JS_HeartFailureDiagnosis

Ioannis Skiadas

2023-12-02

Contents

OVERVIEW

This classification tackling project aims to assess various machine learning algorithms in predicting Heart Failure, a detrimental event to mostly middle aged and older adults. The candidate models will be trained and then tested in a blinded manner, having not met the evaluation data set before. This project aims to partially fulfill the requirements for the Online Harvard Data Science program.

INTRODUCTION

Cardiovascular Diseases (CVD) present a growing risk worldwide. Overall, CVDs list among the top mortality reasons globaly, heart failue being a common incidence among them (Davide Chicco 2020). Different machine learning techniques have been recently tried to tackle the disease prediction problem. In this work different supervised learning algorithms will be utilized with the aim to get an accuracy of over 0.85. The following models will be evaluated: *knn, Logistic Regression, XgBoost, Neural Nets, and Random Forests*

Data Set

The data set comprises eleven features of patients, with or without a Heart Failure condition. It is publicly available from kaddle (fedesoriano 2021). The aim of the project is to assess different algorithms to accurately classify the morbidity. Such systems may be instrumental to the prevention of severe disease. Thus the Heart Failure feature will be the dependent variable A summary of the set's features is depicted below:

Table 1: Heart Failure Dataset Attributes

index	Attributes
1	Age: age of the patient [years]
2	Sex: sex of the patient [M: Male, F: Female]
3	ChestPainType: chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]
4	RestingBP: resting blood pressure [mm Hg]
5	Cholesterol: serum cholesterol [mm/dl]
6	FastingBS: fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]
7	RestingECG: resting electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria]
8	MaxHR: maximum heart rate achieved [Numeric value between 60 and 202]
9	ExerciseAngina: exercise-induced angina [Y: Yes, N: No]
10	Oldpeak: oldpeak = ST [Numeric value measured in depression]
11	ST_Slope: the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]
12	HeartDisease: output class [1: heart disease, 0: Normal]

METHODS/ANALYIS

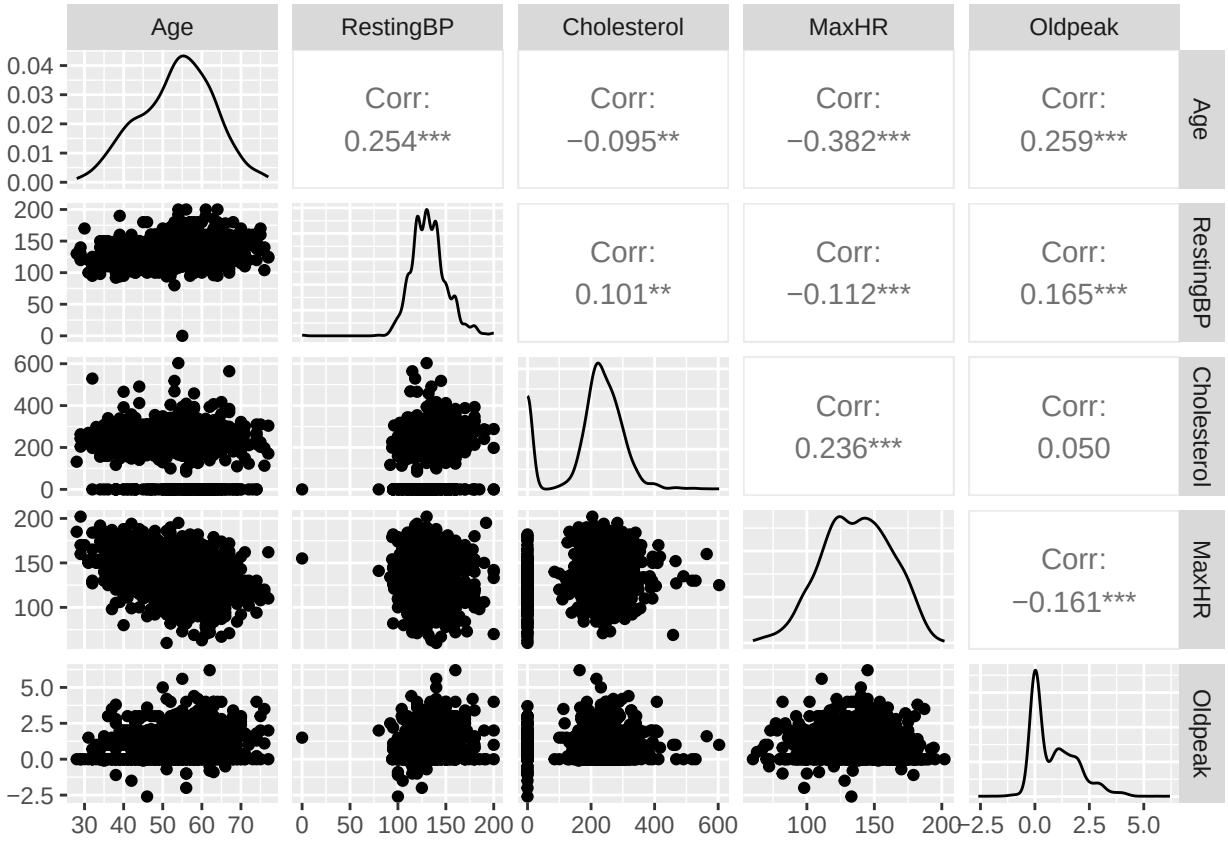
Software and Packages

The analysis will be carried out in R version 4.1.0 (2021-05-18) through the popular IDE 2022.07.2+576 . The needed packages are uploaded:

Load the data set

```
##       Age            Sex            ChestPainType        RestingBP
##  Min.   :28.00   Length:918         Length:918         Min.   :  0.0
##  1st Qu.:47.00   Class :character   Class :character   1st Qu.:120.0
##  Median :54.00   Mode  :character   Mode  :character   Median :130.0
##  Mean   :53.51                                         Mean   :132.4
##  3rd Qu.:60.00                                         3rd Qu.:140.0
##  Max.   :77.00                                         Max.   :200.0
##   Cholesterol      FastingBS       RestingECG           MaxHR
```

```
##  Min.   :  0.0   Min.   :0.0000   Length:918         Min.   : 60.0
##  1st Qu.:173.2   1st Qu.:0.0000   Class :character   1st Qu.:120.0
##  Median :223.0   Median :0.0000   Mode  :character   Median :138.0
##  Mean   :198.8   Mean   :0.2331                      Mean   :136.8
##  3rd Qu.:267.0   3rd Qu.:0.0000                      3rd Qu.:156.0
##  Max.   :603.0   Max.   :1.0000                      Max.   :202.0
##  ExerciseAngina       Oldpeak          ST_Slope          HeartDisease
##  Length:918         Min.   :-2.6000   Length:918         Min.   :0.0000
##  Class :character   1st Qu.: 0.0000   Class :character   1st Qu.:0.0000
##  Mode  :character   Median : 0.6000   Mode  :character   Median :1.0000
##                     Mean   : 0.8874                      Mean   :0.5534
##                     3rd Qu.: 1.5000                      3rd Qu.:1.0000
##                     Max.   : 6.2000                      Max.   :1.0000
```



Correlations among the numeric variables are week. Thus it is not likely that they will be multicollinearity issues in our models.

Converting the character features to categorical

```
##       Age        Sex     ChestPainType   RestingBP      Cholesterol
##  Min.   :28.00   F:193   ASY:496       Min.   :  0.0   Min.   :  0.0
##  1st Qu.:47.00   M:725   ATA:173       1st Qu.:120.0   1st Qu.:173.2
##  Median :54.00           NAP:203       Median :130.0   Median :223.0
##  Mean   :53.51           TA : 46       Mean   :132.4   Mean   :198.8
##  3rd Qu.:60.00                         3rd Qu.:140.0   3rd Qu.:267.0
##  Max.   :77.00                         Max.   :200.0   Max.   :603.0
```

```
##    FastingBS       RestingECG      MaxHR       ExerciseAngina    Oldpeak
##  Min.   :0.0000   LVH   :188   Min.   : 60.0   N:547          Min.   :-2.6000
##  1st Qu.:0.0000   Normal:552   1st Qu.:120.0   Y:371          1st Qu.: 0.0000
##  Median :0.0000   ST    :178   Median :138.0                  Median : 0.6000
##  Mean   :0.2331                Mean   :136.8                  Mean   : 0.8874
##  3rd Qu.:0.0000                3rd Qu.:156.0                  3rd Qu.: 1.5000
##  Max.   :1.0000                Max.   :202.0                  Max.   : 6.2000
##  ST_Slope    HeartDisease
##  Down: 63   Min.   :0.0000
##  Flat:460   1st Qu.:0.0000
##  Up  :395   Median :1.0000
##             Mean   :0.5534
##             3rd Qu.:1.0000
##             Max.   :1.0000
```

create the train and test subsets

Analysis

Knn

The knn algorithm is trained using cross validation and a series of different neighboring areas.

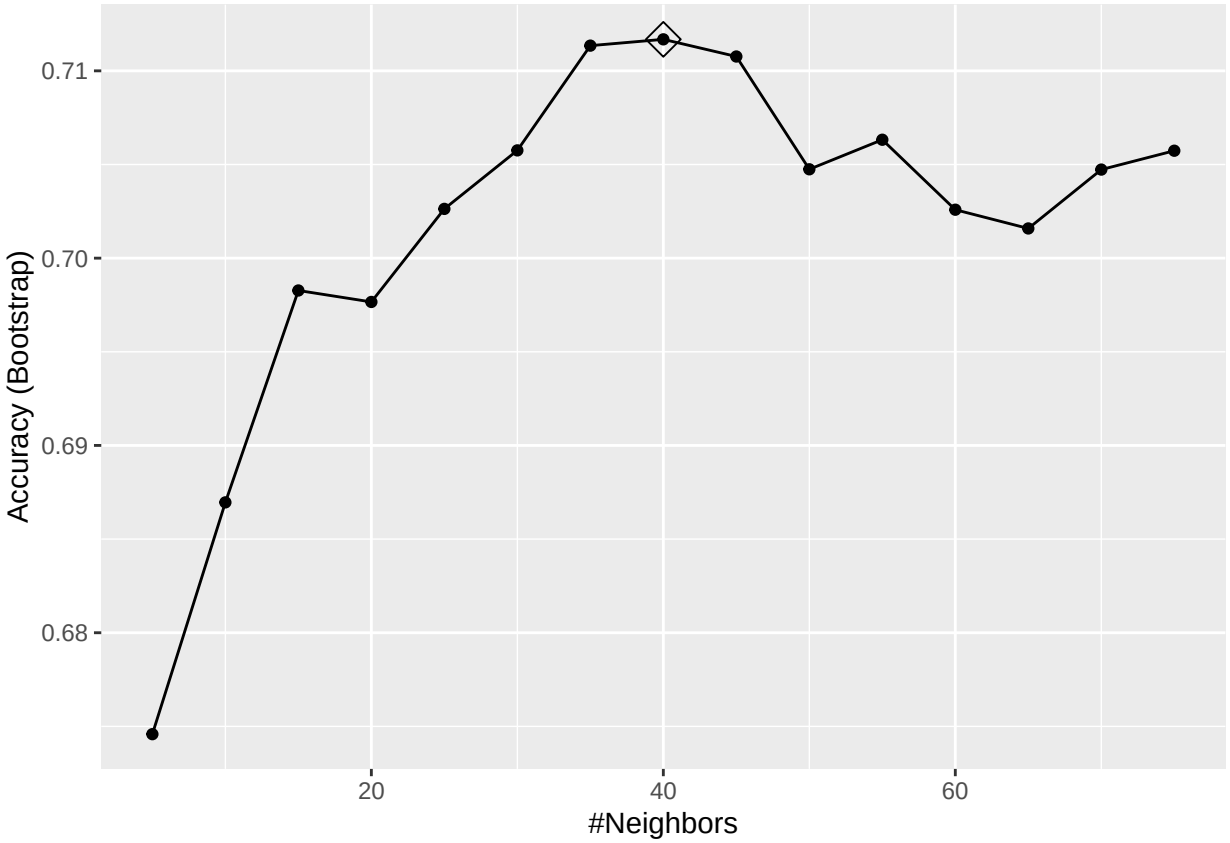


Table 2: Model Accuracy

Model	Accuracy
knn	0.6956522

The most efficient neighboring parameter being 40 .

Logistic Regression

Table 3: Model Accuracy

Model	Accuracy
knn	0.6956522
Logistic Regression	0.8608696

Using the default assumption of a *Gaussian* distribution of errors and the *reweighted least squares, (IWLS)* method for error loss minimization, the model has a robust performance.

eXtreme Gradient Boosting

extreme Gradient Boosting ensembles have often been used for classification problems successfully As a note the training set comprises 688 whereas the testing set, 230.

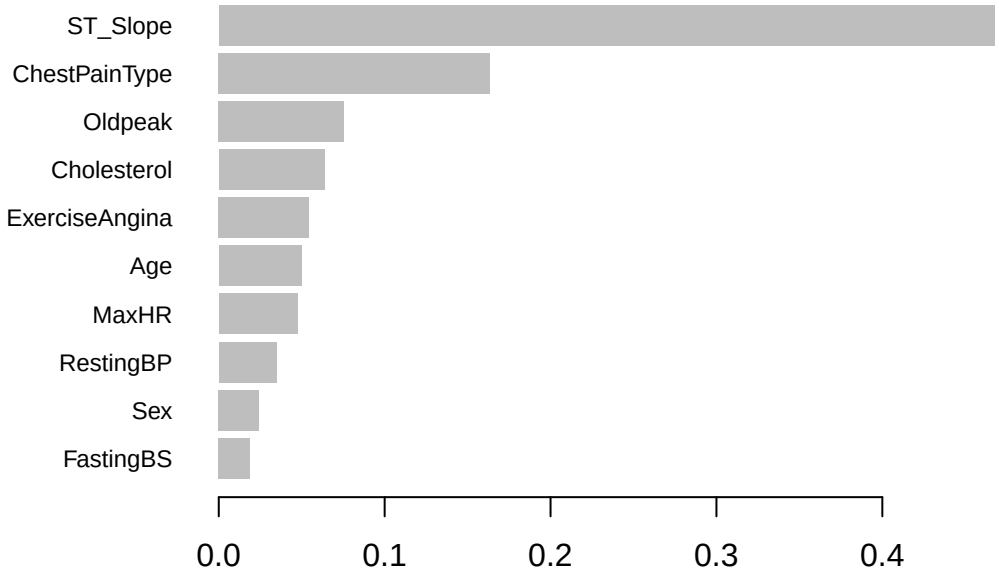


Table 4: Model Accuracy

Model	Accuracy
knn	0.6956522
Logistic Regression	0.8608696
xgb	0.8869565

Electrocardiogram (ECG) findings such as *ST_slope* and the *Oldpeak* parameter seem to be importance.

Neural Nets

Default Parameters

```
## # weights:  86
## initial  value 451.681055
## final  value 431.748718
## converged
```

Table 5: Model Accuracy

Model	Accuracy
knn	0.6956522
Logistic Regression	0.8608696
xgb	0.8869565
nnet default	0.6304348

Parameterized training a grid of node size and weigh decay for regularization and avoiding overfitting

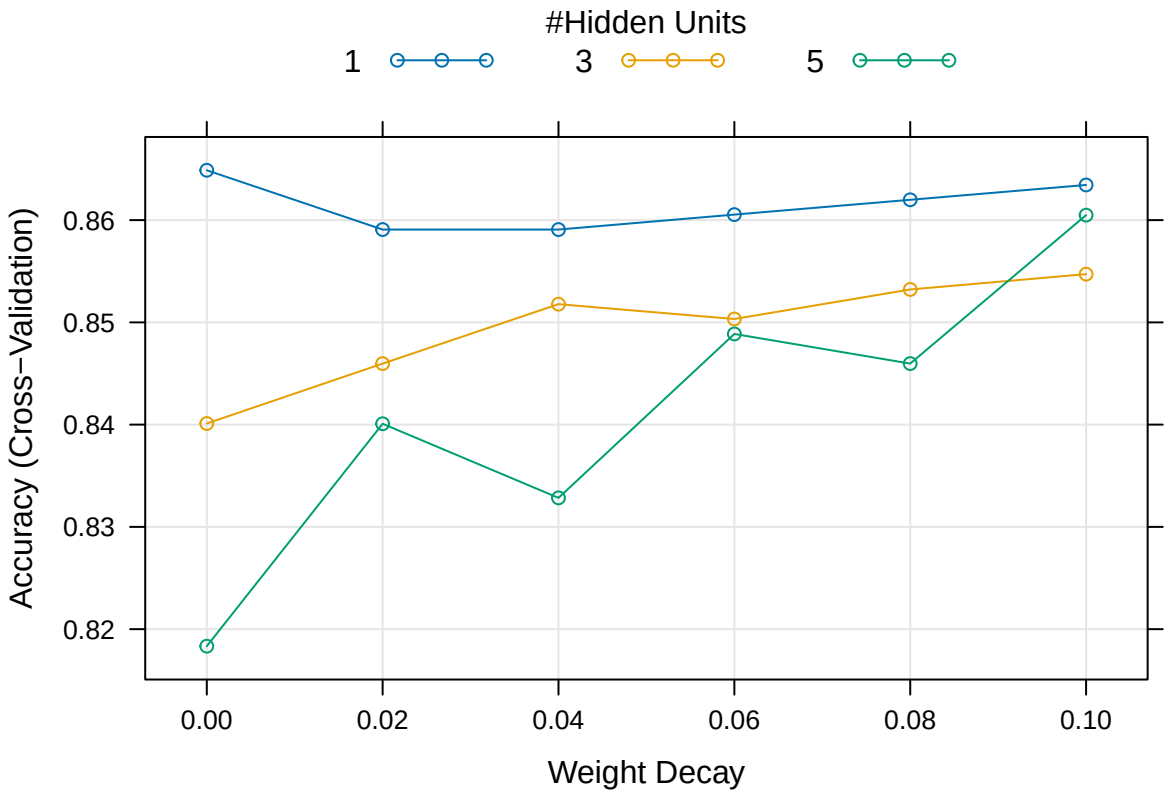


Table 6: Model Accuracy

Model	Accuracy
knn	0.6956522
Logistic Regression	0.8608696
xgb	0.8869565
nnet default	0.6304348
nnet	0.8608696

Table 7: overfitting check

Overall	Train	Test
Accuracy	0.8706395	0.8608696
Kappa	0.7344942	0.7214653
AccuracyLower	0.8432395	0.8093012
AccuracyUpper	0.8948101	0.9028494
AccuracyNull	0.5697674	0.5043478
AccuracyPValue	0.0000000	0.0000000
McnemarPValue	0.1378097	0.0518299

Table 8: NN Variable Importance

	Overall
Age	22.646814
SexM	34.676255
ChestPainTypeATA	60.104691
ChestPainTypeNAP	50.104760
ChestPainTypeTA	59.089372
RestingBP	9.287470
Cholesterol	74.836038
FastingBS	32.829059
RestingECGNormal	3.225281
RestingECGST	0.000000
MaxHR	20.059328
ExerciseAnginaY	29.364390
Oldpeak	100.000000
ST_SlopeFlat	50.488425
ST_SlopeUp	26.083789

The model does not seem to overfit as the proximity of the *Kappa* values reveal in *Table 7*. ALso as the *Figure* shows, accuracy seems to increase with more nodes and higher regularization. Again ECG parameters are of importance, although *Cholesterol* and *ChestPain* are also noteworthy.

Random Forests

Default

Table 9: Model Accuracy

Model	Accuracy
knn	0.6956522
Logistic Regression	0.8608696
xgb	0.8869565
nnet default	0.6304348
nnet	0.8608696
RF_test	0.8826087

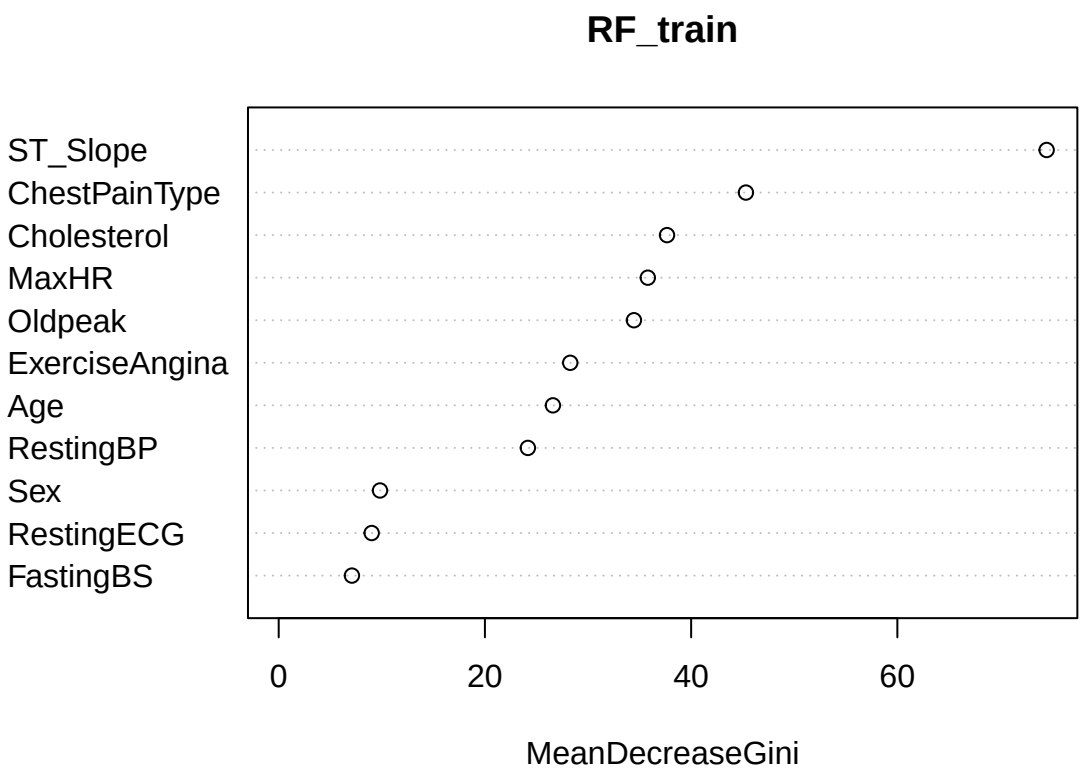
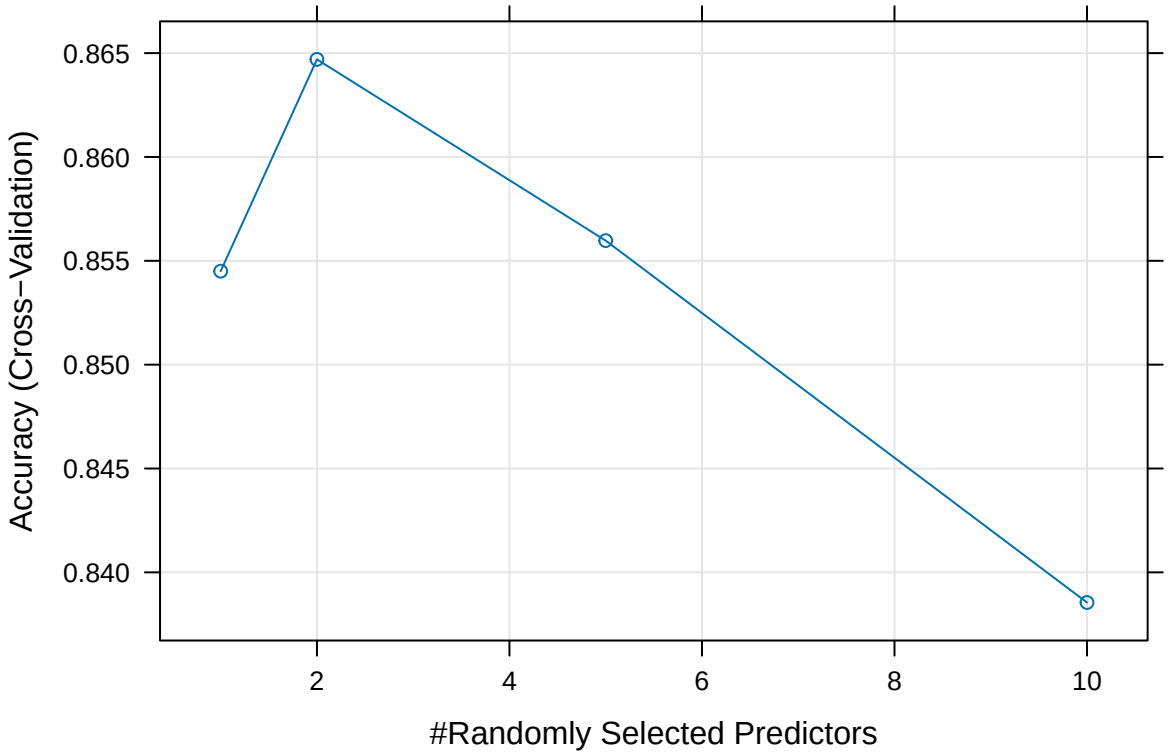


Table 10: RF Variable Importance

	Overall
Age	26.593647
Sex	9.825378
ChestPainType	45.314438
RestingBP	24.165674
Cholesterol	37.659406
FastingBS	7.102225
RestingECG	9.018538
MaxHR	35.799358
ExerciseAngina	28.274591
Oldpeak	34.451150
ST_Slope	74.479204

As regards *Variable Importance*, ECG parameters are again significant with *Cholesterol* and *ChestPain* present as well.

Parameterized



RESULTS

overall:

Table 11: Model Accuracy

Model	Accuracy
knn	0.6956522
Logistic Regression	0.8608696
xgb	0.8869565
nnet default	0.6304348
nnet	0.8608696
RF_test	0.8826087
RF_trained	0.8782609

CONCLUSION

Ensemble approaches seems to have an edge in prediction problems as demonstrated herein with the accuracy, 0.8869565 of the XgBooster model slightly surpassing the rest of the classification algorithms. Electrocardiogram (ECG) parameters such as the *ST_slope* seem to be of particular importance in the prediction of Heart Failure.

REFERENCES

Davide Chicco, Giuseppe Jurman. 2020. "Machine Learning Can Predict Survival of Patients with Heart Failure from Serum Creatinine and Ejection Fraction Alone." *BMC Medical Informatics and Decision Making* 20, 15.

fedesoriano. 2021. "Heart Failure Prediction Dataset. Retrieved 21Sept2023." https://www.kaggle.com/fedesoriano/heart-failure-prediction.