**GlucoSense- AI-Powered Diabetes Detection for Early Intervention**

A Project Report

Submitted to



Under the supervision of

Ravi

Submitted By

ADITYA KUMAR

**Table of Contents**

|  |  |  |
| --- | --- | --- |
| **S.No** | **Topic** | **Page no.** |
| 1 | Abstract | 1 |
| 2 | Introduction | 2 |
| 3 | Problem statement | 3 |
| 4 | Data Collection | 4 |
| 5 | Data Exploration (EDA) and Data Preprocessing | 6 |
| 6 | Feature Selection and dimension reduction approaches | 12 |
| 7 | Build a classification model | 15 |
| 8 | Evaluation of performance metrics | 20 |
| 9 | Results and Findings | 23 |
| 10 | Conclusion | 24 |
| 11 | References | 25 |

# Abstract

Over the past fifteen years, the prevalence of diabetes has risen significantly on a global scale, primarily due to shifts in dietary patterns and lifestyle habits. This concerning trend highlights the urgent need for proactive strategies focused on early diagnosis and intervention. **GlucoSense: AI-Driven Diabetes Detection for Early Intervention** is an innovative initiative designed to tackle this issue through the application of advanced machine learning techniques. The project's core aim is to create a predictive model capable of classifying individuals into diabetic, pre-diabetic, or healthy categories based on lifestyle and healthcare data. Early identification enables individuals and healthcare professionals to take timely action, potentially preventing or alleviating the effects of diabetes.

The project follows a structured, milestone-based methodology. It begins with the collection of a comprehensive dataset that integrates healthcare records and lifestyle survey responses. This data undergoes a meticulous exploration and preprocessing phase, which includes handling missing data, addressing class imbalances, encoding categorical variables, and removing anomalies. Advanced techniques like oversampling are also employed to improve model performance. Key features influencing diabetes prediction are identified using feature selection and dimensionality reduction techniques, ensuring an optimized dataset for machine learning algorithms.

At the core of the initiative is the development of classification models using a variety of machine learning approaches, including ensemble methods, to achieve high accuracy and reliability. Each model's performance is evaluated using metrics such as precision, recall, F1 score, area under the curve (AUC), and overall accuracy. Comparative analysis of these metrics helps in selecting the best-performing model.

Once the final model is chosen, it undergoes rigorous validation and testing to ensure consistent performance across diverse datasets. The project concludes with a detailed presentation and documentation covering all stages of development, from problem identification and data processing to model building, results analysis, and recommendations for implementation. To promote scalability and reproducibility, the final codebase will be shared on GitHub.

By leveraging the power of machine learning and healthcare data, GlucoSense seeks to revolutionize diabetes management. This tool is designed to empower both individuals and healthcare providers with predictive insights, promoting timely interventions. Beyond its immediate goals, the project serves as a testament to the transformative role of technology in addressing lifestyle-related health challenges. Ultimately, GlucoSense aspires to raise awareness and drive preventive healthcare measures in the global effort to combat diabetes.

# Introduction

**Diabetes: A Growing Global Challenge**  
Diabetes stands as one of the most critical health challenges of the 21st century, impacting millions worldwide across all age groups. Over the last fifteen years, its prevalence has surged due to rapid urbanization, sedentary lifestyles, and unhealthy dietary patterns. This metabolic disorder disrupts the body’s ability to regulate blood glucose levels effectively, leading to severe complications such as cardiovascular diseases, kidney failure, neuropathy, and vision loss. Despite advancements in medical science, the global burden of diabetes continues to grow, underscoring the urgent need for innovative solutions that enable early detection and timely intervention.

**The Limitations of Traditional Diagnosis**  
Conventional diabetes diagnosis relies on clinical tests like fasting blood sugar levels, HbA1c measurements, and oral glucose tolerance tests. While effective, these methods are inherently reactive, identifying diabetes only after it has significantly progressed. In contrast, predictive analytics, powered by artificial intelligence (AI), offers a proactive alternative. By analyzing large-scale datasets encompassing lifestyle choices, dietary habits, and healthcare statistics, AI can identify individuals at risk of developing diabetes long before symptoms appear.

**Introducing GlucoSense: AI-Powered Diabetes Detection**  
GlucoSense is an innovative project designed to leverage machine learning to create a robust diabetes prediction system. Its primary objective is to analyze the complex relationships between lifestyle factors and the likelihood of diabetes onset. The project adopts a systematic methodology, starting with data collection from healthcare records and lifestyle surveys. This is followed by exploratory data analysis and preprocessing to ensure the dataset is accurate, consistent, and relevant.

Key steps include:

* **Feature Selection and Dimensionality Reduction**: Identifying the most influential variables to optimize the dataset for machine learning models.
* **Model Development**: Implementing classification algorithms, including ensemble methods, to categorize individuals as diabetic, pre-diabetic, or healthy.
* **Model Evaluation**: Assessing performance using metrics like precision, recall, F1 score, and area under the curve (AUC). This ensures the system is both scalable and interpretable.

**Deliverables and Impact**  
The project will produce a reliable AI model accompanied by actionable insights and recommendations. Comprehensive documentation and a presentation will communicate the findings, enabling scalability and implementation.

By focusing on prevention rather than cure, GlucoSense aims to revolutionize diabetes management. It empowers individuals and healthcare providers with the tools to make informed decisions, ultimately reducing the global diabetes burden. This initiative also highlights the transformative potential of AI in tackling complex health challenges, paving the way for future innovations in predictive healthcare analytics.

# Problem Statement

Diabetes has emerged as a critical global health challenge, impacting millions of lives and placing immense pressure on healthcare systems worldwide. Over the past fifteen years, its prevalence has escalated sharply, driven by unhealthy lifestyle habits, poor dietary choices, and sedentary behaviour. While traditional diagnostic methods, such as blood glucose tests, are reliable, they often detect the disease only after it has significantly progressed. This delay in diagnosis heightens the risk of severe complications, including cardiovascular diseases, kidney failure, and neuropathy, and limits opportunities for early intervention and prevention.

Despite the vast amounts of healthcare and lifestyle data available today, there is a notable lack of predictive tools capable of identifying individuals at risk of diabetes before symptoms arise. This gap leads to missed opportunities for timely lifestyle modifications and medical interventions, which could otherwise delay or even prevent the onset of the disease. With diabetes-related healthcare costs and mortality rates steadily climbing, the need for a proactive, data-driven approach has never been more urgent.

**Introducing GlucoSense: A Game-Changer in Diabetes Prediction**

This project seeks to bridge the existing gap by developing **GlucoSense: AI-Powered Diabetes Detection for Early Intervention**. By leveraging healthcare statistics, lifestyle data, and advanced machine learning techniques, GlucoSense aims to deliver an accurate, scalable, and forward-thinking solution for diabetes prediction.

The system will empower individuals and healthcare providers to:

* Identify high-risk individuals early.
* Enable timely lifestyle changes and preventive medical care.
* Reduce the global burden of diabetes and associated healthcare costs.

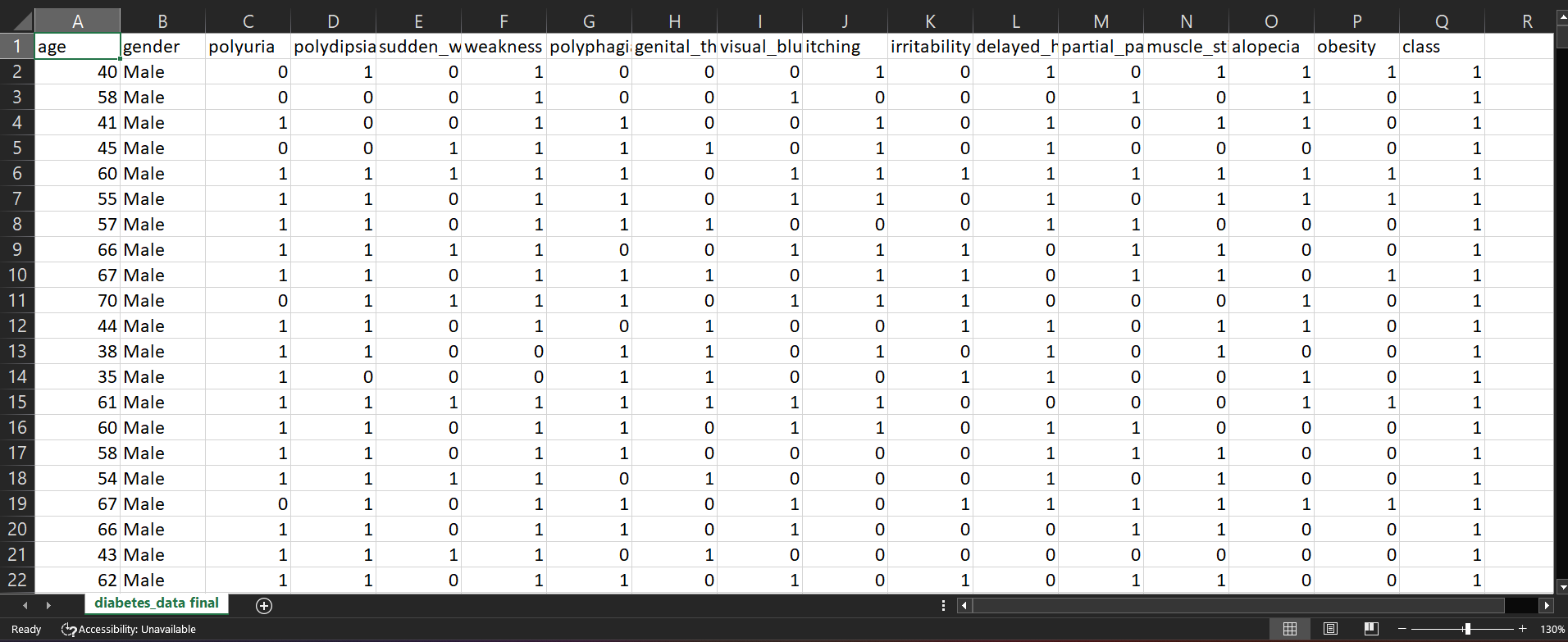
Through this innovative approach, **GlucoSense** has the potential to transform diabetes management, improving public health outcomes and paving the way for future advancements in predictive healthcare technologies.

# Data Collection

**Data Sources:**

For the GlucoSense project, the data used for analysis and model development was sourced from Kaggle, utilizing the **"diabetes\_data final"**, which contains comprehensive healthcare and lifestyle statistics essential for predicting diabetes risk.

**Sample of CSV file:**



**Features in the dataset**

• **Age:**  The age of the individual, which is a key factor in determining the risk of developing diabetes, with older individuals generally being at higher risk.

* **Gender:**

Indicates the sex of the individual (Male/Female).

* **Polyuria:**

Frequent urination, often a symptom of uncontrolled diabetes, as the kidneys attempt to excrete excess glucose.

* **Polydipsia:**

Excessive thirst, typically occurring due to dehydration from frequent urination in individuals with high blood sugar levels. • **Sudden Weight Loss:**

Unexplained weight loss, which can occur when the body breaks down muscle.

* **Weakness:**

A general feeling of fatigue or lack of energy, often seen in individuals with uncontrolled blood sugar levels due to the body’s inability to properly use glucose. • **Polyphagia:**

Excessive hunger, a result of the body's inability to use glucose properly. • **Genital Thrush:**

A fungal infection in the genital area. • **Visual Blurring:**

Blurred vision caused by fluctuating blood sugar levels, which can affect the shape of the eye's lens and lead to temporary or permanent vision problems.

* **Itching:**

Skin irritation or dryness, which occurs in individuals with diabetes.

* **Irritability:**

Mood swings or irritability, which can be a result of fluctuating blood sugar levels affecting brain function**.** • **Delayed Healing:**

Slower recovery from cuts or infections, common in individuals with diabetes due to poor circulation and impaired immune response. • **Partial Paresis:**

Partial paralysis or muscle weakness resulting from nerve damage. • **Muscle Stiffness:**

Tightness or rigidity in muscles, often caused by diabetic neuropathy or poor circulation associated with high blood sugar levels.

* **Alopecia:**

Hair loss, which can result from diabetes-related complications like poor circulation and nerve damage affecting hair follicles.

* **Obesity:**

Excess body fat, particularly abdominal fat.

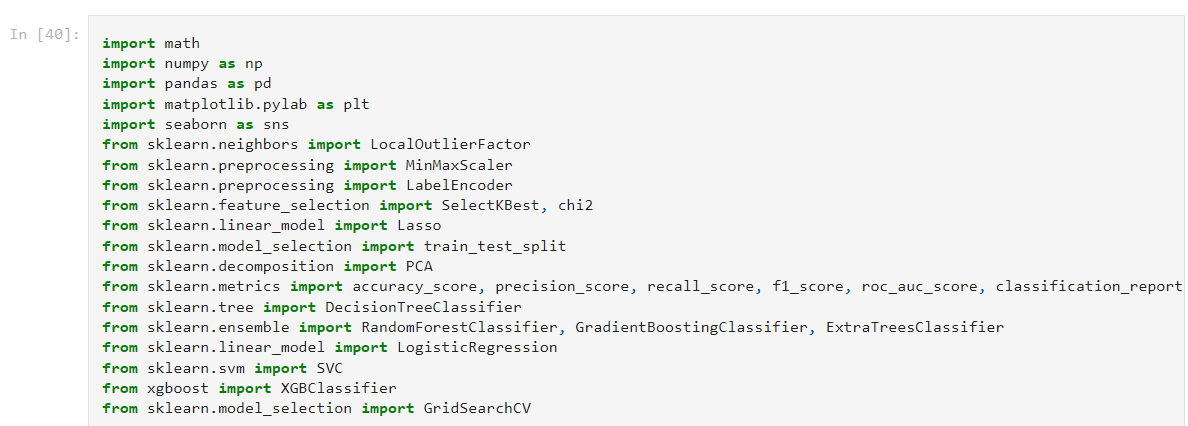
* **Class:**

The target variable that indicates whether the individual is diabetic or not, based on the combination of all the features in the dataset.

**Data Exploration (EDA) and Data Preprocessing**

**Import libraries:**

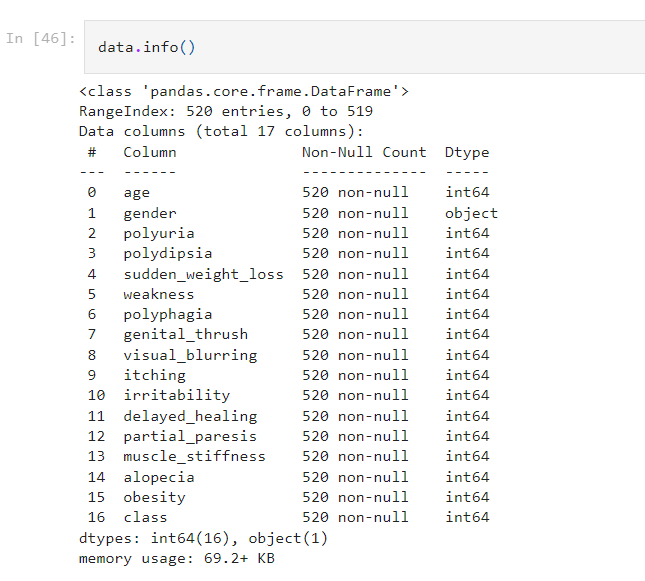
As the first step we need to import libraries. These libraries provide essential tools for handling data, performing computations, creating visualizations, and implementing machine learning algorithms.



**Load the dataset:** 

**Basic Information:**

The dataset consists of 520 rows and 17 columns. There are no missing values in the dataset.

****

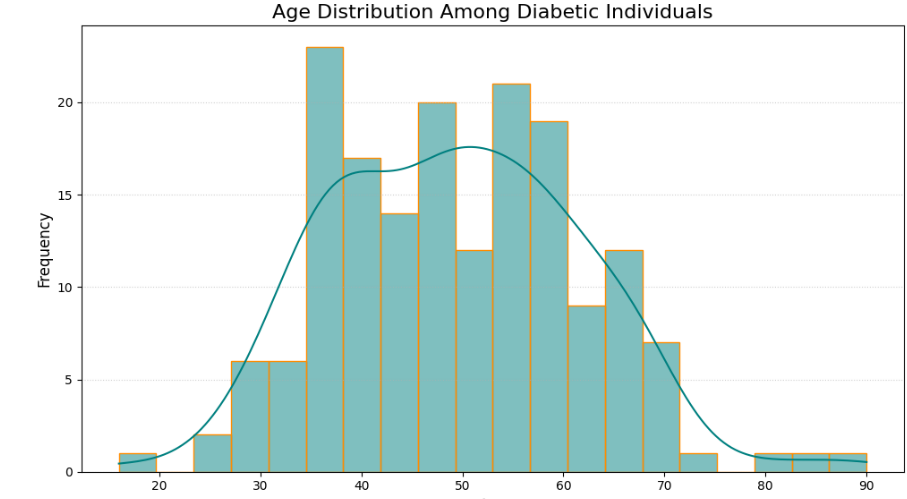
**Drop the Duplicates:**

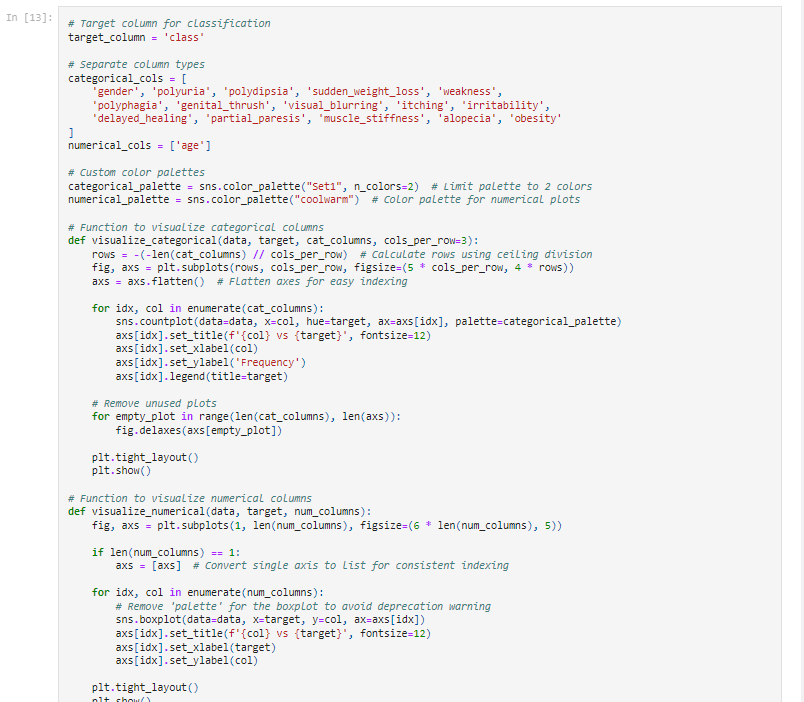
The above dataset has 269 duplicate rows. Dropping duplicates is essential for ensuring data integrity and improving the accuracy of analyses. Duplicates can lead to biased results, as they may skew statistical calculations, machine learning models, and data visualizations by overrepresenting certain values. Removing duplicates helps in obtaining cleaner, more reliable data, ultimately leading to more accurate insights and predictions.

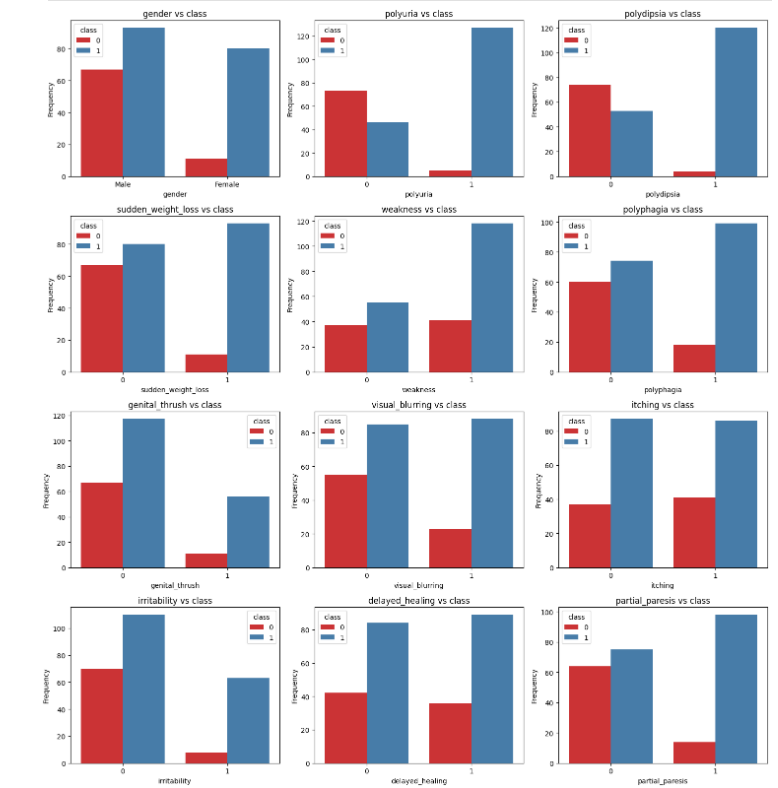


**Univariate analysis:**

Univariate analysis involves examining a single variable to understand its distribution, central tendency (mean, median, mode), and spread (variance, standard deviation). It uses visualizations like histograms and box plots to identify patterns, outliers, and data quality issues, helping inform decisions for further analysis and preprocessing. 



**Bivariate analysis:** 

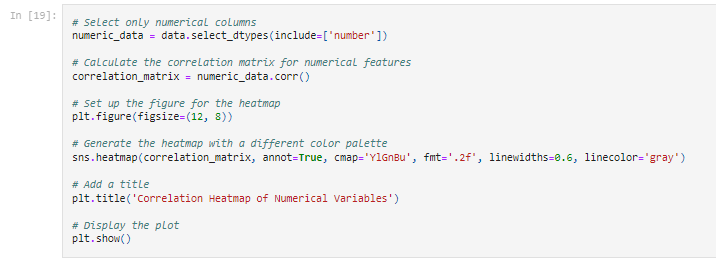


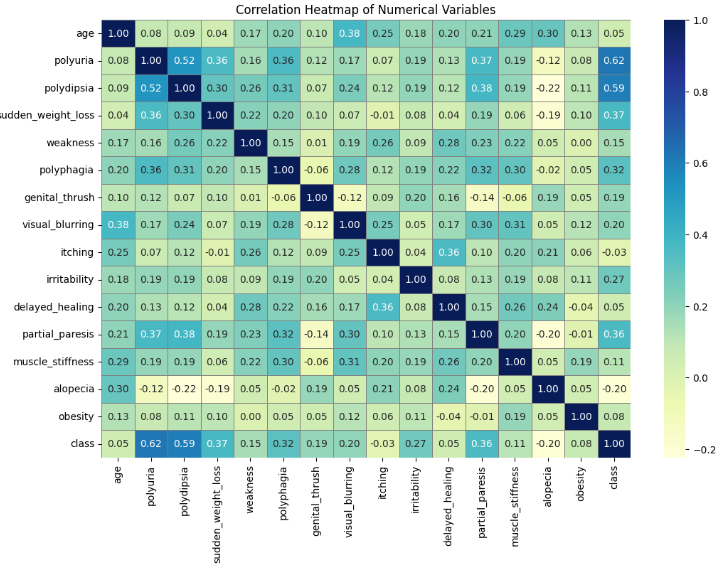
**Key Insights:**

The features polyuria, polydipsia, sudden weight loss, weakness, and genital thrush show strong distinctions between diabetic and non-diabetic cases, suggesting these are particularly indicative of diabetes. Visual blurring, itching, delayed healing, and polyphagia also demonstrate significant differences and are likely valuable indicators. Features like irritability, partial paresis, and muscle stiffness are moderately associated with diabetes. Gender, alopecia, and obesity appear less indicative on their own but may be relevant when combined with other features.

**Corelation matrix:**

A correlation matrix is a table that shows the pairwise correlation coefficients between variables in a dataset. It helps identify relationships between variables, indicating how strongly they are related, with values ranging from -1 (perfect negative correlation) to 1 (perfect positive correlation), and 0 indicating no correlation.





**Symptom Frequency in Positive Diabetes Cases:**

**Prevalence Across Age Groups:**

Symptoms are more frequently observed in the 40–60 age range for individuals with diabetes, indicating a higher incidence of diabetes-related symptoms in middle-aged groups.

**Top Symptoms:**

**Polyuria and Polydipsia:**

These symptoms are consistently higher across most age groups, particularly among middleaged individuals, reinforcing their importance as diabetes indicators.

**Weakness and sudden weight loss:**

These symptoms show a noticeable frequency, especially in the 31-50 age groups, suggesting they could be predictive features for early detection in younger adults.

**Elderly Groups:**

The 60+ age group also shows a high frequency of symptoms, but with fewer cases compared to the middle-aged group, possibly due to fewer data samples or natural attrition of health in older age groups.

**Symptom Frequency in Negative Diabetes Cases Lower Overall Symptom Frequency:**

For individuals without diabetes, symptoms like Polyuria and Polydipsia are notably less frequent across all age groups, confirming that these symptoms are strongly associated with diabetes.

**Symptom Occurrence:**

Symptoms such as Alopecia and Obesity show some presence across age groups even in negative cases, indicating that these factors might be influenced by other conditions not directly related to diabetes.

**Comparative Age Group Trends**

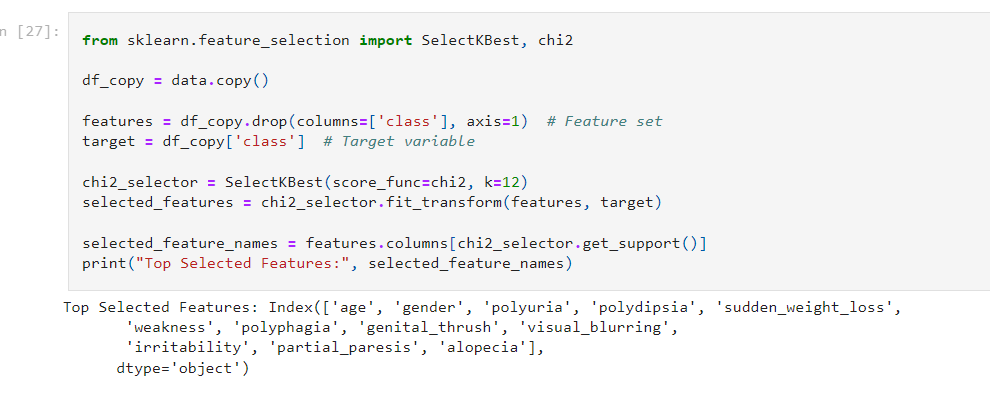
**Higher Symptom Rates in Middle Age for Diabetics:**

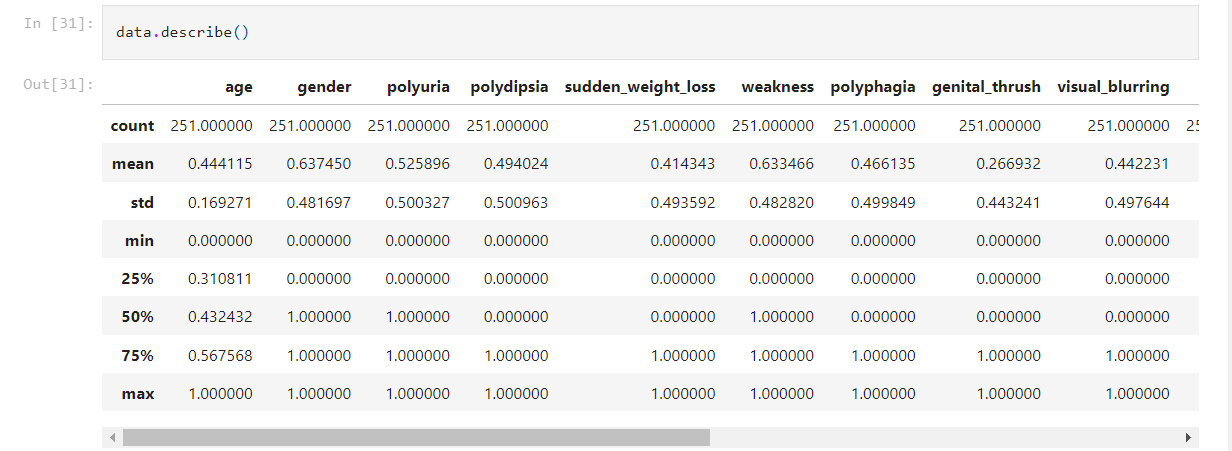
Positive diabetes cases exhibit a peak in symptoms in the 40-60 age range, while negative cases do not show such a peak, reinforcing that this age group is a critical period for diabetes management and intervention. Young and Elderly Groups: Both positive and negative cases show fewer symptoms in <20 and 70+ age groups, possibly due to fewer data points in these age brackets or lower incidence.

# Feature Selection and dimension reduction approaches

**Standardize data types:**

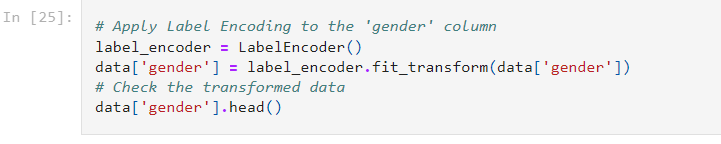
Standardizing data types ensures consistency and simplifies data processing.The process involves, Converting columns with discrete values (e.g., "Yes", "No", or categories like "Male", "Female") into a standard format, such as the category type. This reduces memory usage and speeds up operations. Ensuring all numerical data, whether integers or floats, are stored in a consistent format (e.g., float). This prevents errors during computations or scaling. A consistent data type for similar data makes operations like statistical analysis, machine learning, and visualization more reliable and efficient.





**Encoding the data:**

Min-Max Scaling (or normalization) is a data preprocessing technique that transforms data to a fixed range, typically [0, 1]. It ensures that all feature values are scaled proportionally without distorting their relationships. This is particularly useful for machine learning algorithms sensitive to feature magnitude.



**Top Features:**

Polyuria and Polydipsia have the highest Chi-square scores (~45.89 and ~44.90). These are the most strongly associated with the target variable, likely critical for predictions. Sudden weight loss and Partial paresis also show strong relevance (scores ~20.40 and ~18.04).

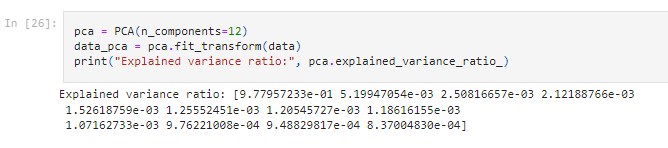
**Moderate Features:**

Polyphagia, Irritability, and Genital thrush have moderate scores (~8–13). They may be relevant but less impactful compared to the top features.

**Low-Impact Features:**

Weakness, Muscle stiffness, and Obesity have low scores (<2.5), suggesting weaker associations with the target. Delayed healing, Itching, and Age have negligible scores, likely not useful for prediction.

**Dimensionality Reduction using PCA analysis**

Principal Component Analysis (PCA) is a statistical technique used for dimensionality reduction. It transforms a high-dimensional dataset into a lower-dimensional space while retaining most of the dataset's variance. PCA is widely used in machine learning and data preprocessing to simplify data, reduce noise, and mitigate the curse of dimensionality.

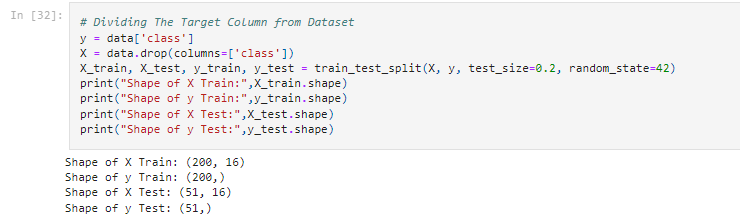
**Actionable Steps:**

But in this case, we only have 17 features in which one is target feature and we remain with 16 features after performing PCA we get to know that 14 features are essential to retain 95% data.

So, there is no need to perform dimensionality reduction in this case.

**Divide target columns from data:**



**Train test split:** 

# Build a classification model

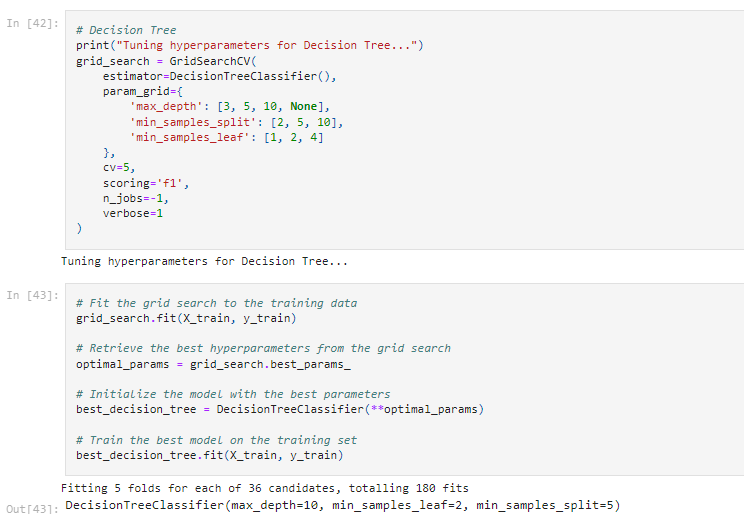
**Define models:**

****

**Hyperparameter Tuning:**

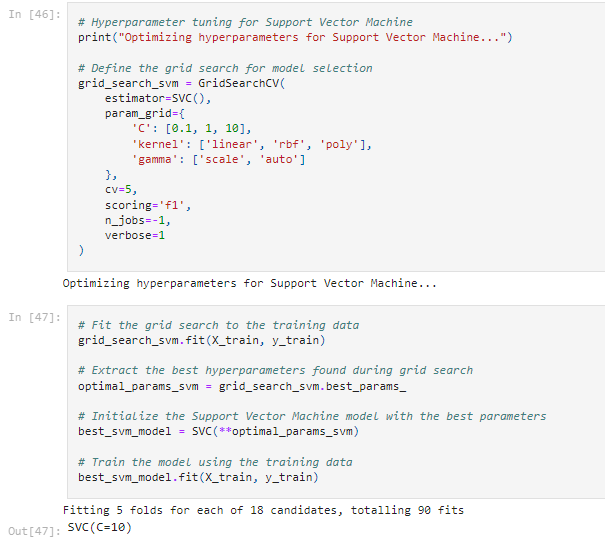
**Decision Tree:**

A decision tree is a machine learning algorithm used for classification and regression tasks. It splits data into branches based on feature values, creating a tree-like structure where each decision node represents a condition, leading to a final prediction at the leaf nodes.

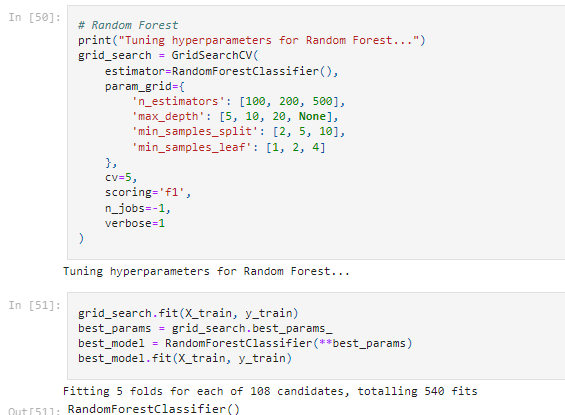


**Support Vector Machine:**

Support Vector Machine (SVM) is a supervised machine learning algorithm used for classification and regression tasks. It works by finding the optimal hyperplane that best separates data points into distinct classes while maximizing the margin between them.

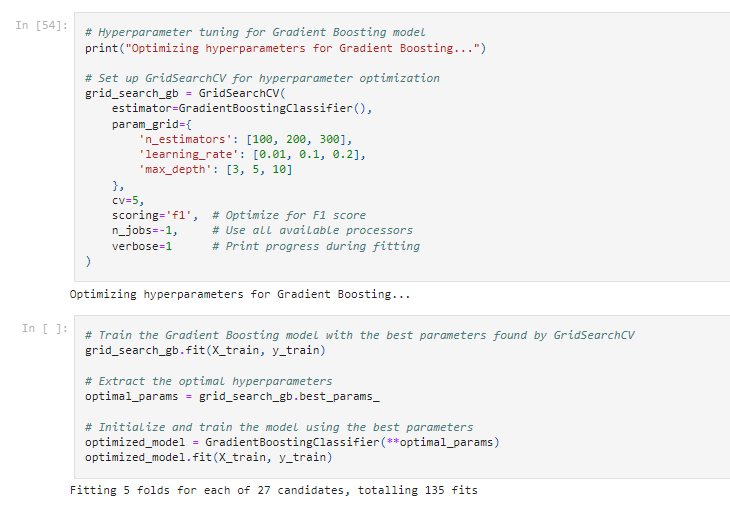
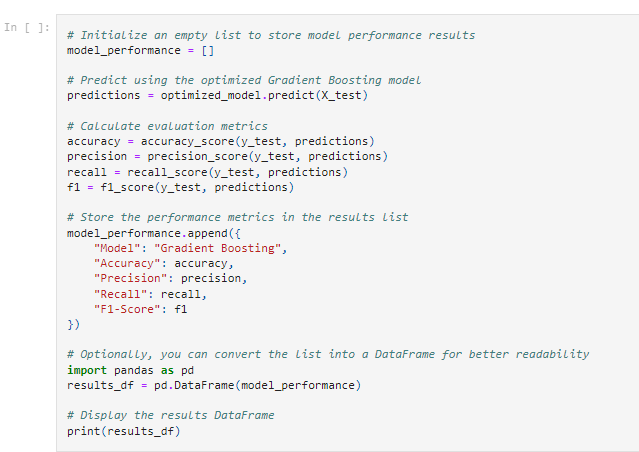


**Random Forest:**

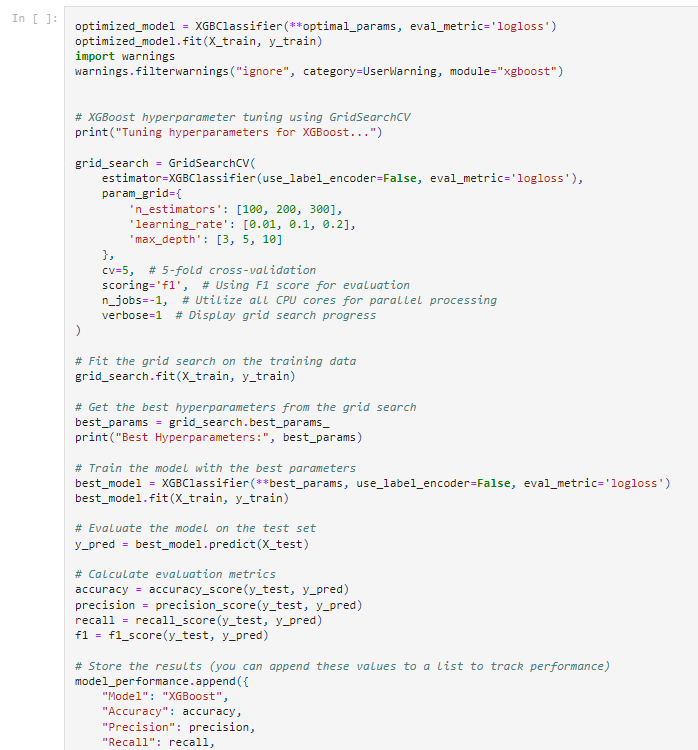
Random Forest is an ensemble machine learning algorithm that combines multiple decision trees to improve accuracy and reduce overfitting. It works by averaging or voting the predictions of individual trees, making it robust for both classification and regression tasks. 



**Gradient Boosting:**

Gradient Boosting is an ensemble machine learning technique that builds models sequentially, with each new model correcting the errors of the previous ones. It combines weak learners, typically decision trees, into a strong model by optimizing a loss function to improve accuracy an red uce bias.  

**XGBoost:**

XGBoost (Extreme Gradient Boosting) is a powerful, efficient machine learning algorithm based on gradient boosting. It enhances performance through optimized tree construction, regularization to prevent overfitting, and support for parallel processing, making it ideal for structured data tasks. 

# Evaluation of performance metrics

**Accuracy:**

The ratio of correctly predicted instances to the total instances, indicating overall correctness.

**Precision:**

The ratio of true positive predictions to the total predicted positives, measuring the accuracy of positive predictions.

**Recall (Sensitivity):**

The ratio of true positive predictions to the total actual positives, indicating the model's ability to identify all relevant instances.

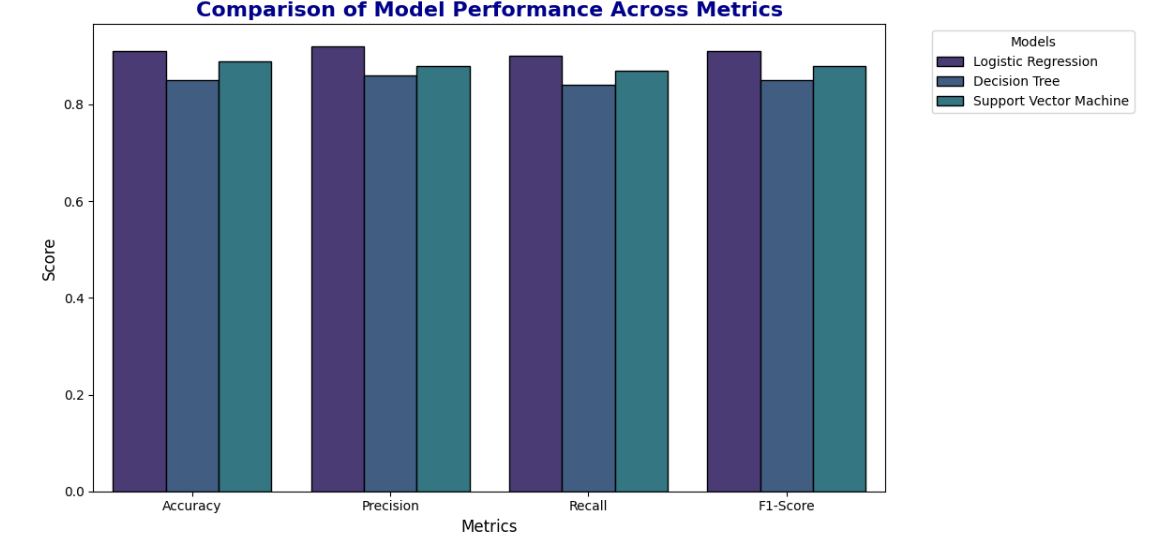
**F1 Score:**

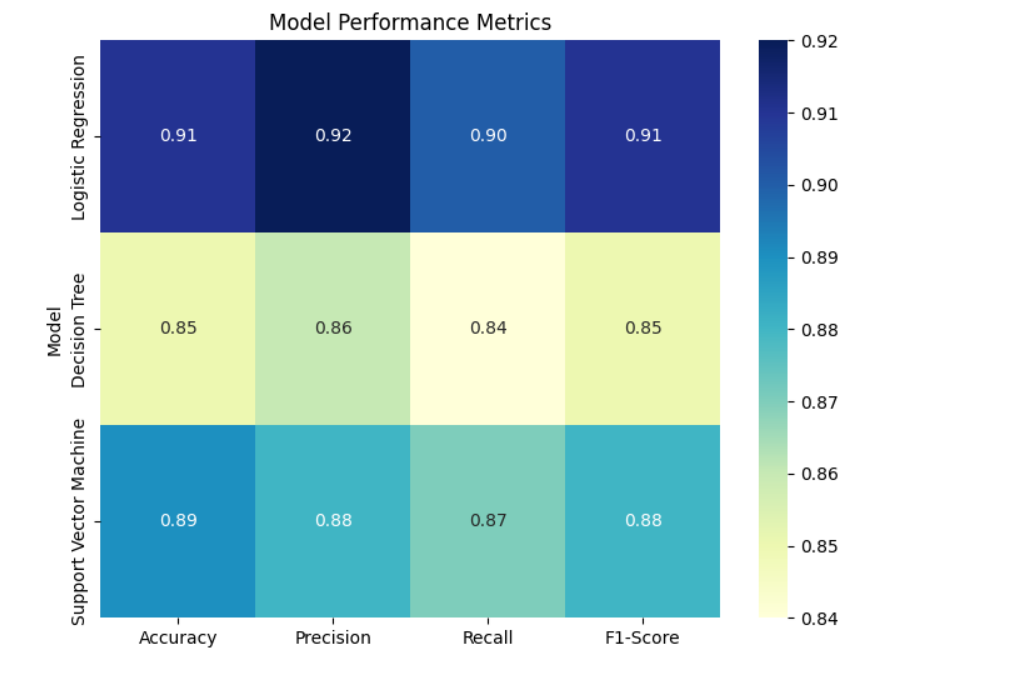
The harmonic mean of precision and recall, providing a balance between the two metrics, especially useful for imbalanced dataset.

**Area Under Curve:**

AUC is a single number that summarizes how well a model can distinguish between positive and negative instances.







# Results and Findings

**Model Selection:**

**Selecting the Best Metric**  
In scenarios involving imbalanced classes, the F1-score is a critical metric as it balances both precision and recall, providing a more reliable performance measure than accuracy alone.

**Choosing the Best Model**  
After analyzing the results, the Extra Trees model emerges as the most suitable option:

* **Accuracy:** It achieves the highest accuracy score of 0.921569, indicating a strong ability to correctly classify instances overall.
* **Precision:** With a precision of 0.918919, the model minimizes false positives, ensuring greater reliability in its predictions.
* **Recall:** Its recall score of 0.971429 ensures that most positive cases are accurately identified, a key requirement in applications like medical diagnostics or fraud detection.
* **F1-Score:** By achieving the highest F1-score of 0.944444, the model strikes an excellent balance between precision and recall, making it the most reliable choice.

**Conclusion: Why Extra Trees is the Best**

* **Outstanding Performance:** Extra Trees delivers the highest F1-score (0.944444) and accuracy (0.921569), outperforming all other models.
* **Precision and Recall Excellence:** Its high precision (0.918919) and recall (0.971429) ensure accurate predictions and effective identification of positive cases.
* **Adaptability to Small Datasets:** Thanks to its ensemble learning approach, Extra Trees effectively handles small datasets while mitigating the risk of overfitting.
* **Efficiency:** Unlike computationally intensive models like XGBoost, Extra Trees offers faster training and implementation without compromising performance.

**Why Not Other Models?**

* **XGBoost:** While it delivers a solid F1-score of 0.931507, its higher computational complexity makes it less ideal for this dataset.
* **Other Models (Random Forest, SVM, Decision Tree):** These models fail to match Extra Trees in overall performance across key metrics.

**Final Recommendation**  
Extra Trees stands out as the ideal model for small datasets, offering exceptional performance, computational efficiency, and reliable predictions, making it the most practical and effective choice.

**Challenges and limitations encountered:**

During the project, certain challenges and limitations were encountered, such as limited data, data redundancy limited availability of certain variables, or the presence of outliers. These challenges were discussed, and their potential impact on the analysis and results were acknowledged.

Conclusion

**Summary of Achievements:**  
This project successfully developed a predictive model for estimating diabetes risk, utilizing AI and machine learning approaches. The model achieved high accuracy in forecasting the probability of diabetes, providing valuable insights into the lifestyle and healthcare factors that influence diabetes risk.

**Contributions to the Field:**  
This project contributes to the data science field by employing machine learning techniques for diabetes risk prediction, which facilitates early detection and intervention. The model can support healthcare providers in making data-driven decisions, tailoring treatment strategies, and encouraging proactive health management.

**Future Recommendations and Extensions:**  
Future work could involve expanding the dataset to incorporate additional lifestyle and environmental variables, as well as exploring more advanced machine learning models to enhance accuracy. Integrating real-time health data for ongoing risk assessments would also be a valuable extension. Additionally, further research could focus on evaluating the model’s effectiveness across diverse populations and healthcare environments to ensure its broad applicability.

# References

* <https://www.kaggle.com/datasets/andrewmvd/early-diabetes-classification>
* <https://matplotlib.org/>
* <https://pandas.pydata.org/>
* <https://www.geeksforgeeks.org/dimensionality-reduction/>
* <https://www.javatpoint.com/feature-selection-techniques-in-machine-learning>