## Machine Learning – HLTWK6 (w/c 8th November 2021)

### Machine Learning Algorithm – Decision Tree.

A decision tree can be explained by having 2 entities. Decision nodes and leaves. The leaves will give an output of the decision, whilst the decision nodes are where Data is split. Since a decision tree is a form of supervised learning, it requires an input of training data to help it predict the outcome of the new data.

There are 2 types of trees; Classification trees and Regression trees.

Classification trees are commonly used when an outcome is a variable. For example, a decision tree to help predict the risk level of an individual could rely on nodes such as: What is the age of the person? Does the person have debt? Does the person have assets?, the outcome/ leaf that would be derived would be “Risky and Not Risky” based on the Yes/No type questions. On the other hand, a Regression tree, would provide an output of a continuous variable, such as a number.

Decision trees intuitively tell us about the predictability of an event occurring. Therefore, the algorithm uses a measure of certainty and randomness of the data (Entropy). A lower entropy denotes ow randomness in data, whilst a higher entropy denotes greater uncertainty of data.

An example of where a decision tree can be found useful is to predict Salaries. If data sets are used with low entropy, we will have more reliable data to predict if an employee’s salary will be greater than £100k. Data sets obtained could include existing salaries of staff with a specific degree, job title, or company name. In theory, this model has been trained to help predict a salary of anybody working in a particular company, with a specific job title and a type of qualification. The outcome would be Yes / No, either it would predict that you earn more than £100k or not.

This model can also be used greatly during medical trial runs. In particular using historical data sets from the trial runs would help predict the chances of survival, infection rates, side effects on a certain individual based on a number of nodes that considers age, pre-existing health conditions, and survival rate. As the trial runs continue, as the data increases in size and becomes more reliable (ie consistent), scientists would be able to predict the effects of changes to the medicine on a particular individual.