You are given a dataset. The dataset is split into two files. You will need to:

1. Load the dataset and deal with missing values in an appropriate manner. Describe how you handled them and why you did it that way.
2. Determine the 10 most important features of the cleaned dataset, explain how you found them and why you think they are important. This is the dataset you will be using for the rest of the assignment.
3. Get results for RF, pruned DT (make sure to use a validation set for pruning!), unpruned DT, and decision stumps and determine if any are statistically significantly better than others. Why are the worst methods performing so badly compared to the others?
4. Add 20% normal additive noise to the features and train the classifiers from step 3 and determine if any are performing significantly worse/better than on the clean dataset. Give reasons based on your knowledge of the classifiers.
5. Repeat step 4 with 20% normal multiplicative noise. Additionally,explain why you think additive and multiplicative noise influence the classifiers differently.
6. Use 5% class noise (that is, you flip 5% of the labels to the other class) and investigate which classifiers' performance is affected significantly. Why do you think class noise affects the classifiers differently from feature noise?
7. If you split the data into a training and test set first, and only afterwards add 20% normal multiplicative noise to the training set, how differently does your algorithm behave? Try this again and add noise only to the test set. Are the results different? Discuss how each of these approaches affect your results.
8. Bonus task (fewer marks, more interesting): What happens if you raise or lower the different kinds of noise to a higher or a lower percentage? Try to identify how much noise each classifier can tolerate and when they start to break.