ITBA
Instituto Tecnológicc
de Buenos Aires

ANÁLISIS PREDICTIVO

EXÁMEN FINAL

CAMILA COLLADO | 1° CUATRIMESTRE 2022

AGENDA

- DATASET
- OBJETIVO
- VARIABLES RELEVANTES
- ANÁLISIS EXPLORATORIO DE DATOS
- ALGORITMOS
- CONCLUSIONES

DATASET

Contiene datos sobre 2.3m de aplicaciones en Google Play Store.

Origen: **Kaggle** (scrapping de Google Play Store).

Fecha: **junio 2021.**

Dimensión: **2.312.944** filas, **24** columnas.

PROPÓSITO

En Google Play Store hay una gran cantidad de apps que cumplen todo tipo de funciones.

A la hora de decidir que aplicación descargar **la valoración es un dato importante para los usuarios**.

Las valoraciones son brindadas por los usuarios pero podemos intentar anticiparnos a su pensamiento.

OBJETIVO

Predecir el rating de una app.

Para poder completar el dato de rating de las aplicaciones que tiene descargas pero no tiene ninguna valoración de usuarios.

HIPÓTESIS

A partir de los datos básicos de una app se puede intuir cómo será valorada por los usuarios.

SELECCIÓN DE VARIABLES

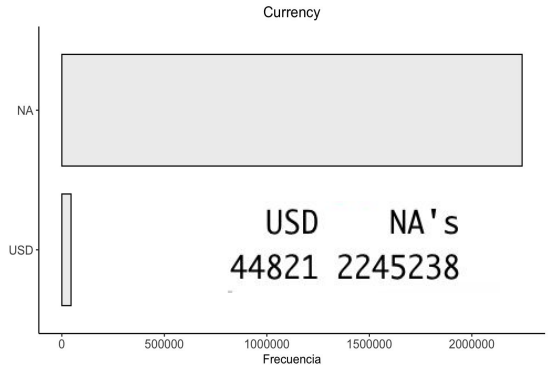
24 VARIABLES

App.id	App.Name	Installs	Developer.Website	Scraped.Time	Privacy.Policy
Released	Last.Updated	Category	Minimum.Android	Developer.Id	Content.Rating
Minimum.Installs	Maximum.Installs	Editors.Choice	Price	Free	Ad.Supported
Currency	Rating	In.App.Purchases	Developer.Email	Rating.Count	Size

VARIABLES QUE NO APORTAN INFORMACIÓN

- Developer.Website
- Developer.Email
- Scraped.Time
- Privacy.Policy
- App.Name
- Currency
- Minimum.Installs
- Installs

199.385 repetidos



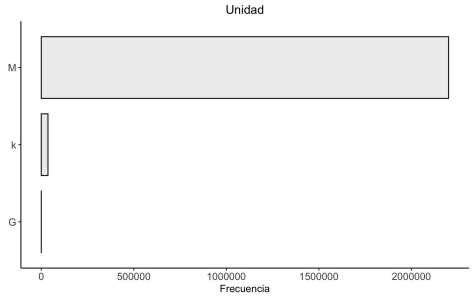
Installs <chr>	Minimum.Installs <dbl>	Maximum.Installs <dbl>
10+	10	15
5,000+	5000	7662
50+	50	58
10+	10	19
100+	100	478
50+	50	89

Eliminadas

Otra variable eliminada: Rating.Count

VARIABLES MODIFICADAS

- Size
- Released
- Last.Updated



Unificar unidad de medida: MB.
Cambiar “Varies with device” por la mediana.

Quitar vacíos.
Cambiar a formato fecha.
Filtrar los registros con Last.Updated < Released.

VARIABLES CREADAS

- Size.mb
- Released.DA
- Last.Updated.DA

Cambiar las variables de fechas por una resta de días para que sea un valor numérico comparable.
DA (days ago): la diferencia de días entre el día que se *scrapeó* el Store (30/6/2021) y la fecha de lanzamiento/última actualización.

VARIABLES FINALES

VARIABLE TARGET

Rating

VARIABLES CATEGÓRICAS

Category
Minimum.Android
Developer.Id
Content.Rating

VARIABLES NUMÉRICAS

Maximum.Installs
Price
ReleasedDA
Last.Updated.DA
Size.mb

VARIABLES LÓGICAS

Free
Ad.Supported
In.App.Purchases
Editors.Choice

ANÁLISIS EXPLORATORIO DE DATOS

LIMPIEZA

Rating: eliminar los NA y 0.

Size.MB: La categoría “Varies with device” se cambió por la mediana (var. numérica).

Last.Updated y Released: Quitar vacíos, cambiar a formato fecha y filtrar los registros con Last.Updated > Released.

Minimum.Android: Cambiar los registros de “Varies with device” por la moda (var. categórica) y quitar NA.

Chequear que no haya **App.id** repetidos ni NA (no había).

App.Name: elimino las vacias o igual a 0.

Chequear que **Installs y Minimum.Installs** sean iguales para poder eliminarlos.

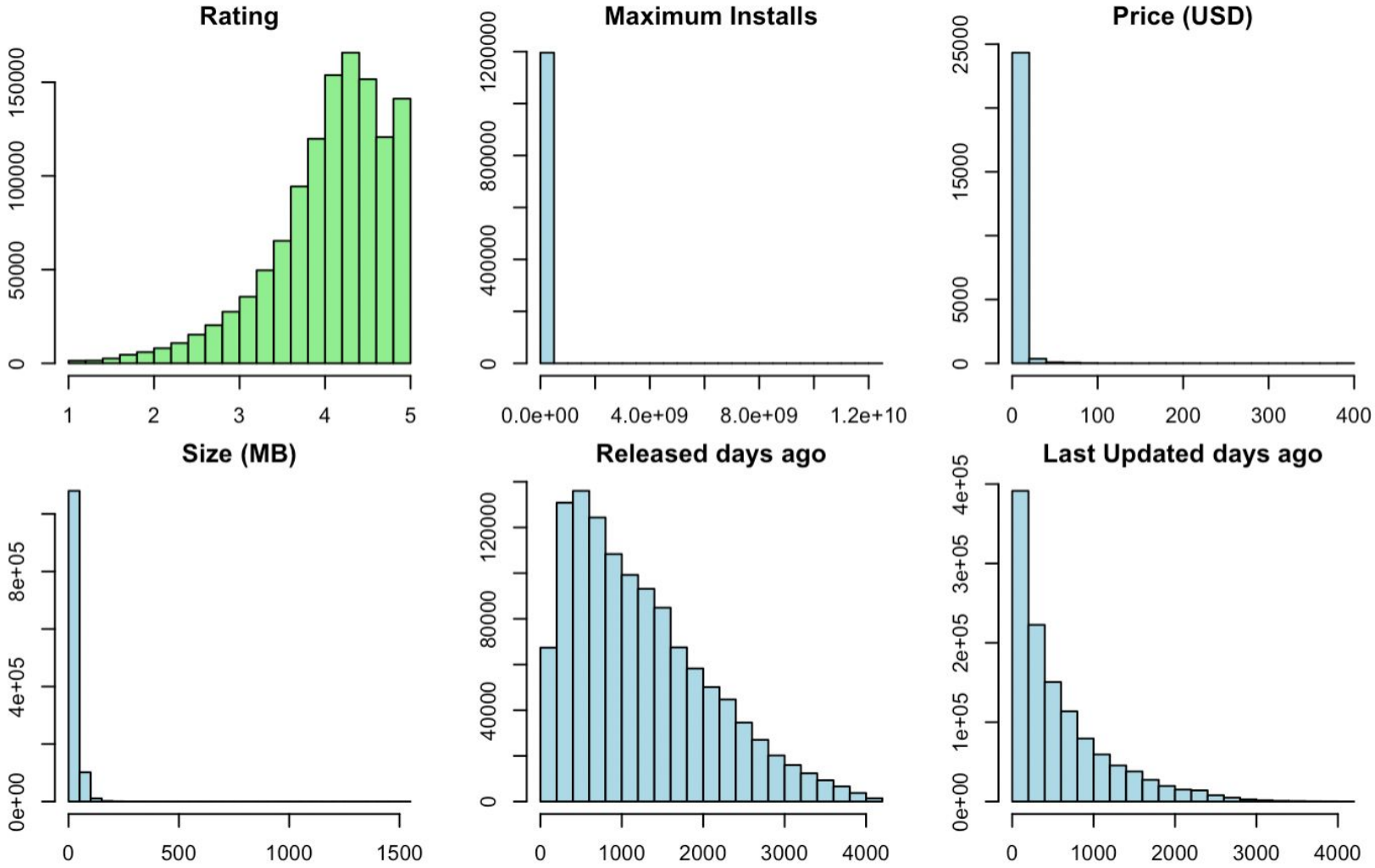
Chequear que **Maximum.Installs** es siempre mayor a **Minimum.Installs**.

Acomodar los registros que tiene **Price**=0 y **Free**=FALSE.

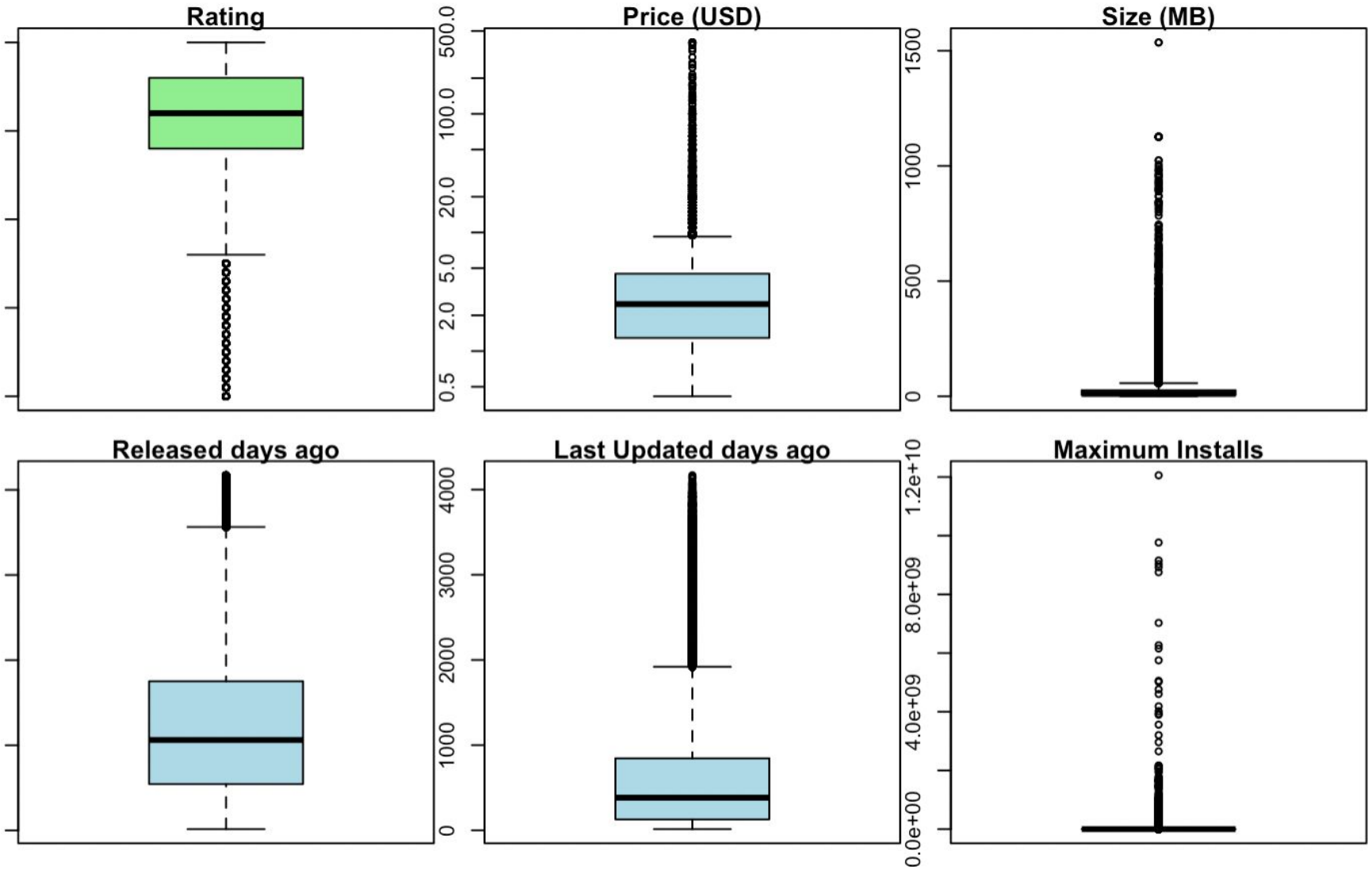
Developer.Id: eliminar NA.

Content.Rating: filtrar la categoría: “Unrated”.

DISTRIBUCIÓN | VARIABLES NUMÉRICAS

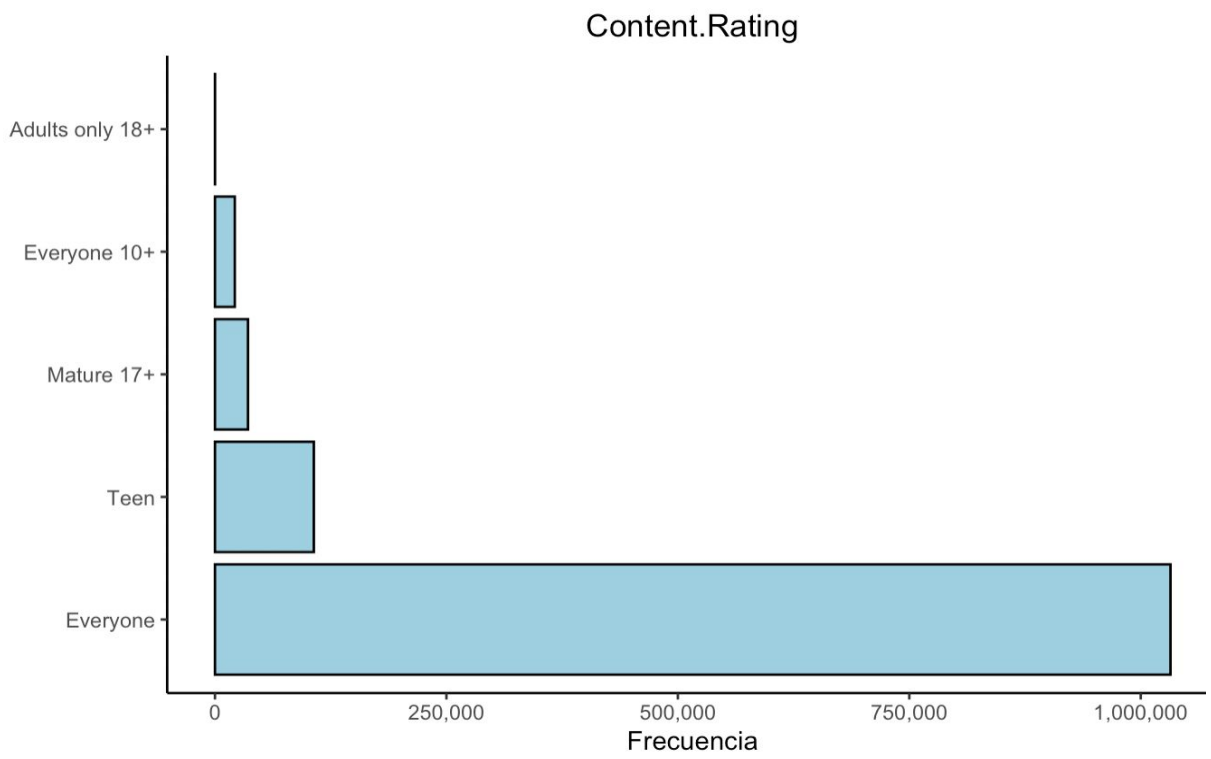


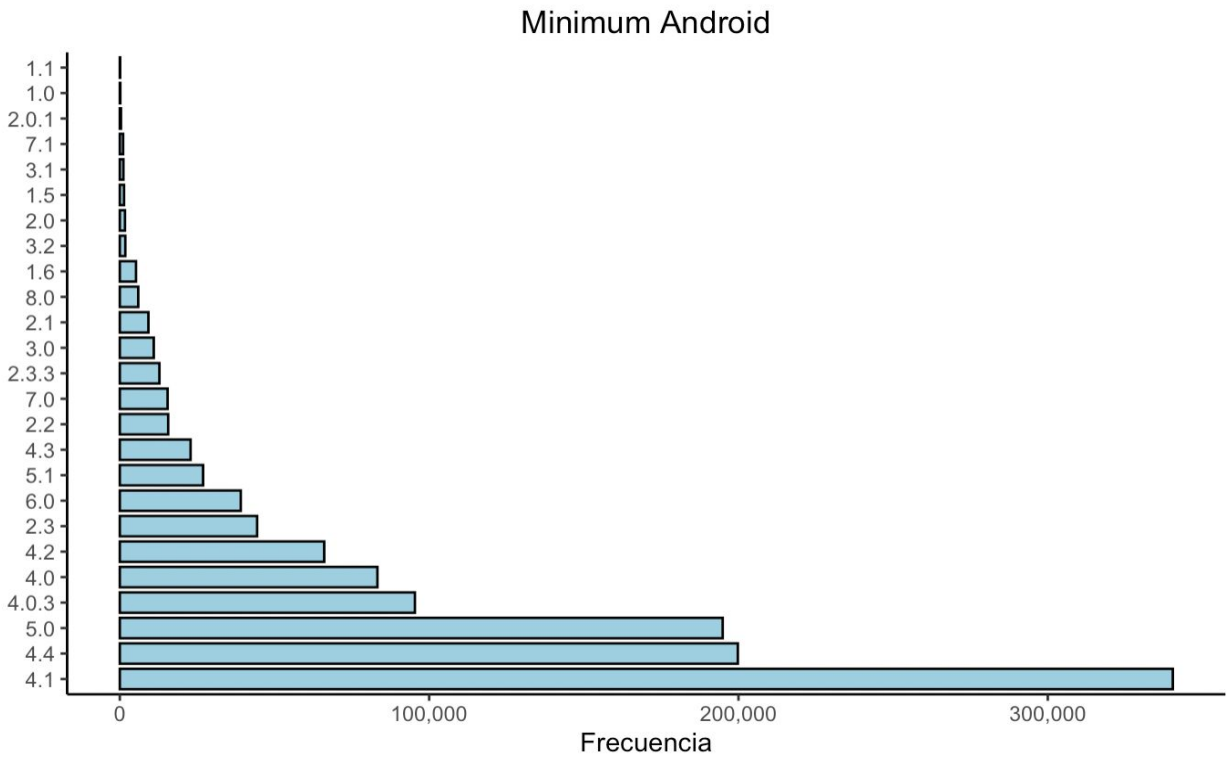
OUTLIERS | VARIABLES NUMÉRICAS



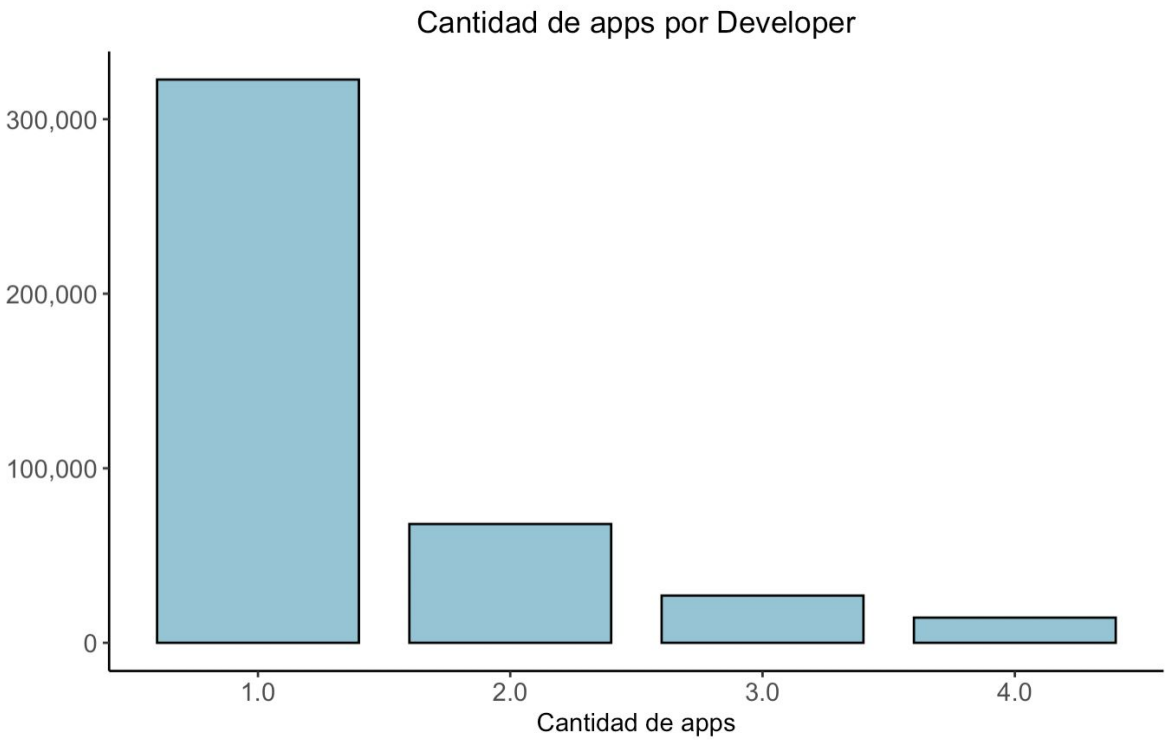
DISTRIBUCIÓN | VARIABLES LÓGICAS

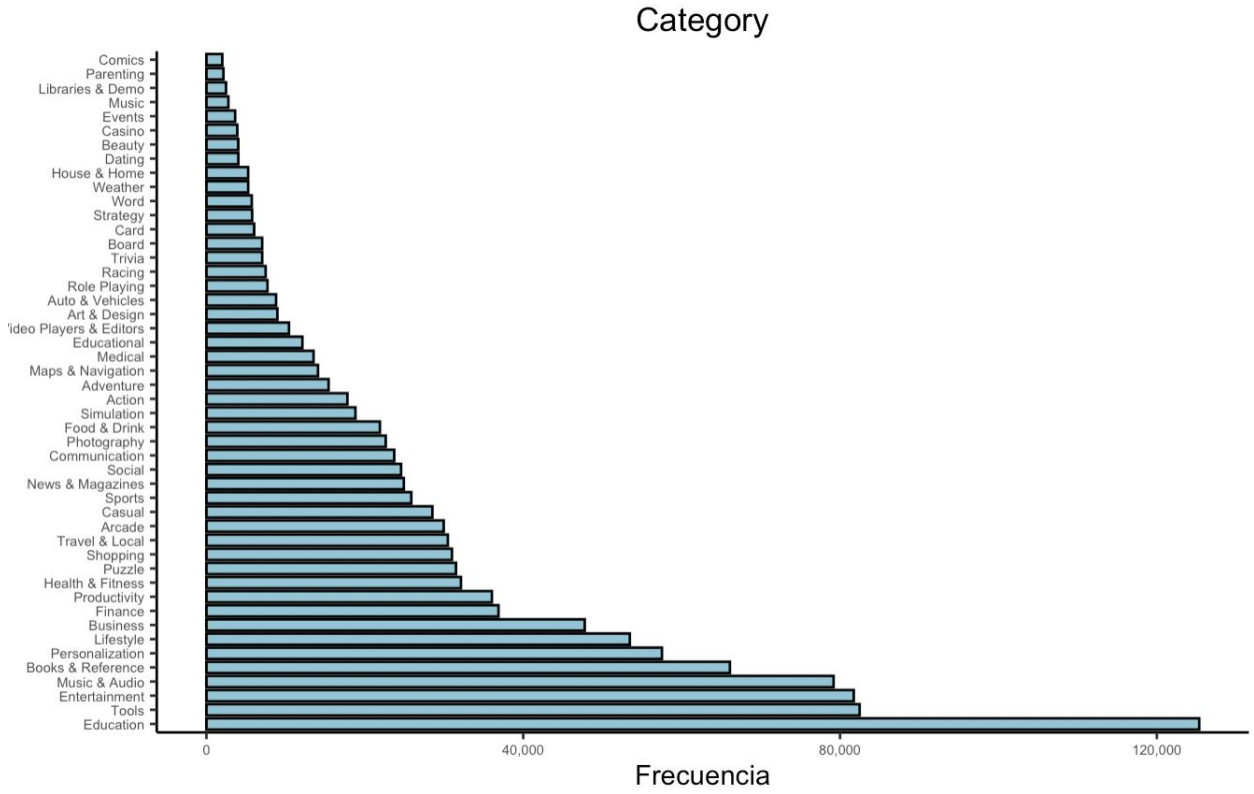
FRECUENCIA | VARIABLES CATEGÓRICAS





FRECUENCIA | VARIABLES CATEGÓRICAS

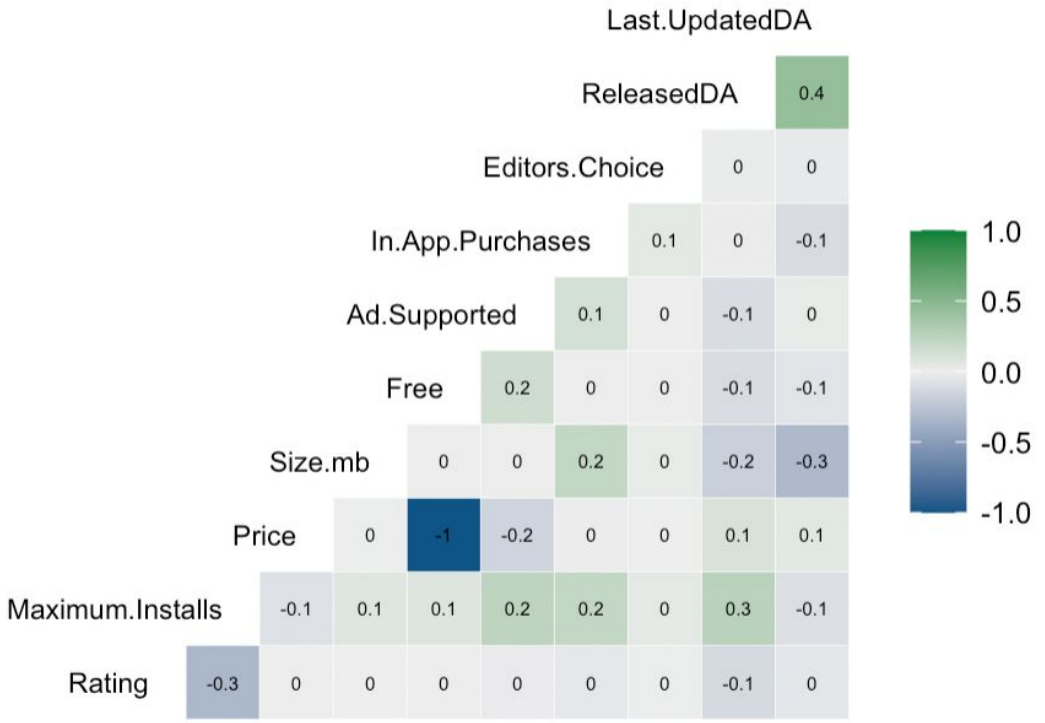




Esta variable tiene **muchos valores únicos.**
Se suma la cantidad de apps por developer y se grafica la distribución para saber si hay "monopolios" o son muchos "pequeños productores" de apps.

CORRELACIÓN

PARA VARIABLES NUMÉRICAS Y CATEGÓRICAS



	Category	Minimum.Android	Developer.Id	Content.Rating
Category	1.0000000	0.0724406	0.8277207	0.2766208
Minimum.Android	0.0724406	1.0000000	0.2766208	0.0381142
Developer.Id	0.8277207	0.8191162	1.0000000	0.8004715
Content.Rating	0.2766208	0.0381142	0.8004715	1.0000000

ALGORITMOS

ENCODING DE VARIABLES CATEGÓRICAS

Category

BINARY ENCODING

Para representar las 45 categorías solo necesito 6 columnas.

Content.Rating

ORDINAL ENCODING

1. Everyone
2. Everyone 10+
3. Teen
4. Mature 17+
5. Adults only 18+

Minimum.Android

FREQUENCY ENCODING

Developer.Id

FREQUENCY ENCODING

Debido a que la mayoría de los desarrolladores producen entre 1-4 apps y hay pocos “grandes productores”.

SEPARACIÓN TRAIN & TEST

Tipo de predicción
Regresión

Variable target
Rating

Train **80%** n = 1.150.398

Test **20%** n = 287.600

MODELOS PROBADOS	SCORE (R^2)
HISTOGRAM GRADIENT BOOSTING	19,0%
ÁRBOL SIMPLE	16,3%
KNN	10,6%
RANDOM FOREST	17,4%
EXTRA TREES	18,5%
X BOOST	20,0%
LIGHT GBM	18,9%

*SIN AJUSTE DE HIPER PARÁMETROS

AJUSTE DE HIPER PARÁMETROS | MODELOS

XGBOOST 21,8%

- learning_rate: 0.1
- n_estimators: 500
- max_depth: 10
- booster: gbtree
- subsample: 1
- gamma: 0

LIGHT GBM 22,0%

- learning_rate: 0.2
- n_estimators: 700
- max_depth: 16
- num_leaves: 31
- boosting_type: gbdt
- min_data_in_leaf: 100
- max_bin: 255

MODELO UTILIZADO

HISTOGRAM GRADIENT BOOSTING

- ✅ USO DE HISTOGRAMAS EN LUGAR DEL DATO PUNTUAL
- ✅ IMPLEMENTACIÓN DE CATBOOST PARA SKLEARN
- ✅ GRID SEARCH + CROSS VALIDATION
- ✅ AJUSTE DE HIPER PARÁMETROS
- ✅ EJECUCIÓN RÁPIDA
- ✅ MEJOR SCORE

AJUSTE DE HIPER PARÁMETROS

Learning rate	Max depth	Min samples leaf	Max leaf nodes
0.01	12	1	11
0.05	14	5	31
0.1	**16**	15	51
0.2	18	50	101
0.3	20	**200**	201
		250	251
		300	301
			100001
			500001
			1000001

AJUSTE DE HIPER PARÁMETROS

Max bins	Max iter	Loss	L2 regularization
205	80	SQUARED ERROR	0.0
245	70	ABSOLUTE ERROR	0.5
255	90	POISSON	1
265	**100**	QUANTILE	10
275	110		15
305			20

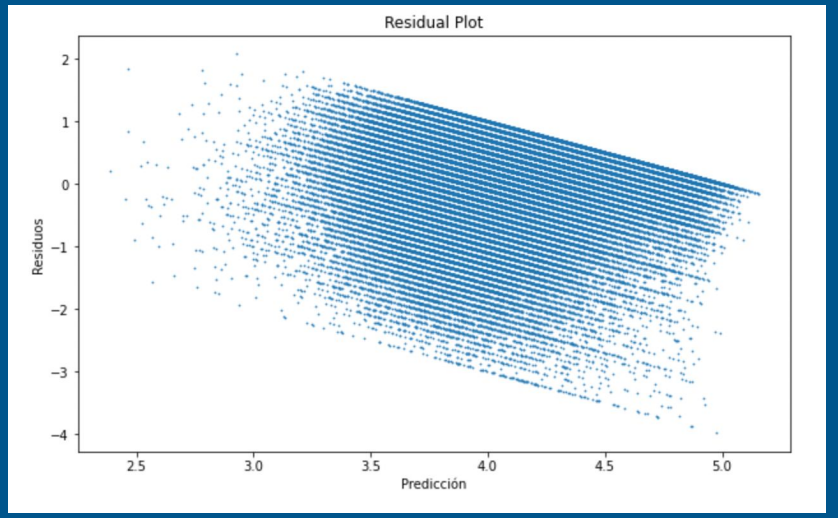
SCORE Y MÉTRICAS

R^2 = 0,2247111

VARIANZA EXPLICADA = 0,2247129

MEAN ABSOLUTE ERROR = 0,4449135

RESIDUOS



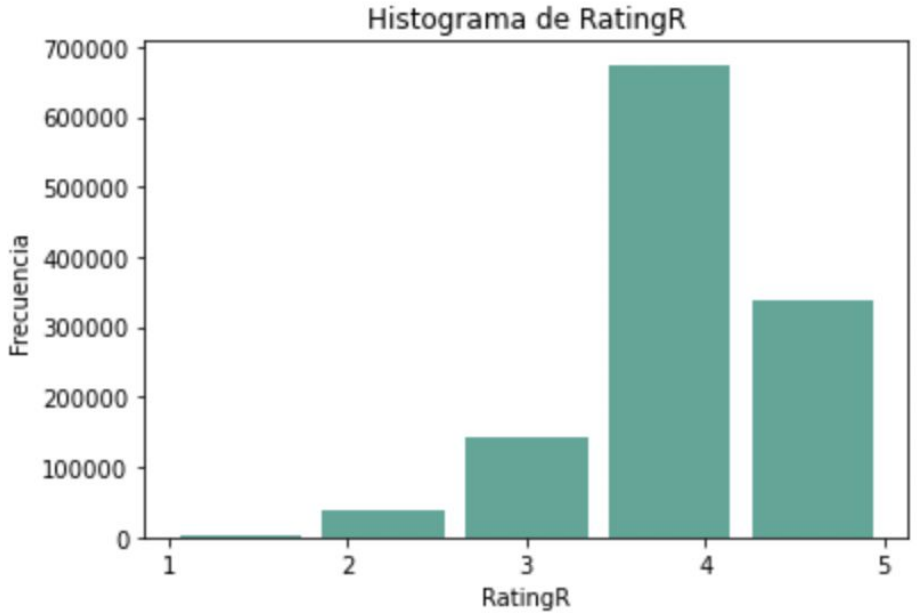
MEJORAR EL SCORE

- **Tipo de predicción:** Clasificación.
- **Variable Target:** ‘RatingR’, variable Rating llevada a número entero.
- **Algoritmo:** Histogram Gradient Boosting classifier.

Clases desbalanceadas: **Oversampling**

Dimensión train y test original vs. *oversampleadas*

(956844, 18)	(2696900, 18)
(956844, 1)	(2696900, 1)
(239211, 18)	(674225, 18)
(239211, 1)	(674225, 1)



Parámetros del modelo

Learning rate = 0.2

Max depth = 12

Min samples leaf = 200

Max leaf nodes = 9001

Class	Recall	Precision
1.0	1.000000	0.988826
2.0	0.900057	0.723207
3.0	0.564785	0.630386
4.0	0.469393	0.562846
5.0	0.684912	0.675051

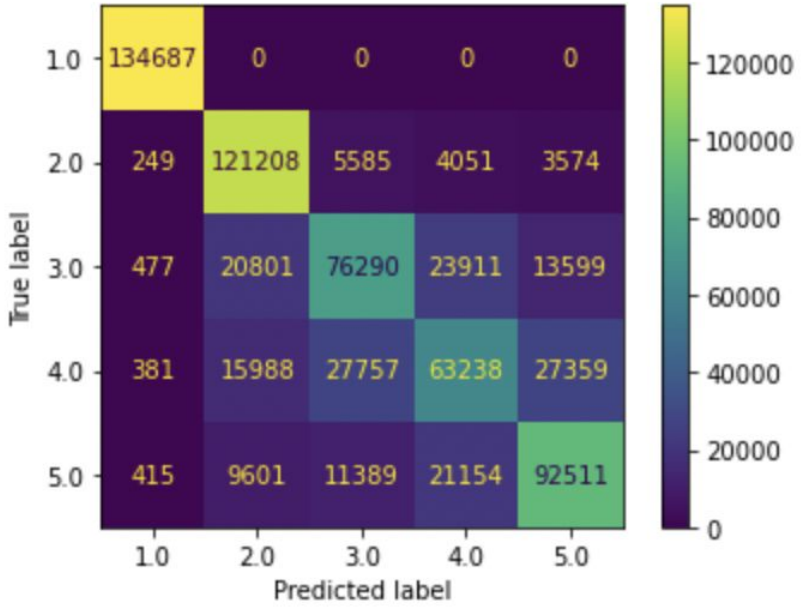
Métricas y score

Accuracy = 0,7237

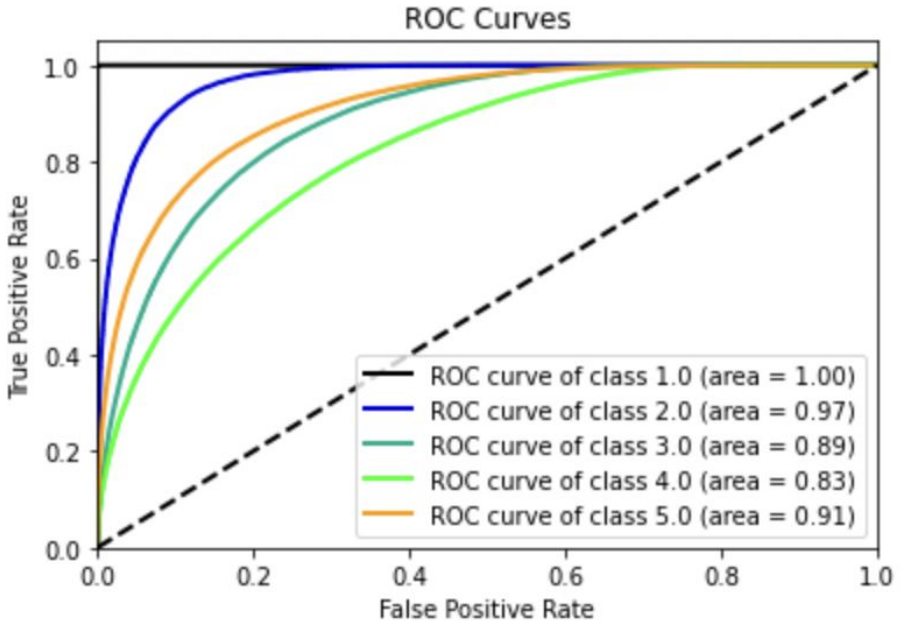
F1 = 0,7167

F2 = 0,7237

Matriz de confusión



Curva ROC



CONCLUSIONES Y HALLAZGOS

Es difícil predecir el rating basándose **solo** en los datos básicos de una app.

Redondeando los valores y haciendo una clasificación se puede obtener una mejor predicción.

¡GRACIAS!