

Reto sobre aplicación de Python para respuesta a preguntas de analítica

Profesores:

Christian Urcuqui (ulcamilo@gmail.com)

Universidad Icesi

Objetivos de esta sesión

1. Usar Python para la exploración de un conjunto de datos.
2. Responder a un conjunto de preguntas a partir de los datos
3. Aplicar Python para el desarrollo de funciones, ciclos, asignación de variables y explorar la información de un dataframe.

Contexto

Análisis de datos

Titanic ML competition

Use machine learning to create a model that predicts which passengers survived the Titanic shipwreck

El 15 de abril de 1912, durante su viaje inaugural, el Titanic, ampliamente considerado "insumergible", se hundió después de chocar con un iceberg. Desafortunadamente, no había suficientes botes salvavidas para todos a bordo, lo que resultó en la muerte de 1502 de los 2224 pasajeros y la tripulación.

Cree un modelo predictivo que responda a la pregunta: "¿Qué tipo de personas tenían más probabilidades de sobrevivir?" utilizando datos de pasajeros (es decir, nombre, edad, sexo, clase socioeconómica, etc.).

<https://www.kaggle.com/c/titanic/overview>

Taller práctico

Variable	Definition	Key
Survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	Age in years
Age	Age in years	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton



PREGUNTAS DE ANÁLISIS

Vamos a utilizar solo el archivo train.csv

Complejidad 1 (25%)

Utilice la mayor cantidad de variables (todas en lo posible)

1. Seleccionar tres algoritmos de clasificación, entrenar de cada uno un modelo y testearlo con el fin de obtener las siguientes métricas:
 1. Matriz de confusión
 2. Accuracy, recall, precision
2. Comparar y concluir cual fue el mejor modelo.

PREGUNTAS DE ANÁLISIS

Vamos a utilizar solo el archivo train.csv



Complejidad 2 (25%)

3. Aplique EDA y seleccione un subconjunto de variables. Explique las razones ya sea con aplicación de estadísticos, visualizaciones u otras técnicas de minería de datos.
4. Aplique la misma aproximación de los puntos 1 y 2

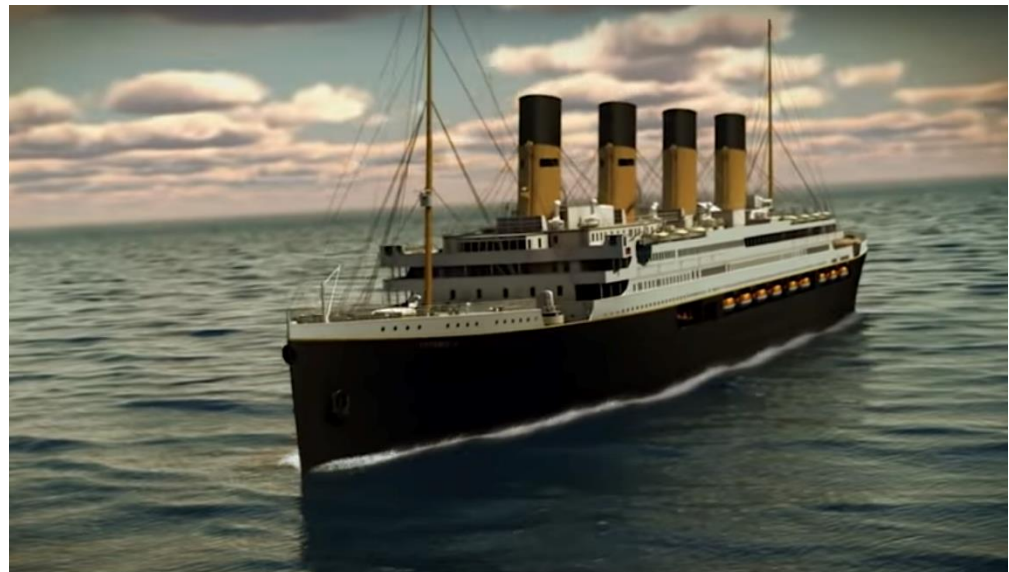
PREGUNTAS DE ANÁLISIS

Vamos a utilizar solo el archivo train.csv

Complejidad 3 (25%)

5. Aplique ingeniería de variables con el fin de crear al menos una e incluirla en el proceso del slide anterior.

6. Aplique la misma aproximación de los puntos 1 y 2



PREGUNTAS DE ANÁLISIS




Complejidad 3 (25%)

7. Compare los resultados de los modelos y seleccione el mejor.


8. Envié los resultados del archivo test.csv a la plataforma de Kaggle, tome captura de la evidencia y anéxelo como imagen al jupyter notebook

All Categories ▾

2 Active Competitions



Titanic - Machine Learning from Disaster
Start here! Predict survival on the Titanic and get familiar with ML basics
Getting Started - Ongoing










1337/30796
Top 5%

Overview Data Code Discussion Leaderboard Rules Team

My Submissions

Submit Predictions

1331	James1987		0.80143	5	3d
1332	Shuliang Zhang		0.80143	8	16h
1333	LeonardChan23		0.80143	10	2d
1334	Cooldude1		0.80143	71	1d
1335	Andrew Ritchie		0.80143	26	4h
1336	Sargis Iskandaryan		0.80143	3	1d
1337	Christian Urcuqui		0.80143	1	~10s

Your First Entry

Welcome to the leaderboard!

Your score represents your submission's accuracy. For example, a score of 0.7 in this competition indicates you predicted Titanic survival correctly for 70% of people.

What next? You've got a few options:

- 👉 Learn skills that can improve your score in [our Intro to Machine Learning course by Dan Becker](#).
- 🔍 Check out [the discussion forum](#) to find lots of tutorials and insights from other competitors.
- 🏆 Find a new challenge by entering one of our [open, active competitions](#) or searching our [public datasets](#).

