Probabilistic supervised classification

Machine Learning

Pablo Bande Sánchez Girón

Introduction

Algorithms:

1. Logistic Regression
2. Tree Augmented Naive Bayes
3. Linear Discriminant Analysis
4. AdaBoost
5. Bagging

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

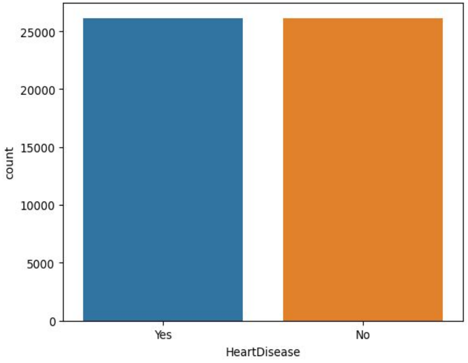
Centers for Disease Control and Prevention. Annual telephone surveys to collect health data of U.S. residents.
319,795 entries and 18 columns.

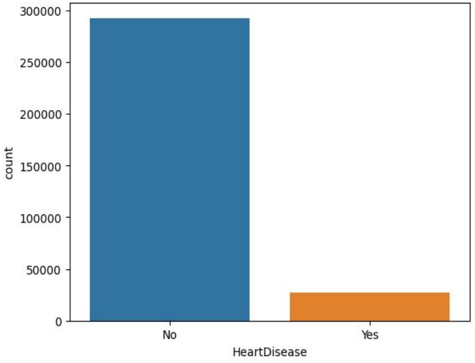
Preprocessing:

- 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 %





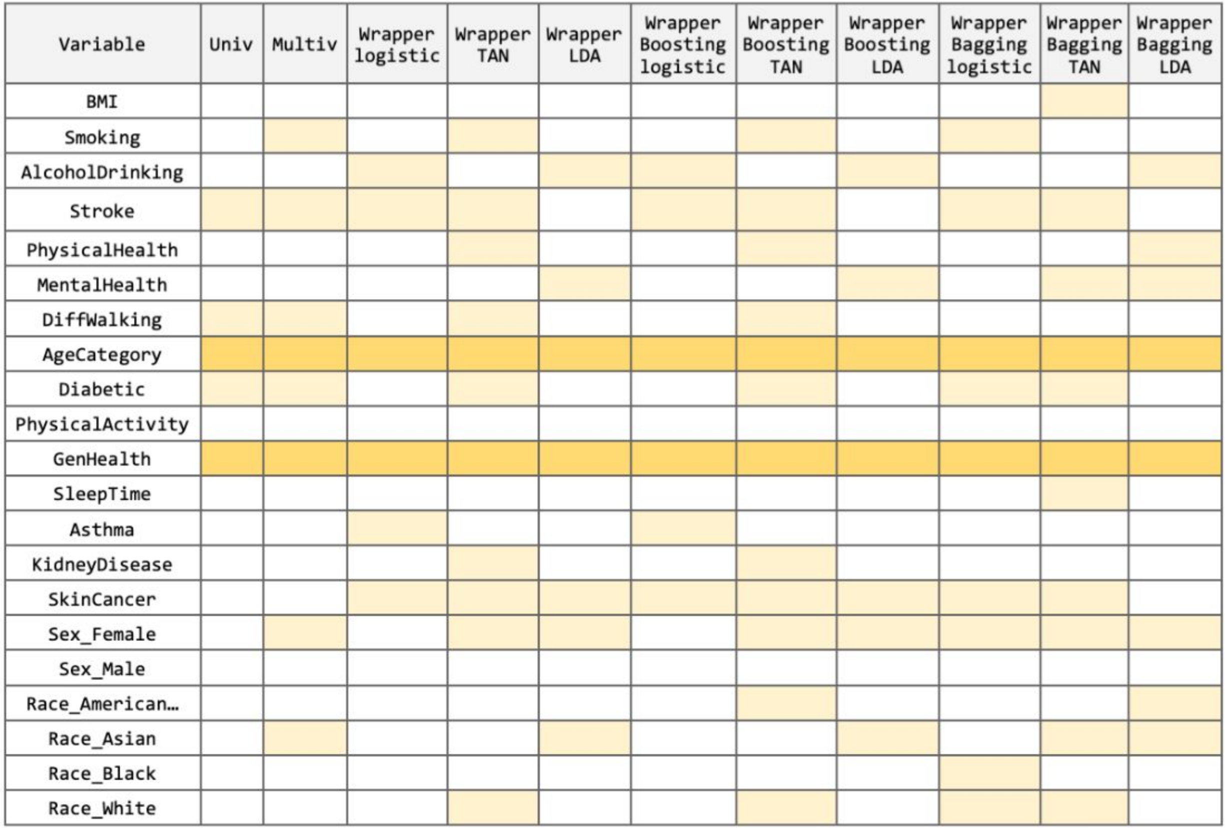
Feature selection

- 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.

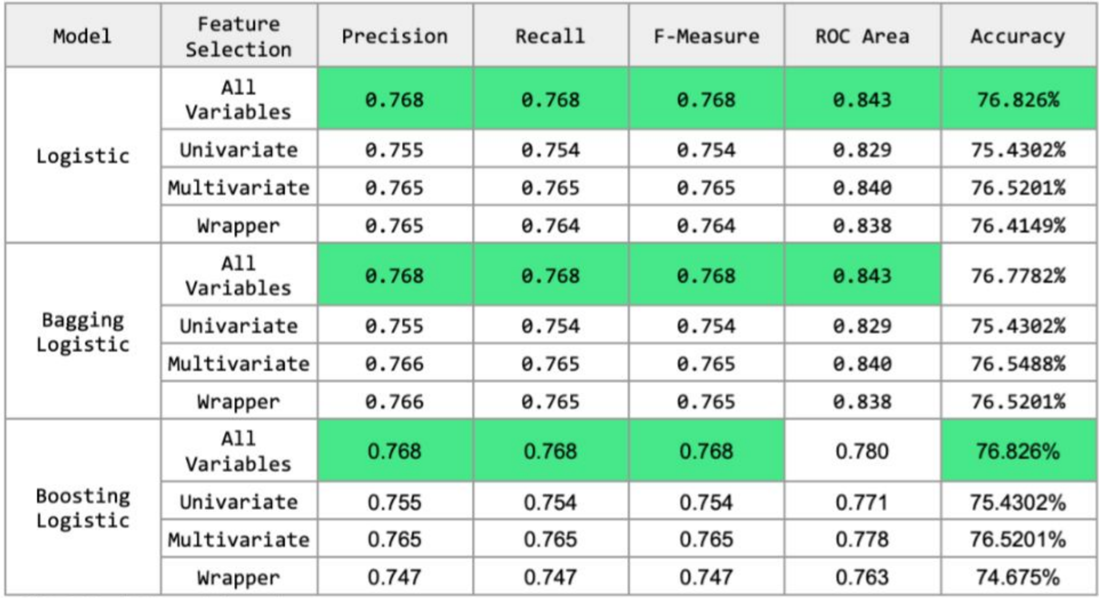


Variable	Univ	Multiv	Wrapper logistic	Wrapper TAN	Wrapper LDA	Wrapper Boosting logistic	Wrapper Boosting TAN	Wrapper Boosting LDA	Wrapper Bagging logistic	Wrapper Bagging TAN	Wrapper Bagging LDA
BMI										●	
Smoking		●		●			●		●		
AlcoholDrinking			●		●	●		●			●
Stroke	●	●	●	●		●	●		●	●	
PhysicalHealth				●			●				●
MentalHealth					●			●		●	●
DiffWalking	●	●		●			●				
AgeCategory	●	●	●	●	●	●	●	●	●	●	●
Diabetic	●	●		●			●		●	●	
PhysicalActivity											
GenHealth	●	●	●	●	●	●	●	●	●	●	●
SleepTime										●	
Asthma			●			●					
KidneyDisease				●			●				
SkinCancer			●	●	●	●	●	●	●	●	
Sex_Female		●		●	●		●	●	●	●	●
Sex_Male											
Race_American…							●				●
Race_Asian		●			●			●		●	●
Race_Black									●		
Race_White				●			●		●	●	

Results: Logistic

- Feature filters did not improve performance.

- Boosting logistic approach exhibits slightly lower performance in terms of correctly identifying positive instances, discriminatory power, and overall prediction correctness.

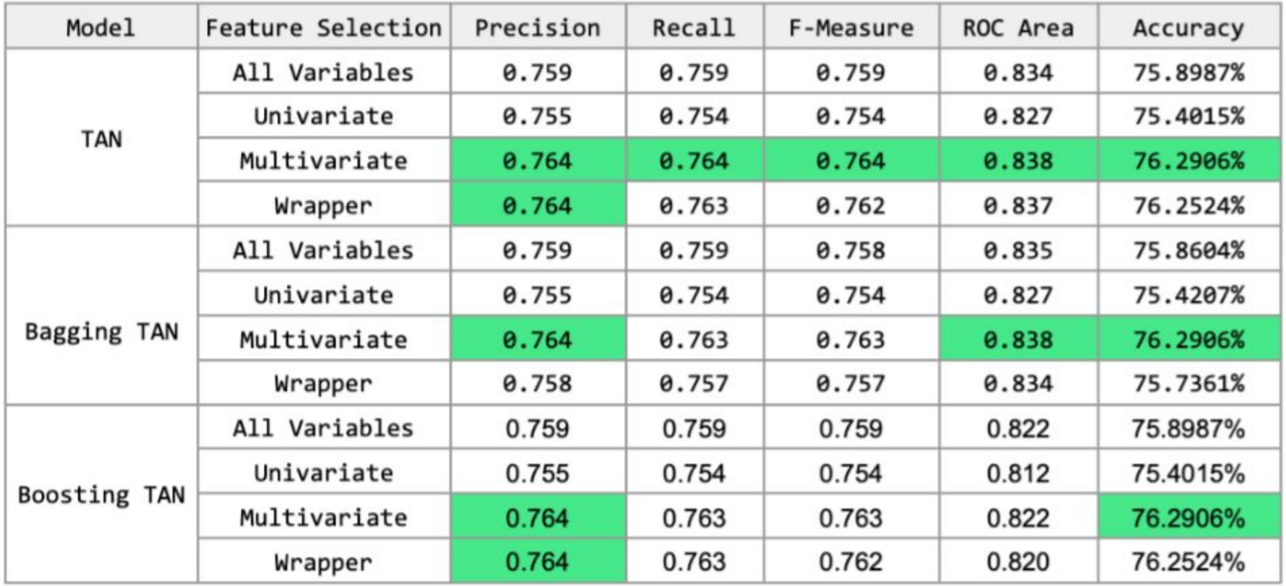


Model	Feature Selection	Precision	Recall	F-Measure	ROC Area	Accuracy
Logistic	All Variables	0.768	0.768	0.768	0.843	76.826%
	Univariate	0.755	0.754	0.754	0.829	75.4302%
	Multivariate	0.765	0.765	0.765	0.840	76.5201%
	Wrapper	0.765	0.764	0.764	0.838	76.4149%
Bagging Logistic	All Variables	0.768	0.768	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%

Results: TAN

- Multivariate filter feature subset selection had a great impact in the performance, leading to an improvement over the rest of the experiments.

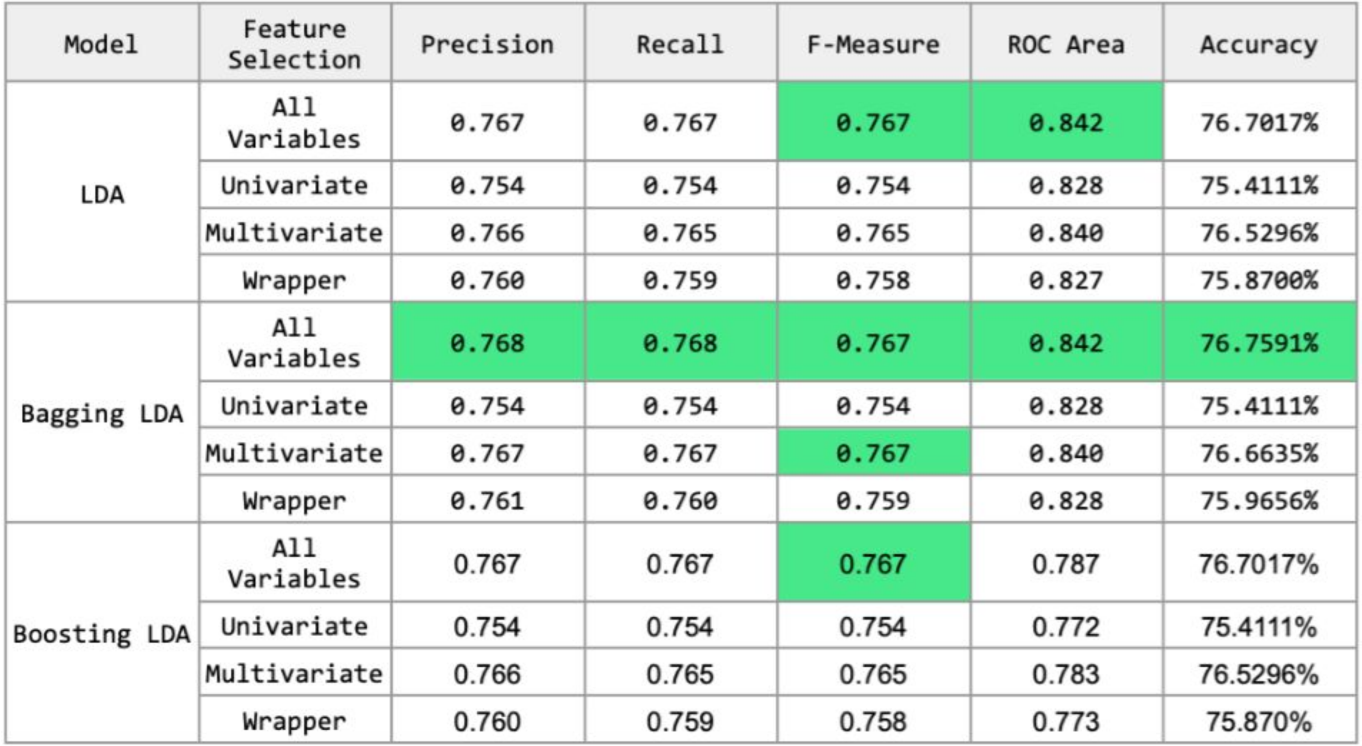
- The model obtained from bagging and boosting has the same results than the base one overall.



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%

Results: LDA

- Similar to the logistic model
- Feature selection had no impact
- Meta Classifiers barely affected results
- Same results overall



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%

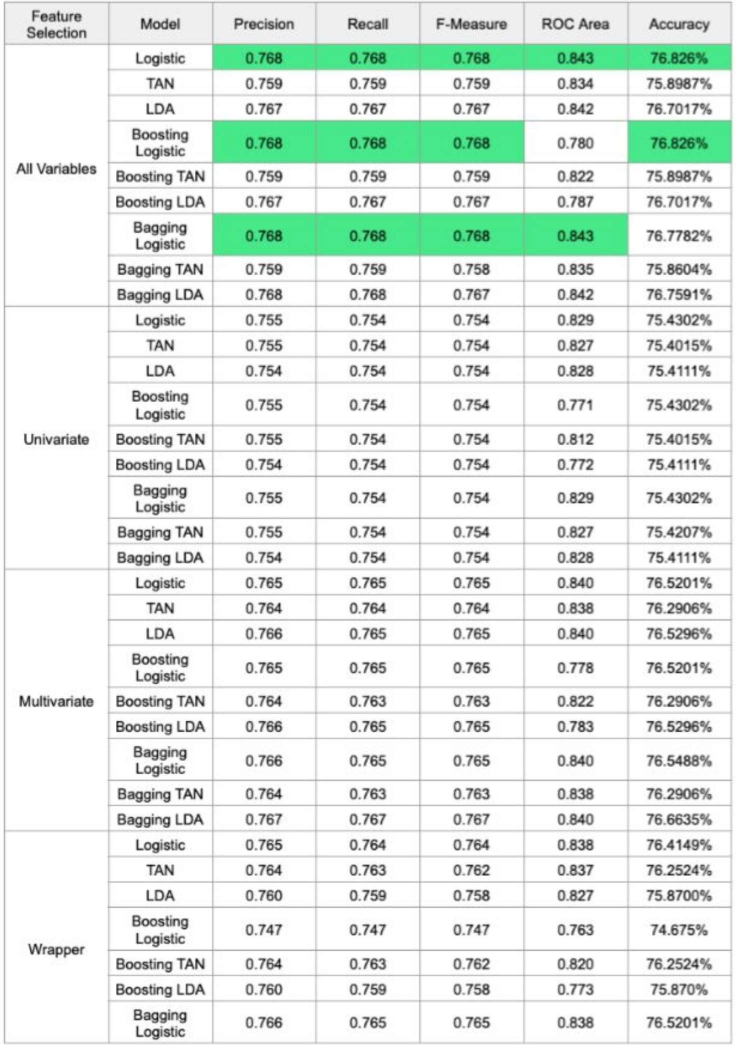
Conclusion

- Similar performance across different feature selection techniques and models.

- Highest accuracy is achieved by the models using all variables.

- Boosting algorithm results have less ROC area, might be due to overfitting.

- Overall best models are Logistic with all attributes.



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%

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.