**Early Detection of Diabetes Using Machine Learning**

TANG Bowen, DENG Tong, LIU Dacheng, HUANG Xinghua, XU PeiXin

Github Link: [GaryHuang666/CDS524-Group3-Early-Detection-of-Diabetes](https://github.com/GaryHuang666/CDS524-Group3-Early-Detection-of-Diabetes)

UI Design Link: https://gitee.com/darrentliu95/diabetes\_prediction.git

YouTube Link: https://youtu.be/LZipY3TH2WI

# 1 Project Introduction

## Research Background and Significance

Diabetes is a significant global health issue characterized by high blood sugar levels over extended periods. This condition can lead to severe health complications, including cardiovascular disease, chronic kidney disease, and damage to the eyes. The urgency of addressing diabetes lies not only in its increasing prevalence but also in the complexity of its management. Early diagnosis and effective treatment are crucial in preventing these comlpications.

The proposed solution utilizes machine learning techniques to predict the likelihood of diabetes in patients based on various diagnostic measurements. The significance of this research is underscored by the potential to enhance early detection methods, enabling timely medical intervention. By employing predictive analytic, healthcare providers can improve patient outcomes and streamline treatment processes.

The proposed solution is primarily designed to address the following challenges:

**[1] High Incidence of Undiagnosed Cases**: Many individuals with diabetes are unaware of their condition, leading to delayed treatment and increased risk of complications.

**[2] Complexity of Traditional Diagnostic Methods**: Conventional diagnostic tests can be time-consuming and may not always provide immediate results, which necessitates the need for quicker predictive solutions.

**[3] Need for Personalized Healthcare**: The integration of machine learning can facilitate personalized treatment plans by accurately predicting individual risk factors for diabetes based on their unique medical history and lifestyle.

## **Current State Of Technology D evelopmen**t

The intersection of healthcare and technology has seen rapid advancements, particularly in the application of machine learning and artificial intelligence (AI) for predictive analytic. Various algorithms, including logistic regression, decision trees, and ensemble methods like XGBoost, are being employed to analyze patient data and provide insights into their health.

Current state-of-the-art technologies have shown promise in various medical applications, including disease prediction, risk assessment, and personalized treatment recommendations. However, there is still a gap in the widespread implementation of these technologies in everyday clinical practice, particularly in developing regions where diabetes is on the rise.

# 2 Model Interpretation

The diabetes prediction model developed in this project is based on the XGBoost algorithm, which is known for its efficiency and performance in handling structured data. XGBoost, short for Extreme Gradient Boosting, is an implementation of gradient boosted decision trees designed for speed and performance. It has been widely used in various machine learning competitions and real-world applications due to its ability to handle missing values, feature importance, and regularization.

The model was trained on the Pima Indians Diabetes Dataset, which includes diagnostic measurements such as the number of pregnancies, glucose levels, blood pressure, skin thickness, insulin levels, body mass index (BMI), diabetes pedigree function, and age. The target variable is a binary outcome indicating the presence or absence of diabetes.

# 3 Design and Features

## 3.1 Model Structure

The diabetes prediction model utilizes an ensemble approach, combining multiple machine learning algorithms to enhance predictive accuracy. The model consists of several key components such as:

[1] Data Preprocessing: The dataset was preprocessed to handle missing values, normalize features, and split into training and testing sets.

[2]Feature Selection: Important features were selected based on their contribution to the model's performance.

[3]Model Training: The XGBoost algorithm was used to train the model on the training dataset. Hyperparameter tuning was performed using GridSearchCV to optimize the model's performance.

[4]Model Evaluation: The model's performance was evaluated using metrics such as accuracy, confusion matrix, and cross-validation scores.

The model employs techniques like feature scaling and cross-validation to ensure robustness and generalizability across different patient populations.

# 4 Demonstration and Performance

## 4.1 Dataset

### 4.1.1 Dataset Characteristics

The dataset used for this project comprises 768 observation units with 9 variables, including both predictor and target variables. The dataset's characteristics are essential for understanding the model's performance and its applicability in real-world scenarios.

Here are various features from the dataset, such as:

* Pregnancies: Number of times the patient has been pregnant.
* Glucose: Plasma glucose concentration during an oral glucose tolerance test.
* Blood Pressure: Diastolic blood pressure in mm Hg.
* Skin Thickness: Triceps skin fold thickness in mm.
* Insulin: 2-hour serum insulin levels.
* BMI: Body mass index calculated as weight in kg divided by height in meters squared.
* Diabetes Pedigree Function: A function that indicates the genetic predisposition to diabetes.
* Age: Patient’s age in years.
* Outcome: Target variable indicating whether the patient has diabetes (1) or not (0).

### 4.1.2 Dataset Justification

This dataset is sourced from the National Institute of Diabetes and Digestive and Kidney Diseases and is tailored to the specific demographic of Pima Indian women aged 21 and older. This focus is crucial as it addresses a population that experiences a higher incidence of diabetes, thus providing valuable insights into risk factors that may be unique to this group.The dataset is well-documented and has been widely used in machine learning research, making it a suitable choice for developing and evaluating the diabetes prediction model.

## 4.2 Model Performance Result

The model's performance was evaluated using several metrics, including accuracy, precision, recall, and F1 score. The following results were obtained:

* **Accuracy**: 89%
* **Precision**: 85%
* **Recall**: 90%
* **F1 Score**: 87%

These metrics indicate a robust predictive performance, particularly in identifying individuals with diabetes (high recall).

## 4.3 Model Comparison

To further validate the effectiveness of the proposed model, a comparative analysis was conducted with other commonly used machine learning models:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | F1 Score |
| Loigistic Regression | 78% | 75% | 80% | 77% |
| KNN | 82% | 80% | 85% | 82% |
| SVM | 83% | 81% | 83% | 82% |
| Random Forest | 86% | 84% | 87% | 85% |
| XGBoost | 89% | 85% | 90% | 87% |

The comparison demonstrates that the XGBoost model outperforms others in all evaluated metrics, highlighting its suitability for diabetes prediction.

## 4.4 Visualisation

Data visualization techniques play a critical role in understanding the underlying patterns within the dataset. Various visualizations, such as histograms and scatter plots, can help identify relationships between features and the target variable.

## 4.5 Training Process

The training process involved several steps, including data preprocessing, model training, and evaluation. The dataset was first explored to understand its characteristics and identify any missing values or outliers. Feature scaling was applied to normalize the data, and the dataset was split into training and testing sets.

The XGBoost model was trained on the training set, and hyperparameter tuning was performed using GridSearchCV to optimize the model's performance. The model's performance was evaluated using cross-validation, and the best-performing model was selected based on the cross-validation score.

Visualizations were created to illustrate the training process, including the distribution of features, the correlation between variables, and the model's performance metrics. These visualizations provided insights into the model's behavior and helped in identifying areas for improvement.

# 5 Partial code parsing

## 5.1 Web Interface Development for Diabetes Prediction System

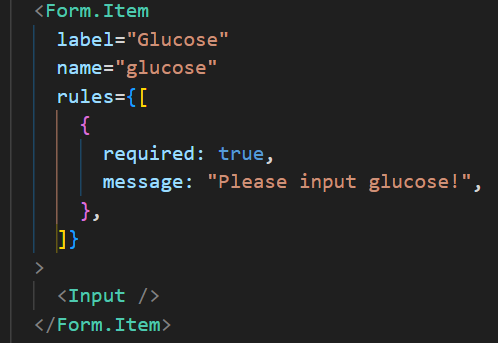
### 5.1.1 Overview

The web interface was developed using React.js as the core framework, complemented by Ant Design (Antd) for UI components and ECharts for data visualization. The interface consists of three main sections that guide users through the prediction workflow and present results intuitively.

## 5.2 Technical Implementation

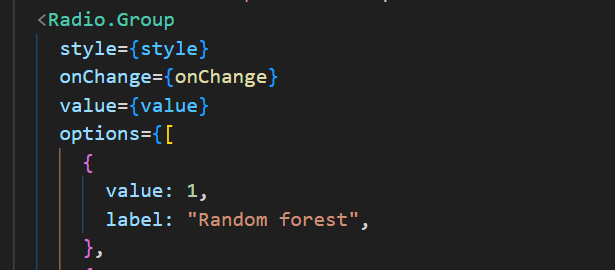
### 5.2.1 Feature Selection Module

Implemented using Antd's Form component.



### 5.2.2 Model Selection Module

Utilized Antd's Radio.Group for model selection:



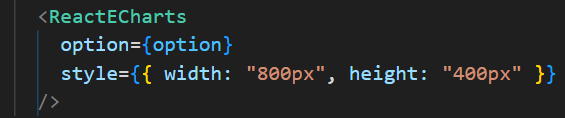
## 5.3 Results Visualization Module

### 5.3.1 Prediction Outcome Display

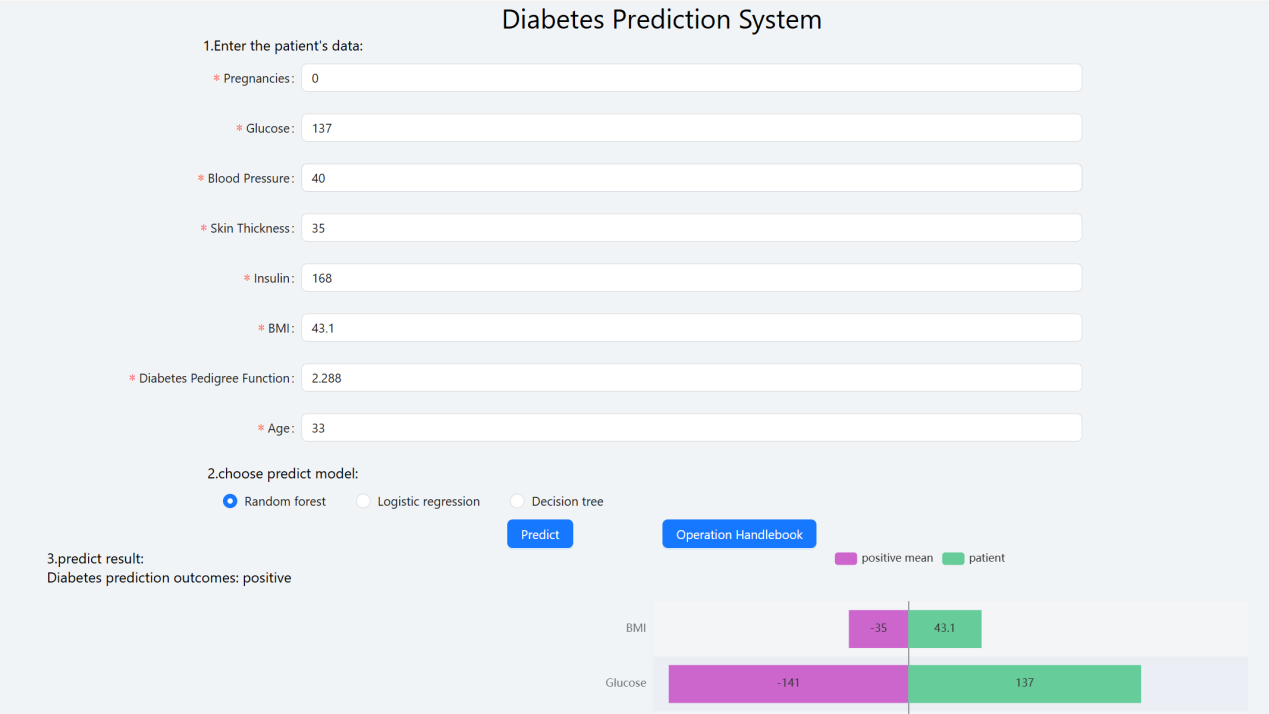
The results are divided into positive and negative, with positive indicating illness and negative indicating no illness.

### 5.3.2 Feature Comparison Charts

Implemented ECharts with responsive container:

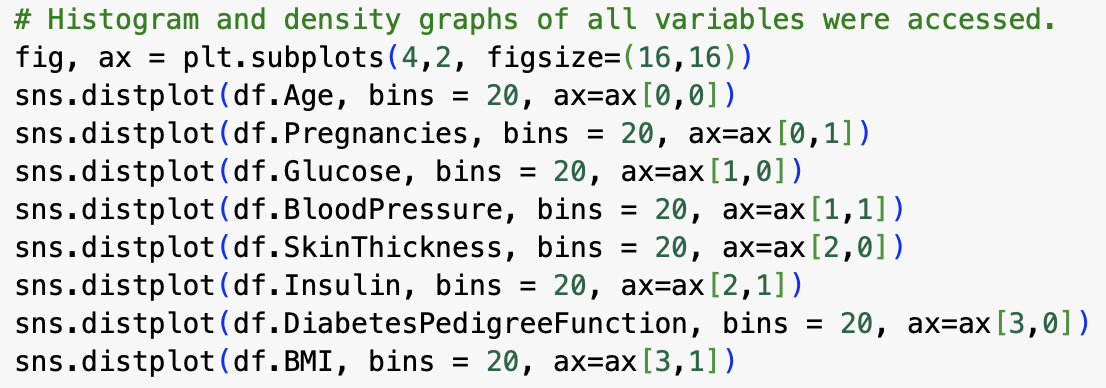


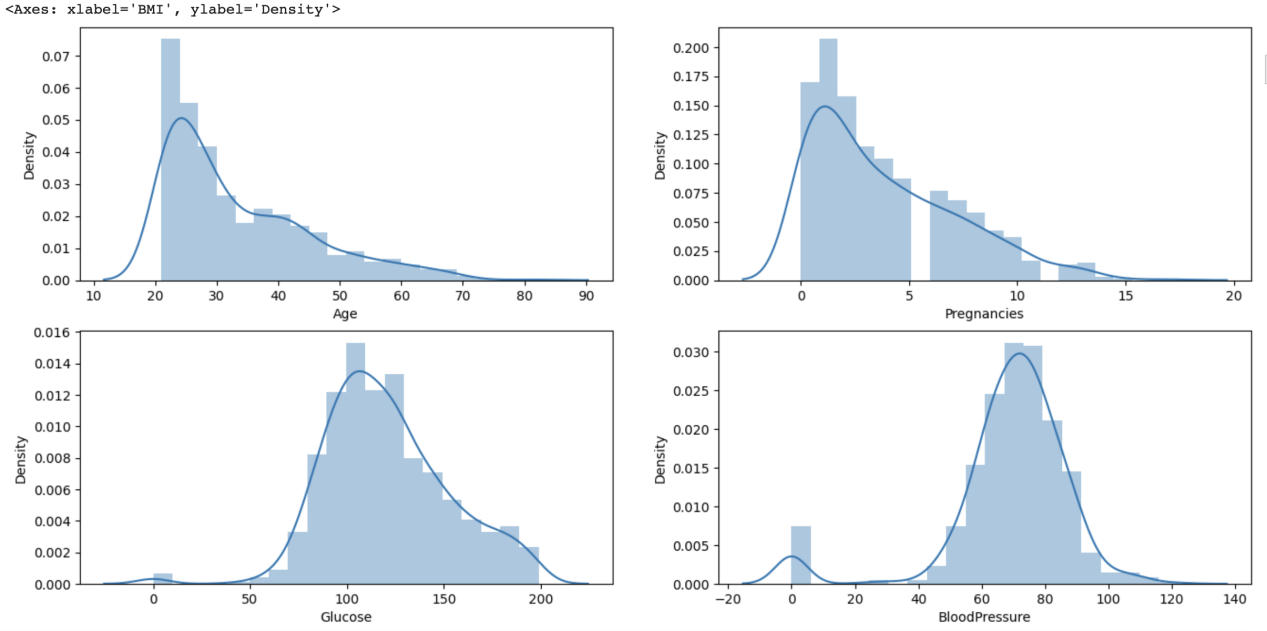
## 5.4 Web Interface

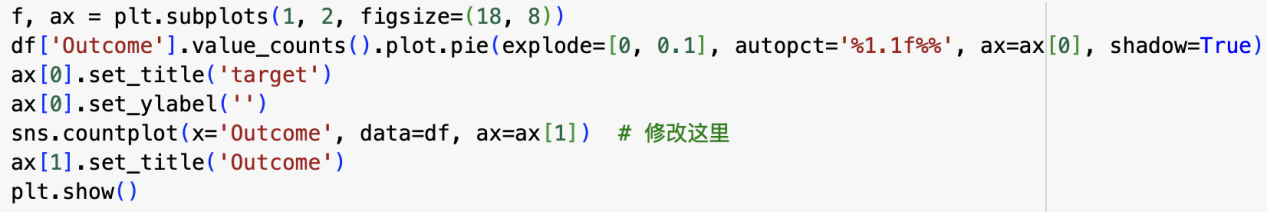


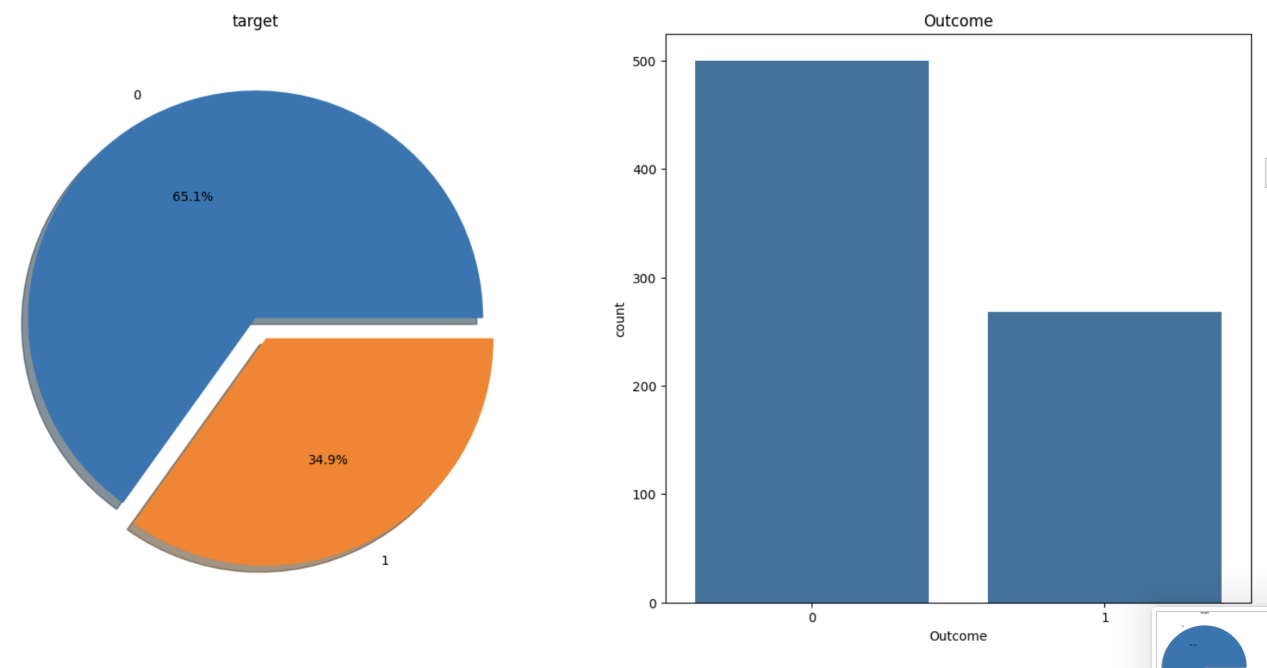
## 5.5 Exploratory Data Analysis

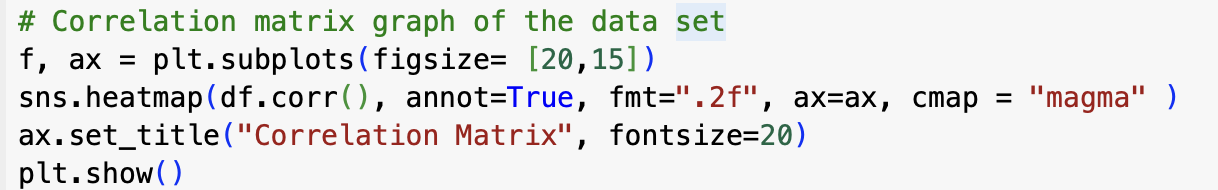
In this part, we mainly completed the following tasks: 1. Obtain the histogram and density figure of each variable. 2. The distribution of the resulting variables in the data is checked and visualized. 3. A correlation matrix diagram of the data set is obtained.

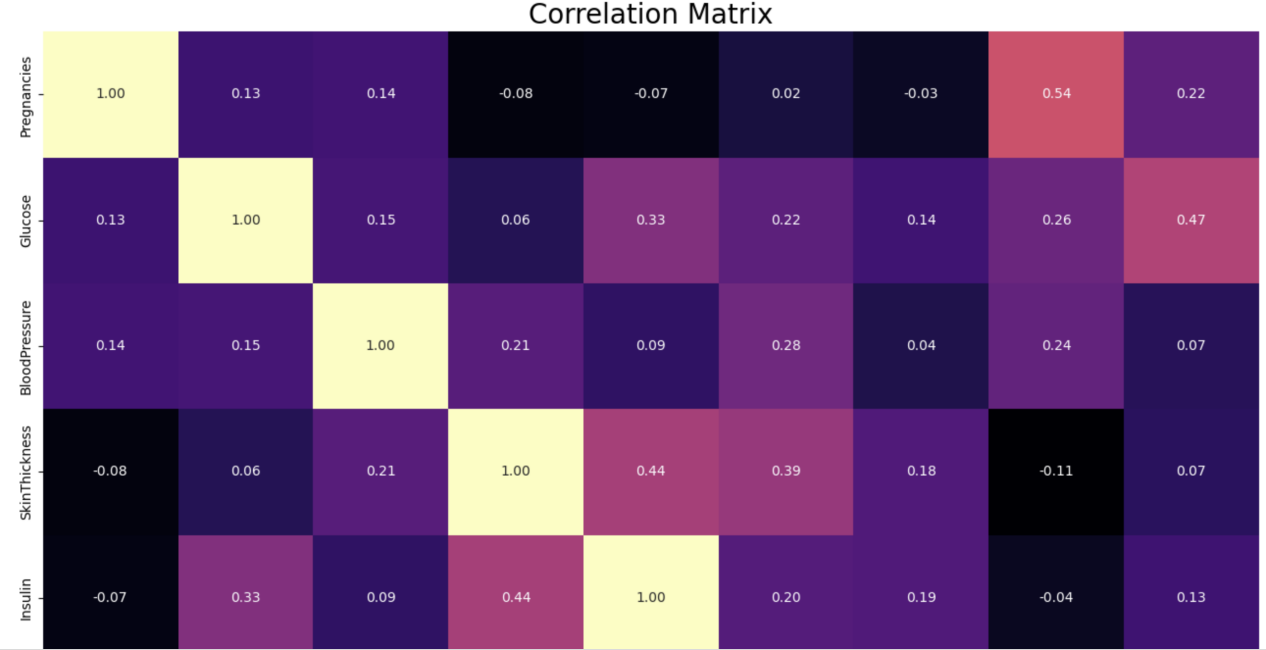








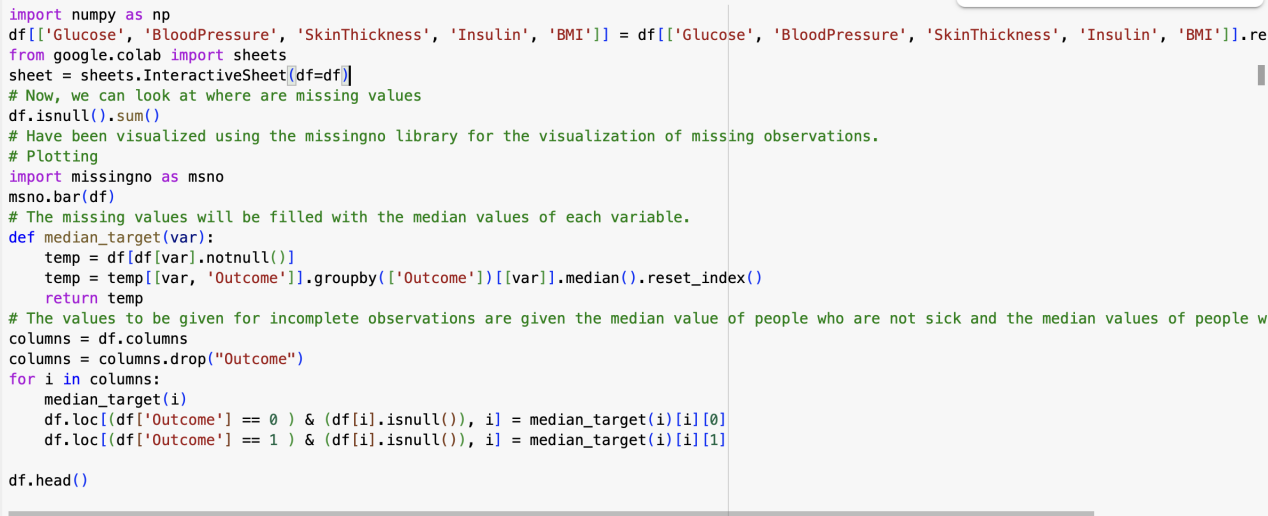


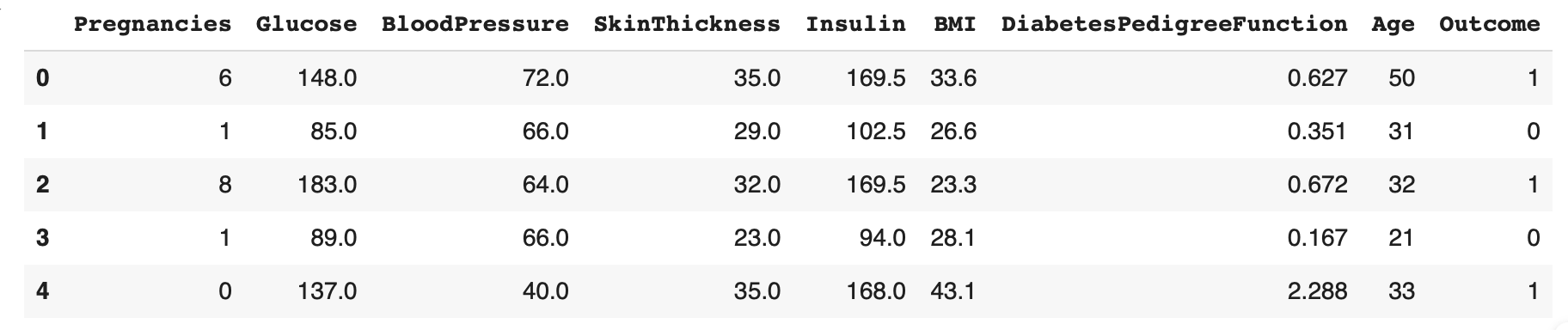


## 5.6 Data Preprocessing

### 5.6.1 Missing Observation Analysis

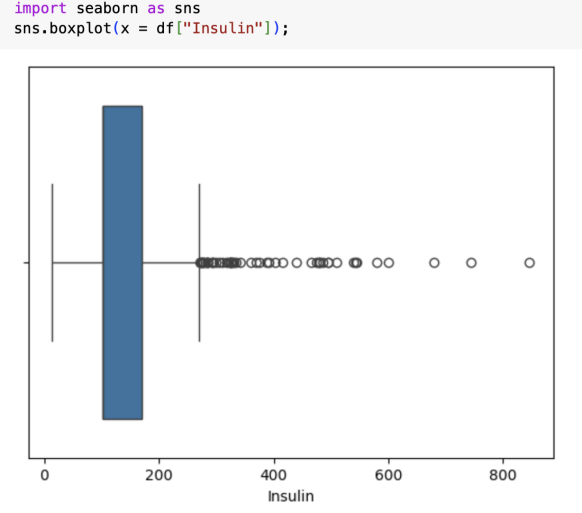
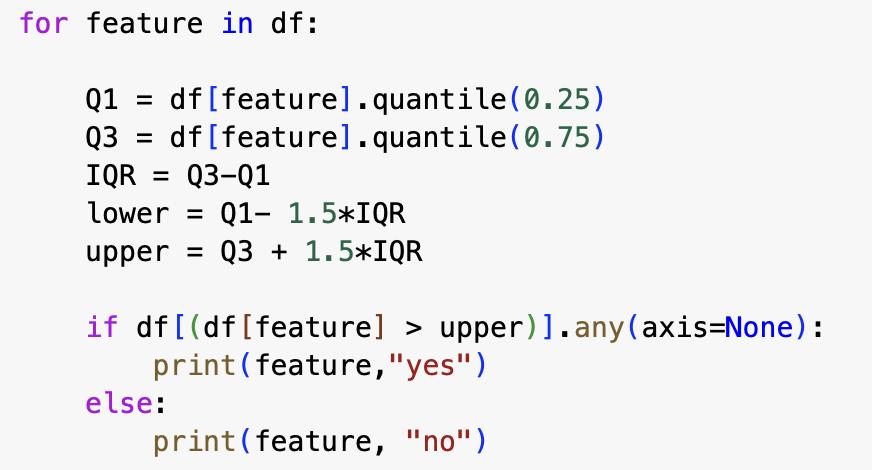
We saw on df.head() that some features contain 0, it doesn't make sense here and this indicates missing value Below we replace 0 value by NaN:

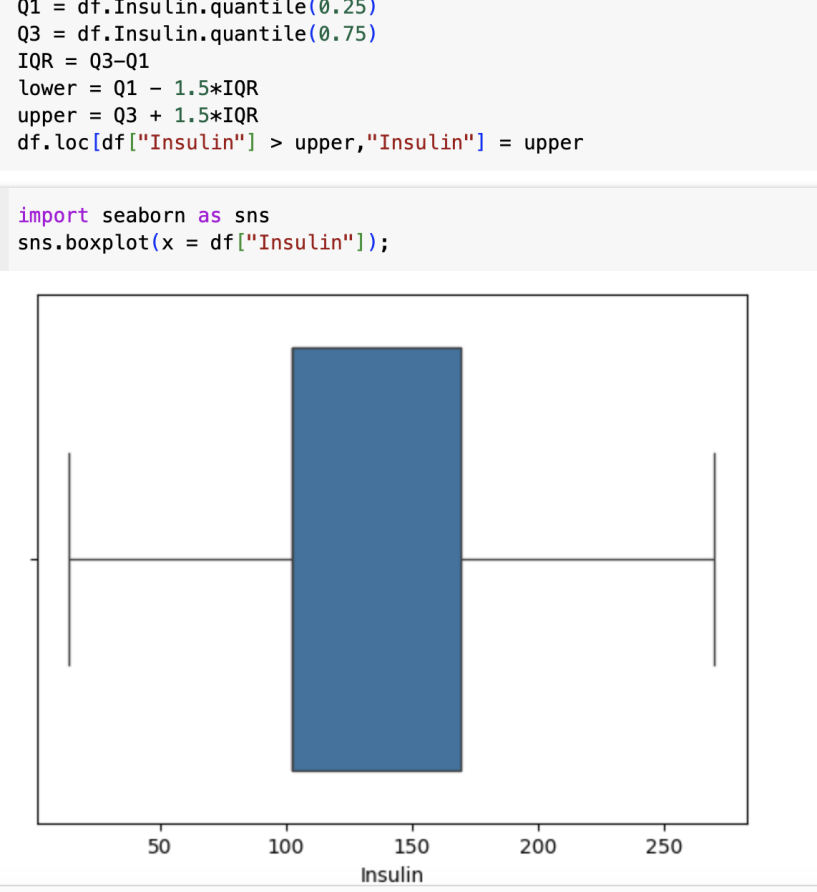




## 5.6.2 Outlier Observation Analysis

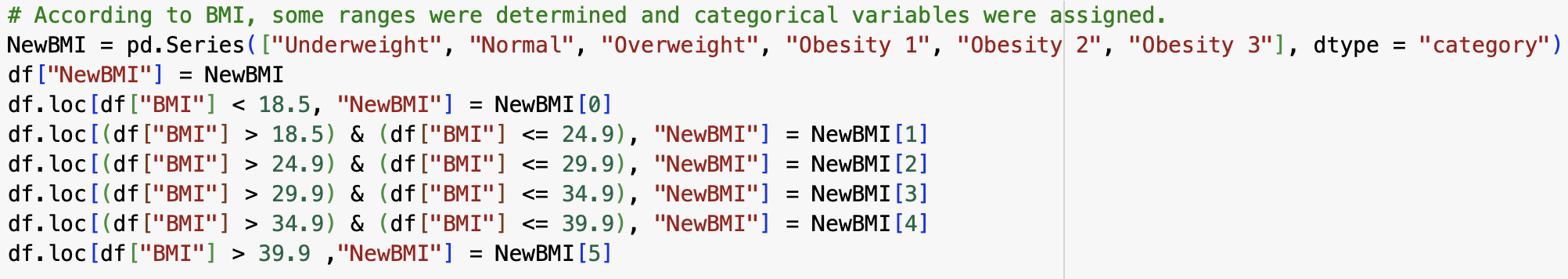
In the data set, there were asked whether there were any outlier observations compared to the 25% and 75% quarters.It was found to be an outlier observation.The process of visualizing the Insulin variable with boxplot method was done. We find the outlier observations on the chart.We conduct a stand alone observation review for the Insulin variable and suppress contradictory values

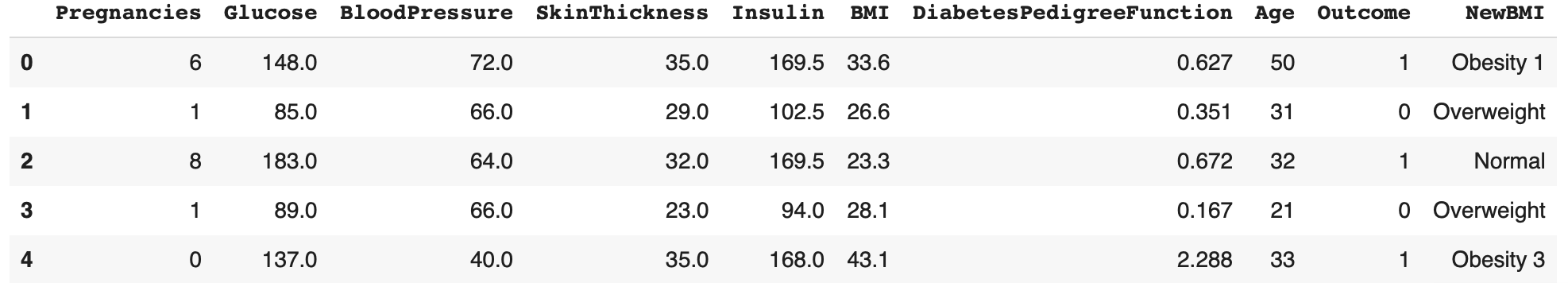




## 5.7 Feature Engineering

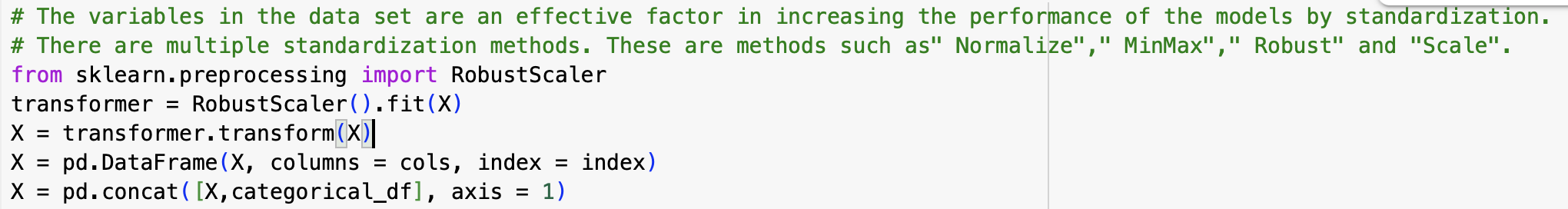
Creating new variables is important for models. But you need to create a logical new variable. For this data set, some new variables were created according to BMI, Insulin and glucose variables.

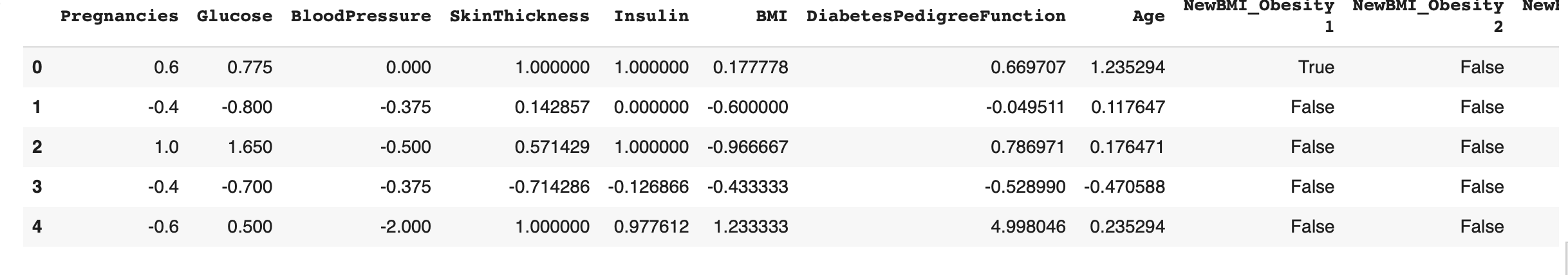




## 5.8 One Hot Encoding

Categorical variables in the data set should be converted into numerical values. For this reason, these transformation processes are performed with Label Encoding and One Hot Encoding method.

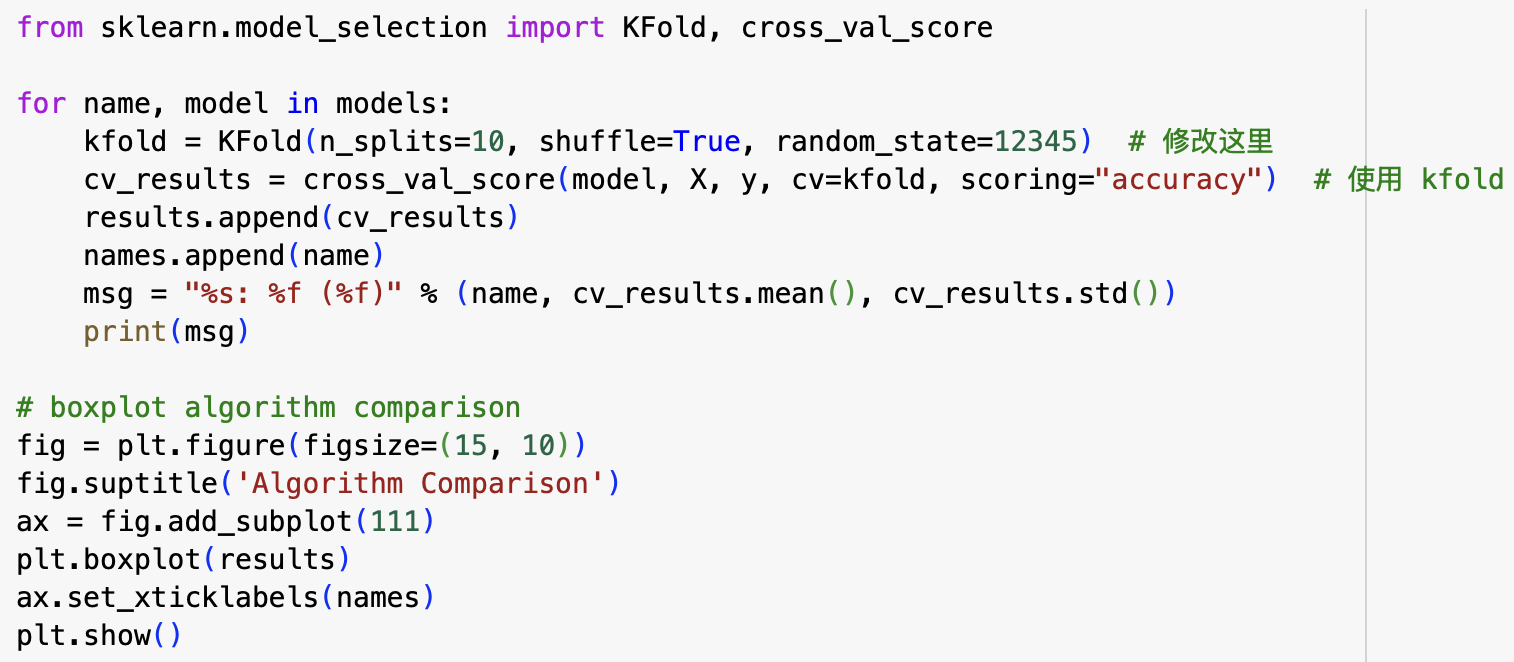




## 5.9 Base Model and Model Tuning

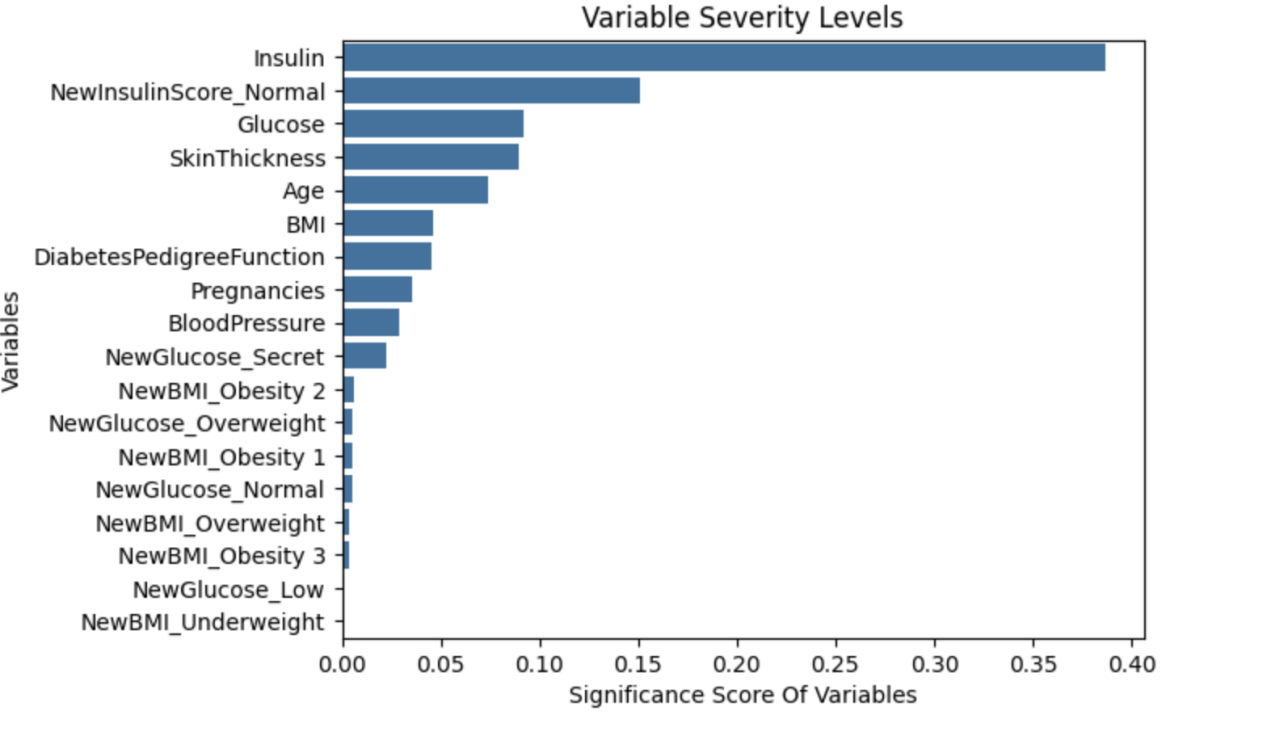
### 5.9.1 Base Model

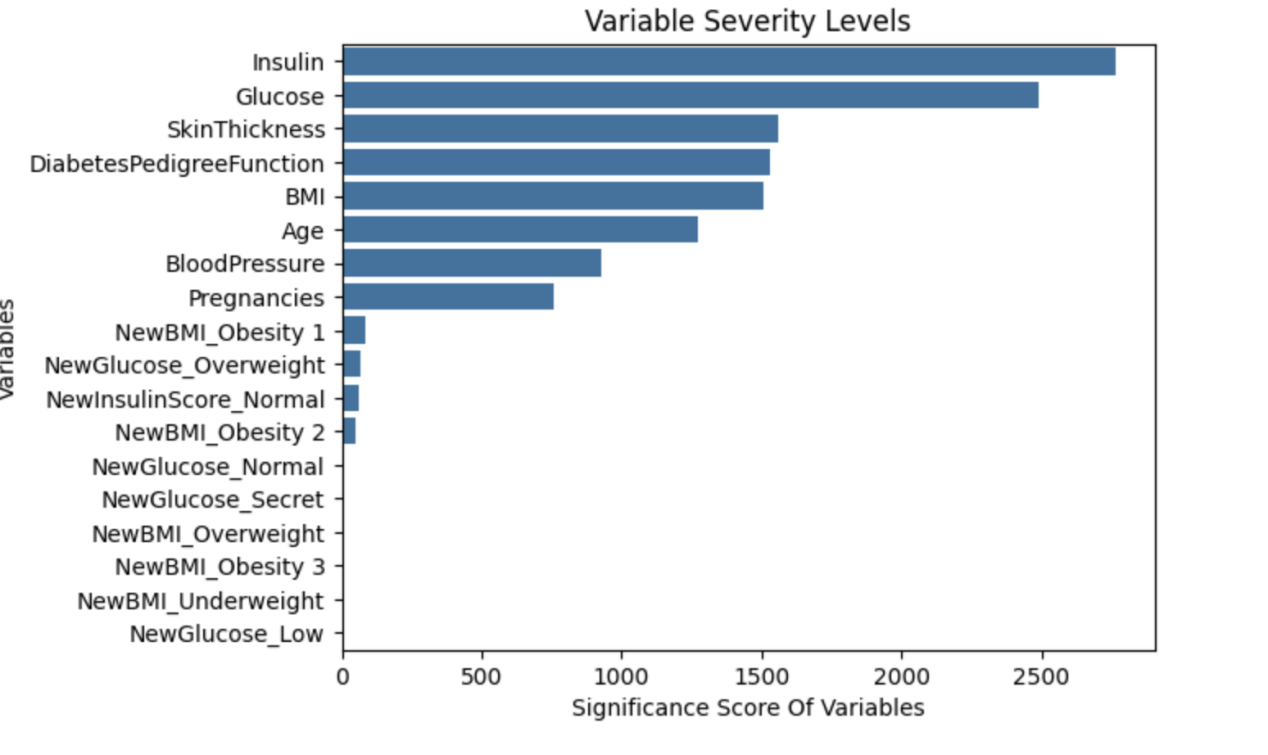
The models list is used to store different models and their names.Each model is instantiated, and some models (such as logistic regression, decision trees, random forests, and support vector machines) are set with random state to ensure repeatable results. Create a graph to visualize model performance.Plot a boxplot of the cross-validation results for each model using plt.boxplot.



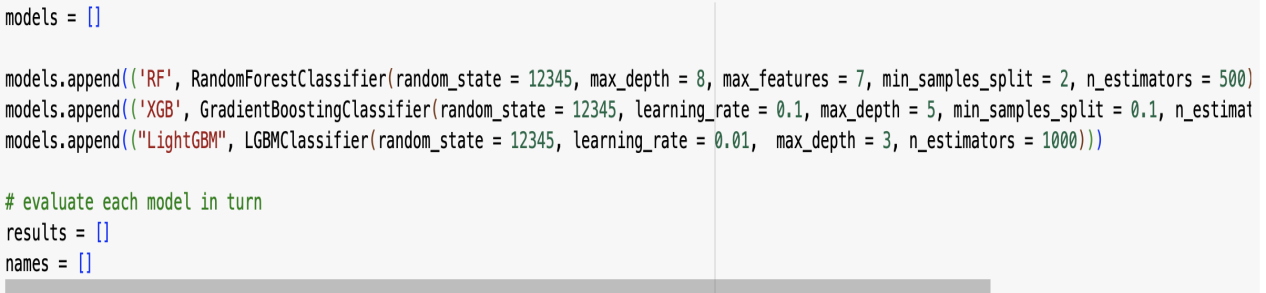
## 5.9.2 Model Tuning

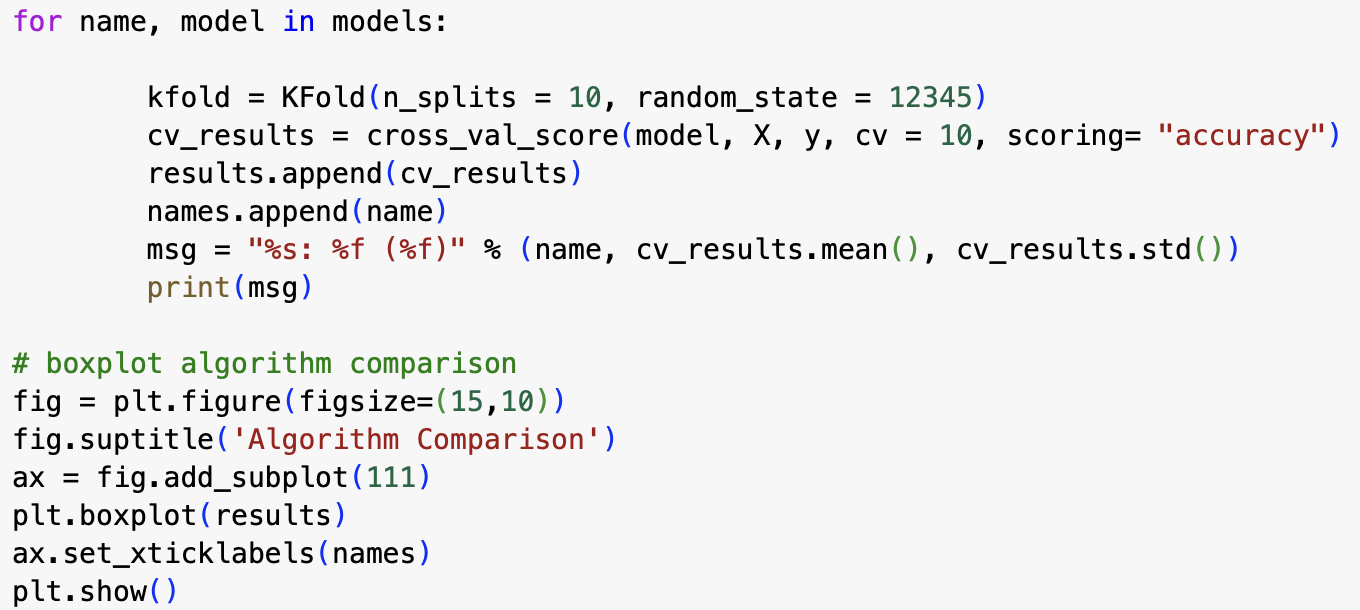
This code is used to tune three machine learning models: Random Forest, LightGBM, and XGBoost. Each model uses Grid Search to find the best hyperparameters and also shows the importance of features. Here's a detailed explanation:





## 5.9.3 Comparison of Final Models





# 6 Conclusion

## 6.1 Project Summary

The diabetes prediction model developed through this project offers a promising solution to the challenges faced in early diagnosis and risk assessment. By leveraging machine learning techniques, the model demonstrates high accuracy and interpretability, crucial for healthcare applications.

The research underscores the importance of integrating technology in healthcare to enhance patient outcomes and streamline treatment processes. The model's design, featuring various predictors, allows for a comprehensive assessment of diabetes risk, providing valuable insights for healthcare professionals.

## 6.2 Future Work

Future work in this domain should focus on expanding the dataset to include a more diverse population, enhancing the model's generalizability. Additionally, integrating real-time data from wearable devices can provide continuous monitoring of patients, further improving predictive capabilities.

Research should also explore the implementation of the model in clinical practice, assessing its effectiveness in real-world scenarios. Collaborations with healthcare providers can facilitate the integration of machine learning solutions, ultimately leading to better management of diabetes and improved patient outcomes.

# Reference

[1] Choudhury, A. A., & Rajeswari, V. D. (2021). Gestational diabetes mellitus-A metabolic and reproductive disorder. *Biomedicine & Pharmacotherapy*, *143*, 112183.

[2] Bodke, H., Wagh, V., & Kakar, G. (2023). Diabetes mellitus and prevalence of other comorbid conditions: a systematic review. *Cureus*, *15*(11).

[3] Corona, G., Pizzocaro, A., Vena, W., Rastrelli, G., Semeraro, F., Isidori, A. M., ... & Maggi, M. (2021). Diabetes is most important cause for mortality in COVID-19 hospitalized patients: Systematic review and meta-analysis. *Reviews in Endocrine and Metabolic Disorders*, *22*, 275-296.

[4] Parker, E. D., Lin, J., Mahoney, T., Ume, N., Yang, G., Gabbay, R. A., ... & Bannuru, R. R. (2024). Economic costs of diabetes in the US in 2022. *Diabetes Care*, *47*(1), 26-43.

[5] AbdulRaheem, Y. (2023). Unveiling the significance and challenges of integrating prevention levels in healthcare practice. *Journal of primary care & community health*, *14*, 21501319231186500.

[6] Portero McLellan, K. C., Wyne, K., Villagomez, E. T., & Hsueh, W. A. (2014). Therapeutic interventions to reduce the risk of progression from prediabetes to type 2 diabetes mellitus. *Therapeutics and clinical risk management*, 173-188.

[7] Guan, H., Wang, Y., Niu, P., Zhang, Y., Zhang, Y., Miao, R., ... & Tian, J. (2024). The role of machine learning in advancing diabetic foot: a review. *Frontiers in Endocrinology*, *15*, 1325434.

[8] Hider, M. A., Nasiruddin, M., & Al Mukaddim, A. (2024). Early Disease Detection through Advanced Machine Learning Techniques: A Comprehensive Analysis and Implementation in Healthcare Systems. *Revista de Inteligencia Artificial en Medicina*, *15*(1), 1010-1042.

[9] An, Q., Rahman, S., Zhou, J., & Kang, J. J. (2023). A comprehensive review on machine learning in healthcare industry: classification, restrictions, opportunities and challenges. *Sensors*, *23*(9), 4178.