The goal of the project is to determine the best model and features to predict if a person has diabetes or not based on diagnostic measurements in the dataset. The algorithms that were run were KNN Classifier, Logistic Regression, SVC, and Random Forest. The process included EDA, feature selection, and hyperparameter tuning of the algorithms to get the best models.

Dataset used: <https://www.kaggle.com/datasets/akshaydattatraykhare/diabetes-dataset>

The process for determining the best model includes:

* Data Overview
* EDA – Statistical Summary and Inference
* EDA – Class Label Imbalance
* EDA - Univariate Analysis and Imputing Zero Values
* EDA – Feature Correlation
* EDA – Feature Selection
* Run models with all features and selected features
* Hyperparameter tuning of algorithms to get best accuracy, precision and recall
* Algorithms Run:
  + kNN Classifier: Tuning Hyperparameter k iteratively (n neighbors)
  + Logistic Regression: Tuning hyperparameter C iteratively (Inverse of regularization strength)
  + SVC: Hyperparameter Tuning for best C and gamma using GridSearchCV
  + Random Forest: Hyperparameter tuning iteratively for N (number of trees) and d(max depth )
* Result Analysis
* Conclusion