Proposal for DATA1030: Hands on Data Science

For my project I hope to explore the Data Set: Student Performance – Student Mat from the UCI Machine Learning Repository. This will be a regression model of supervised Machine Learning. Here is the link to the dataset: http://archive.ics.uci.edu/ml/datasets/Student+Performance

Requirements:

- The question I am hoping to explore is:
  'Can we predict the academic performance in mathematics class currently based on the different features we know about the students in the Mathematics class?'
- The target variable is the overall grade that the students achieved in the class.
- This is a regression problem.
- My motivation to explore this data set is my personal interest in education–secondary and high school education. I want to see the impact of various factors (be it socioeconomic or otherwise) on academic performance – and primarily want to focus on how students are (if at all) motivated to perform better in their classes after having failed previous ones. I personally think it is important to explore the various factors that may or may not impact educational achievement (or lack thereof) in order to better equip educators with the tools to ensure that all students in the classrooms have their diverse academic needs met properly.

Describe the dataset:

- There are 33 features (33 columns, 395 rows) – a total of 13,035 data points.
- The data set provides information about student performance in two classes – Math and Portuguese. I want to focus on the performance in Math data set. Within the math data, we are able to observe information about students from two different schools. There are 30 attributes (not including the three grades – G1, G2, G3) in the data and they are well defined. A project that used this data set in the past analyzed and compared the performance in Math and Portuguese between the two schools as well as the distribution of final grades based on gender. The paper for that project is cited below.

- **Attributes of the data:**
  1. School Name
  2. Sex
  3. Age

4. Address
5. Family size
6. Parents cohabitation status
7. Mothers education
8. Father education
9. Mothers job
10. Fathers job
11. Reason to choose the school
12. Students guardian
13. Home to school travel time
14. Weekly study time
15. Number of past class failures
16. Extra educational support
17. Family educational support
18. Extra paid classes within the course
19. Extra curricular
20. Nursery school attendance
21. Higher education plans
22. Internet access
23. Romantic relationships
24. Family relationship quality
25. Free time after school
26. Going out with friends
27. Daily alcohol consumption
28. Weekend alcohol consumption
29. Current health
30. Absences
31. Grades in first half
32. Grades in second half
33. Final grades

Preprocess the dataset:

- A lot of the data in this data set had been preprocessed already. I noticed that the binary categorical features had not been preprocessed. I applied One Hot Encoding to these variables (because they had binary values). The target variable – i.e. the Final Grades was already label encoded. I applied Min Max scaling to the continuous features i.e. age and number of absences.
- The number of features in the preprocessed data: (395, 59) – 59 total features and a total of 23,305 data points.

Link to Git Hub repository: https://github.com/msyed96/DATA1030-Project
Link to past project: http://www3.dsi.uminho.pt/pcortez/student.pdf