|  |  |  |
| --- | --- | --- |
|  | **Spark Final Assignment** | **A20** |

1. Locate a data set and a problem that can be addressed by using this dataset and using any supervised learning technique like classification, logistic regression or even deep learning.
2. Use the data set to (a) carry out an EDA - Exploratory Data Analysis to understand the nature of the data and then (b) apply machine learning to solve the problem and come out with some good predictions. You should build at least two models, a preliminary model and an improved model that shows better performance.
3. It is important that you work with SPARK and not with a general purpose Python library that can also be used to solve the same problem.
4. You can use the Colab Platform, but using the DataBricks platform will be better.
5. Your report should contain these three parts ( weightage 20:40:40)
   1. A 300 - 400 word executive summary of the problem that you are addressing and the key features of your solution or model. You should not, repeat should not, use any technical terms in this report. Someone with no knowledge of ML, Spark should be able to make sense of your report. This should be in Google Docs
   2. A 15-20 slide Google Slide presentation that will explain the problem, the result of the EDA and technical description of your solution in terms of models, features, accuracy etc. Plus there has to be two slides of final conclusion. What did you achieve through this model? This should match with the executive summary in part a
   3. A notebook ( Colab or Databrick) with public access that has the codes that you used to download the data and perform the analysis. Anyone should be able to reuse the notebook and replicate your conclusions.
6. Submit these three artefacts in Google Classroom.