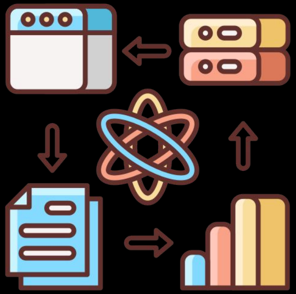
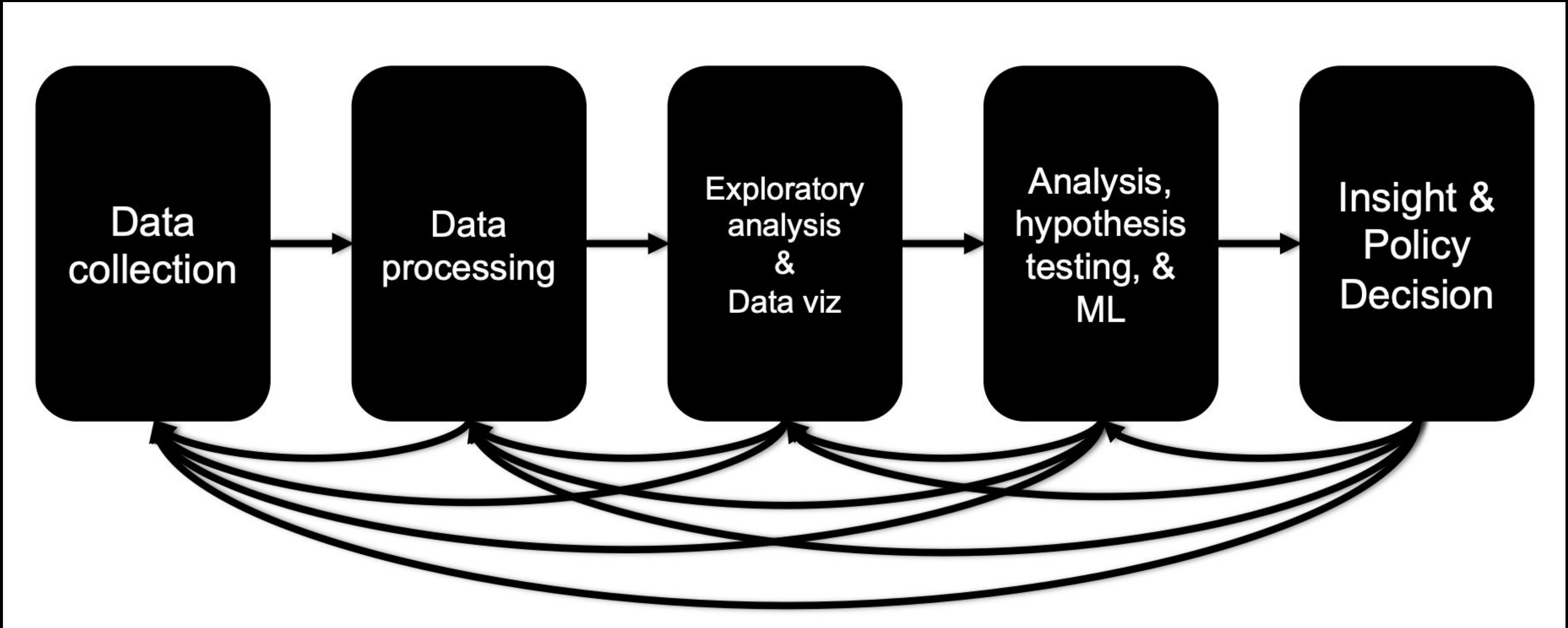


Cont ML Algorithms

CSC 380 - Principles of Data Science

Lecture 7.2

So far in the course



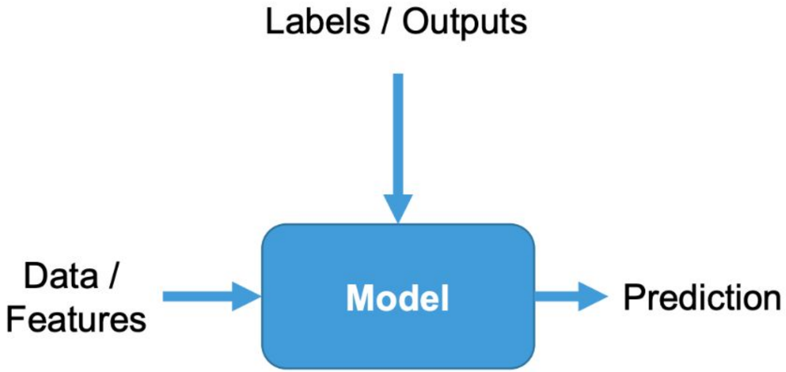
In our last lecture

- Data Preprocessing:
  - Cleaning
  - Integration
  - Transformation
  - Reduction
  - Discretization
  - Normalization
- Machine Learning:
  - Supervised Learning :
    - Linear Regression
    - Bayesian Classification
  - Unsupervised Learning :
    - Clustering - K-means

In this lecture:

- Revise/new terms
- Evaluating models
- Decision Tree Classifier
- Knn (not to be confused with Kmeans from last lecture)
- Logistic regression

Supervised Learning



Unsupervised Learning

ML models distinguished by a number of factors

- Number of parameters needed (parametric / nonparametric)
- Whether they model uncertainty (probabilistic / nonprababilistci)
- Do they model the data generation process? (generative / discriminative)

Revise : Linear v/s Non-Linear

- Linear models generate output as a linear combination of inputs.
    - PCA, linear regression, etc.
- Nonlinear models fit an arbitrary nonlinear function to map inputs to outputs:
    - Neural networks, support vector machine, nonlinear dimensionality reduction

Parametric vs Nonparametric

- Parametric : fixed number of parameters
- Nonparametric : either has an infinite number of parameters, or parameters grow with the amount of data

Parametric Methods	Non-Parametric Methods
Parametric Methods uses a fixed number of parameters to build the model.	Non-Parametric Methods use the flexible number of parameters to build the model.
Parametric analysis is to test group means.	A non-parametric analysis is to test medians.
It is applicable only for variables.	It is applicable for both – Variable and Attribute.
It always considers strong assumptions about data.	It generally fewer assumptions about data.
Parametric Methods require lesser data than Non-Parametric Methods.	Non-Parametric Methods requires much more data than Parametric Methods.
Parametric methods assumed to be a normal distribution.	There is no assumed distribution in non-parametric methods.

Parametric Methods	Non-Parametric Methods
Parametric data handles – Intervals data or ratio data.	But non-parametric methods handle original data.
Here when we use parametric methods then the result or outputs generated can be easily affected by outliers.	When we use non-parametric methods then the result or outputs generated cannot be seriously affected by outliers.
Parametric Methods can perform well in many situations but its performance is at peak (top) when the spread of each group is different.	Similarly, Non-Parametric Methods can perform well in many situations but its performance is at peak (top) when the spread of each group is the same.
Parametric methods have more statistical power than Non-Parametric methods.	Non-parametric methods have less statistical power than Parametric methods.
As far as the computation is considered these methods are computationally faster than the Non-Parametric methods.	As far as the computation is considered these methods are computationally slower than the Parametric methods.
Examples: Logistic Regression, Naïve Bayes Model, etc.	Examples: KNN, Decision Tree Model, etc.

Source : https://www.geeksforgeeks.org/difference-between-parametric-and-non-parametric-methods/

Probabilistic vs Non-Probabilistic

- non-probabilistic generates deterministic outputs / predictions from data ex: k-means
- A probabilistic model represents predictions as random variables, with a distribution that is fit to training data

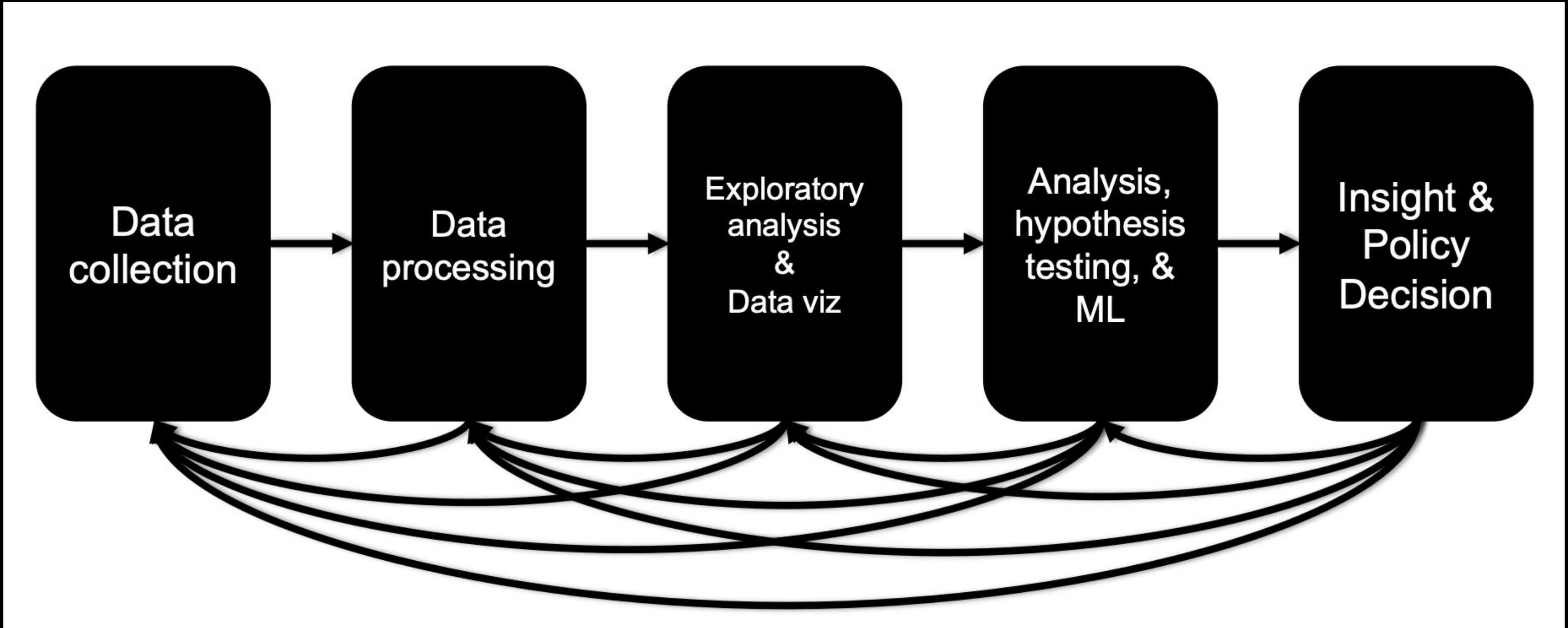
Generative vs Discriminative

- Discriminative models
    - learn the (hard or soft) boundary between classes
    - Lesser assumptions
    - simply providing classification splits (and not necessarily in a probabilistic manner)
    - Assume some functional form for P(Y|X)
    - Estimate parameters of P(Y|X) directly from training data

- Generative models
    - model the distribution of individual classes
    - providing a model of how the data is actually generated
    - make some kind of structure assumptions on your model
    - Assume some functional form for P(Y), P(X|Y)
    - Estimate parameters of P(X|Y), P(Y) directly from training data
    - Use Bayes rule to calculate P(Y |X).

Story Source : https://medium.com/@mlengineer/generative-and-discriminative-models-af5637a66a3

So far in the course



Evaluation

1. Accuracy
2. Confusion matrix
3. Precision
4. Recall
5. F1 score
6. Precision-Recall or PR curve
7. ROC (Receiver Operating Characteristics) curve
8. PR vs ROC curve.

Accuracy

Pros

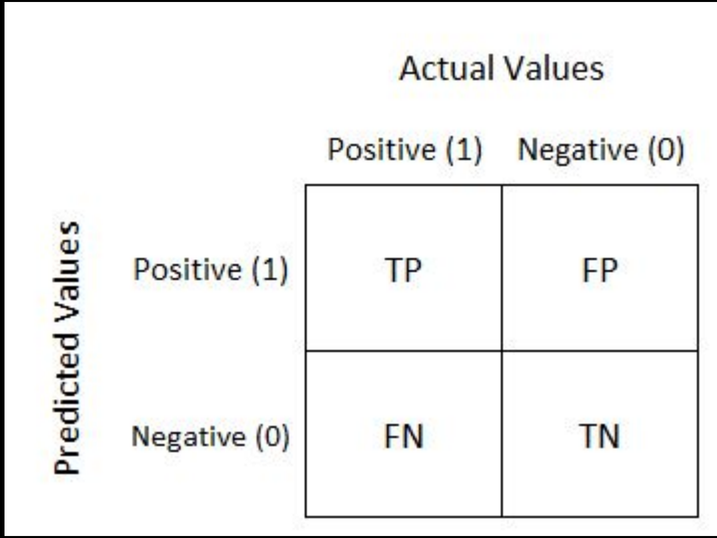
- Easy to interpret

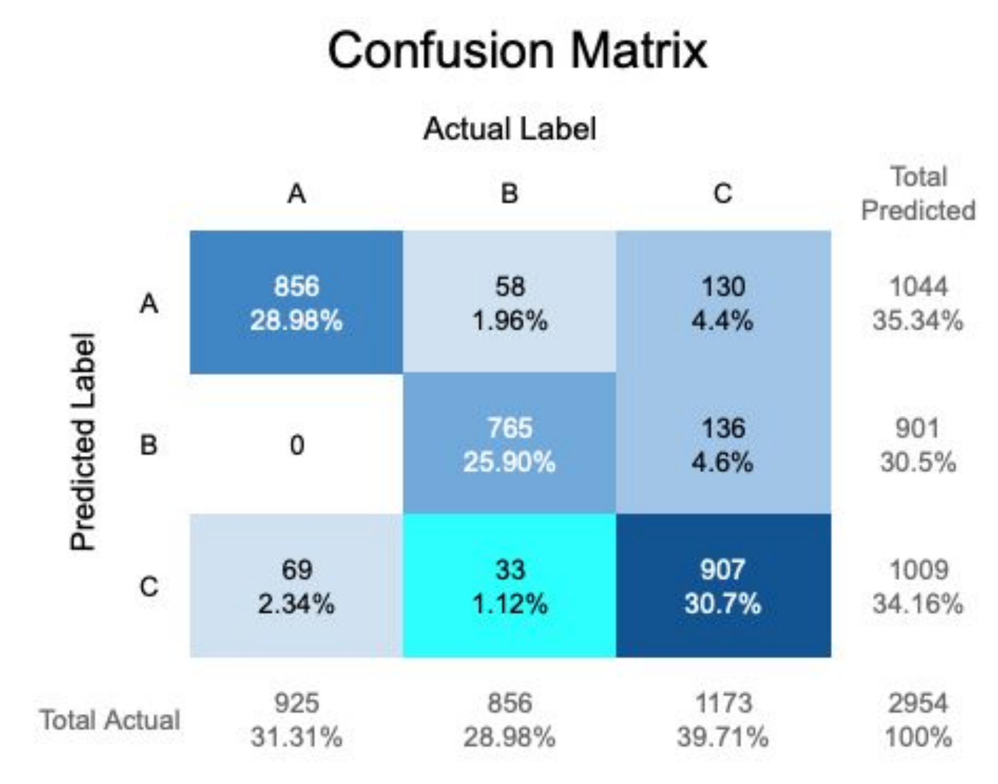
Cons

- Paints an incorrect picture when classes are imbalance
- there are different costs associated with the different mistakes.
- Threshold of accuracy has different costs for small diffs ex: (0.51 and 0.99 is the same) versus (0.49 and 0.51 is not)

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

Confusion Matrix





Recall (or 1 - Sensitivity)

What proportion of actual positives was identified correctly?

- A measure of quality

- Higher recall means that an algorithm returns more of the correct results

$$Recall = \frac{True\ Positive(TP)}{True\ Positive(TP) + False\ Negative(FN)}$$

Precision (or Specificity)

What proportion of positive identifications was actually correct?

- A measure of quantity
- Higher precision means that an algorithm returns more correct results than incorrect ones

$$Precision = \frac{True\ Positive(TP)}{True\ Positive(TP) + False\ Positive(FP)}$$

F1

The harmonic mean encourages similar values for precision and recall.

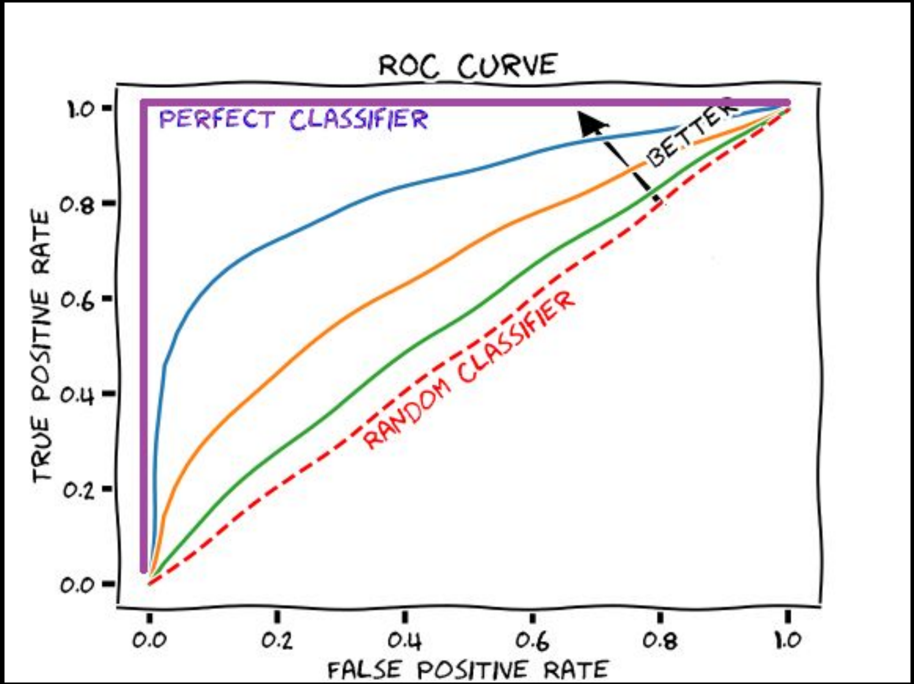
That is, the more the precision and recall scores deviate from each other, the worse the harmonic mean.

$$F_1 = 2 * \frac{precision * recall}{precision + recall}$$

AUCROC

- Area under receiver operating characteristic curve or c-statistic or “concordance statistic.”

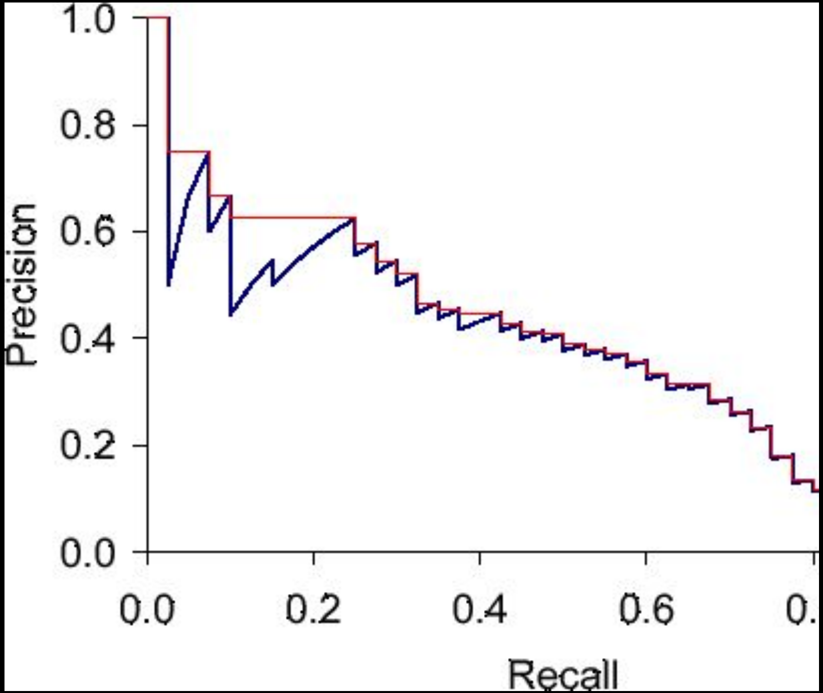
- ROC curves should be used when there are roughly equal numbers of observations for each class.



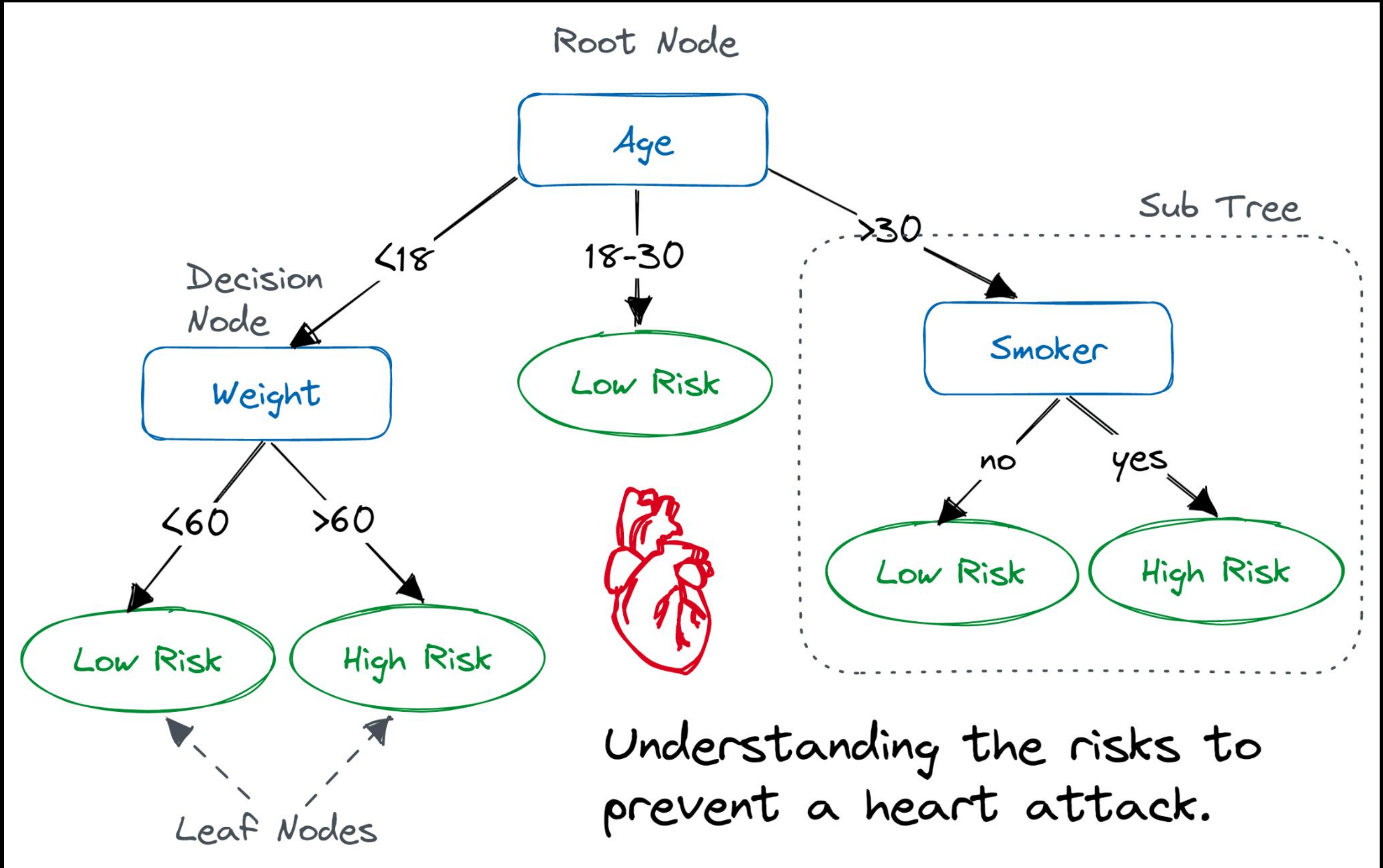
Source : https://glassboxmedicine.com/2019/02/23/measuring-performance-auc-auroc/

Precision-Recall Curve

Precision-Recall curves should be used when there is a moderate to large class imbalance.



Decision Tree Classifier

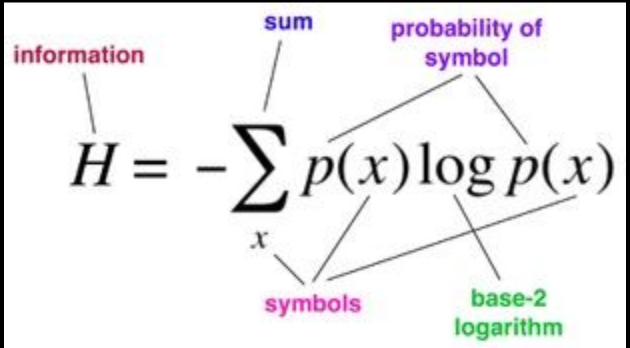


Source : https://www.datacamp.com/tutorial/decision-tree-classification-python

How do we pick which feature becomes the Node?

How do we pick which feature becomes the Node?

Entropy : Measure of information contained in a random variable



Lower Probability , Higher Information

Higher Probability, Lower Information

How do we pick which feature becomes the Node?

Information gain calculates the reduction in entropy parent node to child node.

$$Information\ Gain = E_{parent} - AvgE_{child}$$

	Entropy Node	Average Entropy	Information Gain
Parent	0.9968		
working	0.9183	0.8110	0.1858
Not_work	0.6500		
Bkgrd_Ma	0.9852	0.4598	0.5370
Bkgrd_CS	0.0000		
Bkgrd_oth	0.0000		
online_co	0.9544	0.9688	0.0280
online_no	0.9852		

https://towardsdatascience.com/decision-trees-explained-entropy-information-gain-gini-index-ccp-pruning-4d78070db36c

How do we pick which feature becomes the Node?

Gain Index :

The probability for a random instance being misclassified when chosen randomly. The lower the Gini Index, the better the lower the likelihood of misclassification.

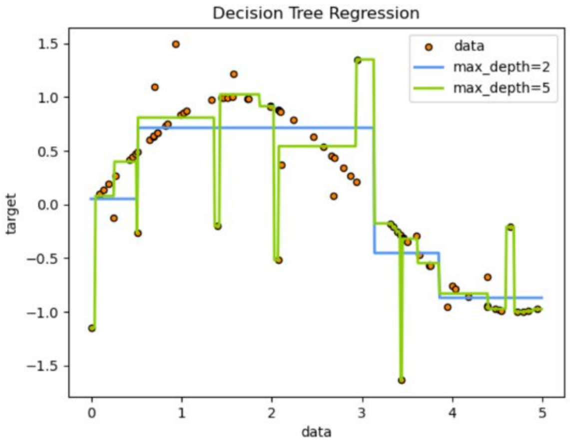
CART (Classification and Regression Tree) uses the Gini method to create split points.

$$Gini = 1 - \sum_{i=1}^{j} P(i)^2$$

1.10. Decision Trees

Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. A tree can be seen as a piecewise constant approximation.

For instance, in the example below, decision trees learn from data to approximate a sine curve with a set of if-then-else decision rules. The deeper the tree, the more complex the decision rules and the fitter the model.



Preventing Overfitting

- Pre pruning or Early stopping: Preventing the tree from growing too big or deep

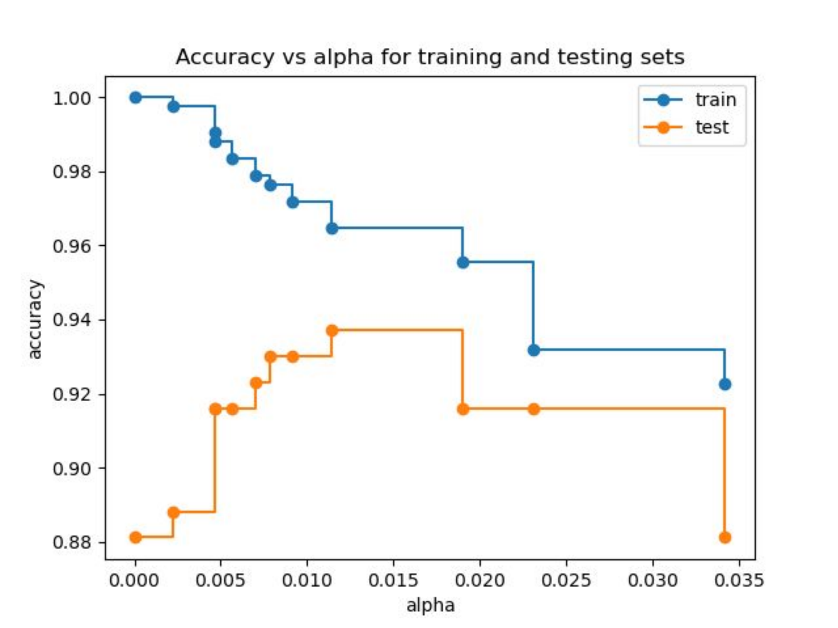
- Post Pruning: Allowing a tree to grow to its full depth and then getting rid of various branches based on various criteria

- Ensembling or using averages of multiple models such as Random Forest

Minimal Cost-Complexity Pruning

ccp_alphano

Complexity parameter used



https://scikit-learn.org/stable/auto_examples/tree/plot_cost_complexity_pruning.html

K Nearest Neighbours

Birds of a feather flock together

Blog - https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761

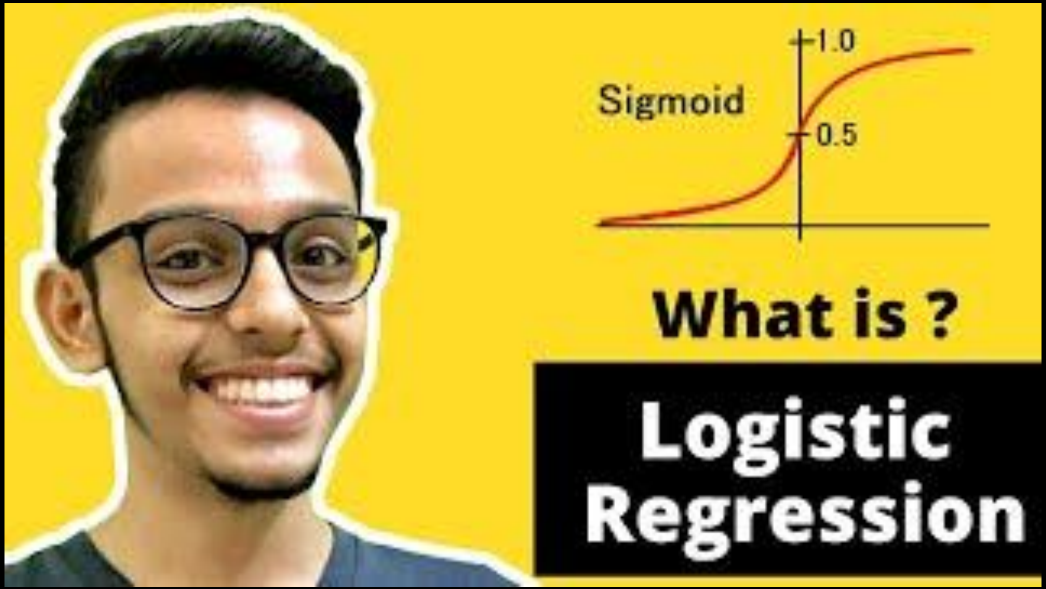
Link - https://www.youtube.com/watch?v=UR2ag4lbBtc

Time - 7m 17s

Logistic Regression

Blog - https://www.geeksforgeeks.org/understanding-logistic-regression/

Video - https://www.youtube.com/watch?v=U1omz0B9FTw



Conclusion