# Random Forest Classifier

Mobin Mustakali Momin*
Master's in Computer Science
Bishop's University
Sherbrooke, Canada.
mobinmomin1994@gmail.com

Vishwas Hasija
Master's in Computer Science
Bishop's University
Sherbrooke, Canada.
hasijavishwas@gmail.com

Dhiraj Wani**
Master's in Computer Science
Bishop's University
Sherbrooke, Canada.
dhirajwani70@gmail.com

UNIVERSITÉ BISHOP'S UNIVERSITY

Master's in Computer Science, Bishop's University

***Abstract*-** Pattern recognition is the process of undermining the pattern within the data using Machine Learning. It can also be stated that pattern recognition is the classification of data based on the knowledge already gained or on statistical information extracted from data itself. 1973 (Duda and Hart) defined the pattern recognition is field concerned with machine recognition of meaning regularities in noisy complex environment. Pattern recognition is used in any area of science and engineering that studies the structure of data. It has many application in the field of manufacturing, health care and military.

***Index Terms*-** Machine Learning, Random Forest Classifier

## I. Introduction

This research paper is a guide to how we implemented our learnings in pattern recognition during the course CS-509 at bishop's university. The main of this project was to implement Random Forest classifier successfully. Radom forest classifier is a supervised ensemble algorithm. Where ensemble means when more than one algorithm is used to make one algorithm. The project was done by a group of three students pursuing Master's in Computer Science program at the Bishop's university.

## I. Machine Learning and Random Forest Classifier

### 1) Machine Learning:

Machine learning is science of making computer learn and act like human. Improve their learning over the time by feeding data and information in terms of information and interaction with real world. "Machine learning at its most basic is the practice of using algorithms to parse data, learn from it and then prediction about something in the real world"-Nvidia. Fundamental goal of machine learning to generalize beyond training sample and interpret the data that it has never seen. Machine learning is the combination of Representation, Evaluation and Optimization. Where choosing a representation for a learner is on par with choosing a classifier that it can possibly learn. This is called hypothesis set and learner cannot learn without hypothesis set, it is called **Representation**. Each algorithm has implicit evaluating function for itself which is completely different from external one. These evaluating function are necessary part of any machine learning algorithm for optimization called **Evaluation**. The **optimization** is the key for the efficiency of learner. Machine learning can be grouped on the basis of learning style (i.e, supervised , unsupervised and semi-supervised ) or on the basis similarity in form or function (i.e, classification, regression, decision tree, clustering and deep learning etc).

### 2) Differentiation on the basis of Learning Style:

Algorithm are differentiated on the basis of their learning style. There is three types of algorithms: Supervised, unsupervised and Semi-Supervised. In **Supervised learning**, algorithms is used to map input data to output data using some function. For example there is input variable (Y) and output variable (X), algorithm is used to learn the mapping function from input to output data.

$$X = f(Y).$$

Data is divided in two parts one part being train and another part being test. Train data is used to learn the mapping function between input and output data. Where test data is used to test the mapping function of an algorithm. Predictive output from algorithm is compared with actual output to check the accuracy of algorithm. **Unsupervised learning**, learns few features from the input data and when new data is entered it used to classify this data

on the basis of previous learnings. Unlike supervised learning there is no teacher to make it learn. It is used to classify input data to output data on the basis of similarities. It is left on their own decisive power. **Semi-Supervised** have the power of both supervised and unsupervised. In semi-supervised type large chunk of data is present out of which some of amount of data have output for its input while other don't, taking example of photo archive where some of the photos only are labelled as it is very cumbersome to label all of them.

**3) Differentiation on the basis of similarity of function:**

**Approximation algorithm** are the algorithm which are used to map input data X to output data (Y) using some function (f). It is used to develop a model using historical data and predict the output for new data. **Classification predictive** modelling is used to approximate mapping function from input data (X) to output data (Y) where output data is labelled or categorized. **Regression Modelling** it is used to model an approximate mapping function from input data to output data, where output is a real integer or floating point value such as amount or quantity.

**4) Random Forest Classifier:**

Random Forest Classifier is ensemble learning method used for classification and regression problems. Ensemble learning paradigm uses more than one method to solve the similar problem. Data used to train the model is divided into in-bag and out-of-bag instances. Instances which are used to train the model are called in-bag instance, these account for sixty four percent of total instances. Whereas left out thirty six percent of instance are called out-of-bag instances, these left out outputs are used to form accurate predictions and using estimated outputs instead of observed outputs increase accuracy in model. Hence, they can be used to reduce error and increase accuracy. Random Forest Classifier is extension of Decision tree, which repetitively divides working data into decision lines forming a tree. Single tree is prone to noise and impurity. On other hand Random Forest Classifier use multiple decision tree as classifiers to reduce noise and increase accuracy. Each tree in model act as a classifier and is used to label the unlabeled instances. This is done by voting where each classifier casts one vote for its class label. Then label with most vote is used to label the instance. Randomization is applied when selecting best node to split from two leave nodes. This randomization is square root of F, where F is number of features selected in data set. Error rate in Random forest classifier depends on the correlation and strength, where error rate is directly proportional to correlation and inversely proportional to strength. Increasing strength between any two trees in random forest also increases the error rate. Where increasing the strength of any tree leads to decrease in error rate. Error rate can be decrease sufficiently by decreasing correlation and increasing strength simultaneously.

Major advantages of Random forest is robustness to noise and overfitting. Overfitting means when a constructed model fits in the data more than wanted. This overfitting leads to poor predictive performance and it lead uncertainty of how well the model make prediction for the cases are not in training set. Other advantage are that it have more accuracy than it counter parts, faster than bagging and bootstrap and it have implicit estimators of error, strength and accuracy.

Applications of Random forest classification, In Ecology (Cutler et al. 2007) random forest classifier is used species data collected from various location in USA. In Autopsy, a new computer-coded verbal autopsy is introduced to predict the cause of death. In agriculture, RF is used to classify the crops and provide spatial information. In Bioinformatics and Computational Biology RF is used to select parameters.

## II. IMPLEMENTATION

We have used Random forest classifier to create a model for a Telecom Industry which predict whether customer will stay with particular service provider in future or not. We have imported this data from kaggle. This problem will fall under the supervised learning and classification problem. As already discussed in supervised learning output data is labelled, in our case output is labeled as customer will stay or not. In classification problem is categorized in groups of the basis of similarity of output, in this case output is categorized on the basis of yes or no. We have decided to implement Random forest in python. Along with random forest classifier we have used numpy, pandas and standard Scaler libraries. We have used panda to standardize the floating point data (Total Charge) into integer value for easy access. We have used Standard scaler to scale down total charges, monthly charge and tenure. Currently accuracy of this model is eighty percent. Accuracy can be improved by taking more fields into consideration for training the model.

## REFERENCES


[1] Machine learning Alogrithm: A review by Ayon Dev, Department of CSE, Gautam Budha University, Greater Noida , Uttar Pradesh, India

[2] A Few Useful Things to Know about Machine learning, Pedro Dominque, Department of CSE, University of Washington, Seattle, USA

[3] Random Forest Classifier from early developments to recent advancements, khaled Fawagreh, Mohamed Medhat Gaber,

[4] Out-of-Bag Estimation, Leo Bremain, Statistics Deparment, University of California, Berkeley, USA

[5] E. H. Miller, "A note on reflector arrays (Periodical style—Accepted for publication)," *IEEE Trans. Antennas Propagat.*, to be published.

[6] J. Wang, "Fundamentals of erbium-doped fiber amplifiers arrays (Periodical style—Submitted for publication)," *IEEE J. Quantum Electron.*, submitted for publication.


## AUTHORS


**Mobin Mustakali Momin-** Pursuing M.S, Bishop's University,mobinmomin1994@gmail.com

**Vishwas Hasija** – Pursuing M.S, Bishop's University, hasijavishwas12@gmail.com

**Dhiraj Wani** – Pursuing M.s, Bishop's University, dhirajwani70@gmail.co