A comparison of Naïve Bayes (NB) and Random Forest (RF) on predicting heart disease

Andra Condurache
200030801

Description and motivation of the problem

The aim of this study is to identify the pacients that might suffer from a heart disease taking into consideration their health condition. There are plenty of features that can lead to a heart disease and there are many types of heart disease that can lead to serious consequences or even death. Two supervised learning algorithms, Naïve Bayes (NB) and Random Forest (RB), are going to be used to solve a binary classification problem. These models will predict patients' health status and we will compare the results.

Data Set and Initial analysis

♦Dataset: Heart Disease UCI (1)
♦The data consists of 4 databases: Cleveland, Hungary, Switzerland and VA Long Beach, focusing on Cleveland database.
♦The dataset contains 303 instances and 75 attributes. We will be using a subset of 14 attributes for this study. Just a few values are null or missing but this should not be an issue for the models.
♦Personal information regarding name and social security numbers was removed.
♦The target value indicates the health condition. This can vary from 0 to 4 but we are interested only in finding the presence or not of the disease. The target value contains 1 if the patient suffers from a heart disease and 0 otherwise.
♦The histogram (Fig 1) shows us there is a small imbalance (138 VS 165, 45.45% VS 54.45%) between the classes but this should not cause a problem.
♦The correlation matrix can conveniently summarizes a dataset. The correlation heatmap (Fig 2) shows us 'exercise induced angina', 'ST depression induced by exercise relative to test', 'number of major vessels', 'resting blood pressure' and 'thalassemia' have a strong influence on the target value. There is a weak correlation between most of features and the target value.
♦The data was grouped by the target value to emphasize the difference between the range of the values among people suffering from heart disease or not. The boxplots (Fig 3) illustrate an extremely high variety between them.

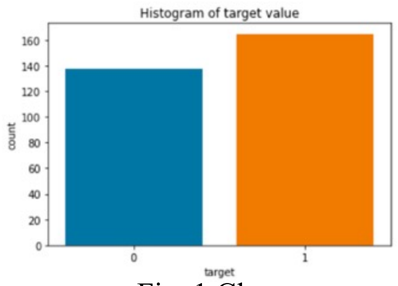

Fig. 1 Classes

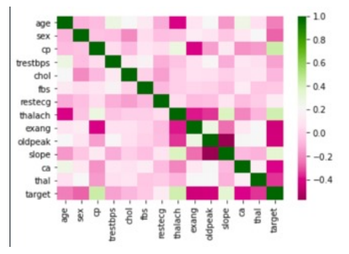

Fig. 2 Correlation

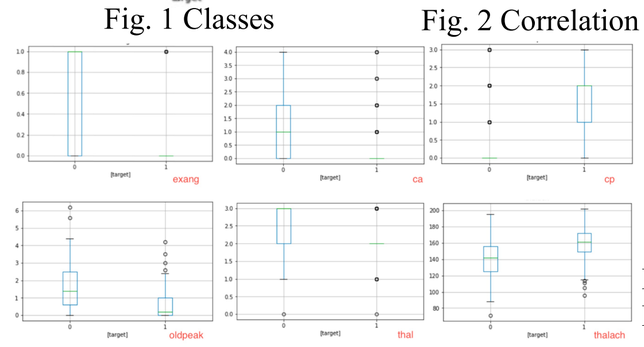

Fig. 3 Boxplots

Naïve Bayes

♦Naïve Bayes is a supervised learning algorithm that solves binary and multi-class classification problems. In our case, a binary classification problem, the existence of a heart condition or not.
♦This implementation is based on Bayes' theorem, using the joint probability of the input X to make predictions regarding Y.
♦It uses prior probability and conditional probability and it can also use Maximum A Posteriori (MAP) to estimate P(Y)
♦It is naïve because it is very simplified assuming there is no correlation between the features, all of them being independent.

PROS

♦Very practical, easy to understand and implement
♦Low running time compared to highly sophisticated methods
♦Useful for large datasets, but it also works well for small amount of data
♦Usually high accuracy

CONS

♦It makes strong assumptions that are not valid in real life

Random Forest

♦Random Forest is a supervised learning algorithm based on ensemble of a large number of individual decision trees. It increases accuracy combining decision trees with flexibility.
♦It uses bagging concept which implies bootstrapping the data and aggregating it. Some of the data in the original dataset will not be included in the bootstrapped dataset.
♦It can measure the accuracy by classifying the out-of-bag samples.

PROS

♦The variety makes random forest more efficient than individual decision trees
♦It handles missing data
♦It solves the issues of overfitting
♦Efficient on large datasets

CONS

♦Not suitable for sparse data
♦It allows duplicates included in the bootstrapped dataset
♦Computationally expensive

Hypothesis Statement

♦RF should achieve better performance than NB, but these two methods should be competitive. RF performs better on average it can vary depending on problems and metrics (2) and AUC values slightly varies (3).
♦The accuracy should be a good measure because the dataset is balanced.
♦RF usually has higher running time than NB because we generate a lot of individual decision trees.

Description of the choice of training and evaluation methodology

♦The dataset will be split 70% for training and 30% for the testing.
♦10 fold cross validation will be used.
♦ Every model will be optimised by tuning the hyperparameters and choosing the best values.
♦The algorithms will be evaluated using accuracy, recall, precision, F1 score and AUC

Choice of Parameters and experimental results

Naïve Bayes

The optimisation focused on finding a suitable distribution and width size. High Width values (mostly greater than 1) performed extremely poor and implicit kernel distribution had poor results.Estimated objective function value=0.18152,Function evaluation time = 0.2208

Random Forest

The model was trained using fitcensemble and it was optimised using OptimizeHyperparameters. The methods were GentleBoost, LogitBoost, RUSBoost. High MinLeafSize and very high NumLearningRateCycles did not perfom well. MaxNumSplits and MinParentSize did not improve the model. NumLearningCycle, LearnRate and MinLeafSize increased the performance significally. The best model was: Method: AdaBoostM1, NumLearningRateCycles:181, LearnRate:0.084653, MinLeafSize: 42, Estimated objective function value = 0.18154, Function evaluation time = 2.1578

Analysis and critical evaluation of results

♦Random Forest is definitely a more flexible method than Naïve Bayes and it is more responsive when tuning the parameters.
♦The next table (Table 1) shows us that NB and RF are highly competitive (Fig 4 and 5).
♦The results obtained using RF vary a lot and the optimization increased substantially the performance.
♦The accuracy proved to be a suitable method to measure the performance as the dataset is balanced.
♦For NB, the total elapsed time was 26.0235 seconds for 30 function evaluations. For RF, the total elapsed time was 50.1006 seconds for 30 function evaluations. As we expected, RF has higher running time than NB.
♦Both of the baseline models had decent results. Tuning the hyperparameters improved a lot the performance for the RF model, while the NB model was not optimized at all. The NB model accuracy decreased and this might be caused by overfitting. The accuracy of the baseline model was 0.83. RF solved the issue of overfitting as we expected.
♦The misclassification rate decreases from 0.25 to 0.15 after 100 Learning Cycles but it increases slightly to 0.16 as more weak learners enter the ensemble. Setting the learning rate to 0.084653 played a huge role during the optimisation. (Fig 6)
♦The algorithms have excellent AUC values, especially RF. Furthermore, both of the algorithms have similar ROC curves.

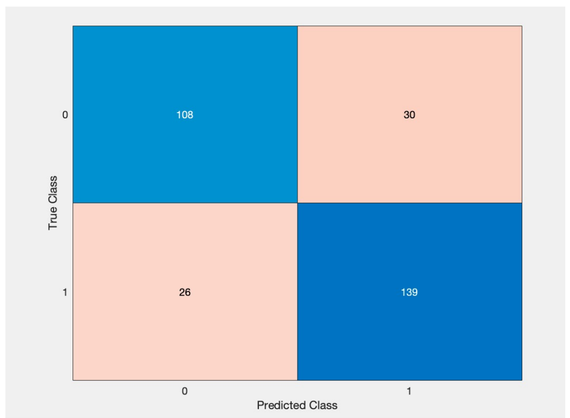

Fig. 4 - NB Confusion matrix

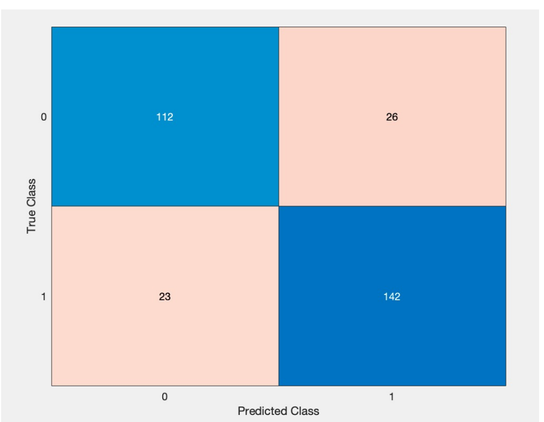

Fig. 5 - RF Confusion matrix

	NB	RF
Accuracy	0.8152	0.8383
Precision	0.8424	0.8606
Recall	0.8225	0.8452
F1 Score	0.8323	0.8529
EstGen Error	0.1848	0.1617
AUC	0.8928	0.9022

Table 1 - Results

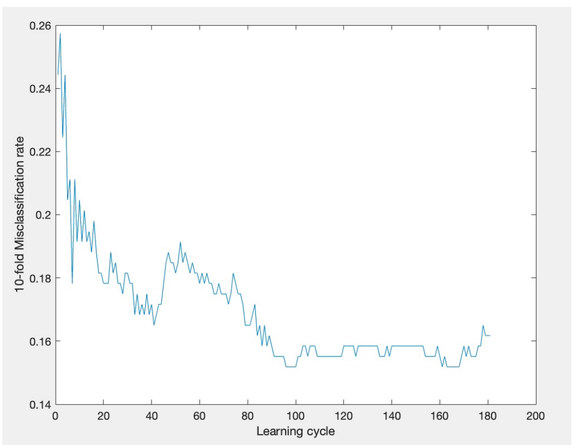

Fig. 6 – RF Generalization Error

Lessons learned and Future work

♦As this is a study related to a medical domain, focusing on True Positive values is crucial. Good results can lead to earlier diagnosis and treatment, and obviously higher recovery/survival rate.
♦The accuracy is not the only way we can evaluate the results. There are many more ways, it depends on the focus of the study.
♦It takes more time to tune the hyperparameters for RF and it is more difficult to implement but it can lead to better results.
♦Feature selection may be a solution to improve the performance
♦EDA and feature engineering could improve substantially the performance

References

(1) Heart Disease Data Set UCI, https://archive.ics.uci.edu/ml/datasets/heart+disease
(2) Caruana, Rich & Niculescu-Mizil, Alexandru. (2006). An Empirical Comparison of Supervised Learning Algorithms. Proceedings of the 23rd international conference on Machine learning - ICML '06. 2006. 161-168. 10.1145/1143844.1143865.
(3) E.M.M. van der Heide, R.F. Veerkamp, M.L. van Pelt, C. Kamphuis, I. Athanasiadis, B.J. Ducro,Comparing regression, naive Bayes, and random forest methods in the prediction of individual survival to second lactation in Holstein cattle, Journal of Dairy Science,Volume 102, Issue 10,2019,Pages 9409-9421