Project Description

- Prediction of Diabetes based on Diagnostic measures.
- Task: **Binary Classification**

Team Members

Banoth Rajshekhar	19BT30008	banothrajkumar9959@gmail.com
Nishant Gahlaut	19BT30015	nishantgahlaut@gmail.com
Venkata Tharun Raj	19BT30033	tharunrajvenkat@gmail.com

Dataset

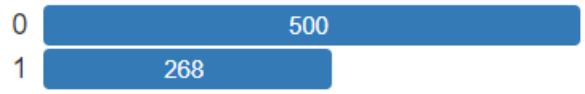
- Given Diabetes dataset contains following Diagnsotic measure
  - Pregnancies Glucose
  - BloodPressure
  - SkinThickness
  - Insulin
  - BMI
  - DiabetesPedigreeFunction
  - Age
- Label Column: Outcome
- Sample Data:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

- Number of datapoints: 768

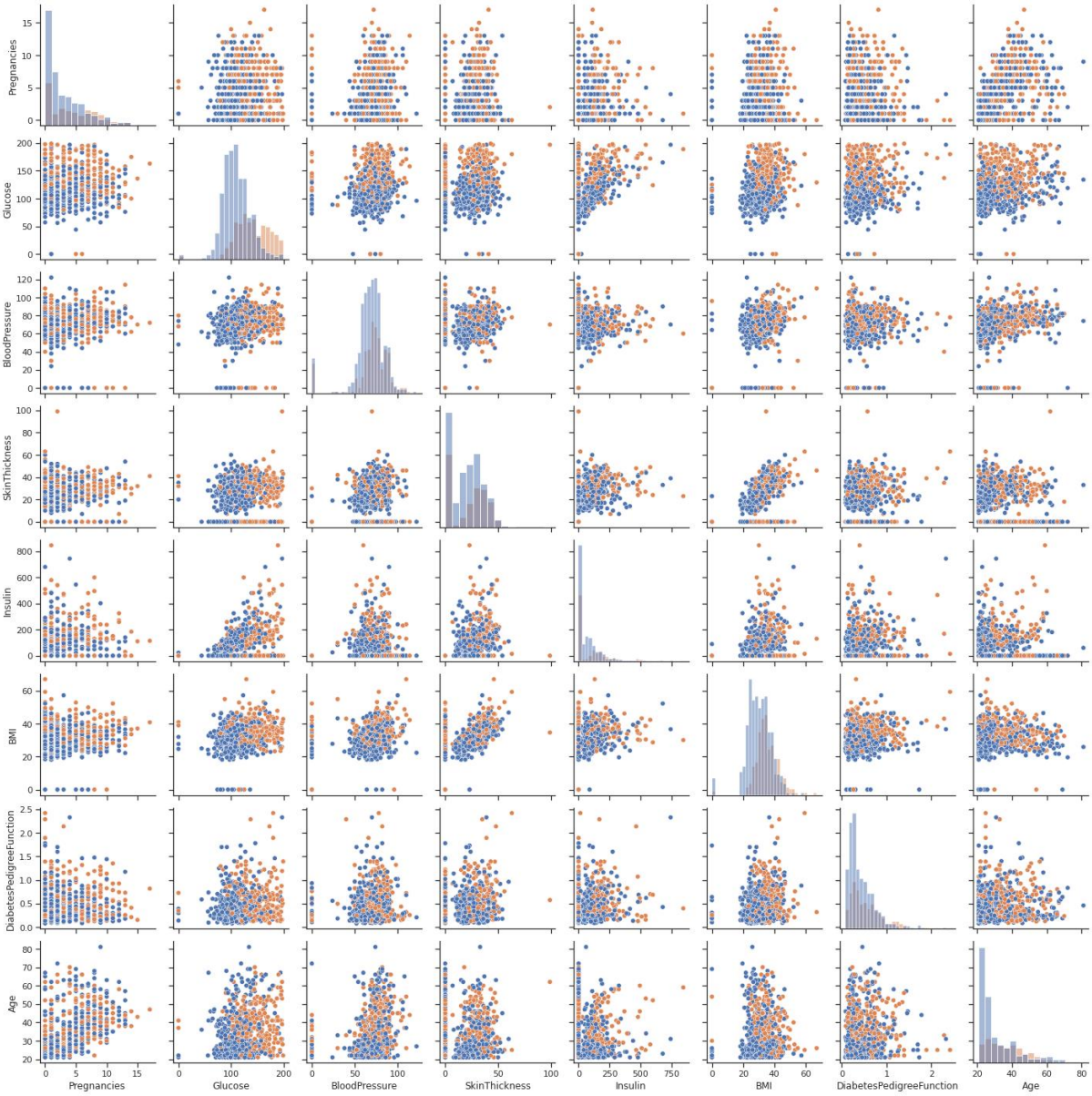
Data Visualization

- Label Distribution:



  - 0: No Diabetes
  - 1: Has Diabetes

- Correlation Plot



Data Pre-processing

Column	Type	Fill NaN	Preprocessing
Pregnancies	Categorical	No NaN	Label Encoding
Glucose	Continuous	No NaN	MinMax Scaling
BloodPressure	Continuous	No NaN	MinMax Scaling
SkinThickness	Continuous	No NaN	MinMax Scaling
Insulin	Continuous	No NaN	MinMax Scaling
BMI	Continuous	No NaN	MinMax Scaling
DiabetesPedigreeFunction	Continuous	No NaN	MinMax Scaling
Age	Continuous	No NaN	MinMax Scaling

ML Model Training

- Data Splitting Details

Training	0.72
Validation	0.18
Testing	0.1

- 5-Fold CV Training
- Library used: scikit-learn
- ML Algorithms Trained
    - Logistic Regression
    - SVM
    - KNN
    - Decision Tree
    - Random Forest
    - Histogram Gradient Boosting
- Evaluation Metric:

Results

Model	Test Data AUROC
Logistic Regression	0.777
SVM	0.657
KNN	0.602
Decision Tree	0.760

Random Forest	0.806
Histogram Gradient Boosting	0.773

- **Best Model: Radom Forest**