University of Bejaia
Faculty of Exact Sciences
Department of Operating Researchers
Academic Year: 2023/2024

First year Master SDAD (S2) Module: **Machine-Learning and Data mining**

TP 3: Classification with KNN (Supervised learning)

1/ Introduction

This lab provides a general overview on how to build and evaluate a supervised learning algorithm such as a classification. We focus on the KNN algorithm.

2/ Learning Objectives

- Train and test data
- Making predictions
- Evaluating predictions

3/ Programming steps

To implement these 3 steps in Python, we need to take the following steps:

- Import all the necessary libraries.
- Preparing the DataSet. The machine receives data characterized by X variables (called features) and annotated with a y variable (called a label or target).
- Select the model (or estimator) the machine needs to learn, specifying the model's hyperparameters ***(KNeighborsClassifier)***.
- Train the model on data X and Y : **model.fit(X,Y)**
- Evaluate the model : **model.score(X,Y)**
- Use the model : **model.predict(x)**

4/ Software tools

You install anaconda (https://www.anaconda.com/), which is the best environment for machine learning coding and the sklearn library. If you use just Jupyter or spyder, you have to install the appropriate libraries.

Duration: 2 hours (+ 2 hours homework)

---------- **Activities** ----------

Activity 1

Suppose we have the height, weight and T-shirt size of some customers and we need to predict the T-shirt size of a new customer based only on the height and weight information we have. The height, weight and T-shirt size data are shown below.

Height (in cm)	Weight (in kgs)	T Shirt Size
158	58	M
158	59	M
158	63	M
160	59	M
160	60	M

163	60	M
163	61	M
160	64	L
163	64	L
165	61	L
165	62	L
165	65	L
168	62	L
168	63	L
168	66	L
170	63	L
170	64	L
170	68	L

1. Create the KNN model in Python with two different solutions : the first using the sklearn library, the second without using sklearn.
2. Predict the size of the T Shirt for values not belonging to the dataset.

Activity 2

Consider a regression problem that generates a real numerical value (1; 3.5; 7.2) as an output. For example, the table below estimates the age of a crab (decimal value label) as a function of its width and mass (predictors).

Width	Mass	Age
50	100	1
150	500	3.5
200	1200	7
225	1800	10
130	250	3

Create a jupyter notebook to predict new values (width and mass) using the KNN algorithm. (The first step is to run the algorithm manually).