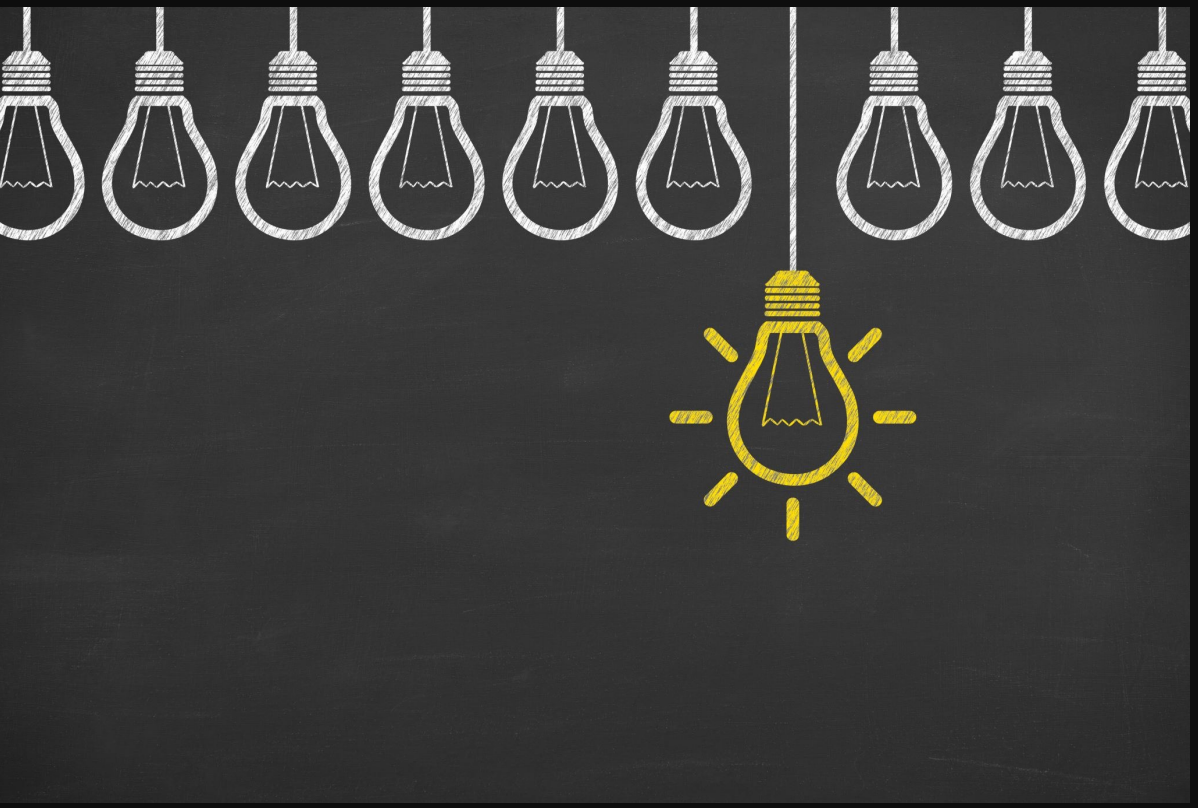
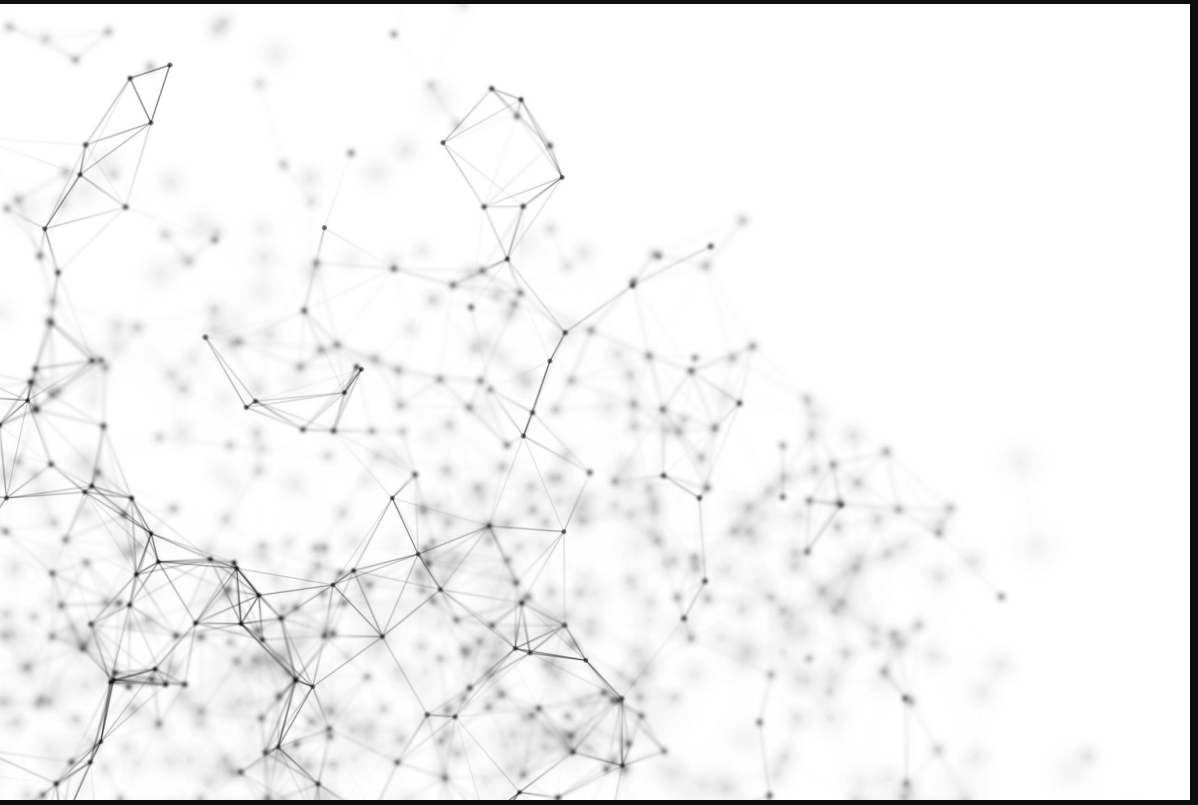
# Supervise learning ‘Diabetes’ Dataset

Kruti kikani



# Process of the project

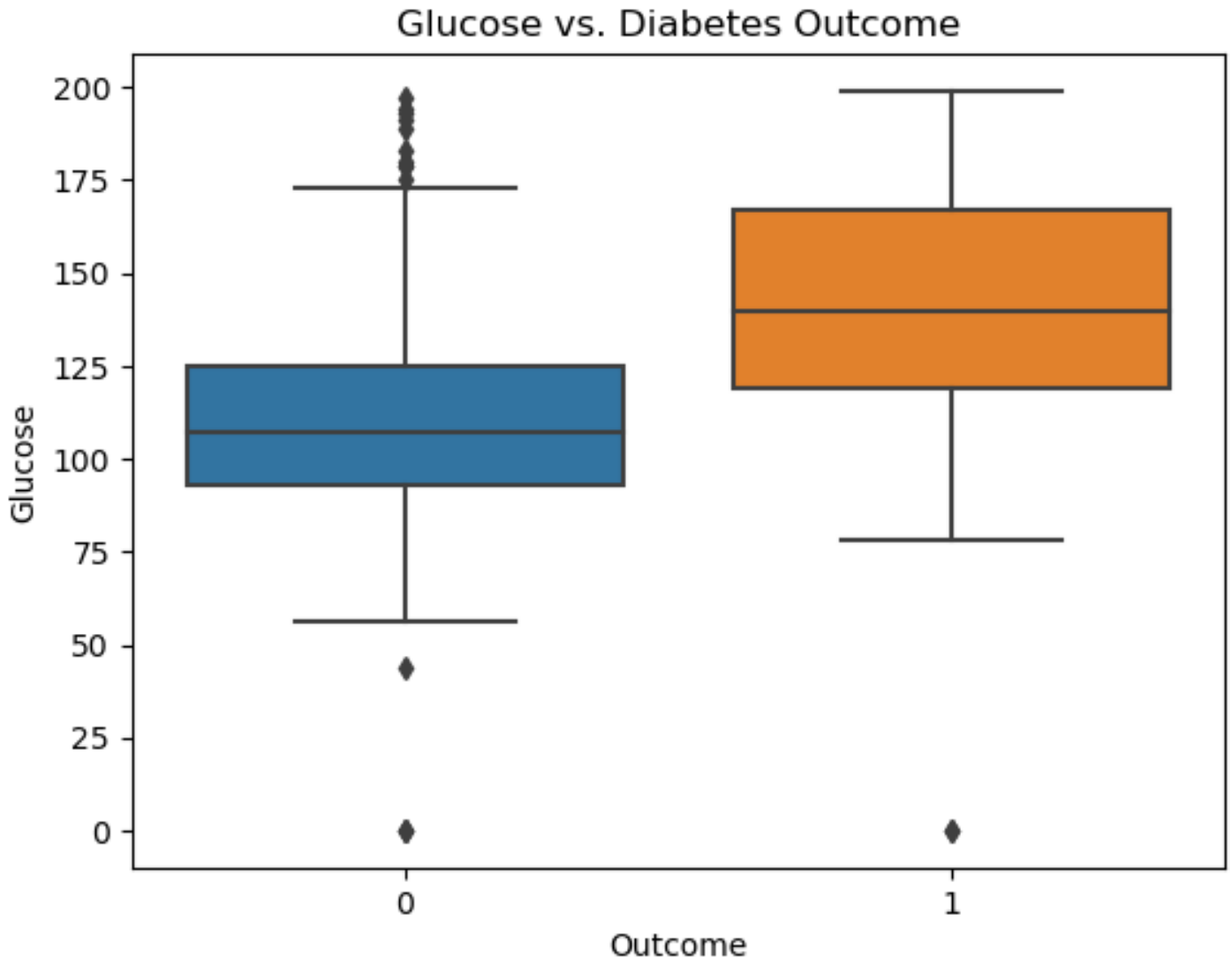
- Project goal
- Exploratory Data Analysis
- Processing and featuring
- Tarining ml model
- conclusion



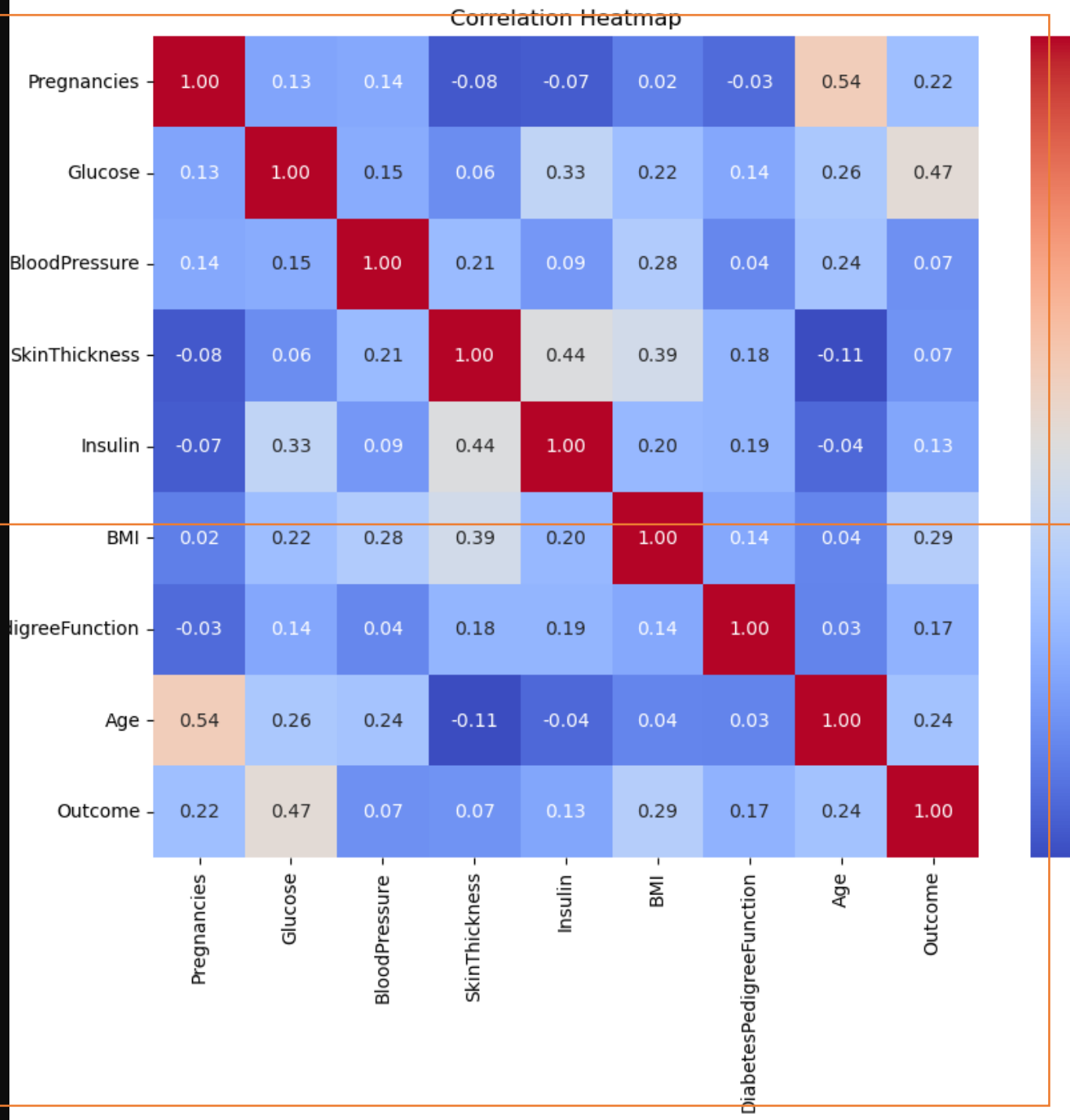
# Project goal

- With using supervised machine learning we have to built a machine learning model and predict where a patient has diabetes or not

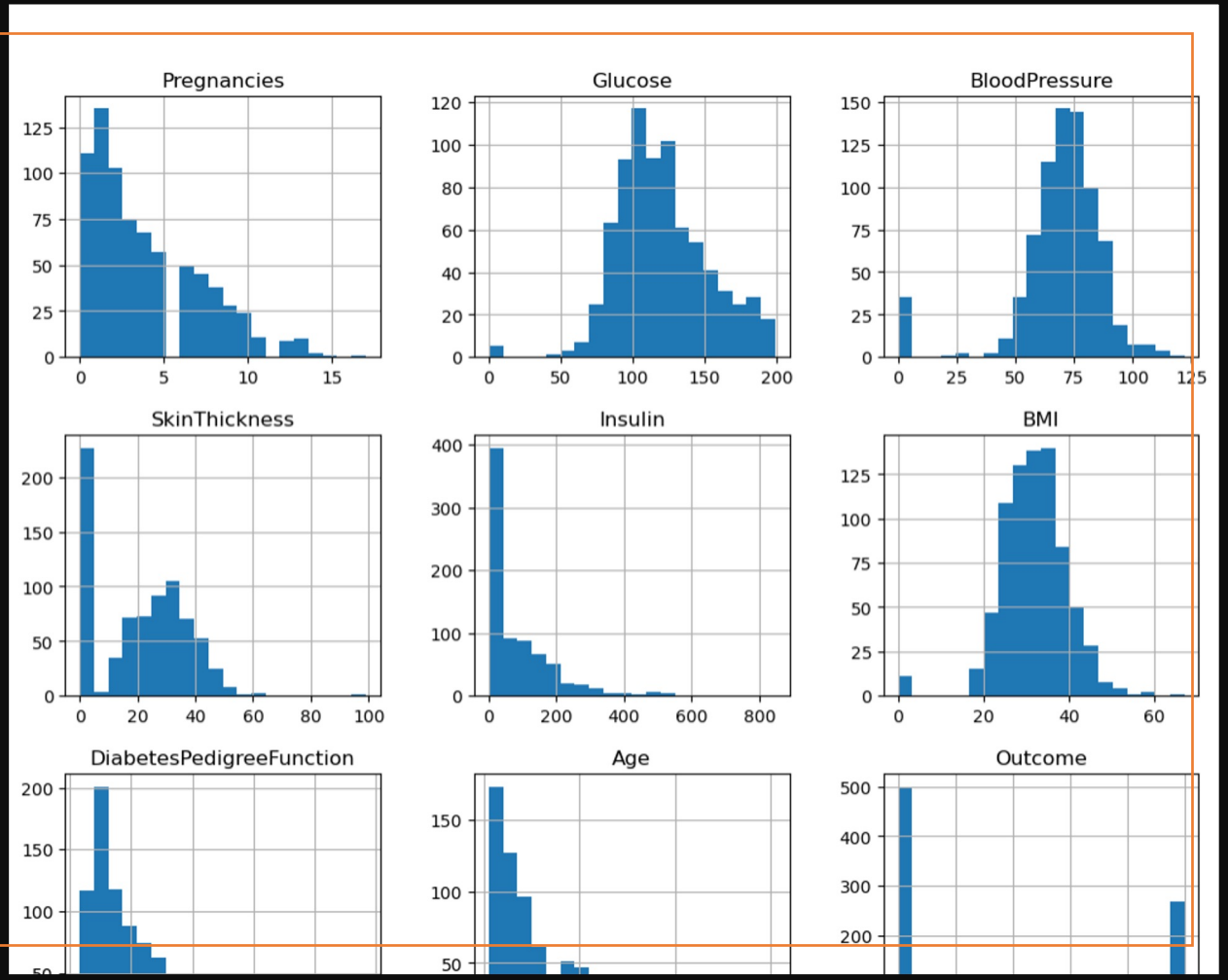
predictor variables are related to the outcome variable



calculate the correlation matrix and visualize it using a heatmap



create histograms to examine the distribution of each predictor variable



# Feature engeineering

- Define age grouping
- BMI categories:
  - Categories:
    - Underweight
    - Normal
    - Overweight
    - obese

# Model building

- Logistic Regression Evaluation:
- Accuracy: 0.73
- Precision: 0.7008547008547008
- Recall: 0.8118811881188119
- F1-score: 0.7522935779816514
- ROC-AUC: 0.7291729172917292

- Random Forest Evaluation:
- Accuracy: 0.825
- Precision: 0.7844827586206896
- Recall: 0.900990099009901
- F1-score: 0.8387096774193549
- ROC-AUC: 0.824232423242324

# conclusin

- the Random Forest model outperforms the Logistic Regression model in most of the metrics:
- Accuracy:
- Random Forest (0.825) > Logistic Regression (0.73)
- Precision:
- Random Forest (0.784) > Logistic Regression (0.701)
- Recall:
- Random Forest (0.901) > Logistic Regression (0.812)
- F1-score:
- Random Forest (0.839) > Logistic Regression (0.752)
- ROC-AUC:
- Random Forest (0.824) > Logistic Regression (0.729)
- The Random Forest model consistently achieves higher values for accuracy, precision, recall, F1-score, and ROC-AUC, indicating better performance overall.
- Therefore, based on the provided evaluation metrics, the Random Forest model appears to be the better choice for predicting diabetes in new data for this specific task. It shows a good balance between precision and recall, which is important for binary classification tasks, especially when the data is imbalanced.

# Thank you