DIABETES PREDICTION

Machine Learning
Side Project

BY HUANG LIN CHUN WALLY

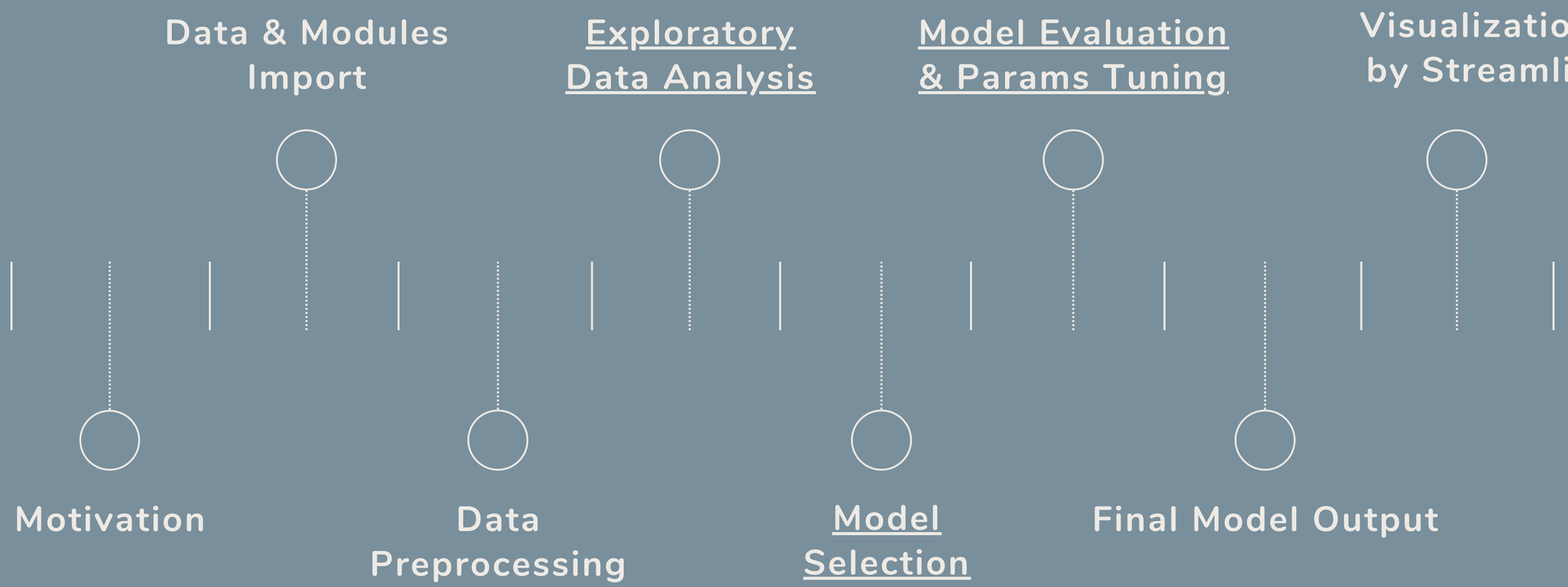
🔗 **link of the visualized interface (Streamlit):**
https://diabetesdetection-2e2nvdfhxvzucsbd3kvrfw.streamlit.app/

🔗 **link of my git hub:**
https://github.com/Wh4130/Diabetes_Prediction

HUANG LIN CHUN, Wally

CATALOG

🔗 link of the visualized interface (Streamlit):
https://diabetesdetection-2e2nvdfhxvzucsbd3kvrfw.streamlit.app/



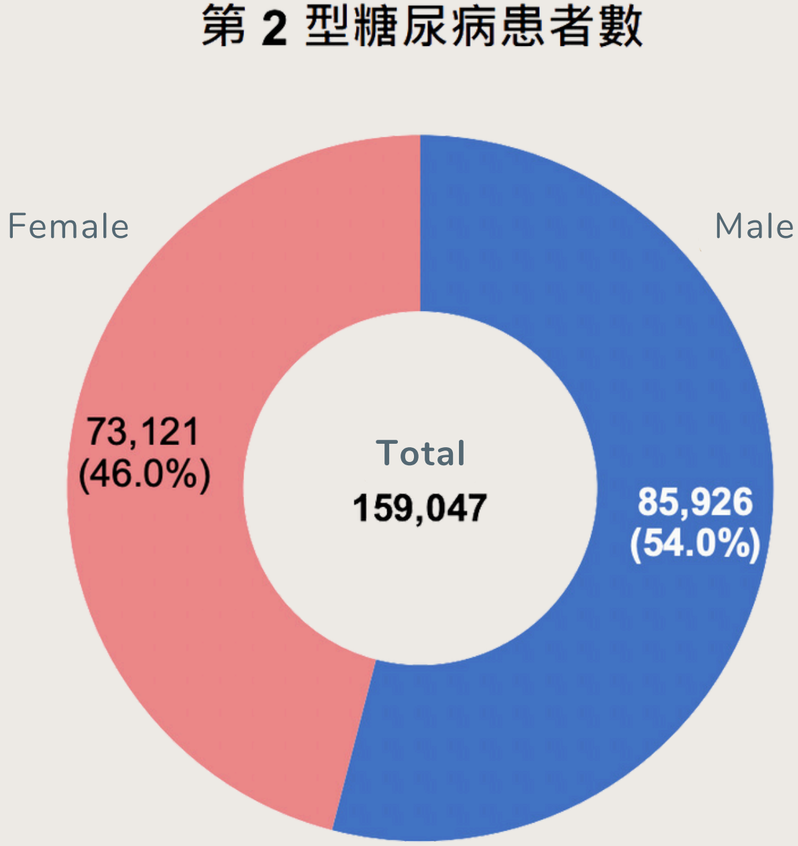
HUANG LIN CHUN, Wally

MOTIVATION

Taiwan is one of the few countries with the highest rate of Diabetes.

According to the government's report, diabetes patients account for 9% of total adults (18 - 64) in 2019, and this astonishingly high amount keeps increasing.

Data Source:
社團法人中華民國糖尿病衛教學會
Taiwanese Association of Diabetes Educators



Number of Patients of Type-2 Diabetes in Taiwan, 2019

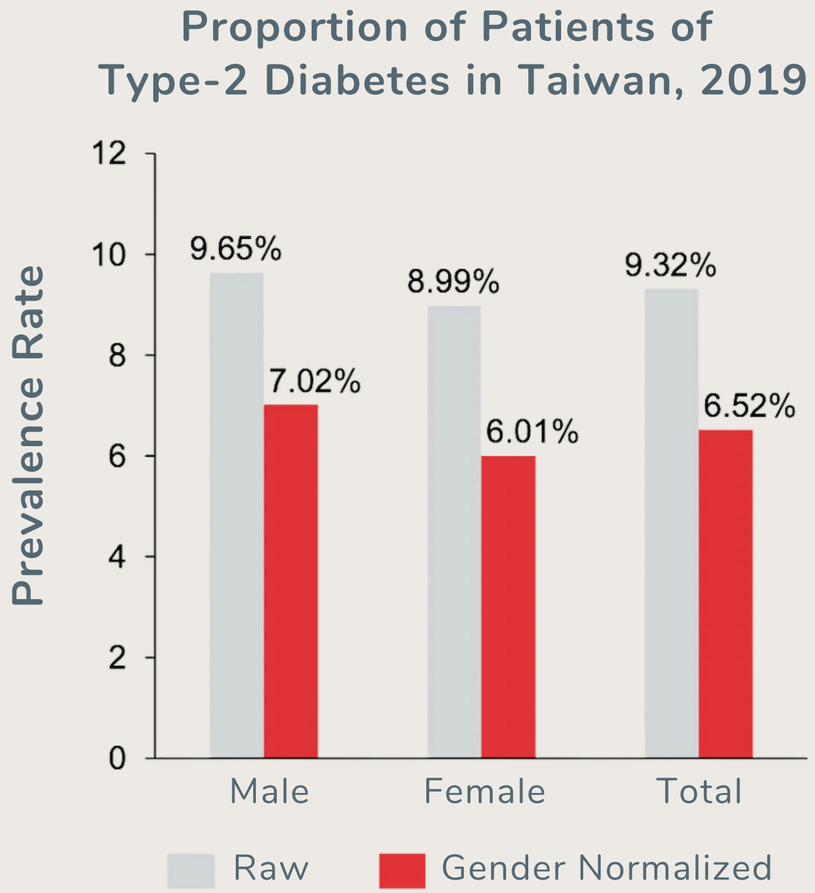
HUANG LIN CHUN, Wally

MOTIVATION

Taiwan is one of the few countries with the highest rate of Diabetes.

According to the government's report, diabetes patients account for 9% of total adults (18 - 64) in 2019, and this astonishingly high amount keeps increasing.

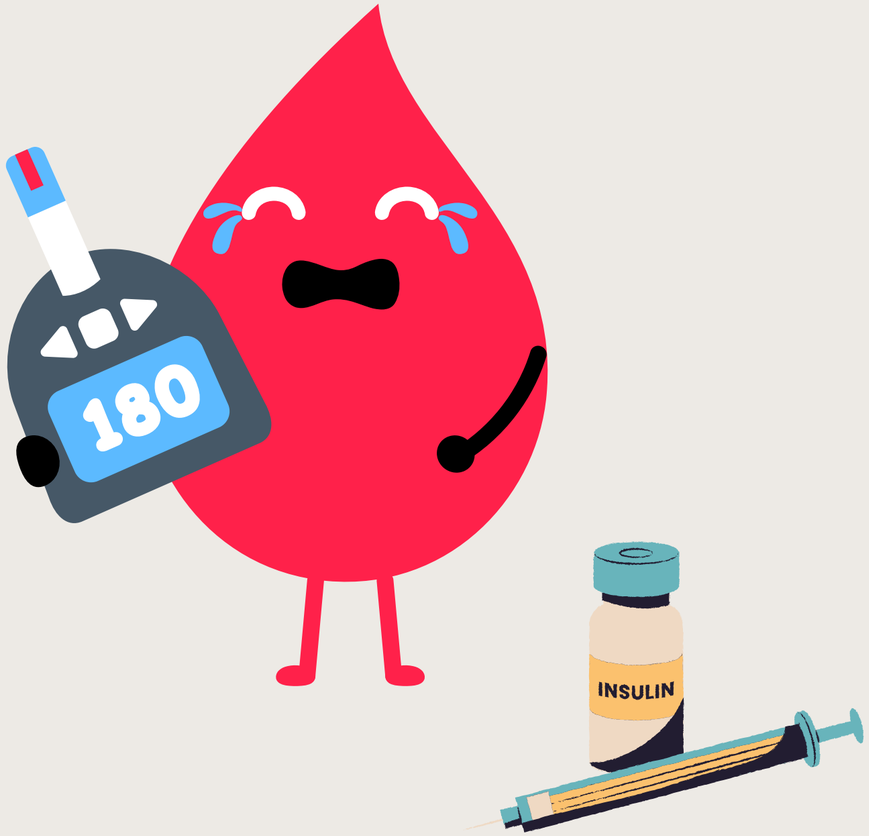
Data Source:
社團法人中華民國糖尿病衛教學會
Taiwanese Association of Diabetes Educators



Proportion of Patients of Type-2 Diabetes among all adults in Taiwan, 2019

MOTIVATION

In light of this, I was inspired to establish a data analysis and machine learning project based on Diabetes survey data on Kaggle.com and build up an interactive user interface to **make the public more aware of the risk and consequence of diabetes**.



🔗 link of the interface (Streamlit):
https://diabetesdetection-2e2nvdfhxvzucsbd3kvrfw.streamlit.app/

DATA & MODULE IMPORT

DATA DESCRIPTION

- Data Source: Kaggle.com → Diabetes Health Indicators Dataset (Survey Data)
- Features

Target

Diabetes Binary
True = 1

21 Features

High Blood Pressure, High Cholesterol, Cholesterol Check, BMI, Smoker, Stroke, Sex, Heart Diseaseor Attack, Physical Activity past 30d, Fruits, Age, Vegetables, Heavy Alcohol Consumption, Any Healthcare Coverage, Not See Doctor due to costs, General Health, Mental Health, Physical Health, Difficulty Walking, Income, Education

DATA & MODULE IMPORT

REQUIRED MODULE & SELECTED ML MODELS

MODULES

Pandas Numpy
Matplotlib pyplot
Seaborn Sklearn
Pickle Streamlit
Gzip

4 Candidate Models

Logistic Regression

Random Forest

Gradient Boosting DT

Neural Network

←

Because:

1. Many binary features
   → Data Scaling Not Proper
2. All these four models
   are versatile and prominent

HUANG LIN CHUN, Wally

DATA PREPROCESSING

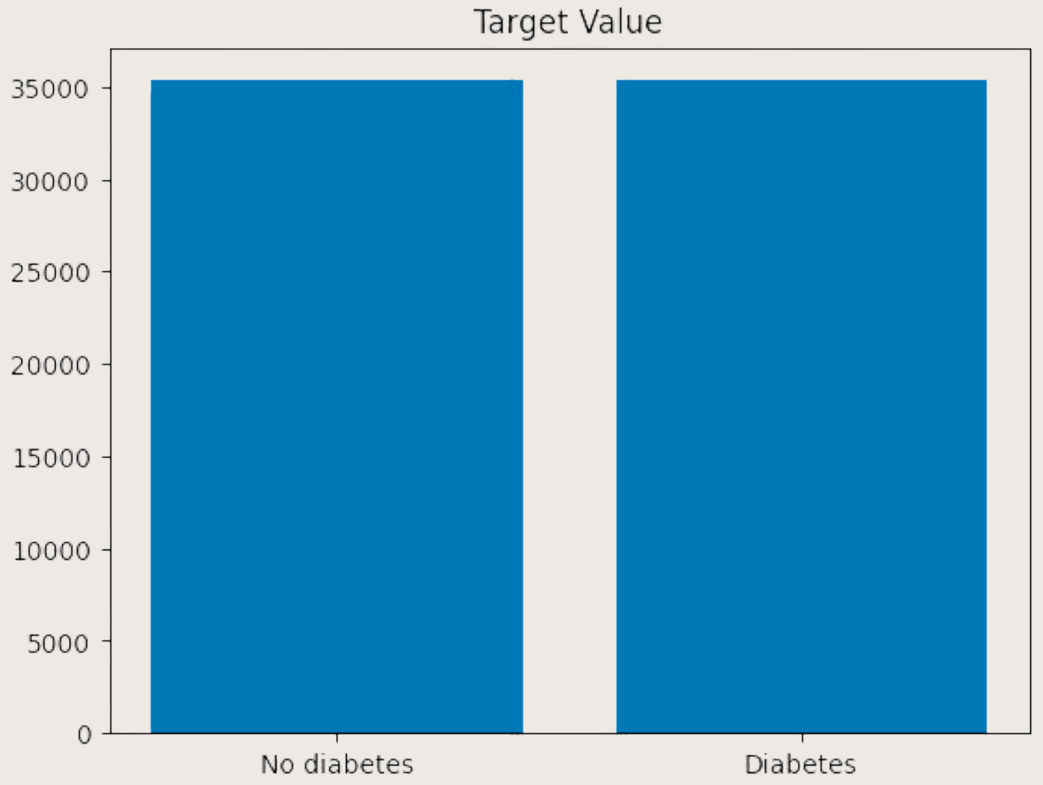
This dataset was already cleaned and available on kaggle

CHECK IF THE DATA IS CLEAN

```
data.isnull().sum()
```

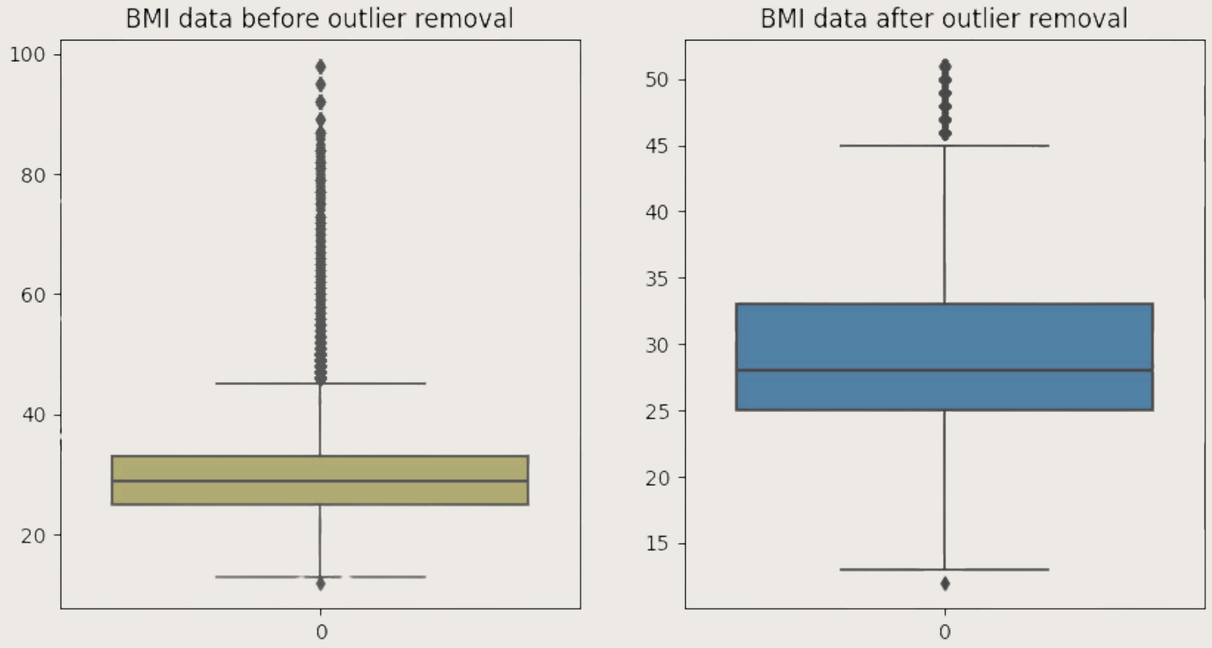
Diabetes_binary	0	HvyAlcoholConsump	0
HighBP	0	AnyHealthcare	0
HighChol	0	NoDocbcCost	0
CholCheck	0	GenHlth	0
BMI	0	MentHlth	0
Smoker	0	PhysHlth	0
Stroke	0	DiffWalk	0
HeartDiseaseorAttack	0	Sex	0
PhysActivity	0	Age	0
Fruits	0	Education	0
Veggies	0	Income	0

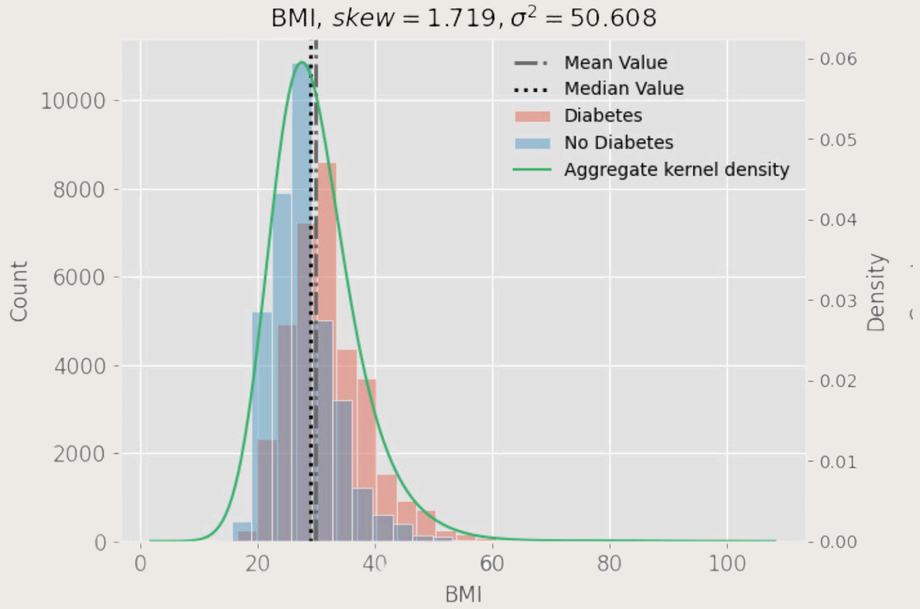
CHEK IF THE DATA IS BALANCED

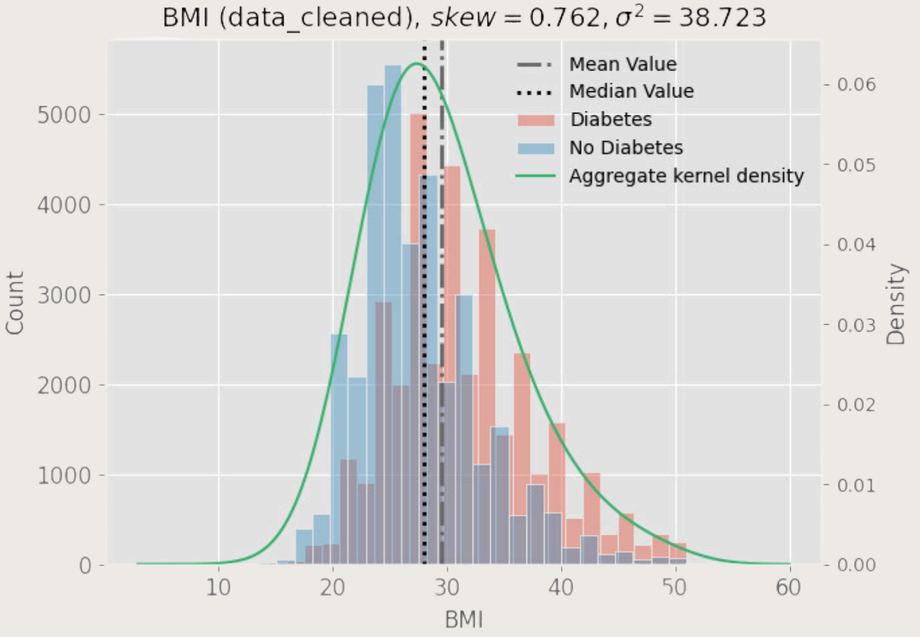


DATA PREPROCESSING

REMOVE OUTLIERS FOR **BMI**





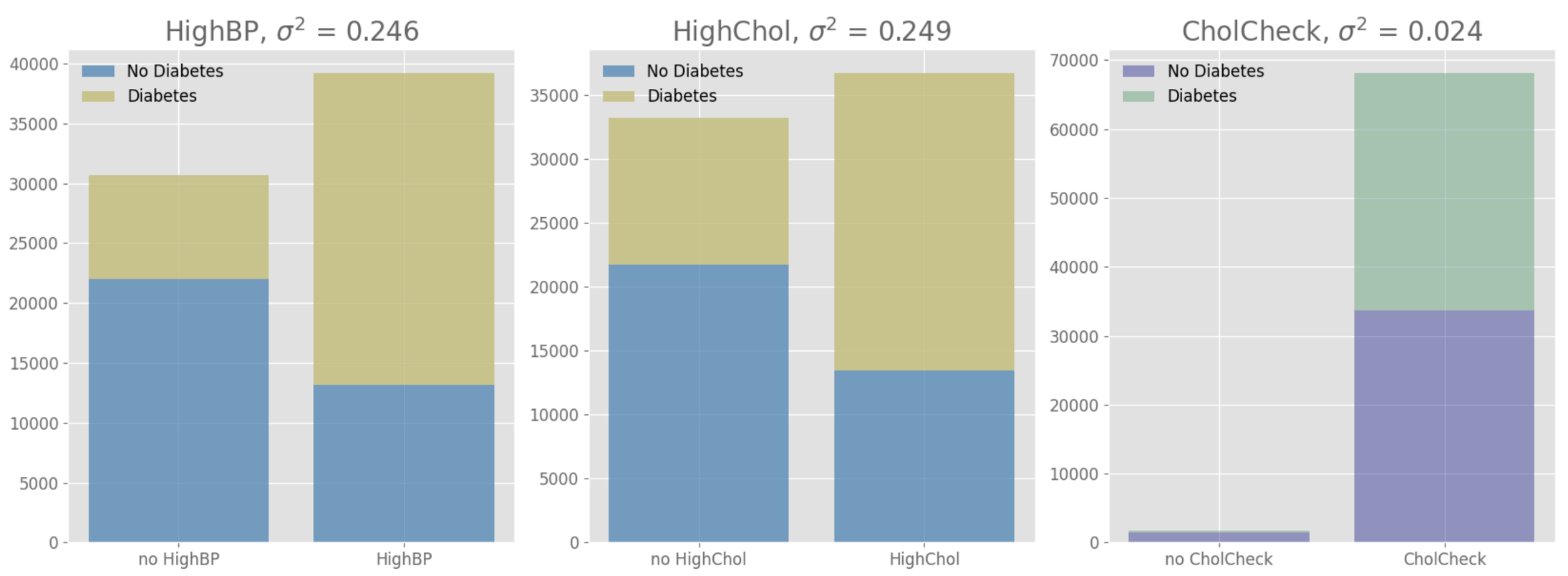


HUANG LIN CHUN, Wally

EXPLORATORY DATA ANALYSIS

BINARY FEATURES

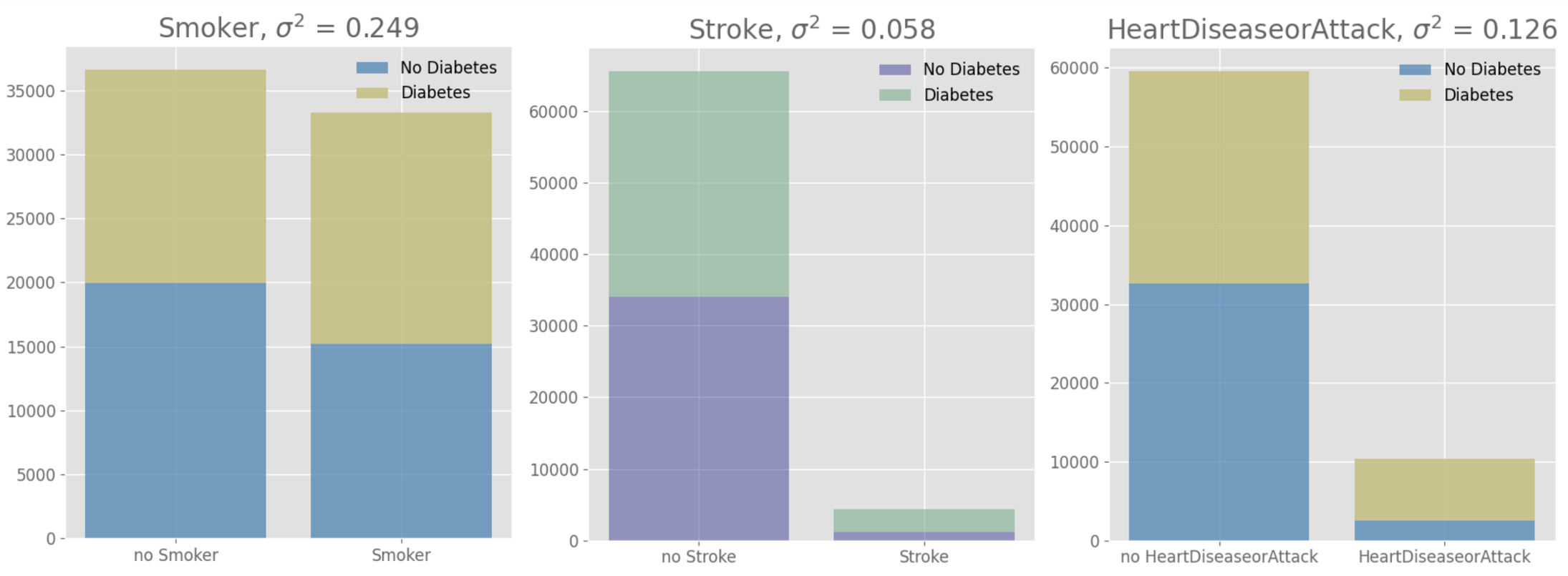
Features with variance < 0.09 were highlighted by green



EXPLORATORY DATA ANALYSIS

BINARY FEATURES

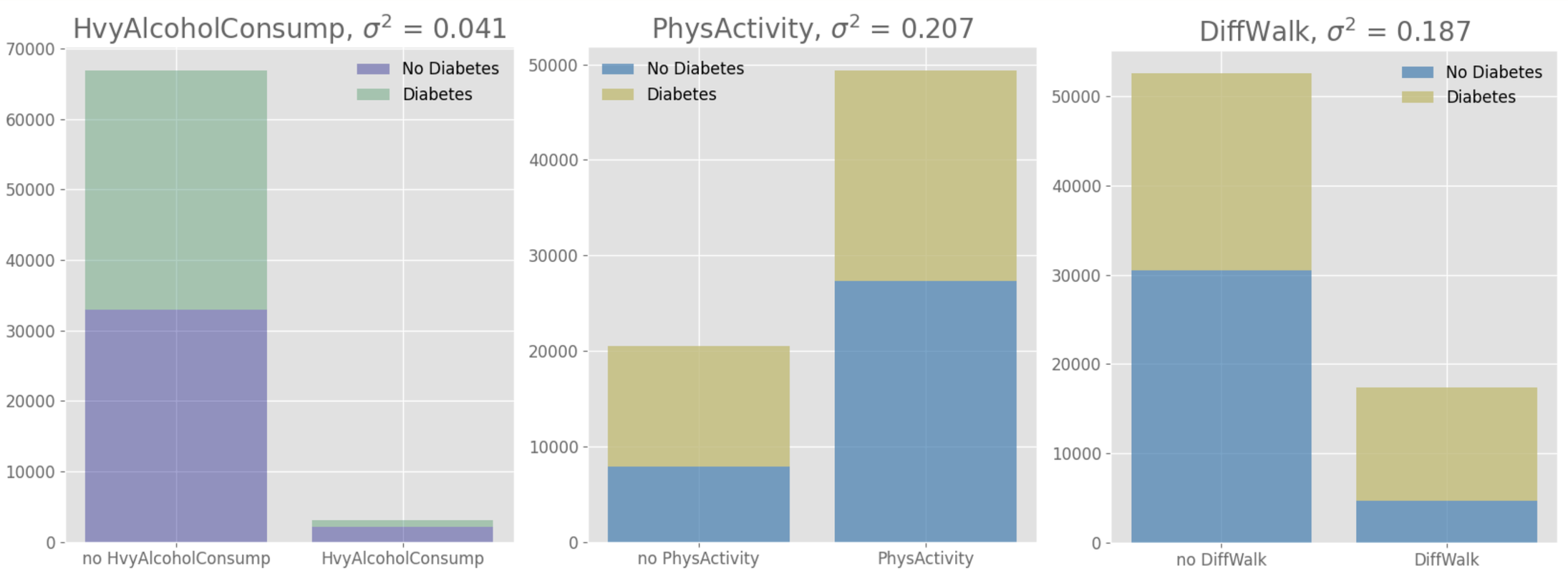
Features with variance < 0.09 were highlighted by green



EXPLORATORY DATA ANALYSIS

BINARY FEATURES

Features with variance < 0.09 were highlighted by green

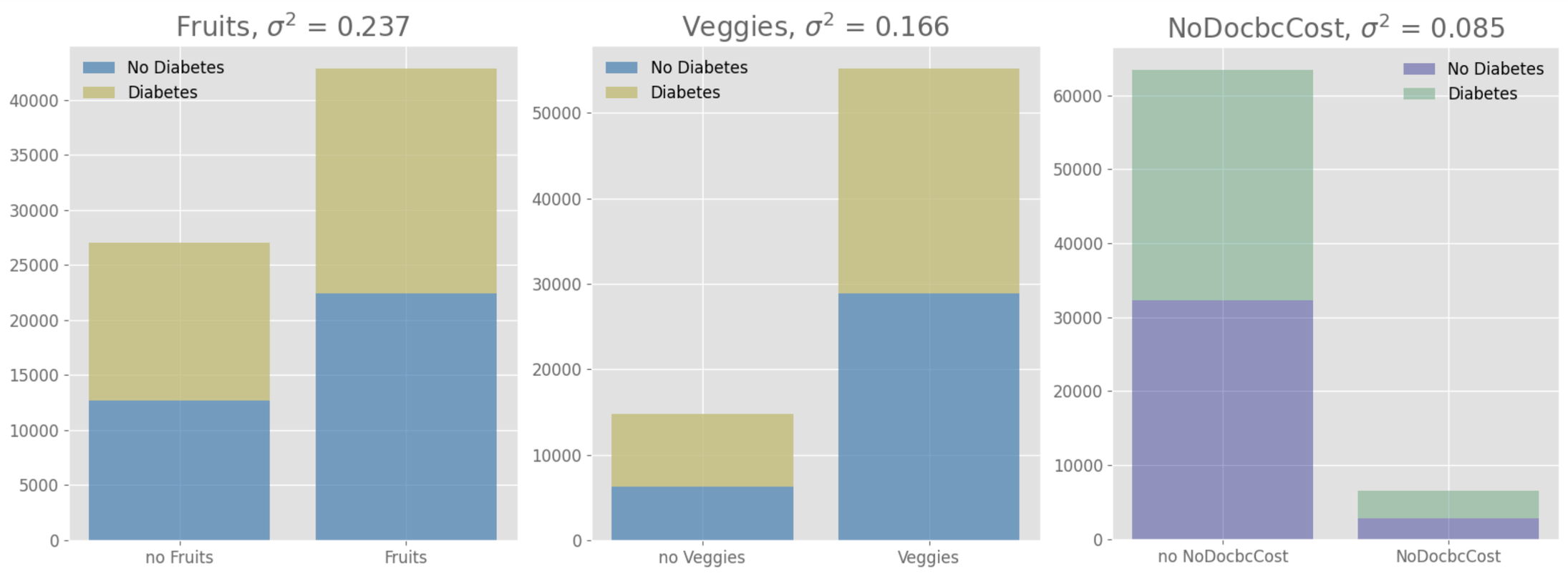


HUANG LIN CHUN, Wally

EXPLORATORY DATA ANALYSIS

BINARY FEATURES

Features with variance < 0.09 were highlighted by green

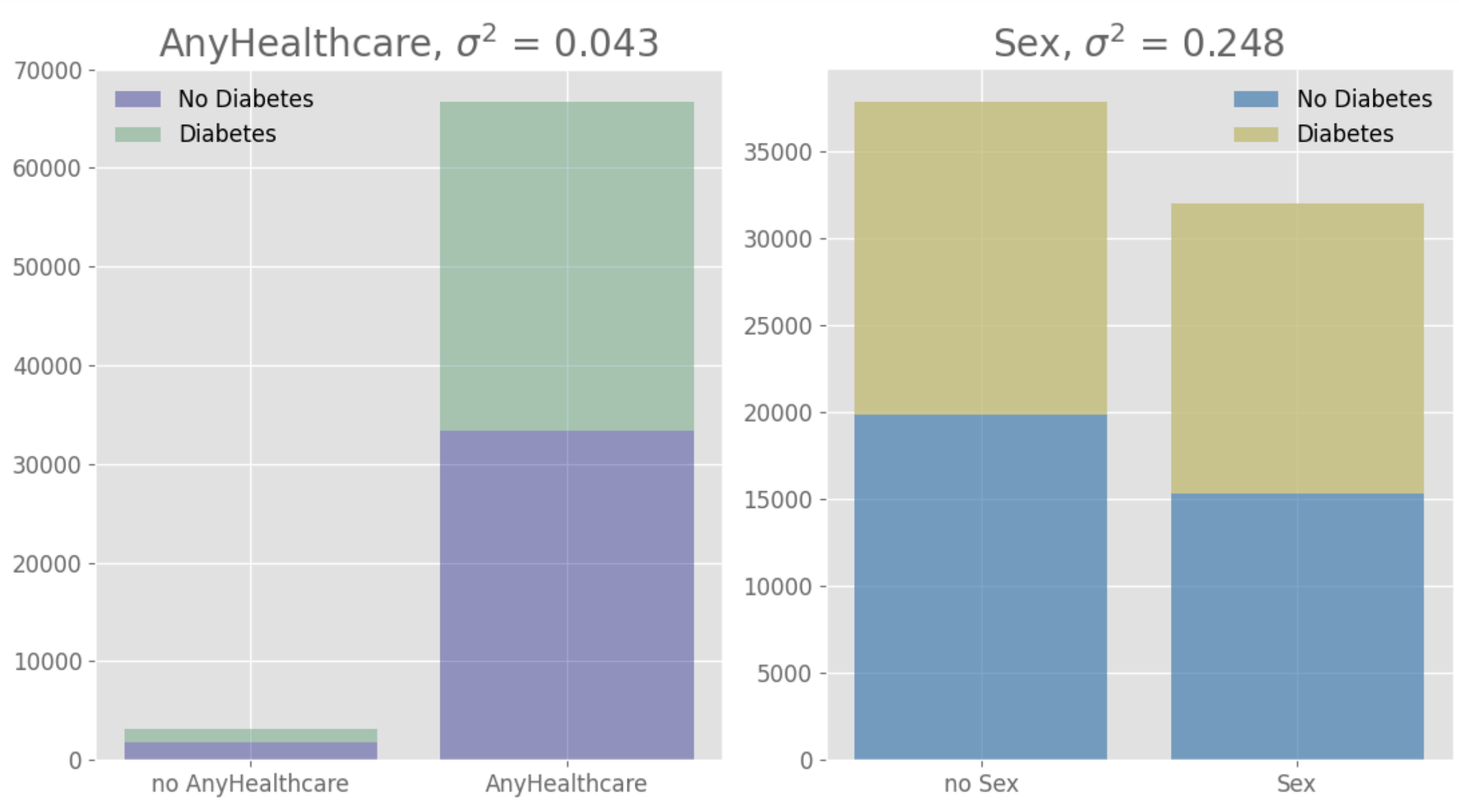


HUANG LIN CHUN, Wally

EXPLORATORY DATA ANALYSIS

BINARY FEATURES

Features with variance < 0.09 were highlighted by green



Features **“Cholesterol Check”, “Stroke”, “Heavy Alcohol Consumption”, “Not See Doctor due to costs”,** and **“Any Healthcare coverage”** have too low variance, which means that they might not be good features.

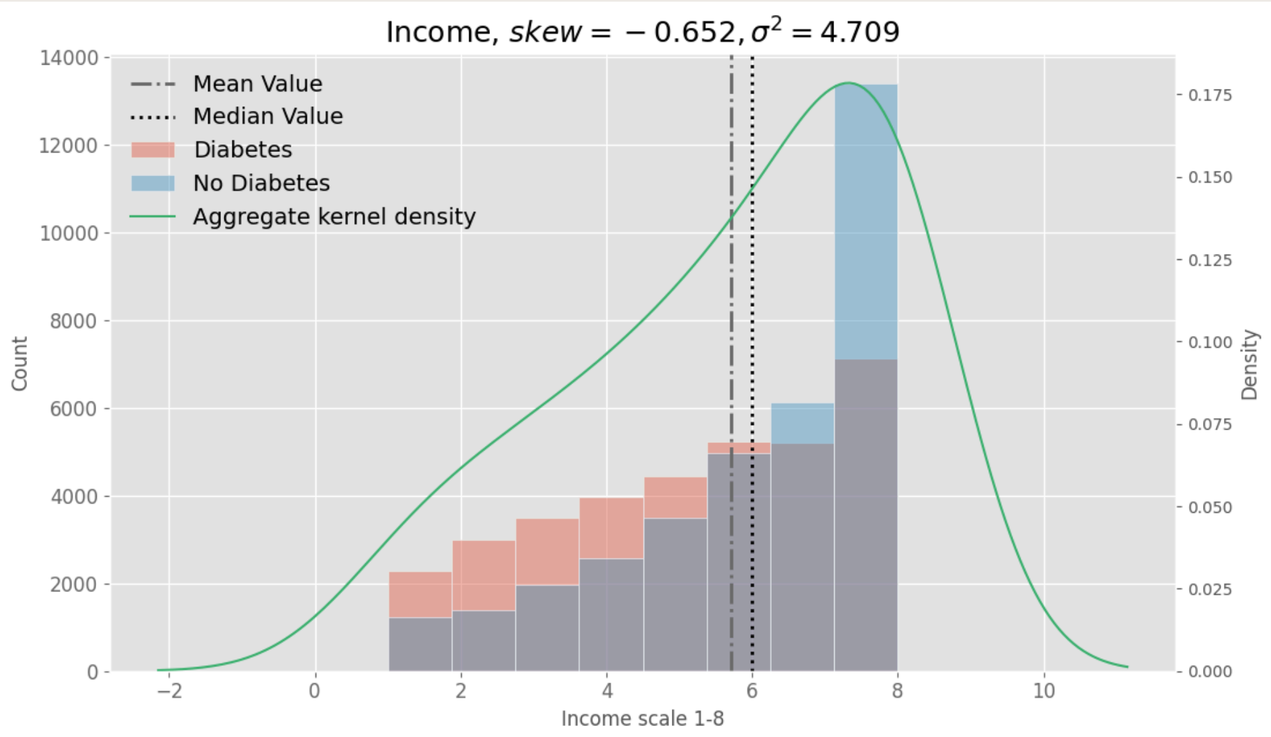
→ We conduct further analysis to determine whether to eliminate these features from our model.

- **Correlation Analysis**
- **Feature Importance Analysis**

HUANG LIN CHUN, Wally

EXPLORATORY DATA ANALYSIS

NON BINARY FEATURES



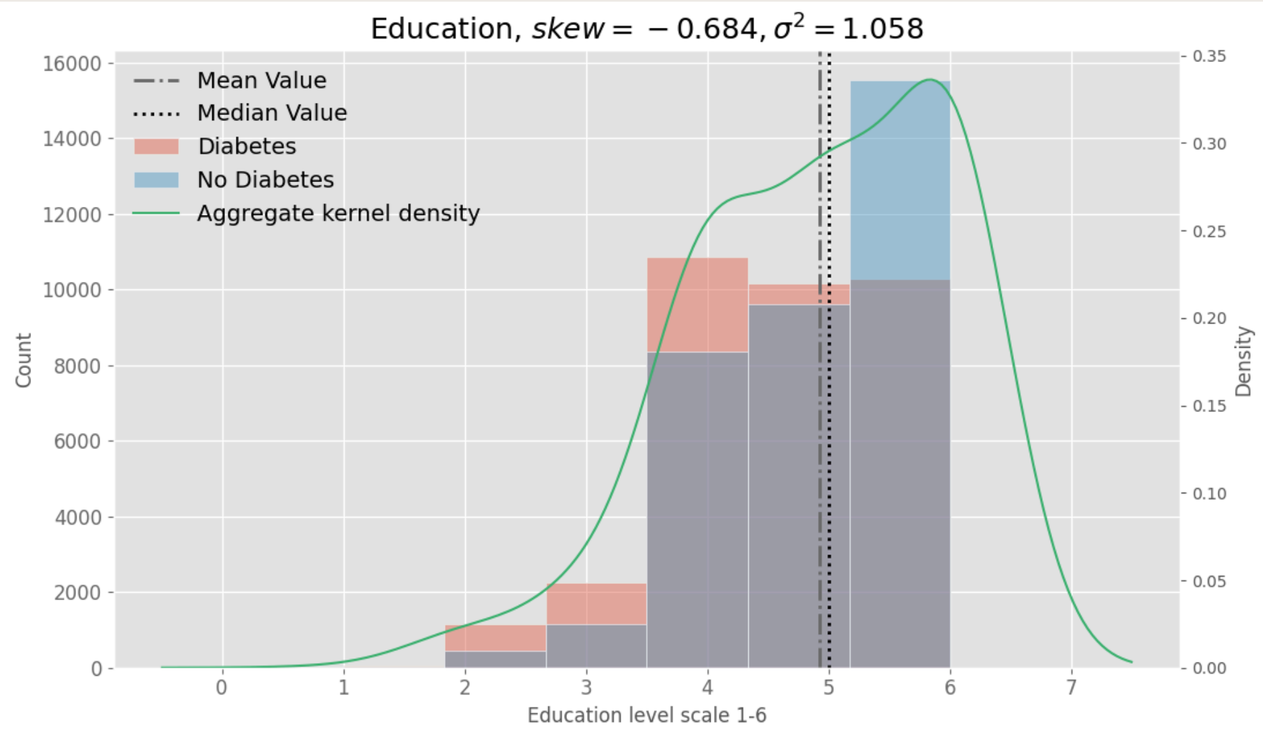
INCOME

- 8 scales income level → Not scalable
- Most of the respondents are in the highest-level group (left-skewed) → biased feature
- It does not have direct predictability on target (more like a confounder)

→ We eliminate Income from our model

EXPLORATORY DATA ANALYSIS

NON BINARY FEATURES



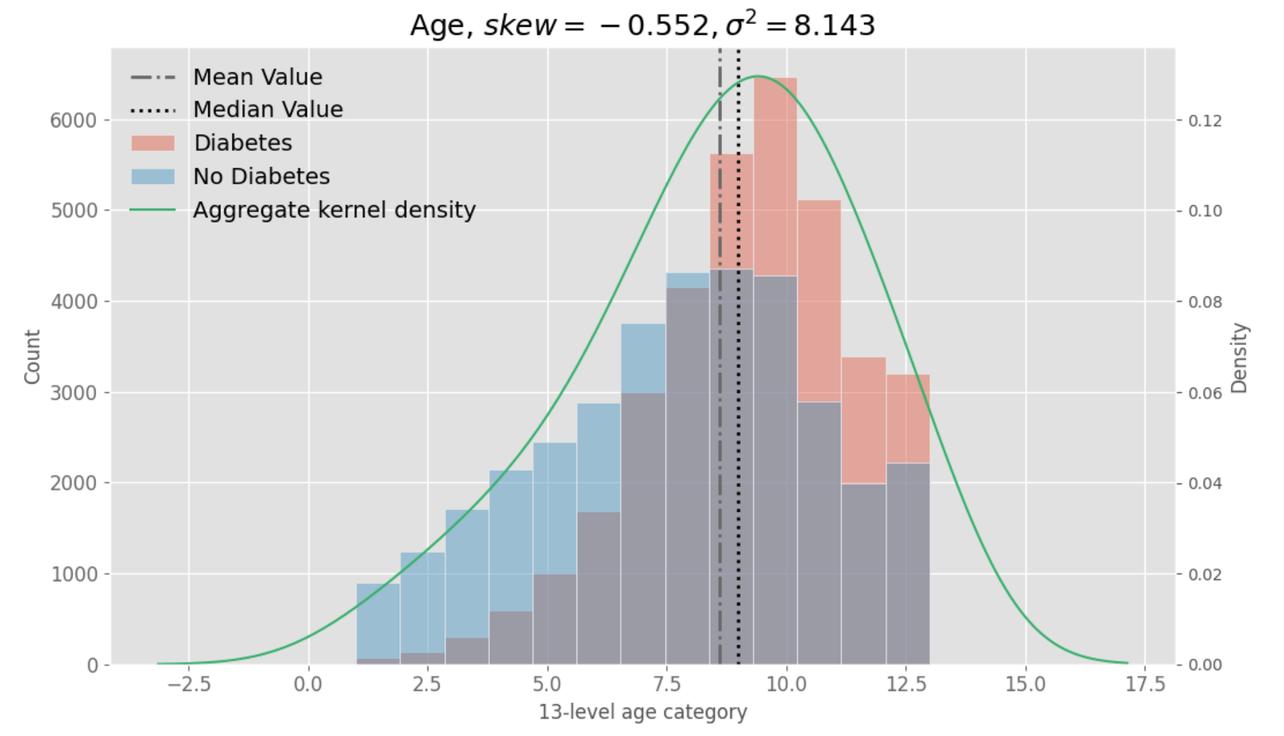
EDUCATION

- 6 scales education level → Not scalable
- Most of the respondents are in the highest-educated group (left-skewed) → biased feature
- It does not have direct predictability on target (more like a confounder)

→ We eliminate Education from our model

EXPLORATORY DATA ANALYSIS

NON BINARY FEATURES



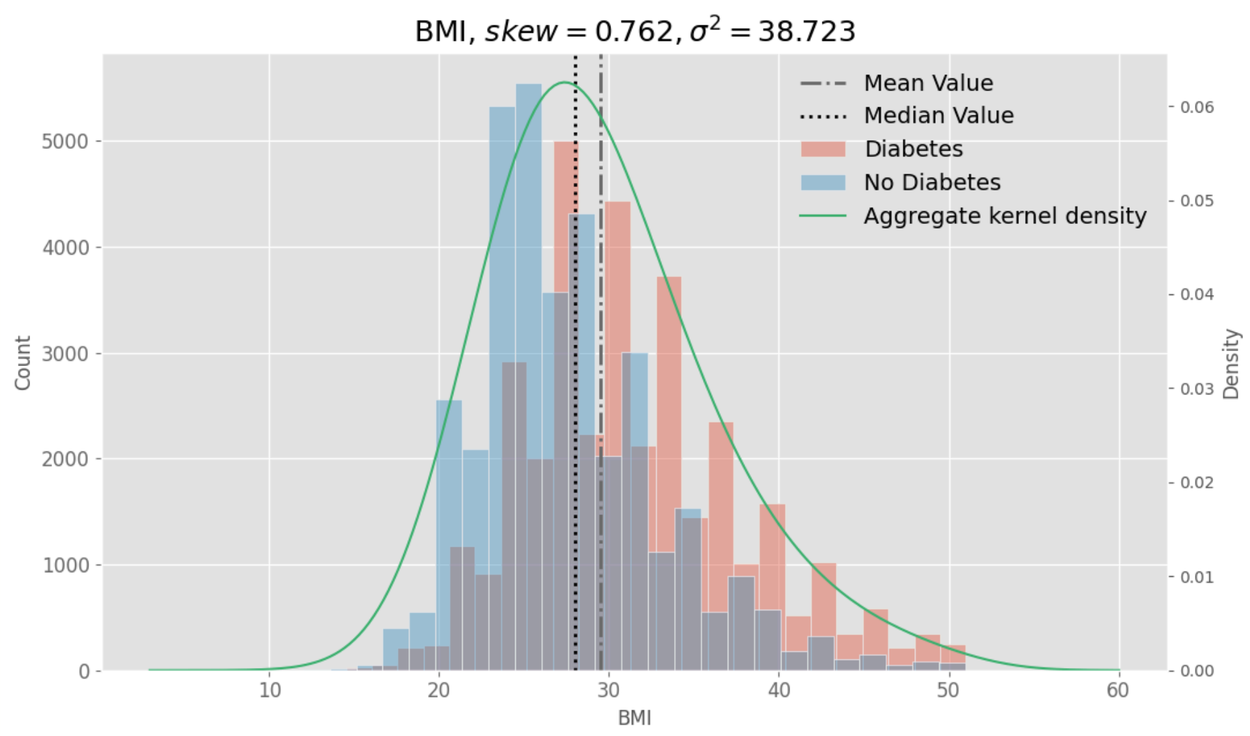
AGE

- 13 scales age level → Not scalable
- Skewness is not severely high
- Variance is large enough
- It separates diabetic and non-diabetic targets well

→ We include Age in our model

EXPLORATORY DATA ANALYSIS

NON BINARY FEATURES



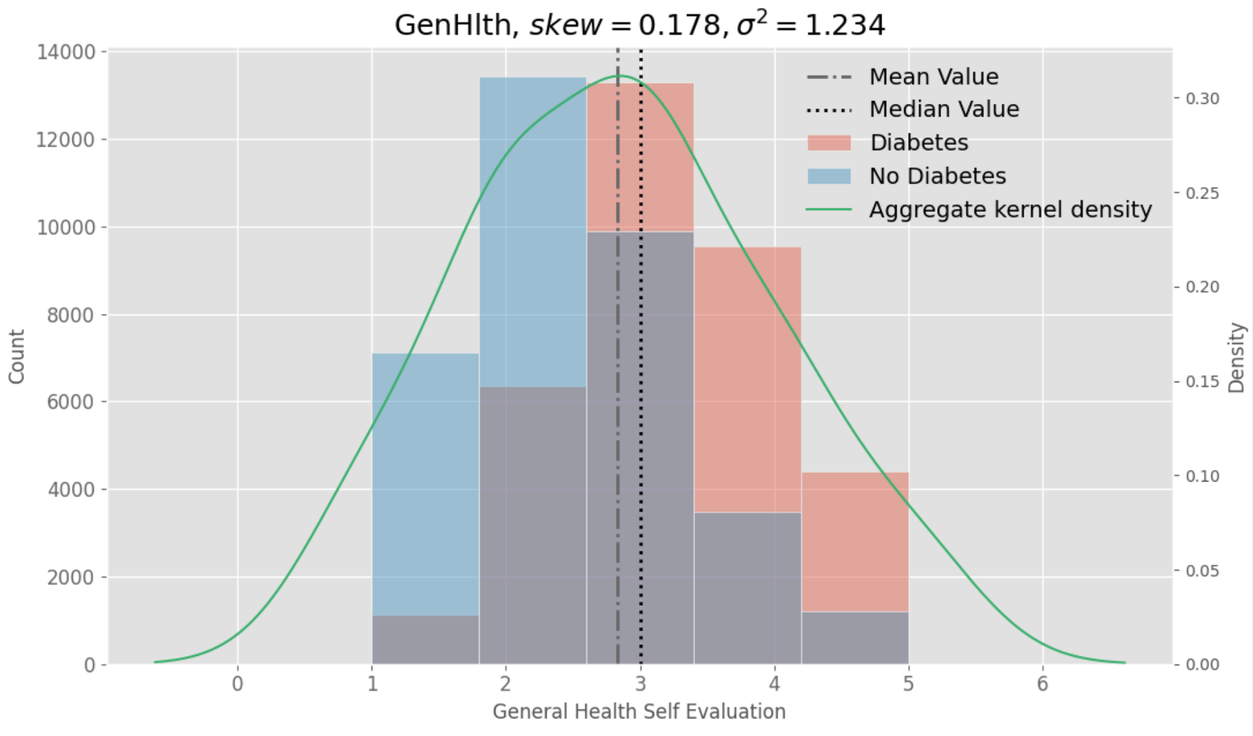
BMI

- Not scaled origin BMI data
- Skewness is not severely high after the removal of outliers
- Variance is large enough
- It separates diabetic and non-diabetic targets well

→ **We include BMI in our model**

EXPLORATORY DATA ANALYSIS

NON BINARY FEATURES



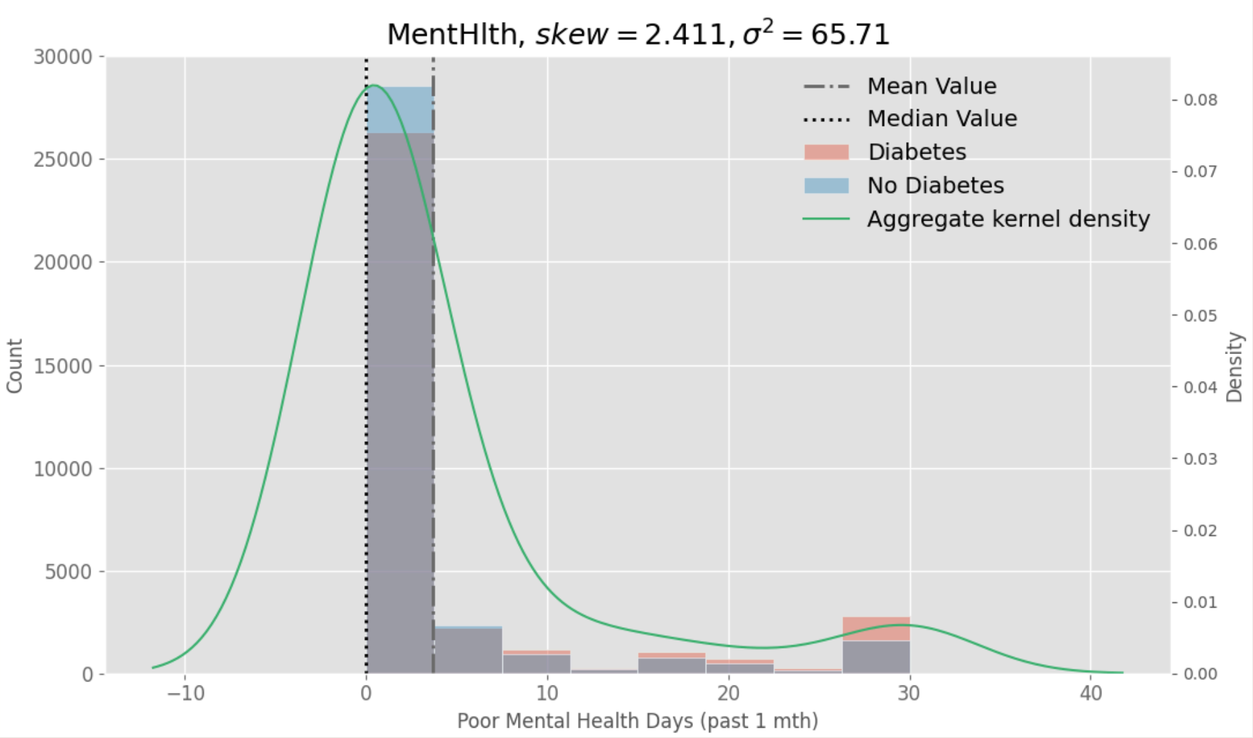
GENERAL HEALTH

- Subjective and abstract self-evaluation
  → likely to be biased (hmm...)
  1 = excellent ; 5 = poor
- Skewness is low
- It separates diabetic and non-diabetic targets well

→ We include GenHlth in our model

EXPLORATORY DATA ANALYSIS

NON BINARY FEATURES



MENTAL HEALTH

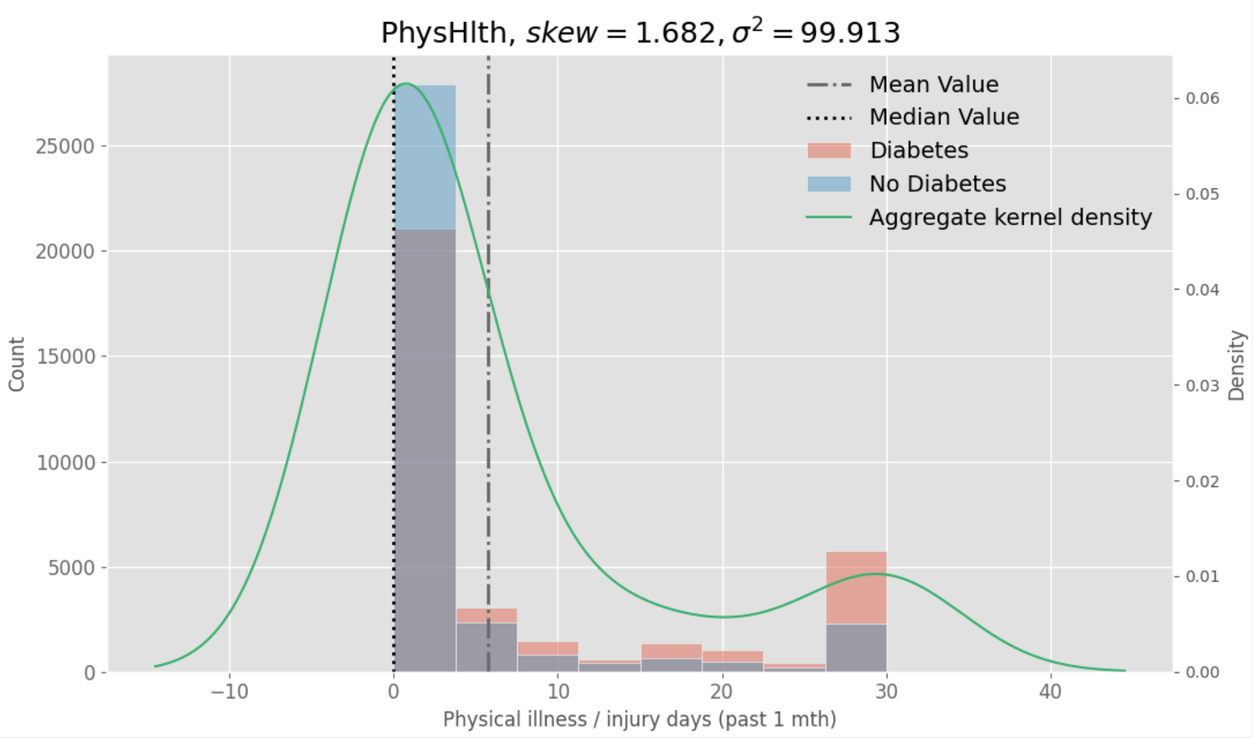
- Quantitative self-evaluation
  How many days during the past 30 days was your mental health not good?
- Severely right skewed
- It does not separate diabetic and non-diabetic targets well

→ We consider if including it to our model

HUANG LIN CHUN, Wally

EXPLORATORY DATA ANALYSIS

NON BINARY FEATURES



PHYSICAL HEALTH

- Quantitative self-evaluation
  How many days during the past 30 days was your physical health not good?
- Severely right skewed
- It separates diabetic and non-diabetic targets well, but might be highly correlated with **General Health**

→ **We consider if including it to our model**

EXPLORATORY DATA ANALYSIS

CORREALTION ANALYSIS

Check if there exists multicollinearity issue

HEATMAP

- General Health is highly correlated with
  - Mental Health
  - Physical Health
  - Difficulty Walking

- But the coefficients of correlation are lower or not too larger than 0.5

→ Acceptable

HUANG LIN CHUN, Wally

EXPLORATORY DATA ANALYSIS

CORREALTION ANALYSIS

Check if there exists multicollinearity issue

HEATMAP

- Income and Education are highly correlated with each other, as expected
- Income and Education are also correlated with General Health

→ Correlation through other feathers that are not observable here

→ Not a good idea to include confounders

EXPLORATORY DATA ANALYSIS

CORREALTION ANALYSIS

CORRELATION WITH TARGET

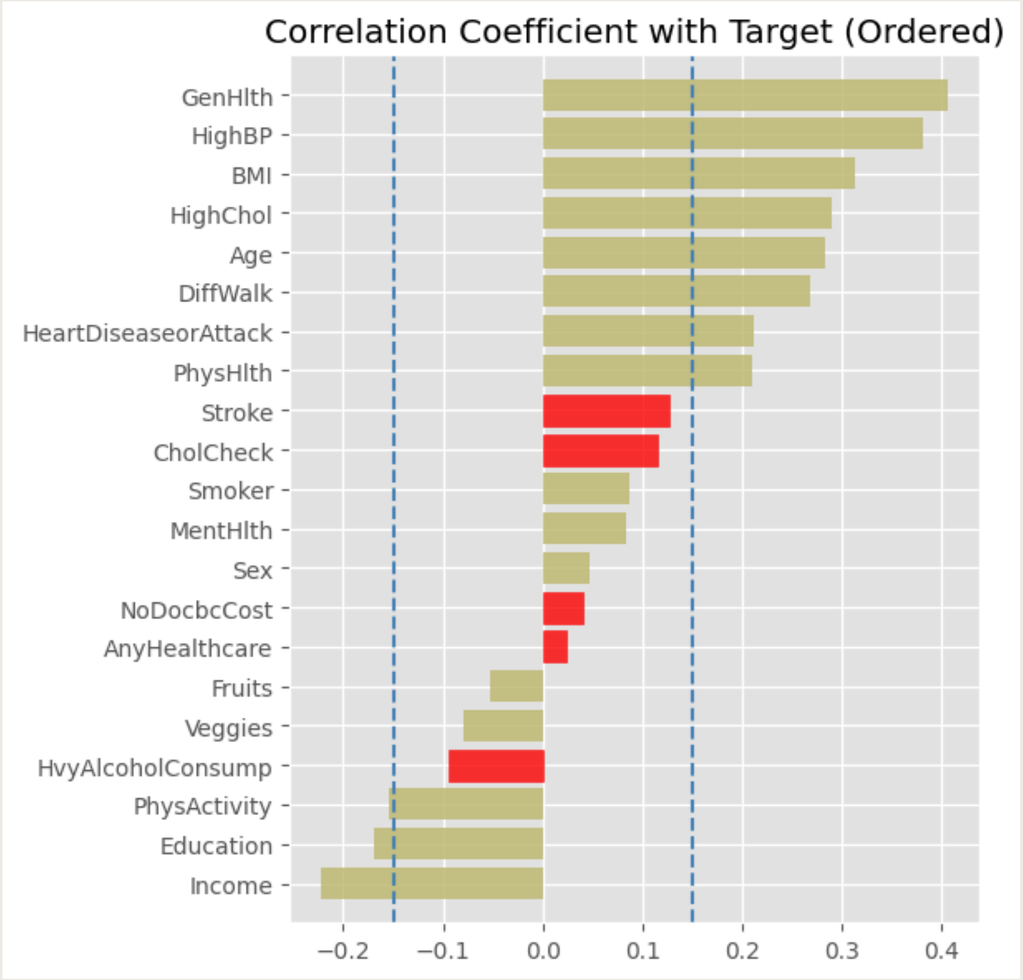
- General Health, BMI, and other main health indicators have high correlation with the target

- Five binary features which have low variances are subtly correlated with the target

  Cholesterol Check | Stroke | Heavy Alcohol Consumption
  Not See Doctor due to costs | Any Healthcare coverage

  - Among which, stroke has the highest correlation
  - And stroke is intuitively correlated with the target

  → We include **stroke**



EXPLORATORY DATA ANALYSIS

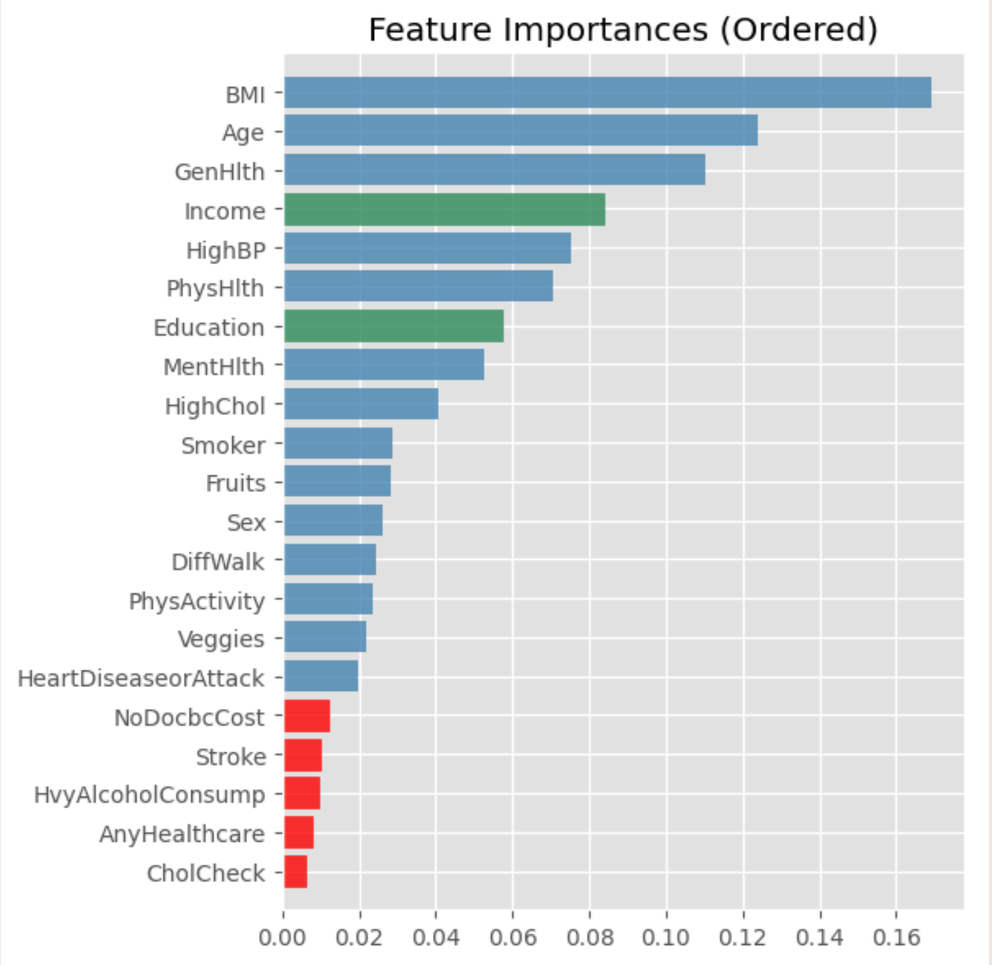
FEATURE IMPORTANCE

RANDOM FOREST FEATURE IMPORTANCES

- Even Income and Education (green bars) have high importances, we exclude them based on the reasons stated before

- The five binary features (red bars) that have low variances and subtly correlated with the target also have the lowest feature importances

→ Eliminate these 7 features



HUANG LIN CHUN, Wally

EXPLORATORY DATA ANALYSIS

FEATURE SELECTION

Based on :

- Variance analysis
- Correlation Analysis
- Feature Importances Analysis
- Other Causal Inference Rationality

→ 6 features are ruled out

15 Features Selected

High Blood Pressure
High Cholesterol
Cholesterol Check
BMI
Smoker
Stroke
Sex
Heart Diseaseor Attack
Physical Activity past 30d
Fruits
Age
Vegetables
Heavy Alcohol Consumption
Any Healthcare Coverage
Not See Doctor due to costs
General Health
Mental Health
Physical Health
Difficulty Walking
Income
Education

MODEL SELECTION

CROSS VALIDATION

- We evaluated four metrics through cross validation across four models
- Among four metrics, we focus on recall (health issue → **minimize false negative** rate

Metrices

Precision

F1_macro

- Recall

ROC_AUC

Model Comparisons based on 5 fold Cross Validation Scoring

	precision	recall	f1_macro	roc_auc
Random Forest Classifier	0.707228	0.754944	0.722935	0.790553
Logistic Regression	0.733449	0.761428	0.744071	0.821454
XGBoost (GBDT)	0.727366	0.788696	0.747969	0.826001
Neural Network	0.732284	0.769098	0.743483	0.824334

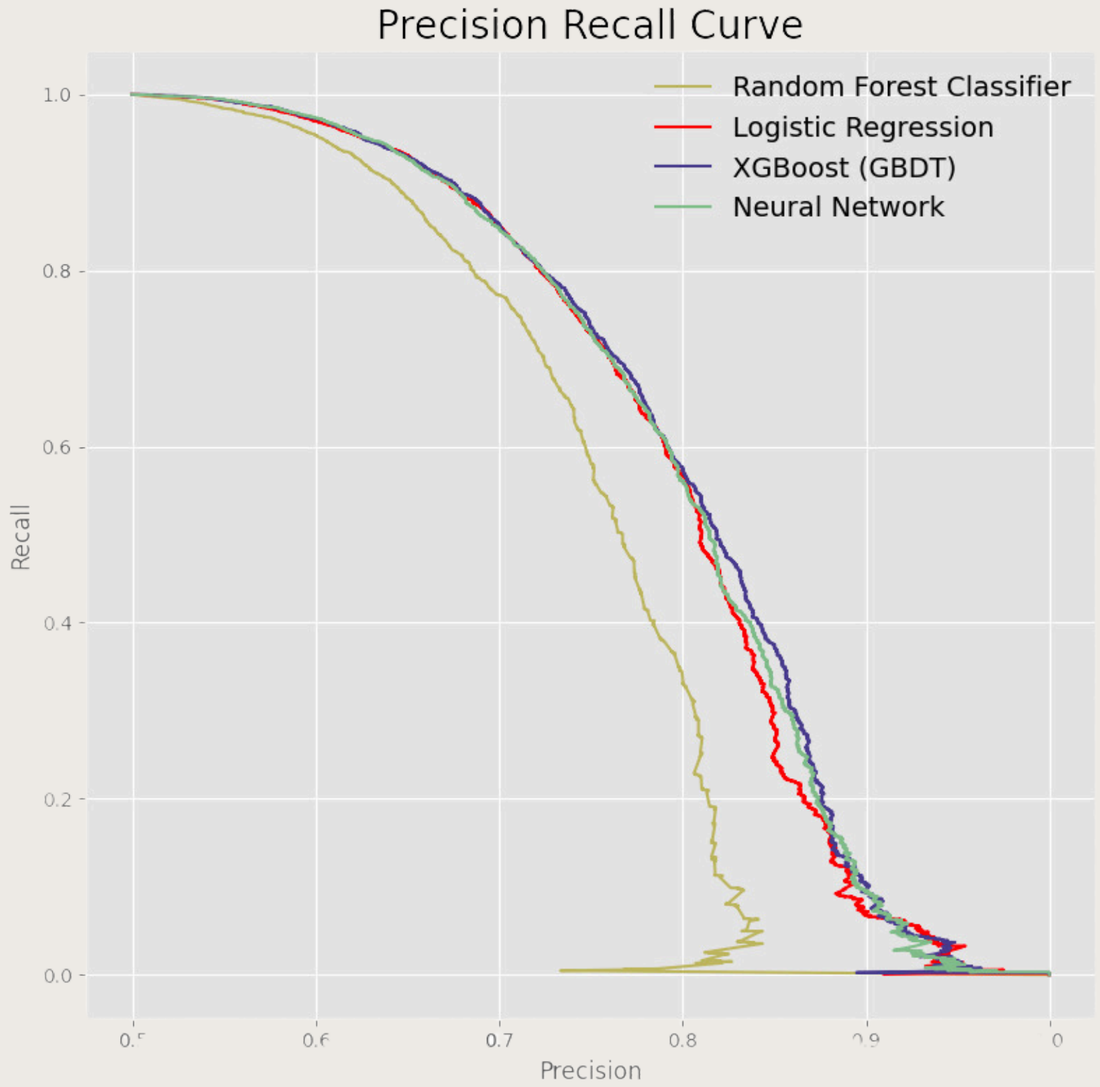
Gradient Boosting Decision Tree has the highest score on three metrics

HUANG LIN CHUN, Wally

MODEL SELECTION

PRECISION-RECALL CURVE

- Precision-recall curve evaluates models' trade-off between precision and recall
- The closer to the point (1, 1), the better the performance
- According to the plot, GBDT also has the best performance

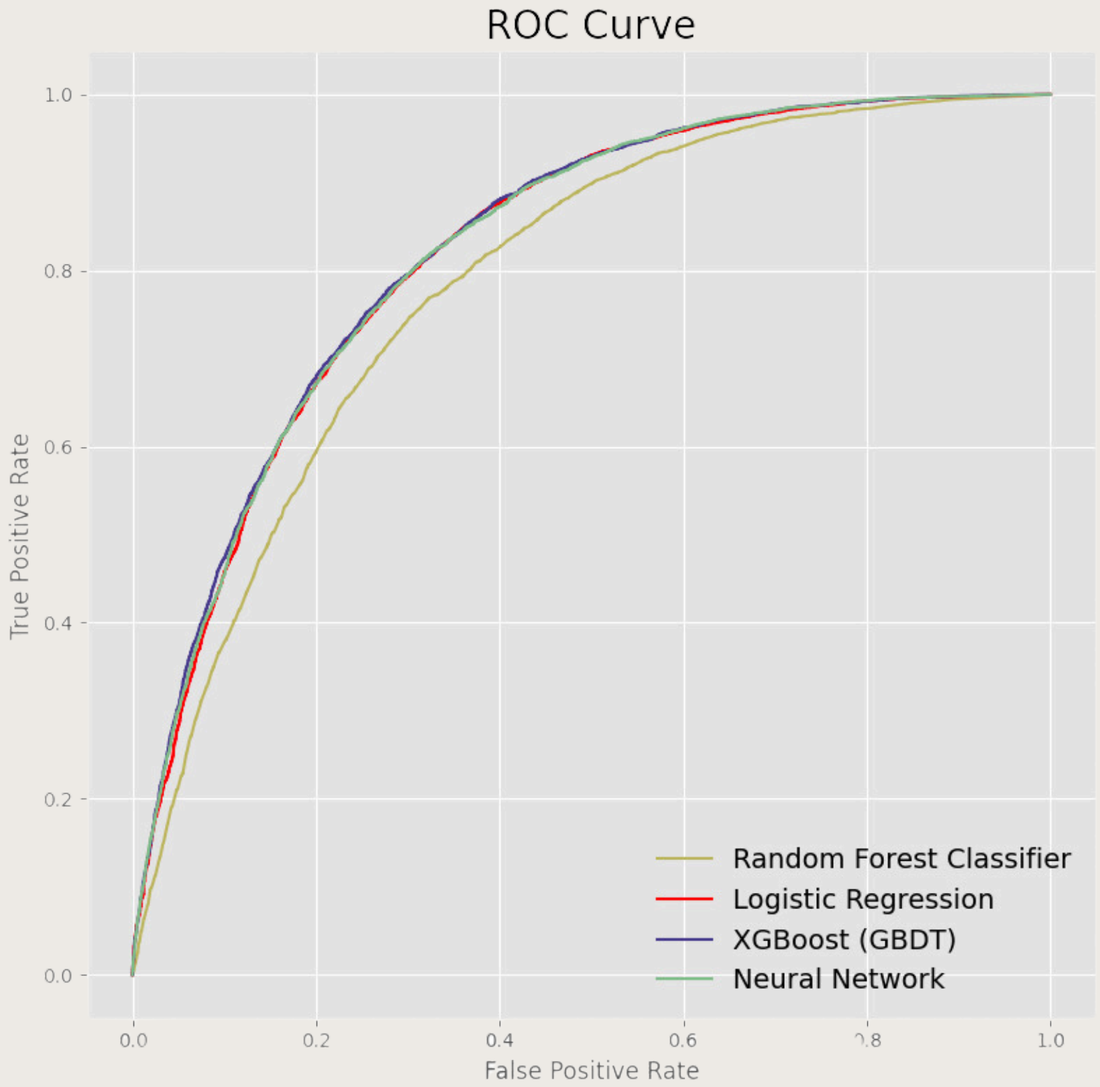


HUANG LIN CHUN, Wally

MODEL SELECTION

ROC CURVE

- ROC curve evaluates models' trade-off between true positive rate (recall) and false positive rate
- The closer to the point (0, 1), the better the performance (the higher the **roc_auc**)
- According to the plot, GBDT also has the best performance



MODEL EVALUATION

HYPER PARAMETERS TUNING

- We exploited **Grid Search** to find the best combination of hyper parameters
- We chose **n_estimators** and **learning_rate** as the objective hyper parameters

```python
clf = GradientBoostingClassifier()
grid_params = {'n_estimators': [10, 50, 100, 250, 500],
               'learning_rate': [0.0001, 0.001, 0.01, 0.1]}
grid_search = GridSearchCV(estimator = clf, param_grid = grid_params, scoring = 'recall', cv = 5).fit
(X_train, y_train)
```

Python

```python
scores_params_dict = grid_search.cv_results_['params']
for i, score in enumerate(grid_search.cv_results_['mean_test_score']):
    scores_params_dict[i]['recall'] = score
scores_params_dict = pd.DataFrame(scores_params_dict).sort_values('recall', ascending = False)
scores_params_dict.iloc[0,:]
```

Python

HUANG LIN CHUN, Wally

MODEL EVALUATION

HYPER PARAMETERS TUNING

- The best combination of the hyper parameters is:
  - learning_rate = 0.1
  - n_estimators = 500
- The average cross-validated recall is 0.791686

```
learning_rate       0.100000
n_estimators      500.000000
recall              0.791686
Name: 19, dtype: float64
```

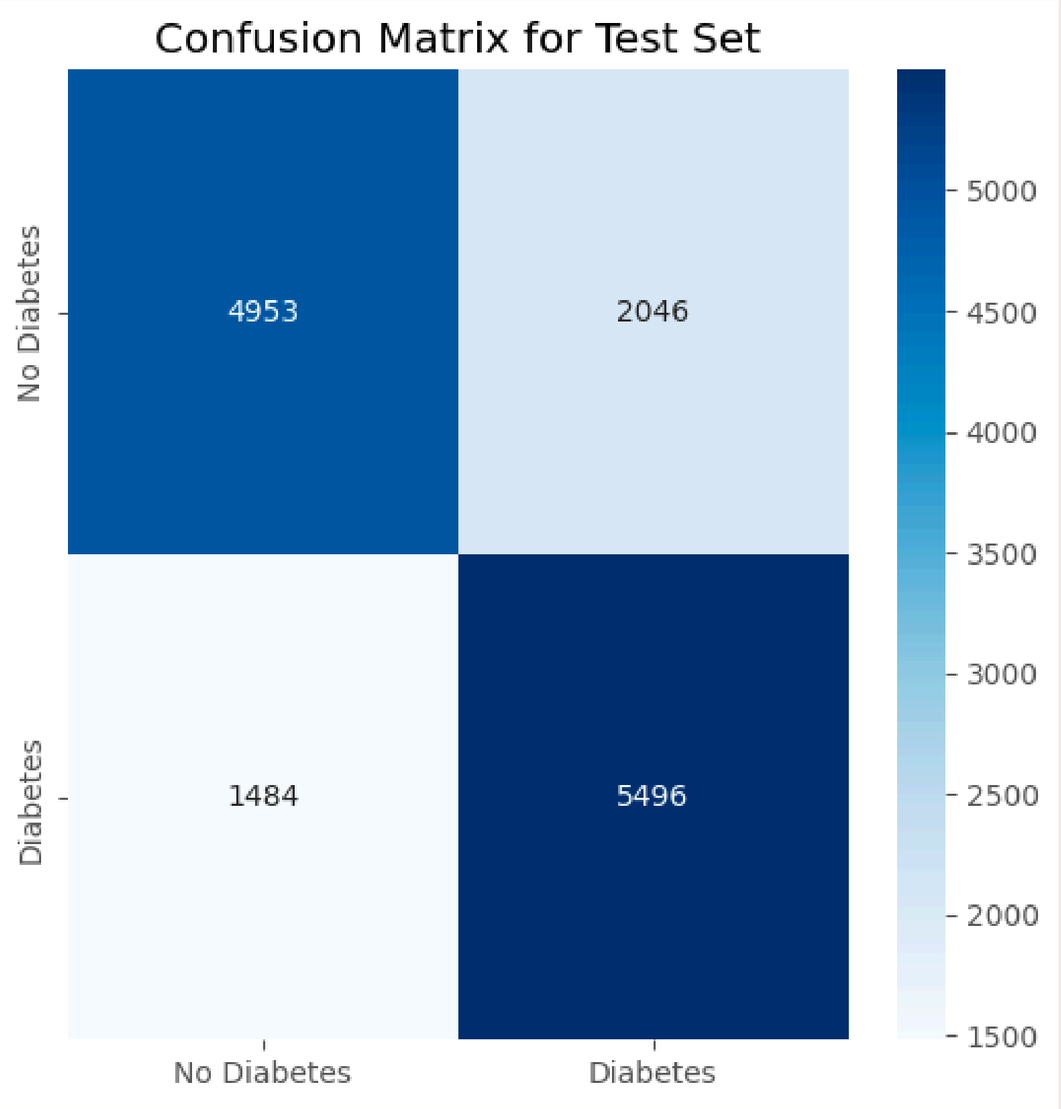
→ We use this combination of hyper parameters as the final model

HUANG LIN CHUN, Wally

MODEL EVALUATION

FINAL MODEL EVALUATION

- We set:
  - learning_rate = 0.1
  - n_estimators = 500
- And trained the **GBDT** model, and then used the predicted values to establish the **confusion matrix** and **classification report**



MODEL EVALUATION

FINAL MODEL EVALUATION

- As shown in the two graphs on the RHS, the final model predicted diabetes with the macro-averaged recall = 0.75
  - **Recall for diabetes = 0.79**
  - Recall for no-diabetes = 0.71
- **However, it is still hard to say that the result is precise and well-performing**

```
-------------- Classification Report ---------------
              precision    recall  f1-score   support

         0.0       0.77      0.71      0.74      6999
         1.0       0.73      0.79      0.76      6980

    accuracy                           0.75     13979
   macro avg       0.75      0.75      0.75     13979
weighted avg       0.75      0.75      0.75     13979
```

MODEL EVALUATION

FINAL MODEL EVALUATION

- **Possible reasons of the low performance:**
  a. features are represented as "level" → less precise (eg: age level instead of real age)
  b. there are many skewed features → might exist selection bias
  c. many features are based on subjective evaluation rather than objective clinical evidence (eg: General Health is based on subjective opinions of health condition)

```
-------------- Classification Report ---------------
              precision    recall  f1-score   support

         0.0       0.77      0.71      0.74      6999
         1.0       0.73      0.79      0.76      6980

    accuracy                           0.75     13979
   macro avg       0.75      0.75      0.75     13979
weighted avg       0.75      0.75      0.75     13979
```

FINAL MODEL OUTPUT

PICKLE

- We trained the model again based on all available data

- ⚠ Because ensemble model from scikit-learn seems not compatible with any web-deployment system, I had no choice but to choose **Logistic Regression** as the final output model

- We saved the model by **pickle** and **gzip** modules

```
X = pd.read_csv("./cleaned_features.csv")
y = pd.read_csv("./cleaned_target.csv")

X.drop(['Unnamed: 0', 'index'], axis = 1, inplace = True)
y.drop(['Unnamed: 0', 'index'], axis = 1, inplace = True)
y = y.values.ravel()
```

```
model = LogisticRegression(max_iter=2000).fit(X, y);
```

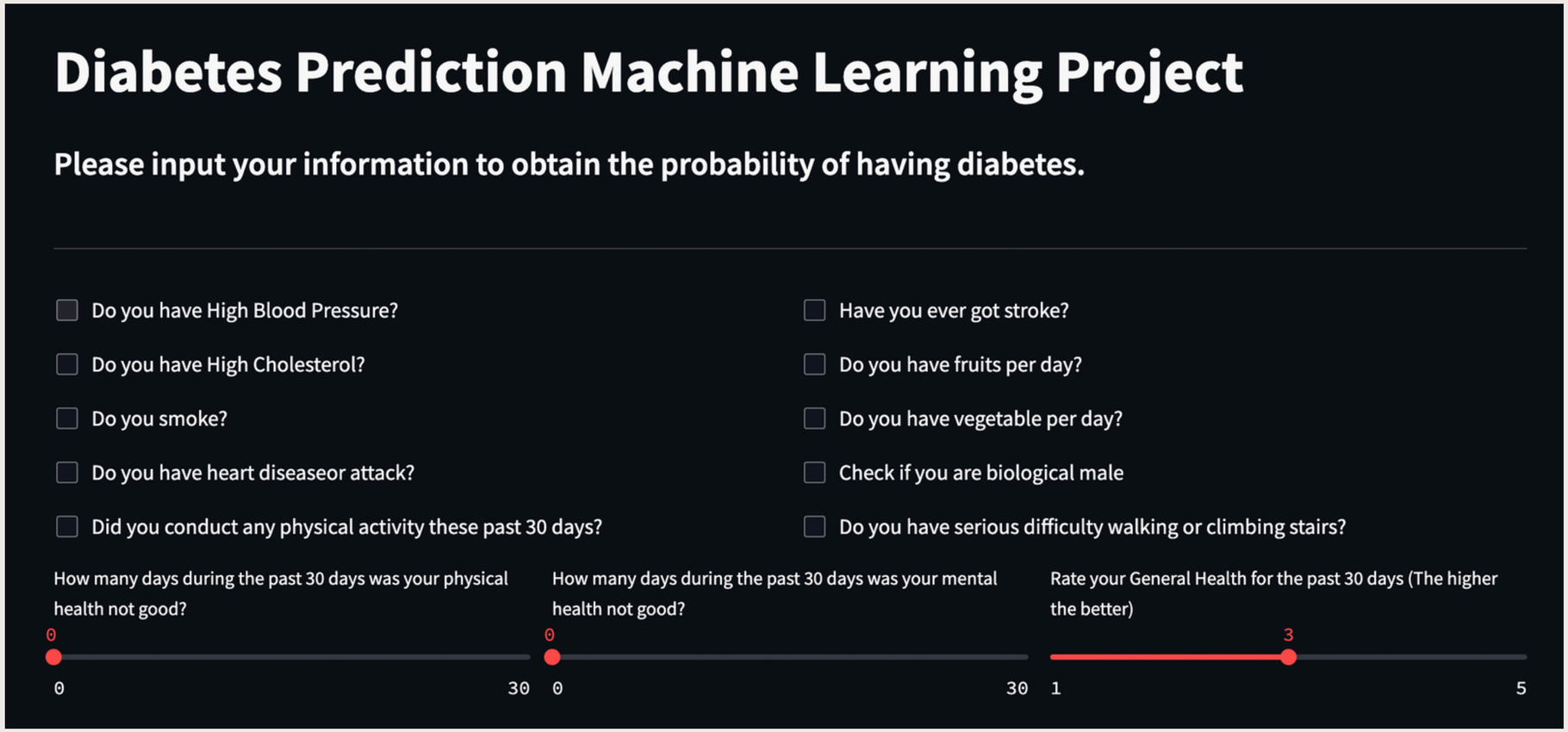
```
with gzip.GzipFile('model.pgz', 'w') as f:
    pickle.dump(model, f)
```

HUANG LIN CHUN, Wally

VISUALIZATION & UI

STREAMLIT

🔗 link of the visualized interface (Streamlit):
https://diabetesdetection-2e2nvdfhxvzucsbd3kvrfw.streamlit.app/



VISUALIZATION & UI

STREAMLIT

🔗 link of the visualized interface (Streamlit):
https://diabetesdetection-2e2nvdfhxvzucsbd3kvrfw.streamlit.app/