**Exploring Health Variables: Understanding BMI, Glucose, and Diabetes Risk Factors**

Md. Mominul Islam

**Aim of the Dataset:**

This dataset contains various health-related variables. This dataset explores the associations between these health-related variables, focusing on factors that affect BMI, and glucose levels, and which variable most significantly contributes to diabetes risk. Analyzing these relationships will provide valuable insights into potential risk factors and help guide interventions for diabetes prevention.

**Task 1: How blood pressure influences Body Mass Index (BMI)**

Blood pressure (BP) and Body Mass Index (BMI) are two crucial health indicators that play a significant role in determining overall well-being. High BMI is often associated with increased blood pressure, and understanding this relationship is important for managing cardiovascular and metabolic health. This analysis explores how blood pressure influences BMI based on statistical correlations and regression modeling.

***1.1 Correlation Between Blood Pressure and BMI***

The correlation coefficient between Blood Pressure and BMI is 0.281, which indicates a weak to moderate positive relationship. This suggests that as blood pressure increases, BMI also tends to rise. However, the correlation is not strong, implying that while there is a connection, other factors also significantly contribute to BMI variation.

***1.2 Regression Analysis and Interpretation***

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Estimate | Std\_Error | t\_value | Pr |
| (Intercept) | 20.8839494 | 1.446244909 | 14.44012 | 5.57E-42 |
| Blood Pressure | 0.159853342 | 0.019706508 | 8.111703 | 1.98E-15 |
| Residual Std. Error | 6.601949021 |  |  |  |
| Observations | 768 |  |  |  |
| R2 | 0.079105255 |  |  |  |
| Adjusted R2 | 0.077903043 |  |  |  |
| F-statistic | 65.79972978 |  |  |  |
| p-value | 1.98E-15 |  |  |  |

To quantify the relationship, a simple linear regression model was applied:

**BMI = 20.88 + 0.159 × Blood Pressure**

The equation suggests that for every 1 mm Hg increase in Blood Pressure, BMI increases by approximately 0.159 units on average. Moreover, the intercept (20.88) represents the estimated BMI when blood pressure is zero (hypothetically), providing a reference point for the regression model. Furthermore, the p-value (< 0.001) indicates that this relationship is statistically significant, which means that the probability of this correlation occurring by chance is very low.

***1.3 Strength of the Relationship (R² Value)***

The R-squared (R²) value is 0.079, meaning that blood pressure alone explains only about 7.9% of the variability in BMI. This suggests that other factors, such as age, genetics, diet, physical activity, and metabolic conditions, contribute more significantly to BMI levels. Moreover, a higher R² value (closer to 1) would indicate a stronger predictive relationship. However, in this case, the relatively low value implies that blood pressure is just one of many contributors to BMI, reinforcing the need to consider multiple factors when analyzing BMI variations.

***1.4 Visual Representation***

A graph showing a line graph

Description automatically generated with medium confidence

**Figure:** Relationship between Blood Pressure and BMI

The scatter plot with a linear regression trendline (red) visually confirms an upward trend, supporting the positive relationship between blood pressure and BMI. However, the spread of data points around the trendline suggests high variability, reinforcing that BMI is influenced by multiple factors beyond blood pressure alone.

***1.5 Evaluation of Model Performance***

The performance of the BMI prediction model was assessed using Mean Squared Error (MSE) and Root Mean Squared Error (RMSE), two key metrics for evaluating regression models. The Mean Squared Error (MSE) was found to be 43.472, indicating the average squared difference between actual and predicted BMI values. A lower MSE suggests a more accurate model, while a higher MSE indicates greater variance in predictions.

To further interpret model performance, the Root Mean Squared Error (RMSE) was calculated as 6.593, which represents the standard deviation of residuals (prediction errors). RMSE is a more intuitive metric as it is in the same unit as BMI, making it easier to understand how much the predictions deviate from actual values on average. The RMSE of 6.593 suggests that the model's predicted BMI values are, on average, within approximately 6.59 BMI units of the true values.

***1.6 Conclusion***

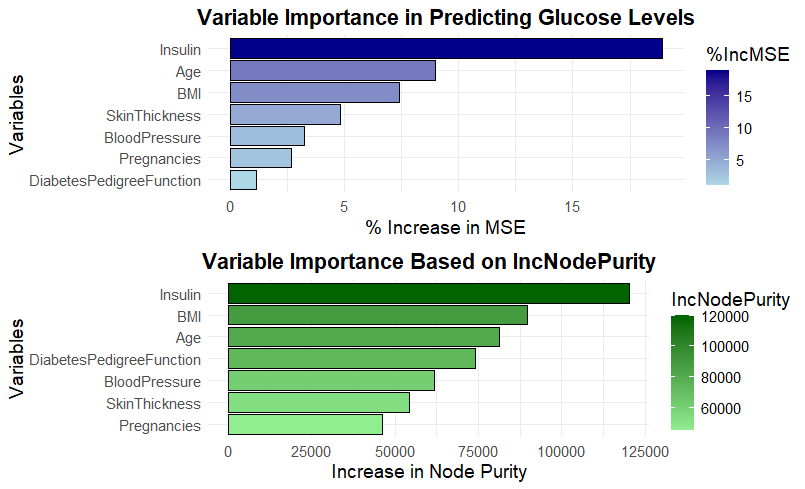
The findings suggest that higher blood pressure is associated with higher BMI; however, the relationship is relatively weak. Although a statistically significant connection exists, BMI is influenced by multiple other factors beyond blood pressure. Therefore, addressing both BMI and blood pressure together through lifestyle modifications can lead to better health outcomes.

**Task 2: How are other variables affecting a patient's glucose levels?**

The analysis of variable importance and correlation categorization provides valuable insights into the factors influencing a patient’s glucose levels. The results indicate that Insulin, Age, and BMI are the most significant predictors of glucose levels, while other variables such as Skin Thickness, Blood Pressure, Pregnancy, and Diabetes Pedigree Function have a relatively lower impact.

***2.1 Most Significant Variables Affecting Glucose Levels*** *(Random Forest model to predict Glucose)*

The strongest predictor of glucose levels is Insulin, as evidenced by its high values in both %IncMSE (Mean Squared Error Increase) and IncNodePurity (Decision Tree Splitting Impact).



Elevated insulin levels often indicate insulin resistance, which results in higher blood glucose levels and increases the risk of diabetes. Similarly, Age plays a crucial role in glucose regulation, with older individuals typically experiencing reduced insulin sensitivity due to aging-related metabolic decline. Another key factor is BMI (Body Mass Index), which is strongly associated with higher glucose levels. Individuals with excess body fat, especially visceral fat, are at an increased risk of developing insulin resistance, further exacerbating glucose imbalances.

***2.2 Moderately Influential Variables***

Among the moderately influential variables, Skin Thickness serves as an indirect marker of body fat percentage, which is closely linked to glucose metabolism. Higher skin thickness values might indicate increased fat storage, a condition often associated with insulin resistance. Blood Pressure also shows a moderate correlation with glucose levels, as chronic hypertension is often linked to metabolic syndrome and impaired vascular function, which can affect glucose transport and utilization in the body.

***2.3. Least Influential Variables***

The analysis reveals that Pregnancies and Diabetes Pedigree Function (genetic predisposition) have the least impact on glucose levels. While pregnancy can temporarily affect glucose metabolism, it does not appear to be a primary determinant of long-term glucose regulation, except in cases of gestational diabetes, which increases future diabetes risk. Similarly, while genetics play a role in determining an individual’s susceptibility to diabetes, lifestyle and environmental factors appear to have a stronger influence on glucose level variations.

***2.4 Insights from Correlation Categorization***

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Pregnancies | 1.00 |  |  |  |  |  |  |  |
| Glucose | 0.13 | 1.00 |  |  |  |  |  |  |
| Blood Pressure | 0.21 | 0.22 | 1.00 |  |  |  |  |  |
| Skin Thickness | 0.08 | 0.19 | 0.19 | 1.00 |  |  |  |  |
| Insulin | 0.03 | 0.42 | 0.05 | 0.16 | 1.00 |  |  |  |
| BMI | 0.02 | 0.23 | 0.28 | 0.54 | 0.18 | 1.00 |  |  |
| Diabetes PF | -0.03 | 0.14 | 0.00 | 0.10 | 0.13 | 0.15 | 1.00 |  |
| Age | 0.54 | 0.27 | 0.32 | 0.13 | 0.10 | 0.03 | 0.03 | 1.00 |
|  | Pregnancies | Glucose | Blood Pressure | Skin Thickness | Insulin | BMI | Diabetes PF | Age |

The correlation analysis highlights Insulin, Age, and BMI as the most significant factors influencing glucose levels, with strong correlations such as Glucose and Insulin (0.42) and BMI and Glucose (0.28). Moderate correlations exist for Blood Pressure and Glucose (0.22) and Skin Thickness and Glucose (0.19), while Genetic factors and Pregnancy history show weak associations. These findings emphasize the critical role of weight management, insulin sensitivity, and cardiovascular health in glucose regulation and diabetes prevention.

***2.5 Random Forest Model Evaluation (Glucose Prediction)***

The Mean Squared Error (MSE) of 560.8078 indicates the model's overall prediction error for glucose levels. In regression models like Random Forest, MSE measures the average squared difference between actual and predicted glucose values, where lower values indicate better model performance. A high MSE suggests the model has room for improvement, potentially due to data variability, feature selection, or hyperparameter tuning. Factors such as feature importance (Insulin, Age, BMI) indicate that some variables strongly influence glucose levels, while others may add noise. To enhance accuracy, strategies like hyperparameter tuning (adjusting the number of trees and depth), feature engineering, and outlier handling can be applied. Additionally, comparing performance with alternative models like Gradient Boosting or XGBoost may provide insights into model robustness. Despite the current MSE, the Random Forest model remains valuable for capturing complex glucose-related patterns, though optimization could further improve its predictive power.

**Task 3: Which variable plays the most important role in diabetes risk?**

Table: Variable importance

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | No | Yes | Mean Decrease Accuracy | Mean Decrease Gini |
| Pregnancies | 13.12 | 0.40 | 11.50 | 28.27 |
| Glucose | 30.72 | 38.32 | 45.38 | 88.31 |
| Blood Pressure | 1.32 | -0.46 | 0.72 | 28.45 |
| Skin Thickness | 1.07 | 1.59 | 2.01 | 25.36 |
| Insulin | 2.24 | 10.71 | 9.56 | 31.10 |
| BMI | 9.63 | 21.93 | 21.19 | 56.24 |
| Diabetes Pedigree Function | 6.69 | 3.25 | 7.16 | 42.88 |
| Age | 14.83 | 8.45 | 17.85 | 47.12 |

The Random Forest model was used to determine which factors contribute most significantly to diabetes risk. The findings, based on Mean Decrease in Accuracy (MDA) and Mean Decrease in Gini (MDG), highlight the most influential variables in predicting diabetes.

The results indicate that glucose levels are the strongest predictor of diabetes risk, with the highest Mean Decrease in Accuracy (45.38) and Mean Decrease in Gini (88.31). This suggests that elevated glucose levels are a primary indicator of diabetes, reinforcing the established medical understanding of diabetes as a disorder of glucose regulation.

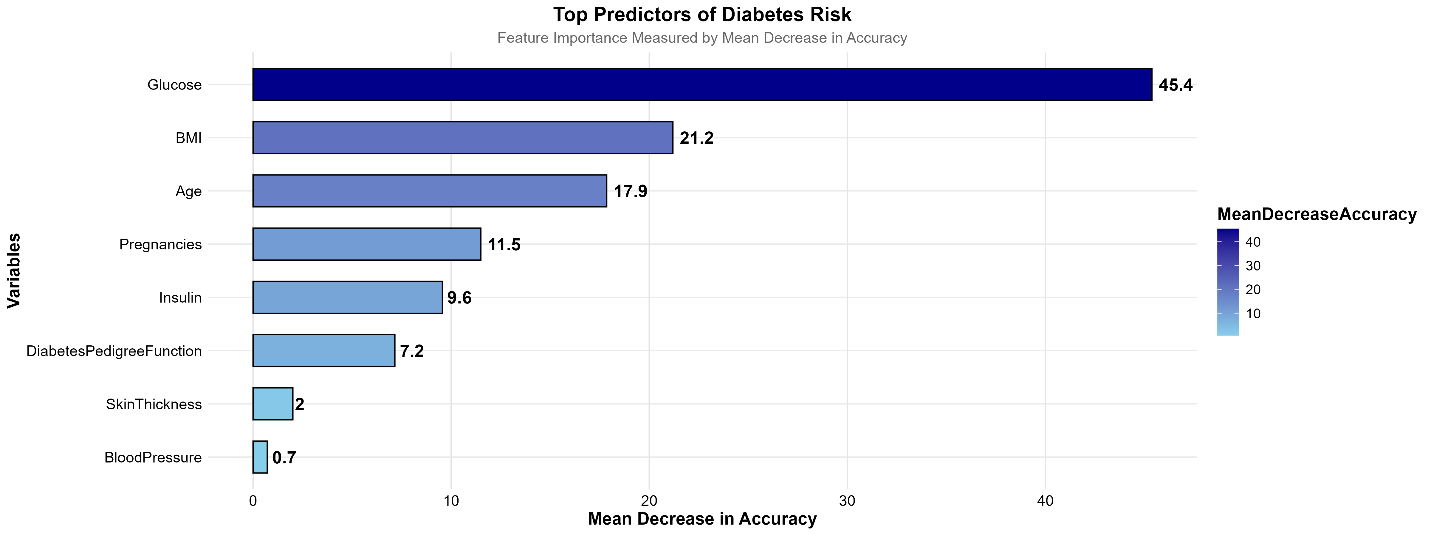


Figure: Feature Importance Measured by Mean Decrease in Accuracy

Beyond glucose, BMI (MDA: 21.19, Gini: 56.23) and Age (MDA: 17.85, Gini: 47.12) emerged as the next most important factors. BMI reflects obesity levels, a key contributor to insulin resistance, while age is associated with declining insulin sensitivity, making older individuals more susceptible to diabetes.

Other variables, including pregnancy history (MDA: 11.49), insulin levels (MDA: 9.56), and genetic predisposition (MDA: 7.15), had moderate importance, suggesting that while they influence diabetes risk, their predictive power is weaker than glucose, BMI, and age. Skin thickness (MDA: 2.00) and blood pressure (MDA: 0.71) were the least important predictors, indicating that these factors alone are not strong indicators of diabetes risk.

Overall, the findings emphasize the critical role of glucose monitoring, weight management, and aging-related risks in diabetes prevention. While genetics and insulin levels contribute to diabetes risk, lifestyle and metabolic factors play a more dominant role in prediction and prevention strategies.