Task One

- The dataset has missing values which have no correlation to specific features but follow a pattern: Each instance has 1 missing value such that all features have missing values.
- All features in the dataset contain continuous values.

Considering the above, the missing values were filled with the mean of each feature. This is because it is unreasonable to remove any features or instances.

Task Two

The scikit learn chi squared test function was unable to be used due to the negative values in the features. Therefore, the features with the highest Pearson's correlation efficient values were used. However, features with similar correlation values occasionally had almost identical feature values (usually 98%). These redundant features were not included in the list of 10 features because it was and could influence the performance of the classifiers. Therefore, cleaned dataset was comprised of unsimilar features with the top 10 correlation efficient values.

This method is likely to be effective because the Pearson correlation efficient examines the co-variance between each individual feature and the target variable. Therefore, the features with higher correlation efficient values would have a stronger correlation with the target variable, thus meaning that the variable is more important to a classifier and is likely to improve performance.

Task Three

The performance by the Random Forest classifier is expected to be great because it is an ensemble classifier. This means that the output of this classifier is done through bagging of multiple decision trees. Similarly, the pruned decision tree classifier is expected great performance because it has been pruned in a similar way to bagging. The decision tree classifier was pruned by comparing accuracy based on different ccp_alpha values on the validation set and taking the ccp_alpha value with highest accuracy. This means that it is likely to perform better than the unpruned tree and decision stump. The decision tree stump is more likely to underfit compared to the other decision tree models because the restriction max_depth =1, a restraint not placed on the other two decision tree variants. Overall, the rankings of the decision tree variants are mostly dependant on the level of underfitting by the decision stump and the level overfitting of the pruned and unpruned decision trees. As such, the decision tree stump performed the worse.

The table below show the Autorank results:

Statistical test	P value	Significant difference
ANOVA	1.578366 e-53	Yes

Task Four

Overall, there was no observable significant different between the cleaned dataset and the dataset with additive noise. This is most likely because the mean of the features is very close to zero, meaning that the 20% noise percentage is not likely to be sufficient to cause significant difference in means.

The table below shows the Autorank results for the classifiers between the cleaned dataset and dataset with additive noise.

Classifier	Statistical test	P_value	Significant difference
Random Forest	t-test	0.786876525	No
Decision Stump	t-test	0.428995675	No
Unpruned decision tree	t-test	0.100402741	No
Pruned decision tree	t-test	0.126069892	No

Task Five

For this specific dataset, the multiplicative noise is more influential because additive noise is dependent on the mean of the features and many of the mean of the features of the dataset are close to zero. Thus, we see significant differences between the mean of accuracy in the cleaned dataset and multiplicated noised dataset in the unpruned and pruned decision tree classifiers. This is because both decision tree variants are susceptible to overfitting, meaning that the variance and error caused by the noise had decreased performance.

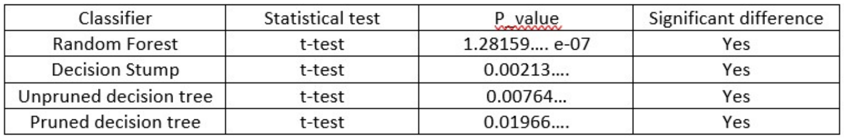
The table below shows the Autorank results for the classifiers between the cleaned dataset and dataset with multiplicative noise.

Classifier	Statistical test	P_value	Significant difference
Random Forest	t-test	0.3028126354910162	No
Decision Stump	t-test	0.7566198484676405	No
Unpruned decision tree	t-test	0.0406747808323915	Yes
Pruned decision tree	t-test	0.01241116212063172	Yes

Task Six

Class noise influences classifier performance more than feature noise. This has been evident in the results shown. The class noise is likely to affect classifier performance at a greater level because the class is the sole dependent variable, whereas for feature noise, there are many features existent in the dataset. This means that depending on the level of error and variance, the classifier can ignore or give weight to specific features in proportion to the level of noise in the dataset. This cannot be executed as effectively with the target variable, ultimately meaning the conditions for the classifier are stricter when there is noise in the target variable.

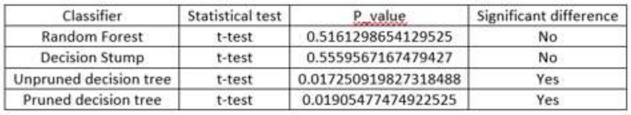
Autorank results between cleaned dataset and target variable noise dataset.



Classifier	Statistical test	P_value	Significant difference
Random Forest	t-test	1.28159…. e-07	Yes
Decision Stump	t-test	0.00213….	Yes
Unpruned decision tree	t-test	0.00764…	Yes
Pruned decision tree	t-test	0.01966….	Yes

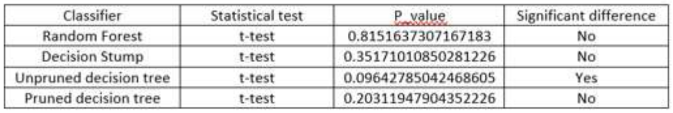
Task Seven

Autorank results between the cleaned dataset and training noise dataset



Classifier	Statistical test	P_value	Significant difference
Random Forest	t-test	0.5161298654129525	No
Decision Stump	t-test	0.5559567167479427	No
Unpruned decision tree	t-test	0.017250919827318488	Yes
Pruned decision tree	t-test	0.01905477474922525	Yes

Autorank results between the cleaned dataset and test noise dataset



Classifier	Statistical test	P_value	Significant difference
Random Forest	t-test	0.8151637307167183	No
Decision Stump	t-test	0.35171010850281226	No
Unpruned decision tree	t-test	0.09642785042468605	Yes
Pruned decision tree	t-test	0.20311947904352226	No

Overall, the unpruned and pruned decision trees had significant differences between the cleaned dataset and the training noise dataset. This is likely because both decision tree variants are unstable classifiers and thus very susceptible to overfitting. Thus, the classifiers must have overfit to the noise, leading to decrease in performance.

Only the unpruned decision tree had significant difference between the cleaned dataset and the test noise dataset. This is likely because the noise in the test data had deviated on the patterns of variable importance and the classifier was too underfit for the test set with noise.

Task Eight

Because noise introduces error and variances to the dataset, a higher percentage of noise will likely decrease the performance of a classifier. However, a certain level of noise is beneficial to prevent overfitting into the training set. For this specific dataset though, the mean of the features is mostly close to zero Therefore, it the level of additive noise has to be very large to have effect on the classifiers. On the other hand, the multiplicative noise level does not need to be as high to influence the performance of the classifiers. The stability of the classifiers is a key factor.

For example, to get the accuracy levels below 55% for this dataset, a noise percentage of around 200% is required for multiplicative noise, whereas a noise percentage of around 2000% is required for additive noise.