Predicting Diabetes: An Exploration

Fantastic Four (Data Analysts Edition): Meihan Chen, Parnika Dandepally, Catherine Maclean, Likitha Temmanaboyina

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

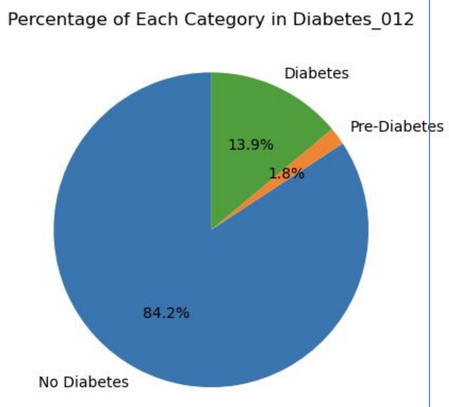
Introduction

Question: What features are most important for determining whether an individual has diabetes?

- This question will help us with identifying key characteristics of diabetes aids in prevention, efficient diagnosis, and overall healthcare efficiency by reducing unnecessary tests and costs.
- **Objective**: Identify health, lifestyle, income, and education characteristics common among people with diabetes and pre-diabetes.
- **Goal**: Determine conditions that place individuals at higher risk of diabetes/pre-diabetes.

Data

- The data was collected by the CDC via a telephone survey on health-related conditions, behaviors, and preventative services.
- The original dataset includes responses from 441,455 American individuals.

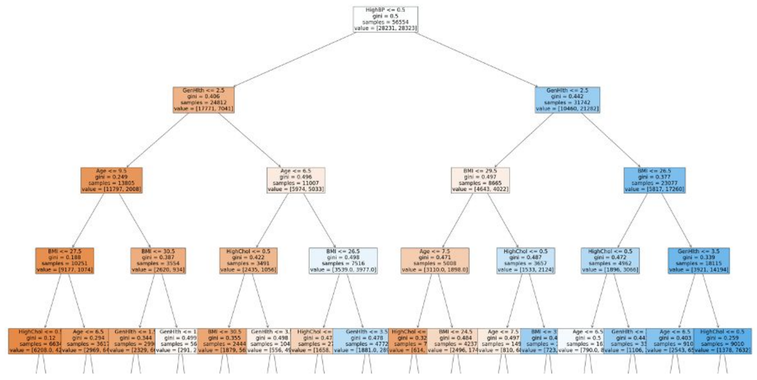


- The cleaned dataset sampled 253,680 individuals.
- Found an imbalance in the dataset: choose to only focus on Diabetes (35346 data points) and No Diabetes (213703 data points).

Methods

Model 1: Decision Tree

- Under sampled non-diabetes group randomly to match the number in diabetes group
- Had 50% of each group in new dataset
- Split 80% and 20% of data points randomly into train and test sets respectively
- Ran the model with tree depth that gives the highest accuracy



Model 2: Random Forest

- Applied SMOTE to balanced the class distribution by generating new synthetic data points for minority class (Diabetes).
- Split the balanced dataset into 80% training and 20% testing sets.
- Trained a Random Forest model with 100 decision trees. It was chosen for its ability to handle both categorical and continuous variables and rank feature importance.

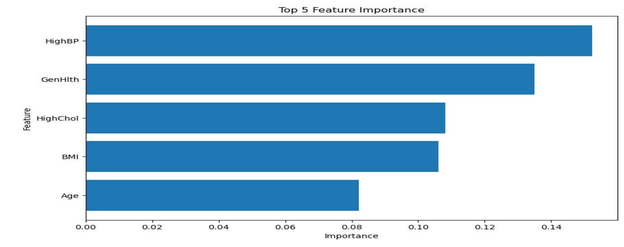
Results

Model 1: Decision Tree

- Achieved an overall Accuracy score of 75%.
- High BP most informative feature
- GenHlth, Age, and BMI were also informative features.
- The best tree depth is 8.

Model 2: Random Forest

- Achieved an overall Accuracy score of 92%.
- Achieved recall scores for No Diabetes and Diabetes of 96% and 88%.
- Ranked the features by importance.



Model Deployment

- The 5 most important features are High blood pressure, General health, High cholesterol, BMI, and Age.
- After a nurse collects patient information, the models can be used to predict whether the patient has diabetes or not, bringing possible diabetes diagnosis to doctor's attention.
- Doctor will have final say in diagnosis but models can help with detecting possible diabetes diagnosis.