# Report of comparing 5 classification algorithms of Machine Learning:

**Author:** Anjali Patel

AITS Machine Learning Engineer intern

Prayagraj,UttarPradesh,India

9161082490 24septanjali@gmail.com

*Abstract*-**This paper is about the comparison of the five most popular classification algorithms of the supervised machine learning(ML). The algorithms are as :**

***Decision Trees**

***Boosted Trees**

***Random Forest**

***Support Machine Learning**

***Neural Networks**

**After the theoretical comparison, I have implemented each of the algorithm on a dataset of Drug_dt based on the features of sex, blood – pressure , cholesterol , Na to K parameters and the target of Drug to be given.**

***Keywords****- machine_learning_algorithms, comparison of five popular ml algorithms, Decision-trees,Boosted-trees,Random Forest, Support-Vector-Machine, Neural-Networks.*

## (I) Introduction:

Machine Learning is subtopic of the Artificial Language, the Machine Learning is an idea of making a machine learnt with the examples and implement it further. The machine learning can be categorised into three parts:-

1.Supervised Learning. 2.Unsupervised Learning. 3. Reinforcement. The supervised learning can further classified into two parts: 1.Classification. 2.Regression. The Unsupervised learning can also classified into two parts: 1.Association. 2.Clustering.Here we would discuss about the algorithms of the Classification, which is a type of the supervised learning.

## (II) Classification Supervised Learning:

The Supervised machine learningalgorithmsearches for the patterns within the value labels assigneddata points. Classification is applied when the output variable is a category.such as ‘High’ or ‘Low’, ‘Normal’ or ‘Abnormal’ , ‘Red’ or ‘Black’. Classification is the process of dividing the datasets into different categories or groups by adding labels.

### (A). Decision Trees.

The Decision tree algorithm is used to solve the classification problem as well as the regression problem, so known as the CART(Classification And regression Tree).

The decision tree uses the tree representation to solve the problems. Using it we can represent any Boolean function on the discrete attributes.

Some assumption that are made while using the decision tree:

- At the beginning , we consider the whole training set as the root.
- Feature values are preferred to be categorical.
- On the basis of attribute values records are distributed recursively.
- We use statical methods for ordering as root or the internal node.

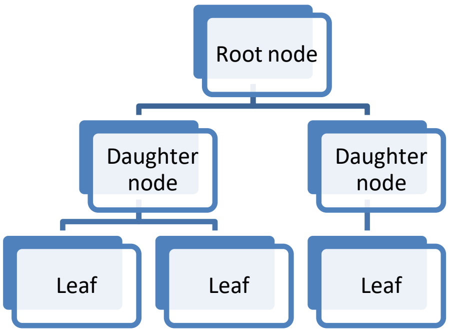


Fig 1. Structure of a decision tree.

**( B). Boosted Trees:**

The Boosted trees are the sequential Ensemble of the tree models. The Boosting tree is the model of machine learning algorithms to combine the weak learner to form the strong learner to improve the accuracy.

Fig 2. Concept of the Boosted Trees

How does the boosting work:

The basic principle behind the working of the boosting algorithm is to generate multiple weak learner and combine their prediction to form one strong rule. Many iteration are used to create the Decision stumps and combine several weak learners to form a strong learner.

STUMPS: These are the trees having the single node and Two leaves.

ADABOOST: The adaboost is used to make the collection of the trees that is the Boosted tree. It combines the stumps to form the boosted tree.

The boosted trees have:

1 Strong predective power , but even less interpredibility than forest. In it each successive tree uses the residue of the previous tree.

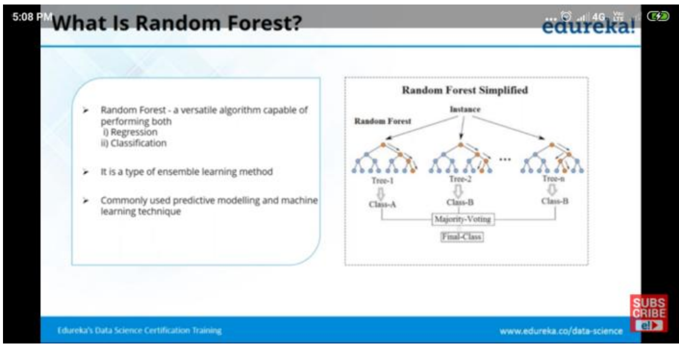
2 Even it have more hyperparameters to control model Building.

**(C). Random Forest:**

The Random forest algorithm is the parallel bagged ensemble of the number trees.This model have strong predective power but lower interpredibility as comparison to the decision tree.

More hyperparameters than the decision trees that control model growth are:

- Number of trees
- Sampling Rate
- Number of the variables to try.



What Is Random Forest?

- Random Forest - a versatile algorithm capable of performing both
  i) Regression
  ii) Classification
- It is a type of ensemble learning method
- Commonly used predictive modelling and machine learning technique

Random Forest Simplified

Edureka's Data Science Certification Training

www.edureka.co/data-science

Fig 3. Random forest

**The concept of Random Forest can be simplified as:**

The Random forest is the combination of the Decision trees. Let there are nimber of the decision trees used to resolve a dataset and there are different outcomes .The outcome which has been repeated most, means having the highest voting is the final consequence of the Random Forest algorithm results.

**(D). Support Vector Machine(SVM):**

A support vector machine (SVM) is a discriminative classifier formally defined by a separating hyperplane . In other word, given labelled training data , the algorithm optimals an optimal hyperplane which categorises new examples. In two dimentional space this hyperplane is a line dividing a plane in two parts where in each class lay in either side.

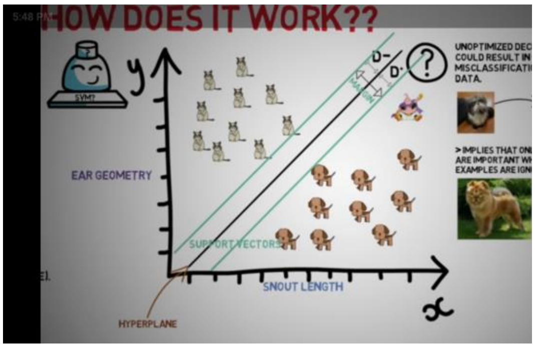

HOW DOES IT WORK??
SVM?
y
EAR GEOMETRY
SUPPORT VECTORS
SNOUT LENGTH
HYPERPLANE
D-
D+
UNOPTIMIZED DEC COULD RESULT IN MISCLASSIFICATI DATA.
>IMPLIES THAT ON ARE IMPORTANT WH EXAMPLES ARE IGN
x

**Fig** 4. Concept of SVM

**KERNELS:** If we have such a data that no any line can separate it into two classes in the X-Y plane .Now, we have to apply the transformation and add one more dimention the Z-axis.Now a line can be drawn which can separate the data into two classes. When we return to the original plane , it maps a circular boundry called kernel.

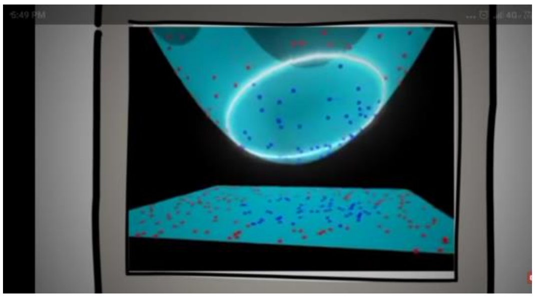


Fig 5. Kernel

**(E). Neural Networks.**

A neural network is a massively parallel distributed Processor that has a natural propensity for storing the experimental knowledge and making it available for.

A neural network , is a collection of layers that transform the input in some way to produce an output.

The perceptron is the basic unit of the neural network. The perceptron is consist of two types of nodes : Input nodes and output nodes, each input node is connected via weighted link to the output node.

ΔW=ή.d.x

d=predicted output-desired output

x= Input data

ή=Learning Rate

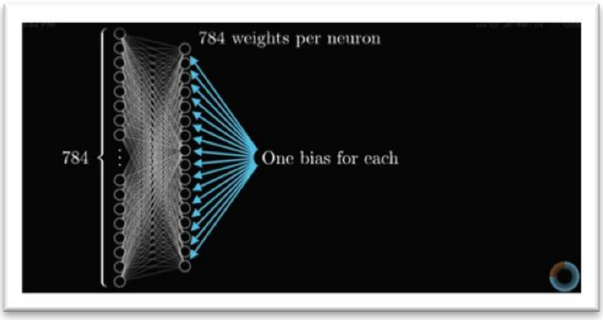


**Fig** 6. Neural Network

**(III) Experiment Exploration:**

In the colab coding section I have implemented the above described algorithms on a dataset Drug_dt.

Imported the dataset using the pandas.then preprocessed the data and then segregate it as follows:

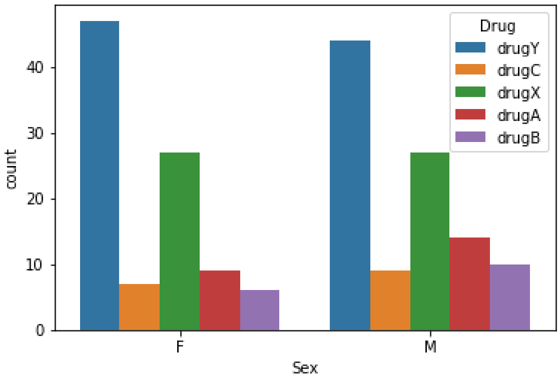


**Fig** 7. Graph count vs sex.

The dataset having the features of sex , BP, Cholestrol, Na-to –K concentration and the target of the Drugs which are the categorical data.

Split the data into train test format of 0.30 to train the various models and then test to it.

- **From sklearn.tree import DecisionTreeClassifier.**
- **From sklearn.ensemble import AdaBoost Classifier.**
- **From sklearn.ensemble import RandomForestClassifier.**
- **From sklearn.svm import Linear SVC.**
- **From sklearn.neural_network import MLP Classifier.**

## (IV) Results:

The results of the different algorithms models are different in the term of the accuracy:

| Serial no. | Algorithms | Accuracy (%) |
|---|---|---|
| 1 | Decision Trees | 98 |
| 2 | Boosted  Trees | 73 |
| 3 | Random Forest | 95 |
| 4 | SVM | 61 |
| 5 | Neural networks | 48 |

Table 1.

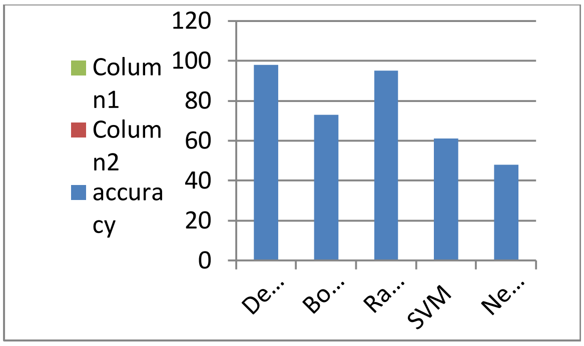


Fig 8. Graph Representation of  results.

## (V) Conclusion:

The conclusion of the above whole discussion and the exploration  is that the five compared algorithms models are used for the classification problems and some are used for the regression problems. For the dataset taken by me the Decision Trees algorithm model is best and then is the RandomForest model.

## (VI).    References:

- https://www.analyticsvidhya.com
- https://www.edureka.co
- www.analyticsindiamag.com
- Data Analytics and Machine Learning  :-By Chandan Verma(author).