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 three different probabilistic supervised classification algorithms (**Logistic Regression, Tree Augmented Naive Bayes** and **Linear Discriminant Analysis**) 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**. Besides, we will also work with meta classifiers such as **AdaBoost** and **Bagging**. The evaluation will consider factors such as accuracy and recall, and the behaviour of the algorithms.

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

2.1.1. Logistic Regression is a statistical model used for binary classification problems. It models the relationship between the input variables and the probability of the outcome belonging to a specific class. Logistic Regression uses the sigmoid function to map the linear combination of the input features to a probability value between 0 and 1. This probability is then used to make predictions using a threshold to classify instances into one of the two classes.

2.1.2. TAN (Tree Augmented Naive Bayes) is an extension of the Naive Bayes algorithm, which is based on Bayes' theorem. Naive Bayes assumes that all features are conditionally independent given the class label. TAN relaxes this assumption allowing to model more complex relationships between the features and the class label. TAN combines the simplicity and efficiency of Naive Bayes with the ability to model feature dependencies.

2.1.3. LDA (Linear Discriminant Analysis) is a dimensionality reduction technique commonly used for classification tasks. LDA aims to find a linear combination of the input features that maximises the separation between classes.

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.

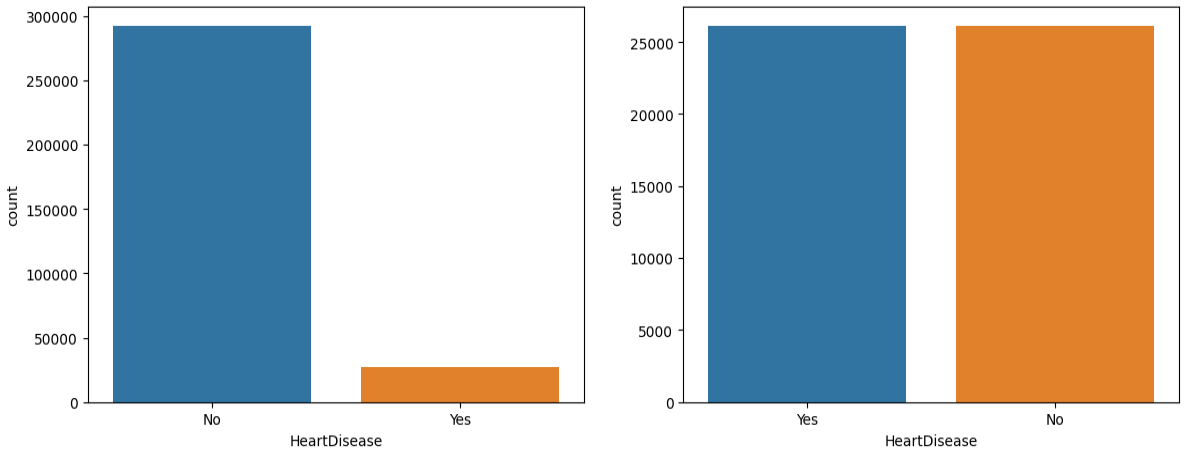

Image 1: Data distribution of target variable before and after balance.

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.

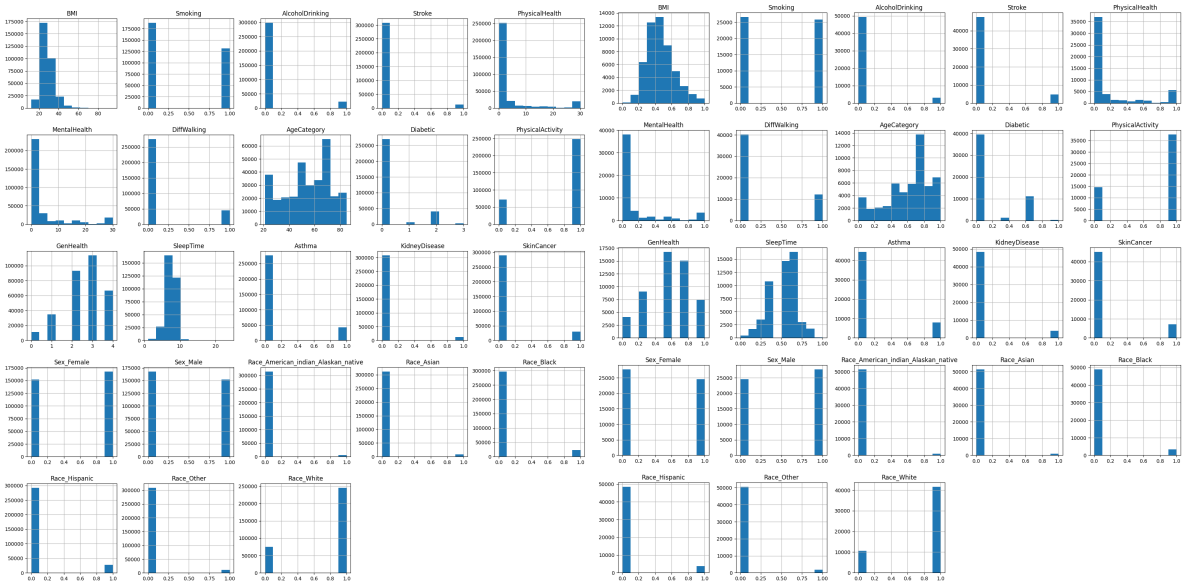

Image 2: Data distribution before and after removing outliers and normalising some columns.

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.

Looking at table 1, it is very clear that some of the variables have more importance than others, for example AgeCategory and GenHealth are present in all feature selections. On the other side, redundant variables like BMI, Physical activity, SleepTime and Sex_Male. Sex_Male might be because it is exactly the opposite of Sex_Female so it is kind of redundant information.

Variable	Univ	Multiv	Wrapper logistic	Wrapper TAN	Wrapper LDA	Wrapper Boosting logistic	Wrapper Boosting TAN	Wrapper Boosting LDA	Wrapper Bagging logistic	Wrapper Bagging TAN	Wrapper Bagging LDA
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											
Race_White											

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.

1. Logistic Model

The feature selection techniques show similar performance in terms of precision, recall, F-measure, ROC area, and accuracy for both logistic regression and bagging logistic. This suggests that the feature selection techniques have minimal impact on the model performance. However, the boosting logistic approach exhibits slightly lower performance in terms of correctly identifying positive instances, discriminatory power, and overall prediction correctness. This indicates that the boosting approach with logistic regression does not perform as good as the other.

Model	Feature Selection	Precision	Recall	F-Measure	ROC Area	Accuracy
Logistic	All Variables	0.768	0.768	0.768	0.843	76.826%
	Univariate	0.755	0.754	0.754	0.829	75.4302%
	Multivariate	0.765	0.765	0.765	0.840	76.5201%
	Wrapper	0.765	0.764	0.764	0.838	76.4149%
Bagging Logistic	All Variables	0.768	0.768	0.768	0.843	76.7782%
	Univariate	0.755	0.754	0.754	0.829	75.4302%
	Multivariate	0.766	0.765	0.765	0.840	76.5488%
	Wrapper	0.766	0.765	0.765	0.838	76.5201%
Boosting Logistic	All Variables	0.768	0.768	0.768	0.780	76.826%
	Univariate	0.755	0.754	0.754	0.771	75.4302%
	Multivariate	0.765	0.765	0.765	0.778	76.5201%
	Wrapper	0.747	0.747	0.747	0.763	74.675%

Table 3: Logistic model results.

2. TAN Model

We can appreciate from table 3 that with these classifiers the multivariate filter feature subset selection had a great impact in the performance, leading to an improvement over the rest of the experiments. Moreover, the model obtained from bagging and boosting has the same results than the base one overall.

Model	Feature Selection	Precision	Recall	F-Measure	ROC Area	Accuracy
TAN	All Variables	0.759	0.759	0.759	0.834	75.8987%
	Univariate	0.755	0.754	0.754	0.827	75.4015%
	Multivariate	0.764	0.764	0.764	0.838	76.2906%
	Wrapper	0.764	0.763	0.762	0.837	76.2524%
Bagging TAN	All Variables	0.759	0.759	0.758	0.835	75.8604%
	Univariate	0.755	0.754	0.754	0.827	75.4207%
	Multivariate	0.764	0.763	0.763	0.838	76.2906%
	Wrapper	0.758	0.757	0.757	0.834	75.7361%
Boosting TAN	All Variables	0.759	0.759	0.759	0.822	75.8987%
	Univariate	0.755	0.754	0.754	0.812	75.4015%
	Multivariate	0.764	0.763	0.763	0.822	76.2906%
	Wrapper	0.764	0.763	0.762	0.820	76.2524%

Table 3: TAN model results.

3. LDA Model

On the linear discriminant analysis it is quite similar to the logistic one, where feature selection did not improve the results. And also the use of feature selection had no impact.

Model	Feature Selection	Precision	Recall	F-Measure	ROC Area	Accuracy
LDA	All Variables	0.767	0.767	0.767	0.842	76.7017%
	Univariate	0.754	0.754	0.754	0.828	75.4111%
	Multivariate	0.766	0.765	0.765	0.840	76.5296%
	Wrapper	0.760	0.759	0.758	0.827	75.8700%
Bagging LDA	All Variables	0.768	0.768	0.767	0.842	76.7591%
	Univariate	0.754	0.754	0.754	0.828	75.4111%
	Multivariate	0.767	0.767	0.767	0.840	76.6635%
	Wrapper	0.761	0.760	0.759	0.828	75.9656%
Boosting LDA	All Variables	0.767	0.767	0.767	0.787	76.7017%
	Univariate	0.754	0.754	0.754	0.772	75.4111%
	Multivariate	0.766	0.765	0.765	0.783	76.5296%
	Wrapper	0.760	0.759	0.758	0.773	75.870%

Table 4: LDA model results.

All data

Feature Selection	Model	Precision	Recall	F-Measure	ROC Area	Accuracy
All Variables	Logistic	0.768	0.768	0.768	0.843	76.826%
	TAN	0.759	0.759	0.759	0.834	75.8987%
	LDA	0.767	0.767	0.767	0.842	76.7017%
	Boosting Logistic	0.768	0.768	0.768	0.780	76.826%
	Boosting TAN	0.759	0.759	0.759	0.822	75.8987%
	Boosting LDA	0.767	0.767	0.767	0.787	76.7017%
	Bagging Logistic	0.768	0.768	0.768	0.843	76.7782%
	Bagging TAN	0.759	0.759	0.758	0.835	75.8604%
	Bagging LDA	0.768	0.768	0.767	0.842	76.7591%
Univariate	Logistic	0.755	0.754	0.754	0.829	75.4302%
	TAN	0.755	0.754	0.754	0.827	75.4015%
	LDA	0.754	0.754	0.754	0.828	75.4111%
	Boosting Logistic	0.755	0.754	0.754	0.771	75.4302%
	Boosting TAN	0.755	0.754	0.754	0.812	75.4015%
	Boosting LDA	0.754	0.754	0.754	0.772	75.4111%
	Bagging Logistic	0.755	0.754	0.754	0.829	75.4302%
	Bagging TAN	0.755	0.754	0.754	0.827	75.4207%
	Bagging LDA	0.754	0.754	0.754	0.828	75.4111%
Multivariate	Logistic	0.765	0.765	0.765	0.840	76.5201%
	TAN	0.764	0.764	0.764	0.838	76.2906%
	LDA	0.766	0.765	0.765	0.840	76.5296%
	Boosting Logistic	0.765	0.765	0.765	0.778	76.5201%
	Boosting TAN	0.764	0.763	0.763	0.822	76.2906%
	Boosting LDA	0.766	0.765	0.765	0.783	76.5296%
	Bagging Logistic	0.766	0.765	0.765	0.840	76.5488%
	Bagging TAN	0.764	0.763	0.763	0.838	76.2906%
	Bagging LDA	0.767	0.767	0.767	0.840	76.6635%
Wrapper	Logistic	0.765	0.764	0.764	0.838	76.4149%
	TAN	0.764	0.763	0.762	0.837	76.2524%
	LDA	0.760	0.759	0.758	0.827	75.8700%
	Boosting Logistic	0.747	0.747	0.747	0.763	74.675%
	Boosting TAN	0.764	0.763	0.762	0.820	76.2524%
	Boosting LDA	0.760	0.759	0.758	0.773	75.870%
	Bagging Logistic	0.766	0.765	0.765	0.838	76.5201%
	Bagging TAN	0.758	0.757	0.757	0.834	75.7361%
	Bagging LDA	0.761	0.760	0.759	0.828	75.9656%

Table 6: All results.

4. Conclusion

Overall, the models exhibit relatively similar performance across the evaluation metrics. The Precision, Recall, F-Measure, and ROC Area values are consistently close for the different models and feature selection techniques. The accuracy values range between approximately 74.68% and 76.82%. The highest accuracy is achieved by the models using All Variables, indicating that including all available features leads to better overall accuracy. This might be because most of the variables are relevant and contain useful information for predicting the target variable. Boosting algorithm results have less ROC area, this might be due to overfitting. Adaboosting has the potential to overfit the training data if the classifiers become too complex or if the algorithm iterates for too long.

The best models obtained were using Logistic model all variables and meta classifiers with Logistic. We can conclude that the best model best suited for this problem is Logistic regression.

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. Logistic Regression.

[3] Pedro Larrañaga, Concha Bielza. Bayesian classifiers discrete.

[4] Pedro Larrañaga, Concha Bielza. Discriminant analysis.