Comparison of Machine Learning Algorithms for Diagnosis of Diabetes

By:
Ojas Mehta
IIT Madras
ojasmehta189@gmail.com

July 31, 2019

- ## Objective:

To compare top ML classifiers on Pima Indians Diabetes Dataset.
The Authors in Paper “Do we Need Hundreds of Classifiers to Solve Real World Classification Problems?” by Manuel Fern´andez-Delgado, Eva Cernadas, et al. after comparing 179 classifiers arising from 17 families on 121 datasets have suggested some top classifiers which performed better than the others. So, six classifiers are taken from the suggested ones and compared on the Pima Indians Diabetes dataset.

- ## The Dataset:

Pima Indians Diabetes Dataset. This dataset is from the National Institute of Diabetes and Digestive and Kidney Diseases.
It is taken from the following link:
https://raw.githubusercontent.com/jbrownlee/Datasets/master/pima-indians-diabetes.data.csv .
The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. Hence a binary classification problem.
The dataset has following attributes:

- **Pregnancies**: Number of times pregnant
- **Glucose** Plasma glucose concentration a 2 hours in an oral glucose tolerance test
- **BloodPressure** Diastolic blood pressure (mm Hg)
- **SkinThickness** Triceps skin fold thickness (mm)
- **Insulin** 2-Hour serum insulin (mu U/ml)
- **BMI** Body mass index (weight in kg/(height in m)^2)
- **DiabetesPedigreeFunction** Diabetes pedigree function
- **Age** Age (years)
- **Outcome** Class variable (0 or 1) 268 of 768 are 1, the others are 0

Total 768 examples are there out of which 268 are positive(1) and 500 negative(0).

- ## Data Preprocessing:

Total 768 instances are there in dataset.

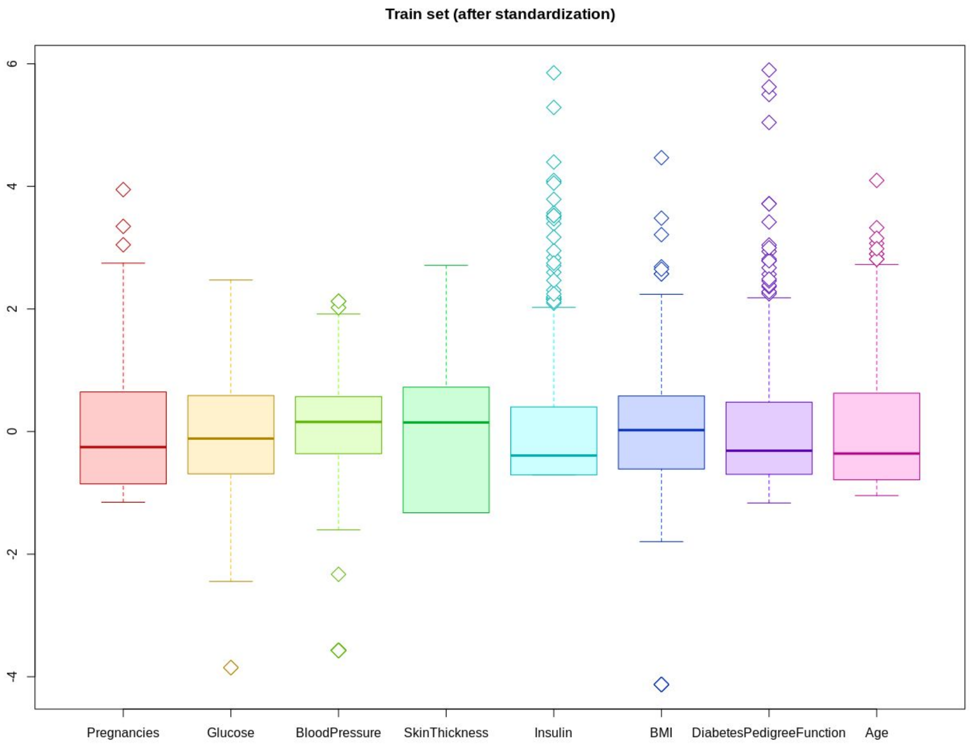
Missing Values: No missing value was found.

The data was split into train-test set in the ratio of 80-20, keeping the similar ratio of positive and negative instances in the train and test set to that of actual dataset.
So, Train set has 615 instances and Test set has 153 instances.
Train set is standardized and its mean and standard deviation are used to standardized the test set.

- ## Data Exploration:

1) Multivariate visualization:

From the image below, features “Insulin” and “DiabetesPedigreeFunction” seems to have maximum outliers.

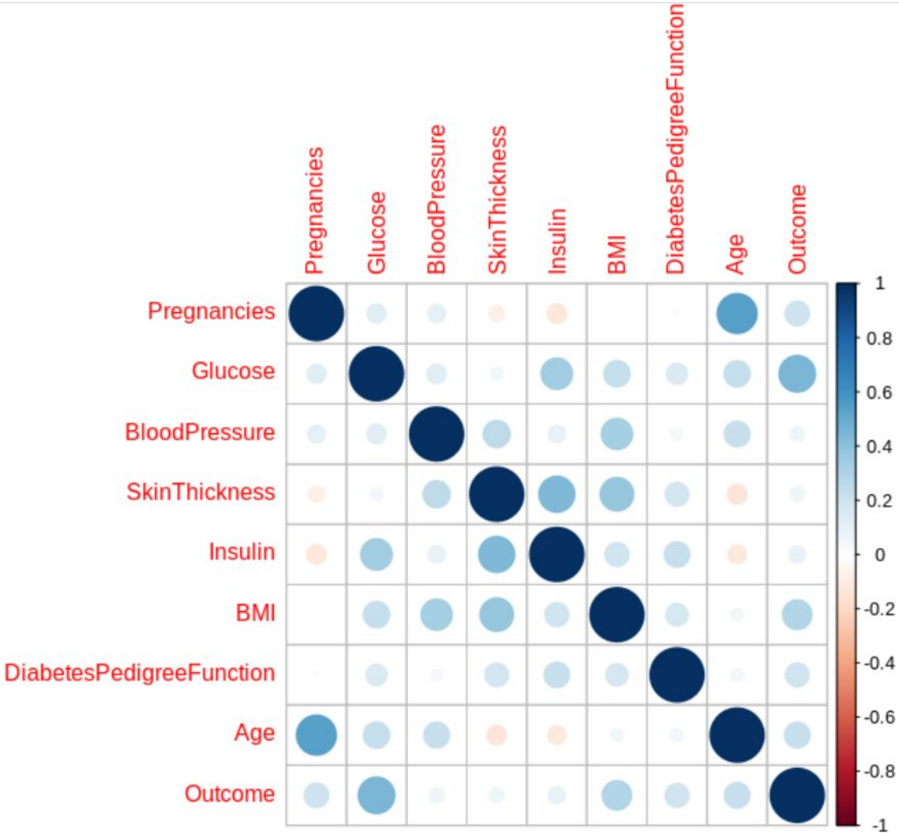


2)Correlation Plot:

Pregnancies and Age are highly correlated.

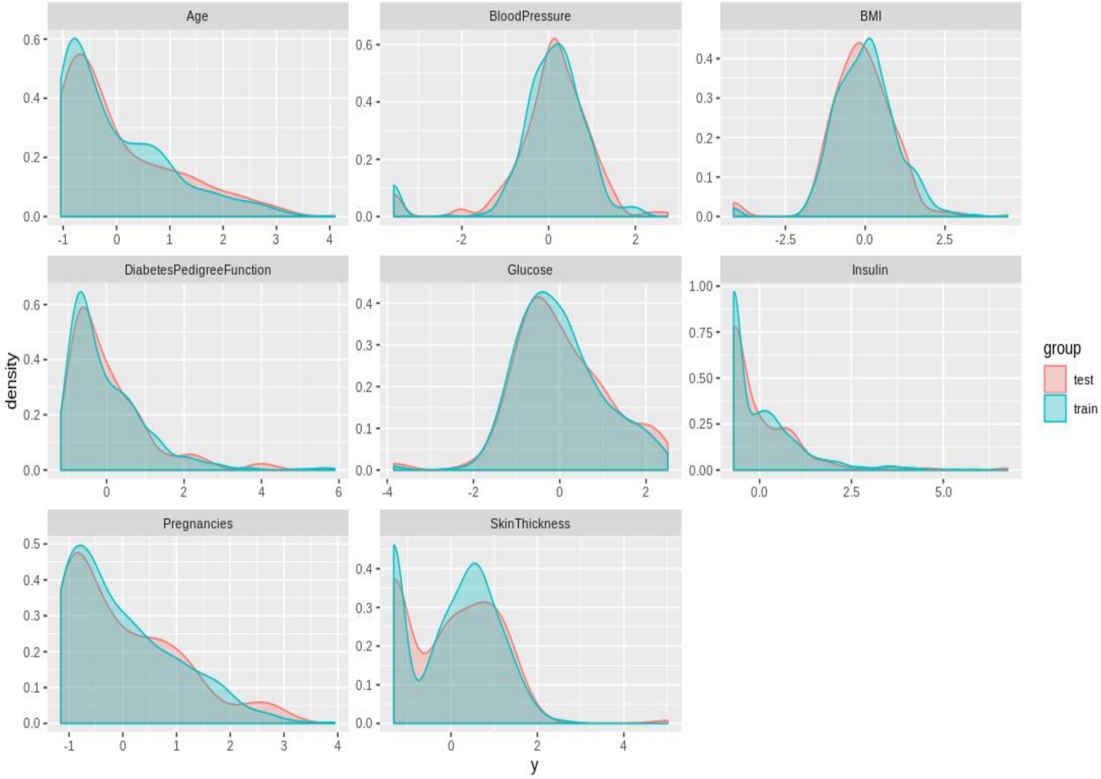
Also, Outcome which is response variable is highly correlated with Glucose followed by BMI. So, higher glucose and higher BMI indicates higher chances of diabetes.

One interesting observation that can be made is that SkinThickness and Insulin are negatively correlated with Age which means that both decreases with the increase in Age.



3)Features density in the Train and Test set:

The following plot shows the distribution of values of each feature. BloodPressure and BMI values are more evenly distributed around the mean.



- **Classifiers:**

  Method:

  - 10-fold cross validation is used to select the best parameters.
  - TuneGrid is used to select the best parameters. The parameters written are chosen after cross validation and searching through the parameter grid.
  - R is used as the programming language for it and the link of the code is: https://github.com/OjasMehta21/ML-Algo-Comparison/blob/master/MLCode.R

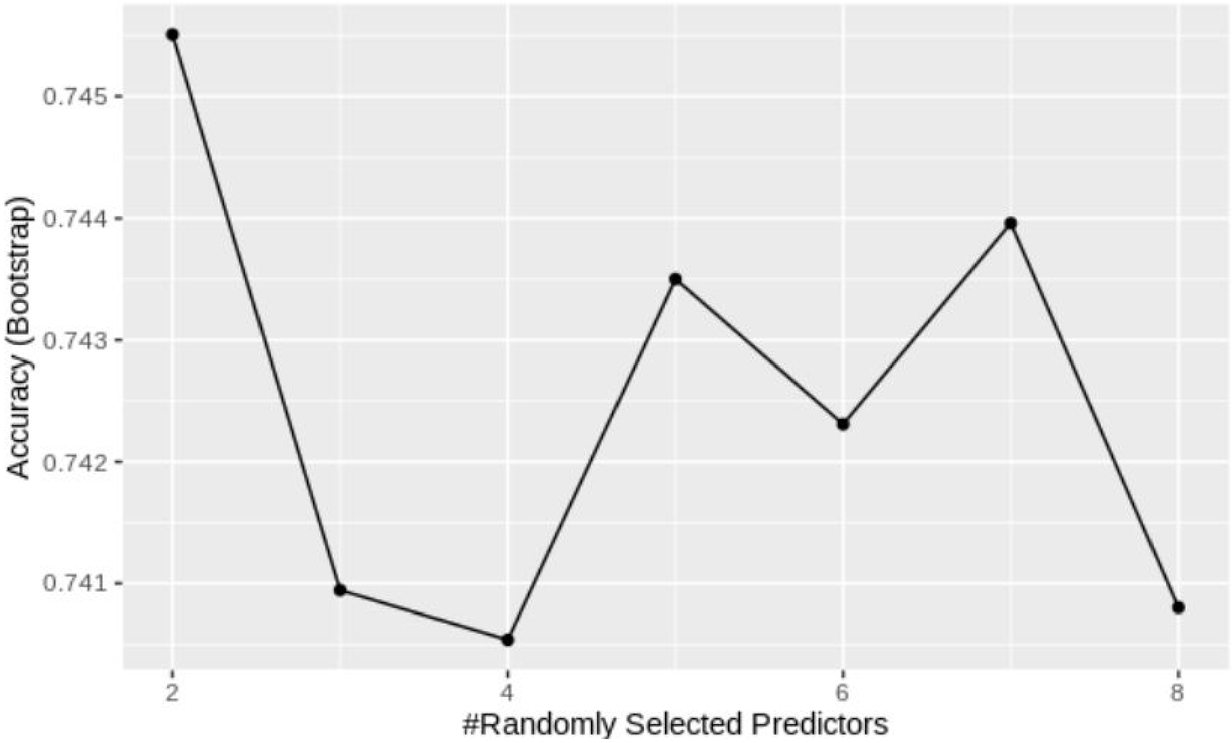
1) Parallel Random Forest (parRF):

a) Training:

Best tune parameters:

Number of trees, **ntree** = 500(default).

Number of Randomly Selected Predictors, **mtry = 2.**



b) Testing:

Accuracy: 78.43%

Confusion matrix		
	0	**1**
0	**85**	**15**
1	**18**	**35**

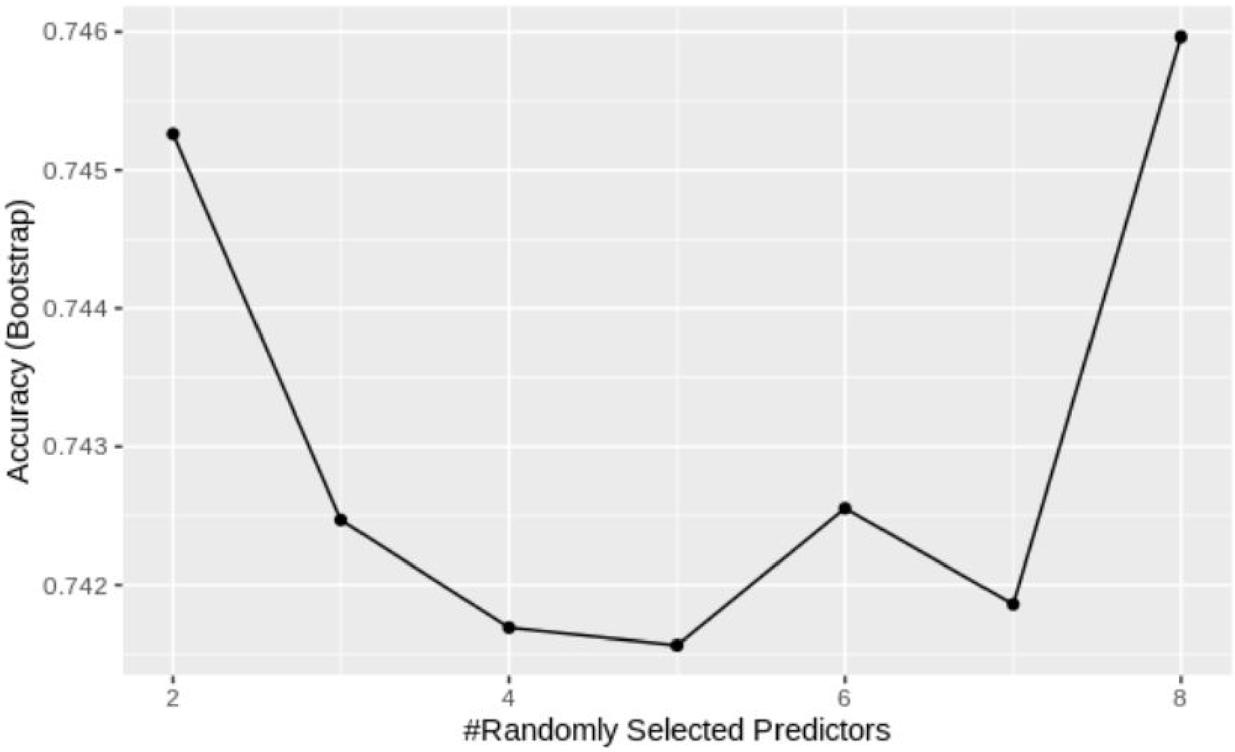
2) Random Forest (rf):

a) Training:

Best tune parameters:

Number of trees, **ntree** = 500(default).

Number of Randomly Selected Predictors, **mtry = 8.**



b) Testing:

Accuracy: 73.20%

Confusion matrix		
	0	**1**
0	**82**	**18**
1	**23**	**30**

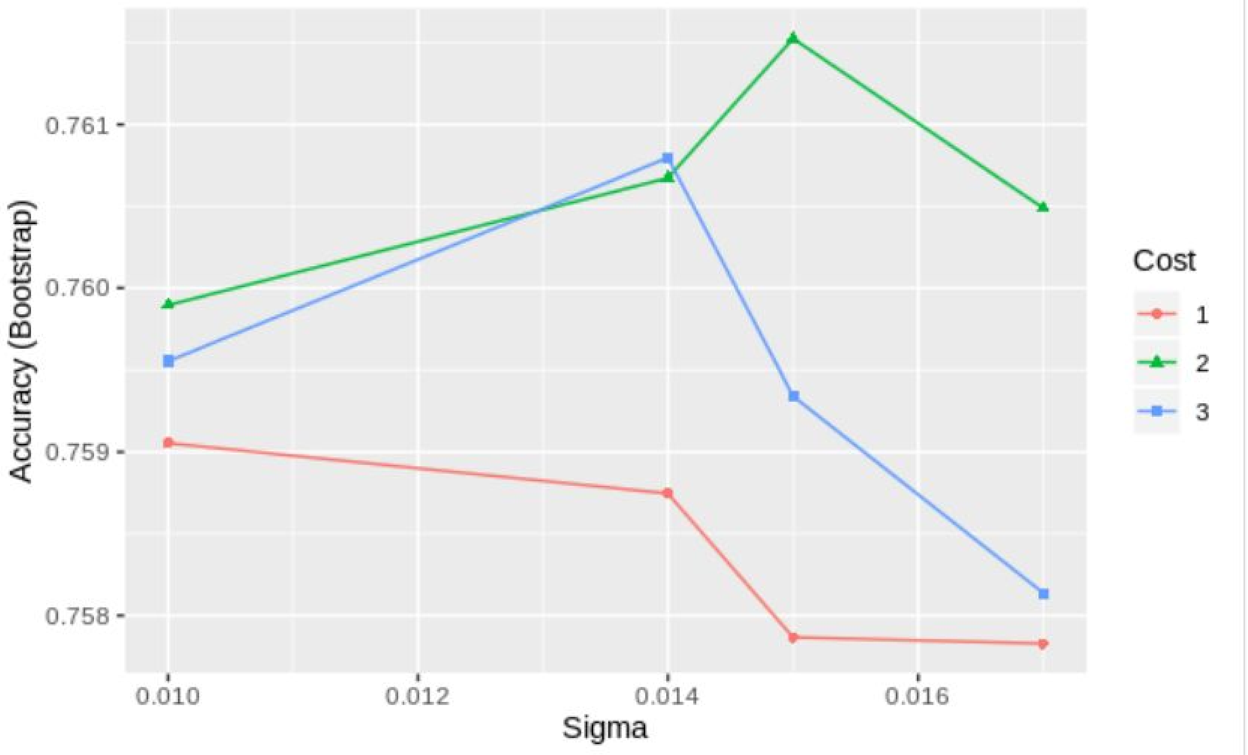
3) Support Vector Machines with Radial Basis Function Kernel(svmRadial):

a) Training:

Best tune parameters:

Sigma, **sigma :** 0.015

Cost**, C :** 2



b)Testing:

Accuracy: 80.39%

Confusion matrix		
	0	**1**
0	**93**	**7**
1	**23**	**30**

4) C5.0 Decision Tree(C5.0):

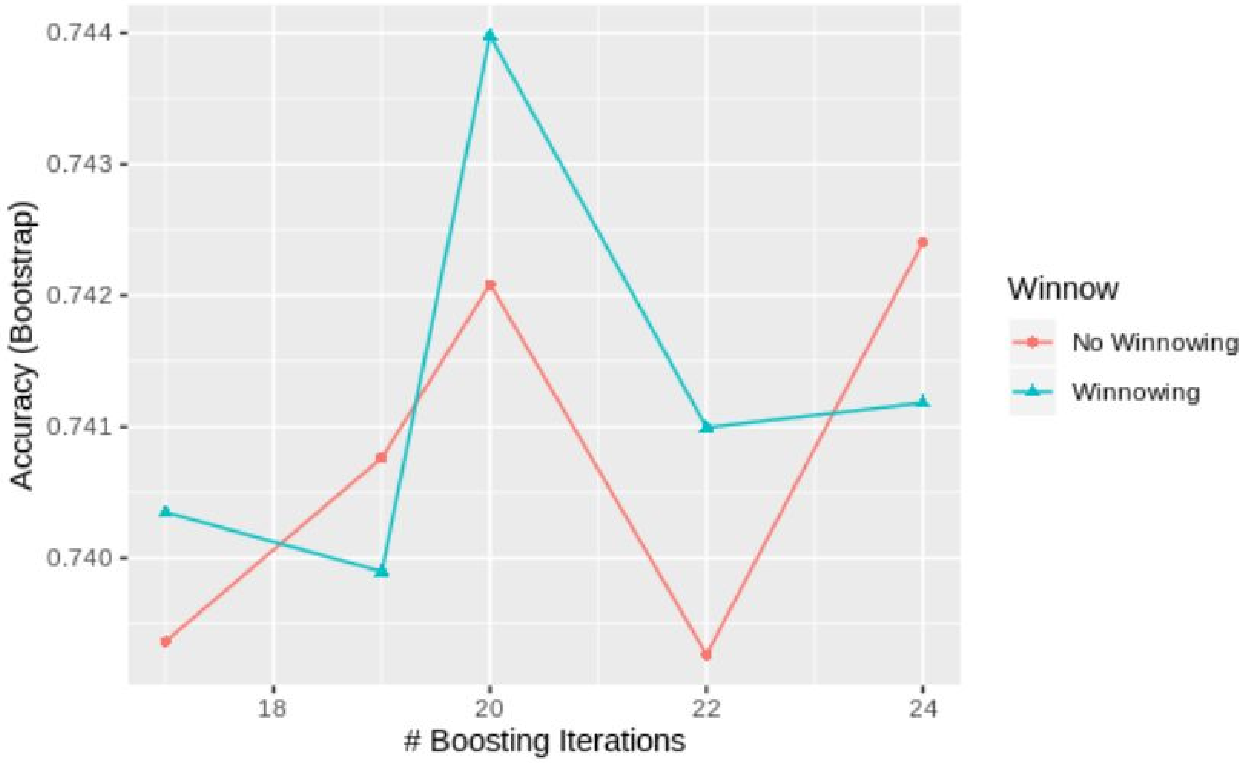
a) Training:

Best tune parameters:

Number of boosting iteration, **Trial:** 20

Type of Model, **Model:** tree

Winnow: TRUE



b)Testing:

Accuracy: 76.47%

Confusion matrix		
	0	**1**
0	**81**	**19**
1	**17**	**36**

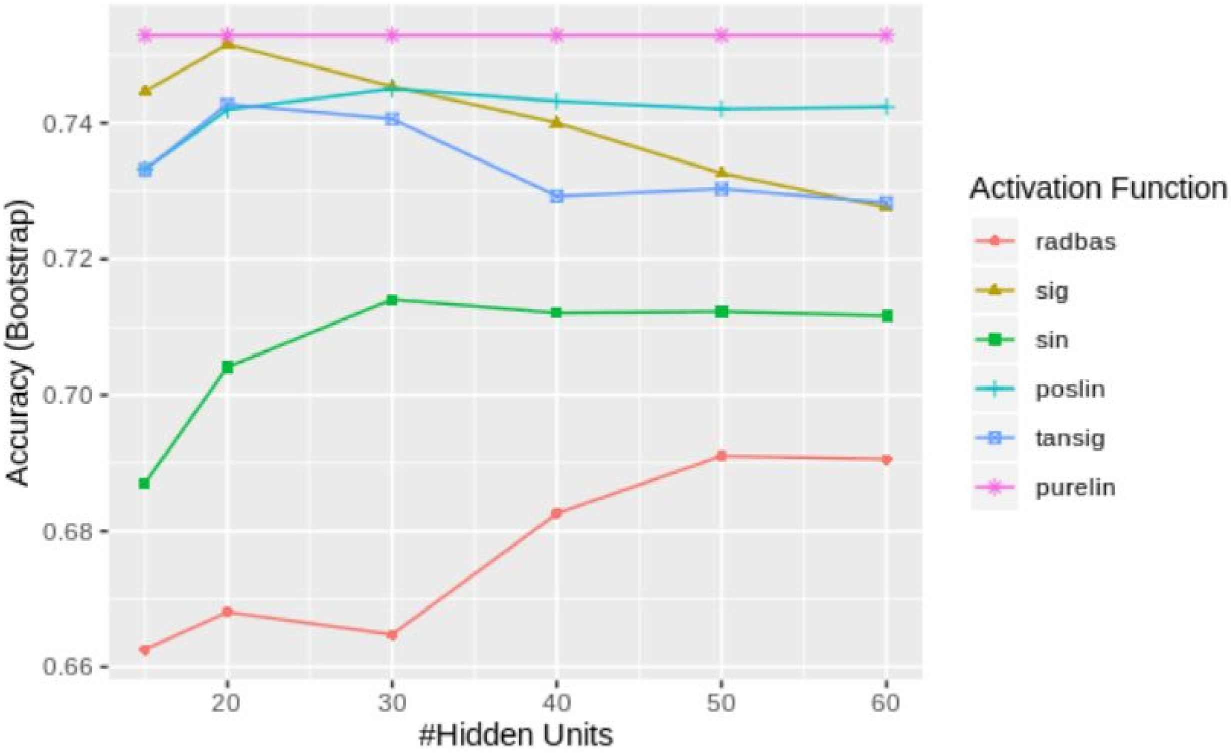
5) Extreme Learning Machine(elm):

a) Training:

Best tune parameters:

Number of Hidden Units, **nhid:** 15

Activation Function, **actfn:** “purelin”



b)Testing:

Accuracy: 77.78%

Confusion matrix		
	0	**1**
0	90	10
1	24	29

6) Model Averaged Neural Network(avNNet):

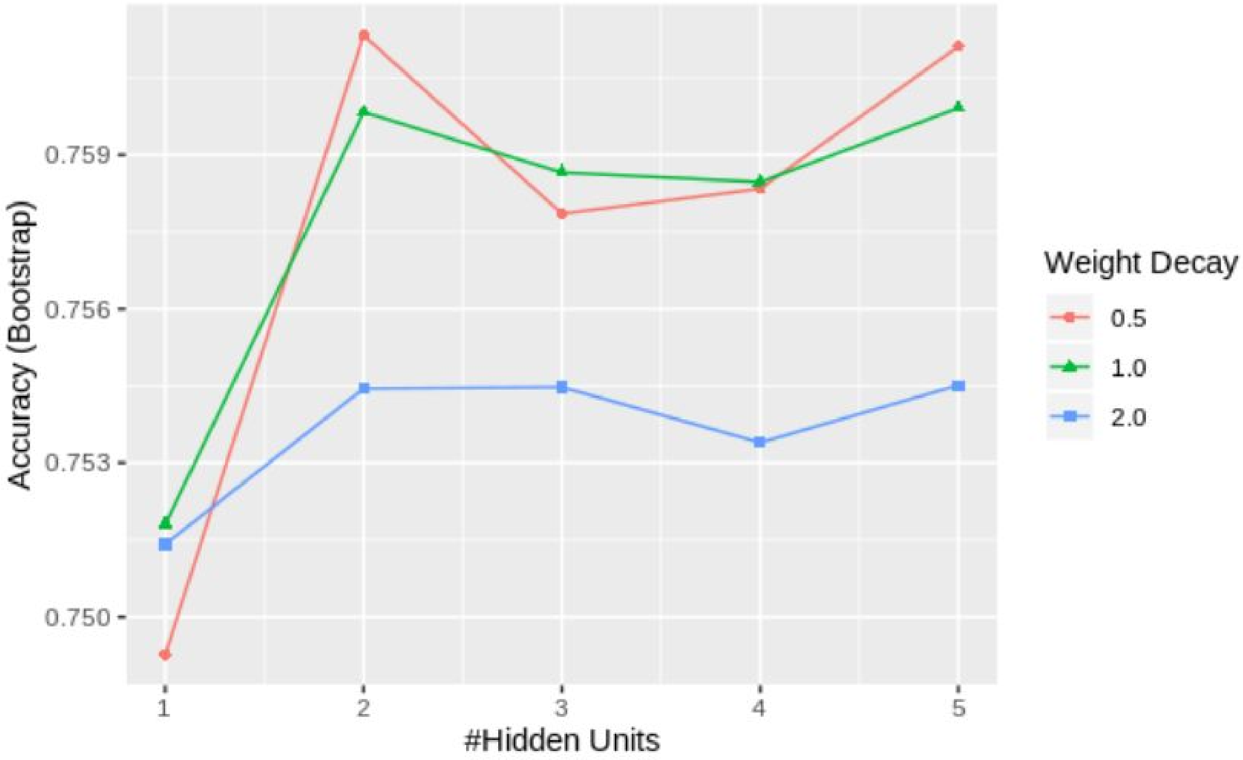
a) Training:

Best tune parameters:

Number of Hidden Units**, Size: 2**

Weight Decay**, Decay : 0.5**

Bagging**, Bag : FALSE**



b)Testing:

Accuracy : 77.78%

Confusion matrix		
	0	**1**
0	86	14
1	20	33

- **Conclusion:**

For the “PIMA Indians Diabetes Dataset”, the classifiers rank tabulated below. The SVM with Radial Basis Function turns out to be the best classifier.

Rank	Classifier	Accuracy
1	SVM with RBF (svmRadial)	80.39%
2	Parallel Random Forest (parRF)	78.43%
3	Model Averaged Neural Network (avNNet)	77.78%
3	Extreme Learning Machine (elm)	77.78%
4	C5.0 Decision Tree (C5.0)	76.47%
5	Random Forest (rf)	73.20%

- **References:**

1. **“**Do we Need Hundreds of Classifiers to Solve Real World Classification Problems?**”** by Manuel Fern´andez-Delgado, Eva Cernadas, Sen´en Barro, Dinani Amorim.

2. Pima Indian Diabetes Dataset:
Smith, J.W., Everhart, J.E., Dickson, W.C., Knowler, W.C., & Johannes, R.S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. *In Proceedings of the Symposium on Computer Applications and Medical Care* (pp. 261--265). IEEE Computer Society Press.
Download:
https://raw.githubusercontent.com/jbrownlee/Datasets/master/pima-indians-diabetes.data.csv