

**ITRI626 Large Project, Phase 1**

**31715982 – Daniel Coetzee**

**30949955 – Lesetja Mojapelo**

**Drug consumption classification**

**Table of Contents**

[**1.** **Preprocessing** 3](#_Toc143344729)

[**2.** **Feature Engineering** 4](#_Toc143344730)

[**3.** **Model Selection** 4](#_Toc143344731)

[**4.** **Random Forest** 5](#_Toc143344732)

[**5.** **Artificial Neural Networks** 5](#_Toc143344733)

[**6.** **Reference List** 7](#_Toc143344734)

# **Introduction**

As we embark on our analytical venture using the drug consumption dataset, which encompasses data from 1,885 individuals, it's pivotal to ensure the data's readiness. This dataset not only provides foundational demographics but extends its depth to intricate personality measurements, ultimately leading to classifications of drug consumption spread across seven distinct categories. Our initial step, Preprocessing, is dedicated to preserving the dataset's authenticity. By judiciously addressing potential pitfalls like missing values, duplicates, and outliers—where the z-score method is employed—we aim to uphold the data's integrity. While feature scaling is a conventional method in data processing, our informed choice to bypass it stems from the potential insight outliers might offer in our models.

# **Preprocessing**

Preprocessing data is the first step before adding data into a model, for this we will be using the drug consumption data set. This dataset includes various data regarding 1885 individuals. The features included within this dataset describe basic information such as age, gender, and location as well as various personality measurements. These personalities measure quantity characteristics such as agreeableness, conscientiousness, and impulsiveness. These features are then accompanied by labels that specify the class of drug consumption for various 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 on the last day).

1. Handling missing values.

The drug classification dataset has no missing values; therefore, this step is unnecessary.

1. Duplicate rows

The drug classification dataset does not have any duplicate rows, thus duplicate removal is not applicable.

1. Outlier detection

The number of outliers that have been detected using z scores are as follows:



Z = (x−μ)/σ​

Where:

Z is the Z-score.

x is the value of the data point.

μ is the mean (average) of the dataset.

σ is the standard deviation of the dataset.

Depending on the chosen model, these may pose an issue should the chosen model be sensitive to outliers in the data. How outliers will be handled will be determined by which models have been chosen.

1. Feature scaling encoding

For our purposes, we have decided not to apply any scaling as the identified outliers may be useful when applied to our models.

1. Encoding categorical variable

No encoding must be done as this data set has quantified all categorical features; thus, features are all numerical.

# **Feature Engineering**

For the drug classification dataset, we have chosen to use random forests as our primary machine learning model, the secondary model will be a deep learning model whose feature engineering will be determined at our discretion.

* Feature Scaling
  + Random forests are unique in that they aren't influenced by the magnitude of numeric variables. Hence, there's no strict need to undertake normalization or standardization, which simplifies our data preparation process. Additionally, the dataset itself shows indications that features have been standardized.
* Feature Creation
  + Random forests are inherently able to model the interactions between features it is modelling, which indicates any manual interaction may lead to redundancy.
* Handling of Outliers
  + Random forests are known to be robust oi nth way they handle outliers. Thus, with the above preprocessing, we do not have to handle any outliers because of this approach.
* Feature Selection
  + We have decided not to withhold any features in the dataset but instead plan to use each feature to perform more than classification.

# **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 to learn patterns and make predictions with data. In this case, we have a collection of decision tree classifiers, each predicting 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).

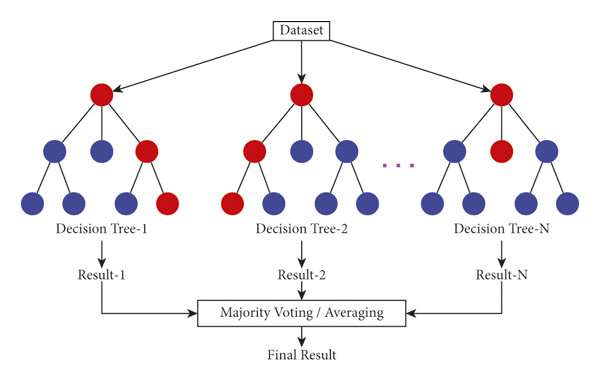


Figure 1: Visual representation of a random forest.

Within the domain of the problem at hand, which is a 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 by Uddin *et al.* (2019), since the random forest model is also an ensemble model, it can outperform the models just by making use of one classifier. The algorithm also improves on the decision tree classifier it is built upon as there is a much lower chance of variance and overfitting occurring when using a random forest model. This would be because of the randomness that is introduced during the training stage using bootstrap sampling and random feature selection to create the individual decision trees.

The random forest model is also easy to use and understand, allowing for easier interpretation of results and application to the problem at hand while not giving way to the prediction performance department (Meltzer, 2021). This will allow us to begin very quickly 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 chosen to use the artificial neural network. These neural networks are these models that take after the functioning of the human brain with thousands or even millions of singular, simple nodes that are densely interconnected to each other. Each node does some kind of operation on the data being fed through the neural network, with the node receiving data from many nodes behind it and outputting data to many nodes in front of it. Generally, these nodes are separated into layers or levels in the neural network. Figure 1 shows an example of an artificial neural network with an initial input layer, many hidden layers and finally an output layer.

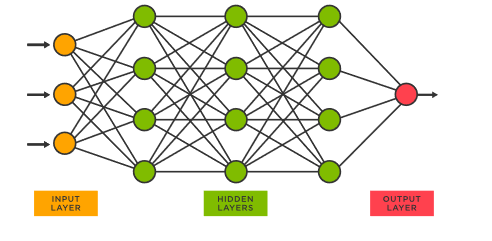


Figure 2: General structure of the neural network with nodes and dense connections between layers.

Each connection from a previous layer to a node will supply a numerical value to the node. With each of these numerical values, the node will assign a weight and then add all these new weighted values together. This arrives at a single number which depending on certain threshold conditions, this node will send this computed value out to all the connections it has with the next layer of nodes. This process is repeated many times within the training phase of the model with information regarding accuracy being given to the neural network through some chosen loss function.

An important ability of neural networks in the case of the drug classification dataset would be the ability to learn complex, non-linear relationships in data. This is important as we can assume that the relationship between human personality and drug-taking habits will most likely not be simple or linear. The hyperparameters associated with Artificial Neural Networks such as the number of hidden layers and number of neurons per layer also will allow an increase in the complexities that can be modeled. This will further our ability to make accurate drug consumption habit predictions even if we encounter very complicated relationships.

# **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.