# Análisis de Spotify

Pablo Toloza Castillo

# Introducción:

Este proyecto se origina desde una pequeña intriga, la cual se remonta a cuando empecé a conocer Spotify. Siempre he escuchado un tipo de musica más local por lo que me excusaba que la música que escuchaba no se encontraba en la aplicación, fue cuando me di cuenta de que, aunque una parte no estaba, otra parte muy grande se encontraba de una manera mucho más cómoda que como lo hacía antiguamente que era descargar el disco o la canción.

Pase por muchas aplicaciones desde Ares hasta youtube-to-MP4, ya que cada una tiene sus ventajas y desventajas y al no ser una fuente confiable donde conseguir descargar esta musica siempre me encontraba con errores o el cierre de estas aplicaciones. Poco después de llegar a la capital de mi país fue cuando un amigo el cual admiro mucho y siempre confió en mí para llegar a desarrollar los proyectos que puedo realizar actualmente, me facilitó una APK de Spotify (Omitiré información sobre hacking ético, etc.) la cual me funcionaba muy bien ya que no tenía las restricciones de un usuario no premium, podía cambiar de canción, buscar todos mis discos preferidos, crear playlist, guardar toda mi música en un lugar y descargar la musica que quisiera para escucharla sin conexión.

Para mí que vengo de un pueblo pequeño donde hace unos años recién llegó el primer "Burger King" y así con varias empresas internacionales que no tuve la oportunidad de conocer hasta llegar a la capital, poder acceder a todo este contenido para mi desconocido hasta ese momento era un mundo nuevo. Comencé a escuchar Podcast más en concreto uno de una persona con la cual me siento muy representada ya que igual es de mi ciudad natal por lo que hace varias alusiones que me recuerdan a mi familia y amigos, el cual se llama TVM o Tomás Va a Morir, este Podcast siempre me ha ayudado a pasar el tiempo y hasta encontrar lo que me llama la atención que es en lo cual estoy trabajando ahora mismo. La disyuntiva la cual mencioné en el principio fue cuando llegó el momento en que me robaron mi celular, había perdido todo ya que con la APK no se me guardaba en el Cloud de Spotify, las playlist que hice, los capítulos que ya había escuchado, hasta canciones las cuales hasta el día de hoy no he vuelto a encontrar.

En este momento me hice la pregunta de si realmente con la subscripción solo estoy pagando el servicio, donde llegue a la conclusión que, cómo mucha gente y donde nace el fin de este proyecto es si realmente los usuarios están dispuestos a aceptar una subscripción premium en una aplicación que es de uso gratuito, utilizando una serie de modelos de Machine Learning en los cuales respecto a sus metricas de desempeño definiré cual es el que cumple con lo esperado y sobre ese sacaré una pequeña conclusión.

# Análisis exploratorio:

Esta base de datos contiene datos de una encuesta realizado a varios usuarios, la cual contiene las siguientes columnas, con su respectiva definición.

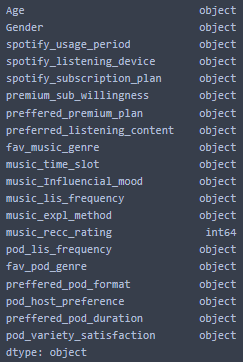
|  |  |
| --- | --- |
| Age | Rango de edad |
| Gender | Genero |
| Spotify\_usage\_period | Corresponde al rango de tiempo que el usuario lleva utilizando la aplicación. |
| Spotify\_subscription\_plan | Corresponde al plan de subscripción que está utilizando el usuario. |
| Premium\_sub\_willingness | Corresponde a sí el usuario esta dispuesto a contratar una subscripción en el futuro. |
| Preffered\_premium\_plan | Corresponde al plan de subscripción que tiene o estaría dispuesto a obtener el usuario. |
| Preferred\_listening\_content | Corresponde a qué prefiere escuchar el usuario. |
| Fav\_music\_gender | Corresponde al genero musical preferido por el usuario. |
| Music\_time\_slot | Corresponde al momento del día en el cual el usuario escucha musica más habitualmente. |
| Music\_influencial\_mood | Cuando el usuario escucha musica, ¿Cuál de los siguientes estados de ánimo o situaciones es mas influyente en su elección musical? |
| Music\_lis\_frequency | Donde o cuando el usuario escucha música más a menudo. |
| Music\_expl\_method | Como el usuario descubre nueva música en Spotify. |
| Music\_recc\_rating | Del 1 al 5 como puntuaría las recomendaciones de Spotify. |
| Prod\_lis\_frequency | Cuan seguido escucha Podcast. |
| Pod\_host\_preference | El usuario está mas inclinado a escuchar Podcast de personas desconocidas o prefiere de personas más conocidas. |
| Preffered\_pod\_format | Que formato de podcast el usuario prefiere. |
| Pod\_Host\_Preference | El usuario está más inclinado a escuchar episodios mas cortos (Menos de 30 min) o más largos (Más de 30 min) de Podcast. |
| Pod\_variety\_satisfaction | El usuario esta satisfecho con la variedad de Podcast disponibles en Spotify |

# En este punto me gustaría recalcar que mi vector objetivo será la variable “Premium\_Sub\_Willingness” ya que está directamente relacionada con mi enigmática.

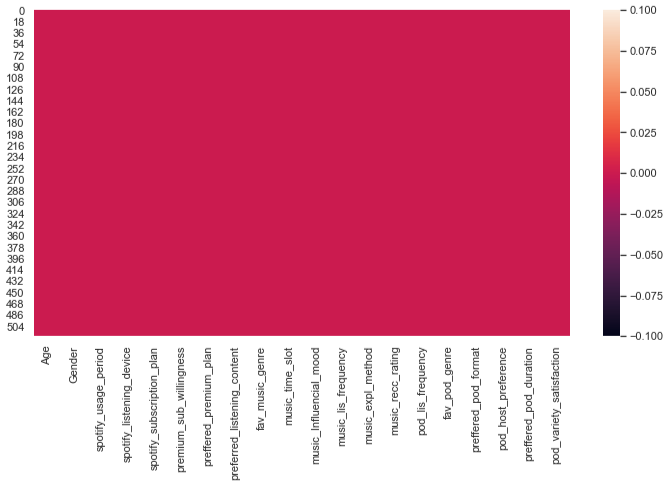
# Análisis de la BDD:

Para continuar generaré una serie de gráficos los cuales nos ayudaran a entender de una mejor manera el estado de nuestros datos:

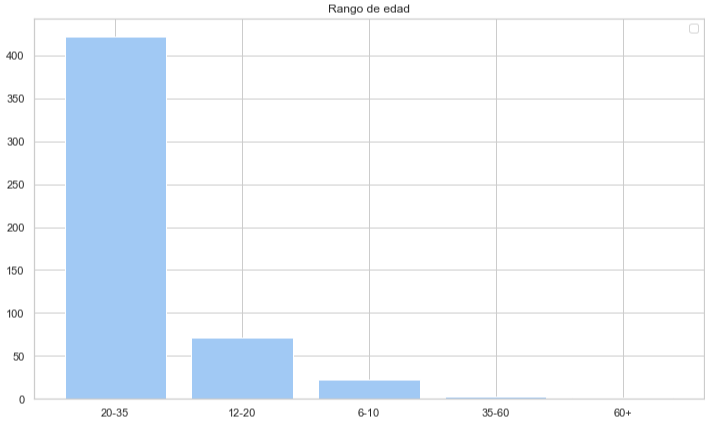
* **Primero verificamos los tipos de nuestras variables:**

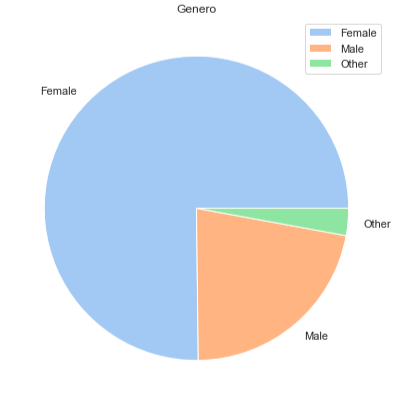


* **Datos perdidos:**



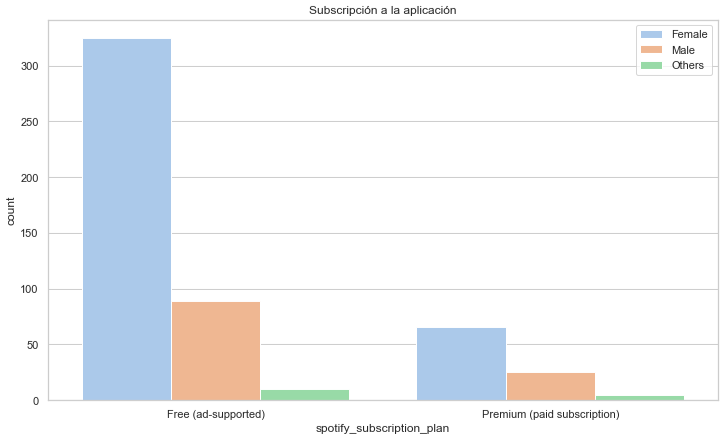
# Análisis de la BDD:

* **Columna “Age”:** 
* **Columna “Gender”:**

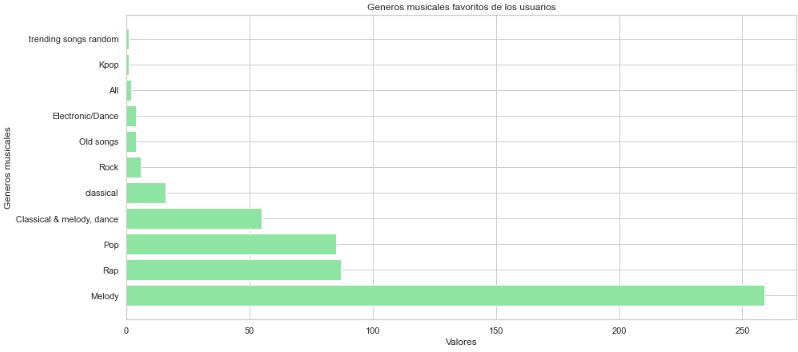


# - Análisis de la BDD:

* **Columna “Spotify\_subscription\_plan” organizado por “Gender”:**

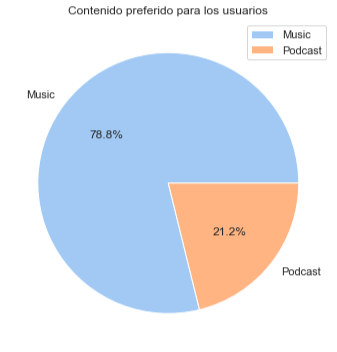


* **Columna “Fav\_music\_genre”:**

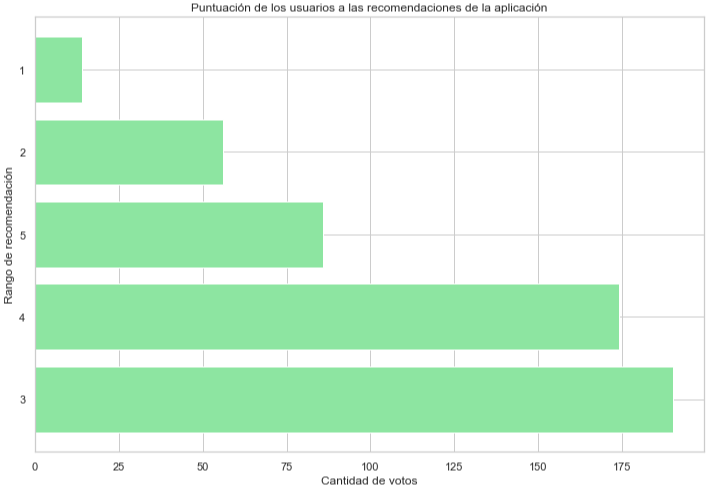


# - Análisis de la BDD:

* **Columna “Preferred\_listening\_content”:**

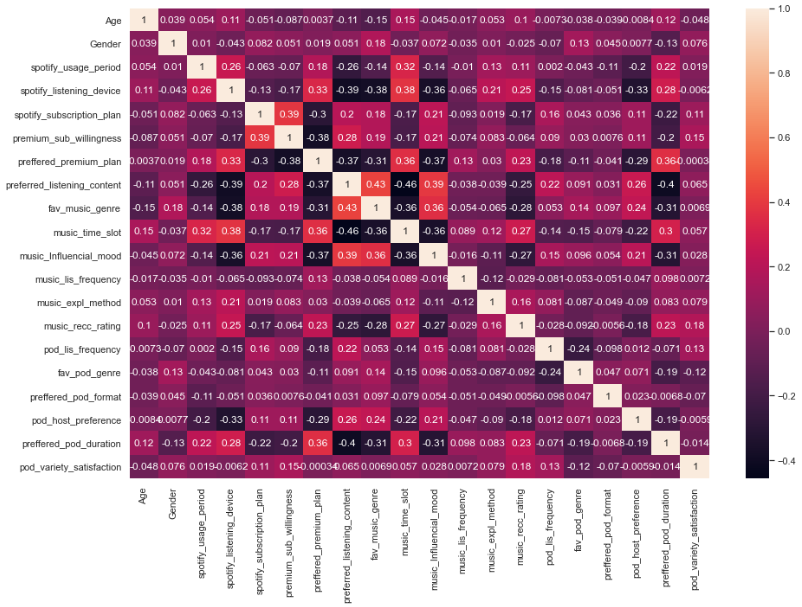


* **Columna “Music\_recc\_rating”:**



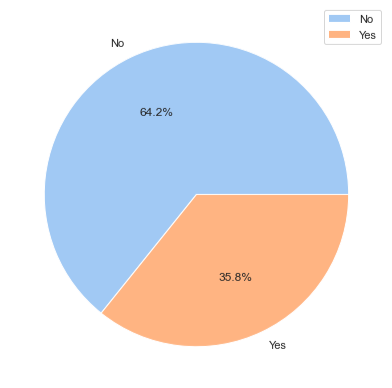
# - Análisis de la BDD:

* **Para continuar pasé la BDD a numérico para realizar un mapa de calor entre todas las variables:**

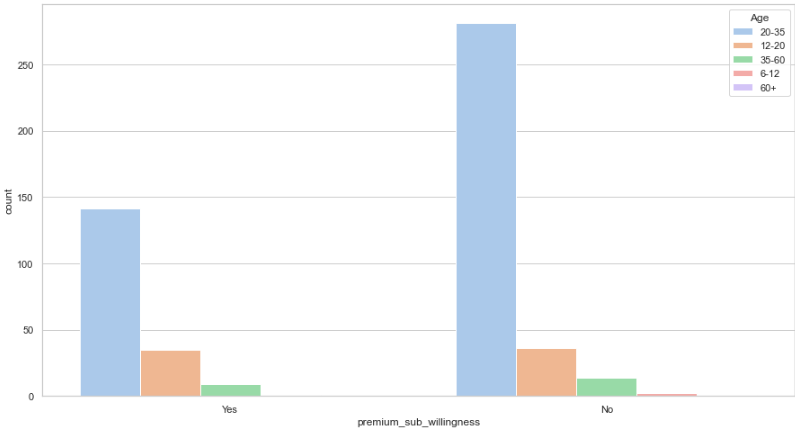


# Análisis del vector objetivo:

* **Primero graficamos la columna “Premium\_sub\_willingness”:**

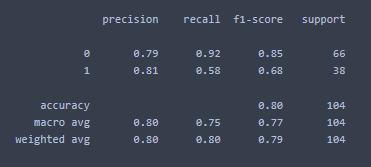


* **Ahora graficamos la columna “Premium\_sub\_willingness” por “Age”:**

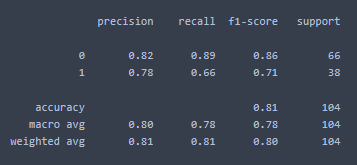


# Modelamiento y obtención de métricas:

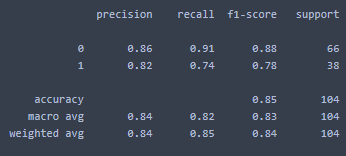
* **Modelo Linear Discriminant Analisys:**

****

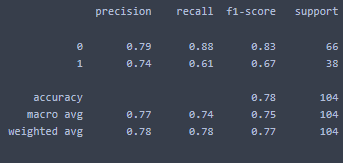
* **Modelo Gradient Boosting:**

****

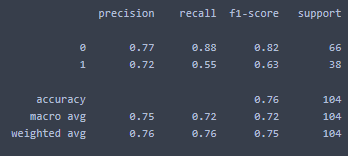
* **Modelo Random forest:**

****

* **Modelo Adaboost:**



* **Modelo Decission tree Classifier:**

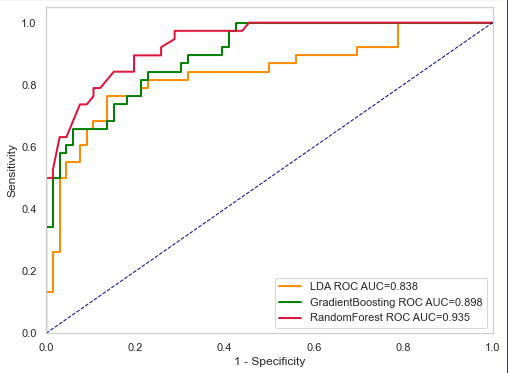


# Conclusiones

Para concluir este proyecto utilizaré mis 3 mejores modelos respecto a sus metricas de desempeño y generare una curva ROC, la cual nos muestra la relación entre los verdaderos positivos y la tasa de falsos positivos del modelo. Por lo tanto, mientras más hacia la izquierda superior esté la curva ROC, mejor será nuestro modelo.

Los modelos que utilizaré son:

* LinearDiscriminantAnalysis
* GradientBoostingClassifier
* RandomForestClassifier

A continuación, muestro la curva ROC:

En conclusión, nuestro modelo de Random Forest está un poco por sobre nuestros modelos de Linear Discriminant Analisys y Gradient Boosting.

Para concluir este proyecto quería enfatizar que quede muy conforme con los resultados, solo quedaría utilizar el modelo para un caso real para ver su variabilidad y asi inferir si funciona para una mayor cantidad de registros.