# Random Forest

## I. Abstract

A random forest is a controlled machine learning method built from decision tree techniques. In machine learning, it may be utilised for both classification and regression issues. It is based on ensemble learning, which is a method of integrating several classifiers to solve a complicated issue and increase the model's performance. This algorithm is used to anticipate behaviour and results in a variety of sectors, including banking and e-commerce.

This document describes the algorithm and how it operates. The article will describe the algorithm's characteristics as well as how it is used in real-world situations. It also discusses the algorithm's benefits and drawbacks.

## II. Introduction

A random forest is a machine learning approach for solving regression and classification issues. It makes use of ensemble learning, which is a technique that combines multiple classifiers to solve complicated problems. A random forest method is made up of a large number of decision trees. The random forest algorithm's 'forest' is trained via bagging or bootstrap aggregation. Bagging is a meta-algorithm ensemble that enhance the reliability of machine learning algorithms.Based on the predictions of the decision trees, the random forest algorithm determines the outcome. It forecasts by averaging or averaging the output from different trees. The precision of the result improves as the number of trees grows. The constraints of a decision tree algorithm are eliminated with a random forest. It reduces dataset overfitting and boosts precision. It provides forecasts without the need for several package settings [**1**].

## III. Characteristics of a Random Forest Algorithm

Random Forest outperforms the decision tree algorithm in terms of accuracy. It gives an efficient method of dealing with missing data. Random forest is capable of producing a fair forecast without the need of hyper-parameter tweaking. It eliminates the problem of overfitting in decision trees. Random forest the node's splitting point in every random forest tree, a subset of characteristics is chosen at random [**2**].

## IV. Working of Algorithm

To have a better knowledge of random forest, we should first learn about decision trees. A random forest algorithm's building components are decision trees. A decision tree is a decision-making approach with a tree-like structure. A review of decision trees will assist us in comprehending how random forest algorithms function. A decision tree is made up of three parts: decision nodes, leaf nodes, and a root node. A decision tree method separates a training dataset into branches, which are then subdivided into subbranches. This procedure is repeated until a leaf node is reached. The leaf node cannot be further separated. The decision tree's nodes indicate qualities that are used to forecast the outcome. The decision nodes connect to the leaves, Figure 1 depicts the three types of decision tree nodes. [**3**].

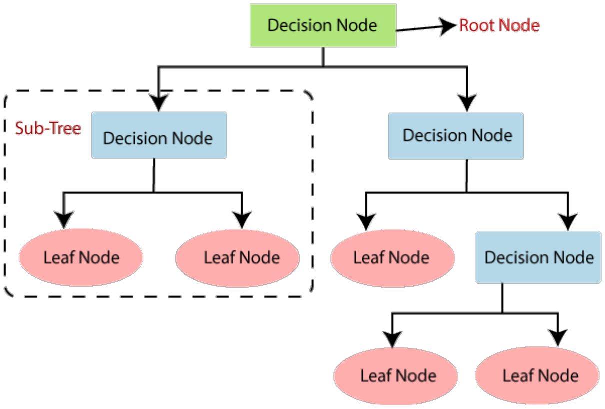

Fig. 1. General structure of a decision tree [**3**]

A supervised learning algorithm is random forest. The ”forest” it creates is an ensemble of decision trees, which are often trained using the ”bagging” approach. The bagging approach is based on the premise that combining learning models improves the final output. Simply said, random forest constructs many decision trees and blends them to get a more accurate and consistent forecast. Random forest has a significant benefit in that it can be utilised for both classification and regression tasks, which comprise the majority of contemporary machine learning systems. Let's look into random forest in classification, because classification is frequently said to be the foundation of machine learning. Random forest has roughly identical hyperparameters as decision trees and bagging classifiers. Fortunately, there is no need to combine a decision tree with a bagging classifier because the random forest classifier-class may be used instead. You may also use random forest to handle regression jobs by employing the algorithm's regressor. While growing the trees, Random Forest adds more randomization to the model. When splitting a node, it looks for the best feature from a random group of characteristics rather than the most essential feature. As a result, there is a wide range of variability, which leads to a better model in general [**4**].

## V. Random forest classification

Random forest classification uses an ensemble technique to get the desired result. Various decision trees are trained using the training data. This dataset contains observations and characteristics that will be chosen at random when nodes are divided [13].

Various decision trees are used in a rain forest system. There are three types of nodes in a decision tree: decision nodes, leaf nodes, and the root node. Each tree's leaf node represents the ultimate result produced by that particular decision tree. The final product is chosen using a majority-voting procedure. In this situation, the ultimate output of the rain forest system is the output chosen by the majority of decision trees. A basic random forest classifier is shown in figure 2 below [13]. Because the random forest mixes numerous trees to forecast the dataset's class, some decision trees may correctly predict the output while others may not. However, when all of the trees are combined, the proper result is predicted. As a result, two assumptions for a better Random forest classifier are as follows [6]:

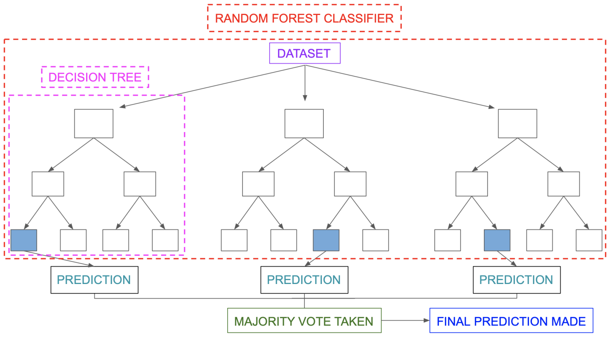

Fig. 2. basic random forest classifier [13]

- The dataset's feature variable should have some real values so that the classifier can predict correct outcomes rather than a guess.
- Each tree's predictions must have very low correlations.

## VI. Example to Understand the Concept

Consider a training datasetas shown in figure 3 made up of diverse fruits such as bananas, apples, pineapples, and mangoes. This dataset is divided into subsets by the random forest classifier. In the random forest system, these subsets are assigned to each decision tree. Each decision tree generates a distinct outcome. For example, apple is predicted for trees 1 and 2.

Another decision tree (n) indicated that the outcome would be banana. To offer the final prediction, the random forest classifier accumulates the majority voting. The apple has been predicted by the majority of the decision trees. As a result, the classifier selects apple as the final prediction[12].

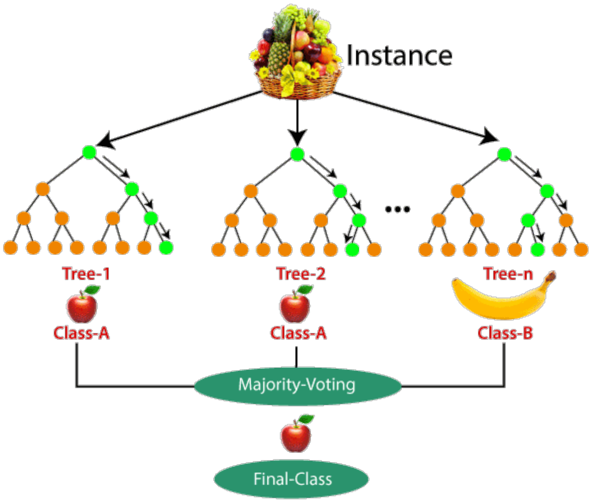

Fig. 3. The working of the algorithm [12]

## VII. Random Forest Regression

A random forest algorithm also does regression. A random forest regression is based on the notion of simple regression. The random forest model takes into account the values of dependent features and independent variables. Random forest regressions may be conducted in a variety of tools, including Statistical Analysis System and Python. In a random forest regression, each tree makes a unique forecast. The regression result is the average forecast of the individual trees. This is in contrast to random forest classification, where the output is dictated by the mode of the decision trees class [5].

Although the concepts of random forest regression and linear regression are similar, they differ in terms of functions. Linear regression is defined as y=bx + c, where y is the dependent variable, x is the independent variable, b is the estimate parameter, and c is a constant. A complicated random forest regression function is similar to a blackbox [1].

## VIII. Random forest applications

Random Forest Analysis may be applied in a variety of disciplines. Random Forest has a wide range of applications in a variety of industries.

### A. Banking Sector

There are many devoted customers, as well as victims of deception. Random forest analysis is used to determine if a client is loyal or dishonest. It can simply determine whether a client is dishonest or loyal using a random forest machine learning technique. A framework employs a random set of algorithms to categorise fraudulent transactions based on a sequence of patterns.

- Detection of Credit Card Fraud
  Credit card firms should detect fraudulent credit card purchases so that consumers are not charged for things they did not purchase. However, it is a very challenging process because there might be only thousand occurrences of fraud in over a one million transactions, representing just 0.1 percent of the dataset, resulting in highly skewed datasets. When trained on unbalanced datasets, Machine learning algorithms are more likely

to produce erroneous classifiers because they appear to favour the majority class, perceiving the minority class as noise in the dataset. There is also no useful statistic for imbalanced categorization because of the ”accuracy” class imbalance. The programme could anticipate almost all situations belonging to the majority class, yielding a high accuracy score. For this, we may still utilise the Random Forest Classifier [**6**].

### *B. Healthcare and Medicine*

Random forest algorithms are used by doctors to diagnose patients. Patients are diagnosed by reviewing their medical history. Previous medical data are evaluated to determine the appropriate dose for the patients. Medicine need a complicated blend of certain substances. As a result, Random Forest may be utilised to identify the perfect blend of medications. With the use of a machine learning system, it is now possible to detect and forecast a medication's drug sensitivity. It also assists in recognising the patient's condition by analysing the patient's medical record.

- Prediction of heart disease
  Using the random forest method, researchers were able to predict heart disease with an accuracy of 86.9 percent, a sensitivity of 90.6 percent, and a specificity of 82.7 percent. The diagnostic rate for heart disease prediction utilising random forest is 93.3 percent, according to the receiver operating characteristics [**7**].
- Diabetes Prediction
  Diabetes is a metabolic disorder characterised by a shortage of insulin induced by a faulty pancreas. Diabetic coma, weight loss, and pathological death of pancreatic beta cells are all possible outcomes of diabetes. Outlier rejection, missing values, data standardisation, K-fold validation, and several Machine Learning classifiers were all included in a robust framework for the early identification of diabetes. Random Forest outperforms all other classifiers in this area [**6**].

### *C. Stock Market*

Machine learning is also used to evaluate the stock market. The Random Forest technique may be used to investigate the stock market's behaviour. The predicted profit or loss that might be made when acquiring a specific stock can also be depicted.

- Stock Market Predication
  Predicting the stock market is a method of estimating future inventory costs. Because stock values fluctuate on a daily basis, determining the ideal moment to purchase and sell stocks is tricky. Since its creation, it has been an interesting issue for scholars and investors. Machine Learning generates a wide range of algorithms, one of which, Random Forest, has been shown to be quite good at predicting future stock values [**8**].

### *D. E-Commerce*

It's tough to propose or recommend what kinds of items a customer should view. This is where a forest-based random method might be useful. A machine learning system may be used to propose items that a buyer is more likely to have. You may offer comparable items to your clients by using a template and tracking a client's interest in the product.

- Product Suggestions
  Amazon has demonstrated that product reviews are effective. The Recommendation Engine accounts for more than 30 percent of the company's sales. Identifying the proper patterns in product sales and purchasing behaviour, on the other hand, necessitates a significant amount of computer capacity.
  Machine learning can help with this. Machine learning can smoothly quantify purchase behaviour over and over again, delving deeper into patterns each time. It's powerful to recommend something to the clients that they didn't even realise they wanted. [**9**].
- Optimization of Price
  Pricing is quite important. The importance of pricing on the internet cannot be overstated. Machine learning technologies can change pricing to account for several variables at once. Competitor rates, demand, time of day, and consumer type might all have an impact on pricing. Machine learning technology allows for pricing changes to be made as needed.

## IX. Not ideal in the following circumstances

In the following cases, random forest algorithms are not ideal. Extrapolation is ranked first, In terms of data extrapolation, random forest regression isn't optimal. Unlike linear regression, which utilises existing observations to estimate values outside of the observation range, nonlinear regression employs existing observations to estimate values outside of the observation range. This explains why the majority of random forest applications are related to categorization. and yet another is Data that is sparse When the data is scarce, random forest does not generate good results. In this situation, the bootstrapped sample and the subset of features will result in an invariant space. This will result in ineffective divides, which will have an impact on the outcome [**10**].

## X. Random forest benefits and drawbacks

## XI. Conclusion

## XII. References

[1] Onesmus Mbaabu. (2020, December 11). Introduction to Random Forest in Machine Learning. Https://Www.Section.Io/. https://www.section.io/engineering-education/introduction-to-random-forest-in-machine-learning/

[2] R, S. E. (2021, June 24). Random Forest — Introduction to Random Forest Algorithm. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/

[3] Machine Learning Decision Tree Classification

Algorithm - Javatpoint. (n.d.). Www.Javatpoint.Com. https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm

[4] Donges, N. (2021, September 17). A Complete Guide to the Random Forest Algorithm. Built In. https://builtin.com/data-science/random-forest-algorithm

[5] Bakshi, C. (2021, December 14). Random Forest Regression - Level Up Coding. Medium. https://levelup.gitconnected.com/random-forest-regression-209c0f354c84

[6] Meena, M. (2020, December 8). Applications of Random Forest. OpenGenus IQ: Computing Expertise and Legacy. https://iq.opengenus.org/applications-of-random-forest/

[7] Pal, Madhumita Parija, Smita. (2021). Prediction of Heart Diseases using Random Forest. Journal of Physics: Conference Series. 1817. 012009. 10.1088/1742-6596/1817/1/012009.

[8] Zakariya, M. (2022, February 2). Predicting Stock Prices Using Random Forest Model. Medium. https://medium.com/@maryamuzakariya/project-predict-stock-prices-using-random-forest-regression-model-in-python-fbe4edf01664

[9] Khanvilkar, Gayatri Vora, Deepali. (2019). Product Recommendation using Sentiment Analysis using Random Forest Approach Gayatri. 8. 146-152.

[10] Shammeer, M. (2022, March 30). Random Forest Fails - The Startup. Medium. https://medium.com/swlh/random-forest-fails-a8ca2d46c312

## XIII. Declaration