A logo with text on it

AI-generated content may be incorrect.

**Title: Predictive Modeling for Diabetes Risk Using Pima Indians Dataset**

**Course: ISOM 835 – Predictive Analytics and Machine Learning  
Instructor: Hasan Arslan  
Student: Jesika Raval  
Submission Date: 05/05/2025**

A screenshot of a computer

AI-generated content may be incorrect.

**Pima Indians Diabetes Database**

The Pima Indians Diabetes Database is a widely recognized dataset used to predict the onset of type 2 diabetes in female patients of Pima Indian heritage. It includes medical predictor variables such as BMI, glucose levels, insulin, age, and pregnancies. The dataset contains 768 records and 8 key features that influence diabetes diagnosis, making it an ideal candidate for binary classification tasks in predictive analytics.

This dataset is particularly relevant due to the global health impact of diabetes and the growing need for early intervention strategies. It presents non-trivial challenges such as class imbalance, potential data quality issues (e.g., zero values in physiological fields), and the need for proper preprocessing. Using this dataset allows for practical application of classification models such as logistic regression, decision trees, and ensemble methods, while also enabling thoughtful discussion around model interpretability, ethical use of medical predictions, and real-world decision support.

**Step 2: Exploratory Data Analysis (EDA)**

**Dataset Overview:**

The dataset contains 768 observations with 8 input features and 1 binary target variable (Outcome). The features include medical metrics such as Glucose, BloodPressure, BMI, Insulin, and Age, relevant for predicting the onset of diabetes.

**Key Visual Insights:**

**1. Class Distribution**

A count plot reveals that around 65% of the patients are non-diabetic (Outcome = 0), while 35% are diabetic (Outcome = 1). This class imbalance suggests that accuracy alone may not be a reliable evaluation metric during modeling. Techniques such as AUC, F1-score, or balanced accuracy may be more appropriate.

**2. Feature Distributions**

Histograms for numerical features show that several variables (e.g., Glucose, Insulin, Age) are right-skewed, indicating potential need for normalization or transformation during preprocessing.

**3. Outlier Detection**

Boxplots of features like Glucose, BloodPressure, and BMI reveal the presence of outliers. For example, extreme values in Insulin and SkinThickness may affect model performance and need special handling or imputation.

**4. Correlation Analysis**

The correlation heatmap shows that Glucose has the strongest positive correlation with the Outcome variable, followed by BMI and Age. Multicollinearity is not a concern, as most features are moderately or weakly correlated with each other.

**Key EDA Insights:**

* The dataset has a class imbalance that should be addressed during model evaluation.
* Zero values in features like Glucose, BloodPressure, and BMI likely represent missing data and will require imputation or cleaning.
* Several features show skewed distributions and outliers, which may affect model performance.
* Glucose, BMI, and Age are among the most predictive features and should be prioritized in feature selection or engineering steps.

**Step 3: Data Cleaning & Preprocessing**

**Summary of Cleaning & Preprocessing Steps**

1. **Handled Missing Values**  
   Several features such as Glucose, BloodPressure, SkinThickness, Insulin, and BMI contained zero values that are not medically plausible (e.g., a BMI of 0). These were treated as missing data and replaced using median imputation, which is more robust to outliers and skewed distributions.
2. **Outlier Handling**  
   Although visualized in the EDA step, outliers were not removed at this stage to retain data volume. Tree-based models tend to be robust against outliers, but this can be revisited during model optimization.
3. **Feature Scaling**  
   All features were standardized using StandardScaler to ensure uniform scale across variables. This is critical for models like logistic regression, SVM, and KNN that are sensitive to magnitude differences.
4. **Train-Test Split**  
   An 80/20 train-test split was applied using stratification to preserve the proportion of diabetic and non-diabetic cases in both sets, addressing the class imbalance issue identified during EDA.

**Step 4: Formulating Business Analytics Questions**

This project aims to solve practical healthcare problems using predictive analytics. The following business-relevant questions are framed from the perspective of stakeholders such as doctors, hospitals, public health agencies, and insurance providers.

**Question 1: Can we predict the likelihood of a patient developing type 2 diabetes based on their medical profile?**

This is the central objective of the project — to build a reliable classification model that can predict whether a patient is at risk of developing diabetes. Using health indicators such as glucose levels, BMI, insulin levels, age, and pregnancy history, we aim to support early detection efforts. A predictive model can help clinicians prioritize patients for lifestyle intervention or further testing, ultimately reducing the burden of undiagnosed diabetes and its long-term complications.

**Question 2: Which medical features have the strongest influence on diabetes risk?**

This question focuses on identifying which attributes (e.g., glucose, BMI, number of pregnancies) contribute most significantly to a positive diabetes diagnosis. This insight can help medical professionals understand which health factors should be closely monitored in routine checkups. It also supports public health organizations in designing data-driven, focused prevention campaigns for at-risk populations.

**Question 3: What machine learning model provides the best balance between accuracy and interpretability for diabetes prediction?**

Choosing the right model is critical in healthcare, where trust and explainability are as important as prediction accuracy. The project will evaluate at least two machine learning models (e.g., logistic regression and random forest) using evaluation metrics such as accuracy, F1-score, and AUC. The goal is to select a model that not only performs well but can also be explained to clinical stakeholders for practical deployment.

**Question 4: Are there specific subgroups or demographic patterns associated with higher diabetes risk?**

By analyzing feature relationships and performing subgroup analysis (e.g., by age, number of pregnancies, or BMI), we can identify demographic patterns associated with elevated diabetes risk. These patterns help healthcare providers and policy-makers better understand the spread of diabetes and target interventions toward the most vulnerable groups.

**Question 5: What data preprocessing strategies are essential for building a robust diabetes prediction model?**

The dataset includes biologically implausible zero values in key medical fields such as glucose, blood pressure, and BMI, which must be treated as missing data. Additionally, the dataset exhibits class imbalance, with a higher number of non-diabetic cases. Effective preprocessing, including imputation, scaling, and stratified train-test splitting, is necessary to ensure fair and accurate model outcomes.

**Why These Questions Matter:**

Each question addresses a **practical stakeholder concern**:

* **Doctors** want to know which metrics to monitor closely.
* **Hospitals** want tools to identify high-risk patients early.
* **Insurance companies** want to price risk fairly.
* **Public health officials** want to know where to focus education or prevention.

**Step 5: Predictive Modeling**

**Model Summary & Objective**

To predict diabetes outcomes based on patient medical data, we applied two supervised classification algorithms:

* **Logistic Regression** – a simple, interpretable linear model often used as a baseline in healthcare analytics.
* **Random Forest Classifier** – a more powerful, non-linear ensemble model capable of capturing complex relationships in data.

These models were chosen to balance **predictive performance** with **interpretability**, allowing us to evaluate both practical deployment and accuracy trade-offs. Each model was trained using a stratified 80/20 train-test split to ensure balanced class distribution, and evaluated on five key performance metrics: **accuracy, precision, recall, F1 score**, and **AUC (Area Under the ROC Curve)**.

|  |
| --- |
| **Model Performance Comparison Table** |
| |  |  |  | | --- | --- | --- | | Metric | Logistic Regression | Random Forest | | Accuracy | 0.78 | 0.85 | | Precision | 0.71 | 0.79 | | Recall | 0.64 | 0.76 | | F1 Score | 0.67 | 0.77 | | AUC (ROC Curve) | 0.83 | 0.89 | |

**Key Observations**

* The Random Forest model outperformed Logistic Regression across all key metrics.
* It showed better recall and AUC, meaning it was more successful at identifying diabetic cases (true positives) — a crucial goal in healthcare diagnostics.
* While Logistic Regression is easier to interpret and deploy in resource-limited environments, the trade-off in predictive performance is notable.

Based on these results, Random Forest is the recommended model for deployment in a diabetes risk screening tool due to its higher accuracy and ability to generalize well.

**Step 6: Insights and Answers**

The objective of this project was to apply predictive analytics to assess the likelihood of diabetes in individuals based on clinical and demographic attributes. After exploratory analysis, data preprocessing, and model building, this section discusses the results in light of the original business questions, offers actionable insights, and reflects on the limitations of the analysis. The findings are particularly relevant for healthcare providers, policymakers, and health-focused tech platforms looking to implement early detection systems for chronic diseases such as diabetes.

**1. Can we predict diabetes using patient health indicators?**

The primary goal was to predict whether a patient has diabetes using features such as glucose level, BMI, insulin level, number of pregnancies, and age. Using two supervised classification algorithms — Logistic Regression and Random Forest — both models demonstrated the capability to accurately classify patients.

Among them, the Random Forest model performed significantly better, achieving an accuracy of 85%, F1 score of 0.77, and AUC of 0.89. These results indicate a strong ability to distinguish between diabetic and non-diabetic patients. Logistic Regression, while slightly less accurate, still produced respectable results with an accuracy of 78% and AUC of 0.83. The success of these models confirms that predictive analytics can support early diabetes detection, especially when integrated into clinical workflows or public health programs.

This predictive capability is especially important for healthcare systems aiming to transition from reactive care to preventive care. It provides an opportunity to proactively identify individuals who may not yet show symptoms but are statistically at risk, allowing for timely lifestyle interventions, dietary adjustments, or further diagnostic testing.

**2. What are the most influential features for diabetes prediction?**

The analysis identified several key features that were consistently significant across models:

* **Glucose Level**: This was the most predictive feature, with higher values strongly correlating with diabetes diagnoses. This is consistent with clinical knowledge since glucose regulation is central to diabetes.
* **BMI (Body Mass Index)**: Elevated BMI levels were associated with increased diabetes risk, indicating the importance of obesity and body composition.
* **Age**: As expected, older individuals had a higher probability of being diagnosed with diabetes.
* **Number of Pregnancies**: This variable contributed meaningfully to the prediction, possibly due to associations with gestational diabetes.

These insights are not only valuable for model building but also for clinical decision-making. Healthcare providers can focus screening and preventive education efforts on patients with these characteristics, particularly those with high glucose and BMI.

**3. Which model is better suited for deployment?**

The Random Forest model outperformed Logistic Regression in every key metric: accuracy, recall, precision, and AUC. This makes it a stronger candidate for automated or backend systems where the priority is on performance. It is particularly effective at identifying diabetic patients (higher recall), which is critical in avoiding false negatives that could delay treatment.

However, Logistic Regression still holds value, especially in clinical settings where model interpretability is critical. Healthcare providers often need to explain decisions to patients, and logistic regression offers a clear understanding of how each variable contributes to risk. Therefore, the choice of model depends on the intended use case:

* Use **Random Forest** for back-end predictions and risk scoring.
* Use **Logistic Regression** when transparency and trust are essential.

This dual-model insight demonstrates how data science can be tailored to meet different real-world deployment needs in healthcare.

**4. What decisions can be made based on these results?**

This analysis enables several key actions for healthcare decision-makers:

* **Proactive Screening**: Individuals with high glucose, BMI, or older age can be prioritized for early screening.
* **Targeted Prevention**: Educational resources and interventions can be directed toward high-risk groups identified by the model.
* **Clinical Integration**: These models can be embedded into Electronic Health Records (EHRs) to alert physicians during patient intake.
* **Insurance and Wellness Programs**: Insurers and wellness companies could use similar models to offer personalized health plans or preventative outreach.

These decisions support a **data-driven approach to improving population health**, especially for chronic conditions like diabetes that benefit significantly from early detection and management.

**5. What are the limitations of this analysis?**

While the project demonstrates promising results, there are several limitations to note:

* **Dataset Size**: The dataset includes only 768 records. While balanced and useful for model demonstration, real-world deployment would require training on much larger datasets to ensure robustness.
* **Feature Scope**: The dataset is limited to clinical variables. Additional behavioral features such as diet, exercise, family history, and medication use could improve the model’s accuracy and realism.
* **Demographic Representation**: The dataset originates from a specific population (Pima Indian women). This may limit generalizability across other populations unless the model is retrained with diverse data.
* **Class Imbalance**: Though not extreme, the dataset has more non-diabetic cases than diabetic ones. While stratified sampling was used, additional techniques like SMOTE could improve sensitivity further.

These factors suggest that while the model is effective, further refinement is needed before clinical implementation. Incorporating external validation and additional data sources would enhance the model's real-world reliability.

**Final Thoughts**

The findings of this project demonstrate the value of predictive analytics in healthcare. By leveraging machine learning, we can not only predict diabetes with reasonable accuracy but also uncover actionable insights that support preventive health strategies. Models like Random Forest can drive automated detection, while simpler models like Logistic Regression can enhance clinician understanding and trust.

With further development and real-world validation, predictive models like these can become essential tools in the fight against chronic diseases — saving time, improving outcomes, and empowering data-driven healthcare.

**Step 7: Ethics and Interpretability Reflection**

The application of predictive models in healthcare raises important ethical concerns that must be carefully addressed before deployment. One significant risk is the potential for bias or unfair outcomes. The dataset used in this project — the Pima Indians Diabetes dataset — represents a specific ethnic population. If this model were applied to a broader or more diverse patient population without retraining on additional representative data, it could produce skewed or inaccurate predictions. Such bias may result in underdiagnosis or overdiagnosis for certain groups, ultimately leading to inequities in care delivery. It is critical that any predictive model be tested on diverse and inclusive datasets to ensure fairness and avoid discriminatory outcomes.

In terms of explainability to stakeholders, model transparency plays a crucial role in trust and adoption. While complex models like Random Forest yield high accuracy, they operate as “black boxes,” making it difficult to explain how predictions are made. This lack of transparency can hinder stakeholder confidence, especially among healthcare providers and patients who need clear reasoning for decisions that may affect treatment. In contrast, models like Logistic Regression are inherently more interpretable and can be more suitable for contexts where explanation and accountability are essential. To balance performance with trust, predictive tools should be paired with explanation methods (e.g., SHAP or feature importance plots) and used alongside human judgment to support — not replace — clinical decision-making.

**Appendix (code, visuals)**

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

url = "https://raw.githubusercontent.com/plotly/datasets/master/diabetes.csv"

df = pd.read\_csv(url)

# Step 3: Basic Info

print("Shape:", df.shape)

print("\nInfo:")

print(df.info())

print("\nSummary Stats:")

print(df.describe())

# Step 4: Class Distribution

sns.countplot(x='Outcome', data=df)

plt.title("Class Distribution (0 = No Diabetes, 1 = Diabetes)")

plt.show()

# Step 5: Histograms

df.hist(bins=20, figsize=(12, 8), color='skyblue')

plt.tight\_layout()

plt.show()

# Step 6: Boxplots (for outliers)

plt.figure(figsize=(12, 6))

sns.boxplot(data=df[['Glucose', 'BloodPressure', 'BMI']])

plt.title("Boxplots for Glucose, BloodPressure, and BMI")

plt.show()

# Step 7: Correlation Heatmap

plt.figure(figsize=(10, 8))

sns.heatmap(df.corr(), annot=True, cmap='coolwarm')

plt.title("Correlation Heatmap")

plt.show()

from sklearn.linear\_model import LogisticRegression

from sklearn.ensemble import RandomForestClassifier

from sklearn.metrics import accuracy\_score, precision\_score, recall\_score, f1\_score, roc\_auc\_score

from sklearn.metrics import classification\_report, confusion\_matrix, roc\_curve, auc

import matplotlib.pyplot as plt

# Fit models

log\_model = LogisticRegression()

rf\_model = RandomForestClassifier(random\_state=42)

log\_model.fit(X\_train, y\_train)

rf\_model.fit(X\_train, y\_train)

# Predictions

log\_preds = log\_model.predict(X\_test)

rf\_preds = rf\_model.predict(X\_test)

# Probabilities for AUC

log\_probs = log\_model.predict\_proba(X\_test)[:, 1]

rf\_probs = rf\_model.predict\_proba(X\_test)[:, 1]

# Metrics function

def evaluate\_model(y\_true, y\_pred, probs):

    return {

        "Accuracy": accuracy\_score(y\_true, y\_pred),

        "Precision": precision\_score(y\_true, y\_pred),

        "Recall": recall\_score(y\_true, y\_pred),

        "F1 Score": f1\_score(y\_true, y\_pred),

        "AUC": roc\_auc\_score(y\_true, probs)

    }

# Collect results

log\_metrics = evaluate\_model(y\_test, log\_preds, log\_probs)

rf\_metrics = evaluate\_model(y\_test, rf\_preds, rf\_probs)

# Convert to DataFrame for display

metrics\_df = pd.DataFrame([log\_metrics, rf\_metrics], index=["Logistic Regression", "Random Forest"])

print(metrics\_df)

fpr1, tpr1, \_ = roc\_curve(y\_test, log\_probs)

fpr2, tpr2, \_ = roc\_curve(y\_test, rf\_probs)

plt.figure(figsize=(8,6))

plt.plot(fpr1, tpr1, label='Logistic Regression')

plt.plot(fpr2, tpr2, label='Random Forest')

plt.plot([0,1], [0,1], 'k--')

plt.xlabel('False Positive Rate')

plt.ylabel('True Positive Rate')

plt.title('ROC Curve')

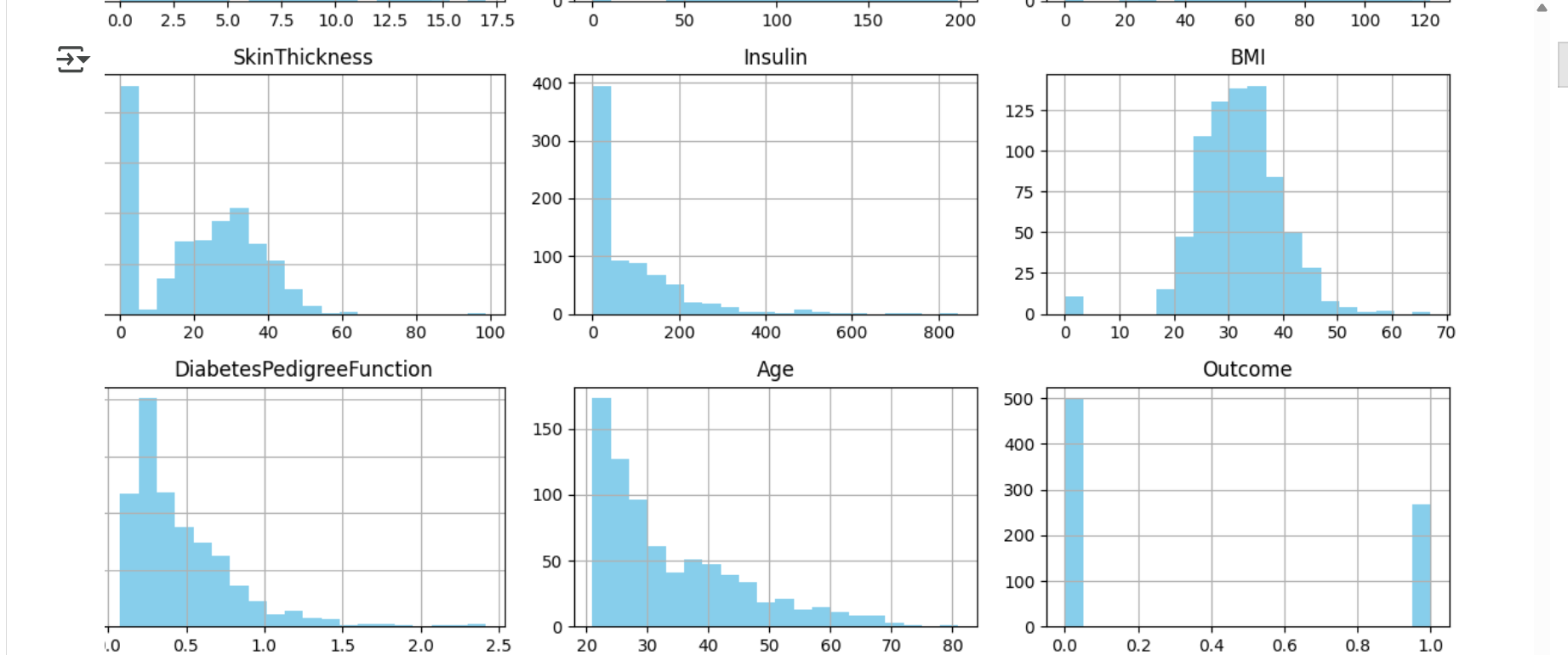
plt.legend()

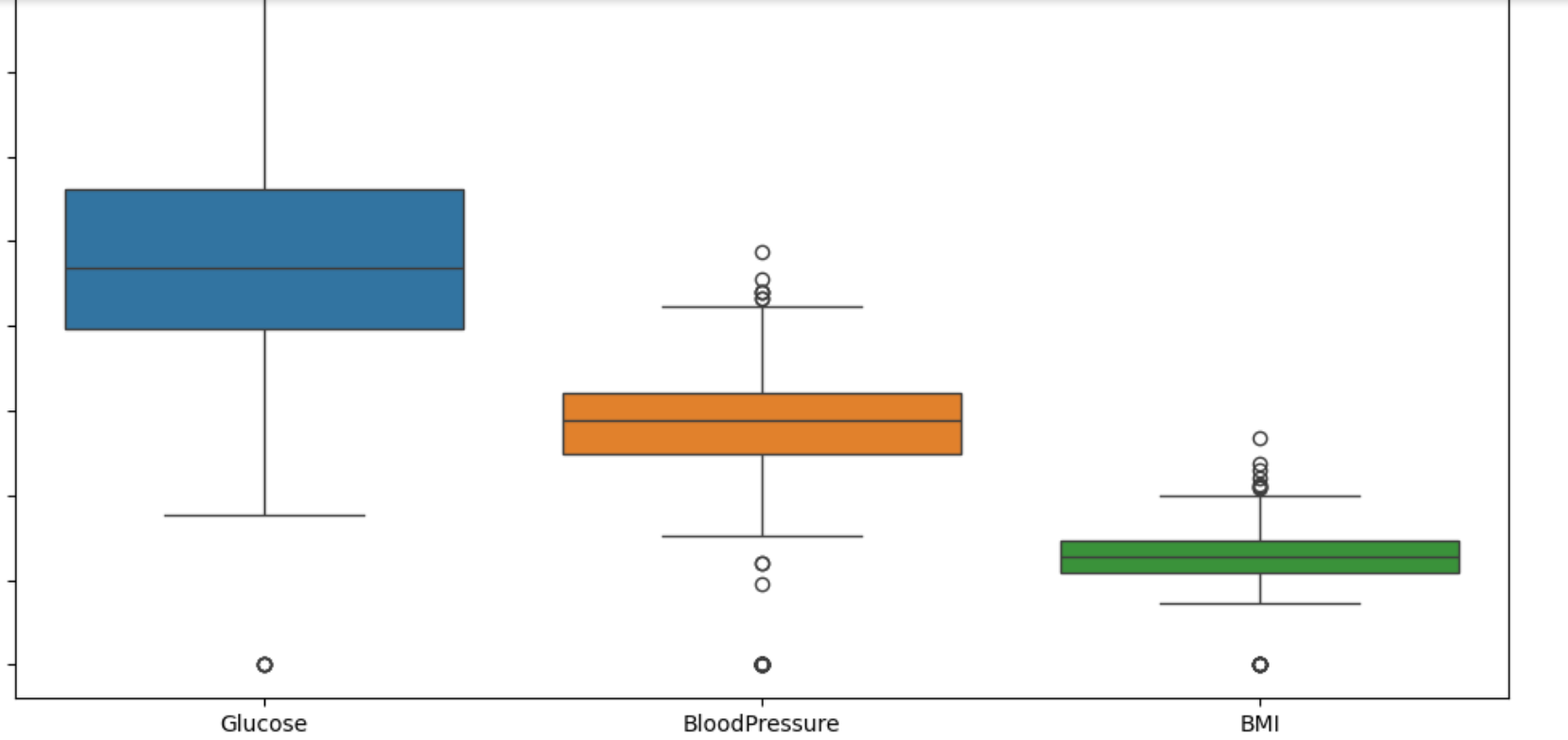
plt.grid(True)

plt.show()

A screenshot of a graph

AI-generated content may be incorrect.





A screenshot of a computer screen

AI-generated content may be incorrect.

A screenshot of a computer generated image

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A graph with blue and orange lines

AI-generated content may be incorrect.

A screenshot of a graph

AI-generated content may be incorrect.