Heart Disease Analysis with Brane – Group 12

Hsiang-ling Tai
University of Amsterdam
hsiang-ling.tai@student.uva.nl

Yung-sheng Tu
University of Amsterdam
yung-sheng.tu@student.uva.nl

Vishwamitra Mishra
Vrije Universiteit Amsterdam
v.mishra@student.vu.nl

1 Introduction

Heart disease is one of the most common diseases worldwide, affected by multiple factors, including genetics, lifestyle, underlying health conditions, and related diseases. Identifying the key factors can help individuals to mitigate the risk of heart disease by implementing early-stage interventions, such as lifestyle modifications.

This project aims to identifying the key factors of heart disease, utilising a data analysis pipeline within the Brane framework. The pipeline is comprised of five stages. Firstly, preprocess the raw heart disease dataset into a proper format for training machine learning models. Secondly, train three tree-based models with the preprocessed dataset. Thirdly, compute and visualise the ranking of feature importance based on the models. Fourthly, leveraging the feature importance figures, analyse and visualise the significant features. Lastly, collect all figures and generate the well-structured final report. With the final report, we can conclude the key factors of heart disease.

2 Background

2.1 Heart Disease Dataset

The heart disease dataset utilised in this project was sourced from Kaggle: Personal Key Indicator of Heart Disease [1]. The dataset comprises survey data collected from 319,795 adults concerning their health status in 2020. As shown in Figure 1, the dataset has 18 features.

Feature Name	example	Description
HeartDisease	No	Respondents that have ever reported having coronary heart disease (CHD) or myocardial infarction (MI)
Sex	Female	Are you male or female?
AgeCategory	55-59	Fourteen-level age category
Race	White	Imputed race/ethnicity value
Smoking	Yes	Have you smoked at least 100 cigarettes in your entire life? [Note: 5 packs = 100 cigarettes]
AlcoholDrinking	No	Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week
PhysicalActivity	Yes	Adults who reported doing physical activity or exercise during the past 30 days other than their regular job
SleepTime	5	On average, how many hours of sleep do you get in a 24-hour period?
BMI	16.6	Body Mass Index (BMI)
GenHealth	Very good	What would you say that in general your health is? (Poor, Fair, Good, Very good, Excellent)
PhysicalHealth	3	Thinking about your physical health, for how many days during the past 30 days was your physical health not good? (0-30 days)
MentalHealth	30	Thinking about your mental health, for how many days during the past 30 days was your mental health not good? (0-30 days)
DiffWalking	No	Do you have serious difficulty walking or climbing stairs?
Stroke	No	Ever told you had a stroke?
Diabetic	Yes	Ever told you had diabetes?
Asthma	Yes	Ever told you had asthma?
KidneyDisease	No	Not including kidney stones, bladder infection or incontinence, were you ever told you had kidney disease?
SkinCancer	Yes	Ever told you had skin cancer?

Figure 1: Heart disease dataset from Kaggle.

2.2 Brane

According to the user guide of Brane framework [2], 'Brane makes use of containerization to encapsulate functionalities as portable building blocks' and 'end-users with limited or no programming experience are empowered to compose applications by themselves, without having to deal with the underlying technical details.' There are several roles within the Brane framework. Among those roles, *Brane software engineers* and *Brane scientists* are the two relevant roles that would be mentioned in this paper. *Brane software engineers* are responsible for developing Brane packages, and *Brane scientists* define the workflow to solve the task by using the functions defined in Brane packages.

3 Implementation

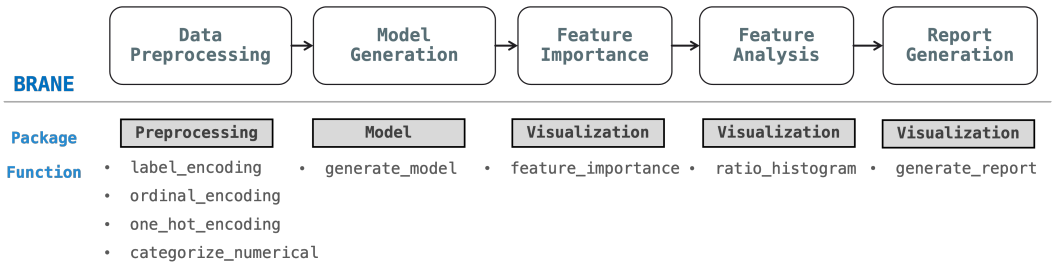


Figure 2: The pipeline of heart disease analysis in the Brane framework.

In Figure 2, we provide an overview of the pipeline as well as the corresponding Brane packages and functions used in each stage. To summarise, the pipeline takes the raw dataset on heart disease from Kaggle as input and produces a HyperText Markup Language (HTML) report containing figures on model performance, feature importance, and feature positive ratio as output. The pipeline consists of five stages:

- *Data Preprocessing*: The purpose of this stage is to generate the dataset for training machine learning models and data analysis from the raw heart disease dataset. The corresponding Brane package in this stage is the *Preprocessing* package. This package contains three encoding functions and one feature engineering function. In the workflow, Brane scientists use encoding functions to transform the categorical feature into numerical ones, as machine learning models can only work with numerical values. These functions take either the raw dataset or the intermediate result of an earlier function call and the target feature names as inputs. The outputs of all these functions are the dataset after being transformed into the *IntermediateResult* type. The feature engineering function, *categorize_numerical*, categorises the numerical feature into given categories. For example, Body Mass Index (BMI) is one of the numerical features in the raw dataset. In the workflow, Brane scientists categorise BMI into four categories: Underweight, Normal, Overweighted, and Obesity. The categorised features are for feature analysis purposes instead of machine learning purposes.

- *Model Generation*: The purpose of this stage is to generate three tree-based classification models based on the dataset generated from *Data Preprocessing* stage. The corresponding Brane package in this stage is the *Model* package. The only function contained in this package is the *generate_model* function. In the workflow, Brane scientists take the dataset after encoding in a previous stage and the label name as inputs. The outputs are three tree-based classification models: Decision Tree, Random Forest, and XGBoost. There are several steps underlying the *generate_model* function. Firstly, it splits the given dataset into training data and testing data. Secondly, it balances the training data by oversampling the minority class through the Random Over-Sampling method. Thirdly, it trains the models based on the training data. Fourthly, it tests the performance of the models using the testing data and keeps

this report in the model. Lastly, it dumps all the models as the output of this function in the *IntermediateResult* type.

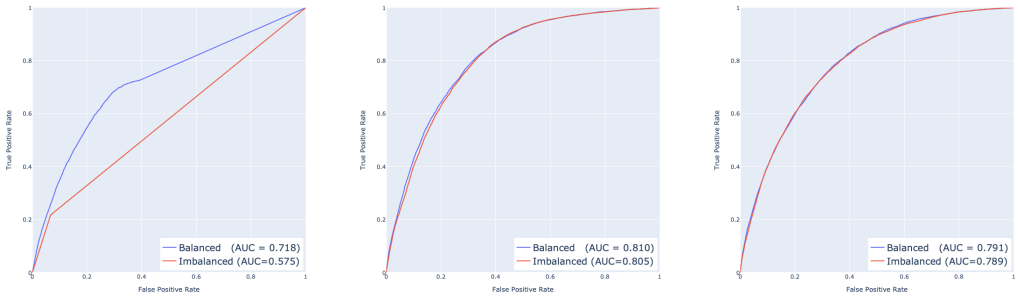
- *Feature Importance*: The purpose of this stage is to visualise the feature importance ranking of the features in the heart disease dataset based on the models generated from the *Tree-based Classification Models Generation* stage. The corresponding Brane package in this stage is *Visualization* package. This package contains three functions, but we would only use the *feature_importance* function in this stage. In the workflow, Brane scientists take the models generated in the *Model* stage as inputs, and the outputs are the feature importance figures of each model. For both Decision Tree and Random Forests models, we use Mean Decrease in Impurity (MDI) method to compute feature importance. For the XGBoost model, we use the Weighted Gini Importance method to compute feature importance.

- *Feature Analysis*: The purpose of this stage is to visualise the detailed statistic analysis of important features. The corresponding Brane package is also the *Visualization* package. In this stage, we use the *ratio_histogram* function. In the workflow, Brane scientists take the names of important features as input, and the outputs are the ratio histograms of the input features. The histogram shows the ratio of positive count to the category count in each category. For example, suppose the input feature is BMI, the histogram will illustrate the positive ratio of the Underweight, Normal, Overweighted, and Obesity classes respectively. With this histogram, scientists are able to identify the high-risk group for heart disease.

- *Report Generation*: The purpose of this stage is to accumulate the figures generated in the previous stages into a single HTML report. The corresponding Brane package in this stage is once again the *Visualization* package. In this stage, we use the *generate_report* function. In the workflow, Brane scientists take the figures as input, and the output is the HTML report that includes all the input figures in an organised manner.

4 Results

We conducted the classification analysis using three different models: Decision Tree, Random Forest, and XGBoost. To evaluate their performance, we examined the models both with and without over-sampling, utilising Receiver Operating Characteristic (ROC) curves and Area Under the Curve (AUC). The corresponding ROC curves are illustrated in Figure 3. It is noteworthy that all the models trained with a balanced dataset demonstrated superior performance, as indicated by higher AUC values. Therefore, during the *Model Generation* stage, we employed the Random Over-Sampling method to balanced the dataset.

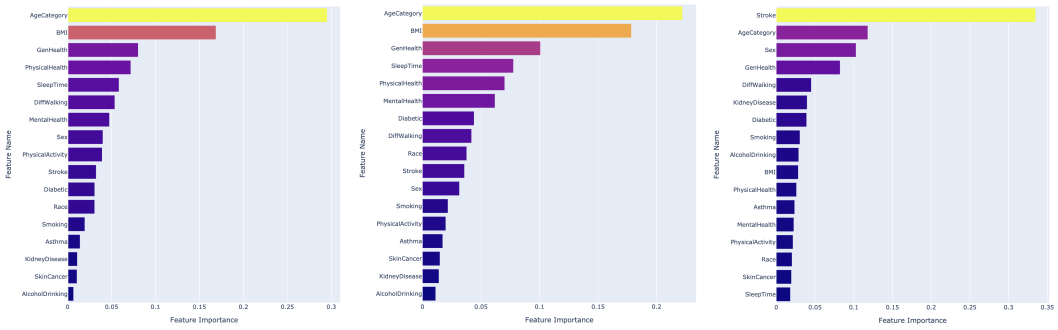


(a) Decision Tree (b) Random Forest (c) XGBoost

Figure 3: ROC curves with balanced and imbalanced datasets in the proposed models

Upon training the aforementioned models, we obtained the feature importances from each of them, as depicted in Figure 4. Notably, *AgeCategory* and *GenHealth* consistently emerged as sig-

nificant features among the top five in all of these models. Additionally, *BMI*, *SleepTime* and *PhysicalHealth* are also prominent features within the top five for Decision Tree and Random Forest. However, *Stroke*, *Sex* and *DiffWalking* exclusively appeared in the top five of the XGBoost model. This information provides valuable insights into the factors that influence the models' predictions and should be further considered.



(a) Decision Tree (b) Random Forest (c) XGBoost

Figure 4: Feature importance of the proposed models

Given that our primary objective is to identify the key factors of heart disease, it is crucial to select the most effective model to accurately classify individuals who are affected by this condition and constitute the minority in the dataset. Considering the AUC values highlighted in Figure 3 and the classification report presented in Figure 5, we deduce that the Random Forest model is the optimal choice for accurately classifying whether an individual is affected by heart disease or not.

	precision	recall	f1-score	support
0	0.95	0.77	0.85	74619
1	0.20	0.61	0.30	7039
accuracy			0.76	81658
macro avg	0.58	0.69	0.58	81658
weighted avg	0.89	0.76	0.81	81658

(a) Decision Tree

	precision	recall	f1-score	support
0	0.95	0.87	0.91	74619
1	0.27	0.49	0.35	7039
accuracy			0.84	81658
macro avg	0.61	0.68	0.63	81658
weighted avg	0.89	0.84	0.86	81658

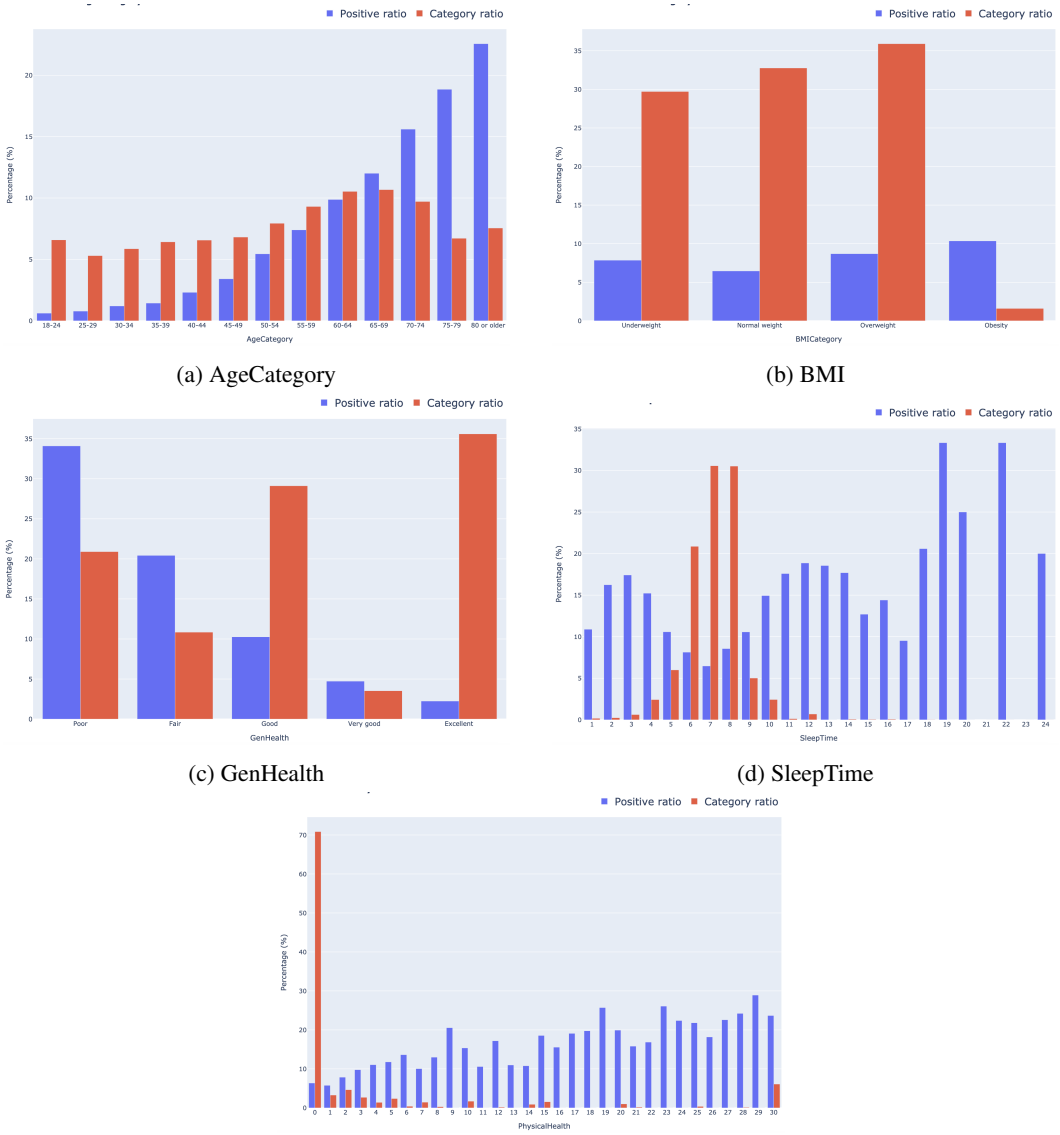
(b) Random Forest

	precision	recall	f1-score	support
0	0.94	0.91	0.93	74619
1	0.29	0.36	0.32	7039
accuracy			0.87	81658
macro avg	0.61	0.64	0.62	81658
weighted avg	0.88	0.87	0.87	81658

(c) XGBoost

Figure 5: Classification report of the proposed models

Moving on to our analysis of the top five features of Random Forest in Figure 6, we observe several trends. In Figure 6a, the distribution of age appears fairly uniform, but the proportion of positive heart disease cases increases with age. Therefore, older individuals have a higher likelihood of having a positive heart disease outcome. Figure 6b shows that individuals who are not in the normal status have a slightly higher likelihood of testing positive for heart disease. Figure 6c clearly indicated that individuals with generally poorer health conditions are more prone to have a positive heart disease outcome. Furthermore, in Figure 6d, the distribution of sleep duration among individuals is concentrated between 4 to 10 hours. Consequently, considering other durations as outliers, we observed an increased proportion of positive heart disease cases as sleep time either increases or decreases. Referring to Figure 6e, which represents the number of days within 30-day timeframe when an individual's physical health is not in good condition, a clear upward trend in the positive ratio of heart disease is evident as the number of such days increases.



(a) AgeCategory

(b) BMI

(c) GenHealth

(d) SleepTime

(e) PhysicalHealth

Figure 6: Positive Proportions and Category Proportions of various features

5 Conclusion

This project aims to discover the key factor of heart disease. Based on the Brane pipeline report and the AUC values, we adopt the feature importance of the Random Forest model. According to the feature importance and our analysis, the key factors of heart disease are: old age, abnormal weight, poor general health, insufficient or excessive sleep time, and worst physical health condition.

References

[1] Kamil Pytlak. *Personal Key Indicator of Heart Disease*. URL: `https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease`.

[2] Onno Valkering et al. *Brane framework*. URL: `https://wiki.enablingpersonalizedinterventions.nl/user-guide/`.