75.50 – Introducción a los Sistemas Inteligentes

Cátedra Ochoa

Trabajo Práctico Final

2do cuatrimestre 2018

Padrón	Nombre	Email
96626	María Florencia Prado	pradomflorencia@gmail.com
97404	Matías Rozanec	rozanecm@gmail.com

Facultad de Ingeniería
Universidad de Buenos Aires

Índice

Fase 1: Comprensión del negocio

Determinar los objetivos del negocio

Escenario actual

Al momento de escribir este informe, la empresa en cuestión tiene interés en desarrollar una app para la plataforma iOS, y requiere medir de alguna forma las probabilidades de éxito de la app en desarrollo.

Objetivos del negocio

El objetivo es en este caso poder determinar cuáles son las condiciones que una app debe cumplir para que la misma sea exitosa en el mercado AppStore de Apple.

Criterios de éxito del negocio

El proyecto se considerará exitoso si se llegan a detectar las variables clave que influyen en que una aplicación sea o no exitosa, y en qué nivel influye cada una de ellas.
Se considerará que una app es exitosa si la misma obtiene un rating de 4 puntos o más (estrellitas).

Evaluación de la situación

Inventario de recursos

Para desarrollar el presente proyecto se cuenta con una amplia gama de recursos que asegura un desarrollo de calidad y confianza del mismo. En primer lugar, se cuenta con un set de datos extraído de la API de iTunes Search, lo que asegura que se cuenta con una materia prima muy confiable. Además se cuenta con herramientas de software de análisis y visualización de datos líderes en el mercado, como lo son Pandas[1], matplotlib[2], seaborn[3] y sklearn[4] por nombrar solamente las principales. Por último, se cuenta con personal altamente calificado para la correcta interpretación de los mismos.

Requisitos, supuestos y restricciones

Requisitos

Contar con datos suficientes y sobre todo representativos del canon de apps.

[1] https://pandas.pydata.org/
[2] https://matplotlib.org/
[3] https://seaborn.pydata.org/
[4] https://scikit-learn.org/stable/

Supuestos

Los datos en estudio son lo suficientemente correctos como para poder sacar conclusiones confiables a partir de ellos.

Restricciones

Se cuenta solamente con datos del app store de Apple. No se asegura que los resultados sean válidos para otras plataformas.

Riesgos y contingencias

Si bien se puede llevar a cabo un análisis lo más riguroso posible, siempre cabe la posibilidad de que aún cumpliendo con todos los requisitos necesarios para que una app sea exitosa según el análisis, hay que tener en cuenta que cada app es un mundo, y si bien se cumplen patrones, no hay nada que asegure al 100% el éxito o no de una app.
En caso de detectar que la app no está teniendo el éxito esperado, habrá que recurrir a las diversas métricas que pueda ofrecer Apple a desarrolladores y analizar la situación desde ese punto de vista.

Terminología

Glosario de términos del negocio

Aplicación: es un programa informático diseñado como herramienta para permitir a un usuario realizar uno o diversos tipos de tareas.
Categoría: define el nombre de un grupo de aplicaciones con cualidades comunes. Tenemos como posibles categorías a: juegos, entretenimiento, educación, fotos y videos, utilities, salud y fitness, productividad, redes sociales, música, compras, comida, médicas.
User rating: puntuación promedio que los usuarios han otorgado a una aplicación.
Versión: el versionado de software es el proceso de asignación de un nombre, código o número único, a un software para indicar su nivel de desarrollo.
Aplicación exitosa: aquella cuyo user rating es mayor o igual a 4.
Dispositivos soportados: son los dispositivos electrónicos para los cuales la aplicación ha sido desarrollada. Esto implica que son aquellos en los que se puede ejecutar la misma.
Descripción de aplicación: breve texto que apunta a vender la aplicación junto con una pequeña definición de para qué y/o quienes ha sido desarrollada.

Glosario de términos de la minería de datos

Atributo: dato sobre alguna característica de las observaciones.
Atributo relevante: atributo que juega un papel principal en la clasificación, por lo que la clase dependerá en alguna medida de qué valor tenga.
Registro: fila que representa una observación, está compuesto de atributos.
Dataset: conjunto de datos a ser utilizados para la ejecución de los algoritmos de Data Mining, está compuesto de registros.
Filtrado de atributos: específica al dataset formado considerando sólo los atributos relevantes.

Regla: es una implicación, que representa una acción mediante una condición. Sigue la estructura “Si..., entonces...”.
Soporte: es la relación entre la cantidad total de registros del dataset que cumplen la regla y la cantidad de observaciones procesadas.
Confianza: es la relación entre la cantidad total de observaciones de la clase mayoritaria que cumplen la regla y la cantidad de observaciones que fueron afectadas por esa misma regla.
Captura: es la relación entre la cantidad de observaciones de la clase mayoritaria que cumplen la regla y la cantidad de observaciones procesadas pertenecientes a esa misma clase.
Coeficiente de correlación de Pearson: es la estadística de prueba que mide la relación estadística, o asociación, entre dos variables continuas. Es conocido como el mejor método para medir la asociación entre variables de interés porque se basa en el método de covarianza. Da información sobre la magnitud de la asociación, o correlación, así como la dirección de la relación

Costos y beneficios

El beneficio del proyecto es evidente: detectar las características que hacen a una app exitosa, todo el proceso dedicado a pensar el desarrollo de apps puede ser aprovechado muchísimo mejor, dado que se conocen las exigencias que habrá que cumplir para maximizar la probabilidad de éxito.
Al ser un trabajo final educativo, no hay costos.

Determinar los objetivos de la minería de datos

Objetivo de minería de datos

El objetivo de minería de datos es el análisis de los datos obtenidos a partir de la información disponible, buscando obtener así información relevante que permita predecir las condiciones bajo las cuales una aplicación es exitosa.

Criterios de éxito de minería de datos

Selección de al menos 4 reglas con soporte mayor o igual al 20%.

Realizar el plan del proyecto

Plan de proyecto

- Recolección de datos: 5 horas.
- Preparación de datos: 20 horas.
- Ejecución del algoritmo de Inducción: 3 horas.
- Análisis de resultados de algoritmo de Inducción: 3 horas.
- Combinación de resultados: 4 horas.
- Elaboración de reporte: 12 horas.

Validación inicial de las herramientas

Para el desarrollo del proyecto se utilizarán las siguientes herramientas:

- Pandas
- Matplotlib
- Seaborn
- Sklearn
- Jupiter Notebook
- Python

Fase 2: Comprensión de los datos

Recolección de los datos iniciales

Los datos utilizados durante el transcurso del proyecto fueron obtenidos gratuitamente en el sitio de Kaggle y el mismo se encuentra en formato csv.

Reporte de la recolección de datos iniciales

Con millones de aplicaciones en la actualidad, el conjunto de datos se ha convertido en la clave para obtener las mejores aplicaciones en la tienda de aplicaciones iOS. Este conjunto de datos contiene más de 7000 detalles de aplicaciones móviles de Apple iOS. Los datos se extrajeron de la API de búsqueda de iTunes en el sitio web de Apple Inc.

Fecha de recolección de datos (de API); Julio 2017

Descubrir datos

Reporte de descripción de datos

Se cuenta con dos datasets: por un lado se tienen datos estadísticos y generales de las apps descritos en 16 columnas; por otro lado hay un dataset con las descripciones de las apps. A continuación, los nombres de los features con su correspondiente descripción y tipo de dato.

Nombre variable	Tipo de dato	Descripción
id	discreta	indica el id de aplicación. Cada aplicación tiene un id único e intransferible
track_name	alfanumérica	el nombre de la aplicación
size_bytes	discreta	tamaño de la aplicación expresado en bytes
currency	alfanumérica	indica la moneda en que está expresado el precio
price	numérica	precio de la aplicación
rating_count_tot	discreta	cantidad total de usuarios que puntuaron la aplicación a lo largo de todas sus versiones
rating_count_ver	discreta	cantidad total de usuarios que puntuaron la aplicación en su última versión
user_rating	numérica	puntuación de los usuarios

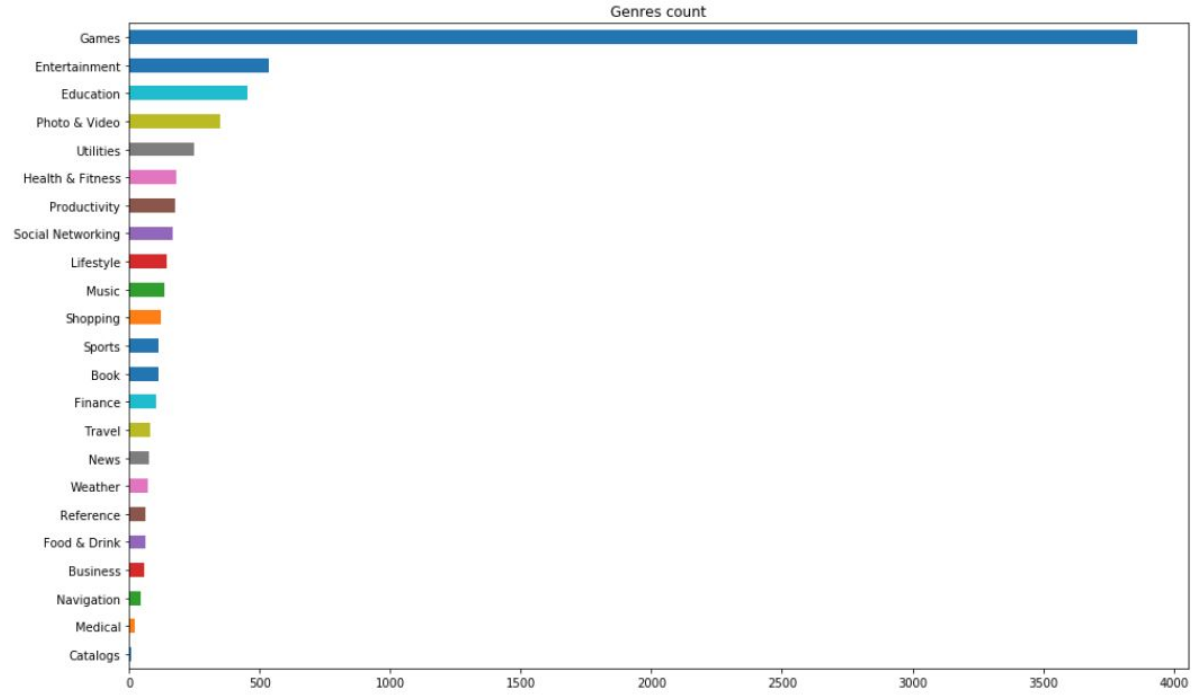
user_rating_ver	numérica	puntuación de los usuarios tomando solamente las puntuaciones referentes a la última versión
ver	alfanumérica	versión de la aplicación
cont_rating	alfanumérica	calificación del contenido. Indica si es apto para para mayores de 4, 7, 12 o 17 años
prime_genre	alfanumérica	género principal de la app
sup_devices.num	discreta	cantidad de dispositivos soportados
ipadSc_urls.num	discreta	cantidad de screenshots disponibles para mostrar en el store
lang.num	discreta	cantidad de idiomas soportados
vpp_lic	discreta	indica si la app soporta licencia vpp[5]
description	alfanumérica	breve descripción de la app que se le proporciona a los usuarios en el Store

[5] *The Apple Volume Purchase Program (VPP) is a service that allows organizations that have registered for the Apple VPP to purchase iOS apps in bulk, but not at discounted prices.*
After making a bulk purchase, the organization receives redemption codes for each app bought. The organization can then distribute app codes to individual users, who use the codes to "purchase" the app from the Apple App Store.
source:
https://searchmobilecomputing.techtarget.com/definition/Apple-Volume-Purchase-Program-Apple-VPP

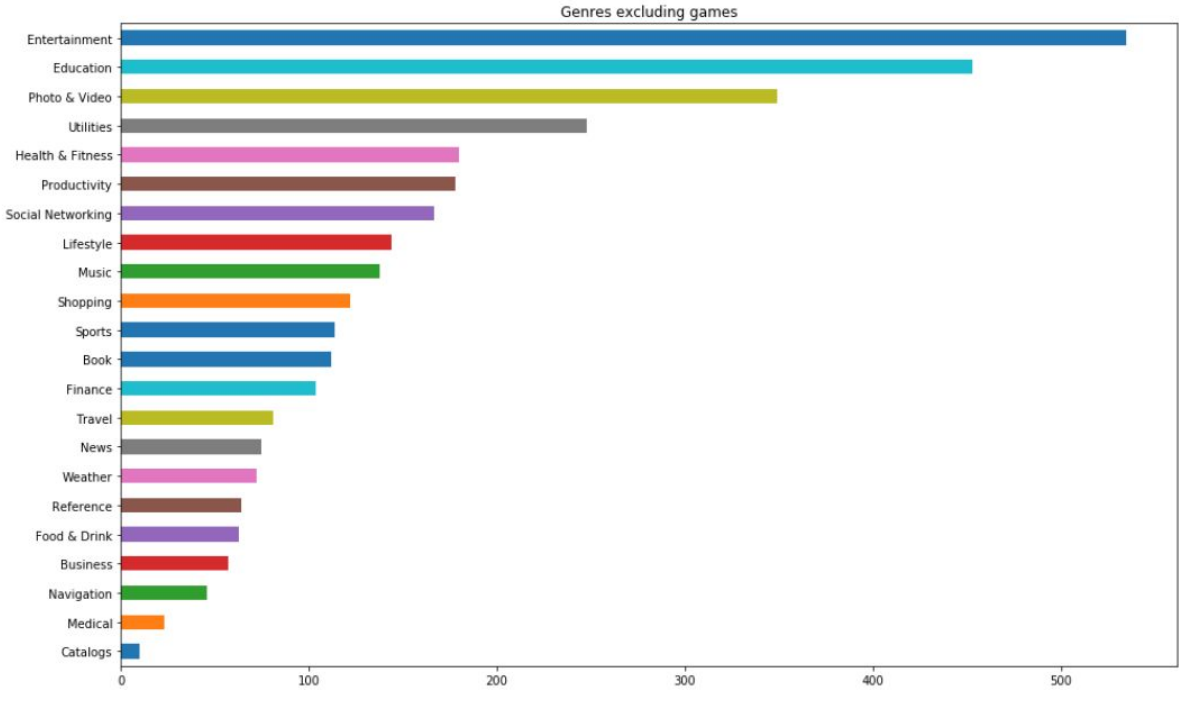
Exploración de los datos

Reporte de exploración de datos

Análisis de categorías



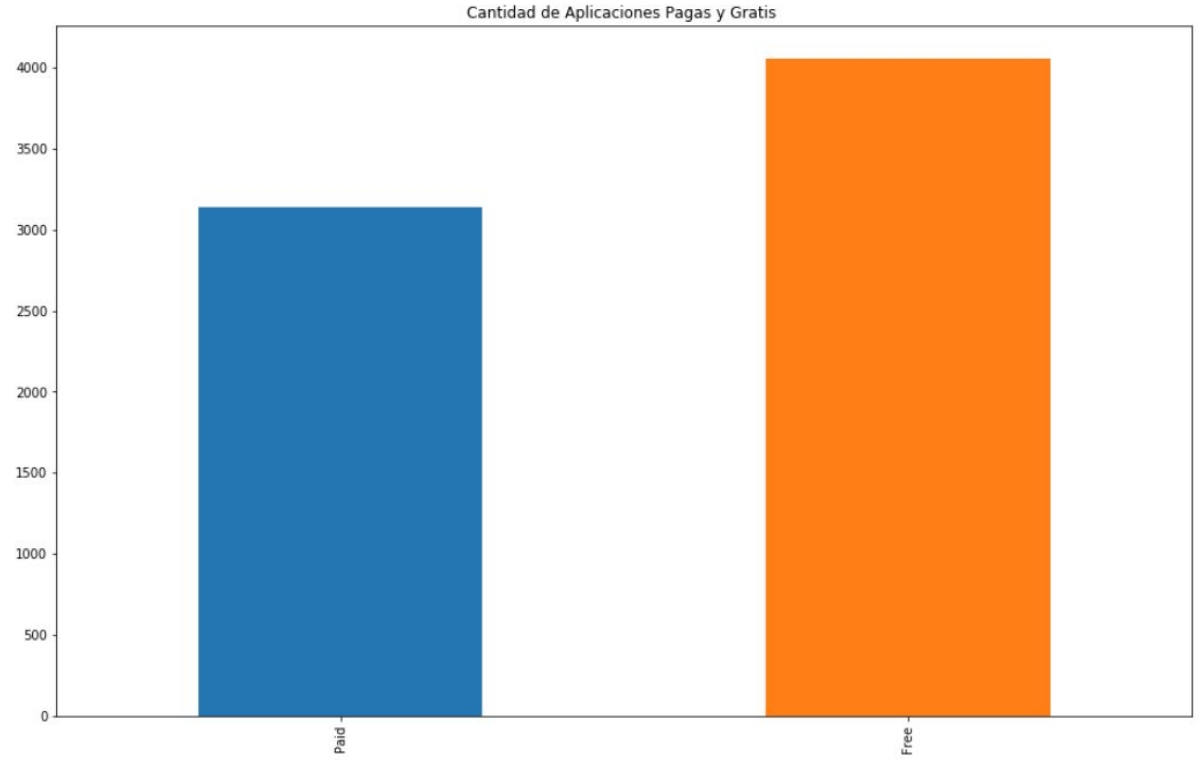
Cantidad de aplicaciones por categoría



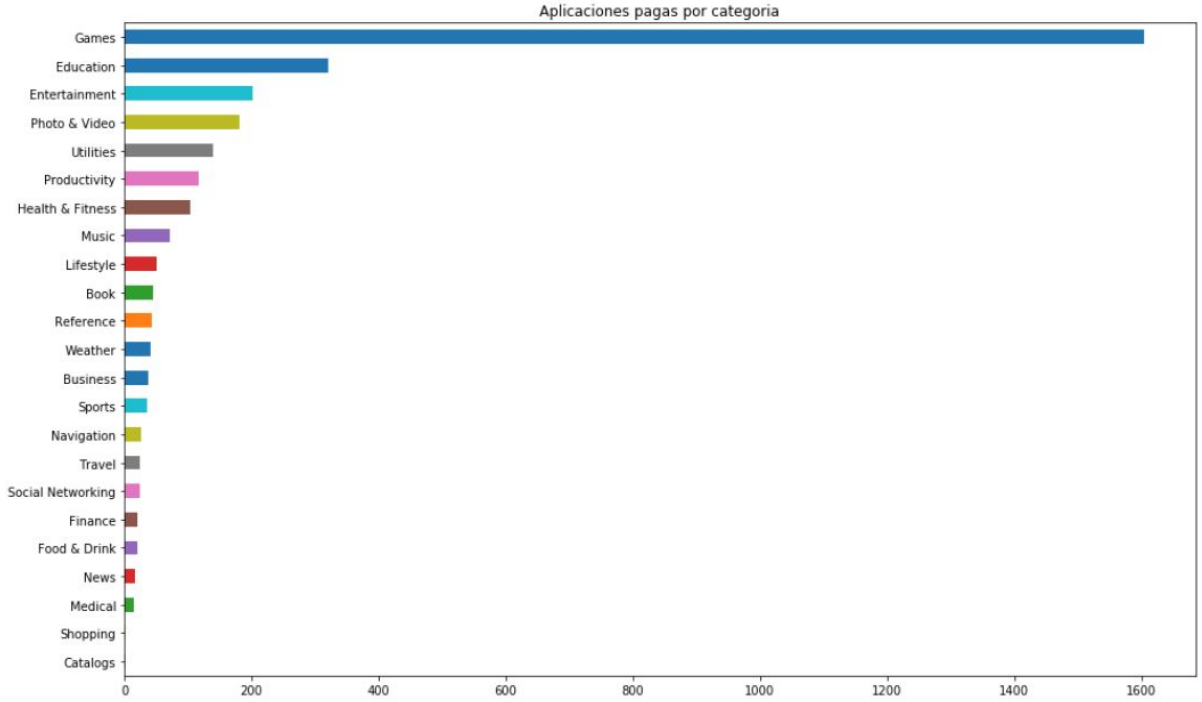
Cantidad de aplicaciones por categoría excluyendo “Games”

Análisis de precios

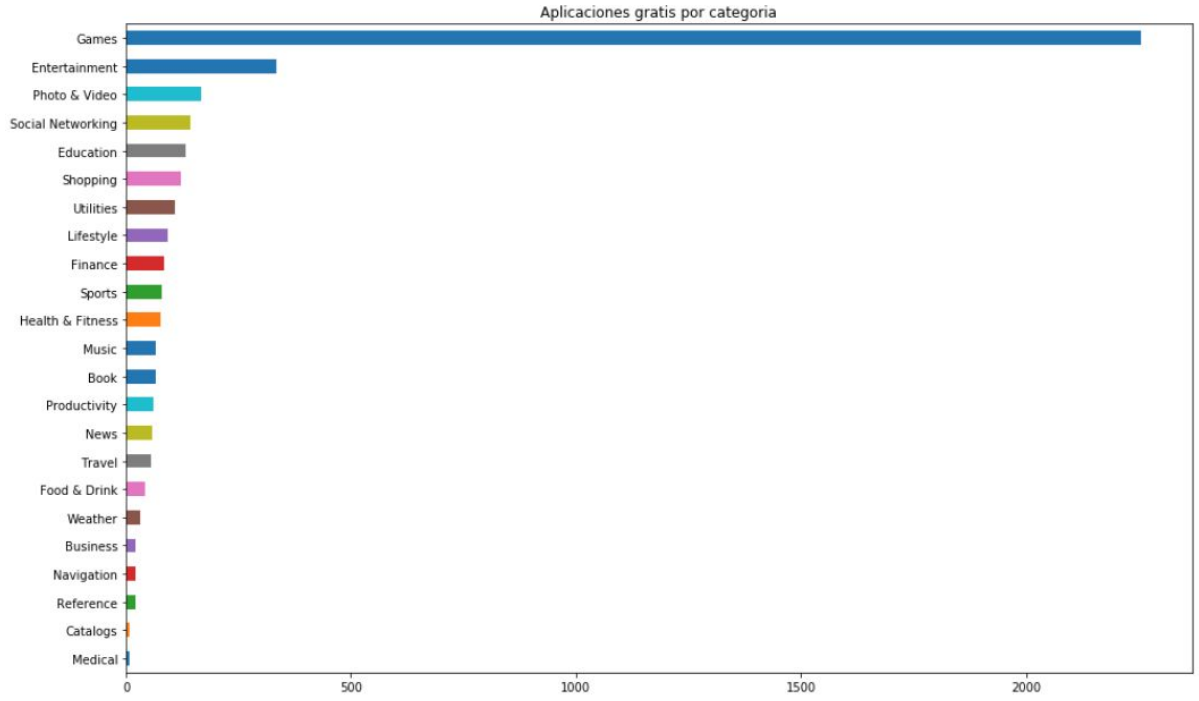
Todos los precios están fijados en dólares, por lo que la columna “Currency” es irrelevante.



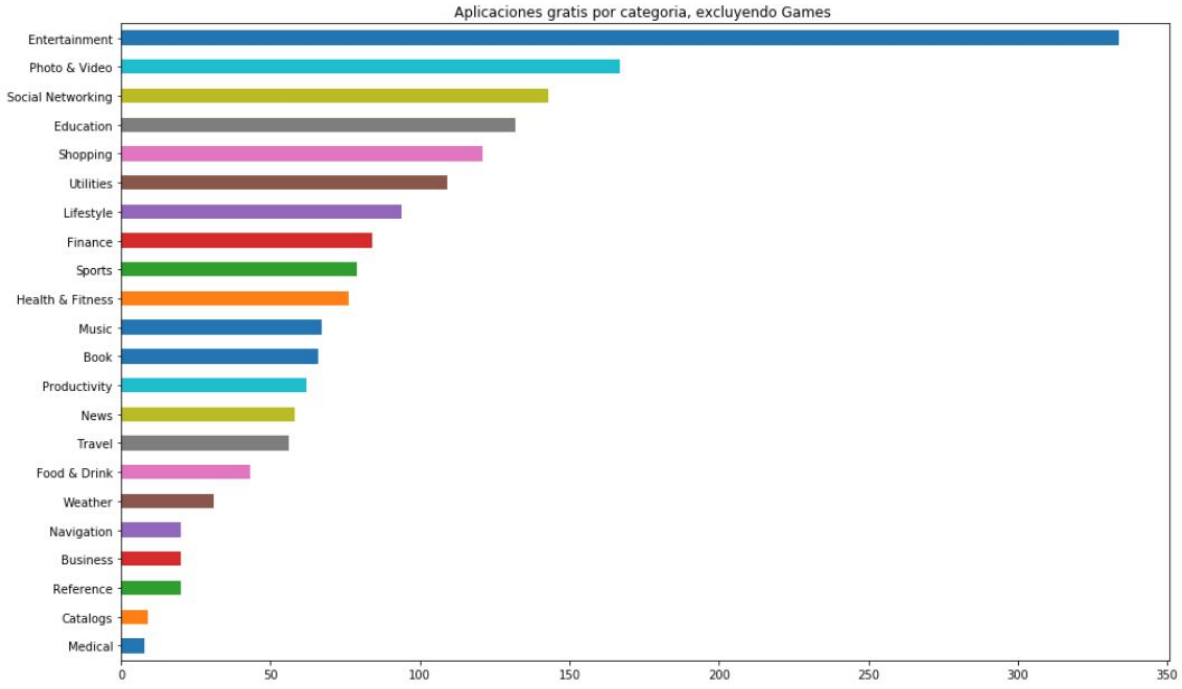
Cantidad de aplicaciones pagas vs. gratuitas



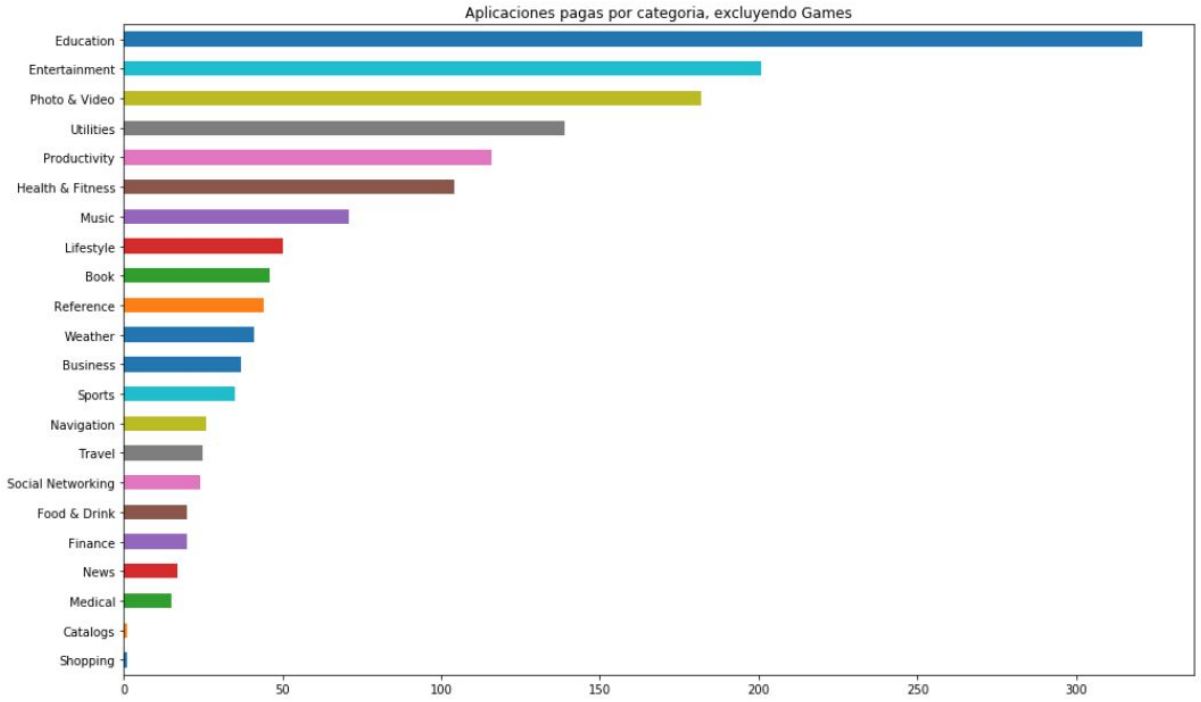
Cantidad de aplicaciones pagas de acuerdo a cada categoría.



Cantidad de aplicaciones gratis de acuerdo a cada categoría

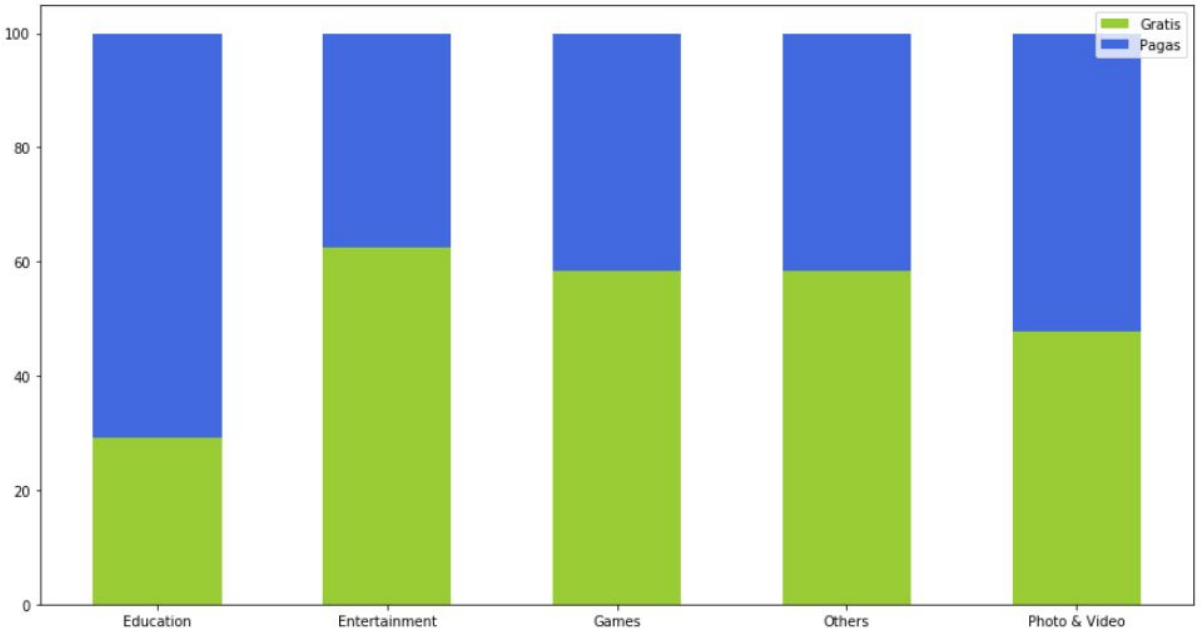


Cantidad de aplicaciones gratuitas de acuerdo a cada categoría, excluyendo “Games”

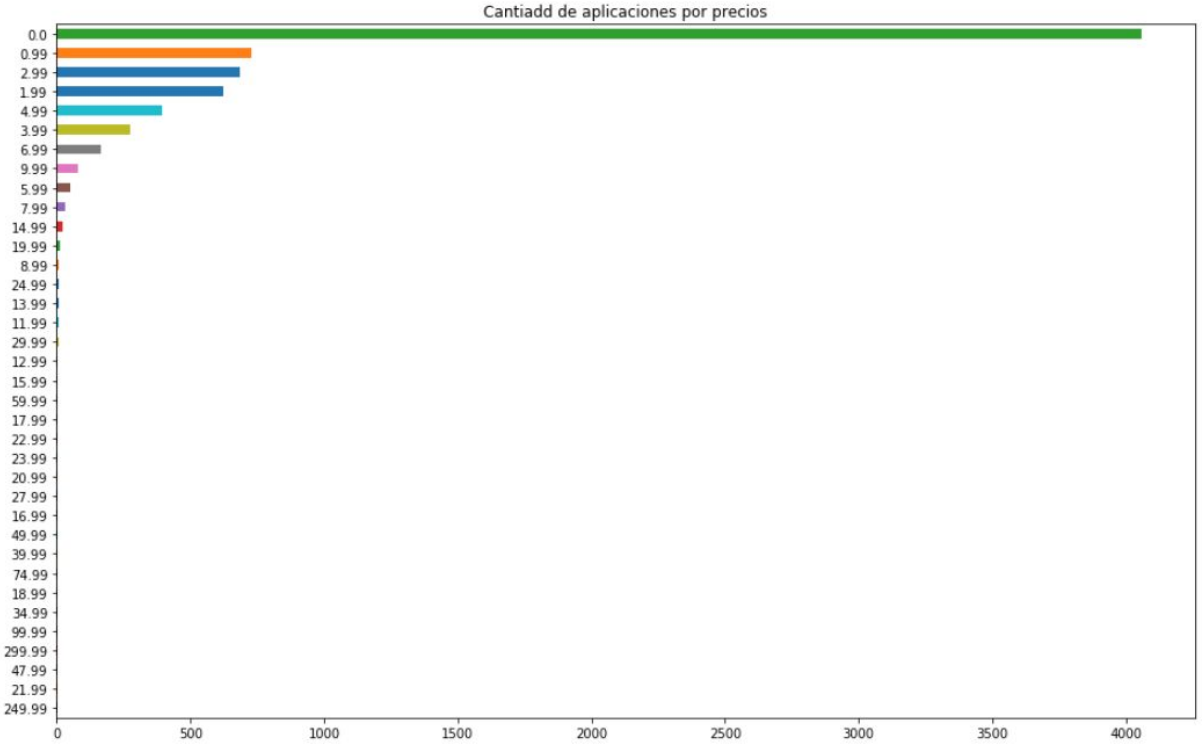


Cantidad de aplicaciones pagas de acuerdo a cada categoría, excluyendo “Games”

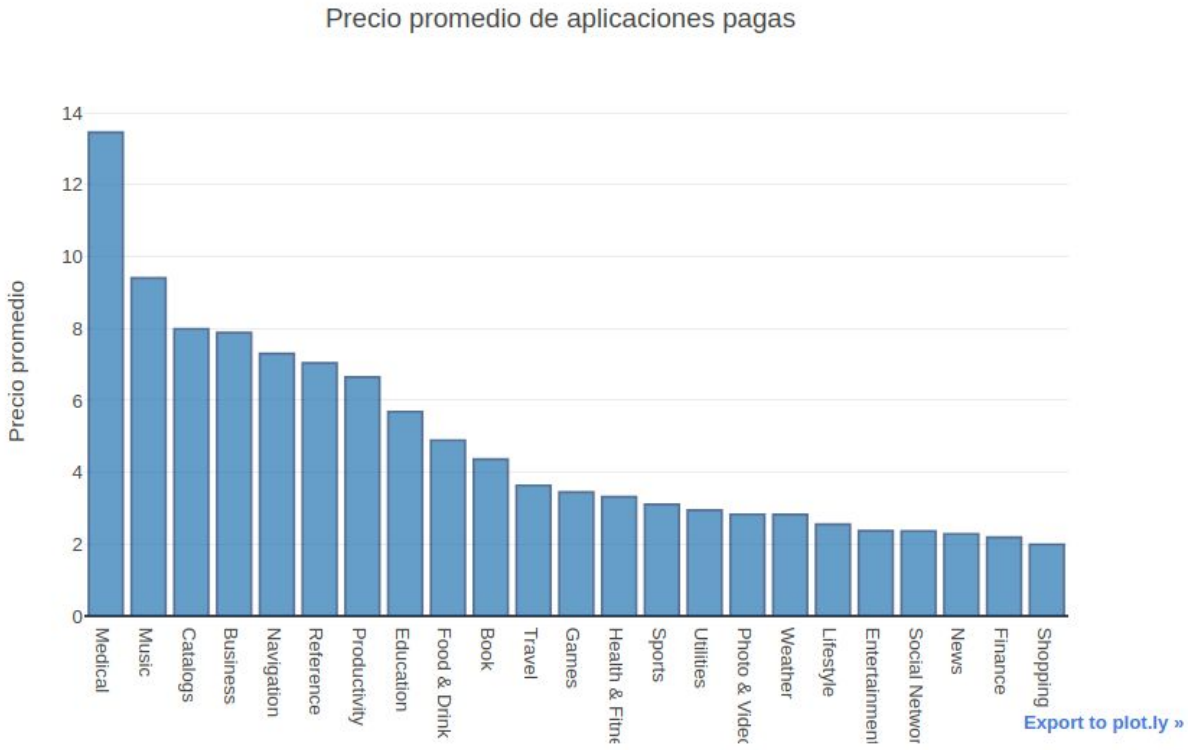
	Gratis	Pagas	Total	Porcentaje Pagas %	Porcentaje Gratis %
Education	132	321	453	70.860927	29.139073
Entertainment	334	201	535	37.570093	62.429907
Games	2257	1605	3862	41.558778	58.441222
Others	1166	832	1998	41.641642	58.358358
Photo & Video	167	182	349	52.148997	47.851003



Vemos que para la única categoría que la cantidad de aplicaciones pagas supera a la cantidad de aplicaciones no pagas es Eduación



Como se puede ver, la gran mayoría de apps son gratis; de las pagas los precios van hasta los USD 10, a partir de ahí ya hay muy pocos casos como para generalizar.



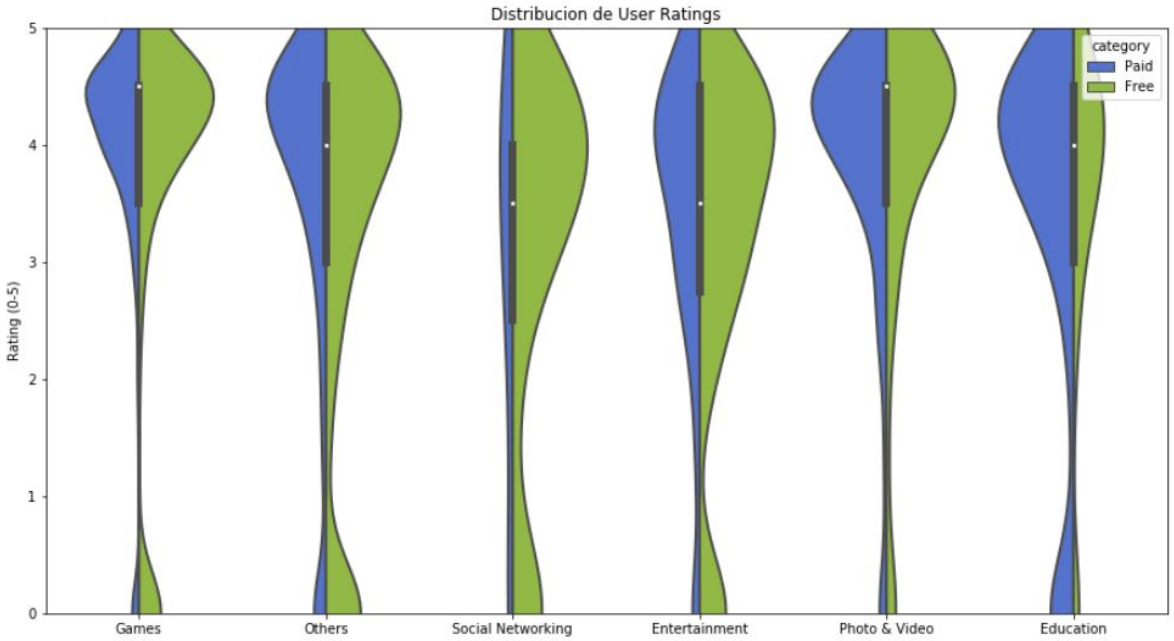
Dentro de las aplicaciones pagas, las aplicaciones Médicas son las más caras, con un promedio de casi 14 dólares. Y si analizamos aquellas categorías con mayor cantidad de aplicaciones en

el store mantienen precios menores a 6 dólares en promedio, como lo son Games, Entretainment, Education y Photo & Video.

Análisis de User Ratings

Una vez analizadas la cantidad de aplicaciones pagas y gratis por categoría nos preguntamos si las aplicaciones pagas son realmente buenas. Esto lo vamos a analizar siguiendo la opinión de los usuarios.

El campo que define la opinión de los usuarios es USER RATING, que tiene valores entre 0 y 5, siendo 5 la mejor puntuación.



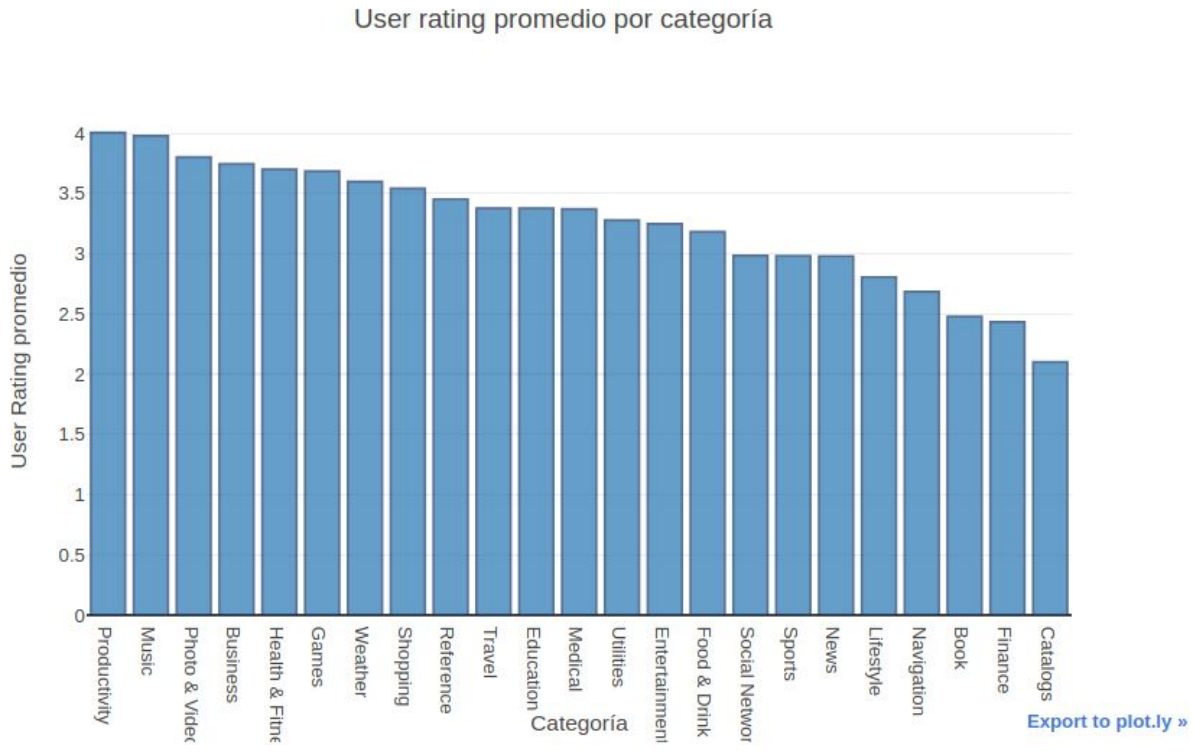
Los resultados varían parecido en el caso de las apps pagas y de las gratuitas. En general vemos que la distribución de las puntuaciones de usuarios son parecidas en todas las categorías.

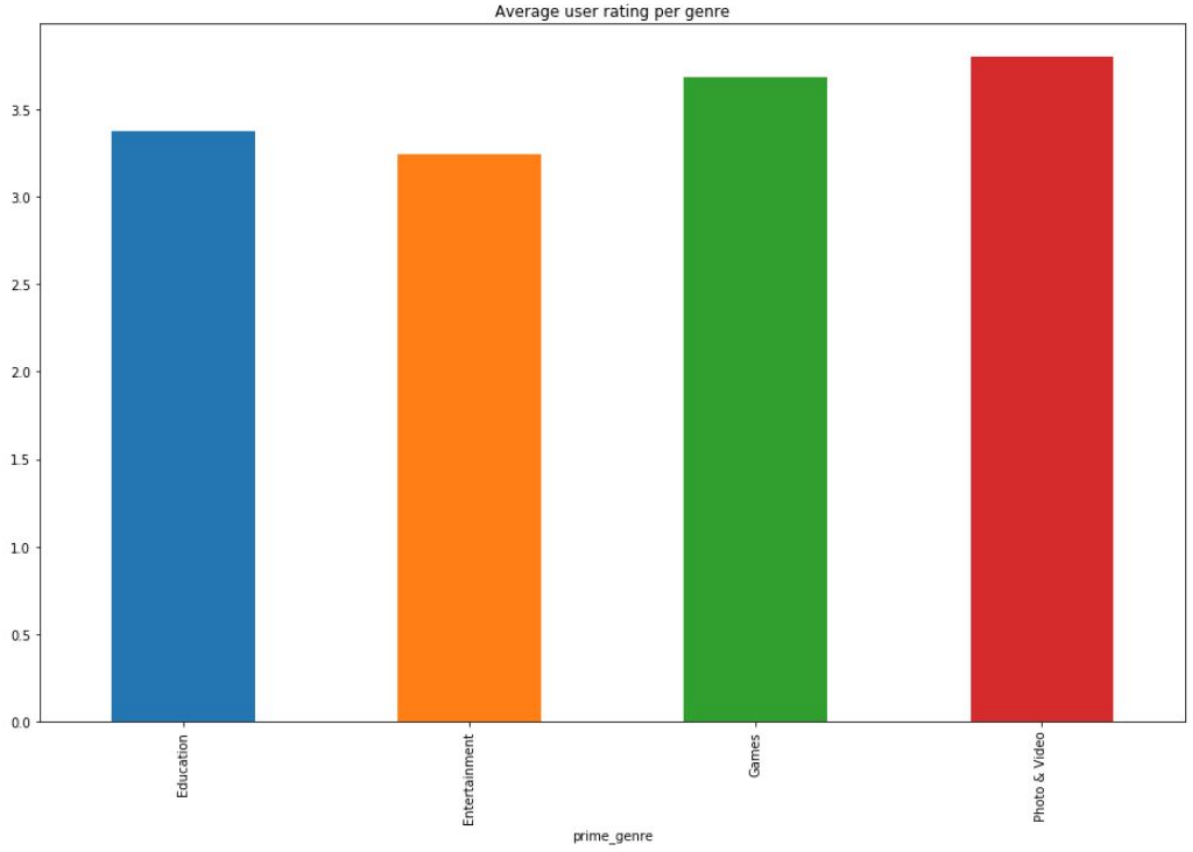
Se puede destacar que para el caso de entretenimiento y otros sí se aprecian puntuaciones entre 1 y 2, mientras que en el resto de las categorías las puntuaciones más presentes son 0 y luego entre 3 y 5, generalmente los valores entre 1 y 2 son poco usados. De todos modos las distribuciones son parecidas tanto para pagas como no pagas.

Se puede notar que en las apps pagas, en el género *Education* hay mayor concentración en la puntuación 4 para las pagas, mientras que para las no pagas, entre los valores 3 y 5, la distribución es más pareja.

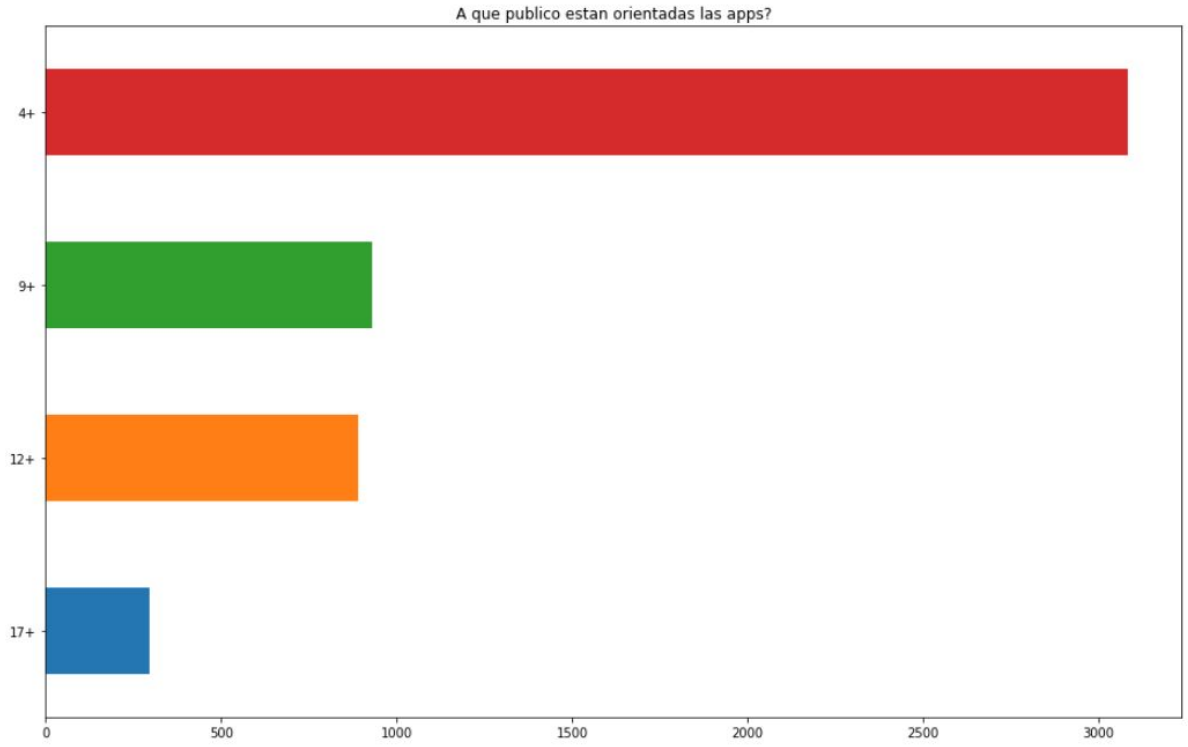
La mayoría de apps relacionadas a Redes sociales por amplia diferencia es gratis. No sería popular tener que pagar por una red social; además las redes sociales generan ingresos mediante publicidades, por lo que conviene tener una base amplia de usuarios, por más que eso signifique que ellos estén 'gratis'.

No perder de vista que todas las apps también pueden generar ingresos mediante *in app purchases*.



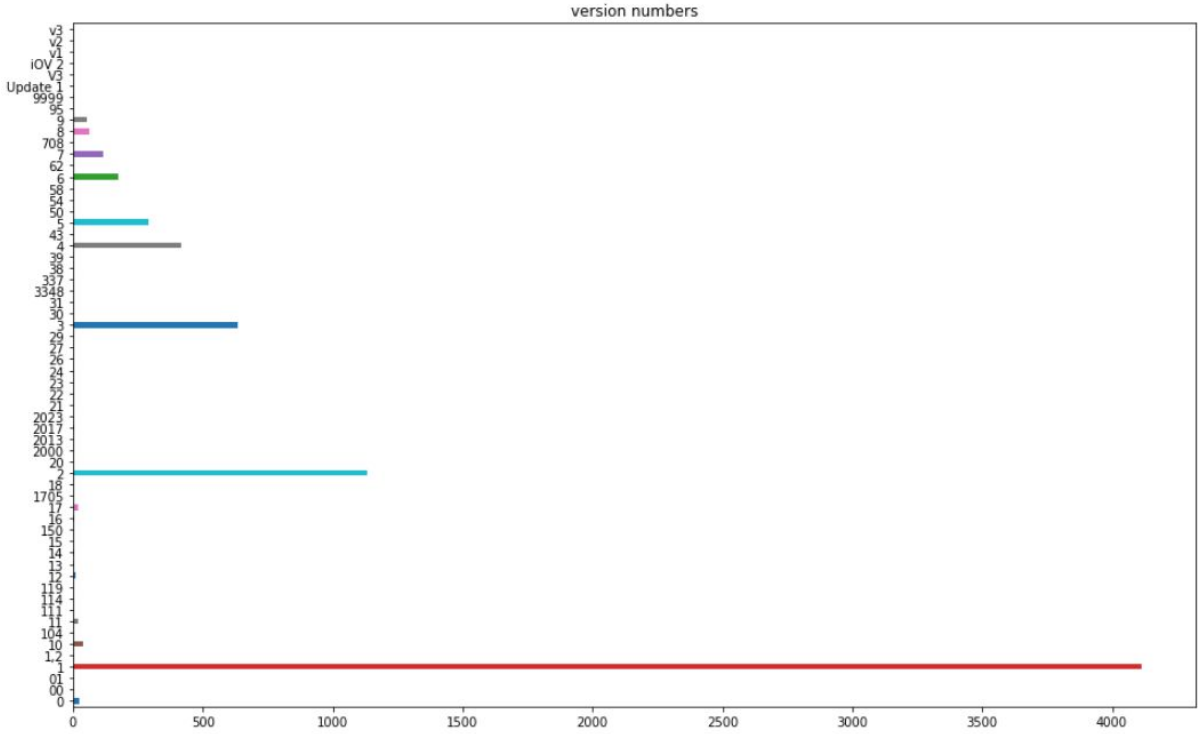


Análisis de Cont Rating



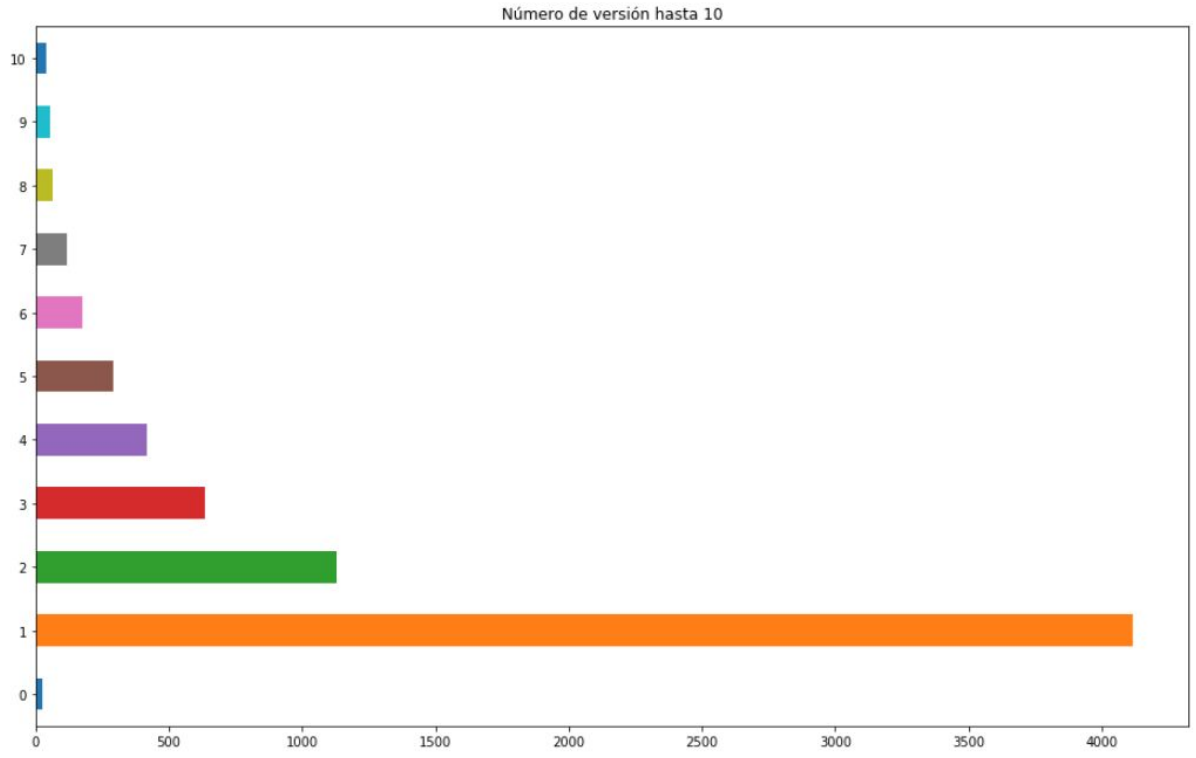
Análisis de versión

Nos quedamos con la *major version* de las apps.



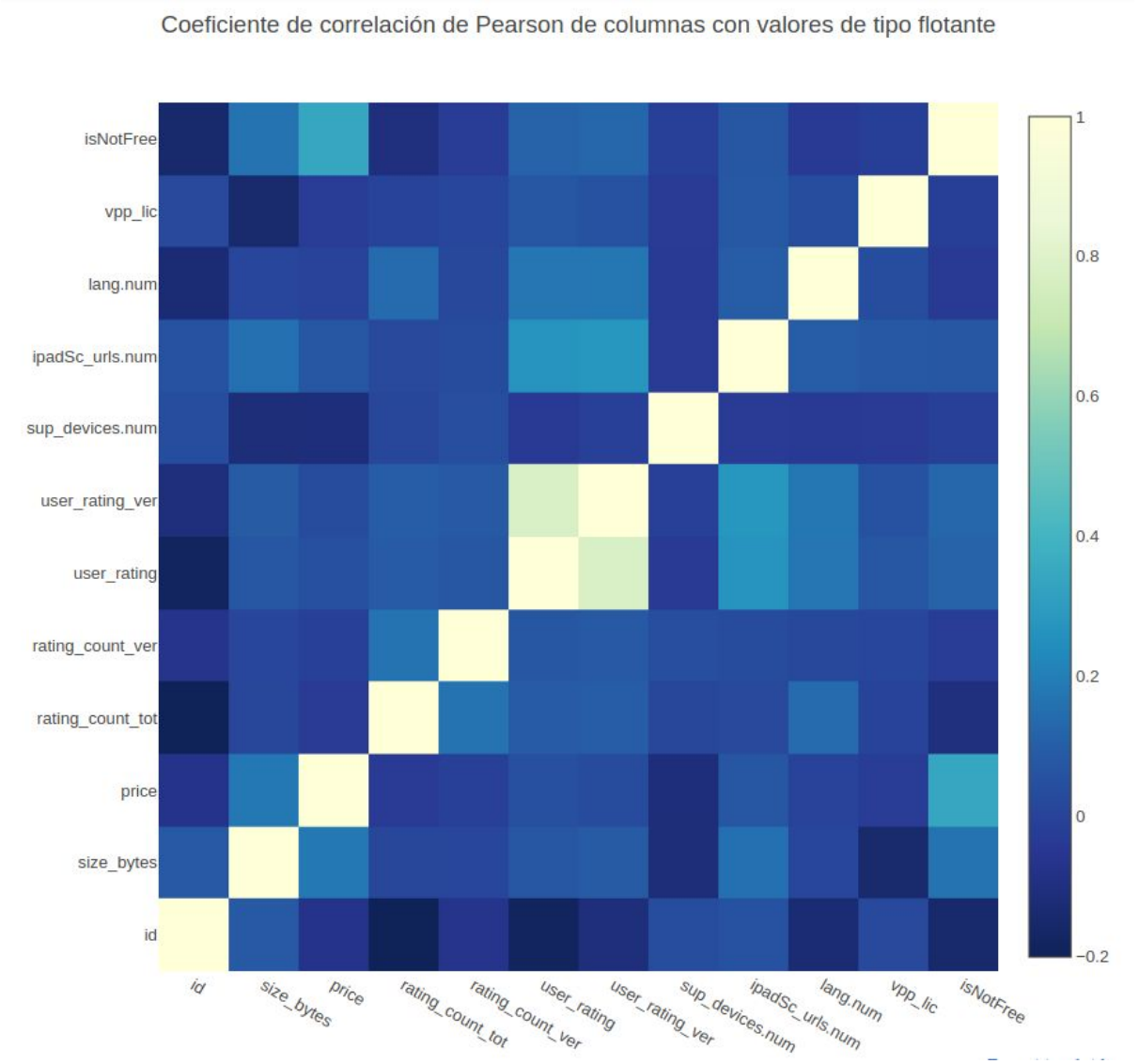
Se observan versiones un poco raras.... como `1,2` o `iOV2`, `9999`, `2000`.... no es posible que esto sean números de versiones reales.

Más allá de eso, la gran mayoria esta en su primera versión, llegando hasta la 9. Más allá de la 9



Correlación de features

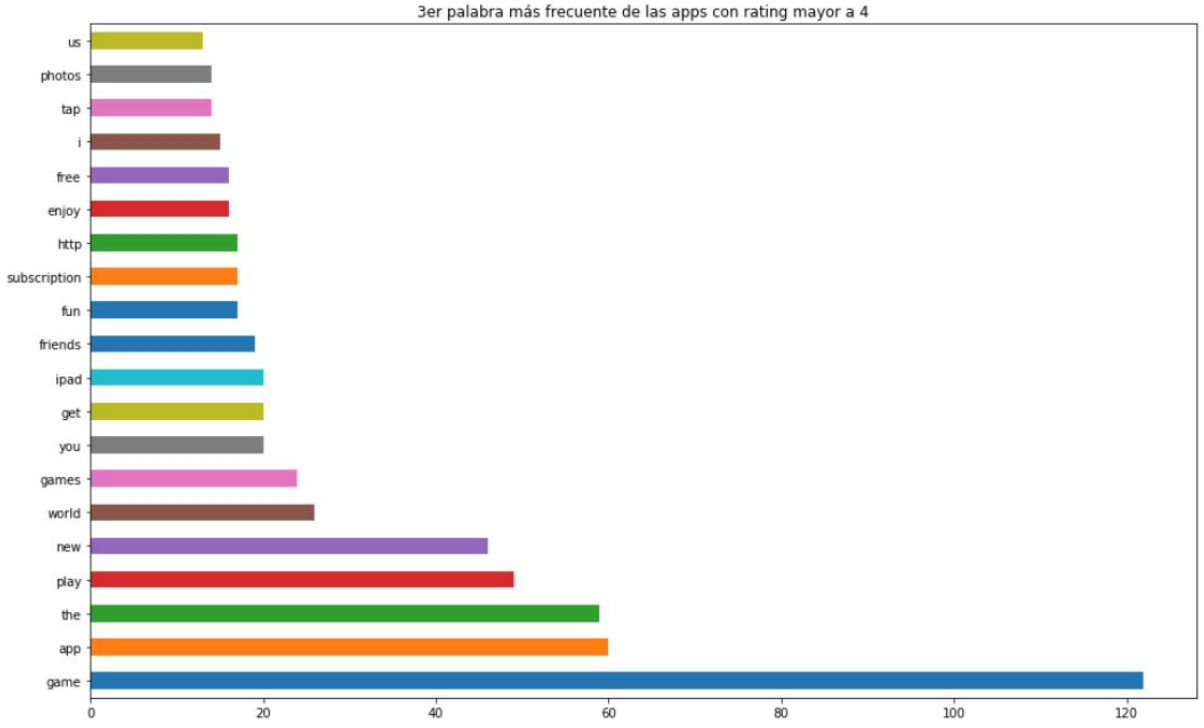
Luego de mergear los dos set de datos analizamos la correlación



El coeficiente de correlación de Pearson es la estadística de prueba que mide la relación estadística, o asociación, entre dos variables continuas. Es conocido como el mejor método para medir la asociación entre variables de interés porque se basa en el método de covarianza. Da información sobre la magnitud de la asociación, o correlación, así como la dirección de la relación.

Análisis de descripción de las aplicaciones

Se agregan tres columnas con las tres palabras más frecuentes en la descripción de las apps, omitiendo stopwords.



Verificación de la calidad de los datos

Reporte de calidad de datos

En su gran mayoría los datos están completos y son coherentes.
Una excepción a esto se da con una cantidad muy reducida de registros que no cuentan con información de rating.
Por otro lado, el campo que indica la versión de la app tiene una gran cantidad de casos anómalos[6], sin embargo esto no impide una generalización para entender los datos.

[6] Se registran casos con nros. de versión por arriba de 1000, o strings que no cumplen con el formato esperado para indicar la versión de una app.

Fase 3: Preparación de los datos

Selección de los datos

Los datos que se terminaron utilizando fueron todos los de tipo numérico y discreto, principalmente porque el algoritmo utilizado trabaja con este tipo de datos. Las variables categóricas podrían incluirse también, pero habría que mapearlas a valores numéricos previamente, cosa que excede esta materia y el presente proyecto.

Limpiar los datos

Como se mencionó anteriormente, el campo rating no estaba completo en una cantidad pequeña de registros: se decidió eliminar dichos registros del dataset.
Dados los casos anómalos en el campo de versión, este campo sirvió para un entendimiento general de los datos, pero no se incluyó finalmente en el algoritmo.
Se eliminó la columna “Currency” dado a que todos los registros tenían el mismo valor, USD.

Estructurar los datos

Derivación de atributos

- Se generó un campo indicando la categoría, pero limitando los valores posibles a las tres principales categorías, indicando como *Otros* en caso de cualquier otra categoría. Esto permitió un mejor entendimiento de los datos.
- Se creó un campo derivado del precio indicando si la app es paga o no para contar con esta información independientemente del precio que pueda llegar a tener una app paga.
- Se agregó un campo indicando la *major version* de la app, que no es más que un split del campo *version* original que guarda el mismo campo hasta el primer punto.
- Se ha analizado la frecuencia de las palabras en las descripciones: por cada app se agregaron 3 features indicando las 3 palabras más frecuentes respectivamente.
- Se generó un campo para registrar la cantidad de puntuaciones existentes hasta la versión actual, o sea, a la cantidad total se le sustrajo la cantidad correspondiente a la versión actual.
- Finalmente, se agregó el feature de clasificación, indicando True en caso de que el rating registrado sea mayor o igual a 4.

Generación de registros

No hubo necesidad de generar registros.

Integración de los datos

Se unieron las datos de los dos datasets en uno solo, contando de esta forma con un solo set conteniendo tanto la información general y estadística de las apps como sus descripciones.

Formato de los datos

Para el análisis de las descripciones se han pasado todas a *lowercase* y se han omitido las llamadas *stopwords* para obtener un análisis más significativo.
En cuanto al orden de los features, no hubo necesidad de aplicar ningún orden en particular, ya que los algoritmos usados no lo requieren.

Fase 4: Modelado

Selección de una técnica de modelado

Técnica de modelado

Para el modelado se ha decidido usar árboles de decisión.

Supuestos de modelado

Todos los datos usados para el modelado son numéricos.

Generación de diseño de las pruebas

Por un lado se tomaron todos los features numéricos, dejando de lado el resto. Se tomó aparte el feature de clasificación, esto es, el que indica si una app es exitosa o no.

Construcción del modelo

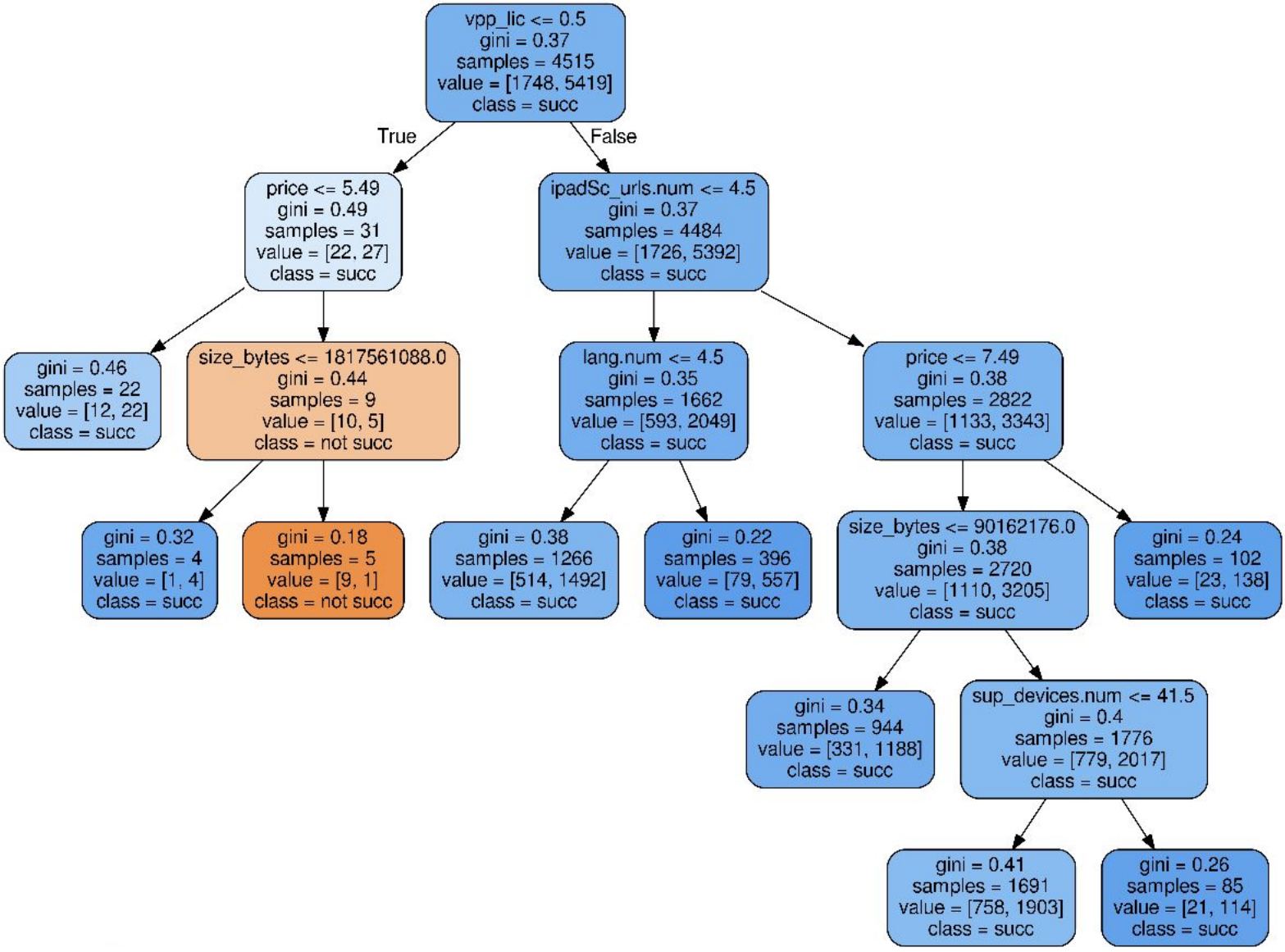
Configuración de parámetros

Para obtener un árbol legible y que sea significativo, sin que describa absolutamente cada caso por separado, hubo que ajustar la profundidad máxima del árbol a 4. Se ha probado ajustar también la cantidad mínima de muestras para hacer el split y la cantidad mínima de muestras que debe quedar en un nodo hoja, pero finalmente se comprobó que con solo ajustar la profundidad máxima se obtenía el resultado óptimo.

Evaluación del modelo

Evaluación del modelo

Árbol 1

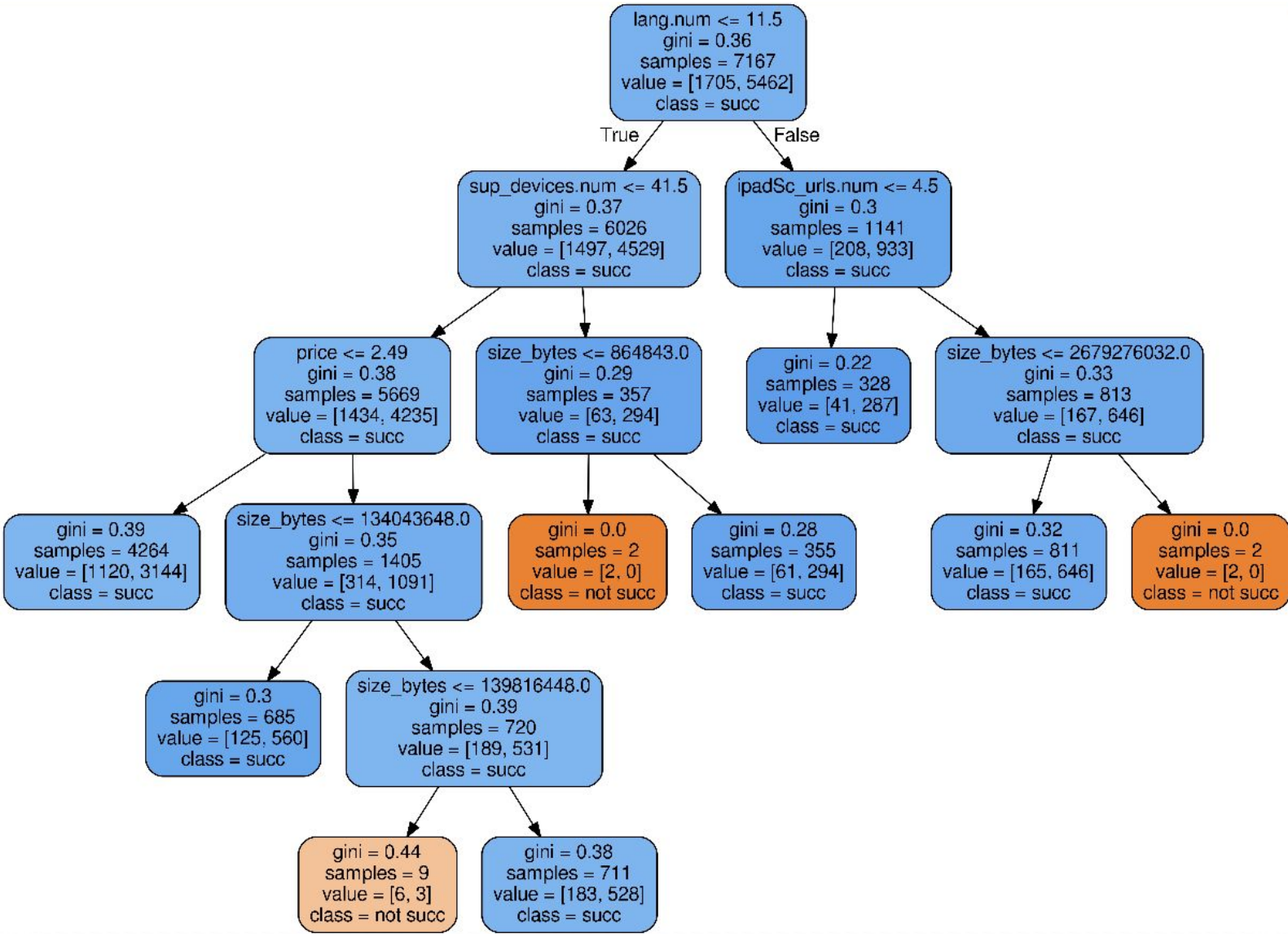


Reglas

- Regla 0: Si la app no tiene licencia vpp, y el precio es menor a 5.50 dólares, entonces será exitosa.
    - Confianza: 64%
    - Captura: 0.4%
    - Soporte: 0.47%
- Regla 1: Si la app no tiene licencias vpp pero su precio supera los 5.50 dólares, entonces será exitosa solamente si su tamaño no supera los 1.7 Gb.
    - Confianza: 80%
    - Captura: 0.07%
    - Soporte: 0.07%
- Regla 2: Si la app no tiene licencias vpp, su precio supera los 5.50 dólares y además su tamaño excede los 1.7 Gb, la app no será exitosa.
    - Confianza: 90%
    - Captura: 0.5%
    - Soporte: 0.14%
- Regla 3: Si la app sí tiene licencia vpp, tiene menos de 5 screenshots para mostrar y soporta menos de 5 idiomas, entonces será exitosa.
    - Confianza: 74%
    - Captura: 27%
    - Soporte: 28%
- Regla 4: Si la app sí tiene licencia vpp, tiene menos de 5 screenshots para mostrar y soporta más de 4 idiomas, entonces será exitosa.
    - Confianza: 88%
    - Captura: 10%
    - Soporte: 9%
- Regla 5: Si la app sí tiene licencia vpp, tiene más de 4 screenshots para mostrar, y supera los 7.50 dólares, entonces será exitosa.
    - Confianza: 86%
    - Captura: 2.55%
    - Soporte: 2.25%
- Regla 6: Si la app sí tiene licencia vpp, tiene más de 4 screenshots para mostrar, no supera los 7.50 dólares y tiene un tamaño menor a 90 Mb, entonces será exitosa.
    - Confianza: 78%
    - Captura: 22%
    - Soporte: 21%
- Regla 7: Si la app sí tiene licencia vpp, tiene más de 4 screenshots para mostrar, no supera los 7.50 dólares, tiene un tamaño mayor a 90 Mb y soporta menos de 42 dispositivos, entonces será exitosa.
    - Confianza: 72%
    - Captura: 35%
    - Soporte: 37%
- Regla 8: Si la app sí tiene licencia vpp, tiene más de 4 screenshots para mostrar, no supera los 7.50 dólares, tiene un tamaño mayor a 90 Mb y soporta más de 41 dispositivos, entonces será exitosa.

- Confianza: 84%
- Captura: 2%
- Soporte: 2%

Árbol 2



Reglas

- Regla 0: Si la cantidad de idiomas soportados es menor a 12, la cantidad de dispositivos soportados menor a 42 y el precio es menor a 2.50, entonces la app será exitosa.
    - Confianza: 74%
    - Captura: 58%
    - Soporte: 59%
- Regla 1: Si la cantidad de idiomas soportados es menor a 12, la cantidad de dispositivos soportados menor a 42, el precio supera los 2.50 dólares y el tamaño es inferior a 134 Mb, entonces la app será exitosa.
    - Confianza: 82%
    - Captura: 10%
    - Soporte: 10%
- Regla 2: Si la cantidad de idiomas soportados es menor a 12, la cantidad de dispositivos soportados menor a 42, el precio supera los 2.50 dólares y el tamaño es inferior a 140 Mb, entonces la app no será exitosa.
    - Confianza: 66%
    - Captura: 0.35%
    - Soporte: 1.3%
- Regla 3: Si la cantidad de idiomas soportados es menor a 12, la cantidad de dispositivos soportados menor a 42, el precio supera los 2.50 dólares y el tamaño es superior a 139 Mb, entonces la app será exitosa.
    - Confianza: 74%
    - Captura: 10%
    - Soporte: 10%
- Regla 4: Si la cantidad de idiomas soportados es menor a 12, la cantidad de dispositivos soportados mayor a 41, tamaño menor a 865 Kb, entonces la app no tendrá éxito.
    - Confianza: -%
    - Captura: -%
    - Soporte: 0%
- Regla 5: Si la cantidad de idiomas soportados es menor a 12, la cantidad de dispositivos soportados mayor a 41, tamaño mayor a 864 Kb, entonces la app tendrá éxito.
    - Confianza: 83%
    - Captura: 5%
    - Soporte:6%
- Regla 6: Si la cantidad de idiomas soportados es mayor a 11 y la cantidad de screenshots disponibles es menor a 5, entonces la app tendrá éxito.
    - Confianza: 88%
    - Captura: 5%
    - Soporte:5%
- Regla 7: Si la cantidad de idiomas soportados es mayor a 11, la cantidad de screenshots disponibles es mayor a 4 y el tamaño no supera los 2.7 Gb, entonces la app tendrá éxito.

    - Confianza: 80%
    - Captura:12%
    - Soporte:11%
- Regla 8: Si la cantidad de idiomas soportados es mayor a 11, la cantidad de screenshots disponibles es mayor a 4 y el tamaño supera los 2.7 Gb, entonces la app no tendrá éxito.
    - Confianza: 100%
    - Captura:0.1%
    - Soporte:0.02%

Análisis de las reglas

Se detallan a continuación las reglas respaldadas por una cantidad de observaciones razonable, tomando las que tengan soporte mayor o igual al 20%.

- Regla 3 (árbol 1): Si la app sí tiene licencia vpp, tiene menos de 5 screenshots para mostrar y soporta menos de 5 idiomas, entonces será exitosa.
    - Confianza: 74%
    - Captura: 27%
    - Soporte: 28%
        - A partir de esta regla se puede inferir que la licencia vpp juega un papel muy importante en el éxito de una app, ya que aún teniendo pocos screenshots y soportando relativamente pocos idiomas alcanza el éxito. Hay que tener en cuenta que no todo tipo de apps es adecuada para tener una licencia de este tipo: suele darse mucho en apps de educación.
- Regla 6 (árbol 1): Si la app sí tiene licencia vpp, tiene más de 4 screenshots para mostrar, no supera los 7.50 dólares y tiene un tamaño menor a 90 Mb, entonces será exitosa.
    - Confianza: 78%
    - Captura: 22%
    - Soporte: 21%
        - Esta regla indica que las apps con licencia vpp que no superan los 7.50 dólares son exitosas si se logra que tengan un tamaño inferior a los 90 Mb. Un menor tamaño del descargable tiene relación con la infraestructura telecomunicacional presente en el área de negocio: es lógico entonces ver que un precio bajo que se corresponde con un tamaño menor tenga éxito.
- Regla 7 (árbol 1): Si la app sí tiene licencia vpp, tiene más de 4 screenshots para mostrar, no supera los 7.50 dólares, tiene un tamaño mayor a 90 Mb y soporta menos de 42 dispositivos, entonces será exitosa.
    - Confianza: 72%
    - Captura: 35%
    - Soporte: 37%
        - En caso de superar los 90Mb, la aplicación aún puede ser exitosa si soporta menos de 42 dispositivos, lo que probablemente tenga que ver con que una nueva aplicación hará uso de APIs más nuevas

compatibles con menos dispositivos, y al ser más nuevas también ocupen más lugar.

- Regla 0 (árbol 2): Si la cantidad de idiomas soportados es menor a 12, la cantidad de dispositivos soportados menor a 42 y el precio es menor a 2.50, entonces la app será exitosa.
    - Confianza: 74%
    - Captura: 58%
    - Soporte: 59%
        - La regla indica que si el precio es bajo, aún si no se soporta gran cantidad de idiomas, la app puede tener éxito.

Revisión de configuración de parámetros

Para lograr una buena construcción del árbol, se limitó la cantidad de nodos hoja a 9. Además, para poder reproducir el resultado nuevamente, se usó una seed = 144 en un caso y seed = 115 en el otro.
Además de la cantidad de nodos hoja, se probó variar la cantidad mínima de muestras requerida para hacer un split, la cantidad mínima de muestras que debe haber en un nodo hoja, la cantidad máxima de niveles del árbol.
Los features que se usaron finalmente fueron:

- size_bytes
- price
- cont_rating
- sup_devices.num
- ipadSc_urls.num
- lang.num
- vpp_lic

Fase 5: Evaluación

Evaluación del resultado

Valoración de los resultados de minería de datos

Luego de haber aplicado las técnicas introducidas previamente, se han llegado a inferir algunas reglas que dejan al descubierto algunas caracterśticas que aumentan la probabilidad de éxito de una app. Si bien queda lugar a más análisis debido a que con las reglas no se llega a cubrir un panorama completo de las cuestiones a tener en cuenta en el desarrollo de una app, se puede confirmar que se ha obtenido información que de otra forma sería difícil obtener.
En general, revisando las reglas encontramos algunos patrones en común que categorizarían a una aplicación como exitosa:

- Tamaño de 100 Mb entendemos que es un tamaño medio.
- Precio bajo, convenientemente gratuita.
- Que soporte 5 o más idiomas
- Que tenga al menos 4 screenshots
- Si se cuenta con licencia vpp, el resto de las reglas pierde algo de peso.

Por todo esto, se puede considerar que el proyecto de investigación fue exitoso.

Modelo aprobado

Ambos árboles dieron información útil; cabe notar que visualizando dos árboles se obtuvo un mejor panorama de los criterios buscados.

Proceso de revisión

A partir del análisis exploratorio de datos, seguido de la generación de reglas, se ha llegado a cumplir con el objetivo planteado, por lo que en principio no quedan actividades pendientes.

Determinación de próximos pasos

Como próximos pasos se podría dejar algún análisis más profundo una vez definida la categoría sobre la que se quiera desarrollar una app, ya que se supone que las reglas de éxito puede variar de categoría en categoría. Esto podría dar reglas muchísimo más claras.

Conclusiones

En base al set de datos, los resultados obtenidos, y las conclusiones a las que llegamos a partir de ellas, podemos decir que se cumple el objetivo de la metodología, para luego, a partir de esto tomar acciones para predecir si una app será exitosa dadas ciertas características.