AMI23B – Business Intelligence

This lab includes two tasks:

Task1: Descriptive Analysis, Unsupervised Learning – IKEA

This task is about finding and evaluating clusters that contain data with similar properties.

Your task: is to discover some new places here in Sweden that may be suitable for IKEA department stores. You will do this by using the *k-means* method. You have a text file, *ikea_kommun_data.txt*, which contains essential features for many of Sweden's municipalities to aid you in your findings. The English term *municipality* translated to Swedish is *kommun*.

IKEA stores are already available in the following municipalities: Borlänge, Gävle, Göteborg, Haparanda, Helsingborg, Jönköping, Kalmar, Karlstad, Linköping, Malmö, Stockholm, Sundsvall, Uddevalla, Umeå, Uppsala, Västerås, Älmhult, and Örebro. Some of these municipalities are missing in the ikea_data.txt file. The following link shows a map of Sweden's municipalities, https://www.scb.se/contentassets/1e02934987424259b730c5e9a82f7e74/kommunkarta09.pdf

The general steps are data exploration, data transformation, data reduction, and the k-means clustering method.

Task2: Predictive Analysis, Supervised Learning – Titanic

This task is about classifying a large set of data based on a set of pre-classified samples.

Your task: is to predict whether a passenger survived the Titanic shipwreck or not. You will use both a *Decision Tree Classifier* and a *Support Vector Machine* to do this and compare the results.

The general steps are data exploration and analysis, data pre-processing and transformation (handling missing values, converting categorical features into numeric, converting discrete features into binary, etc.), and implementing your classifier.

The classic Titanic dataset provides information on the fate of passengers on the Titanic, summarised according to economic status (class), sex, age, and survival.

You will find two data files:

- Training set (titanic_train.csv) should be used to build your ML models.
- Test set (titanic_test.csv) should be used to see how well your model performs on unseen data.

Data Description and Notes:

Pclass: A proxy for Socio-Economic Status (SES).

- 1st = Upper
- 2nd = Middle
- 3rd = Lower

Age: Age in years. It is fractional if less than 1. If the age is estimated, it is in the form of xx.5.

SibSp: The number of siblings/spouses aboard the Titanic. The dataset defines family relations in this way:

- Sibling = brother, sister, stepbrother, stepsister
- Spouse = husband, wife (mistresses and fiancés were ignored)

Parch: The number of parents/children aboard the Titanic. The dataset defines family relations in this way:

- Parent = mother, father
- Child = daughter, son, stepdaughter, stepson
- Some children travelled only with a nanny, therefore Parch = 0 for them.

Embarked: The port of embarkation, C = Cherbourg, Q = Queenstown, S =Southampton.

Ticket: The ticket number.

Fare: The passenger fare.

Cabin: The cabin number.

Main Python libraries to use:

- scikit-learn (a Python library that features various classification, regression, and clustering algorithms) https://scikit- learn.org/stable/
- pandas https://pandas.pydata.org/docs/
- NumPy https://numpy.org/
- Matplotlib https://matplotlib.org/
- seaborn: statistical data visualisation https://seaborn.pydata.org/

"You can have data without information, but you cannot have information without data."

~ Daniel Keys Moran