**3.2.3     Learning Algorithm Selection (8%)**

Choose three machine learning algorithms, which are not necessarily taught in this submodule. They cannot be based on deep learning or reinforcement learning. Explain why the chosen algorithms are appropriate for the project.

The task of predicting participants’ gender input should be treated as a classification task [1], as the output are a range of discrete values represented by integers. While having the classes predefined and data labelled means clustering algorithms are not applicable. [2]

In this case, we are doing a multiclass classification since three types of gender are considered. They are 1: male 2: female and 3: other.

We will be using the following three algorithms:

1. Decision Trees

Decision trees are the fundamental components of random forests, it can perform both regression and classification tasks as well as multioutput tasks. [3]

This algorithm is chosen since it is a typical machine learning algorithm that is easy to use, having an overall prediction complexity of O(log\_2(n)), which made the process very fast. For this task, with less than a thousand instants for training sets, pre-sort in Scikit-Learn is also available to speed up the process. [4] One-versus-all strategy will be used here to predict the gender.

2. Random Forest

The Random Forest classifier is an ensemble of decision trees algorithm, aim to give a better prediction by taking the most voted class from individual trees.

This algorithm is chosen because it can directly classify instants into multiple classes, it reduces overfitting problem in decision trees and variance. [5] This way we can compare the performance of these two algorithms in a straightforward manner.

3. K-Nearest Neighbours (KNN)

­­­KNN can perform both regression and classification. The object is assigned to the class most common among its k nearest neighbours. [6]

This algorithm is appropriate because it does not require any additional assumptions on data. The database used for this task doesn’t have any missing data, and have less than a thousand instances, therefore some disadvantages such as the need to deal with missing data will not hinder its performance. [7]

Weights will be assigned uniformly to contribution of the neighbours.

[1]: <https://en.wikipedia.org/wiki/Statistical_classification>

[2]: <https://blog.bismart.com/en/classification-vs.-clustering-a-practical-explanation>

[3]: https://scikit-learn.org/stable/modules/tree.html

[4]: hands on machine learning with scikit-learn and tensorflow, page 222

[5]: <http://theprofessionalspoint.blogspot.com/2019/02/advantages-and-disadvantages-of-random.html>

[6] https://en.wikipedia.org/wiki/K-nearest\_neighbors\_algorithm

[7] https://medium.com/analytics-vidhya/summary-of-knn-algorithm-when-used-for-classification-4934a1040983