**Preprocessing :**

Dropped the rows that have any Null / Missing values (Using the dropna built-in functions).

Dropped the unnecessary columns (Using the drop built-in functions).

Applying the One-hot encoder technique on the “KEY” column (How ? import preprocessing from sklearn and then using one-hot encoder algorithm to apply it on our column).

Applying the Label encoder technique on the “Artists” column (How ? import preprocessing from sklearn and then using Label encoder algorithm to apply it on our column).

**Analysis :**

Using the correlation technique to find the top related features (Using heatmap).

We found the top features affected on the popularity.

Frist one “year” feature affected with 0.85 on the popularity.

Second one “energy” feature affected with 0.47 on the popularity.

Third one “loudness” feature affected with 0.45 on the popularity.

Fourth one “explicit” feature affected with 0.18 on the popularity.

Fifth one “danceability” feature affected with 0.17 on the popularity.

We selected the top three features (year , energy , loudness).

Graphical user interface

Description automatically generated with medium confidence

**Regression techniques :**

Frist Model : We used multiple liner regression with three features (Using it to reduce the mean squared error more than simple liner regression and to give us model more related / accurate to make relation between our features and the target).

Second Model : We used polynomial regression with three features with degree 2 (Using it to reduce the mean squared error more than liner regression because it makes the model fit more on our data using it’s polynomial equation not liner equation).

**Differences between 2 Models :**

|  |  |  |
| --- | --- | --- |
|  | **Model1 (Multiple liner regression)** | **Model2 (Polynomial)** |
| **Mean squared error** | **~ 123** | **~ 121** |
| **Root mean squared error** | **11.12** | **11.03** |
| **Accuracy** | **0.73** | **0.74** |
| **Training Time (Minute)** | **0.0127** | **0.057** |

**Features :**

Used features : year , energy , loudness

Discarded features : name , id , release date , valence , acousticness , artists , danceability , duration\_ms , explicit , instramentalness , key , liveness , tempo , speechiness and mode.

**Training / Test sets Size :**

Training set : 70% from the whole dataset.

Test set : 30% from the whole dataset.

This is a suitable splitting to make the model make good training and to have enough data to test the model with it to ensure the model work in a good way.

**Screenshots :**

**Chart, scatter chart

Description automatically generatedChart

Description automatically generatedChart, scatter chart

Description automatically generated**

**Conclusion :**

The dataset is very large and scattered on a large scale and this not lead to an excellent model with a lager accuracy but still get good acceptable accuracy and this is proved from trying to fit the model with a large degree but the model can’t make a change on the MSE.