**Machine Learning Algorithms**

Aswathy Kuttisseril Jewel (C0813455)

Gayathri Ravi Nath (C0818959)

Prasanth Moothedath Padmakumar (C0796752)

Cloud Computing for Big Data, Lambton College

CBD 3335 – Data Mining and Analysis

Parisa Naraei

May 17, 2022

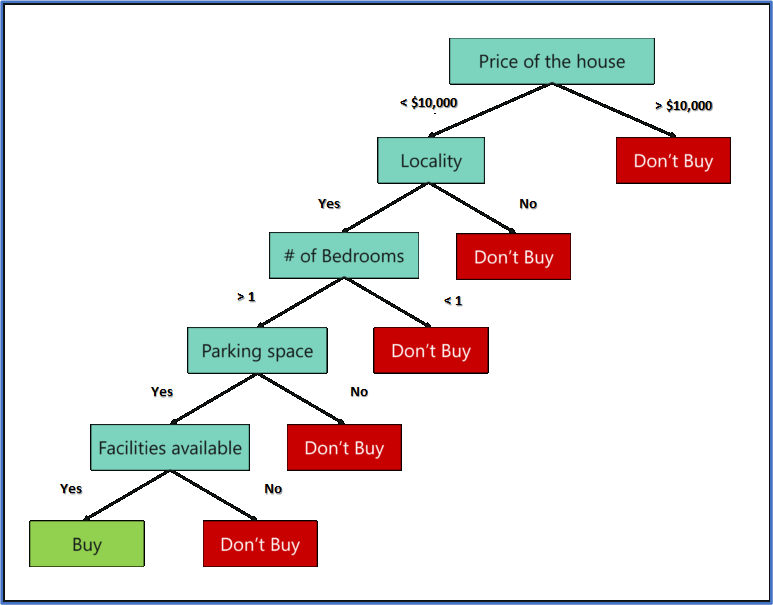
**Random Forest**

Random forest is one of a machine learning algorithm used for both classification and regression analysis. Rather than using just one decision tree, a number decision trees are used to predict and the results are aggregated to provide one result. The problem of using just one decision tree is that it is prone to overfitting, which means it is highly sensitive to the training data set. A slight change in the dataset can lead to change in the prediction outcome [1].

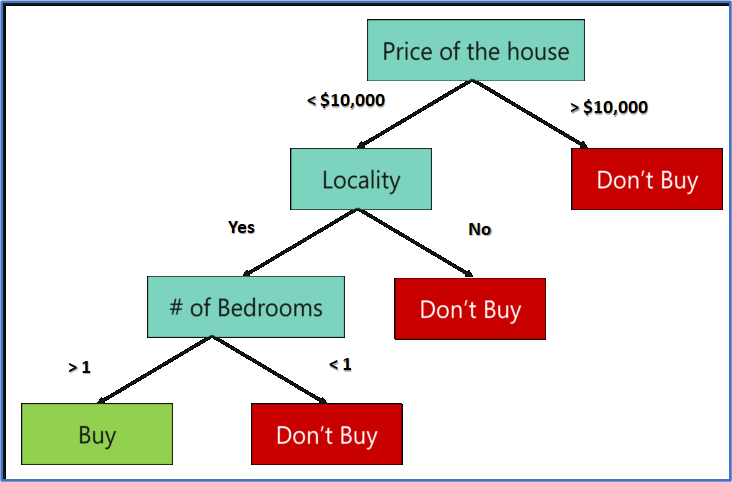
When choosing the different decision trees, we have to make sure that each of the features are present in at least one of the tree and not all of the rows and all of the features does not appear in one single tree. The number of feature in each of these sub trees will be same. This will ensure that the different trees are not correlated.

Following is a scenario explaining random forest algorithm, when one wants to buy a house he will have will a have set of questions to decide whether he needs to buy, these parameters are our predictors. A decision tree with all the questions incorporated will be the base decision tree and from that we derive different sub trees and aggregate each result to get our final prediction

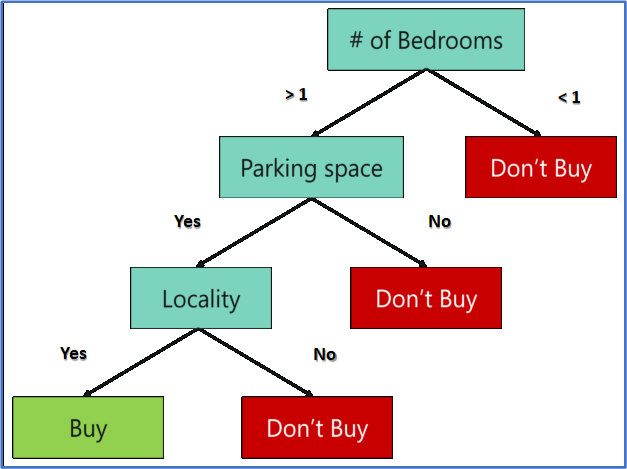
Base decision tree is



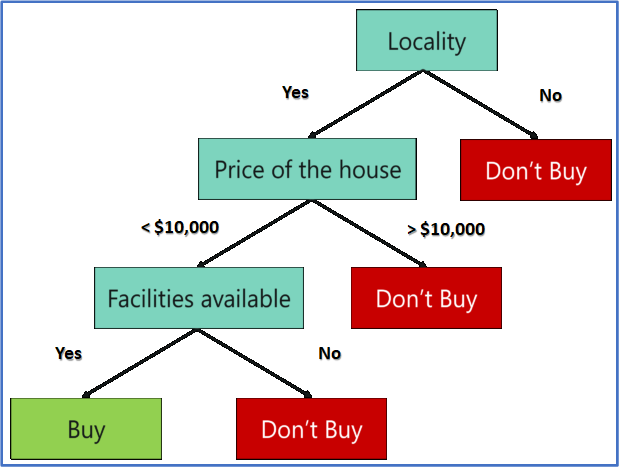
From all the decision tree randomly select some feature and form a sub tree, following is one such sub tree, here only three features from the total five features are selected



The second sub tree will also have three features



The third sub tree is



After using this approach to create numerous Decision trees, each tree selects or votes on the class (in this case, the decision trees will choose whether or not a house is bought). The predicted class is the one that receives the most votes by a simple majority. To summarise, decision trees are constructed utilising all predictor factors and the whole data set. Random Forests, on the other hand, are used to generate several decision trees, each of which is based solely on a portion of the data set.

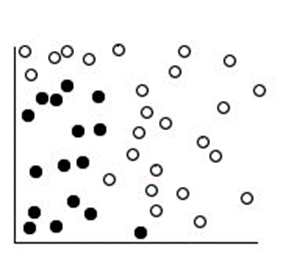
**Support Vector Machine**

A support vector machine (SVM) is a deep learning technique that uses supervised learning for data group classification or regression. It used to classify two data groupings into categories that are similar. The algorithms use hyperplanes (decision boundary that differentiates the two classes in SVM) to divide the groupings into categories based on patterns.

SVM categorizes data points by mapping them to a high-dimensional feature space, even when the data is not otherwise linearly separable. After finding a separator between the categories, the data are converted so that the separator can be drawn as a hyperplane. Following that, fresh data features can be utilized to predict which category a new record should belong to. [2]

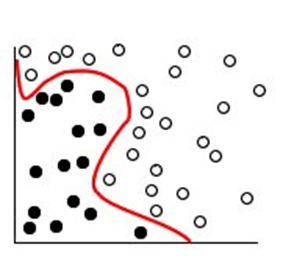
For example, consider the following figure, in which the data points fall into two different categories.

Figure 1. Original dataset



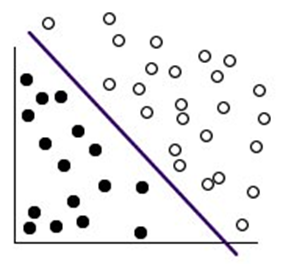
The two categories can be separated with a curve, as shown in the following figure.

Figure 2. Data with separator added



After the transformation, the boundary between the two categories can be defined by a hyperplane, as shown in the following figure.

Figure 3. Transformed data



**XGBoost**

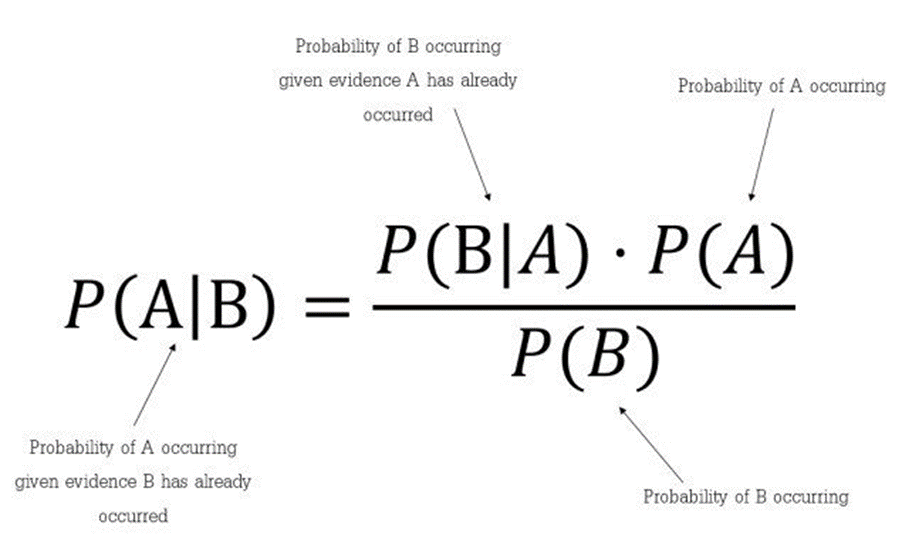
XGBoost is an open-source implementation of the gradient boosted trees algorithm that is popular and efficient. Gradient boosting is a supervised learning algorithm, which attempts to accurately predict a target variable by combining the estimates of a set of simpler, weaker models. When using gradient boosting for regression, the weak learners are regression trees, and each regression tree maps an input data point to one of its leafs that contains a continuous score. XGBoost minimizes a regularized (L1 and L2) objective function that combines a convex loss function (based on the difference between the predicted and target outputs) and a penalty term for model complexity (in other words, the regression tree functions). The training proceeds iteratively, adding new trees that predict the residuals or errors of prior trees that are then combined with previous trees to make the final prediction. It's called gradient boosting because it uses a gradient descent algorithm to minimize the loss when adding new models. [3]

**Naive Bayes Classifier**

A classifier is a machine learning model that is used to discriminate different objects based on certain features.

Naïve Bayes Classifier is a probability based machine learning model used to perform the classification task.

The equation is as follows,

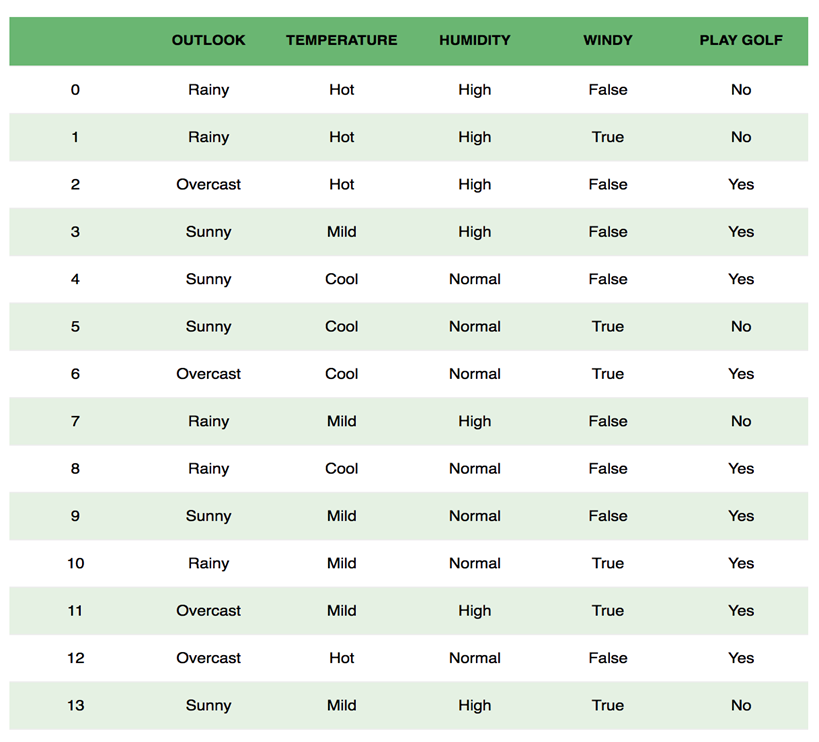


**Bayes Theorem:**

With the Bayes theorem we can predict the chances of B happening with the given fact that A has already occurred. Here B is the evidence and A is the hypothesis. Here, the predictors/features are supposed to be independent. That is, the existence of one feature does not affect the presence of another. So it is called Naive.

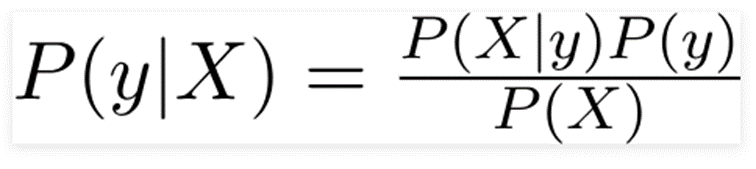
For an example we take the situation of playing golf.

The dataset is as represented below,

****

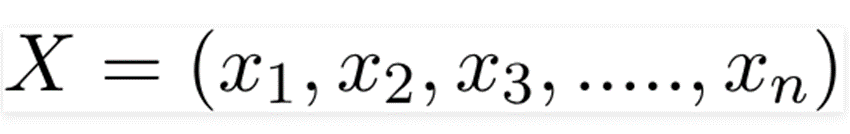
We determine whether the day is suitable for playing golf, considering the features of the day that affects the weather. The columns represents the individual entries and rows represents the features. When we take the first row, we can see that the day is not suitable to play golf since the outlook is rainy, temperature is hot, humidity is high and it is not windy. As we discussed earlier, the predictors are independent. That if the temperature is high, it is not necessary to be the humidity have to be high. Another assumption is that all the predictors make an equal effect on the outcome.

According to this example, Bayes theorem can be rewritten as:



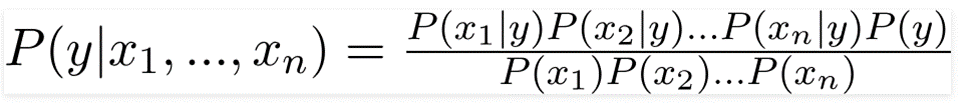
The variable **y** is the class variable(play golf), which represents if it is suitable to play golf or not given the conditions. Variable **X** represent the parameters/features.

X is given as,

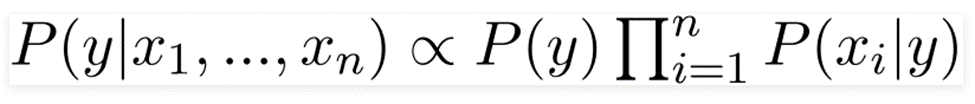


X represents the features, the factors determine the weather of the day.

By substituting for X we get,



You can now seek up the values for each from the dataset and plug them into the equation. The denominator does not vary for all entries in the collection; it remains constant. As a result, the denominator can be eliminated and proportionality introduced.



**References**

<https://towardsmachinelearning.org/random-forest/>

<https://www.ibm.com/docs/it/spss-modeler/SaaS?topic=models-how-svm-works>

<https://docs.aws.amazon.com/sagemaker/latest/dg/xgboost-HowItWorks.html>

<https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c>