Eppens, González del Solar, Reiman

SPOTIFY

ITBA - 2do cuatrimestre 2023

CONTENIDO

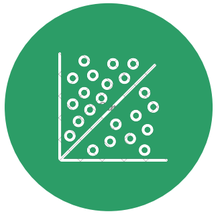
1. SELECCIÓN DEL DATASET

DATASET: POPULARIDAD EN SPOTIFY

Fuente Kaggle

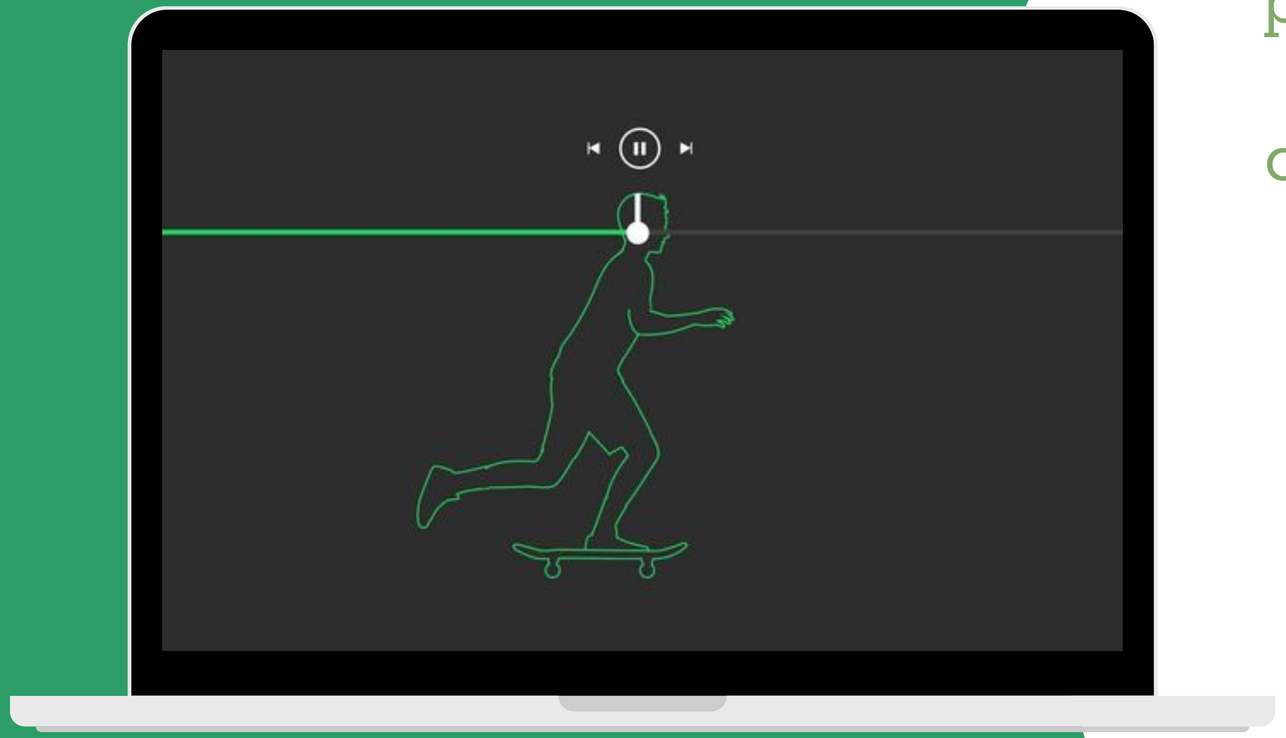
5 columnas sobre la publicación de la canción, el resto sobre sus cualidades musicales

Variable objetivo: popularidad, es una variable numérica que va del 0 al 100



Tipo de modelo requerido: Regresión

MODELO DE NEGOCIO



Identificar con precisión qué canciones tienen el potencial de convertirse en éxitos populares, permitiéndoles invertir recursos en promocionar las canciones correctas y, en última instancia, aumentar sus ingresos y su visibilidad en la industria musical.

2. ANALISIS EXPLORATORIO DE DATOS

ANÁLISIS DESCRIPTIVO

En esta etapa, realizamos el EDA de la base. Miramos missings y outliers, contamos la cantidad de valores únicos en columnas categóricas, y observamos los estadísticos despcriptivos de las columnas numéricas

TAMAÑO DE LA BASE

Filas: 129172
Columnas: 17

COLUMNAS CATEGÓRICAS

['artists', 'name']

COLUMNAS BINARIAS

['explicit', 'mode']

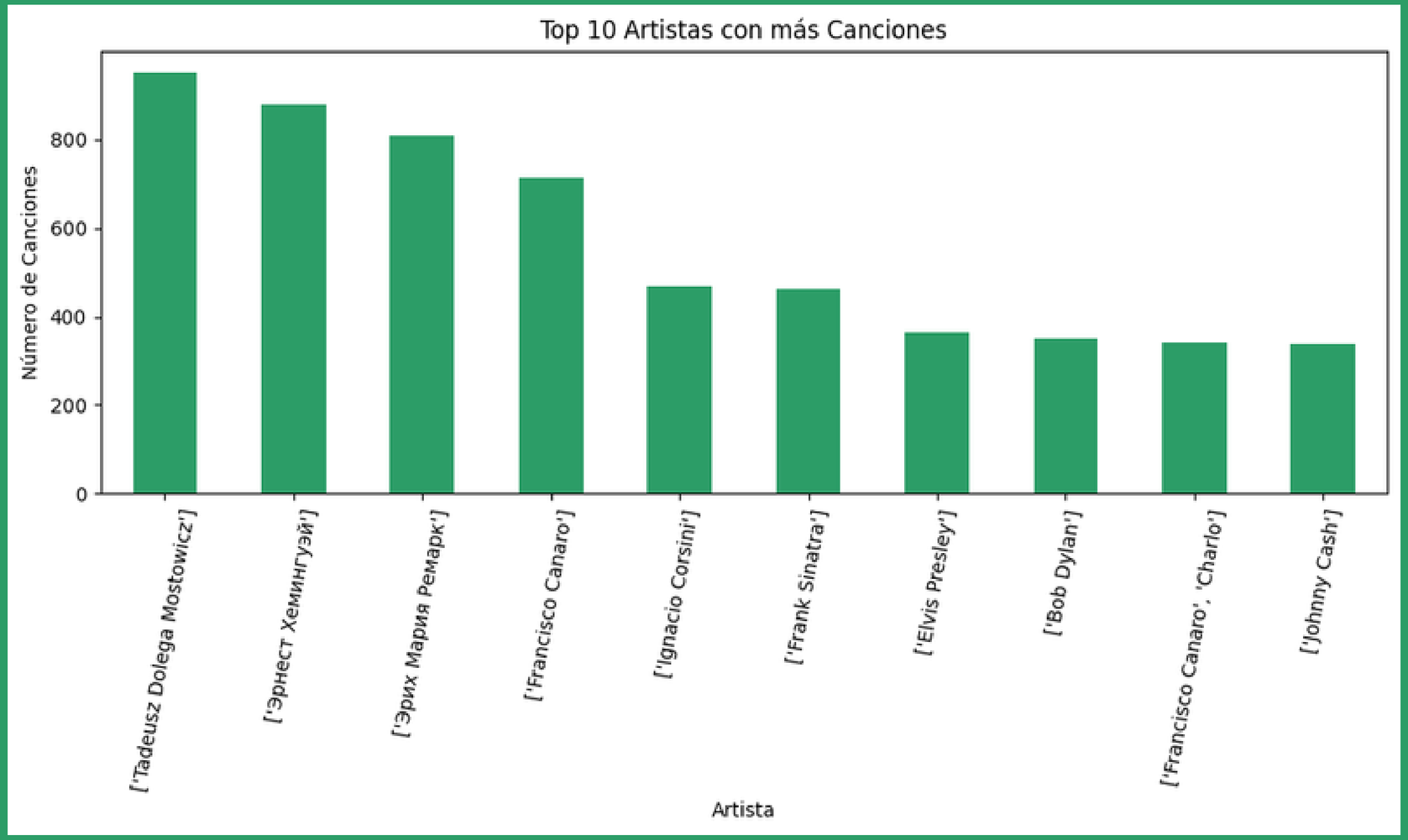
COLUMNAS NUMÉRICAS

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

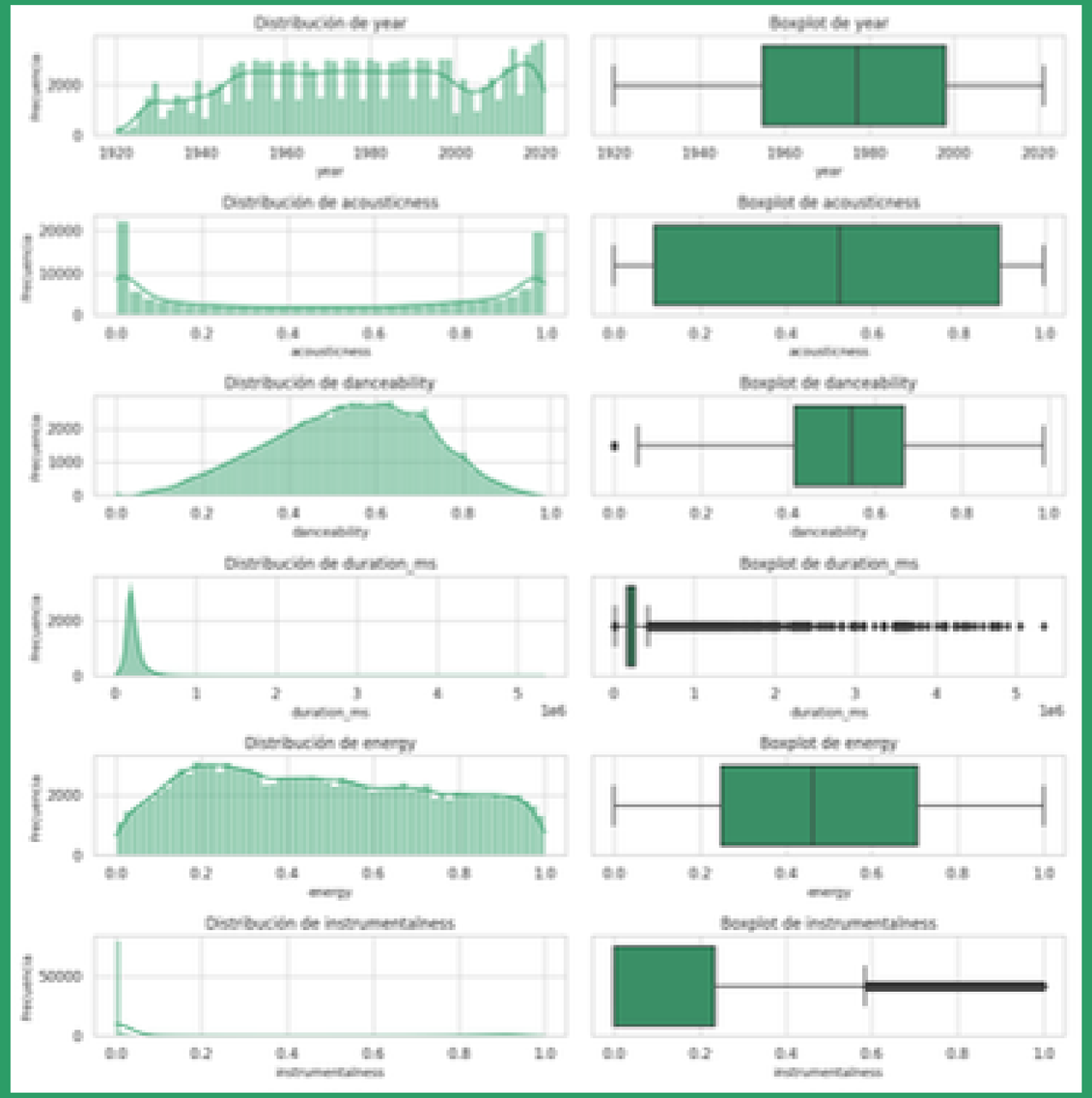
VARIABLES

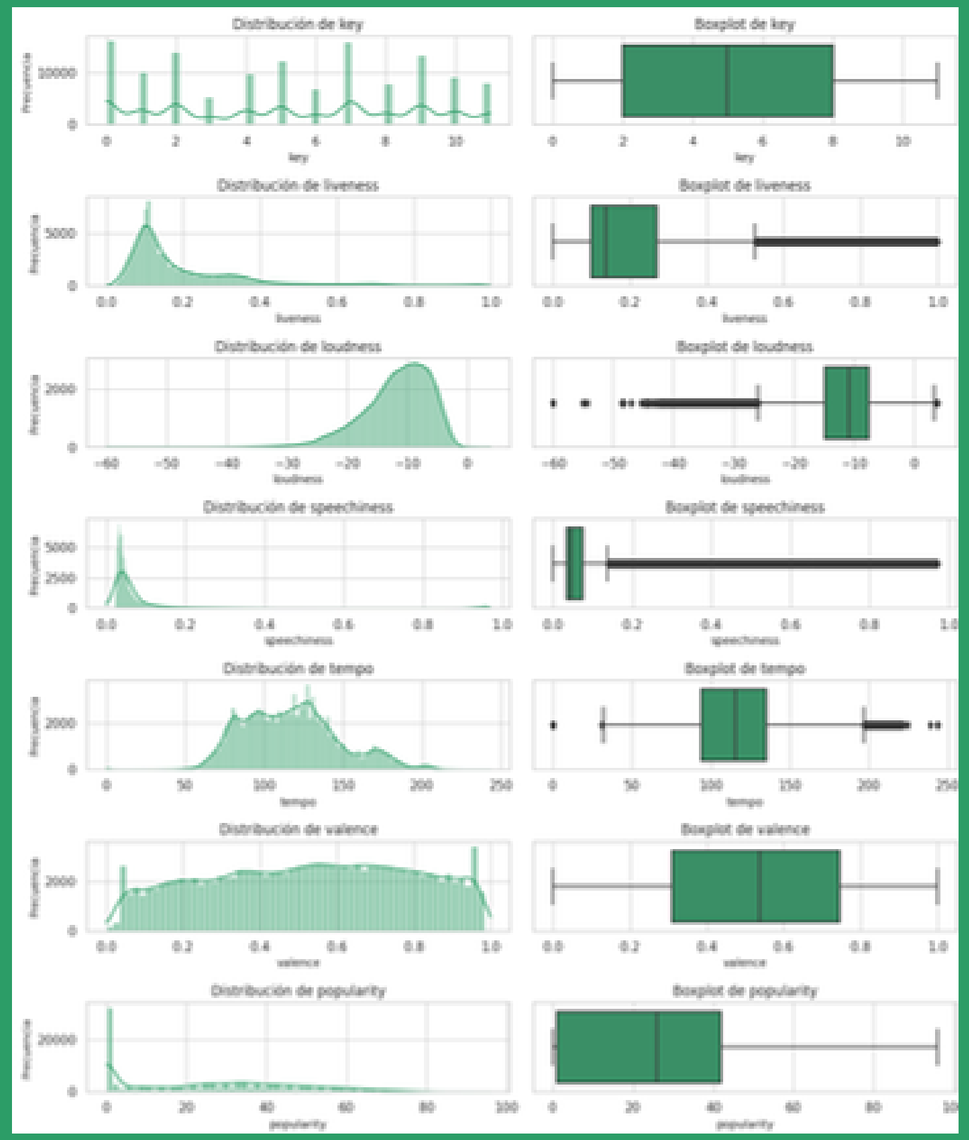
id
artists
name
year
acousticness
danceability
duration_ms
energy
explicit
instrumental
ness
key
liveness
loudness
mode
speechiness
tempo
valence
popularity

TOP 10 ARTISTA CON MÁS CANCIONES

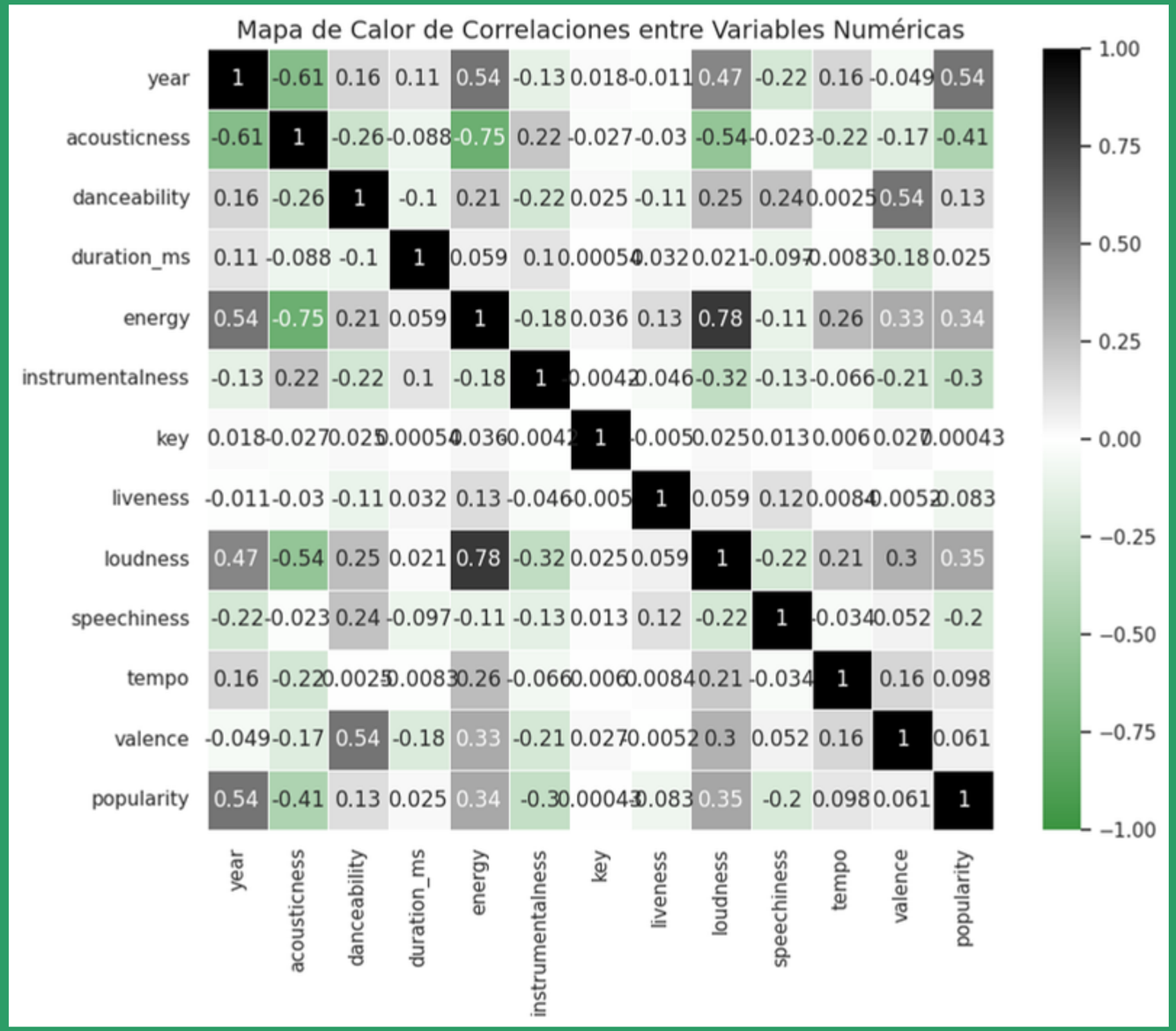


DISTRIBUCIÓN VARIABLES NUMÉRICAS





MATRIZ DE CORRELACIONES



3. VARIABLES AGREGADAS

VARIABLES AGREGADAS

IDIOMA DE LA CANCION

Usamos la librería “langdetect” para completar una columna con el idioma de la canción o desconocido en caso de ser imposible detectarlo.

FRENCUENCIA DE ARTISTAS

Reemplazamos la columna artistas por la cantidad de artistas que participaron en la cancion

FRECUENCIA DE IDIOMA

Reemplazamos la columna idioma por su frecuencia para representar l ainformación de forma numérica y evitar problemas con los modelos

4. PIPELINE

PIPELINE

Usamos un pipeline para encadenar y automatizar una secuencia de pasos del procesamiento de datos y del modelado.

¿CUÁLES SON LOS USOS DEL PIPELINE EN EL CÓDIGO?

1. Agregar la cantidad de artistas
2. Codificar el idioma de la canción según su frecuencia
3. Ajustar el modelo de regresión
4. StandardScaler()

CLASE 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
```

CLASE 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. SELECCIÓN DEL MODELO PREDICTIVO

SELECCIÓN DE MODELOS

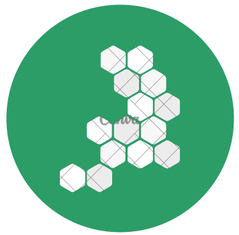
Corrimos diversos modelos
para ver con cual se obtenía
un mayor R2 y un mayor MSE

XGBoost

Hist gradian boosting

LASSO regression

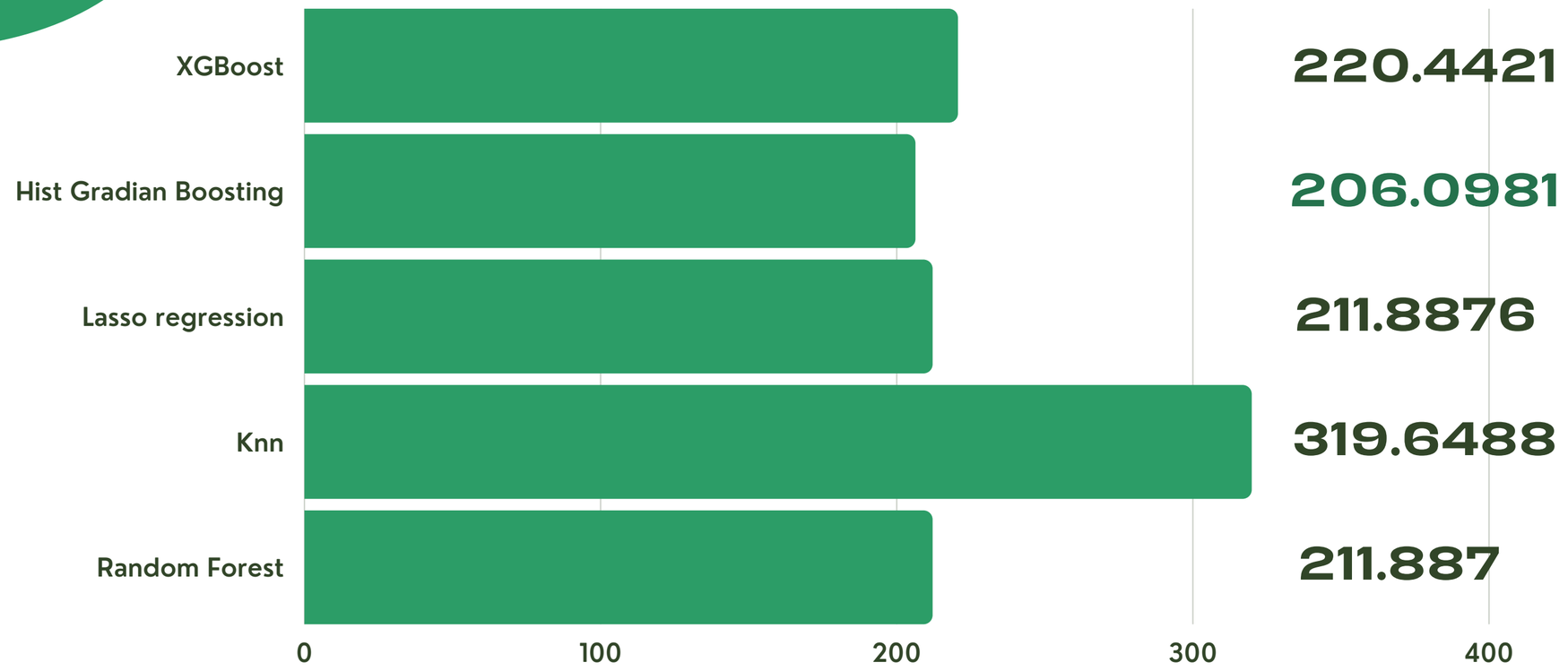
KNN



Random forest

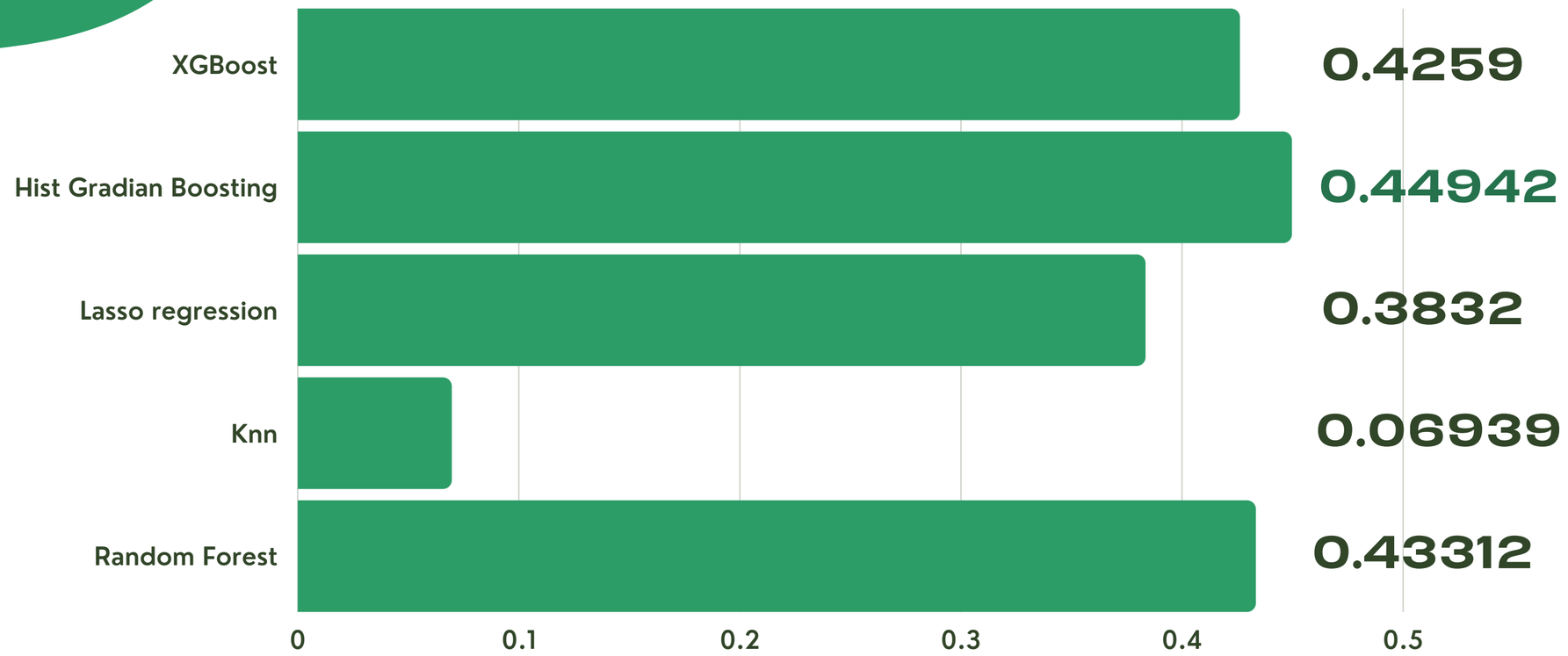
¿CUÁL ES EL MEJOR MODELO SEGÚN EL MSE?

(CROSS VALIDATION)



¿CUÁL ES EL MEJOR MODELO SEGÚN EL R2?

(CROSS VALIDATION)



Hist Gradian Boosting

UTILIZACIÓN DE HISTOGRAMAS

El modelo se basa en la creación y uso de histogramas para acelerar el proceso de entrenamiento.

ÁRBOLES DE DECISIÓN PEQUEÑOS

Usa árboles de decisión pequeños„ más simples y menos propensos al overfitting.

OPTIMIZACIÓN GRADUAL

Optimiza gradualmente un conjunto de árboles para minimizar la función de pérdida. Cada árbol se ajusta para corregir los errores cometidos por los árboles anteriores en el conjunto.

PRECISIÓN Y ROBUSTEZ

Logra una alta precisión en la predicción. Además, es robusto frente a outliers, lo que lo convierte en una elección sólida para una variedad de problemas de aprendizaje automático.

6. OPTIMIZACIÓN DE HIPERPARÁMETROS

OPTIMIZACIÓN DE HIPERPARÁMETROS

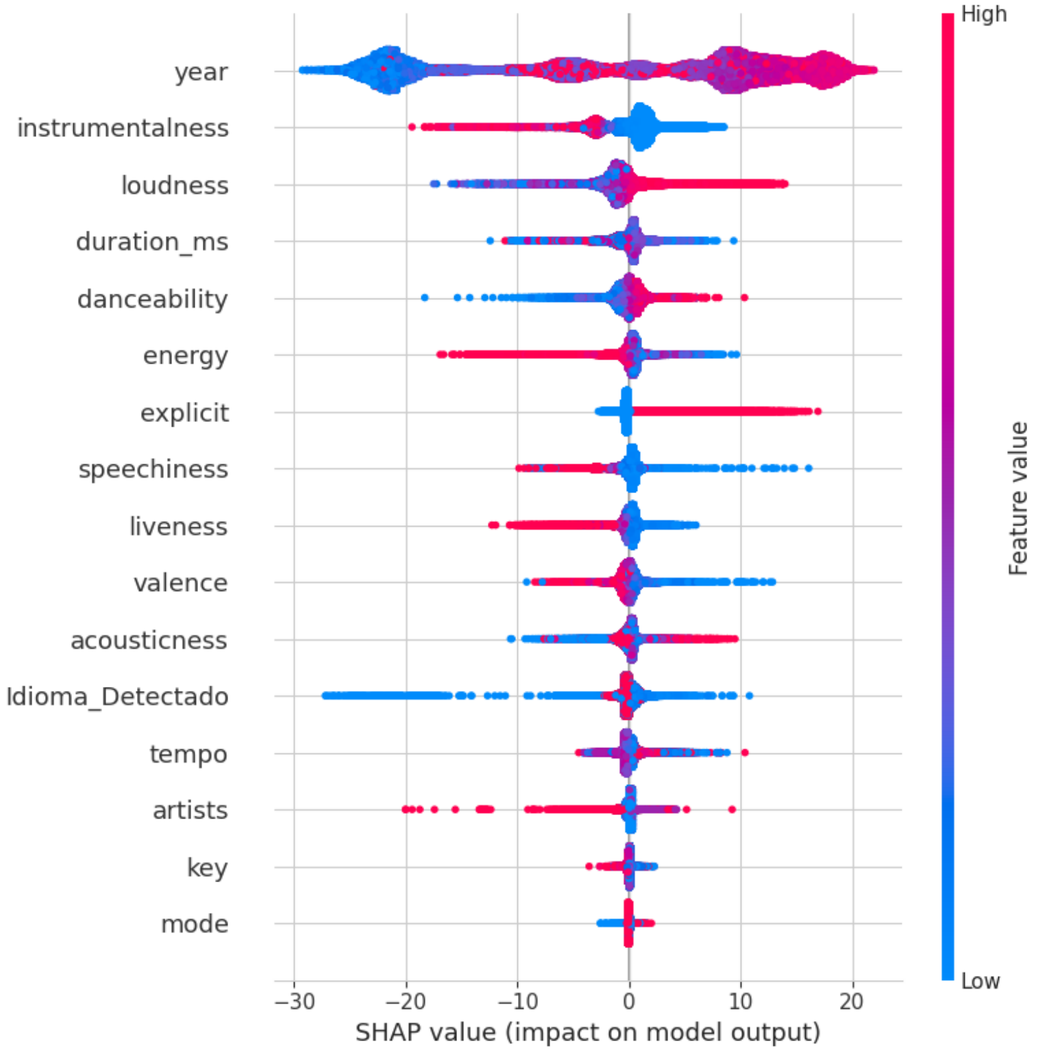
Usamos distintas técnicas de optimización de hiperparámetros y estos fueron los resultado que obtuvimos, para ccada hiperparámetro, según cada una de lás técnicas:

	GRID SEARCH	HYPEROPT	RANDOM SEARCH	SCIKIT OPTIMIZE
MSE en el conjunto de prueba	151.8576	153.244	154.335	151.830
R2 en el conjunto de prueba	0.68168	0.67878	0.67649	0.68174

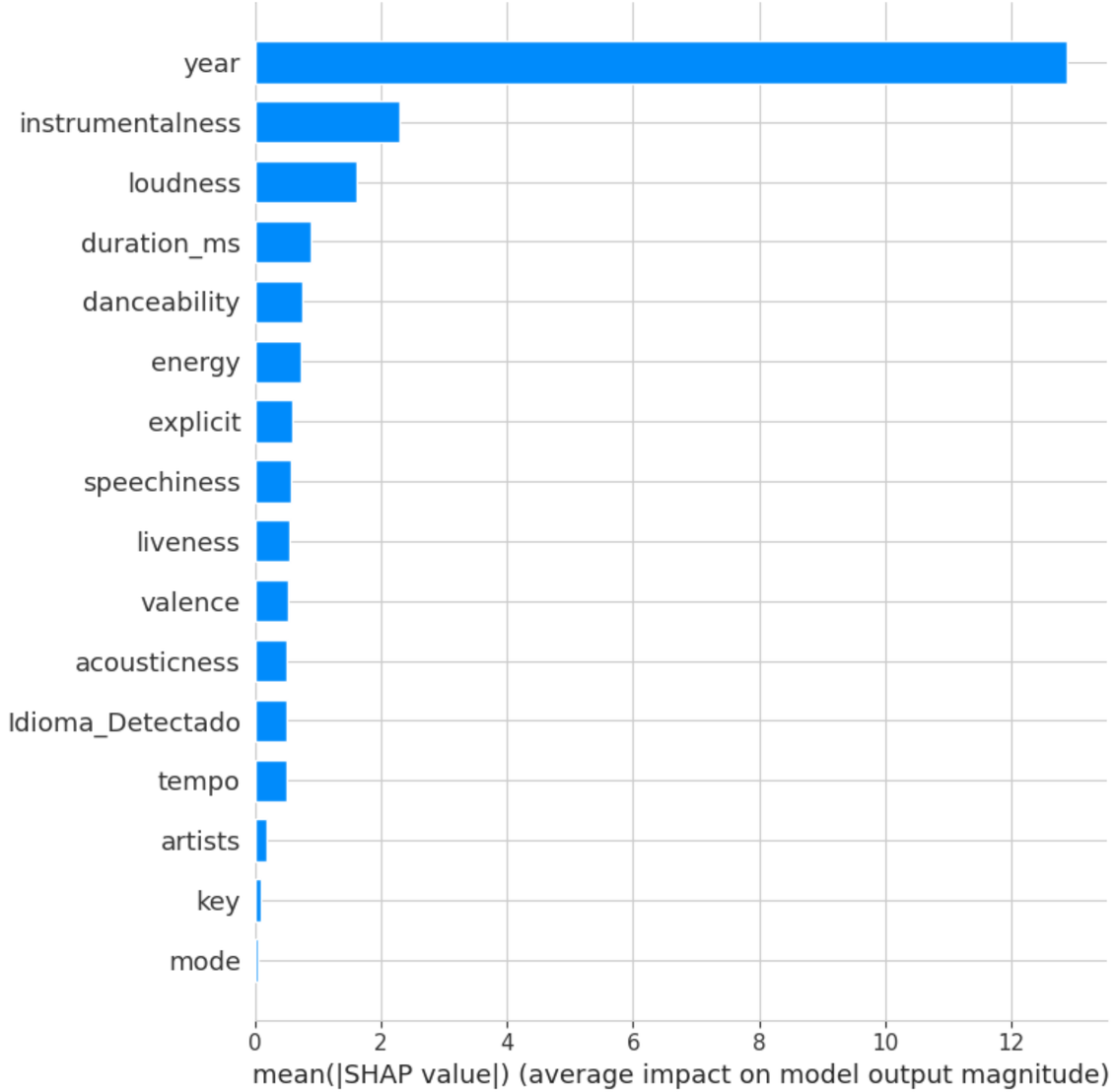
7. SHAP VALUES

SUMMARY PLOT

Summary plot: sirve para ver de forma desagregada, la importancia de cada variable para *cada caso en particular.*

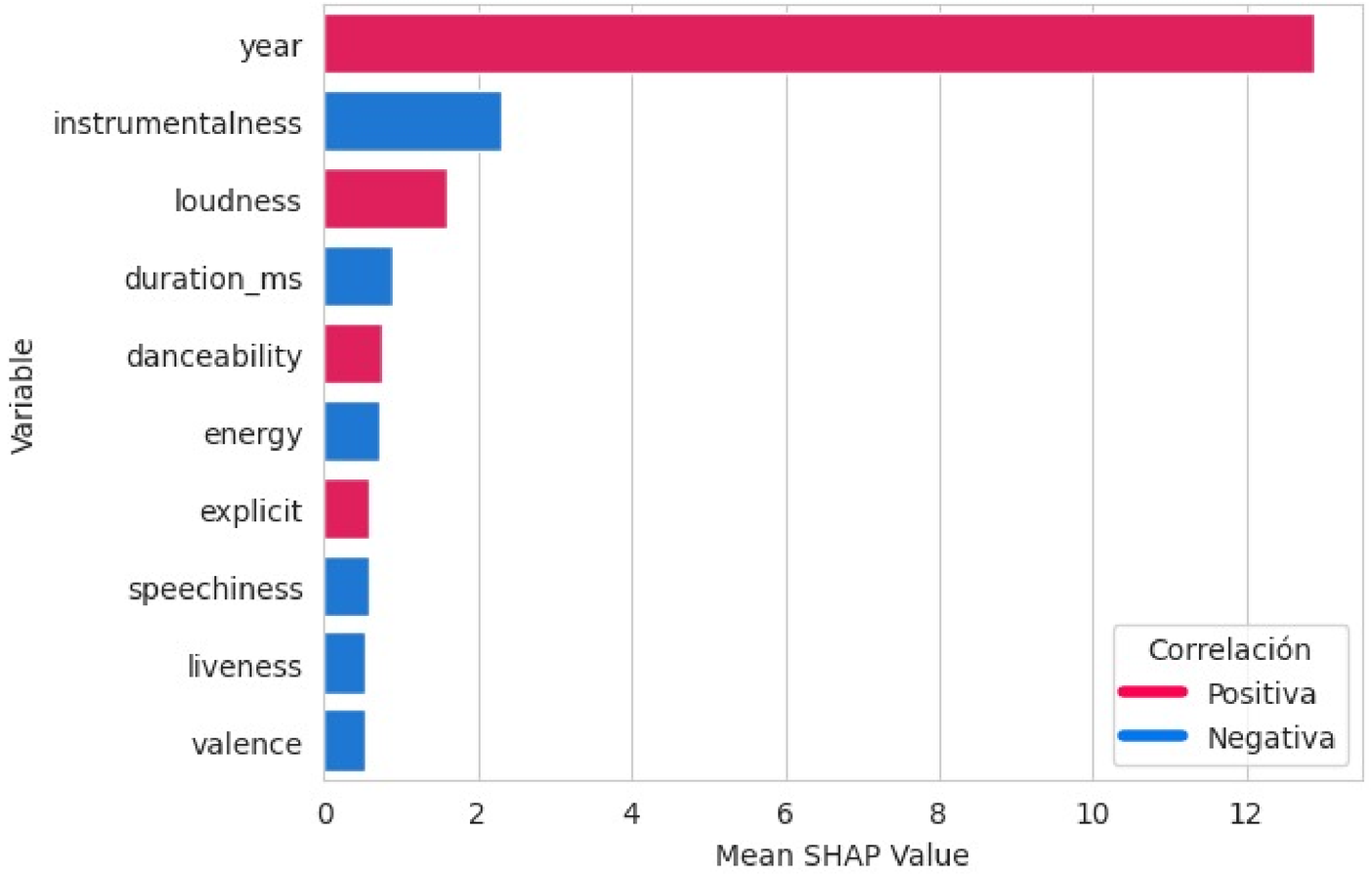


¿CUÁL ES EL APORTE MEDIO DE CADA VARIABLE AL MODELO?



¿CUÁL ES EL APORTE MEDIO DE CADA VARIABLE AL MODELO?

En esta variante, vemos el mismo gráfico pero incorporando **el sentido de la correlación**. Puede ser engañoso cuando la relación no es lineal, pero es útil si queremos enriquecer el gráfico anterior



8. CONCLUSIONES

CONCLUSIÓN

- Se preparó el dataset para hacer la regesión: Se agregaron y se eliminaron columnas.
- Se usó una pipeline para recodificar una columna y ajustar el modelo de regresión.
- Se implementó distintos modelos de regresion y se comparó su resultado según dos métricas (mse y r2) y se determinó que el mejor modelo es el Hist Gradian Boosting (r2 = 0.44942)
- Se implementaron distintas técnicas para optimizar hiperparámetros y se eligió la mejor combinación hallada (con con Scikit Optimize r2 = 0.68174)
- Se realizó el análisis de SHAP values para relevar la importancia de cada variable en el modelo.

ITBA

2do cuatrimestre - 2022

TRABAJO PRÁCTICO I
ANÁLISIS PREDICTIVO AVANZADO

Integrantes

- Magdalena Eppens
- Sofía Gonzalez del Solar
- Nicole Reiman