Classification

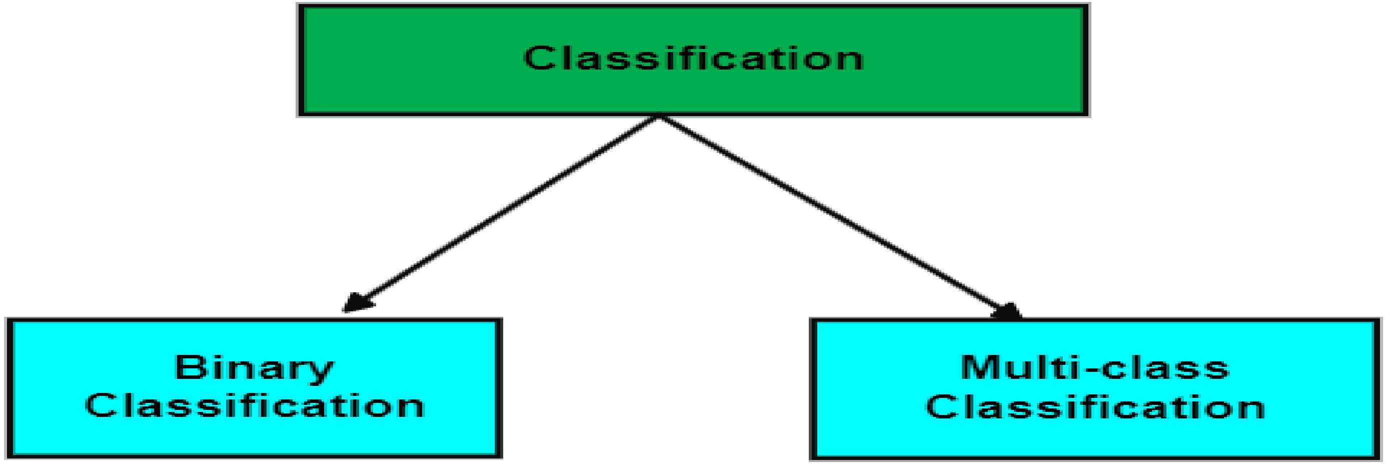
Public

Classification

Classification is a supervised machine learning technique which categorizes the data into different classes.

It is used when the output/outcome variable is categorical/ordinal/discrete in nature

Classification



In binary classification labels have two unique values
For Ex. Yes / No
0 /1
Spam/Ham

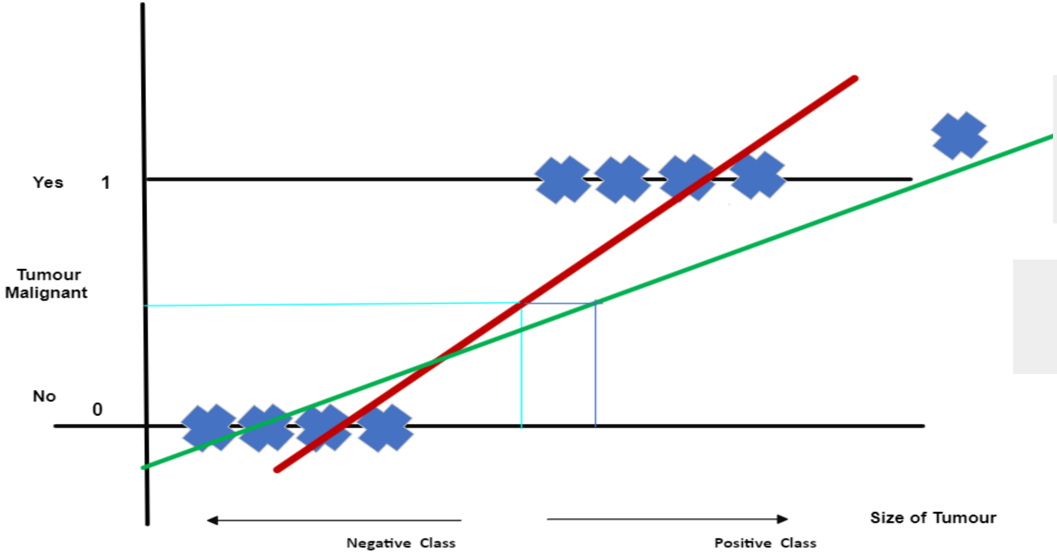
In multi-class classification labels have more than two unique values

For Ex. Setosa/Virginica/Versicolor

Logistic Regression

	Tumour Size (X)	Malignant (Y)
1	0.1	No (0)
2	0.2	0
3	0.3	0
4	0.4	0
5	0.6	Yes (1)
6	0.7	1
7	0.8	1
8	0.9	1
9	2	1

Logistic Regression

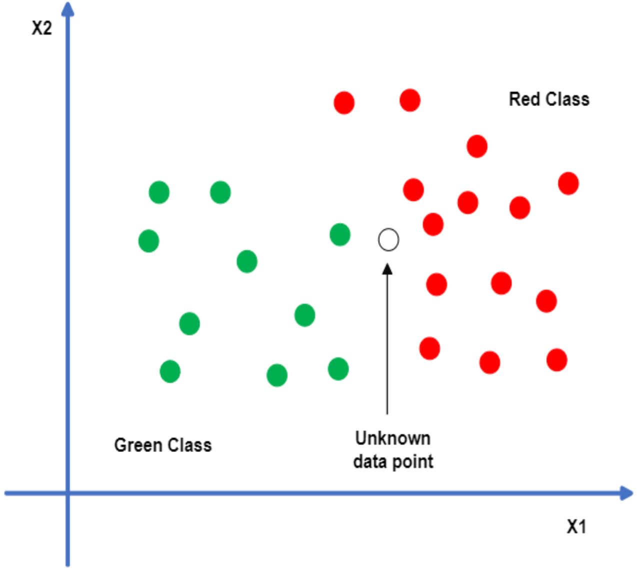


$$Sigmoid = 1/1 + e^{-z}$$

$$y = 1/1 + e^{-}(b0 + b1x_1)$$

Example Source : Coursera

K-Nearest Neighbour



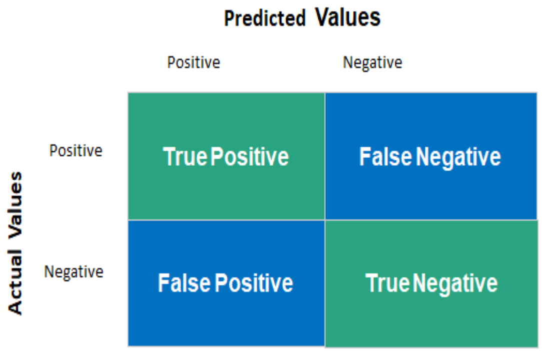
Training Algorithm :- Copying of Training data (features and labels) into memory.

Prediction Algorithm :-

- Decide the value of k.
- Compute the distance between unknown point and all the training points.
- Once the distance is calculated, sort the data in ascending order on the basis of distance.
- Choose Top k values .
- Perform election and assign the label as per majority voting.

Evaluation Metrics

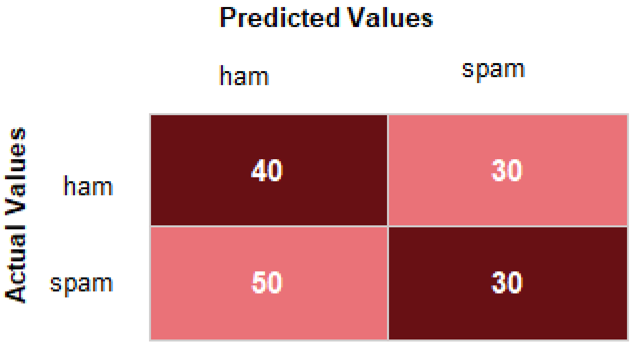
Confusion Matrix



		Predicted Values	
		Positive	Negative
Actual Values	Positive	True Positive	False Negative
	Negative	False Positive	True Negative

Confusion Matrix : It is a matrix used to evaluate the performance of your classification model.

For Ex.



		Predicted Values	
		ham	spam
Actual Values	ham	40	30
	spam	50	30

No of records for ham = 70

No of records for spam = 80

Support(ham) = 70

Support(spam) = 80

Evaluation Metrics

Accuracy : It is ratio of correct predictions over the total no of predictions.

Accuracy = (TP + TN) / (TP+TN+FP+FN)

Precision : It is ratio of correct predictions over the total no of predictions for positive class

Precision = TP/(TP+FP)

F1-Score : It is a harmonic mean of precision and recall.

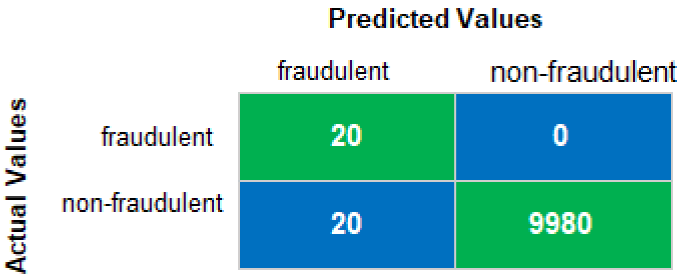
F1-Score = 2*P*R/(P + R)

Recall : It is a ratio of correct predictions over the total no of correct items.

Recall = TP/(TP+FN)

Evaluation Metrics

Confusion Matrix



		Predicted Values	
		fraudulent	non-fraudulent
Actual Values	fraudulent	20	0
	non-fraudulent	20	9980

Accuracy = 10,000/10,020 = 99.8%

Precision = 20/40 = 50%

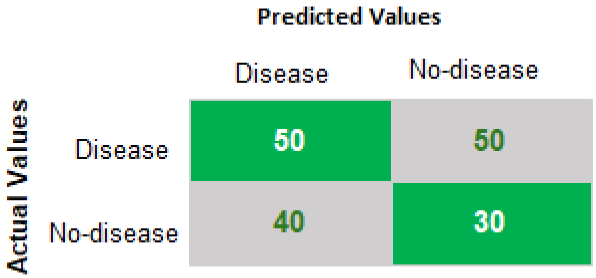
Recall = 20/20 = 100%

In this example, the accuracy of model is 99.8%. But model is not doing great job.

Accuracy is measure which is preferred for a balanced dataset.

We need to minimize the false positives so that precision will be improved.

Evaluation Metrics



		Predicted Values	
		Disease	No-disease
Actual Values	Disease	50	50
	No-disease	40	30

In this example, we need to minimize the false negatives so recall is important

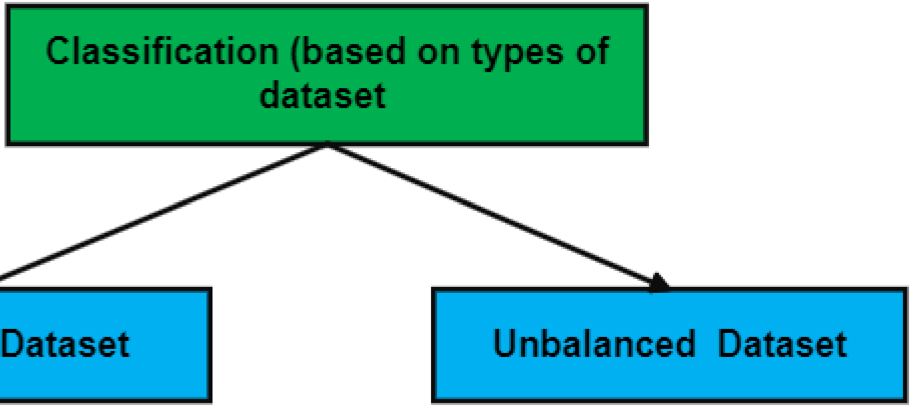
Accuracy = 50 +30/50+50+40+30 = 47%

Precision = 50/50+40 = 55%

Recall = 50/50+50 = 50%

Evaluation Metrics

When to use which metric ?



Accuracy

Precision
Recall
F1-score

wipro

Thank You !!!