Data Analytics Project Report

A.Y. 2023/2024

Claudia Citera

claudia.citera@studio.unibo.it

Riccardo Evangelisti

riccardo.evangelist6@studio.unibo.it

INDICE

da aggiungere

1 Introduzione

The following project was developed as part of the **Data Analytics** course of the Master's Degree in Computer Science at the University of Bologna.

The objective of the project is to carry out a data analytics study, which involves the implementation of all the analytical pipeline phases studied during the course:

- Data Acquisition
- Data Visualization
- Data Preprocessing
- Modeling
- Evaluation

The main purpose of this study is to recognize the year in which a song was published based on the features of its audio track.

The following functionalities were developed:

- Traditional non-deep supervised Machine Learning techniques
- Supervised ML techniques based on neural networks
- Supervised ML technique with deep models for TabularData

2 Data Acquisition

The data acquisition phase involves collecting the data that needs to be analyzed. Data can be acquired in various ways, including static acquisition, which was used in this project.

The dataset used in this project consists of a single csv file with 252175 rows and 91 columns. One of the columns represents the year of publication of the song, ranging from 1956 to 2009. All other columns contain floating-point numbers related to the audio track, making the entire dataset continuous. For this reason, regression models were used to solve the problem.

Before proceeding with further operations, the dataset was analyzed to check for missing or duplicate values. It was found that there are no missing values and only 52 duplicate rows. Since they are very few and mainly related to the most represented classes, we decided to remove them.

3 Data Visualization

Data visualization is a technique that aims to communicate data in a clear and effective manner through the use of graphs. It is useful for exploring data efficiently and discovering possible relationships.

In this section of the project, several graphs and visual representations were produced to provide an overview of the dataset in question. Since the dataset has a high number of features, we chose to focus only on some key information and represent it visually to make it easier to understand and analyze the data.

As a first step, we checked the correlation between the various columns to see if it was possible to remove any of them.

From Figure 3.1 we can notice that the dataset shows a low degree of correlation among its various features. Even the correlation with the Year column is very low, so it is irrelevant for our investigation. However, a closer examination of the top-left quadrant reveals a cluster of variables exhibiting stronger correlation. A zoomed-in view of this region is provided below (Figure 3.2) for further analysis.

> IMMAGINE CORRELAZIONE GRANDE come Figura 3.1
>
> > TITOLO Correlation between features

> IMMAGINE CORRELAZIONE PICCOLA come Figura 3.2
>
> > TITOLO Correlation close up

A very relevant piece of information is the distribution of the Year attribute, which reveals that the dataset is unbalanced. In fact, more recent songs (from around 1990 to 2009) are much more represented than older ones (from 1990 backwards). It is very likely that we will attempt Under Sampling and Over Sampling.

Then we checked the distribution of the other varibles and we noticed that all the attributes exhibit a **Gaussian distribution** with varying skewness and their means are centered around 0 (or close to it), except for the attribute `S0`.
The plots indicate that the distribution ranges differ a lot across the various columns. Therefore, it's probable that we will require a preprocessing model to **rescale** the data. Since there are 90 columns (all except *Year*), for space reasons we decided not to include this plot in the report, but it is present in the notebook entitled *0_DataAcquisition+Visualization.ipynb*.