Chapter 1

Introduction

In this report we will discuss about the whole data pipeline of a project that uses different techniques of machine learning to classify some tabular data. In particular, we choose the first project proposal that aims to predict the average mark of a film from its features. The first step of the data pipeline requires to download, save and load in memory the MovieLens [?], TMDB [?] and IMDb [?] datasets. Consequently, the datasets were elaborated with the objective to generate a unified dataset that can be used as an input for some machine learning algorithms. Afterwards, during the modelling phase we built an MLP model thanks to the PyTorch framework [?]. In addition, we used other techniques like SVM, tree methods and naive bayes methods that are available into the Scikit-Learn library [?]. In order to find a good configuration during the performance analysis, it was mandatory to define a large enough hyperparameters space for all the models that we defined. Moreover, in the last phase, the cross validation technique was used which resulted in more robust statistics that has been compared between the trained models.