Assignment 2

Consider the Student Performance Data Set from this website,
https://archive.ics.uci.edu/ml/datasets/Student+Performance#

This data approach student achievement in secondary education of two Portuguese schools. The data attributes include student grades, demographic, social, and school-related features and it was collected by using school reports and questionnaires. Two datasets are provided regarding the performance in two distinct subjects: Mathematics (mat) and Portuguese language (por).

Let say you are only interested to predict the grade of the students (G1, G2, and G3) for Mathematics subject.

By using a decision tree:

1. Build a regression tree where the aim is to predict variable G1.
2. Let say you transform the variable G2 into new variable, G2T which have five categories, such as
    - 0 to 4 into E,
    - 5 to 8 into D
    - 9 to 12 into C
    - 13 to 16 into B
    - 17 to 20 into A

    Build a classification tree to classify the G2T variable.
3. By using Random Forest, rank the variables according to its importance.

If there are any missing values, use appropriate technique to handle it, such as mean or mode for that particular variable.

Make sure to split the data set into training and testing sets where the train size is 80% and the random state is 1.

Write a report in Word file and specify which variables are important for predicting the grades.