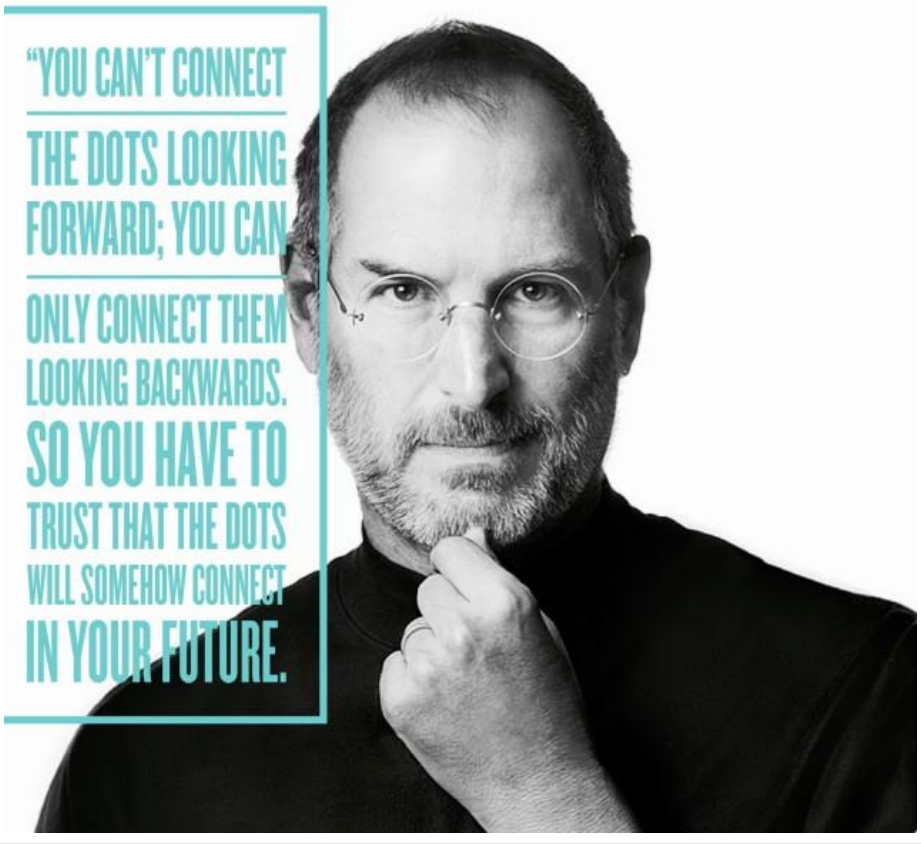


Stay Hungry Stay Foolish!!

Machine Learning

It is well for the heart to be naive and for the mind not to be.
Anatole France

Setting the Context : Machine Learning, Data Mining etc…

Machine learning is the practice of applying algorithmic models to data, in an iterative manner, so that your computer discovers hidden patterns or trends that you can use to make predictions. It's also called *algorithmic learning.* Machine learning has a vast and ever-expanding assortment of use cases, including

- Real-time Internet advertising
- Internet marketing personalization
- Internet search
- Spam filtering
- Recommendation engines
- Natural language processing and sentiment analysis
- Automatic facial recognition
- Customer churn prediction
- Credit score modeling
- Survival analysis for mechanical equipment

"I am always ready to learn although I do not always like being taught."
Winston Churchill

Basic Concept of Classification (Data Mining)

Data Mining: Data mining in general terms means mining or digging deep into data which is in different forms to gain patterns, and to gain knowledge on that pattern. In the process of data mining, large data sets are first sorted, then patterns are identified and relationships are established to perform data analysis and solve problems.

Classification: Classification is the problem of identifying to which of a set of categories (subpopulations), a new observation belongs to, on the basis of a training set of data containing observations and whose categories membership is known.

What Is Expected from ML/DS in the business?

- Predicting whether an email message is spam or not
- Predicting whether a credit card transaction is fraudulent
- Predicting which advertisement a shopper is most likely to click on
- Predicting which football team is going to win the Super Bowl

TYPES OF MODEL

Supervised models (in which there is a set of data labeled with the correct answers to learn from)

Unsupervised models (in which there are no such labels).

Semi supervised (in which only some of the data are labeled)

Reinforcement learning (in which decision is taken based on previous step)

Understanding the Data Types

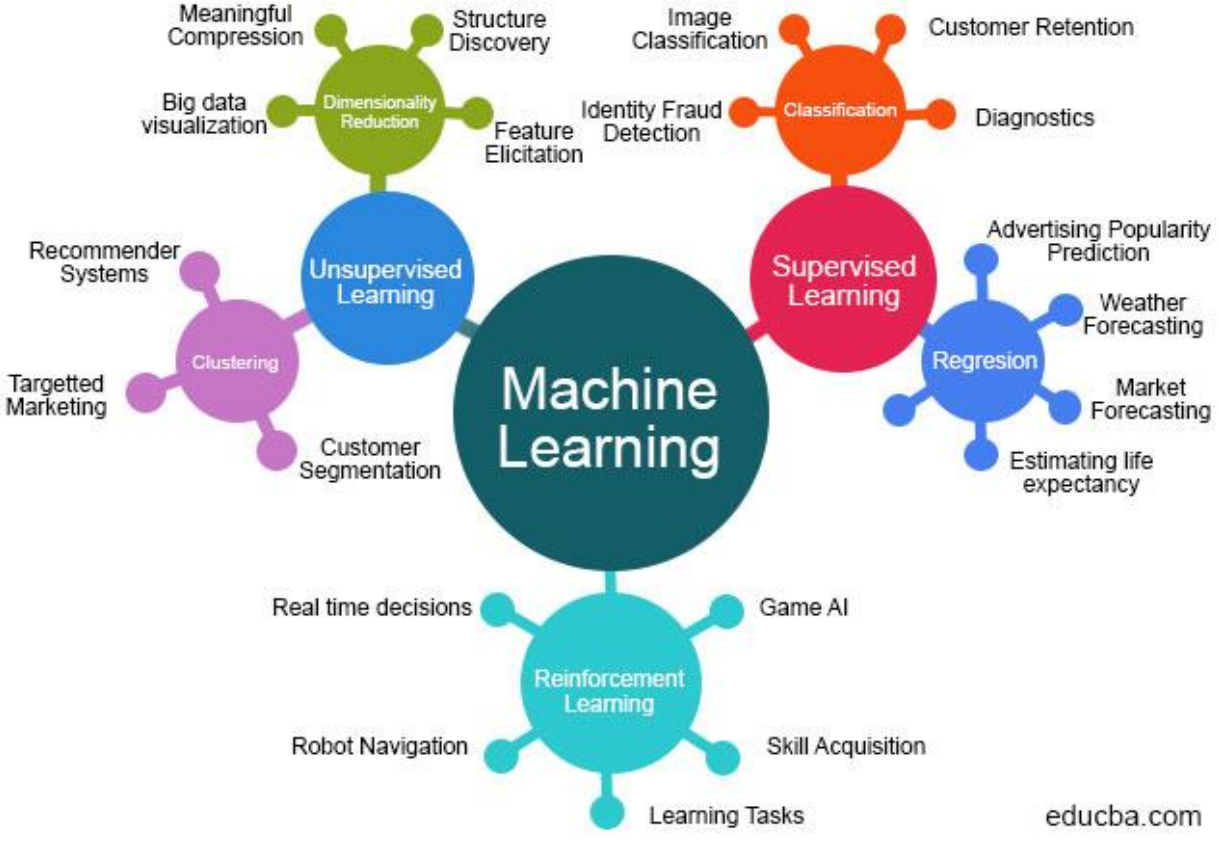
- **Nominal**: When more than two outcomes are possible. It is in Alphabet form rather than being in Integer form.
  Example: One needs to choose some material but of different colors. So, the color might be Yellow, Green, Black, Red.
  Different Colors: Red, Green, Black, Yellow

- **Ordinal**: Values that must have some meaningful order.
  Example: Suppose there are grade sheets of few students which might contain different grades as per their performance such as A, B, C, D
  Grades: A, B, C, D

- **Continuous**: May have infinite number of values, it is in float type
  Example: Measuring weight of few Students in a sequence or orderly manner i.e. 50, 51, 52, 53
  Weight: 50, 51, 52, 53

- **Discrete**: Finite number of values.
  Example: Marks of a Student in a few subjects: 65, 70, 75, 80, 90
  Marks: 65, 70, 75, 80, 90

Machine Learning Algorithms



Selecting algorithms:

Naïve Bayes method: If you want to predict the likelihood of an event occurring based on some evidence in your data

Instance-based algorithm: If you want to use observations in your dataset to classify new observations based on similarity, you can use this type. To model with instances, you can use methods like k-nearest neighbor classification

Clustering algorithm: You can use this type of unsupervised machine learning method to uncover subgroups within an unlabeled dataset

Decision tree: A tree structure is useful as a decision-support tool. You can use it to build models that predict for potential fallouts that are associated with any given decision

Association rule learning algorithm: This type of algorithm is a rule-based set of methods that you can use to discover associations between features in a dataset.

Naïve Bayes

It is well for the heart to be naive and for the mind not to be.
Anatole France

Cluster Analysis

Process of grouping data into classes/clusters, so that objects within a cluster have high similarity in comparison to one another but are very dissimilar to objects in other clusters

- Clusters defined based on similarity among the objects in a cluster
- Clusters are also defined based on difference among the objects across separate cluster
- Cluster analysis as such is not an automatic task, but an iterative process of knowledge discovery

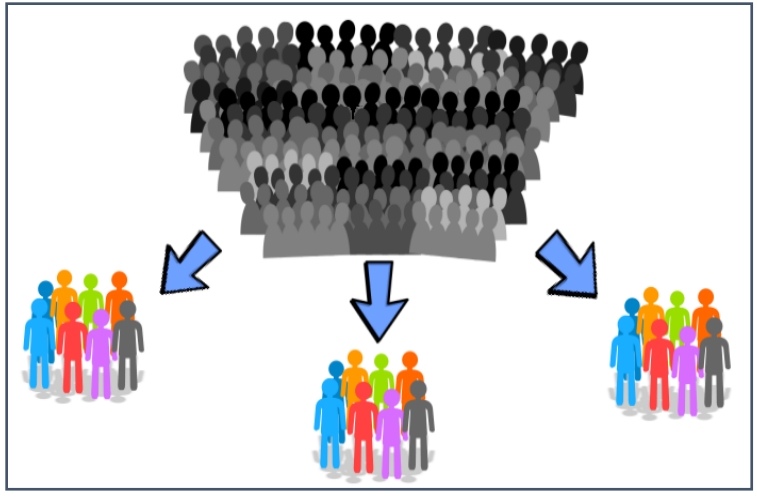
E.g. May be used to identify clusters between online shopping consumers

Business Questions and Answer through Cluster Analysis

- Market **segmentation** involves dividing a broad target market into subsets of consumers who have common needs, characteristics or behaviors.
- Helps firms identify sections of the markets that they can serve best and maximize their returns on investment.
- Critical to identify the optimal balance of satisfying diverse customer needs in different segments.

Types of Segmentation methods

Post-hoc Segmentation



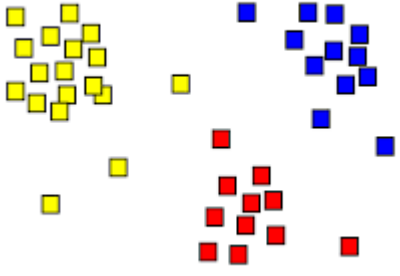
K-Means, Hierarchical,
Two-Step, Latent Class

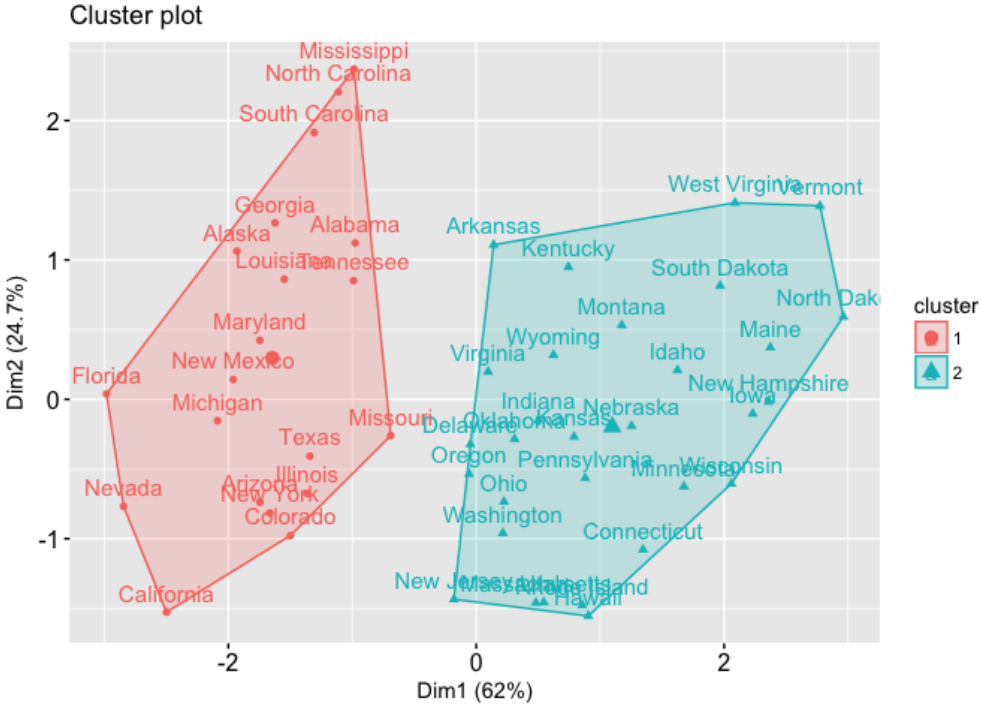
Apriori Segmentation

Discriminant

Types of Cluster Analysis

K-Means,
Hierarchical,
Two-Step





K-Means Clustering

Step 1
- K starting points are selected from within the data, where k is equal to the number of clusters anticipated

Step 2
- The Euclidean distance between a case and each starting point is calculated and the case is assigned to the starting point to which they are closest.

Step 3
- Once all cases are classified, variable means are calculated for each cluster.

Step 4
- The starting points are then replaced with these means for each cluster.

Step 5
- Steps 2 through 4 are repeated, with cases being reclassified as appropriate. These steps are repeated until further iterations will not lead to reclassifying any of the cases.

K-Means Clustering – Simple Example

$$Distance\ measure\ for\ cluster\ identification = \sqrt[2]{\sum (x_1 - x_2)^2}$$

Let a student be considered as a vector of different scores

Student Score: (HS Score, UG Score, CAT Score)
– Ashis (92%, 82%, 95%)
– Neha (96%, 75%, 98%)
– Sumod (78%, 75%, 96%)
– Javed (95%, 78%, 99%)
– Megha (92%, 75%, 98%)

Conduct an iteration of cluster analysis on the student scores.

K-Means Clustering – Simple Example

Let a student be considered as a vector as Student(HS Score, UG Score, CAT Score)
– Ashis (92%, 82%, 95%), Neha (96%, 75%, 98%), Sumod (78%, 75%, 96%), Javed (95%, 78%, 99%), Megha (92%, 75%, 98%)

- Let Ashis (92, 82, 95) and Neha (96, 75, 98) be the initial cluster centroids 1 & 2
- Try checking Sumod's RMS distance from these cluster centers

– Sumod – Ashis = [(92-78)2 + (82-75)2 + (95-96)2]^0.5 = 15.93
– Sumod – Neha = [(96-78)2 + (75-75)2 + (98-96)2]^0.5 = 18.00

- New cluster centroids are as follows:

– Cluster 1 = (85, 78.5, 96.5), Cluster 2 = (96, 75, 98)

- Try checking Javed's RMS distance from these clusters

– Javed – Cluster 1 = 10.31
– Javed Cluster 2 = 3.31

- New cluster centroids are as follows:

– Cluster 1 = (85, 78.5, 96.5), Cluster 2 = (95.5, 76.5, 98.5)

- Which cluster should Megha fall under?

– Megha-C1: 7.97; Megha - C2: 3.84
– New C2 [94.3, 76.0, 98.3]

- Continue the clustering iteratively, till there is no change in cluster allocations

20-

RMS – Root Mean Square

Hierarchical clustering

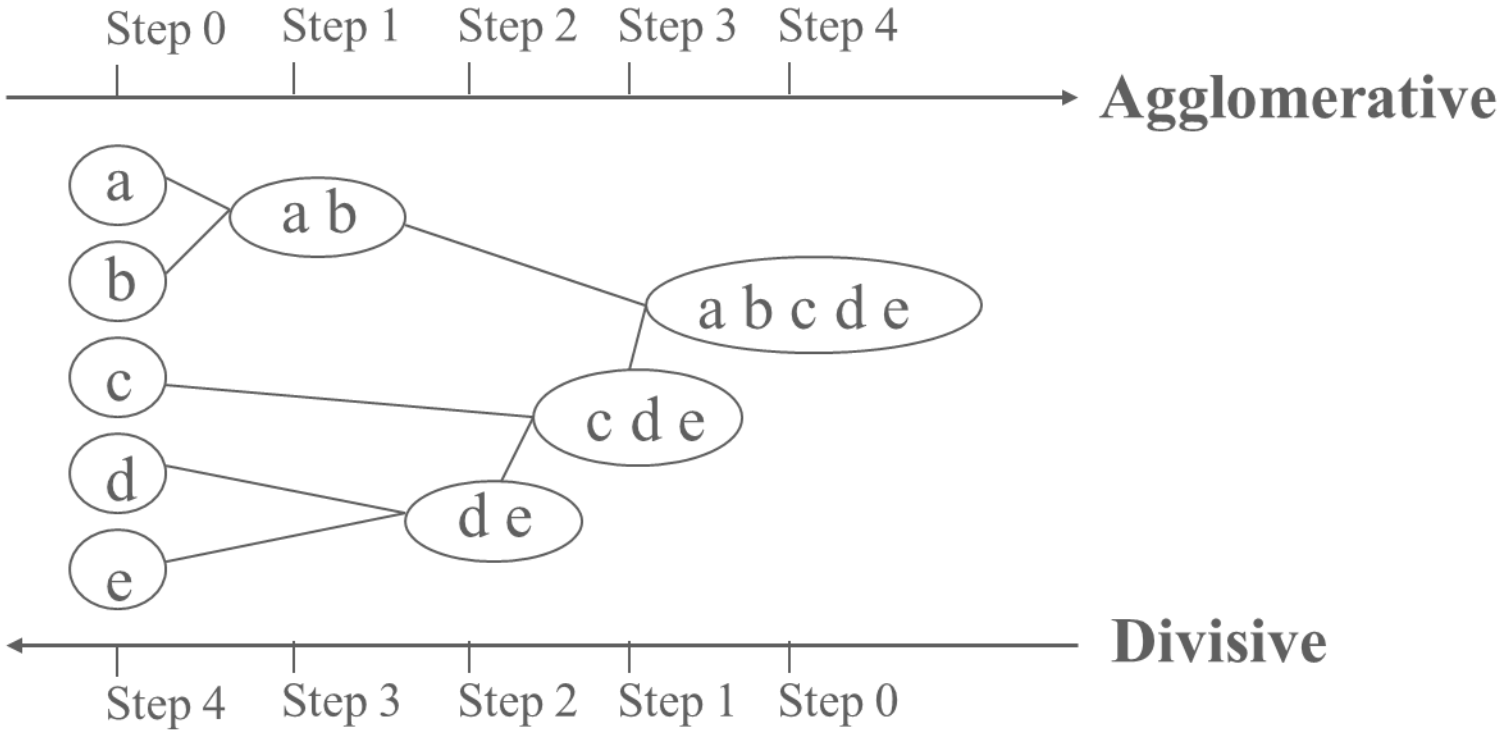
- Hierarchical clustering is a method of Cluster Analysis which seeks to build a hierarchy of clusters.

- Algorithms for hierarchical clustering generally fall into two types –

  - Agglomerative: This is a "bottom up" approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.

  - Divisive: This is a "top down" approach: all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.

Hierarchical clustering



Hierarchical clustering

- The basic criterion for any clustering is distance. Objects that are near each other should belong to the same cluster, and objects that are far from each other should belong to different clusters. For a given set of data, the clusters that are constructed depend on your specification of the following parameters:

  - Cluster method defines the rules for cluster formation. For example, when calculating the distance between two clusters, you can use the pair of nearest objects between clusters or the pair of furthest objects, or a compromise between these methods.

  - Measure defines the formula for calculating distance. For example, the Euclidean distance measure calculates the distance as a "straight line" between two clusters. Interval measures assume that the variables are scale; count measures assume that they are discrete numeric; and binary measures assume that they take only two values.

  - Standardization allows you to equalize the effect of variables measured on different scales.

Hierarchical clustering

- Advantages of using Hierarchical Clustering –
  - ➢ Works well for relatively smaller datasets
  - ➢ Provides the capability of clustering of cases as well as variables

Two-step clustering

- Step 1: Pre-clustering
  - The first step of the two-step procedure is the formation of “pre-clusters”. The goal of pre-clustering is to reduce the size of the matrix that contains distances between all possible pairs of cases.
  - As a case is read, the algorithm decides, based on a distance measure, if the current case should be merged with a previously formed pre-cluster or start a new pre-cluster. When pre-clustering is complete, all cases in the same pre-cluster are treated as a single entity. The size of the distance matrix is no longer dependent on the number of cases but on the number of pre-clusters.

- Step 2: Clustering of Pre-clusters
  - In the second step, SPSS uses the standard hierarchical clustering algorithm on the pre-clusters. Forming clusters hierarchically lets you explore a range of solutions with different numbers of clusters.

Two-step clustering

- Advantages of using Two-Step Clustering –
  - ➢ Handles both continuous and categorical variables
  - ➢ Uses a two-step approach that works faster for large datasets
  - ➢ Provides the capability to automatically find the optimal number of clusters

Association Rules

Association Rule Mining

Association Rule Mining is used when you want to find an association between different objects in a set, find frequent patterns in a transaction database, relational databases or any other information repository. The applications of Association Rule Mining are found in Marketing, Market Basket Analysis) in retailing, clustering and classification. It can tell you what items do customers frequently buy together by generating a set of rules called **Association Rules**. In simple words, it gives you output as rules in form if this then that. Clients can use those rules for numerous marketing strategies:

- Changing the store layout according to trends
- Customer behavior analysis
- Catalogue design
- Cross marketing on online stores
- What are the trending items customers buy
- Customized emails with add-on sales

The concept & context

Market Basket Analysis takes data at transaction level, which lists all items bought by a customer in a single purchase. The technique determines relationships of what products were purchased with which other product(s). These relationships are then used to build profiles containing If-Then rules of the items purchased.

Other Application Areas

- Analysis of credit card purchases.
- Analysis of telephone calling patterns.
- Identification of fraudulent medical insurance claims. (Consider cases where common rules are broken).
- Analysis of telecom service purchases.

Decision Trees

Decision trees

Decision trees are a series of sequential steps designed to answer a question and provide probabilities, costs, or other consequence of making a particular decision

Decision tree algorithms are perfect to solve **classification** (where machines sort data into classes, like whether an email is spam or not) and **regression** (where machines predict *values*, like a property price)

Regression Trees are used when the dependent variable is **continuous** or **quantitative** (e.g. if we want to estimate the probability that a customer will default on a loan)

Classification Trees are used when the dependent variable is **categorical** or **qualitative** (e.g. if we want to estimate the blood type of a person).

Application of Decision Trees

- Healthcare industry to improve the screening of positive cases in the early detection of cognitive impairment, and also to identify the main risk factors of developing some type of dementia in the future.

- Chatbots use these algorithms for gathering data from customers through the application of innovative surveys and friendly chats.

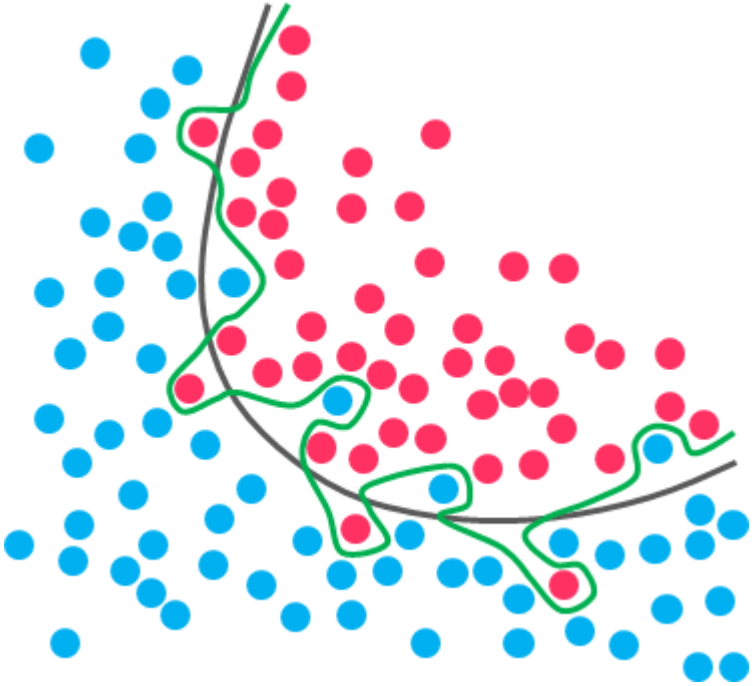
- To perform sentiment analysis of texts

- These are used to improve financial fraud detection

Overfitting

Overfitting refers to a model that learns the training data(the data it uses to learn) so well that it has problems to generalize to new (unseen) data.

In other words, the model learns the detail and noise (irrelevant information or randomness in a dataset) in the training data to the extent that it negatively impacts the performance of the model on new data. This means that the noise or random fluctuations in the training data is picked up and learned as concepts by the model.



While the black line fits the data well, the green line is overfitting

Pruning to avoid overfitting?

Pruning is a technique used to deal with overfitting, that reduces the size of Decision Trees by removing sections of the Tree that provide little predictive or classification power.

The goal of this procedure is to reduce complexity and gain better accuracy by reducing the effects of overfitting and removing sections of the DT that may be based on noisy or erroneous data. There are two different strategies to perform pruning

Pre-prune: When you stop growing Decision Trees branches when information becomes unreliable.

Post-prune: When you take a fully grown Decision Trees and then remove leaf nodes only if it results in a better model performance. This way, you stop removing nodes when no further improvements can be made.

Algorithms used in Decision Trees - CART

CART is an algorithm that produces binary Classification or Regression Trees, depending on whether the dependent (or target) variable is categorical or numeric, respectively.

It handles data in its raw form (no preprocessing needed), and can use the same variables more than once in different parts of the same Trees, which may uncover complex interdependencies between sets of variables.

In the case **of Classification Trees**, CART algorithm uses a metric called Gini Impurity to create decision points for classification tasks. Gini Impurity gives an idea of how fine a split is (a measure of a node's "purity"), by how mixed the classes are in the two groups created by the split.

In the case of **Regression Trees**, CART algorithm looks for splits that minimize the Least Square Deviation (LSD), choosing the partitions that minimize the result over all possible options. The LSD (sometimes referred as "variance reduction") metric minimizes the sum of the squared distances (or deviations) between the observed values and the predicted values. The difference between the predicted and observed values is called "residual", which means that LSD chooses the parameter estimates so that the sum of the squared residuals is minimized.

Algorithms used in Decision Trees - CHAID

The CHAID stands for Chi-squared Automatic Interaction Detection (CHAID).
This is one of the oldest algorithms methods that produces multiway Decision Tress inn which the splits can have more than two branches

For **Classification Trees** (where the dependent variable is categorical in nature), CHAID relies on the Chi-square independence tests to determine the best split at each step. Chi-square tests check if there is a relationship between two variables, and are applied at each stage of the Decision Tree to ensure that each branch is significantly associated with a statistically significant predictor of the response variable.

For **Regression Trees** (where the dependent variable is continuous), CHAID relies on F-tests (instead of Chi-square tests) to calculate the difference between two population means. If the F-test is significant, a new partition (child node) is created (which means that the partition is statistically different from the parent node). On the other hand, if the result of the F-test between target means is not significant, the categories are merged into a single node.

Algorithms used in Decision Trees – ID3 and C4.5

The Iterative Dichotomiser 3 (ID3) is mostly used for classification tasks

ID3 splits data attributes (dichotomizes) to find the most dominant features, performing this process iteratively to select the DT nodes in a top-down approach.

For the splitting process, ID3 uses the Information Gain metric to select the most useful attributes for classification. Information Gain is a concept extracted from Information Theory, that refers to the decrease in the level of randomness in a set of data: basically it measures how much "information" a feature gives us about a class. ID3 will always try to maximize this metric, which means that the attribute with the highest Information Gain will split first.

ID3 has some disadvantages: it can't handle numeric attributes nor missing values, which can represent serious limitations.

C4.5 is the successor of ID3 and represents an improvement in several aspects. C4.5 can handle both continuous and categorical data, making it suitable to generate Regression and Classification Trees.

Disadvantages of Decision Trees

- Decision Trees tend to overfit on their training data, making them perform badly if data previously shown to them doesn't match to what they are shown later.
- Decision Trees suffer from high variance, which means that a small change in the data can result in a very different set of splits, making interpretation somewhat complex
- In Classification Trees, the consequences of misclassifying observations are more serious in some classes than others.
- Decision Trees can create biased Trees if some classes dominate over others.
- In the case of Regression Trees, Decision Trees can only predict within the range of values they created based on the data they saw before, which means that they have boundaries on the values they can produce.
- Decision Trees algorithms grow Trees one node at a time according to some splitting criteria and don't implement any backtracking technique

Why Use Decision Trees?

Ensemble methods combine several Decision Trees to improve the performance of single Decision Trees

The Two most common techniques to perform ensemble Decision Trees are Bagging and Boosting.

The idea is to train multiple models using the same learning algorithm to achieve superior results.

Bagging and Boosting

Bagging (or Bootstrap Aggregation) is used when the goal is to reduce the variance of a DT. Variance relates to the fact that DTs can be quite unstable because small variations in the data might result in a completely different Tree being generated. So, the idea of Bagging is to solve this issue by creating in parallel random subsets of data (from the training data), where any observation has the same probability to appear in a new subset data. Next, each collection of subset data is used to train DTs, resulting in an ensemble of different DTs. Finally, an average of all predictions of those different DTs is used, which produces a more robust performance than single DTs.

Random Forest is an extension over Bagging, which takes one extra step: in addition to taking the random subset of data, it also takes a random selection of features rather than using all features to grow DTs.

Boosting is another technique that creates a collection of predictors to reduce the variance of a DT, but with a different approach. It uses a sequential method where it fits consecutive DTS, and at every step, tries to reduce the errors from the prior Tree. With Boosting techniques, each classifier is trained on data, taking into account the previous classifier success. After each training step, the weights are redistributed based on the previous performance. This way, misclassified data increases its weights to emphasize the most difficult cases, so that subsequent DTs will focus on them during their training stage and improve their accuracy.

Classification and Regression Trees

Context

We use regression techniques to predict variables.

For example: Categorical or continuous variable can be predicted from one or more predictor variables using logistic and linear regression, respectively.

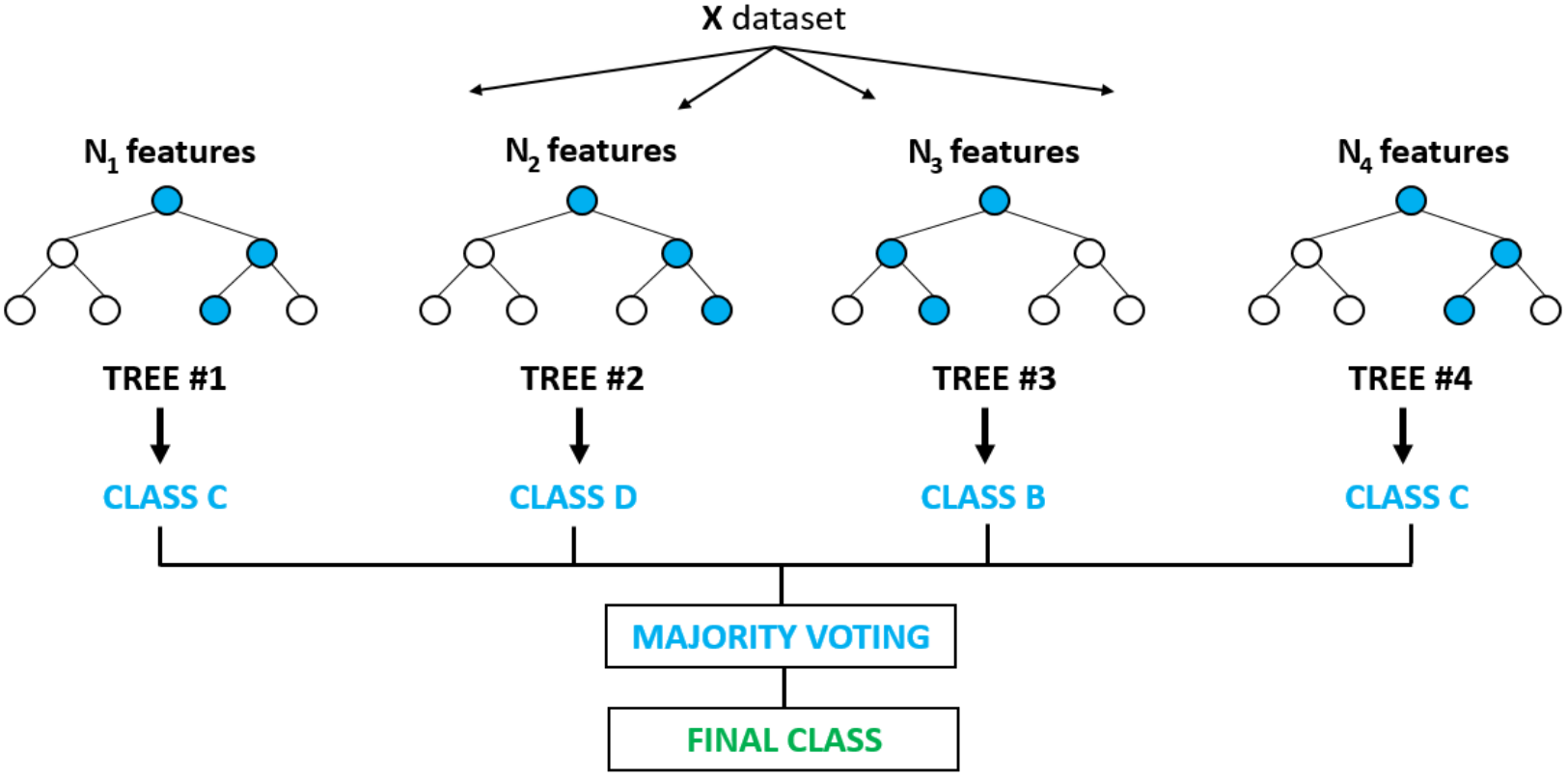
CART is a predictive model which helps to find a variable based on other labeled variables.

This is simple but powerful approach to prediction.

Unlike logistic and linear regression, CART does not develop a prediction equation. Instead, data are partitioned along the predictor axes into subsets with homogeneous values of the dependent variable

a process represented by a decision tree that can be used to make predictions from new observations.

Random Forests Explained



Random Forests and Boosted Trees

- Examples of “ensemble” methods, “Wisdom of the Crowd” (Chap 13)
- Predictions from many trees are combined
- Very good predictive performance, better than single trees (often the top choice for predictive modeling)
- Cost: loss of rules you can explain implement (since you are dealing with many trees, not a single tree)
  - However, RF does produce “variable importance scores,” (using information about how predictors reduce Gini scores over all the trees in the forest)

Random Forests

Random forest classifier creates a set of decision trees from randomly selected subset of training set. It then aggregates the votes from different decision trees to decide the final class of the test object.

In short, using Random Forest we generate multiple small decision trees from random subsets of the data (hence the name “Random Forest”).

- Each of the decision tree gives a biased classifier since it is based on subset of the data
- For solving the classification problem using Random Forest we take majority vote to classify the data
- For Regression we consider average of all trees as out prediction

Random Forests

Pros:
One of the most accurate decision models.
Works well on large datasets.
Can be used to extract variable importance.
Do not require feature engineering (scaling and normalization)
Cons:
Over fitting in case of noisy data.
Unlike decision trees, results are difficult to interpret.

Applications
Random forests have successfully been implemented in a variety of fields. Some applications include:

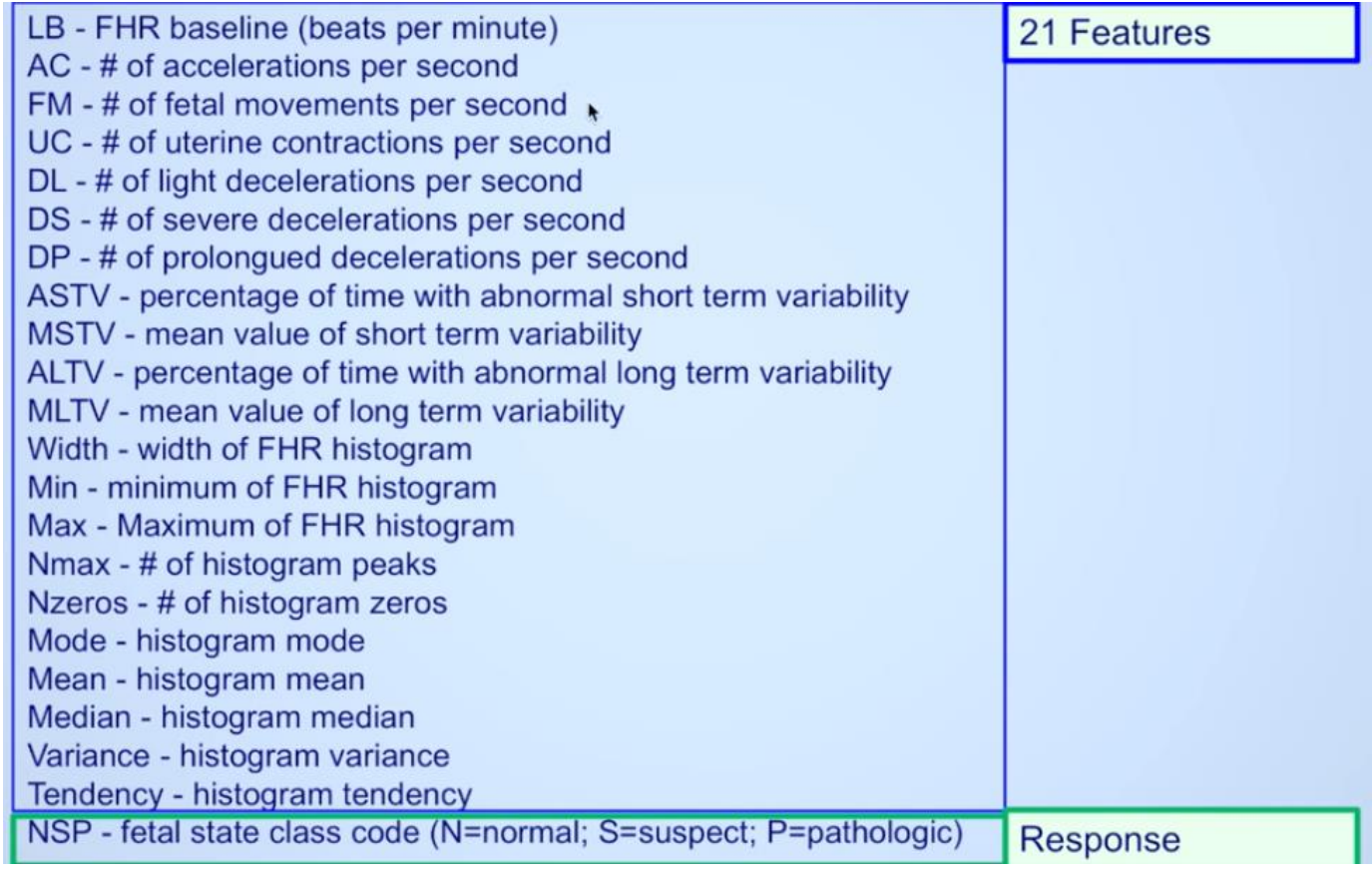
- Object recognition.
- Molecular Biology (Analyzing amino acid sequences)
- Remote sensing (Pattern recognition)
- Astronomy (Star Galaxy classification, etc)

Example of Random Forest – Data – CTG.csv

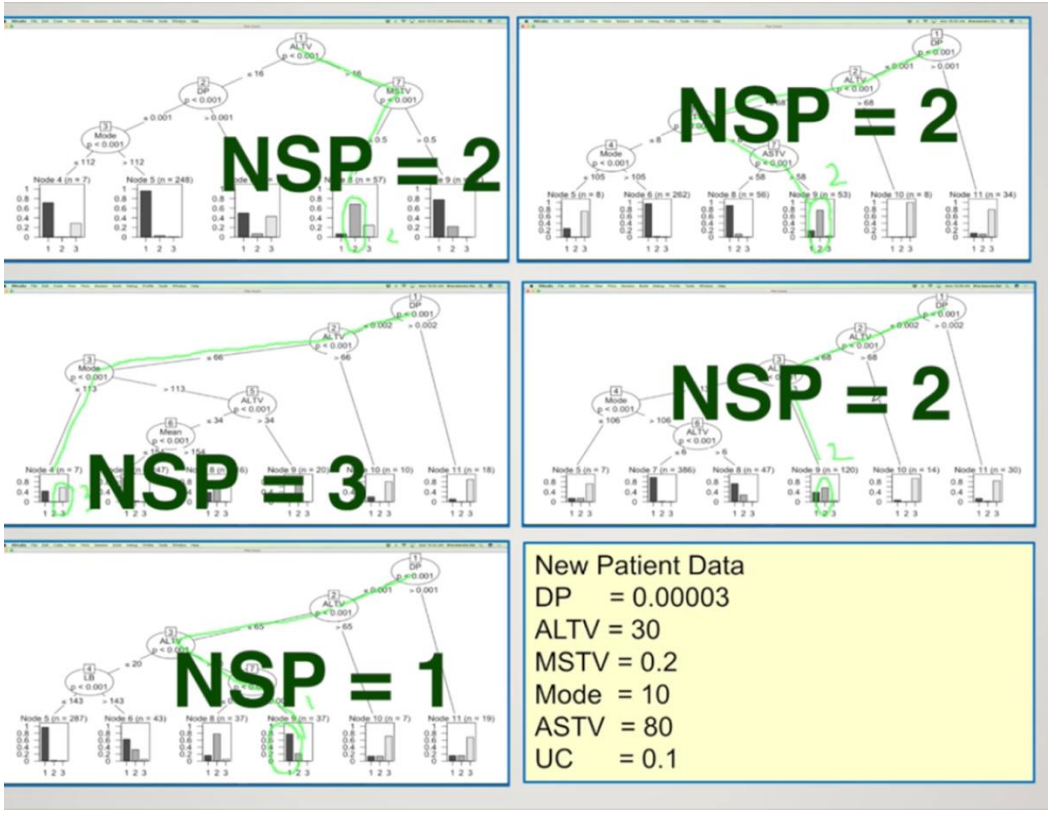
- Measurements of fetal heart rate (FHR) and uterine contraction (UC) features on cardiotocograms
- 2126 fetal cardiotocograms (CTGs) automatically processed and diagnostic features measured
- CTGs classified by three expert obstetricians and consensus classification label as Normal, Suspect, or Pathologic.

Example of Random Forest – Data – CTG.csv

CTG.csv



Selection


NSP = 2
NSP = 2
NSP = 3
NSP = 2
NSP = 1
New Patient Data
DP = 0.00003
ALTV = 30
MSTV = 0.2
Mode = 10
ASTV = 80
UC = 0.1