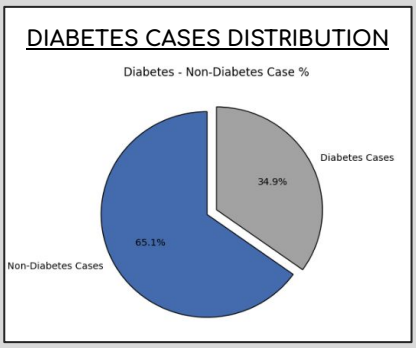
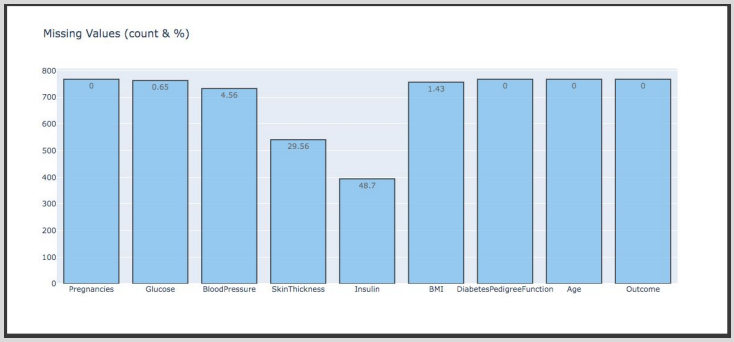


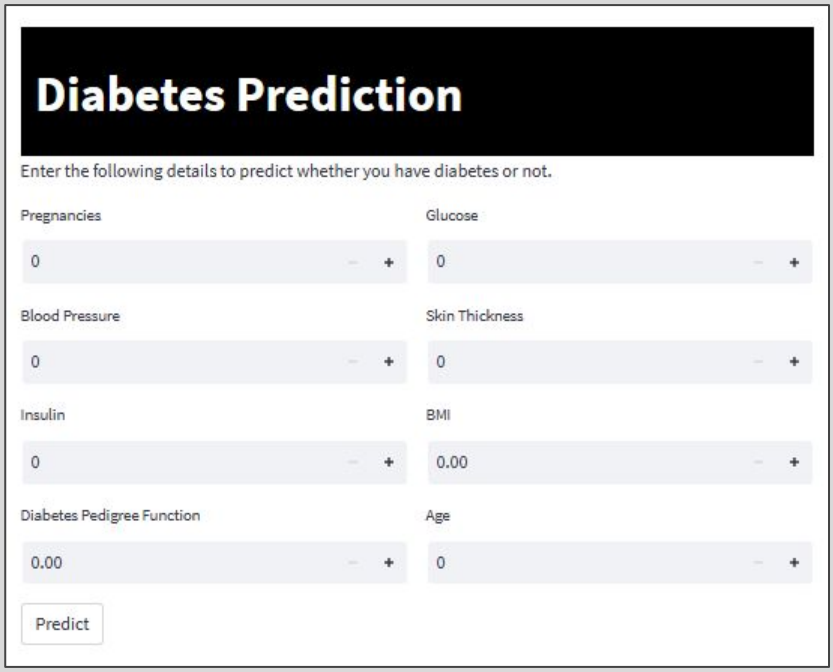


- According to data, the distribution numbers are imbalanced. There were as many as 34.9% of diabetes cases and 65.1% non-diabetes cases. The total number of cases amounted to 768.
- Imputation of missing values is performed using mean and median of the data.
- Performing data preprocessing and data balancing using SMOTE Technique.
- Constructing data using Standard Scaler for Standardization.

➢ Best Model Machine Learning

Train Test 60:40	Accuracy Training	Accuracy Testing	Precision	Recall	F-1 Score
XGBoost	90.1%	89.3%	88%	88%	88%

➢ Deployment Model



- The data was divided into train:test sets in ratios of 80:20, 70:30, and 60:40.
- Algorithms of the modeling process included: Logistic Regression, XGBoost, Support Vector Classifier, K-Nearest Neighbors Classifier, Decision Tree Classifier, and Stacking (Random Forest, AdaBoost, Logistic Regression).
- The best model was obtained from a 60:40 train-test split using the XGBoost algorithm.

PROJECT EXPLANATION

This project was developed for the Final Project in the Digital Skola course. The objective of the project is to construct a model for diagnostic prediction of diabetes in patients based on the measurement of predictor variables in the Pima Indians dataset. My team successfully completed this project within a timeframe of one month.