MACHINE LEARNING LAB 01

STEFANO FOCHESATTO

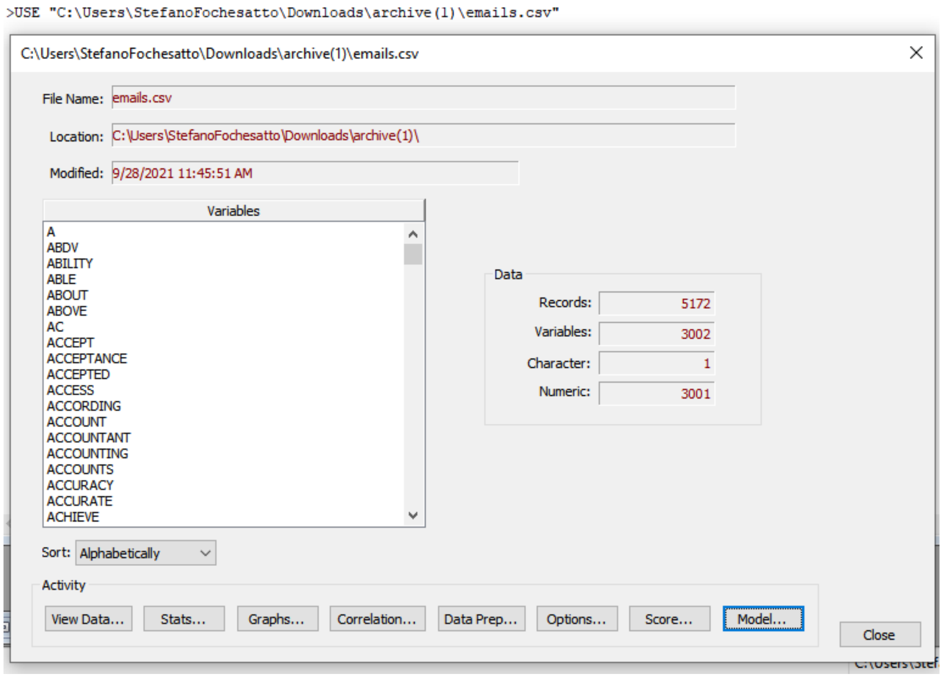
1. Introduction

In this lab we will build a CART model to detect whether or not an email is spam. In doing so we will explore SPM's CART Decision Tree Analysis Engine, and develop a data workflow with SPM. The data used throughout this lab can be found on Kaggle and is free to use for educational purposes under the Open Data Commons Open Database License. The data comes in the form of a .csv file with 5172 rows and 3002 columns. Each row represents data for a single email, the first column is used to number each email, the last column contains our prediction information(1 for spam, 0 for not spam) and the rest of the columns contain word counts for the 3000 most common words found across all emails.

Preliminary Analysis (Loading Data). Upon loading the data into SPM using either the file menu, or the USE"" command, we are met with a variety of options to perform exploratory analysis and data preparation in the 'Activity' pane. For the most part the most important tools for exploratory analysis are the View Data, Stats, Graphs, and Correlation options in this menu (these options can also be found under ther View and Explore menues). With these options we can view histograms, evaluate correlations, and compute univariate statistics like the Mean, Min, Max and IQR. Not entirely relevant to our lab but usefully nonetheless.

Date: September 29, 2021.

Figure 1. SPM Activity Pane



```
>USE "C:\Users\StefanoFochesatto\Downloads\archive(1)\emails.csv"
```

C:\Users\StefanoFochesatto\Downloads\archive(1)\emails.csv

File Name: emails.csv

Location: C:\Users\StefanoFochesatto\Downloads\archive(1)\

Modified: 9/28/2021 11:45:51 AM

Variables

A
ABDV
ABILITY
ABLE
ABOUT
ABOVE
AC
ACCEPT
ACCEPTANCE
ACCEPTED
ACCESS
ACCORDING
ACCOUNT
ACCOUNTANT
ACCOUNTING
ACCOUNTS
ACCURACY
ACCURATE
ACHIEVE

Data

Records: 5172

Variables: 3002

Character: 1

Numeric: 3001

Sort: Alphabetically

The Data Prep option allows us to do any preparation like handling missing data, duplicate data, unwanted outliers, or even fixing structural errors in our data using built in SPM commands. The Options menu gives us access to the content of our Model's text report, as well as the random number seeds for each SPM modeling engine.

2. SPM CART Model

3. Performance Analysis

4. Conclusion