**Delivery Approach:**

**Questions of Interest:**

**Main Question:** Is it possible to classify songs into the genres of 'Hip-Hop' or 'Rock' using data analysis without listening to the songs?

**Secondary Questions:**

* What song features are most relevant for classification?
* What is the performance of machine learning models in this classification task?
* How can music recommendations for users be improved?

**Type of Problem:**

The type of problem this project belongs to is a supervised *classification problem*. The goal is to assign each song to one of the two classes: 'Hip-Hop' or 'Rock', based on data features without listening to the songs.

**Background:**

The background domain of the problem is the field of music. In this case focusing on classifying songs into two genres: hip-hop and rock. To do this, audio data from the songs is used, such as bit rate, acousticness,danceability,energy among others.

**-Song classification is an important task in the music industry.** Streaming companies use the rating to recommend songs to users.

**-Song classification can be done manually or automatically.** Manual sorting is expensive. Automatic classification uses machine learning algorithms to classify songs. Common features include fundamental frequency, energy, rhythm, and complexity.

**-Machine learning algorithms have proven to be effective for song classification.** However, these algorithms require a large training data set that must be representative of the population of songs you want to classify. Music genres are not clearly defined and can be ambiguous.

**Academic Research**

A 2012 study (Eronen & Virtanen, 2012) evaluated the performance of different audio features for music genre classification. The researchers found that the most relevant audio characteristics for musical genre classification were timbre, fundamental frequency and energy.

A 2013 study (Hu & Wang, 2013) proposed a new machine learning algorithm for music genre classification. The algorithm was based on combining audio features and text features. The algorithm was shown to be superior to existing algorithms in terms of accuracy and robustness.

A 2023 study (Zhang et al., 2023) investigated the use of deep learning for music genre classification. The researchers trained a deep learning model on a dataset of songs tagged with their genre. The model was shown to be superior to existing algorithms in terms of accuracy and efficiency.

**Solutions implemented by third parties**

Spotify for Artists, Spotify Discover and Apple Music Replay stand out.

These common elements are the basis of the classification of musical genres. Audio features are used to represent the properties of a song, and machine learning algorithms are used to learn patterns in audio features that are associated with different musical genres.

**Metrics to use:**

-Precision: Precision is the proportion of songs that are classified correctly.

-Recall: Recall is the proportion of songs of a specific genre that are classified correctly.

-F1-score: The F1-score is a combination of precision and recall.

-Execution time: Execution time is the time it takes a system to classify a song.

**Databases:**

Mainly we have two files that act as databases, a csv file and a json file. Both files have different attributes, they only share one in common which is the TRACK\_ID.

**Step by Step:**

The first thing we did after reading the statement was to perform a preliminary analysis of both documents, we extracted important information such as the weight of each of the documents, how many rows it had, what the columns were, what type of data each of the columns were, if it had nulls or duplicate elements and a preliminary view of the first and last five elements of each of the two documents. In the code in google collabo you can verify these operations.

Then the next thing we did was to join the two documents by the track\_ID and by performing this operation we had several interesting observations, these are that the number of rows that were in the first document which was the Csv file were reduced from 17,000 to 4000 since the Json file (which is very important because it adds attributes that will be of vital importance when making a decision) contains a smaller number of rows that share the same track\_ID. For now we consider to continue working with this reduced number of data that we have of rows and in next deliveries we will improve and we will plan a method of collection of more data. but, for now we are going to guide us with this data that we have when joining the two documents (dff) once the two documents are joined, we proceed to do the same steps that we did with each one but this time already with the two Joins. We took the descriptive statistics of each one of the elements of this new set and we could also appreciate that there

were several attributes that had null, it is worth mentioning that there was no duplicated attribute, then between us we decided to delete some of the attributes that had many null, because based on our research we consider that they are not of vital importance when deciding if a song is of the rock or hip hop genre.

We finished this first part of the project by carrying out the strategy of collecting new data, as well as analyzing the impact of artificial intelligence in general on the planet.

**Second delivery**

This second installment was aimed at obtaining new data and making the first classification models.

To obtain new data, we searched different sites that contain databases, in order to match the disparity between Rock and Hip-Hop data. So, we focused on databases that could provide us with more Hip-Hop data. We found these two databases on Kaggle: <https://www.kaggle.com/datasets/mrmorj/dataset-of-songs-in-spotify> and <https://www.kaggle.com/datasets/fedeschmidt/music-genre-classification-ia> .

**Step by Step:**

First, once we obtained the databases, we continued to perform an in-depth exploration of the database. We focused on finding which database had the most attributes in common with our data set. Furthermore, this data was really relevant and did not have a large number of null values. In the end we decided on the Spotify dataset as it had the attributes we were looking for. Next, we proceeded to integrate this new set of data with the one we already had. At this point we encountered different problems. The first obstacle we had was that the new dataset, although it had the attributes we needed, had some columns that were unnecessary. and others had a different name than the one that appeared in the dataset we already had. We solved this by eliminating unnecessary attributes and renaming the attributes that were necessary. Next, we perform a concatenation operation to join our two datasets. Then, we explored the new dataset, here we found several relevant insights. Since we were able to equate the number of rock songs with hip hop songs, we also showed that several attributes had many outlier values and their distribution between the hip hop and rock genres presented a bias, so we proceeded to eliminate these through the use of quartiles.

When we finished integrating the data, we moved on to developing the first classification models. For this delivery we implemented logistic regression, knn and random forest models. In the process of creating a logistic regression model, we began by loading and preprocessing the data, including cleaning and converting categorical variables into numerical ones. Subsequently, the data set was divided into training data and test data at 80% and 20% respectively. Then, the logistic regression model was built and trained using scikit-learn. The evaluation of the model was performed based on the test data, calculating the performance metrics mentioned above in the first part of this report. The results were visualized using graphs, including representation of the decision boundary in a scatter plot. Based on the results, they were extracted and a table was generated where the results were recorded. .Also, we decided to use Random Forest as one of the models to solve the problem; it seemed more appropriate to use all the columns of the dataframe as predictors. First we trained the model and took out the confusion matrix and the metrics, and we realized that there was overtraining with 100% in all metrics in the training data and 50.57% in the test and also in the Confusion matrix in the test data could be noticed as half of the data were false negatives. To fix overtraining we modify 3 hyperparameters of the model separately. The first was the n\_estimators we say make several changes at once to see which gave the best probability and fit between the training and test data. In the case of this hyper parameter we try with 1,2,4,8,16,32,64,100 and 200 estimators. For the second hyper parameter we tested with the max\_depth and in this case we tested with depths from 1 to 16 and, finally, it was the max\_features that we tested between 1 to 8 features. The results that seemed the best to us were in the max\_depth hyper parameter. Because it was the one in which the vast majority of its results showed better proportionality between training and test results, avoiding overtraining. Continuing the process we approach the KNN algorithm, first addressing the baseline, and then choosing the predictors to address. For this algorithm we choose the following: all attributes, then liveness, speechiness, then acousticness, danceability, tempo, valence and finally danceability, tempo, valence. The next step was to convert the target variable (genre) into a categorical variable, then applying one hot coding. We split the training sets in an 80/20 ratio respectively. We build the model with k equivalent to 5 and proceed to make the corresponding predictions. We obtain the metrics and the corresponding confusion matrix.

Finally, the analysis of the impact of the implementation of the solution was carried out in multiple dimensions: global, environmental, social and economic. Globally, it was determined that the solution could optimize the organization of music catalogs on streaming services, resulting in a reduction in operational costs and an improvement in the user experience. In addition, the potential to generate new forms of musical consumption based on personalized recommendations was identified.

From an environmental perspective, it was concluded that the solution could contribute to the reduction of energy consumption by automating musical organization tasks.

In social terms, the solution's ability to promote musical inclusion and diversity was highlighted by incorporating music genres from around the world into the model. However, it was emphasized that the model should not be used to stereotype people based on their musical preferences.

In terms of economic impact, the possibility of generating employment in the technology sector was identified, although concerns were also raised about the possible loss of jobs for those currently performing manual music classification. Additionally, the potential of the model to help artists reach a broader audience through personalized recommendations and to facilitate the work of music producers in differentiating songs of similar genres was highlighted. These conclusions arose from the comprehensive analysis of the aforementioned aspects.

**Final delivery**

In the final delivery, we continued to refine and enhance the classification models based on the insights gained from the second installment. We also conducted a more comprehensive analysis of the impact of the solution on various dimensions.

**Model refinement:**

After the initial implementation of logistic regression, KNN and random forest models in the second installment, we made further adjustments to improve model performance. This involved the use of PCA for dimensionality reduction, we performed tests before and after using PCA, we also used Grid Search to find out the best predictors for the models. For each model we evaluate its advantages and disadvantages compared to the others, in order to have an idea of which could be the best model for the context of the problem.

**Evaluation metrics:**

We continue to use the metrics mentioned above (precision, recall, F1 score, and execution time) to evaluate the performance of the refined models. The evaluation metrics provided valuable information about the effectiveness of each model.

**Impact Analysis:**

The impact analysis was expanded to cover in more detail the global, environmental, social and economic dimensions.

**Global Impact:**

The implementation of the solution offers significant advantages to streaming services, enhancing the efficiency of music catalog organization. This not only reduces operational costs but also greatly improves the overall user experience. Furthermore, the solution has the potential to revolutionize music consumption by introducing personalized recommendations based on audio data, creating new opportunities for artists and the music industry. It has the capacity to transform the way people consume, sell, create, and distribute songs, ushering in a new era of innovation in the music domain.

**Environmental Impact:**

By enabling streaming services to automate music sorting without human intervention, the solution contributes to a substantial reduction in energy consumption. This not only aligns with sustainability goals but also directly reduces the carbon footprint associated with manual music organization processes. Additionally, the decrease in physical storage needs results in less waste, reducing the environmental impact of outdated or obsolete music collections.

**Social Impact:**

The solution plays a crucial role in promoting inclusion and diversity in music. By incorporating new genres into the model, it introduces users to music from diverse cultures and regions, fostering a more interconnected global music community. However, it is imperative to mitigate potential negative impacts, such as avoiding the use of the model for stereotyping individuals based on their musical preferences. The solution contributes to reducing gender disparity by challenging traditional stereotypes in music, creating a more inclusive environment for artists to express diverse musical styles without limitations based on gender expectations.

**Economic Impact:**

The development and implementation of the solution present a dual economic impact. On one hand, it has the potential to generate new jobs in the technology sector, reflecting the growth and demand for expertise in artificial intelligence and data analysis. On the other hand, the automation of music sorting processes may lead to job losses for individuals currently performing manual tasks. Nevertheless, the model's ability to help artists reach a wider audience through personalized recommendations, and facilitate the work of music producers in categorizing complex songs, can contribute to economic growth in the music industry.

**Taking into account what has been developed in this project, possible advances are directed in future iterations.**

**Dynamic User Preferences:**

Introduce mechanisms to adapt the model in real-time based on evolving user preferences. This could involve continuous learning from user interactions and feedback, allowing the model to stay current with changing tastes.

Incorporating Multimodal Data:

**Personalized Playlists:**

Develop a feature that generates personalized playlists based on a user's mood, activity, or context. This could involve analyzing additional metadata, such as tempo and energy of songs, to create playlists tailored to specific situations.

**Collaborative Filtering and Social Integration:**

Implement collaborative filtering techniques to incorporate user similarities and social network data. This can enhance recommendations by considering the preferences of users with similar tastes or integrating social interactions into the recommendation algorithm.

**Adaptive Model Training:**

Implement adaptive training strategies to allow the model to quickly adapt to new genres, artists, or trends in the music industry. This can be crucial in keeping the recommendations relevant and reflective of the dynamic nature of the music landscape.

Cross-Cultural Recommendations:

**Localized Recommendations:**

Tailor recommendations based on geographical locations and local music scenes. This can help users discover and appreciate regional artists and genres that may not be widely known on a global scale.