Problem Definition

Heart disease remains one of the leading causes of death globally, and early identification of individuals at high risk is essential for timely prevention and treatment. In the healthcare industry, the ability to predict health risks using patient data has become increasingly important.

Despite advancements in digital health platforms and the increasing availability of health data, many individuals and even healthcare professionals may lack a clear understanding of how various lifestyle, demographic, and medical factors contribute to heart risk. While some risk factors like age, weight, or chronic conditions are well known, the complex interactions between physical health, mental health, behaviors like smoking and drinking, and historical health events often go unnoticed.

Predicting heart risk is challenging due to several factors:

- Health outcomes are influenced by a wide variety of features, including chronic diseases, test results, medical history, age, BMI, and daily habits, making accurate prediction a non-trivial task.
- There is a growing need for interpretable, data-driven tools that can assist healthcare providers in identifying at-risk individuals and making informed clinical decisions.

This project aims to develop machine learning models that predict whether an individual is at **high risk** of heart-related issues using a rich dataset of health indicators. By identifying the most influential factors affecting heart risk, this report will contribute to building a predictive tool that can assist in early intervention and improve overall public health outcomes.

Exploratory Data Analysis

The dataset used in this project includes various demographic, behavioral, and health-related features collected from individuals across different U.S. states. The primary goal of the analysis is to understand how these features relate to the target variable: **HighRisk**, which indicates whether an individual is at high risk of heart-related health conditions.

Overview of Dataset

- **Number of observations**: *315607*
- **Number of features**: 39
- **Target variable**: HighRisk (binary: Yes/No)

The dataset includes:

- **Demographics**: State, Sex, Age Category, HeightInMeters, WeightInKilograms, BMI

- **Mental & Physical Health**: GeneralHealth, PhysicalHealthDays, MentalHealthDays, LastCheckupTime, PhysicalActivities, ChestScan, HIVTesting, FluVaxLast12, PneumoVaxEver, TetanusLast10Tdap, CovidPos,
- **Chronic Conditions**: HadHeartAttack, HadAngina, HadStroke, HadAsthma, HadSkinCancer, HadCOPD, HadDepressiveDisorder, HadKidneyDisease, HadArthritis, HadDiabetes
- **Lifestyle & Behavior**: SleepHours, SmokerStatus, ECigaretteUsage, AlcoholDrinkers,
- **Disabilities & Difficulties**: RemovedTeeth, DeafOrHardOfHearing, BlindOrVisionDifficulty, DifficultyConcentrating, DifficultyWalking, DifficultyDressingBathing, DifficultyErrands

	State	Sex	GeneralHealth	PhysicalHealthDays	MentalHealthDays	LastCheckupTime
0	Texas	Female	Good	5.0	15.0	Within past 5 years (2 years but less than 5 y...
1	Delaware	Male	Good	NaN	NaN	NaN
2	Florida	Male	Excellent	0.0	0.0	Within past year (anytime less than 12 months ...
3	Maryland	Female	Good	0.0	0.0	Within past year (anytime less than 12 months ...
4	Georgia	Female	Excellent	0.0	0.0	Within past 2 years (1 year but less than 2 ye...
...	...	...	...	...	...	...
315602	Florida	Male	Very good	0.0	3.0	Within past year (anytime less than 12 months ...
315603	Utah	Female	Good	0.0	2.0	Within past year (anytime less than 12 months ...
315604	Texas	Male	Poor	30.0	0.0	Within past 2 years (1 year but less than 2 ye...
315605	Ohio	Male	Very good	0.0	0.0	Within past year (anytime less than 12 months ...
315606	Maryland	Female	Good	0.0	0.0	Within past year (anytime less than 12 months ...

Figure 1: HeartRisk.csv DataFrame

	PhysicalHealthDays	MentalHealthDays	SleepHours	HeightInMeters	WeightInKilograms	BMI
count	308270.000000	309547.000000	312010.000000	308536.000000	298730.000000	294464.000000
mean	4.362653	4.398857	7.021922	1.702695	83.179090	28.566633
std	8.684399	8.375863	1.483322	0.107070	21.459567	6.565876
min	0.000000	0.000000	1.000000	0.910000	22.680000	12.020000
25%	0.000000	0.000000	6.000000	1.630000	68.040000	24.130000
50%	0.000000	0.000000	7.000000	1.700000	80.740000	27.440000
75%	4.000000	5.000000	8.000000	1.780000	95.250000	31.750000
max	30.000000	30.000000	24.000000	2.410000	292.570000	99.640000

Figure 2: Summary of the dataset in the HeartRisk.csv

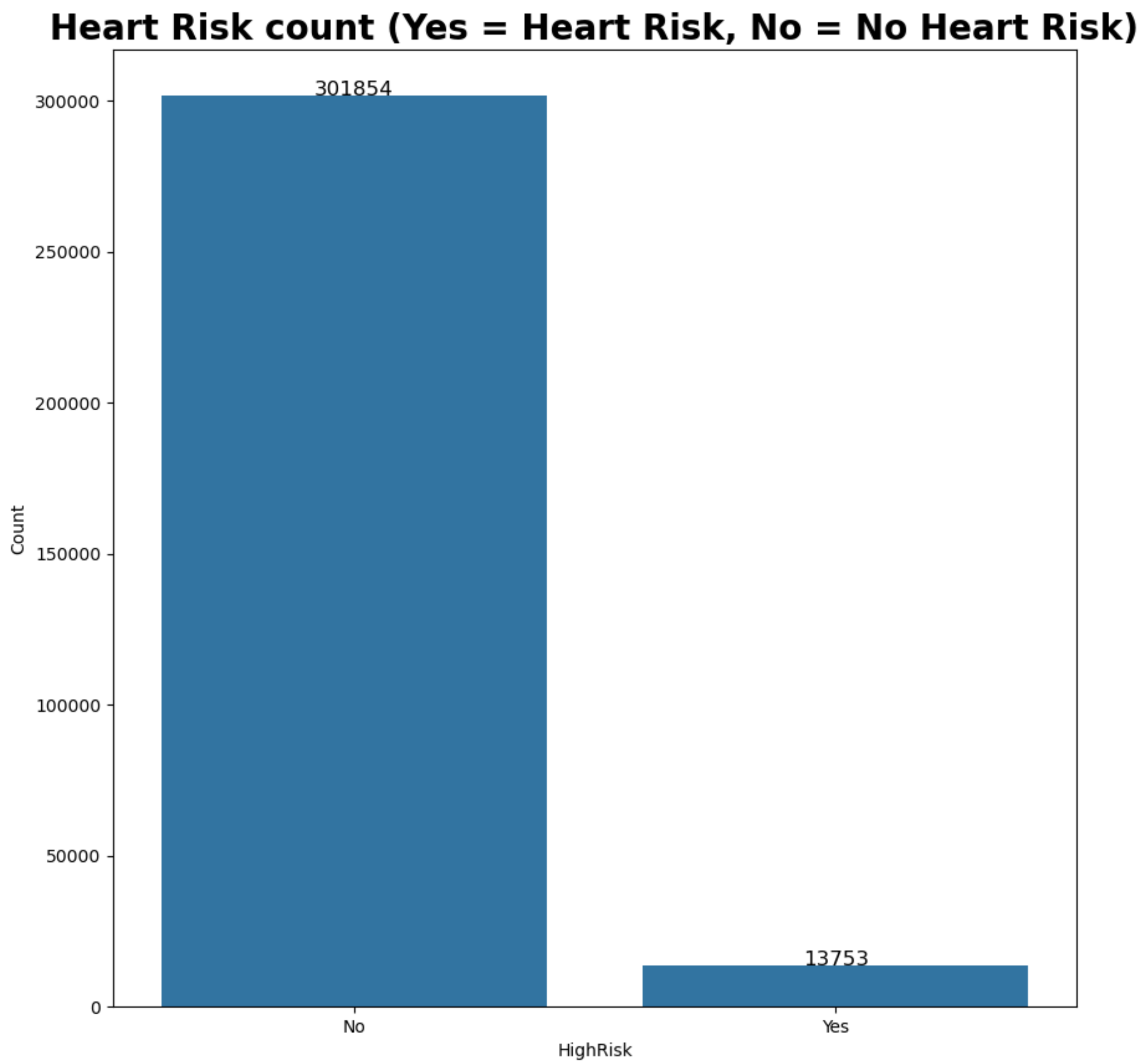

Figure 3: The bar plot showing the imbalance in target column (HighRisk) before SMOTE

Only around 4.36% of data rows are classified as Yes. Therefore, there is an imbalance in this dataset so I will apply SMOTE to handle imbalance.

After handling the imbalance, the HighRisk column look like this

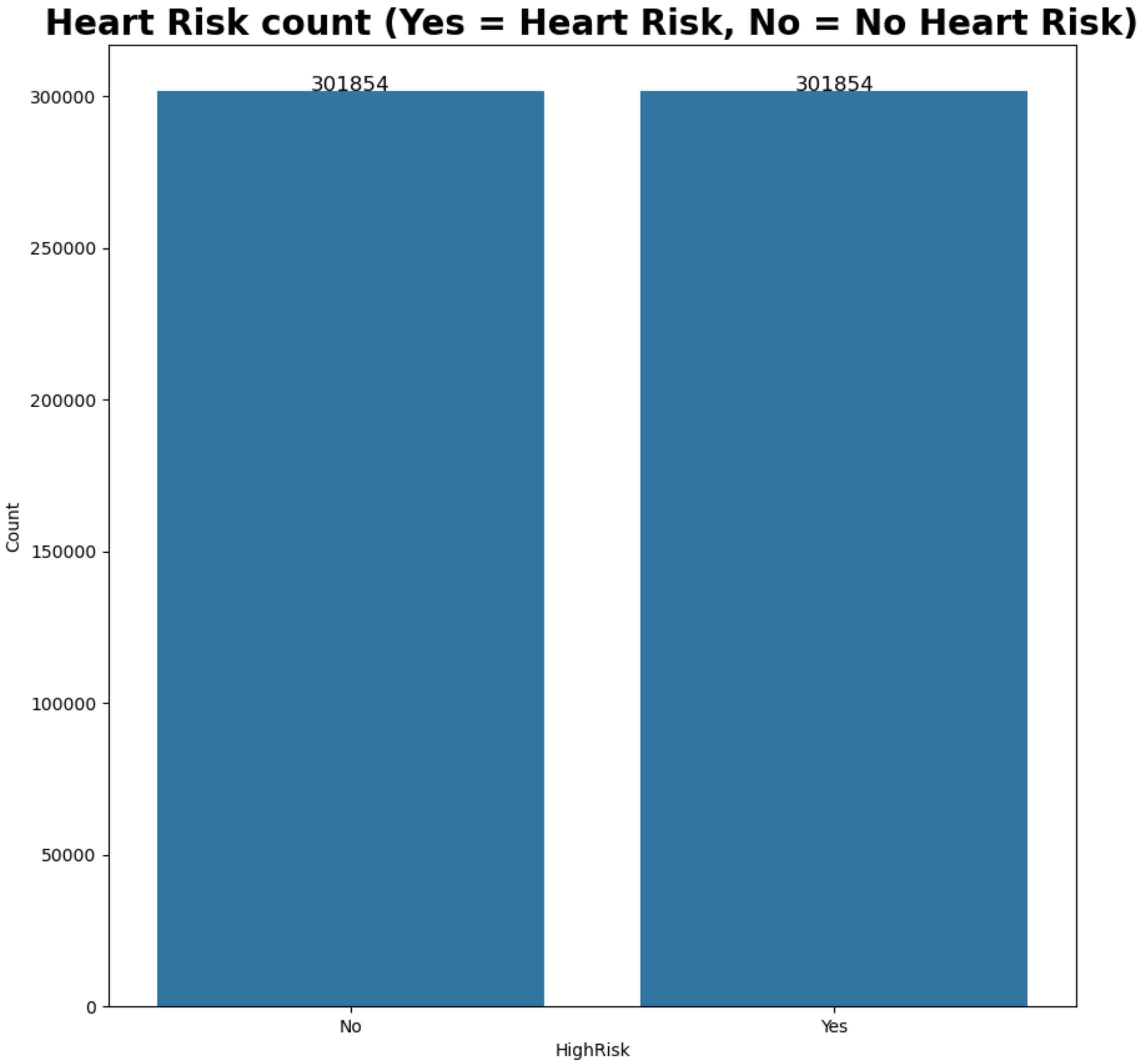

Figure 4: The bar plot showing the imbalance in target column (HighRisk) after SMOTE

Correlation

This section will illustrate the correlation between the target and all predictor variables which are used to predict.

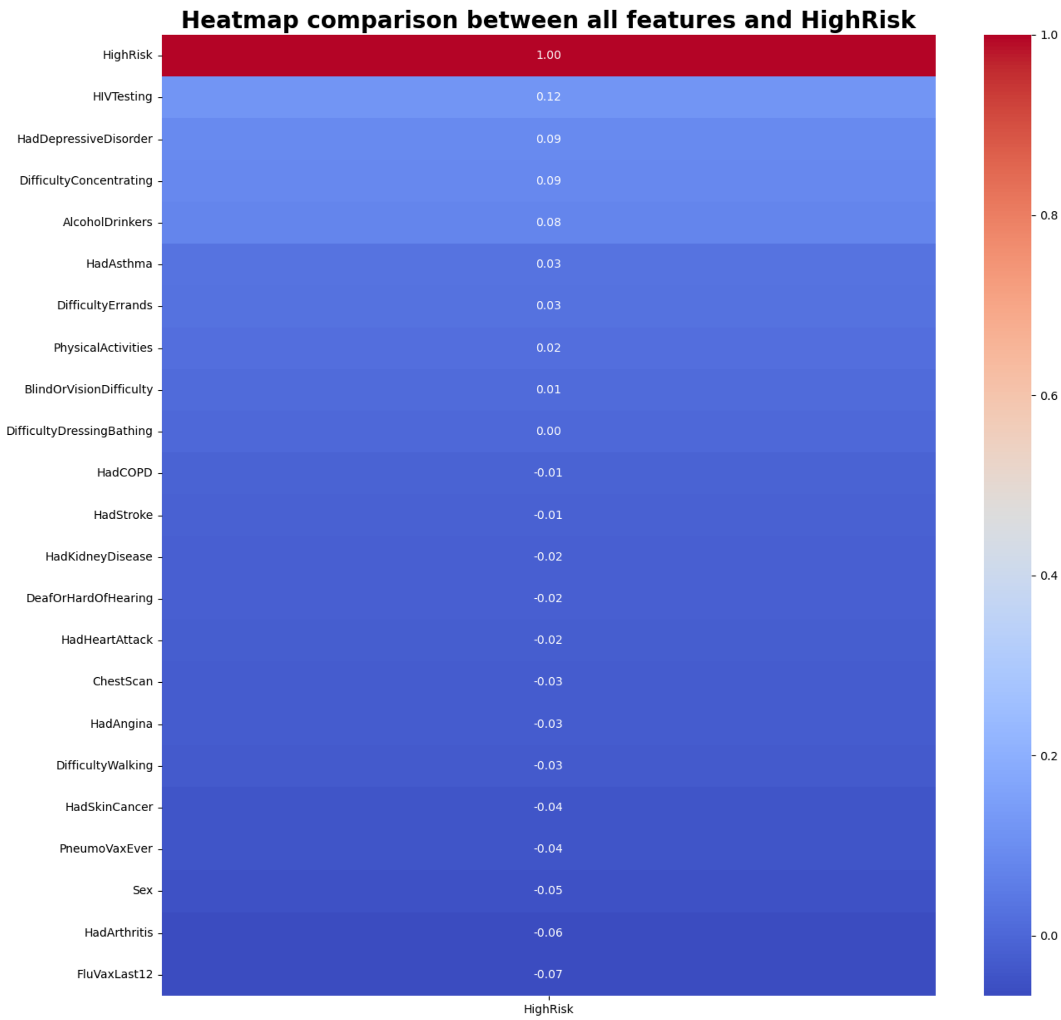

Figure 5: The heatmap showing the correlation between predictor variables and target – HighRisk

Model Development and Tuning

Model Type	Accuracy	Precision	Recall	F1-score	Number of Features	Data Treatments Applied
Logistic Regression	0.9680	0.9724	0.9949	0.9835	70	‘Convert Categorical’, ‘Dummy Variables’, ‘Drop Unused’, ‘Impute Missing’, ‘Scaling’, ‘SMOTE Balacing’, ‘Drop Outliers’
Decision Tree Classifier	0.9630	0.9736	0.9884	0.9809		
Random Forest Classifier	0.9716	0.9737	0.9974	0.9854		
SGD Classifier	0.9650	0.9661	0.9987	0.9821		
Support Vector Classifier	0.9680	0.9696	0.9980	0.9836		
K Neighbors Classifier	0.9675	0.9702	0.9969	0.9833		
Ridge Classifier	0.9621	0.9621	1	0.9807		
Ada Boost Classifier	0.9644	0.9650	0.9992	0.9818		
Gradient Boosting Classifier	0.9645	0.9648	0.9995	0.9819		
XG Boost Classifier	0.9627	0.9628	0.9999	0.9810		
Artificial Neural Network	0.9714	0.9802	0.9903	0.9852		
MLP Classifier	0.9635	0.9808	0.9899	0.9843		
Stacked Model	0.9715	0.9741	0.9969	0.9854		

I choose Artificial Neural Network Model but not MLP Classifier for the following reasons:

- Accuracy is higher (0.9714 > 0.9635)
- Recall is higher (0.9903 > 0.9899)
- F1 score is higher (0.9852 > 0.9843)

In my heartrisk_predictions.py, I use ANN Model to predict the HighRisk in the Test dataset (HeartRisk_Test.csv), then fit the predictions to the Stacked Model to get final result, and the accuracy is 97.19% proving that this ANN and Stacked Model works efficiently.

```
Accuracy of predictions: 0.9719105458923032
```

Figure 6: The accuracy of ANN and Stacked Model when making predictions with Test dataset

Code Appendix

```
Evaluate Stacked Model: Logistic Regression

Confusion Matrix
Predicted    0      1
Actual
0          803   1651
1          194  62145
              precision    recall  f1-score   support

           0       0.81      0.33      0.47      2454
           1       0.97      1.00      0.99     62339

    accuracy                           0.97     64793
   macro avg       0.89      0.66      0.73     64793
weighted avg       0.97      0.97      0.97     64793


Accuracy: 0.9715247017424722
Precision: 0.9741206345225406
Recall: 0.9968879834453552
F1: 0.9853728148412415
```

Figure 7: Summary of Stacked Model

```
Evaluate ANN Model:

Confusion Matrix
Predicted     0      1
Actual
0          1205   1249
1           605  61734
              precision    recall  f1-score   support

           0       0.67      0.49      0.57      2454
           1       0.98      0.99      0.99     62339

    accuracy                           0.97     64793
   macro avg       0.82      0.74      0.78     64793
weighted avg       0.97      0.97      0.97     64793


Accuracy: 0.9713857978485331
Precision: 0.9801692520203864
Recall: 0.9902949999197934
F1: 0.9852061090630536
```

Figure 8: Summary of ANN Model