**Email Classification By Machine Learning**

**Will Smith**

**Melika Shekarriz**

**Adam Ismaeili**

**Introduction**

What is classification?

Classification is technique to categorize our data into a desired and distinct number of classes where we can assign label to each class.

One of the application of classification is email classification.

We can use different types of machine learning algorithms such as SVM,KNN,Decision Tree and Random Forest(RF).

The one we used to classify emails is Decision Tree.

**Methodology**

Decision Tree:

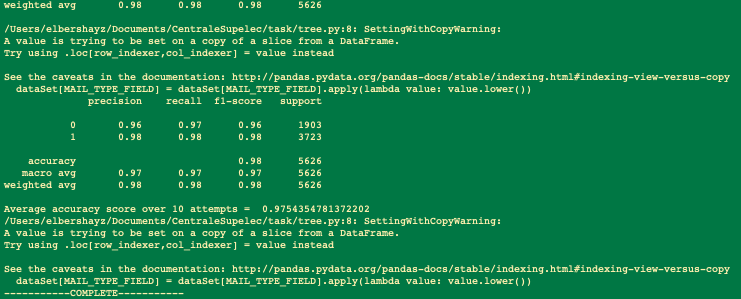
Given a data of attributes together with its classes, a decision tree produces a sequence of rules that can be used to classify the data. Decision Tree, as it name says, makes decision with tree-like model. It splits the sample into two or more homogeneous sets (leaves) based on the most significant differentiators in your input variables. To choose a differentiator (predictor), the algorithm considers all features and does a binary split on them (for categorical data, split by cat; for continuous, pick a cut-off threshold). It will then choose the one with the least cost (i.e. highest accuracy), and repeats recursively, until it successfully splits the data in all leaves (or reaches the maximum depth).

**Code:**

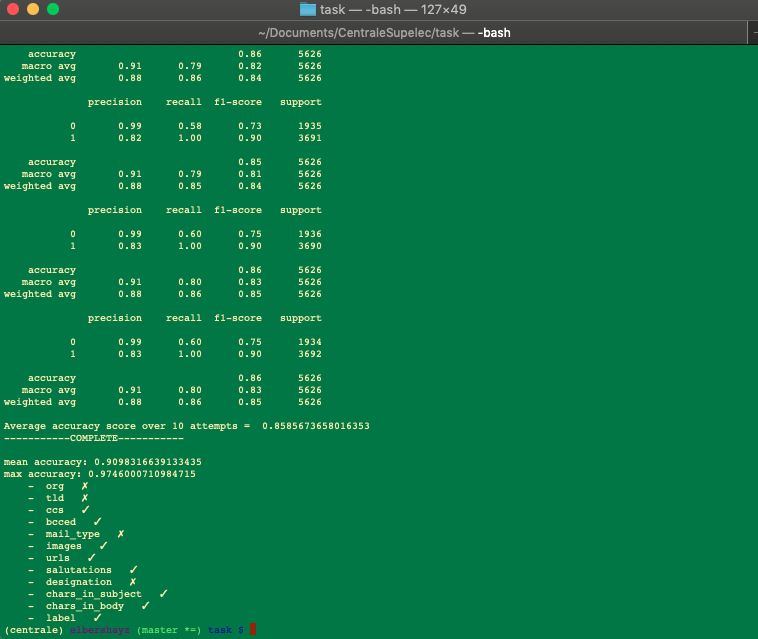
**Brief observation of what we have done:**

At first we ran the sample code and got (0.23)accuracy,then we tried decision tree without date, org, mail\_type, tld and got (0.9809) accuracy after that in the third submission we did all the same minus designations field from powerset test and got (0.97855),in 4th submission we got all features included except date and we got (0.99146) accuracy which is our best results. And at last we tried all fields except date for logistic regression and multilayer perceptron.

1. Results of first test tree without categorical features:



2. Result of testing every combination of fields (without categorical ones):



As we can see we will get better results and accuracy if we use check marked features.

**Results:**

So we got the best result when all the features are included except the date by using a decision tree and ran through a powerset of all the quantitative and tests found.

**Future optimizations:**

We will work on Random Forest and we will see how we can improve the accuracy by including the date information of the emails.