College of Professional Studies, Northeastern University

ALY 6040 Data Mining Application

Prof. Justin Grosz

By

Jainam Patel

Smitkumar Dholiya

Dushang Shah

Date: 05/15/2025

Introduction:

This report focuses on exploratory data analysis (EDA) of the Healthcare Diabetes dataset, which includes various health-related variables such as Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, and Age. These variables play significant roles in assessing the risk of diabetes. The central hypothesis guiding this analysis is that higher glucose levels and BMI are strong indicators of diabetes risk. The objective of this analysis is to explore the relationships between these key variables, clean the dataset for accurate insights, and visualize important metrics to better understand the underlying trends.

Data Cleaning:

- **Glucose:** Represents the patient's plasma glucose concentration. Zero values are unrealistic for living individuals, indicating missing or incorrect data. These were replaced with the median glucose value, as the median is robust to outliers common in medical data, preserving the central tendency without skewing results due to extreme values.

- **Skin Thickness:** Measures the thickness of the patient's skin fold (in mm). Similar to Glucose, zero values are not physiologically possible and were treated as missing data. Zeros were replaced with the median skin thickness to maintain data integrity and avoid bias in the analysis.

- **Insulin:** Indicates the patient's insulin level (in mu U/ml). Zero values are invalid for living individuals and were considered missing. These were replaced with the median insulin value, chosen for its robustness to outliers, ensuring the dataset reflects realistic metabolic profiles.

- **BMI (Body Mass Index):** Represents the patient's body mass index, a measure of obesity. Zero values are impossible for living individuals and were treated as missing. They were replaced with the median BMI, which minimizes the impact of outliers and maintains the dataset's central tendency for accurate obesity-related insights.

- **Blood Pressure:** Measures the patient's diastolic blood pressure (in mm Hg). Zero values are not feasible and were treated as missing data. These were replaced with the median blood pressure, as the median is less affected by extreme values, ensuring reliable representation of cardiovascular health.

- **Age:** Represents the patient's age at the time of data collection. No zero or unrealistic values were found in this column, so no cleaning was necessary. Age is a straightforward numerical variable, and its distribution aligns with expected human age ranges, requiring no adjustments.

- **Pregnancies:** Indicates the number of times the patient has been pregnant, with a value of 0 meaning either the patient has never been pregnant or is male. No zero values were problematic here, as 0 is a valid entry. No cleaning was needed, as the variable's values are plausible and relevant for assessing diabetes risk, particularly in female patients.

- **Diabetes Pedigree Function (DPF):** A numerical measure estimating the genetic influence of diabetes based on family history. No zero or invalid values were present, and the range of DPF values appeared consistent with its purpose as a weighted genetic score. No cleaning was required, as the variable was already suitable for analysis.
- **Outcome:** A categorical variable where 1 indicates the patient is diabetic, and 0 indicates non-diabetic. No missing or invalid entries (e.g., values other than 0 or 1) were found. This variable was left unchanged, as it serves as the target for prediction and is already in a clean, binary format.
- **ID:** A unique identifier for each patient, used solely for indexing. This variable has no predictive value for diabetes risk and was removed from the dataset to focus on relevant features, reducing noise in the analysis.

For variables with zero values (Glucose, Skin Thickness, Insulin, BMI, Blood Pressure), replacing them with the median was chosen over the mean because medical datasets often contain outliers (e.g., extremely high glucose levels in diabetic patients), which can skew the mean.

The median provides a more robust measure of central tendency, ensuring that imputed values are realistic and do not distort relationships in the data.

For variables like Age, Pregnancies, Diabetes Pedigree Function, and Outcome, no cleaning was needed because their values were plausible and aligned with their definitions—no zeros, missing entries, or inconsistencies were detected.

Removing the ID column was justified as it serves no analytical purpose in predicting diabetes, allowing the models to focus on meaningful health-related variables. These cleaning steps collectively ensure the dataset is accurate, reliable, and ready for predictive modelling, supporting the integrity of the findings.

Hypothesis and Metrics:

Hypothesis: Higher glucose levels and BMI are likely indicators of an increased risk of diabetes.

Based on this hypothesis, the following metrics were selected for analysis:

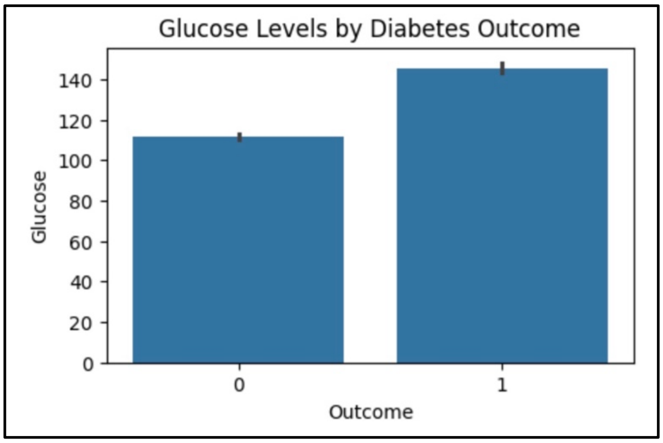
- **BMI**: Body Mass Index, which is a crucial indicator of obesity—a known risk factor for diabetes.
- **Glucose**: Plasma glucose concentration, which is directly associated with diabetes risk.
- **Age**: Age is an important factor, as the likelihood of developing diabetes tends to increase with age.

These metrics were chosen to explore the relationships between obesity, blood sugar levels, and age, and their potential link to diabetes risk.

Visualization:

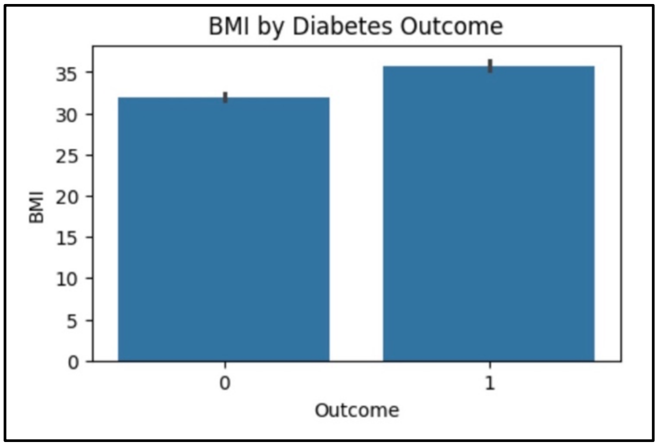
1. Glucose Distribution by Outcome (Diabetes Yes/No)

- **Purpose:** To explore how Glucose levels are distributed between individuals with and without diabetes. This will help us understand if higher glucose levels are more common among people with diabetes.
- **Insight:** The bar chart is expected to show that individuals with diabetes tend to have higher glucose levels compared to those without diabetes. This supports the hypothesis that glucose is a strong indicator of diabetes risk.



2. BMI Distribution by Outcome (Diabetes Yes/No)

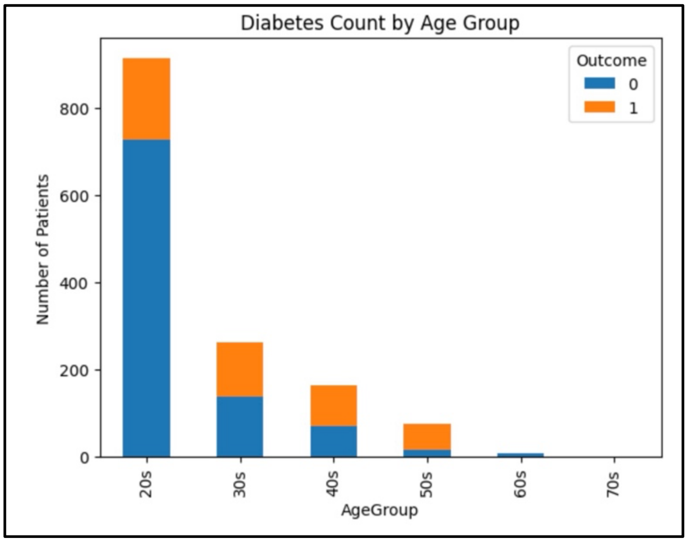
- **Purpose:** To visualize the BMI distribution for individuals diagnosed with and without diabetes. This can help determine if individuals with higher BMI are more likely to have diabetes.
- **Insight:** The bar chart will likely show that individuals with higher BMI values are more common among those with diabetes, supporting the relationship between obesity (BMI) and diabetes.



3. Age Distribution by Outcome (Diabetes Yes/No)

- **Purpose:** To examine how Age is distributed for individuals with and without diabetes. This will allow us to assess whether age plays a role in the likelihood of diabetes.

- **Insight:** The bar chart should reveal that older individuals are more likely to have diabetes, confirming age as a risk factor for the disease.



Key Findings:

- **Glucose Distribution by Outcome**: The bar chart depicting **Glucose** levels shows a clear distinction between individuals with and without diabetes. It indicates that those with diabetes generally have higher glucose levels compared to those without diabetes. This supports the hypothesis that elevated glucose levels are a significant indicator of diabetes risk.

- **BMI Distribution by Outcome**: The bar chart illustrating the **BMI** distribution for individuals with and without diabetes demonstrates that higher BMI values are more common among those with diabetes. Although, optimal BMI level should be around 25, patients with BMI between 25 and 30 are categorized as overweight and above 30 are categorized as Class 1 obesity. Thus, having Class 1 obesity directly increases chance of having diabetes. This finding reinforces the well-established link between obesity (measured by BMI) and the increased likelihood of developing diabetes.

- **Age Distribution by Outcome**: The bar chart showing the **Age** distribution reveals that older individuals are more likely to be diagnosed with diabetes. During the age of 20s and 30s number of people having diabetes is less compared to number of people not having diabetes during that age gap. But from the age of 40s, diabetes is present in half of the total number of people present in any age gap. This represents that from the age of 40 onwards, risk of diabetes increases by 50%. The data highlights a significant correlation between age and diabetes risk, with the likelihood of diabetes increasing as age advances.

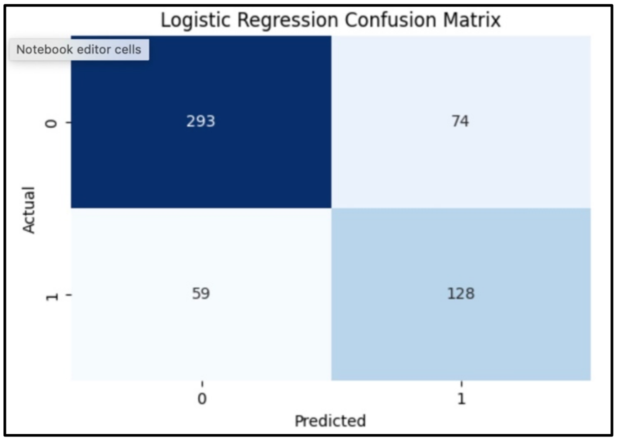
Analysis

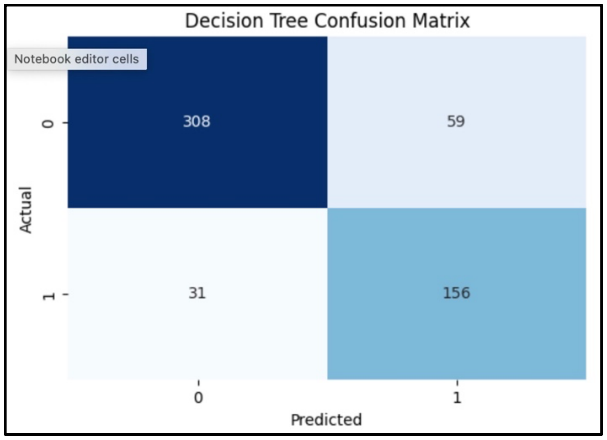
Steps and Methodologies: The analysis progressed through several stages.

1. **Exploratory data analysis (EDA):** Based on hypothesis, initial focus was on BMI, age and glucose levels and used visualizations to identify trends.
2. **Feature expansion:** VIF values were determined to check multicollinearity and prompted to include all the features.
3. **Correlation analysis:** Pearsons's correlations showed significant relationships with diabetes.
4. **Model development:** Applied logistic regression and decision tree models to predict diabetes using all the features.
   - **Tools:** Python (scikit-learn, matplotlib, seaborn, pandas)
   - **Techniques:** VIF for multicollinearity, Pearson correlation analysis, Logistic regression for linear classification, Decision tree for non-linear patterns, evaluation via confusion matrices and ROC curves.
   - **Why these models:** Decision tree captures complex relationships which are suitable for medical data with potential non-linearities among variables which is common in medical datasets. Its ability to split data based on features thresholds, allows it to model these patterns effectively.
   - While logistic regression provides interpretability, providing clear coefficients that indicate the relevance and direction of each features' influence on diabetes risk. Stakeholders at healthcare providers need to understand the working and results of the model's reasoning.
   - Logistic regression model serves as a baseline model, offering simple and linear perspective to compare against decision tree's ability to handle complex and non-linear relationships and its robustness to multicollinearity and predictive power.

Results: Following are models, and their results obtained.

Model/Results	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**ROC**
Logistic Regression	0.76	0.63	0.68	0.65	0.83
Decision Tree	0.83	0.72	0.83	0.77	0.90





Logistic Regression Confusion Matrix:

- It correctly identified 293 people who don't have diabetes (true negatives) and 128 who do suffer from diabetes (true positives).
- It missed 59 people who actually have diabetes, labelling them as not having it (false negatives). This is worrying since missing a diabetes case can be serious. It also wrongly flagged 74 people as diabetic when they weren't diabetic in real (false positives).

Decision Tree Confusion Matrix:

- This model did better overall, correctly spotting 308 people without diabetes (true negatives) and 156 with diabetes (true positives).
- It missed fewer diabetic cases—only 31 (false negatives), which is better. It had fewer mislabelling with 59 people wrongly labelled as diabetic (false positives).

The decision tree caught more people with diabetes (156 vs. 128) and missed fewer cases (31 vs. 59), which is important for a medical diagnosis. It also made fewer wrong predictions on people who don't have diabetes (59 vs. 74). So, the decision tree model feels more reliable and safer for patients.

Insights:

- Decision tree performance with AUC = 0.90 suggests that non-linear interactions enhanced prediction better than the logistic regression.
- Multicollinearity exists impacting the performance of logistic regression more than the decision tree.
- Glucose, BMI and age remain the key predictors while 'diabetespedigreefunction' (correlation = 0.161) and 'insulin' (correlation = 0.199) adding more value.

Connections to EDA insights: Results confirm that hypothesis: Higher glucose and BMI levels along with age indicates diabetes risk.

Questions arises due to multicollinearity found that including all the features in the model increased model robustness.

- **Glucose:** High glucose levels in diabetes (correlation = 0.489) aligns with model reliance
- **BMI:** Higher BMI in diabetes (correlation = 0.29) supports, though multicollinearity with skin thickness is noted
- **Age:** Diabetes risk increases with increasing age (correlation = 0.237)

The decision tree model's achieving higher AUC (0.90), and fewer false negatives (31) indicates that it effectively separates diabetes cases, leveraging all features' interactions. Visualizations like confusion matrix and ROC curve highlights this improvement over logistic regression where AUC is 0.83 and has 59 false negatives. This also presents multifaceted risk beyond glucose, age and BMI.

Interpretations:

1. Logistic regression achieves moderate performance having F1-score of 0.65, AUC of 0.83 with 128 true positives but 59 false negatives. Multicollinearity limits its effectiveness.
2. Decision tree outperforms with higher accuracy (0.83), F1-score of 0.77 and AUC of 0.90 simultaneously reducing false negatives to 31 and increasing true positives to 156. This reflects its ability to capture non-linear relationships and minimize effect of multicollinearity.

Implications for Healthcare Provider: The results validate glucose and BMI as primary factors which aligns with the hypothesis while 'insulin' and 'diabetespedigreefunction' enhances risk assessment. The decision tree model is reliable as it screens and minimize missed diagnosis (false negatives). This implies that apart from glucose and BMI, monitoring additional factors like insulin and diabetes-pedigree-function improves diabetes prevention and early intervention strategies.

Recommendations

1. Adopt and use decision tree model considering all the features for diabetes risk prediction as it improves performance over logistic regression, capturing non-linear patterns.
2. To capture broader risk of diabetes profiles, adding lifestyle factors would be more useful.
3. Focus on healthcare efforts is required like early monitoring of glucose and BMI levels should be prioritized as early detection or signs can prevent other diseases that maybe caused due to diabetes.
4. Incorporate Diabetes Pedigree Function (DPF) and insulin levels into standard risk assessments to better capture genetic predisposition and metabolic health. With correlations of 0.161 and 0.199, respectively, DPF and insulin provide valuable insights into hereditary risk and metabolic dysfunction, enabling earlier detection of at-risk individuals who may not yet show high glucose or BMI levels. This approach is particularly useful for younger patients or those with a family history of diabetes, allowing healthcare providers to tailor preventive strategies, such as genetic counselling or metabolic monitoring, to mitigate risk.
5. Wearable health devices and mobile health apps are widely available and can track real-time data on glucose levels, physical activity, and dietary habits. Integrate these technologies into the data collection process to provide dynamic, longitudinal data that can enhance the model's ability to predict diabetes risk over time. For example, continuous glucose monitors can provide more granular data than single-point measurements, offering deeper insights into glucose trends and variability.
6. Given Age's significant role (correlation: 0.237, with a 50% increased risk after 40), implement age-specific screening protocols, prioritizing patients over 40 for more frequent glucose and BMI checks. Additionally, since Pregnancies (correlation: 0.227) is relevant for female patients, incorporate gender-specific analyses to explore how pregnancy history influences diabetes risk, particularly for women with a history of gestational diabetes, which is a known precursor to type 2 diabetes.
7. Refine the dataset by addressing multicollinearity, particularly between BMI and Skin Thickness (VIF: 16.246216 for Skin Thickness), which share overlapping information about obesity. Consider consolidating these features (e.g., using BMI alone) or applying techniques like Principal Component Analysis (PCA) to reduce redundancy. This will create a cleaner dataset, improving the interpretability and efficiency of future models while maintaining focus on the most impactful predictors like glucose and insulin.

Conclusions

The analysis confirms that glucose, BMI, and age are significant indicators of diabetes risk, aligning with the initial hypothesis that higher glucose levels and BMI increase the likelihood of diabetes, with age amplifying this risk as patients cross the 40-year threshold. Glucose, with its strong correlation of 0.489, stands out as the most critical marker, while BMI (correlation: 0.290) underscores the role of obesity, particularly for patients exceeding a BMI of 30 (Class 1 obesity). Age, correlated at 0.237, highlights a 50% increased risk in older individuals, making it a key factor for long-term monitoring.

The inclusion of additional features like insulin (correlation: 0.199) and Diabetes Pedigree Function (correlation: 0.161) reveals the importance of metabolic and genetic factors, painting a more comprehensive picture of diabetes risk. The Decision Tree model, with its superior performance (AUC: 0.90, F1: 0.776) compared to Logistic Regression (AUC: 0.83, F1: 0.658), enhances prediction by capturing complex, non-linear interactions among these variables, reducing missed diagnoses (false negatives: 31 vs. 59).

This multifaceted risk assessment provides healthcare providers with a practical, data-driven tool for early detection, enabling timely interventions to prevent or manage diabetes. The model’s effectiveness can be further improved by incorporating lifestyle factors such as diet (e.g., sugar intake, carbohydrate consumption), food preferences (e.g., processed vs. whole foods), physical activity levels, and smoking habits into the dataset. These additions would offer a more holistic view of risk, potentially uncovering new patterns and enhancing the model’s ability to identify at-risk individuals before clinical symptoms emerge, ultimately supporting better health outcomes.