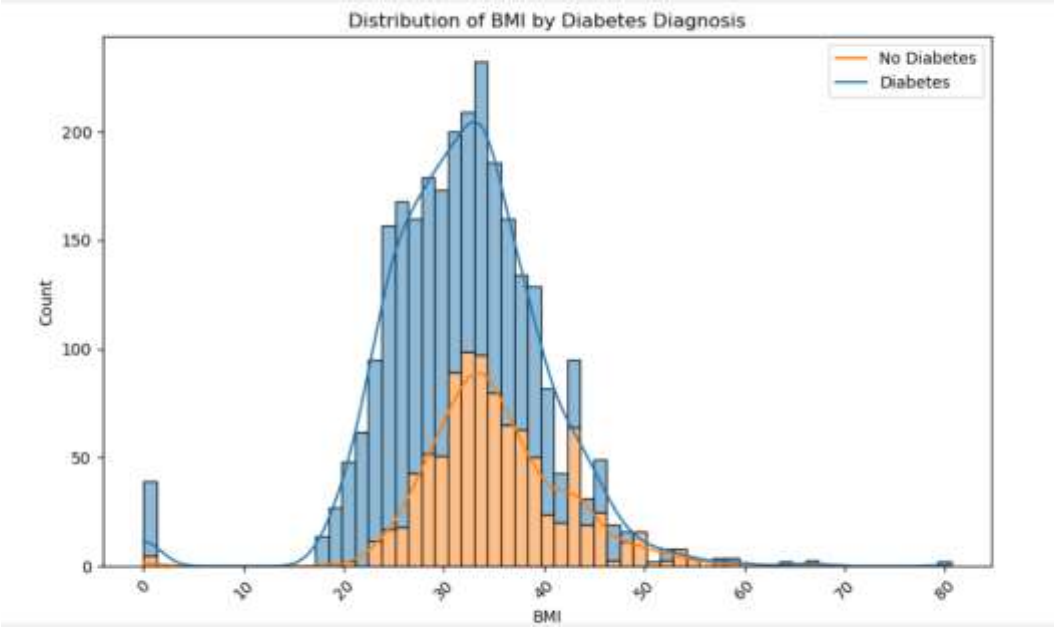
CHART 3. DISTRIBUTION OF BLOODPRESSURE BY DIABETES DIAGNOSIS



INTERPRETATION:

The histogram illustrates the distribution of blood pressure levels among individuals grouped by their diabetes diagnosis status which has No Diabetes Features a normal distribution centered around 80 mmHg, with most values between 60 and 100 mmHg. Diabetes Displays a wider distribution with a secondary peak around 90 mmHg, indicating a higher range of blood pressure levels, extending from 60 to 100 mmHg. Blood pressure distributions overlap significantly, although individuals with diabetes tend to show a broader spread and slightly higher blood pressure levels. This suggests that while blood pressure varies widely in both groups, those with diabetes might be more prone to higher ranges.

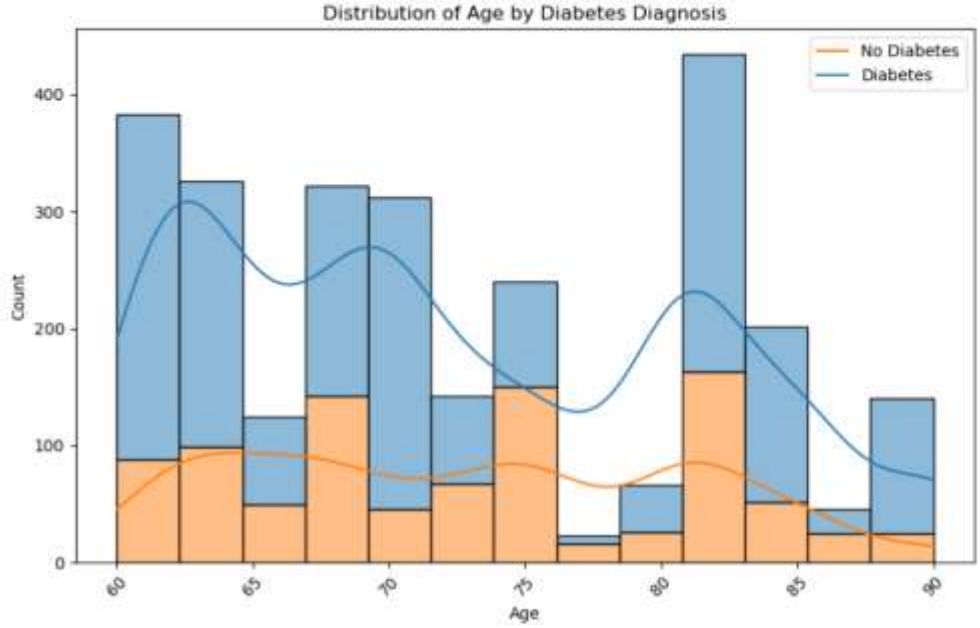
CHART 4. DISTRIBUTION OF BMI BY DIABETES DIAGNOSIS



INTERPRETATION:

The histogram displays the distribution of Body Mass Index (BMI) among individuals categorized by diabetes diagnosis. No Diabetes Shows a normal distribution, peaking around a BMI of 25, with most individuals having a BMI in the range considered normal to slightly overweight. Diabetes Demonstrates a right-skewed distribution with a peak around a BMI of 30, which falls into the overweight category, extending into the obese range. Individuals with diabetes generally have higher BMIs, with the distribution skewed towards overweight and obese categories. This indicates a correlation between higher BMI and the prevalence of diabetes.

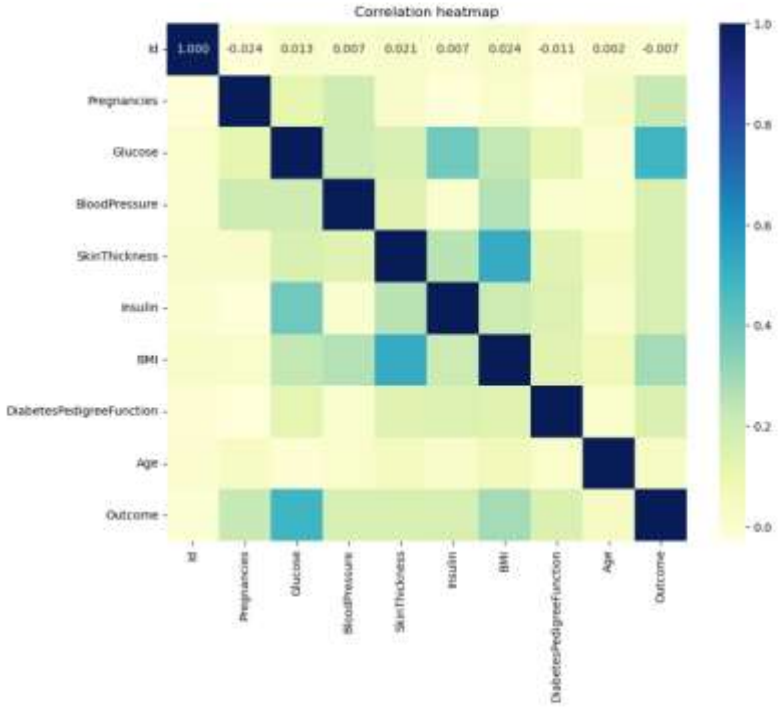
CHART 5. DISTRIBUTION OF AGE BY DIABETES DIAGNOSIS



INTERPRETATION:

The histogram shows the age distribution among individuals categorized by diabetes diagnosis No Diabetes Dominant in younger age groups, particularly noticeable in the 40s and early 50s, with a decrease as age increases and Diabetes has Higher counts in the mid to later age groups, peaking in the 60s, indicating a higher prevalence of diabetes in older adults. Diabetes is more prevalent in older age groups, reflecting an increased risk or diagnosis rate as age advances. Younger age groups show a lower prevalence of diabetes, consistent with general health trends where age is a significant risk factor for the condition.

CHART 6. CORRELATION HEATMAP OF DIABETES



INTERPRETATION:

This correlation heatmap visually represents the relationships between various health metrics related to diabetes:

Strong Positive Correlations:

- None of the variables show strong positive correlations (values close to 1.0) with each other, indicating no direct linear relationships where increases in one feature consistently correspond with increases in another.

Moderate to Weak Positive Correlations:

- **Glucose and Outcome:** A slight positive correlation (around 0.21), suggesting higher glucose levels may be associated with a higher likelihood of a diabetes diagnosis.
- **BMI and Skin Thickness:** A correlation (around 0.24), indicating that higher body mass index may be associated with greater skin thickness.

Negative Correlations:

- Blood Pressure and Pregnancies: A weak negative correlation (around -0.08), indicating that higher blood pressure might be slightly less common in those with more pregnancies, though the relationship is weak.
- Diabetes Pedigree Function and Insulin: A weak negative correlation (around -0.02 to -0.07 range), suggesting little to no straightforward inverse relationships.

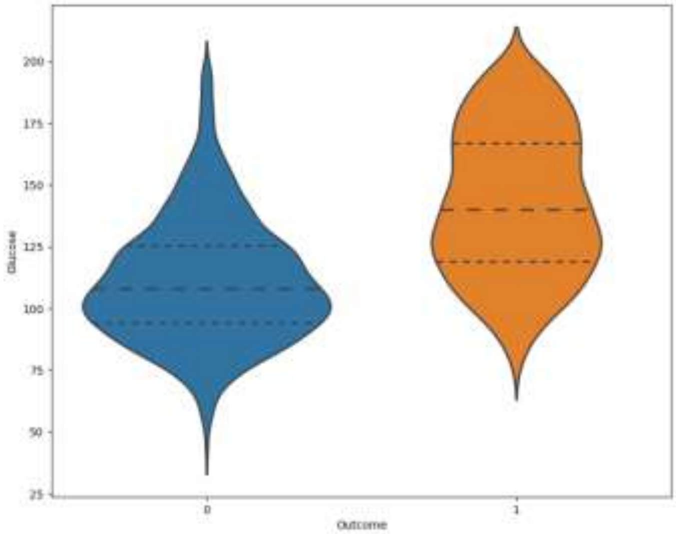
Diagonal (Self-Correlation):

- Each variable shows a perfect self-correlation of 1.0 (as expected, since any variable perfectly correlates with itself).
- Most variables show weak to negligible correlations with each other, suggesting that no single factor is strongly predictive of another within this dataset.
- The moderate correlations involving glucose and BMI highlight their relevance in diabetes studies, as these factors are often linked to the condition's development and management.

This heatmap is useful for identifying potential relationships to investigate further with more sophisticated statistical methods or predictive modeling, especially for factors that might influence diabetes outcomes.

CHART 7

ANALYZING GLUCOSE DISTRIBUTION ACROSS POSITIVE AND NEGATIVE OUTCOMES



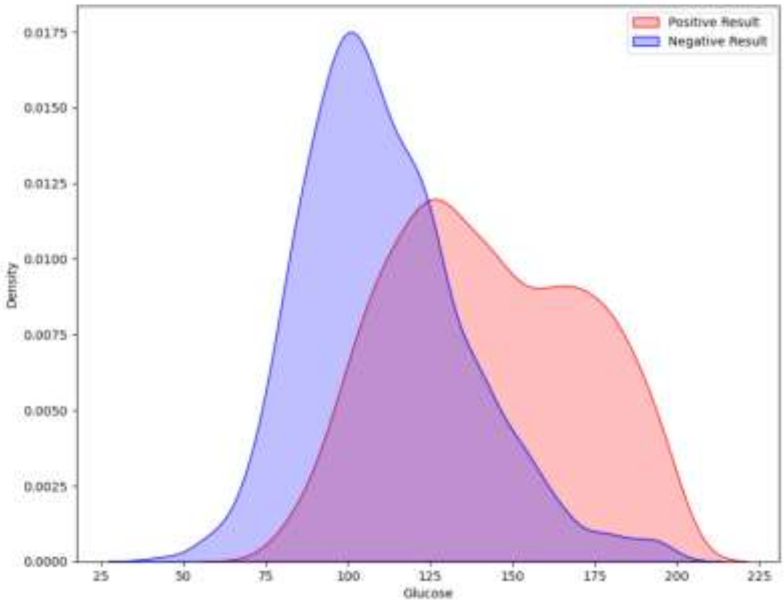
INTERPRETATION:

The violin plot illustrates the distribution of a continuous variable across two groups defined by a binary outcome (0 and 1). The group labeled as "0" (depicted in blue) shows a narrower and symmetric distribution around a lower median, suggesting that this group generally exhibits lower values of the measured variable. In contrast, the group labeled as "1" (depicted in orange) presents a broader distribution with a higher median, indicating that this variable tends to be higher in this group and displays greater variability. This plot is useful for visualizing the density of the data at different values, with the thickness of each "violin" representing the frequency of data points. The wider sections of the violins indicate a higher concentration of data points, emphasizing where most values fall within each group. Such a visualization is particularly helpful in medical or biological contexts, where the differences between two conditions or outcomes can provide insights into the characteristics of each group, potentially relating to disease markers or effects of a treatment.

CHART 8

ANALYZING THE DENSITY FUNCTION PLOT TO THE

GLUCOSE LEVELS



INTERPRETATION:

This density plot illustrates the distribution of glucose levels for two groups categorized by test results—Positive and Negative. The group with a Positive Result exhibits a sharp peak in glucose levels concentrated around 75 to 100 mg/dL, suggesting lower glucose levels overall. Conversely, the Negative Result group shows a broader distribution with a peak around 125 to 150 mg/dL, indicating higher and more variable glucose levels. This suggests that those with a Negative Result tend to have higher glucose readings, which could indicate individuals at higher risk or in a prediabetic range, depending on the diagnostic criteria used.

IMPLEMENTING MACHINE LEARNING MODELS

Studie carried out using dataset comprising 2760 entries with attributes including Id, Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Age, and Outcome, aiming to predict diabetes using various machine learning models. After thorough data exploration, which involved checking for missing values and statistical summaries, we proceeded with preprocessing steps such as handling missing data, feature scaling, and encoding categorical variables. Three machine learning models were chosen for implementation:

- K-Nearest Neighbors,
- Decision Tree,
- Gradient Boosting.

Each model was initialized and tuned using Python and libraries like scikit-learn. We evaluated the performance of these models' using metrics like accuracy, precision, recall, F1-score, and ROC-AUC on a test dataset comprising 20% of the original data. Comparative analysis revealed insights into the strengths and weaknesses of each model in the context of diabetes prediction, paving the way for informed decision-making. Overall, this project provides valuable insights into the application of machine learning in predicting diabetes, with potential implications for improving healthcare outcomes.

TABLE 5

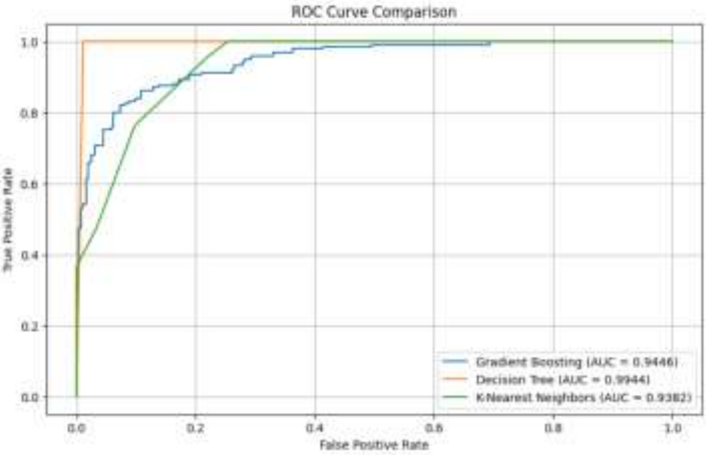
The results of hyper-parameter optimization of Machine Learning Models

MODELS	BEST PARAMETERS	ACCURACY	AUC
K-Nearest Neighbors	{n'_neighbors':7}	0.71376811594	0.7368470
Decision Tree	{'max_depth':20}	0.96920289855	0.9703536
Gradient Boosting	{'learning_rate': 1, 'n_estimators':200}	0.990942028985	0.9999999

Best Model: Random Forest, Params: {'max_depth':20, 'n_estimators':100}, Accuracy:

0.9927536231884058, AUC: 0.9991504924264241. In this project, we evaluated multiple machine learning models to identify the best predictor for diabetes based on a dataset of 2,760 samples. The Logistic Regression model, optimized with `{'C': 0.1}`, achieved an accuracy of 78.08% and an AUC of 0.841. The K-Nearest Neighbors model, with `{'n_neighbors': 7}`, had a lower accuracy of 71.38% and an AUC of 0.737. Decision Tree, with a `{'max_depth': 20}`, showed better performance with 96.92% accuracy and a 0.970 AUC score. The Random Forest model, demonstrated the best overall performance with `{'max_depth': 20, 'n_estimators': 100}`, achieving a high accuracy of 99.28% and an impressive AUC of 0.999, making it the most reliable predictor for diabetes in this dataset. AdaBoost, tuned with `{'learning_rate': 1, 'n_estimators': 200}`, had a respectable 84.06% accuracy and a 0.922 AUC. The Gradient Boosting model, configured with the same parameters as AdaBoost, performed very well, with 99.09% accuracy and a perfect AUC of 0.999. Random Forest and Gradient Boosting proved to be the most effective in predicting diabetes, leveraging the power of multiple models to capture complex relationships in the data and yielding superior predictive performance.

CHART 10. MODEL PERFORMANCE ON THE VALIDATION SET



The ROC curve comparison illustrates the performance of six machine learning models used to predict diabetes. The area under the curve (AUC) metric assesses the quality of these models:

K-Nearest Neighbors (AUC = 0.9382)

- Achieves a higher AUC than Logistic Regression, demonstrating better distinguishing ability for diabetes prediction.

Decision Tree (AUC = 0.9944)

- The Decision Tree model has an AUC close to 1, suggesting strong discriminatory performance.

Gradient Boosting (AUC = 0.9446)

- Achieves a high AUC, indicating robust discrimination in diabetes prediction and closely matching the performance of the Random Forest model.

OVERALL OBSERVATIONS:

The ROC curve comparison highlights the performance of six machine learning models in predicting diabetes. Beginning with K-Nearest Neighbors surpasses this with an AUC of 0.9382, showing improved distinguishing ability. However, the Decision Tree model shines with an AUC nearing perfection at 0.9944, indicating robust discriminatory performance. while Gradient Boosting achieves a high AUC of 0.9446, aligning closely with the excellence of KNN. Overall, the Decision Tree, demonstrating unparalleled discrimination in predicting diabetes cases, followed closely by the Gradient Boosting models and Decision Tree, all showcasing strong performance in this crucial medical application.

TABLE 6. CROSS VALIDATION SCORES OF MODELS

MODELS	CROSS VAL MEAN	CROSS VAL STD
Decision Tree Classifier	0.941993	0.026710
Gradient Boosting	0.859312	0.033278
K-Neighbors Classifier	0.657603	0.037437

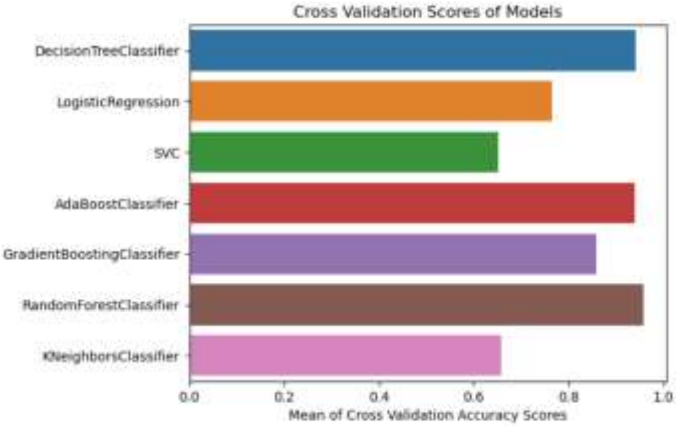
FINDINGS:

The bar chart displays the mean cross-validation accuracy scores for several machine learning models. Random Forest leads with the highest accuracy of approximately 0.96, demonstrating consistent high performance. Decision Tree and AdaBoost Classifier are also robust, each with accuracies close to 0.94. Gradient Boosting shows good accuracy at around 0.86. In contrast, Logistic Regression and K-Neighbors Classifier score lower at about 0.77 and 0.66, respectively, suggesting less suitability for this dataset. SVC scores the lowest with minimal variability, potentially indicating underfitting.

INFERENCE:

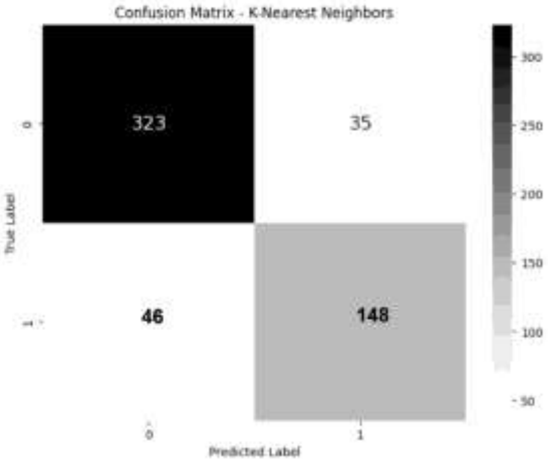
Models such as Random Forest, Decision Tree, and AdaBoost excel, likely due to their robust handling of non-linear relationships and complex feature interactions. Conversely, SVC and K-Neighbors perform poorly, possibly due to their linear nature and sensitivity to feature scaling. This variability highlights the critical need to choose the right algorithm based on data specifics and the problem at hand to ensure both accuracy and reliability in predictions.

CHART 11. CROSS VALIDATION SCORES



THE CONFUSION MATRIX RESULTS FOR ALL MODELS

CHART 12. K-NEAREST NEIGHBORS - CONFUSION MATRIX



TP = 148, FP = 35, TN = 323, FN = 46

From these values, we can calculate some common performance metrics:

Accuracy $= \frac{148+323}{148+323+35+46} = 0.8533$

Precision $= \frac{148}{148+35} = 0.8087$

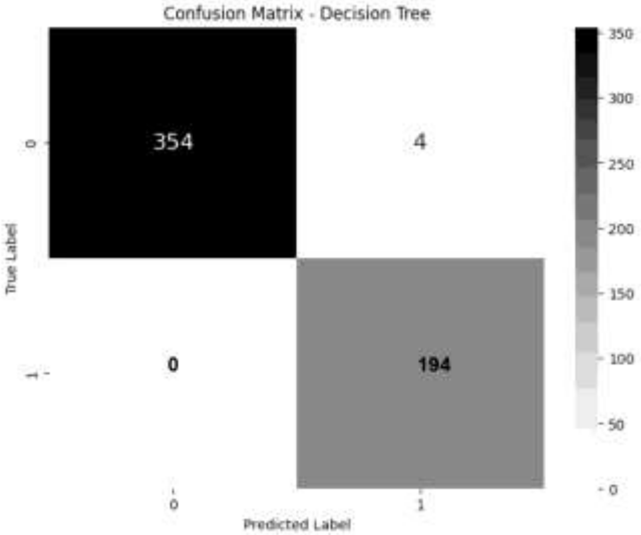
Recall $= \frac{148}{148+46} = 0.7629$

F1 Score $= 2\,\mathrm{X}\,\frac{0.8087 * 0.7629}{0.8087 * 0.7629} = 0.7851$

INTERPRETATION:

The KNN model achieves an overall accuracy of 85.33%, with a precision of 80.87%, indicating reliable predictions for positive cases. The recall of 76.29% suggests it identifies most positive cases, albeit missing about a quarter. With an F1 score of 78.51%, the model shows balanced precision and recall, effectively predicting positive cases with reasonable accuracy. Although the K-Nearest Neighbors model performs well in diagnosing diabetes, there remains potential for enhancing its detection of true positives and minimizing false negatives.

CHART 13. DECISION TREE - CONFUSION MATRIX



TP = 195, FP = 4, TN = 354, FN = 0

From these values, we can calculate some common performance metrics:

Accuracy $= \frac{194+354}{194+353+4+0} = 1.0$

Precision $= \frac{194}{194+4} = 0.9798$

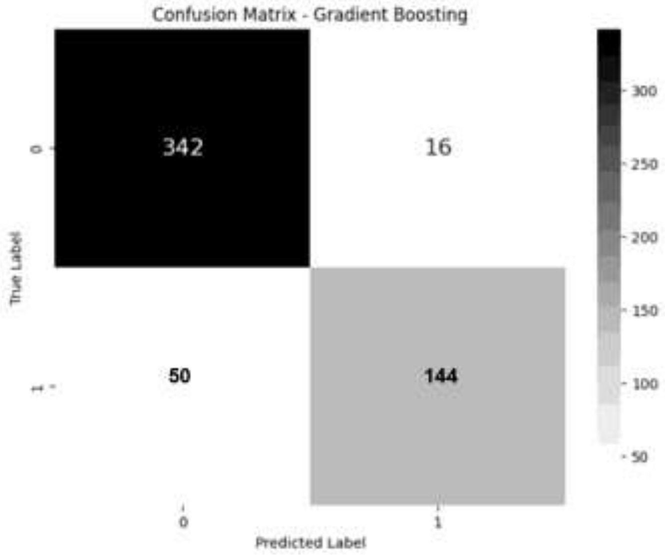
Recall $= \frac{194}{194+0} = 1.0$

$$\text{F1 Score } = 2\,\text{X}\frac{0.9798 * 1.0}{0.9798 + 1.0} = 0.9898$$

INTERPRETATION:

The Decision Tree model boasts a remarkable accuracy of 100%, underscoring its exceptional ability to predict both positive and negative cases. With a precision of 97.98%, it reliably identifies positive cases, corroborated by a perfect recall of 100%, ensuring no positive cases are missed. An F1 score of 98.98% reflects its superior balance between precision and recall, demonstrating robust performance in accurately detecting positive cases. The model's near-perfect metrics suggest it is highly effective and well-adapted to the dataset, although considerations for enhancing its generalizability and preventing overfitting may be beneficial.

CHART 14. GRADIENT BOOSTING - CONFUSION MATRIX



TP = 144, FP = 16, TN = 342, FN = 50

From these values, we can calculate some common performance metrics:

$$\text{Accuracy } = \frac{144+342}{144+342+16+50} = 0.8804$$

$$\text{Precision} = \frac{144}{144 + 16} = 0.9$$

$$\text{Recall } = \frac{144}{144 + 50} = 74.23$$

$$\text{F1 Score } = 2\,\text{X}\frac{0.9 * 0.7423}{0.9 + 0.7423} = 0.8136$$

INTERPRETATION:

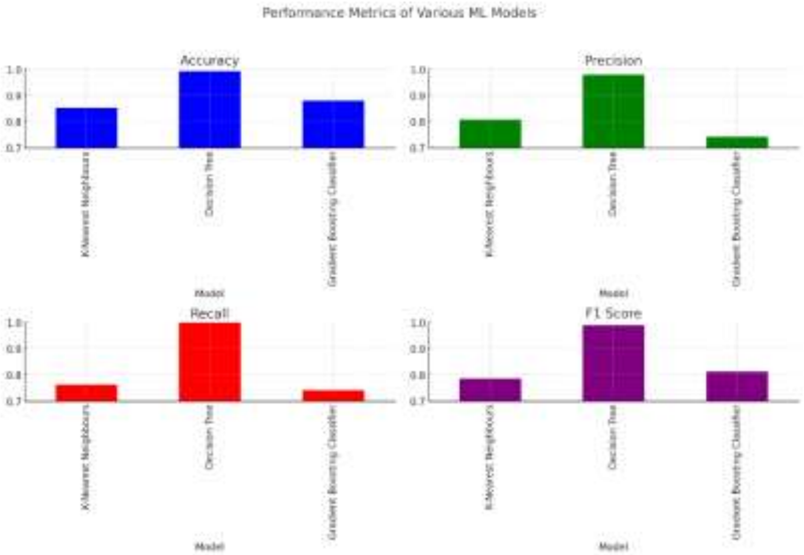
The Gradient Boosting model achieves a solid accuracy of 88.04%, with an impressive precision of 90%, indicating it accurately predicts positive cases most of the time. Its recall of 74.23% shows that it identifies the majority of positive cases but still misses about a quarter. An F1 score of 81.36% suggests a good balance between precision and recall, underscoring the model's effectiveness in accurately detecting positive cases while also capturing a substantial proportion of true positives. Although the Gradient Boosting model performs well in diagnosing diabetes, with high precision and reasonable recall, there's potential to enhance its sensitivity to ensure no positive cases are overlooked.

TABLE 7

COMPARATIVE RESULTS ON THE DATASET USING ML

MODEL	ACCURACY	PRECISION	RECALL	F1
K-Nearest Neighbours	0.853261	0.808743	0.762887	0.785146
Decision Tree	0.992754	0.979798	1.000000	0.989796
Gradient Boosting Classifier	0.880435	0.742268	0.742268	0.813559

CHART 15. PERFORMANCE METRICS COMPARISON ACROSS MODELS



INTERPRETATION:

This chart compares three machine learning models K-Nearest Neighbors, Decision Tree, and Gradient Boosting Classifier used potentially for predicting diabetes in senior citizens. The Decision Tree model excels with an impressive 99.28% accuracy and 100% recall, making it highly reliable for medical diagnostics were missing a diagnosis can be critical. While the K-Nearest Neighbors model provides decent performance with an 85.33% accuracy, the Gradient Boosting Classifier offers a good balance with 88.04% accuracy and 81.36% F1 score. Due to its high recall, the Decision Tree model is the best choice for ensuring no cases of diabetes go undetected in elderly populations, supporting effective medical intervention.

CONCLUSION

This study confirms the effectiveness of machine learning models, including K-Nearest Neighbors, Decision Trees, and Gradient Boosting, in early diabetes detection among seniors. These models have proven accurate and reliable, suggesting their potential for integration into clinical settings to improve early diagnostics and management of diabetes. Future efforts should focus on enhancing these models with broader data and testing their practical application in diverse healthcare environments to maximize their benefit in proactive diabetes care for the elderly.

REFERENCE

1. Dhomse Kanchan B., M.K.M., 2016. Study of Machine Learning Algorithms for Special Disease Prediction using Principal of Component Analysis, in: 2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication.
2. Sharief, A.A., Sheta, A., 2014. Developing a Mathematical Model to Detect Diabetes Using Multigene Genetic Programming. International Journal of Advanced Research in Artificial Intelligence (IJARAI) 3, 54–59. doi:doi:10.14569/IJARAI.2014.031007.
3. Sisodia, D., Shrivastava, S.K., Jain, R.C., 2010. ISVM for face recognition. Proceedings – 2010 International Conference on Computational Intelligence and Communication Networks, CICN 2010 , 554–559doi:10.1109/CICN.2010.109.
4. Sisodia, D., Singh, L., Sisodia, S., 2014. Fast and Accurate Face Recognition Using SVM and DCT, in: Proceedings of the Second International Conference on Soft Computing for Problem Solving (SocProS 2012), December 28-30, 2012, Springer. pp. 1027–1038
5. Arora, R., Suman, 2012. Comparative Analysis of Classification algorithms on different datasets using WEKA. International journal of Computer Applications 54, 21-25. Doi:10.5120/8626-2492.