**BU - Business understanding**: What is the aim of the project? What is its business relevance? (What is the research question?) How can it be translated to a data science question?

The objective of the project is to detect the students who have high potential to suffer academically and to drop out, which will allow us to take precautionary actions to improve the chance that these students will graduate successfully. If we can implement this project successfully, it will certainly benefit these students. Furthermore, if every student does well academically and is not at risk of dropping out, it will generate a better studying environment within the college. In the long run, the project can potentially increase the reputation of the college.

**DU – Data understanding**: What data do we have? Can we collect more data? What is the quality of the data? What do the features mean?

We have personal data from the first-year students as well as personal data from the previous groups of students who have entered this college. The features can include but not limited to race, age, sex, address, high school results (GPA will be the main indicator here), standardized testing results, whether they are first-gen in college or not, financial status (whether they are on financial aid or not, the amount of aid they are receiving). These are all data that the college should have on hand as these data are common data collected in college admission. Assuming that no students lied on their college application, we could consider the data we have high quality data. The only issue would probably be missing features from the data collected a long time ago (the college might not have asked about students’ age 5 years ago for example). If we need more data, we could potentially identify other colleges similar to our college to get access to their data. To decide which colleges are similar to our college, we could take into consideration location, demographic, size, college ranking. However, I do not think it would be likely that other colleges would be allowed to share this kind of data with another college for privacy reasons.

**DP – Data preparation**: The process of transforming and mapping data from one "raw" data form into another format with the intent of making it more appropriate and valuable for analytics. Data cleaning, finding relevant attributes.

We need to perform some data cleaning on the data we have including: fixing the problem of missing data, transforming features, and trying to drop irrelevant features. My main concern with the data we have are missing data fields from previous years. We can potentially drop these years if they are too far back and we have enough data. We can also keep them depends on whether the missing fields are relevant attributes or not. We can decide whether the features are relevant or not based on some visualization or checking the correlations between each feature with another and with our target variable. Then, we can drop the features that do not have high correlation with our target variable (irrelevant) or the features that have high correlation with each other (redundant, i.e. high school GPA and standardized testing result). Finally, we want to perform some transformations on our features. Address could be transformed into categorical based on some kind of classification (i.e., urban, suburban, countryside) which could be transformed into binary using one-hot encoding. All other categorial variables can also be transformed into binary variables using the same method. The amount of aid the student is receiving should be made into percent in regard to the specific year’s tuition. Furthermore, high school GPA can be on different scale (international students might come in with a GPA that is calculated on a different scale than domestic students’), so it needs to be standardized. Similarly, standardized testing scores need to be normalized due to many different forms of available standardized tests (SAT and ACT are common ones in the U.S. and SAT is scored out of 2400 while ACT is scored out of 36).

**M -Modeling**: Finding the best performing model, fitting the parameters.

The problem is a binary classification problem. It will be a supervised learning task due to the fact that the label of the data is known as we knew whether the students of previous years dropped out or not. The problem is a classification task because the goal is to predict a discrete value for each incoming student based on their profile. Specifically, we are predicting whether that student will drop out or not so there are two possible values for our label making it a binary classification task. All of the approaches we have discussed so far in class could be applied here (kNN-neighbor, decision tree, linear regression). We can do a k-fold cross validation to choose the best performing model.

**E – Evaluation**: Evaluating the model. How does it perform? Is it good enough to achieve our goal?

We will choose the model that performs the best in our indicators. Assuming that our label is 1 or positive for if a student will drop out and 0 otherwise. Then, the main indicator we want to look at is Recall, since we want to avoid type I error with this kind of project. In other words, we do not mind classifying more students as potential dropouts as we are more worried about not catching the potential dropouts. We can also use F1-score as the indictor if we want a balance between Recall and Precision here.

**D – Deployment**: Implementing the model, embedding it to the system. Communicating the results. Writing the report/research paper.

We could design a support system for the students being classified as potential dropouts. We could ask the teachers to pay more attention to these students in class. We should also tailor the support systems for each student. We need to try to figure out why each specific case was classified as a potential dropout and solve the problem that way. For example, if a student was classified as a potential dropout due to financial reason, we could provide more aids for that student. All of this should be done in a way that the students do not notice the supportive system as it could leave a negative impact on the students.

I do think we have met all the requirements for a successful data science project here. We as students have some domain knowledge in the matter as we ourselves have a good sense of what will allow a student to success in college, and what the stories of dropout students usually look like. That being said, we also need to contact education experts as they have done concrete research on topic like this for many years. I think we do have good enough data to do the project. We should have enough observations in our data as thousands of students go in a college each year. Also, as I have mentioned, if we actually lack data to perform the project, we could reach out to other similar colleges to ask for their collaboration in a joint-project. I think I have listed a decent number of relevant features that could be extracted from the students’ college applications for the project. The data we have should be clean after our data processing steps given that the students did not lie on their applications. Even if there are students who did lie on their applications, I presume that the percent would not be high enough to significantly affect the data analysis. The data would not be unbiased as we would be using all the data from freshman-year students in the college. In other words, we are using the population itself. We do have the capacity to act as our action plan are all things that could be carried out by the college. We have a clear measurement for Return of Investment as we would just need to wait for the group of freshman-year college students to go through college to see whether we have succeeded in increasing the graduation rates as well as the academic results of the students who we identified as potential dropouts.

Ethical questions I think being raised in this project is whether we could use students’ personal data this way and whether it is fair for other students that we provide extra help for only a group of students. For the first issue, I think it is ethical to use the students’ data this way since it is an institution’s duty and responsible to help their students to success. We should still include some kind of clause in the college application form to ask for the students’ permission to use their data though. On the second issue, I think we could avoid if by making sure all students could access all the resources and extra help that we are giving to the students we determined as potential dropouts. Hence, the help would be there for all the students who think they need it. A good utilization would be the personalized help system that I have mentioned. Another good way to use the result could be to trying to fix the students’ problem on a bigger scale. For example, if we are seeing all the students from one specific neighborhood showing high potential of dropping out, we could work with the local government to examine the neighborhood itself outside of helping the students from that neighborhood. A bad utilization would be to tell the students the results of the project as it will not help them in any shape or form. The students who are predicted to drop out might just drop out. The students who are predicted to do well academically might get too confident and might ignore their studies. Another bad way to use the result would be to purposefully ignore the students who are potential drop outs. For example, a school can potentially ignore a student asking for extra help due to the fact that they think the student will likely drop out anyway and is not worth the investment. I think this is just unethical if a school do this, but it could definitely happen. For example, colored students in the U.S. are sometimes ignored in schools due to the fact that teachers seeing them as not as academically capable as their peers.