**Leveraging Machine Learning to Predict Diabetes: Project Report**

1: **Data Collection and Exploration:**

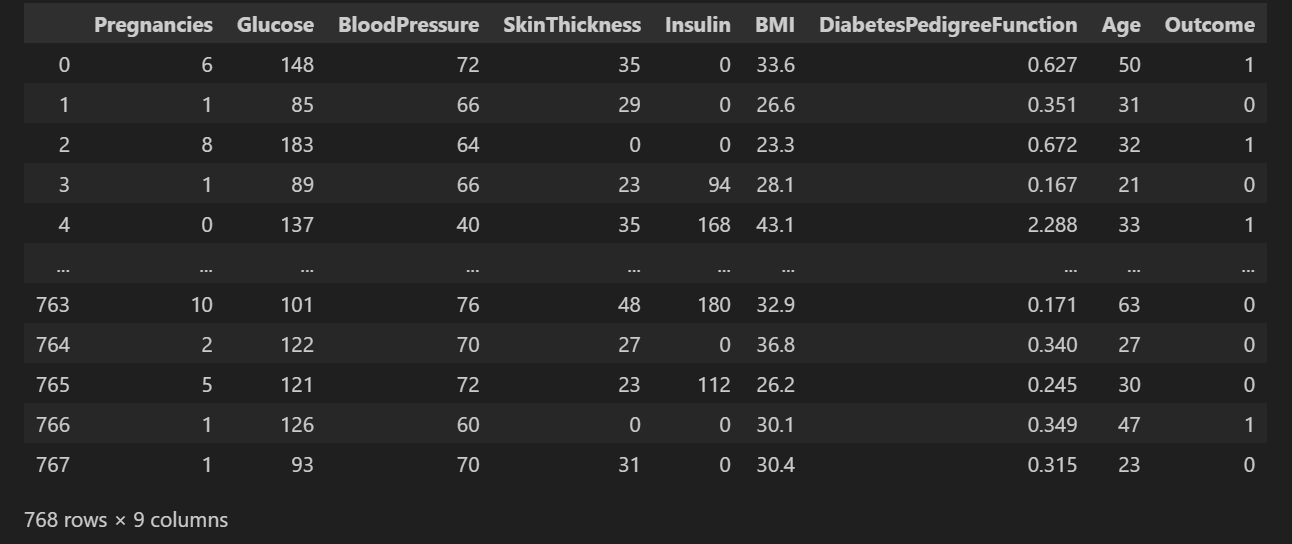
We downloaded the dataset from the UCI Machine Learning Repository.

This dataset contains the sign and symptoms data of newly diabetic or would be diabetic patient.

This has been collected using direct questionnaires from the patients of Sylhet Diabetes

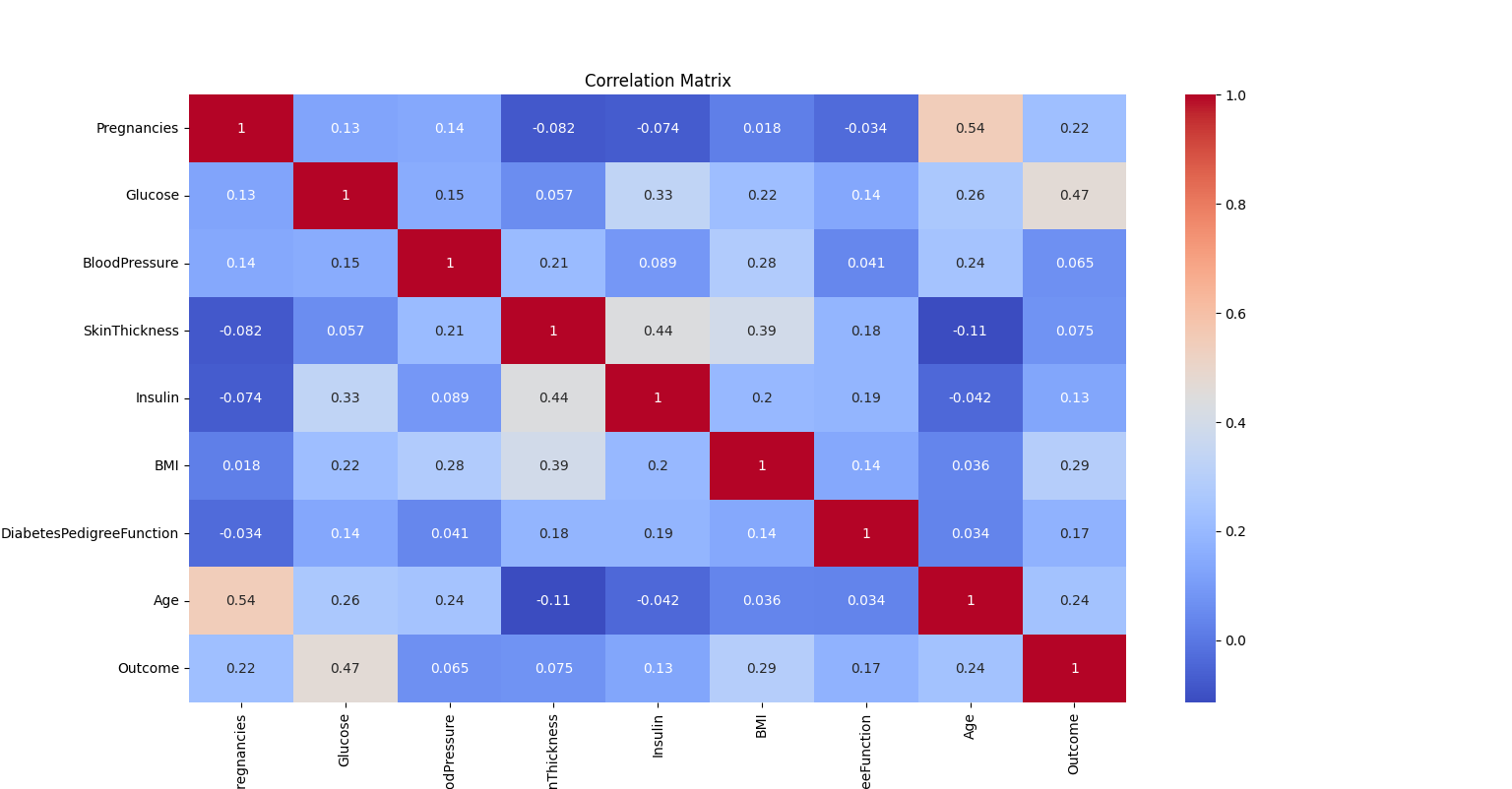
Hospital in Sylhet, Bangladesh and approved by a doctor

We loaded the dataset using Pandas.



We explored the dataset’s structure, features, and distribution of classes using pyplot

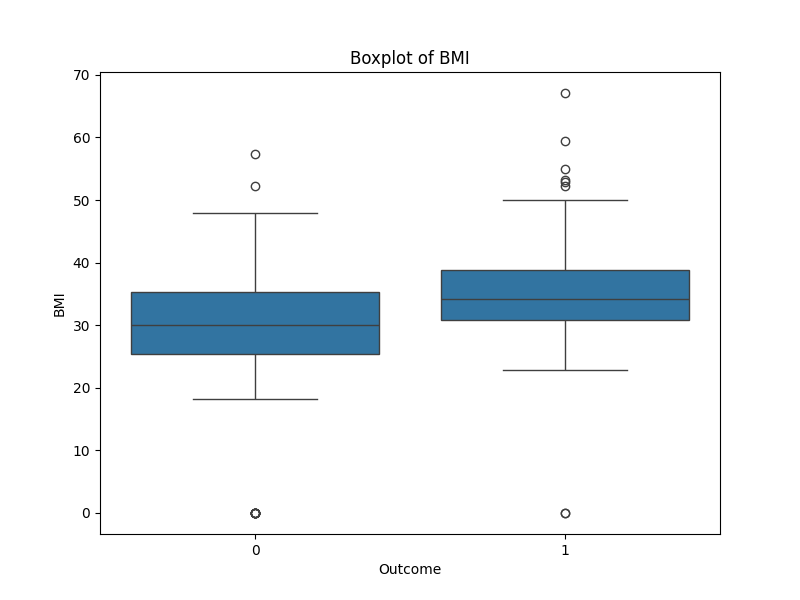
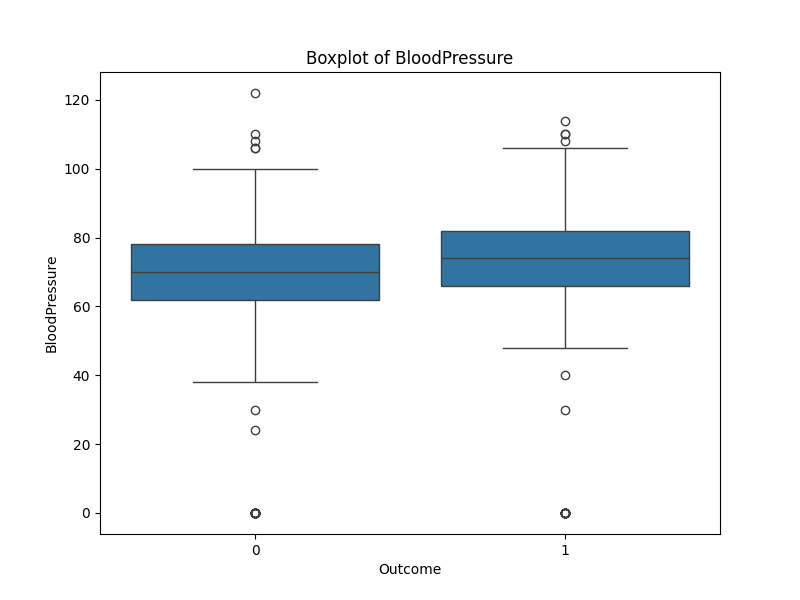
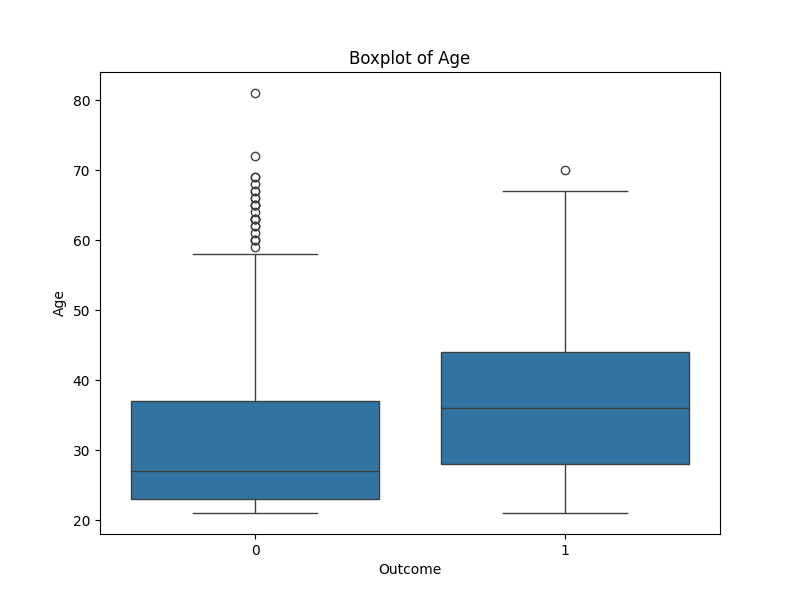
The matrix below show the correlation between different features of the dataset which was generated during our data xploration



Relationship between Age feature and the target variable.

The box plot below shows the relationship between ‘Age’ and the target ‘Outcome’ variable

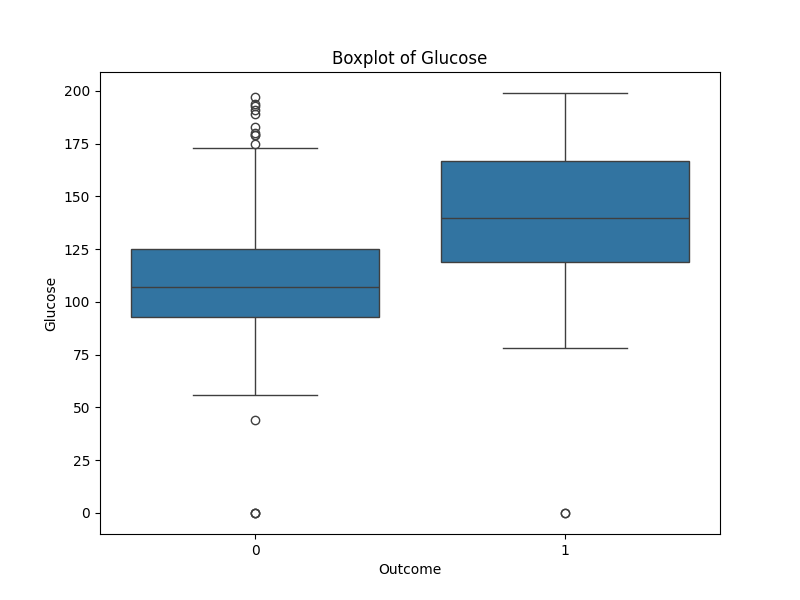
During information gain analysis, the ‘Age’ feature provided an information gain of 0.04282841 making an important features to be considered in the model.

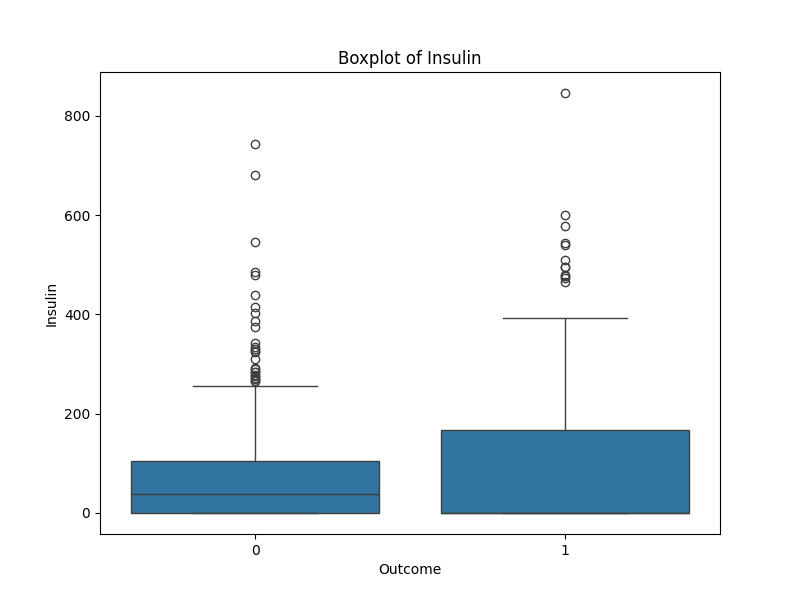


Relationship between ‘Glucose’ feature and the outcome :

The boxplot below shows the relationship betwee glucose level and the diabetes test outcome.

During information gain analysis, the ‘glucose’ feature had an information gain of 0.11977819 making it the most important feature that is being used in the prediction.





We handled missing values and performed basic statistical analysis (e.g., mean, median, standard deviation) to understand the nature of the data using pandas Dataframe methods.

2: **Data Preprocessing:**

-We preprocessed the data by scaling numerical features to a standard range using StandardScaler.

-We spitted the dataset into training and testing sets (e.g., 70/30 or 80/20 split) to prepare for model training and evaluation.

3: **Feature Engineering:**

- Conduct feature selection using techniques like correlation analysis or feature importance from tree-based models (e.g., Random Forest).

- Implement dimensionality reduction techniques such as Principal Component Analysis (PCA) to reduce the number of features while preserving information.

4: **Model Selection and Training:**

-We tried machine learning algorithms suitable for binary classification tasks such as:

* Logistic Regression
* Random Forest,
* Support Vector Machines
* Kneighbors Classifier
* Ridge Classifier

-We trained these models using the training dataset with optimized hyper-parameters obtained through GridSearchCV.

-We evaluated the performance of each model using evaluation metrics such as accuracy, precision, recall, F1-score, and AUC-ROC curve.

5: **Model Evaluation and Validation:**

- Validate the trained models using the testing dataset to assess their generalization performance.

- Perform k-fold cross-validation to ensure the robustness and reliability of the models.

- Analyze the results and compare the performance of different models to identify the best-performing algorithm for diabetes detection.

6: **Model Interpretation and Fine-Tuning:**

-We interpreted the results to understand the importance of individual features and their contribution to the predictive performance of the models.

-We fine-tuned the selected model by adjusting hyper-parameters or incorporating additional features to improve its performance further.

7. **Future Plans:**

Given more time and resources, we plan to:

1: Deployment and Integration:

- Deploy the selected model into a production environment using frameworks like Flask or FastAPI.

- Integrate the model into a user-friendly interface or application, allowing healthcare professionals to input patient data and obtain predictions conveniently.

- Ensure proper documentation and version control of the deployed model for future updates and maintenance.

2: Monitoring and Iteration:

- Monitor the model’s performance in real-world scenarios and collect feedback from users.

- Continuously iterate and improve the model based on new data, emerging trends, and user feedback to enhance its accuracy and reliability over time.

3: Ethical Considerations:

- Ensure compliance with data privacy regulations (e.g., GDPR, HIPAA) and ethical guidelines regarding the use of patient data.

- Prioritize transparency and accountability in model development and decision-making processes.

4: Collaboration and Knowledge Sharing:

-We need collaborate with domain experts, healthcare professionals, and stakeholders throughout the implementation process to incorporate their insights and expertise.

-We also plan to share our findings, methodologies, and learnings with the broader community through research papers, presentations, and open-source contributions.