Non-probabilistic supervised classification

Machine Learning

Pablo Bande Sánchez Girón

1. Introduction

1.1. Problem

This experiment aims to predict heart diseases using a dataset obtained from the Centers for Disease Control and Prevention. The dataset is part of the Behavioral Risk Factor Surveillance System, which conducts annual telephone surveys to collect health data of U.S. residents. The objective is to implement five different algorithms (**Neural Network, k-Nearest Neighbors, Rule Induction, Support Vector Machines, and Classification Trees**) and evaluate their performance in various scenarios, including using **all variables**, applying **univariate** and **multivariate** filter feature subset selection, and utilising multivariate **wrapper feature subset selection**. The evaluation will consider factors such as accuracy and recall, and the behaviour of the algorithms, such as examining the tree structure or rules generated.

1.2. Dataset

The initial dataset consisted of 319,795 entries and 18 columns. Each column is described as follows:

- HeartDisease: Indicates the presence or absence of heart disease.
- BMI: Body Mass Index, a numerical value representing body composition.
- Smoking: Indicates smoking habits.
- AlcoholDrinking: Indicates alcohol consumption habits.
- Stroke: Indicates the presence or absence of a stroke.
- PhysicalHealth: Numeric value representing physical health.
- MentalHealth: Numeric value representing mental health.
- DiffWalking: Indicates difficulty in walking.
- Sex: Indicates the gender of the individual.
- AgeCategory: Indicates the age category of the individual.
- Race: Indicates the race of the individual.
- Diabetic: Indicates if the individual has diabetes.
- PhysicalActivity: Indicates the level of physical activity.
- GenHealth: Indicates general health status.
- SleepTime: Numeric value representing the duration of sleep.
- Asthma: Indicates if the individual has asthma.
- KidneyDisease: Indicates if the individual has kidney disease.
- SkinCancer: Indicates if the individual has skin cancer.

2. Methodology

2.1. Algorithms

- **K-Nearest Neighbors** (KNN): It predicts outcomes based on the proximity of data points, assigning the majority class among the k-nearest neighbours to a new data point. It's commonly used for predicting heart diseases, with k being an odd number and typically set as sqrt(N)/2, where N is the number of total samples.

- **Rule Induction** (RIPPER): algorithm that generates rules from data. It iteratively refines and prunes rules based on training data, aiming to produce accurate rules for classifying instances correctly.

- **Support Vector Machines** (SVM): finds an optimal hyperplane to separate different classes in a high-dimensional space. It identifies a decision boundary maximising the margin between positive and negative instances.

- **Artificial Neural Network** (ANN): consists of interconnected neurons organised in layers. It learns by adjusting weights between neurons to minimise error. In heart disease prediction, it takes input features, processes them through hidden layers, and produces a binary output indicating the presence or absence of heart disease.

- **Classification Tree** (J48)**:** builds a tree-like model for decision-making based on features. Internal nodes represent tests on features, branches represent outcomes, and leaves correspond to class labels. The algorithm partitions data based on features to maximise information gain. In heart disease prediction, it makes a series of feature-based decisions to determine the likelihood of a patient having heart disease.

2.2. Measurements:

To evaluate the heart disease detection model, the following measurements were considered:

- Accuracy: Measures the overall correctness of the model's predictions.
- Precision: Represents the proportion of true positive predictions to the total number of positive predictions.
- Recall: Measures the proportion of true positive predictions to the total number of actual positive instances.
- F1 Score: The harmonic mean of precision and recall, providing a balance between the two.
- AUC-ROC: Area Under the ROC Curve, used to assess the model's discrimination ability in distinguishing between positive and negative instances.

2.3. Dataset preprocessing

To prepare the data for analysis, the features were converted into numerical representations. Label encoding was used for columns with "Yes" and "No" values, ordinal encoding was used for the "GenHealth" column with values mapped from "Poor" to "Excellent," and age intervals were replaced with the midpoint value. For the "Sex" and "Race" columns, one-hot encoding was applied to create new columns with binary values.

It was clearly unbalanced at first due to the nature of the dataset because the amount of patients that suffer heart diseases is clearly on the lower side. In order to get a more solid model and use common metrics that are negatively affected by this lack of parity between variable values I decided to balance it by resampling the dataset and reducing the number of samples to work with lower computational resources in a reasonable execution time.

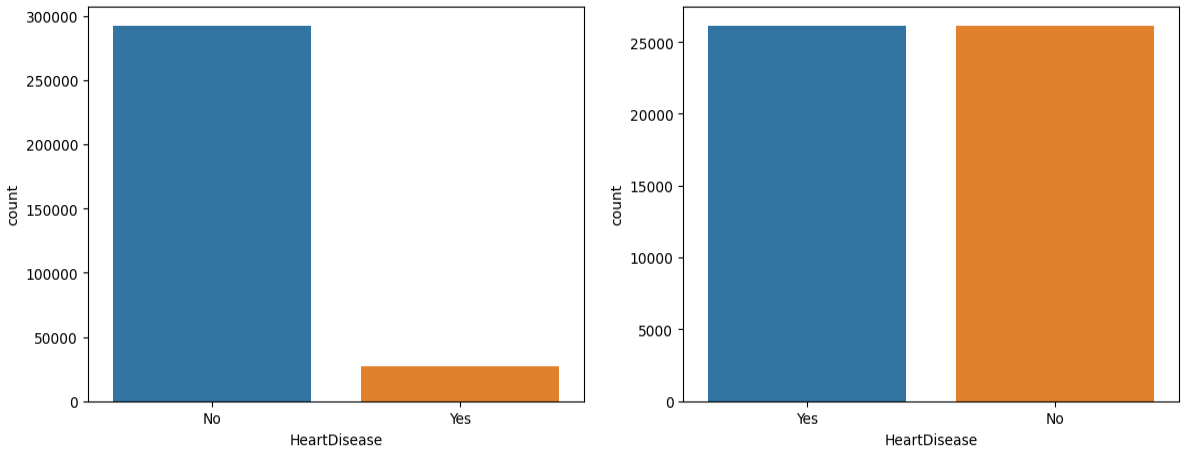

Image 1: Data distribution of target variable before and after balance.

Outliers were removed using the z-score method, and the data was normalised to address potential interference with certain algorithms that will be used such as KNN, SVM, ANN and Decision trees.

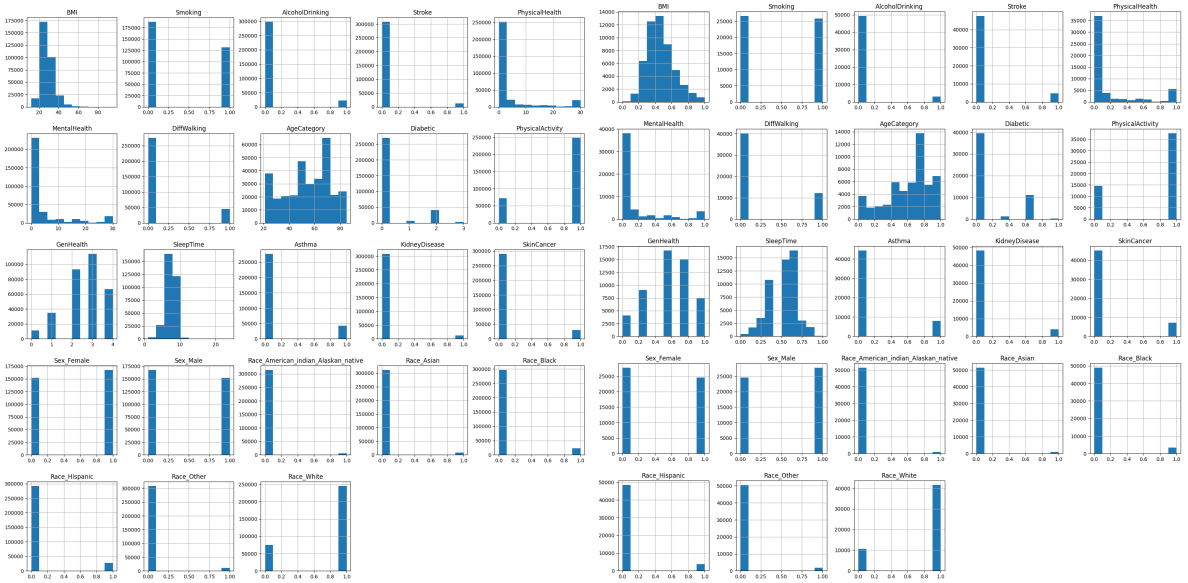

Image 2: Data distribution before and after removing outliers and normalising some columns.

After preprocessing, the dataset consisted of 24 columns, including the target variable (HeartDisease), and 52,296 samples. The dataset was randomly split into training and test sets with a proportion of 20% and 80%, respectively.

2.4. Feature selection

Various feature selection methods were applied, including:

- Whole dataset: All variables were used for analysis.
- Univariate filter feature subset selection: Subset selection based on gain ratio evaluation and a ranker with a threshold of 0.05 p-value.
- Multivariate filter feature subset selection: Subset selection using the CFS (Correlation-based Feature Selection) and GreedyStepwise algorithms.
- Multivariate wrapper feature subset selection: Wrapper approach using the ClassifierSubsetEval with a GreedyStepwise algorithm. Due to computational limitations, a subset of the data of size 5000 was used for wrapping. A GreedyStepwise approach was chosen instead of BestFirst due to the large amount of data and computational complexity involved.

Variable	Univariate	Multivariate	Wrapper KNN	Wrapper RIPPER	Wrapper SVM	Wrapper ANN	Wrapper TREE
BMI							
Smoking							
AlcoholDrinking							
Stroke							
PhysicalHealth							
MentalHealth							
DiffWalking							
AgeCategory							
Diabetic							
PhysicalActivity							
GenHealth							
SleepTime							
Asthma							
KidneyDisease							
SkinCancer							
Sex_Female							
Sex_Male							
Race_American...							
Race_Asian							
Race_Black							

Table 1: Feature selection results,marked variables that were obtained in each method

3. Results

In this section we will compare the results obtained for each model and talk about the best one for each algorithm.

KNN Model

For all but wrapper K = 103 using the criteria explained earlier. In the wrapper due to the data dimension reduction I ended up using 51 as K value. Filters definitely improved results but did not have a high impact overall.

Feature Selection	Precision	Recall	F-Measure	ROC Area	Accuracy
All variables	0.755	0.755	0.755	0.827	75.4685%
Univariate	0.757	0.753	0.752	0.829	75.3442%
Multivariate	0.767	0.750	0.749	0.821	76.4532%
Wrapper	0.767	0.765	0.765	0.838	76.5392%

Table 2: KNN results.

RIPPER Model

JRIP rules:

```
(GenHealth >= 0.75) and (AgeCategory <= 0.09375) and (Stroke <= 0) => HeartDisease=No (1980.0/88.0)
(GenHealth >= 0.75) and (AgeCategory <= 0.40625) and (Stroke <= 0) => HeartDisease=No (3840.0/323.0)
(GenHealth >= 0.5) and (AgeCategory <= 0.25) and (Stroke <= 0) => HeartDisease=No (1577.0/154.0)
(GenHealth >= 0.5) and (AgeCategory <= 0.484375) and (Stroke <= 0) => HeartDisease=No (3114.0/773.0)
(AgeCategory <= 0.71875) and (GenHealth >= 1) and (Diabetic <= 0) and (Stroke <= 0) => HeartDisease=No
(1457.0/283.0)
(GenHealth >= 0.75) and (AgeCategory <= 0.71875) and (Diabetic <= 0.333333) and (Stroke <= 0) =>
HeartDisease=No (3182.0/1041.0)
(AgeCategory <= 0.796875) and (Diabetic <= 0) and (GenHealth >= 1) and (Stroke <= 0) =>
HeartDisease=No (468.0/126.0)
(GenHealth >= 0.5) and (AgeCategory <= 0.5625) and (Stroke <= 0) and (DiffWalking <= 0) =>
HeartDisease=No (935.0/407.0)
(Diabetic <= 0) and (AgeCategory <= 0.875) and (Stroke <= 0) and (GenHealth >= 1) and (KidneyDisease
<= 0) => HeartDisease=No (260.0/106.0)
(GenHealth >= 0.75) and (Diabetic <= 0) and (AgeCategory <= 0.796875) and (Stroke <= 0) =>
HeartDisease=No (1215.0/583.0)
(DiffWalking <= 0) and (AgeCategory <= 0.328125) and (Stroke <= 0) => HeartDisease=No (570.0/164.0) =>
HeartDisease=Yes (23238.0/6364.0)

Number of Rules : 12
```

From these rules, we can deduce the following information:

- General health score (GenHealth) appears to be an important factor in predicting heart disease. Higher values of GenHealth are associated with a lower likelihood of heart disease.
- Age also plays a role in predicting heart disease. Younger individuals are generally less likely to have heart disease.
- The absence of a previous stroke is another indicator of lower risk for heart disease.
- Other variables such as diabetic status, difficulty in walking, and kidney disease were not explicitly mentioned in the summarised rules, suggesting that they may have less influence on the prediction of heart disease in this particular set of rules.

Feature Selection	Precision	Recall	F-Measure	ROC Area	Accuracy
All variables	0.764	0.761	0.761	0.788	76.109%
Univariate	0.753	0.750	0.749	0.781	75.0191%
Multivariate	0.765	0.763	0.762	0.790	76.2906%
Wrapper	0.763	0.761	0.760	0.790	76.0803%

Table 3: RIPPER results.

SVM Model

Feature Selection	Precision	Recall	F-Measure	ROC Area	Accuracy
All variables	0.770	0.769	0.769	0.769	76.9216%
Univariate	0.755	0.754	0.754	0.754	75.392%
Multivariate	0.767	0.766	0.766	0.766	76.6157%
Wrapper	0.766	0.765	0.765	0.765	76.4818%

Table 3: SVM results.

ANN Model

One hidden layer multi perceptron.

Feature Selection	Precision	Recall	F-Measure	ROC Area	Accuracy
All variables	0.764	0.764	0.763	0.839	76.3576%
Univariate	0.746	0.742	0.742	0.826	74.2447%
Multivariate	0.761	0.758	0.758	0.840	75.8317%
Wrapper	0.758	0.755	0.755	0.839	75.5354%

Table 4: ANN results.

J48 Model

The tree generated is way too complex to get solid knowledge so I tried to tweak some hyperparameters but still could not get a simpler one with the same or better performance.

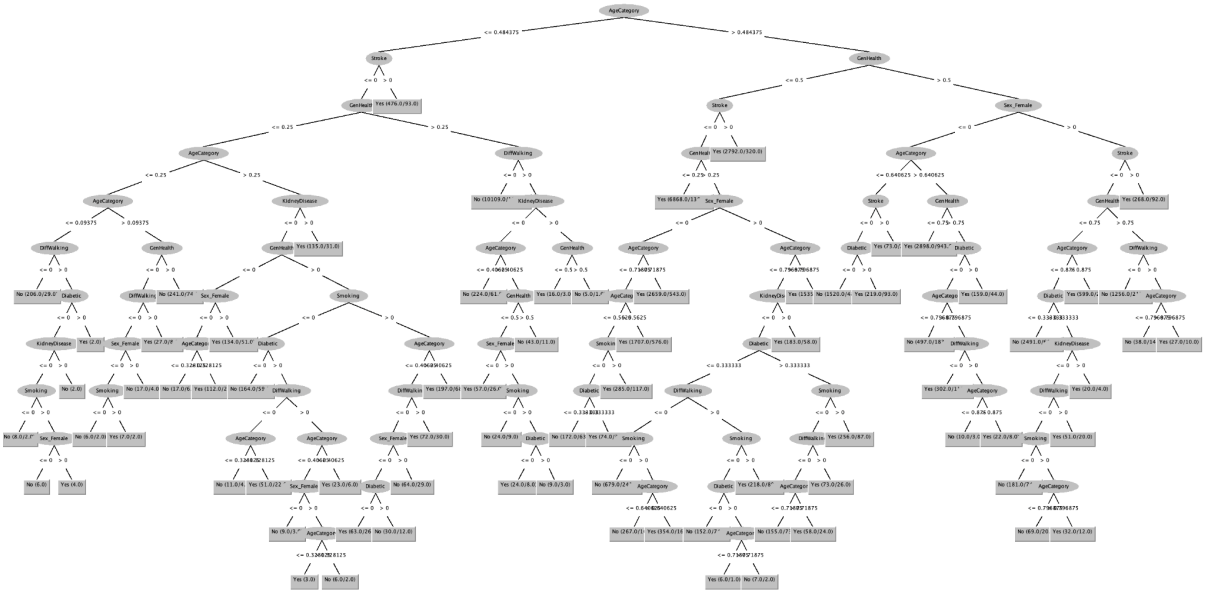

Image 3: Classification tree visual representation.

- AgeCategory: AgeCategory is the first variable used to split the data in the tree. It appears multiple times in different branches of the tree, indicating its significance in determining the outcome.
- GenHealth: GenHealth is another important variable considered in the decision tree. It is used in multiple splits, suggesting that it has a significant impact on the final prediction.
- Stroke: Stroke is also used as a splitting criterion in the tree. While it does not appear as frequently as AgeCategory and GenHealth, it still plays a role in the classification process.
- DiffWalking: DiffWalking is used in a few splits in the tree. Its inclusion suggests that it provides additional information for making predictions, although it may not be as influential as AgeCategory or GenHealth.
- KidneyDisease: KidneyDisease is considered in a couple of splits in the tree, indicating its relevance in certain scenarios.
- Diabetic: Diabetic is used in some splits, but its importance seems to be relatively lower compared to AgeCategory and GenHealth.

Feature Selection	Precision	Recall	F-Measure	ROC Area	Accuracy
All variables	0.748	0.747	0.747	0.785	74.6845%
Univariate	0.756	0.765	0.764	0.838	75.3537%
Multivariate	0.768	0.766	0.765	0.821	76.5679%
Wrapper	0.754	0.753	0.753	0.794	75.3155%

Table 5: J48 results.

All data

Feature Selection	Model	Precision	Recall	F-Measure	ROC Area	Accuracy
All variables	KNN	0.755	0.755	0.755	0.8274	75.4685 %
	RIPPER	0.764	0.761	0.761	0.788	76.109 %
	SVM	0.770	0.769	0.769	0.769	76.9216 %
	ANN	0.764	0.764	0.763	0.839	76.3576 %
	J48	0.748	0.747	0.747	0.785	74.6845 %
Univariate	KNN	0.757	0.753	0.752	0.829	75.3442 %
	RIPPER	0.753	0.750	0.749	0.781	75.0191 %
	SVM	0.755	0.754	0.754	0.754	75.392 %
	ANN	0.746	0.742	0.742	0.826	74.2447 %
	J48	0.756	0.765	0.764	0.838	75.3537 %
Multivariate	KNN	0.767	0.750	0.749	0.821	76.4532 %
	RIPPER	0.765	0.763	0.762	0.790	76.2906 %
	SVM	0.767	0.766	0.766	0.766	76.6157 %
	ANN	0.761	0.758	0.758	0.840	75.8317 %
	J48	0.768	0.766	0.765	0.821	76.5679 %
Wrapper	KNN	0.767	0.765	0.765	0.838	76.5392 %
	RIPPER	0.763	0.761	0.760	0.790	76.0803 %
	SVM	0.766	0.765	0.765	0.765	76.4818 %
	ANN	0.758	0.755	0.755	0.839	75.5354 %
	J48	0.754	0.753	0.753	0.794	75.3155 %

Table 6: All results.

4. Conclusion

The models generally exhibit similar performance across different feature selection techniques. The metrics such as precision, recall, F-measure, and ROC area are relatively close for each model. The impact of feature selection techniques on model performance is not significant. The metrics for these techniques are very similar to the results obtained when using all variables. This might be because most of the variables are relevant and contain useful information for predicting the target variable.

The models demonstrate reasonable accuracy in classifying positive instances, as indicated by precision values ranging from 0.746 to 0.770. They also exhibit the ability to identify positive instances from the total number of actual positives, with recall values ranging from 0.742 to 0.769. The F-measure values, ranging from 0.742 to 0.769, represent a balanced measure of precision and recall. The models show good discrimination and predictive ability, as reflected in the ROC area values ranging from 0.785 to 0.839. Higher values suggest effective differentiation between positive and negative instances. The accuracy values range from 74.2447% to 76.9216%, serving as a representative metric for evaluating model performance in this balanced dataset.

Overall, the results indicate that the models perform well according to the metrics used. The J48 model generally achieves lower results compared to other models, meanwhile the overall best one is SVM using all attributes.

5. References

[1] KamilPytlak. Personal Key Indicators of Heart Disease. Data retrieved from Kaggle, https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease. 2021.

[2] Pedro Larrañaga, Concha Bielza. Feature Subset Selection.

[3] Pedro Larrañaga, Concha Bielza. K-nearest neighbours.

[4] Pedro Larrañaga, Concha Bielza. Rule induction.

[5] Pedro Larrañaga, Concha Bielza. Support vector machines.

[6] Pedro Larrañaga, Concha Bielza. Artificial neural networks.

[7] Pedro Larrañaga, Concha Bielza. Classification trees.