

**PRODUCT DELIVERY ESTIMATION**

Submitted by

Vishvam Ganesh(231801189)

Soorya G(231801169)

Siva Murugan(231801167)

Shyam Prasad S(231801166)

AI23331 Fundamentals of Machine Learning Department of Artificial Intelligence and Data Science

**Rajalakshmi Engineering College, Thandalam**



BONAFIDE CERTIFICATE

NAME………………………………………………………………………………................................................

ACADEMIC YEAR………………SEMESTER………….BRANCH………………………..

UNIVERSITY REGISTER No.

Certified that this is the bonafide record of work done by the above students in the Mini Project titled " **PRODUCT DELIVERY ESTIMATION**" in the subject AI23331 – FUNDAMENTALS OF

**MACHINE LEARNING during the year 2023 - 2024.**

**Signature of Faculty – in – Charge**

**Submitted for the Practical Examination held on \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Internal Examiner External Examiner**

|  |  |  |
| --- | --- | --- |
|  |  |  |

# **Table of Contents**

* Objective
* Dataset Overview
* Machine Learning Pipeline
* Importing dataset, packages, and libraries
* Data Cleaning and Analysis
* Data Visualization
* Statistical Research
* Data Preparation
* Data Split
* Random Forest Classifier
* Decision Tree Classifier
* Confusion Matrix
* Classification Report
* Accuracy comparison between Random Forest Classifier and
* Conclusion
* Further Improvements
* References

Objective

An international e-commerce company “**ecommerce 365”** into selling of electronic products wants to discover key insights from their customer database.The goal of this project is to understand key factors in predicting delivery time for a product and developing a model to accurately make these predictions. These types of predictive models could be used by companies to develop a better understanding regarding their supply chain, customer response time, and any issues that would slow down the delivery of a product. It is also critical to the business, as there is a good chance of incurring significant losses if the delivery execution is sub-optimal.We have considered the available data set to build our model from the below link

<https://www.kaggle.com/prachi13/customer-analytics>

Also a comparison is made between the performance of Random Forest and Decision Tree classifiers in our analysis.

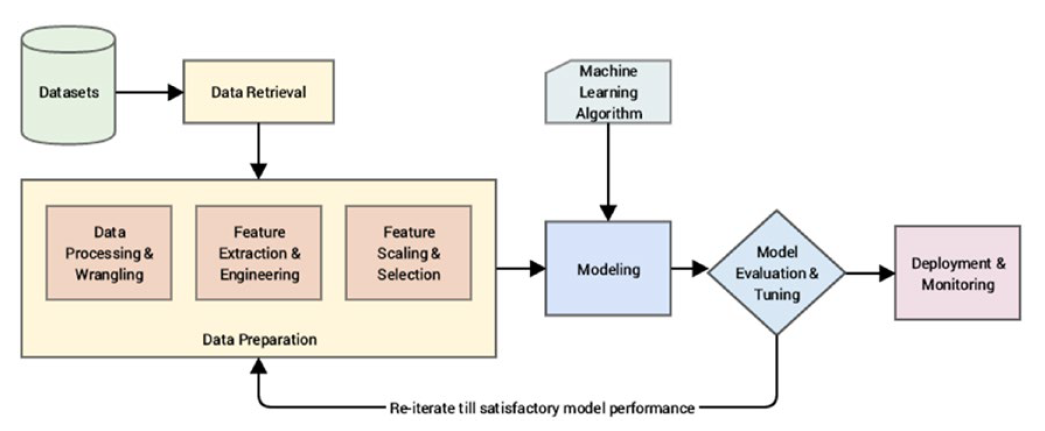
# **Data Set overview**

The dataset used for model building contained 10999 observations of 12 variables. The data contains the following information:

* ID: ID Number of Customers.
* Warehouse block: The Company have big Warehouse, which is divided in to block such as A, B, C, D, E.
* Mode of shipment: The Company Ships the products in multiple way such as Ship, Flight and Road.
* Customer care calls: The number of calls made from enquiry for enquiry of the shipment.
* Customer rating: The company has rated from every customer. One is the lowest (Worst), 5 is the highest (Best).
* Cost of the product: Cost of the Product in US Dollars.
* Prior purchases: The Number of Prior Purchase.
* Product importance: The Company has categorized the product in the various parameter such as low, medium, high.
* Gender: Male and Female.
* Discount offered: Discount offered on that specific product.
* Weight in gms: It is the weight in grams.
* Reached on time: It is the target variable, where 1 Indicates that the product has NOT reached on time and 0 indicates it has reached on time.

**Data Source:** <https://www.kaggle.com/prachi13/customer-analytics>

# **Machine Learning Pipeline**



1. Domain Exploration & Problem Formulation

- Perform detailed domain exploration; understand the relevant business process, common challenges, and beliefs

- Formulate the business problem to solved as an ML problem, plan the modelling of target variable, prepare the plan

2. Data Collection & Data Exploration

- collect data from relevant units, perform data modelling, and prepare the dataset

- Perform a generic exploration to understand common data quality challenges

3. Data Cleaning

- Handle duplicate entries

- handle missing values

- handle outliers / unnatural values

4. Feature Engineering

- Feature Extraction - creating new useful features from existing variables

- Feature Selection

- Data visualization - univariate, bivariate and multivariate

- Statistical research - correlation analysis, chi square test etc.

5. Preprocessing of features

- Encoding, Normalization

- Train test split

6. Apply ML to train a model

- pick the right algorithms based on requirement and train an ML model

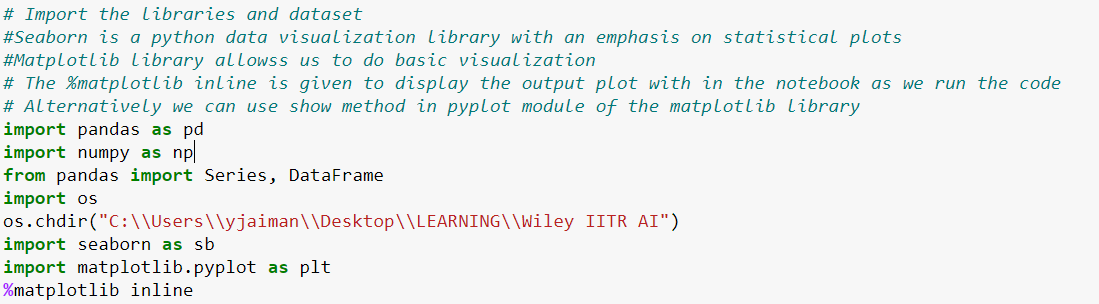
- We used Random Forest and Decision Tree

7. Performance Analysis

- Identify how good the ML model is using appropriate metric

- Detailed performance analysis

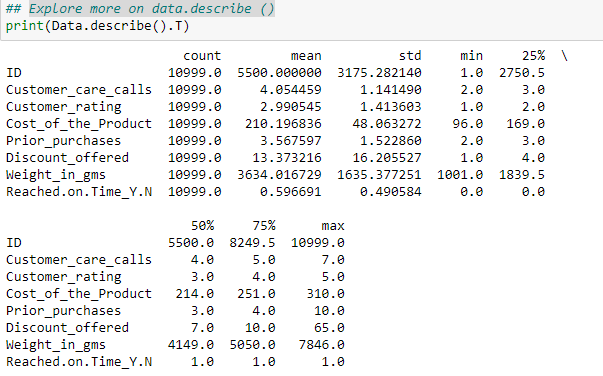
# **Importing dataset, Packages and Libraries**

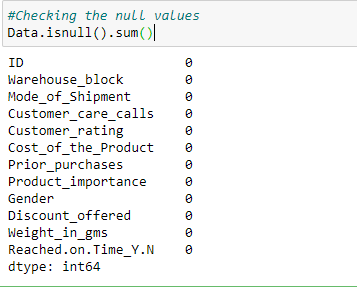




# **Data Analysis & Cleaning**

* How Many observations and variables are there?
* What is the different type of variables?
* What values are contained in the variables?
* Explore more on data. Describe ()
* check for duplicated rows
* Checking the null values





# **Data visualization**

The preliminary analysis of data to discover relationships between measures in the data and to gain an insight on the trends, patterns, and relationships among various entities present in the data set with the help of statistics and visualization tools is called Exploratory Data Analysis (EDA).

Exploratory data analysis is cross - classified in two different ways where each method is either graphical or non-graphical. Then, each method is either univariate, bivariate or multivariate.

### **Univariate Analysis**

Univariate analysis is the simplest form to analyze data. Uni means one and this means that the data has only one kind of variable. The major reason for univariate analysis is to use the data to describe. The analysis will take data, summaries it, and then find some pattern in the data.

**Use case 1**

Observation -

* Warehouse block F holds most items.
* Maximum deliveries are done via Ship
* There are lot of low importance products and medium importance relative to the high importance products.

**Use case 2**

Observation - 40% of the total deliveries are not reached on time

### **Bi-Variate analysis**

Bi means two and variate means variable, so here there are two variables. The analysis is related to cause and the relationship between the two variables.

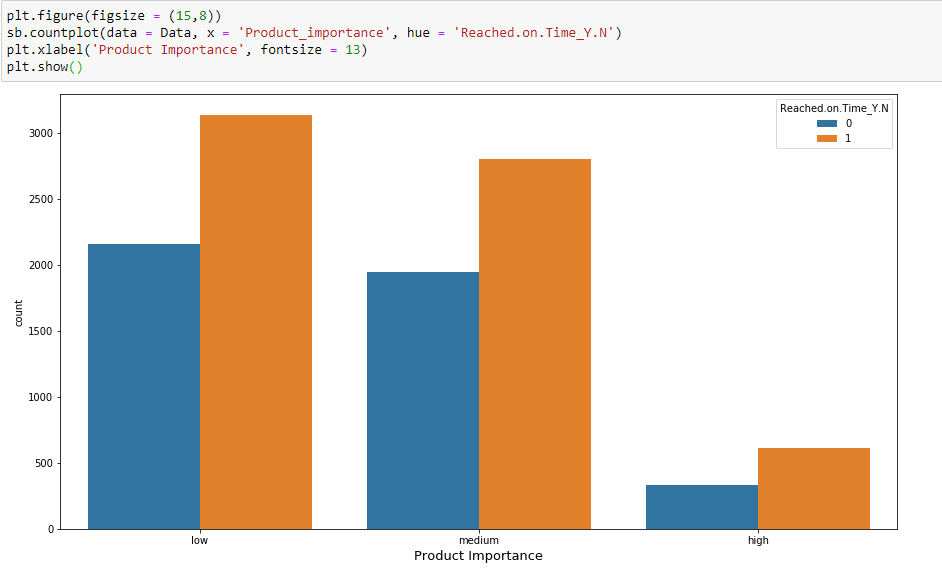
## **Use case 1:** Relation between Product Arrived on Time - Discount Offered



Observation -

* Indicates the product did not reach in time and 0 when it does.
* We can see that on a normal basis a discount is given in both scenarios. between the range of 0 to 10.
* An observation can made that most of the time when deliveries are delayed discounts of more than 20$ are given

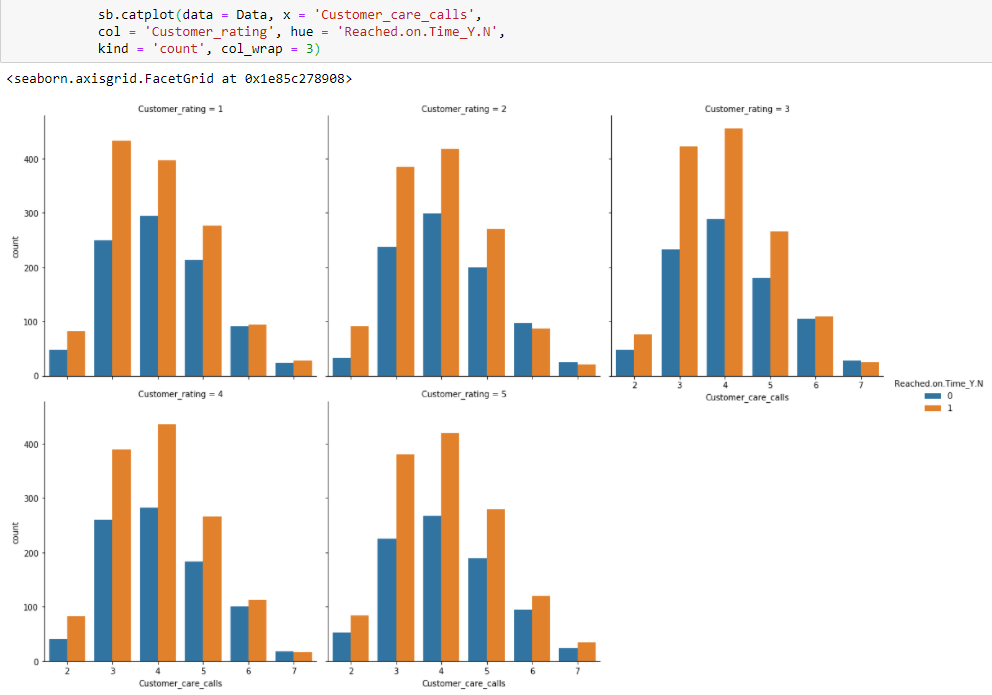
## **Use case 2**: Relation between Product Arrived on Time – Product Importance



Observation-

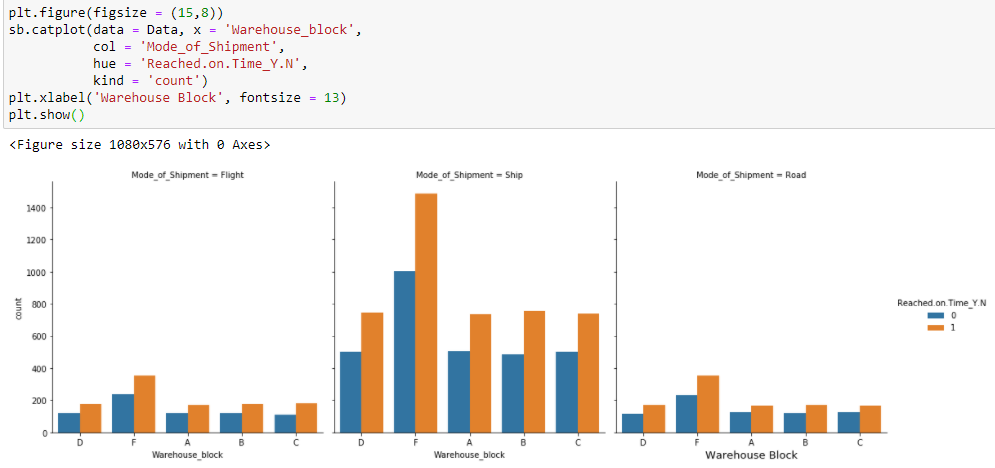
* An observation can be made that regardless of what product importance, the products are delivered late, nonetheless.
* Late delivery of products prompts the customers to leave a bad review and high discounts are needed to be given.
* Higher number of deliveries falls under low product importance.
* Very less number of highly important products delivered.
* It means customers are ordering a greater number of low importance products from this ecommerce group

## **Use case 3**: Relation between Product Arrived on Time – Customer Care calls – customer rating



Observation - Almost every plot looks the same and relevant insight cannot be drawn

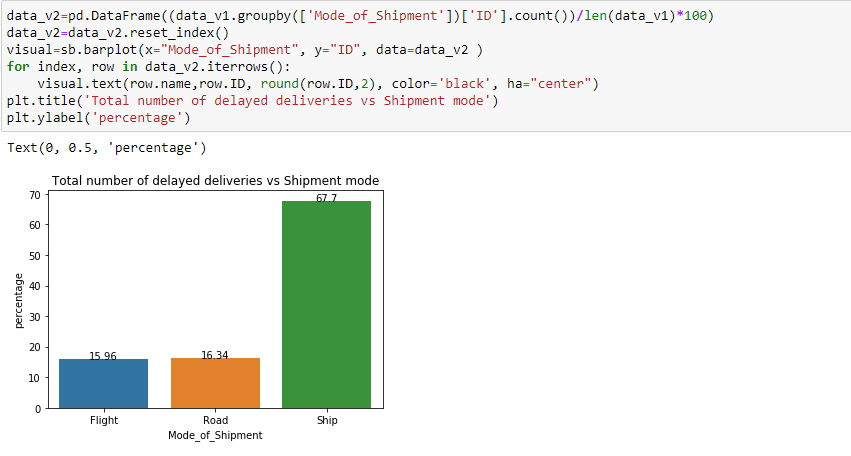
## **Use case 4**: Relation between Product Arrived on Time – Warehouse block – Shipment mode



Observation -

* Since Warehouse F holds the maximum number of products, we can see that there is a delay from it the most.
* Warehousing processing would need a quality assurance/check for better functioning.

## **Use case 5:** Relation between Product Arrived on Time – shipment mode

Observation -

Around 68% of the delayed deliveries are caused when ships are used as a mode of shipments.

So, Alternate options like Flight and Road services might be considered to reduce the delayed deliveries

## **Use case 6:** Relation between Product Arrived on Time – Gender

# 

# Statistical Research

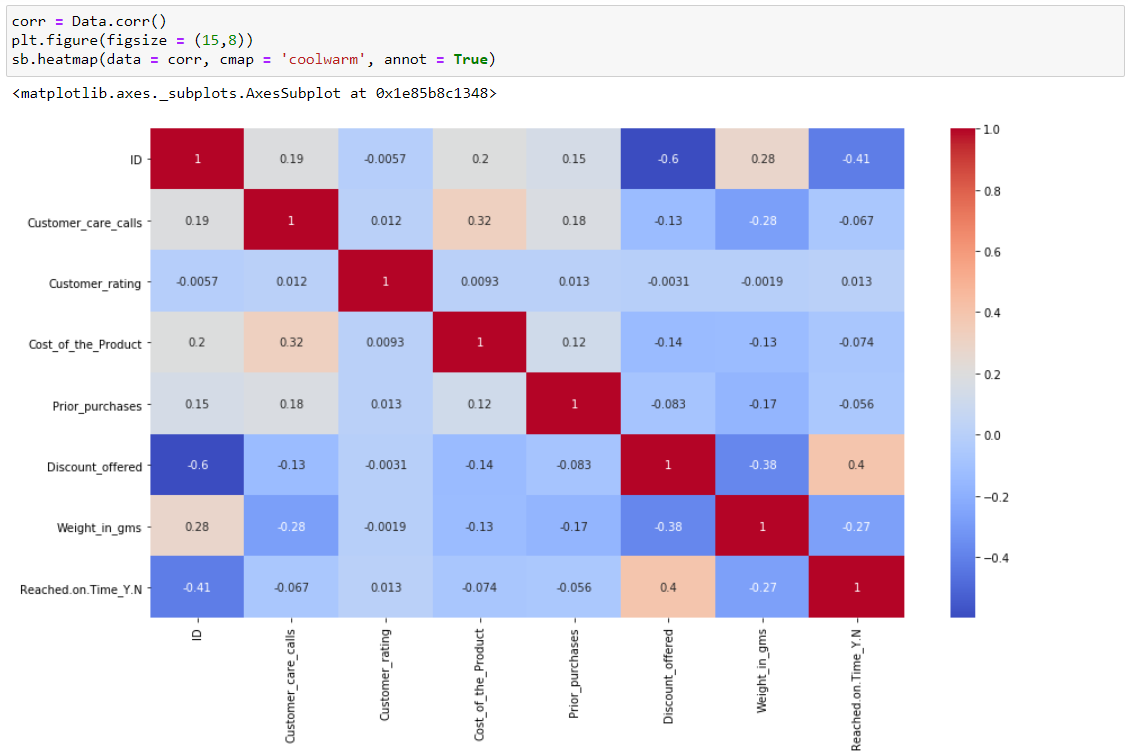
## Correlation between Numerical Values

#### **Understanding Correlation**

Correlation shows the strength of a relationship between two variables and is expressed numerically by the correlation coefficient. The correlation coefficient's values range between -1.0 and 1.0. A perfect positive correlation means that the correlation coefficient is exactly 1. This implies that as one security moves, either up or down, the other security moves in lockstep, in the same direction. A perfect negative correlation (-1.0) means that two assets move in opposite directions, while a zero correlation implies no linear relationship at all.

Calculated correlation (r) between two variables value:

* r<0.20 and values close to zero are not related or indicates a very weak relationship.
* Weak correlation between 0.20-0.39
* Moderate correlation between 0.40-0.59
* A high level of correlation between 0.60-0.79
* If it is 0.80-1.0, it is interpreted as a very high correlation.



Positive correlation can be seen between the following variables:

* Reached on time & Discount offered 0.4
* Customer Care calls & cost of the product 0.32
* Prior Purchases & customer care calls 0.18
* Cost of the product and prior purchase 0.12
* Reached on time and customer rating 0.013
* Customer rating and customer care calls 0.012
* There is a slight positive correlation between 'cost of products' and 'customer care calls'

# **Data Preparation**

**Drop the columns -**

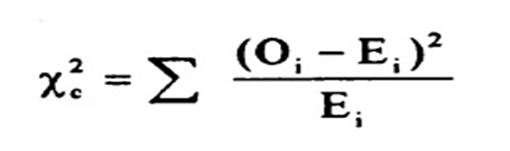
High cardinality refers **to columns with values that are very uncommon or unique**.



## **Chi-square Test of Independence**

The chi-square test is used for determining the association between categorical variables. It is calculated based on the difference between expected frequencies and the observed frequencies in one or more categories of the frequency table. A probability of zero indicates a complete dependency between two categorical variables and a probability of one indicates that two categorical variables are completely independent.

Here, subscript c indicates the degrees of freedom, O indicates observed value, and E indicates expected value.

****



Observation –

From chi square test, we can conclude that three of the independent categorical features are not related to the response variable because the p value is greater than 0.05.

Therefore, features ( warehouse\_block , Mode\_of\_shipment , Gender) can be removed and product importance feature can be included in feature list.

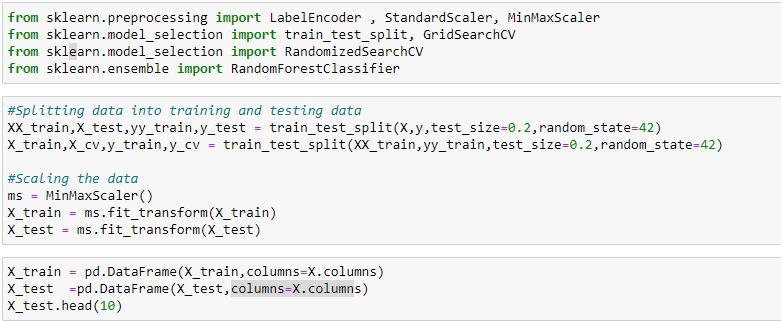
We have dropped below columns -

* ID
* Reached on Time
* Warehouse\_block
* Mode\_of\_Shipment
* Gender



## **Data Split**

We spilt the data in training & test data set in ratio of 80:20 respectively and choose random sampling.



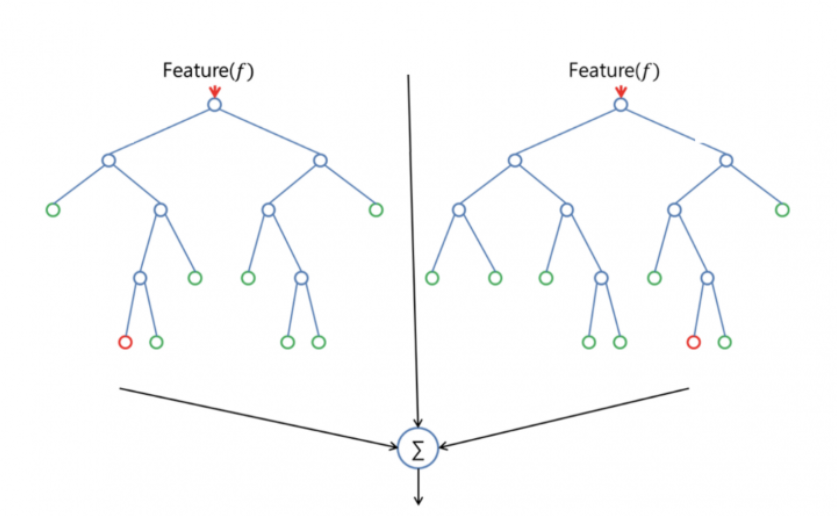
# **Random forest classifier**

Random forest is a supervised learning algorithm. The "forest" it builds, is an ensemble of decision trees, usually trained with the “bagging” method. The general idea of the bagging method is that a combination of learning models increases the overall result

**Put simply: random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction.**

One big advantage of random forest is that it can be used for both classification and regression problems, which form the majority of current machine learning systems.

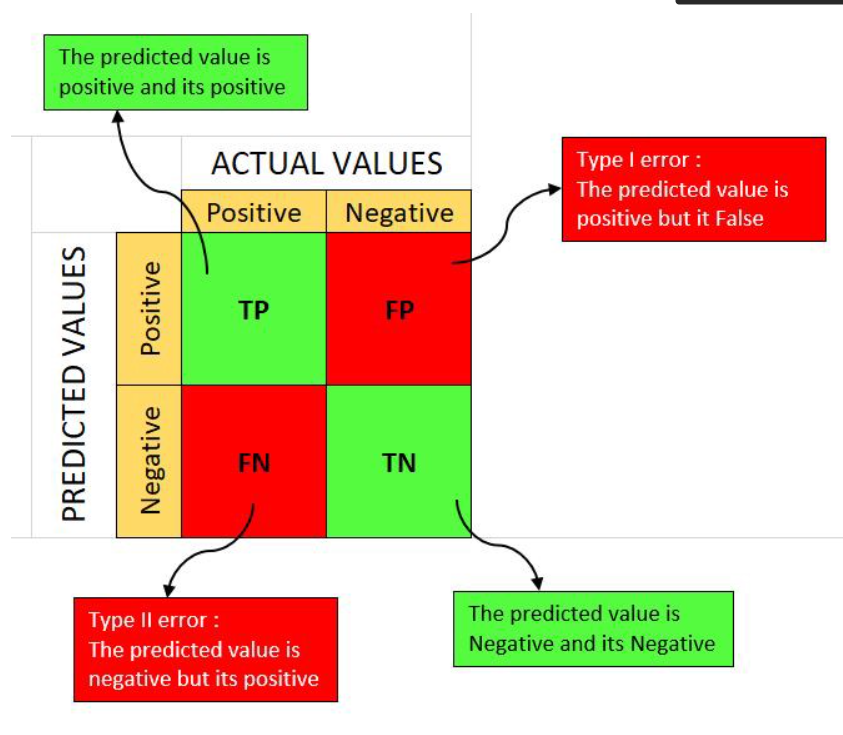
Let's look at random forest in our classification problem –



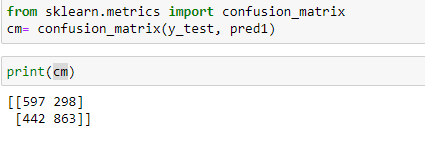


### **Confusion Metrics**

A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known.



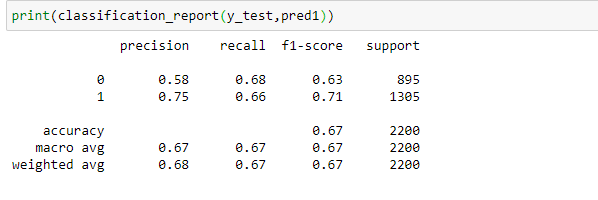
A **good model** is one which has ***high TP and TN rates***, while ***low FP and FN rates***.



**Classification report**

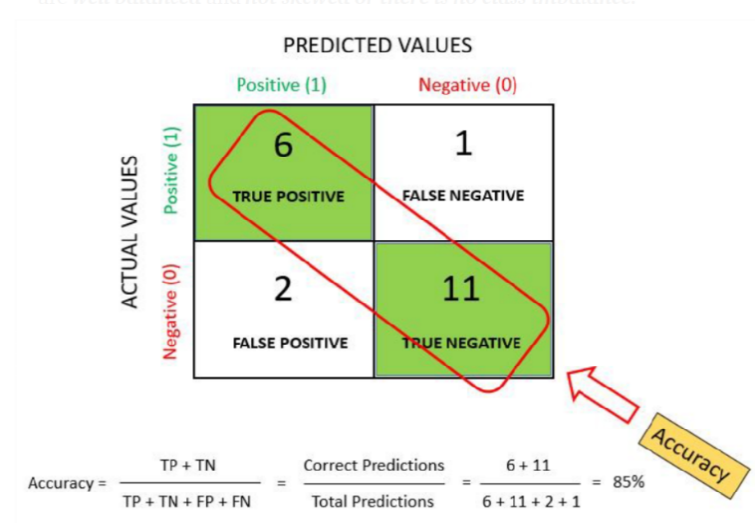
**classification\_report()** takes in the list of actual labels, the list of predicted labels, and an optional argument to specify the order of the labels. It calculates performance metrics like precision, recall, and support.

The report shows the main classification metrics precision, recall and f1-score on a per-class basis. The metrics are calculated by using true and false positives, true and false negatives.



**Accuracy** simply measures how often the classifier makes the correct prediction.

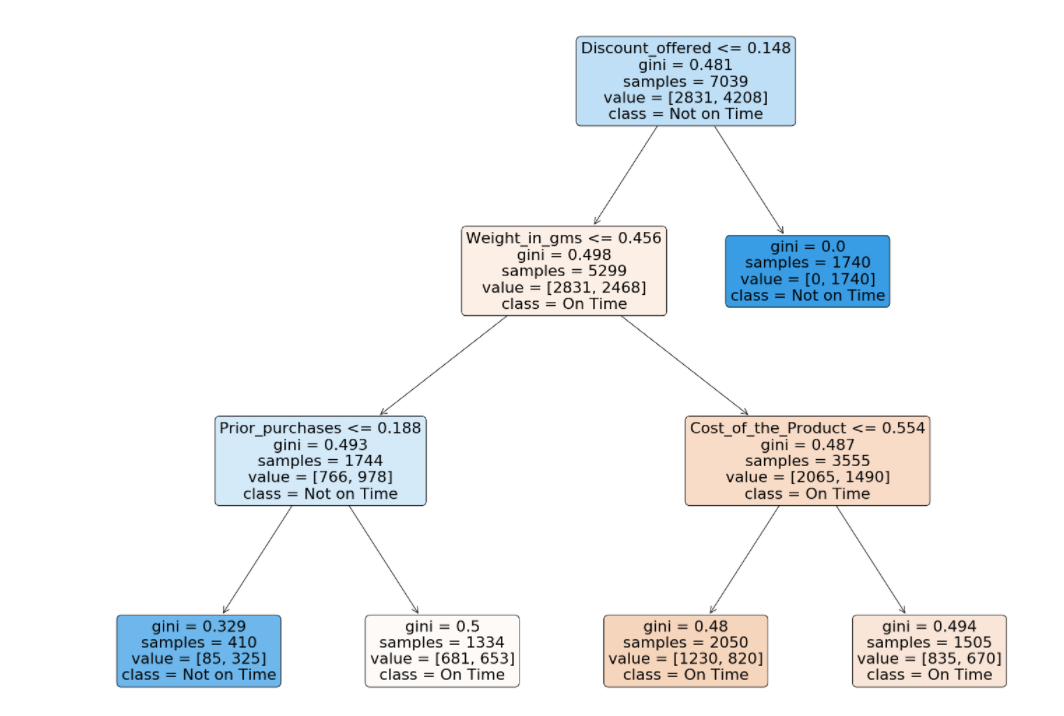
It’s the ratio between the number of correct predictions and the total number of predictions.

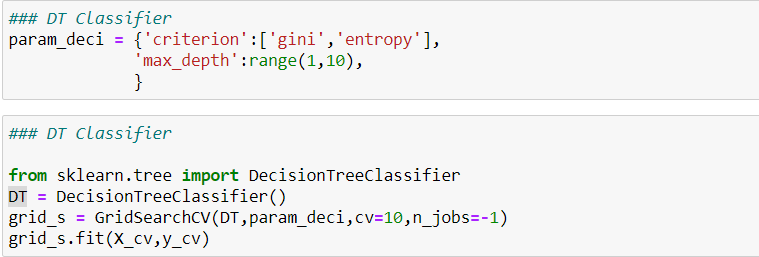


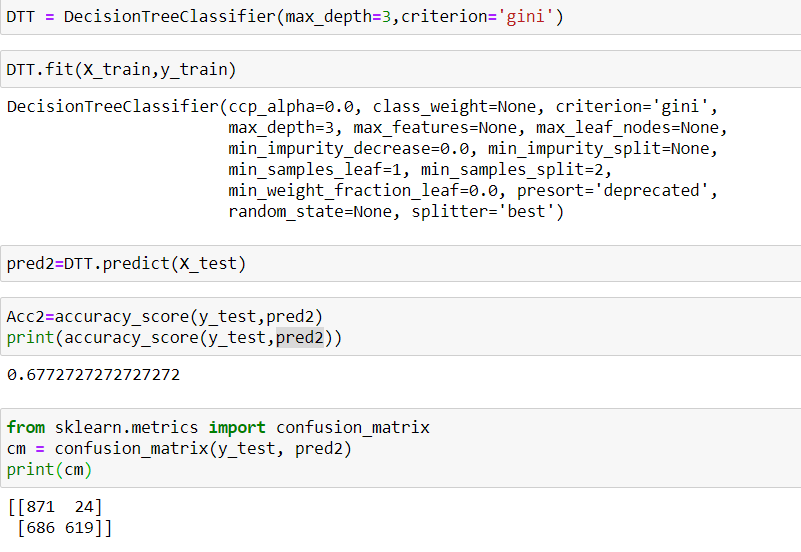
**Accuracy of our model is 67 %**

# **Decision Tree Classifier**

A decision tree is a flowchart-like tree structure where an internal node represents feature (or attribute), the branch represents a decision rule, and each leaf node represents the outcome. The topmost node in a decision tree is known as the root node. It learns to partition based on the attribute value. It partitions the tree in recursively manner call recursive partitioning. This flowchart-like structure helps you in decision-making. It is visualization like a flowchart diagram, which easily mimics the human level thinking.

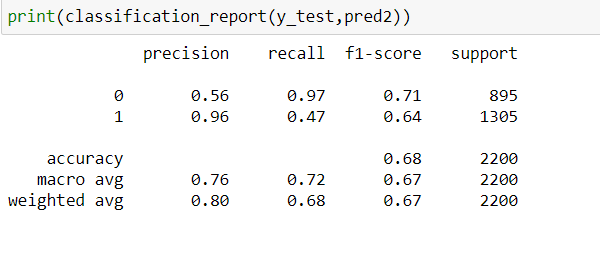




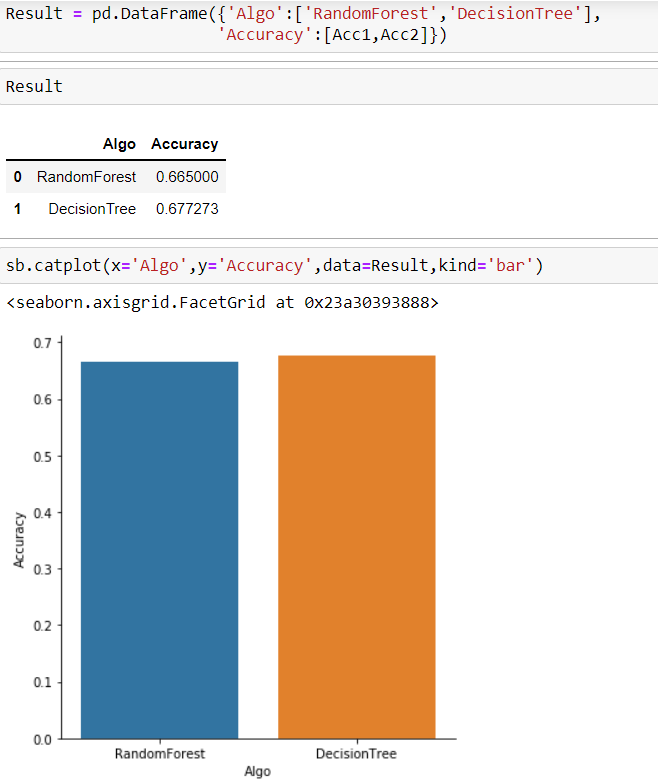


**Confusion metrics** shows that 710 observations have been classified as false.

**Classification Report:**



# **Accuracy Comparison of Random Forest and Decision Tree classifier**



# **Conclusion**

In this project, we have done a comparison between Decision Tree and Random forest for predicting the delivery time of the product and found the accuracy level to be 67% in both. The classification results show that both the models give almost the same level of accuracy for the same number of attributes and data set without any significant difference.

# **Further** **Improvements**

In this model, primarily, the data set is incomplete. Without mentioning about the distance from the warehouse to delivery location, it’s really hard to predict the delivery time. With the addition of the distance from the warehouse to delivery location, the data can be visualized better and prediction that is more accurate can be made. However, we believe that we need few more input data variable in terms of predicting delivery status accurately including

* + Order placement time & date
  + Dispatch time & date from Warehouse
  + Last mile -delivered time & date

# **References:**

1. Kaggle
2. Willey Canvas
3. Google