Course Project #2 Assignment AY 22-2

Instructions: Complete each problem below, showing all your work. This is intended to be a group project; if your group receives assistance, please document accordingly. Turn in an electronic copy of your python solution in a Jupyter notebook with all document strings, your writeup (4-page max not including any appendices), and signed cover sheet in cocalc before COB of the date in cocalc (28 March). Homework for this block is due on 22 March (on the day of the class drop for the project).

Introduction: Complete the following questions and turn in via your 4-page paper (you can have as many appendices as you need).

Problem 1: Mathematically explain the difference between Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), and Naïve Bayes.

Problem 2: Wrangle the data in the 4 csv files (airports, flights, planes, and weather) so that you can build the best possible model for predicting if a particular flight will have a delay in its takeoff time.

Problem 3: Perform some exploratory data analysis (EDA) that will assist in understanding the relationship between a departure delay at the airport and some of the potential predictors. Show one graph or table that shows a relationship between the predictor variable(s) and the response variable.

Problem 4: Use the training data to create three models for predicting if a flight was going to have a long delay (more than 60 minutes), a short delay (1-60 minutes), or no delay. Explain how you will compare the models using the training data.

Problem 5: Select a model from problem 4 and use it for predicting a departure delay in a flight. Use your model to create predictions for the test data and turn these in as a .csv in the same order in which they are given. Your instructor should be able to compare your answers (0 – no delay, 1 – short delay, and 2 for long delay) easily and determine your misclassification rate.

Bonus (you can use an appendix to explain your model): Create a classification technique of your own (not using sklearn) using information from the LDA, QDA, logistic regression, SVM, or kNN to predict the departure delay. Hint: you will need to understand the mathematics behind the methods that you are using to create this new technique. Explain your model and turn in the predictions for the training set using this technique.

Turn in your writeup, documentation, code, and .csv with predictions in one of your cocalc folders.