

Data Mining Project

Group 8 - Team Members:

Manoj Nagarajan – mn393

Kabilan Senapathy – ks879

Srijan Deo – sd759

Sirisha Bojjireddy – sb2423

Submission Requirements

- Part I: Introduction. Discuss the problem at hand, the background of the data set. Who collected/created it, for what purpose, etc.
- Part II: Data. Show summary stats for your data. Number of rows, columns, median, mean, standard deviation, etc. RapidMiner could be a very useful help for this task.
- Part III: Mining Algorithms. Introduce at least TWO CLASSIFICATION / PREDICTION algorithms covered in our class. Show screenshot of the mining workflow.
- Part IV: Evaluation. The most important part. You will address the following issues:
 - a. Do you choose precision or recall as the main measure for your task? Why?
 - b. Show the confusion matrix for the two algorithms. Which one is better?

Part I: Problem At Hand

- Background of the data set
 - Who collected/created it
 - For what purpose
-
- Dataset chosen – Microsoft HR dataset

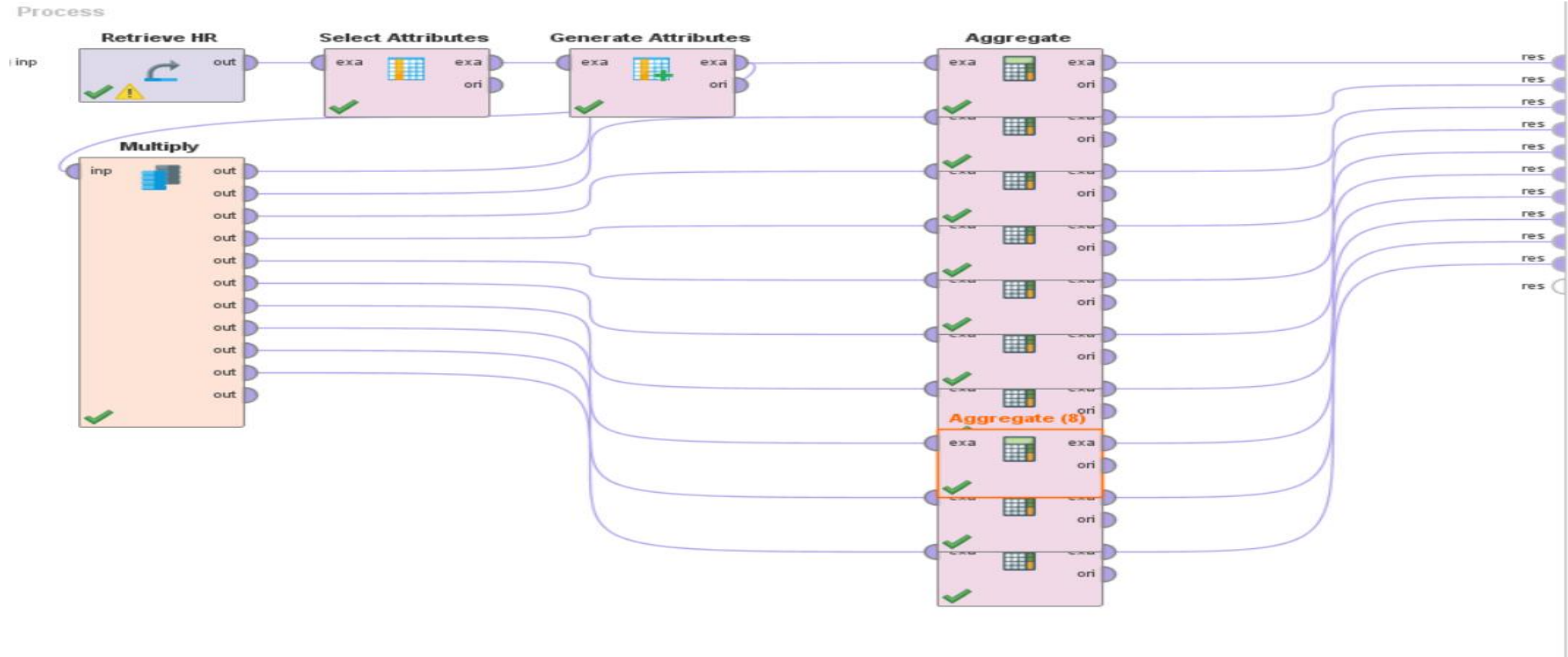
Introduction

- The purpose of the project is to analyze how attrition affects a company's employees? Is attrition one of the main reasons for employees to quit?
- Our team chose to use the data set created by Human Resources department at Microsoft for this analysis.

Part II: Summary Statistics

- Number of rows
- Number of columns
- Median
- Mean
- Standard Deviation

Summary Statistics in Rapidminer



Year At Company

Open in



Turbo Prep



Auto Model

Row No.	Attrition	count(Years...	minimum(Ye...	median(YearsAtCom...	maximum(YearsAtCompany)	standard_deviation(YearsAtCompany)
1	No	1233	0	6	37	6.096
2	Yes	237	0	3	40	5.950

Age

Open in



Turbo Prep



Auto Model

Row No.	Attrition	median(Age)	minimum(A...	average(Age)	maximum(A...	standard_d...	count(Age)
1	No	36	18	37.561	60	8.888	1233
2	Yes	32	18	33.608	58	9.689	237

Business Travel

Open in





Turbo Prep



Auto Model


Row No.	Attrition	BusinessTr...	count(Busin...
1	No	Non-Travel	138
2	Yes	Non-Travel	12
3	No	Travel_Frequ...	208
4	Yes	Travel_Frequ...	69
5	No	Travel_Rarely	887
6	Yes	Travel_Rarely	156

Daily Rate

Open in  Turbo Prep  Auto Model

Row No.	average(Dail...	minimum(D...	median(Dail...	maximum(D...	standard_d...	count(DailyR...
1	802.486	102	802	1499	403.509	1470

Department

Open in  Turbo Prep  Auto Model

Row No.	Attrition	Department	count(Depar...
1	No	Human Reso...	51
2	Yes	Human Reso...	12
3	No	Research & ...	828
4	Yes	Research & ...	133
5	No	Sales	354
6	Yes	Sales	92

Distance From Home

Open in



Turbo Prep



Auto Model

Row No.	Attrition	average(Dis...	minimum(Di...	median(Dist...	maximum(D...	standard_d...	count(Dista...
1	No	8.916	1	7	29	8.013	1233
2	Yes	10.633	1	9	29	8.453	237

Education

Open in



Turbo Prep



Auto Model

Row No.	Attrition	average(Edu...	minimum(Ed...	median(Edu...	maximum(E...	standard_d...	count(Educa...
1	No	2.927	1	3	5	1.027	1233
2	Yes	2.840	1	3	5	1.008	237

Education Field

Open in



Turbo Prep



Auto Model

Row No.	Attrition	EducationField	count(EducationField)
1	No	Human Resources	20
2	Yes	Human Resources	7
3	No	Life Sciences	517
4	Yes	Life Sciences	89
5	No	Marketing	124
6	Yes	Marketing	35
7	No	Medical	401
8	Yes	Medical	63
9	No	Other	71
10	Yes	Other	11
11	No	Technical Degree	100
12	Yes	Technical Degree	32

Gender

Open in



Turbo Prep

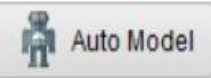
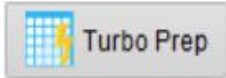


Auto Model

Row No.	Attrition	Gender	MaritalStatus	count(Gender)
1	No	Female	Divorced	108
2	Yes	Female	Divorced	9
3	No	Female	Married	241
4	Yes	Female	Married	31
5	No	Female	Single	152
6	Yes	Female	Single	47
7	No	Male	Divorced	186
8	Yes	Male	Divorced	24
9	No	Male	Married	348
10	Yes	Male	Married	53
11	No	Male	Single	198
12	Yes	Male	Single	73

Hourly Rate

Open in



Row No.	Attrition	average(HourlyRate)	minimum(HourlyRate)	median(HourlyRate) ↑	maximum(HourlyRate)	standard_deviation(HourlyRate)
1	No	65.952	30	66	100	20.381
2	Yes	65.574	31	66	100	20.100

Part III: Mining Algorithms

- Classification Algorithms our team chose
 - ☐ Naïve Bayes
 - ☐ Logistics Regression

Comparison of Naïve Bayes and LogisticRegression

- Naïve Bayes and Logistic regression are two popular models used to solve numerous machine learning problems, in many ways the two algorithms are similar, but at the same time very dissimilar.
- Both algorithms are used for classification problems,
- The learning mechanism is a bit different between the two models, where Naive Bayes is a generative model and Logistic regression is a discriminative model.
- Naïve Bayes assumes all the features to be conditionally independent. Logistic regression splits feature space linearly, and typically works reasonably well when some of the variables are correlated.

Algorithm 1 – Naïve Bayes

- A naive Bayes classifier is a probabilistic machine learning algorithm that uses Bayes' theorem to classify objects.
- Naive Bayes classifiers assume strong, or naive, independence between attributes of data points.
- Popular uses of naive Bayes classifiers include spam filters, text analysis and medical diagnosis.
- We are using Nominal to Numerical. The Nominal to Numerical operator is used for changing the type of non-numeric attributes to a numeric type. This operator not only changes the type of selected attributes but also maps all values of these attributes to numeric values.
- Here we are converting nominal attribute attrition to numerical for analysis purpose.

Algorithm 1 – Naïve Bayes

Bayes Rule is a way to go from $P(X | Y)$ to find $P(Y | X)$

$$\underset{\text{Known}}{P(X | Y)} = \frac{P(X \cap Y)}{P(Y)} \quad 1$$

$P(\text{Evidence} | \text{Outcome})$
(Known from training data)

$$\underset{\text{UnKnown}}{P(Y | X)} = \frac{P(X \cap Y)}{P(X)} \quad 2$$

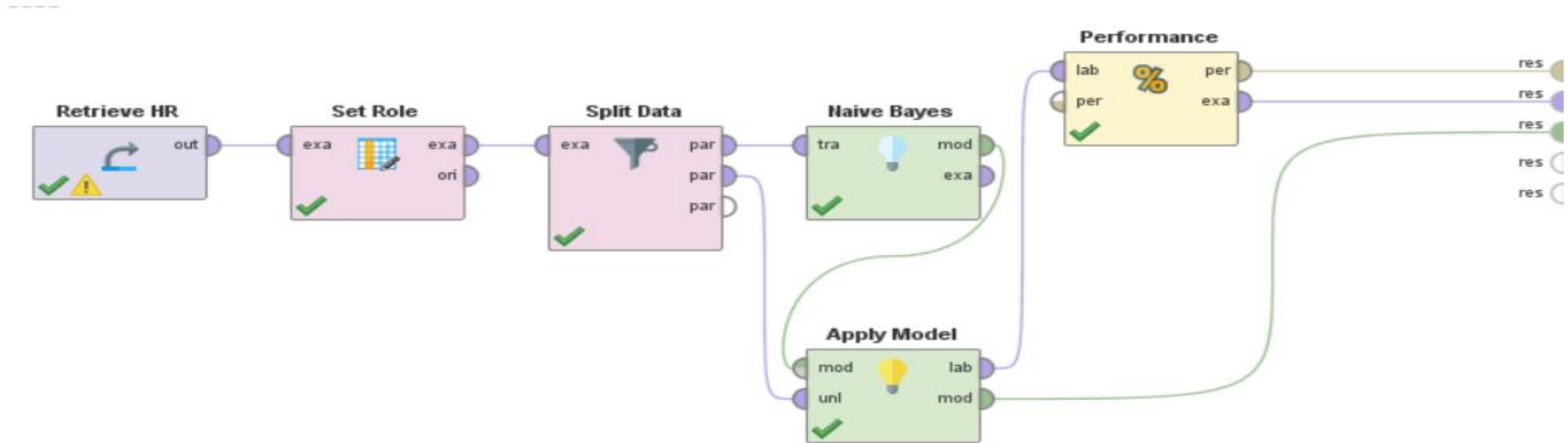
$P(\text{Outcome} | \text{Evidence})$
(To be predicted for test data)



Bayes Rule

$$P(Y | X) = \frac{P(X | Y) * P(Y)}{P(X)}$$

Algorithm 1 – Workflow Screenshot



Results

accuracy: 82.65%

	true Yes	true No	class precision
pred. Yes	48	54	47.06%
pred. No	31	357	92.01%
class recall	60.76%	86.86%	

Accuracy is determined by the count of number of correct prediction to the total number of predictions made

So, the accuracy obtained by Navie Bayes is **82.65%**.

We trained the model with 70% of the dataset. And to validate the performance of the model, 30% of the dataset was used as test data. The test data output had 82 correct predictions for every 100 predictions of Attrition.

Precision:

$$\text{precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}} = \frac{\text{terrorists correctly identified}}{\text{terrorists correctly identified} + \text{individuals incorrectly labeled as terrorists}}$$

- Among all the records we predict as positive, precision is the percent of records that are actually positive.
- According to the confusion matrix precision is **92.01%** which means among all the predicted values of negative 92.01% were negative.

Recall

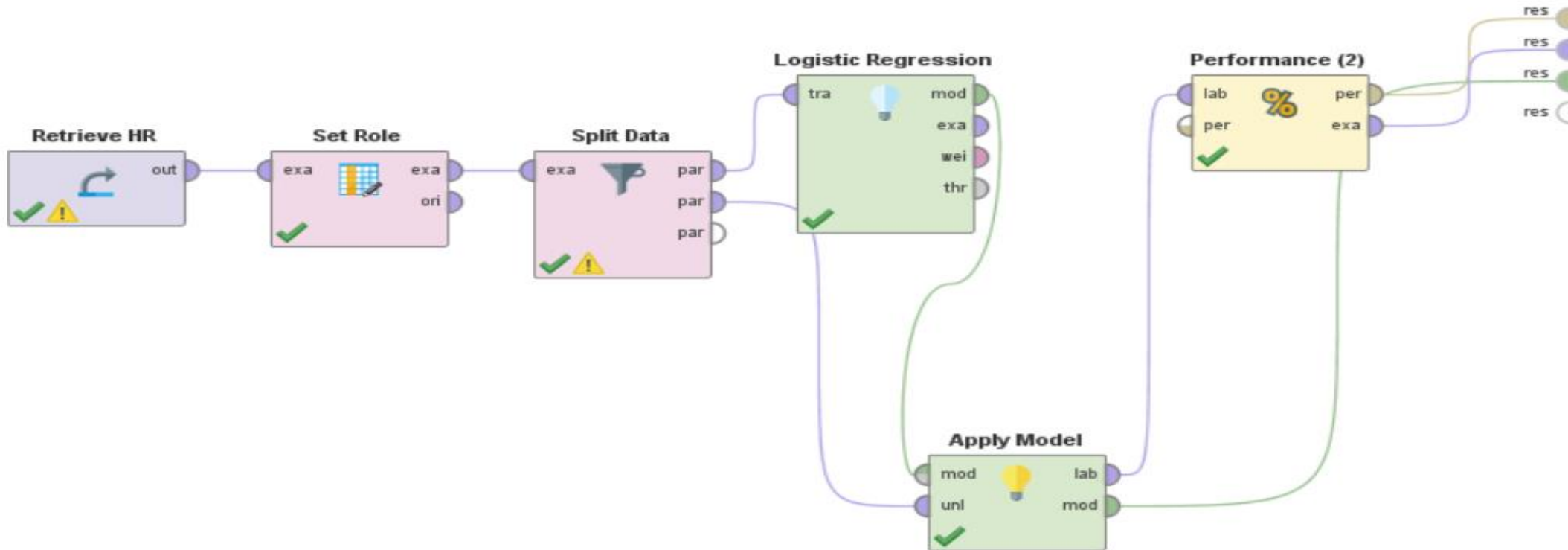
$$\text{recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}} = \frac{\text{terrorists correctly identified}}{\text{terrorists correctly identified} + \text{terrorists incorrectly labeled as not terrorists}}$$

- Recall: Among all the actual positive records, how many percent records has been predicted as positive.
- Here among the 411 negative records, **86.86%** records have been actually predicted as negative. This means when Attrition is negative it has been classified as negative.

Algorithm 2 – Logistics Regression

- Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary).
- Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.
- We use a sigmoid or Logistic Function which gives values exactly between 0 to 1.
- We take a log of the linear regression function to obtain the classification of the model using sigmoid function.
- Thus Logistic regression gives more accurate results compared to Naïve Bayes.

Algorithm 2 – Workflow Screenshot



Result:

accuracy: 88.66%

	true Yes	true No	class precision
pred. Yes	36	13	73.47%
pred. No	42	394	90.37%
class recall	46.15%	96.81%	

Accuracy is determined by count the number of correct prediction to the total number of predictions made

So, the accuracy obtained by Logistic Regression is **88.66%**.

We trained the model with 70% of the dataset. And to validate the performance of the model, 30% of the dataset was used as test data.

The test data output had 88 correct predictions for every 100 predictions of Attrition.

Precision:

$$\text{precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}} = \frac{\text{terrorists correctly identified}}{\text{terrorists correctly identified} + \text{individuals incorrectly labeled as terrorists}}$$

- Among all the records we predict as positive, precision is the percent of records that are actually positive.
- According to the confusion matrix precision is **90.37%** which means among all the predicted values of negative 90.37% were negative.

Recall

$$\text{recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}} = \frac{\text{terrorists correctly identified}}{\text{terrorists correctly identified} + \text{terrorists incorrectly labeled as not terrorists}}$$

- Recall: Among all the actual positive records, how many percent records has been predicted as positive.
- Here among the 407 negative records, **96.81%** records have been actually predicted as negative. This means when Attrition is negative it has been classified as negative.

Part IV: Evaluation

According to the confusion matrix we obtained from the results, precision and recall, we came up with a conclusion that Logistic Regression classification algorithm is better compared to Naïve Bayes Classification Algorithm.