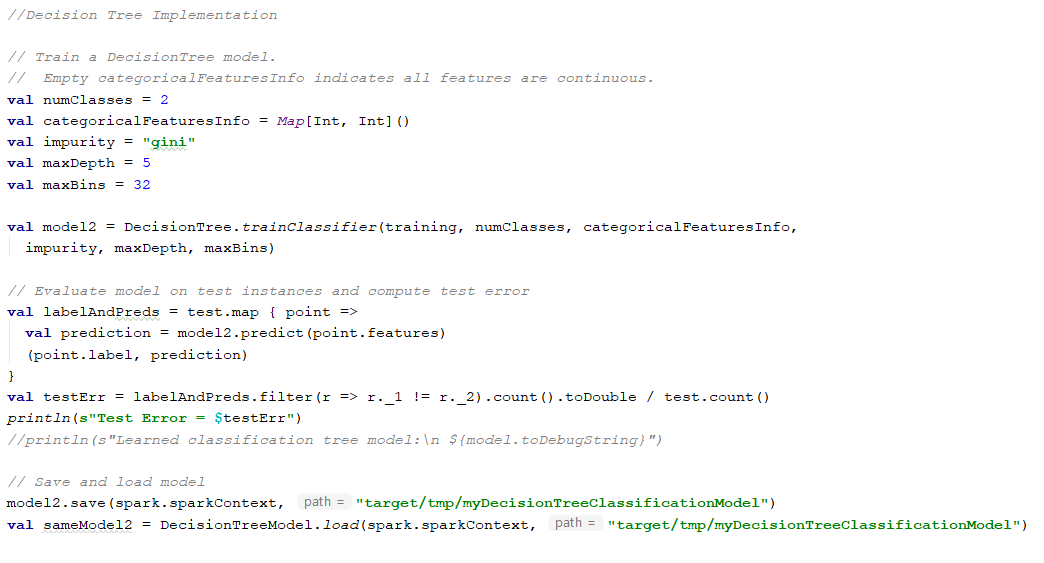
**Lab 4 Part 2 : Classification Models**

For this lab I got Spark to run three different classification models. Doing this I ran into many issues and in the end, am not sure that I have the desired output. But, nonetheless, I have some output that I am ready to show you today. I look forward to learning this in the coming class, as I feel that I was unable to get a solid grasp on it during the days that I have spent trying to get the code working.

With all this code it is pretty much just the code from spark.apache.org. It was a challenge just to get the example code running properly. I spent many an hour trying to get the models to work using one of the provided datasets, but the format of the datasets didn’t allow for easy molding into the algorithms. I also didn’t know which field we were trying to predict for either of the two datasets, so I was left a little confused there. I have left the code for my attempts commented below the sample code that I have managed to get running. I will try and explain the example code in the following lines.



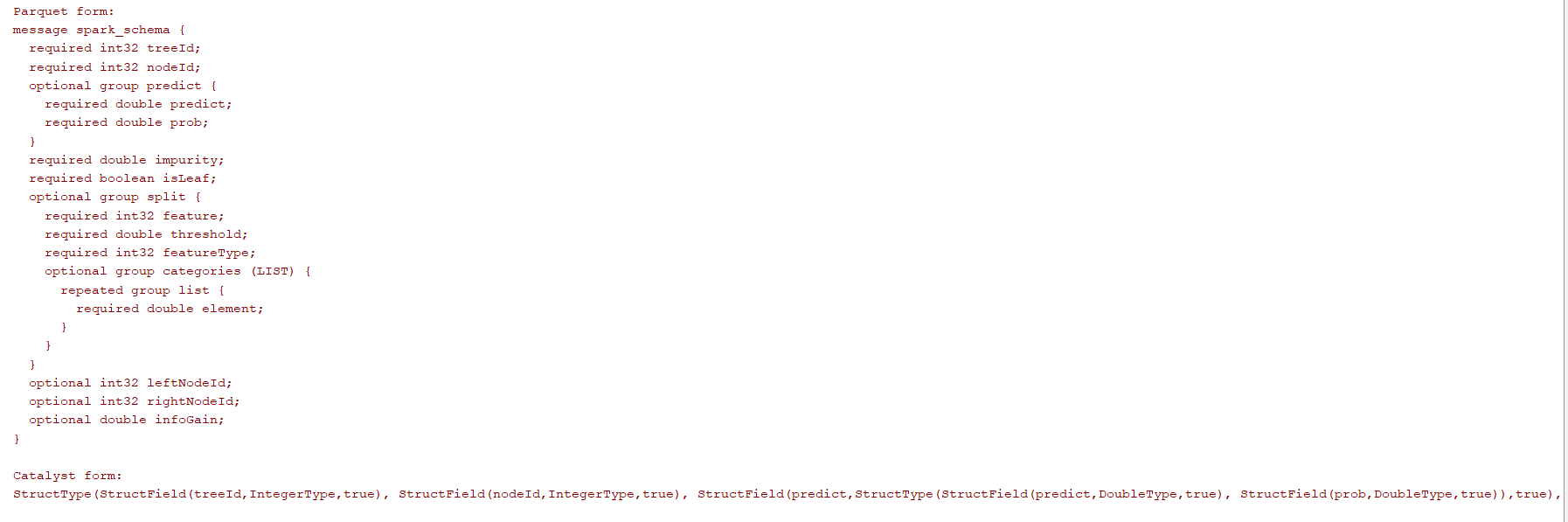
Naïve Bayes was the simplest of the three algorithms to implement. With all these the initial dataset will be split into two portions. A training portion and a test portion. This is so that the algorithm can try and classify the training data and then you can compare and test the test data to see how well it performed.



With decision tree and random forest as well there are a few other variables that are required for input into the algorithm. But ultimately, you’re still using training and test datasets in order to try and predict the prediction field.



Here is the cdoe for the Random Forest implementation. It is very similar in structure to the decision tree logic.



Regarding the output. The way that the code was structured it produces parquet files as output. I ran these parquet files with parquet-files tools in hdfs and read the files, and they didn’t seem to have any useful information that I could interpret. I’m sure there was something more there, as there were just a bunch of columns that contained numerical data that I’m sure could have been interpreted by a program, but alas, I don’t have the time in order to decipher the output in order to obtain the metrics in order to compare.

With classification models which one to choose depends largely on the dataset. Naïve bayes is a simpler model, so its performance is typically inferior to the other two models we have chosen. Although simplicity can sometimes be a positive feature in an algorithm when maybe a rough answer is needed quickly as apposed to a more robust solution that would take longer to produce.