# Machine Learning Assignment 2

In this study, the proposed approach used is the Decision Tree algorithm using SKLearn.

## Decision Tree:

Decision tree is one of the predictive modelling approaches used in Machine Learning. It can be used for both a classification problem and regression problem.

### How does a decision tree work?

The logic behind the decision tree, shows a tree-like structure. Decision trees classify instances by sorting them down from the top of the tree down to leaf node, which provides the classification of the instance. Each node in the tree specifies a test of some attribute of the data. Each branch descending from a node corresponds to one of the possible values for the attribute.

### Advantages of decision Tree:

* Highly intuitive and ease to understand
* Less number of data preparation
* Highly versatile algorithm and can perform multiple roles apart from the standard predictions.

### Why?

I mainly chose decision tree as it is the classifier that I understand the most but also because its advantages and because it works greatly with binary answers such as the one the prediction is trying to find, “typeA or typeB”

## How you solved the problem and any decisions you had to make:

1. The features given on the data description were added to the training and queries as the headers for these data sets. Also, the “unknown” values were replaced by NaN values; once replaced by NaN Values, these values were once more replaced by zeroes.

2. A function has been made to replace from categorical to continuous data, this returns the values of the data e.g., JobCat1 => 1

3. All the data inputted are non-null values, meaning that we have a value for every column.

4. Divide the training data between data and the outcome of the data.

5. Some of the features were dropped as these columns were not needed or this did not affect the accuracy values for the prediction; the columns that remained were: age, Job, marital education, balance, housing, loan, and month. These new features were added to both the training and queries data sets as the new headers removing all the non-mentioned columns.

6. The data was sent through the decision tree using entropy as the criteria, for the tree to decide on the different nodes to create.

## Testing data

### Division of data

The training data was split in three parts:

1. Training, this sample was used to train the model.
2. Validation, this sample is used to verify the output of the data
3. Test, used to verify the accuracy of prediction in the queries. This sample is obtained by measuring the length of the queries dataset, using this length I randomly selected (2700) rows from the training data.

### Issues with data

I created a loop that repeated 10 times to test the accuracy of the prediction and get the average harmonic accuracy percentage of these predictions. I then manually tested the accuracy of the data by removing columns in the datasets and achieve an accuracy percentage in a range between 75% and 80%. Below are the reasons why I removed or leave certain features.

#### Continuous data:

Id; removed is not needed for the output target.

Age; I changed the values using a range: younger or equal than 25=1; younger or equal than 45=2; younger or equal than 65=3 Older than 65=4. This classification was done to reduce the number of values in the age category and increase accuracy of the prediction.

Balance; although, it had a wide range from negative to positive values, I believe this is an important feature to keep data more accurate. If this column is removed the accuracy goes too high meaning the prediction is not accurate.

Day: removed, accuracy greatly decrease if left on the training set.

Duration; removed because all data on this column was null or 0.

Campaign; removed because accuracy slightly decreases.

Pdays; removed because accuracy slightly decreases.

Previous; removed because accuracy slightly decreases.

#### Categorical data:

Job; not changed, I tried to remove it, but accuracy greatly decrease.

Marital; not changed, if remove it did not affect the accuracy. But it is an important feature to keep.

Education; not changed, if remove the accuracy greatly decreased.

Default, removed accuracy did not change once remove.

Housing, no changes if removed the accuracy greatly decreased.

Loan; no changes if remove the accuracy was greatly decreased.

Contact; too many unknown values approximately 30%, also when testing the program without this column there was not significant change to accuracy.

Month; No changes if remove the accuracy was decrease.

Poutcome; too many unknown this column had approximately 82% missing values, therefore it is not needed for the output target.