Proyecto Final Data Science 2021

1 FEBRERO

Gallo, Florentina
Grillo, Gian Franco
Pizzingrilli, Federico
Quiroga, Gerardo

Contenido

CONSIGNA

Objetivos Generales:

- Entender el problema de negocio e identificar los elementos a ser considerados para el planteamiento de un Modelo de *Data Science*.

- Describir los datos de negocio y las relaciones entre datos mediante el Análisis Exploratorio de Datos.

- Construir una presentación ejecutiva para la alta gerencia mostrando los resultados obtenidos.

Objetivos Específicos:

- Desarrollar las instancias de *Data Acquisition* y *Data Wrangling* en tu trabajo final.

- Lograr una articulación en equipo y una división de tareas adecuadas a los objetivos.

- Realizar Filtrado.

- Describir qué significa cada variable, cómo se comporta.

- Especificar las distribuciones y relaciones (géneros, sexo, edad, IVA, tipo de empresa).

Se debe entregar:

- Presentación de la empresa, organización o problema específico.

- Preguntas y objetivos de la investigación.

- Conformación del equipo de trabajo.

- Indicación de la fuente del *dataset* y los criterios de selección (*Data Acquisition*).

- Generación del primer *Data Wrangling* y EDA, apuntado a sus datos (*insights*) univariado, bivariado y multivariado.

- Análisis de componentes principales.

- Contar la historia de sus datos

- Filtros aplicados a los datos. Distribución. *Dataset* final para analizar.

- *Plottear* objetivos u objetivo para esos datos.

PRESENTACIÓN

El equipo está conformado por:

- Gallo, Florentina
- Grillo, Gian Franco
- Pizzingrilli, Federico
- Quiroga, Gerardo

Los temas de interés principales del equipo son Música, Deportes, y Salud.

Con esta premisa se comenzó con la búsqueda de diferentes *datasets* de las principales temáticas de interés.

Si bien se barajaron diferentes opciones, se decidió abordar la temática musical.

Para esto, se encontró el set de datos "The Spotify Hit Predictor" con pistas de entre 1960 y 2019, clasificadas en si fueron un hit o no.

Se estudiará y analizará el mismo, únicamente en la década de 2010's, para cumplir con los requisitos de volumen de datos solicitados.

Se espera obtener un modelo de clasificación binaria que clasifique pistas en 'Hit' o 'Flop' (No Hit).

DATASET

The Spotify Hit Predictor Dataset (1960-2019).

Este *dataset* contiene *features* de canciones obtenidas mediante la API de *Spotify*. Las pistan están etiquetadas como '1' o '0' ('Hit' o 'Flop') dependiendo de algunos criterios del autor.

Este set de datos contiene información de las pistas, tanto características, como artista y nombre, como de índole "musical" de las mismas, como la energía y la acusticidad.

Este conjunto de datos se puede usar para hacer un modelo de clasificación que predice si una pista sería un 'Hit' o no. Las pistas son clasificadas como 'Hit' si es que formaron parte de la lista Hot-100 de los *Billboard* al menos 1 vez.

(Nota: el autor no considera objetivamente una pista inferior, mala o un fracaso si está etiquetada como *'Flop'*. *'Flop'* aquí simplemente implica que es una pista que probablemente no podría considerarse popular en el *mainstream*).

Si bien se cuenta con pistas entre 1960 y 2019, se utilizarán únicamente pistas entre 2010-2019 por el alto volumen del *dataset*.

Fuente: **The Spotify Hit Predictor Dataset (1960-2019).**

DESCRIPCIÓN DE VARIABLES

- track: Nombre de la canción.

- artist: Nombre del artista.

- uri: Identificador del recurso de la canción.

- danceability: Capacidad de baile. Describe qué tan adecuada es una pista para bailar en función de una combinación de elementos musicales que incluyen el tempo, la estabilidad del ritmo, la fuerza del ritmo y la regularidad general. Un valor de 0.0 es menos bailable y 1.0 es más bailable

- energy: la energía es una medida de 0.0 a 1.0 y representa una medida perceptiva de intensidad y actividad. Por lo general, las pistas enérgicas se sienten rápidas, altas y ruidosas. Por ejemplo, el *death metal* tiene mucha energía, mientras que un preludio de Bach tiene una puntuación baja en la escala. Las características de percepción que contribuyen a este atributo incluyen rango dinámico, volumen percibido, timbre, frecuencia de inicio y entropía general.

- key: Clave general estimada de la pista. Los enteros se asignan a los tonos utilizando la notación estándar de clase de tono. P.ej. 0 = C, 1 = C?/D?, 2 = D, y así sucesivamente. Si no se detectó ninguna clave, el valor es -1.

- loudness: Volumen general de una pista en decibelios (dB). Los valores de sonoridad se promedian en toda la pista y son útiles para comparar la sonoridad relativa de las pistas. La sonoridad es la cualidad de un sonido que es el principal correlato psicológico de la fuerza física (amplitud). Los valores típicos oscilan entre -60 y 0 db.

- mode: Indica la modalidad (mayor o menor) de una pista, el tipo de escala de la que se deriva su contenido melódico. Mayor está representado por 1 y menor es 0.

- speechiness: El habla detecta la presencia de palabras habladas en una pista. Cuanto más exclusivamente parecida a un discurso sea la grabación (por ejemplo, programa de entrevistas, audiolibro, poesía), más cercano a 1.0 será el valor del atributo. Los valores superiores a 0,66 describen pistas que probablemente estén compuestas en su totalidad por palabras habladas. Los valores entre 0,33 y 0,66 describen pistas que pueden contener tanto música como habla, ya sea en secciones o en capas, incluidos casos como la música rap. Los valores por debajo de 0,33 probablemente representen música y otras pistas que no se parecen al habla.

- acousticness: Medida de confianza de 0.0 a 1.0 de si la pista es acústica. 1.0 representa una alta confianza en que la pista es acústica.

- instrumentalness: Predice si una pista no contiene voces. Los sonidos “Ooh” y “aah” se tratan como instrumentales en este contexto. Las pistas de rap o de palabra hablada son claramente "vocales". Cuanto más cercano esté el valor de instrumentalidad a 1.0, mayor será la probabilidad de que la pista no contenga contenido vocal. Los valores superiores a 0,5 están destinados a representar pistas instrumentales, pero la confianza es mayor a medida que el valor se acerca a 1,0.

- liveness: Detecta la presencia de una audiencia en la grabación. Los valores de *liveness* más altos representan una mayor probabilidad de que la pista se haya interpretado en vivo. Un valor superior a 0,8 proporciona una gran probabilidad de que la pista esté "en vivo".

- valence: Una medida de 0.0 a 1.0 que describe la positividad musical que transmite una pista. Las pistas con valencia alta suenan más positivas (p. Ej., Feliz, alegre, eufórico), mientras que las pistas con valencia baja suenan más negativas (p. Ej., Triste, deprimido, enojado).

- tempo: El tempo global estimado de una pista en tiempos por minuto (BPM). En terminología musical, el tempo es la velocidad o el ritmo de una pieza determinada y se deriva directamente de la duración media del tiempo.

- duration_ms: Duración de la pista en milisegundos.

- time_signature: Una signatura de tiempo total estimada de una pista. Convención de notación para especificar cuántos tiempos hay en cada compás.

- chorus_hit: Esta es la mejor estimación del autor de cuándo comenzaría el coro para la pista. Es la marca de tiempo del inicio de la tercera sección de la pista. Esta función se extrajo de los datos recibidos por la llamada API para el análisis de audio de esa pista en particular.

- sections: El número de secciones que tiene la pista en particular. Esta función se extrajo de los datos recibidos por la llamada a la API para el análisis de audio de esa pista en particular.

- target: La variable objetivo de la pista. Puede ser '0' o '1'. '1' implica que esta canción ha aparecido en la lista semanal (emitida por *Billboards*) de pistas Hot-100 en esa década al menos una vez y, por lo tanto, es un 'Hit'. '0' Implica que la pista no es un 'Hit' ('Flop').
La condición del autor para que una pista sea 'Flop' es la siguiente:
    - La pista no debe aparecer en la lista de 'hits' de esa década.
    - El artista de la pista no debe aparecer en la lista de 'hits' de esa década.

- La pista debe pertenecer a un género que pueda considerarse no *mainstream* y/o vanguardista.
- El género de la pista no debe tener una canción en la lista de "hits".
- La pista debe tener 'US' como uno de sus mercados.

STORYTELLING

¿Qué pasaría si en una fiesta con amigos quisiera sorprenderlos poniendo música que aún no escucharon pero que en poco tiempo comenzará a sonar en todos lados y será escuchada por muchas personas?

¿Qué pasaría si fuera un músico amateur y quisiera grabar una canción que tenga buena adaptación por las personas que la escuchen, y por qué no, poco a poco comenzar a sonar en radios y diferentes plataformas de transmisión de contenidos?

Ya sea para quedar cómo un genio musical y sorprender a amigos o compañeros de trabajo, o para producir canciones que tengan un alto impacto en los oyentes y poder obtener ganancias, es necesario hacer un análisis sobre canciones que son o fueron un 'Hit' en el *mainstream*. Luego, con este análisis tendremos mayor probabilidad de saber con antelación si una canción será considerada también un *Hit*.

Con estas premisas, y un alto interés en el impacto de la música en la sociedad, nuestro equipo se propuso poder identificar canciones que serán un *Hit*, todo esto, utilizando modelos de *Machine Learning*.

Para esto se utilizó el set de datos "*The Spotify Hit Predictor"*, específicamente, las canciones a partir de la década de 2010's, que fue cuando se produjo un *"Boom"* de las plataformas de reproducción de música vía *streaming*.

Este set de datos nos provee información de cada pista, como el o los artistas que la compusieron, y otras características de índole musical calculadas a partir de la pista de audio. Estas características musicales son la energía, clave de la pista, capacidad de baile, volumen general, modalidad, presencia de habla, nivel acústico, instrumentalidad, presencia de audiencia, positividad, tempo en BPM, signatura, duración, secciones, y la identificación de si es un *Hit* o no.

El criterio empleado para determinar si una pista es un hit o no, es el siguiente:

- Debe aparecer en la lista de *Hits* de la lista semanal de los *Billboard* en, al menos, una semana.
- Debe ser de un género considerado *mainstream.*
- Debe contener a los Estados Unidos como uno de sus mercados.

Con un total de casi 6400 pistas para poder estudiar se comenzó un análisis exploratorio de datos para poder obtener información relevante que nos permita predecir si una canción será un hit.

Estudiando todas las pistas disponibles, se detectó con no hay faltante de información para ninguna de las variables, pero sí se encontraron 20 casos (<1% del total de datos) donde se presentaba una canción de un artista y con el mismo URI (Identificador único). Estos casos se descartaron para poder preservar la unicidad e integridad del conjunto de datos.

Luego se analizó la distribución de los datos de forma aislada y sin relacionarlos entre sí. Con este análisis se pudo entender más en profundidad ciertas características de la distribución de las pistas, como por ejemplo:

- Muy poca probabilidad de que una pista sea en vivo. (Fig. 1)
- La gran mayoría de canciones no duran más de 4 minutos. (Fig. 2)
- Bajo nivel acústico. (Fig. 3)

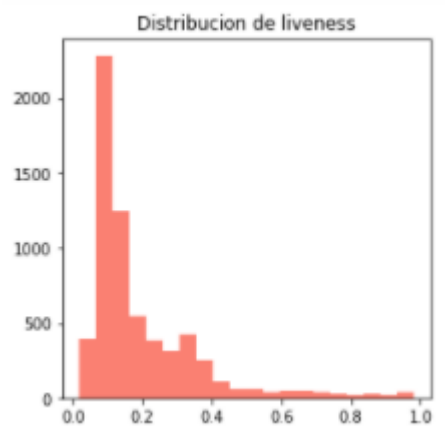

Fig. 1. Presencia de audiencia.

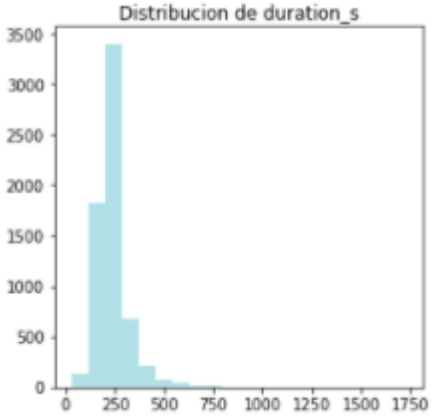

Fig. 2. Duración en segundos.

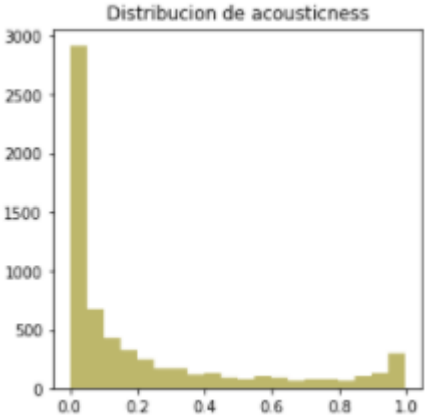

Fig. 3. Nivel acústico.

- Nivel de energía alto. (Fig. 4)
- Alta capacidad de baile, concentrándose la mayoría en >= 0.5. (Fig. 5)
- Poca instrumentalidad, es decir, no hay sólo instrumentos en las pistas. (Fig. 6)

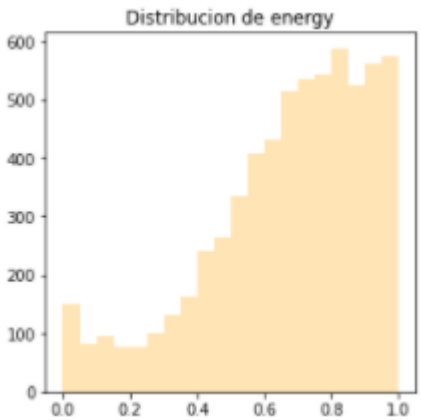

Fig. 4. Niveles de energía.

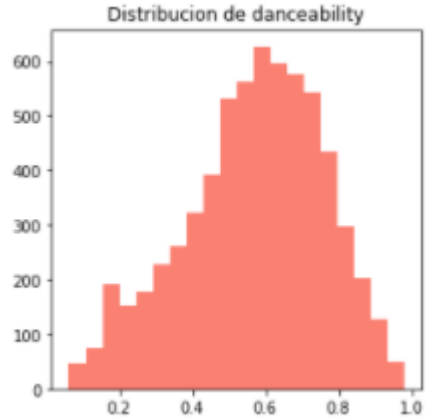

Fig. 5. Capacidad de baile.

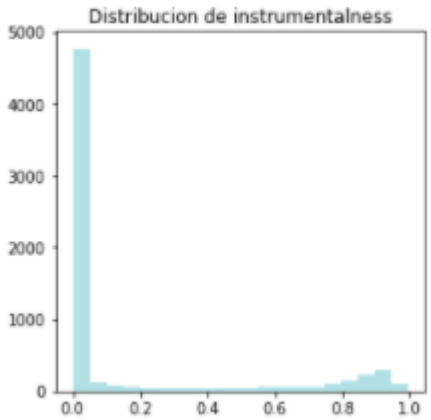

Fig. 6. Instrumentalidad.

- Pistas concentradas en torno a los -10 dB de volumen general. (Fig. 7)
- Positividad musical entre 0,2 y 0,6, ni muy negativa ni muy positiva. (Fig. 8)
- Misma cantidad de pistas ‘Hit’ que ‘Flop’ (No hit). (Fig. 9)

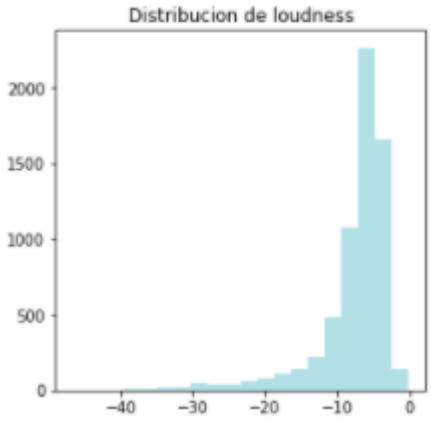

Fig. 7. Volumen general en dB.

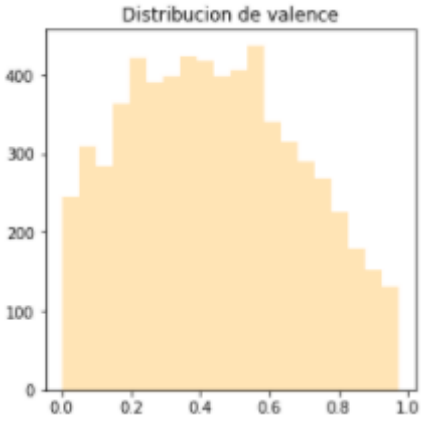

Fig. 8. Positividad musical.

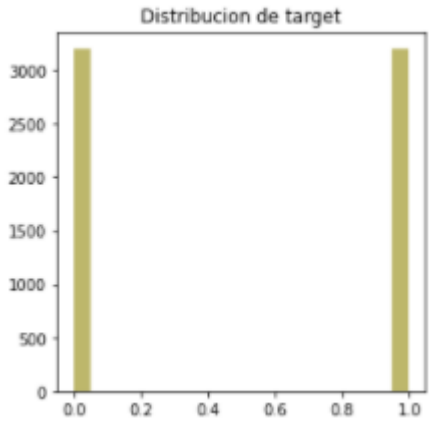

Fig. 9. Pistas Hit o *Flop* (No Hit).

- Casi el doble de pistas de modalidad Mayor respecto a las de modalidad Menor. (Fig. 10)
- Velocidad o Tempo entre 75 y 175 BPM’s para el “grueso” de las pistas. (Fig. 11)
- Poca presencia del habla, es decir, no son relatos o cuasi relatos (por ej. Pistas de RAP). (Fig. 12)

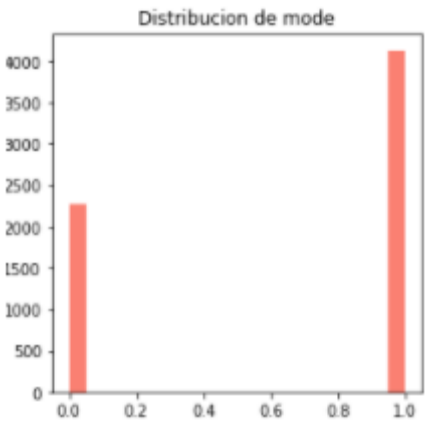

Fig. 10. Modalidad (Mayor o Menor).

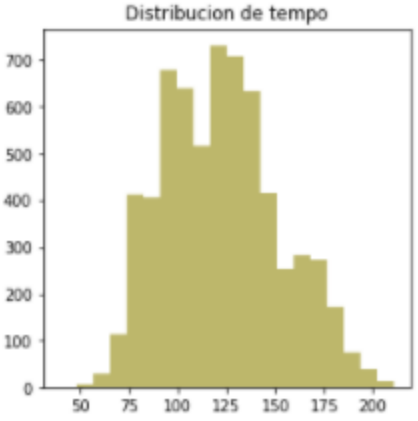

Fig. 11. Tempo en BPM.

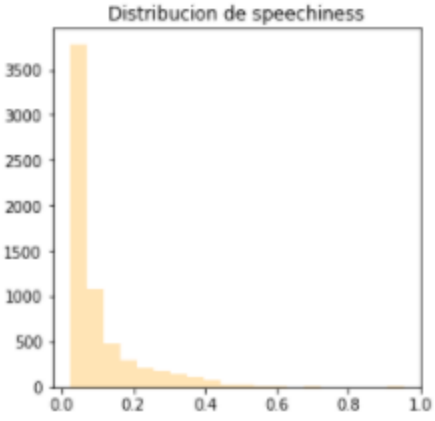

Fig. 12. Presencia del habla.

Estos análisis nos permitieron comprender la distribución de cada variable respecto a la cantidad de pistas.

Luego se pasó a estudiar el comportamiento de las variables, comparando una respecto a la otra.

Se detectó una alta relación entre energía y volumen general (*loudness*), nivel acústico y energía, nivel acústico y volumen general (*loudness*), secciones y duración en segundos. (Fig. 13).

Además, si bien no hay un alto nivel de correlación, se detectó cierta relación entre las variables capacidad de baile, instrumentalidad y volumen general (*loudness*). Estos niveles nos indican que los comportamientos de estas 3 variables son, de cierta forma, dependientes entre sí.

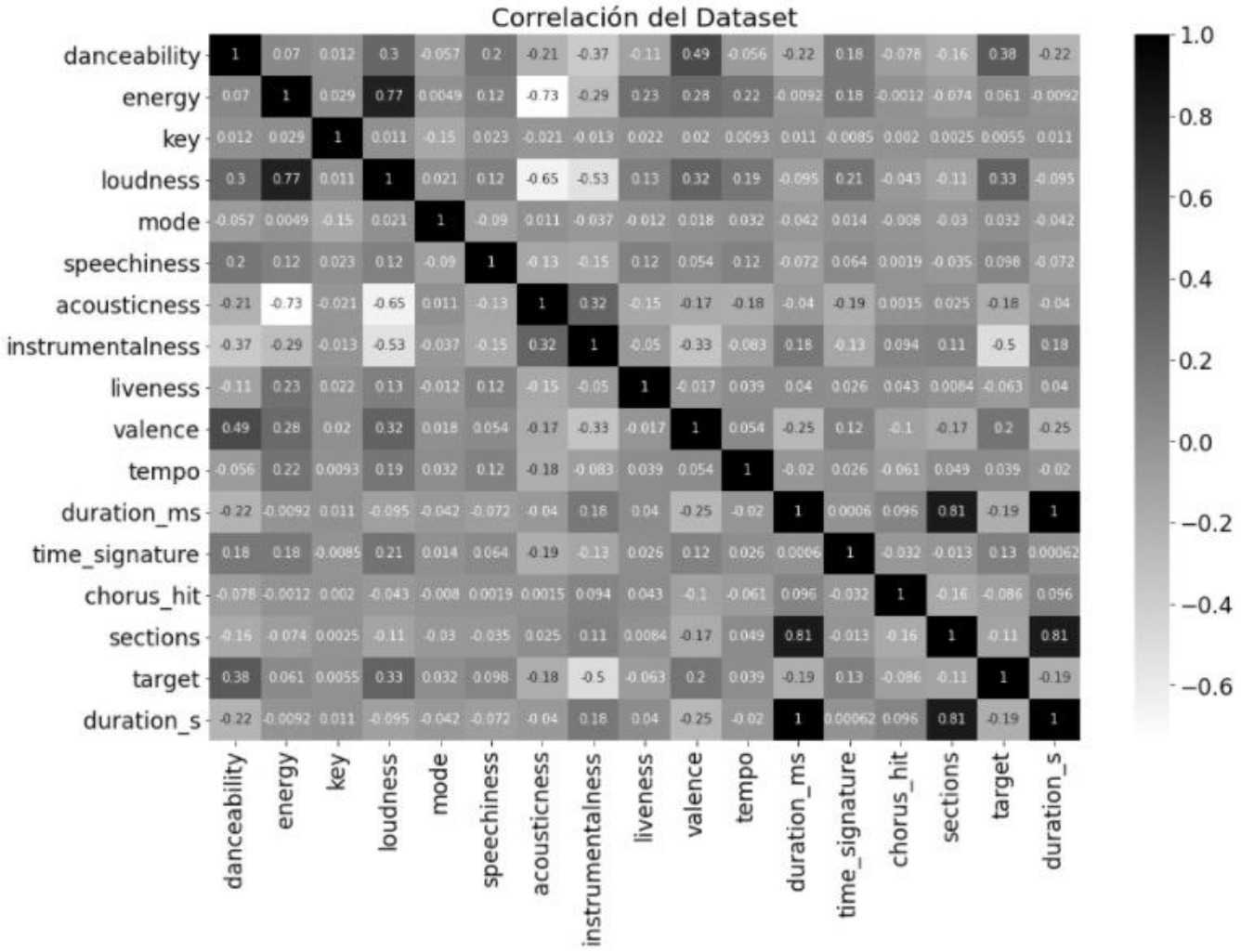

Fig. 13. Correlación de variables del set de datos.

NOTEBOOK

Para ver el análisis completo con el código utilizado, gráficos, y comentarios, ingresar a:
Notebook de la 1° entrega del proyecto final

SELECCIÓN Y EVALUACIÓN DE ALGORITMOS

Se han evaluado diferentes algoritmos de clasificación, con diferentes configuraciones de híper-parámetros, y utilizando diferentes variables en los mismos.

Los algoritmos evaluados fueron *Decision Tree Classifier, Random Forest Classifier, Support Vector Classifier, y Gradient Boosting Classifier.*

La evaluación fue realizada en base a la matriz de confusión. La métrica objetivo definida fue la exactitud o *accuracy*. Se definió utilizar esta métrica como comparación ya que nos permite evaluar los elementos clasificados correctamente, tanto las canciones que fueron hits como las que no, y al tener las clases de la muestra balanceadas (misma cantidad de *hits* que no *hits)* no había problema con la misma.

$$\text{VP} = \text{Verdaderos Positivos}$$
$$\text{VN} = \text{Verdaderos Negativos}$$
$$\text{FP} = \text{Falsos Positivos}$$
$$\text{FN} = \text{Falsos negativos}$$

$$accuracy = \frac{VP + VN}{VP + VN + FP + FN}$$

Fig. 14. Fórmula para calcular el accuracy.

Además, se determina como objetivo reducir al máximo posible los falsos positivos, es decir, las canciones que fueron clasificadas como *hit* cuando no lo eran. Este objetivo planteado se debe a que, si queremos invertir en producir una canción o un video musical, buscamos tener la mayor certeza posible de que una canción será un *hit* de forma que nuestra inversión quede asegurada y que podamos obtener ganancias luego de invertir en las mismas. Buscando el modelo con la menor cantidad de falsos positivos y mejor *accuracy* tendremos mayor certeza de que el modelo no clasificará canciones como *hits* cuando no lo son, por lo que es una confianza extra que se brinda para tomar la decisión de invertir o no en una producción musical.

El modelo con un desempeño considerablemente malo respecto al resto de los modelos fue el SVC (*Support Vector Classifier)*. Dicho modelo obtuvo un *accuracy* del 68%, siendo este el mejor resultado luego de haber probado diferentes combinaciones de híper-parámetros

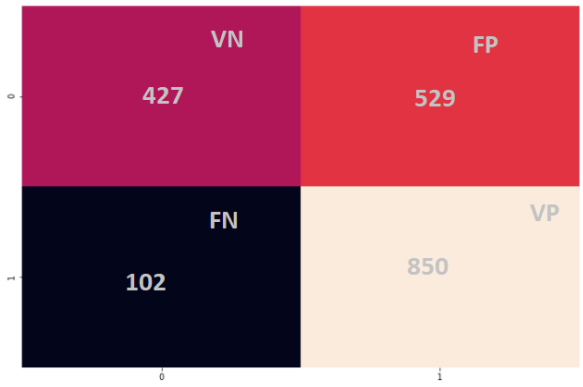

Fig. 15. Matriz de confusipon del *Support Vector Classifier*

Luego, el siguiente modelo que, si bien tuvo un accuracy considerablemente mejor, quedó un 5% del mejor modelo encontrado, fue el *Decision Tree Classifier.* Con este modelo se obtuvo un *accuracy* del 80%. Este *accuracy* se obtuvo luego de haber probado diferentes combinaciones de híper-parámetros.
Otro enfoque que se probó fue el de haber utilizado las 4 variables más significativas, siendo que entre ellas sumaban un 94% de representación de los datos. Sin embargo, esto no produjo que se mejorara el resultado, sino que empeoró en un 0,3%.

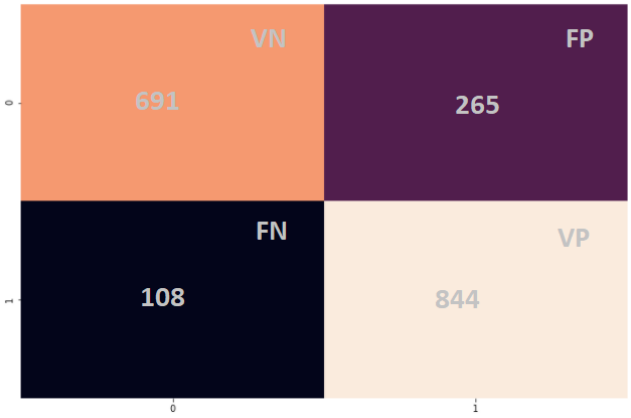

Fig. 16. Matriz de confusipon del *Decission Tree Classifier*

El tercer modelo, cuyo *accuracy* fue de 84,28%, quedando solamente 0,3% por debajo del *accuracy* del mejor modelo, fue el *XGBoost Classifier (eXtreme Gradient Boosting Classifier)*. Dado que la diferencia es ínfima, tranquilamente se podría utilizar como el modelo clasificador, por lo que se podrán evaluar y comparar los tiempos de entrenamiento, y tamaño de los modelos para tomar la decisión final.
En este caso, se lo deja como segunda opción ya que se obtuvieron 170 falsos positivos, es decir, 170 canciones que fueron clasificadas como *Hit* cuando no lo eran.

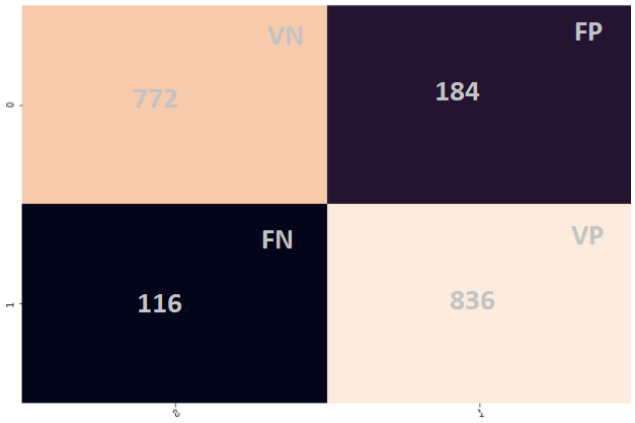

Fig. 17. Matriz de confusión del *XGBoost Classifier*

El modelo con mayor *accuracy* y menor cantidad de Falsos Positivos fue el *Random Forest Classifier*.
Al igual que con los otros modelos, se probó utilizar las variables más significativas y diferentes combinaciones de híper-parámetros.
El *accuracy* fue de 84.5%, con tan sólo 170 falsos positivos.
Los parámetros del modelo son: 200 estimadores, 15 muestras mínimas para aperturar un nodo, y 1 muestra como mínimo para que un nodo sea considerado raíz.
Además, las otras métricas calculadas para este modelo fueron el F1 *Score* para medir el *accuracy* y la sensibilidad siendo este de 94%, y el *recall* o sensibilidad para determinar la cantidad de verdaderos positivos clasificados en función del total de los positivos de la muestra, obteniendo un 94%.

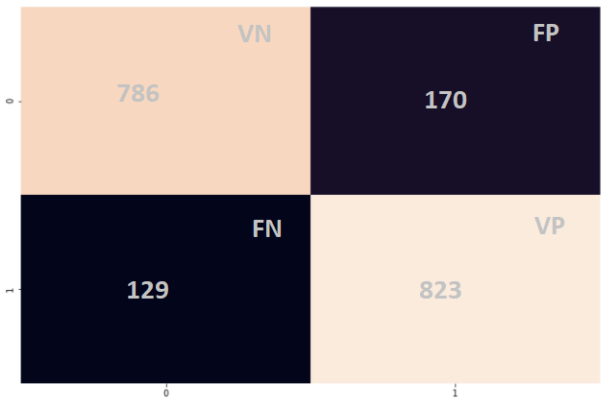

Fig. 18. Matriz de confusión del *Random Forest Classifier*

CONCLUSIÓN

El modelo recomendado para utilizar es un Random Forest Classifier, con los parámetros:

- n_estimators = 200
- max_depth = None
- min_samples_leaf = 1
- min_samples_split = 15

Con este modelo obtenemos las siguientes métricas:

- *Accuracy* o precisión = 84.5%
- *Recall* o sensibilidad = 94%
- *F1 Score* = 94%

Este modelo fue el que mejores métricas nos arrojó, siendo el más preciso y el que menos casos clasificó como *Hit* cuando no lo eran.

Podemos concluir que :

- Tenemos un 84.5% de acierto, tanto para clasificar las canciones en *Hits* o no *Hits*.

- Tenemos un 94% de acierto al identificar las canciones que realmente son *Hits*, es decir, solamente en el 6% de los casos se identificaron como que no eran *hits* cuando sí lo eran.

- Tenemos un 94% de acierto al combinar las métricas de precisión y sensibilidad.