

Resolviendo tu primera competencia de Kaggle

Francisco Alfaro

16 de Junio del 2023



Tabla de Contenidos

Introducción

Motivación

Mundo del Data Science

Esquema de Trabajo

Manos a la Obra

Caso de Estudio + Ejemplos

Mejorarando Resultados

Conclusiones

Resultados



¿Qué es Kaggle? k

kaggle

Kaggle es una plataforma acceder a conjuntos de datos, participar en desafíos y colaborar con otros en proyectos de aprendizaje automático.





Datasets

Explore, analyze, and share quality data. [Learn more](#) about data types, creating, and collaborating.

[+ New Dataset](#)[Your Work](#) Search datasets[Filters](#)[All datasets](#)[Computer Science](#)[Education](#)[Classification](#)[Computer Vision](#)[NLP](#)[Data Visualization](#)[Pre-Trained Model](#)

❖ Trending Datasets

[See All](#)

Ranking of United States cities by area

D Rahulsingh · Updated a day ago
Usability **10.0** - 5 kB
1 File (CSV)



Used Car Listings: Features and Price Prediction

Tugberk Karan · Updated 11 hours ago
Usability **9.4** - 1 MB
2 Files (CSV)



⚡️Tesla Supercharge Locations Globally⚡️

Omar Sobhy · Updated 7 days ago
Usability **10.0** - 305 kB
1 File (CSV)



⚡️r/Fitness Posts & Comments⚡️

PeachOrange · Updated 3 hours ago
Usability **10.0** - 2 MB





Learn

Gain the skills you need to do independent data science projects.



② Your Courses

Active



Intro to Programming

Next up: [Exercise: Arithmetic and Variables](#)



Python

Next up: [Exercise: Syntax, Variables, and Numbers](#)



Intro to Machine Learning

Next up: [Exercise: Explore Your Data](#)



Puede encontrar más detalles en el siguiente link:
⇒ [fralfaro/kaggle-courses](https://github.com/fralfaro/kaggle-courses)





Kaggle Courses

Kaggle Courses

Intro to Programming >

Python >

Pandas >

Data Visualization >

Data Cleaning >

Intro to Machine Learning >

Intermediate to Machine Learning >

Feature Engineering >

Time Series >

Intro to Deep Learning >

SQL >

Advanced SQL >

Kaggle Courses

Gain the skills you need to do independent data science projects.

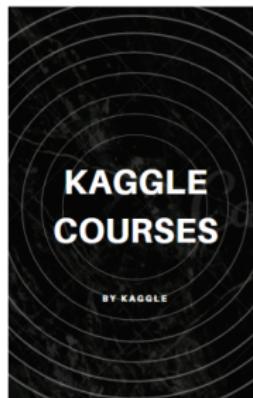
We pare down complex topics to their key practical components, so you gain usable skills in a few hours (instead of weeks or months).

The courses are provided at **no cost to you**, and you can now [earn certificates](#).

Note: This new documentation was developed by [fralfaro](#).

You can find the repository at the following link:

fralfaro/kaggle-courses.





Kaggle Courses

Kaggle Courses

[Open in Kaggle](#)

Intro to Programming



Python



Hello, Python!

Functions and Getting Help

Booleans and Conditionals

Lists

Loops and List

Comprehensions

Strings and Dictionaries

Working with External Libraries

Pandas



Data Visualization



Data Cleaning



Hello, Python!

This course covers the key Python skills you'll need so you can start using Python for data science.

We'll start with a brief overview of Python syntax, variable assignment, and arithmetic operators. If you have previous Python experience, you can [skip straight to the hands-on exercise.](#)

Python was named for the British comedy troupe [Monty Python](#), so we'll make our first Python program a homage to their skit about Spam.



Competitions



Competitions

Grow your data science skills by competing in our exciting competitions. Find help in the [documentation](#) or learn about [Community Competitions](#).

[Host a Competition](#)[Your Work](#) Search competitions[Filters](#)[All Competitions](#)
Everything, past & present[Featured](#)
Premier challenges with prizes[Getting Started](#)
Approachable ML fundamentals[Research](#)
Scientific and scholarly challenges[Community](#)
Created by fellow Kagglers[Playground](#)
Fun practice problems

Active Competitions

[Hotness](#)

Vesuvius Challenge - Ink Detection

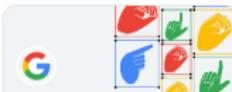
Resurrect an ancient library from the ash...

Featured

Code Competition - 1243 Teams

\$1,000,000

3 days to go



Google - American Sign Language Fingerspelling...

Train fast and accurate American Sign La...

Research

Code Competition - 255 Teams

\$200,000

2 months to go



2023 Kaggle AI Report

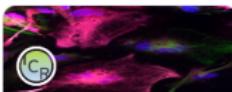
Essays on the state of machine learning ...

Analytics

82 Teams

\$70,000

24 days to go



ICR - Identifying Age-Related Conditions

Use Machine Learning to detect condition...

Featured

Code Competition - 2679 Teams

\$60,000

2 months to go



Por qué debería participar ?

- Aprendizaje y mejora de habilidades
- Acceso a conjuntos de datos y problemas interesantes
- Colaboración y aprendizaje compartido
- Reconocimiento y oportunidades profesionales



Por qué debería participar ?

- Aprendizaje y mejora de habilidades
- Acceso a conjuntos de datos y problemas interesantes
- Colaboración y aprendizaje compartido
- Reconocimiento y oportunidades profesionales



Por qué debería participar ?

- Aprendizaje y mejora de habilidades
- Acceso a conjuntos de datos y problemas interesantes
- Colaboración y aprendizaje compartido
- Reconocimiento y oportunidades profesionales



Por qué debería participar ?

- Aprendizaje y mejora de habilidades
- Acceso a conjuntos de datos y problemas interesantes
- Colaboración y aprendizaje compartido
- Reconocimiento y oportunidades profesionales



Tabla de Contenidos

Introducción

Motivación

Mundo del Data Science

Esquema de Trabajo

Manos a la Obra

Caso de Estudio + Ejemplos

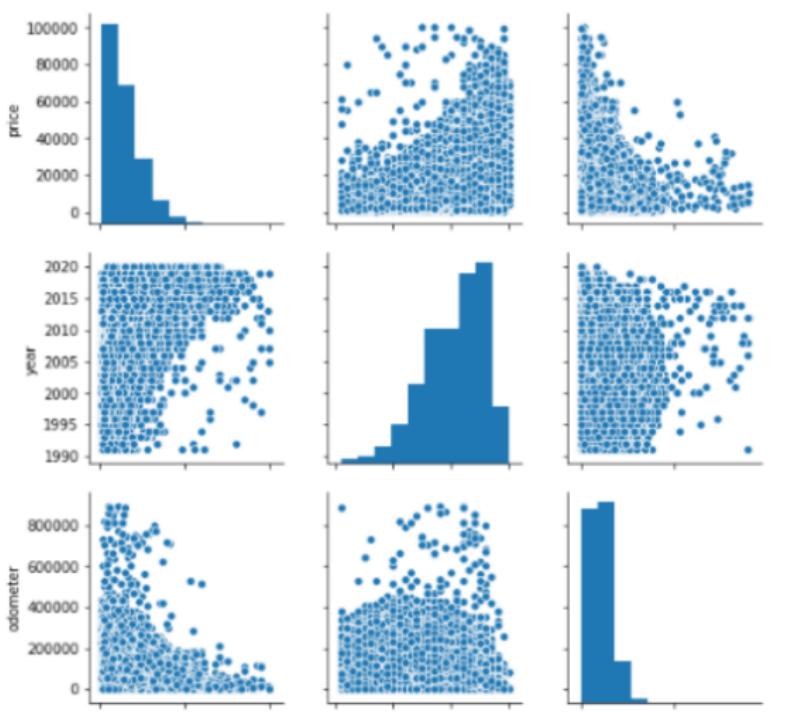
Mejorarando Resultados

Conclusiones

Resultados









- Resumen estadístico
- Visualización de datos
- Tratamiento de datos faltantes o duplicados
- Análisis de valores atípicos





- Resumen estadístico
- Visualización de datos
- Tratamiento de datos faltantes o duplicados
- Análisis de valores atípicos





- Resumen estadístico
- Visualización de datos
- Tratamiento de datos faltantes o duplicados
- Análisis de valores atípicos

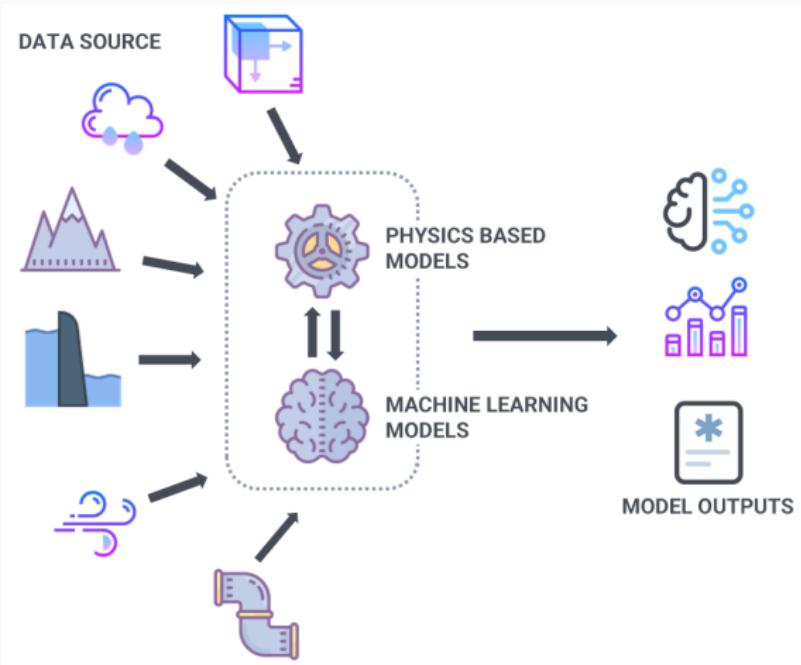




- Resumen estadístico
- Visualización de datos
- Tratamiento de datos faltantes o duplicados
- Análisis de valores atípicos



Machine Learning



Machine Learning

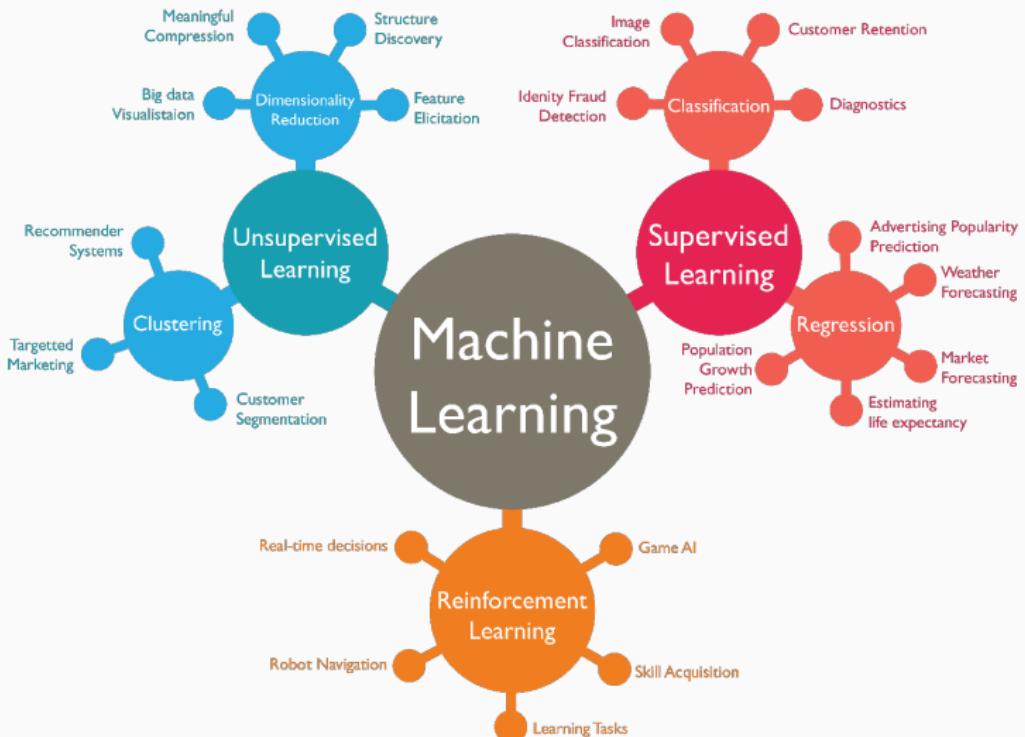


Tabla de Contenidos

Introducción

Motivación

Mundo del Data Science

Esquema de Trabajo

Manos a la Obra

Caso de Estudio + Ejemplos

Mejorarando Resultados

Conclusiones

Resultados





Resolver competencia de **Titanic** mediante notebooks.

- Código este en Github
- Ejemplo del Notebook
- Subir resultados

Ocuparemos el repositorio: [fralfaro/kaggle-competitions](#)





Resolver competencia de **Titanic** mediante notebooks.

- Código este en Github
- Ejemplo del Notebook
- Subir resultados

Ocuparemos el repositorio: [fralfaro/kaggle-competitions](#)





github.com/fralfaro/kaggle-competitions

📁 competitions/titanic	titanic v1.0	1 hour ago
📄 .gitignore	first commit	6 hours ago
📄 LICENSE	first commit	6 hours ago
📄 README.md	titanic v1.01	1 hour ago

☰ README.md ✍

Kaggle Competitions

Kaggle Competitions is a platform hosted by Kaggle, a data science community and platform. Kaggle Competitions provide a space for data scientists, machine learning practitioners, and enthusiasts to participate in various data-related challenges and competitions.

| ⓘ Note: For more information, see the following [link](#).

Description

Sprint	Description	Google Colab
Titanic - Machine Learning from Disaster	markdown	Open in Kaggle



Caso de Estudio

kaggle-competitions/tree/main/competitions/titanic

[kaggle-competitions](#) / [competitions](#) / [titanic](#) / 



fralfaro titanic v1.0

Name	Last commit message
 ..	
 data	add titanic data
 description.md	first commit
 solution.ipynb	titanic v1.0





notebook7c0ac83041 Draft saved

File Edit View Run Add-ons Help

Share

Save Version 0

+ Run All Markdown ▾

Draft Session off (run a cell to start)

Open in Kaggle

Titanic Data Science Solutions

This notebook is a companion to the book [Data Science Solutions](#).

The notebook walks us through a typical workflow for solving data science competitions at sites like Kaggle.

There are several excellent notebooks to study data science competition entries. However many will skip some of the explanation on how the solution is developed as these notebooks are developed by experts for experts. The objective of this notebook is to follow a step-by-step workflow, explaining each step and rationale for every decision we take during solution development.

Workflow stages

Notebook

Data

Models

Notebook options

ACCELERATOR

None

LANGUAGE

Python

PERSISTENCE





- Definir Problema
- Datos entrenamiento y pruebas
- Discutir, preparar, limpiar los datos
- Analizar y explorar los datos
- Modelar, predecir y resolver el problema
- Visualizar resultados y dar solución final
- Enviar los resultados



Definir Problema

www.kaggle.com/competitions/titanic/overview

The screenshot shows the 'Getting Started Prediction Competition' for the 'Titanic - Machine Learning from Disaster'. The page features a large background image of the Titanic ship at night. At the top left is a competition icon. The title 'Titanic - Machine Learning from Disaster' is prominently displayed, followed by the subtitle 'Start here! Predict survival on the Titanic and get familiar with ML basics.' Below this, a 'Kaggle' logo indicates 15,761 teams are ongoing. A navigation bar includes links for Overview, Data, Code, Discussion, Leaderboard, Rules, Team, Submissions, Submit Predictions, and more. The main content area has a 'Overview' tab selected, showing a 'Description' section with a welcome message: 'Ahoy, welcome to Kaggle! You're in the right place.' It explains the competition is the legendary Titanic ML challenge, designed to introduce users to ML competitions and the Kaggle platform. It also describes the goal of creating a model to predict passenger survival. A note encourages reading a video or tutorial for more details.

Getting Started Prediction Competition

Titanic - Machine Learning from Disaster

Start here! Predict survival on the Titanic and get familiar with ML basics.

Kaggle · 15,761 teams · Ongoing

Overview Data Code Discussion Leaderboard Rules Team Submissions Submit Predictions ...

Overview

Description

Ahoy, welcome to Kaggle! You're in the right place.

This is the legendary Titanic ML competition – the best, first challenge for you to dive into ML competitions and familiarize yourself with how the Kaggle platform works.

The competition is simple: use machine learning to create a model that predicts which passengers survived the Titanic shipwreck.

Read on or watch the video below to explore more details. Once you're ready to start competing, click on the "Join Competition" button to create an account and gain access to the competition data. Then check out [Alexis Cook's Titanic Tutorial](#) that walks you through step by step how to make your first submission!



Datos entrenamiento y pruebas

www.kaggle.com/competitions/titanic/data

The screenshot shows the Kaggle interface for the "Getting Started Prediction Competition" titled "Titanic - Machine Learning from Disaster". The page features a large image of the Titanic ship at night. Below the image, the text reads "Start here! Predict survival on the Titanic and get familiar with ML basics". A "Kaggle" logo indicates 15,761 teams are ongoing. The navigation bar includes links for Overview, Data (which is underlined), Code, Discussion, Leaderboard, Rules, Team, Submissions, Submit Predictions, and more.

Dataset Description

Overview

The data has been split into two groups:

- training set (train.csv)
- test set (test.csv)

The **training set** should be used to build your machine learning models. For the training set, we provide the outcome (also known as the "ground truth") for each passenger. Your model will be based on "features" like passengers' gender and class. You can also use feature engineering to create new features.

The **test set** should be used to see how well your model performs on unseen data. For the test set, we do not provide the ground truth for each passenger. It is your job to predict these outcomes. For each passenger in the test set, use the model you trained to predict whether or not they survived the sinking of the Titanic.

We also include **gender_submission.csv**, a set of predictions that assume all and only female passengers survive, as an example of what a submission file should look like.

Files

3 files

Size

93.08 kB

Type

CSV



Discutir, preparar, limpiar los datos

```
train_df.describe()
```

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057	49.693429
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	223.500000	0.000000	2.000000	20.125000	0.000000	0.000000	7.910400
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	668.500000	1.000000	3.000000	38.000000	1.000000	0.000000	31.000000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200



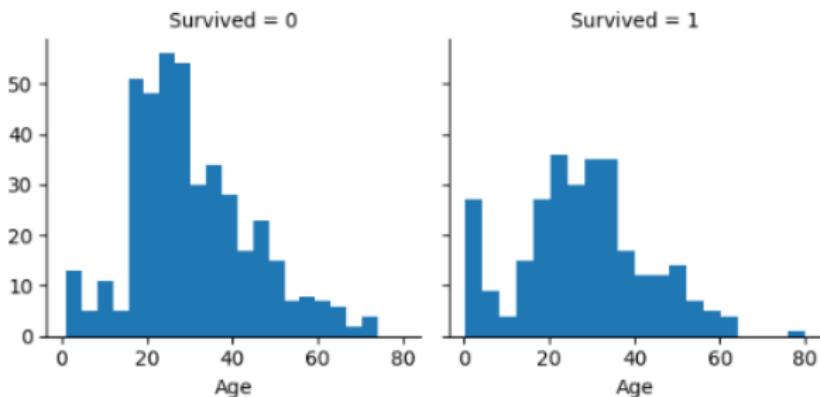
Se utiliza la librería de **Pandas** para leer, agrupar y manipular los datos.



Analizar y explorar los datos

```
g = sns.FacetGrid(train_df, col='Survived')
g.map(plt.hist, 'Age', bins=20)
```

```
<seaborn.axisgrid.FacetGrid at 0x198bf4cc430>
```



Se utiliza **Matplotlib** o **Seaborn** para el análisis univariado y multivariado.



Modelar, predecir y resolver el problema.

```
# Decision Tree

decision_tree = DecisionTreeClassifier()
decision_tree.fit(X_train, Y_train)
Y_pred = decision_tree.predict(X_test)
acc_decision_tree = round(decision_tree.score(X_train, Y_train) * 100, 2)
acc_decision_tree
```

86.76

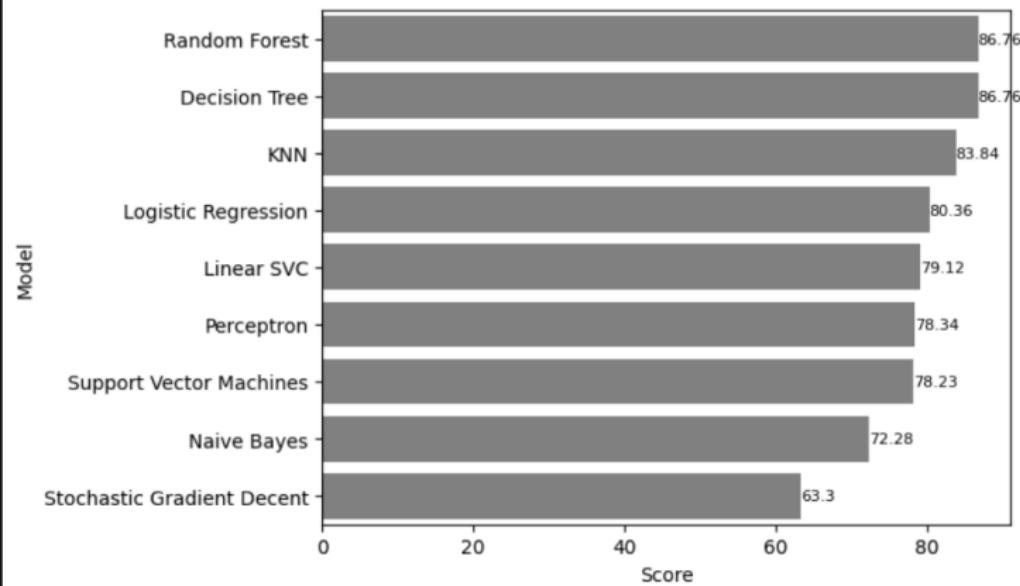


Se utiliza la librería de **scikit-learn** para los modelos de machine learning.



Visualizar resultados y dar solución final

```
plotting = sns.barplot(data=models, y="Model", x="Score",color = 'gray')
for container in plotting.containers:
    plotting.bar_label(container,fontsize=8)
```



Enviar los resultados

Overview Data Code Discussion Leaderboard Rules Team

Submissions

Submit Predictions

...

X Submit to Competition

File Upload Notebook



Titanic - Machine Learning from Disaster

You have 10 submissions remaining today. This resets in 3 hours.

Uploaded File

submission.csv (3 KiB)



Your submission should be a CSV file with 418 rows and a header. You can upload a zip/gz/7z archive.

DESCRIPTION

submission file

15 / 500

```
>_ kaggle competitions submit -c titanic -f submission.csv -m "Mes_
```



Enviar los resultados

www.kaggle.com/competitions/titanic/submissions

Submissions

	Submission and Description	Public Score	Recent
	submission.csv Complete · now · submission file	0.75837	
	submission.csv Complete · 1mo ago · my first attempt	0.77511	



Tabla de Contenidos

Introducción

Motivación

Mundo del Data Science

Esquema de Trabajo

Manos a la Obra

Caso de Estudio + Ejemplos

Mejorarando Resultados

Conclusiones

Resultados



Modelos dominantes - Gradient Boosting

- **LightGBM:**

- Funciona por defecto con variables categóricas.
- Muy rápido, incluso en CPU.

- **XGBoost:**

- Debemos hacer encoding de variables categóricas.
- Rápido en GPU.

- **CatBoost:**

- Funciona por defecto con variables categóricas.
- Más rápido que XGBoost, más lento que LightGBM.



Modelos dominantes - DeepLearning

TabNet (2020): Mecanismo de atención secuencial. Selecciona features en cada iteración (ganando interpretabilidad). [Ver Paper](#)

Name	Rossman	CoverType	Higgs	Gas	Eye	Gesture	YearPrediction	MSLR	Epsilon	Shruthime	Blastchar
XGBoost	490.18	3.13	21.62	2.18	56.07	80.64	77.98	55.43	11.12	13.82	20.39
NODE	488.59	4.15	21.19	2.17	68.35	92.12	76.39	55.72	10.39	14.61	21.40
DNF-Net	503.83	3.96	23.68	1.44	68.38	86.98	81.21	56.83	12.23	16.80	27.91
TabNet	485.12	3.01	21.14	1.92	67.13	96.42	83.19	56.04	11.92	14.94	23.72
1D-CNN	493.81	3.51	22.33	1.79	67.90	97.89	78.94	55.97	11.08	15.31	24.68
Simple Ensemble	488.57	3.19	22.46	2.36	58.72	89.45	78.01	55.46	11.07	13.61	21.18
Deep Ensemble w/o XGBoost	489.94	3.52	22.41	1.98	69.28	93.50	78.99	55.59	10.95	14.69	24.25
Deep Ensemble w XGBoost	485.33	2.99	22.34	1.69	59.43	78.93	76.19	55.38	11.18	13.10	20.18

TabNet

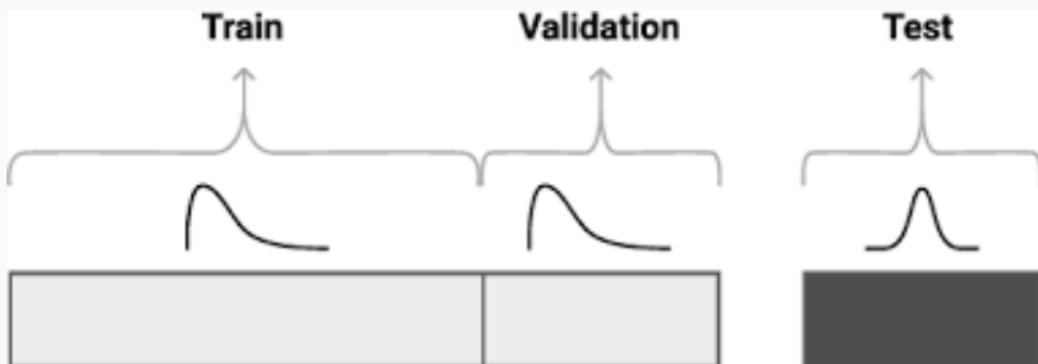
DNF-Net

NODE

New datasets



Validación adversarial



La idea es comparar el nivel de similitud entre **Train** y **Test** en términos de la distribución de las features.



Validación adversarial

- Si son difíciles de distinguir, entonces probablemente las distribuciones son similares y técnicas de validación usuales pueden ser aplicadas.
- Si no, debemos buscar la(s) features que nos están dando problema(s) y arreglarlas (esto normalmente es complejo) o eliminarlas.



NVIDIA RAPIDS



- **cuML:** ¿Sklearn en GPU?. [Ver algoritmos en github](#)
- **cuDF:** Dask en GPU. Competencia en Kaggle Accelerating Trading on GPU.



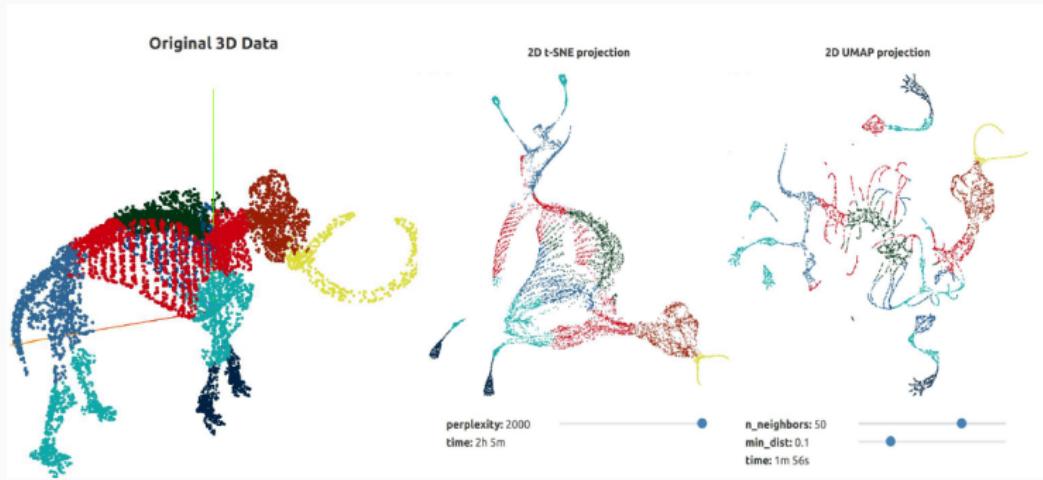
Reducción Dimensionalidad - TSNE



Visualizing Data using t-SNE (2008)



Reducción Dimensionalidad - UMAP



UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction (2018)



Leave One Feature Out (LOFO)

```
kf = KFold(n_splits=10, shuffle=True, random_state=42)

features = ['cat0', 'cat1', 'cat2', 'cat3', 'cat4', 'cat5', 'cat6', 'cat7',
           'cat8', 'cat9', 'cont0', 'cont1', 'cont2', 'cont3', 'cont4', 'cont5',
           'cont6', 'cont7', 'cont8', 'cont9', 'cont10', 'cont11', 'cont12',
           'cont13']

model = XGBRegressor(tree_method ="gpu_hist", gpu_id = 0, predictor = "gpu_predictor")
dataset = Dataset(df=train, target="target", features=features)
lofo_imp = LOFOImportance(dataset, cv=kf, scoring="neg_root_mean_squared_error", model = model)
importance_df_xgb = lofo_imp.get_importance()
plot_importance(importance_df_xgb, figsize=(12, 20))
```

□ □ 100% | 24/24 [05:00<00:00, 12.49s/it]

- Evalúa la performance del modelo utilizando todas las features, luego, de manera iterativa remueve features una a una.
- Generaliza bien a test set no vistos.
- Es agnóstica al modelo.



Optuna

```
%%time
study = optuna.create_study(direction='minimize')
optuna.logging.set_verbosity(optuna.logging.WARNING)
study.optimize(objective, n_trials=1024)

best_trial=study.best_trial.params
print('Number of finished trials:', len(study.trials))
best_trial

Number of finished trials: 1024
CPU times: user 52min 20s, sys: 1min 55s, total: 54min 15s
Wall time: 9min 4s
```

- Optimización de hiperparámetros mediante técnicas de optimización bayesiana y algoritmos genéticos.
- Pruning Callbacks y más!



Machine Learning Competitivo

Autor: Daniel Pereda

Machine Learning Competitivo: top 1% en Kaggle - Youtube



Tabla de Contenidos

Introducción

Motivación

Mundo del Data Science

Esquema de Trabajo

Manos a la Obra

Caso de Estudio + Ejemplos

Mejorarando Resultados

Conclusiones

Resultados



Resultados

- Conocer plataforma de Kaggle.
- Primer Desafío en el mundo DS.
- Mejoras continuas en competiciones.



Resultados

- Conocer plataforma de Kaggle.
- Primer Desafío en el mundo DS.
- Mejoras continuas en competiciones.



Resultados

- Conocer plataforma de Kaggle.
- Primer Desafio en el mundo DS.
- Mejoras continuas en competiciones.



Referencias

- Machine Learning Competitivo: top 1% en Kaggle - YouTube
- KC: A Beginner's Guide to Winning (Rob Mulla) - YouTube
- Introducción a Kaggle - YouTube



Resolviendo tu primera competencia de Kaggle

Francisco Alfaro

16 de Junio del 2023

