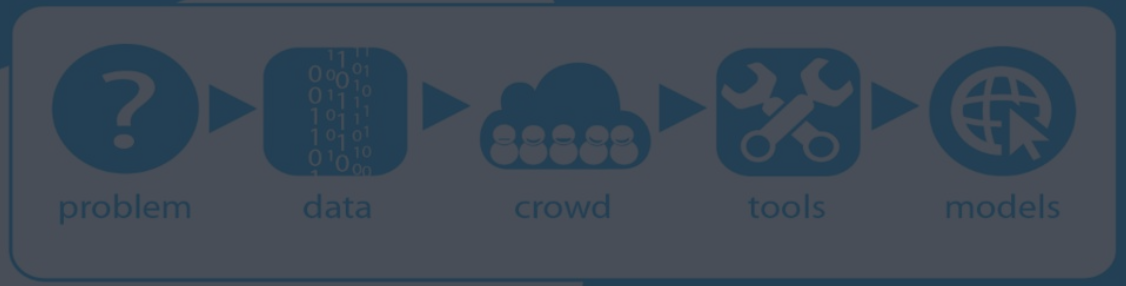
Paradigma

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

github.com/marcusRB

marcusRB

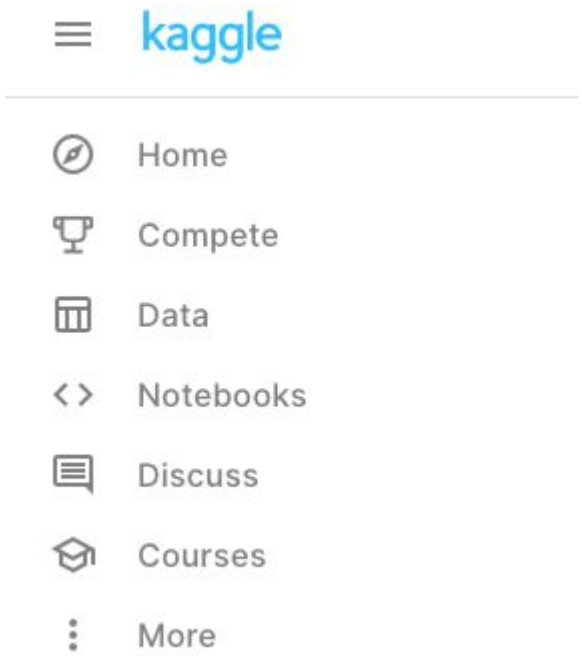
01.01

Introducción

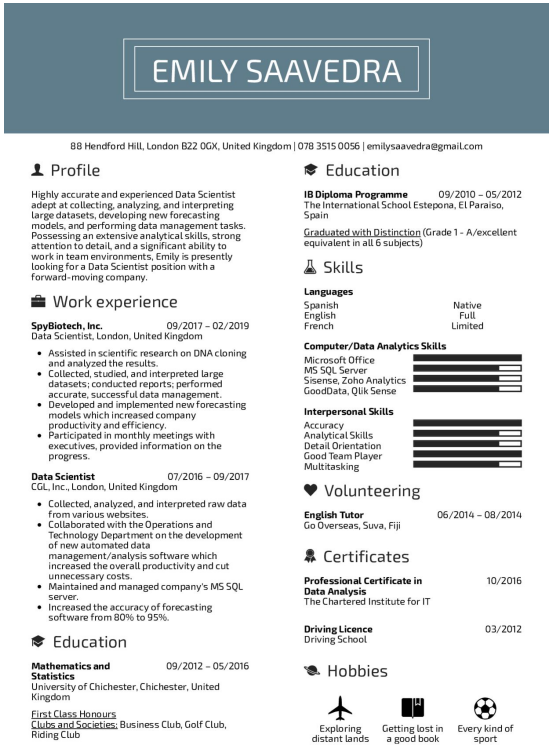
Visión general.

¿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**



¿Qué es Kaggle?



EMILY SAAVEDRA

88 Hendford Hill, London B22 0GX, United Kingdom | 078 3515 0056 | emilysaavedra@gmail.com

Profile

Highly accurate and experienced Data Scientist adept at collecting, analyzing, and interpreting large datasets, developing new forecasting models, and performing data management tasks. Possessing an extensive analytical skills, strong attention to detail, and a significant ability to work in team environments, Emily is presently looking for a Data Scientist position with a forward-moving company.

Work experience

SpyBiotech, Inc. 09/2017 – 02/2019
Data Scientist, London, United Kingdom

- Assisted in scientific research on DNA cloning and analyzed the results.
- Collected, studied, and interpreted large datasets; conducted reports; performed accurate, successful data management.
- Developed and implemented new forecasting models which increased company productivity and efficiency.
- Participated in monthly meetings with executives, provided information on the progress.

Data Scientist 07/2016 – 09/2017
CGL, Inc., London, United Kingdom

- Collected, analyzed, and interpreted raw data from various websites.
- Collaborated with the Operations and Technology Department on the development of new automated data management/analysis software which increased the overall productivity and cut unnecessary costs.
- Maintained and managed company's MS SQL server.
- Increased the accuracy of forecasting software from 80% to 95%.

Education

Mathematics and Statistics 09/2012 – 05/2016
University of Chichester, Chichester, United Kingdom

First Class Honours
Clubs and Societies: Business Club, Golf Club, Riding Club

Education

IB Diploma Programme 09/2010 – 05/2012
The International School Estepona, El Paraiso, Spain

Graduated with Distinction (Grade 1 - A/excellent equivalent in all 6 subjects)

Skills

Languages

Spanish	Native
English	Full
French	Limited

Computer/Data Analytics Skills

Microsoft Office
MS SQL Server
Sisense, Zoho Analytics
GoodData, Qlik Sense

Interpersonal Skills

Accuracy
Analytical Skills
Detail Orientation
Good Team Player
Multitasking

Volunteering

English Tutor 06/2014 – 08/2014
Go Overseas, Suva, Fiji

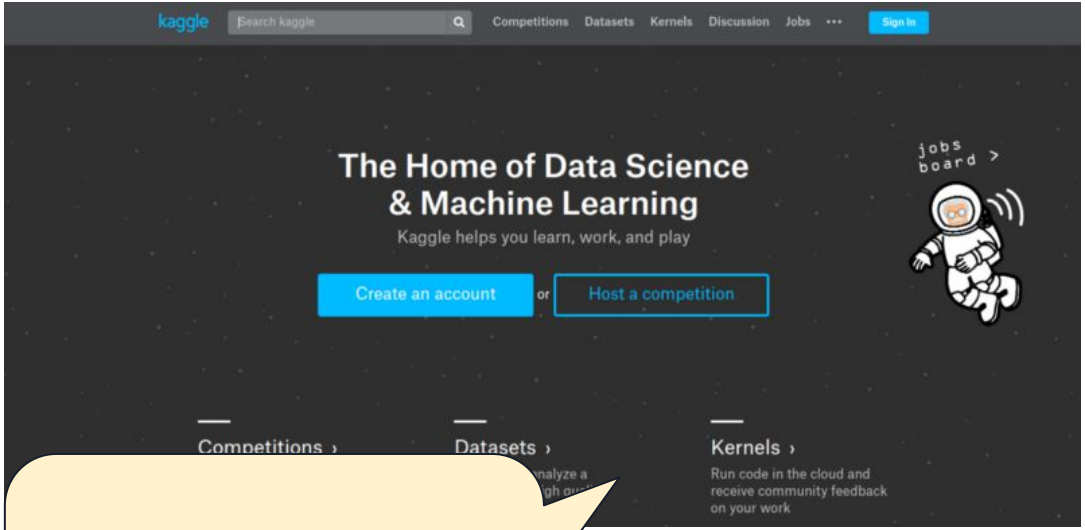
Certificates

Professional Certificate in Data Analysis 10/2016
The Chartered Institute for IT

Driving Licence 03/2012
Driving School

Hobbies

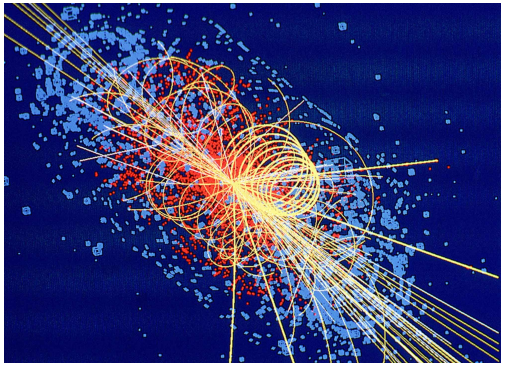
Exploring distant lands
Getting lost in a good book
Every kind of sport



Tu **portfolio**

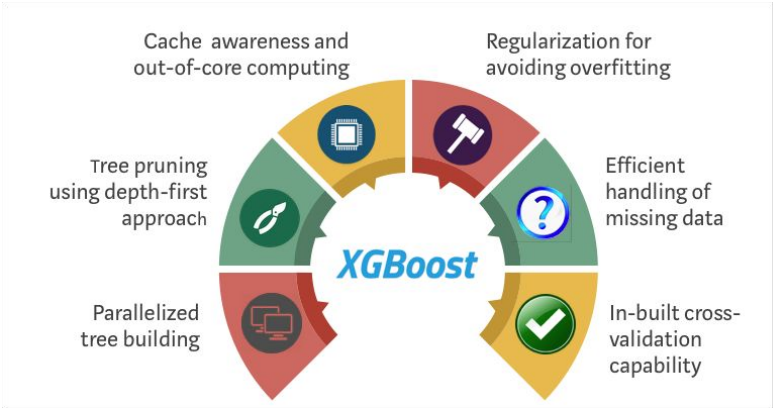
¿Qué es Kaggle?

fundada en 2010 y adquirida por Google en 2017



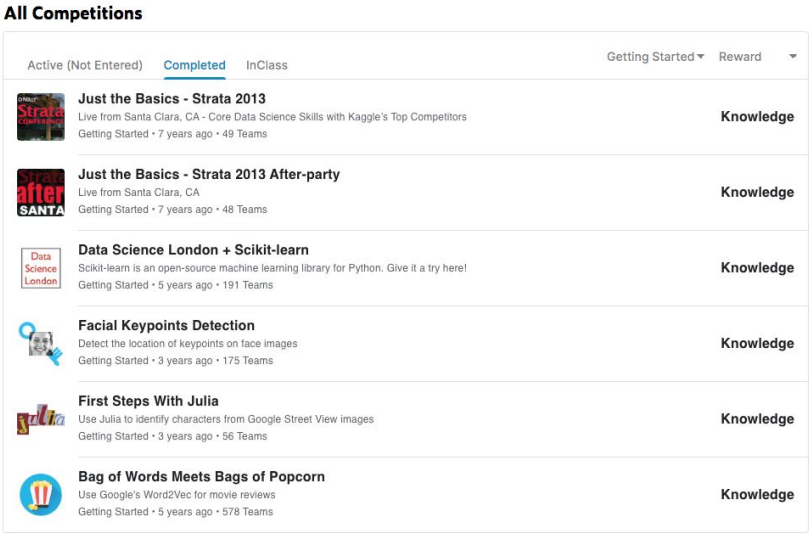
Walmart

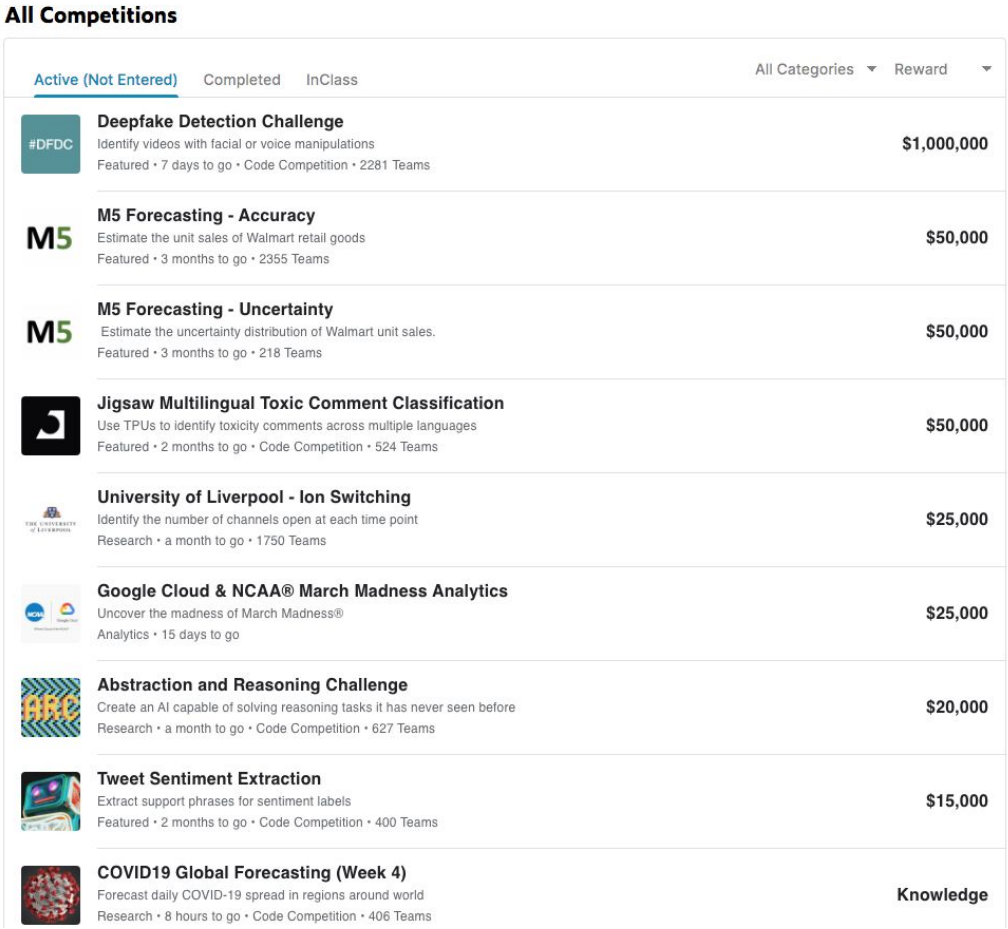
KINECT



¿Qué es Kaggle?

¿Qué es Kaggle?





¿Qué es Kaggle?

bit.ly/2XD71Mz

Datathon, Competiciones internas, académicas, etc.

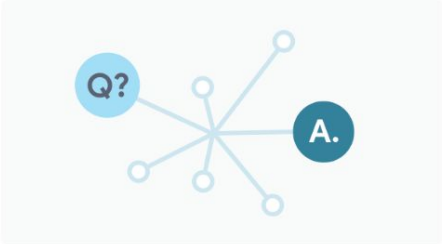
Kaggle InClass competitions make machine learning fun.

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.

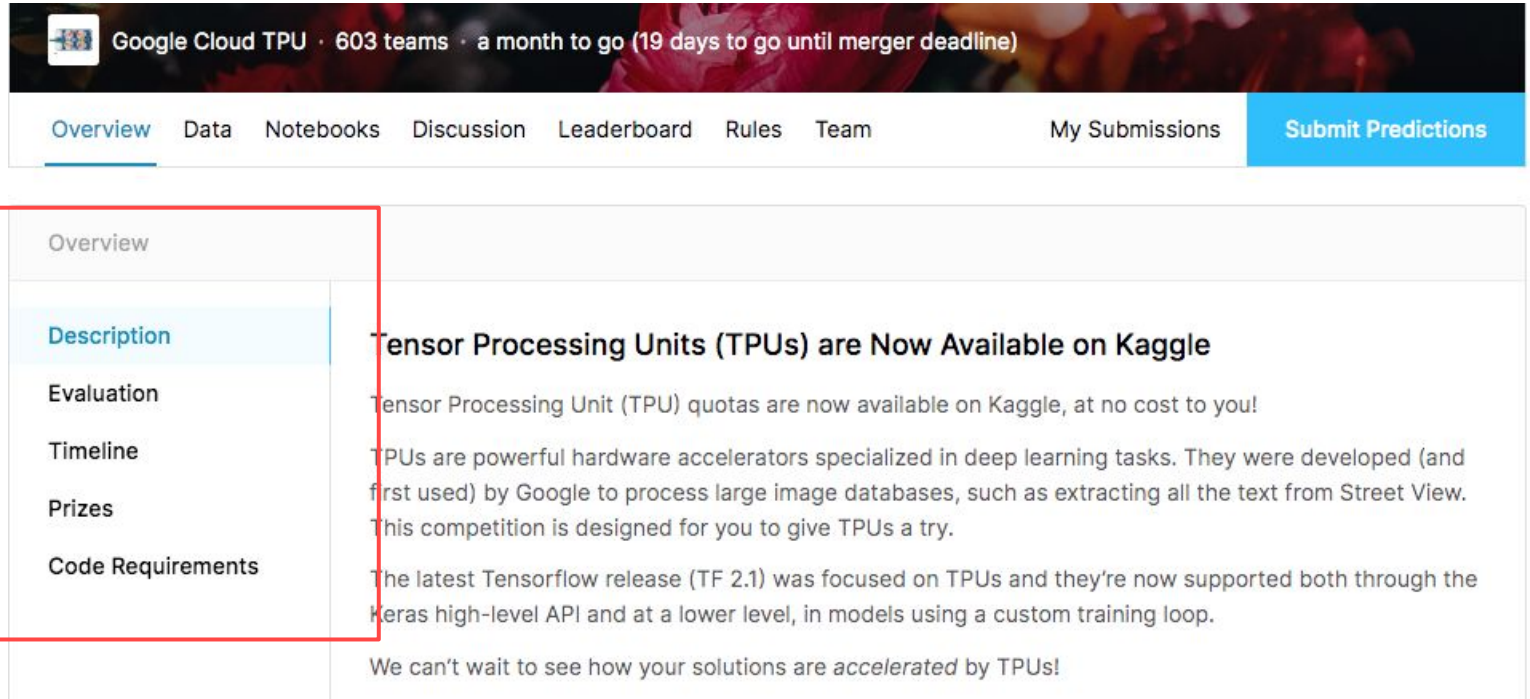
603	666	5,882
Teams	Competitors	Entries

Points This competition does not award standard ranking points

Tiers This competition does not count towards tiers

Tags image data plants macrofscore

Condiciones.



Métricas evaluación.

Cost Functions

Root Mean Squared Error (RMSE)

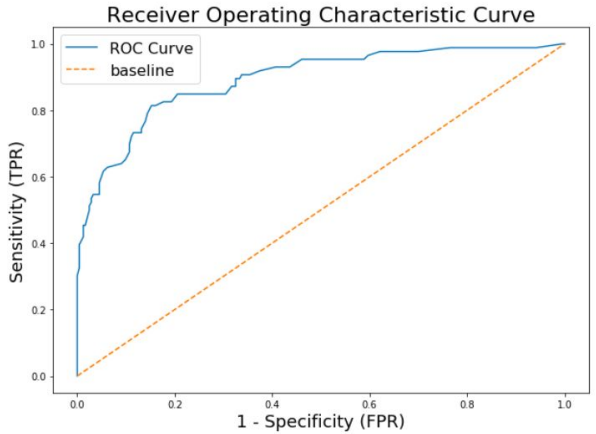
$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Root Mean Squared Log Error (RMSLE)

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

prediction

actual



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)

	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	dd8b1aaca	0
9	8c39d1b41	0
10	e2fdce920	0
11	3af4c8105	0
12	2c73bd62a	0

Leaderboard.

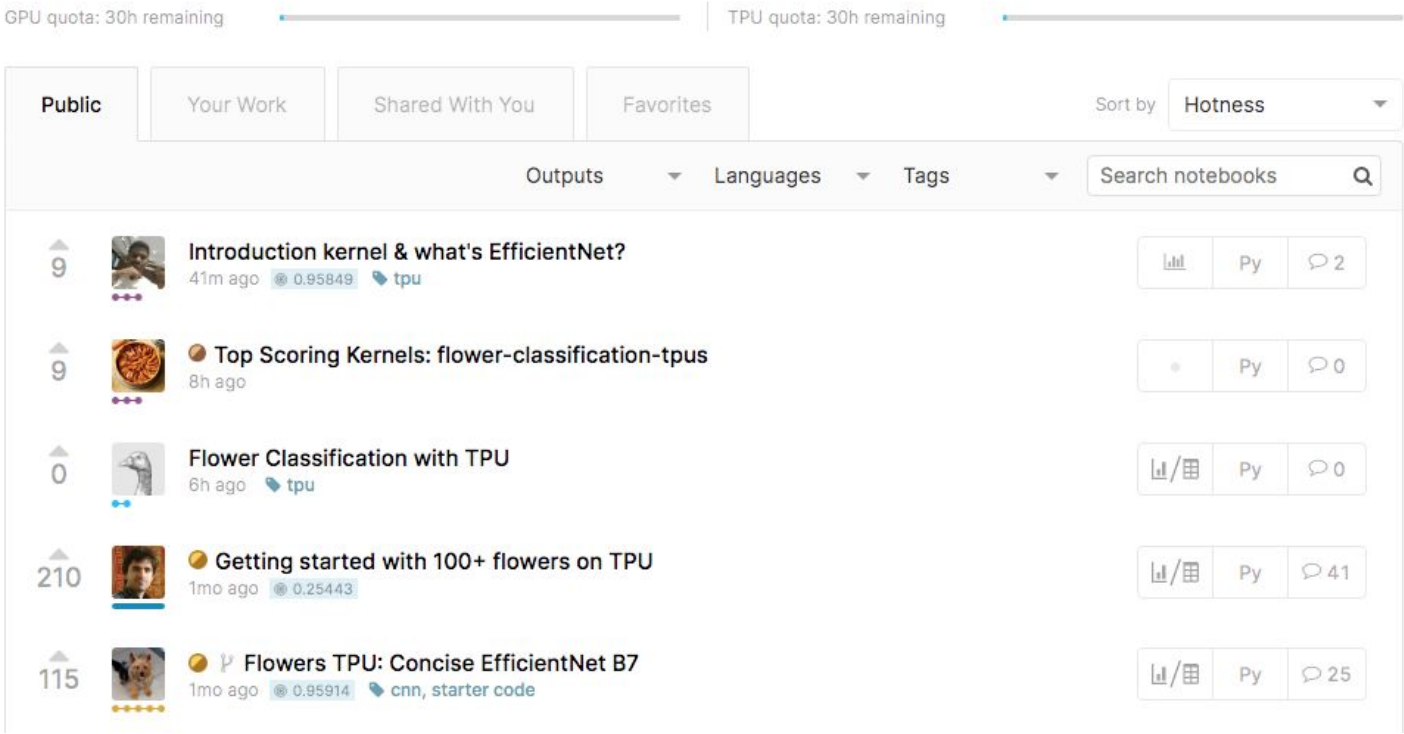
Public Leaderboard | Private Leaderboard

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.

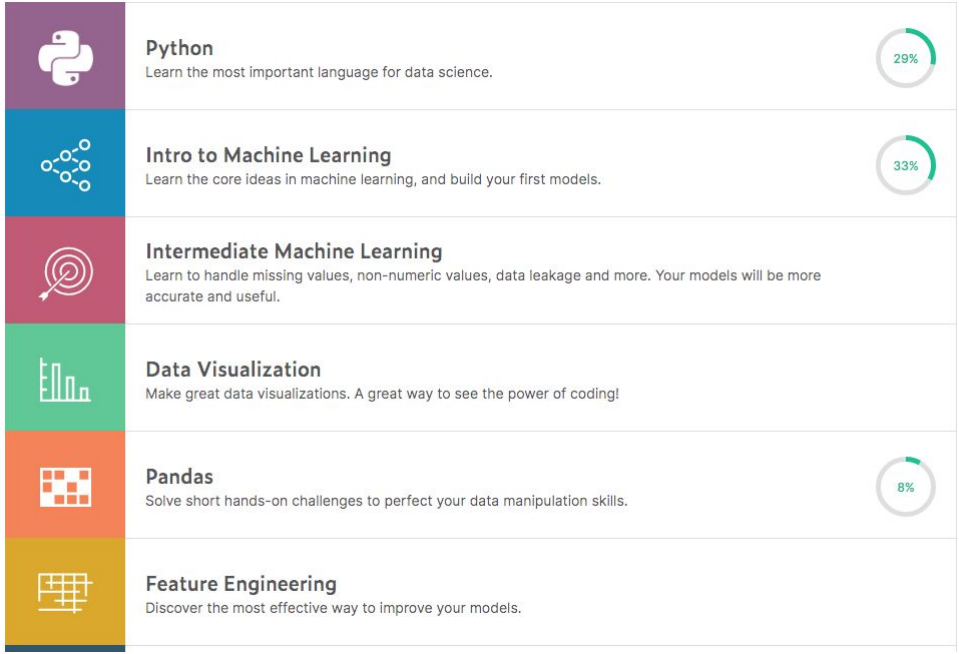
Raw Data | Refresh

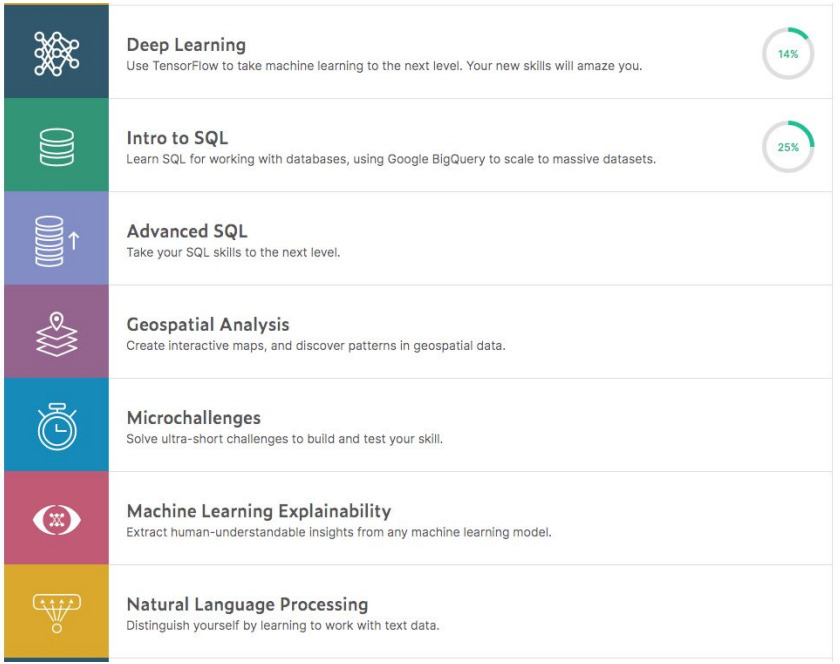
#	Team Name	Notebook	Team Members	Score	Entries	Last
1	HengCherKeng+huiqin			0.98424	102	3d
2	Loc Truong			0.97457	21	6d
3	Chris Deotte			0.97109	56	1mo
4	David Young			0.96947	29	6h
5	Dr.Young			0.96946	5	5h
6	Zeeshan Arif			0.96935	3	10d
7	Mathurin Aché			0.96933	44	1h
8	Md. Redwan Karim Sony			0.96921	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
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.
29%
Intro to Machine Learning
Learn the core ideas in machine learning, and build your first models.
33%
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.
8%
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.
14%
Intro to SQL
Learn SQL for working with databases, using Google BigQuery to scale to massive datasets.
25%
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.

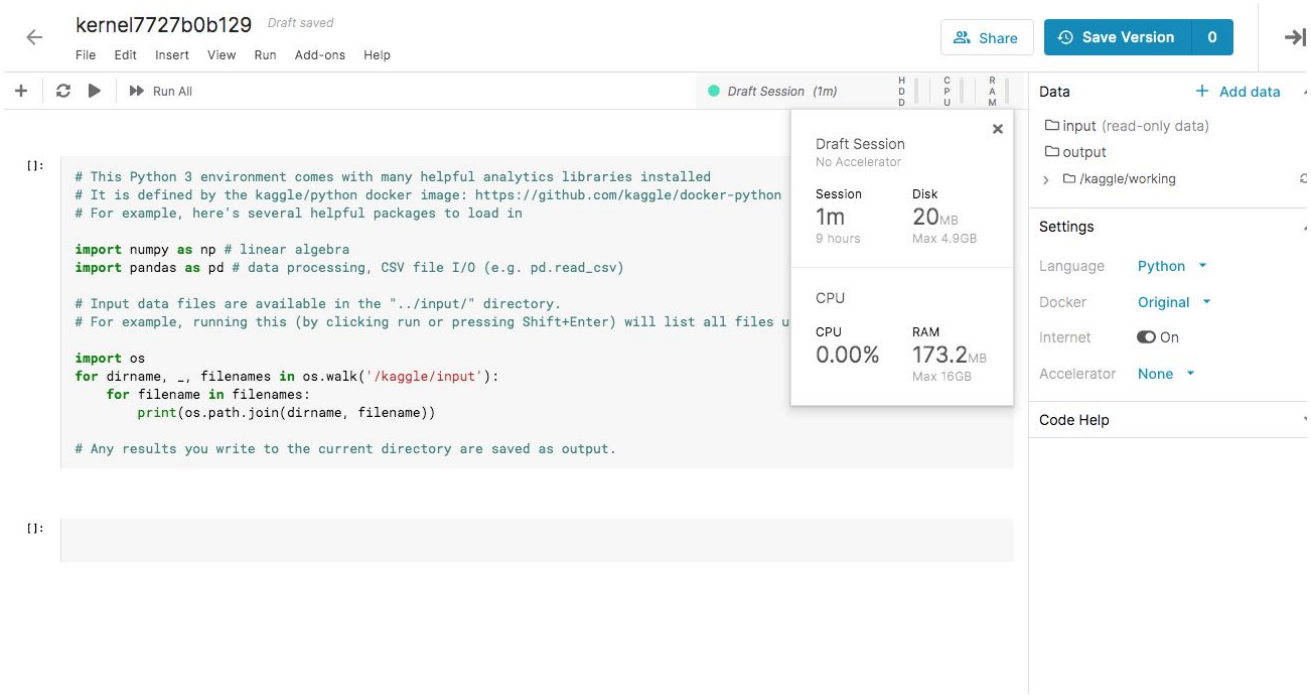
Lo ideal es diseñar un flujo de algoritmos, ajustando parámetros y ensayando, evaluar los resultados y mejorarlos.

Despliegue.

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

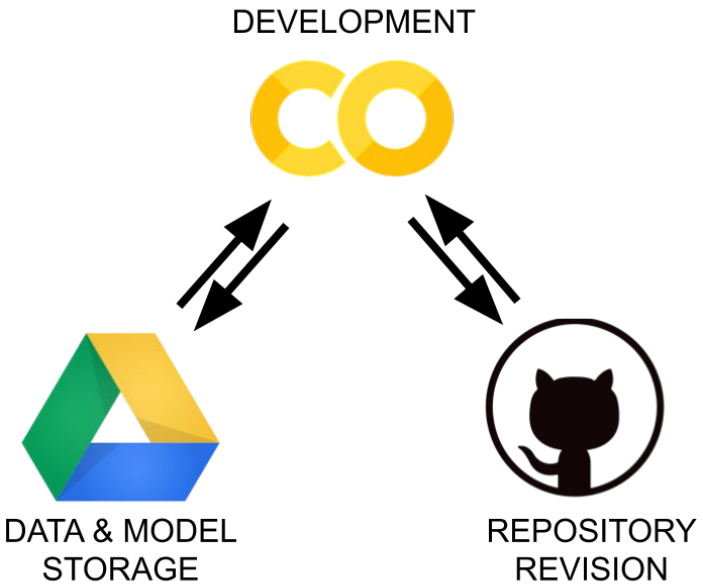
Entornos de trabajo - starter.

bit.ly/2yj4VH8



Entornos de trabajo - starter.

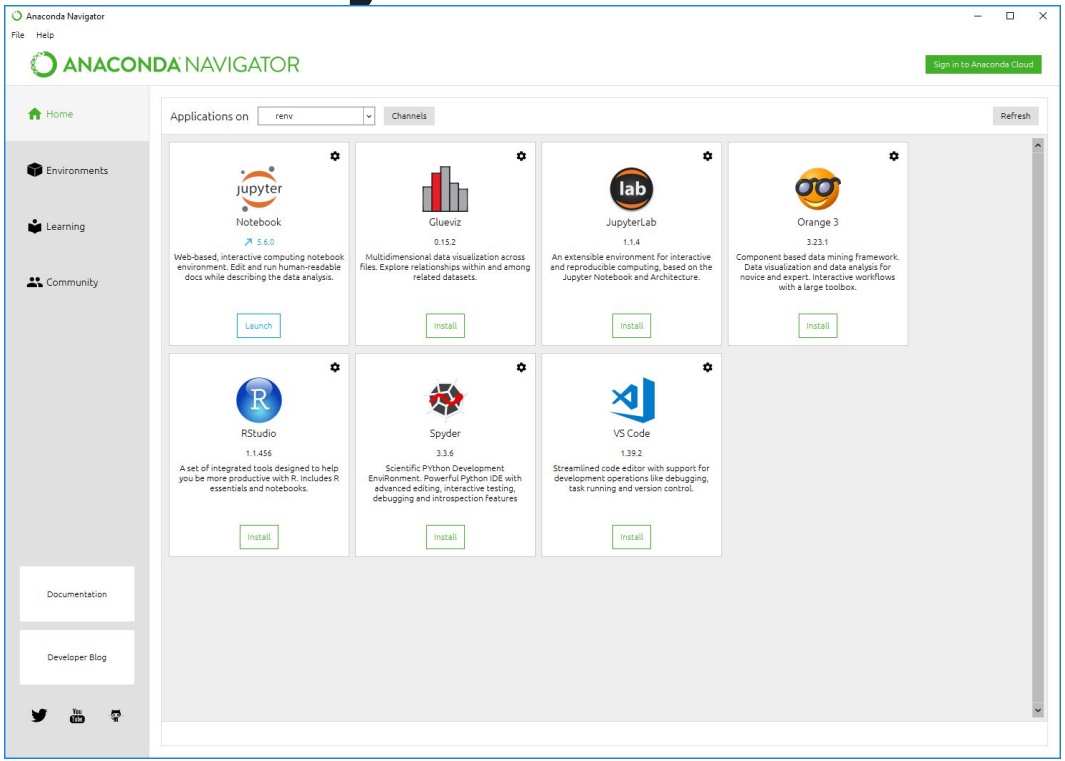
bit.ly/34DoZ30



Entornos de trabajo - newbie.

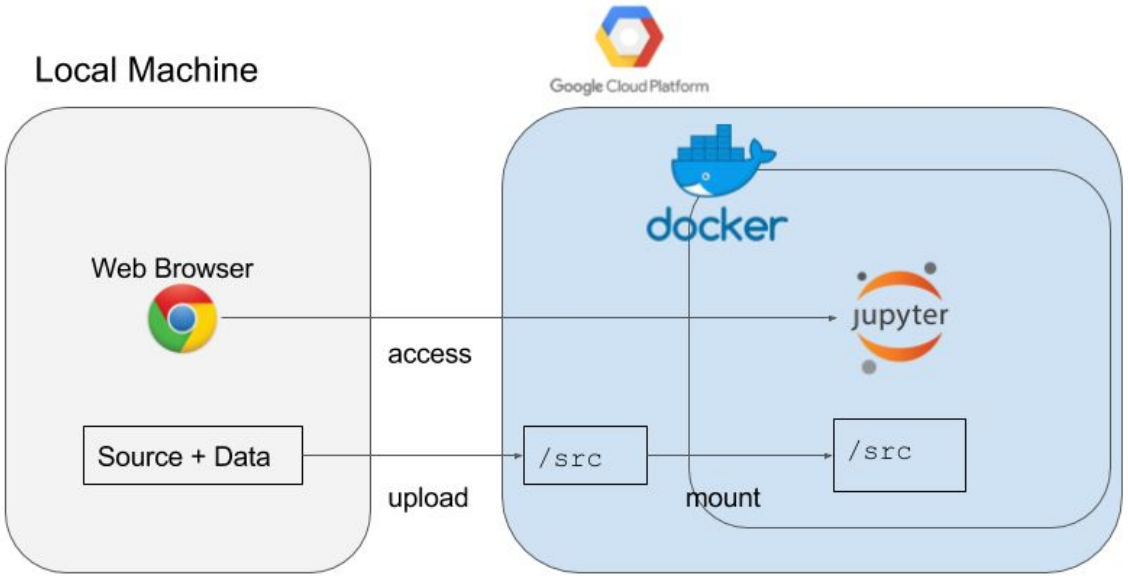
ANACONDA®

bit.ly/2XDVVac



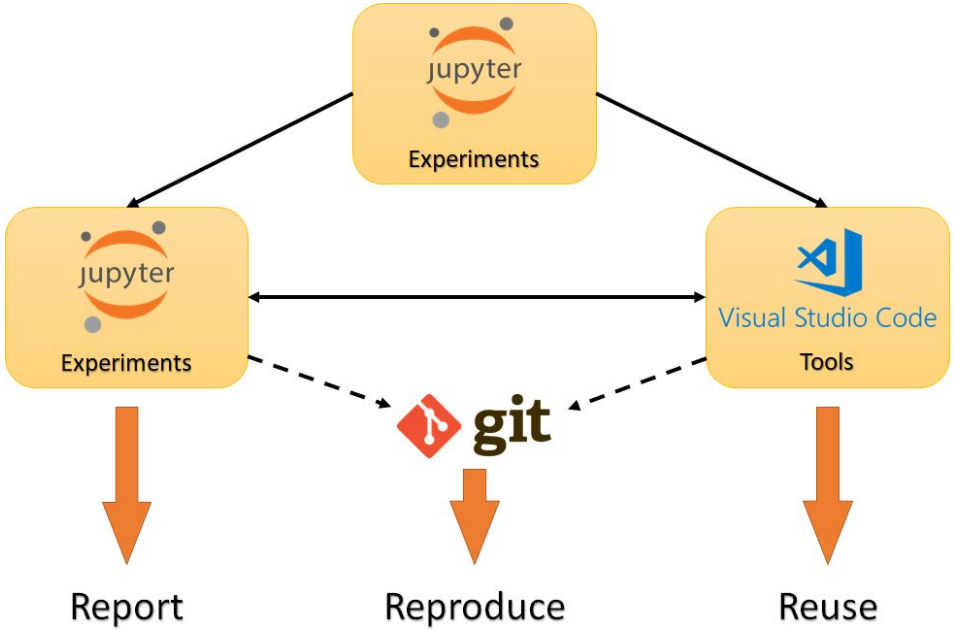
Entornos de trabajo - mid.

dockr.ly/3ewelQr

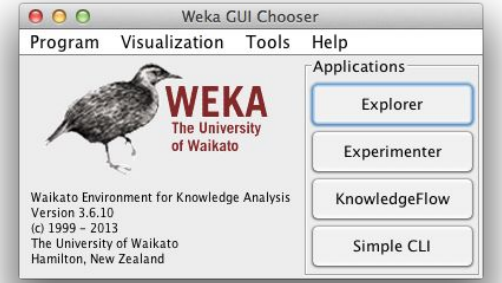


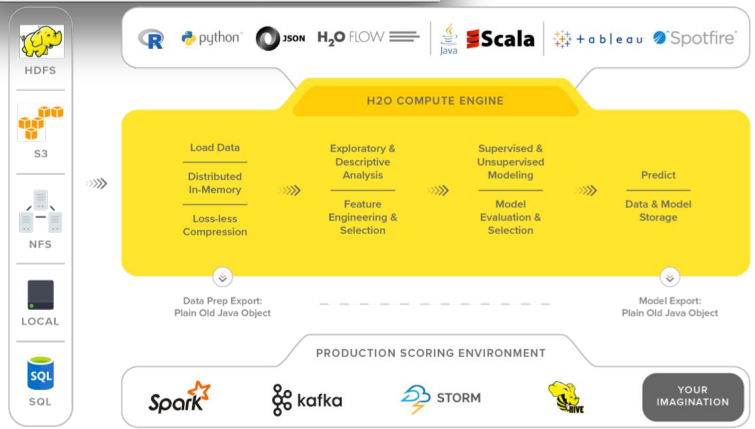
Entornos de trabajo - pro.

Visual Studio Code



Entornos de trabajo - biz.





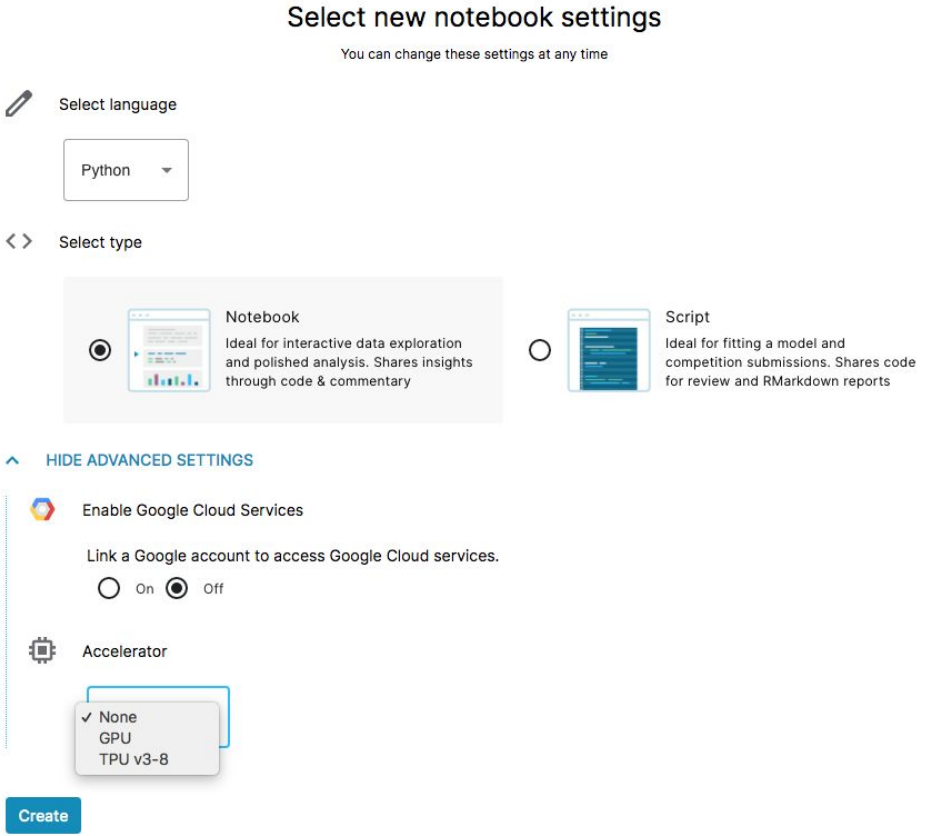
¿Y, los recursos?

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



Librerías.

NumPy

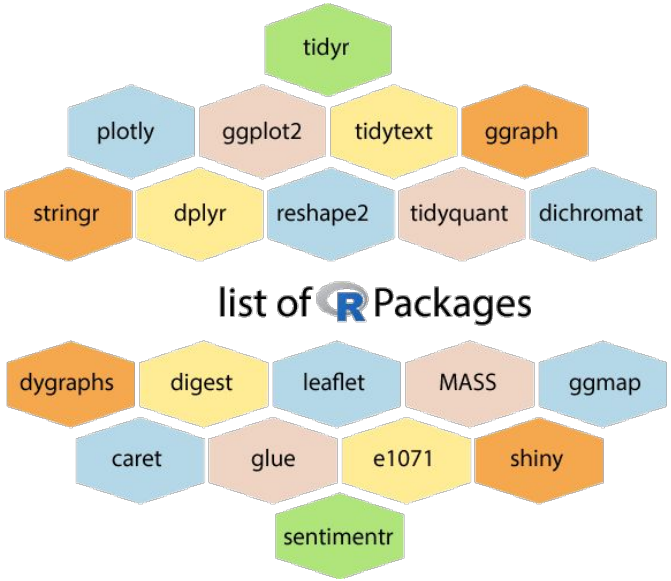
SciPy

matplotlib

IP[y]: IPython
Interactive Computing

pandas

scikit learn



Librerías.


Keras
Caffe
TensorFlow
PyTorch


spaCy
flair
jupyter
NLTK
Reference Guide
AllenNLP
NLP ARCHITECT

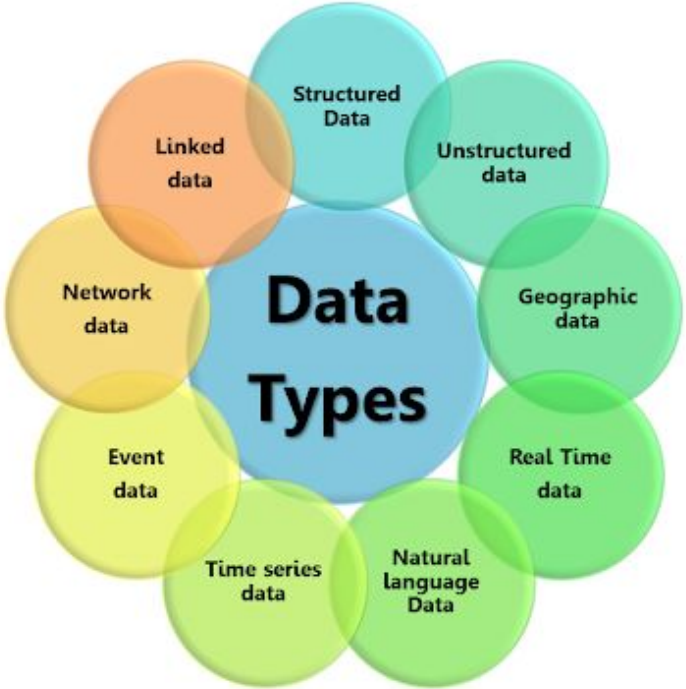

Google Cloud
DATA LAKE

Datasets - tamaño.

Zip ~1 MB

Zip ~6 GB

Datasets - tipo de dato.



...además de mucha paciencia....

01.04

Introducción

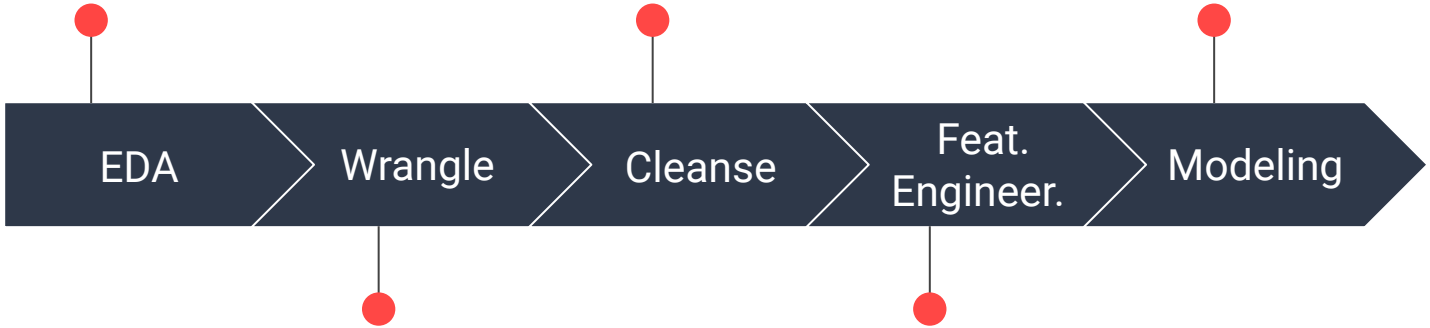
Flujo de trabajo.

Flujo de trabajo.

Check feature one by one new patterns and insights

Check missing values and impute / remove them.

Apply selected features to ML models.

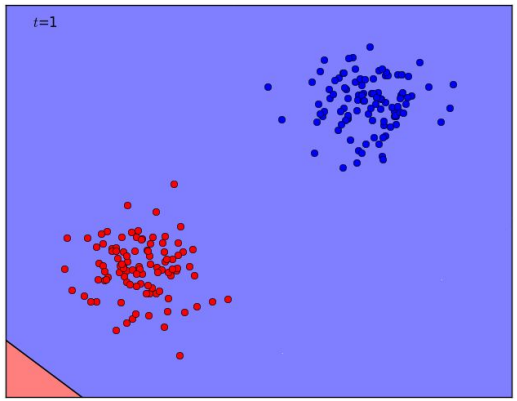


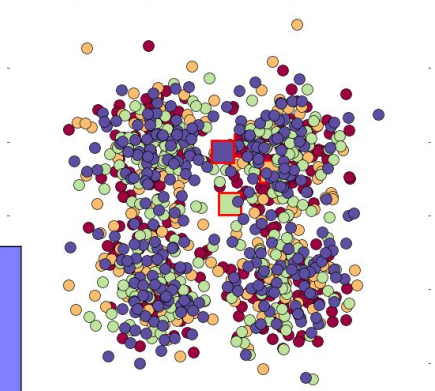
Prepare features for the final model.

Create new features or check cluster / groups to select the best ones.

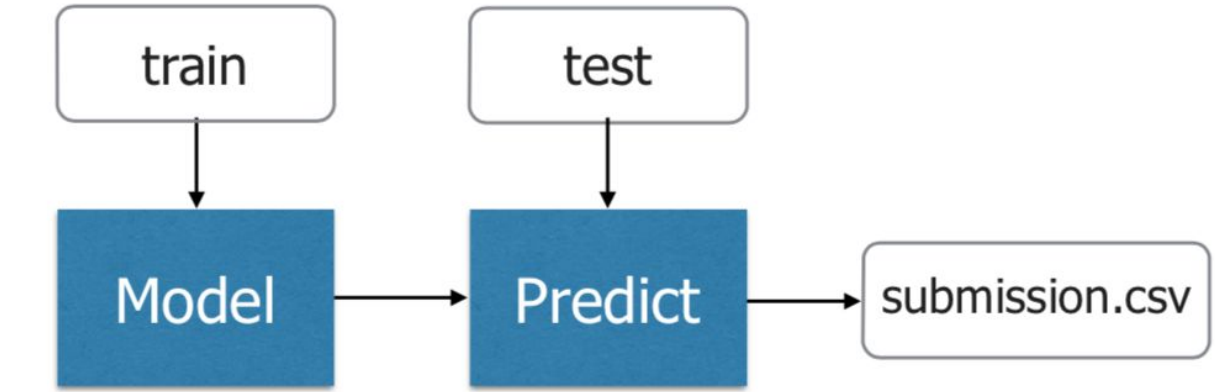
Tipos de ML.

- 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





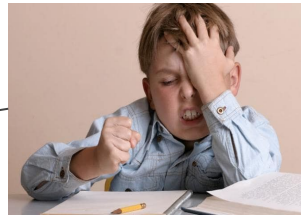
Build prediction model.



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

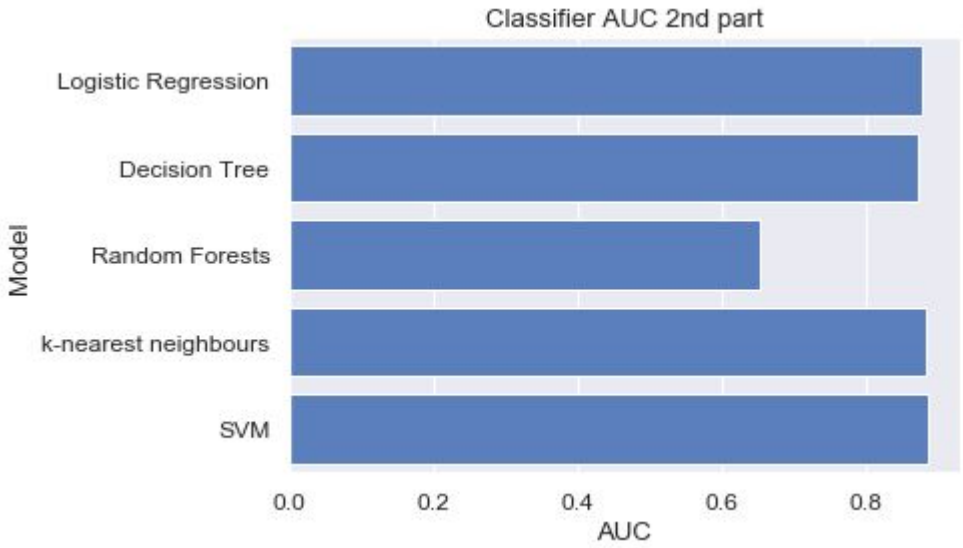


Training

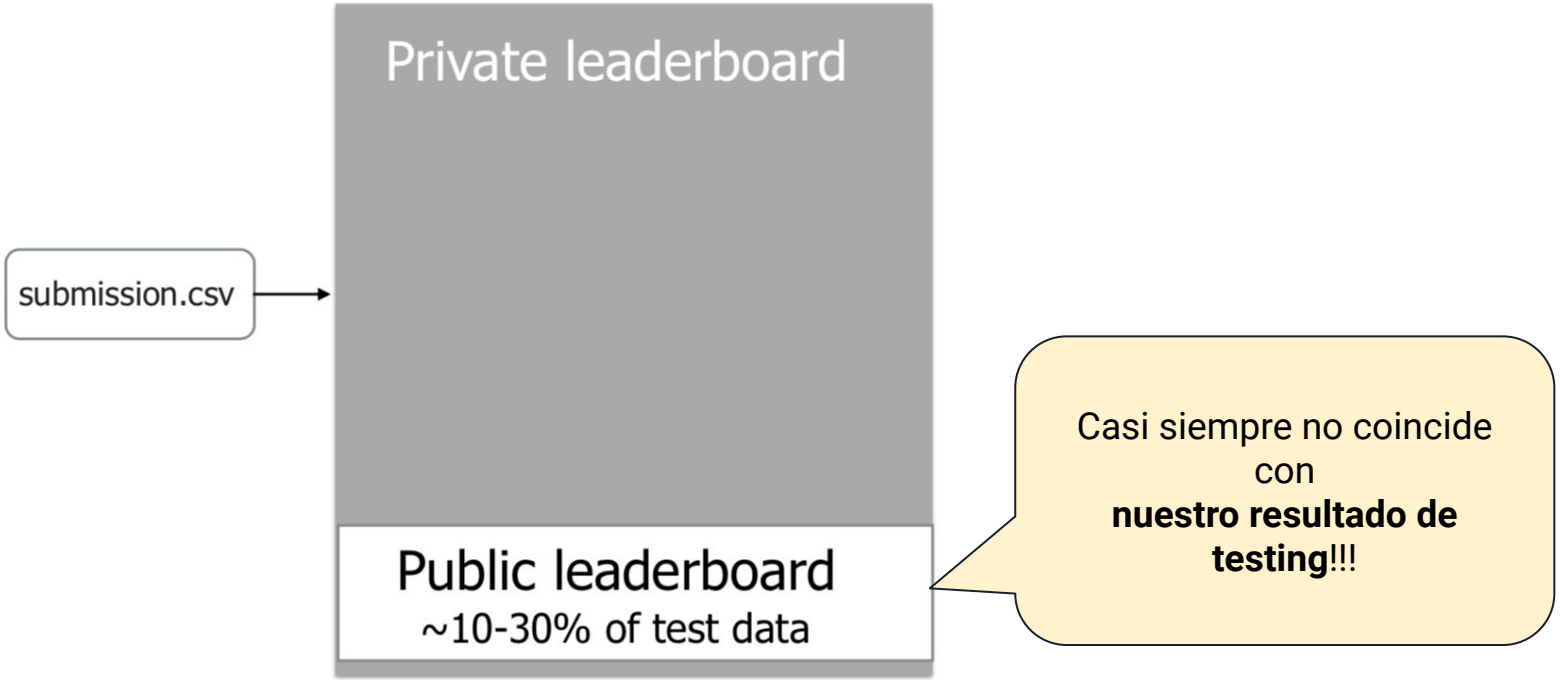
Inference

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.

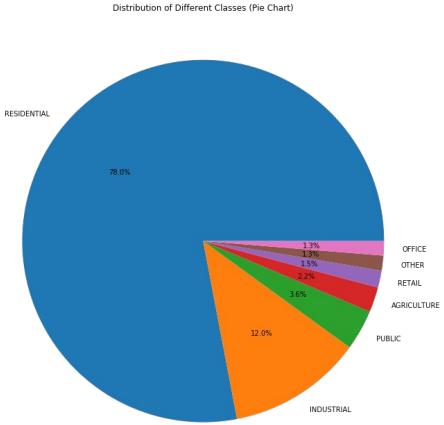

Private leaderboard
submission.csv
Public leaderboard
~10-30% of test data
Casi siempre no coincide con
nuestro resultado de testing!!!

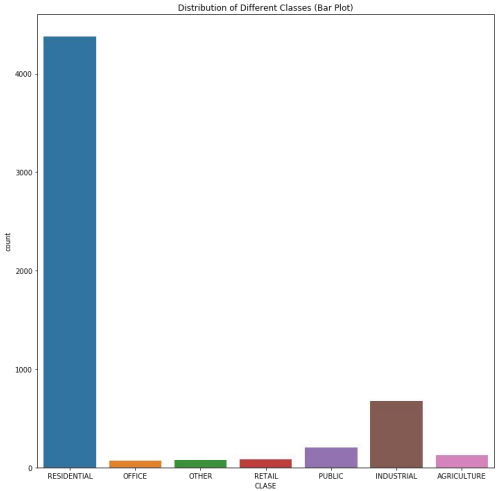
01.05

Introducción

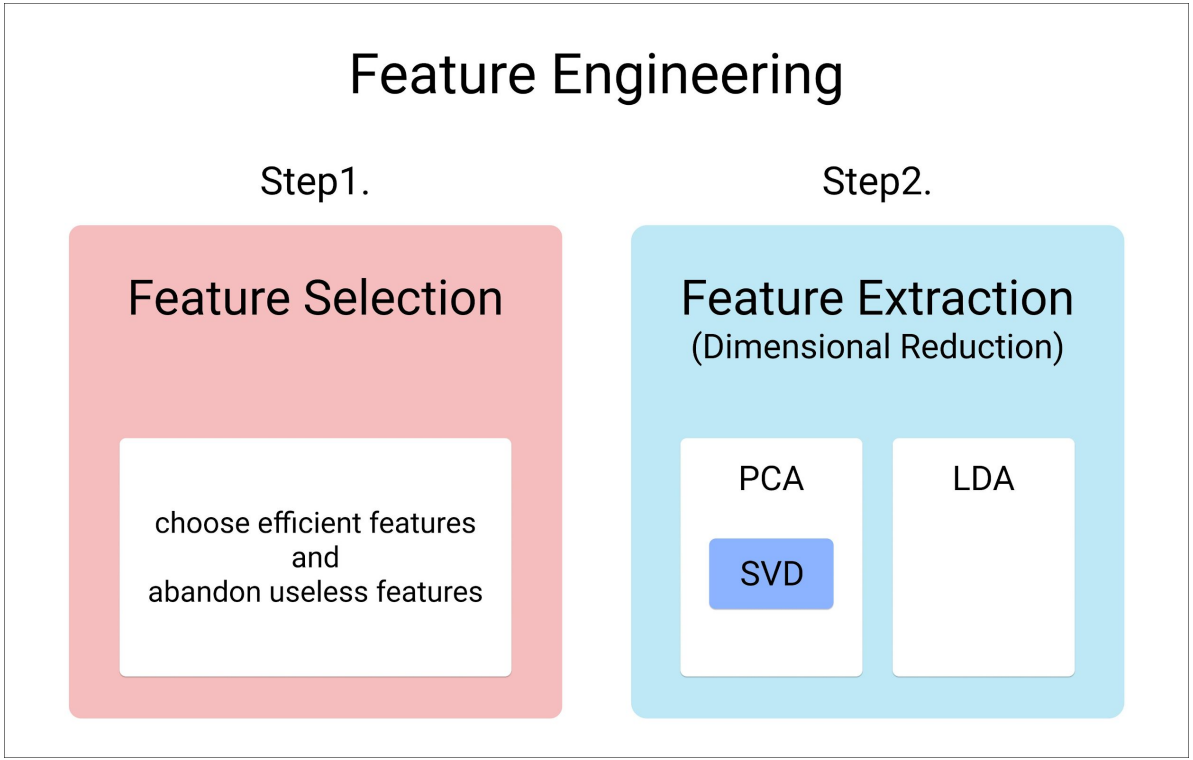
Tips.

EDA y correlaciones.





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!