Eppens, González del Solar, Reiman

SPOTIFY

Buenos Aires Technology Institute 2023

CONTENT

1 Dataset

2 EDA

3 Added variables

4 Pipeline

5 Predictive model choice

6 Hyperparameter Optimization

7 Shap values

8 Key Results

1. DATASET

DATASET: POPULARIDAD EN SPOTIFY

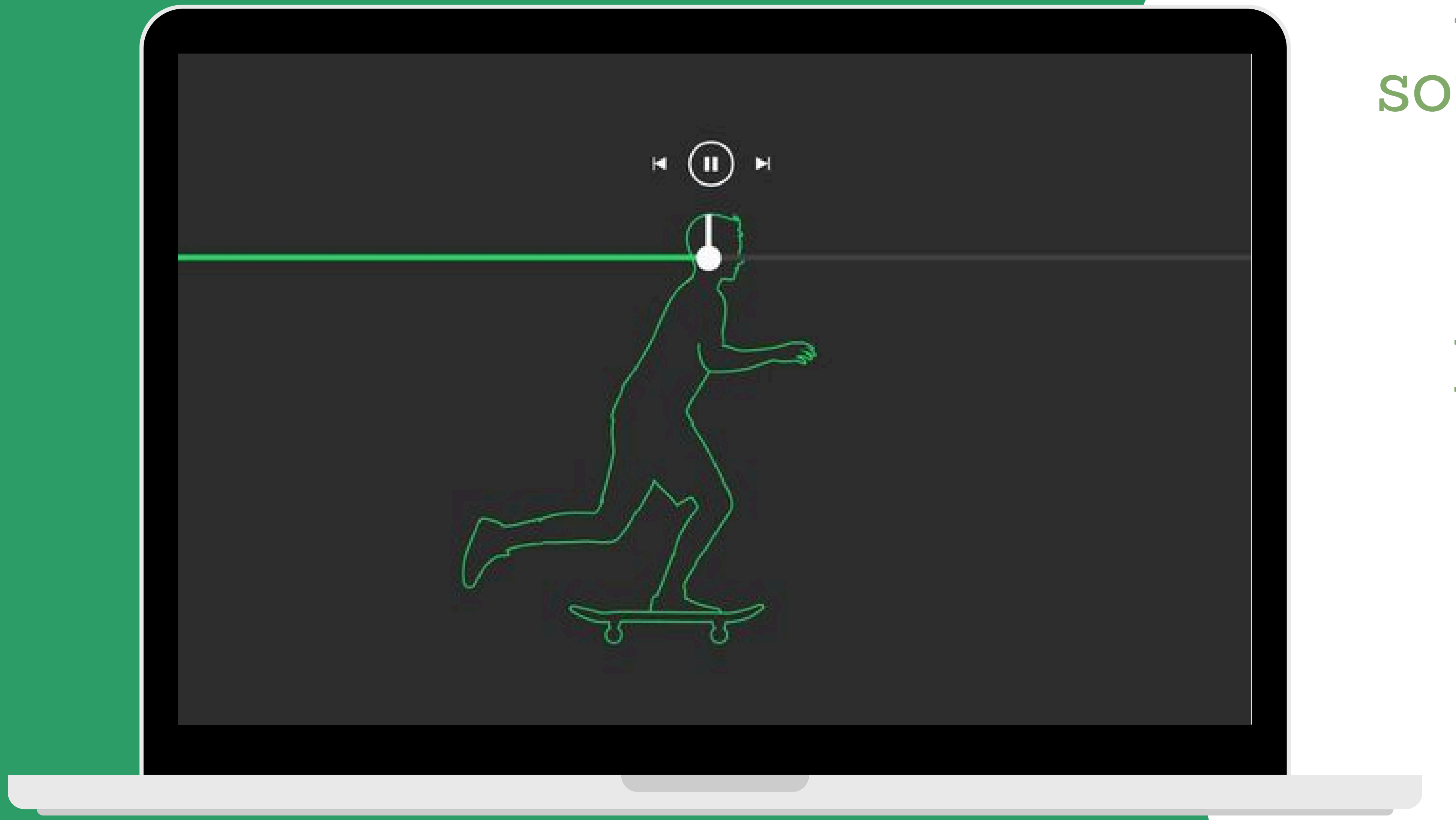
Source Kaggle

Five columns about the song's release details, and the rest describe its musical features

objective variable: popularity (INT 0 ; 100)

Type of model required: Regression

BUSINESS OBJECTIVE



The goal is to help music platforms and producers identify which songs are likely to become hits. By accurately predicting song popularity, stakeholders can better allocate resources to promote tracks with high potential, ultimately boosting visibility and revenue in the music industry

2. EXPLORATORY DATA ANALYSIS (EDA)

ANÁLISIS DESCRIPTIVO

At this stage, we performed exploratory data analysis (EDA) on the dataset. We examined missing values and outliers, counted the number of unique values in categorical columns, and reviewed descriptive statistics for the numerical features.

DATASET SIZE

Rows: 129172
Columns: 17

CATEGORICAL COLUMNS

['artists', 'name']

BINARY COLUMNS

['explicit', 'mode']

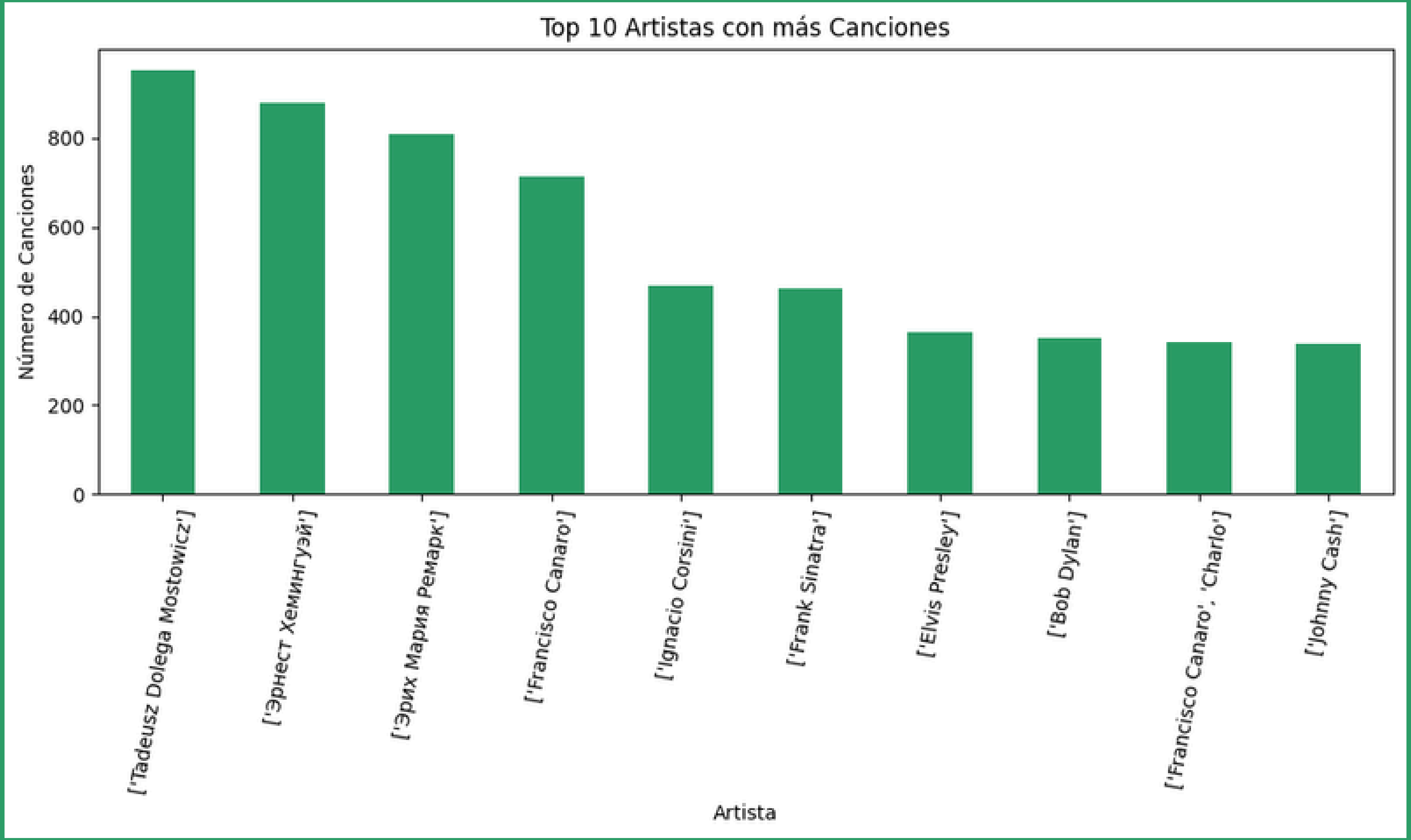
NUMERICAL COLUMNS

['year', 'acousticness', 'danceability', 'duration_ms', 'energy', 'instrumentalness', 'key', 'liveness', 'loudness', 'speechiness', 'tempo', 'valence', 'popularity']

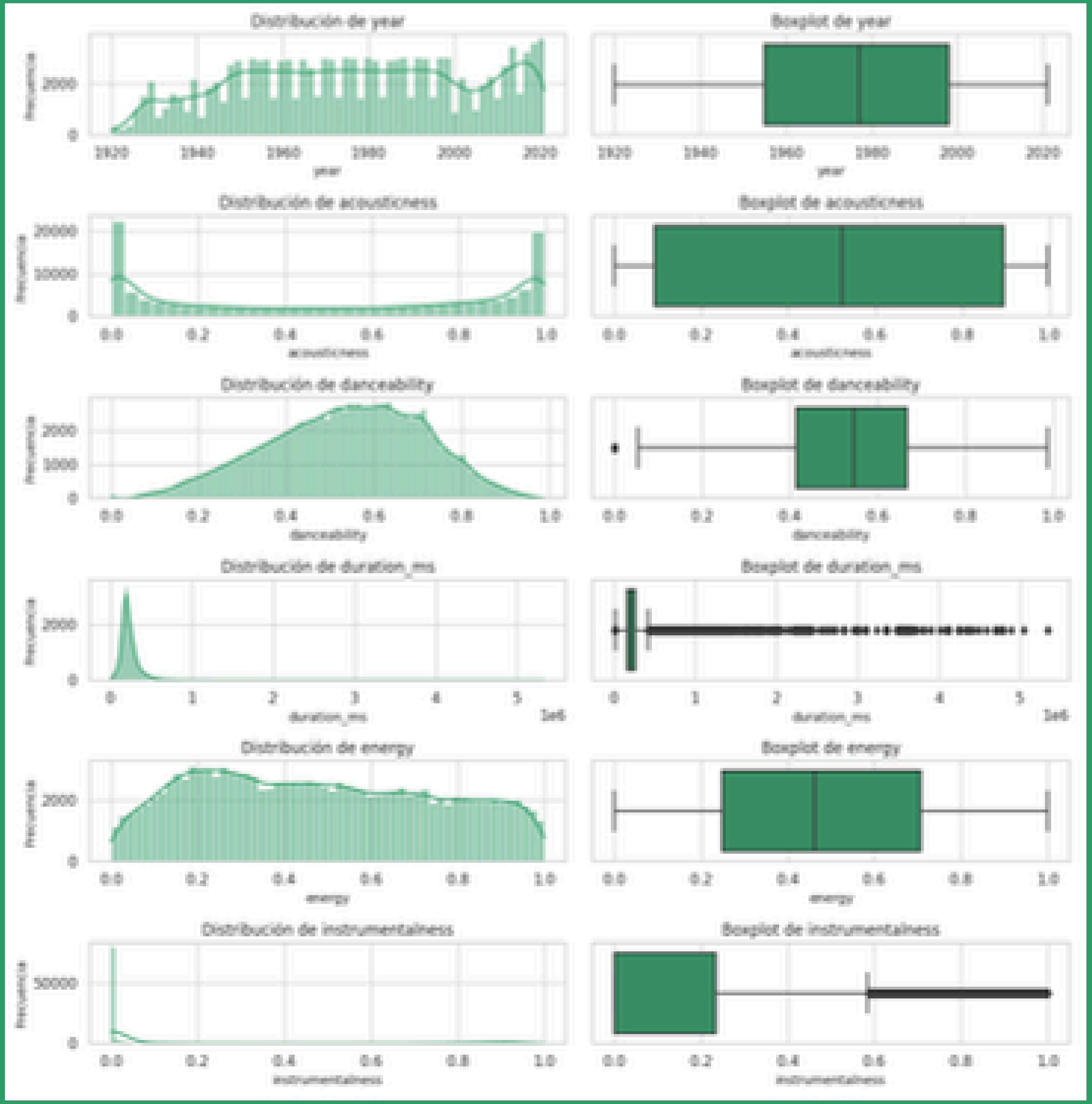
VARIABLES

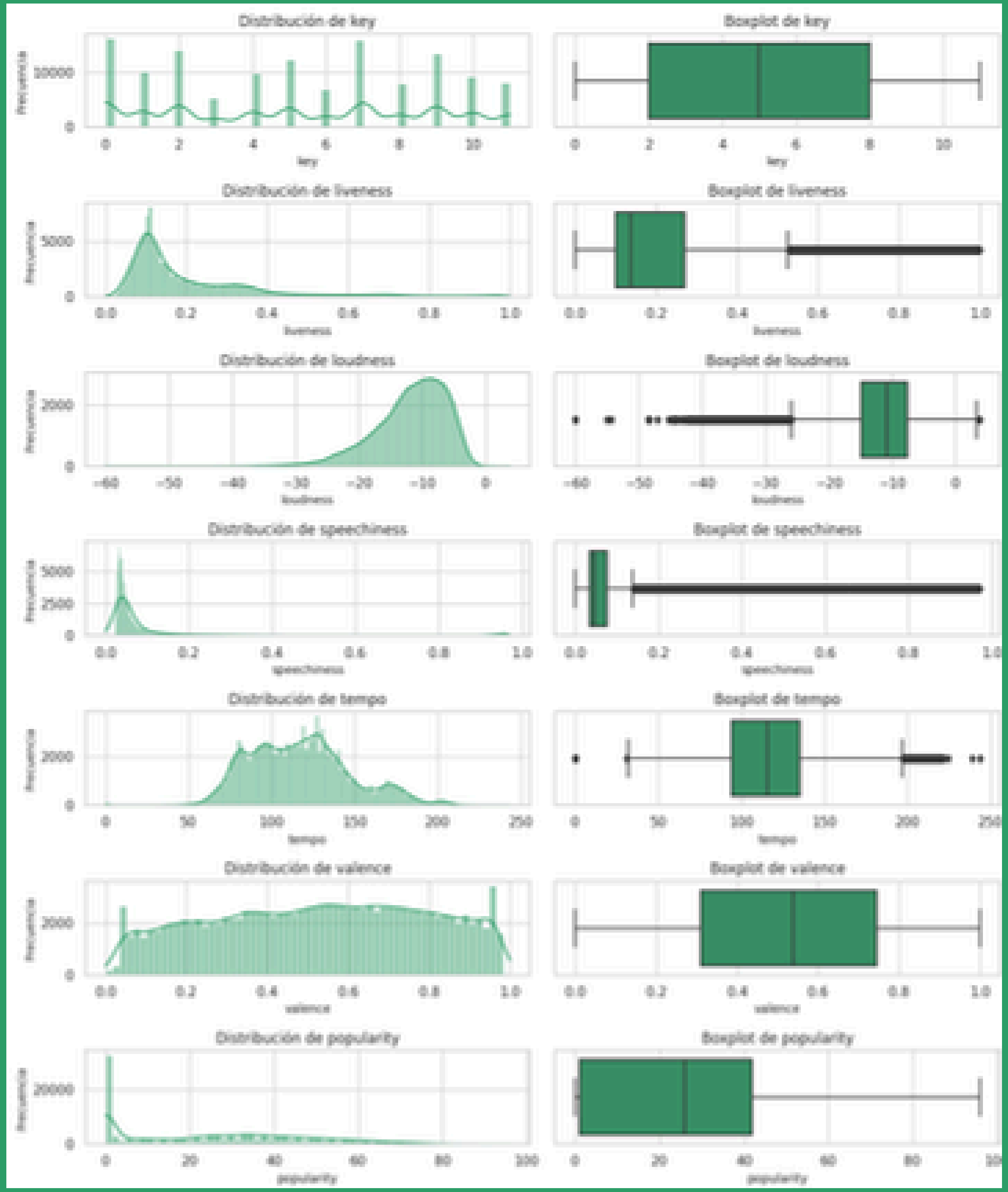
id
artists
name
year
acousticness
danceability
duration_ms
energy
explicit
instrumental
ness
key
liveness
loudness
mode
speechiness
tempo
valence
popularity

TOP 10 ARTISTS WITH THE MOST SONGS

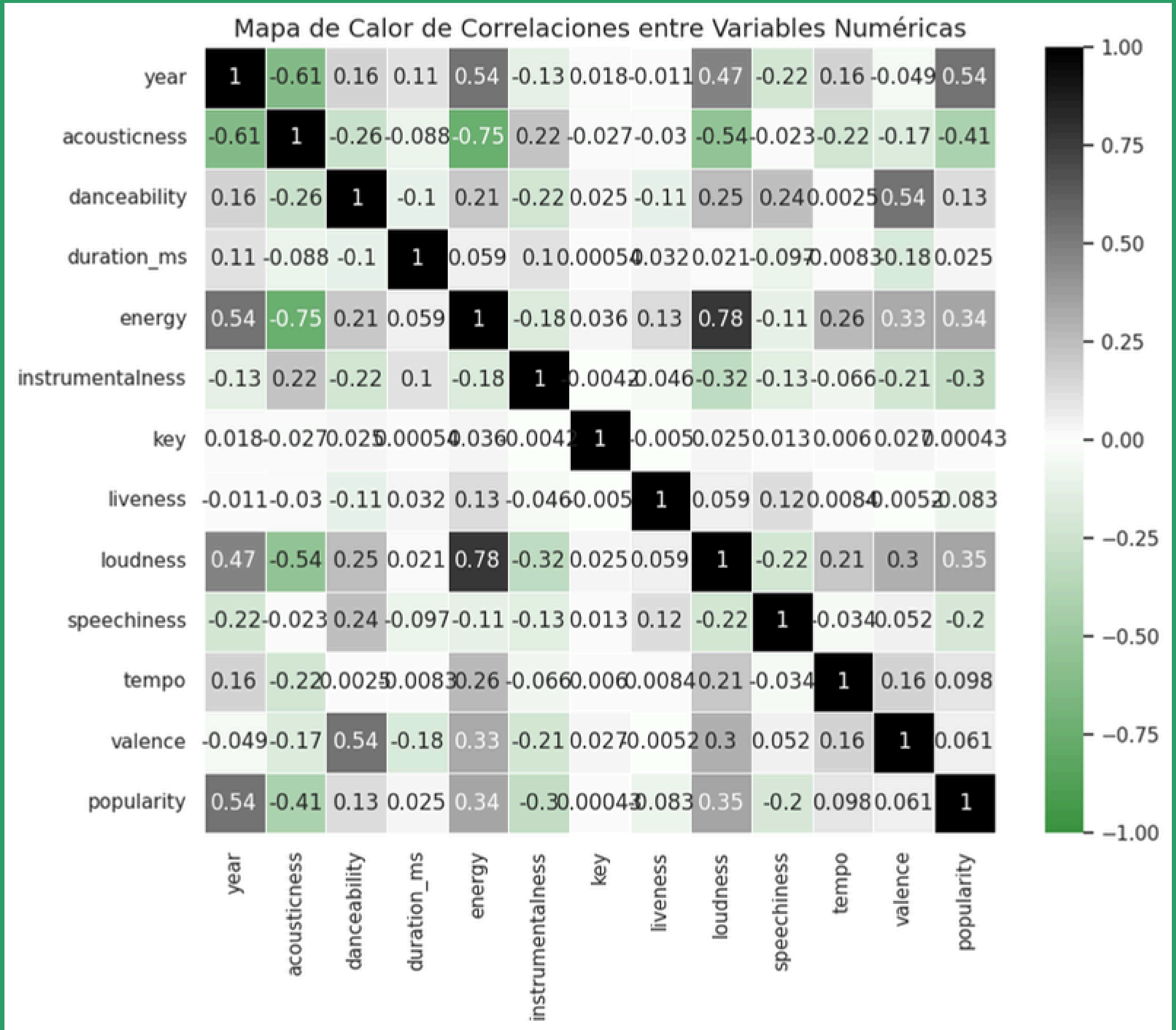


NUMERICAL FEATURE DISTRIBUTIONS





CORRELATION MATRIX



3. ADDED VARIABLES

VARIABLES AGREGADAS

LANGUAGE DETECTION

We used the langdetect library to create a column identifying the language of each song, or marked it as “unknown” when detection was not possible.

ARTIST COUNT

We replaced the original artists column with a new variable indicating the number of artists featured in each song.

LANGUAGE FREQUENCY

The language column was replaced with a frequency-based encoding to represent the information numerically and avoid issues with machine learning models.

4. PIPELINE

PIPELINE

We used a pipeline to chain and automate a sequence of data preprocessing and modeling steps.

WHAT IS THE PIPELINE USED FOR IN THE CODE?

1. Add the number of artists
2. Encode the song's language based on its frequency
3. Fit the regression model
4. StandardScaler()

CLASS 1

```
class FeatureSelectionFrequencyEncoder(BaseEstimator, TransformerMixin):
    def __init__(self, selected_features):
        self.selected_features = selected_features
        self.encoding_dict = defaultdict(int)

    def fit(self, X, y=None):
        for feature in self.selected_features:
                frequencies = X[feature].value_counts().to_dict()
                self.encoding_dict[feature] = frequencies

        return self

    def transform(self, X):
        X_copy = X.copy()
        for feature in self.selected_features:
            if feature in self.encoding_dict:
                X_copy[feature] = X_copy[feature].map(self.encoding_dict[feature])
        return X_copy
```

CLASS 2

```
class AgregarArtistasInvolucrados(BaseEstimator, TransformerMixin):
    def __init__(self, selected_features):
        self.selected_features = selected_features

    def fit(self, X, y=None):

        return self

    def transform(self, X):
        X_copy = X.copy()
        for feature in self.selected_features:
          X_copy[feature] = X_copy[feature].str.count(',') + 1
        return X_copy
```

5. PREDICTIVE MODEL SELECTION

MODEL SELECTION

We ran several models to evaluate which one achieved the highest R^2 and the lowest MSE.

XGBoost

Hist gradian boosting

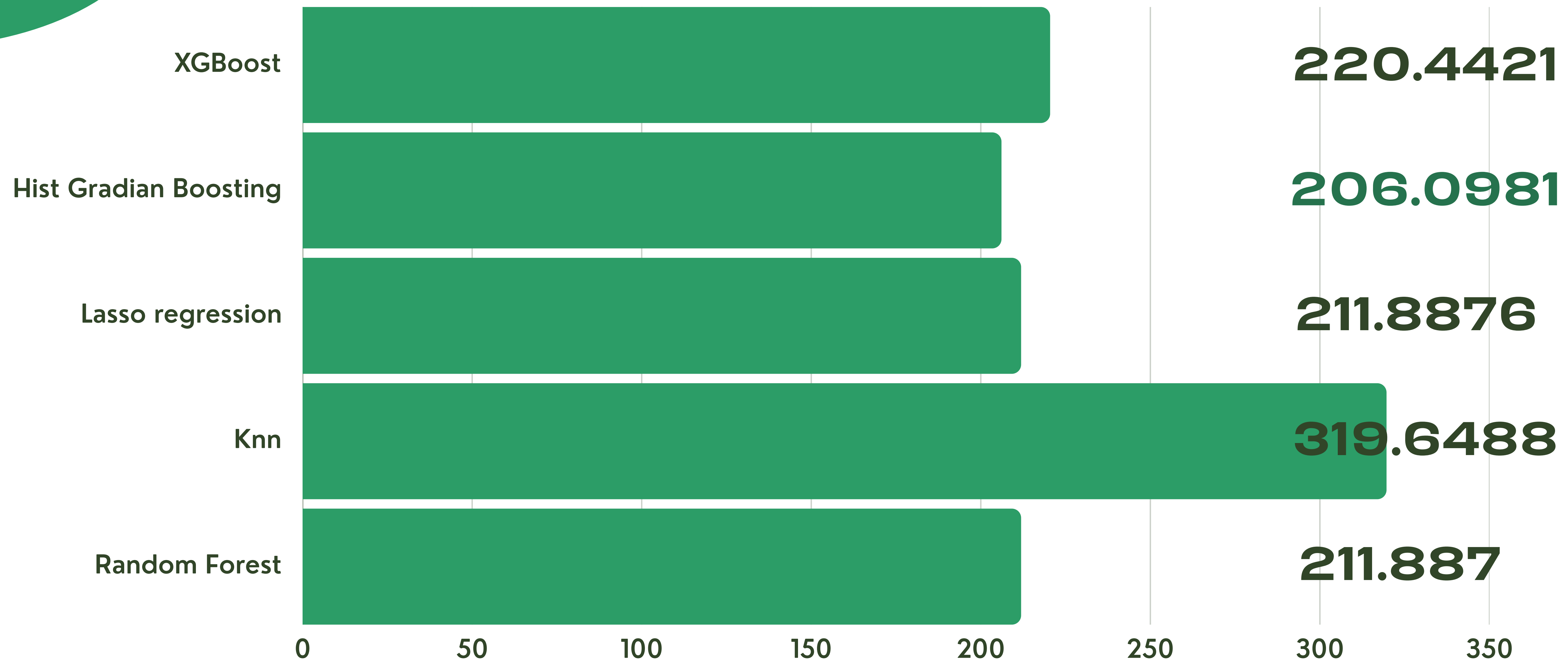
LASSO regression

KNN

Random forest

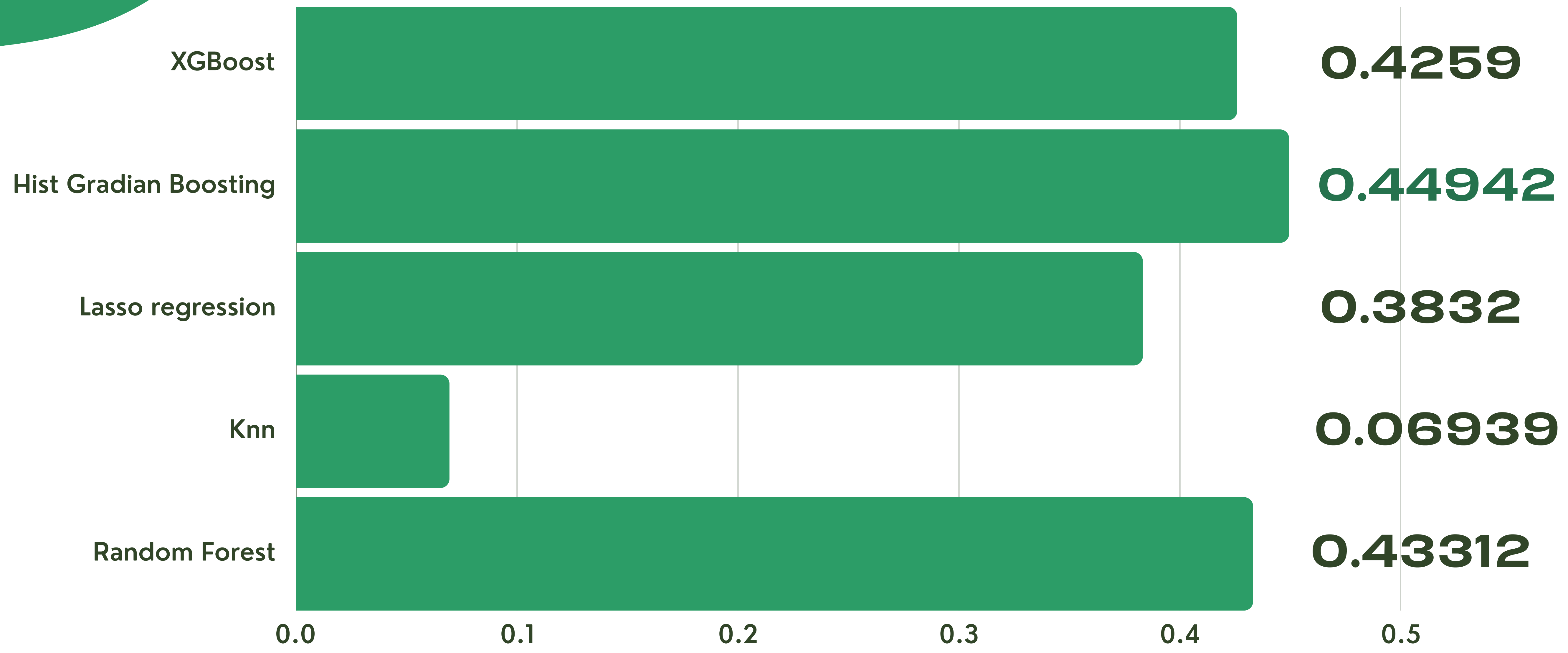
WHICH MODEL ACHIEVED THE LOWEST MEAN SQUARED ERROR (MSE)?

(CROSS VALIDATION)



WHICH MODEL ACHIEVED THE HIGHEST R^2 SCORE?

(CROSS VALIDATION)



Hist Gradian Boosting

HISTOGRAM USAGE

The model relies on the creation and use of histograms to accelerate the training process.

SMALL DECISION TREES

It uses small and simple decision trees, which are less prone to overfitting.

GRADUAL OPTIMIZATION

The model gradually optimizes a set of trees to minimize the loss function. Each new tree is trained to correct the errors made by the previous ones in the ensemble.

ACCURACY AND ROBUSTNESS

It achieves high prediction accuracy. Additionally, it is robust to outliers, making it a solid choice for a wide range of machine learning problems.

6. OPTIMIZACIÓN DE HIPERPARÁMETROS

HYPERPARAMETER OPTIMIZATION

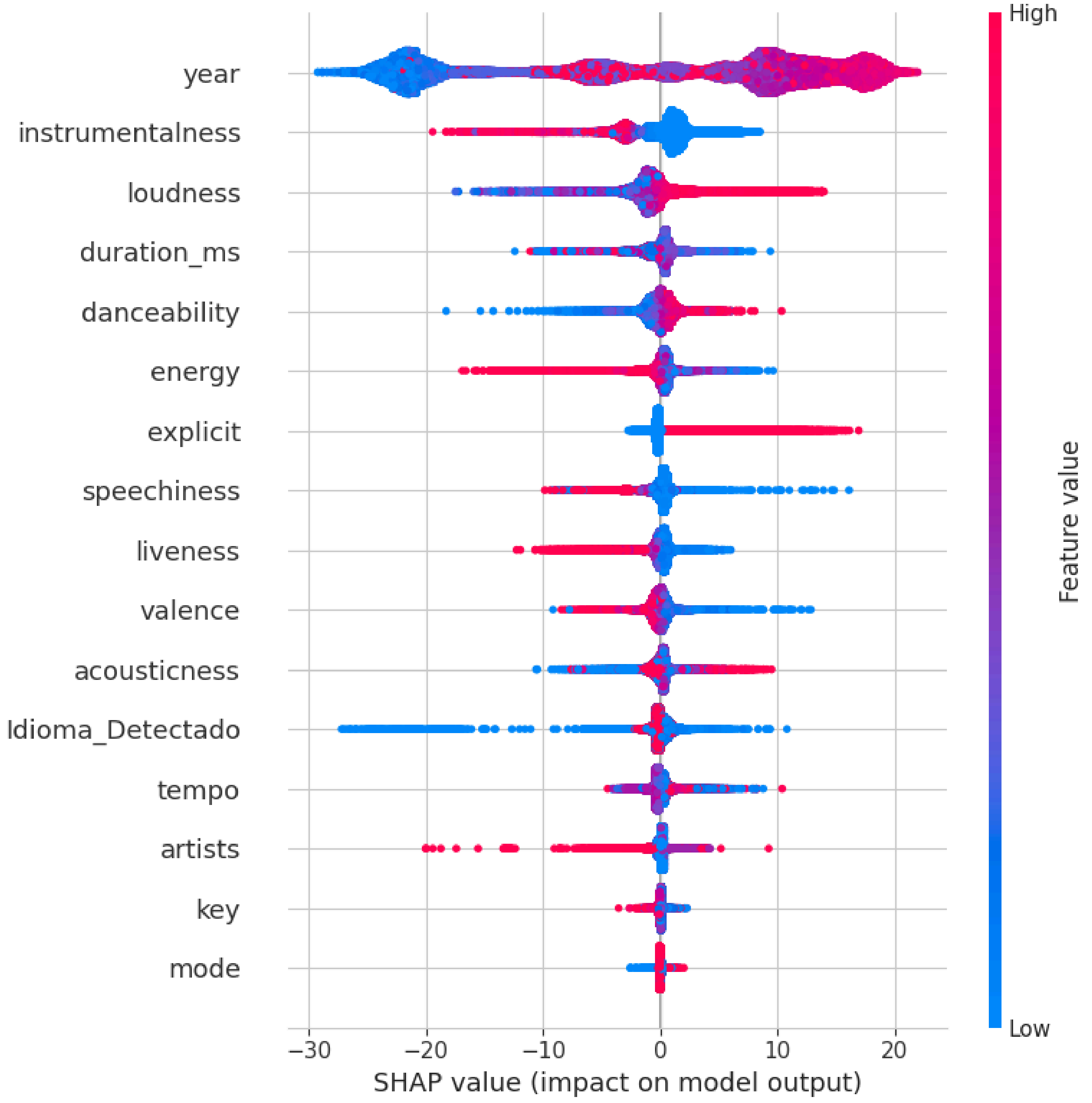
Various hyperparameter optimization techniques were applied, and the table below summarizes the results obtained for each method and hyperparameter configuration.

	GRID SEARCH	HYPEROPT	RANDOM SEARCH	SCIKIT OPTIMIZE
MSE on the test set	151.8576	153.244	154.335	151.830
R^2 on the test set	0.68168	0.67878	0.67649	0.68174

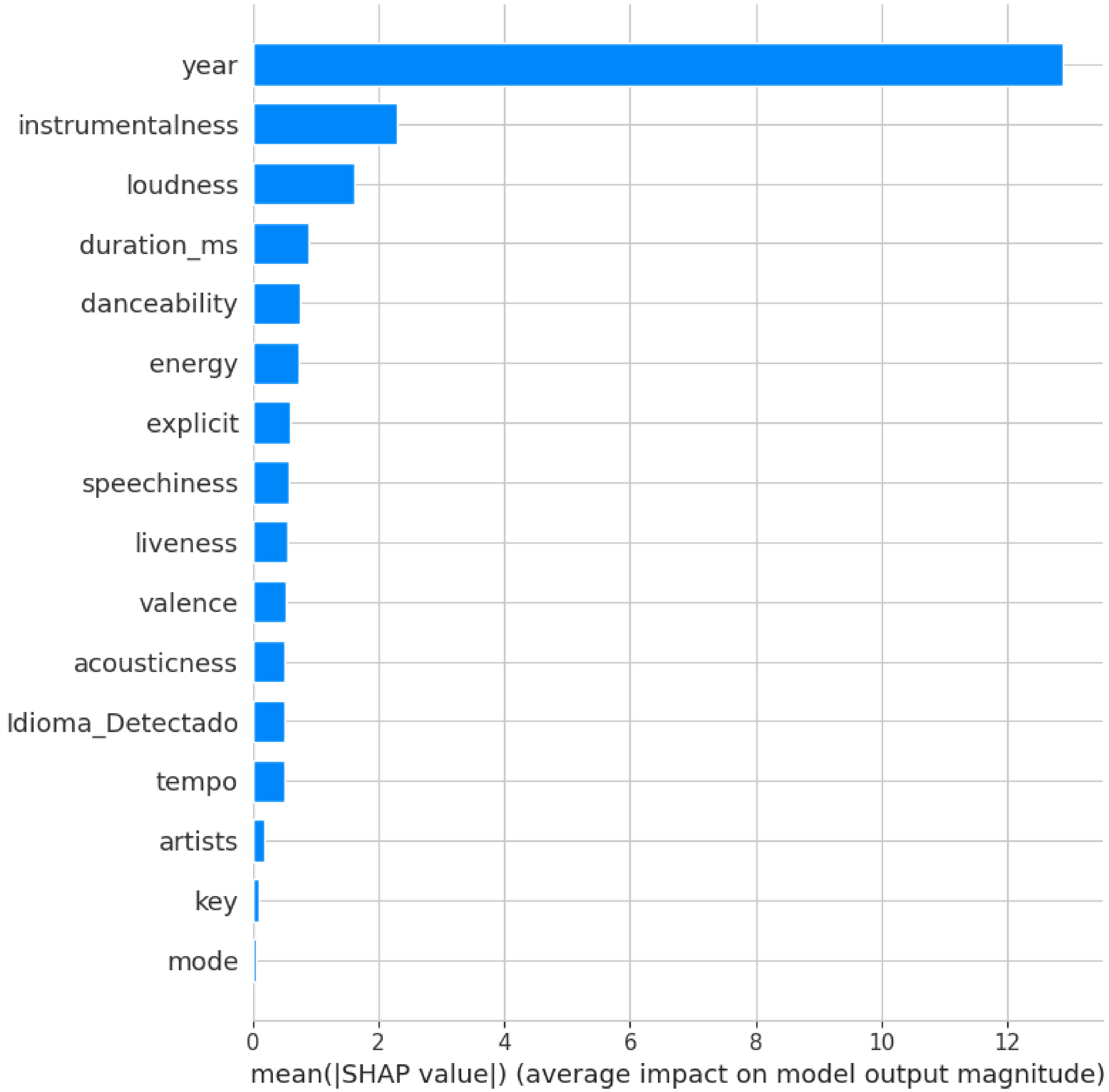
7. SHAP VALUES

SUMMARY PLOT

Summary plot: used to visualize, on an individual level, the importance of each variable for each specific case.

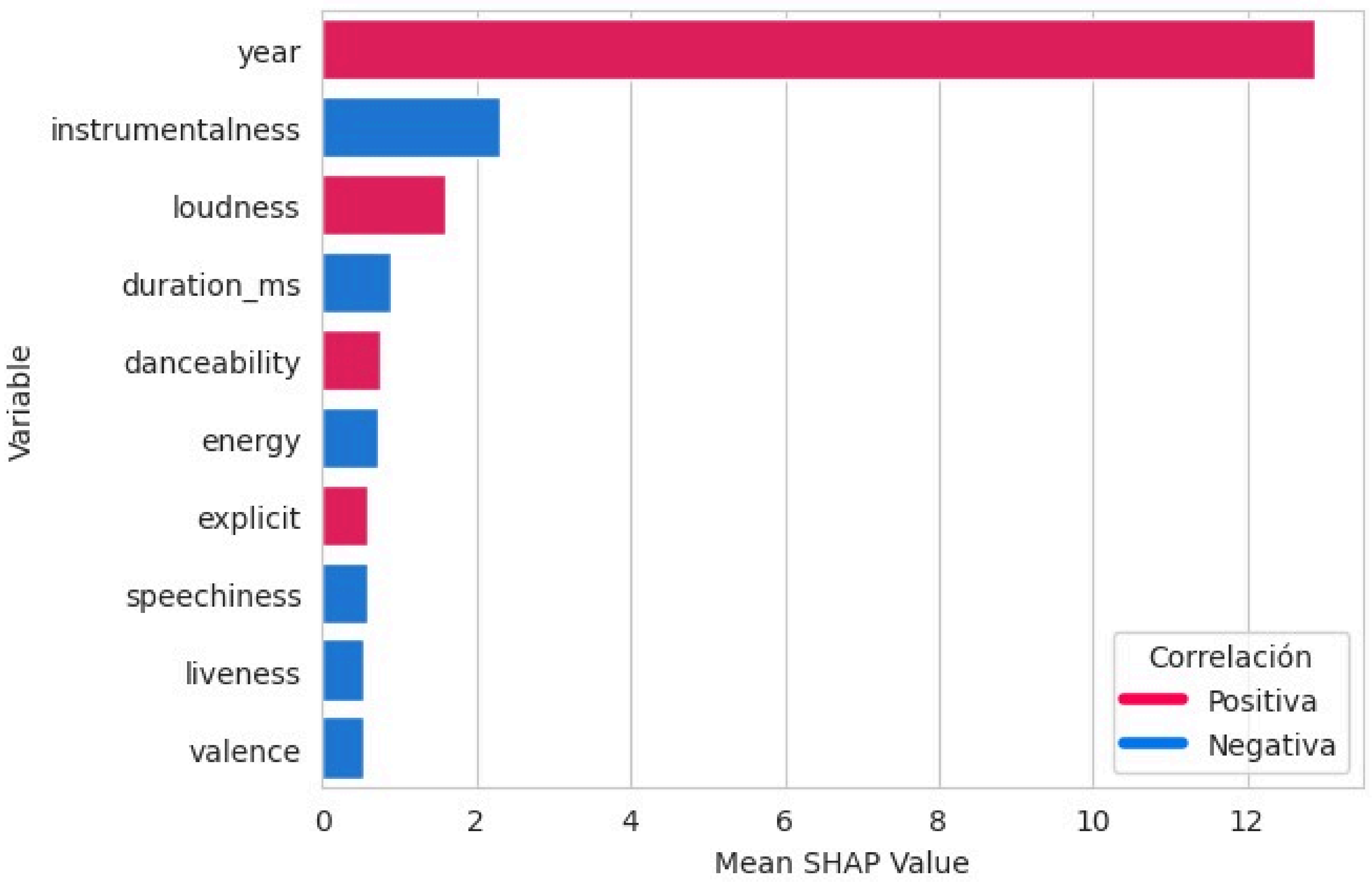


WHAT IS THE AVERAGE CONTRIBUTION OF EACH VARIABLE TO THE MODEL?



WHAT IS THE AVERAGE CONTRIBUTION OF EACH VARIABLE TO THE MODEL?

n this version, we see the same plot but with the direction of the correlation included. While it can be misleading when the relationship is non-linear, it is useful to enrich the interpretation of the previous plot.



8. SUMMARY

SUMMARY

- The dataset was prepared for regression analysis: several columns were added and removed.
- A pipeline was used to recode a column and fit the regression model.
- Multiple regression models were implemented and evaluated using two metrics (MSE and R^2). The best-performing model was HistGradientBoosting (R^2 = 0.44942).
- Different hyperparameter optimization techniques were applied, and the best configuration was achieved using Scikit-Optimize (R^2 = 0.68174).
- SHAP values were analyzed to assess the contribution of each variable to the model's predictions.

Second semester – 2022

TRABAJO PRÁCTICO I
ANÁLISIS PREDICTIVO AVANZADO

Group members

- Magdalena Eppens
- Sofía Gonzalez del Solar
- Nicole Reiman