**Project Title**

**Designing an End-to-End Machine Learning Use Case**

**Team Members**

**Amin Adel**

**Abdullah Ali**

**Mohamed Gamal**

**Moataz Mohamed**

1. Overview of the Dataset

The dataset contains the following columns:

popularity: The popularity score of the song.

acousticness: A confidence measure of whether the track is acoustic.

danceability: How suitable a track is for dancing.

duration\_ms: The duration of the song in milliseconds.

energy: The intensity and activity measure of the song.

instrumentalness: Predicts whether a track contains no vocals.

liveness: The presence of an audience in the recording.

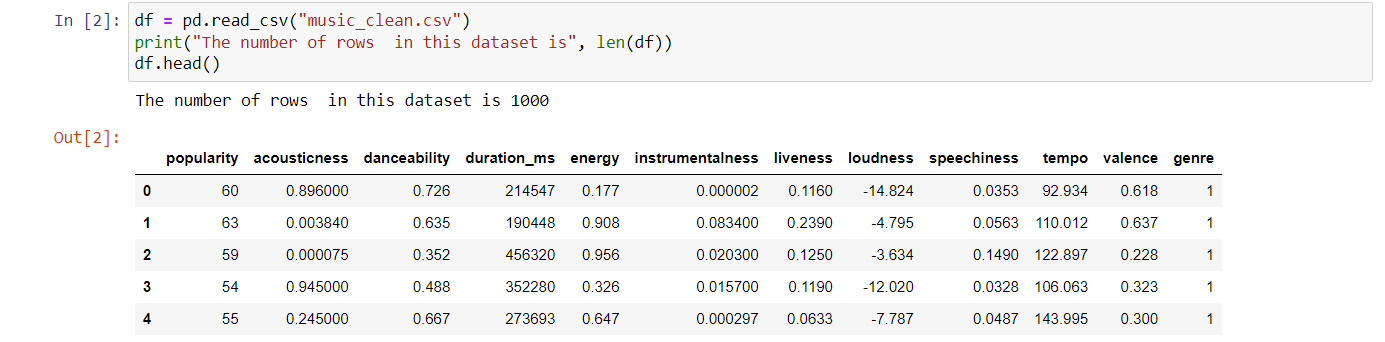
loudness: The overall loudness of the track in decibels (dB).

speechiness: The presence of spoken words in a track.

tempo: The speed or pace of the song, measured in beats per minute (BPM).

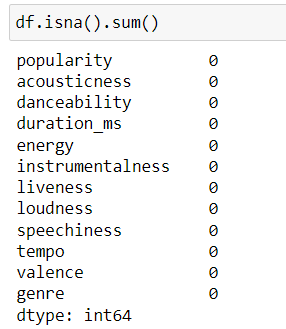
valence: The musical positiveness conveyed by a track.

genre: The genre of the song (all entries are '1,0', which suggests a binary genre focus).



2. Summary Statistics

Let's check if there is null values and compute some basic summary statistics for each column.



After that, I checked to see if the dataset had duplicates, and my dataset had no duplicates.

And this is my dataset Decription

A table with numbers and symbols

Description automatically generated

3. Correlation Analysis

We'll examine the correlation between different variables to identify any significant relationships.

Step 1: Calculate the Correlation Matrix

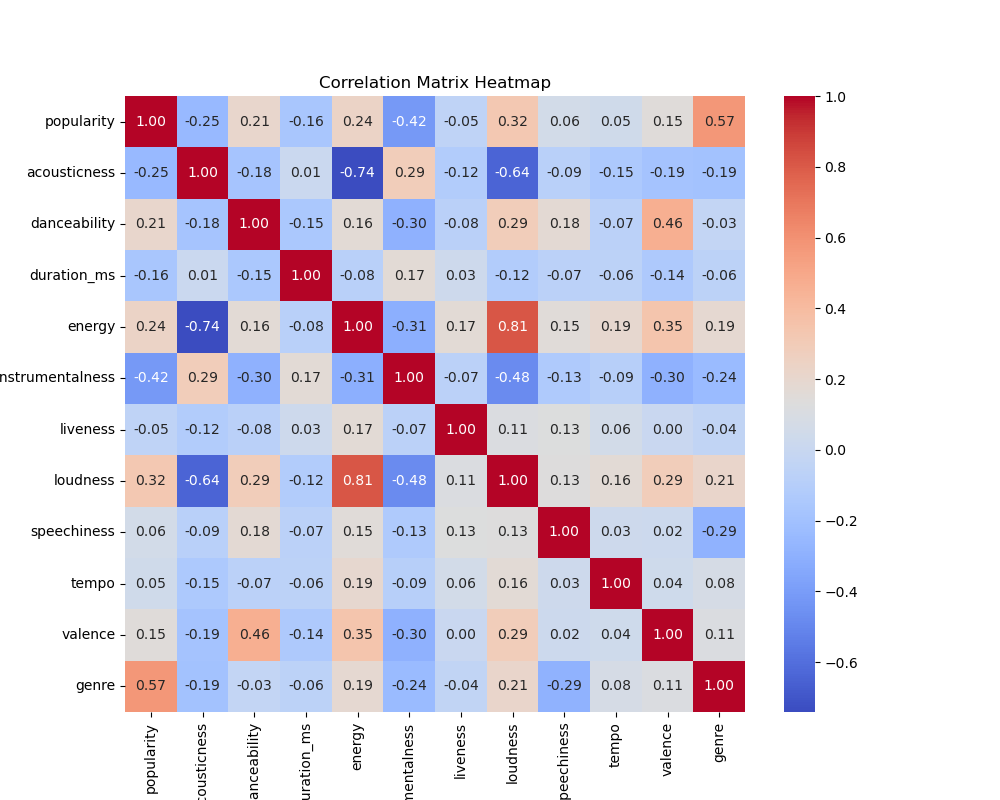
Using a tool like pandas in Python, you can calculate the correlation matrix, which shows the Pearson correlation coefficient between each pair of variables.

A black and white text

Description automatically generated with medium confidence

Step 2: Interpret the Correlation Matrix

The output of the above code is a correlation matrix, where each cell (i, j) represents the correlation coefficient between the ith and jth variables.



Key Observations

energy and loudness: A correlation of 0.81 indicates a strong positive relationship.

Popularity and genre: A correlation of 0.57 indicates a moderate positive relationship.

valence and danceability: A correlation of 0.46 indicates a moderate positive relationship.

Acousticness and energy: A correlation of -0.74 shows a strong negative relationship.

loudness and Acousticness: A correlation of -0.64 shows a strong negative relationship.

loudness and instrumentalness: A correlation of -0.48 shows a moderate negative relationship.

Popularity and instrumentalness: A correlation of -0.42 shows a moderate negative relationship.

4. Histogram

Histograms give a clear visual representation of the distribution of data in each column. You can see if the data is normally distributed or not.

Histograms can help in identifying extreme values that may need further investigation or handling.

A group of blue graphs

Description automatically generated

5. Outlier Handling

We will try the model with outlier handling, then try it without outlier handling. Because the outliers may be useful for my dataset, the accuracy of the model will determine that.

To handle the outliers, we will make a boxplot to see if there is an outlier or not, and if there are outliers, we will use the IQR method to handle them. If there are no outliers, we will ignore this column and check the next one.

Let’s start with the 1st column named popularity.

A diagram of a box plot

Description automatically generated

We found that there are outliers, so we will handle it using the IQR method.

After handling the outliers, the figure will change and be like this:A blue and black box plot

Description automatically generated

Let’s continue with the 2nd column named acousticness.

A diagram of a box plot

Description automatically generated

There are no outliers.

Let’s continue with the 3rd column named Danceability.

A diagram of a danceability

Description automatically generated

We found that there are outliers, so we will handle it using the IQR method.

After handling the outliers, the figure will change and be like this:

A diagram of a danceability

Description automatically generated

Let’s continue with the 4th column named duration\_ms.

A graph with a line and a line

Description automatically generated with medium confidence

We found that there are outliers, so we will handle it using the IQR method.

After handling the outliers, the figure will change and be like this:

A diagram of a box plot

Description automatically generated

Let’s continue with the 5th column named Energy.

A diagram of energy and energy

Description automatically generated

There are no outliers.

Let’s continue with the 6th column named instrumentalness.

A graph of a box plot

Description automatically generated

We found that there are outliers, so we will handle it using the IQR method.

After handling the outliers, the figure will change and be like this:

A diagram of a box

Description automatically generated

Let’s continue with the 7th column named Liveness.

A diagram of a box plot

Description automatically generated

We found that there are outliers, so we will handle it using the IQR method.

After handling the outliers, the figure will change and be like this:

A diagram of a box plot

Description automatically generated

Let’s continue with the 8th column named Loudness.

A diagram of a box plot

Description automatically generated

We found that there are outliers, so we will handle it using the IQR method.

After handling the outliers, the figure will change and be like this:

A diagram of a box plot

Description automatically generated

Let’s continue with the 9th column named speechiness.

A graph of a speech plot

Description automatically generated with medium confidence

We found that there are outliers, so we will handle it using the IQR method.

After handling the outliers, the figure will change and be like this:

A diagram of a box plot

Description automatically generated

Let’s continue with the 10th column named Tempo.

A diagram of a box plot

Description automatically generated

We found that there are outliers, so we will handle it using the IQR method.

After handling the outliers, the figure will change and be like this:

A diagram of a box plot

Description automatically generated

Let’s continue with the 11th column named Valence.

A blue and black box diagram

Description automatically generated

There are no outliers.

6. Preparing the data columns and target column.

7.Splitting the data to 60% for training, 20% for validation and 20% for testing.

8.Feature scaling

I used feature scaling in my data preprocessing stage to ensure that all features contribute equally to the machine learning model's performance. Feature scaling standardizes the range of independent variables or features of data, bringing them to a common scale without distorting differences in the ranges of values. This is crucial because support vector machines (SVM) are sensitive to the magnitudes of feature values. By scaling the features, the optimization process becomes more efficient, leading to faster convergence and improved model accuracy.

9. Hyper Tuning

Hyperparameter tuning is crucial because it allows the model to find the most optimal settings for its parameters, leading to improved performance and generalization on unseen data. Grid search systematically works through multiple combinations of parameter tunes, cross-validating as it goes to determine which tune gives the best performance.

10. Train the models with outlier handling applied

10.1 Logistic regression model:

The validation accuracy is 0.88.

The test accuracy is 0.90.

This is the classification report:

A screenshot of a screen

Description automatically generated

10.2 Support Vector Machine model:

The validation accuracy is 0.93.

The test accuracy is 0.94.

This is the classification report:

A screenshot of a report

Description automatically generated

10.3 Random Forest model:

The validation accuracy is 0.99.

The test accuracy is 1.0.

This is the classification report:

A screenshot of a computer

Description automatically generated

11. Train the models without outlier handling

11.1 Logistic regression model:

The validation accuracy is 0.91.

The test accuracy is 0.91.

This is the classification report:

A screenshot of a computer screen

Description automatically generated

11.2 Support Vector Machine model:

The validation accuracy is 0.93.

The test accuracy is 0.94.

This is the classification report:

A screenshot of a computer screen

Description automatically generated

11.3 Random Forest model:

The validation accuracy is 0.99.

The test accuracy is 1.00.

This is the classification report:

A screenshot of a computer screen

Description automatically generated