◆ Regression Task

I. *Comparing accuracy*

Algorithm 1: Decision Tree Regressor
Algorithm 2: Random Forest Regressor

For this task, R^2 score is used to measure how well the model fit the data. The closer it is to 1, the better is the result.
For the first algorithm we get an overall score of 0,99 for the training set and 0,67 for the test set.
For the second one we get a score of 0,97 for the training set and 0,84 for the test set.
For the neural network we get a score of 0,86 for the training set and 0,83 for the test set.

The result above show that there is an over-fitting of the data when it comes to the decision tree.
And for the Random Forest Regressor to be better in the test set is pretty much justified.
The Neural Networks is doing not bad, but can further be improved with some tuning.

II. *How the models work*

Algorithm 1: Decision Tree Regressor

Decision Tree belong the family of supervised learning algorithm.
The algorithm produces a model that can predict value by learning simple decision rules based on the features inferred from the training data.
how does it build such rule-based mapping?
We start with a root node, which represents the entire sample. This node will be split in a two or more sub-

nodes. When a node is split into further sub-nodes it become a decision node. Eventually we will get leaf node which are terminal node.
The process of splitting can be done using many algorithms like ID3(extension of D3).

ID3 algorithm:

1) Starting the training set S as the root node
2) On each iteration of the algorithm, it iterates trough the unused attribute of the set S and calculate Entropy or Information gain of this attribute
3) Then it selects the one with the smallest entropy or the largest information gain
4) The set S is then split by the selected attribute to produce a subset of the data
5) Recursively, the same thing is done in each subset

Criteria like Entropy, Information gain, Gini index, Reduction in Variance, Chi-Square, Gain Ration are used for attribute selection.
Entropy is a measure of the randomness in the information being processed. The smaller it is, the easier it will be to draw conclusion from the information.

$$E(S)=\sum_{i=1}^{n} - p_i \log_2 p_i$$

p_i : probability of an event i of state S
Entropy for multiple attributes:

$$E(T,X)=\sum_{c \in X} P(c) E(c)$$

T: current state
X: Selected attribute

Reference: How do Decision Trees work?

Algorithm 2: Random Forest Regressor

Random Forest is also a supervised machine learning algorithm.
The algorithm builds decision trees on different sample and take the majority vote for classification and average in case of regression.
Random Forest is a part of the family of ensemble learning algorithm, which means combining multiple models.

It uses bagging method, which consist of training different subset from the training data and use the average in the case of prediction of continuous value as the output.

Reference: Understanding Random Forest

III. *Tuning models*

Models can be tuned by evaluation features importance, and removing irrelevant one.

In the case of our first model, we have an over-fitting. The result can be improved by updating the hyper parameters like the max_depth, min_samples_leaf, min_samples_split.

K-fold cross-validation can be used in the first two algorithms with GridSearchCV to optimize hyper parameters.

The multilayer perceptron can be improved by using K-fold cross-validation with RandomizedSeachCV to optimize the hyper parameters.

◆ Classification Task

I. *Comparing accuracy*

Algorithm 1: Decision Tree Classifier
Algorithm 2: Random Forest Classifier

We have a 100% accuracy in the 3 for the training set.
When it come to the test set the Random Forest Classifier do a better Job with a 96% against 78%.
The Neural Network have an overall accuracy of 95%.
We can see that we have almost the same results as in the regression but the classification is a little better.

II. *How the models work*

Algorithm 1: Decision Tree Classifier
Same logic in the regression task, but predicting discrete value by majority voting

Algorithm 2: Random Forest Classifier
Random Forest of Decision Tree Classifier
For this task, the accuracy score is used:
$(TP+TN)/(TP+TN+FP+FN)$
TP: True Positives
FP: False Positives
FN: False Negatives
FN: False Negatives

III. *Tuning models*

Same thing in regression.