WEATHER PREDICTION USING RANDOM FOREST ALGORITHM

A MAJOR PROJECT REPORT

Submitted by

CH.EN.U4AIE20001	ANBAZHAGAN E
CH.EN.U4AIE20053	RAMYA POLAKI
CH.EN.U4AIE20069	TRINAYA KODAVATI
CH.EN.U4AIE20031	SMITHIN REDDY K
CH.EN.U4AIE20035	ABHIRAM KUNCHAPU

in partial fulfilment for the award of the degree

Of

BACHELOR OF TECHNOLOGY

IN

PYTHON FOR MACHINE LEARNING

AMRITA SCHOOL OF ENGINEERING, CHENNAI

AMRITA VISHWA VIDYAPEETHAM

CHENNAI – 601103, TAMIL NADU

AMRITA VISHWA VIDYAPEETHAM

AMRITA SCHOOL OF ENGINEERING, CHENNAI, 601103

श्रद्धावान् लभते ज्ञानम्

BONAFIDE CERTIFICATE

This is to certify that the major project report entitled “**WEATHER PREDICTION USING RANDOM FOREST ALGORITHM”** submitted by

CH.EN.U4AIE20001	ANBAZHAGAN E
CH.EN.U4AIE20053	RAMYA POLAKI
CH.EN.U4AIE20069	TRINAYA KODAVATI
CH.EN.U4AIE20031	SMITHIN REDDY K
CH.EN.U4AIE20035	ABHIRAM KUNCHAPU

in partial fulfilment of the requirements for the award of the **bachelor of Master of Technology** in **COMPUTER SCIENCE ENGINERRING** is a bonafide record of the work carried out under my guidance and supervision at Amrita School of Engineering, Chennai.

Signature

Vigneshwaran Muralidaran

CSE Dept professor

This project report was evaluated by us on ……………...

INTERNAL EXAMINER EXTERNAL EXAMINER

ACKNOWLEDGEMENT:

I offer my sincere pranams at the lotus feet of Universal guru, MATA AMRITANANDAMAYI DEVI who blessed me with her grace to make this a successful project.

I express my deep sense of gratitude to Dr. P Shankar, Principal, Amrita School of Engineering, Chennai for his kind support. I would like to extend my gratitude and heartfelt thanks to Mr. vigneshwaran Muralidaran CSE Department and Dr. A Manikandan, , , for his constant help, suggestions and inspiring guidance. I am grateful to my guide, Department of EEE , for his invaluable support and guidance during the course of the major project work. I am also indebted to Dr. Prasanna Kumar, Chairperson of Department of AI, ASE, Chennai for his guidance.

I specially thank our director Sri I B Manikandan who supported us in every possible manner.

INDEX:

Topic **Page. No**

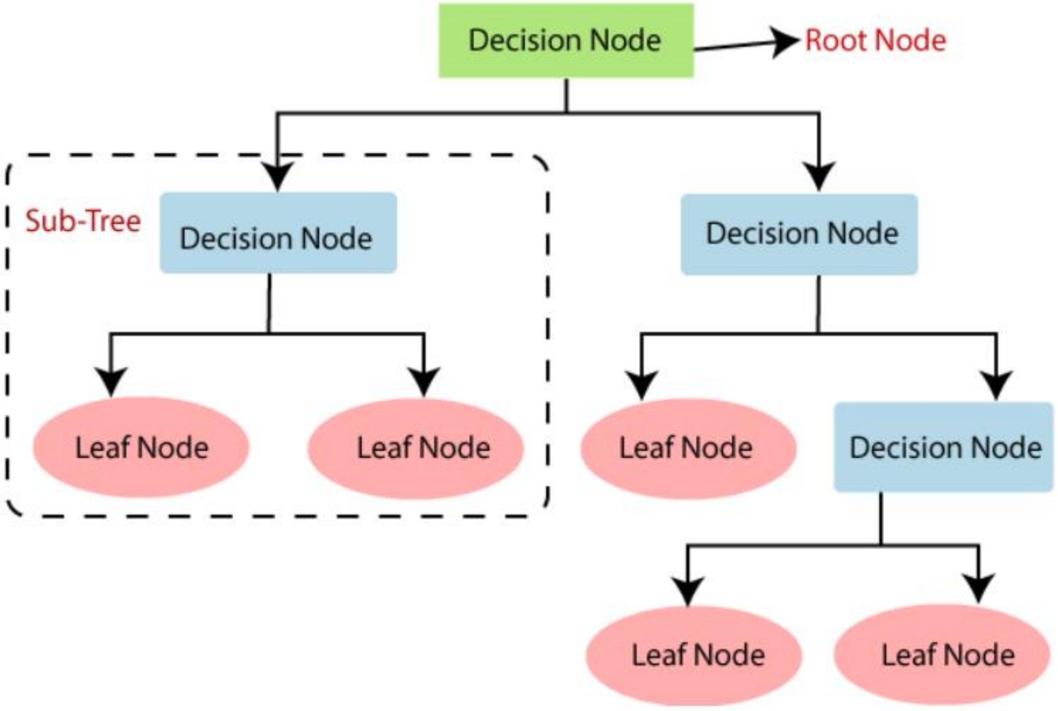
1. ABSTRACT:

One of the most difficult challenges that the meteorological department tackles is effectively forecasting weather. These forecasts are significant because they affect daily living as well as the economics of a state or even a nation. Weather forecasting is particularly vital since it is the first line of defence against natural disasters, which can mean the difference between life and death. They also aid in reducing resource loss and limiting the mitigation procedures that must be implemented following a natural disaster. Machine learning algorithms learn complicated mappings from input to output solely through samples and with little resources. Because of the presence of dynamic behaviour in the atmosphere, precise prediction of weather conditions is a difficult undertaking. This project work focuses on using the Random Forest Algorithm (RFA)to predict the weather conditions based on chosen datasets like humidity ,wind speed , temperature, etc., This random forest algorithm helps us to evaluate with high performance and accuracy besides with lower variance , in contrast to Decision tree Algorithm even though it sets the base for RFA.

2. INTRODUCTION:

2.1 Decision Trees:

A random forest algorithm's building components are decision trees. A decision tree is a decision-making technique with a tree-like structure. It is a supervised learning technique that may be used to solve classification and regression problems.



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 sub-branches. This sequence is repeated until a leaf node is reached. The leaf node cannot be further separated. The decision tree's nodes indicate attributes that are used to forecast the outcome.

It is a graphical representation of all possible solutions to a problem/decision given certain parameters. It is named a decision tree because, like a tree, it begins with the root node and then develops on subsequent branches to form a tree-like structure.

In a decision tree, for predicting the class of the given dataset, the algorithm starts from the root node of the tree. This algorithm compares the values of root attribute with the record (real dataset) attribute and, based on the comparison, follows the branch and jumps to the next node.

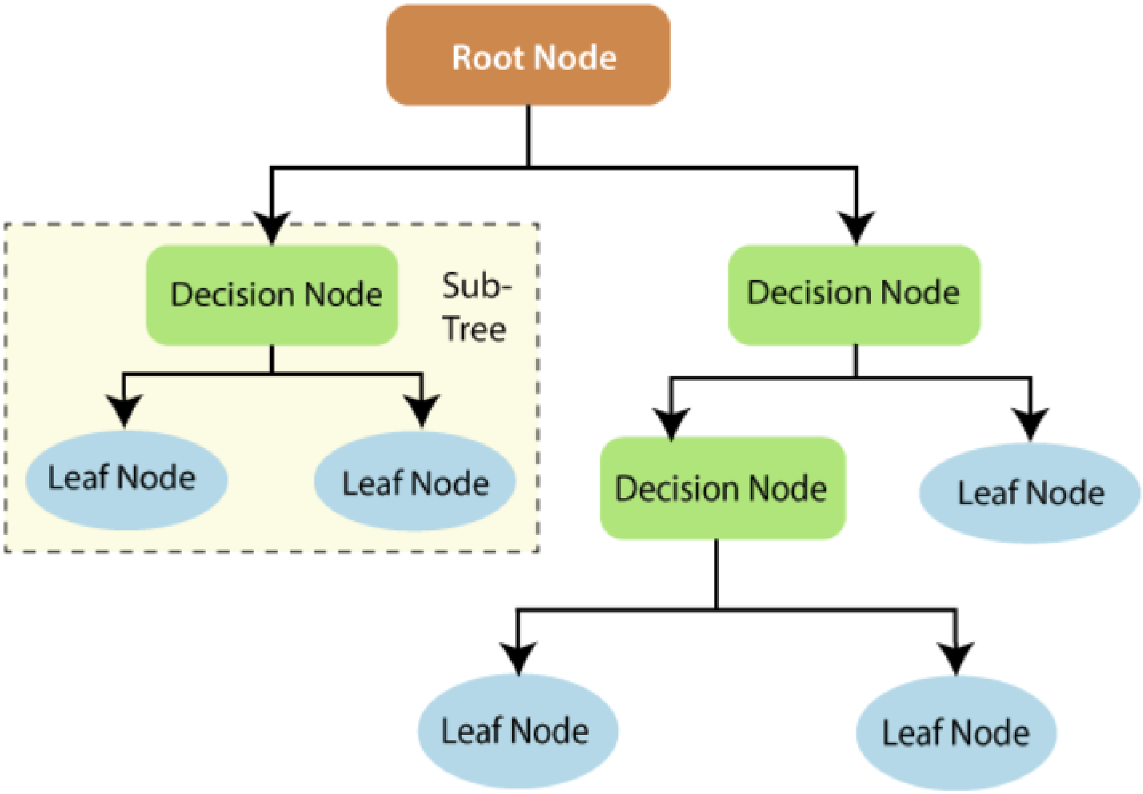
For the next node, the algorithm again compares the attribute value with the other sub-nodes and move further. It continues the process until it reaches the leaf node of the tree.

2.2 Random Forest:

The Random forest algorithm is a supervised learning algorithm which builds multiple decision trees and merges them together to get a more accurate and stable prediction. The "forest" it creates is an ensemble of decision trees, which are often trained using the "bagging" method. The bagging method is based on the idea that combining learning models improves the final output.

Features of a Random Forest Algorithm:

- It outperforms the decision tree algorithm in terms of accuracy.
- It gives an efficient method of dealing with missing data.
- It is capable of producing a reasonable prediction without the use of hyper-parameter tuning.
- It eliminates the problem of overfitting in decision trees.
- At the node's splitting point in every random forest tree, a subset of features is chosen at random.



The algorithm establishes the outcome based on the predictions of the decision trees. It predicts by taking the average or mean of the output from various trees. Increasing the number of trees increases the precision of the outcome.

2.2.1 Bagging

Bagging (bootstraping + aggregating) is the usage of an ensemble of models in which each model uses a bootstrapped data set and the predictions of the models are aggregated.

Bootstraping:

Each tree in a random forest learns from a random selection of data points during training. The samples are drawn utilising replacement, which implies that certain samples will be used several times in a single tree. The theory is that by training each tree on diverse samples, even if each tree has a high variation with regard to a specific set of training data, the overall variance of the forest will be smaller, but not at the expense of increasing the bias.

Aggregation:

An overall forecast or estimate is created using random forest methods by aggregating predictions given by individual decision trees.

Forecasts are made at test time by averaging the predictions of each decision tree. Bagging, short for bootstrap aggregating, is the process of training each individual learner on distinct bootstrapped subsets of the data and then averaging the predictions.

3. LITERATURE REVIEW

Predicting weather properly is one of the most difficult tasks the meteorological service undertakes. These forecasts are significant because they have an impact on daily living as well as the economy of a state or even a country. Weather forecasts are especially important since they serve as the first line of defense against natural disasters, which might mean the difference between life and death. They also aid in reducing resource loss and minimizing the mitigation measures that must be implemented in the aftermath of a natural catastrophe.

Weather forecasting has a wide range of effects on society and our everyday lives, ranging from improvement to catastrophe preparedness. Previous weather forecasting and prediction models relied on a complex set of mathematical instruments that were insufficient to achieve a high categorization rate. Using machine learning techniques, Recent Researches present new unique approaches for estimating monthly rainfall in this study. Weather predictions are generated by gathering quantitative data about the current state of the atmosphere. Machine learning algorithms learn complicated mappings from input to output solely through samples and with little resources. Because of the presence of dynamic behaviour in the atmosphere, precise weather forecasting is a difficult undertaking. Weather conditions from the previous year should be utilized to forecast future weather conditions. There is a very good likelihood that it will fall inside the range of an adjacent fortnight from the previous year. Random forest algorithm and linear regressions are used to examine characteristics such as wind, temperature, and humidity for weather forecasting systems. The suggested model's weather forecasting is based on historical data. As a result, this forecast will be quite accurate. The model's performance is more accurate than traditional medical analysis since it employs a higher-quality merged picture.

Some other researchers says that Weather forecasting is the application of science and technology to anticipate the weather in a specific location. It is one of the world's most challenging problems. The goal of those researches is to use predictive analysis to forecast the weather. As a result, prior to applying, a thorough examination of various data mining processes is required. Those studies provide a classifier strategy for weather prediction and demonstrates how Naive Bayes and the Chi square algorithm may be used for classification. That system is a web application with a graphical user interface that works well. The user will enter information such as the current weather forecast, temperature, humidity, and wind speed. The system will use this parameter to forecast weather after comparing the input data to the data in the database. As a result, two basic operations will be performed: classification (training) and prediction (testing).

Researchers also look into how data mining techniques may be used to forecast qualities like maximum and minimum temperatures. Some of the researches was based on using Decision Tree algorithms and meteorological data collected from several cities between 2010 and 2017. Complex meteorological phenomena make it difficult to anticipate the weather. Many elements of weather events, such as maximum temperature, lowest temperature, humidity, and wind speed, are hard to count and quantify. They have used the Decision Tree Algorithm to delete unwanted data from available datasets. In general, maximum and minimum temperatures are the most important factors in weather forecasting. They anticipate a full cold, full hot, or full snowfall based on the proportion of these characteristics. That work develops a decision tree-based model to anticipate meteorological events such as extreme cold, extreme heat, and snowfall, which can be life-saving information. On a final note, Decision tree and Random Forest Algorithms have made deeper foot prints through years on manifesting mind blowing results on predicting weather.

4. METHODOLOGY

Random Forest operates in two stages: the first is to generate the random forest by mixing N decision trees, and the second is to make predictions for each tree generated in the first phase.

The Working process can be explained in the below steps and diagram:

Step 1: Choose K data points at random from the training set.

Step 2: Build the decision trees linked with the data points you've chosen (Subsets).

Step 3: Decide on the number N for the number of decision trees you wish to create.

Step 4: Repeat Steps 1 & 2.

Step 5: Find the forecasts of each decision tree for new data points and assign the new data points to the category with the most votes.

4.1 Implementation of Random Forest Algorithm in Python:

4.1.1 Data cleaning and Data Pre-processing

Data pre-processing is required tasks for cleaning the data and making it suitable for a machine learning model which also increases the accuracy and efficiency of a machine learning model. The process of converting raw data into a comprehensible format is known as data pre-processing. Before using machine learning or data mining methods, the data quality should be evaluated.

The process of finding, eliminating, and/or replacing inconsistent or erroneous information from a database is known as data cleansing. This technique assures that processed data is of excellent quality and reduces the possibility of incorrect or misleading findings. It removes major errors and inconsistencies that are inevitable when multiple sources of data are being pulled into one dataset.

4.1.2 Fitting the Random forest algorithm to the Training set

Model fitting is the essence of machine learning. It is an automatic process of fitting models to data and analysing the accuracy of the fit. It ensures your machine learning models have the individual parameters that are best suited to solve our problem with high accuracy.

If the differences between the observed values and the model's predicted values are minimal and unbiased, the model fits the data well. A model fits the data well if the differences between the observed values and the model's predicted values are small and unbiased.

4.1.3 Predicting the test result

After our model is fitted to the training set, we can predict the test result. By checking the prediction vector and test set real vector, we can determine the incorrect predictions done by the classifier.

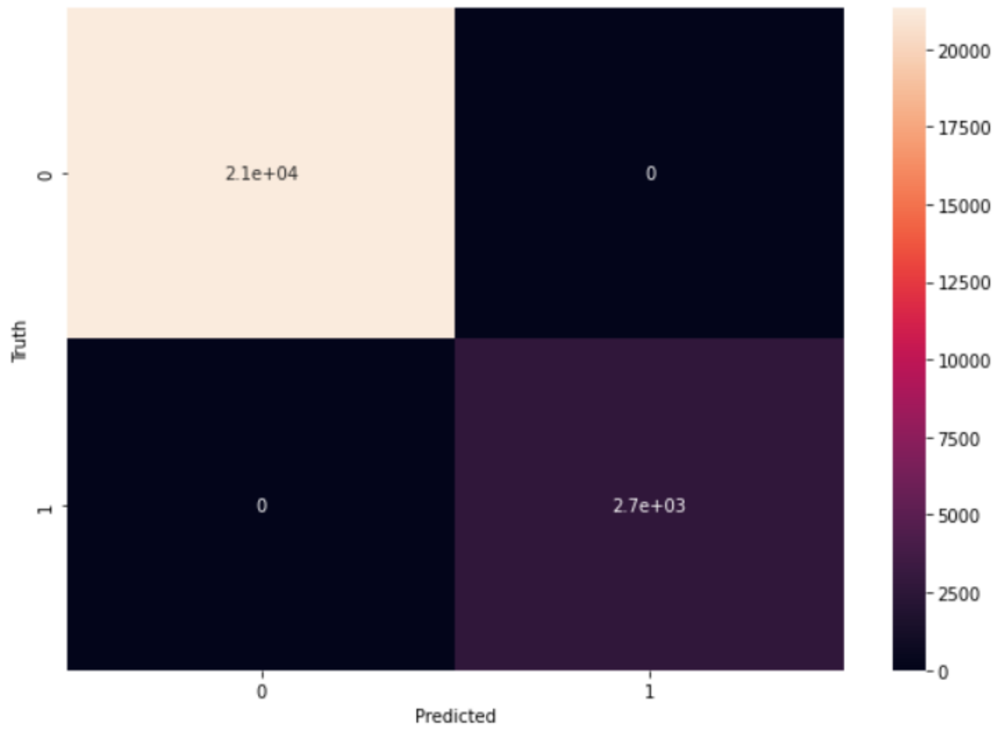
4.1.4 Matrix Validation

We validate the algorithm using confusion matrix. A Confusion matrix is a N x N matrix that is used to assess the effectiveness of a classification model, where N is the number of target classes.

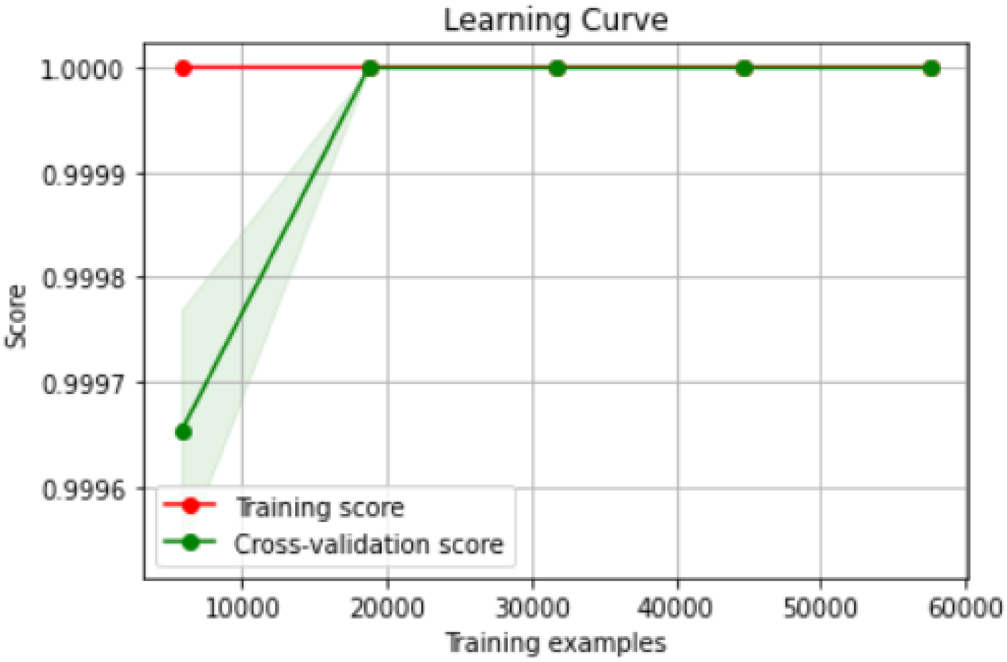
The matrix compares the actual goal values to the machine learning model's predictions. We are validating the model for verifying the accuracy score i.e. how far our model is predicting accurately.

5. DISCUSSION AND RESULT

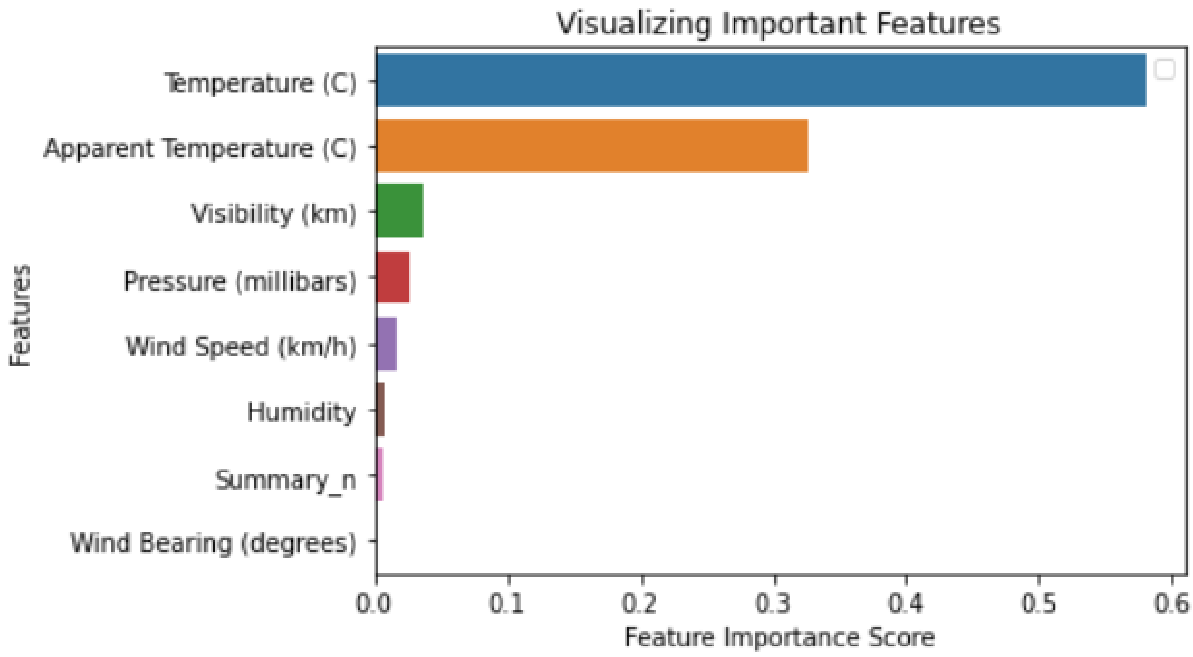
The dataset being using in the model consists of the data over a span of 10 years (2006-2016). It comprises of the precipitation type, temperature, humidity, wind speed and bearing, visibility and the pressure on the particular days. This data is being used to predict weather conditions (rainy, snowy or null) on a day given with the necessary inputs, using the random forest algorithm.



From the resulting graph, we interpret that the true positive and true negative values are 2.1e+0.4 and 2.7e+0.3 respectively, and the false positives and false negatives are both 0.



The learning curve shows us that the training score and the cross-validation score are around the maximum.



The above graph shows which features were of more importance and took more priority in predicting the weather conditions. We can see that temperature played the major role and the wind bearing was the least participant in the prediction process

6. CONCLUSION

To harness the potential of massive amounts of data, Random Forest can be used to forecast dependent variables such as fog and rain.

A Random Forest -equipped software can bring artificial intelligence to a machine. Such software can be used by hikers, mountaineers, and drivers to help them make decisions, because there are many locations where telemetric data does not exist, and if it does exist, it is too location sensitive to generalise to a large geographical area. It is fairly cost effective in this information age to equip a machine with sensors and artificial intelligence software such that the machine demonstrates intelligence.

7. REFERENCES

https://machinelearningmastery.com/random-forest-for-time-series-forecasting/

https://www.youtube.com/watch?v=9yl6-HEY7_s&list=PLeo1K3hjS3uvCeTYTeyfe0-rN5r8zn9rw&ab_channel=codebasics

https://acadpubl.eu/jsi/2018-118-20/articles/20a/34.pdf

https://research.ijcaonline.org/volume76/number2/pxc3890620.pdf

https://sebastianraschka.com/faq/docs/bagging-boosting-rf.html

https://www.javatpoint.com/machine-learning-random-forest-algorithm

https://www.section.io/engineering-education/introduction-to-random-forest-in-machine-learning/#:~:text=A%20random%20forest%20is%20a,to%20predict%20behavior%20and%20outcomes.