Confusion Matrix – It's a table that helps describing the performance of classification model.

measures for (metrics) classification

- Accuracy
- Precision
- Recall
- F1 Score

	S	Ve	Vi
S	9	0	0
Ve	0	7	1
Vi	0	0	13

ratio of correctly predicted observation to the total observations

Accuracy for iris use-case

accuracy = $\frac{9+7+13}{(9+0+0+0+7+1+0+0+13)}$ (entire model)

$= \frac{29}{30} = 0.9666$

	S	Ve	Vi
S	9	0	0
Ve	0	7	1
Vi	0	0	13

ratio of the correctly (tve) PREDICTED observations to the total observations (tve) of a given label class. (predicted) (true and false)

precision for iris use-case (for each unique label)

Precision (setosa) $= \frac{9}{9+0+0} = 1$

Precision (versicolor) $= \frac{7}{0+7+0} = 1$

Precision (virginica) $= \frac{13}{0+1+13} = \frac{13}{14} = 0.928$

Recall for iris use-case (for each unique label)

	S	Ve	Vi
S	9	0	0
Ve	0	7	1
Vi	0	0	13

$$\text{Recall (setosa)} = \frac{9}{9+0+0} = 1$$

$$\text{Recall (versicolor)} = \frac{7}{0+7+1} = 0.875$$

$$\text{Recall (Virginica)} = \frac{13}{0+0+13} = 1$$

ratio of correctly predicted observations to all observations of the true class

F1 Score

$$\frac{2 * (\text{Recall} * \text{Precision})}{\text{Recall} + \text{Precision}}$$

	S	Ve	Vi
S	9	0	0
Ve	0	7	1
Vi	0	0	13

$$\text{F1 score (setosa)} = \frac{2*(1)}{2} = 1$$

$$\text{F1 score (versicolor)} = \frac{2*(0.875 \times 1)}{1.875}$$

$$= 0.9333$$

Harmonic mean of precision and recall of each label class.

$$\text{F1 score (virginica)} = \frac{2 \times (1 \times 0.9285)}{1.9285}$$

$$= 0.962924$$

When to use which metric for model selection?

① If your data is a balanced data, then check only the accuracy part of the metric as a model selection criteria.

Balanced data is any data that has an equal class weightage.

```
iris.species.value_counts()
```

Setosa	50
Versicolor	50
Virginica	50

} equal values, therefore the data is a balanced data.

Since the data is the balanced dataset, therefore consider only ACCURACY. In this case, accuracy is 96%, therefore our model is apprael!

② If your data is an unbalanced data, in that case check for the domain requirement and based on the domain requirement, select the correct metric you need to consider.

The typical industrial standard for defining unbalanced data is any thing that doesn't satisfy the below ratio:

2 labels

55 : 45 .... ±5% is considered as balanced.

Setosa	60
Versicolor	40
Virginica	50

} → Balanced / Unbalanced

$$x = \frac{\text{Total observations}}{\text{no of unique classes}} \Rightarrow \frac{150}{3} \Rightarrow 50$$

$\boxed{x = 50}$

Use this x to identify the ratio. If in any one of the ratio, the data is not satisfying $\pm 5\%$ formula, consider entire dataset as unbalanced.

Setosa : 60 $\longrightarrow$ $60/50 \Rightarrow 1.2$

Versicolor : 40 $\longrightarrow$ $40/50 \Rightarrow 0.8$

Virginica : 50 $\longrightarrow$ $50/50 \Rightarrow 1$

unbalanced

$\| -5 \|$

$\boxed{95\% \rightarrow \text{unbalanced}}$

for most of the odd labels dataset, the data is always unbalanced.

for most of the even label dataset, the data is mostly balanced.

Email spam classification

Domain → model that can help removing spams.

true \ predicted	spam	notspam
spam	100	170
not spam	30 *	700

accuracy = 80%.

precision (spam) = 0.7692

precision (not spam) = 0.8046

recall (spam) = 0.3704

recall (not spam) = 0.9589

If a not spam emails goes to SPAM

Domain:- A hospital wants you to create a model that can classify whether the patient has critical diseases or not.

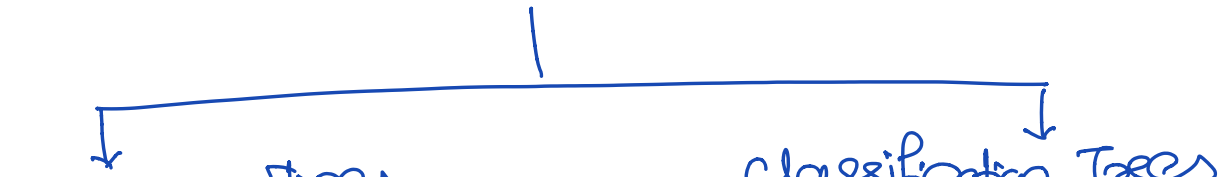
	diagnosed sick	diagnosed healthy
sick	1000	200 **
healthly	800	9000

** is worse case

precision (diagnosed healthy) ⇒ 0.978

CART

(Classification and Regression Trees)



Regression Trees

Label → continous

- Decision Tree Regressor
- Randomforest Regressor

Classification Trees

label → categorical.

- Decision Tree classifier
- Randomforest Classifier