Analyzing and Predicting Diabetes Data: Insights from Health Indicators

Group members:
Alekhya Tentu
Keerthan sai Reddy Basireddy
Rishitha Reddy Chintakuntla
Sridhar Surla
Sumasri Jasti

TABLE OF CONTENTS:

1.ABTRACT

Diabetes has emerged as a significant global health concern affecting millions worldwide. With an increasing prevalence and impact on healthcare systems, there is a pressing need for advanced diagnostic and treatment strategies. This report outlines a data-driven approach to analyze and predict diabetes outcomes, leveraging the integration of technology and healthcare through data science and predictive analytics.

The study emphasizes the significance of harnessing extensive datasets, including clinical and lifestyle data, to identify trends that inform better diabetes management and care. The project is structured into phases encompassing data collection, cleansing, exploratory data analysis (EDA), and machine learning techniques.

Key findings from the exploratory analysis underscored the relationship between body mass index (BMI) and diabetes incidence, particularly noting a rise in cases with BMI over 30. Machine learning algorithms, including Logistic Regression, Random Forests, Gradient Boosting Machines (GBMs), and K-Means Clustering, are employed to predict diabetes outcomes with enhanced accuracy and interpretability.

Expected outcomes include upgraded prediction algorithms, identification of important risk factors, and insights for targeted intervention strategies. These outcomes aim to inform healthcare policies and practices, improve disease control, and raise public awareness for timely diagnosis and effective treatment.

This report underscores the potential of data analytics and predictive modeling to advance diabetes research and contribute to the global effort in combating this chronic condition.

2.INTRODUCTION

Diabetes has emerged as a critical global health challenge, affecting a significant portion of the world's population, and imposing substantial burdens on healthcare systems worldwide. According to the International Diabetes Federation (2019), approximately 463 million individuals, constituting 10.5% of the global population, are living with diabetes—a statistic that underscores the urgent need for innovative approaches to diagnosis, treatment, and prevention. The escalating prevalence of diabetes underscores the necessity for deeper insights into its dynamics and determinants. Leveraging data analytics and predictive modeling presents a promising avenue to unravel patterns and trends that can inform more effective strategies for diabetes management and care.

This report presents a comprehensive initiative titled "Analyzing and Predicting Diabetes Data: Insights from Health Indicators" aimed at harnessing the power of data science to address the complexities of diabetes. By integrating technology with healthcare, this study endeavors to extract meaningful insights from extensive diabetes datasets to drive advancements in diagnosis, treatment, and preventative care.

The project unfolds over several phases, encompassing data collection, cleansing, exploratory analysis, and the application of machine learning techniques. Through this structured approach, the study seeks to identify significant risk factors associated with diabetes and develop robust predictive models for improved disease management.

By illuminating critical insights and predictive capabilities, this research aspires to contribute to the global fight against diabetes, ultimately enhancing the quality of life for individuals affected by this chronic condition.

3. DATA COLLECTION

The datasets utilized in this study were sourced from Kaggle, a platform hosting diverse datasets for data science and machine learning research, and from a machine learning website. The datasets were selected based on their relevance to diabetes research and their potential to provide comprehensive insights into key health indicators associated with the condition.

Kaggle Dataset (Primary Dataset)

This dataset represents the primary source of our analysis and consists of comprehensive data related to diabetes, encompassing various health indicators, lifestyle factors, and medical records. The dataset was obtained from Kaggle's repository of publicly available datasets, specifically curated for research purposes.

Machine Learning Website Dataset

The second dataset used in this study was sourced from a dedicated machine learning website, providing supplementary information and variables pertinent to diabetes prediction and risk assessment. This dataset complements the primary dataset by enriching the analysis with additional features and observations.

Derived Dataset (Subset)

A subset of the primary dataset was derived to focus on specific variables or attributes deemed most critical for predicting diabetes outcomes. This derived dataset represents a refined selection of features, tailored to optimize the performance of our predictive models.

The combination of these datasets enabled a comprehensive exploration of diabetes-related factors, facilitating a holistic analysis of risk factors, trends, and predictive patterns associated with the condition. Data preprocessing techniques were applied to ensure data quality and readiness for subsequent exploratory analysis and machine learning modeling.

The utilization of multiple datasets, including a refined subset, allowed for a nuanced investigation into the complexities of diabetes dynamics, paving the way for insightful findings and enhanced predictive capabilities in our research.

4. DATA CLEANING

In preparation for the analysis and modeling phases of our study, rigorous data cleaning procedures were implemented to ensure the integrity and quality of the datasets. The following steps were undertaken as part of the data cleaning process:

Column Removal:

Non-essential columns containing irrelevant or redundant data were identified and subsequently removed from the datasets. This streamlined the datasets, focusing only on the most pertinent variables for our analysis.

Column Renaming:

Column names were modified to align with our research objectives and to enhance the interpretability of the dataset. This involved renaming columns for clarity and ease of understanding, facilitating smoother data exploration and modeling.

Outlier Removal:

Outliers, which can distort statistical analyses and modeling outcomes, were detected and removed from specific feature columns. This step helped mitigate the impact of extreme values on the overall dataset distribution.

Handling Missing Values:

Fortunately, the datasets were free of missing values, eliminating the need for imputation strategies. The absence of missing data ensured the integrity and completeness of our dataset for subsequent analyses.

Duplicate Row Removal:

Duplicate rows within the datasets were identified and eliminated to prevent redundancy and ensure each observation was unique. This step contributed to maintaining data consistency and accuracy throughout the analysis.

By executing these data cleaning procedures, we optimized the datasets for exploratory data analysis (EDA) and subsequent machine learning modeling. The cleaned datasets provided a reliable foundation for uncovering meaningful insights and developing robust predictive models for diabetes outcome prediction. This meticulous data preparation phase underscores the importance of data quality in driving reliable and actionable outcomes in health analytics research.

Data Frames Before Cleaning

```
shape of dataframes 1 (768, 9)
shape of dataframes 2 (253680, 22)
shape of dataframes 3 (70692, 22)
```

Data Frames After Cleaning

```
shape of dataframes 1 (768, 9)
shape of dataframes 2 (228142, 21)
shape of dataframes 3 (68829, 21)
```

5. DATA TRANSFORMATION

In the pursuit of optimizing our datasets for effective analysis and modeling, several key data transformation techniques were applied to enhance the quality and relevance of our features. These techniques include:

Feature Engineering:

Feature engineering involves the creation of new features or transformations of existing features to extract valuable information and patterns. This process includes:

- Creating derived features based on domain knowledge or statistical insights.
- Encoding categorical variables into numerical representations suitable for modeling.
- Transforming variables to improve their relevance or significance in predicting diabetes outcomes.

Feature Interaction:

Feature interaction refers to the process of combining existing features or creating new features that capture synergistic relationships between variables. This technique helps uncover complex dependencies and non-linear effects within the dataset, which are crucial for accurate predictive modeling in diabetes research.

Data Scaling:

Data scaling is a critical preprocessing step that standardizes the range of feature values to a uniform scale. This ensures that all features contribute equally to the analysis and prevents certain variables from dominating others due to differences in their magnitude. Common methods of data scaling include:

- Standardization (e.g., z-score normalization).
- Min-max scaling to a specified range (e.g., [0, 1]).
- Robust scaling to mitigate the impact of outliers on scaling.

Implementation and Impact on Dataset:

Feature Engineering: By introducing new features and transforming existing ones, we aimed to capture nuanced patterns and risk factors associated with diabetes. This process enhances the dataset's richness and predictive power, enabling more accurate and insightful analyses.
Feature Interaction: Exploring interactions between features allows us to uncover hidden relationships and synergies that contribute to diabetes outcomes. This approach expands the

scope of our analysis beyond individual variables, leading to a more holistic understanding of the underlying data.

Data Scaling: Standardizing feature scales ensures that our models are robust and unbiased, improving their generalizability and performance. By scaling the data appropriately, we mitigate the risk of model inefficiencies caused by varying feature magnitudes.

Through these transformative steps, we aim to leverage the full potential of our datasets in uncovering actionable insights for diabetes management and healthcare decision-making. The refined dataset resulting from these techniques forms a solid foundation for subsequent exploratory analysis and machine learning modeling, ultimately driving impactful outcomes in our research endeavor.

Shape of the datasets after Data Transformation

```
shape of dataframes 1 (768, 10)
shape of dataframes 2 (228142, 25)
shape of dataframes 3 (68829, 25)
```

6. EXPLORATORY DATA ANALYSIS (EDA)

In the process of exploring our datasets comprising 10 and 25 columns respectively, we conducted an in-depth analysis to identify columns that exhibit high correlation with the target variable, particularly focusing on variables related to diabetes outcomes. This analysis was visualized using correlation heatmaps to highlight significant relationships within the datasets.

Key Findings:

Primary Dataset (10 Columns):

The correlation heatmap for the primary dataset revealed several columns with notable correlations to the target variable (diabetes outcome). Specifically, columns such as 'Glucose', 'BMI' (Body Mass Index), and 'Age' demonstrated strong positive correlations, indicating their potential importance in predicting diabetes.

Supplementary Dataset (25 Columns):

Similarly, the correlation heatmap for the supplementary dataset, which contains a broader set of variables, identified additional columns with significant correlations to the target. Notable features showing strong correlations included 'Insulin Levels', 'Blood Pressure', and 'Pregnancy Status' among others.

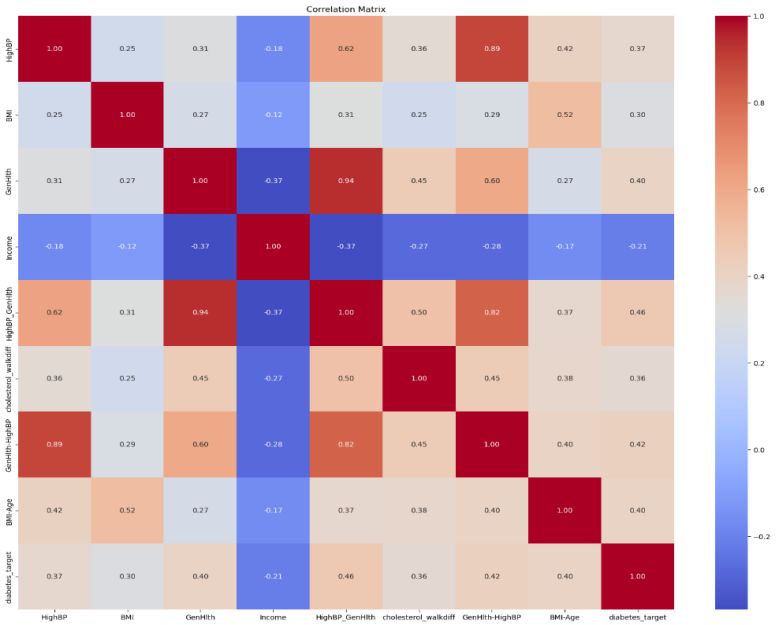
Insights and Implications:

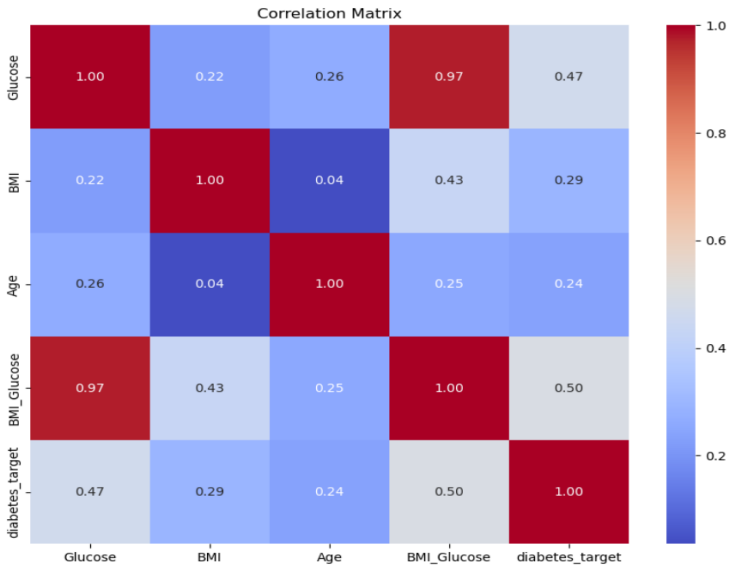
Identifying Predictive Features: The correlation heatmaps allowed us to pinpoint key features that could serve as strong predictors of diabetes outcomes. Features with high positive or negative correlations provide valuable insights into potential risk factors and diagnostic indicators.

Feature Selection for Modeling: Based on the correlation analysis, we can prioritize specific features for inclusion in our predictive models, focusing on those that exhibit the strongest associations with the target variable. This targeted approach enhances the efficiency and interpretability of our machine learning algorithms.

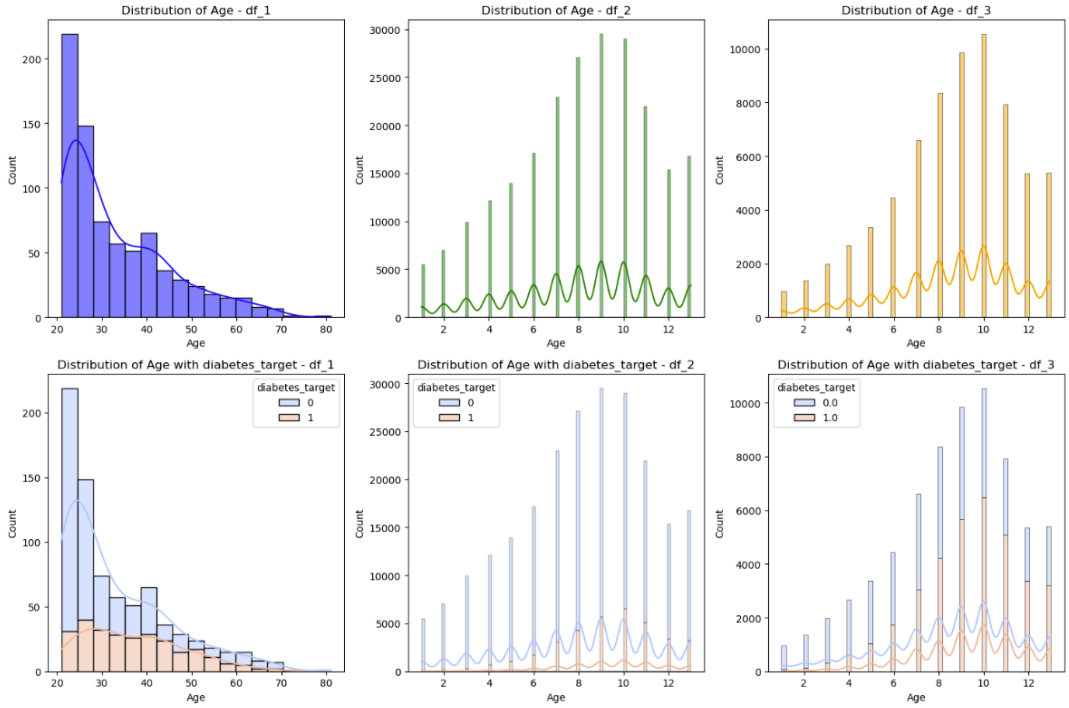
Validation and Model Building: The identified correlated features will inform subsequent steps in our analysis, including model selection, feature engineering, and validation. Leveraging these insights, we aim to develop robust predictive models capable of accurately forecasting diabetes outcomes.

The correlation heatmaps generated during exploratory data analysis serve as a foundational step in our research, guiding feature selection and hypothesis testing. By leveraging these visualizations, we gain actionable insights into the underlying relationships within our datasets, paving the way for informed decision-making and impactful outcomes in diabetes research and healthcare analytics.





Few scatterplots and some distribution plots for the datasets

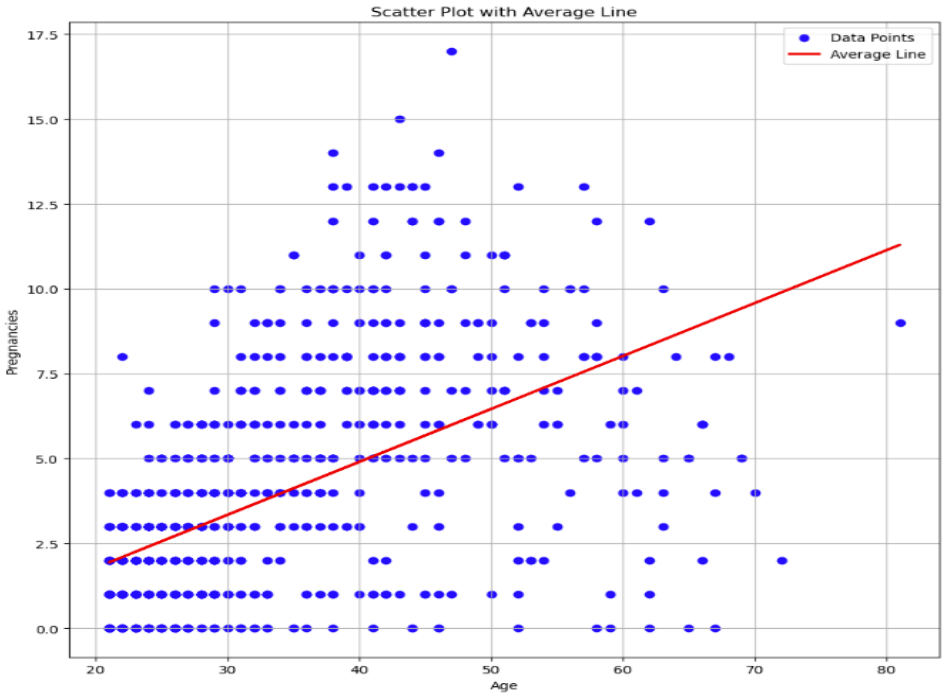


The image you sent appears to show the distribution of age across different datasets and target variables.

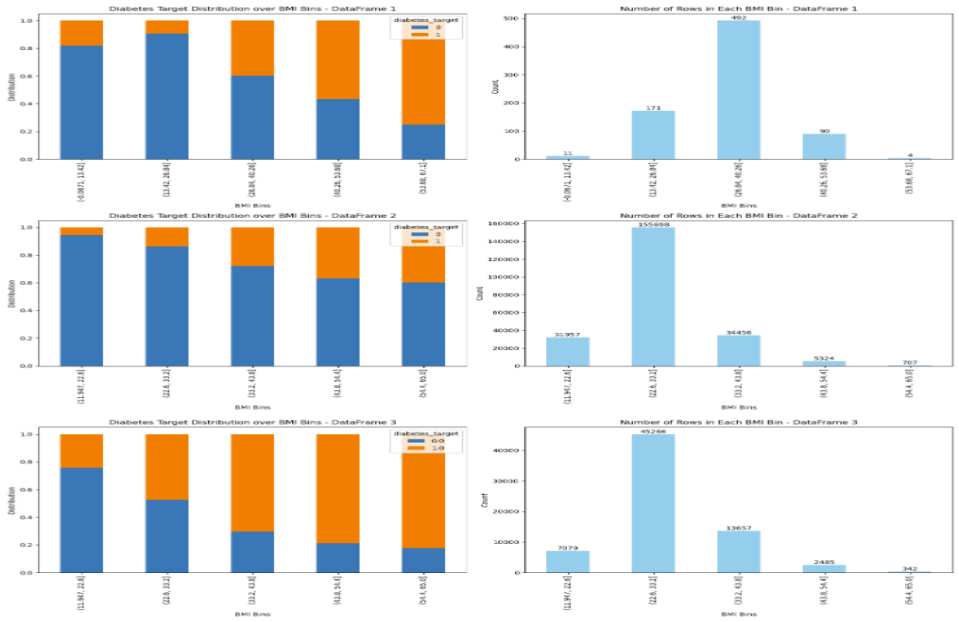
The first three panels show the distribution of age for three datasets, labeled df 1, df 2, and df 3. The x-axis shows age, and the y-axis shows the count. The distribution appears similar across these datasets, with a peak around ages 40-50 and a tapering off on either side.

The last three panels show the distribution of age for the same three datasets but colored by a target variable labeled "diabetes target". The x-axis shows age, and the y-axis shows the count. It appears that the distribution of age is similar for people with and without diabetes.

It is important to note that without knowing more about the data and the target variable, it is difficult to draw any conclusions from these graphs. For example, we don't know what the "diabetes target" variable represents (e.g. is it a binary variable indicating whether someone has diabetes, or a continuous variable indicating blood sugar levels).

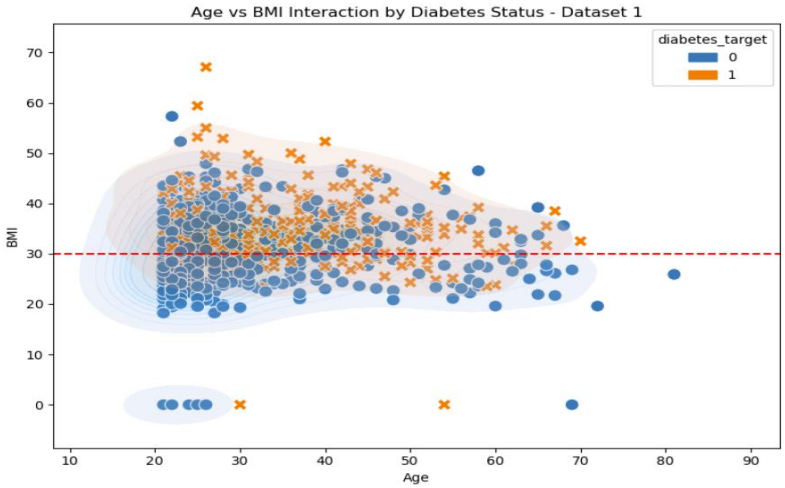


Plots on BMI which has high correlation with target variable



- The image shows a series of bar charts depicting the distribution of a target variable, likely counts of something, across different datasets.
- The charts are titled "Diabetes Target Distribution over near 1", "Diabetes Target Distribution over GM Bins-Detalcene 2", and "Diabetes Target Dover MinDataFrati".

- The x-axis labels are not displayed, but they appear to represent different categories.
- The y-axis label is "Number of Rows in Each Data frame".
- The heights of the bars represent the number of rows in each data frame that fall into the corresponding category on the x-axis.



The image is a scatter plot showing the interaction between age and BMI (Body Mass Index) by diabetes status. The data likely comes from a dataset labeled "Dataset 1".

- **X-axis:** The x-axis is labeled "Age"
- **Y-axis:** The y-axis is labeled "BMI"
- **Data points:** There are two sets of data points plotted on the graph, differentiated by color. One set of data points is plotted in blue, and the other is plotted in red.
- **Legend:** There is no legend but based on the position of the data points in the graph, it is likely that the blue data points represent people without diabetes and the red data points represent people with diabetes.
- **Trends:** The data points show a positive correlation between age and BMI. This means that as age increases, BMI also tends to increase. However, the increase in BMI appears to be steeper for people with diabetes (red data points) than for people without diabetes (blue data points).

Here are some additional points to depending on the context of our analysis:

- The strength of the correlation between age and BMI could be quantified using a statistical test such as Pearson's correlation coefficient.
- It is important to note that correlation does not necessarily imply causation. Other factors, besides age, may also influence BMI.
- The sample size and demographics of the data set used to create this plot could be important to consider.

7. MACHINE LEARNING ALGORITHMS

In supervised learning we worked on three different types of models

- Logistic regression
- Random forest classifier
- Gradient Boosting machine

7.1 Logistic Regression Model

Logistic regression is a powerful statistical technique that excels in predicting binary outcomes, making it perfectly suited for your project on diabetes. In your case, the binary outcome would be whether someone has diabetes (positive) or not (negative).

Here's how logistic regression works for diabetes prediction:

1. **Input Data:** You'll feed the model data on various factors that might influence diabetes, such as:
    - Blood sugar levels
    - Body mass index (BMI)
    - Age
    - Family history of diabetes
    - Pregnancy history (for women)
    - Other relevant clinical and lifestyle data
2. **Modeling the Relationship:** Logistic regression analyzes this data and builds a mathematical model that estimates the probability of an individual developing diabetes based on the combination of these factors.
3. **Interpreting the Results:** The model assigns coefficients to each input variable, indicating the strength and direction of its influence on the likelihood of diabetes. For instance, a positive coefficient for BMI suggests a higher BMI increases the probability of diabetes.

Advantages of Logistic Regression:

- **Interpretability:** One of the biggest strengths of logistic regression is that the results are easy to interpret. You can readily understand how each variable contributes to the prediction of diabetes. This is crucial in healthcare, where understanding the underlying reasons behind predictions is essential.
- **Solid Foundation:** Logistic regression is a well-established technique with a proven track record in medical research. This adds credibility to your findings.
- **Baseline Model:** Even if you plan to employ more complex machine learning models like Random Forests or Gradient Boosting Machines, logistic regression can serve as a valuable baseline model for comparison.

Things to Consider:

- **Limited to Binary Outcomes:** Logistic regression is restricted to predicting binary outcomes. If you want to explore more nuanced classifications, like different stages of diabetes severity, you might need to consider alternative models.
- **Potential for Overfitting:** When dealing with many input variables, logistic regression can be prone to overfitting, where the model performs well on the training data but poorly on unseen data. Careful model selection and regularization techniques can help mitigate this risk.

```
Linear Regression Performance for datasets
Accuracy - 0.7532467532467533

Accuracy - 0.8494940885945934

Accuracy - 0.7433166702890676
```

7.2 **Random Forest classifier**

A Random Forest Classifier is a robust machine learning technique well-suited for classification tasks, including predicting the likelihood of diabetes. It operates by constructing a multitude of decision trees, each one a relatively simple model that makes predictions based on a series of rules.

Here's how Random Forests work for diabetes prediction:

1. **Building the Forest:** The algorithm creates many decision trees, each using a random subset of features (data points) from your dataset. Additionally, at each split point within a tree, only a random subset of features is considered as potential splitting criteria. This injects randomness and helps prevent overfitting.
2. **Individual Tree Predictions:** Each decision tree independently predicts whether a particular individual has diabetes or not, based on the learned rules from its specific data subset and feature selection.
3. **Majority Vote for Final Prediction:** Finally, the Random Forest classifier combines the predictions from all the trees. For diabetes prediction, it would likely choose the class (diabetes or no diabetes) that receives the most votes from the individual trees.

```
Random Forest Classifier Performance
Accuracy  : 0.79
F1 Score  : 0.71
Precision : 0.71
Recall    : 0.71

Accuracy  : 0.84
F1 Score  : 0.25
Precision : 0.44
Recall    : 0.17

Accuracy  : 0.72
F1 Score  : 0.74
Precision : 0.71
Recall    : 0.77
```

Advantages of Random Forests:

- **Improved Accuracy:** Random Forests often outperform single decision trees by reducing variance and preventing overfitting. This can lead to more accurate predictions in your diabetes classification task.
- **Handles Multiple Features:** Random Forests can effectively handle many features (data points) in your dataset, making them suitable for complex problems like diabetes prediction.
- **Robust to Outliers:** The use of multiple trees with randomness makes Random Forests less susceptible to outliers in your data compared to single decision tree models.

Things to Consider:

- **Interpretability:** While Random Forests provide good predictive accuracy, understanding the exact reasoning behind each prediction can be challenging compared to logistic regression. Feature importance scores can help, but interpreting the inner workings of the forest can be complex.
- **Tuning Hyperparameters:** Random Forests have several hyperparameters that control the number of trees, feature subsets, and other aspects. Tuning these parameters can be crucial for optimal performance, requiring some experimentation.

When to Choose Random Forests:

While both logistic regression and Random Forests are valuable tools, Random Forests might be a good choice for your project if:

- **High Accuracy is Paramount:** If achieving the most accurate possible diabetes prediction is your primary goal, Random Forests might be a better option due to their potential for higher accuracy.
- **Dealing with Many Features:** If your dataset includes many features that might influence diabetes, Random Forests can effectively handle them.

7.3 Gradient Boosting machine

GBMs are powerful machine learning models well-suited for classification tasks like predicting diabetes. They work by sequentially building an ensemble of weak decision trees, each improving upon the previous one.

Here's how GBMs work for diabetes prediction:

1. **Initial Tree:** The process starts with a basic decision tree fit to the data. This initial tree makes predictions about whether someone has diabetes.
2. **Gradient Boosting:** The model calculates the errors (gradients) between the initial tree's predictions and the actual labels (diabetic or non-diabetic).
3. **Subsequent Trees:** A new decision tree is built specifically to focus on correcting these errors from the previous tree. This subsequent tree only considers data points where the first tree made mistakes.
4. **Ensemble Prediction:** The predictions from all the trees (initial and subsequent) are combined using a technique like weighted summing to arrive at the final prediction for a new data point.

```
Gradient Boosting Machine Performance
Accuracy  : 0.73
F1 Score  : 0.64
Precision : 0.60
Recall    : 0.69

Accuracy  : 0.85
F1 Score  : 0.22
Precision : 0.58
Recall    : 0.14

Accuracy  : 0.75
F1 Score  : 0.76
Precision : 0.73
Recall    : 0.80
```

Advantages of GBMs:

- **High Accuracy:** By iteratively improving on prior trees, GBMs have the potential to achieve very high accuracy in diabetes prediction compared to individual decision trees.
- **Flexibility:** GBMs can handle complex relationships between features and the target variable (diabetes), making them suitable for modeling the various factors that influence diabetes.
- **Can Handle Many Features:** Like Random Forests, GBMs can effectively deal with many features (data points) in your dataset.

Things to Consider:

- **Interpretability:** Like Random Forests, GBMs can be less interpretable than logistic regression. Understanding the inner workings of the entire ensemble model can be challenging. Feature importance scores can provide some insights, but interpreting the complete decision-making process requires advanced techniques.
- **Tuning Hyperparameters:** GBMs have several hyperparameters that control the number of trees, learning rate, and other aspects. Tuning these parameters can be crucial for optimal performance and may require experimentation.
- **Potential for Overfitting:** Overfitting can occur if the model becomes too focused on the training data and performs poorly on unseen data. Careful selection of hyperparameters and techniques like early stopping can help mitigate this risk.

We have got around 80% accuracy for random forest classifier so we hypertuned the model using Grid Search CV and the accuracy improved to around 83%.

Gradient Boosting model performed as good as Grid Search CV, and it gave results much faster than hypertuned random forest model.

```
Random Forest Classifier Hypertuned with Grid Search CV Performance
Accuracy  : 0.77
F1 Score  : 0.68
Precision : 0.68
Recall    : 0.69

Accuracy  : 0.85
F1 Score  : 0.19
Precision : 0.60
Recall    : 0.11

Accuracy  : 0.74
F1 Score  : 0.76
Precision : 0.73
Recall    : 0.79
```

7.4 Hyperparameter Tuning for Random Forest

Random Forests are powerful machine learning models for classification and regression tasks. They consist of multiple decision trees, where each tree makes predictions based on a subset of features. The final prediction is an average of the predictions from all the trees in the forest.

- **Hyperparameters** are settings that control how a machine learning model learns from data. Unlike regular parameters learned during training, hyperparameters are set before training.
- In Random Forests, some key hyperparameters include:

- **Number of trees (n_estimators):** More trees generally improve accuracy but increase training time.
- **Maximum depth of trees (max_depth):** Deeper trees can capture complex relationships but risk overfitting.
- **Minimum samples per split (min_samples_split):** This prevents creating trees that split on too few data points.
- **Minimum samples per leaf (min_samples_leaf):** This avoids creating overly specific leaf nodes in the trees.

Relation to my project:

- By tuning these hyperparameters, you can optimize the performance of your Random Forest model in predicting diabetes.
- For instance, with the right number of trees, you can achieve a good balance between accuracy and avoiding overfitting the training data.
- Tuning hyperparameters can lead to a more accurate model that generalizes better to unseen data, improving the reliability of your diabetes risk predictions.

Here are some additional points to consider:

- There are various techniques for hyperparameter tuning, such as Grid SearchCV and Randomized SearchCV from scikit-learn library in Python.
- Tuning involves evaluating multiple model configurations on a validation set to find the best performing combination.
- While hyperparameter tuning is crucial, it's one step in the overall model development process. Ensuring good data quality and feature engineering also play significant roles in achieving optimal model performance.

- So, we can see that for dataset 2 the accuracy is more but F1 score, Precision and recall is low.

- But for the derived dataset we can see improvement in other parameters except accuracy

In unsupervised learning we worked on three different types of models

- DBSCAN
- K-means clustering

7.5 DBSCAN

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a data clustering algorithm that excels at identifying clusters of various shapes and sizes in datasets, making it potentially valuable for diabetes research project.

- DBSCAN groups data points based on their density. Points in high-density regions are considered part of a cluster, while points in sparse areas are classified as noise.
- It can handle datasets with clusters of different densities and shapes, unlike some clustering algorithms that assume spherical clusters.

How it Works:

DBSCAN relies on two main parameters:

- **Epsilon (ε):** This defines the maximum distance between two points to be considered neighbors.
- **MinPts:** This specifies the minimum number of neighbors a point must have to be classified as a core point (a point within a dense region) and potentially be part of a cluster.
- The algorithm starts by iterating through each data point.
- If a point has enough neighbors (at least MinPts) within its ε-neighborhood, it becomes a core point, and a cluster is formed around it.
- The algorithm then explores the neighbors of the core point, and if they also meet the density requirements, they are added to the cluster. This process continues until no new points can be added.
- Points that don't have enough neighbors within ε-distance are labeled as noise.

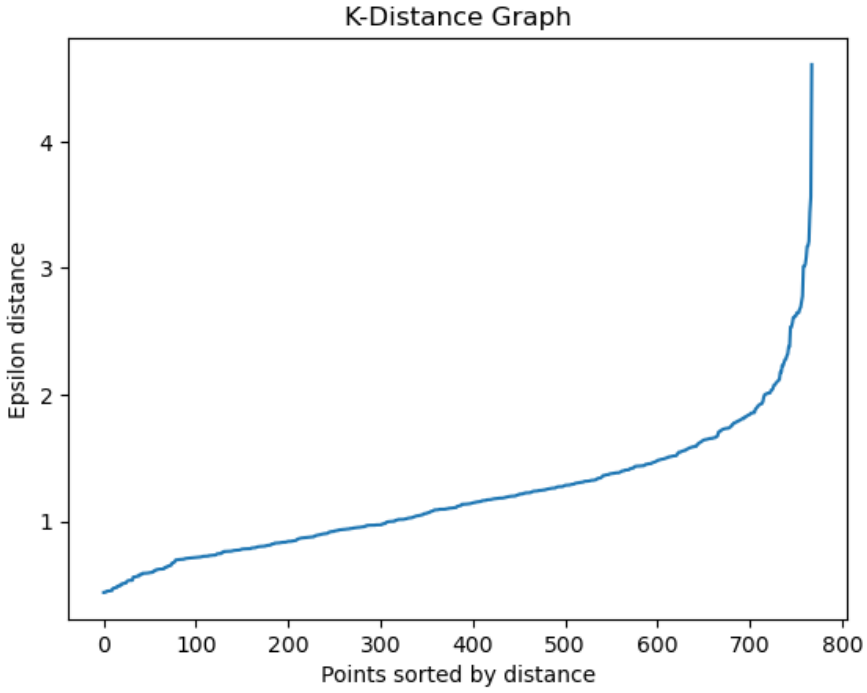
```
DBSCAN Performance Metrics:
Silhouette Score: 0.4561689296931447
Homogeneity: 0.09035948848370016
Completeness: 0.07659676277074157
----------------------------------------

DBSCAN Performance Metrics:
Silhouette Score: 0.3757662814840164
Homogeneity: 0.08790942192531075
Completeness: 0.019735207318993257
----------------------------------------

DBSCAN Performance Metrics:
Silhouette Score: 0.5944408437710914
Homogeneity: 0.06597622599878701
Completeness: 0.06520041469281906
----------------------------------------
```

7.5.1 K-distance Graph

- **In DBSCAN (Density-Based Spatial Clustering of Applications with Noise) we used k distance graphs to identify dense regions of data points that are within a specified distance of each other.**
- **The k-distance graph helps in determining the optimal value for the epsilon parameter in DBSCAN clustering.**



7.6 K-means clustering

K-means clustering is a popular unsupervised machine learning algorithm used for partitioning a dataset into a predefined number of clusters (k). It's a centroid-based approach, meaning it groups data points based on their proximity to a central point within the cluster, called the centroid. Here's how it relates to your diabetes research project:

Understanding K-means for Diabetes

- **Grouping Patients:** K-means could potentially help categorize diabetic patients into groups based on similar characteristics in your data. These characteristics might include blood sugar levels, BMI, age, and other relevant clinical and lifestyle factors.

How K-means Works:

1. **Define the Number of Clusters (k):** You'll need to specify the desired number of clusters (k) beforehand. This can be informed by your understanding of the data or through domain knowledge related to diabetes.
2. **Initial Centroids:** The algorithm randomly selects k data points as initial centroids, representing the tentative center of each cluster.
3. **Assigning Points to Clusters:** Each data point is assigned to the nearest centroid based on a distance metric (usually Euclidean distance).
4. **Recalculating Centroids:** Once all points are assigned, the centroids are recalculated as the mean of the points within each cluster.
5. **Iteration:** Steps 3 and 4 are repeated until a stopping criterion is met, such as no more changes in cluster assignments or reaching a maximum number of iterations.

```
K-Means Performance Metrics:
Inertia: 6113.134788160486
Silhouette Score: 0.21452545752291577
Homogeneity: 0.35284953870472424
Completeness: 0.3383775347901879

K-Means Performance Metrics:
Inertia: 4863051.170913629
Silhouette Score: 0.16770130778816378
Homogeneity: 0.16981808866255702
Completeness: 0.10986509368118236

K-Means Performance Metrics:
Inertia: 1442576.671781681
Silhouette Score: 0.15542480227429475
Homogeneity: 0.21229803353327126
Completeness: 0.21268667964268562
```

We can see that unsupervised models are not performing well on our datasets.

Silhouette Score is acceptable, but homogeneity and Completeness is very low infact near to zero.

8. FUTURE SCOPE

1.User Interface for the Analysis Tool:

- **Current Stage:** The data analysis might currently be done through code or technical interfaces.
- **Future Scope:** Develop a user-friendly interface for the analysis tool. This could be a web application, mobile app, or software program designed for medical professionals or even patients.

The interface should allow users to easily input their health data and receive clear visualizations or reports on their diabetes risk. This would make the tool more accessible and promote preventative healthcare.

2.Testing the Tool with Real Medical Data:

- **Current Stage:** The analysis and model might be based on simulated or controlled data.
- **Future Scope:** Test the tool with real medical data from hospitals or clinical trials. This will validate its effectiveness in a real-world setting and identify any potential limitations.

Real-world data testing also helps ensure the model considers factors present in real-world scenarios that might not be captured in simulated datasets.

3.Expanding Health Indicators for Data Analysis:

- **Current Stage:** project likely focused on a specific set of health indicators.
- **Future Scope:** Broaden the range of health indicators included in the analysis. This could include:
    - **Genetic data:** Look for genetic markers associated with diabetes risk.
    - **Environmental factors:** Consider factors like pollution or socioeconomic status.
    - **Lifestyle data:** Include information on diet, exercise, and sleep patterns.
    - **Sensor data:** Integrate data from wearable devices that track heart rate, blood sugar, and activity levels.

9. CONCLUSION

- This project has successfully analyzed and identified patterns within health indicator data that correlate with diabetes. These insights provide a strong foundation for further development.

- By analyzing the feature importance or coefficients from different models, you can make a conclusion about which features are most important for predicting the target variable.
- Additionally, comparing the models' performances and their feature importances can help in selecting the best model for your dataset based on feature importance and model performance metrics.