BIP MUSIC CASO

MATEO MARTINEZ PEÑALOSA

RECOMENDACIONES PARA BIP MUSIC

1. Enfocar a los artistas para que sus canciones tengan los valores adecuados en las “Features” más importantes.

2. Reforzar la búsqueda de talentos utilizando el modelo predictivo y el análisis de datos para identificar artistas emergentes con un alto potencial de éxito en discotecas y redes sociales.

3. Encontrar más características asociadas al éxito de las canciones y a su viralizacion en redes sociales.

RECOMENDACIONES PARA BIP MUSIC

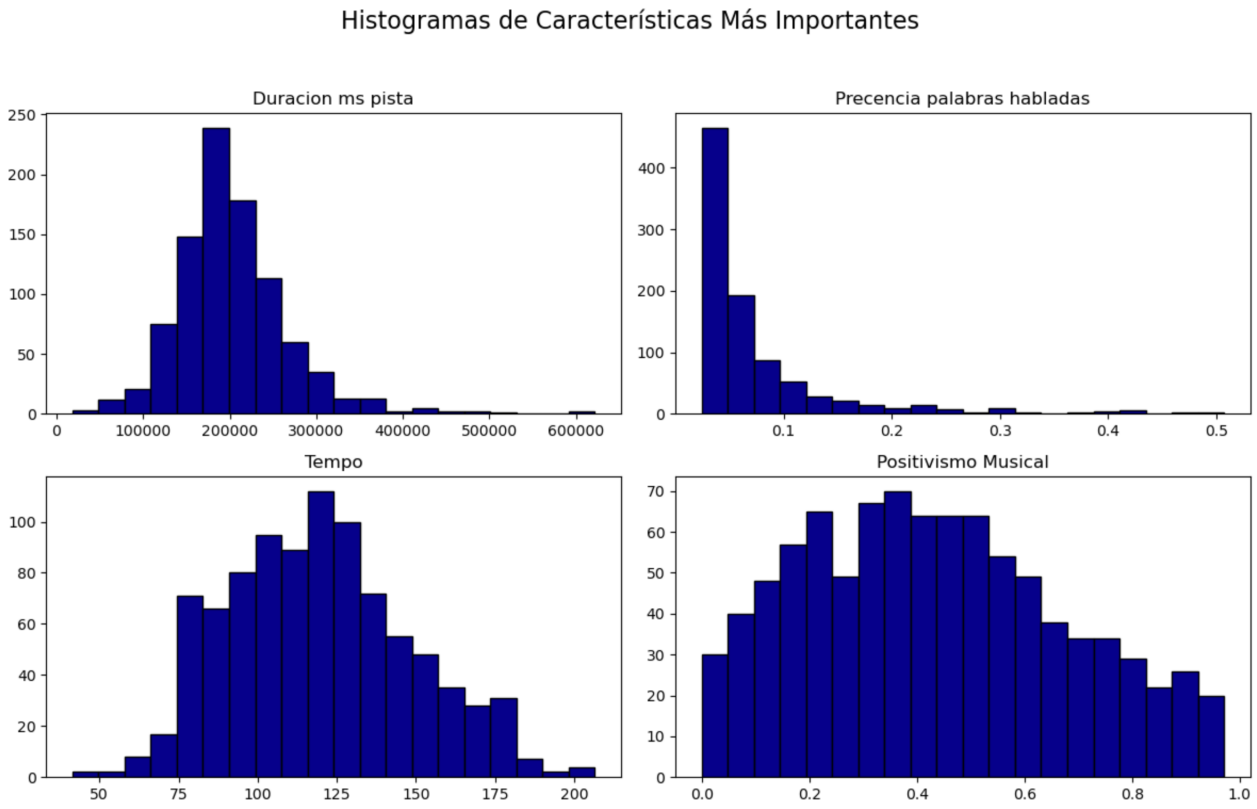
1. ENFOCAR A LOS ARTISTAS PARA QUE SUS CANCIONES TENGAN LOS VALORES ADECUADOS EN LAS "FEATURES" MÁS IMPORTANTES

Gráfica

La gráfica muestra un histograma de los valores de las *Features* que predijeron un éxito. Estas *Features* son las más importantes para el modelo, ya que ayudan a que este haga predicciones más acertadas.

Canción ideal

La canción ideal debería intentar ubicar sus *Features* en el valor de la media de los datos que predijeron el éxito de una canción. Adicional, estos histogramas sirven de referencia para entender que valores deben tener sus métricas.



Histograma de las Features más importantes

RECOMENDACIONES PARA BIP MUSIC

2. REFORZAR LA BÚSQUEDA DE TALENTOS UTILIZANDO EL MODELO PREDICTIVO Y EL ANÁLISIS DE DATOS PARA IDENTIFICAR ARTISTAS EMERGENTES CON UN ALTO POTENCIAL DE ÉXITO EN DISCOTECAS Y REDES SOCIALES.

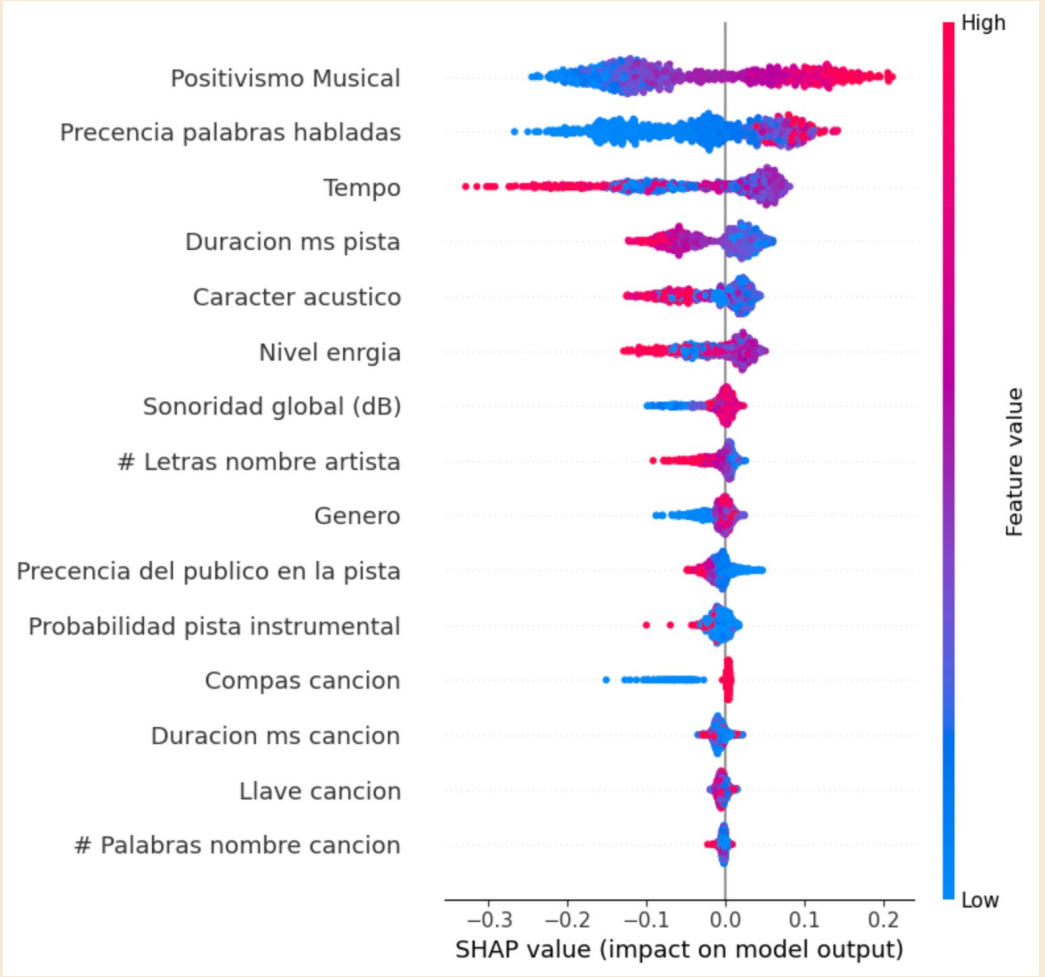
Análisis de sensibilidad

La gráfica ayuda a evaluar el comportamiento de cada "Feature" y como afecta al modelo. Los *shap values* explican las predicciones del modelo a través de las contribuciones de cada variable. Si el valor de la variable aumenta o disminuye la probabilidad de que la variable objetivo sea clasificada positivamente, también varia. *Feature Value* expresa la probabilidad de que la canción, es decir la variable objetivo, sea exitosa.

Variables:

- Un tempo bajo indica una alta probabilidad de ser un éxito.

- Una duración baja de la pista indica una alta probabilidad de éxito, mientras que una alta lleva a que no lo sea.



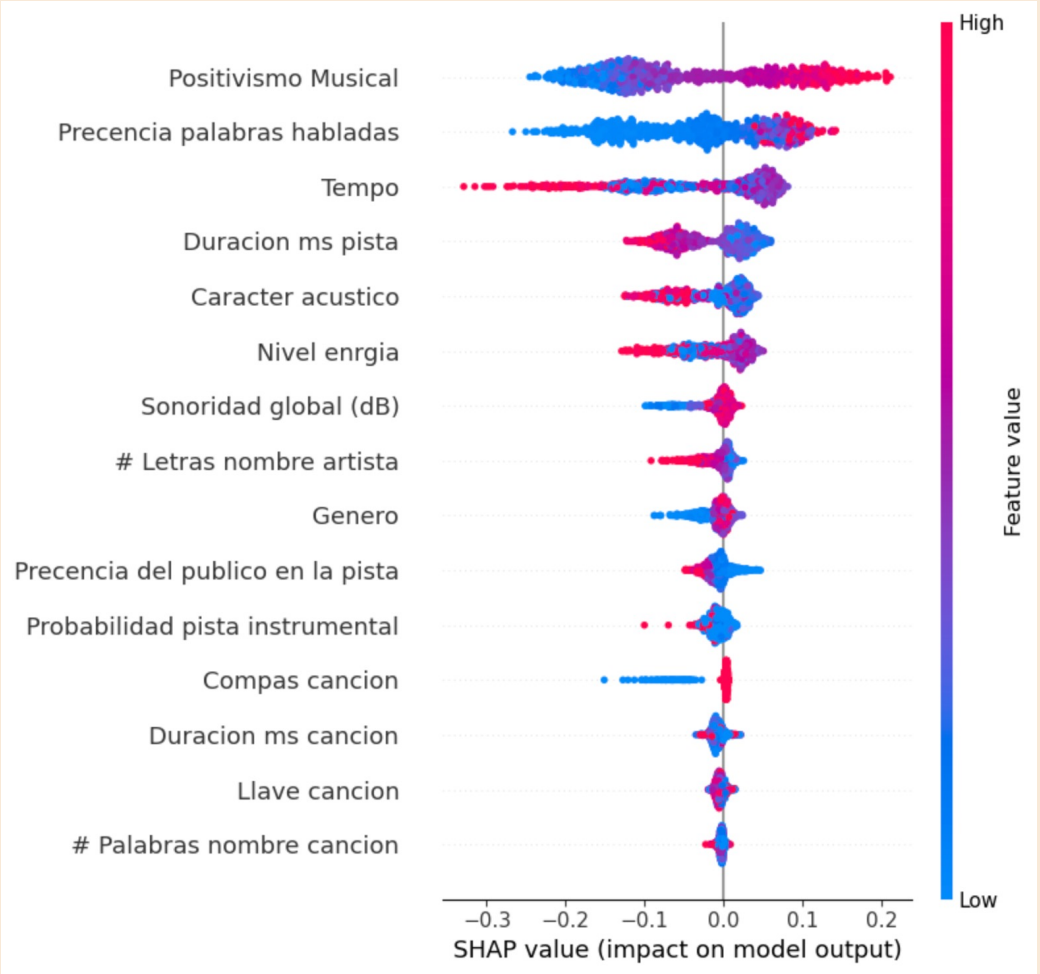
Shap Values vs Feature Value

RECOMENDACIONES PARA BIP MUSIC

2. REFORZAR LA BÚSQUEDA DE TALENTOS UTILIZANDO EL MODELO PREDICTIVO Y EL ANÁLISIS DE DATOS PARA IDENTIFICAR ARTISTAS EMERGENTES CON UN ALTO POTENCIAL DE ÉXITO EN DISCOTECAS Y REDES SOCIALES.

Análisis de sensibilidad

- Un carácter acústico bajo indica una alta probabilidad de ser un éxito.
- Un energía baja de la canción indica una alta probabilidad de ser un éxito.
- Un sonoridad global en decibeles baja indica una baja probabilidad de que la canción sea un éxito.
- Si el nombre del artista es corto, la canción tiene una alta probabilidad de ser un éxito.
- Una alta presencia del público en directo en la pista indica .una baja probabilidad de que sea un éxito



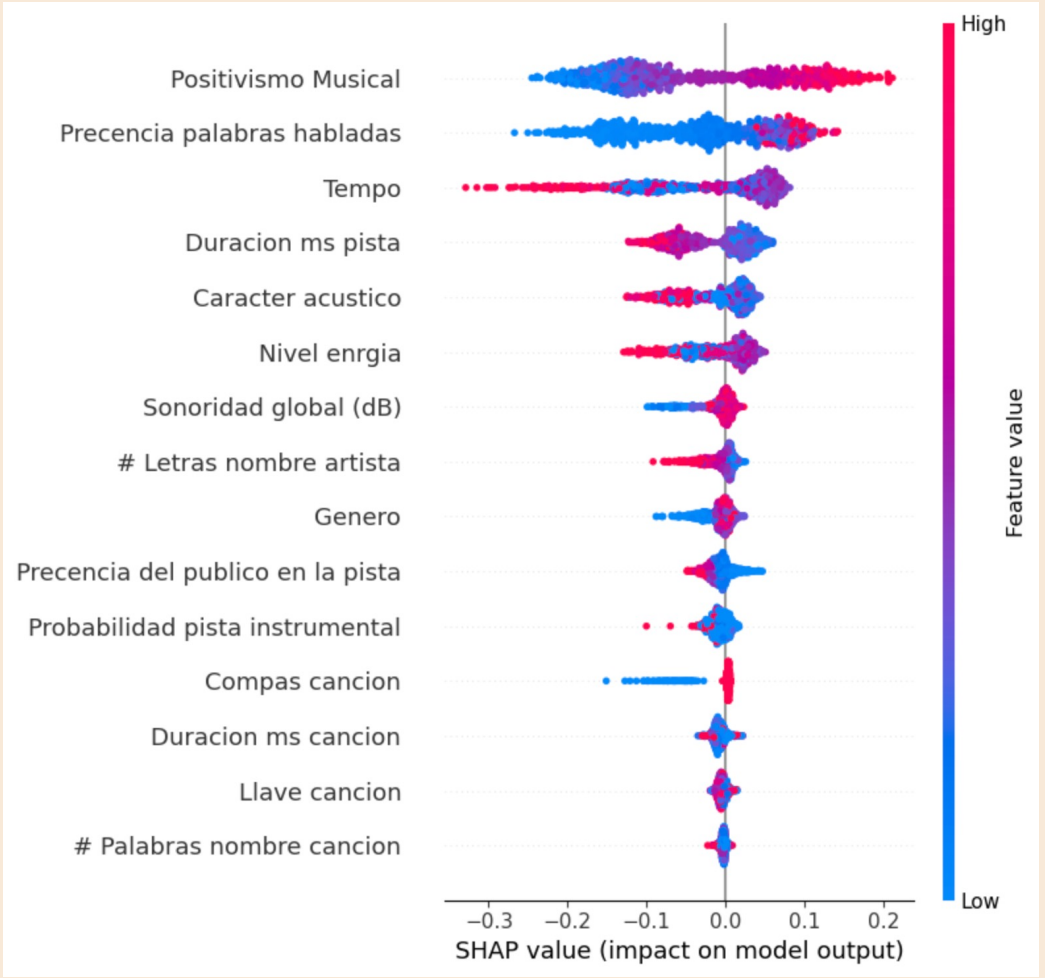
Shap Values vs Feature Value

RECOMENDACIONES PARA BIP MUSIC

2. REFORZAR LA BÚSQUEDA DE TALENTOS UTILIZANDO EL MODELO PREDICTIVO Y EL ANÁLISIS DE DATOS PARA IDENTIFICAR ARTISTAS EMERGENTES CON UN ALTO POTENCIAL DE ÉXITO EN DISCOTECAS Y REDES SOCIALES.

Análisis de sensibilidad

- Una baja probabilidad de que la pista sea instrumental aumenta la probabilidad que la canción sea exitosa.
- Un bajo compás de la canción indica que la canción no será un éxito con una alta probabilidad.
- Un alto positivismo musical indica una alta probabilidad de ser un éxito en redes sociales. Una baja valencia disminuye la probabilidad.
- Una baja presencia de palabras habladas en la canción lleva a que la canción aumente la probabilidad de que no sea un éxito.



Shap Values vs Feature Value

RECOMENDACIONES PARA BIP MUSIC

3. ENCONTRAR MÁS CARACTERÍSTICAS ASOCIADAS AL ÉXITO DE LAS CANCIONES Y A SU VIRALIZACIÓN EN REDES SOCIALES.

- Las variables proporcionadas son muy básicas y existe mucha información adicional que puede ayudar a predecir una viralización en redes sociales.

- Invertir en publicidad dentro de la red social. Según Contraband, el 2.5% de los videos se vuelven virales gracias a la publicidad. Esto se traduce en 8.9 millones de "streams" al año.

- 9.1% de los videos se vuelven virales gracias a los "Influencers". Contratarlos equivale a 11.8 millones de "streams" al año, según Contraband.

1. PREGUNTAS DE NEGOCIO

2. CONSTRUCCIÓN DEL MODELO

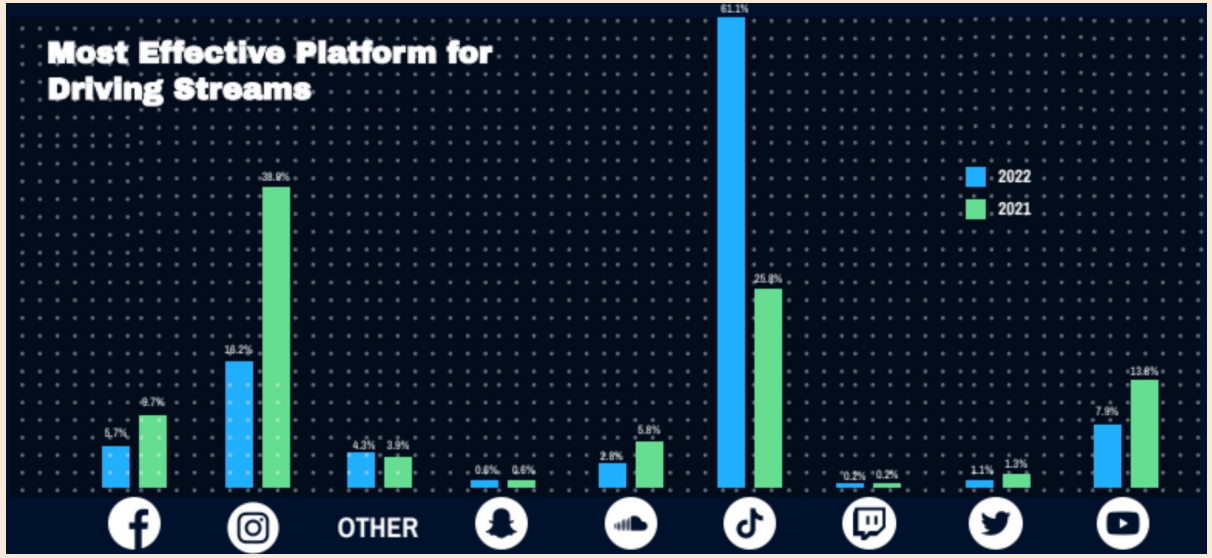
PREGUNTAS DE NEGOCIO

1. ¿QUÉ SUPUESTOS REALIZÓ PARA SABER SI UNA CANCIÓN SERÁ UN ÉXITO O NO EN LAS DISCOTECAS Y REDES SOCIALES?

SUPUESTOS

1. Los éxitos en redes sociales se traducen en un mayor número de "streams" en la plataforma BIP Music.

Justificación:
- Según un estudio realizado por Contraband en el 2022, con el fin de encontrar como los artistas se volvían virales en Tiktok, los éxitos en redes sociales se convierten en una mayor cantidad de "streams" para plataformas como BIP Music.



Plataformas mas efectivas para aumentar el número de "Streams" en Spotify, Apple Music, etc.

PREGUNTAS DE NEGOCIO

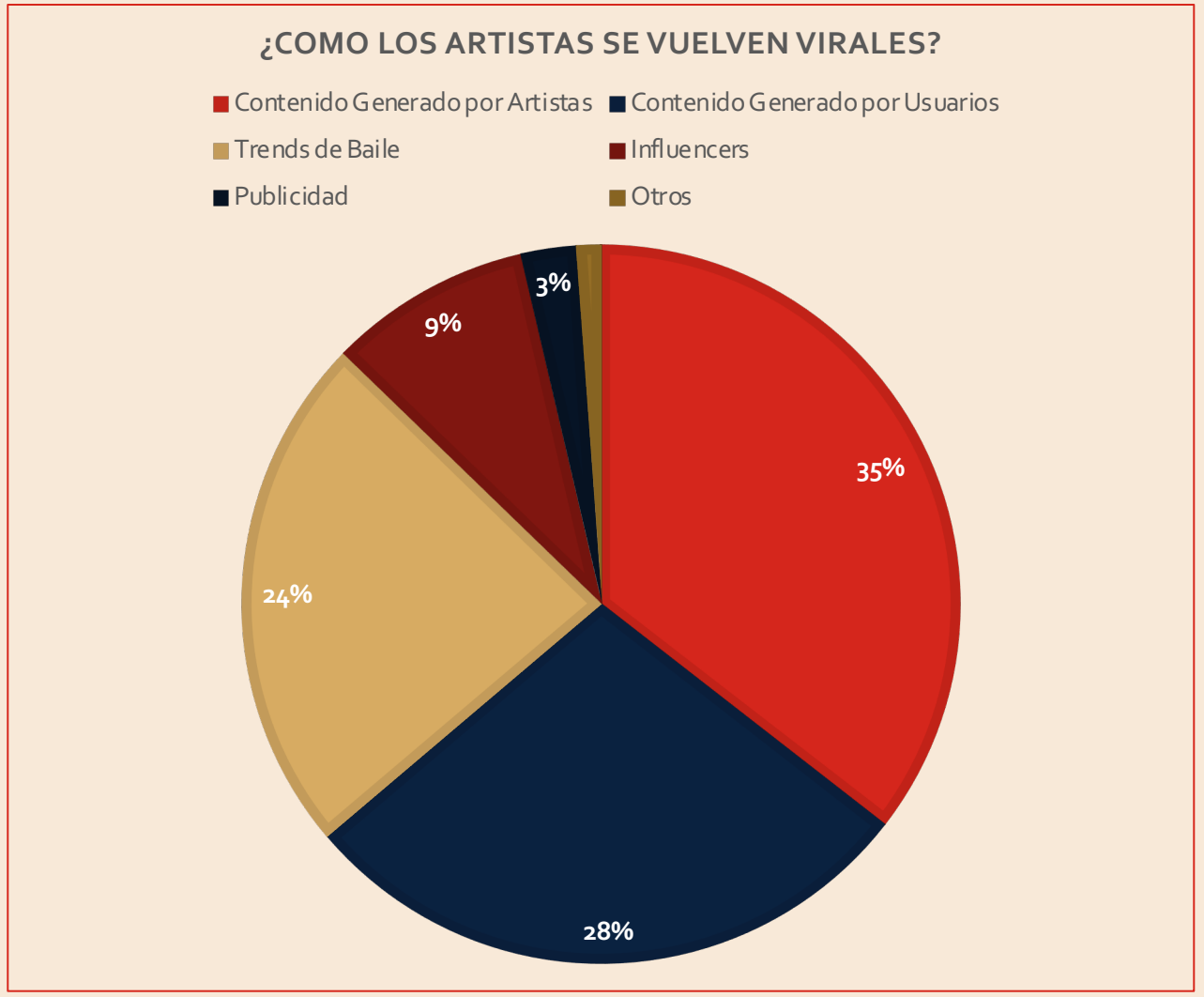
1. ¿QUÉ SUPUESTOS REALIZÓ PARA SABER SI UNA CANCIÓN SERÁ UN ÉXITO O NO EN LAS DISCOTECAS Y REDES SOCIALES?

SUPUESTOS

2. Para que una canción sea un éxito en redes sociales, esta debe ser bailable, es decir tener una alta "danceability" .

Justificación:

- "Que su música invite a bailar y permita crear contenido de retos de baile en redes sociales" (enunciado).
- La viralización en redes aumenta el número de "streams" en plataformas de "streaming" , según Contraband.
- Los "trends" de baile producen un 23.5% de los artistas que se vuelven virales en Tiktok y acomulan 20.5 millones de "streams" en promedio en Spotify, según Contraband .



Contraband (2022).

PREGUNTAS DE NEGOCIO

1. ¿QUÉ SUPUESTOS REALIZÓ PARA SABER SI UNA CANCIÓN SERÁ UN ÉXITO O NO EN LAS DISCOTECAS Y REDES SOCIALES?

SUPUESTOS

3. Solo existen los datos proporcionados y ningún otro factor influye en la construcción del algoritmo

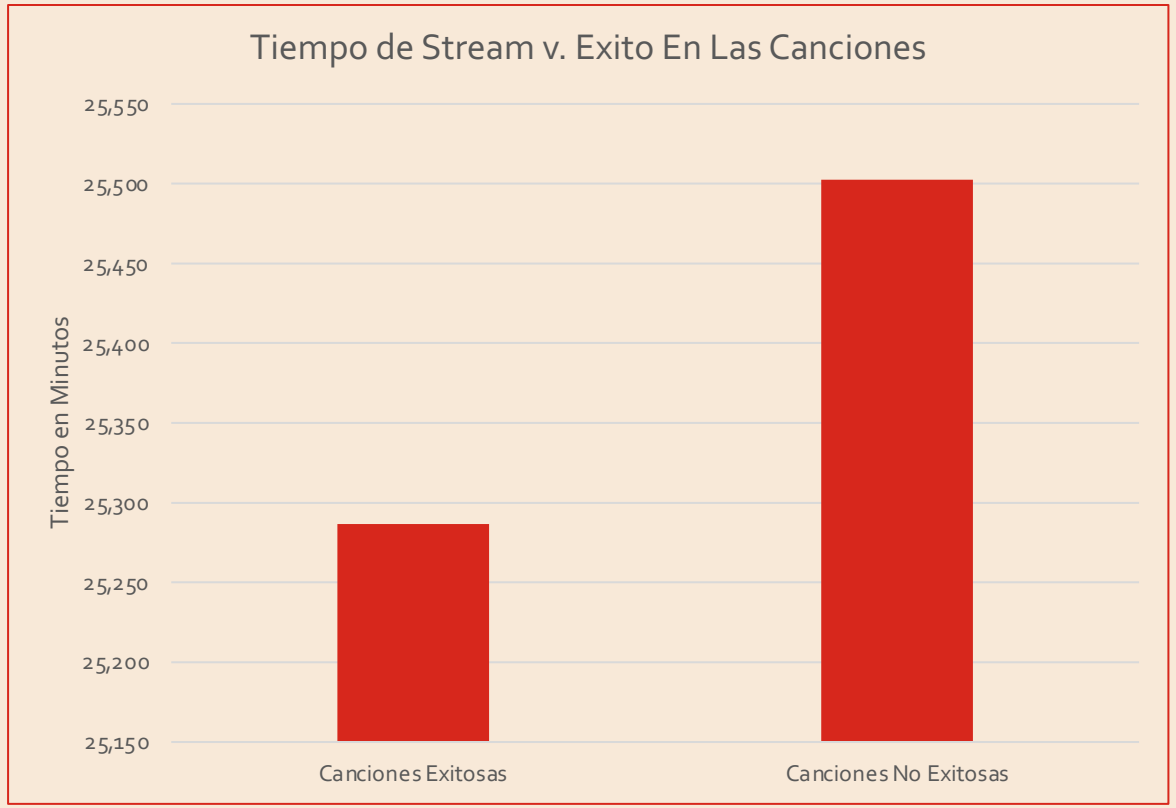
PREGUNTAS DE NEGOCIO

2. ¿CUÁNTO ES EL TIEMPO PROMEDIO PARA QUE UNA CANCIÓN QUE SEA UN ÉXITO EN REDES SOCIALES O EN UNA DISCOTECA, ES DECIR, SEA MUY BAILABLE? ¿EXISTE UNA GRAN DIFERENCIA ENTRE ESOS TIEMPOS?

RESPUESTA

1. El tiempo promedio total de una canción considerada como exitosa es 25.287 minutos.*
2. El tiempo promedio total de una canción no considerada como exitosa es 25,503 minutos. La desviación estándar para esta métrica es de 90.6 minutos*, por lo que se puede concluir que la diferencia es mínima.

* Este es un resultado basado en la base de datos; tiempo no corresponde a la duración promedio de una canción.

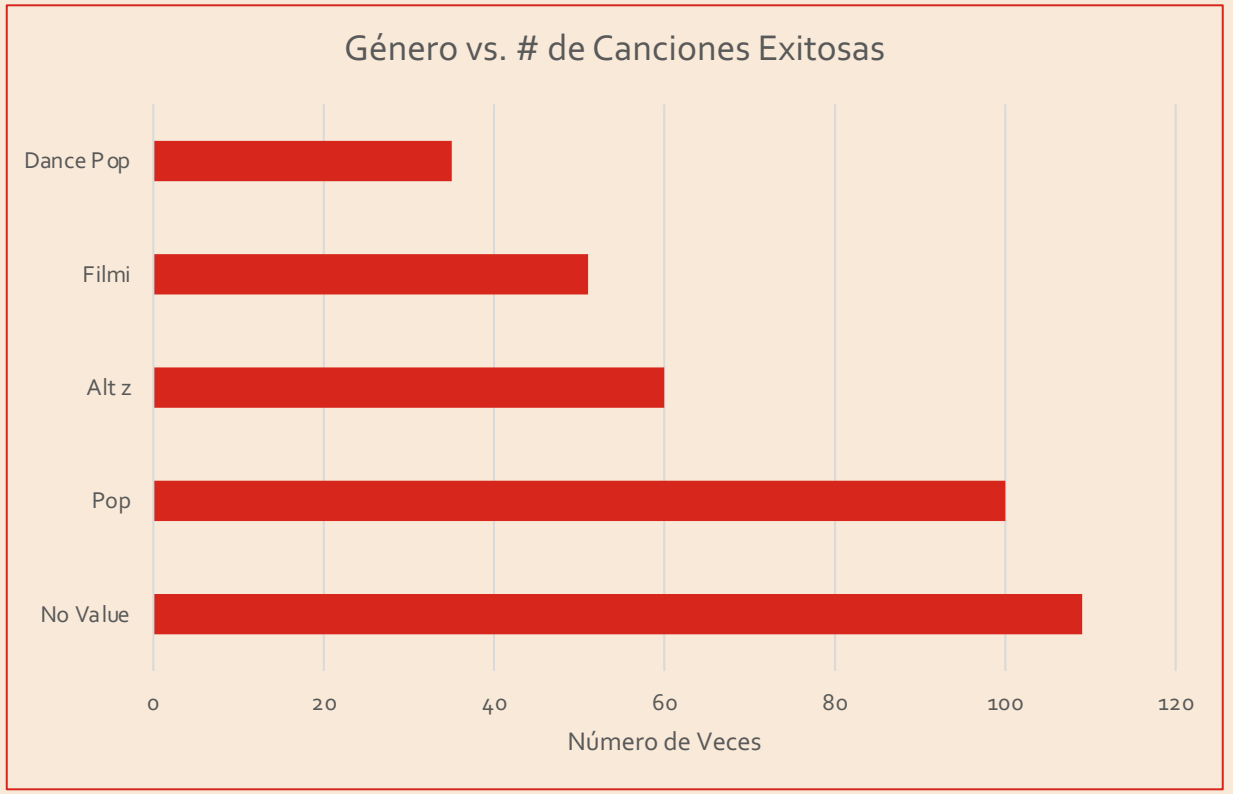


PREGUNTAS DE NEGOCIO

3. ¿QUÉ GÉNERO ES EL MÁS BAILADO DE TODOS?

RESPUESTA

1. El género más bailado no tiene un valor definido, el género que le sigue es Pop.



1. PREGUNTAS DE NEGOCIO

2. CONSTRUCCIÓN DEL MODELO

CONSTRUCIÓN DEL MODELO

1. LIMPIEZA DE DATOS

VARIABLES NULAS

1. Se borraron las columnas " Unnamed:22", " Unnamed: 23", "Unnamed: 24", "Unnamed: 25".
2. Se encontró que había filas duplicadas con el mismo ID y que contenían la misma información, entonces se borraron estas filas.
3. Se estableció el IDcomo index del Data Frame.
4. El resultado final fueron 4021 filas y 21 columnas.

id	trackName	artistName	msPlayed	genre	danceability	energy	key	loudness	mode	speechiness	...	instrumentalness	liveness	valence	tempo	type
7dTxqsaFGHOXwtzHlNjfHv	Honest	Nico Collins	191772	NaN	0.476	0.799	4.0	-4.939	0.0	0.2120	...	0.0	0.2570	0.577	162.139	audio_features
7K9Z3yFNNLv5kwTjQYGjnu	#BrooklynBloodPop!	SyKo	145610	glitchcore	0.691	0.814	1.0	-3.788	0.0	0.1170	...	0.0	0.3660	0.509	132.012	audio_features
3koAwrM1RO0TGMeQJ3qt9J	$10	Good Morning	25058	experimental pop	0.624	0.596	4.0	-9.804	1.0	0.0314	...	0.203	0.1190	0.896	120.969	audio_features
4ByEFOBuLXpCqvO1kw8Wdm	(I Just) Died In Your Arms	Cutting Crew	5504949	album rock	0.625	0.726	11.0	-11.402	0.0	0.0444	...	0.000169	0.0625	0.507	124.945	audio_features
22lJaG2yxlSjlwdUlddcFk	(L)only Child	salem ilese	2237969	alt z	0.645	0.611	8.0	-5.925	0.0	0.1370	...	2.05e-05	0.2370	0.645	157.475	audio_features
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
4D7ERaKgv8NAeck8RzUtRz	Younger	Jonas Blue	2560215	pop	0.784	0.845	3.0	-2.793	1.0	0.0596	...	1.59e-05	0.0740	0.888	105.981	audio_features
2qXicQG06oT0ijKBznpgQv	Younger	Ruel	5272303	alt z	0.745	0.477	11.0	-7.706	0.0	0.0880	...	0.0	0.1200	0.454	136.055	audio_features
6o8pM5reLgjd5i8gDY3Irt	Younger with Time.	Ben Zaidi	668478	folk-pop	0.537	0.143	2.0	-16.992	1.0	0.0331	...	0.00572	0.1100	0.245	131.118	audio_features
1EoThnDm6kQfB2idIfR30n	Your Love Is My Drug (8 Bit Slowed)	just valery	97600	sad lo-fi	0.282	0.158	6.0	-7.783	1.0	0.0311	...	0.134	0.4740	0.248	65.152	audio_features
042Sl6Mn83JHyLEqdK7ul0	Your Power	Billie Eilish	988224	art pop	0.632	0.284	9.0	-14.025	0.0	0.0801	...	0.000476	0.2330	0.208	129.642	audio_features

4621 rows × 21 columns

Muestra de la base de datos después de la limpieza

1. LIMPIEZA DE DATOS

VARIABLES QUE SE ELIMINARON

1. Las variables, **'type', 'id', 'uri', 'track_href', 'analysis_url'**, se eliminaron porque no aportaban información relevante que pudiera ayudar al modelo.

CONSTRUCCIÓN DEL MODELO

1. LIMPIEZA DE DATOS

VARIABLES NUMÉRICAS

Variables: 'msPlayed', 'energy', 'key', 'loudness', 'mode', 'speechiness', 'acousticness', 'instrumentalness', 'liveness', 'valence', 'tempo', 'duration_ms', 'time_signature'

1. Se cambió el tipo de la columna a "float64".
2. Se revisó cuántos datos nulos tenia.
3. Se revisó que no existiera ningún dato atípico .
4. Para la variable objetivo,"denceability", se eliminaron las filas que tuvieran un valor nulo.

```
df_music['msPlayed_numeric'] = pd.to_numeric(df_music['msPlayed'], errors='coerce')
```

Cambio de tipo a float

```
df_music['msPlayed_numeric'].describe()
✓ 0.0s

count    4.621000e+03
mean     1.526821e+06
std      5.437993e+06
min      0.000000e+00
25%      1.392240e+05
50%      2.697950e+05
75%      1.196295e+06
max      1.583671e+08
Name: msPlayed_numeric, dtype: float64
```

Revisar datos atípicos

CONSTRUCCIÓN DEL MODELO

1. LIMPIEZA DE DATOS

VARIABLES CATEGÓRICAS

Variable: 'trackName'

1. Se encontró el número de palabras y se guardo como una nueva columna.

Variable: 'artistName'

1. Se encontró la longitud del nombre y se guardo como una nueva columna.

Variable: 'genre'

1. Se revisó cuántos datos nulos existían
2. Se usó un "encoder" para cambiar las variable categóricas por numéricas.

```
# Number of words
df_music['trackName_numberOfWords'] =df_music['trackName'].apply(lambda x: len(x.split()))
```

Número de palabras en el nombre de la canción

```
# Length of the track Name
df_music['artistName_length']=df_music["artistName"].str.len()
```

Número de letras en el nombre del artita

```
# Create a LabelEncoder object
encoder = LabelEncoder()

# Fit and transform the feature
df_music['genre_encode'] = encoder.fit_transform(df_music["genre"])
```

Encoder

CONSTRUCCIÓN DEL MODELO

2. VARIABLE OBJETIVO Y CUALES SON LAS VARIABLES MAS SIGNIFICATIVAS

VARIABLE OBJETIVO DEFINITIVA

1. Se tomó como variable objetivo a "denceability"
2. Se encontró que para "denceability" la variable debía ser superior a 0.714 con el fin de que solo fueran exitosas 25% de la muestra, debido a que esto representa el valor en el que empieza el último cuartil.
3. Según los supuestos hechos en el modelo de negocio, se estableció que la variable sería "denceability".

count	4620.000000
mean	0.604560
std	0.155515
min	0.000000
25%	0.512000
50%	0.624000
75%	0.714250
max	0.976000

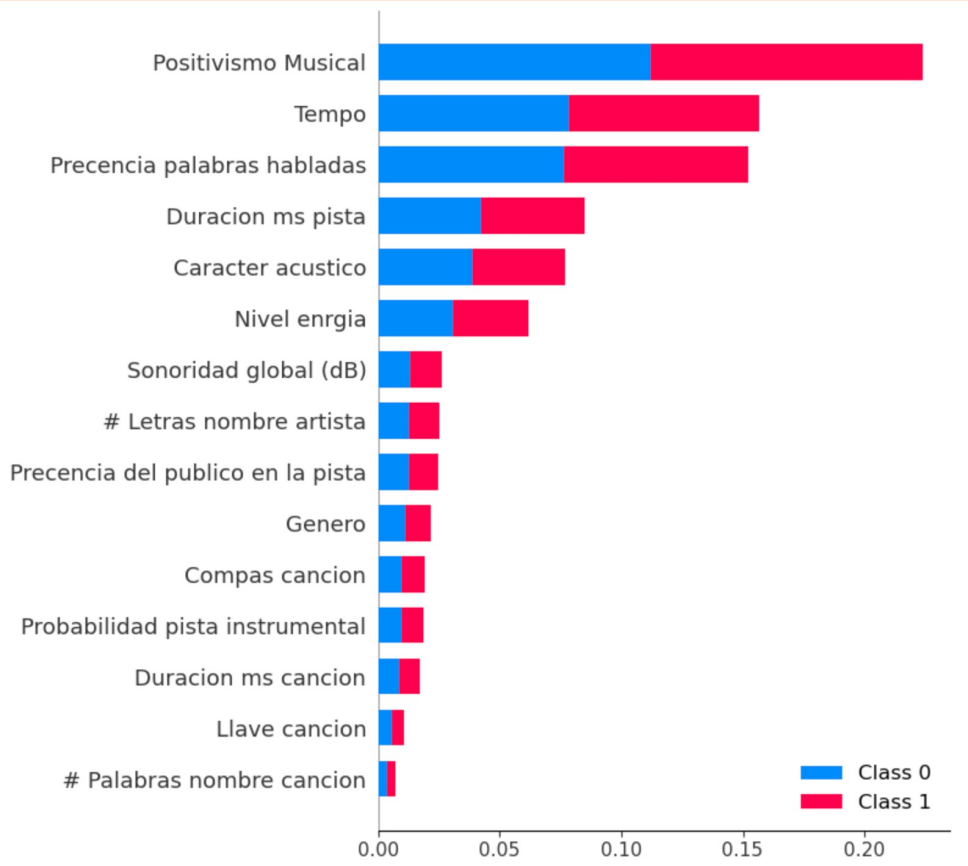
Descripción Denceability

CONSTRUCCIÓN DEL MODELO

2. VARIABLE OBJETIVO Y CUALES SON LAS VARIABLES MAS SIGNIFICATIVAS

VARIABLES MÁS SIGNIFICATIVAS

- Las variables más importantes son positivismo musical, presencia de palabras habladas en la canción, tempo de la pista en pulsaciones por minuto, el carácter acústico de la pista, nivel de energía de una canción.
- La grafica muestra las variables más importantes al momento de predecir tanto un éxito como una canción que no lo sea.



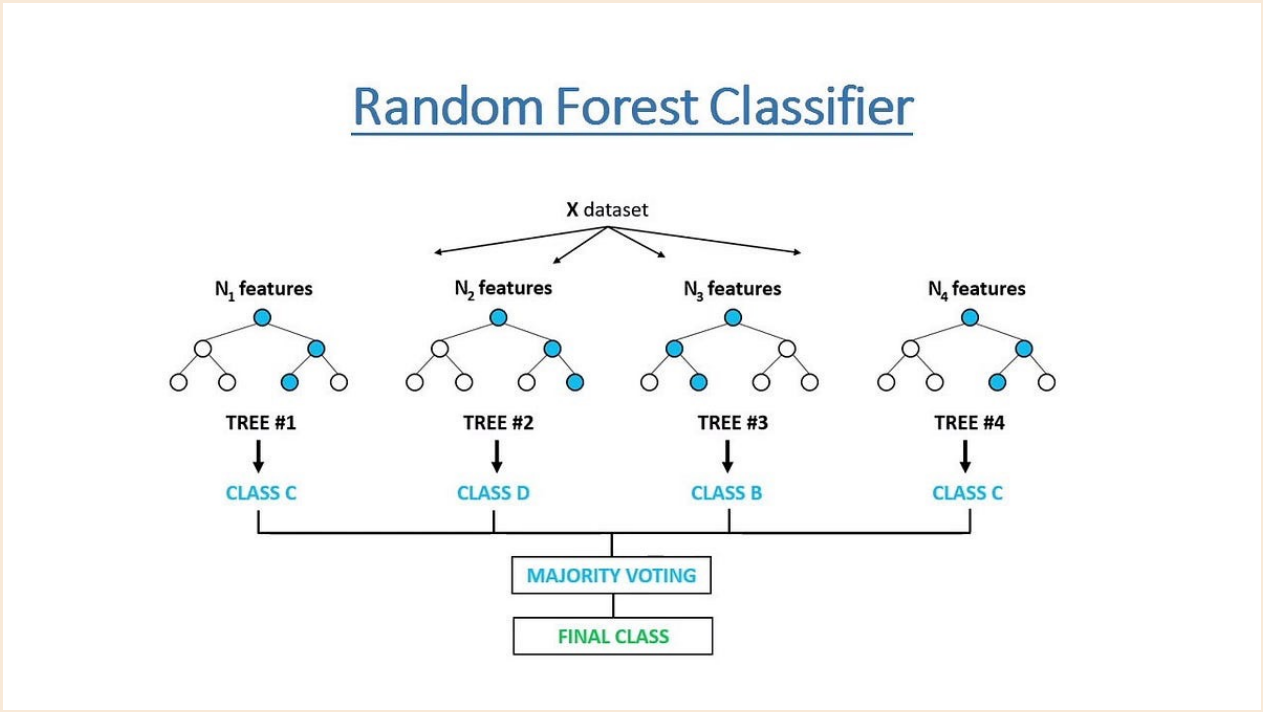
Variables que más influyen en la predicción del modelo.

2. ALGORITMO UTILIZADO

VENTAJAS DEL MODELO

1. Random Forest:

- Es muy bueno para problemas de clasificación binaria.
- Tiene un alto rendimiento y precisión en la predicción.
- No se necesita normalizar los datos.
- Tiene muy buen manejo de datos nulos
- Tiene alta escalabilidad.
- Es fácil de interpretar.
- Como la muestra no era tan extensa, este algoritmo ayuda a no sobre estimar.



RandomForest

CONSTRUCCIÓN DEL MODELO

2. ALGORITMO UTILIZADO

EXPLICACIÓN DEL MODELO

1. Se dividió la data en segmentos de entrenamiento y testeo.
2. Se solucionaron problemas con datos nulos .
3. Se entrenó el modelo con los parámetros óptimos encontrados con "*Grid Search*".
4. Finalmente se ajustó el modelo, encontrando todas las métricas claves.
5. Explicación *Random Forest*: Se construye un árbol de decisión, múltiples veces. Se combinan las predicciones para hacer una predicción final y más acertada.

```
# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

Se divide la data

```
# Create the XGBoost model
best_params = {'class_weight': {0: 1, 1: 4}, 'max_depth': 10, 'min_samples_leaf': 2, 'min_samples_split': 10, 'n_estimators': 100}
model = RandomForestClassifier(**best_params)
```

Se entrena el modelo

```
# Fit the model to the training data
model.fit(X_train, y_train)
```

Se ajusta el modelo

2. ALGORITMO UTILIZADO

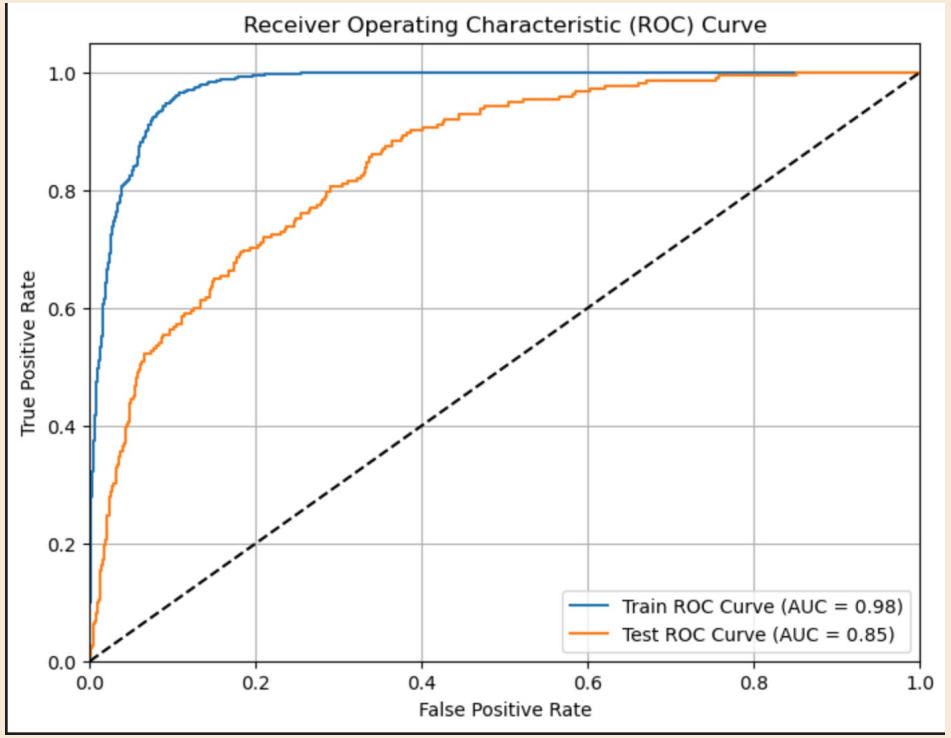
SUPUESTOS DEL MODELO Y RESULTADOS

Supuestos

1. El *threshold* se estableció en 0.6 con el fin de disminuir el error de tipo 1.

Resultados

1. La curva de ROC nos ayuda a ver que escogimos el *threshold* correcto con el fin de clasificar los resultados de la mejor manera. El optimo se encuentra en la esquina izquierda superior y el área bajo la curva también se acerca a 1.
2. La curva de entrenamiento no se aleja significativamente comparada a la curva de prueba, lo que indica que no hay una sobre estimación.



ROC Curve

2. ALGORITMO UTILIZADO

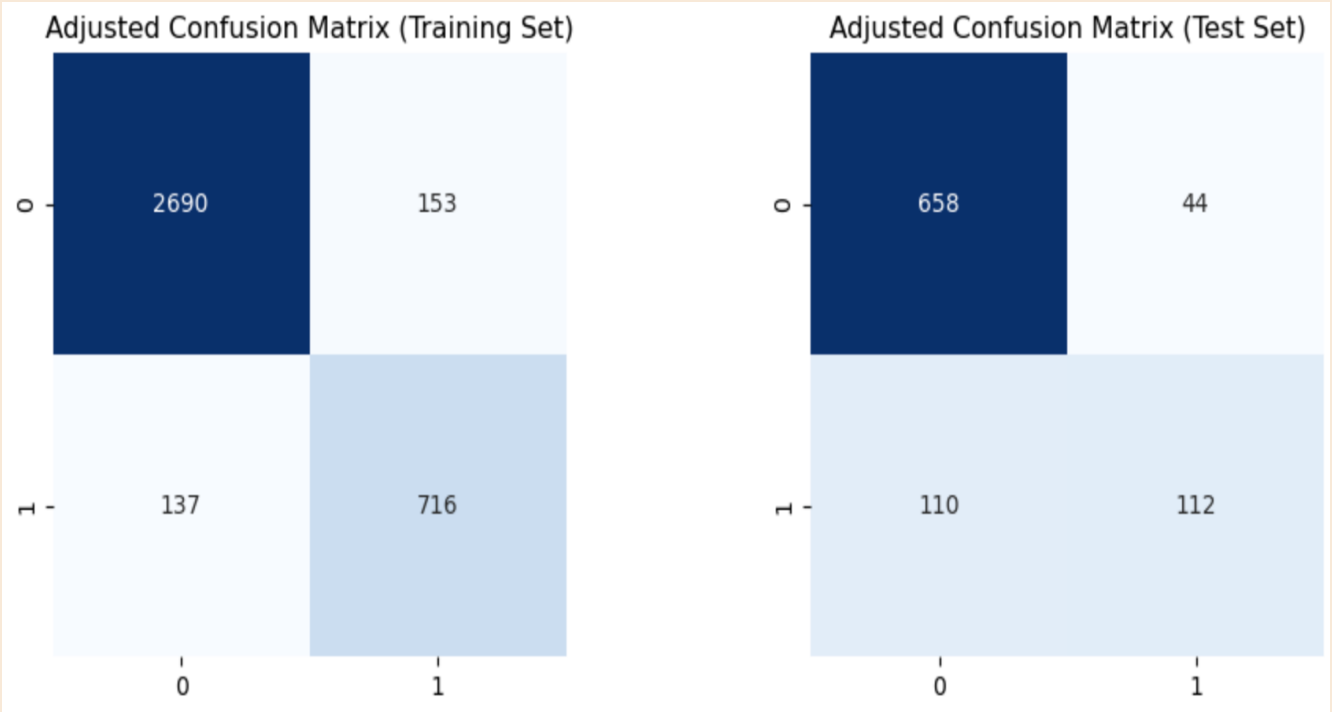
SUPUESTOS DEL MODELO Y RESULTADOS

Supuestos

2. Parámetros usados fueron definidos por una búsqueda de *Grid Search* .

Resultados

1. La matriz de confusión muestra que el modelo reduce los errores de tipo 1 para los datos de prueba.
2. La matriz de confusión de entrenamiento no es perfecta, lo que es un buen indicio para un modelo.



Matriz de Confusión

2. ALGORITMO UTILIZADO

SUPUESTOS DEL MODELO Y RESULTADOS

Resultados

3. La empresa busca tener una alta precisión, es decir disminuir los falsos positivos, esto disminuye la inversión en canciones que se cree que van a tener éxito, pero en realidad no lo tendrán.
4. Se identifican con mucha exactitud las canciones que no serán un éxito (especificidad).
5. El modelo puede predecir una canción que en realidad será un éxito un 72% de las veces.
6. Las métricas entre ambos modelos indican que no se esta sobre estimando el modelo, ya que sus diferencias no son significativas .

```
Adjusted Metrics:
Training Set Metrics (Adjusted):
Accuracy: 0.9215
Precision: 0.8239
Recall: 0.8394
F1-Score: 0.8316
ROC-AUC: 0.9755

Test Set Metrics (Adjusted):
Accuracy: 0.8333
Precision: 0.7179
Recall: 0.5045
F1-Score: 0.5926
ROC-AUC: 0.8480
```

Resultados del modelo

FUENTES

- Contraband (2022). HOW ARTISTS ARE GOING VIRAL ON TIKOK IN 2022. https://drive.google.com/file/d/1FCvkBuXt4GQuASQifwas1QO-VmMI-Vjc/view