Diabetes Prediction Using Medical and Demographic Factors

1st Kanika Saxena
ksaxena2@buffalo.edu

2nd Rajashree Shanmuganathan
rajashre@buffalo.edu

3rd Alankriti Dubey
adubey2@buffalo.edu

I. Problem Statement

This report aims to identify individuals at higher risk of developing diabetes based on demographic and medical factors, such as obesity, age, and lifestyle choices like smoking. By analyzing these factors, the report seeks to pinpoint specific populations more susceptible to diabetes and highlight the conditions that contribute to its onset, helping to guide targeted intervention and prevention efforts.

A. Background

Diabetes affects millions of people worldwide, and many of those cases are driven by medical and demographic factors. Predictive models using basic medical and demographic data can offer an early warning system, allowing timely detection and reducing patient complications. This is significant, especially in settings where routine testing may not be feasible, as well as early detection and timely treatment of potential patients.

B. Contribution to the Problem Domain

This project has the potential to significantly enhance diabetes management by:

- Early Risk Identification: Helping healthcare providers focus on individuals at high risk, enabling preventive procedures.
- Healthcare Efficiency: Reducing the reliance on extensive testing through accurate, data-driven predictions using basic medical and demographic data.
- Personalized Treatment: Supporting tailored healthcare plans based on each individual's unique risk profile.
- Research Insights: Advancing understanding of the relationships between medical and demographic factors and diabetes risk.

II. Data Source

The dataset used for this project is the Diabetes Prediction Dataset obtained from Kaggle. It contains around 100,000 rows and 9 columns, with the target variable being Diabetes, where a value of 1 indicates the presence of diabetes and 0 indicates its absence. The dataset includes several medical and demographic features such as age, gender, BMI, hypertension, heart disease, HbA1c level, and blood glucose level.

This dataset has been utilized in over 240 Kaggle projects, and has also been referenced in various research papers focusing on diabetes prediction and machine learning applications in healthcare. Its large size and diverse features make it a valuable resource for understanding the factors contributing to diabetes risk.

The dataset can be accessed at: Kaggle Diabetes Prediction Dataset

III. Data Cleaning

In this section, we describe the various data cleaning steps undertaken to prepare the dataset for analysis.

A. Dropping Duplicate Rows

Duplicate rows were removed to ensure data integrity. This was done using the following command:

```
df = df.drop_duplicates()
```

The shape of the dataset after removing duplicates was checked to confirm the changes.

B. Converting Categorical Labels to Numerical Format

Categorical labels in the dataset were converted to a numerical format to facilitate analysis. Specifically, the `gender` column was encoded using a label encoder:

```
label_encoder = LabelEncoder()
df['gender'] = label_encoder.fit_transform(df
    ['gender'])
```

C. Handling Outliers for Age

Outliers in the `age` column were managed by ensuring that all age values were integers:

```
df = df[df['age'] == df['age'].astype(int)]
```

D. Handling Outliers for BMI

To manage outliers in the `bmi` column, the interquartile range (IQR) method was used. The outliers were managed by clipping the `bmi` values to the lower and upper limits.

E. Filtering Invalid Biological Values

Rows with biologically invalid values were removed. The filtering conditions were:

- Age greater than 1
- BMI greater than 10
- Blood glucose level of at least 80

F. Cleaning Text Data

The `smoking_history` column was processed as follows:

- The first letter of each string was capitalized.
- Entries labeled as `No info` were replaced with the mode of the column.

G. Normalizing Blood Glucose and HbA1c Levels

The `blood_glucose_level` and `HbA1c_level` columns were normalized using a standard scaler.

H. Dropping Null Values

To address missing data, rows with null values were dropped:

```
df = df.dropna()
```

I. Categorizing BMI into Ranges

The BMI values were categorized into different ranges to classify individuals as underweight, normal, overweight, or obese.

J. Saving the Cleaned Dataset

Finally, the cleaned dataset was saved to a CSV file:

```
df.to_csv('diabetes_cleaned.csv', sep='\t')
```

IV. Exploratory Data Analysis (EDA)

A. Distribution by Gender

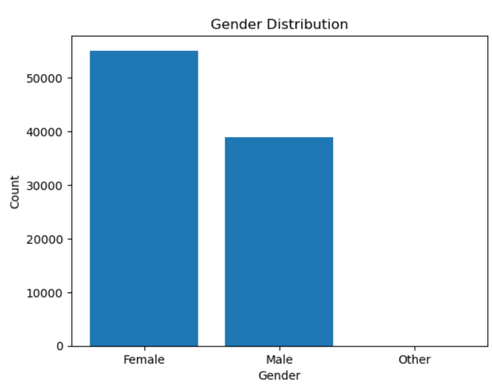

Fig. 1. Distribution by Gender

The histogram shows that a higher proportion of diabetic patients are female compared to males, suggesting a potential gender-related factor in diabetes commonness.

B. Distribution of BMI

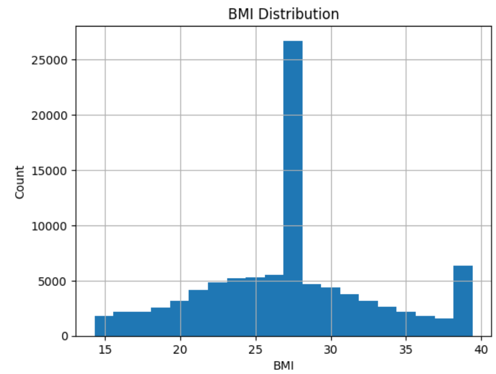

Fig. 2. Distribution of BMI

Approximately 70% of patients in the dataset are classified as either overweight or obese, highlighting a significant relationship between weight and diabetes risk. Given this concerning observation, it is crucial to maintain a healthy weight in order to reduce the likelihood of developing diabetes and similar health complications.

Below is the weight distribution of patients in the table:

Classification	**Count**
Overweight	41,701
Obesity	23,524
Healthy Weight	21,770
Underweight	7,055

C. Smoking Habits of Patients

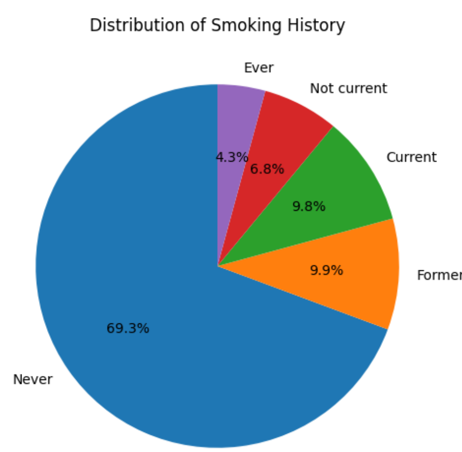

Fig. 3. Smoking Habits of Patients

The visualization of smoking history shows that approximately 70% of patients have never smoked, while 9.7% are current smokers. While the data shows that many patients have never smoked, this doesn't mean that smoking is irrelevant to diabetes risk. Current and former smokers may face higher complications related to diabetes.

D. Glucose Levels vs BMI

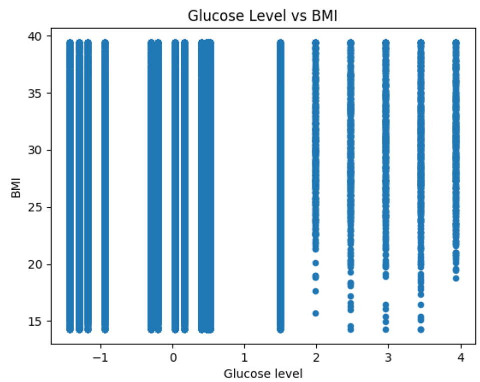

Fig. 4. Glucose Levels vs BMI

This graph concludes that glucose levels can vary independently of BMI values, as shown by the similar glucose levels across a range of BMI classifications. However, higher glucose levels are more commonly observed in patients with a high BMI, indicating a potential association between obesity and glucose dysregulation.

E. Diabetes Rates with Age

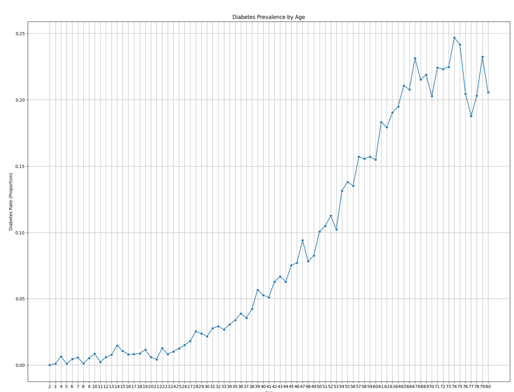

Fig. 5. Diabetes Rates with Age

The line chart clearly demonstrates that the rate of diabetes increases with age, showing a general upward trend despite some slight fluctuations. This highlights the growing risk of diabetes among older populations.

F. BMI Distribution Across Age Groups

The line chart clearly demonstrates that the rate of diabetes increases with age, showing a general upward trend despite some slight fluctuations. This highlights the growing risk of diabetes among older populations.

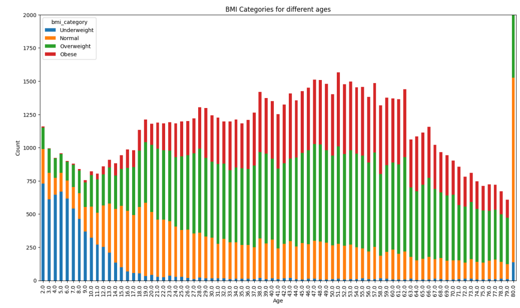

Fig. 6. BMI Distribution Across Age Groups

G. BMI vs Diabetes Risk

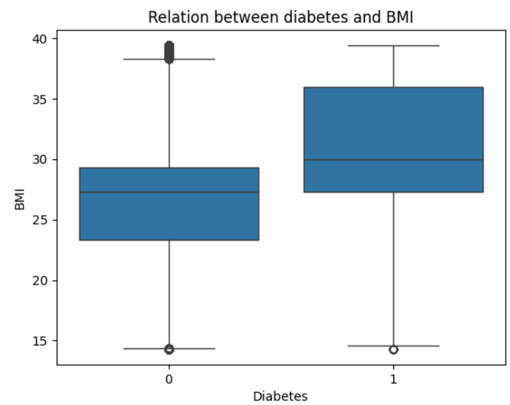

Fig. 7. BMI vs Diabetes Risk

The box plot illustrates that diabetic patients tend to have higher BMIs, with the median BMI falling between 30 and 35. In contrast, non-diabetic individuals have a lower median BMI, around 25 to 30. This suggests that a higher BMI is strongly associated with an increased likelihood of diabetes.

V. Key Outcomes and Insights

The analysis shows that higher BMI (27-37 range), older age (especially 40+), and female gender are strongly linked to a higher risk of diabetes. Additionally, a portion of diabetic patients have a history of smoking.

Key measures include promoting weight management, early screening for older populations and women, and reinforcing smoking cessation programs to reduce diabetes risk.

This analysis will inform the feature selection process for predictive modeling, with BMI, age, and gender identified as the most significant predictors of diabetes risk.

VI. Machine Learning Models

In this section, we present the outcomes of our analysis using six different machine learning models. Each model was selected and tuned to address our specific problem statement, and the results of their performance are detailed below. We provide a comprehensive evaluation of each model, including accuracy, precision, recall, F1-score, and other relevant metrics, along with insights gained from their application.

A. *Logistic Regression*

*1) **Justification for Choosing Logistic Regession**:* We have chosen the Logistic Regression algorithm based on the following reasons –

1) **Nature of the Problem :** The problem at hand—predicting diabetes based on medical and demographic factors—is fundamentally a binary classification task. Logistic regression is well-suited for such tasks as it models the probability of a binary outcome, making it a natural choice for predicting the presence or absence of diabetes (0 or 1).
2) **Interpretability:** One of the critical advantages of logistic regression is its interpretability. The model coefficients can provide insight into the relationship between each predictor and the likelihood of diabetes, allowing for meaningful medical and demographic interpretations. This is crucial in healthcare, where understanding risk factors can inform treatment and prevention strategies.
3) **Handling of Predictor Types:** Logistic regression effectively handles both continuous predictors (like age and BMI) and categorical predictors (like gender). The ability to include both types of variables without extensive preprocessing makes it a practical choice for this analysis.

*2) **Model Training and Tuning***

: 1. Data Preprocessing :

Before training the model, essential preprocessing steps were taken, including:

- Handling Missing Values: Any missing data in the dataset was addressed to ensure a complete dataset for training.
- Handling Data Inconsistency: Removing data inconsistency by making all string values in title case.
- Encoding Categorical Variables: Gender and smoking history were encoded using techniques such as one-hot encoding to convert categorical data into a numerical format suitable for logistic regression.

2. Model Training :

- **Splitting the Data:** The dataset was split into training and test sets (typically an 80-20 split) to evaluate model performance effectively.
- **Fitting the Model**: The logistic regression model was trained using the training set. Hyperparameters such as regularization were adjusted using cross-validation to prevent overfitting and enhance generalization.

*3) **Effectiveness of the Algorithm**:*

1. Performance Metrics

Several metrics were utilized to evaluate the effectiveness of the logistic regression model:

label=•

- **Accuracy:** Measures the overall correctness of the model. Given the class distribution in the dataset, accuracy alone might be misleading.
- **Precision and Recall:** These metrics are crucial in healthcare applications where false negatives (missed diabetes cases) can have serious consequences. High precision indicates a low false positive rate, while high recall indicates a low false negative rate.

2. Model Performance Results

Upon evaluation, the logistic regression model achieved:

label=•

- **Accuracy:** Approximately 95.68%, suggesting that the model performs well in classifying diabetes cases.
- **Macro Average:** label=–
  - **Precision:** 0.91 indicates the average precision across classes, treating all classes equally, indicating that when the model predicts a patient has diabetes, it is correct 91% of the time.
  - **Recall:** 0.80 indicates the average recall, again treating all classes equally, which means the model correctly identifies 80% of actual diabetes cases.
  - **F1-Score:** 0.85 is the average F1-score across classes, indicating a balanced trade-off between precision and recall.
- **Weighted Average:** label=–
  - **Precision:** 0.95, reflecting the average precision while considering the number of instances in each class.
  - **Recall:** 0.96, indicating a strong overall recall.
  - **F1-Score:** 0.95 shows a strong balance between precision and recall when accounting for class imbalance.

Insights Gained from the Algorithm

1) **Key Predictors:** The analysis revealed that higher BMI, older age, and female gender are significantly associated with an increased risk of diabetes. This supports public health initiatives focusing on obesity and age as critical factors in diabetes management.
2) **Impact of Lifestyle Choices:** While the dataset indicated that a majority of patients are non-smokers, the presence of diabetes in current and former smokers emphasizes the need for continued awareness of lifestyle choices.
3) **Data-Driven Interventions:** The insights from the predictive model allow for tailored interventions, such as: label=–
   - Targeted screenings for older populations and individuals with higher BMI.
   - Implementing educational programs focused on weight management and smoking cessation.
   - Developing personalized treatment plans that take into account individual risk profiles based on demographic and medical factors.
4) **Future Research Directions:** The analysis opens avenues for further research, such as investigating the impact of other lifestyle factors (e.g., diet, physical activity)

on diabetes risk. It may also provide a basis for exploring more complex models (like ensemble methods) to see if predictive performance can be improved.

B. Decision Tree

*1) **Justification for Choosing Decision Tree Algorithm**:* We have chosen the Decision Tree algorithm based on the following reasons –

1) **Interpretability**: Decision Trees are highly interpretable compared to many other machine learning models. The flowchart-like structure allows stakeholders (e.g., healthcare providers) to understand how decisions are made based on specific features. This transparency is crucial in healthcare, where understanding the rationale behind predictions can inform clinical decisions.
2) **Handling Non-Linear Relationships**: The nature of the data in diabetes prediction often includes non-linear relationships between features and the target variable. Decision Trees can capture these relationships effectively without requiring extensive preprocessing or transformation of the data.
3) **Feature Importance**: Decision Trees inherently provide a mechanism to assess feature importance, allowing us to identify which demographic and medical factors are most influential in predicting diabetes risk. This is particularly useful for informing further research and preventive strategies.

*2) **Model Training and Tuning***
: 1. Data Preprocessing :
Before training the model, essential preprocessing steps were taken, including:

- Missing values were checked and handled accordingly.
- Categorical features were converted to numerical format using ordinal encoding.
- The dataset was split into training and testing subsets to evaluate model performance accurately.

2. Model Training :

- The initial model was trained using the default parameters of the DecisionTreeClassifier.
- The model was fitted to the training data using `dt_classifier.fit()`.

3. Feature Selection :

- Features were assessed for their importance, which helped refine the feature set and remove irrelevant or redundant features, improving model accuracy and interpretability.

*3) **Effectiveness of the Algorithm**:*

1. Performance Metrics

Several metrics were utilized to evaluate the effectiveness of the logistic regression model:
label=•

- **Accuracy:** Measures the overall correctness of the model. Given the class distribution in the dataset, accuracy alone might be misleading.
- **Precision and Recall:** These metrics are crucial in healthcare applications where false negatives (missed diabetes cases) can have serious consequences. High precision indicates a low false positive rate, while high recall indicates a low false negative rate.

2. Model Performance Results

Upon evaluation, the Decision Tree model achieved:
label=•

- **Accuracy:** The model achieved an accuracy of approximately 95%, indicating a high level of correct predictions overall.
- **Precision:**
  - For Class 0 (non-diabetic): 0.97 (This means 97% of the instances predicted as non-diabetic were correct.)
  - o For Class 1 (diabetic): 0.72 (This means 72% of the instances predicted as diabetic were correct.)
- **Recall:**
  - For Class 0 (non-diabetic): 0.97 (This means 97% of the actual non-diabetic cases were correctly identified.)
  - For Class 1 (diabetic): 0.74 (This means 74% of the actual diabetic cases were correctly identified.)
- **F1-Score:**
  - For Class 0 (non-diabetic): 0.97 (This indicates a very strong performance for predicting non-diabetic patients.)
  - For Class 1 (diabetic): 0.73 (This shows moderate performance for predicting diabetic patients.)

Insights Gained from the Algorithm

- Feature importance analysis revealed that certain factors (e.g., BMI, age, and gender) were significantly associated with diabetes risk.
- The decision tree visualization helped identify specific thresholds for features (e.g., BMI ¿ 30) that are critical indicators of diabetes, which can guide public health interventions and individual risk assessments.
- The model's performance on different classes provided insights into the need for targeted interventions for populations at risk, especially older individuals and those with high BMI.

C. K-Nearest Neighbors (KNN)

*1) **Justification for Choosing KNN**:* Using K-Nearest Neighbors (KNN) for predicting diabetes in this dataset is justified for these reasons:

1) **Binary Classification:** KNN is ideal for binary classification tasks, directly predicting diabetes status (0 for non-diabetic, 1 for diabetic).
2) **Non-Linear Relationships:** KNN effectively captures complex, non-linear relationships between features like

age, BMI, and HbA1c level, which are common in medical data.
3) **No Distribution Assumptions:** As a non-parametric method, KNN does not assume any specific distribution for input features, making it suitable for diverse medical data.
4) **Mixed Data Types:** KNN handles both categorical (e.g., gender, smoking history) and continuous features (e.g., age, BMI), allowing for a comprehensive analysis.
5) **Hyperparameter Tuning:** The ability to optimize k (the number of neighbors) through cross-validation enhances model performance.
6) **Focus on Local Patterns:** KNN emphasizes local data structure, effectively identifying patterns within demographic groups relevant to diabetes risk.

*2) **Work Done for Training the Dataset**:* The main hyperparameter for KNN is the number of neighbors (k). To determine the optimal value of k, a range of values (1 to 25) was tested, and the model's accuracy was evaluated for each. The dataset was divided into training and testing sets in an 80:20 ratio using a random state of 42.

- Number of training rows: 75,240
- Number of testing rows: 18,810

Choosing a suitable k is crucial because a smaller k can lead to overfitting, while a larger k might cause underfitting.

*3) **Effectiveness**:* The KNN model demonstrated strong overall performance in predicting diabetes, achieving a maximum accuracy of **95.09%** for k = **7** on the test dataset, indicating that it correctly classified a substantial majority of the values.

***Key Metrics**:*

- **Accuracy:** 0.9509
- **Precision:** 0.9107 (for positive class)
- **Recall:** 0.5085 (for positive class)
- **F1 Score:** 0.6526 (for positive class)

Detailed Classification Report: **For Class 0 (No Diabetes):**

- **Precision:** 0.95: This indicates that 95% of the instances predicted as "No Diabetes" were correctly classified.
- **Recall:** 1.00: The model identified all actual "No Diabetes" cases correctly.
- **F1 Score:** 0.97: Harmonic mean of precision and recall, showing a very high performance for this class.

For Class 1 (Diabetes):

- **Precision:** 0.91: 91% of the instances predicted as "Diabetes" were indeed diabetes cases.
- **Recall:** 0.51: The model only identified about 51% of actual diabetes cases, indicating it misses a considerable number of true positives.
- **F1 Score:** 0.65: This score reflects a balance between precision and recall but suggests room for improvement in detecting diabetes.

Conclusion The KNN algorithm achieved a high **accuracy** of 95%, indicating that it correctly predicted most of the outcomes. However, the model's **recall** (0.51) was lower compared to precision. This is also reflected in the **confusion matrix**, which shows 838 false negatives. The **ROC curve area** of 0.89 demonstrates that the model has a strong ability to differentiate between positive and negative cases, but there is still room for improvement in detecting positive cases.

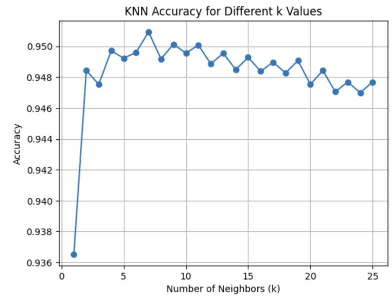

Fig. 8. Accuracy Accuracy for k-values (Best $k = 7$)

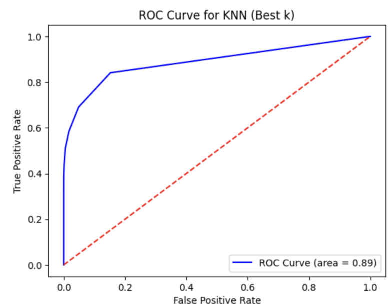

Fig. 9. ROC Curve for KNN (Best $k = 7$)

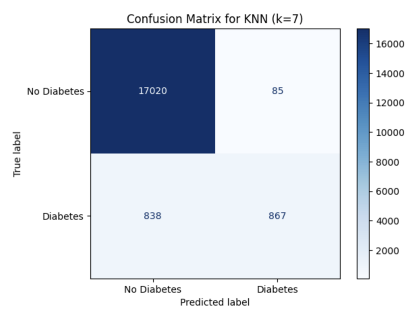

Fig. 10. Confusion Matrix for KNN (Best $k = 7$)

***Intelligence Gained from KNN Prediction**:* The application of the KNN algorithm provided several insights into the diabetes prediction task:

- **Optimal K Value:** The best value of k was found to be 7, which yielded the highest accuracy of approximately 95.09%. This emphasizes the significance of hyperparameter tuning in achieving optimal model performance.
- **High Precision for Non-Diabetic Cases:** The model demonstrated a precision of 0.95 for predicting non-diabetic cases, indicating that 95% of the predicted instances for this class were accurate.

- **Challenges in Detecting Diabetic Cases:** The recall for diabetic cases was only 0.51, suggesting that the model missed a considerable number of true positives, highlighting the need for model improvement.
- **F1 Score Indicating Trade-Offs:** An F1 score of 0.65 for diabetic cases reflects the trade-off between precision and recall, highlighting room for improvement in the model's ability to identify positive instances.

D. Random Forest Algorithm

*1) **Justification for Choosing Random Forest Algorithm**:* Using Random Forest Algorithm for predicting diabetes in this dataset is justified for these reasons:

1) **Efficient for Large Datasets**: With 95,000 rows, Random Forest processes large datasets effectively by building multiple trees in parallel.
2) **Suitable for Binary Classification**: It excels in binary tasks (0: No Diabetes, 1: Diabetes) by combining predictions from multiple decision trees for better accuracy.
3) **Robust to Outliers**: The ensemble method reduces sensitivity to outliers, such as varying blood glucose or BMI values.
4) **Feature Importance**: Identifies key factors (e.g., age, BMI, HbA1c level) that significantly impact diabetes prediction, offering useful insights.

*2) **Work Done for Training the Dataset**:* The Random Forest model was configured with 100 decision trees (*n_estimators* = 100), chosen to balance performance and computational efficiency. Setting a fixed random state ensured reproducibility of results. This configuration helped achieve accurate predictions for the binary classification of diabetes. The dataset was split into training and testing sets using an 80:20 ratio, with a random state of 42.

- **Training set**: 75,240 samples
- **Testing set**: 18,810 samples

*3) **Effectiveness**:* The Random Forest model demonstrated strong overall performance on the diabetes prediction dataset. Key evaluation metrics include:

- **Accuracy:** 96.60%
- **Precision:** 92.84% (for positive class)
- **Recall:** 67.68% (for positive class)
- **F1 Score:** 78.29% (for positive class)

Detailed Classification Report: For Class 0 (No Diabetes):

- **Precision:** 97% — The model correctly identified 97% of cases predicted as "No Diabetes."
- **Recall:** 99% — Almost all actual "No Diabetes" cases were accurately detected.
- **F1 Score:** 98% — High precision and recall led to a strong F1 score for this class.

For Class 1 (Diabetes):

- **Precision:** 93% — The model accurately identified 93% of instances predicted as "Diabetes."
- **Recall:** 68% — The model detected 68% of actual diabetes cases, indicating it missed some true positives.
- **F1 Score:** 78% — Reflects a balance between precision and recall but also suggests further tuning could improve sensitivity.

Conclusion The Random Forest model achieved an **accuracy** of 96.60%, indicating its effectiveness in predicting diabetes outcomes. The model is reliable in identifying true positives; however, it missed a high proportion of actual diabetes cases. Additionally, the **ROC curve area** of 0.96 indicates discriminative ability between diabetic and non-diabetic patients, highlighting the model's overall robustness with potential for improvement in detecting positive cases.

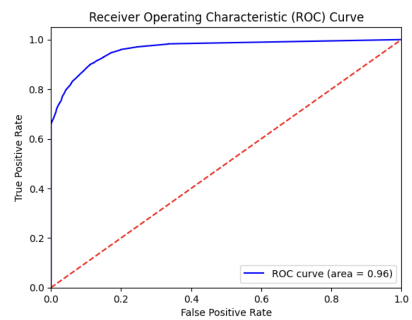

Fig. 11. ROC Curve for Random Forest

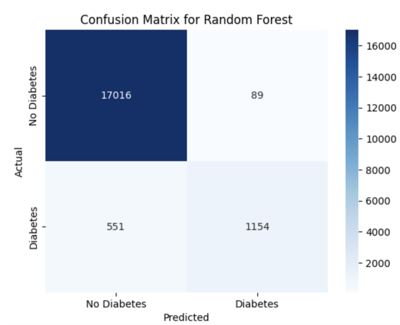

Fig. 12. Confusion Matrix for Random Forest

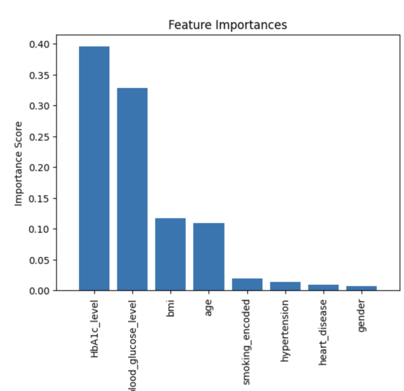

Fig. 13. Important Features for Prediction

Intelligence Gained from Random Forest Prediction: The Random Forest model provided valuable insights for diabetes prediction:

- **High Accuracy:** The model achieved an impressive accuracy of approximately 96.6%, effectively classifying the majority of diabetic and non-diabetic patients, and showcasing its reliability in real-world applications.
- **Precision-Recall Dynamics:** While the precision for diabetes cases was 92.8%, the recall of 67.7% indicates that a high number of diabetic patients may have been overlooked, highlighting the need for model refinement.
- **Feature Importance Insights:** Key predictors such as HbA1c level, blood glucose level, BMI, and age significantly influence diabetes classification, providing a basis for targeted interventions and risk assessment strategies.
- **Robust Class Handling:** The model maintained a high precision of 97.0% and a recall of 99.0% for non-diabetic cases, demonstrating its capability to effectively differentiate between the two classes and reduce the likelihood of false positives.

E. SVM

*1) **Justification for Choosing SVM:*** The SVM model is used for the following reasons:

- **Effective for Non-linear Data:** SVM is effective for handling non-linear correlations between features using the kernel method.
- **High Dimensional Data:** SVM is effective in high-dimensional spaces and can handle both linear and non-linear relationships. For diabetes prediction, it's useful in drawing decision boundaries in a complex feature space.
- **Resistance to Over-fitting:** SVM's margin-based method to classification reduces the risk of over-fitting, particularly in high-dimensional fields. SVM can better generalize to previously unseen data by maximizing the margin between classes,

*2) **Work Done to Train the Model:***

1) **Data Preprocessing:** The data is made sure that it is preprocessed by handling missing values,encoding categorical values
2) **Standardization:** Required feature scaling to ensure the effectiveness of SVM's distance-based decision boundaries.
3) **Splitting the Data:** The data was split into training and testing sets using an 80:20 ratio.
4) **Model Tuning:** The linear kernel was selected due to its effectiveness in high-dimensional spaces and its ability to create a clear decision boundary for linearly separable data.
5) **Hyper-parameters tuning:** The C (regularization) and gamma(kernel coefficient) were also tuned.
6) **Splitting the Data:** The data was split into training and testing sets using an 80:20 ratio.

*3) **Effectiveness:*** The following metrics were evaluated on the SVM model .The accuracy of the model is **95%** on the test dataset, indicating that it classified a majority of the values.

Key Metrics: **For Class 0 (No Diabetes):**

- **Precision:** 96% This indicates that most of the cases are correctly classified.
- **Recall:** 99% This indicates that 99% of the instances predicted as "No Diabetes" .It has almost classified correctly.
- **F1 Score:** 98% This indicates that the model has performed very well

For Class 1 (Diabetes):

- **Precision:** 91% This high precision indicates that the model is good at reducing false positives, making it useful for identifying diabetic individuals.
- **Recall:** 58% This relatively low recall indicates that the model misses a considerable proportion of true diabetes patients.
- **F1 Score:** 71% An F1 score of 71% shows a reasonable balance of precision and recall. While the precision is excellent, the reduced recall lowers the F1 score, indicating that the model's capacity to detect all diabetic patients might be improved.

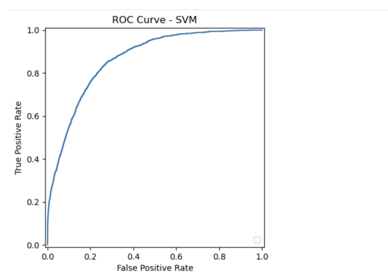

Fig. 14. ROC Curve For SVM

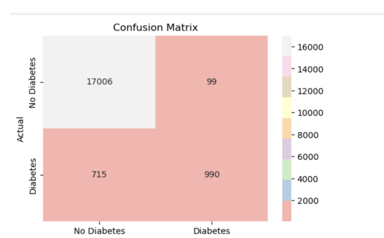

Fig. 15. Confusion Matrix for SVM

Intelligence Gained from SVM Prediction: The following insights are gained from the SVM diabetes prediction task:

- **Effectiveness of Linear Separation:** Using a linear kernel, SVM successfully distinguished between diabetes and non-diabetic patients. This implies that a definite decision limit exists in the feature space, allowing for precise patient classification based on measurable characteristics.

- **Feature Importance:** SVM's performance can help identify which features are most important for diabetes prediction. Analyzing feature weights can assist in identifying crucial indicators, such as glucose levels and BMI, allowing healthcare providers to focus on these parameters during patient examinations.
- **Performance:** The model achieved high precision (91%) but had a lower recall (58%). This efficiency in precision suggests that while the model performs incredibly well in reducing false positives, it may require alters to boost sensitivity and discover more real positives.

F. Naive Bayes

*1) **Justification for Choosing Naive Bayes**:* The Naive Bayes model is used for the following reasons:

- **Probabilistic Interpretation:** Naive Bayes succeeds at providing probabilistic results, such as the likelihood of a patient having diabetes (0 for non-diabetic, 1 for diabetic). This can be useful in a medical situation to inform the classification decision.
- **Handling of Categorical and Continuous Data:** Diabetes datasets typically include both continuous (e.g., blood pressure, glucose levels) and categorical variables (e.g., family history of diabetes). Naive Bayes can be used to process both types of data.
- **Interpretability:** Naive Bayes models are easily interpretable, making them useful for medical applications such as diabetes prediction.

*2) **Work Done to Train the Model**:*

1) **Handling Categorical Features:** The dataset contained categorical variables which were transformed into numerical values using label encoding. This method assigns each unique category a distinct integer.
2) **Tuning:** Since the dataset included continuous features such as age, BMI, and blood glucose levels, Gaussian Naive Bayes was chosen. This variant of Naive Bayes assumes that continuous features follow a normal (Gaussian) distribution, making it suitable for handling these kinds of attributes in the dataset.
3) **Splitting the Data:** The data was split into training and testing sets using an 80:20 ratio.
4) **Training the Model:** The model was trained using the features provided. During this process, the model learned the conditional probabilities of each feature with respect to the target label (diabetes or non-diabetes).

*3) **Effectiveness**:* The following metrics were evaluated on the Naive Bayes model .The accuracy of the model is **90%** on the test dataset, indicating that it correctly classified a substantial majority of the values.

Key Metrics: **For Class 0 (No Diabetes):**

- **Precision:** 96% This indicates that most of the cases are correctly classified.
- **Recall:** 93% This indicates that 93% of the instances predicted as "No Diabetes" were correctly classified.
- **F1 Score:** 95% This indicates that the model has performed well

For Class 1 (Diabetes):

- **Precision:** 48% Precision suffered slightly, as Naive Bayes is prone to more false positives in this context.
- **Recall:** 66% The model only identified about 66% of actual diabetes cases.
- **F1 Score:** 55% the F1 score of the diabetes prediction model is 55%, this indicates that the model's balance between precision and recall is somewhat lacking and can be improved in its ability to make accurate predictions

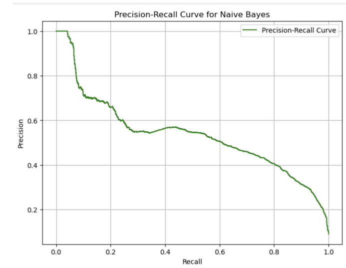

Fig. 16. Precision Recall Curve

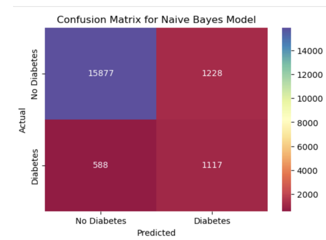

Fig. 17. Confusion Matrix for Naive Bayes

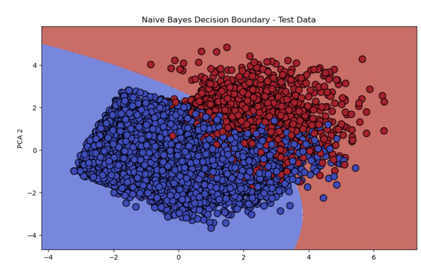

Fig. 18. Naive Bayes on Test Data

***Intelligence Gained from Naive Bayes Prediction**:* The following insights are analysed from the Naive Bayes diabetes prediction:

- **Feature Independence Assumption:** Naive Bayes makes the assumption that all features are independent. But, moderate performance indicates that some features—such as BMI and glucose levels—may really interact with one another, something that Naive Bayes does not account for.

- **Essential Attributes for Diabetes Prediction:** The higher probability of characteristics like age, BMI, and glucose levels indicate that these factors had a greater influence on diabetes prediction. This implies that these are important variables to consider when assessing a person's risk of developing diabetes.
- **Performance:** The Naive Bayes model had slightly lower precision, recall, and F1-score than SVM and Logistic Regression. This suggests that while Naive Bayes provides a fast and interpretable solution, it may not be as accurate for complex problems

REFERENCES

[1] Mustafa, Mohammed. *Diabetes Prediction Dataset: A Comprehensive Dataset for Predicting Diabetes with Medical & Demographic Data*. Kaggle, 2023. Available at: https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset/data

[2] Random Forest: Breiman, L. *Random Forests*. Machine Learning, vol. 45, no. 1, pp. 5-32, 2001. Available at: https://doi.org/10.1023/A:1010933404324

[3] Decision Tree: Quinlan, J.R. *Induction of Decision Trees*. Machine Learning, vol. 1, pp. 81-106, 1986. Available at: https://doi.org/10.1023/A:1022643204877