**Aprendizagem Automática Avançada - Project Proposal - 2022/23**

**by Martim Silva 51304 and Alexandre Sobreira 59451**

**1- Goal of the Project**

Our proposed work will be inspired by a paper from Hiremath et al (2022) that used a support vector machine (SVM) to determine human gender using face photographs of various ages. In this paper the authors obtained a model with 91.63% accuracy. The SVM used in this paper was a SVM-Classifier using a RBF Kernel.

In this project, the technique used on the above mentioned paper will be replicated, and the results compared. Also, other techniques, like Convolutional Neural Networks and SVMs using other kernel functions, will be applied to try to solve the same problem and compare the result in order to see if it is a better algorithm to this type of problem and data. Another analysis is proposed which is the classification by age apart from gender which is done on the mentioned paper.

**2- Libraries, tools used to reach goal**

The tools used for this project will be using Python3 through Jupyter Notebook to enact preprocessing of the data, selection of samples, we will also use scikit-learn, pandas and numpy for things like evaluating models (model metrics), we will employ matplotlib/seaborn for visualising results, to create models we will use scikit-learn for SVM, Keras from Tensorflow for Convolutional Neural Networks among other libraries like PIL for things like displaying images, transforming them into arrays and so on since the dataset is comprised of only images.

**3- Dataset that is going to be used for that**

To attend this project goal, the same dataset used in mentioned paper will be used.

The dataset used will be UTKFace, it is a data set that consists of more than 20,000 pictures of faces that vary in terms of gender (0 for male, 1 for female), age (from 0 to 116 years old) and ethnicity, with this characteristic detailed in each image’s file name. The images cover large variation in pose, facial expression, illumination, occlusion, resolution, etc.

The dataset will be taken from the Kaggle platform and is available through this link: <https://www.kaggle.com/datasets/jangedoo/utkface-new?datasetId=44109>

**4- Dataset characteristics, problems and how is intended to solve them**

Some problems we are bound to face include the sheer volume of, not rows of numbers or strings, but of images which by default are larger in how much memory it takes to store each one, but also the fact that this set holds more than 20,000 of these. We intend to make an analysis of the distribution of the photographs basing ourselves on the two relevant factors for our goal, the age and the gender of the person and have this distribution be as balanced as necessary for each range of ages and whether the person is male or female. We will bin the photos according to predefined age intervals for both genders and use a subsample of each to achieve this goal.