VIETNAM GENERAL CONFEDERATION OF LABOR

**TON DUC THANG UNIVERSITY**

**FACULTY OF INFORMATION TECHNOLOGY**



**INTRODUCTION TO MACHINE LEARNING**

**MID-TERM**

*Lecturer*: **Mr. LE ANH CUONG**

*Students*: **HO QUOC CUONG – 519H0146**

**HUYNH ANH KHOI – 518H0376**

*Class:* **19H50204**

*Course:* **23**

**HO CHI MINH CITY, MARCH 2023**

VIETNAM GENERAL CONFEDERATION OF LABOR

**TON DUC THANG UNIVERSITY**

**FACULTY OF INFORMATION TECHNOLOGY**



**INTRODUCTION TO MACHINE LEARNING**

**MID-TERM**

*Lecturer*: **Mr. LE ANH CUONG**

*Students*: **HO QUOC CUONG – 519H0146**

**HUYNH ANH KHOI – 518H0376**

*Class:* **19H50204**

*Course:* **23**

**HO CHI MINH CITY, MARCH 2023**

**THANK YOU**

During my working on this report, we have been confronted with many difficulties but with the help of Mr. Bui Thanh Hung, we could widen more knowledge to make this report better. Therefore, we sincerely thank you for giving me such useful advice and spending your time instructing me.

Though being made by my most carefulness, the report cannot avoid some mistakes. If you find any errors in our report, please give me some advice so that i can be able to make a better report next time.

Once again, sincerely thank you!

**PROJECT COMPLETED AT TON DUC THANG UNIVERSITY**

I hereby declare that this is my own project and is guided by Mr. Bui Thanh Hung; The content research and results contained herein are central and have not been published in any form before. The data in the tables for analysis, comments and evaluation are collected by the main author from different sources, which are clearly stated in the reference section.

In addition, the project also uses some comments, assessments as well as data of other authors, other organizations with citations and annotated sources.

**If something wrong happens, I’ll take full responsibility for the content of my project.** Ton Duc Thang University is not related to the infringing rights, the copyrights that I give during the implementation process (if any).

*Ho Chi Minh City, 19th March, 2023*

*Author*

*Ho Quoc Cuong*

*Huynh Anh Khoi*

**EVALUATION OF INSTRUCTING LECTURER**

**The confirmation of the instructor**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Ho Chi Minh city, day month 3 year 2023

(Sign and write full name)

**The assessment of grading teacher**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Ho Chi Minh city, day month 3 year 2023

(Sign and write full name)

**TABLE OF CONTENT**

[**1 – EXERCISE 1 7**](#_heading=h.gjdgxs).

[**1.1 – DATA 7.**](#_heading=h.30j0zll)

[**1.2 – PRE-PROCESSING**](#_heading=h.30j0zll) **7.**

[**1.3 – CORRELATION METHOD 8**](#_heading=h.30j0zll).

[**1.4 – MODEL DECISION TREE 7.**](#_heading=h.30j0zll)

[**2 – EXERCISE 2 8.**](#_heading=h.1fob9te)

[**2.1 – DATASET 8**](#_heading=h.30j0zll).

[**2.2 – DECISION TREE 9**](#_heading=h.3znysh7).

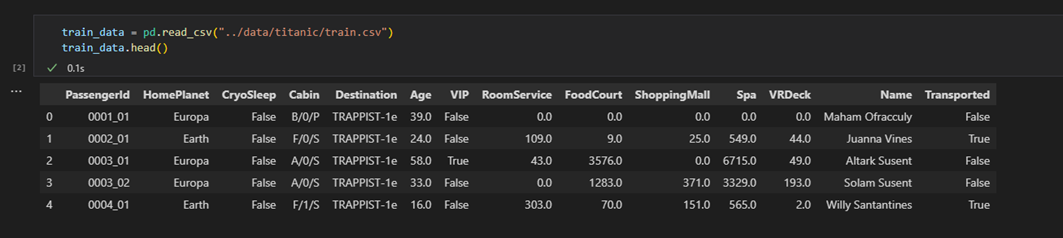
[**2.2 – RANDOM FOREST 9**.](#_heading=h.3znysh7)

[**2.2 – ANALYSIS 9**.](#_heading=h.3znysh7)

1. **Exercise 1:**
   1. Data:

The titanic dataset is a dataset containing passenger information recovered from the ship's damaged computer system. Based on this data set can we predict whether a passenger was transported to another dimension during the collision of the Titanic Spaceship?

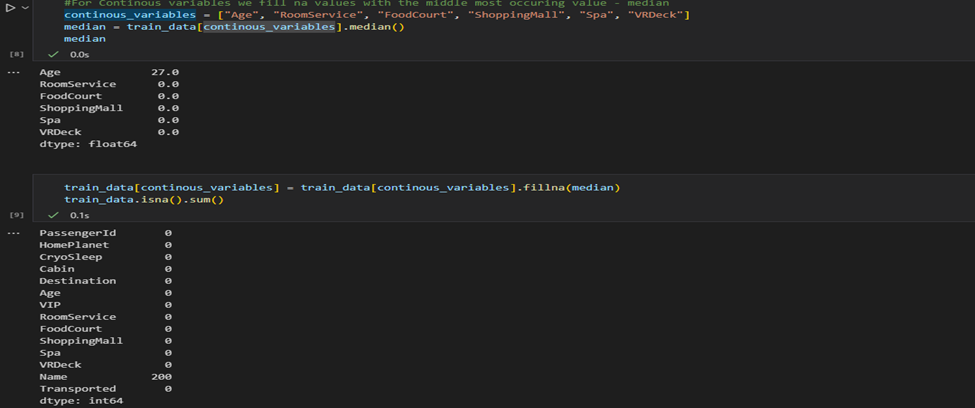
Tabular data has the following properties:



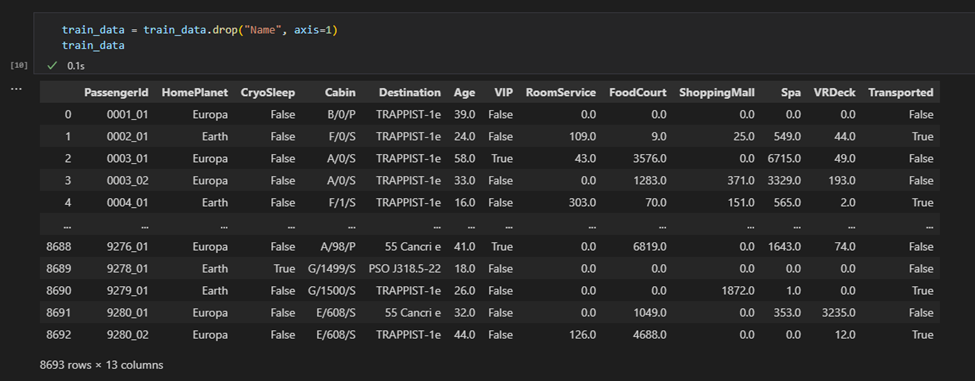
* 1. Pre-processing:

The initial data may contain null or null values that affect the model training for prediction.

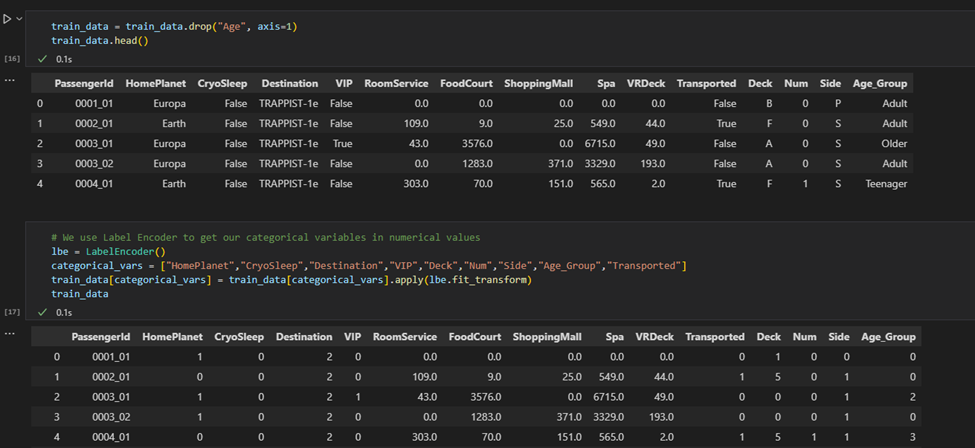
Calculate the average value of each attribute to fill in empty cells.



Remove attributes that are not valid for training, such as the "Name" attribute.



The model only learns the attributes with numeric values, so the attributes must be converted to digits, here we use the Label encoder method to encode and convert the values.



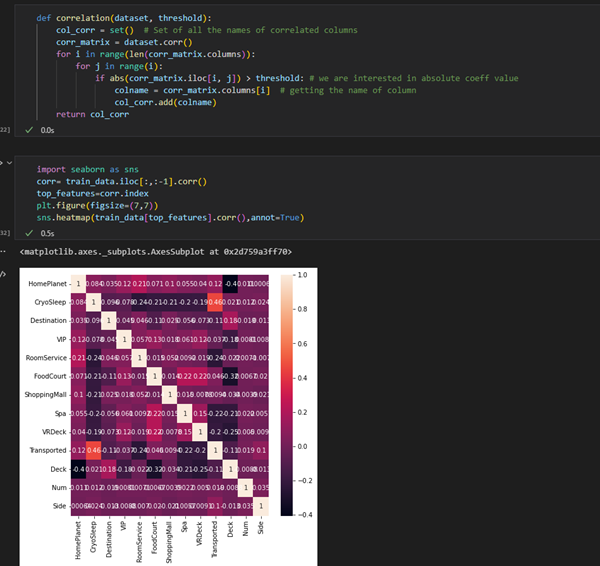
Example : True -> 1; False -> 0

* 1. Correlation method:

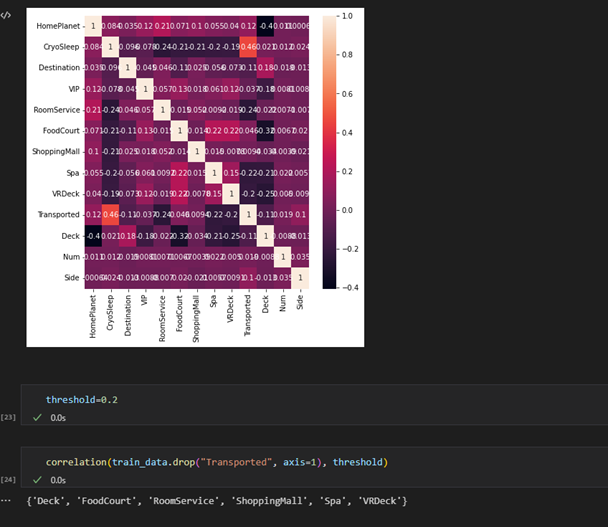
Correlation is a standardized statistical measure that expresses the degree to which two variables are linearly related to each other (that is, they change together at a constant rate).

In machine learning, correlation can be used to understand the relationship between different features (or attributes) in a dataset. There are several ways to calculate correlation between variables in Python, including using libraries such as NumPy, SciPy, and pandas.

The strength and directional association of the relationship between two variables is determined by the correlation, and it ranges from -1 to +1. Similar to covariance, a positive value indicates that both variables move in the same direction while a negative value tells us that they move in opposite directions.



After a relative analysis between features, we find features that are valuable for training such as:



* 1. Model Decision tree:

A decision tree (DT) is a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of the target variable by learning simple decision rules inferred from data features. A tree can be thought of as a partial constant approximation.

In other words, a Decision Tree consists of decision nodes and leaf nodes. The decision nodes are used to partition the data based on a certain criterion, such as the Gini index or entropy. At each decision node, the algorithm selects the feature with the highest information gain or the best split that separates the data into different regions with high purity. The leaf nodes are the final output of the tree and represent the prediction for a given input.

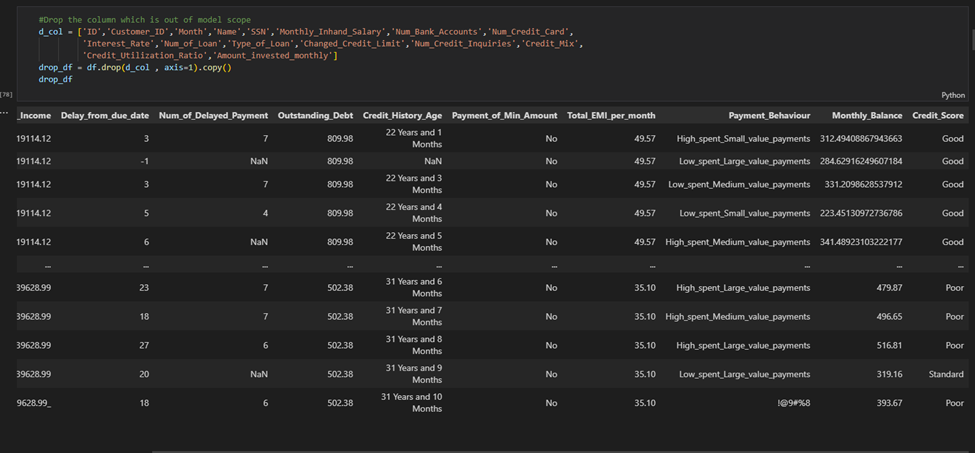
In Python, the scikit-learn library provides a DecisionTreeClassifier class that can be used to construct a decision tree model. The class provides several hyperparameters that can be tuned to improve the performance of the model, such as the maximum depth of the tree, minimum samples for a node to be split, and the split criterion.

A decision tree is a white box type of ML algorithm. It shares internal decision-making logic, which is not available in the black box type of algorithms such as with a neural network. Its training time is faster compared to the neural network algorithm.

The time complexity of decision trees is a function of the number of records and attributes in the given data. The decision tree is a distribution-free or non-parametric method which does not depend upon probability distribution assumptions. Decision trees can handle high-dimensional data with good accuracy.

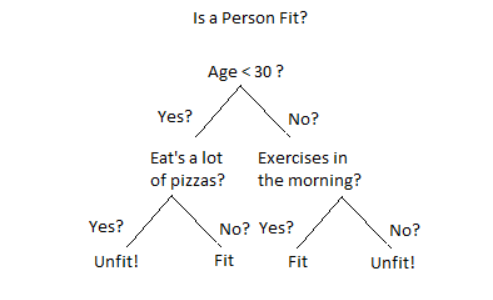
1. **Exercise 2:**
   1. Data set:

This is a data set containing customer credit information to predict credit scores. Output value is “Score Credit” including Good, Standard, Poor



* 1. Decision tree model:

Decision Trees are a type of Supervised Machine Learning (that is you explain what the input is and what the corresponding output is in the training data) where the data is continuously split according to a certain parameter. The tree can be explained by two entities, namely decision nodes and leaves. The leaves are the decisions or the final outcomes. And the decision nodes are where the data is split. Decision Trees modified



An example of a decision tree can be explained using the above binary tree. Let’s say you want to predict whether a person is fit given their information like age, eating habit, and physical activity, etc. The decision nodes here are questions like ‘What’s the age?’, ‘Does he exercise?’, ‘Does he eat a lot of pizzas’? And the leaves, which are outcomes like either ‘fit’, or ‘unfit’. In this case this was a binary classification problem (a yes no type problem). There are two main types of Decision Trees:

1. Classification trees (Yes/No types)

What we’ve seen above is an example of a classification tree, where the outcome was a variable like ‘fit’ or ‘unfit’. Here the decision variable is Categorical.

ID3 is the most popular algorithm in building decision tree for classification in particular, where all attributes are in category

ID3 (Iterative Dichotomiser 3) Algorithm was jointly developed by Quinlan in AI and Breiman, Friedman, Olsen and Stone in statistics. ID3 is a simple but successful learning algorithm in many fields. ID3 is a good algorithm because of its learned representation, its approach to managing complexity, its heuristic for selecting candidate concepts, and its potential for handling noisy data.

ID3 represents concepts in the form of decision trees. This representation allows us to determine the classification of an object by examining its values on certain attributes.

Thus, the task of the ID3 algorithm is to learn a decision tree from a set of training examples, also known as training data.

Input: A set of examples. Each example includes attributes that describe a situation, or an object, and its categorical value.

Output: The decision tree is capable of properly classifying the examples in the training data set, and hopefully, the correct classification for the examples that have not yet been encountered in the future.

The ID3 algorithm to build a decision tree is presented as follows:

Repeat:

1. Choose A <= “best” decision attribute for the next node

2. Assign A to be the decision attribute for node

3. For each value of A, create a new sub-branch of node

4. Classification of training samples for nodes

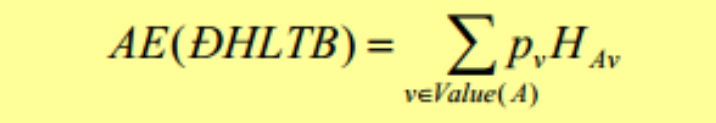
5. If the training samples are completely classified then STOP,

Otherwise, iterate with the new nodes.

The best property here is the one with the lowest mean entropy according to the resulting attribute with the Entropy calculated as follows:

* Let S be the set of training samples
* Let p be the proportion of positive samples in S
* We have H ≡ – p.log2p – (1 – p).log2(1 – p)

The mean entropy of an attribute is the proportional average of the entropy of the branches:

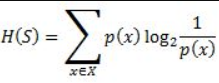


1. Regression trees (Continuous data types)

Here the decision or the outcome variable is Continuous, e.g. a number like 123. Now that we know what a Decision Tree is, we’ll see how it works internally. There are many algorithms out there which construct Decision Trees, but one of the best is called the ID3 Algorithm. ID3 Stands for Iterative Dichotomiser 3. Before discussing the ID3 algorithm, we’ll go through a few definitions.

* Entropy:

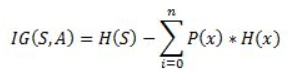
Entropy, also called Shannon Entropy is denoted by H(S) for a finite set S, is the measure of the amount of uncertainty or randomness in data.



Intuitively, it tells us about the predictability of a certain event. Example, consider a coin toss whose probability of heads is 0.5 and probability of tails is 0.5. Here the entropy is the highest possible, since there’s no way of determining what the outcome might be. Alternatively, consider a coin which has heads on both sides, the entropy of such an event can be predicted perfectly since we know beforehand that it’ll always be heads. In other words, this event has no randomness hence its entropy is zero. In particular, lower values imply less uncertainty while higher values imply high uncertainty.

* Information Gain:

Information gain is also called Kullback-Leibler divergence denoted by IG(S,A) for a set S is the effective change in entropy after deciding on a particular attribute A. It measures the relative change in entropy with respect to the independent variables. 

Alternatively, 

where IG(S, A) is the information gained by applying feature A. H(S) is the Entropy of the entire set, while the second term calculates the Entropy after applying the feature A, where P(x) is the probability of event x.

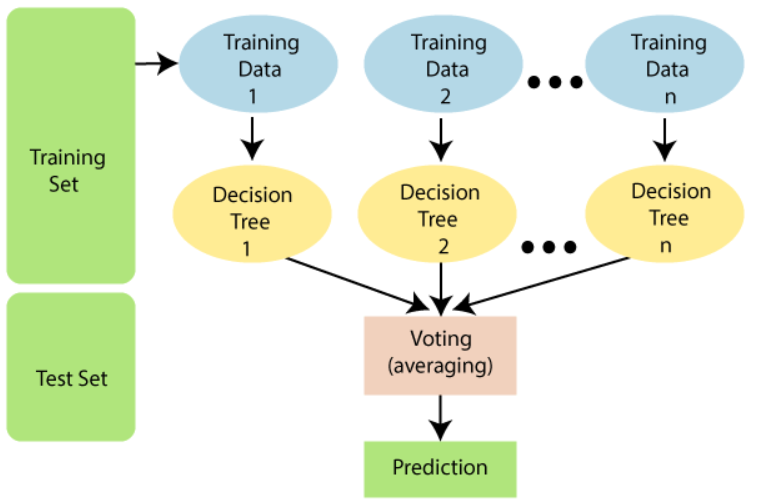
* 1. Random Forest:

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

The below diagram explains the working of the Random Forest algorithm:



Assumptions for Random Forest:

Since the random forest combines multiple trees to predict the class of the dataset, it is possible that some decision trees may predict the correct output, while others may not. But together, all the trees predict the correct output. Therefore, below are two assumptions for a better Random forest classifier:

* There should be some actual values in the feature variable of the dataset so that the classifier can predict accurate results rather than a guessed result.
* The predictions from each tree must have very low correlations.

Why use Random Forest?

Below are some points that explain why we should use the Random Forest algorithm:

* It takes less training time as compared to other algorithms.
* It predicts output with high accuracy, even for the large dataset it runs efficiently.
* It can also maintain accuracy when a large proportion of data is missing.

Random Forest works in two-phase first is to create the random forest by combining N decision trees , and second is to make predictions for each tree created in the first phase.

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

Step-1: Select random K data points from the training set.

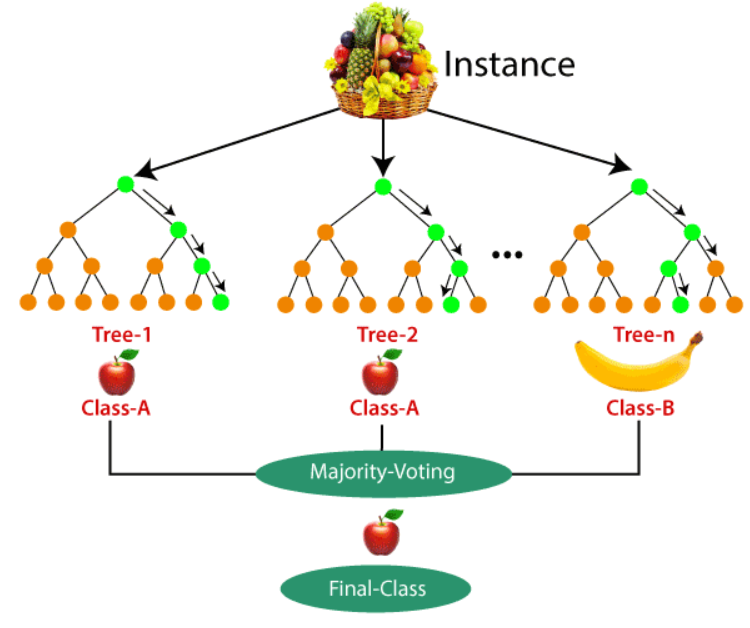
Step-2: Build the decision trees associated with the selected data points (Subsets).

Step-3: Choose the number N for decision trees that you want to build.

Step-4: Repeat Step 1 & 2.

Step-5: For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes.

Example: Suppose there is a dataset that contains multiple fruit images. So, this dataset is given to the Random forest classifier. The dataset is divided into subsets and given to each decision tree. During the training phase, each decision tree produces a prediction result, and when a new data point occurs, then based on the majority of results, the Random Forest classifier predicts the final decision. Consider the below image:



Applications of Random Forest:

There are mainly four sectors where Random forest mostly used:

* Banking: Banking sector mostly uses this algorithm for the identification of loan risk.
* Medicine: With the help of this algorithm, disease trends and risks of the disease can be identified.
* Land Use: We can identify the areas of similar land use by this algorithm.
* Marketing: Marketing trends can be identified using this algorithm.

Advantages of Random Forest:

* Random Forest is capable of performing both Classification and Regression tasks.
* It is capable of handling large datasets with high dimensionality.
* It enhances the accuracy of the model and prevents the overfitting issue.

Disadvantages of Random Forest

Although random forest can be used for both classification and regression tasks, it is not more suitable for Regression tasks.

* 1. Analysis:

After using models such as Naïve Bayes, KNN, Decision Tree, Random Forest, Linear Regression, we find that the Decision Tree model has the most positive results with the most correctness.

We using this data for the Linear Regression model is only for predicting error < 0.5 units: for example y\_test = 1 -> y\_pedict = 1.233 ; y\_test = 2 -> y\_pedict = 1.77

