

**College of Professional Studies, Northeastern University**

**ALY 6040 Data Mining Application**

Prof. Justin Grosz

By

Jainam Patel

Smitkumar Dholiya

Dushang Shah

Date: 05/11/2025

**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.

The variables ‘glucose’, ‘skin thickness’, ‘insulin’, ‘BMI’ and ‘blood pressure’ indicates the levels of those metrics in the patient’s body.

The ‘age’ represents patient’s current age at the time of data collected.

The ‘pregnancies’ variable is the number of times the patient has been pregnant. However, count of 0 indicates that either lady hasn’t been pregnant till the data was collected or the patient might be a male.

The variable ‘ID’ is nothing but just the indexing/unique patient count.

The variable ‘Diabetes pedigree function’ is the numerical measure that estimates the genetic influence of diabetes on an individual based on their family history.

The variable ‘Outcome’ is the categorical variable indicating that patient is diabetic if it is 1 and non-diabetic if it is 0.

**Data Cleaning:**

To ensure the accuracy of the analysis, several data cleaning steps were undertaken:

1. **Handling Missing and Invalid Data**:  
   The dataset contained zero values in several columns, including **Glucose**, **Blood Pressure**, **Skin Thickness**, **Insulin**, and **BMI**. Since, zero values are unrealistic and not possible for living individuals and were thus treated as missing data. These invalid entries were replaced with the **median** of their respective columns.

* **Why the Median**: The median was selected over the mean for replacing the zero values because it is more robust to outliers, which are common in medical datasets. Using the median helps preserve the central tendency of the data while minimizing the impact of extreme values that might skew the results.
* **Columns Cleaned**:
  + **Glucose**: Zeros were replaced with the median value of glucose levels.
  + **Blood Pressure**: Zeros were replaced with the median value of blood pressure.
  + **Skin Thickness**: Zeros were replaced with the median value of skin thickness.
  + **Insulin**: Zeros were replaced with the median value of insulin levels.
  + **BMI**: Zeros were replaced with the median value of BMI.
* **Justification for Data Cleaning Choices**:  
  The zero values were assumed to be due to missing or incorrectly recorded data, rather than actual measurements. By replacing these zeros with the median, we ensured that the analysis would reflect more accurate relationships without being biased by invalid entries. This method of cleaning helps maintains the integrity of the dataset and supports the reliability of the findings.
* The column ‘ID’ was removed since it is just for indexing and is does not have any significance in predicting diabetes.

**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:

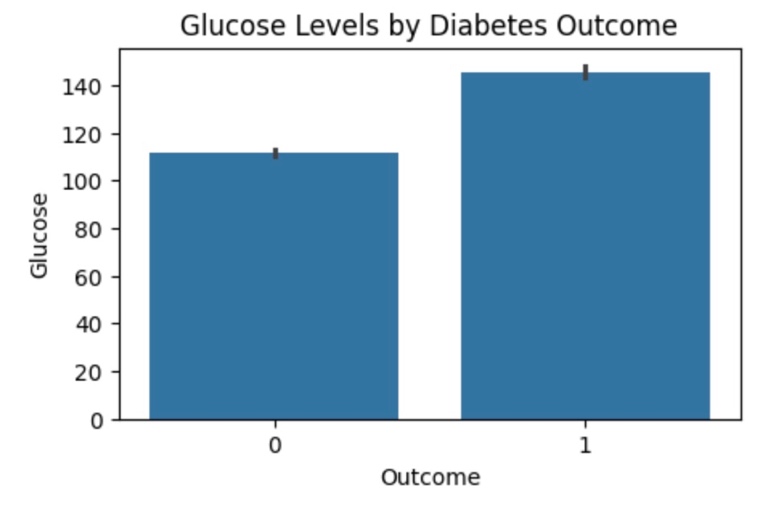
* **BMI**: Body Mass Index, which is a crucial indicator of obesity—a known risk factor for diabetes.
* **Glucose**: Plasma glucose concentration, which is directly associated with diabetes risk.
* **Age**: Age is an important factor, as the likelihood of developing diabetes tends to increase with age.

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

**Visualization:**

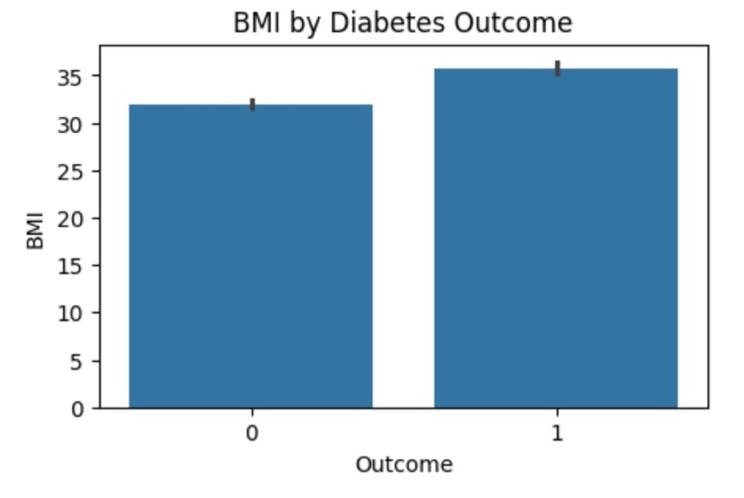
**1. Glucose Distribution by Outcome (Diabetes Yes/No)**

* **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.



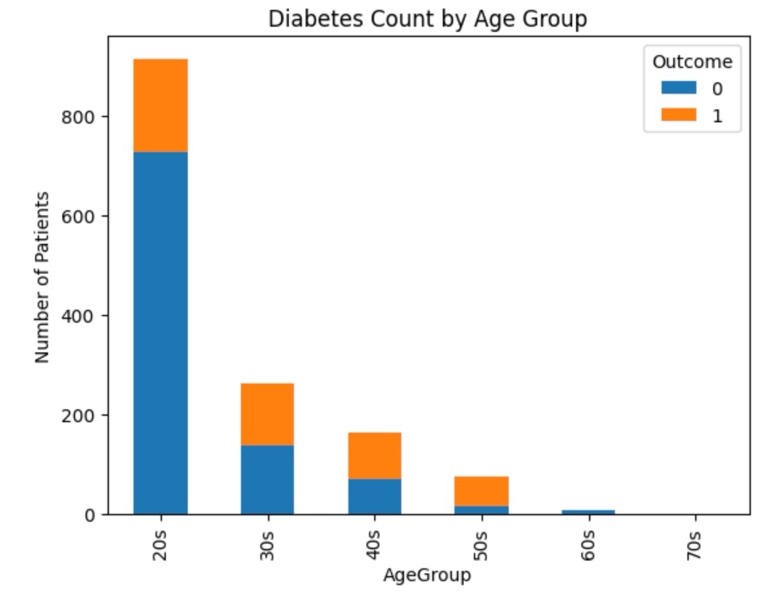
**2. BMI Distribution by Outcome (Diabetes Yes/No)**

* **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 while logistic regression provides interpretable coefficients.

**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 |

**A graph showing the number of cells

AI-generated content may be incorrect. A graph showing a number of cells

AI-generated content may be incorrect.**

**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 improve logistic regression model’s performance, applying principal component analysis (PCA) or feature selection can be relevant.
3. To capture broader risk of diabetes profiles, adding lifestyle factors would be more useful.
4. For model validation, conducting k-fold cross validation and hyperparameter tuning can ensure robustness.

**Conclusions**

The analysis confirms that glucose, BMI and age are significant diabetes risk indicators while decision tree model enhances prediction over logistic regression model. This supports a multifaceted risk assessment, offering healthcare providers a practical tool for early detection and it can even be further improved by incorporating lifestyle (diet, food preferences) of the patient into the dataset.