## What is clustering?

Clustering is an unsupervised machine learning technique, where the goal is to define clusters of datapoints that are relatively close to each other. This also means to divide the data into groups that are useful or meaningful, while still capturing the "natural" way of the data structure (Nathiya, Punitha, Punithavalli, 2010). This way of grouping data points together is important because of one crucial characteristic of unsupervised machine learning. We do not have the labels for the available data. This means that we do not know which class or classes that a certain data point belongs to, which is why it is necessary to cluster related data points together, to be able to create your own labels and groups within the data. Usually, this will not give as specific labeling as a supervised machine learning problem, where we clearly know the labels of our data, but it serves to both understand the data and to analyze it.

## Use cases for clustering in the industry

Clustering within unsupervised machine learning has a wide range of applications. It can be used for things such as customer segmentation, anomaly detection, search image and semi-supervised machine learning (Géron, 2019). All these applications, apply a wide variety of clustering techniques, where some look for densely packed data points that can be classified as a cluster, whereas others look to locate centroids of the data, that can act as centers for the clusters. As Nathiya et al. (2010), mentions in their article, clustering can also be useful in terms of noticing heart diseases in patients. They have used the Heart Spect dataset to apply clustering techniques, to be able to categorize new and incoming data. The clustering algorithm of their choice is the K-means algorithm. The principle of the K-means algorithm is to make an initial guess on the clusters in the dataset based on a given hyperparameter, which tells the model how many clusters it should guess on. The initial guesses are random, meaning that the algorithm will place the given number of centroids at any place in the dataset (Géron, 2019). Then the instances closest to the centroids are labeled as belonging to that cluster. Then the centroids are updated based on how well they describe the datapoints around them. This iterative process is continued until the clusters have made the best guesses.

This algorithm is used in the paper to try and predict whether a given patient has a heart disease. They reached the conclusion that within the field of medicine and clustering diseases a great deal of work still needed to be done to fully understand and be able to use it in the industry, however, this has been a good first step, towards using unsupervised machine learning and clustering analysis in the medicinal sector.

## How did we use clustering in the assignment?

During our assignment we utilized clustering to be able to cluster the dataset Forest Fires and find interesting patterns. When doing unsupervised machine learning, exploratory data analysis is essential to get a feeling on which model to use. When clustering is used, it is important to have every feature as a numeric feature. This means converting categorical features into numbers, which can be done in a few different ways. The one that we chose, was to replace the months and the days, with numbers from respectively 1-12 for the months and 1-7 for the days. This allows us to explore the relationships between the features. Having looked at the relationships between our features, we decided to move forward with our analysis by analyzing the conditions affecting the likelihood and the area of the fires. Further inspection of the data made us choose our final columns for the analysis, Drought Code and Relative Humidity and analyzing these against the affected burned area.

To be able to use clustering for the chosen features, we decided to scale them using logarithms, which makes it easier to compare them and better to work with. After, we utilize the K-Means clustering algorithm to train our model with 4 initial clusters. This creates 4 clusters in a 3-dimensional space, which we can later use when introducing new data into the model.

## Reference List

Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow (2nd ed.). O'Reilly Media Inc.

Nathiya, G., Punitha, S. C., Punithavalli, M. (2010). An Analytical Study on Behavior of Clusters Using K Mean, EM and $K^*$ Means Algorithm. Internation Journal of Computer Science and Information Security, Vol. 7, No. 3, March 2010.