Proyecto Google Play Store Apps

1st Esneider Pantoja
Facultad de ciencias naturales e ingenieria
Universidad De Bogota Jorge Tadeo lozano
Bogota D.C
esneider.pantojad@utadeo.edu.co

2nd Juan Naranjo
Facultad de ciencias naturales e ingenieria
Universidad De Bogota Jorge Tadeo lozano
Bogota D.C
juanf.naranjog@utadeo.edu.co

3rd Cristian Hernández
Facultad de ciencias naturales e ingenieria
Universidad De Bogota Jorge Tadeo lozano
Bogota D.C
cristianc.hernandezs@utadeo.edu.co

Resumen

En el presente proyecto se contempla el proceso de analisis datos, basados en Google Play Store App.Una base de datos en la cual nos entregan 10 mil datos acerca de apps que ofrese play store. partiendo de una depuracion y organizacion de datos utilizando las herramientas de pyton como pandas, se logra una mejor visualizacionde los resultados que se presentan en el siguiente reporte

1. Introduccion

Si bien muchos conjuntos de datos públicos (en Kaggle y similares) proporcionan datos de la App Store de Apple, no hay muchos conjuntos de datos equivalentes disponibles para las aplicaciones de Google Play Store en la web. Al profundizar, descubrí que la página de la App Store de iTunes implementa una estructura similar a un apéndice bien indexada para permitir un raspado web simple y fácil. Por otro lado, Google Play Store utiliza sofisticadas técnicas modernas (como la carga dinámica de páginas) utilizando JQuery, lo que hace que el scraping sea más desafiante.

2. Método de visualización

Partiendo de la necesidad de tener resultados coherente y acordes a las necesidades del entorno se trabajo mediante la metodologia CRISP-DM (Cross Industry Standard Process for Data Mining) para evaluar la trazabilidad de los datos entregados por la pagina kaggle: Google play store. Para ello se aplicaron los 6 pasos guia de dicha metodologia (comprensión del problema, comprensión de datos, preparación de datos, modelado, evaluación del modelo e implementación del mismo).

Con un proceso inicial de limpieza, transformacion y revision de los datos usamos las herramientas entregadas en clase para tal fin.Con PYTHON como herramienta principal se uso (NUMPY) con sus extencion (PANDAS). para lograr un tratamiento de los datos entregados y asi obtener una visualizacion coherente y optima de los datos utilizables y operacionales para un analisis postumo, logrando una toma de desiciones acertiva.

3. Análisis de los datos

3.1. Comprensión del problema:

Se hizo entrega de una base de datos para nuestro archivo nos indica un total de 10841 registros divididos en 13 columnas C̈aracteristicasd̈iferentes para un total de 140933 entradas.Es evidente la necesidad de aplicacion de tecnicas de Machine learning,para la depuracion de datos obsoletos o que esten fuera de contexto para las necesidades de nuestro proyecto.

Nuestro objetivo es Identificar el tipo de applicaciones mas demandas dentro de la plataforma Google Play.bajo delimitaciones especificas de rangos de publico,como rango de edades de los ususrios, zonas geograficas de mayor demanda y aplicaciones de preferencia.

3.2. Comprensión de datos:

Nos dirigimos a la pagina kaggle: Google play store. Descargamos el conjunto de datos referentes al proyecto en la opcion D̈ownload (2 MB)l̈a cual contiene los dataset del proyecto. Conjunto de datos: -googleplaystore.csv (1.3 MB) Este archivo contiene los detalles de las aplicaciones en Google Play. con 13 características que describen una aplicación dada. -googleplaystore$_u ser_r eviews.csv$(7,31MB)$Estearchivocontienelasprimeras100rese\ddot{n}asm\ddot{a}srelevantesd\ddot{e}cadaaplicación.Cadatext$

3.3. Preparación de datos:

Con la implementacion de Python lo primero en hacerse fue la importacion de las librerias: numpy es una libreria que da soporte para crear vectores y matrices grandes multidimensionales, junto con una gran colección de funciones matemáticas de alto nivel para operar con ellas. pandas Es una liberia como extensión de NumPy para manipulación y análisis de datos, en particular ofrece estructuras de datos y operaciones para manipular tablas numéricas y series temporales.

- Para examinar el contenido del archivo ejecutamos la funcion head() que nos indica los 5 primeros registros del dataSet.

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
0	Photo Editor & Candy Camera & Grid & ScrapBook	ART_AND_DESIGN	4.1	159	19M	10,000+	Free	0	Everyone	Art & Design	January 7, 2018	1.0.0	4.0.3 and up
1	Coloring book moana	ART_AND_DESIGN	3.9	967	14M	500,000+	Free	0	Everyone	Art & Design;Pretend Play	January 15, 2018	2.0.0	4.0.3 and up
2	U Launcher Lite – FREE Live Cool Themes, Hide ...	ART_AND_DESIGN	4.7	87510	8.7M	5,000,000+	Free	0	Everyone	Art & Design	August 1, 2018	1.2.4	4.0.3 and up
3	Sketch - Draw & Paint	ART_AND_DESIGN	4.5	215644	25M	50,000,000+	Free	0	Teen	Art & Design	June 8, 2018	Varies with device	4.2 and up
4	Pixel Draw - Number Art Coloring Book	ART_AND_DESIGN	4.3	967	2.8M	100,000+	Free	0	Everyone	Art & Design;Creativity	June 20, 2018	1.1	4.4 and up

Figura 1. “5 primeros registros del dataSet.” [**?**]

- Para saber todas las columnas de nuestro archivo ejecutamos el siguiente atributo de dataset colums esto nos permitira conocer todos las caracteristicas de nuestro archivo.

```
Index(['App', 'Category', 'Rating', 'Reviews', 'Size', 'Installs', 'Type',
       'Price', 'Content Rating', 'Genres', 'Last Updated', 'Current Ver',
       'Android Ver'],
      dtype='object')
```

Figura 2. “columnas del archivo.” [**?**]

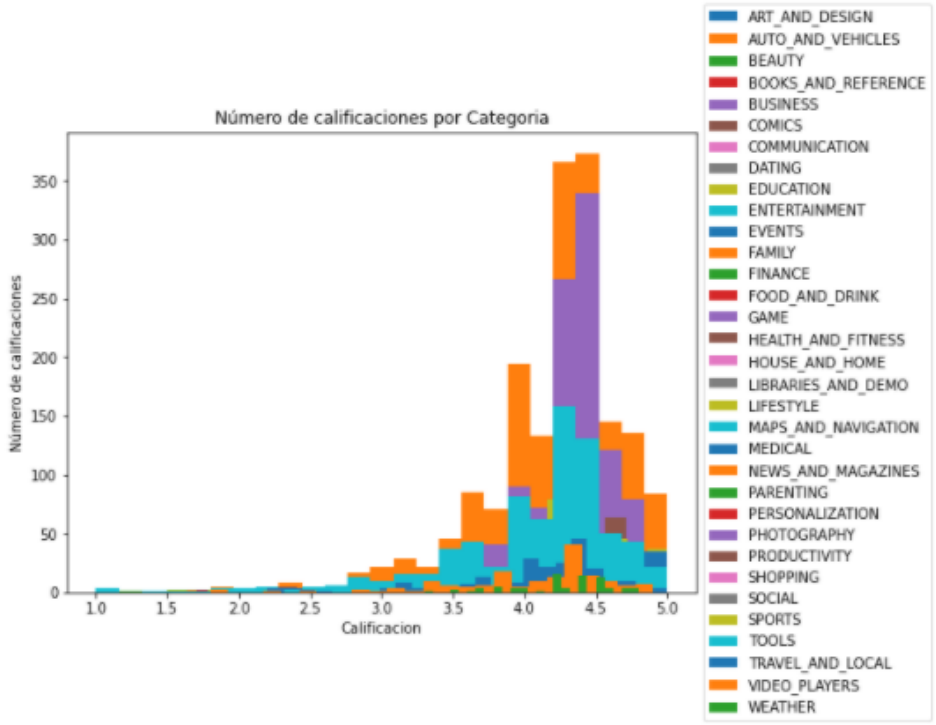
- Para conocer el tipo de datos de las columnas ejecutamos info().
  Column Non-Null Count Dtype
  — —— ————— ——
  0 App 10841 non-null object
  1 Category 10841 non-null object
  2 Rating 9367 non-null float64
  3 Reviews 10841 non-null object
  4 Size 10841 non-null object
  5 Installs 10841 non-null object
  6 Type 10840 non-null object
  7 Price 10841 non-null object
  8 Content Rating 10840 non-null object
  9 Genres 10841 non-null object

- Agrupamos nuestro conjunto de datos por C̈ategory:.

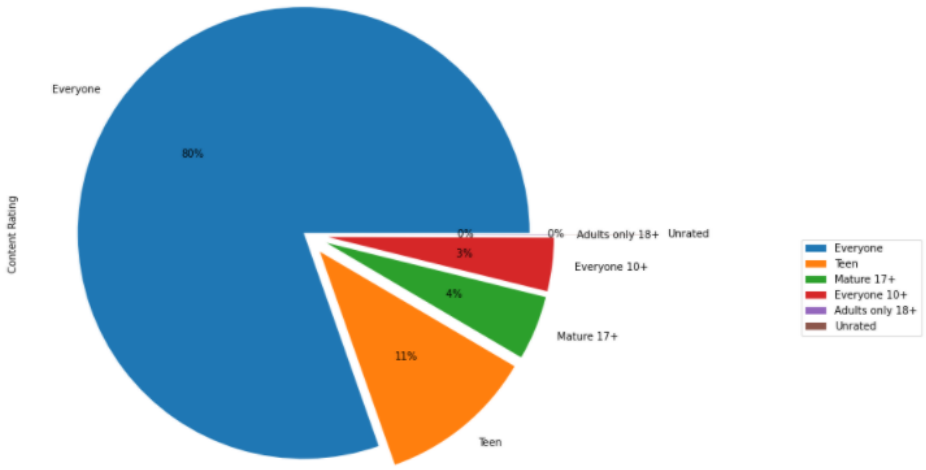
	Rating							
	count	mean	std	min	25 %	50 %	75 %	max
Category								
ART$_A$ND_DESIGN	62.0	4.358.065	0.358297	3.2	4.100	4.4	4.700	5.0
AUTO$_A$ND_VEHICLES	73.0	4.190.411	0.543692	2.1	4.000	4.3	4.600	4.9
BEAUTY	42.0	4.278.571	0.362603	3.1	4.000	4.3	4.575	4.9
BOOKS$_A$ND_REFERENCE	178.0	4.346.067	0.429046	2.7	4.100	4.5	4.600	5.0
BUSINESS	303.0	4.121.452	0.624422	1.0	3.900	4.3	4.500	5.0
COMICS	58.0	4.155.172	0.537758	2.8	3.825	4.4	4.500	5.0
COMMUNICATION	328.0	4.158.537	0.426192	1.0	4.000	4.3	4.400	5.0
DATING	195.0	3.970.769	0.630510	1.0	3.700	4.1	4.400	5.0
EDUCATION	155.0	4.389.032	0.251894	3.5	4.200	4.4	4.600	4.9
ENTERTAINMENT	149.0	4.126.174	0.302556	3.0	3.900	4.2	4.300	4.7
EVENTS	45.0	4.435.556	0.419499	2.9	4.200	4.5	4.700	5.0
FAMILY	1747.0	4.192.272	0.508026	1.0	4.000	4.3	4.500	5.0
FINANCE	323.0	4.131.889	0.642108	1.0	4.000	4.3	4.500	5.0
FOOD$_A$ND_DRINK	109.0	4.166.972	0.548070	1.7	4.000	4.3	4.500	5.0
GAME	1097.0	4.286.326	0.365375	1.0	4.100	4.4	4.500	5.0
HEALTH$_A$ND_FITNESS	297.0	4.277.104	0.617822	1.4	4.100	4.5	4.600	5.0
HOUSE$_A$ND_HOME	76.0	4.197.368	0.368411	2.8	4.000	4.3	4.500	4.8
LIBRARIES$_A$ND_DEMO	65.0	4.178.462	0.378522	3.1	4.000	4.2	4.400	5.0
LIFESTYLE	314.0	4.094.904	0.693907	1.5	3.800	4.2	4.600	5.0
MAPS$_A$ND_NAVIGATION	124.0	4.051.613	0.519926	1.9	3.775	4.2	4.400	4.9
MEDICAL	350.0	4.189.143	0.663581	1.0	4.000	4.3	4.600	5.0
NEWS$_A$ND_MAGAZINES	233.0	4.132.189	0.536707	1.7	3.900	4.2	4.500	5.0
PARENTING	50.0	4.300.000	0.517845	2.0	4.100	4.4	4.675	5.0
PERSONALIZATION	314.0	4.335.987	0.352732	2.5	4.200	4.4	4.600	5.0
PHOTOGRAPHY	317.0	4.192.114	0.462896	2.0	4.000	4.3	4.500	5.0
PRODUCTIVITY	351.0	4.211.396	0.504931	1.0	4.100	4.3	4.500	5.0
SHOPPING	238.0	4.259.664	0.404577	1.6	4.100	4.3	4.500	5.0
SOCIAL	259.0	4.255.598	0.413809	1.9	4.100	4.3	4.500	5.0
SPORTS	319.0	4.223.511	0.427857	1.5	4.100	4.3	4.500	5.0
TOOLS	734.0	4.047.411	0.616143	1.0	3.800	4.2	4.400	5.0
TRAVEL$_A$ND_LOCAL	226.0	4.109.292	0.504691	2.2	3.900	4.3	4.400	5.0
VIDEO$_P$LAYERS	160.0	4.063.750	0.551098	1.8	3.800	4.2	4.400	4.9
WEATHER	75.0	4.244.000	0.331353	3.3	4.050	4.3	4.500	4.8

- aplicaciones con calificación de 5.

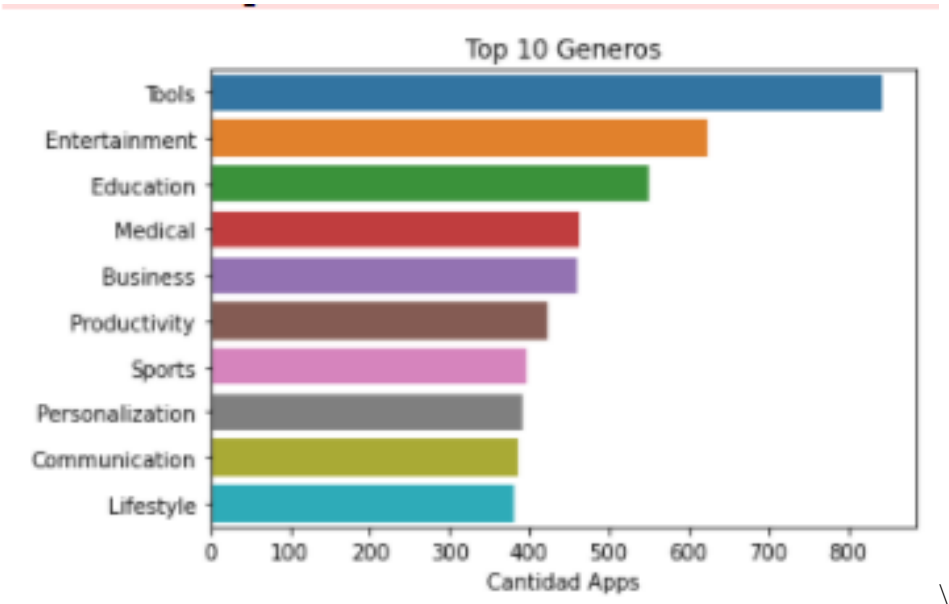
Category	Installs	App	Rating	
739	BUSINESS	5+	EB Cash Collections	5.0
740	SOCIAL	50+	UP EB Bill Payment	Details
5.0				
741	SOCIAL	10+	DN Blog	5.0
742	SOCIAL	5+	CB Heroes	5.0
743	GAME	50+	Axe Champs! Wars	5.0
...	...	...	...	...
1003	GAME	10+	211:CK	5.0
1004	DATING	500+	Spine- The dating app	5.0
1005	DATING	100+	Online Girls Chat Group	5.0
1006	TOOLS	100+	BK Formula Calculator	5.0
1007	TOOLS	100+	Jabbla BT	5.0



$H]$ $\backslash numerodecalificacionesporcategoria.$"1001



$H]$ $\backslash tiposdepersonasquecalificanlasapps.$"1001



$H]$ $\backslash Appsporgenero.$"1001[?]

4. Conclusiones

- El lenguaje de programación Python, nos brinda herramientas de gran utilidad para el manejo de datos, brindando la facilidad de trabajar con datos verdaderamente importante para nuestro análisis. llevando a una toma de decisiones asertivas.
- Las personas jóvenes representan el 17 por ciento del total de calificaciones en total de las apps.
- las aplicaciones mas calificadas, no siempre son las de mejor puntuacion, en muchas ocaciones las personas le ven mayor motivacion a dar una calificacion negativa por un mal servicio.

Referencias

[1] Google Play Store Apps. (2019, 3 febrero). Kaggle.

[2] GitHub: Where the world builds software. (2021). GitHub.

[3] Project Jupyter. (2021). Try Jupyter. https://jupyter.org/try

[4] Aplicación de metodología CRISP-DM para segmentación geográfica de una base de datos pública. (2020, agosto). http://www.scielo.org.mx/scielo.php?script=sci$_arttextpid = S1405 - 77432020000100008 Project Jupyter.(2021).TryJupyter.https : //jupyter.org/try$