Project 1

The question deals with dataset called 'authorship' and it is provided to you in two different versions: .csv and .txt. The dataset is coming from the book 'Analysing Categorical Data' by Jeffrey S. Simonoff, Springer. The dataset contains the 841 rows and 71 columns, where each row is a written chapter or scene of a book from one of four known authors. More specifically, the following table gives the list of the authors and a breakdown of the sample points in the dataset.

	Number of Books	Number of Chapters/Scenes
Austin	5	317
London	6	296
Milton	2	55
Shakespeare	12	173
		841

There 71 variables from which the first 69 correspond to 69 *function* words, such that: a, the, and, or, as, also, etc. The variables in particular measure the number of times this word appears at the specific play which corresponds to the row number. For example the first variable corresponds to word 'a' and value 46 at position (1,1) means that we measured 46 times the word 'a' at the play 1. Variable #70 with label 'BookID' is an ID for the book and the last variable (#71) is a variable which called 'Author' and gives the author for each play.

We may use the data set to see if the authors do discriminate with respect to the variables selected and if we can construct a discriminant rule able to predict the author of a play having the measurement on the 70 function words.

Explore a range of classification methods available to propose the prediction model for the author. Write a report to present your analysis providing with the steps you follow and supporting each step with tables/plots that are necessary. Propose the best candidate for the prediction model based on your analysis at the conclusion part of your report.