In this project for my three methods I chose Decision Tree modeling, KNN clustering, and Gaussian Naive Bayes. I wanted to do three different kinds of modeling and ones that I thought would be interesting.

To start with the Data, I used only a few features which I thought were most important for my models. The features I used were:

Medu and Fedu - I used these features because there has been significant research that the education of your parents impacts the quality of your education.

Traveltime - I think travel time could have a significant effect on performance because if travel time is higher, that makes studying harder.

Studytime - I think studytime has a significant impact on performance. By studying more you are better prepared in your class and will perform better that way.

Failures - failures have a significant impact on showing you the kind of student you are. If the student has lots of failures their performance won't be good. This is because getting a failure means that your performance in a class wasn't sufficient.

SchoolSup - Extra educational support is very important in the performance of a student. If a student has lots of support they will be able to do better in school.

FamSup - Family educational support is also something that can have a significant impact on performance. If in the household the student is able to receive lots of educational support that means that the student will be better prepared in school and will therefore have a better performance.

Higher - I think if a student wants higher education or not it has a significant impact on how they want to perform which then impacts their performance. If a student wants to go to higher education they are more likely to work harder and then they are more likely to perform better.

Internet - If a student has access to the internet they are able to access many more resources than someone without internet. This would allow them to perform better in school.

Freetime, Goout, Dalc, Walc - All of these are related to how much a student is studying. If a student is going out more, has more free time, and is consuming more alcohol on the weekends and weekdays than that means the student is focusing on school as much and then performance will suffer.

Absences - A higher number of absences will indicate that the student is missing more school which will mean the student hasn't been in class much and will have had less classroom time which would reduce the performance of a student.

I randomized the data and then used the first 40 for my testing set and then the rest for training data. I then made all features into binary. So if they had yes or no, I turned that to 0 or one. Then for performance I made it into -1 - low, 0 - normal, 1 - high.

I chose the algorithms I did because I thought they all were interesting. Decision tree modeling is one which I have some experience and wanted to try to do again. I thought it would be a good place to start out on and is something that I am mildly comfortable in. I also think that Decision Tree modeling is very interesting. I think the fact that each node is being created based on a mathematical set of ideas is interesting. I then chose KNN clustering because it is a very interesting algorithm. I find it interesting that the model is changing for every input that is received. For the number of neighbors I chose 3 because I thought that would make most sense as we had three different outputs we could receive. This was also the default value and many

places used this as a default value. Lastly I used Gaussian Naive Baynes modeling. This kind of modeling is very simple to understand. I also think even though it is a relatively easy concept it is very useful. One more thing with Naive Baynes is that many of the variables I used should be independent so by assuming that(which Naive Bayes does) it may help the model be better.

For my evaluation metrics I used both accuracy and R^2 score. I feel like both of these are pretty simple to understand but are very useful. Accuracy is how many times the model gets the correct answer and R^2 is essentially a metric to show you how far the predicted values are from the actual values. Both of these do a good job to depict how good the model is and they are simple. This is why I used them. My metrics are below.

	Accuracy	R^2
Decision Tree (DT)	0.3	-0.759
K-Nearest Neighbors(3) (KNN)	0.325	-0.528
Gaussian Naive Bayes (GNB)	0.45	-0.435

As we can clearly see GNB was the best model out of the three. It had the closest R^2 to 0 and also had the highest accuracy with 0.45. The Decision Tree model had the worst performance out of all of the models having the lowest accuracy with only 0.3 and the worst R^2 score. KNN was in the middle of the pack having an accuracy of 0.325 and a R^ of -0.528. To conclude, Gaussian Naive Bayes was the best model for predicting the performance and Decision Tree was the worst with KNN being in the middle of both of them.

There were many limitations to the models. Starting with Decision Tree it has a tendency to over fit the model and that definitely could have happened.It also is prone to small changes in the data making a big difference on the outcome. It can be Biased towards features that come first in the data or have lots of features. KNN also has some limitations where it is sensitive to feature scaling. If the features have different units or scales some of the features may dominate others in the model. KNN also requires a careful selection of K and this makes a large impact on the performance of the model. I could have probably chosen a better K that would have made the model better. In GNB a limitation is that it assumes all variables are independent. Some of the variables in our data were independent of each other and therefore this assumption would be incorrect. GNB can also only model linear relationships between input and output. It isn't really able to capture more complicated relationships like interactions between features or nonlinear relationships. GNB is also very sensitive to outliers. In terms of my data I used all supervised learning models. The features I selected had a significant impact on the models. I probably selected some features which didn't really impact performance or weren't good indicators of performance and this would make all of the models not as great. Another limitation was that I could have pre-processed my data differently.

For future work I would like to try an unsupervised learning model. I would like to try working with neural networks in order to create a model. I want to see how it would do with all the data and what kinds of results it would produce. I also would like to do some more research on what features would be the best to use and that way I could make all my models better. Another thing I would like to try is just using new models to learn how they work and their nuances.