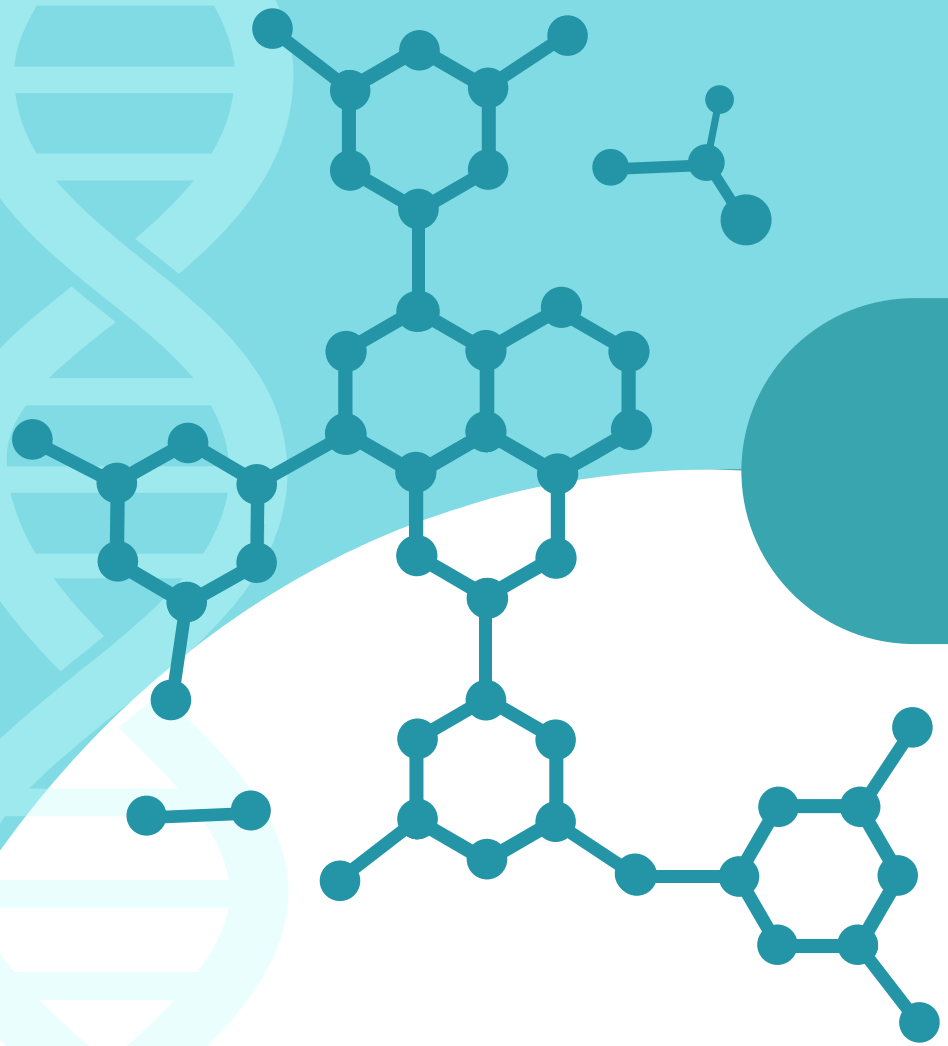
Diabetes Prediction and Feature Importance Analysis

By Nurul Khasanah - DS 24

Outline

Project Background and Objective Goals



What is Diabetes?

Diabetes is a chronic disease that increases risk for stroke, kidney failure, renal complications, peripheral vascular disease, heart disease, and death (Xie, 2019)

The International Diabetes Federation(IDF) estimates that at the current growth, **693 million** people will have diabetes worldwide by 2045.

The Centers for Disease Control and Prevention(CDC) recorded that in 2012, **29.1 million** people in the United State were diagnosed diabetes. This condition put high financial burden for government because of medical cost and decreased productivity.

According to IDF data, in 2021, Indonesia is in the fifth position with **19.5 million** diabetes cases and estimated will increase to **28.6 million** in 2045.

In 1984 CDC initiated the Behavioral Risk Factor Surveillance System (BRFSS), an ongoing, state-based, random-digit–dialed telephone survey of noninstitutionalized US adults aged 18 years or older to identify the risk factors for a variety of human diseases.

What’s the purpose?

1. To determine the features that affect diabetes, the government can provide appropriate preventive measures
2. To develop predictive modelling which could help facilitate early diagnosis and intervention


Data Understanding and
Data Cleaning

Dataset Source :

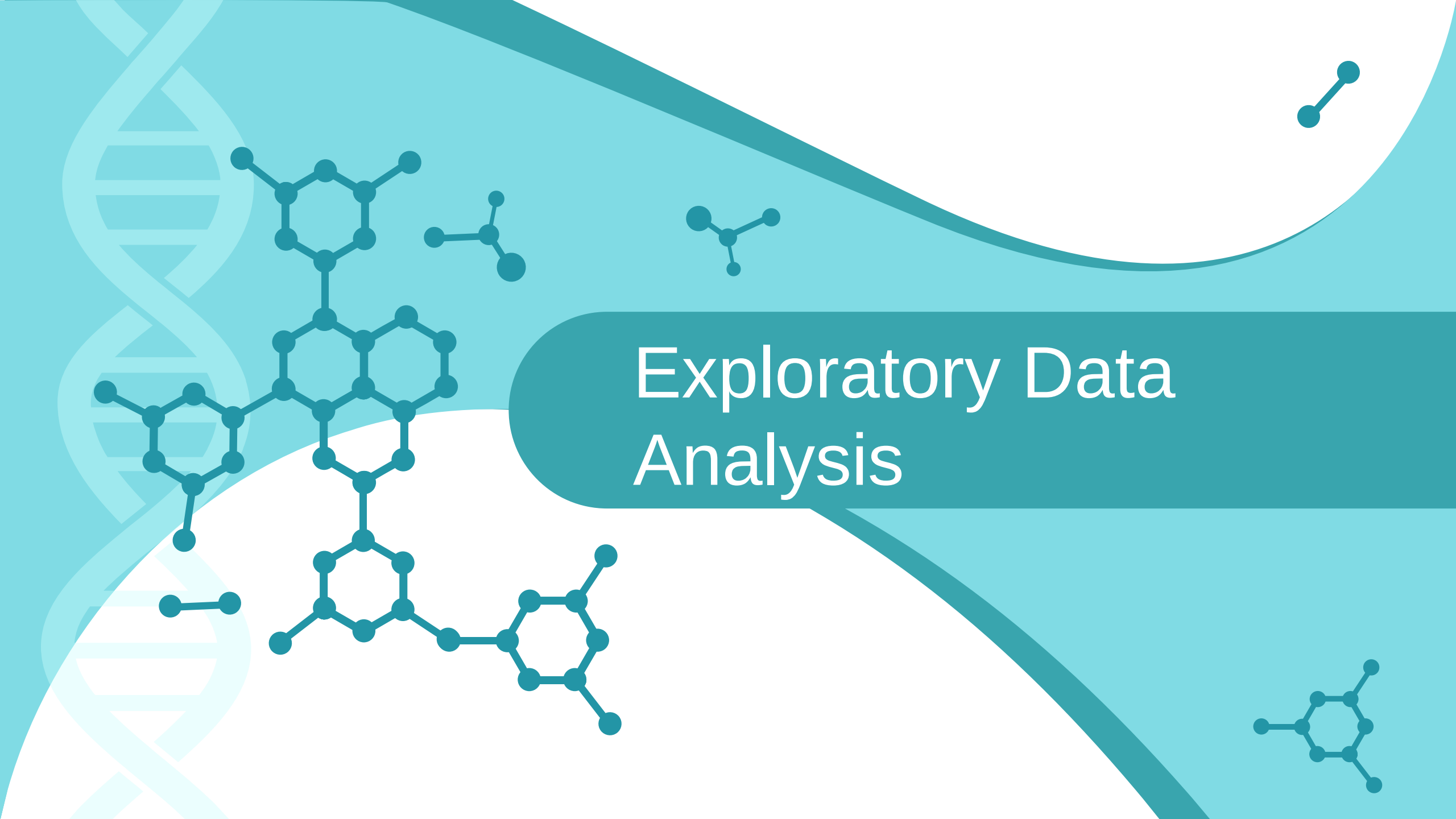
https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset/data

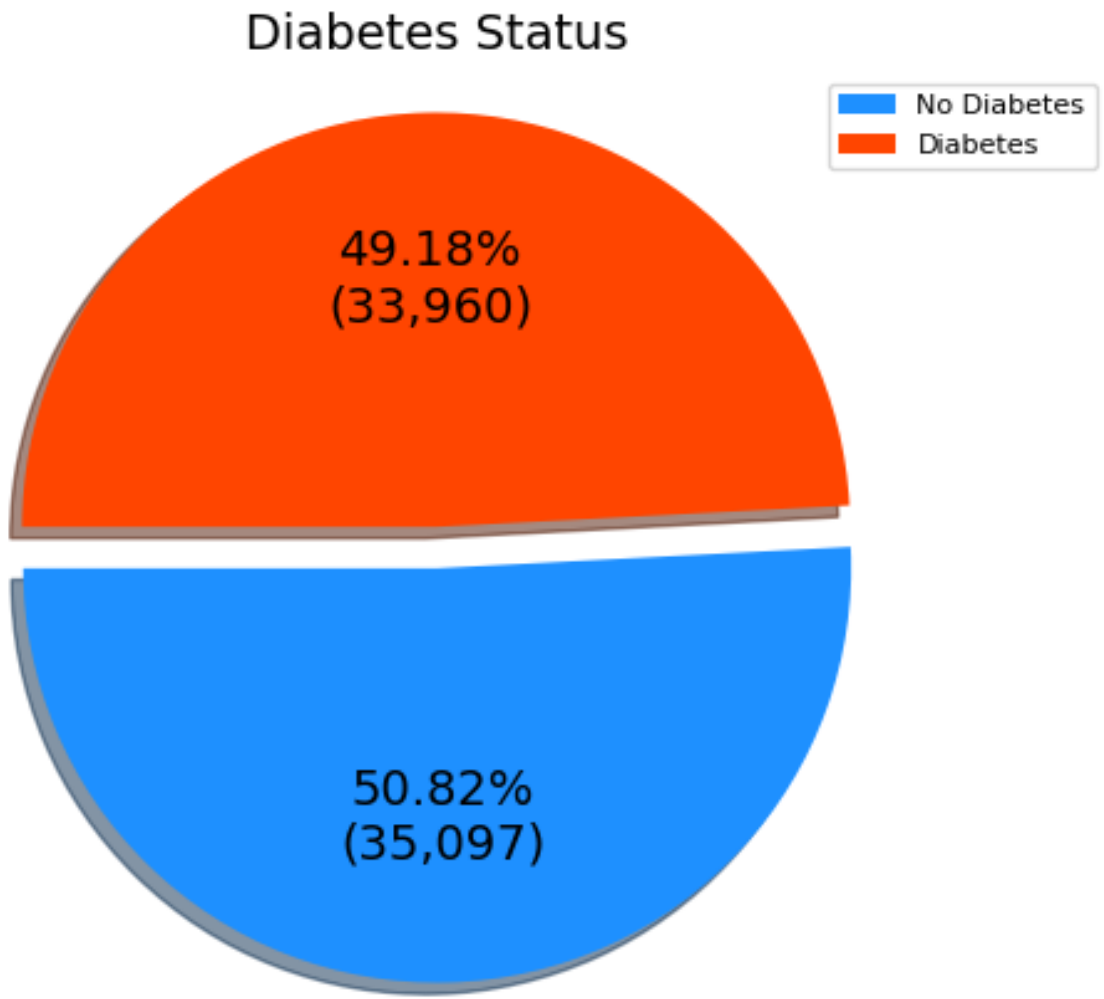
- This dataset comes from survey Behavioral Risk Factor Surveillance System (BRFSS) in the US.
- It contains 22 columns and 70,692 rows
  15 columns with binary data (0,1)
  1 columns is continuous
  4 columns with ordinal
  2 columns with discrete

Description for each feature
https://www.cdc.gov/pcd/issues/2019/19_0109.htm

Data Cleaning

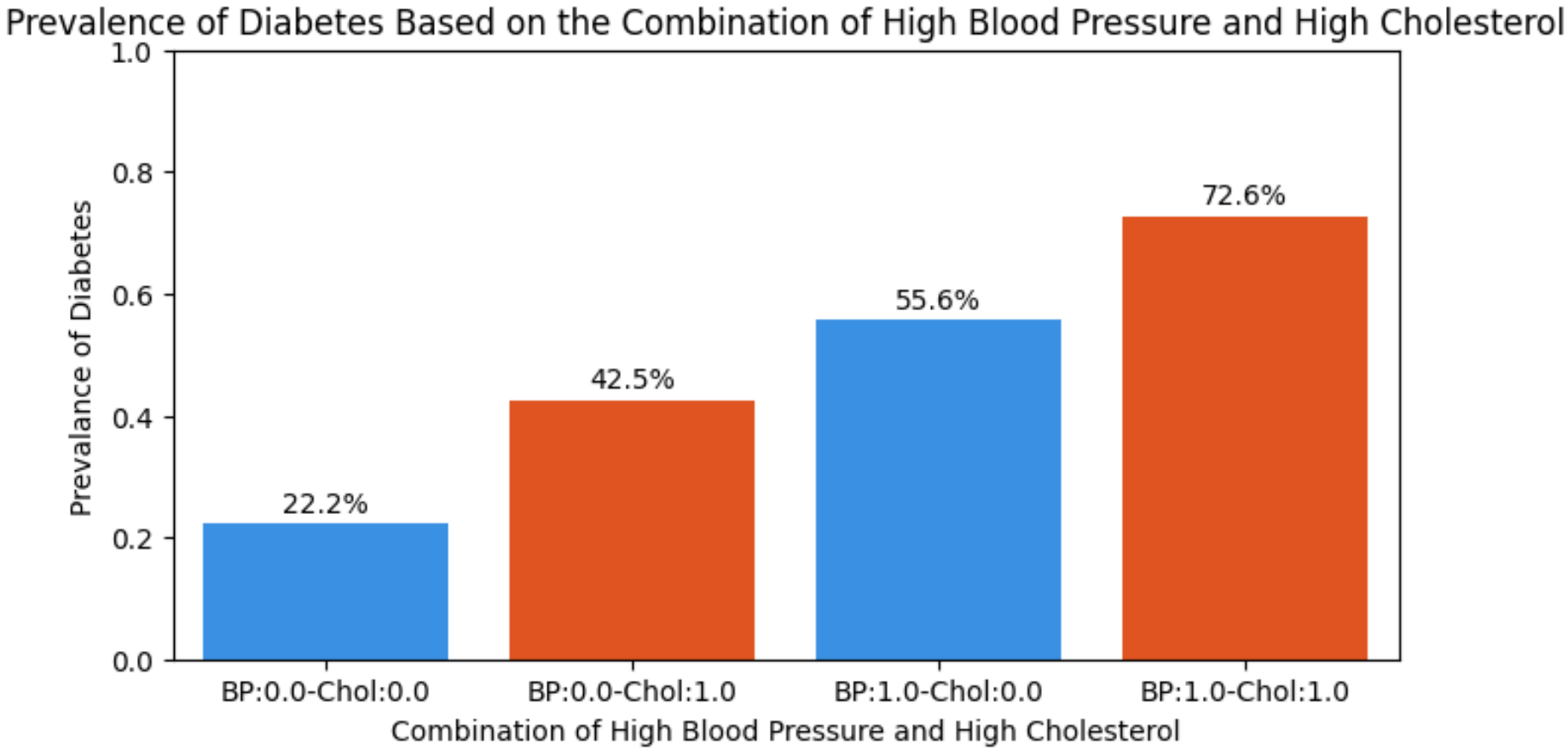
- **No Missing Value**
- **There are 1635 duplicate rows**


Exploratory Data
Analysis



Overall, there are 49.18% respondents who have diabetes, and 50.82% don't have diabetes. It means that the dataset is quite balance

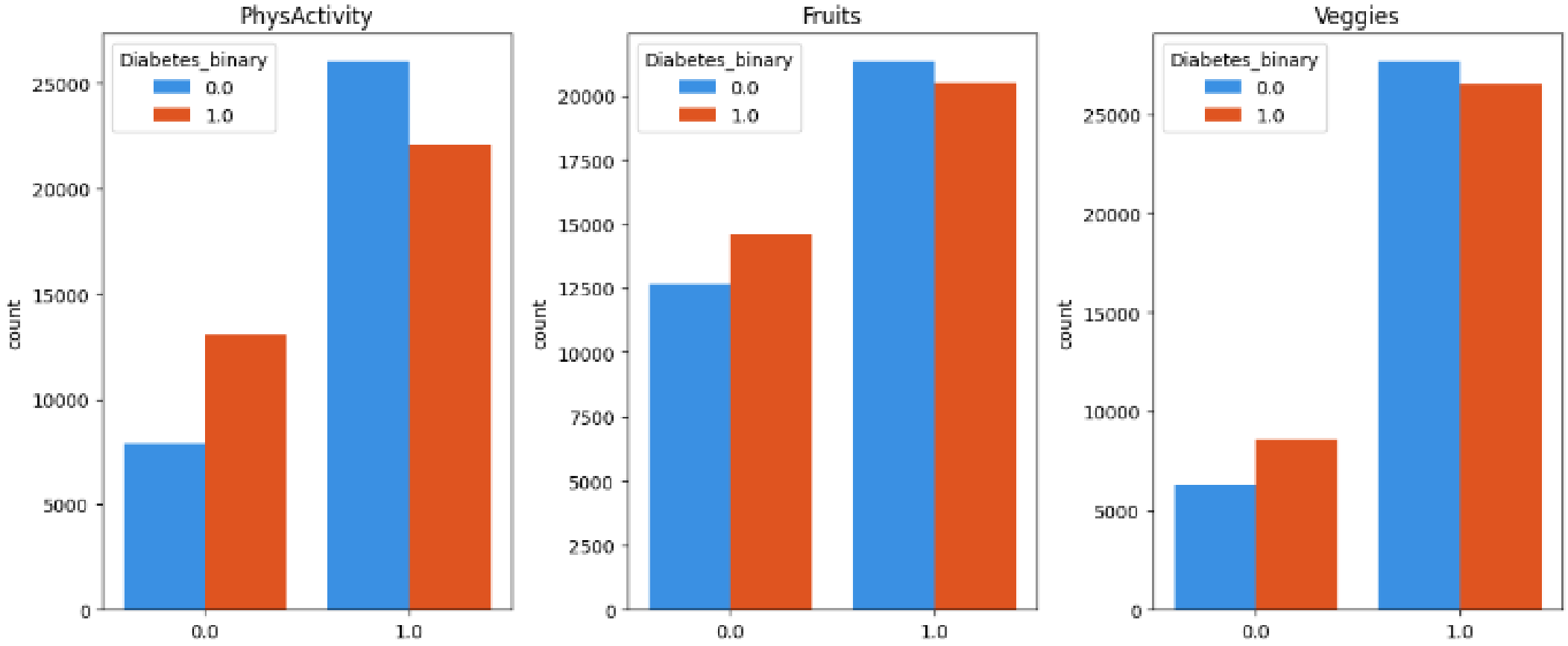
Are people who have high cholesterol and high blood pressure susceptible to diabetes?



Prevalence of diabetes for people who have high blood pressure and high cholesterol is 72.6%

Do healthy people's habits affect diabetes?

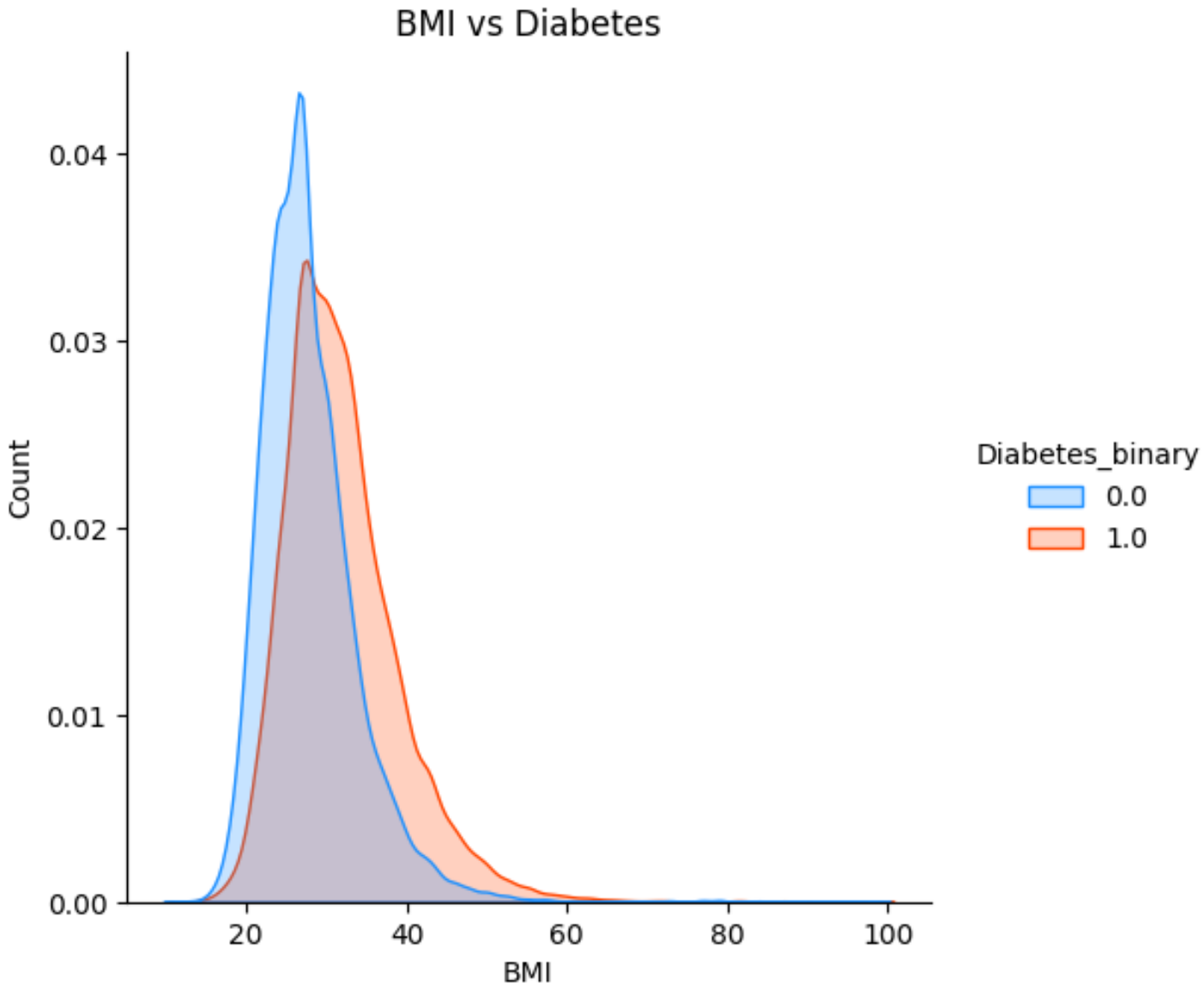
People who do physical activity outside of work tend not to be at risk of diabetes. However, there is no significant difference between people who consume fruits and vegetables



Are obese people at greater risk of diabetes?

Obese = BMI > 30

The plot shows that although there is some overlap between the two curves, the red curve is more spread out to the right suggests a tendency for individuals with diabetes to have a higher BMI than those without diabetes.




Data Preprocessing

Data Pre-processing

1. Drop rows if Cholcheck =0
   0 = no cholesterol check in 5 years, so the data in HighChol is not relevant

1. Drop Cholcheck because the data has been represent in column HighChol

1. Check correlation between features using VIF score
   The VIF score is about 1, it means that no high correlation between independent features.

Feature	VIF Score
HighBP	1.32661
HighChol	1.161452
BMI	1.170568
Smoker	1.07622
Stroke	1.091595
HeartDiseaseorAttack	1.190057
PhysActivity	1.159091
Fruits	1.09534
Veggies	1.098271
HvyAlcoholConsump	1.023087
AnyHealthcare	1.078701

Feature	VIF Score
NoDocbcCost	1.130973
GenHlth	1.850239
MentHlth	1.262823
PhysHlth	1.690836
DiffWalk	1.579181
Sex	1.090481
Age	1.324524
Education	1.316245
Income	1.521131

Data Pre-processing

4. Feature selection based on correlation each feature to target
   The feature with correlation less than 0.05 will be dropped : Fruits, Sex, NoDocbcCost, AnyHealthcare

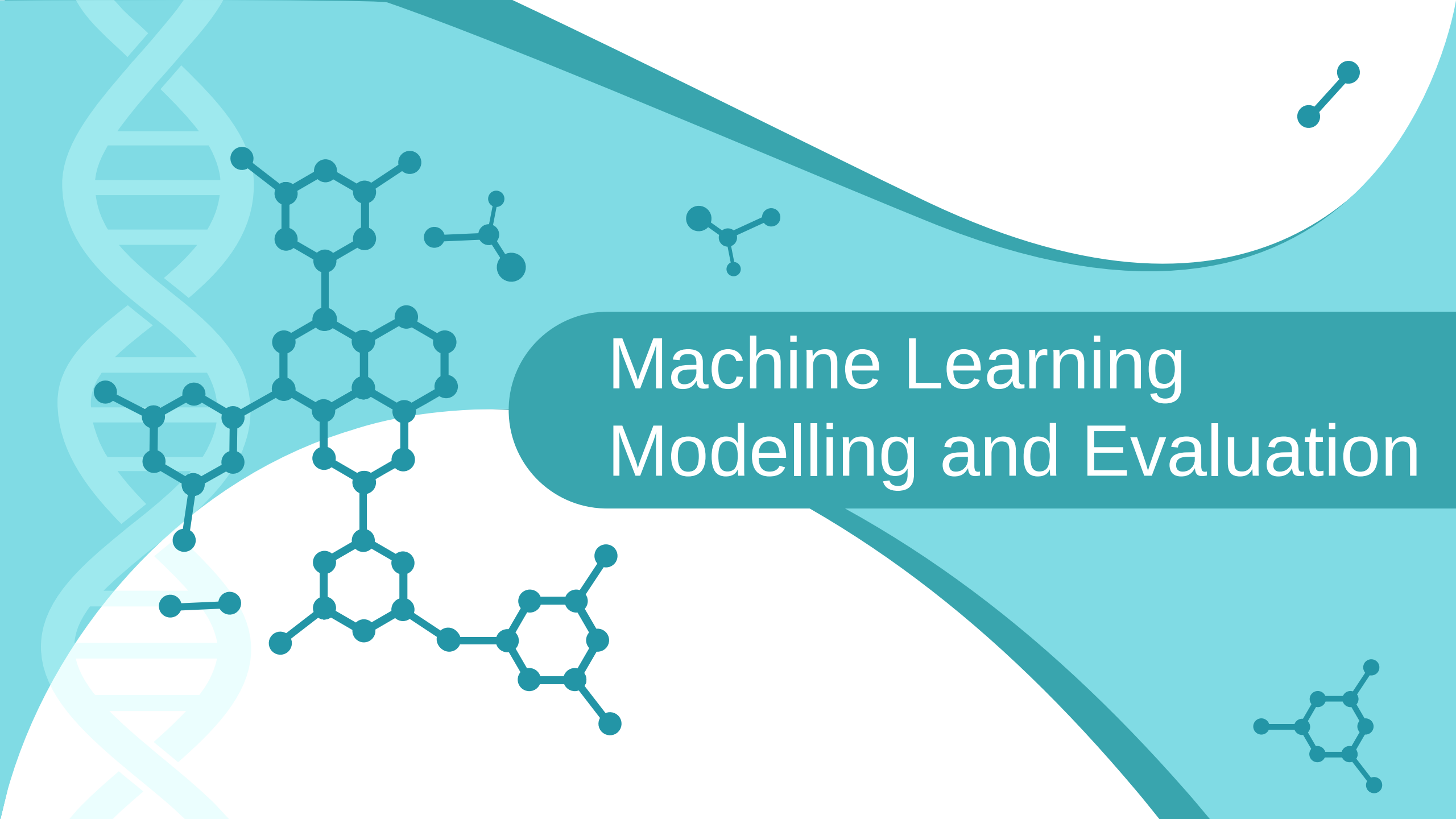
4. Feature Scalling using MinMax Scaler because the data has various range data

4. Split data into data train data test
   Data train = 80%
   Xtrain (53853, 16)
   ytrain (53853,)

   Data test = 20%
   Xtest (13464, 16)
   ytest (13464,)

Feature	Correlation
Diabetes_binary	1
GenHlth	0.393242
HighBP	0.36473
BMI	0.283732
HighChol	0.273975
Age	0.266386
DiffWalk	0.263705
Income	0.214717
PhysHlth	0.203523
HeartDiseaseorAttack	0.202517
Education	0.158577

Feature	Correlation
PhysActivity	0.150078
Stroke	0.120452
HvyAlcoholConsump	0.097422
MentHlth	0.082333
Smoker	0.076613
Veggies	0.072748
Fruits	0.046335
Sex	0.044002
NoDocbcCost	0.041968
AnyHealthcare	0.01789


Machine Learning
Modelling and Evaluation

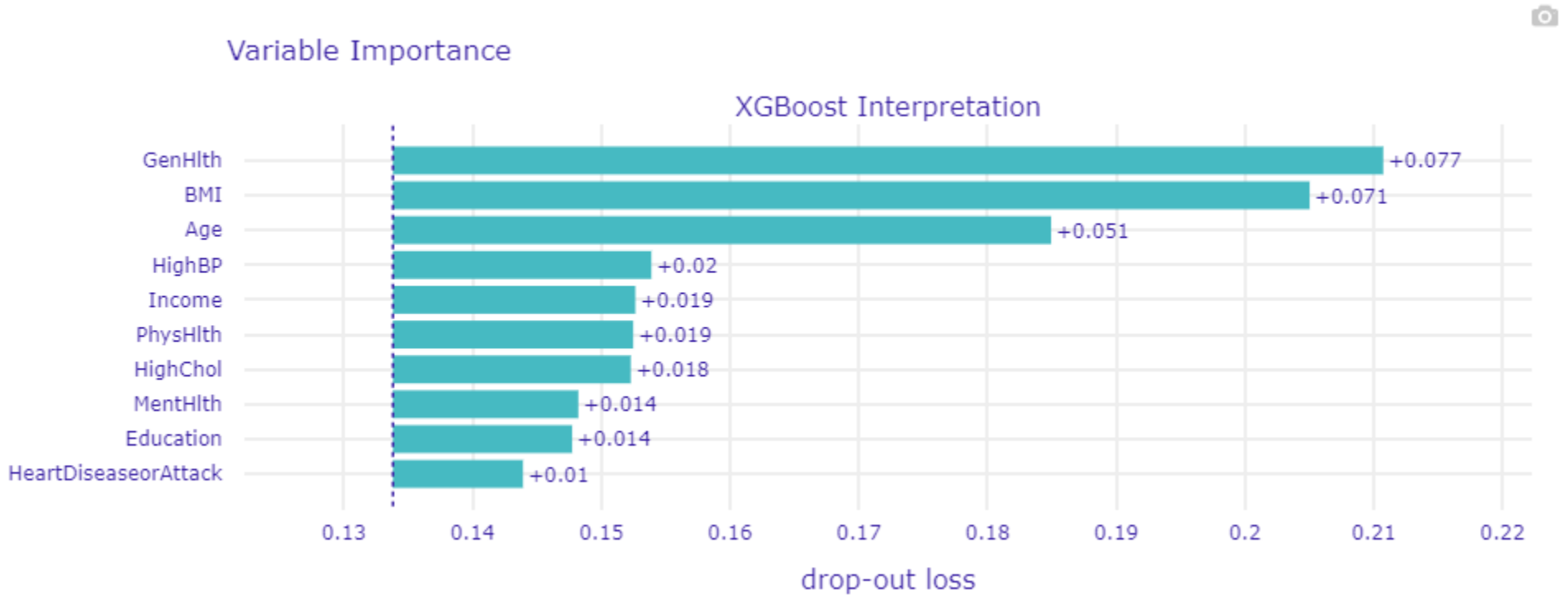
Model Evaluation

Model	recall	precision	f1-score	accuracy
Logistic regression	0.776321	0.741159	0.758333	0.742053
Random Forest Classifier	0.774327	0.716640	0.744368	0.742276
XGBoost Classifier	**0.730844**	**0.800399**	**0.764042**	**0.737718**

- f1-score difference between baseline model and machine learning modelling is not significant. It's only +-1% both before or after hyperparameter tuning
- f1-score of XGBoost Classifier is the highest, 0.764042. **It means that the best model is XGBoost Classifier**

Feature Importance Analysis

Permutation Feature Importance



General health, BMI, Age, High blood pressure, income, physical health, high cholesterol, mental health, education, and heart disease or heart attack have positive affect to diabetes. The higher this feature, the greater risk for someone to have diabetes.

Conclusion

XGBoost classifier is the best predictive model with f1-score 76.98%. There are 5 important features that affect diabetes positively, namely General health, BMI, Age, High blood pressure, difficulty walking, income, heart disease or attack, Mental health, and Heavy alcohol consumptions.

Actionable Recommendation

- Government promotes routine medical check up to diagnose a person's health, if diabetes is indicated then it can be indicated at an early stage
- Provide educational programs about maintaining a healthy lifestyle, including proper diet and regular exercise
- Enhance blood pressure control by recommending dietary modifications, such as reducing sodium intake and increasing potassium-rich foods

Model Deployment

https://predict-diabetes-status-by-nurul.streamlit.app/

Thank You

Link Notebook
https://github.com/nurullkhasanah/predict-diabetes-status/blob/main/diabetes_status_prediction/modelling.ipynb
Link Kaggle :
https://www.kaggle.com/code/nurulk/diabetes-prediction-by-nurul/edit