STA4100 Final Project: Heart Attack Prediction in Indonesia

Zhecheng Ren 任喆程 121090464

April 29, 2025

Table of Contents

1. Dataset Description

This dataset provides a detailed health profile of individuals in Indonesia, focusing on heart attack prediction.

Indonesia has seen a rising trend in cardiovascular diseases, making early prediction and prevention crucial. This dataset includes key demographic, clinical, lifestyle, and environmental factors associated with cardiovascular risks. It reflects real-world health trends in Indonesia, considering factors such as hypertension, diabetes, obesity, smoking, and pollution exposure.

The main task of this final project is to predict heart attack risks. In this way does this project support the public health research and epidemiological studies in Indonesia.

1.1 Dataset Loading & Variable Description

In this section, we load the dataset and describe the basic information of the varibles in this dataset.

The dataset contains 28 variables and $158,355$ records. The variables can be divided into 6 categories:

1. Demographics
2. Clinical Risk Factors
3. Lifestyle & Behavioral
4. Environmental & Social
5. Medical Screening
6. Target Variable

The following table summarizes the variable name, data type, description, Chinese name, values/units of each variable:

Variable Name	Type	Description	Chinese Name	Values/Units
Demographics				
age	int	Age of the individual	年龄	25-90 years
gender	str	Gender	性别	Male, Female
region	str	Living area	居住区域	Urban, Rural
income_level	str	Socioeconomic status	收入水平	Low, Middle, High
Clinical Risk Factors				
hypertension	int	Presence of high blood pressure	高血压	1=Yes, 0=No
diabetes	int	Diagnosed diabetes	糖尿病	1=Yes, 0=No
cholesterol_level	int	Total cholesterol level	总胆固醇水平	mg/dL
obesity	int	BMI >30	肥胖	1=Yes, 0=No
waist_circumference	int	Waist circumference measurement	腰围	cm
family_history	int	Family history of heart disease	心脏病家族史	1=Yes, 0=No
Lifestyle & Behavioral				
smoking_status	str	Smoking habit	吸烟状态	Never, Past, Current
alcohol_consumption	str	Alcohol intake level	饮酒情况	None, Moderate, High
physical_activity	str	Physical activity level	身体活动水平	Low, Moderate, High
dietary_habits	str	Diet quality assessment	饮食习惯	Healthy, Unhealthy

Environmental & Social				
air_pollution_exposure	str	Exposure to air pollution	空气污染暴露程度	Low, Moderate, High
stress_level	str	Perceived stress level	压力水平	Low, Moderate, High
sleep_hours	float	Average nightly sleep duration	睡眠时长	3-9 hours
Medical Screening				
blood_pressure_systolic	int	Systolic blood pressure measurement	收缩压	mmHg
blood_pressure_diastolic	int	Diastolic blood pressure measurement	舒张压	mmHg
fasting_blood_sugar	int	Fasting blood glucose level	空腹血糖水平	mg/dL
cholesterol_hdl	int	High-density lipoprotein (HDL) cholesterol	高密度脂蛋白胆固醇	mg/dL
cholesterol_ldl	int	Low-density lipoprotein (LDL) cholesterol	低密度脂蛋白胆固醇	mg/dL
triglycerides	int	Triglyceride level	甘油三酯水平	mg/dL
EKG_results	str	Electrocardiogram results	心电图结果	Normal, Abnormal
previous_heart_disease	int	History of heart disease	既往心脏病史	1=Yes, 0=No
medication_usage	int	Use of heart-related medications	用药情况	1=Yes, 0=No
participated_in_free_screening	int	Participation in free health screening program	参与免费筛查项目	1=Yes, 0=No
Target Variable				
heart_attack	int	Occurrence of heart attack	心脏病发作	1=Yes, 0=No

1.2 Data Visualization & Summary

In this section, we visualize the dataset by ploting the histogram for each variable. For variables with continuous values, we conduct density estimation using the Parzen's kernel estimator:

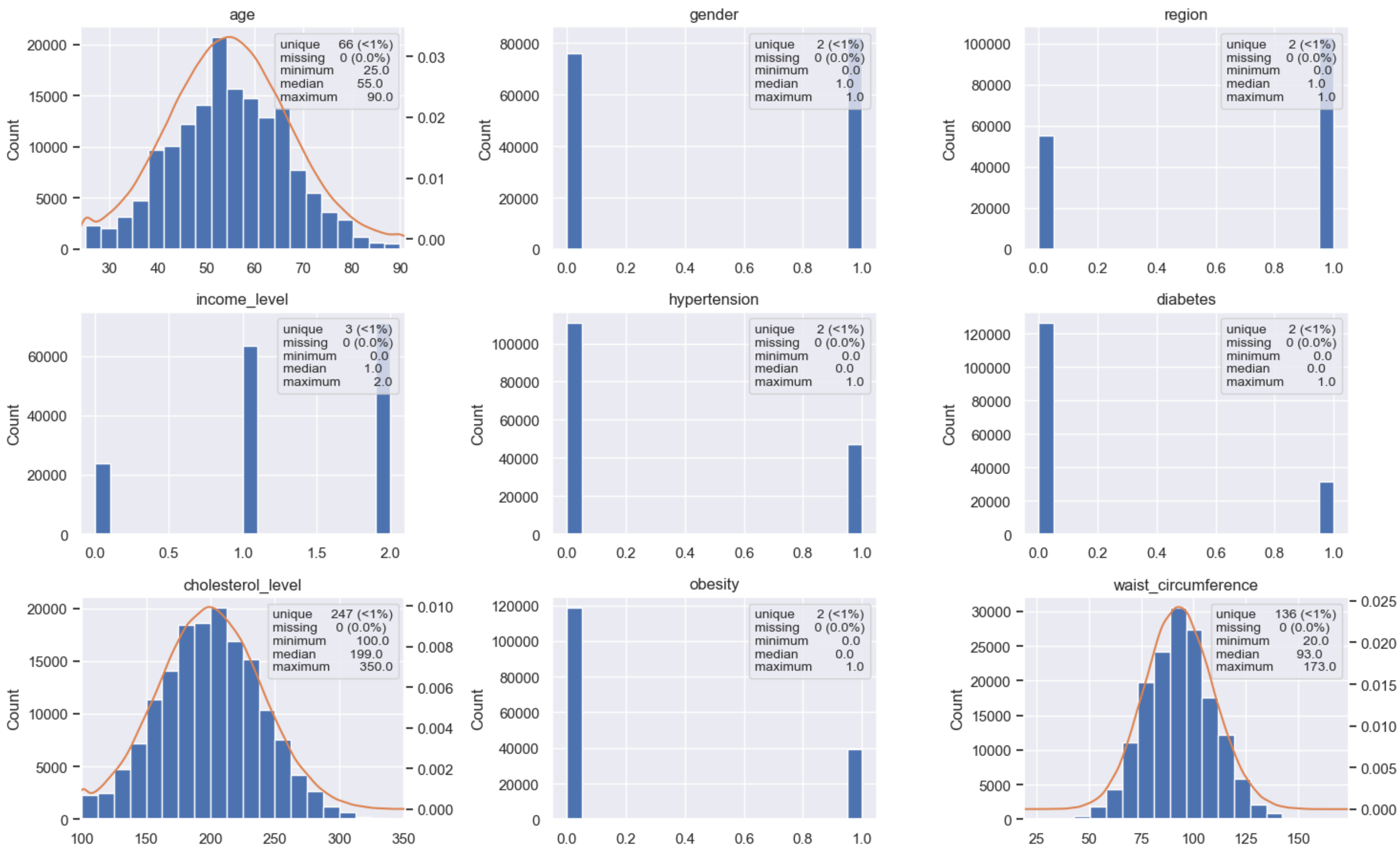
$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{x - X_i}{h}) \tag{1}$$

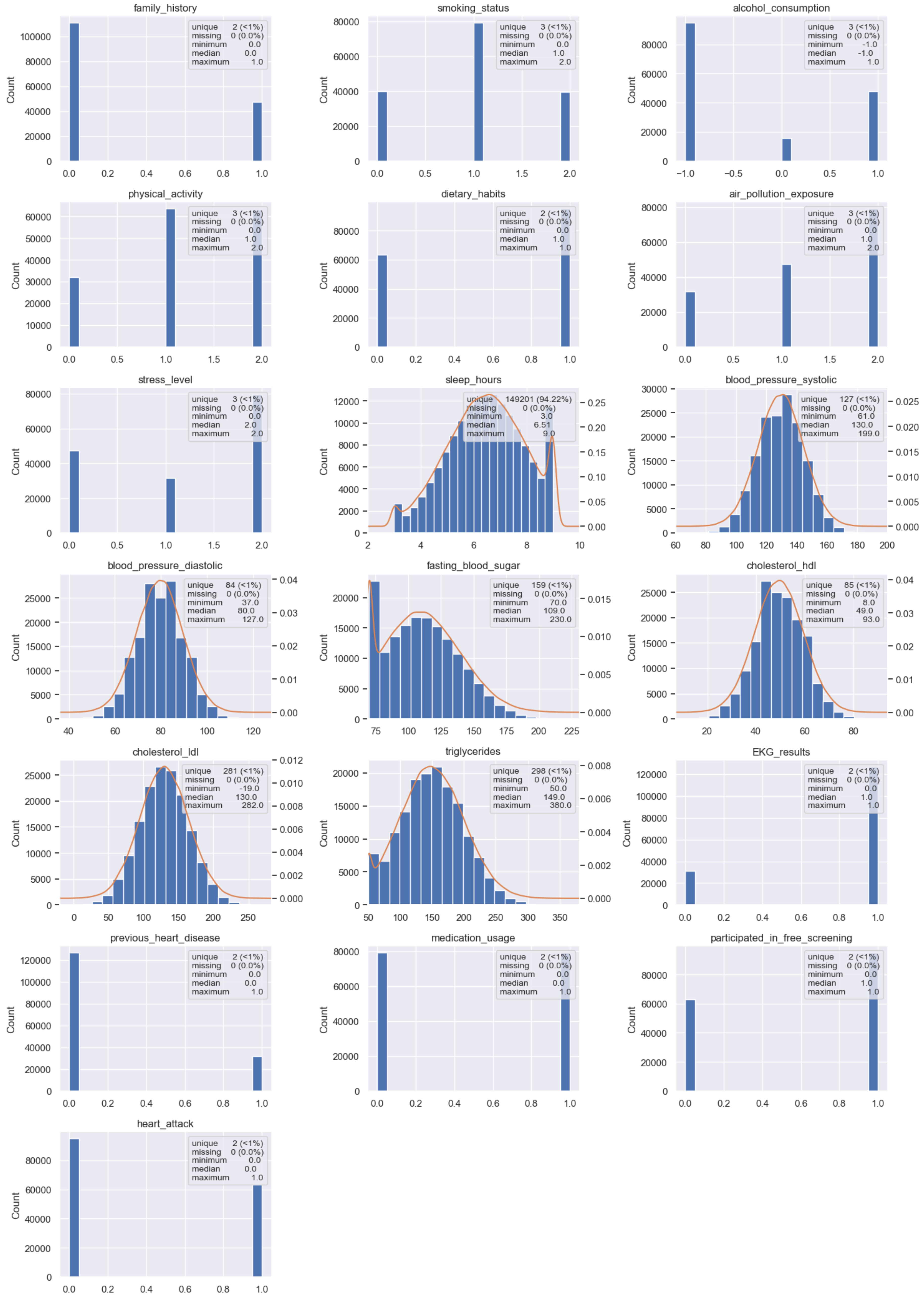
where the Normal kernel is picked as the kernel function $K(\cdot)$; the bandwidth h is picked according to the Silverman's rule of thumb.

A summary is attached for each variable, which includes the variable's number of unique and missing values, minimum, median, and maximum.

Remark:

Some variables take categorical values (e.g. `income_level`: Low-Middle-High). These variables are converted into the numerical data type based on the categorical levels (e.g. `income_level`: 0-1-2).





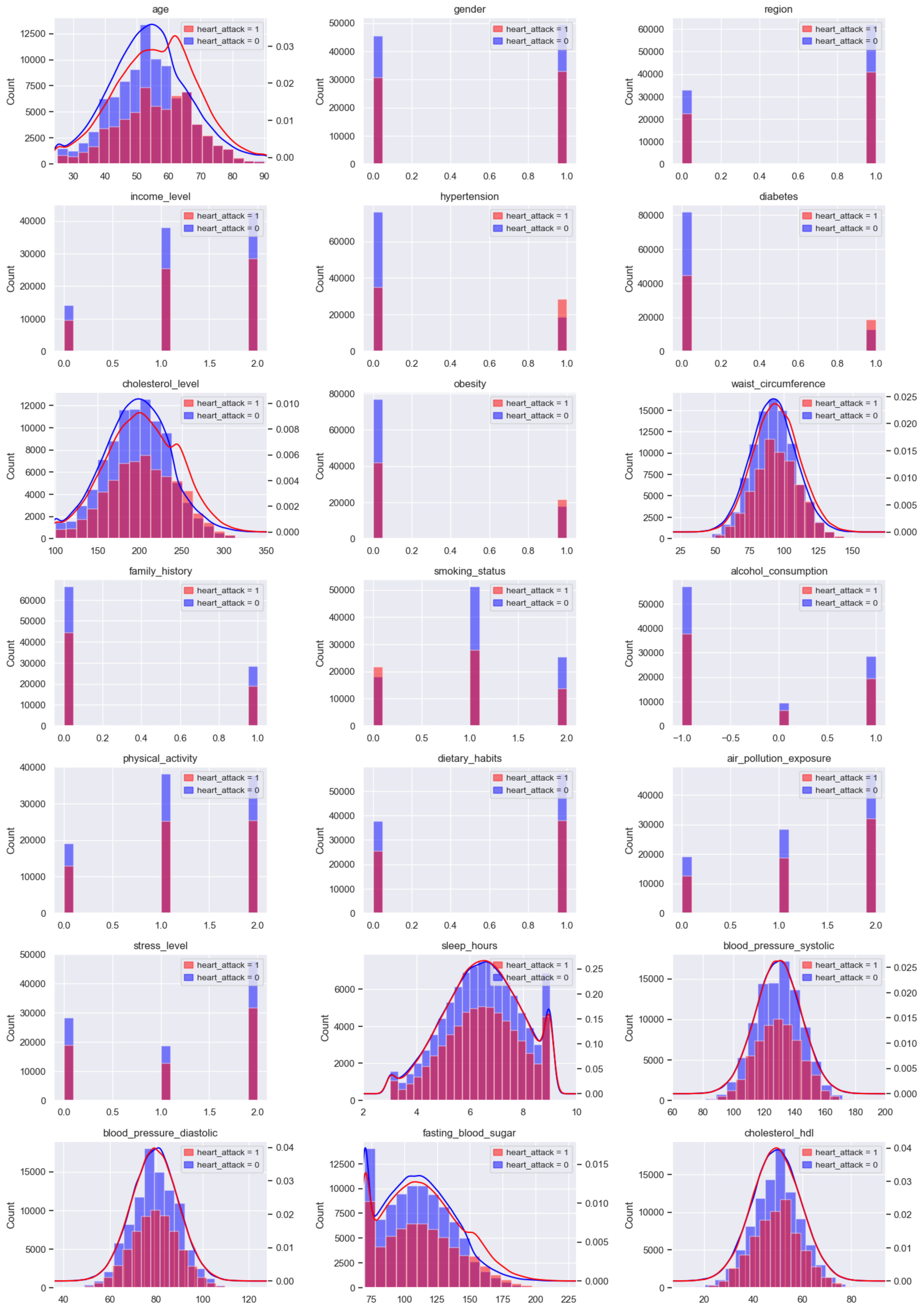
2. Data Preprocessing

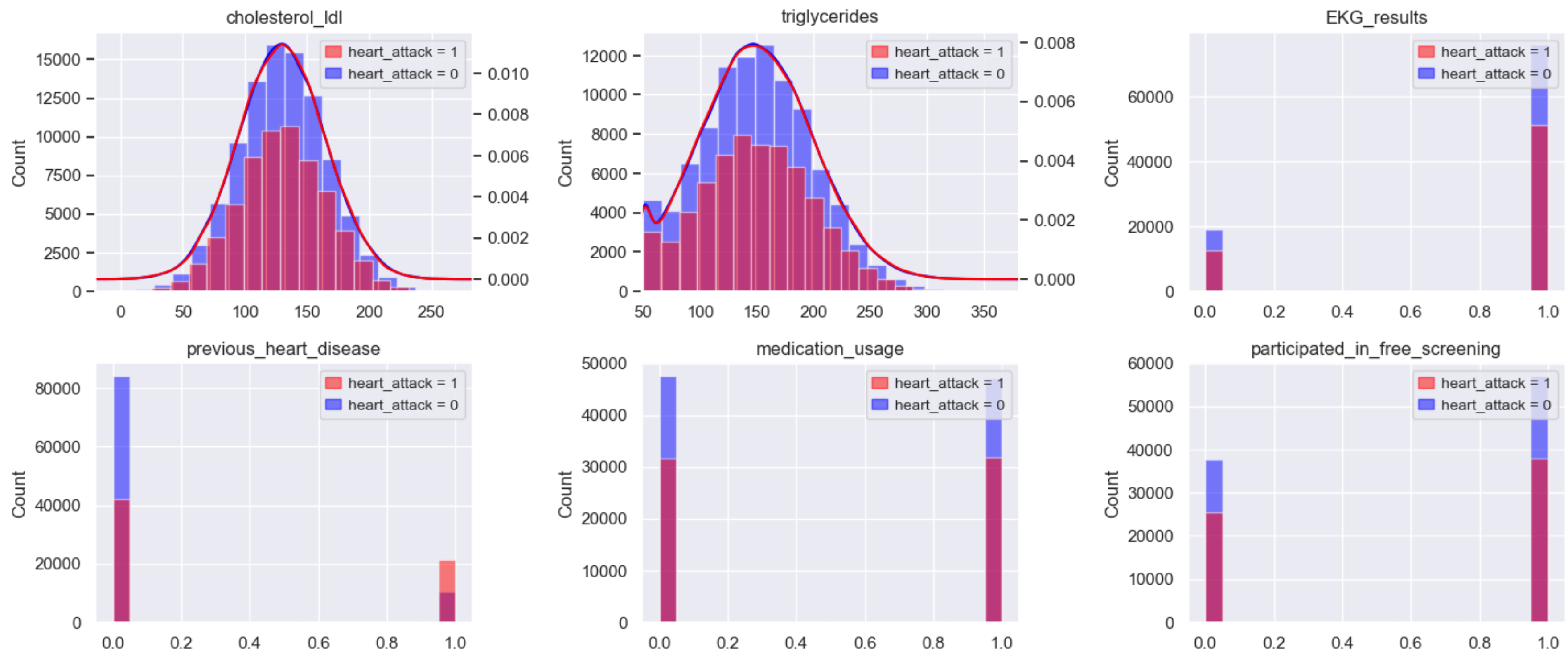
In this project, the dataset first undergoes the preprocsing procedure, **which includes feature selection, train-test split, and data scaling**.

2.1 Feature Selection (By Checking for Correlation via Density Estimation and Visualization)

The original features include 27 different variables, some of which may be irrelavent to our target variable `heart_attack`.

We first investigate the relationship between each feature variable and `heart_attack` through visualization. The whole dataset is splitted into two according to the value of `heart_attack`. The histogram plot and density estimation (with Normal kernel and rule-of-thumb bandwidth) are conducted for each feature variable in both datasets.





One key observation is that, for feature variables like `gender`, `blood_pressure_diastolic`, `sleep level`, their distributions bear almost no difference under different values of `heart_attack`. This observation indicates that, these variables may make little contribution to the prediction of heart attack.

Remark:

This observation may sounds counter-intuitive at first glance. For example, some medical research has revealed the strong relationship between heart attack and abnormal blood pressure. How can `blood_pressure_diastolic` *be irrelavent with* `heart_attack`*?*

My speculation is that for diastolic blood pressure, it will be considered abnormal when its value larger than 100*. Such data may indeed imply the risk of heart attack, but they are also very rare in the whole population. Therefore, in a general sense,* `blood_pressure_diastolic` *looks like irrelavent with* `heart_attack`.

Also, a notable observation is that `hypertension` *is related to* `heart_attack`.

2.2 Feature Selection (By Checking for Correlation via Hypothesis Testing)

Simply doing visualization does not provide us a concrete threshold for feature selection. We conduct feature selection through the following procedures:

1. Compute the (absolute) Pearson's correlation between each feature and `heart_attack`.
2. For each feature, conduct a Spearsman's ρ correlation coefficient test with the following hypothesis. Since it's a multiple hypothesis testing, we control the FDR through Benjamini-Hochberg procedure.

$$H_{0i}: \text{feature i is independent with heart attack} \quad v.s. \quad H_{1i}: otherwise \tag{2}$$

3. For each feature, conduct a two-sample Kolmogorove-Smirnov test with the following hypothesis. Since it's a multiple hypothesis testing, we control the FDR through the Benjamini-Hochberg procedure.

$$H_{0i}: \text{the distribution of feature i bears no difference under heart attack} = 0 \text{ or heart attack} = 1 \quad v.s. \quad H_{1i}: otherwise \tag{3}$$

We select a feature as long as any one of the two tests shows statistical significance.

Remark:

Why do I control the FDR instead of FWER?

Since it's in the preprocessing procedure, compared to missing potentially useful features, including potentially useless features is more affordable. Therefore, we prefer a less conservative procedure to control the multiple hypothesis testing.

It also explains why I select a feature even when only one of the tests shows significance.

	Correlation	Abs_correlation	KS_p_adj	Spearman_rou_p_adj	Selected
previous_heart_disease	0.274775	0.274775	0.000000e+00	0.000000e+00	True
hypertension	0.269261	0.269261	0.000000e+00	0.000000e+00	True
diabetes	0.194512	0.194512	0.000000e+00	0.000000e+00	True
obesity	0.171720	0.171720	0.000000e+00	0.000000e+00	True
smoking_status	-0.139962	0.139962	0.000000e+00	0.000000e+00	True
age	0.105756	0.105756	0.000000e+00	0.000000e+00	True
cholesterol_level	0.092611	0.092611	0.000000e+00	3.384134e-263	True
fasting_blood_sugar	0.069826	0.069826	1.390236e-135	1.610945e-115	True
waist_circumference	0.067883	0.067883	1.034631e-114	6.174598e-150	True
alcohol_consumption	0.005742	0.005742	1.787037e-01	4.227025e-02	True
region	-0.005585	0.005585	4.999502e-01	6.442110e-02	False
dietary_habits	-0.005271	0.005271	4.999502e-01	8.084381e-02	False
medication_usage	0.004694	0.004694	6.684006e-01	1.283408e-01	False
air_pollution_exposure	0.003909	0.003909	4.999502e-01	1.588202e-01	False
participated_in_free_screening	-0.003656	0.003656	9.207673e-01	2.468746e-01	False
gender	-0.003502	0.003502	9.207673e-01	2.596670e-01	False
stress_level	-0.003429	0.003429	7.401756e-01	2.468746e-01	False
EKG_results	0.002583	0.002583	1.000000e+00	4.560036e-01	False
income_level	-0.001941	0.001941	1.000000e+00	5.847724e-01	False
blood_pressure_systolic	-0.001644	0.001644	9.207673e-01	7.489109e-01	False
family_history	0.001374	0.001374	1.000000e+00	7.493712e-01	False
physical_activity	-0.000751	0.000751	1.000000e+00	9.101568e-01	False
triglycerides	-0.000709	0.000709	9.940624e-01	7.493712e-01	False
sleep_hours	0.000673	0.000673	7.401756e-01	9.101568e-01	False
cholesterol_hdl	0.000648	0.000648	9.207673e-01	8.857649e-01	False
cholesterol_ldl	0.000632	0.000632	1.000000e+00	8.744611e-01	False
blood_pressure_diastolic	-0.000301	0.000301	8.206312e-01	7.493712e-01	False

The above table is ordered according to the absolute value of Pearson's correlation.

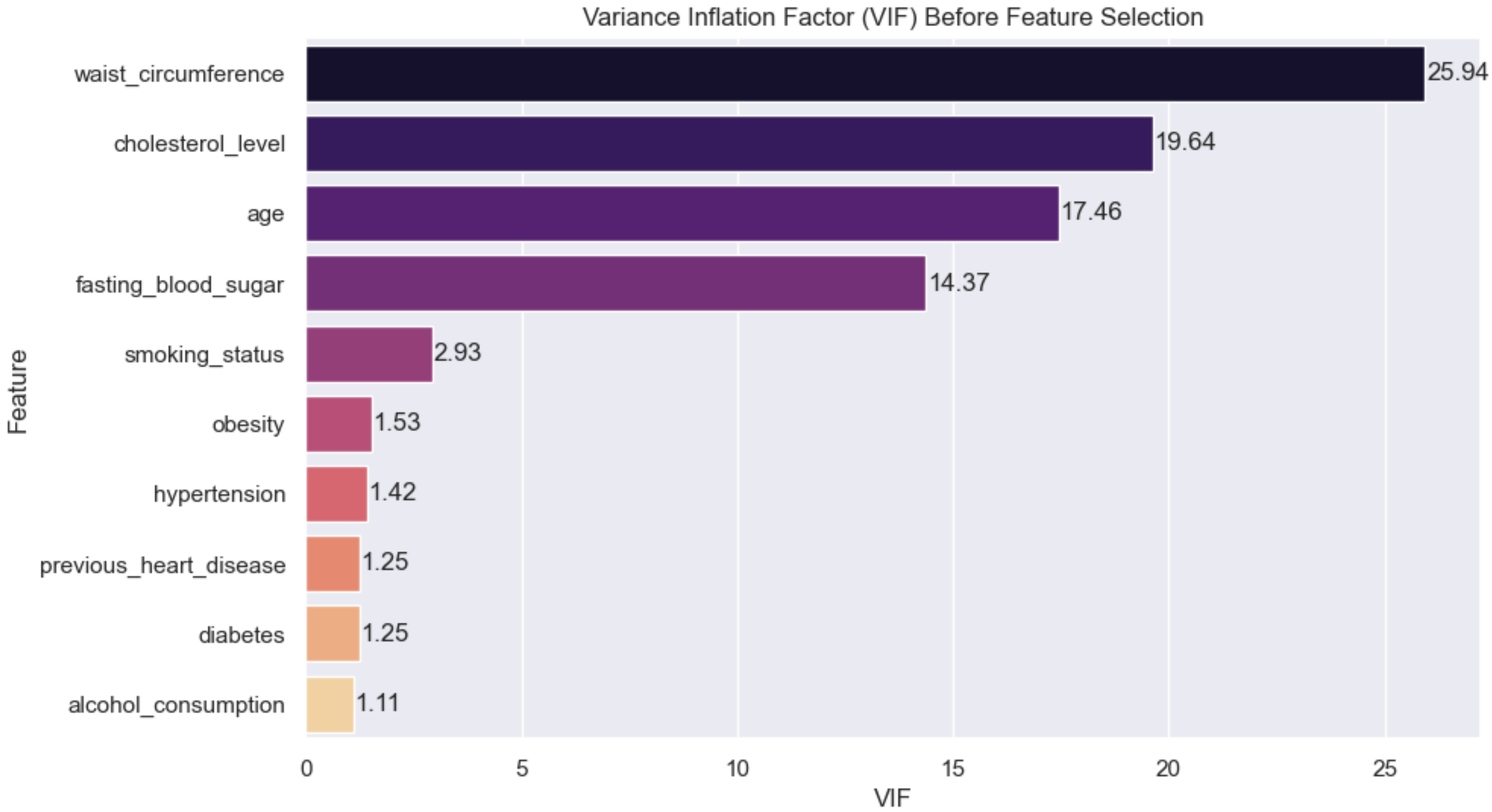
After this procedure, the following 10 features are selected to conduct subsequent analysis: `previous_heart_attack`, `hypertension`, `diabetes`, `obesity`, `smoking_status`, `age`, `cholesterol_level`, `fasting_blood_suger`, `waist_circumference`, `alcohol_consumption`.

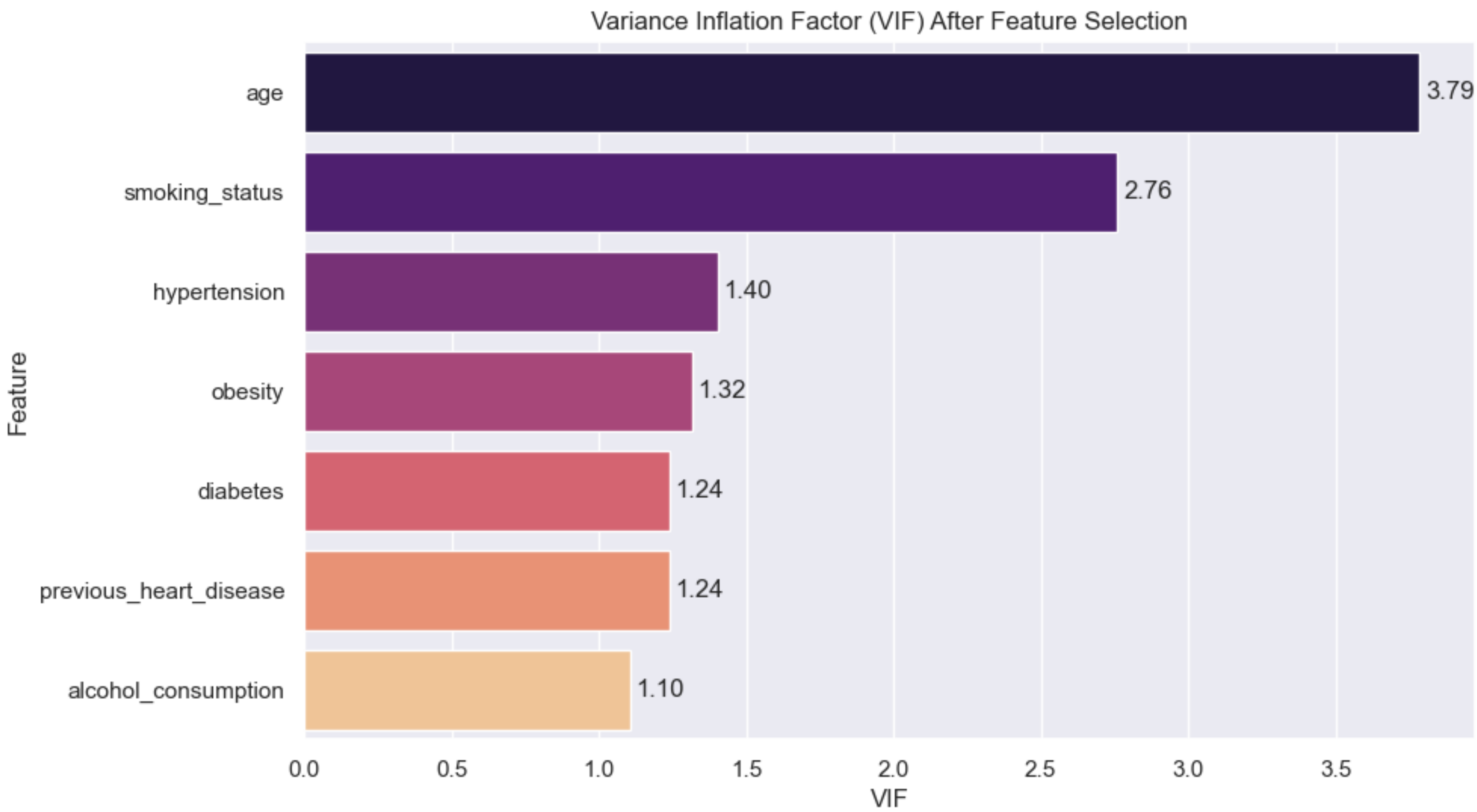
2.3 Feature Selection (By Checking for Multicollinearity)

Notice that for the selected 10 features, some of them may have strong correlation with each other (e.g. `waist_circumference` and `obesity`, `fasting_blood_sugar` and `diabetes`). To enhance the explanatory power of our analysis, we consider checking for the multicollinearity.

We compute the variance inflation factor (VIF) for each feature, and consider filtering out features with VIF > 10.

After filtering out `waist_circumference`, `cholesterol_level`, and `fasting_blood_sugar`, the VIF for each feature becomes less than 4, which is acceptable.





2.4 Train-Test Split & Data Scaling

To evaluate the effectiveness of our prediction, we split the whole dataset into a training set (which accounts for 80 of the data) and a testing set (which accounts for 20 of the data).

A Z-score normalization is also performed for each feature.

3. Prediction

We consider using **nonparametric regression** to predict heart attack.

3.1 Nonparametric Regression with Different Kernels

Under the nonparametric regression setting, the model is given by

$$Y_i = m(X_{1i}, X_{2i}, \dots, X_{7i}) + \epsilon_i, \quad i = 1, \dots, n \tag{4}$$

where $\{\epsilon_i\}_{1 \le i \le n}$ are i.i.d. random errors with mean 0 and variance σ^2.

Consider using the **Product Kernel**

$$K(u_1, u_2, \dots, u_7) = K_1(u_1)K_2(u_2)\dots K_7(u_7). \tag{5}$$

Given independent samples $\{(X_{1i}, X_{2i}, \dots, X_{7i}, Y_i)\}_{1 \le i \le n}$, the Nadaraya-Watson kernel estimator of $m(x)$ is given by

$$\hat{m}(x_1, x_2, \dots, x_7) = \frac{\sum_{i=1}^n Y_i K(\frac{x_1 - X_{1i}}{h_1}, \dots, \frac{x_7 - X_{7i}}{h_7})}{\sum_{i=1}^n K(\frac{x_1 - X_{1i}}{h_1}, \dots, \frac{x_7 - X_{7i}}{h_7})} \tag{6}$$

where $K(\cdot)$ is the product kernel and h is the bandwidth. For simplicity, we consider all $K_1, \dots, K_7$ to be the same kernel function, and all $h_1, \dots, h_7$ to be the same bandwidth.

The kernel function is chosen from:

- Normal kernel: $K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}$
- Double exponential kernel: $K(u) = \frac{1}{2} e^{-|u|}$
- Triangle kernel: $K(u) = (1 - |u|)\mathbf{1}_{\{|u| \le 1\}}$
- Uniform kernel: $K(u) = \mathbf{1}_{\{|u| \le 0.5\}}$
- Triweight kernel: $K(u) = (1 - u^2)^3 \mathbf{1}_{\{|u| \le 1\}}$

$\hat{m}(x_1, x_2, \dots, x_7)$ can be understood as a heart attack score (not necessarily meaning probability). And we predict `heart_attack` $= 1$ if $\hat{m}(x_1, x_2, \dots, x_7) > 0.5$, `heart_attack` $= 0$ otherwise.

Remark: Here we do not bother too much about the bandwidth here.

```
Accuracy of non-parametric regression with normal kernel: 0.7229
Accuracy of non-parametric regression with double exponential kernel: 0.7230
Accuracy of non-parametric regression with triangular kernel: 0.7222
Accuracy of non-parametric regression with uniform kernel: 0.7221
Accuracy of non-parametric regression with triweight kernel: 0.7234
```

3.2 Nonparametric Regression with Cross Validation

To pick the best bandwith, we consider using the cross validation.

```
kr_model = kr.KernelReg(endog=y_train, exog=X_train, var_type='o'*X_train.shape[1])
results = kr_model.fit(X_test)
y_pred_lst = results[0]
y_pred_lst = np.where(y_pred_lst > 0.5, 1, 0)
accuracy_bagging = accuracy_score(y_test, y_pred_lst)
print(f"Accuracy of non-parametric regression with cross validation: {accuracy_bagging:.4f}")
```

Remark: Unfortunately, the code has been running for over ten hours and the result has not shown up till now.

3.3 Bagging (Bootstrap Aggregation)

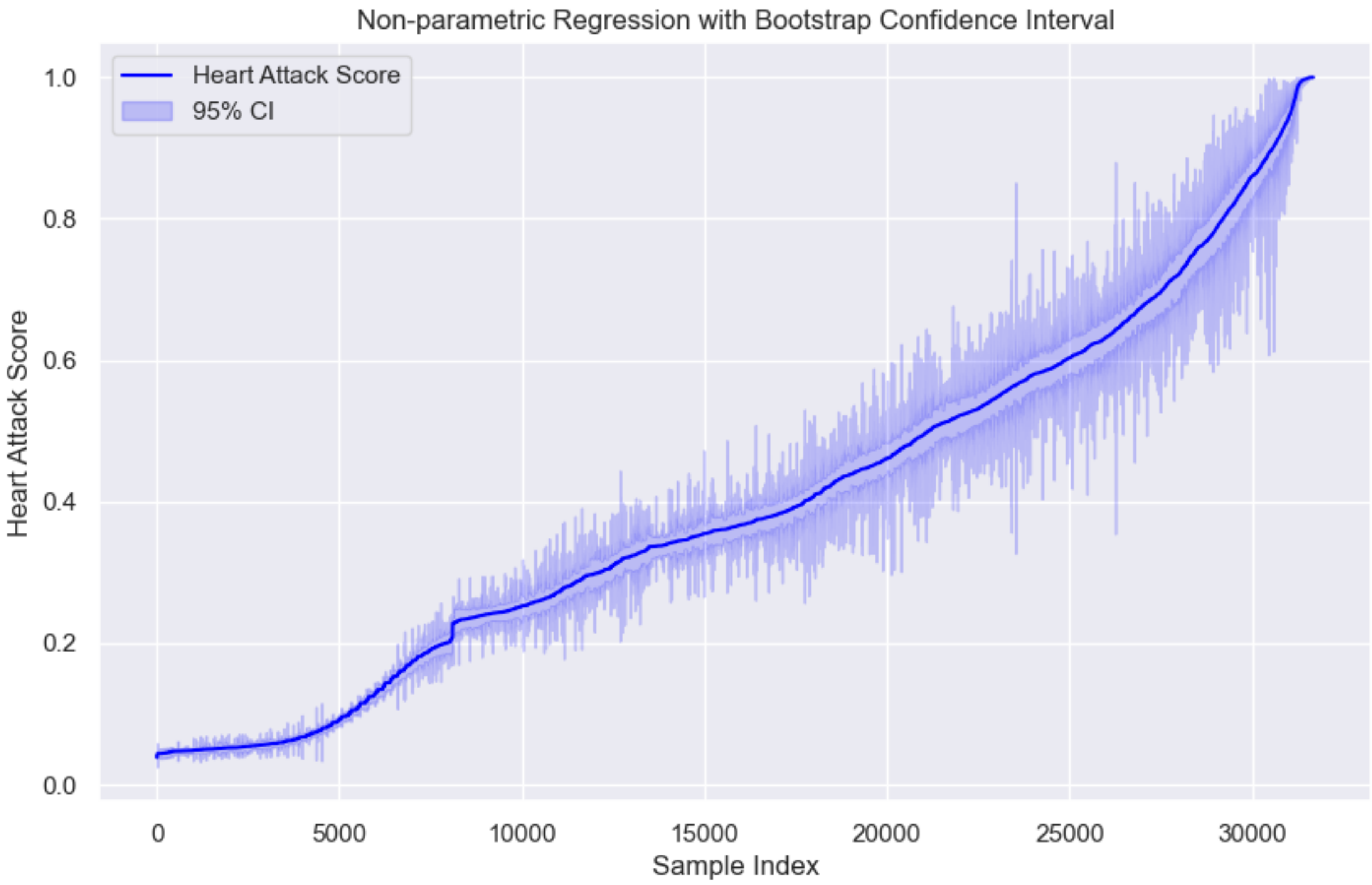
Different from Bootstrap, bagging is a method in ensemble learning. By aggregating several weak predictors (that may be easy to overfit and influenced by outliers), bagging can reduce the variance of the prediction.

But in principle, bagging relies on bootstrap (resampling). Here we resample the training set 10 times (without replacement) and train 10 different nonparametric regression predictor. We obtain the final prediction by taking the mean of the results of 10 predictors.

```
Accuracy of non-parametric regression with normal kernel and bagging: 0.7228
```

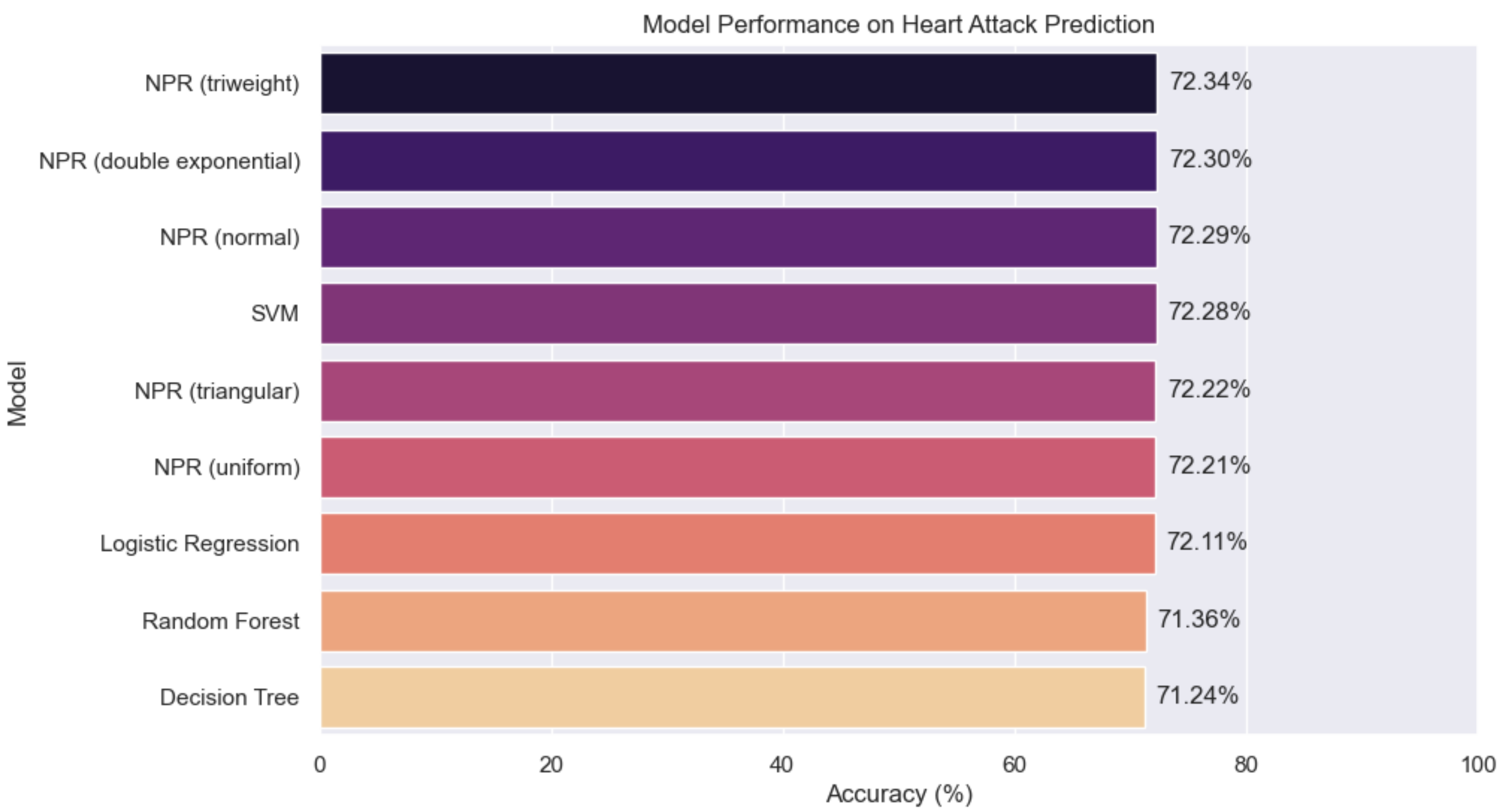
3.4 Bootstrap Confidence Interval for Heart Attack Score

We consider construct a bootstrap confidence interval for our heart attack score prediction $\hat{m}(x_1, x_2, \dots, x_7)$ (using normal kernel and bandwidth 0.5).



4. Evaluation

To evaluate the performance of our prediction algorithms, we also consider implementing several benchmark methods. Here we implemente logistic regression, decision tree, random forest, and SVM (with RBF kernel) and derive their accuracies.



After comparing with these benchmark methods, we can see that non-parametric regression has a slightly better performance.