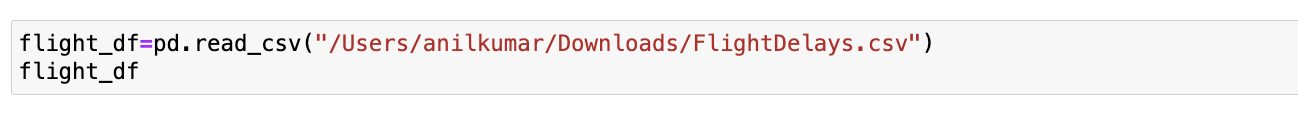
**DATA Mining Case Study**

**1 a)**

**The FlightDelays dataset can be uploaded in the following way:**



**Dimensions of the dataset**

****

From above the dimensions of the dataset: the rows and columns of the dataset are 2201 rows and 11 columns.

**1 b)**

# Removal of the DEST and ORIGIN variable from flight\_df and identifying data types

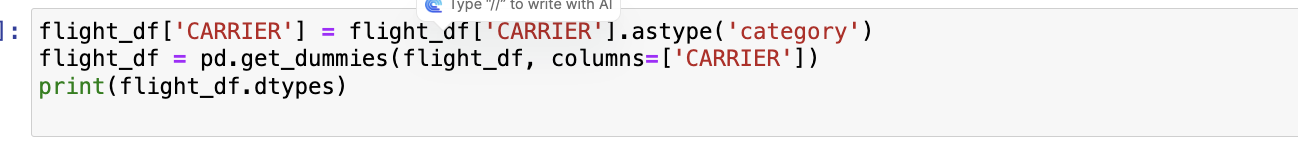
A screenshot of a computer

Description automatically generated

we can see that there are two categorical types with **object** data types and the rest all are the numerical data types. We will convert the **CARRIER** variable into dummies and leave **out FL\_STATUS** as FL\_STATUS is our outcome variable.

**1 c)**

**Changing of the Categorical Variable Carrier**



**Updated data types after converting**

A white background with black text

Description automatically generated

**1 d) First 10 records of the data frame.**

A screenshot of a computer

Description automatically generated

After dropping the unwanted variables, i.e., **DEST and ORIGIN we will be left with 9 variables** in our data frame. There is two categorical variables FL\_STATUS and CARRIER variables in our data frame. We have just changed the **CARRIER** variable in our data frame into the category form, leaving the **FL\_STATUS** untouched which is the outcome variable.

After converting the CARRIER variable, we will have a total of **2201 rows and 15 columns**. So, after converting we have a **total of 14 predictor variables, and FL\_STATUS being the outcome variable.**

**2 a)**

A screenshot of a computer code

Description automatically generated

the

predictors = ['SCH\_TIME', 'DEP\_TIME', 'DISTANCE', 'FL\_NUM', 'WEATHER', 'WK\_DAY', 'MTH\_DAY',

'CARRIER\_CO', 'CARRIER\_DH', 'CARRIER\_DL', 'CARRIER\_MQ', 'CARRIER\_OH',

'CARRIER\_RU', 'CARRIER\_UA', 'CARRIER\_US']

outcome = 'FL\_STATUS'

**Decision Tree:**

Maximum Depth = 5, Minimum Impurity Decrease = 0.001,Minimum Samples Split = 10

**A screenshot of a computer

Description automatically generated**

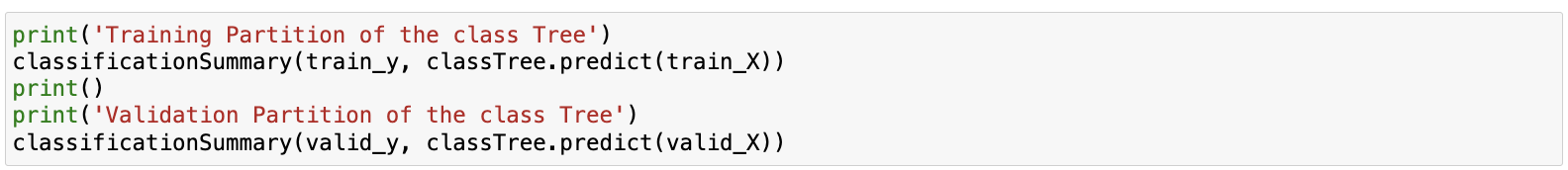
2b) A screenshot of a computer

Description automatically generated

From the above DecisionTreeClassification, If the weather condition is marked as good, the decision node evaluates the **DEP\_Time**. In this scenario, the **DEP\_Time** is 1605 (4:05 pm), which is less than 2147.5 (the criteria of the decision node). Consequently, it proceeds to the node where **DEP\_Time** is <= 1504.5. Here, the **DEP\_TIME** is 1605 (4:05 pm), which is not less than 1504.5 (3:04 pm), thus resulting in a False condition. The process then moves to **SCH\_TIME** <= 1517.5. The given scheduled time is 15:10 (3:10 pm), which satisfies the condition, leading to the next evaluation**: MTH\_DAY** <= 30.5. As the given month day is 28, which is less than 30, this condition holds. Therefore, the **classification** will be **delayed**.

2c)

**Confusion Matrices for Validation and Training.**

****

**A screenshot of a computer

Description automatically generated**

**Predictions for 0's (Actual Negative Class):**

**Training Partition:**

93 instances of actual 0's were correctly predicted as 0's.

211 instances of actual 0's were incorrectly predicted as 1's.

**Validation Partition:**

39 instances of actual 0's were correctly predicted as 0's.

85 instances of actual 0's were incorrectly predicted as 1's.

**Predictions for 1's (Actual Positive Class):**

**Training Partition:**

11 instances of actual 1's were incorrectly predicted as 0's.

1225 instances of actual 1's were correctly predicted as 1's.

**Validation Partition:**

19 instances of actual 1's were incorrectly predicted as 0's.

518 instances of actual 1's were correctly predicted as 1's.

Training Partition:

Accuracy: 85.58% (0.8558)

Misclassification Rate: 14.42% (1 - 0.8558)

Validation Partition:

Accuracy: 84.27% (0.8427)

Misclassification Rate: 15.73% (1 - 0.8427)

Since the difference between the Accuracy and Misclassification rate between the training and validation partition is quite less, we can say that this model has no chance of **Overfitting.**

2 d)

A screenshot of a computer program

Description automatically generated

**Output:**

A number of words on a white background

Description automatically generated

From the above data and with the help of our trained classification tree, the flight status of two new records is classified as delayed.

From the above decision tree

For First Record = If WEATHER <= 0.5 AND DEP\_TIME > 2147.5 -> Classification = DELAYED

For Second Record = If WEATHER <= 0.5 AND DEP\_TIME <= 2147.5, AND DEP\_TIME > 1504.5, AND SCH\_TIME <= 1517.5 -> Classification = DELAYED

The criteria given to the new flight records are checked with the criteria of our trained classification tree by checking the criteria of WEATHER, DEP\_TIME, SCH\_Time, and FL\_NUM as they are marked as decision nodes in our classification tree.

After checking the conditions and traveling the tree path both our new flight records are marked as delayed.

**3 a)**

**GridSearchCV ()**

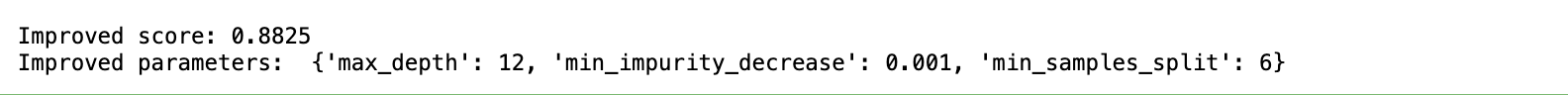
In the above function we will be using the criteria maximum depth is b/w 2 to 30, **minimum impurity is in of 0, 0.0005 and 0.001 and minimum number of nodes being 5 to 30**, with no considering the initial guess grid search.

**Finding of the Improved Score and Parameters**

A screenshot of a computer code

Description automatically generated

**Output**



In the above, we can see that the **improved score is 0.8825** and the parameters are, **maximum depth being 12**, **minimum impurity decrease of 0.001,** and minimum number of nodes or samples per split is **6**

**Code for Classification Tree.**

**A screenshot of a computer code

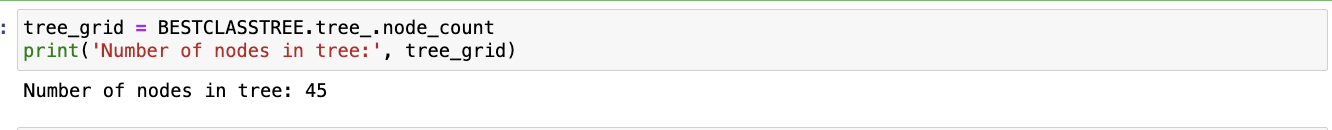
Description automatically generated**

**Classification Tree.**

A diagram of a company

Description automatically generated

**Calculating number of nodes**



**Confusion Matrix of Best Classification Tree**

**A screenshot of a computer

Description automatically generated**

**3b)**

Training Partition Confusion Matrix:

Accuracy: 91.69% (0.9169)

-Misclassification Rate: 8.31% (1 - 0.9169)

-Number of Correct Predictions:

- Actual 0 predicted as 0: 173

- Actual 1 predicted as 1: 1231

- Number of Incorrect Predictions:

- Actual 0 predicted as 1: 131

- Actual 1 predicted as 0: 5

Validation Partition Confusion Matrix:

Accuracy:89.41% (0.8941)

Misclassification Rate: 11.04% (1 - 0.8841)

- Number of Correct Predictions:

- Actual 0 predicted as 0: 65

- Actual 1 predicted as 1: 523

- Number of Incorrect Predictions:

- Actual 0 predicted as 1: 59

- Actual 1 predicted as 0: 14

**For Class Tree**

Validation Partition of the class Tree

Confusion Matrix (Accuracy 0.8427)

Prediction

Actual 0 1

0 39 85

1 19 518

**Misclassification Rate for Class Tree: 15.73% (1 - 0.8427)**

**For bestclasstree**

Validation Partition

Confusion Matrix (Accuracy 0.8896)

Prediction

Actual 0 1

0 69 55

1 15 522

**Misclassification Rate for Bestclasstree is: 10.74% (1 - 0.8926)**

We can see from the

The accuracy is higher for **Bestclass tree** with **88.96 percent** compared with **Class Tree** which has **84.27 percent** and misclassification is **10.74 percent for Bestclasstree** which is lower than **Class Tree, which has 15.73 percent**. Higher Accuracy and Lower Misclassification rate is a sign of better prediction, Therefore BestclassTree is considered a better prediction model.