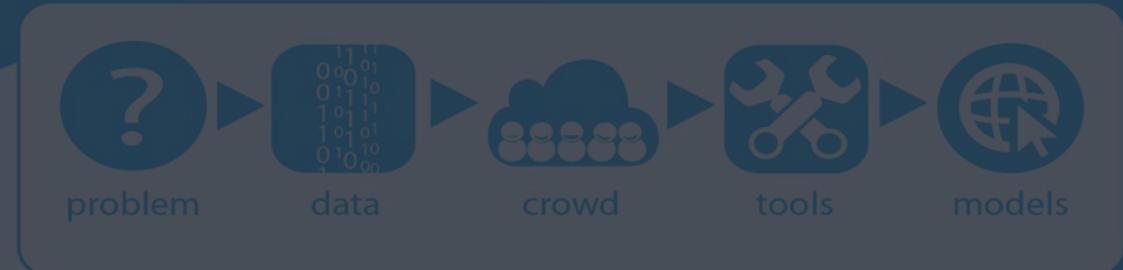


WEBINAR - abril 16 2020

Kaggle de Data Science. Nivel 1.

ponente: Marco Russo



Qué vamos a ver.

1. Organización del entorno de trabajo
2. Aplicación de Metodologías para la exploración de datos
3. Entrenamiento de tu primer modelo de aprendizaje automático
4. Enfrentarse a las competiciones de "Primeros pasos..."
5. Competir para maximizar los aprendizajes

Quién soy.

- **Consultor en Data** en Paradigma Digital, con más de 7 años como docente para importantes escuelas de negocios y profesor colaborador en la UOC.
- **Especializado** en data mining, optimización de modelos y machine learning en área del Marketing, Retail y Banca-Finanzas entre otras. Además de especialista en analítica digital, SEO y PPC en digital marketing y visualización de datos - BI.
- **Apasionado** de IoT, datos y robótica, dedico el tiempo con mi familia y a mi deporte favorito, bici de carretera.



Marco Russo (aka marcusRB)



[@rb_marcus](https://twitter.com/rb_marcus)



github.com/marcusRB



[marcusRB](https://www.linkedin.com/in/marcusRB)

01.01

•••

Introducción

Visión general.

¿Qué es Kaggle?

- **Comunidad** de data scientist y cientos de miles aficionados en aprendizaje automático
- **Kernels** o notebooks, script de casos de usos de negocio / investigaciones
- **Datasets** (aprox. 34k)
- **Competiciones** (privadas, públicas, inClass, \$\$\$)
- **Cursos** de python, R, SQL, machine learning, deep learning
- Espacio **ofertas de empleo**

≡ kaggle



Home



Compete



Data



Notebooks



Discuss



Courses



More

¿Qué es Kaggle?



The Home of Data Science & Machine Learning

Kaggle helps you learn, work, and play

Create an account or Host a competition

Competitions Datasets Kernels Discussion Jobs ... Sign In

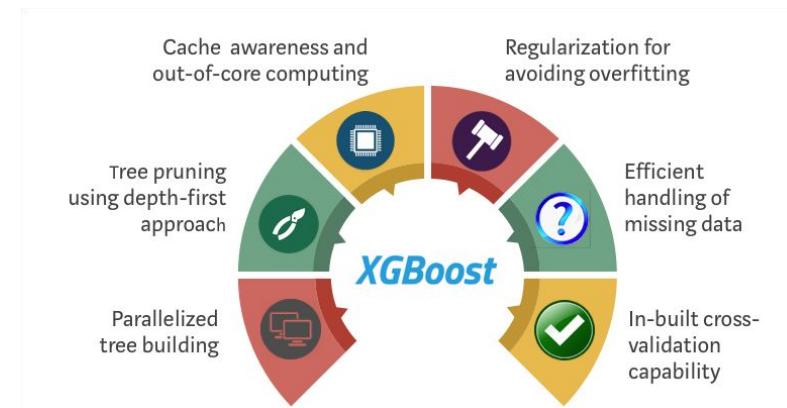
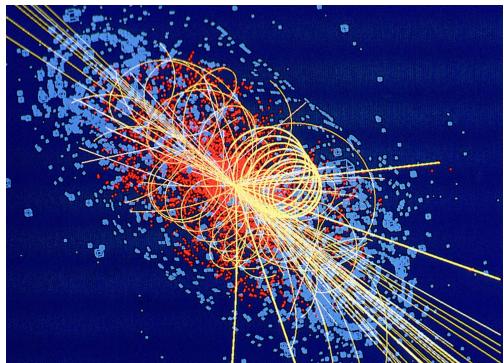
jobs board >

jobs board >

Tu portfolio

¿Qué es Kaggle?

fundada en 2010 y adquirida por Google en 2017



¿Qué es Kaggle?



¿Qué es Kaggle?

All Competitions

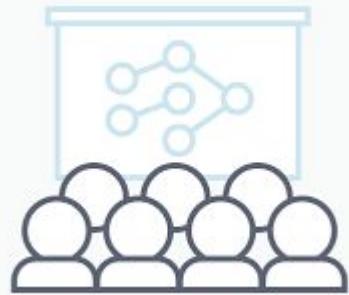
		Getting Started ▾	Reward ▾
		Active (Not Entered) Completed InClass	
	Just the Basics - Strata 2013 Live from Santa Clara, CA - Core Data Science Skills with Kaggle's Top Competitors Getting Started • 7 years ago • 49 Teams	Knowledge	\$1,000,000
	Just the Basics - Strata 2013 After-party Live from Santa Clara, CA Getting Started • 7 years ago • 48 Teams	Knowledge	\$50,000
	Data Science London + Scikit-learn Scikit-learn is an open-source machine learning library for Python. Give it a try here! Getting Started • 5 years ago • 191 Teams	Knowledge	\$50,000
	Facial Keypoints Detection Detect the location of keypoints on face images Getting Started • 3 years ago • 175 Teams	Knowledge	\$25,000
	First Steps With Julia Use Julia to identify characters from Google Street View images Getting Started • 3 years ago • 56 Teams	Knowledge	\$25,000
	Bag of Words Meets Bags of Popcorn Use Google's Word2Vec for movie reviews Getting Started • 5 years ago • 578 Teams	Knowledge	\$20,000

All Competitions

		Getting Started ▾	Reward ▾
		Active (Not Entered) Completed InClass	
	#DFDC Deepfake Detection Challenge Identify videos with facial or voice manipulations Featured • 7 days to go • Code Competition • 2281 Teams	\$1,000,000	
	M5 Forecasting - Accuracy Estimate the unit sales of Walmart retail goods Featured • 3 months to go • 2355 Teams	\$50,000	
	M5 Forecasting - Uncertainty Estimate the uncertainty distribution of Walmart unit sales. Featured • 3 months to go • 218 Teams	\$50,000	
	Jigsaw Multilingual Toxic Comment Classification Use TPU to identify toxicity comments across multiple languages Featured • 2 months to go • Code Competition • 524 Teams	\$50,000	
	University of Liverpool - Ion Switching Identify the number of channels open at each time point Research • a month to go • 1750 Teams	\$25,000	
	Google Cloud & NCAA® March Madness Analytics Uncover the madness of March Madness® Analytics • 15 days to go	\$25,000	
	Abstraction and Reasoning Challenge Create an AI capable of solving reasoning tasks it has never seen before Research • a month to go • Code Competition • 627 Teams	\$20,000	
	Tweet Sentiment Extraction Extract support phrases for sentiment labels Featured • 2 months to go • Code Competition • 400 Teams	\$15,000	
	COVID19 Global Forecasting (Week 4) Forecast daily COVID-19 spread in regions around world Research • 8 hours to go • Code Competition • 406 Teams		Knowledge



¿Qué es Kaggle?



bit.ly/2XD71Mz

Datathon, Competiciones internas, académicas, etc.

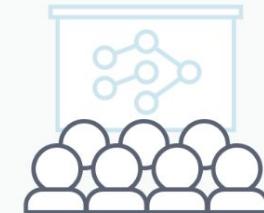
Kaggle InClass competitions make machine learning fun.

Use our free, self-service platform to create classroom competitions that engage and inspire your students.

How it Works

Start here to learn how to setup your first competition and how to best take advantage of the platform.

[Learn more](#)



FAQs

We've compiled the feedback and wisdom of other InClass hosts so you don't have to reinvent the bell curve.

[Learn more](#)

01.02

•••

Introducción

Entender su estructura.

Detalles.



Points This competition does not award standard [ranking points](#)

Tiers This competition does not count towards [tiers](#)

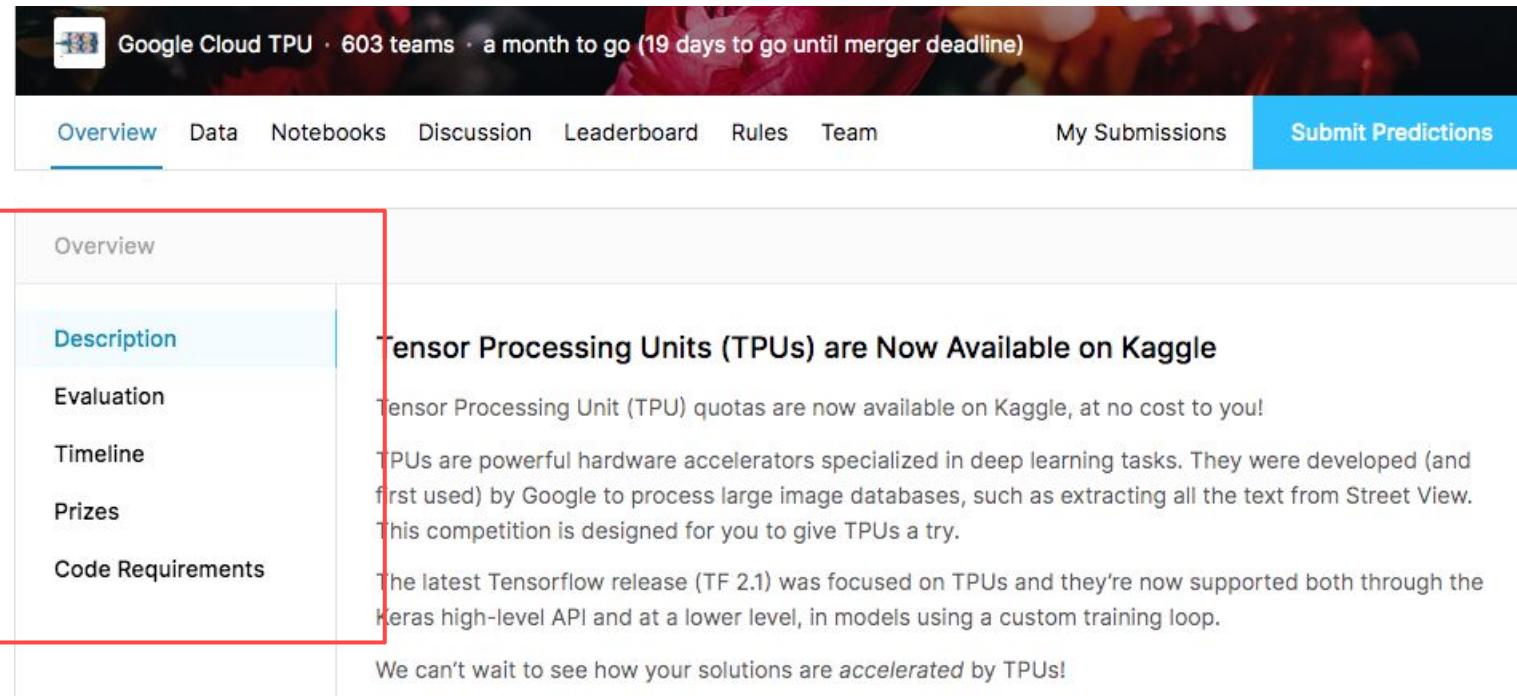
Tags

[image data](#)

[plants](#)

[macrofscore](#)

Condiciones.



The screenshot shows a competition page for "Google Cloud TPU". The top navigation bar includes links for Overview, Data, Notebooks, Discussion, Leaderboard, Rules, Team, My Submissions, and Submit Predictions. The "Overview" tab is selected. A red box highlights the left sidebar containing sections for Description, Evaluation, Timeline, Prizes, and Code Requirements. The main content area displays a bold announcement about TPUs being available on Kaggle, followed by detailed descriptions of each section.

Google Cloud TPU · 603 teams · a month to go (19 days to go until merger deadline)

Overview Data Notebooks Discussion Leaderboard Rules Team My Submissions Submit Predictions

Overview

Description

Tensor Processing Units (TPUs) are Now Available on Kaggle

Evaluation

Tensor Processing Unit (TPU) quotas are now available on Kaggle, at no cost to you!

Timeline

TPUs are powerful hardware accelerators specialized in deep learning tasks. They were developed (and first used) by Google to process large image databases, such as extracting all the text from Street View. This competition is designed for you to give TPUs a try.

Prizes

Code Requirements

The latest Tensorflow release (TF 2.1) was focused on TPUs and they're now supported both through the Keras high-level API and at a lower level, in models using a custom training loop.

We can't wait to see how your solutions are accelerated by TPUs!

Métricas evaluación.

Cost Functions

Root Mean Squared Error (RMSE)

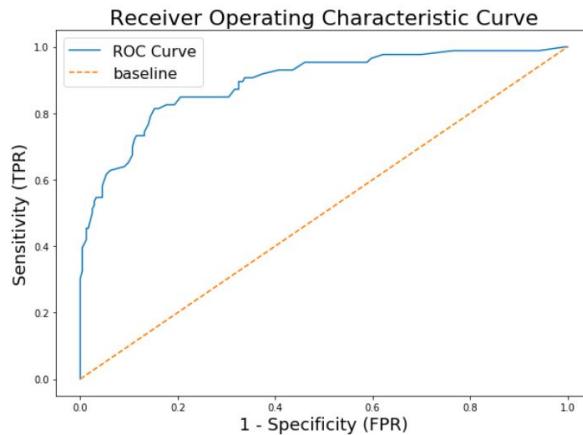
$$\sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2}$$

actual

prediction

Root Mean Squared Log Error (RMSLE)

$$\sqrt{\frac{1}{n} \sum_{i=1}^n (\log(p_i + 1) - \log(a_i + 1))^2}$$



Submissions are evaluated on [macro F1 score](#).

F1 is calculated as follows:

$$F_1 = 2 * \frac{precision * recall}{precision + recall}$$

where:

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

In "macro" F1 a separate F1 score is calculated for each class / label and then averaged.

Submission.



This is a Code Competition

Submissions to this competition must be made by an output from a Kaggle Notebook/Script.

- Notebook run-time per session capped at 3 hours.
- CPU, GPU, or TPU all allowed. Only TPU submissions eligible to win [prizes](#).
- The submission file must contain predictions for the test set, following the format of the [sample_submission.csv](#) file on the [Data Page](#).
- Your code will not be re-run, as there is no completely hidden test set.

Please see the [Code Competition FAQ](#) for more information on how to submit.

sample_submission.csv (86.52 KB)		
A	id	# label
	7382	
	unique values	
	0	0
1	b48c962e0	0
2	a13d3dfa4	0
3	94269c190	0
4	bcb18c6e4	0
5	d15a4d94c	0
6	914b0e71b	0
7	065c9b5be	0
8	dd8b1aac0	0
9	8c39d1b41	0
10	e2fdce920	0
11	3af4c8105	0
12	2c73bd62a	0

Leaderboard.

Public Leaderboard Private Leaderboard

This leaderboard is calculated with approximately 70% of the test data.
The final results will be based on the other 30%, so the final standings may be different.

[Raw Data](#) [Refresh](#)

#	Team Name	Notebook	Team Members	Score ⓘ	Entries	Last
1	HengCherKeng+huiqin		 	0.98424	102	3d
2	Loc Truong		 	0.97457	21	6d
3	Chris Deotte		 	0.97109	56	1mo
4	David Young		 	0.96947	29	6h
5	Dr.Young		 	0.96946	5	5h
6	Zeeshan Arif		 	0.96935	3	10d
7	Mathurin Aché		 	0.96933	44	1h
8	Md. Redwan Karim Sony		 	0.96921	8	13h
9	Kamal Chhirang		 	0.96833	25	4d
10	Thomas Brekk Unnvik		 	0.96819	17	1mo

Notebook.

GPU quota: 30h remaining | TPU quota: 30h remaining

Public Your Work Shared With You Favorites Sort by Hotness

Outputs Languages Tags Search notebooks

#	User	Title	Age	Score	Tags	Outputs	Languages	Tags	Comments
9		Introduction kernel & what's EfficientNet?	41m ago	0.95849	tpu		Py		2
9		Top Scoring Kernels: flower-classification-tpus	8h ago				Py		0
0		Flower Classification with TPU	6h ago		tpu		Py		0
210		Getting started with 100+ flowers on TPU	1mo ago	0.25443			Py		41
115		Flowers TPU: Concise EfficientNet B7	1mo ago	0.95914	cnn, starter code		Py		25

Cursos.

	Python Learn the most important language for data science. 
	Intro to Machine Learning Learn the core ideas in machine learning, and build your first models. 
	Intermediate Machine Learning Learn to handle missing values, non-numeric values, data leakage and more. Your models will be more accurate and useful.
	Data Visualization Make great data visualizations. A great way to see the power of coding!
	Pandas Solve short hands-on challenges to perfect your data manipulation skills. 
	Feature Engineering Discover the most effective way to improve your models.

	Deep Learning Use TensorFlow to take machine learning to the next level. Your new skills will amaze you. 
	Intro to SQL Learn SQL for working with databases, using Google BigQuery to scale to massive datasets. 
	Advanced SQL Take your SQL skills to the next level.
	Geospatial Analysis Create interactive maps, and discover patterns in geospatial data.
	Microchallenges Solve ultra-short challenges to build and test your skill.
	Machine Learning Explainability Extract human-understandable insights from any machine learning model.
	Natural Language Processing Distinguish yourself by learning to work with text data.

01.03

•••

Introducción

Metodologías.

Cómo participar a una competición.

Es un flujo de buenas prácticas ideal.



Identificar el problema.

El paso más importante para poder participar a una competición. Realizar conocimiento del contexto general.

Colaboración.

Realizar tareas en equipo multidisciplinar para poder avanzar sin obstáculos.

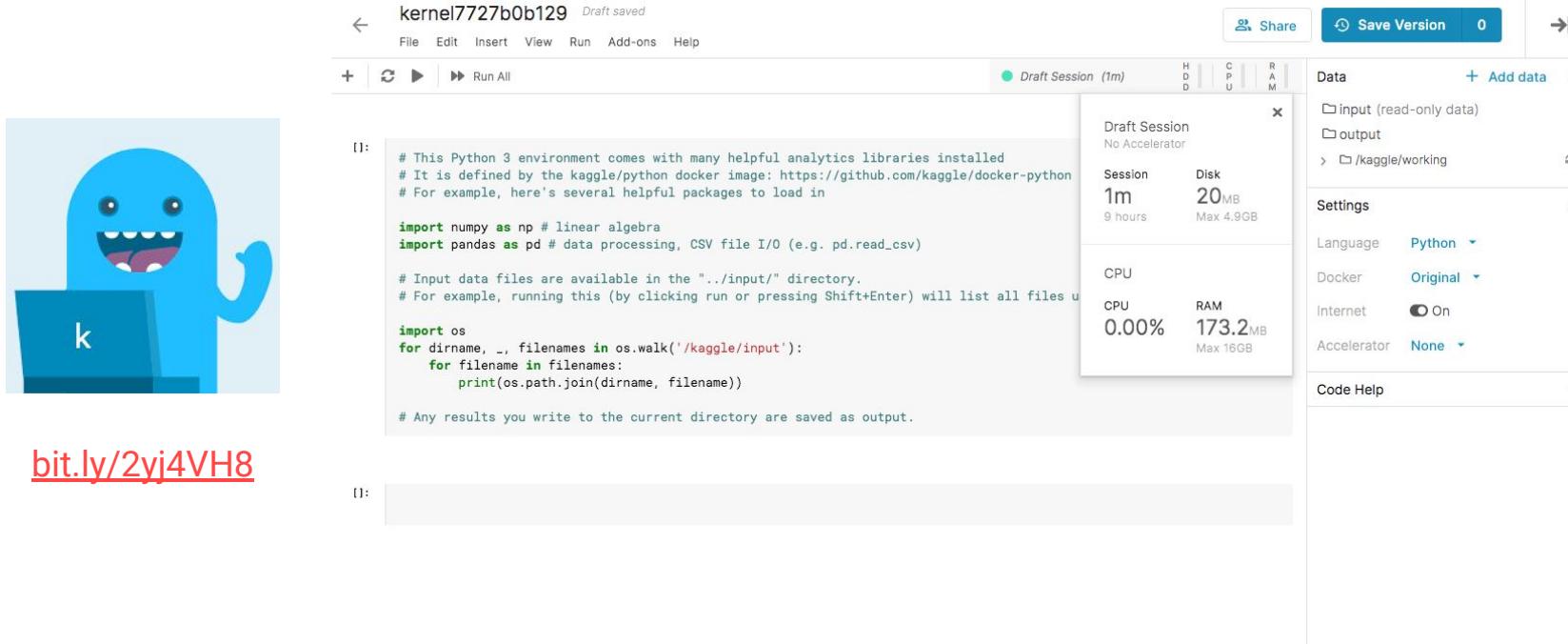
Prueba-Error.

Lo ideal es diseñar un flujo de algoritmos, ajustando parámetros y ensayando, evaluar los resultados y mejorarllos.

Despliegue.

Enviar los resultados, poner en práctica nuevas técnicas y participando activamente en la comunidad científica.

Entornos de trabajo - starter.



Draft saved

File Edit Insert View Run Add-ons Help

Share Save Version 0

Draft Session (1m) H D C P U R A M

Draft Session
No Accelerator

Session 1m Disk 20 MB
9 hours Max 4.9GB

CPU

CPU RAM
0.00% 173.2 MB
Max 16GB

Data + Add data

input (read-only data)

output

/kaggle/working

Settings

Language Python

Docker Original

Internet On

Accelerator None

Code Help

```
# This Python 3 environment comes with many helpful analytics libraries installed
# It is defined by the kaggle/python docker image: https://github.com/kaggle/docker-python
# For example, here's several helpful packages to load in

import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)

# Input data files are available in the "../input/" directory.
# For example, running this (by clicking run or pressing Shift+Enter) will list all files under the input directory

import os
for dirname, _, filenames in os.walk('../kaggle/input'):
    for filename in filenames:
        print(os.path.join(dirname, filename))

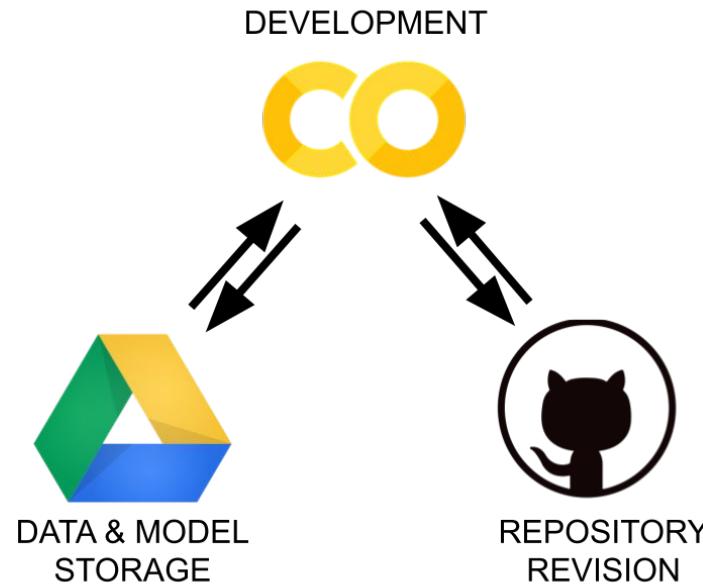
# Any results you write to the current directory are saved as output.
```

bit.ly/2yj4VH8

Entornos de trabajo - starter.



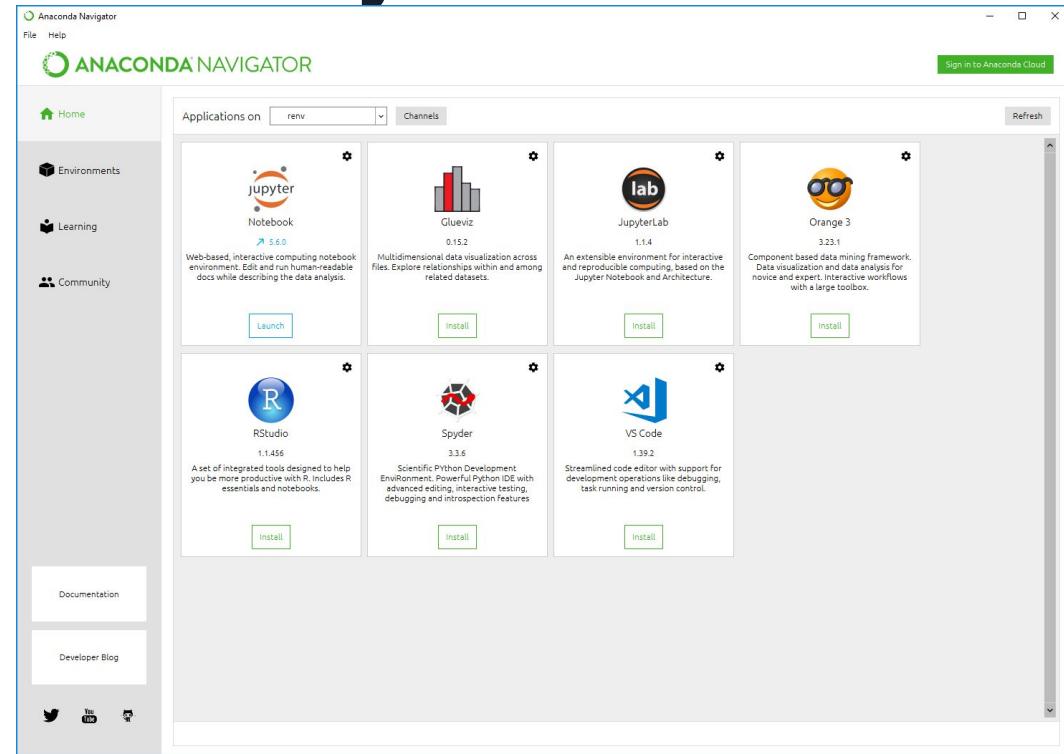
bit.ly/34DoZ30



Entornos de trabajo - newbie.



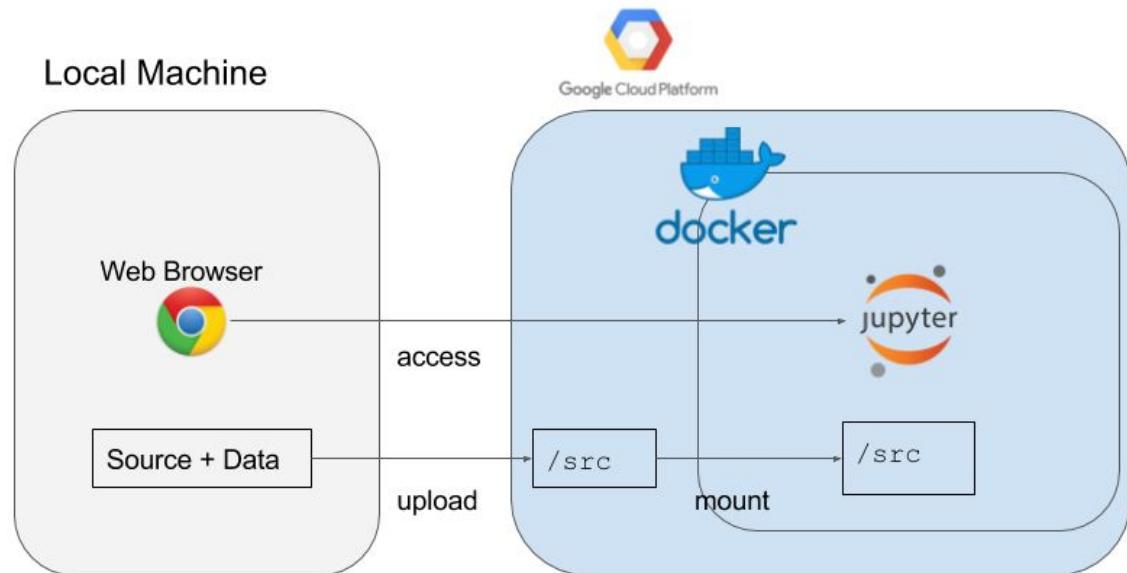
bit.ly/2XDVVac



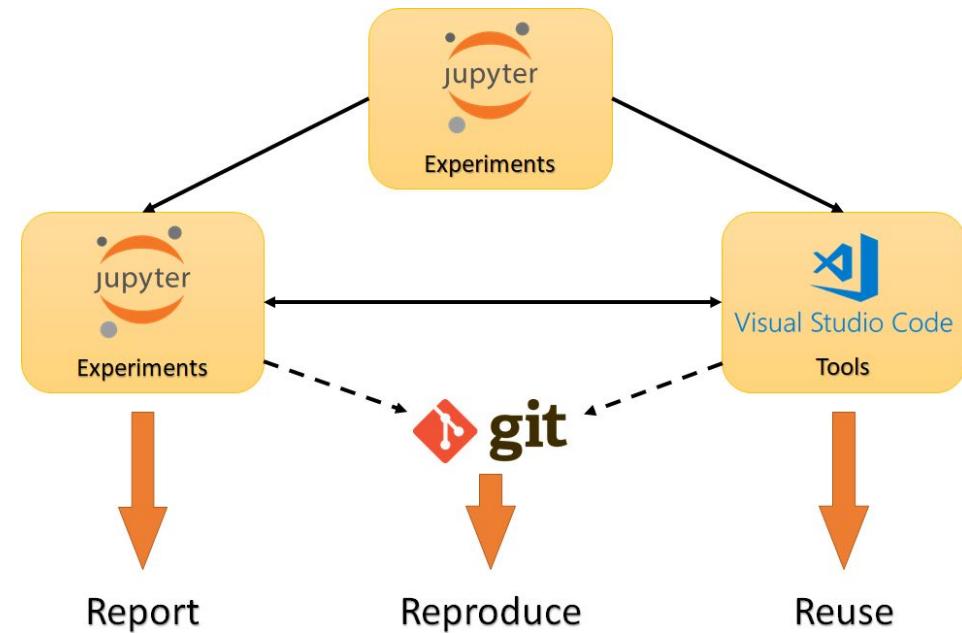
Entornos de trabajo - mid.



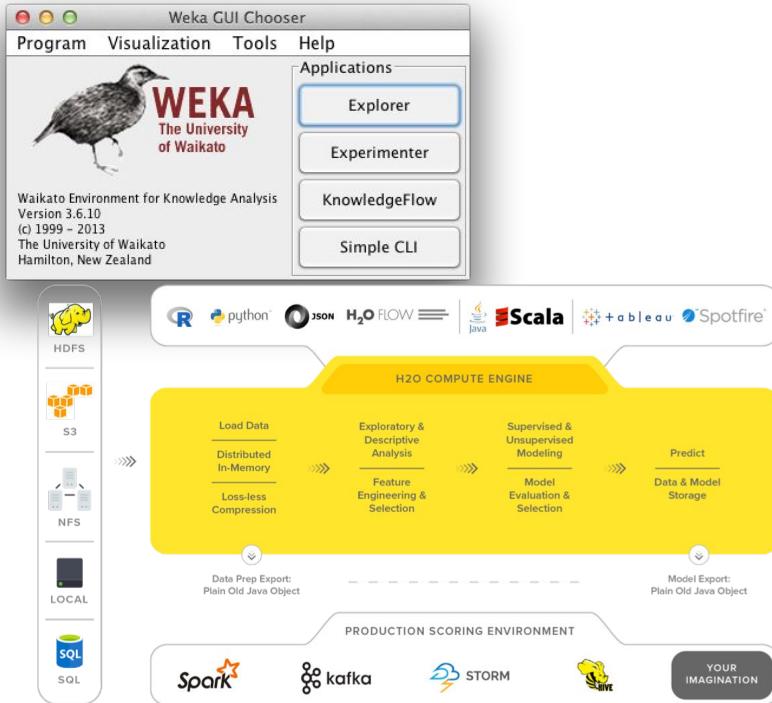
dockr.ly/3ewelQr



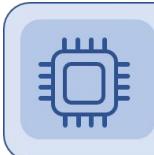
Entornos de trabajo - pro.



Entornos de trabajo - biz.

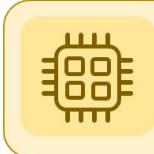


¿Y, los recursos?



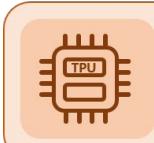
CPU

- Small models
- Small datasets
- Useful for design space exploration



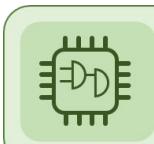
GPU

- Medium-to-large models, datasets
- Image, video processing
- Application on CUDA or OpenCL



TPU

- Matrix computations
- Dense vector processing
- No custom TensorFlow operations



FPGA

- Large datasets, models
- Compute intensive applications
- High performance, high perf./cost ratio

Select new notebook settings

You can change these settings at any time



Select language

Python ▾



Select type



Notebook

Ideal for interactive data exploration and polished analysis. Shares insights through code & commentary



Script

Ideal for fitting a model and competition submissions. Shares code for review and RMarkdown reports

HIDE ADVANCED SETTINGS



Enable Google Cloud Services

Link a Google account to access Google Cloud services.



On Off



Accelerator



- ✓ None
- GPU
- TPU v3-8

Create

Librerías.

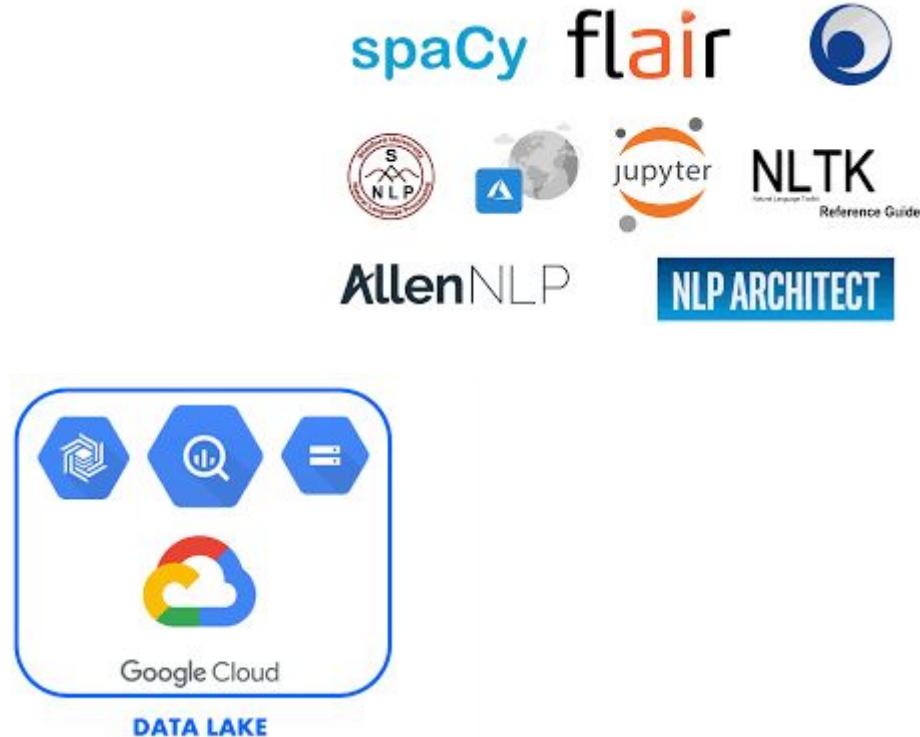


IP[y]: IPython
Interactive Computing



Librerías.

K Keras TensorFlow
Caffe PyTorch



Datasets - tamaño.

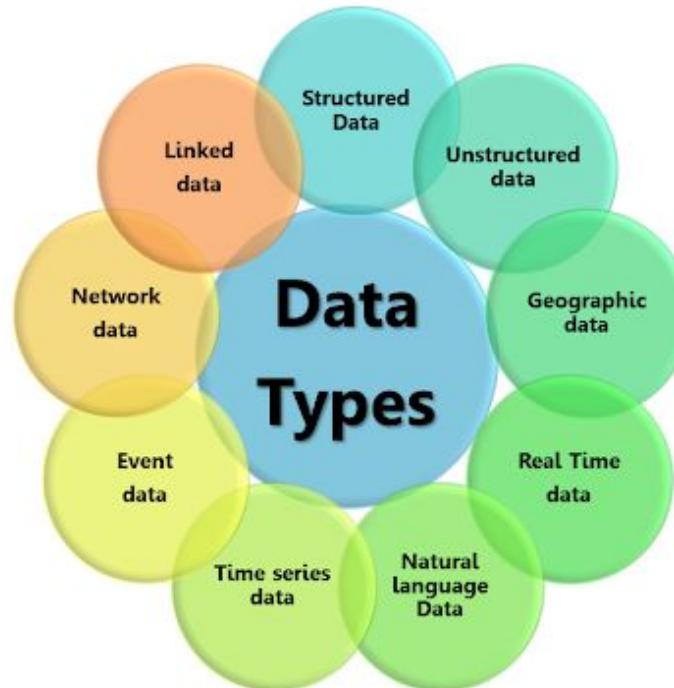


Zip ~1 MB



Zip ~6 GB

Datasets - tipo de dato.



...además de
muchas paciencias....

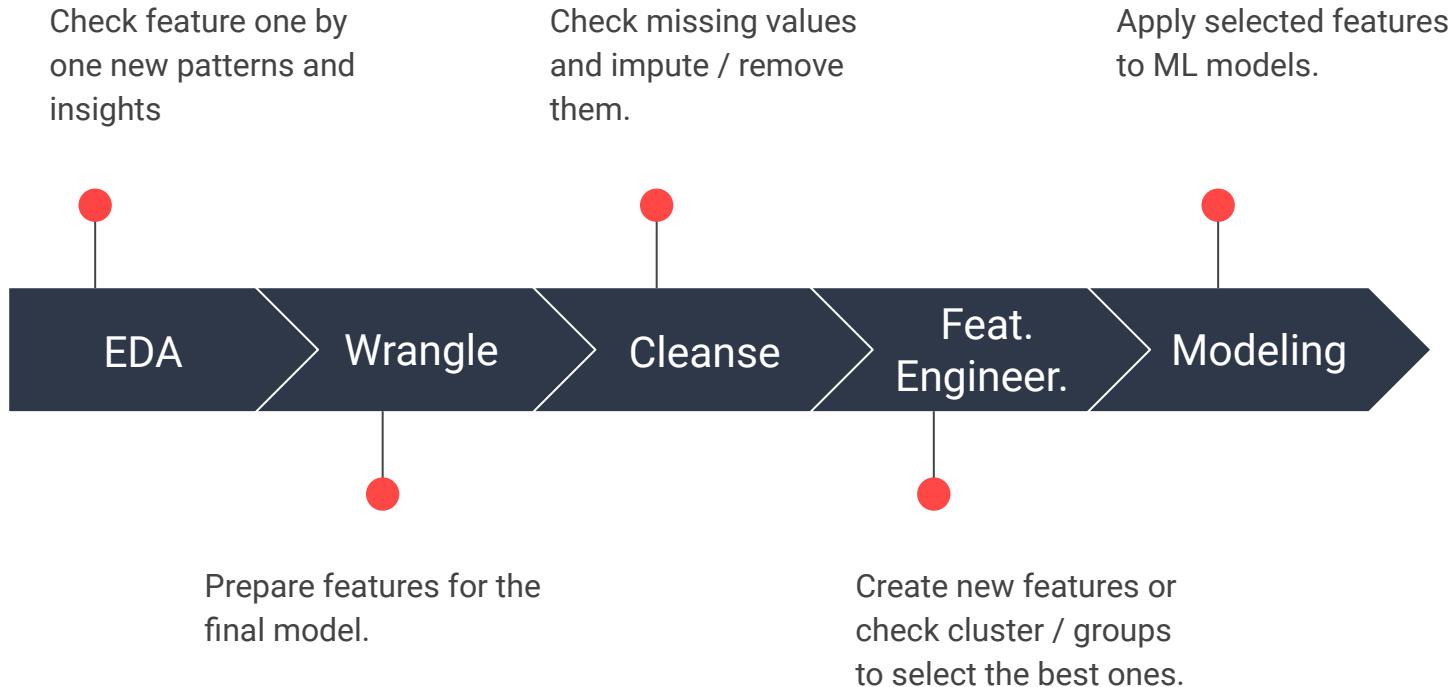


01.04

••• Introducción

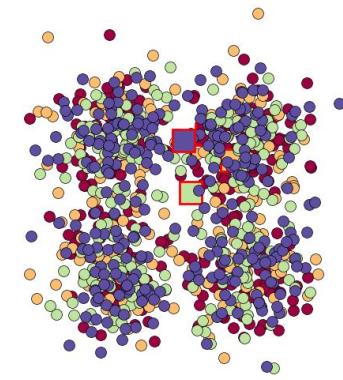
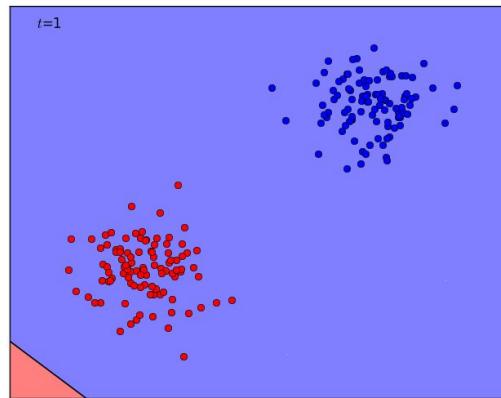
Flujo de trabajo.

Flujo de trabajo.

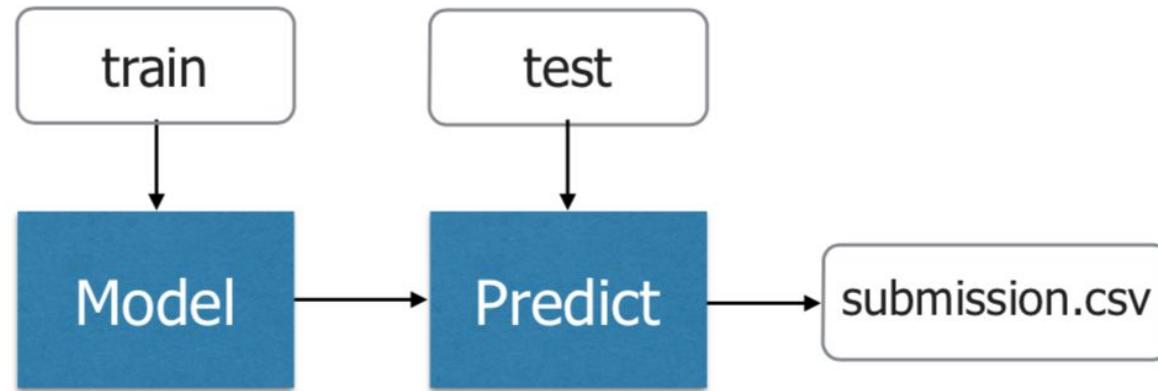


Tipos de ML.

- Supervisado
 - All labeled data
 - E.g. classification
 - Is this a chihuahua or a muffin?
- Semi-supervisado
 - Some labeled data, lots of unlabeled data
 - When labeling is expensive
- No Supervisado
 - Discover patterns without labels
 - E.g. clustering



Build prediction model.



Calculate CV to
cross-validate

Cómo funciona.

- “Train” a **model** on lots and lots of data
 - Start with poor predictions
 - Make little tweaks to improve
 - Like child doing homework!
- Infer predictions on new data



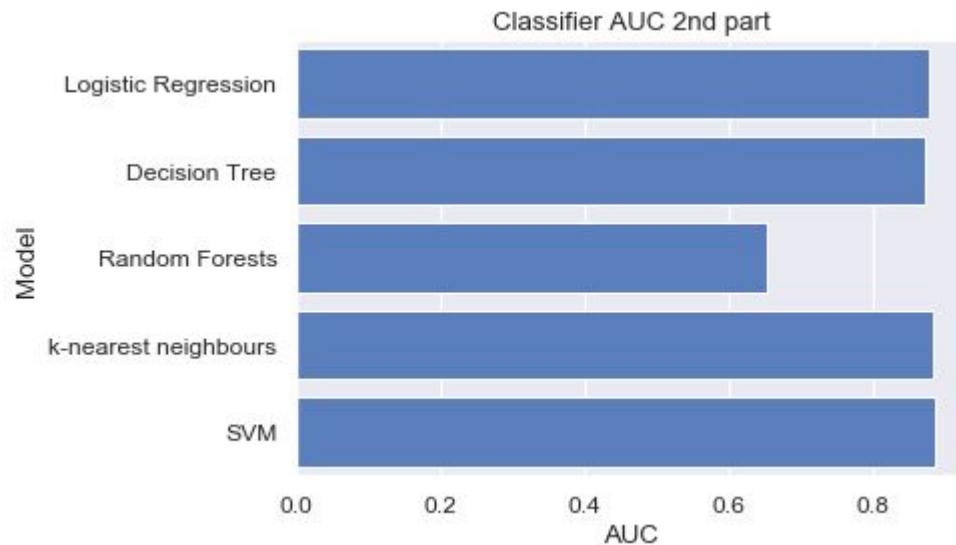
Training



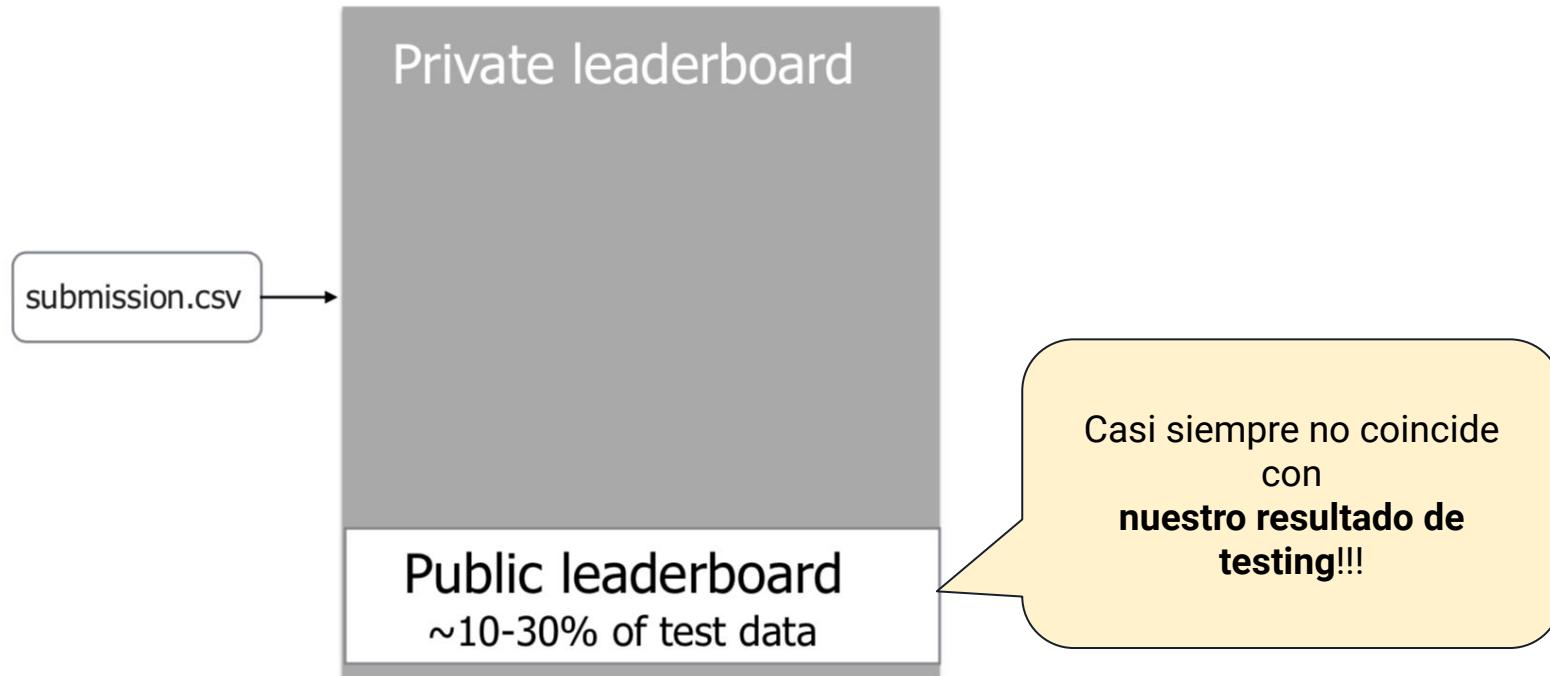
Inference

Mejora continua.

	Model	Score_1st
4	Decision Tree	99.99
1	KNN	98.27
0	Support Vector Machines	97.69
2	Logistic Regression	97.08
3	Random Forest	93.86



Submit.



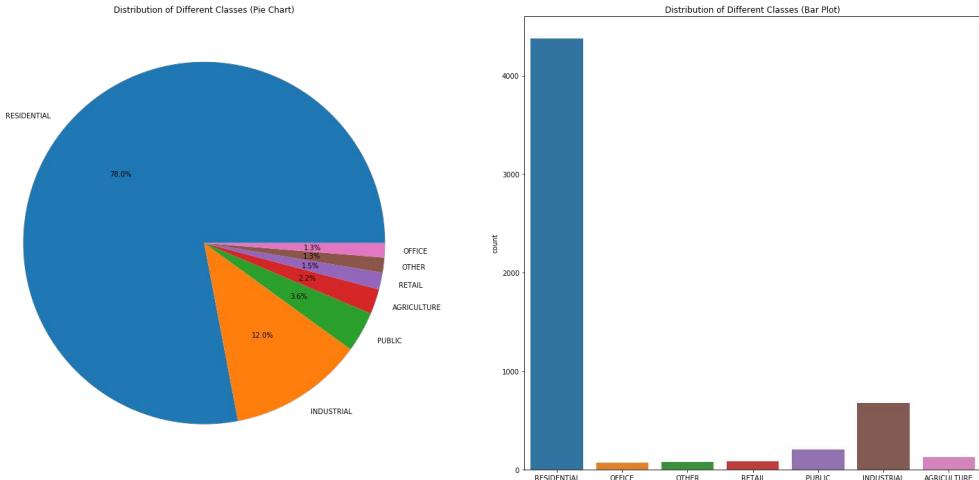
01.05

• • •

Introducción

Tips.

EDA y correlaciones.



Pearson Correlation Matrix

	CLASE_grp	AREA	MAXBUILDINGFLOOR	Q_B_2_0_9	Q_B_2_0_8	Q_G_3_0_9	Q_NIR_8_0_0	Q_G_3_0_6	Q_B_2_0_7	Q_G_3_0_7	Q_R_4_0_9	Q_G_3_0_6	Q_B_2_0_6	Q_R_4_0_8	Q_G_3_0_5	Q_G_3_0_0
CLASE_grp	1.00	0.20		0.10	0.10	0.09	0.09	0.08	0.08	0.08	0.07	0.07	0.07	0.06	0.06	0.05
AREA	0.20	1.00		0.08	0.08	0.08	-0.00	0.07	0.07	0.07	0.06	0.06	0.07	0.06	0.06	-0.00
MAXBUILDINGFLOOR				0.10	0.08	0.96	0.97	0.96	0.92	0.92	0.89	0.85	0.84	0.84	0.80	-0.02
Q_B_2_0_9	0.10	0.08		1.00	0.97	0.96	-0.11	0.92	0.92	0.89	0.85	0.84	0.84	0.80	0.79	-0.02
Q_B_2_0_8	0.10	0.08		0.97	1.00	0.97	-0.12	0.96	0.98	0.94	0.90	0.90	0.92	0.87	0.85	-0.02
Q_G_3_0_9	0.09	0.08		0.96	0.97	1.00	-0.13	0.98	0.95	0.95	0.94	0.92	0.91	0.90	0.87	-0.02
Q_NIR_8_0_0	0.09	-0.00		-0.11	-0.12	-0.13	1.00	-0.12	-0.14	-0.13	-0.17	-0.13	-0.15	-0.16	-0.13	0.42
Q_G_3_0_8	0.08	0.07		0.92	0.96	0.98	-0.12	1.00	0.97	0.99	0.95	0.96	0.94	0.93	0.92	-0.02
Q_B_2_0_7	0.08	0.07		0.92	0.98	0.95	-0.14	0.97	1.00	0.96	0.93	0.93	0.98	0.93	0.89	-0.03
Q_G_3_0_7	0.08	0.07		0.89	0.94	0.95	-0.13	0.99	0.96	1.00	0.94	0.99	0.95	0.94	0.97	-0.02
Q_R_4_0_9	0.07	0.06		0.85	0.90	0.94	-0.17	0.95	0.93	0.94	1.00	0.92	0.94	0.98	0.89	-0.03
Q_G_3_0_6	0.07	0.06		0.84	0.90	0.92	-0.13	0.96	0.93	0.99	0.92	1.00	0.95	0.94	0.99	-0.01
Q_B_2_0_6	0.07	0.07		0.84	0.92	0.91	-0.15	0.94	0.98	0.95	0.94	0.95	1.00	0.95	0.92	-0.03
Q_R_4_0_8	0.06	0.06		0.80	0.87	0.90	-0.16	0.93	0.93	0.94	0.98	0.94	0.95	1.00	0.92	-0.03
Q_G_3_0_5	0.06	0.06		0.79	0.85	0.87	-0.13	0.92	0.89	0.97	0.89	0.99	0.92	0.92	1.00	-0.01
Q_G_3_0_0	0.05	-0.00		-0.02	-0.02	-0.02	0.42	0.02	-0.03	-0.02	-0.03	-0.01	-0.03	-0.03	-0.01	1.00

Features.

Feature Engineering

Step1.

Feature Selection

choose efficient features
and
abandon useless features

Step2.

Feature Extraction (Dimensional Reduction)

PCA

SVD

LDA

DEMO TIME.



Webinar.

<https://youtu.be/UzwRO4hj8c8>





¡Muchas
Gracias!