**Problem Statement:**

ABC bank has decided to select a model that helps in identifying whether the loan should be granted to an applicant or not based on the dataset available with the bank.

Using Bank Dataset, Apply decision tree to predict whether loan can be granted or not.

**Data Demonstration**

Based on the given data set, we can calculate the following statistics.

[**Count**](https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.count.html#pandas.DataFrame.count)**:** Count number of non-NA/null observations.

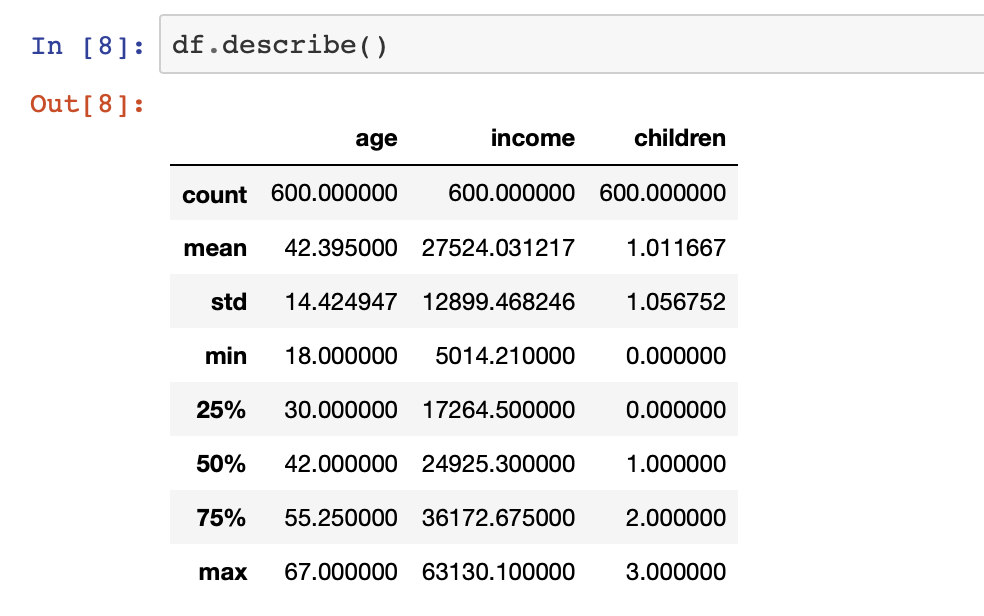
[**Max**](https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.max.html#pandas.DataFrame.max)**:** Maximum of the values in the object.

[**Min**](https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.min.html#pandas.DataFrame.min)**:** Minimum of the values in the object.

[**Mean**](https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.mean.html#pandas.DataFrame.mean)**:** Mean of the values.

[**Std**](https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.std.html#pandas.DataFrame.std)**:** Standard deviation of the observations.

[**Select\_dtypes**](https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.select_dtypes.html#pandas.DataFrame.select_dtypes)**:** Subset of a Data Frame including/excluding columns based on their dtype.



**Data Analysis**

1. **Check for missing values**

In real world data, there are some instances where a particular element is absent because of various reasons, such as, corrupt data, failure to load the information, or incomplete extraction. [Handling](https://analyticsindiamag.com/get-started-preparing-data-machine-learning/) the missing values is one of the greatest challenges.

Some methods to handle the missing data:

* Deleting the entire Row
* Replacing with Mean/Median/Mode
* Assigning a unique category(for categorical values)

Based on the given data, there are missing values in the data set.

1. **Data Transformation**

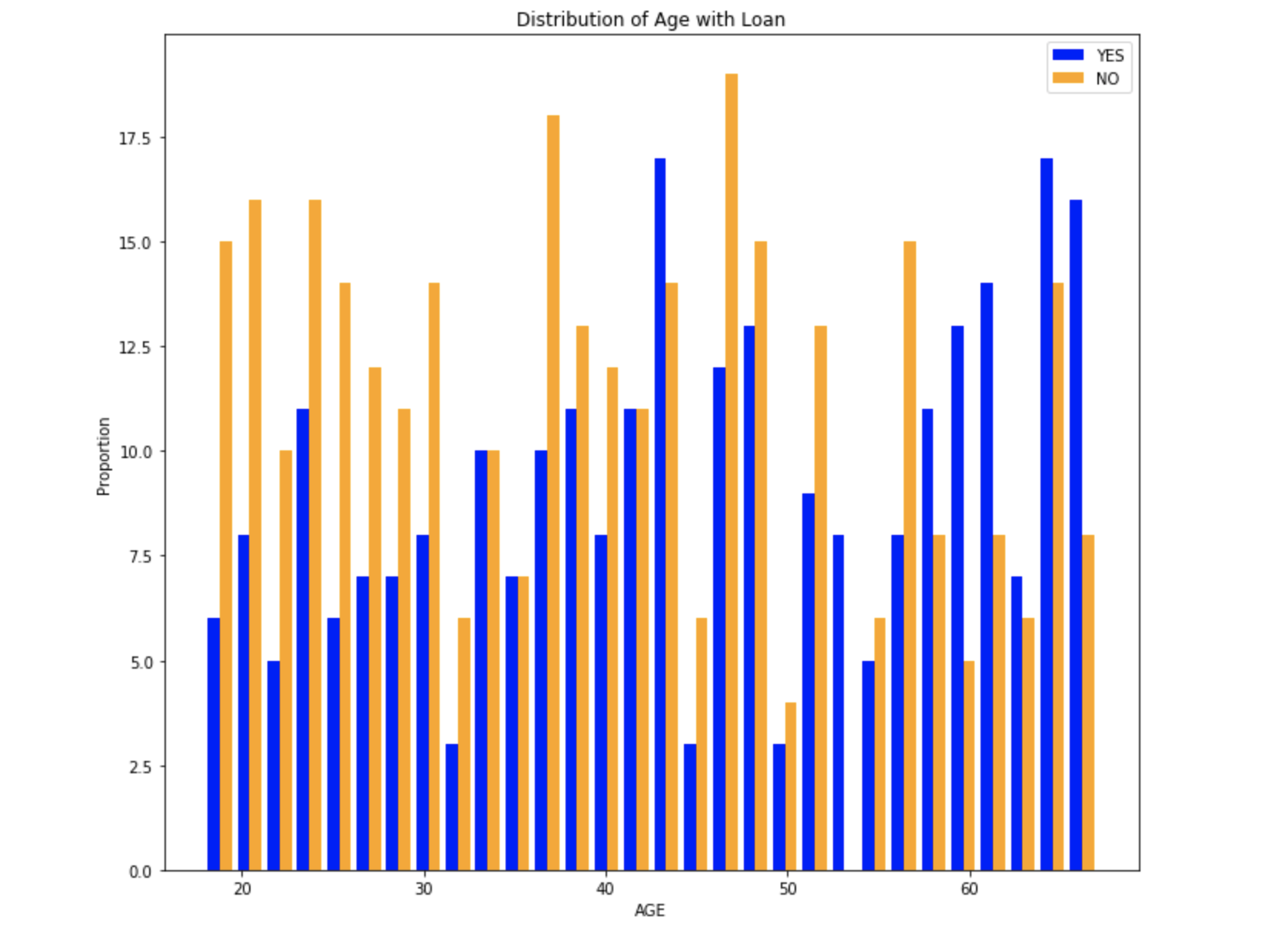
In the given data set we can observe that there are few categorical variables like sex, region, married, car, saving\_account, current\_account, mortgage, Loan.

To build the model we need to transform these categorical values into numbers.

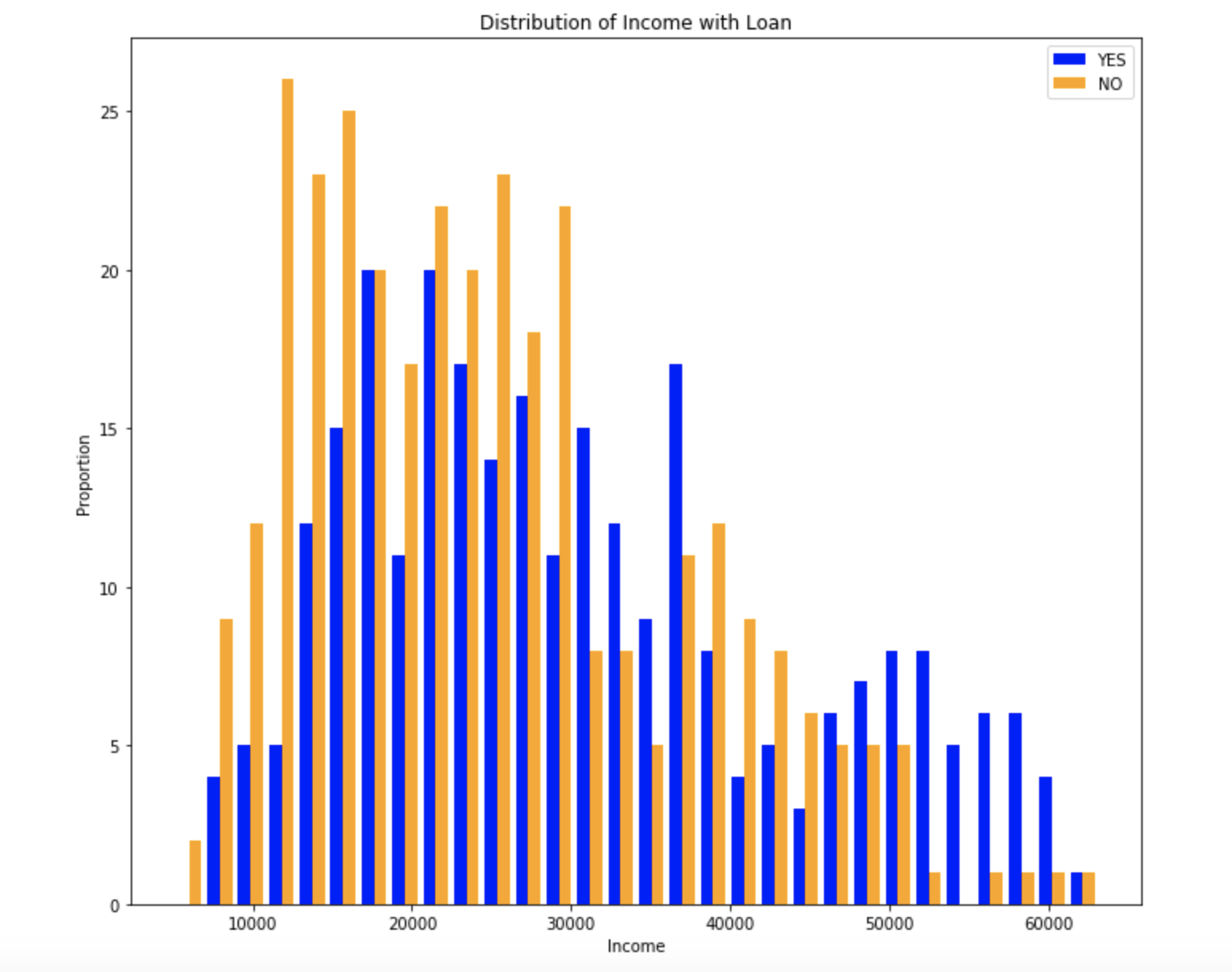
Label Encoding is the one of the technique used to transform the categorical values into numbers.

1. **Relation between dependent variables and independent variable**

**Distribution if Age w.r.t loan:**

****Based on the above graphs we converted the age into three categories

1. Age < 35
2. 35<= Age<= 55
3. Age>55

**Distribution of Income w.r.t Loan:** ****

Based on the above graph we converted the age into 4 categories:

1. Income < 5000
2. 5000 <= Income <= 33000
3. 33000 < Income < 61000
4. Income > 61000

**Feature Selection**

Based on the given data set, Data transformation we have consider the all the features except “**CustomerID”** to predict the loan feature for customer.

**Selecting the Training and Test data**

**Training Dataset**: The sample of data used to fit the model.

**Test Dataset**: The sample of data used to provide an unbiased evaluation of a final model fit on the training dataset.

In the experiment we split the data into training and testing data in the ratio of 80% and 20% respectively.

**Predictions and Reporting**

1. **Classification Accuracy:** Classification accuracy is the number of correct predictions made as a ratio of all predictions made. This is the most common evaluation metric for classification problems.

In the experiment we got **85%** accuracy when training data and test data are in 70:30 % ration, we got **91%** accuracy when training and test data are in 80:20 % ratio.

1. **Confusion Matrix:** The confusion matrix is a handy presentation of the accuracy of a model with two or more classes. The table presents predictions on the x-axis and accuracy outcomes on the y-axis. The cells of the table are the number of predictions made by a machine learning algorithm.

array([[79, 13],

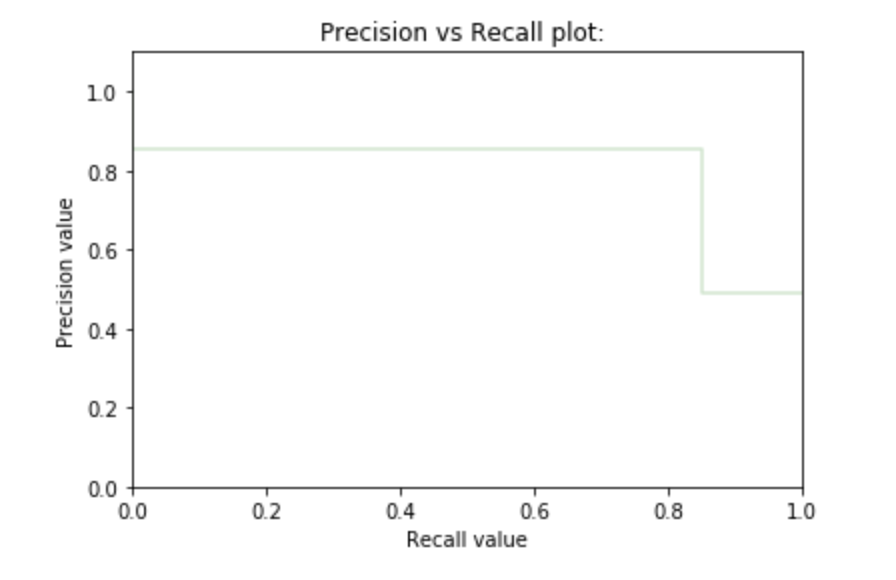
[13, 75]])

1. **Precision Score**: TP – True Positives, FP – False Positives

Precision – Accuracy of positive predictions.  
Precision = TP/(TP + FP)

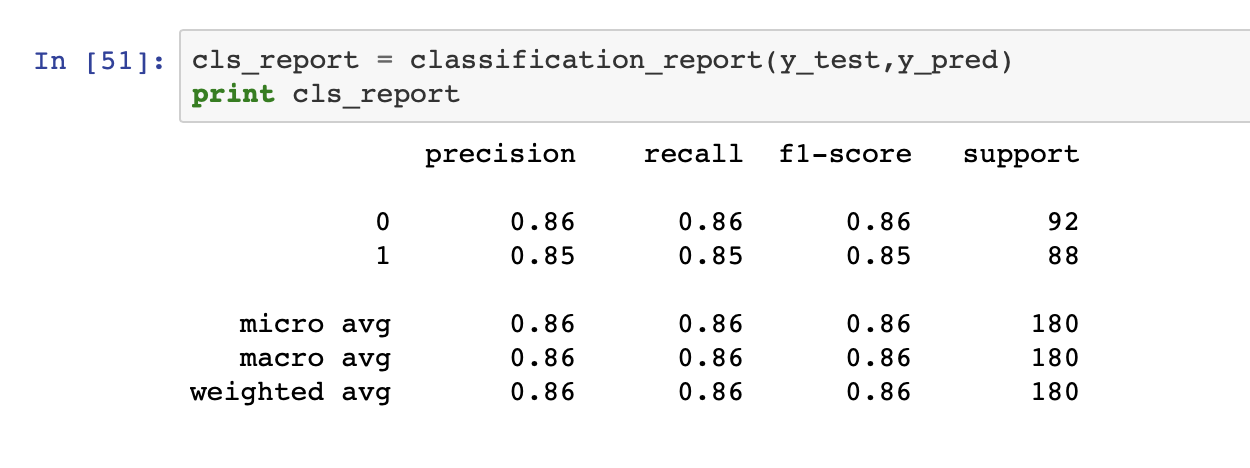
1. **Recall Score:**  FN – False Negatives

Recall: Fraction of positives that were correctly identified.  
Recall = TP/ (TP+FN)

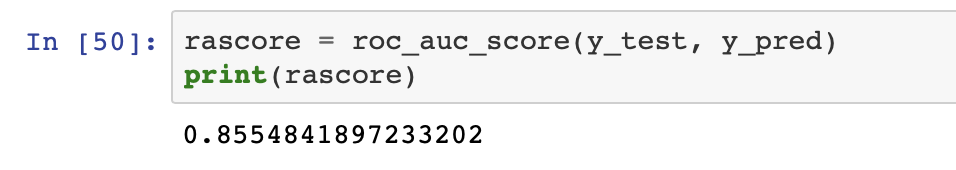
****

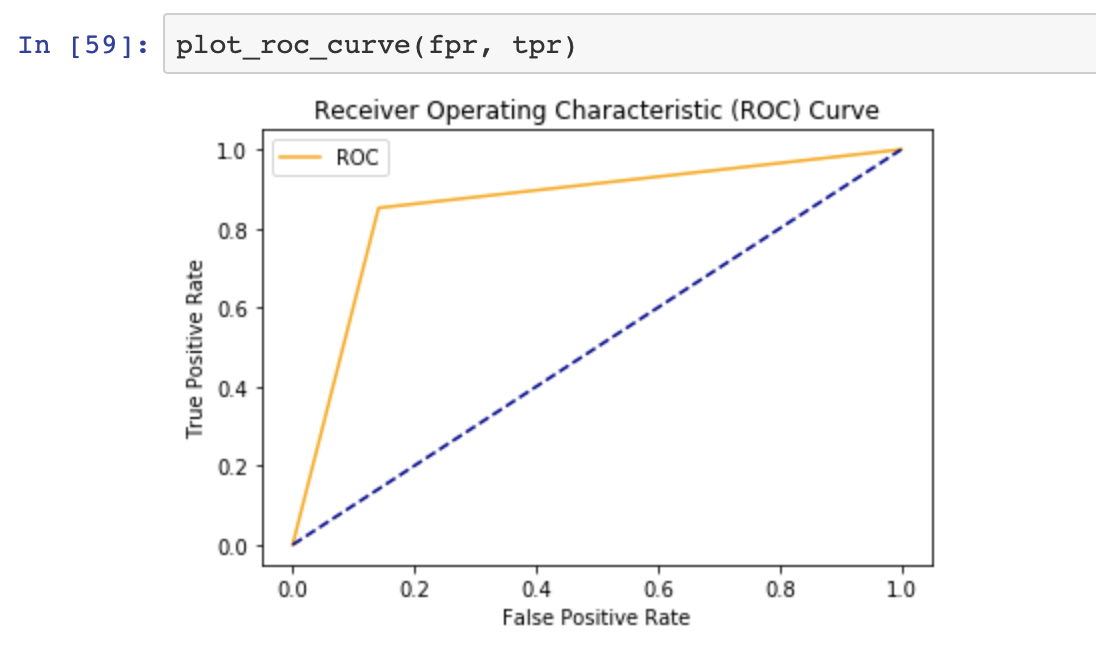
1. **F1 Score–** A helpful metric for comparing two classifiers. F1 Score takes into account precision and the recall. It is created by finding the harmonic mean of precision and recall.

**F1 = 2 x (precision x recall)/ (precision + recall)**

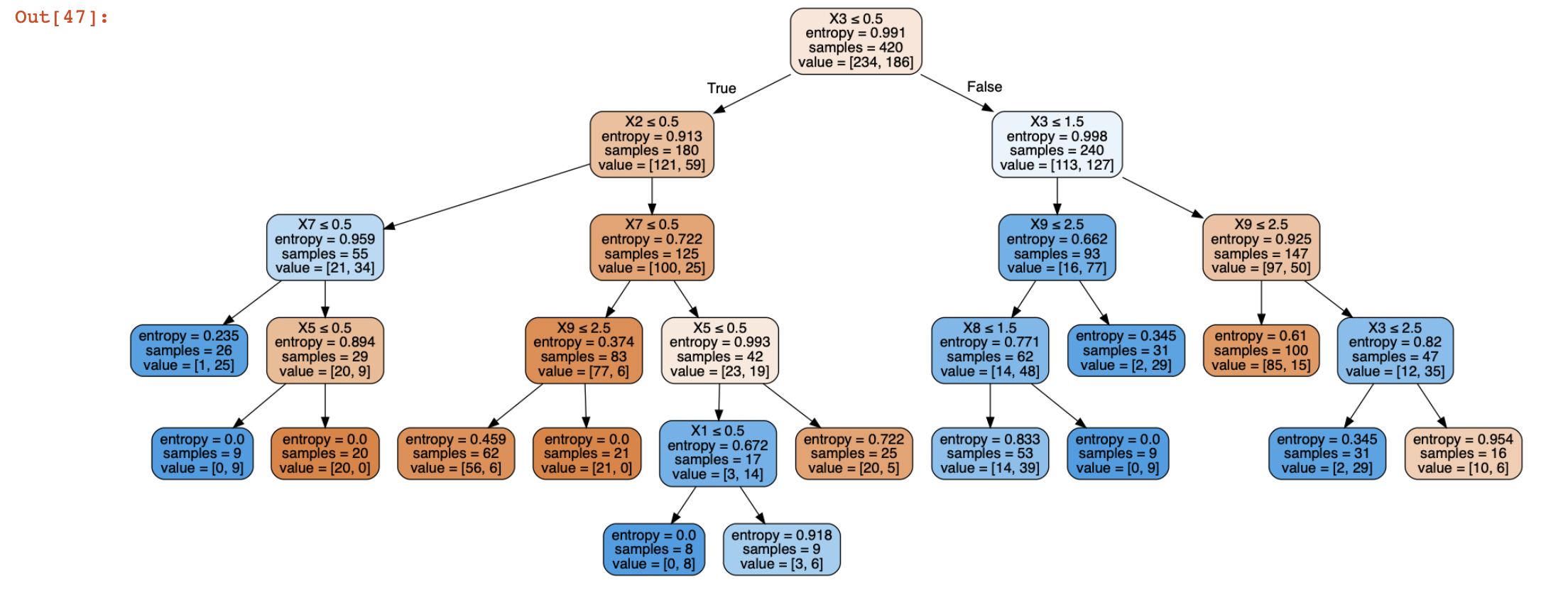


1. **ROC Curve**: The Receiver Operation Characteristic(ROC) curve is created by plotting the [true positive rate](https://en.m.wikipedia.org/wiki/True_positive_rate) (TPR) against the [false positive rate](https://en.m.wikipedia.org/wiki/False_positive_rate) (FPR) at various threshold settings. AUC is good for classification problems with a class imbalance.

****

****

**Decision Tree**

****