Task One

There are many ways to handle missing data such as ignoring the tuple, using a global constant or imputing an average. We did not use the method that involved ignoring the tuple as if there are many missing values, then we are compromising too much of the data and the process which generates the data could be related to the process that causes data to be missing, meaning we may bias the model if we ignore all tuples. Using a global constant could be an issue if the process that causes data to be missing is too complex to generalise all the missing data into a single value. Imputing the mean allows us to retain data and replace the missing values from the given dataset without over-generalisation or much domain knowledge.

Task Two

Feature selection is done to reduce the amount of attributes and regulate dimensionality by eliminating features that are irrelevant, redundant or less useful than other features. Univariate selection works by selecting the best features based on univariate statistical tests and removing all but the highest-scoring features. In particular, we used the f\_classif argument for classification which is based on the F-test and estimates the degree of linear dependency between two random variables.

Task Three

Table

Description automatically generated

Decision stumps will likely always underfit, whereas decision trees are most likely to overfit. From the autorank results, we can derive that the decision tree performs the worst, whereas the simplified versions (Decision stumps and Pruned Decision Tree) of the Decision Tree model, perform better. This implies that the Decision tree is overfitting to the training data quite significantly and is affecting their performance.

Task Four

Table

Description automatically generated

Noise is defined as random error or variance in measured variables that may be attributed to technology limitation, naming convention inconsistency and etc in data. The model which performed better on the noisy dataset, as opposed to the clean dataset is the Decision Stump model. Where Decision Trees are likely to overfit and ‘learn’ the noise of the training data, Decision Stumps are considerably less complex and refrains from memorising the noisy data, thus enabling them to preform better on unseen data in the test set.

Task Five

Table

Description automatically generated

Additive noise benefits the performance of models which underfit such as that of decision, whereas it appears that multiplicative noise improves the performance of models who overfit like that of decision trees. This could be due to the fact that additive noise adds noisy data on top of the existing data, and the probability of overfitting on noise increases as a tree gets deeper, whereas multiplicative noise permeates throughout the data and is consequently learned by the decision tree in training and accounted for in testing as both are noisy data. You will notice the similarity in overall performance of each classifier between the holistically cleaned dataset and this holistically noisy dataset.

Task Six

Table

Description automatically generated

Pruned Decision Trees perform significantly worse with the class noise, due to the fact that the pruning process involves limiting tree depth by eliminating the weakest sub-trees which do not contribute to classification power in the cost-complexity method. We know that noisy labels degrade the generalisation performance of models, thus may be why Decision stumps slightly outperforms Decision trees as it typically has an over-generalised fit as is.

Task Seven

Table

Description automatically generated

The Decision Tree outperforms the Pruned Decision tree as the Decision Tree learns on the clean training data well, such that when it is tested against unforeseen noisy data, it would overfit to the standards and classifications of the clean set as opposed to Decision Stumps and Pruned Decision Trees which are less sensitive to the particular training set and more suited to fitting the noisy test set. It all harkens back to the fundamental trade-off and reaching a balance between bias and variance to minimise total error.