# Music Genre Classification Project Documentation

## Introduction

As part of the Shai for AI machine learning engineering internship, we participated in a Kaggle competition aimed at classifying music genre based on various features. This project highlighted the importance of simplicity and a proper understanding of the process.

## Introduction Team D's approach involved each member independently developing the project from start to finish. Throughout the process, members shared information on their initial actions, dataset insights, interesting discoveries, and challenges encountered.

## Dataset Overview

The dataset contains the prices and other attributes of almost 54,000 diamonds, which makes it a great dataset for beginners learning data analysis and visualization.

## Data Description

* Artist: Name of the artist.
* Song: Name of the track.
* Popularity: Popularity score of the track (higher values indicate more popularity). Danceability: A score representing how suitable the track is for dancing.
* Energy: A measure of intensity and activity of the track.
* Key: The musical key of the track (e.g., 0 = C, 1 = C♯/D♭).
* Loudness: The overall loudness of the track in decibels.
* Mode: Modality of the track (1 for major, 0 for minor).
* Speechiness: Indicates the presence of spoken words.
* Acousticness: Measures the confidence that the track is acoustic.
* Instrumentalness: Predicts if the track contains no vocals.
* Liveness: Detects if the track was performed live.
* Valence: Measures the musical Positiveness of the track.
* Tempo: The tempo of the track in beats per minute (BPM).
* Duration\_in\_milliseconds: Duration of the track in milliseconds.
* Time\_signature: Time signature of the track.
* Class: Genre of the track (target variable).

## Evaluation Metric

The F1 score was used as the evaluation metric for this classification task. The F1 score is the harmonic mean of precision and recall, with a value ranging from 0 (worst) to 1 (best). The formula for the F1 score is: 𝐹 1 - Score = 2 × precision × recall / precision + recall​

## Data Import and Initial Inspection

## In this phase, we imported the necessary libraries, read the training and test datasets, and performed an initial inspection of the data. We checked the dataset's length, datatypes, summary statistics for numerical features, and assessed missing values and duplicates.

## Key Takeaways:

## Training Dataset: 14,395 rows with 18 columns.

## Test Dataset: 3,600 rows with 17 columns.

## Missing Values: The training and test datasets contained missing values in the 'Popularity', 'key', and 'Instrumentalness' columns.

## Data Cleaning and Preprocessing

We conducted several steps to clean and preprocess the data, including handling missing values, dropping irrelevant columns, and addressing duplicate records.  
  
Key takeaways:

* Irrelevant Features: The 'Track Name' feature was dropped, as it was not predictive of the genre.
* Imputation: Missing values in the 'Popularity', 'key', and 'Instrumentalness' columns were replaced with the respective column's median.

## Exploratory Data Analysis (EDA)

During the EDA phase, we analyzed the distribution of features and examined the correlations between them. Visualizations such as histograms, boxplots, and correlation heatmaps were used to understand the data better.  
  
Key takeaways:

* Feature Distributions: Features such as ' Speechiness’, 'Acousticness', and 'Loudness' showed skewness.
* Boxplots: The Duration\_in\_milliseconds feature showcased many outliers conveyed through an inspection of boxplots plotted for features.
* Correlations: The correlation heatmap revealed that certain features had strong correlations, which may influence model performance.

Given these takeaways and constraints, it was important to utilize a model that is not significantly affected by outliers and skewness.

## Feature Engineering

The CatBoost model was chosen for its ability to handle categorical variables implicitly through one-hot encoding. For other models, frequency encoding was utilized to encode the categorical variables and monitor performance.

## Model Selection and Evaluation

We trained several machine learning models, including Logistic Regression, Decision Tree, Random Forest, and catBoost, to predict the genre of the tracks. We split the data into training and testing sets and evaluated the models using the F1 score.

Key takeaways:

* Model Selection: The CatBoost classifier was chosen for its ability to handle categorical features effectively.
* F1 Score: The models were evaluated based on the macro F1 score, which considers the balance between precision and recall across all classes.

## Final Prediction and Submission

The best model was used to make predictions on the test set. The final predictions were saved to a CSV file with ID and price as required columns for submission.

## Conclusion

This project demonstrated the process of building a music genre classification model using various audio features. The successful application of machine learning techniques enabled us to classify tracks into multiple genres with reasonable accuracy. Further improvements could include addressing class imbalance and exploring advanced feature engineering techniques.

Shai for ai – Kaggle’s Music Genre Classification Competition, Team D’s submission