AMI23B – Business Intelligence

This lab includes two tasks:

Task1: Descriptive Analysis, Unsupervised Learning – IKEA

This task is about finding and evaluating clusters that contains 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. To aid you in your findings, you have a text file, ikea_data.txt, which contains important features for many of Sweden's municipalities. 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

General steps to follow: data exploration, data transformation, data reduction and then implement 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.

General steps to follow: data exploration and analysis, data pre-processing and transformation (handle missing values, convert categorical features into numeric, convert discrete features into binary etc.), implement your classifier.

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

You will find two data files:

- Training set (train.csv), should be used to build your ML models.
- Test set (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: In years. Age is fractional if less than 1. If the age is estimated, is it in the form of xx.5

sibsp: 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: 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: Port of Embarkation C = Cherbourg, Q = Queenstown, S = Southampton.

Ticket: Ticket number

Fare: Passenger fare

Cabin: Cabin number

Main Python libraries to use:

- Sci-Kit 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 visulasation https://seaborn.pydata.org/

"You can have data without information, but you cannot have information without data." ~ Daniel Keys Moran