Feature Subset Evaluation Criteria

An evaluation criterion is a process that uses a variety of techniques to extract the relevant feature from the feature sets. The study utilized filter method as the feature subset evaluation for the models using Chi^2 test.

Filter Method

In this method, features are selected based on their relation to the output, or how they are correlating to the output.

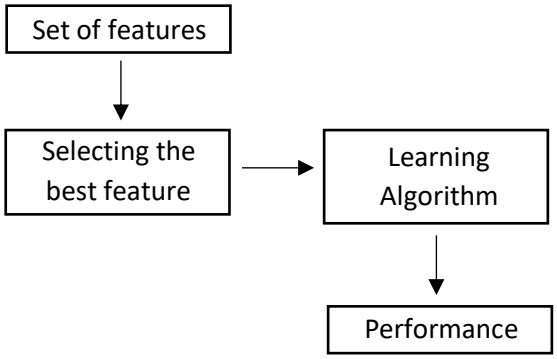

Figure 1. Structure of filter method.

Chi-Squared Test

A statistical hypothesis test used extensively in analyzing categorical data. Its primary purpose is to determine whether there's a statistically significant difference between observed data and expected data. It helps evaluate if a hypothesis about the relationship between two categorical variables holds. The null hypothesis typically states that there is no relationship between the variables you are analyzing. The chi-square statistic (χ^2) is calculated by summing up the squared differences between the observed (O) and expected (E) frequencies, divided by the expected frequency (E) for each category.

$$x^2 = \sum_{k=1}^{n} \frac{(O_k - E_k)^2}{Ek}$$

Logistic Regression

Logistic regression assigns a coefficient to each feature in the model. The magnitude and direction of the association between the feature and the target variable are shown by these coefficients. Features with large coefficients (positive or negative) are likely more influential in

predicting the outcome. By analyzing these coefficients, you can identify the most relevant features that significantly impact the model's predictions. Logistic regression contrasts a binary dependent feature or variable against an independent one using the logistic function. To model binary dependent characteristics or variables against the independent one, use the following equation.

$$f(x) = \frac{1}{1 + e^{-x}}$$

Dichotomous variables (0 or 1) can be predicted by using Logistic regression.

In this study Chi^2 test and logistic regression is utilized in the feature selection process. Chi^2 test is used to determine the features with statistically insignificant relationship with the target variable. Logistic regression is later on utilized to for more advance and deeper understanding of the remaining features.

The chi-squared scores revealed strong associations between several features and the outcome variable. Age has the strongest association, followed by diabetes, stroke, and high blood pressure (HBP). Even features like body mass index (BMI), sex, physical activity (PA), and heavy alcohol consumption (HAC) show statistically significant associations, though to a lesser degree.

Feature No.	Feature Name	Chi2 Statistics Score	p-value
9	Age	14469.840724	0.000000e+00
5	Diabetes	13533.426473	0.000000e+00
4	Stroke	10029.967794	0.000000e+00
0	HBP	6349.125913	0.000000e+00
1	HBC	4773.620749	0.000000e+00
3	Smoker	1850.013591	0.000000e+00
2	BMI	1092.568061	1.362007e-239
8	Sex	1052.373263	7.421522e-231
6	PA	470.679946	2.279333e-104
7	HAC	201.223900	1.129155e-45

Figure 2. Chi-Squared Scores and p-value

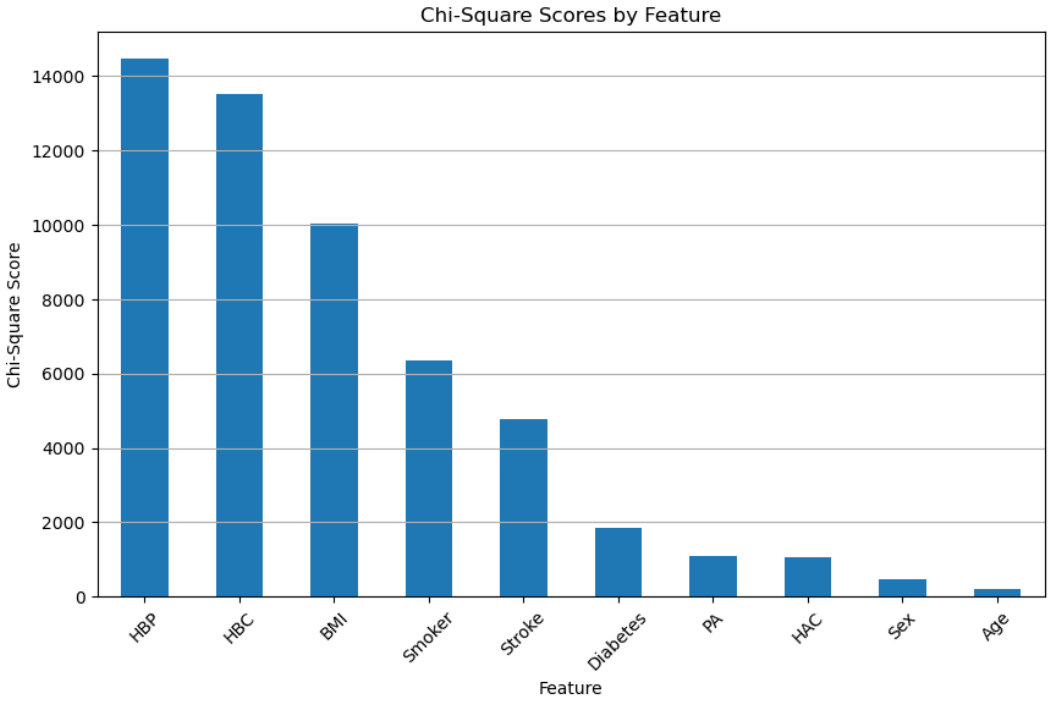

Figure S. Chi-Squared Scores by Feature

By analyzing the logistic regression feature coefficients, the researcher gains valuable insights into feature importance of the dataset. The output of the feature coefficient helped the researchers to select the most relevant features for building a robust and interpretable model. With having a heart disease or attack feature (HAD) feature as the target variable. Having a history of stroke, high blood pressure, high blood cholesterol is associated with the greatest increase in the log-odds of the target variable. Making the said features to have a positive influence or stronger positive association on the likelihood of the target variable. On the other hand, physical activity (PA) and heavy alcohol consumption (HAC) has negative influence based on the feature coefficient data. These two features are considered to have a negative association with a decrease in the log-odds of the target variable.

Feature	Feature Coefficient Scores
Stroke	1.3136
HBP	0.7025
HBC	0.6675
Sex	0.6094
Smoker	0.4988
Diabetes	0.2927
Age	0.2423
BMI	0.0103
PA	-0.2814
HAC	-0.4413

Figure 4. Logistic Regression Feature Coefficient Scores

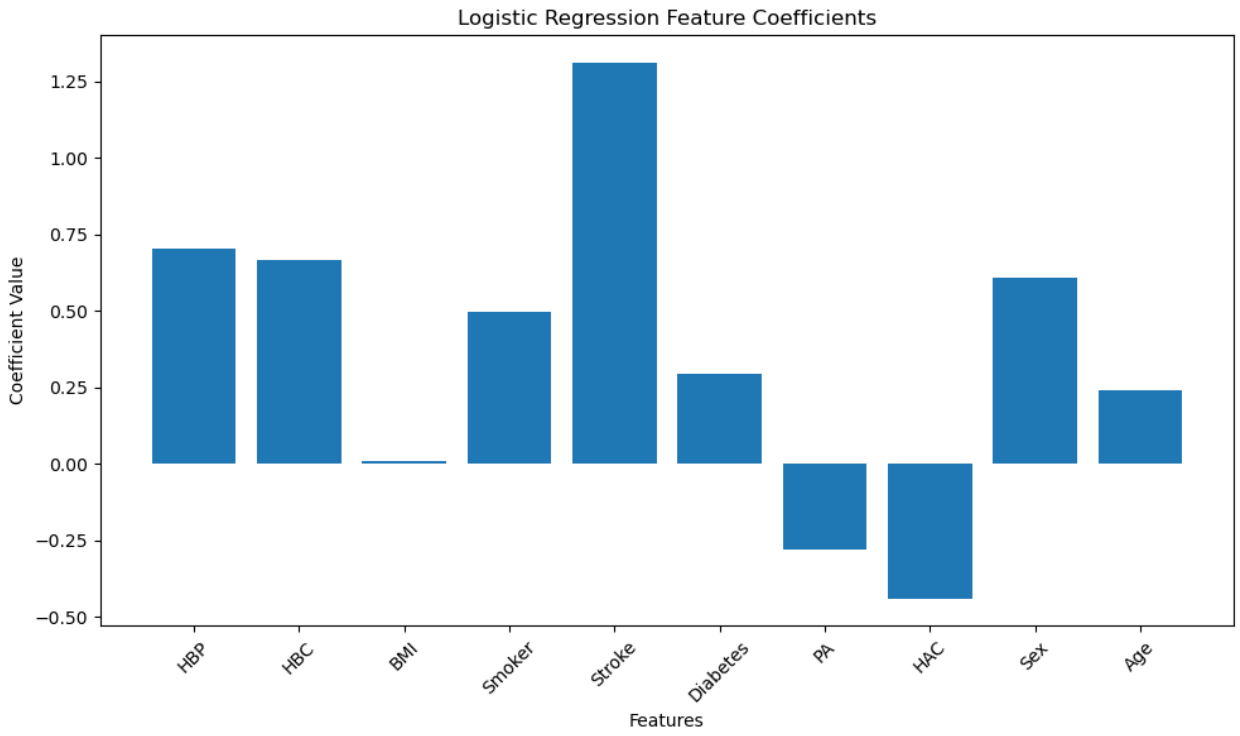

Figure 5. Logistic Regression Feature Coefficient by Feature