Shrey Agrawal DS HW#3 Ex #2

Problem: Using predictive analytics methods, we would like to identify first-year college students at risk of dropping out and academic failure.

a) Design the analysis, think through the stages of the analysis following CRISP-DM methodology! How can you state the problem as a data science problem? What machine learning approaches could be used here? (See Lecture 02!)

We can look at this problem as a classification problem - whether the student is at risk or not. A possible machine learning approach here can be building a supervised learning model. The details are outlined below:

CRISM-DM stages:
Business understanding: The aim of the project is to find which students are at risk of dropping out. This is a data science problem because we'll be predicting a target variable using explanatory variable and are goal is to minimize error and have high accuracy in our prediction of whether a student is at risk of dropping out or not
Data understanding: Well, I'm not sure what kind of data we have, but assuming the data is about a student and their academic performance, financial status, and other such explantory variables that can contribute to the likelyhood of someone dropping out. It will be valuable for us to have multiple attributes that can be used to determine dropout risk, and also it will help to have a big data set with lots of records so that we can have enough data even after the test/train split. It is also vital that our data is collected in a way that is representative of the population and has no bias so that it can be generalized to other students if wanted.
Data preparation: Since we will perform classification regression, it will be useful to clean the data of any inconsistencies, na values, outliers. It will also be vital to quantify any descriptive data and convert them to numeric data based on its weight. Eg. excellent = 5, poor = 1.
Modeling: we can build a model using supervised learning from training data that will predict dropout or not for each of the students. A knn classifier model would work by grouping students with lower scores (or other attributes often tied to students with a higher dropout rate), and then if there is a new student that has those attributes and falls in the same neighbourhood, they would be classified as being at risk of dropout.
Evaluation: we should check the accuracy, sensitivity, etc of the data - possibly use a confusion matrix and see how the model performed on the test data. If satisfactory, move on, else retrain model with new parameters.
Deployment: If the model is satisfactory, next step will be to finilize implementation, write a report/dashboard on the results so that they can be communicate to the stakeholders

b) Do you think that the requirements of a successful data science projects are met? Go through the 7 requirements that we have covered in class! (See Lecture 02!)
• Having domain knowledge or consulting with domain experts:
    - Yes, if we can have a conversation about people more experienced in dealing with education and students in general, then it would help us understand the problem

• Big data (many observations):

- Assuming we have enough data points, they would be useful to build an accurate model, and that way we are less likely to get connections that are there just by chance.

• Many features
  - Having a lot of features that help us train the classifier and minimizing MSE would be great. Our dataset has potential of having many features since there are lot of explanatory variables for something like dropout risk
• Clean data
  - Taking steps to clean the data to make it appropriate for building our model is vital and will help in creating an accurate model
• Unbiased data
  - If our data was taken from students of various backgrounds and is of a diverse population, then it will be unbiased
• The capacity of act
  - Here, it will be quite possible to take appropriate actions with students that are at a risk of droupout
• Measurability of Return of Investment (ROI)
  - Depending on how accurately our model is able to identify students that are at a risk of dropout

c) What ethical questions are raised in this project? Mention utilizations that you think are useful and ethical and give examples of bad applications as well!

<u>Useful:</u>

- It may be useful for faculty to identify students that may need more support
- If a student knows they are at a risk of dropout, they can take steps early on to ensure it doesn't happen. This might mean putting in extra hours of work, or managing financial resources, or addressing other reasons why they may be at risk of dropping out.

<u>Bad applications:</u>

- It could bias a teacher's opinion and treatment of a particular student if they know they are at a risk of dropout
- This information should be kept confdential and no unauthorized person should have access to this information, otherwise it can be used against the student
- If the model is not so accurate, it might falsely predict someone is at risk of dropout and add unneccasry stress for the student and involved parties. Alternatively, if it falsely predicts someone is not at a risk when they are, the student may not take appropriate steps to address the risk in time.