Within this machine learning project, we will be using the drug consumption data set. This dataset includes various data regarding 1885 individuals. The features included within this dataset describes basic information such as age, gender and location as well as various personality measurements. These personality measurements quantity characteristics such as agreeableness, conscientiousness and impulsiveness. These features are then accompanied by labels that specify the class of drug consumption for various different drugs such as alcohol, amphetamines and cannabis. The actual drug consumption classes specify a range of drug consumption habits in 7 unique classes going from CL0 (Never used the drug) to CL6 (Used the drug in the last day).

# **Model Selection**

In the case of model selection for the chosen dataset, we have identified 1 deep learning and 1 machine learning algorithm. The machine learning algorithm chosen will be the Random Forest classifier and the deep learning algorithm will be the Artificial Neural Network. Within this section, these models will be discussed to understand the important factors and exactly how to make use of both in the mission of creating a quality prediction model. A rationale for the choices of these models will also be included, highlighting their advantageous properties which can be applied to the problem at hand.

# **Random Forest**

The random forest classification algorithm is an ensemble machine learning algorithm, meaning that it combines more than one model in order to learn patterns and make predictions with data. In this case, we have a collection of decision tree classifiers, each making a prediction on the class that an input to the model will belong to. After each decision tree makes a vote for the class that the input belongs to, the class that was predicted the most will then be the final predicted class for said input (Kulkarni & Sinha, 2013).

Within the domain of the problem at hand, which is classification of drug consumption levels, this problem can be identified as a classification problem. The first reason for the choice of random forest then would be that it is a classification model and so will be able to be applied to the problem. As shown in Uddin *et al.* (2019), since the random forest model is also an ensemble model, it is able to out perform the models just making use of one classifier. The algorithm also improves on the decision tree classifier which it is built upon as there is a much lower chance of variance and overfitting occurring when making use of a random forest model. This would be because of the randomness that is introduced during the training stage through the use of bootstrap sampling and random feature selection to create the individual decision trees.

The random forest model is also shown to be easy to use and understand, allowing for easier interpretation of results and application to the problem at hand while not giving way in the prediction performance department (Meltzer, 2021). This will allow us to very quickly begin creating a quality solution with powerful python libraries such as sklearn which provides functionality for random forest classification.

# **Artificial Neural Networks**

Within the field of deep learning algorithms, we have made the choice to use the artificial neural network.

# **Reference List**

Kulkarni, V.Y. & Sinha, P.K. 2013. Random forest classifiers: a survey and future research directions. *Int. J. Adv. Comput*, 36(1):1144-1153.

Meltzer, R. 2021. *What Is Random Forest?* <https://careerfoundry.com/en/blog/data-analytics/what-is-random-forest/#what-are-the-advantages-of-random-forest> Date of access: 16 August.

Uddin, S., Khan, A., Hossain, M.E. & Moni, M.A. 2019. Comparing different supervised machine learning algorithms for disease prediction. *BMC medical informatics and decision making*, 19(1):1-16.