Non-probabilistic supervised classification

Machine Learning

Pablo Bande Sánchez Girón

Introduction

Algorithms:

1. Artificial Neural Network
2. k-Nearest Neighbors
3. Rule Induction
4. Support Vector Machines
5. Classification Trees

Evaluation:

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

Feature selection:

1. All variables
2. Univariate filter
3. Multivariate filter
4. Wrapper

Goal: Predict Heart disease

Dataset

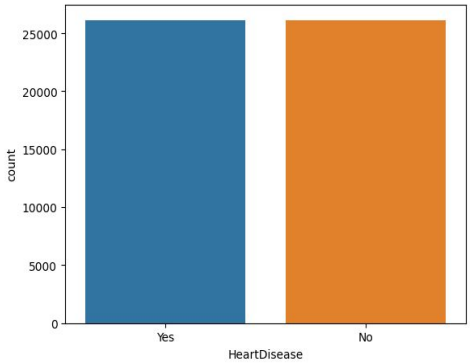
Centers for Disease Control and Prevention. Annual telephone surveys to collect health data of U.S. residents.
319,795 entries and 18 columns.

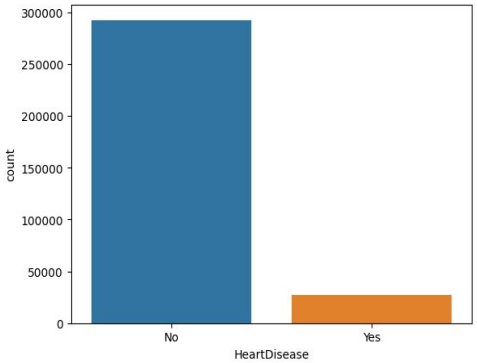
Preprocessing:

- Numerical to categorical using Label encoding and One-hot encoding
- Removed outliers
- Normalized

Due to the nature of the dataset, it was unbalanced so i balanced and still had enough data to work with

Result:
52,296 entries and 24 columns
Split into training and test sets 80 / 20 %





Feature selection

- All variables
- Univariate filter: Gain ratio evaluation and a ranker with a threshold of 0.05 p-value.
- Multivariate filter: Correlation-based Feature Selection and GreedyStepwise.
- Wrapper: ClassifierSubsetEval with a GreedyStepwise algorithm. Subset of the data of 5000 entries.

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							

Results: KNN

Two classes: K an odd number and set as sqrt(N)/2.

- K = 103
- K = 51

Filters improved results but not 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%

Results: RIPPER

```
(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
```

- Higher values of GenHealth are associated with lower likelihood of heart disease.
- Younger individuals are generally less likely to have heart disease.
- Absence of a previous stroke is indication of lower risk for heart disease.
- Other variables have less influence on the prediction of heart disease.

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%

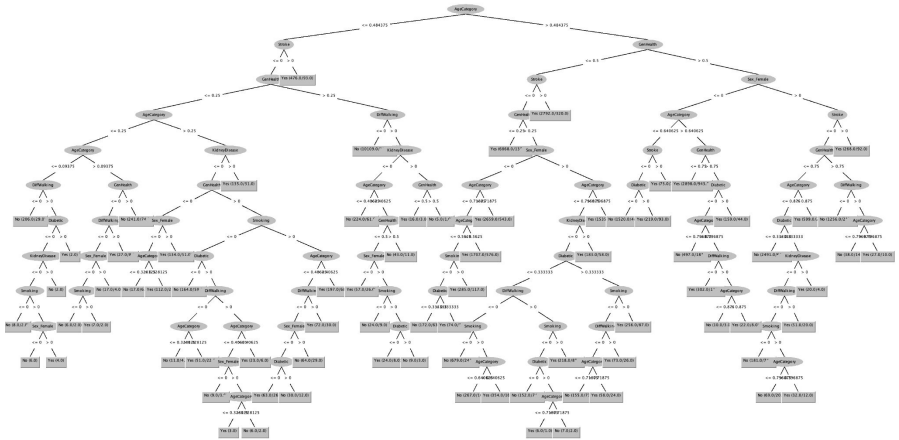
Results: ANN and SVM

One hidden layer multilayer 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%

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%

Results: J48



- AgeCategory: first to split the data. Multiple times in different branches of the tree, significant impact.
- GenHealth: Used in multiple splits, significant impact.
- Stroke, DiffWalking, KidneyDisease, Diabetic: Few splits in the tree. Provides additional information.

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%

Conclusion

- Similar performance across different feature selection techniques.
- The impact of feature selection techniques on model performance is not significant. Most of the variables are relevant.
- The J48 model generally lower results.
- Overall best one is SVM using all attributes.

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 %

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.