The Application of Machine Learning on Risk Classification of Heart Disease

Xiaoyu Lin Kyoka Ono

Abstract

GitHub Link: https://github.com/kyokaono8811/biostat625finalproject.git

Heart disease is a leading cause of death for adults in the U.S, and early detection of key risk factors is essential for prevention. In this study, we apply multiple machine learning methods, including logistic regression, random forest, GAM, ikNN, XGBoost, and RNN identify the significant risk factors of heart disease. We then evaluate model performance and compare variable importance across methods using evaluation Metrics.

(Some results conclusion)

Contributions

- **Xiaoyu Lin** — "Data description, Data cleaning..."
- **Kyoka Ono** — "..."

Setup

```
#install.packages("reticulate")
library(reticulate)
```

```
#reticulate::py_install("matplotlib")
#reticulate::py_install("pandas")
#reticulate::py_install("seaborn")
```

```
import pandas as pd
import numpy as np
import matplotlib as mpl
import matplotlib.pyplot as plt
import seaborn as sns
pd.set_option('future.no_silent_downcasting', True)
```

Introduction

Heart disease affects millions of individuals in the U.S, making early identification of risk factors a high public-health priority. Traditional statistical approaches have identified several predictors, but machine learning methods can capture more complex, nonlinear relationships.

This project applies several different machine learning algorithms to the `heart_2022_no_nans.csv` dataset from the CDC Behavioral Risk Factor Surveillance System (BRFSS) on Kaggle. Our goal is to determine which model outputs the highest evaluation metric for predicting heart attack, and which variables significantly associated with having had a heart attack?

Methods

Data Source

The dataset contains 40 variables for over 200,000 survey participants (all complete cases).

```
heart_data = pd.read_csv("data/heart_2022_no_nans.csv")

# Number of participants (rows) and variables (columns)
print("\nNumber of participants:", heart_data.shape[0])
#>
#> Number of participants: 246022
print("Number of variables:", heart_data.shape[1])
#> Number of variables: 40
```

Preprocessing

Variable Selection

We choose 10 covariates out of 40 for our models based on literature review and use `HadHeartAttack` as the predicted variable

```
heart_data = heart_data[[
        "HadHeartAttack",
        "Sex",
        "PhysicalActivities",
        "SleepHours",
        "HadStroke",
        "HadDiabetes",
        "SmokerStatus",
        "RaceEthnicityCategory",
        "AgeCategory",
        "BMI",
        "AlcoholDrinkers"]
]
```

Outcome Variable:

- `HadHeartAttack`: Binary indicator (Yes/No) of whether a doctor diagnosed the respondent with a heart attack.

Predictor Variables:

- `Sex`: Biological sex of the participant (Male/Female).
- `PhysicalActivities`: Whether the participant engaged in physical activities in the past month (Yes/No).
- `SleepHours`: Average number of hours of sleep per night (numeric).
- `HadStroke`: Whether had a stroke (Yes/No).
- `HadDiabetes`: Whether had a diabetes (Yes/No/Yes, but only during pregnancy (female)/No, pre-diabetes or borderline diabetes).
- `SmokerStatus`: Smoking status of the participant (Former smoker/Never smoked/Current smoker – now smokes every day/Current smoker – now smokes some days/No).
- `RaceEthnicityCategory`: (White only, Non-Hispanic/Black only, Non-Hispanic/Other race only, Non-Hispanic/Multiracial, Non-Hispanic/Hispanic)
- `AgeCategory`: Age group (18–24, 25–29, . . . , 80+).
- `BMI`: Body mass index (numeric).
- `AlcoholDrinkers`: Whether a participant is a heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per (Yes/No).

Data Cleaning

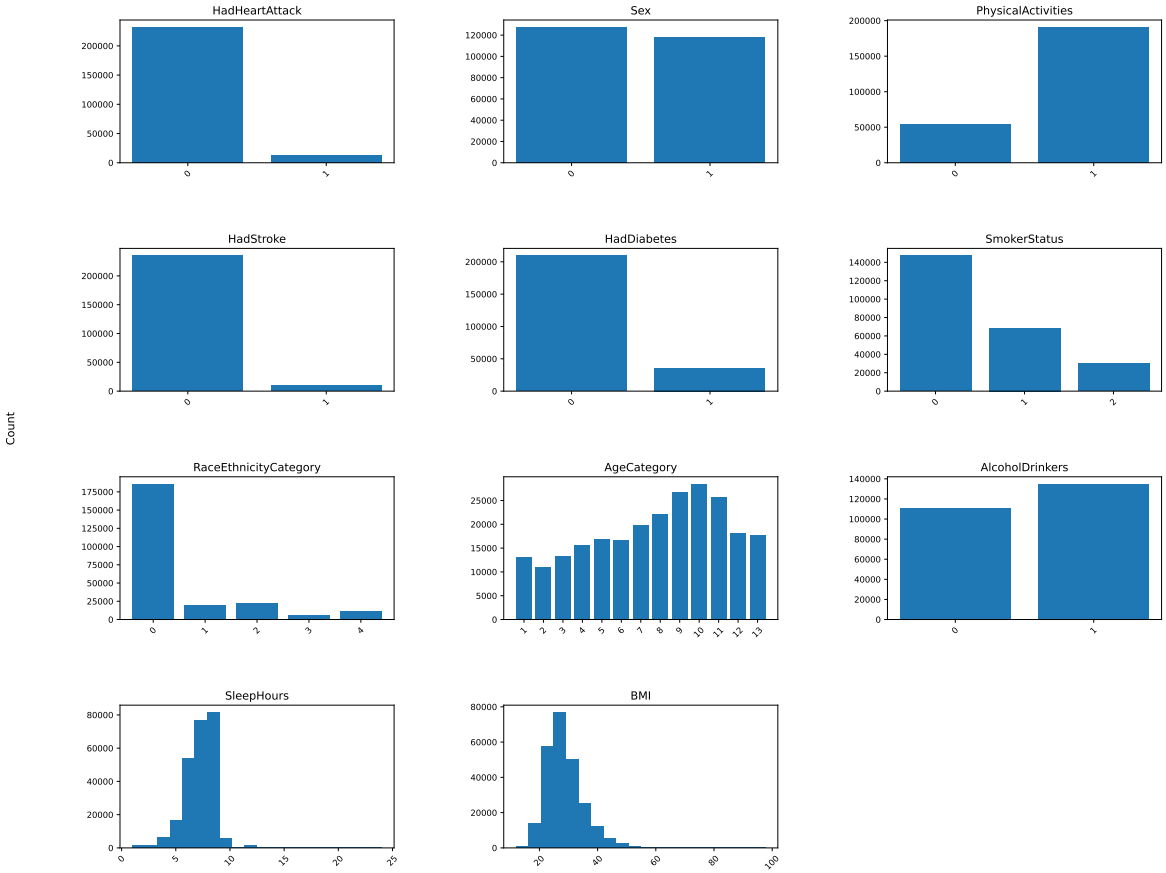
We would like to create dummy variables for categorical variables and merge equivalent levels.

```
from analysis.cleaning import clean_data
heart_clean = clean_data(heart_data)
```

Data Description

Let's look at the distribution of the variables

```
from analysis.description import description
description(heart_clean)
plt.show()
```

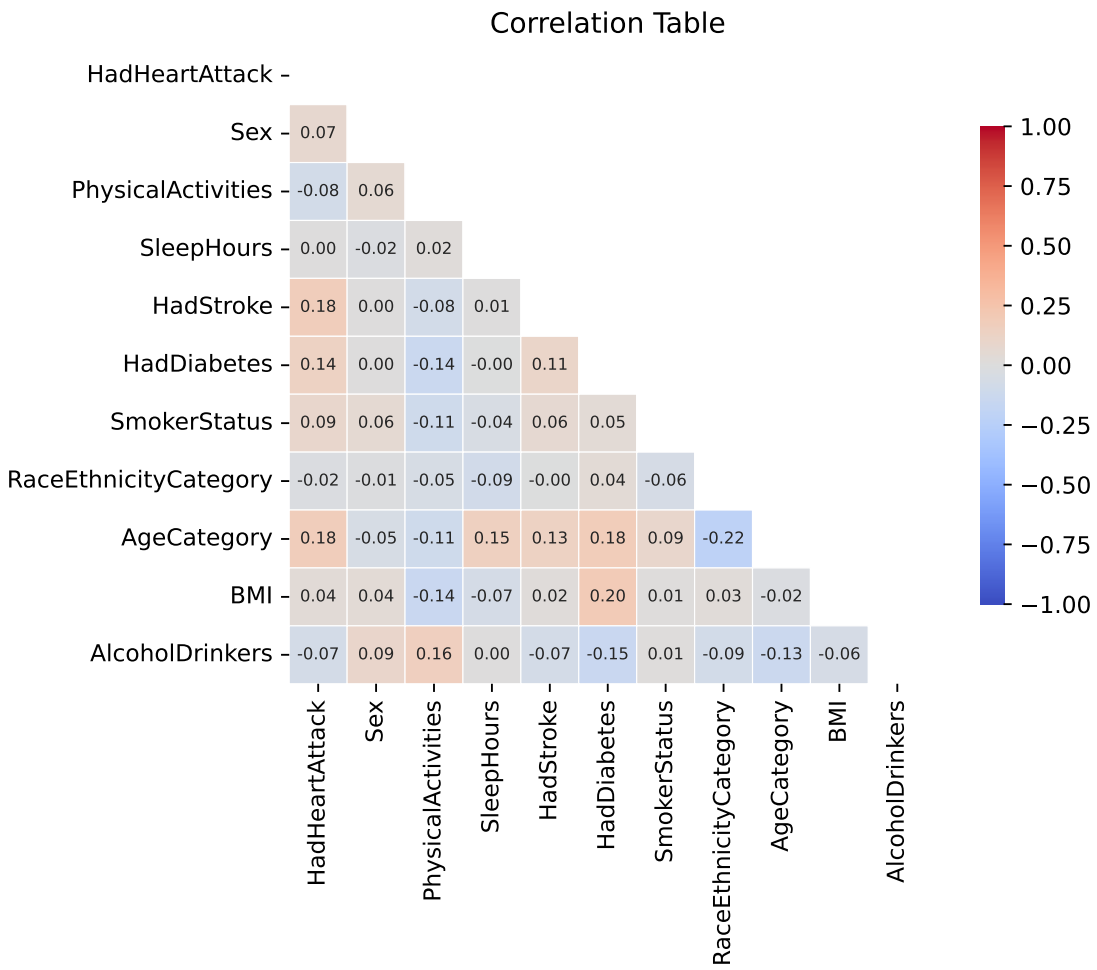


From the graph above, we can see that some categorical variables such as `HadHeartAttack`, `HadStroke`, and `RaceEthnicityCategory` exhibit noticeable class imbalance. Highly imbalanced classes may affect model performance if not addressed, so we will use the method of undersampling the majority class during model training. Other categorical variables are more balanced.

Numerical variables like `AgeCategory`, `SleepHours`, and `BMI` show skewed distributions. To improve model performance, these variables should be standardized or centered, especially for algorithms sensitive to scale (e.g. logistic regression).

Now Let us build a correlation table. We use spearman correlation because the dataset contains a mixture of binary, ordinal, and continuous variables. Spearman does not assume linearity or normality and is appropriate for ranked relationships, making it a better measure of association than Pearson for this dataset.

```
from analysis.correlation import correlation
correlation(heart_clean)
plt.show()
```



All pairwise correlations were relatively weak, indicating that no individual predictor shows a strong linear association with heart attack or with each other. This is expected in multi-factor health datasets, where the outcome is influenced by many small effects rather than a single dominant variable.

Low pairwise correlations do not imply weak predictive power for nonlinear effects, interactions, and combined contributions can still provide meaningful classification performance in multivariate models.

Models Applied

1. **Logistic Regression**
2.

Evaluation Metrics

Results

Conclusion

References

- Pytlak, K. (n.d.). Personal key indicators of heart disease [Data set]. Kaggle. https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease/data
- Centers for Disease Control and Prevention. (2024, December 2). Heart disease risk factors. https://www.cdc.gov/heart-disease/risk-factors/?CDC_AAref_Val=https://www.cdc.gov/heartdisease/risk_factors.htm