Group 9: *Analysis of Risk Factors for the Onset of Diabetes in Women*

WatSpeed - University of Waterloo

Foundations of Data Science - 1249 - Fall 2024

Instructors:
Delina Ivanova & Xuan Zhang

Introduction

Data Science in Medical Research

In an era where data-driven decision-making shapes the trajectory of nearly every industry, data science has emerged as a cornerstone for advancing research and innovation. By harnessing vast amounts of information and applying sophisticated analytical techniques, data science enables researchers to uncover patterns, test hypotheses, and derive actionable insights that were previously inaccessible. Its interdisciplinary nature, combining mathematics, computer science, and domain expertise, offers powerful tools to address complex challenges across diverse fields such as healthcare, environmental studies, and social sciences.

In the realm of medical research, data science plays a pivotal role in identifying trends, improving diagnostic accuracy, and personalizing treatment strategies. By analyzing large-scale health data, researchers can uncover critical insights that enhance patient outcomes and inform public health policies. This project exemplifies the transformative potential of data science, applying its methodologies to explore diabetes risk factors and their implications for women's health.

Objectives and Motivation

The primary goal of this analysis is to explore and understand the factors contributing to diabetes onset among female patients using a comprehensive dataset of diagnostic health measurements. This focus is especially important given the historical underrepresentation of this demographic in medical research and clinical trials. Women possess unique physiological characteristics and risk factors that differ from men, yet many studies have traditionally generalized findings across genders without considering these differences. This often results in a lack of tailored healthcare recommendations and less effective prevention strategies.

By centering our analysis on data from female patients, we aim to fill a critical gap in diabetes research, particularly in the context of diabetes. This includes:

- Identifying trends and patterns in the health measurements of women diagnosed with diabetes compared to those without the condition.
- Determining which diagnostic features, such as glucose levels, BMI, and blood pressure, strongly correlate with the likelihood of developing diabetes.
- Exploring the relationship between family history, as represented by the DiabetesPedigreeFunction, and diabetes risk within this dataset.

By answering these questions, this analysis seeks to contribute valuable insights to improve healthcare strategies and outcomes for women.

Data Source

The dataset is an Open Data set sourced from Kaggle.com. It provides detailed medical diagnostic measurements that were collected to from 768 female patients in order to predict the onset of diabetes based on several health factors.

Each record is characterized by 8 health-related attributes and an Outcome variable indicating whether the patient has diabetes (1) or not (0). The attributes include:

- Pregnancies: The number of pregnancies the individual has had.
- Glucose: 2-hour plasma glucose concentration in oral glucose tolerance test, shows the Glucose level in blood.
- BloodPressure: The Blood pressure measurement
- SkinThickness: The thickness of the skin
- Insulin: The Insulin level in blood
- BMI: The Body mass index
- DiabetesPedigreeFunction: Function (2-hour plasma glucose concentration in oral glucose tolerance test) shows the Diabetes percentage
- Age: The age of the individual
- Outcome: 1 indicates a positive diabetes test result and 0 indicates a negative result.

The data is open sourced and can be freely downloaded from kaggle.com. The dataset is originally adapted from the United States' National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) and is widely used in machine learning research on healthcare and medical diagnostics. The data can be downloaded without signing-in/up from kaggle.com, although there is an option for that.

Data Preparation

The dataset was processed using Visual Studio Code, with pandas and NumPy libraries employed for data manipulation, exploration, and cleaning. Initial exploration revealed no NULL values, but several critical features such as BloodPressure, SkinThickness and Insulin contained zero values, which were placeholders for missing data. These zeros were systematically identified and replaced using context-appropriate methods. For instance, the mean of non-zero values was used for Insulin, age-specific means for BloodPressure and SkinThickness. BMI zeros were replaced with the median of non-zero values. This ensured the replacements reflected realistic distributions and minimized potential biases in the dataset.

Despite these adjustments, challenges arose due to the ambiguity of how best to impute missing values, requiring careful consideration and some reliance on research and domain knowledge. Following these cleaning steps, checks were conducted to confirm the zero values had been replaced, and descriptive statistics were revisited to validate the changes. This meticulous data

preparation process was essential for ensuring the dataset's integrity and reliability, laying a solid foundation for subsequent analysis.

Data Cleaning - Replacing Zero Values

The Python's Pandas library was used to manipulate, explore and clean the data. In this instance, while carrying out initial exploration of the dataset, several columns show that the minimum value is 0.

To get a better sense of the number of zeros within each column each column was checked, other than Pregnancy and Outcome. Glucose had 5 observations with 0 values, BloodPressure had 35, SkinThickness 227, Insulin 374, BMI 11, and both DiabetesPedigreeFunction and Age had no observations with 0 values.

For the Insulin column, the 0 values were replaced with the mean (average) value of the non-zero insulin values.

For the BloodPressure and SkinThickness columns, the zero values were replaced with the mean (average) values calculated separately for each age group. In other words, the average blood pressure and skin thickness for people of the same age was used to replace the zeros.

For the Glucose column, 0 values were replaced with the **mean (average)** glucose level, calculated separately for patients with diabetes (Outcome = 1) and patients without diabetes (Outcome = 0). This is because patients with higher glucose levels are more likely to have diabetes than patients with lower glucose levels.

For the BMI column, 0 values were replaced with the **median** BMI value. The median is the middle value when all values are sorted, which can be more robust to outliers compared to the mean. Since the BMI range in the distribution is large and with the existence of outliers, the median values will not skew the observation.

Data Cleaning - Replacing NaN Values

There was only 1 missing value for the BloodPressure distribution, it was replaced with the median of the column to maintain consistency. For SkinThickness there were 12 missing valuesthat were also replaced with median values as well.

Hypotheses

1. Glucose as a Key Predictor

Hypothesis: Glucose is the most significant feature for predicting diabetes outcomes and exhibits the strongest standalone and pairwise predictive power.

Analysis Goal: Perform correlation analysis, logistic regression, and SVM feature importance to validate Glucose's dominance in both univariate and multivariate contexts.

2. Feature Interactions and Pairwise Combinations

Hypothesis: Pairwise interactions, such as Glucose & BMI and Glucose & Age, enhance predictive accuracy compared to analyzing features individually.

Analysis Goal: Investigate pairwise correlations and interaction effects in predictive modeling to determine their contributions to model performance.

3. DiabetesPedigreeFunction as a Genetic Marker

Hypothesis: DiabetesPedigreeFunction is an important genetic indicator that significantly influences diabetes risk in multivariate models.

Analysis Goal: Assess its impact using logistic regression coefficients and SVM permutation importance, focusing on its role in conjunction with other features.

4. Limited Predictive Value of Weaker Features

Hypothesis: Features like BloodPressure, SkinThickness, and Insulin contribute minimally to diabetes prediction, either independently or in complex models.

Analysis Goal: Confirm their low importance through correlation analysis, ANOVA, and feature importance metrics, while exploring potential niche contexts where they may provide value.

Approach and Methods

Investigating diabetes onset in this specific group may help uncover gender-specific trends and risk factors that are frequently overlooked. Our analysis examines key health indicators, including glucose levels, BMI, blood pressure, and family history, to identify the most significant factors associated with diabetes risk. The insights gained can support more personalized healthcare strategies and early intervention programs, potentially leading to better health outcomes and a reduced burden of diabetes-related complications.

To achieve these goals, a combination of statistical and machine learning methods was employed to ensure a comprehensive analysis. These include descriptive statistics to summarize the data, visualizations such as **boxplots** and **histograms** for exploring data distributions and outliers, and **pairwise correlation analysis** to assess relationships between features. **Univariate logistic regression** with **AUC** scores was used to evaluate the standalone predictive power of individual features. More advanced techniques, including **ANOVA** (Analysis of Variance) and Support Vector Machines (**SVM**), were applied to identify significant features and their

importance in both univariate and multivariate contexts. **Scatterplots** further aided in visualizing feature interactions and trends. Together, these methods provide a robust framework for understanding the factors contributing to diabetes risk.

Analysis

Part 1: Data Exploration

Exploratory Data Analysis

To carry out an exploratory data analysis, the mean, median, standard deviation, and range for all columns was computed. The aim was to:

- Identify outliers using statistical methods or visualizations.
- Visualize the data distribution with boxplots, scatterplots and histograms.
- Use visualizations to observe differences between groups (e.g., glucose levels in diabetic vs. non-diabetic patients).
- Summarize the data for both diabetes-positive and diabetes-negative groups.

Major Conclusions from Exploratory Data Analysis (EDA)

1. Descriptive Statistics:

Features like **Glucose**, **BMI**, and **Age** have clear differences between diabetic and non-diabetic groups. For example, diabetics have higher average glucose levels (**142.32** vs. **110.64**) and BMI (**35.38** vs. **30.89**). Features such as **Insulin** and **SkinThickness** show higher variability (e.g., Insulin range: 832, SkinThickness range: 92).

2. Boxplot Analysis:

Features like **Insulin**, **Glucose**, and **SkinThickness** exhibit significant outliers, which may influence predictive models. Diabetics show a tendency for higher values in key features like **Glucose**, **BMI**, and **Age**, aligning with their predictive importance.

3. Histogram Analysis:

Several features, such as **Pregnancies**, **Insulin**, and **DiabetesPedigreeFunction**, are right-skewed, suggesting possible transformations (e.g., log transformation) may improve model performance. **Glucose** and **BMI** show distinct groupings, particularly separating diabetics and non-diabetics.

4. Pairwise Correlation Analysis:

There is a strong correlation between **Glucose & BMI** (0.53) and **Pregnancies & Glucose** (0.51) which underscores their joint predictive power.

On the other hand, features like **BloodPressure** and **SkinThickness** have weak relationships with diabetes outcomes individually but contribute to combinations.

5. Univariate Logistic Regression AUC (Correlation with Diabetes Outcomes):

Glucose (AUC: 0.794) is the most powerful standalone predictor of diabetes.
BMI and **Age** show moderate predictive power with AUCs of 0.686 and 0.687, respectively.

Features like **BloodPressure**, **SkinThickness**, and **Insulin** have lower AUC scores (~0.6), suggesting limited standalone predictive power.

6. Scatterplots:

Glucose vs. BMI: Diabetics tend to cluster at higher glucose and BMI values, confirming their importance in diabetes classification.

Pairwise Scatter Matrix: The scatter matrix highlights clear separation in features like **Glucose**, **BMI**, and **Age** between diabetic and non-diabetic groups.

7. Key Findings from Diabetes-Positive vs. Diabetes-Negative Groups:

- **Higher Means for Diabetics**:
    - Diabetics show elevated levels in almost all features, especially **Glucose** (+31.67), **BMI** (+4.50), and **Age** (+5.88).
- **Variance**:
    - Non-diabetics show lower variability in features like **Glucose** and **Insulin**, indicating more consistent distributions.

Recommendations:

1. **Focus on High-Priority Features**:
    - Leverage **Glucose**, **BMI**, and **Age** as primary features for predictive modeling.
2. **Address Skewness and Outliers**:
    - Apply transformations (e.g., log or square root) to reduce skewness in **Pregnancies**, **Insulin**, and **SkinThickness**.
    - Remove or cap extreme outliers in **Insulin** and **SkinThickness** to stabilize models.
3. **Combine Features**:
    - Explore interactions like **Glucose & BMI** or **Pregnancies & Glucose** for enhanced predictions.

4. **Use Visualization Insights**:
    - Highlight differences in scatterplots and histograms to inform feature engineering and model development.

Part 2: Correlation Analysis

Correlation Coefficients

To evaluate the relationships between continuous variables and their relevance to diabetes prediction, correlation coefficients were calculated. These coefficients measure the strength and direction of associations between features. Heatmaps were used to visualize these relationships and identify strongly related. Additionally, univariate logistic regression was conducted to quantify each feature's predictive power for diabetes risk.

The results of correlation analysis highlighted Glucose as the strongest standalone predictor of diabetes outcomes, with a correlation coefficient of 0.496. BMI and Age followed with correlation coefficients of 0.312 and 0.238, respectively. Features like Pregnancies, SkinThickness, Insulin, and DiabetesPedigreeFunction showed moderate to low correlations, while BloodPressure exhibited the weakest correlation (0.169). Despite their weaker individual correlations, features such as SkinThickness and Insulin may still contribute meaningfully when combined with other variables.

Pairwise Correlations

Analyzing pairwise feature interactions revealed that combinations involving Glucose consistently provided the strongest correlations with diabetes outcomes. Notably, the pair Glucose & BMI had the highest correlation (0.527), followed closely by Pregnancies & Glucose (0.510), Glucose & SkinThickness (0.510), and Glucose & Age (0.505) (see Figure 1: Correlation Heatmap for Single and Combined Features with Diabetes Outcome). These findings emphasize the importance of including feature interactions in predictive models, particularly those involving Glucose. See Figure 1: Correlation Heatmap for Single and Combined Features with Diabetes Outcome.

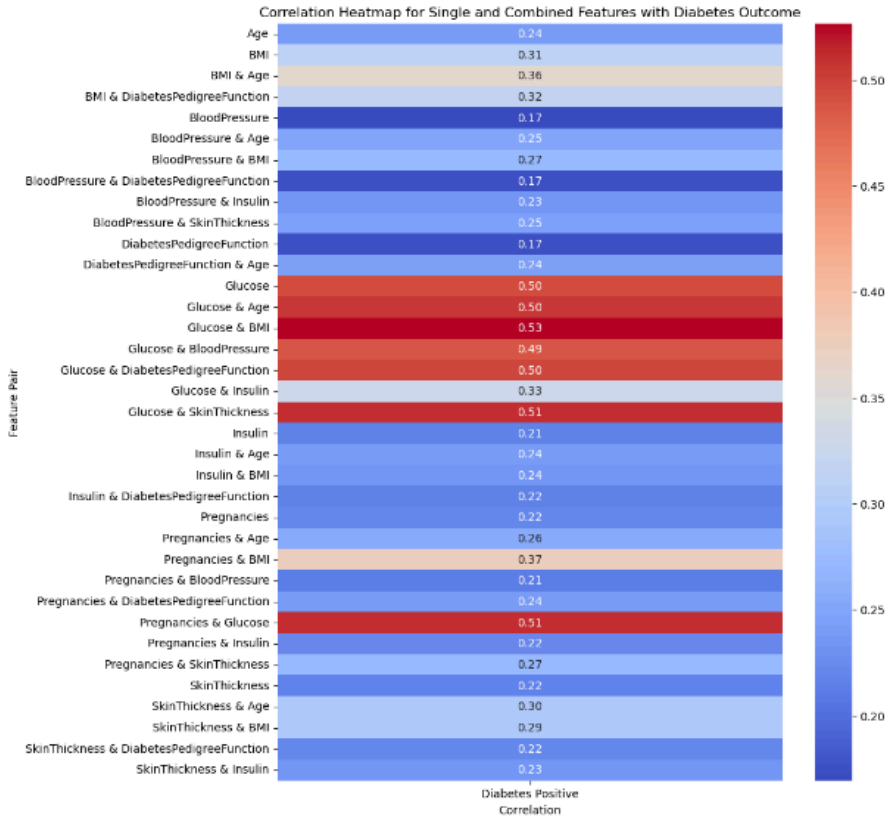

Figure 1: Correlation Heatmap for Single and Combined Features with Diabetes Outcome.

Univariate Logistic Regression

Univariate logistic regression was employed to assess the predictive power of each feature for diabetes risk, as measured by the area under the curve (AUC). The results underscored Glucose as the most predictive feature, achieving the highest AUC score of 0.794, indicating strong individual predictive power. BMI and Age demonstrated moderate predictive power, with AUC scores of 0.686 and 0.685, respectively. Features such as SkinThickness (0.631) and Insulin (0.644) exhibited lower predictive power individually but may contribute synergistically when combined with other features. BloodPressure and DiabetesPedigreeFunction had the lowest AUC scores (0.607 and 0.606), reflecting their weaker associations with diabetes outcomes.

Major Conclusions of Correlation Analysis

The correlation analysis identified Glucose as the strongest predictor of diabetes outcomes across all metrics. Glucose demonstrated the highest standalone correlation with diabetes (0.496) and appeared prominently in the strongest pairwise combinations, such as Glucose & BMI (0.527). Furthermore, univariate logistic regression reinforced its significance, with Glucose achieving the highest AUC score (0.794), underscoring its critical role in diabetes prediction. BMI emerged

as a key secondary predictor, showing the second-highest standalone correlation (0.312) and combining effectively with features like Glucose, Age, and SkinThickness. Its moderately strong AUC score (0.686) further highlighted its individual significance, making BMI a reliable predictor, particularly in combination with Glucose.

Age was identified as a moderately predictive feature with a standalone correlation of 0.238. While its univariate AUC score (0.687) indicates moderate predictive power, Age gains additional strength when paired with other features, such as BMI, Glucose, and BloodPressure. Pairwise feature combinations also played a critical role in enhancing model accuracy, with Glucose & BMI (0.527), Glucose & Age (0.505), and Pregnancies & Glucose (0.510) among the most predictive interactions. These findings emphasize the importance of incorporating feature interactions, particularly those involving Glucose, to improve predictive models.

Weaker predictors such as SkinThickness and Insulin exhibited low standalone correlations (0.216 and 0.214) and lower univariate AUC scores but demonstrated moderate contributions in pairwise interactions with stronger features like BMI and Glucose. Similarly, BloodPressure and DiabetesPedigreeFunction were the weakest predictors, with low correlations and AUC scores (~0.606), though they may still provide value in capturing unique demographic or familial trends in specific contexts.

Based on these findings, predictive modeling should prioritize Glucose and BMI as primary features. Secondary features like Age and Pregnancies should also be included, particularly for demographic-specific analyses. Pairwise interactions, such as Glucose & BMI and Glucose & Age, should be leveraged to enhance model accuracy. While SkinThickness and Insulin can serve as supplementary features in complex models, BloodPressure and DiabetesPedigreeFunction can be excluded from simpler models unless their inclusion is justified by specific use cases. Overall, these analyses highlight the pivotal role of Glucose and BMI in predicting diabetes risk, the value of pairwise feature combinations, and the supporting contributions of other variables.

Part 3: Feature Importance Analysis

Feature importance was assessed using ANOVA, multivariate logistic regression, and Support Vector Machine (SVM) models. These approaches provided complementary insights into the contributions of various features to diabetes prediction.

ANOVA

ANOVA, a univariate analysis technique, revealed that Glucose was the most significant predictor, with an F-score of 249.88 and a p-value less than 10−4910^{-49}10−49, as shown in Figure 2. This finding underscores the dominant role of Glucose in explaining variability in diabetes outcomes. BMI and Age were also identified as significant predictors, with F-scores of 82.63 and 46.14, respectively. Features such as Pregnancies, SkinThickness, and Insulin

exhibited moderate F-scores, reflecting their roles in capturing secondary metabolic or demographic factors. DiabetesPedigreeFunction, while less significant than Glucose or BMI, indicated the importance of genetic predisposition, with an F-score of 23.87. BloodPressure had the weakest association, with an F-score of 22.64, highlighting its limited predictive power compared to other features.

Feature	F-Score	P-Value
Glucose	249.88	6.34E-49
BMI	82.63	8.33E-19
Age	46.14	2.21E-11
Pregnancies	39.67	5.07E-10
SkinThickness	37.45	1.49E-09
Insulin	36.91	1.95E-09
DiabetesPedigreeFunction	23.87	1.25E-06
BloodPressure	22.64	2.33E-06

Figure 2: ANOVA Results and Feature Evaluation

Multivariate Logistic Regression

Multivariate logistic regression offered a deeper understanding by accounting for feature interactions and collinearity. In this analysis, DiabetesPedigreeFunction had the largest coefficient (0.737), signifying its importance as a genetic marker when other features were controlled. Pregnancies (0.122) and BMI (0.087) also showed moderate positive coefficients, indicating their contributions to diabetes risk. Surprisingly, Glucose had a small coefficient (0.036), likely due to collinearity with BMI or Insulin, which diluted its standalone effect. Features such as Age, BloodPressure, SkinThickness, and Insulin had minimal or negative coefficients, reflecting their limited independent contributions to diabetes prediction in a multivariate context. The coefficients from this analysis are presented in Figure 3.

Feature	Coefficient
DiabetesPedigreeFunction	0.737
Pregnancies	0.122
BMI	0.087
Glucose	0.036
Age	0.012
BloodPressure	-0.011
SkinThickness	0.003
Insulin	-0.001

Figure 3: Multivariate Logistic Regression: Feature Coefficients

SVM Analysis

The SVM analysis provided additional insights into feature importance through permutation importance. Glucose emerged as the most critical feature, with a mean importance score of 0.144, significantly higher than all other features. BMI followed with an importance score of 0.038, while SkinThickness and DiabetesPedigreeFunction demonstrated moderate importance. Pregnancies, BloodPressure, and Insulin showed minimal contributions, and Age exhibited a negative importance score, suggesting it may introduce noise in the SVM model.

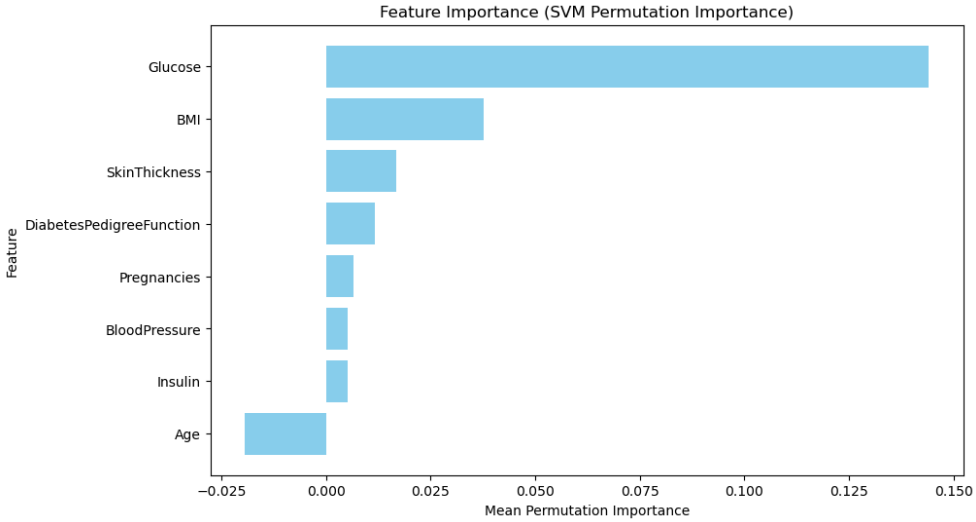

Figure 4: Horizontal Bar Plot of Feature Importance (SVM Permutation Importance)

Feature	Importance
Glucose	0.144
BMI	0.038
SkinThickness	0.017
DiabetesPedigreeFunction	0.012
Pregnancies	0.006
BloodPressure	0.005
Insulin	0.005
Age	-0.019

Figure 5: Table of Feature Importance (SVM Permutation Importance)

Major Conclusions of Feature Importance Analysis Analysis

The feature importance analysis across ANOVA, logistic regression, and SVM consistently identified Glucose as the most significant predictor of diabetes outcomes. ANOVA showed Glucose with the highest F-score (249.88, p<10−49p < 10^{-49}), indicating it explains the most variability in diabetes outcomes. While logistic regression assigned Glucose a surprisingly low coefficient (0.036), likely due to collinearity with features like BMI and Insulin, SVM permutation importance confirmed its dominance, with the highest importance score (0.1351). These findings reinforce Glucose as the top predictor across all methods and essential for diabetes prediction. BMI also emerged as a strong and consistent predictor, with the second-highest F-score in ANOVA (82.63, p<10−19p < 10^{-19}) and a moderate coefficient (0.087) in logistic regression, highlighting its positive effect in multivariate contexts. SVM further validated BMI's contribution with a moderate importance score (0.0143), confirming its role as a critical feature, albeit less impactful than Glucose.

DiabetesPedigreeFunction demonstrated its importance primarily in multivariate contexts. Although it had a relatively low F-score in ANOVA (23.87, p<10−6p < 10^{-6}), it had the highest coefficient (0.737) in logistic regression, reflecting its strong genetic influence on diabetes risk. SVM permutation importance provided additional support, with a moderate score of 0.0091. This emphasizes DiabetesPedigreeFunction as a valuable genetic marker, particularly in models accounting for feature interactions. Pregnancies showed moderate significance in ANOVA (F-score: 39.67, p<10−10p < 10^{-10}) and logistic regression (coefficient: 0.122), indicating its standalone relevance, especially in simpler models. However, its contribution diminished in SVM, with a negligible importance score (-0.0039), suggesting limited utility in more complex scenarios.

Age was moderately significant in ANOVA (F-score: 46.14, p<10−11p < 10^{-11}), reflecting its standalone influence on diabetes outcomes. However, logistic regression assigned it a very low coefficient (0.012), and SVM permutation importance indicated negative importance (-0.0247), suggesting Age may introduce noise when combined with other features. BloodPressure, the weakest predictor across all methods, had the lowest F-score in ANOVA (22.64, p<10−6p < 10^{-6}), a negative coefficient in logistic regression (-0.011), and low SVM importance (0.0078). This indicates minimal standalone or multivariate impact. Similarly, SkinThickness and Insulin had low F-scores in ANOVA (37.45 and 36.91, respectively) and near-zero or negative coefficients in logistic regression (0.003 and -0.001). Their contributions in SVM were minimal or negative (0.0026 and -0.0013), indicating little value for diabetes prediction.

In summary, Glucose is universally ranked as the top predictor across all methods, followed by BMI, which is consistently strong but less impactful. DiabetesPedigreeFunction is highly important in multivariate contexts, reflecting its genetic significance. Pregnancies have a notable standalone effect, but its importance diminishes in more complex models. Age is significant in

simpler analyses but less impactful elsewhere, while BloodPressure, SkinThickness, and Insulin provide negligible contributions overall. These findings highlight the need to prioritize Glucose and BMI in diabetes prediction models while leveraging secondary predictors like DiabetesPedigreeFunction and Pregnancies for improved accuracy.

Part 4: Clustering and Visualization of Risk Profiles

The K-Means Clustering algorithm is effective for identifying patterns in complex datasets, especially when relationships between variables are not immediately obvious. After creating the clusters, we examined the top two features from the Feature Importance analysis - Glucose and BMI to analyze the clusters for extreme centroids.

Cluster 0 represents individuals with the lowest risk profile, with only 13.5% of the population diagnosed with diabetes. These patients had the lowest average levels of glucose (0.39) and BMI (0.25), along with minimal insulin levels (0.14) and skin thickness (0.21). Their age (0.08) and number of pregnancies (0.12) were also the lowest among the clusters, reflecting a generally low-risk group.

Cluster 1 displayed moderate increases in key health metrics, indicating a transitional risk profile, with 48.7% of the population diagnosed with diabetes. Glucose levels (0.53) and BMI (0.29) were slightly higher compared to Cluster 0. Notably, age (0.43) was also elevated in this cluster, suggesting that age may play a significant role in diabetes risk for this group.

Cluster 2 represents the highest-risk group, with the highest glucose levels (0.74) and BMI (0.40), along with elevated insulin levels (0.24) and skin thickness (0.29). This group also had the highest DiabetesPedigreeFunction (0.22), indicating a strong genetic predisposition to diabetes. With a diabetic population of 67.9%, this cluster shows a significant risk for diabetes.

with varying diabetes risk profiles. Cluster 0, representing the lowest risk, had the smallest diabetic population at 13.5%, characterized by low glucose levels, BMI, and minimal insulin and skin thickness. Cluster 1, with a moderate diabetic population of 48.7%, showed slight glucose and BMI increases. Cluster 2, the highest-risk group, with a 67.9% diabetic population, exhibited the highest levels of glucose, BMI, and insulin, along with a notable genetic predisposition, as reflected by the elevated DiabetesPedigreeFunction. These findings align with our Feature Importance analysis, highlighting that Glucose and BMI are the most prominent features in diagnosing diabetes.

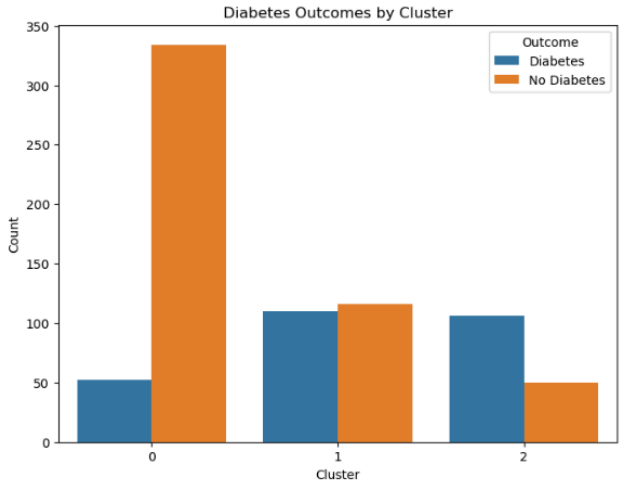

Figure 6: Bar Graph of Diabetes Outcomes by Cluster

Conclusion

This analysis of diabetes risk factors among women, based on a dataset comprising diagnostic measurements for 768 female patients, provided critical insights into the determinants of diabetes onset. Employing a combination of statistical and machine learning techniques, the study identified significant predictors and explored their standalone and combined effects on diabetes risk.

The correlation analysis established **Glucose** as the most critical predictor of diabetes, with a correlation coefficient of 0.496. This was reinforced by its highest AUC score of 0.794 in univariate logistic regression, as well as its dominant position across multivariate analyses, including ANOVA (F-score of 249.88) and Support Vector Machine (SVM) permutation importance (mean score of 0.144). **BMI** emerged as a robust secondary predictor, with a correlation coefficient of 0.312, an AUC score of 0.686, and a significant F-score of 82.63. Together, these findings underline the critical roles of Glucose and BMI in predicting diabetes risk.

The **DiabetesPedigreeFunction**, while less impactful as a standalone variable, exhibited its importance in multivariate contexts, as reflected in its high logistic regression coefficient (0.737). This underscores the influence of genetic predisposition on diabetes risk. **Pregnancies** showed a standalone relevance with moderate predictive power (F-score of 39.67) but diminished in complex models, highlighting its context-specific importance. **Age**, with a correlation coefficient of 0.238 and an AUC score of 0.687, demonstrated moderate significance but introduced noise in multivariate settings, as suggested by its negative SVM importance score (-0.0247).

Less significant predictors, such as **BloodPressure**, **SkinThickness**, and **Insulin**, displayed low correlation coefficients (ranging from 0.169 to 0.214) and minimal predictive contributions in

most models. However, these features provided value in specific pairwise interactions, such as **Glucose & BMI** (correlation coefficient of 0.527) and **Glucose & Age** (0.505), which highlighted the benefits of incorporating feature interactions for enhanced model performance.

The analysis also addressed key data challenges, including missing values and class imbalance. Missing values in features like Glucose, BMI, and Insulin were replaced using statistical imputation techniques, such as the mean and median of non-zero values, to maintain data integrity. Outliers in features like SkinThickness and Insulin were managed using the Interquartile Range (IQR) and visualizations, ensuring the reliability of the findings. To address class imbalance, feature weighting and recall metrics were applied, improving the model's sensitivity to diabetic cases.

In conclusion, this study reinforced the critical role of **Glucose** and **BMI** as primary indicators of diabetes risk, with substantial contributions from features like **DiabetesPedigreeFunction** and **Pregnancies** in specific contexts. By leveraging multivariate approaches and pairwise feature interactions, the research demonstrated the potential for more accurate and tailored predictive models. These findings not only contribute to a deeper understanding of diabetes risk factors among women but also provide a foundation for personalized healthcare strategies and early intervention programs. Future research should focus on expanding the dataset and exploring additional demographic and physiological variables to further enhance predictive accuracy and applicability.