DIABETES ANALYSIS

FDAA Team 1:
Ryan Phua Rui En (U2320499A)
Tan Yichen (U2323280L)
Wong Kwan Kit (U2322309F)

Motivation

Dataset Used

ALEX TEBOUL · UPDATED 2 YEARS AGO

517 | New Notebook | Download (6 MB)

Diabetes Health Indicators Dataset

253,680 survey responses from cleaned BRFSS 2015 + balanced dataset

diabetes_012_health_indicators_BRFSS2015	17/3/2024 10:08 pm	Microsoft Excel Com...	22,206 KB
diabetes_binary_5050split_health_indicators_...	17/3/2024 10:08 pm	Microsoft Excel Com...	6,199 KB
diabetes_binary_health_indicators_BRFSS2015	17/3/2024 10:08 pm	Microsoft Excel Com...	22,206 KB

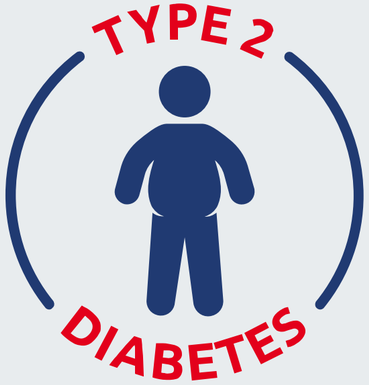
Motivation

- Singapore: No.2 nation with the most diabetics
- Friends diagnosed with high risk levels for diabetes at health checkups
- Lifestyle factors that contribute to this problem are often within our control

Problem Definition

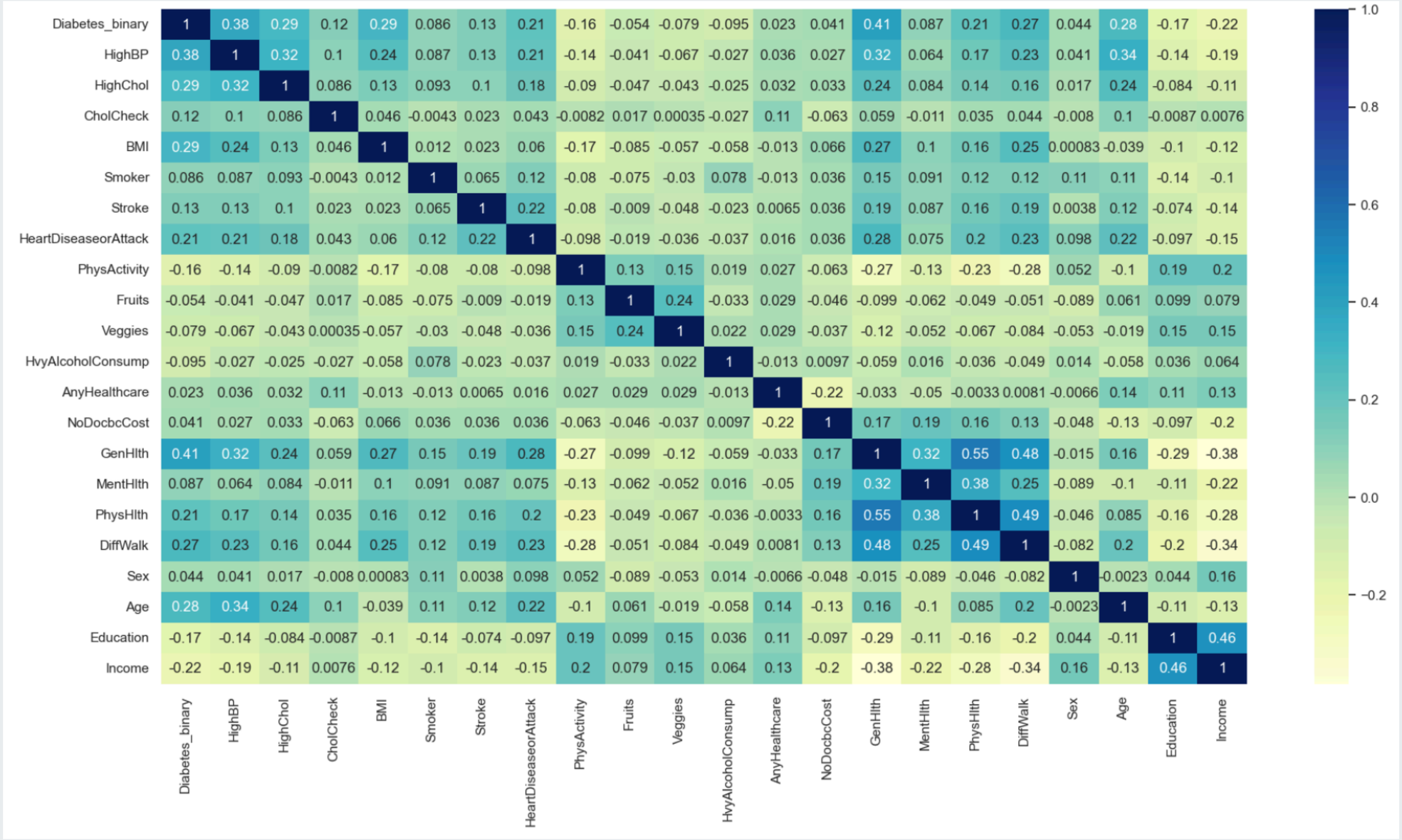
We wish to:

1. identify potential lifestyle factors that has higher association with risk of diabetes
2. make suggestions for lifestyle changes to lower risk of diabetes

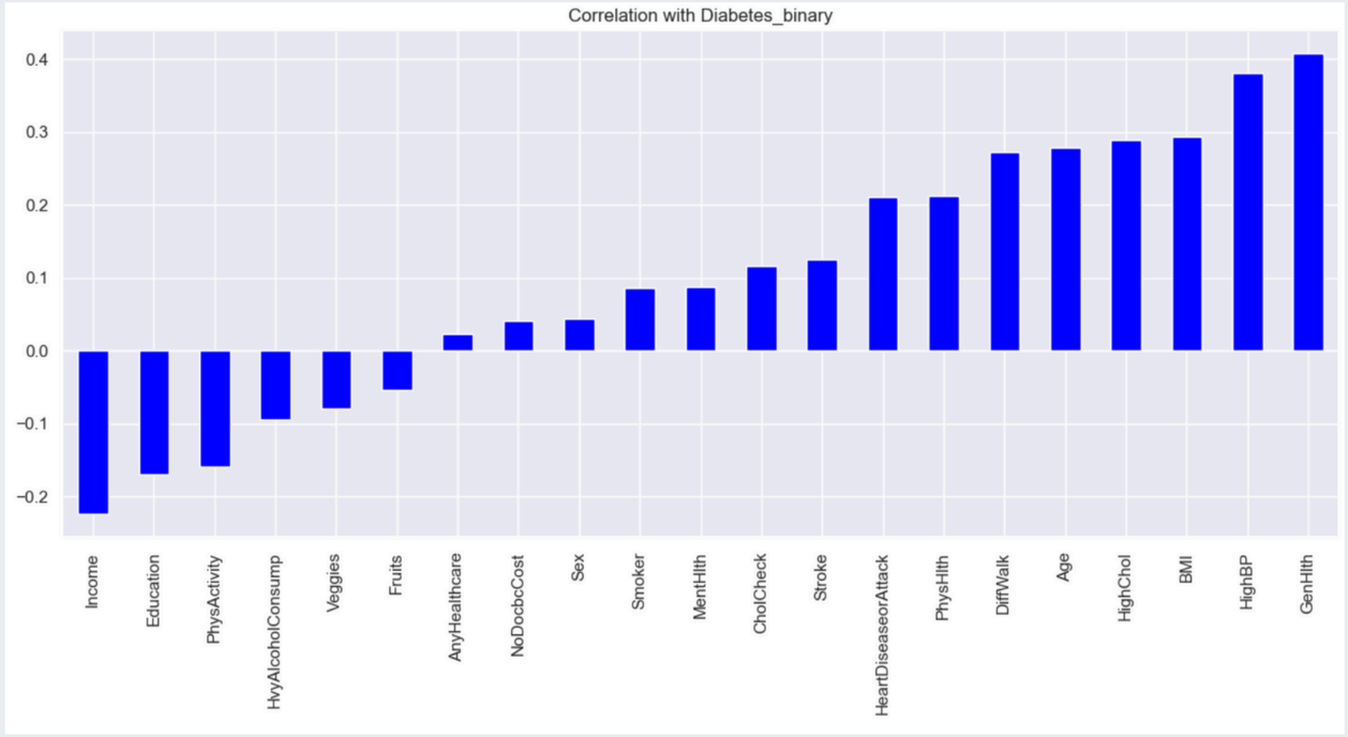


Cleaning the Dataset

Correlation Matrix



Correlation Coefficients



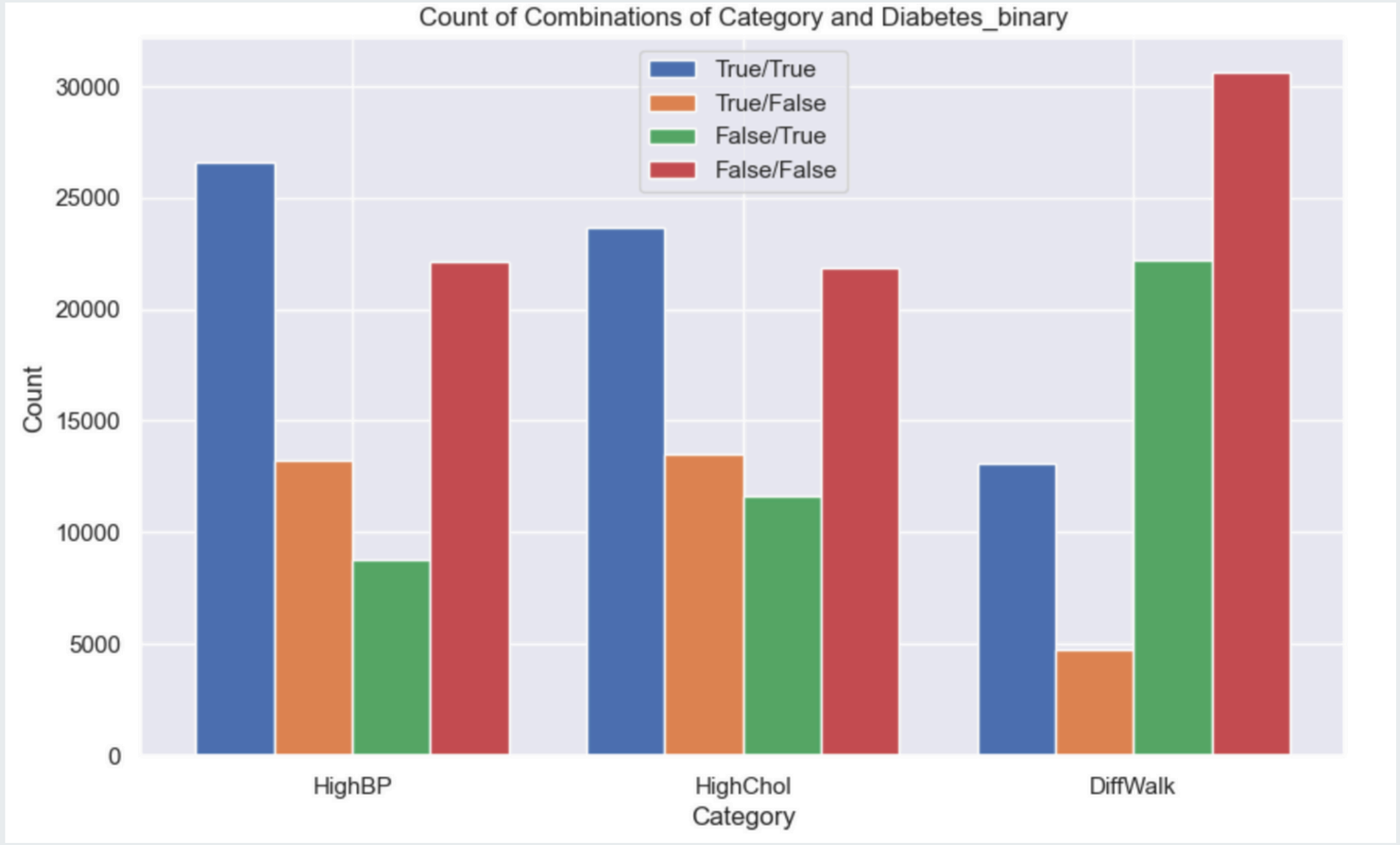
Significant Risk Factors (corr > | 0.2 |)

1. GenHealth (Most Significant)
2. HighBP
3. BMI
4. HighChol
5. Age
6. Income (-*ve* correlation)
7. DiffWalk
8. PhysHealth

Binary Data Analysis

- Predictor Variable 1: HighBP
- Predictor Variable 2: HighChol
- Predictor Variable 3: DiffWalk

Binary Data: Exploratory Data Analysis



Respondents who identified with these factors are likely to have diabetes

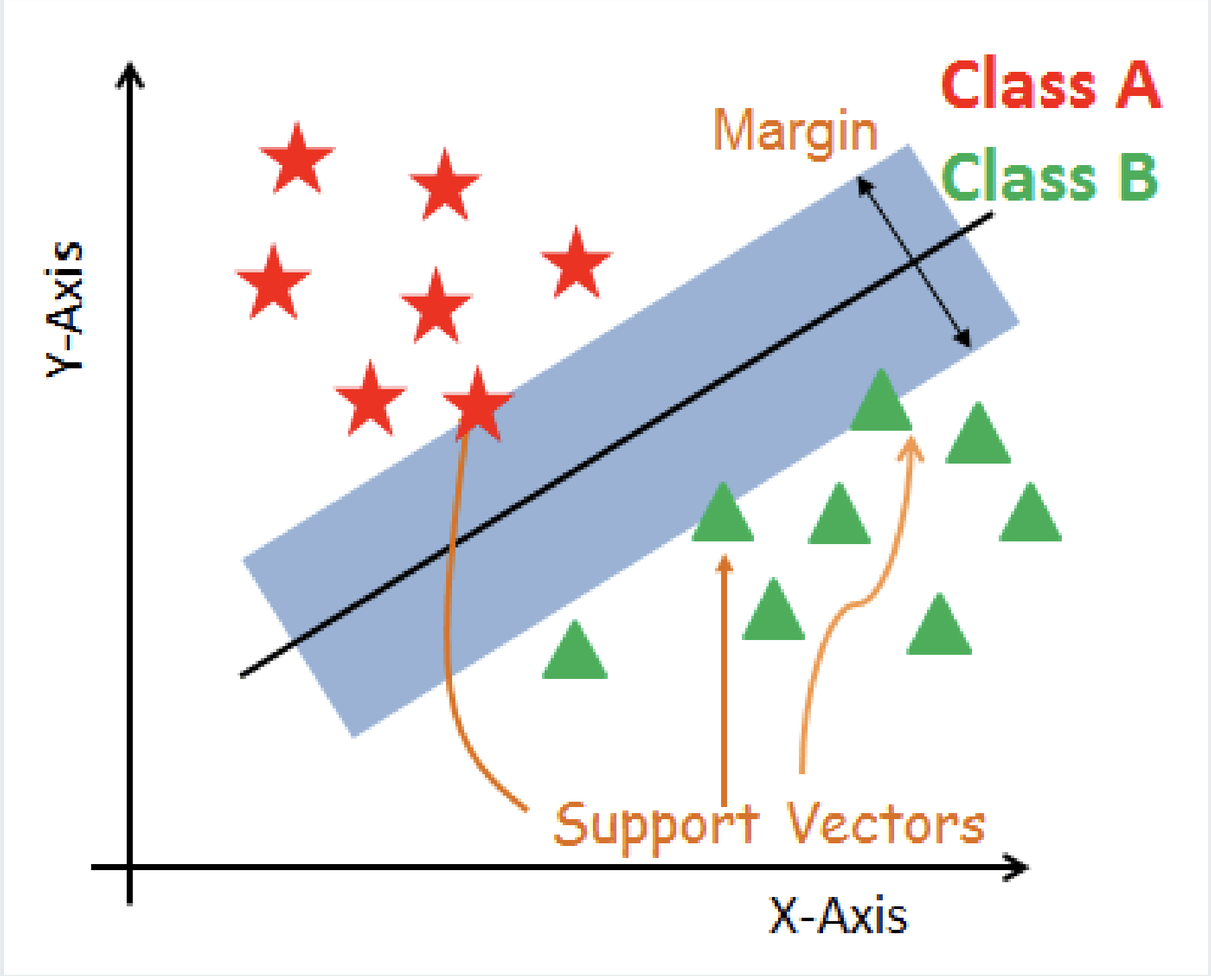
- More people have HighBP and HighChol
- Less people have DiffWalk

Binary Data Analysis: Support Vector Machine

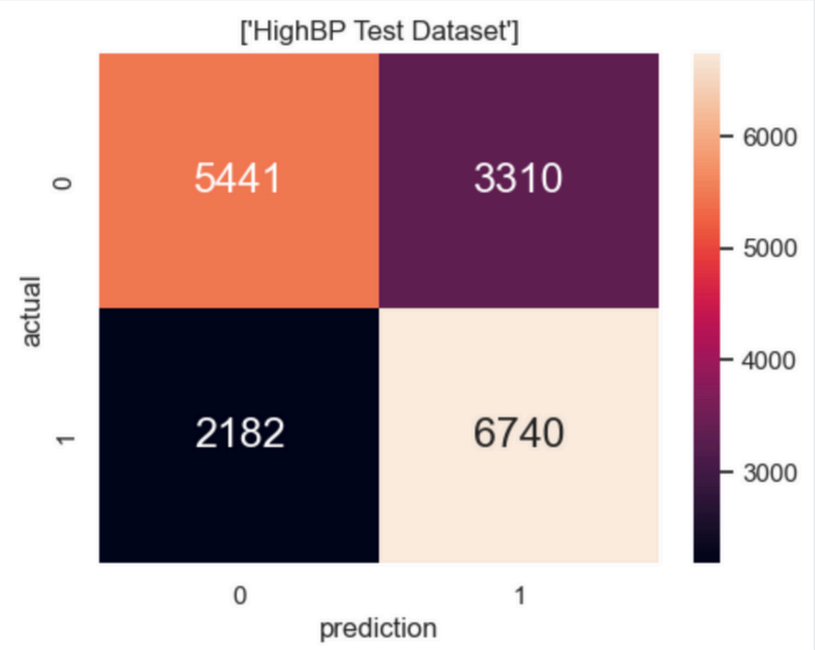
Support Vector Machine: Constructs a hyperplane to separate different classes

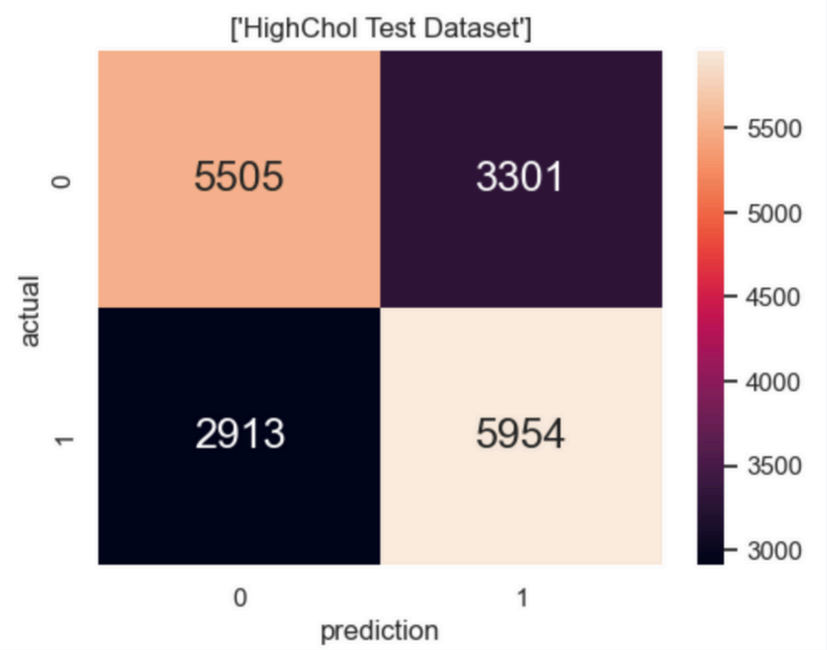
SVM Kernels:

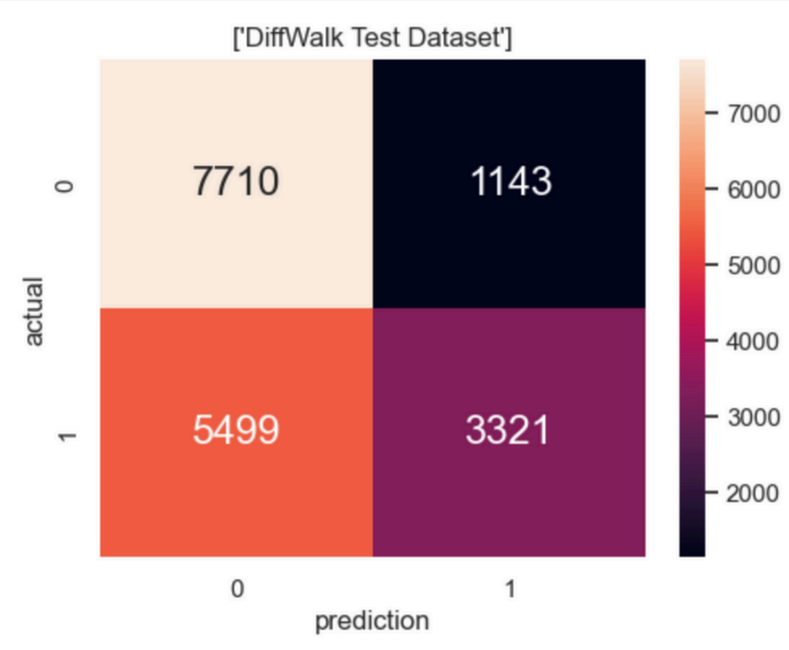
Linear, Polynomial and Radial



Binary Data Evaluation: Support Vector Machine







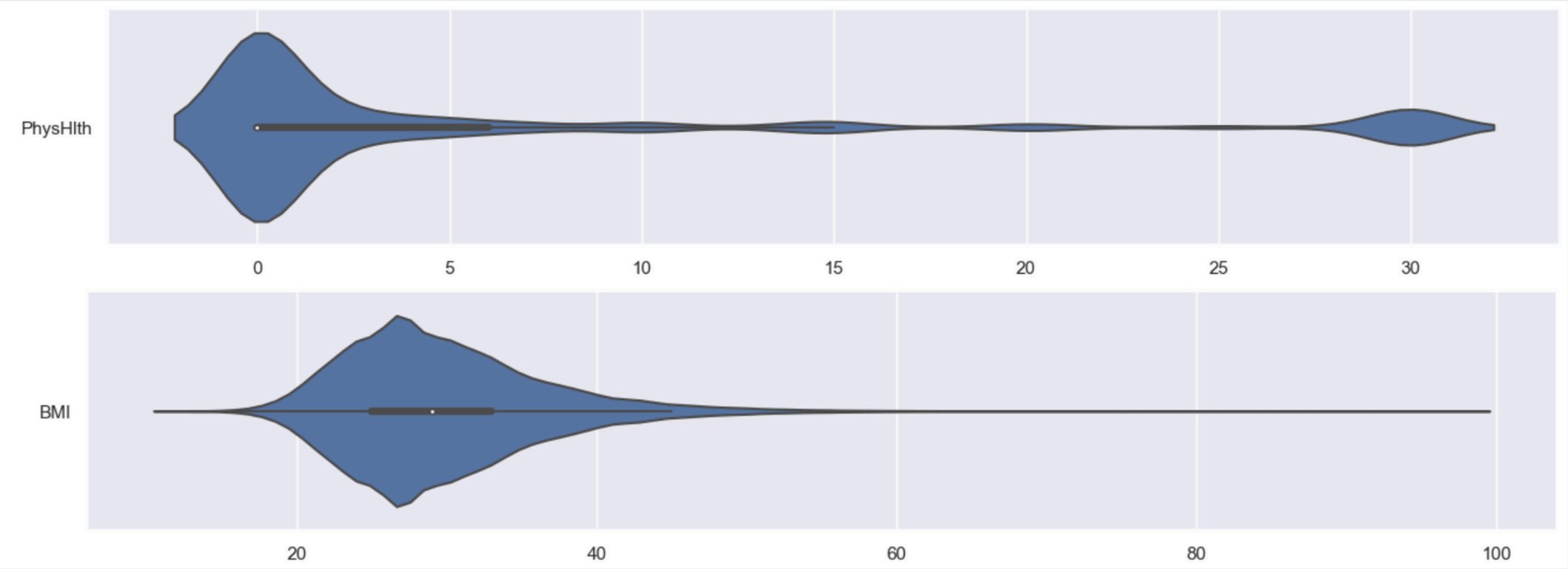
Classification Accuracy:

69% **65%** **62%**

Numerical Data Analysis

- Predictor Variable 1: PhysHlth
- Predictor Variable 2: BMI

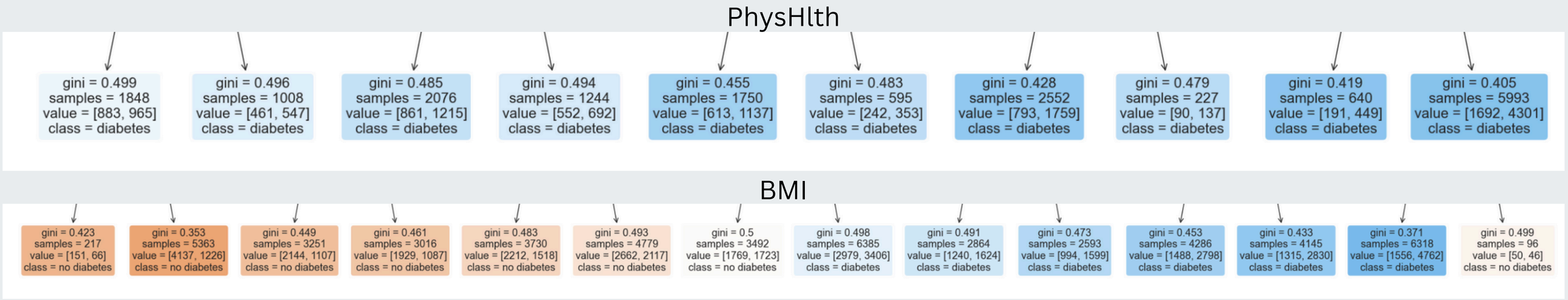
Numerical Data: Exploratory Data Analysis



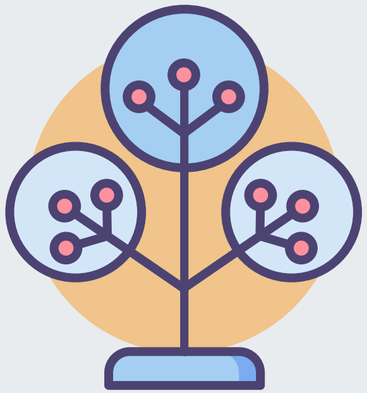
PhysHlth -> imbalanced data

BMI -> right skewed data

Numerical Data Analysis: Classification Tree



Gini Coefficient for most of the leaf nodes are above 0.4 -> High chance for misclassification



Highly skewed/imbalanced data

makes the classification tree inaccurate

Numerical Data Analysis: Logistic Regression

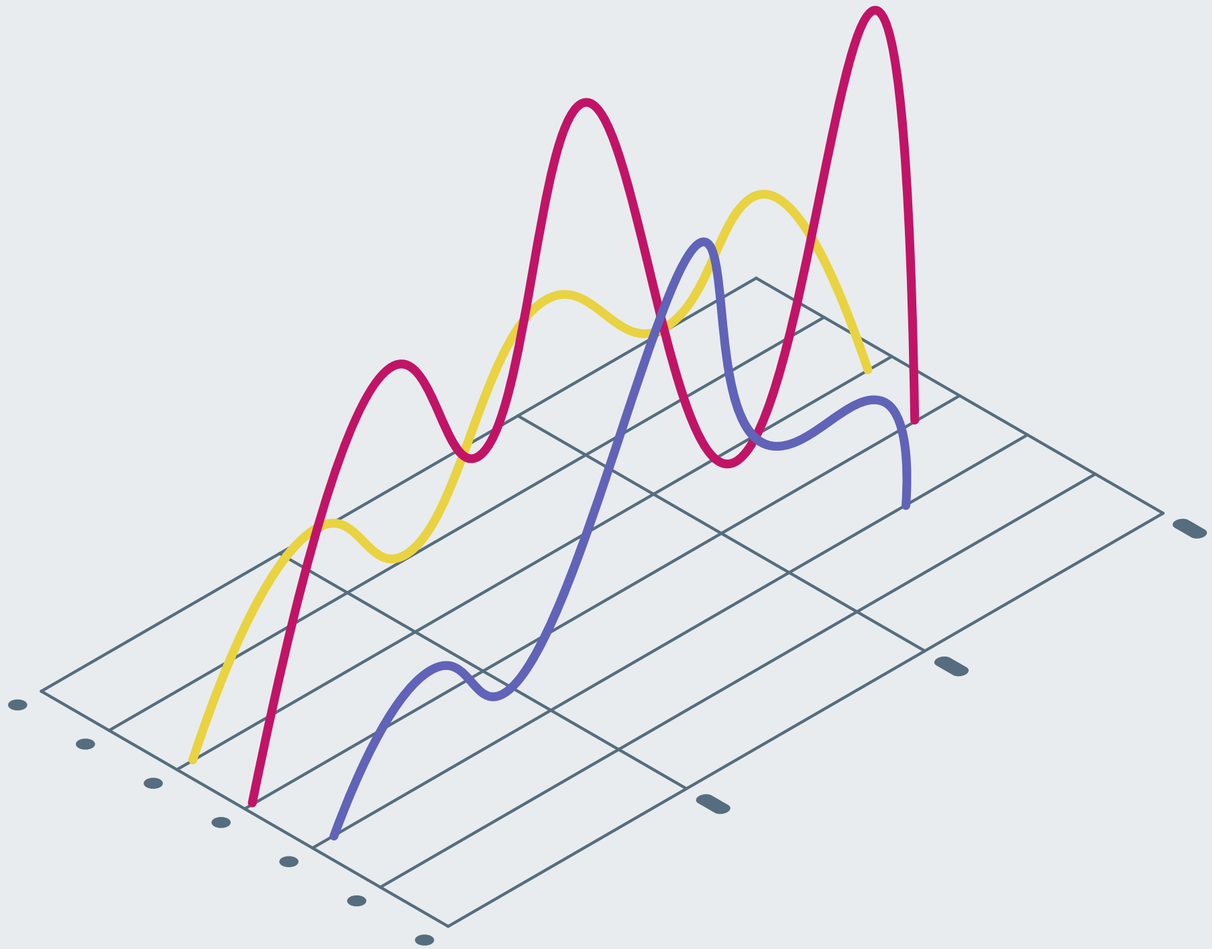
Logistic Regression: Linear Regression with Sigmoid Function

Sigmoid Function:

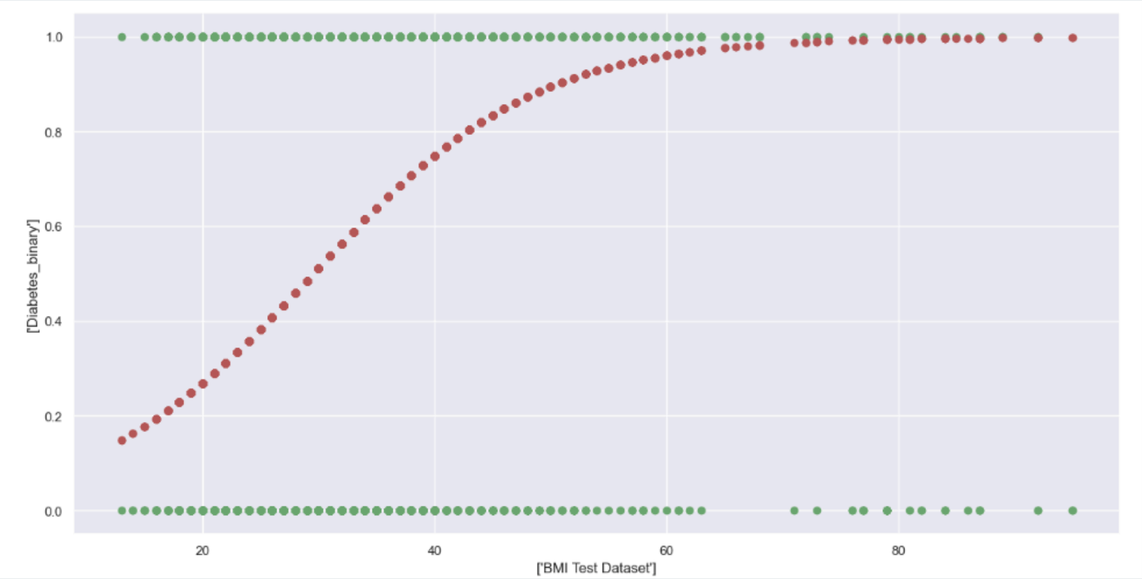
$$p = \frac{1}{1 + e^{-y}}$$

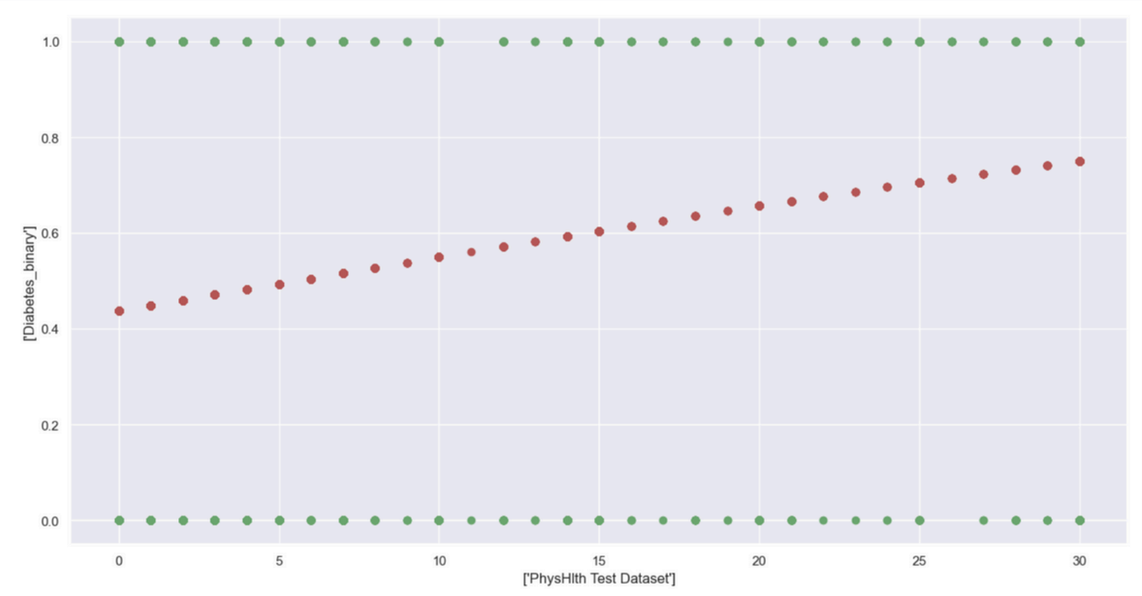
p: Probability

y: Linear Regression Equation



Numerical Data Analysis: Logistic Regression



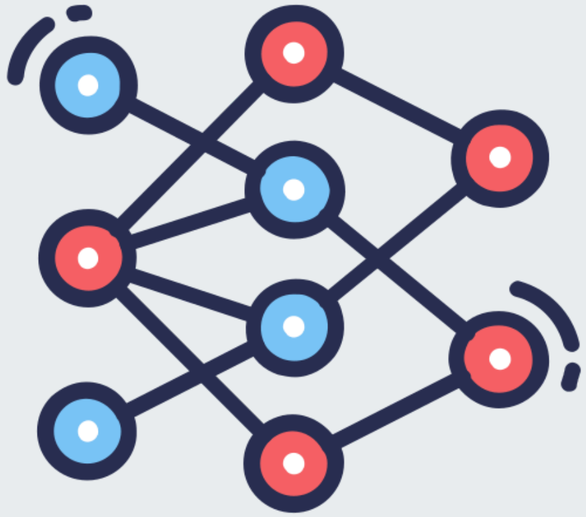


Logistic Regression:

- Able to handle the skewed data of BMI
- Unable to handle the imbalanced data of PhysHlth
- Cannot compare between PhyHlth & BMI

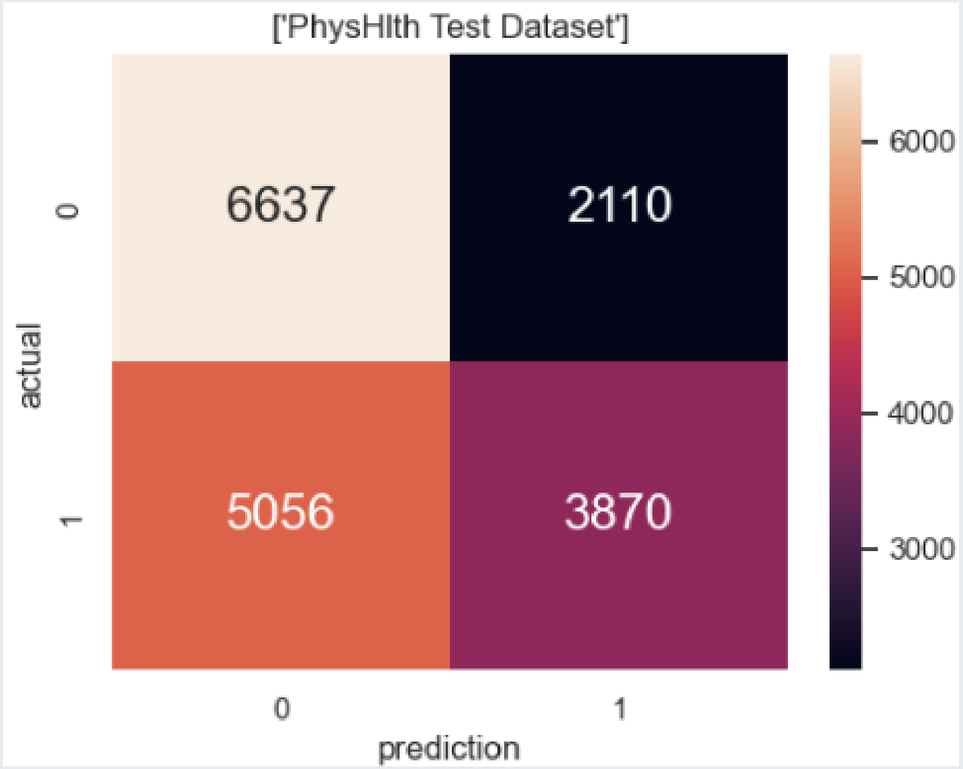
Numerical Data Analysis: Multi-Layer Perceptron

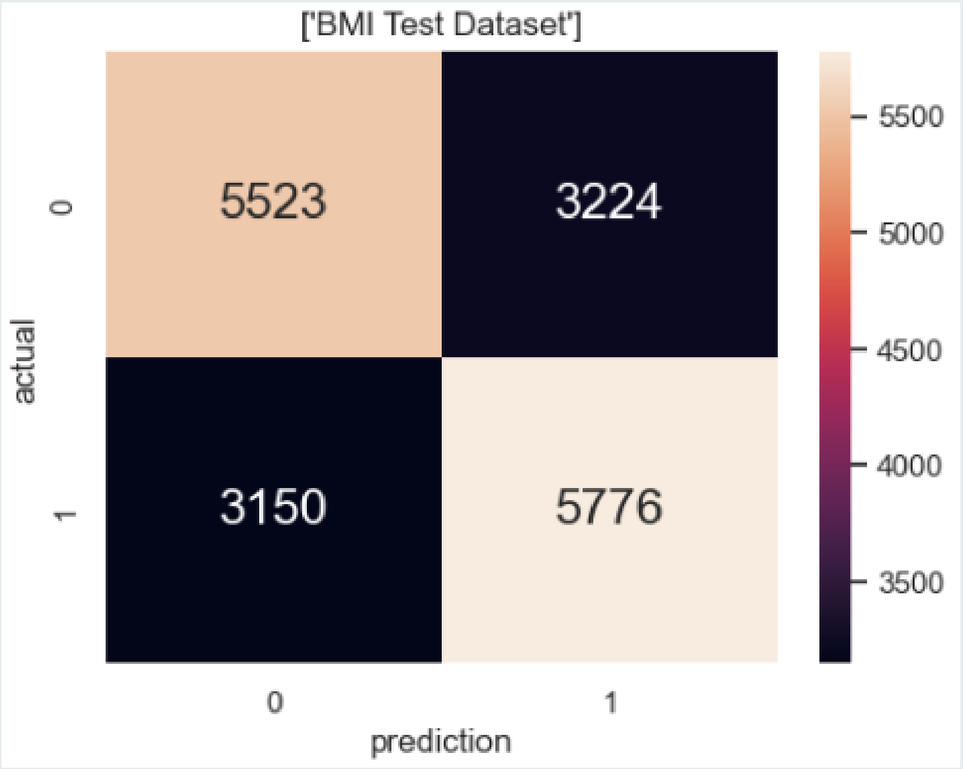
- A feedforward Artificial Neural Network, where neurons in each layer is connected to that in the next layer
- During the training process, the network learns to adjust the weight and biases
- Uses backpropagation



Adaptability of MLP allows it to model underlying patterns, even in the presence of skewed data.

Numerical Data Evaluation: Multi-Layer Perceptron





Classification Accuracy:

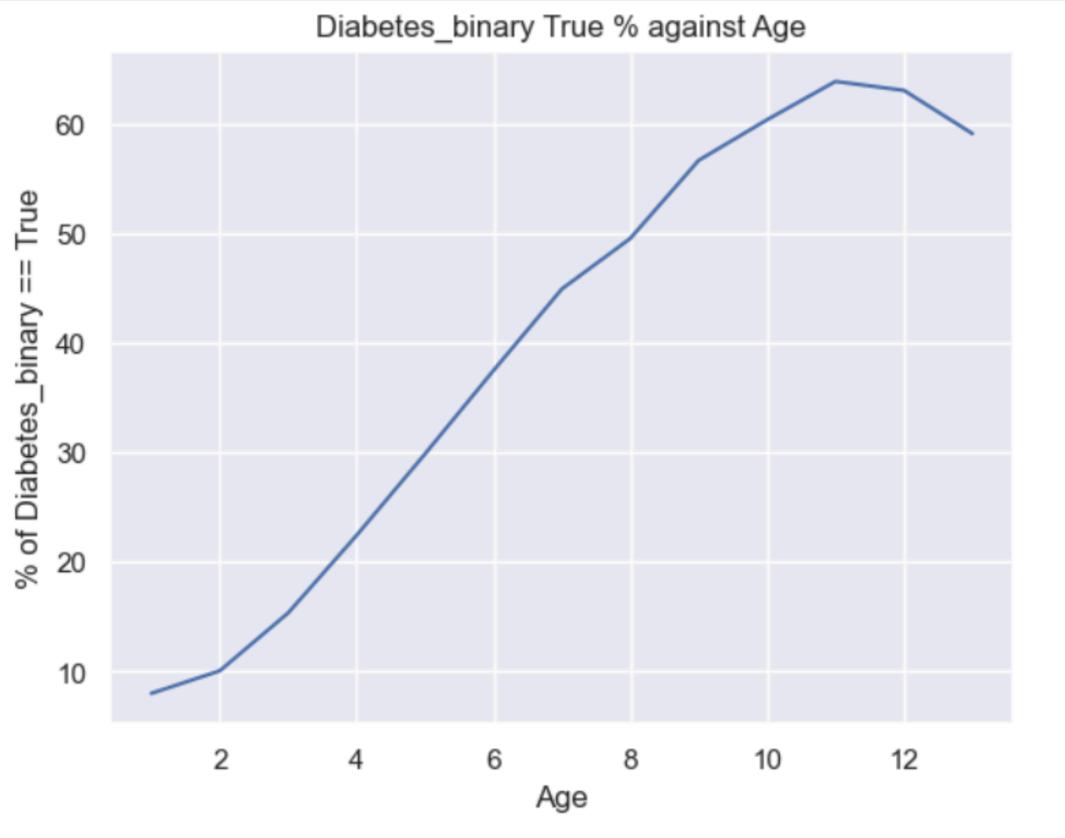
59%

64%

OneHot Data Analysis

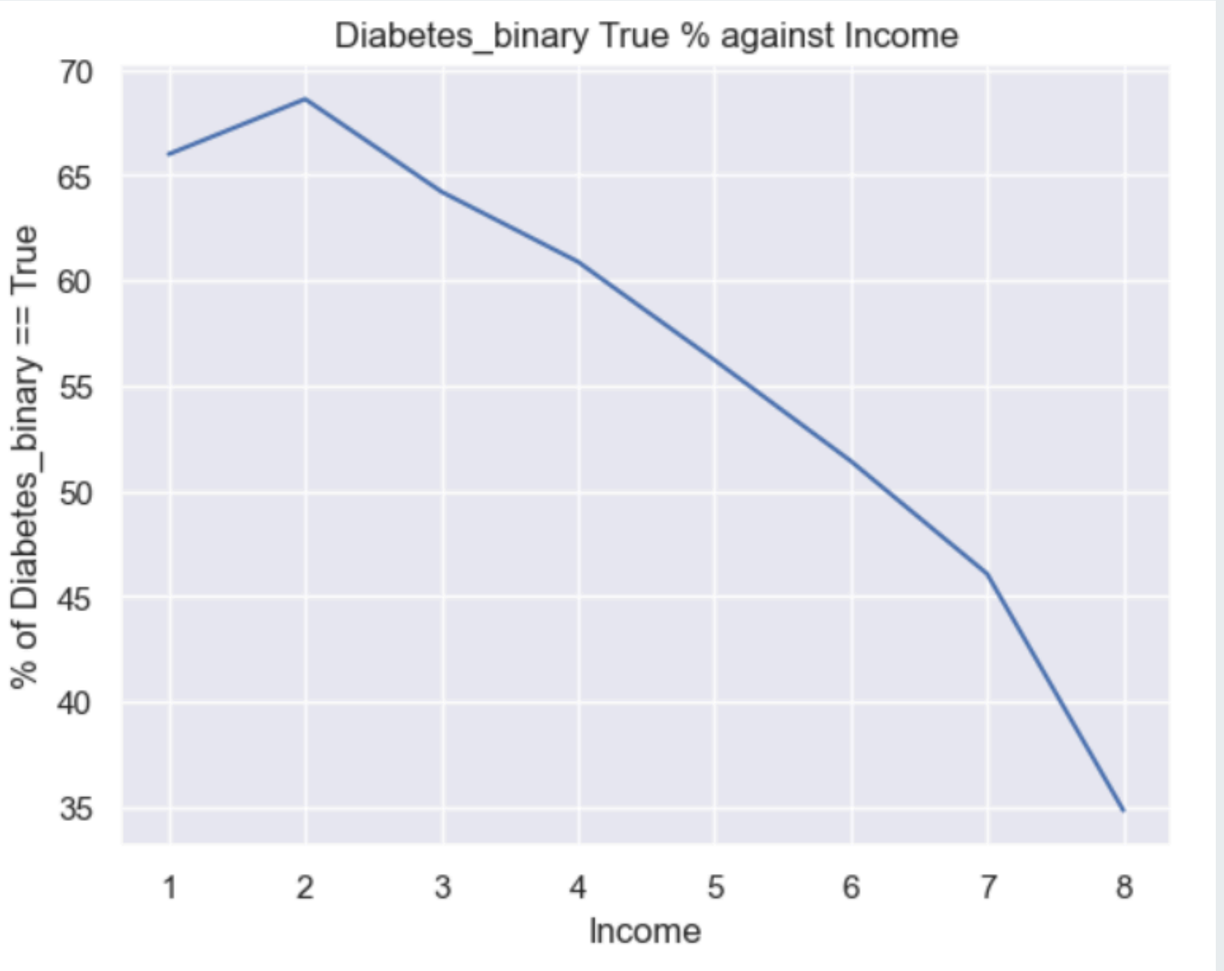
- Predictor Variable 1: Age
- Predictor Variable 2: GenHlth
- Predictor Variable 3: Income

One-Hot Data: Exploratory Data Analysis



Increasing Trend Observed for Age and GenHlth

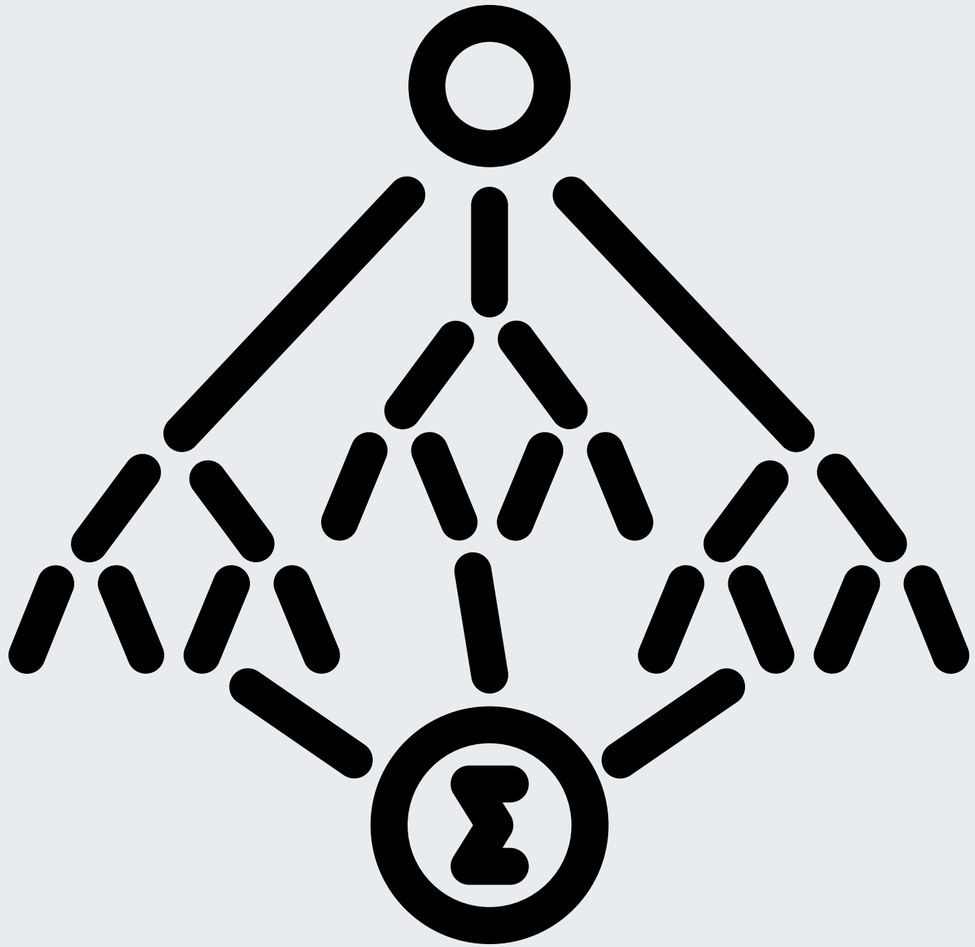
One-Hot Data: Exploratory Data Analysis



Decreasing Trend Observed for Income

One-Hot Data Analysis: Random Forest

- Bootstrap sample: Train-Test Set
- Feature bagging:
    - Reduces correlation among decision trees
    - Adds more diversity to dataset

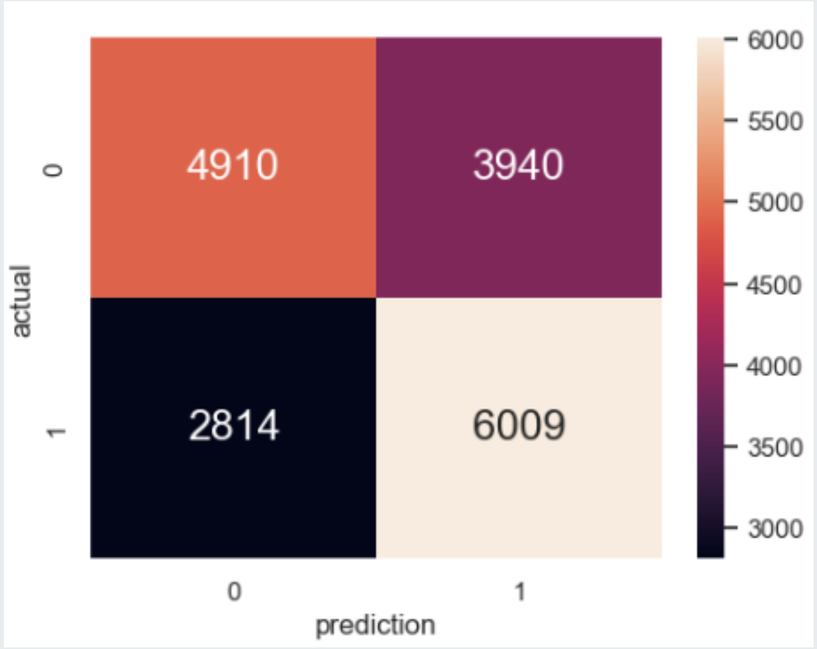


Advantages of Random Forest:

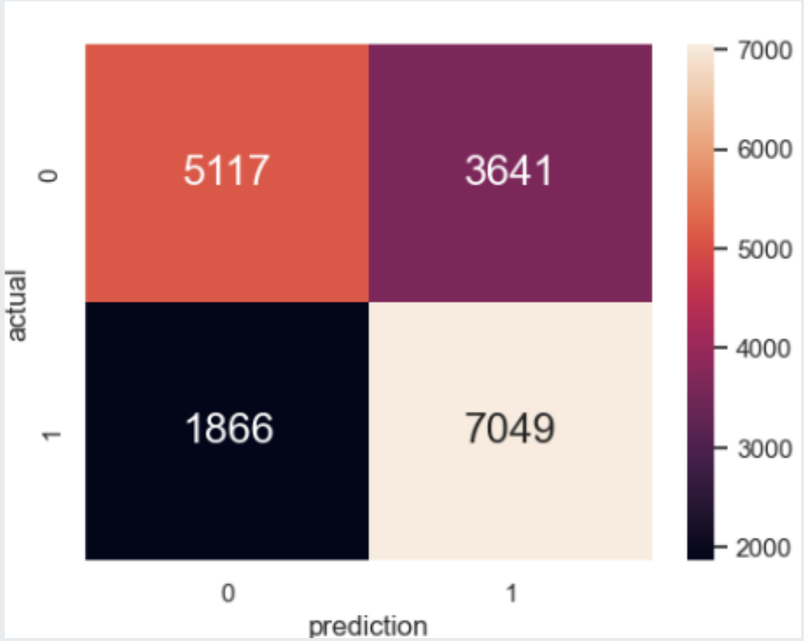
- Reduced risk of overfitting
- Able to handle classification tasks with high degree of accuracy

One-Hot Data Evaluation: Random Forest

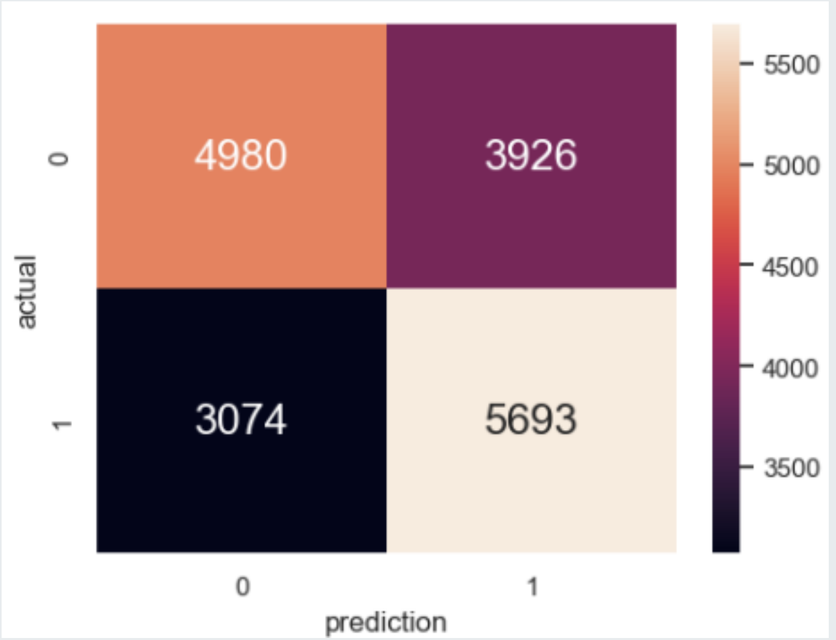
Age Test Dataset



GenHlth Test Dataset



Income Test Dataset



Classification Accuracy:

62%

69%

60%

Outcomes & Insights

Outcomes

Most significant factors of each category

Numerical: BMI
Binary: HighBP
One-hot: GenHlth

Insights

HighBP and BMI are factors that are focused on by medical professionals

GenHlth is a survey of how one perceives their own health status

Lack of exercise & unhealthy eating habits

leads to

high BMI and High BP, which in turn shows poor GenHlth

Suggestions

Lead a healthy lifestyle

↓

Includes exercise and healthy eating habits

↓

Improves general health

&

Lowers intrinsic factors such as BMI and HighBP

THANK YOU!