Project Paragraph

For my project, I selected a data set from the UCI Machine Learning archive that contains many data features about various students in secondary school (school, age, mother’s and father’s income and education, free time, travel time, daily and weekly alcohol consumption), and uses all these factors to determine the student’s final grades (link: <https://archive.ics.uci.edu/ml/datasets/STUDENT+ALCOHOL+CONSUMPTION#>). This dataset really caught my attention as it is something I could really relate to; it would be extremely interesting to see if any of these features are significant enough to actually determine the student’s final grades. Looking at the data closely (descriptive statistics), I found that a lot of the variables were categorical (with a good amount of those being binary) with the rest being discrete numerical variables. Due to this, I transformed the rest of the variables into factors and decided that the best approach to analyzing this data would be a decision tree (perhaps with a random forest to increase accuracy). With this method, I feel like it would be able to accurately separate the data instances while pointing out the less helpful variables in predicting the student’s final grades.