Presentación final del reto

Inteligencia artificial avanzada para la ciencia de datos I

Kaggle Challenge: Driving Behavior

Presentan:
Edgar Daniel Acosta Rosales
Diana Guadalupe García Aguirre
José Herón Samperio León
David Alejandro Velázquez Valdéz

14 de septiembre del 2022

Conózcannos mejor

Daniel

David

Diana

Herón

Definición del problema

TC3006C.101

Driving Behavior Challenge

Ciencia de datos para todos

Kaggle es la plataforma de Ciencia de Datos más grande del mundo, con más de un millón de usuarios.

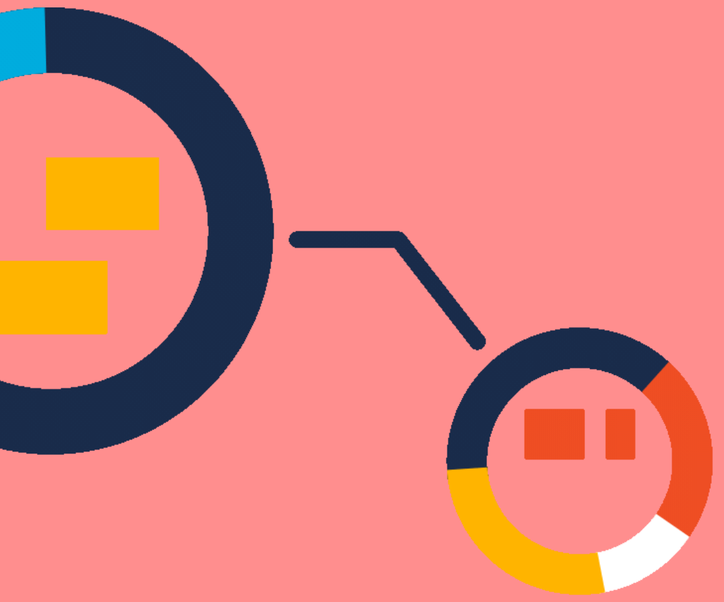
Subsidiaria de Alphabet Inc.

De **principiante** a **Grandmaster.**

IA para la Ciencia de Datos I

El reto:

Competencia internacional en la que personas de todo el mundo buscan resolver un problema en concreto, con impacto en la industria tecnológica.



Metodología CRISP - DM en Ciencia de Datos

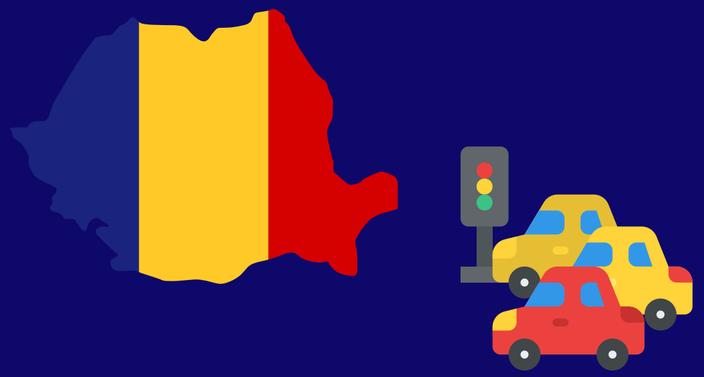
Cross Industry Standard Process for Data Mining

TC3006C.101

Driving Behavior Challenge

Business Understanding

What does the business need?



AAA Foundation for Traffic Safety

Ion Cojocaru, Stefan Popescu y Cristian Mihaescu, University of Craiova, Rumania.

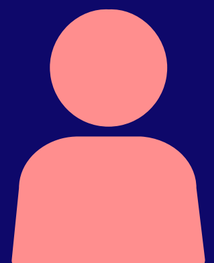
Aplicación de Android que mide y almacena valores de sus sensores durante ciertos intervalos de tiempo.

¿Cómo predecir tipos de conductas de manejo agresivas de manera rápida y precisa?

Data Understanding

What data do we have / need? Is it clean?

There was only one driver (33 years old) and we used a Dacia Sandero 1.4 MPI.
The data was collected with the help of 2 people (driver and assistant). Assistant's role was to note the behavior and keep the phone steady on the armrest.
We also wrote a paper which is currently under review, and we included all the details in it, but if you need any other information, please let us know.

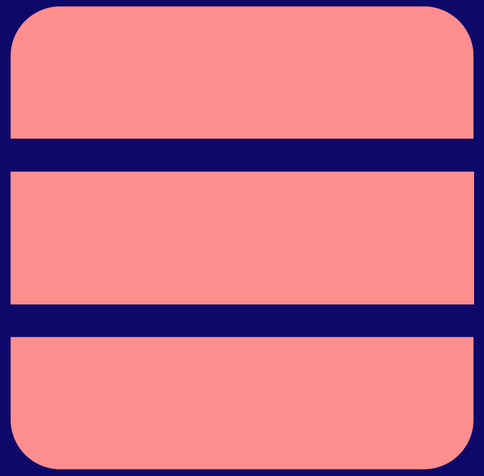


Ion Cojocaru, author

Data Understanding

What data do we have / need? Is it clean?

Dos datasets, para training y testing.

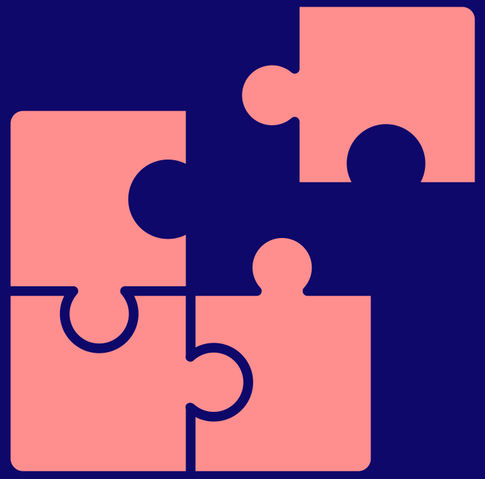


3644 registros - Train set
3084 registros - Test set

AccX, AccY, AccZ, GyroX, GyroY, GyroZ, Timestamp, Class

Data Understanding

What data do we have / need? Is it clean?



No hay datos faltantes.

Proporción de SLOW 36.5%
Proporción de NORMAL 32.9%
Proporción of AGGRESSIVE 30.5%

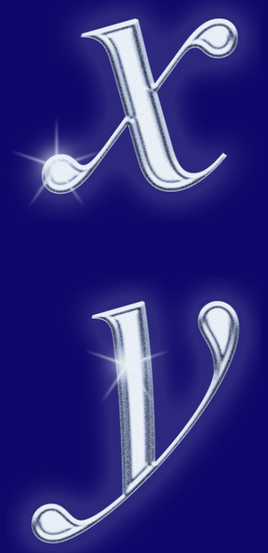
Seis variables de tipo float (continuas).
Una variable de tipo string (categórica).

Seis variables independientes, una variable dependiente.

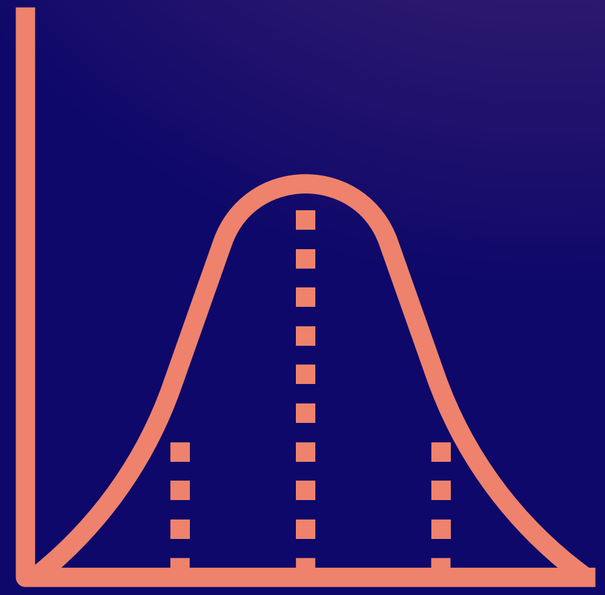
Data Preparation

How do we organize the data for modeling?

Dataframes de training y testing, **sin imputación.**



Variables independientes (X), variable dependiente (y)



Normalización de datos

Data Preparation

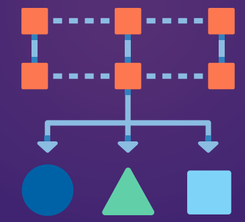
How do we organize the data for modeling?

Dataframes con **las tres clases,** sin modificaciones.

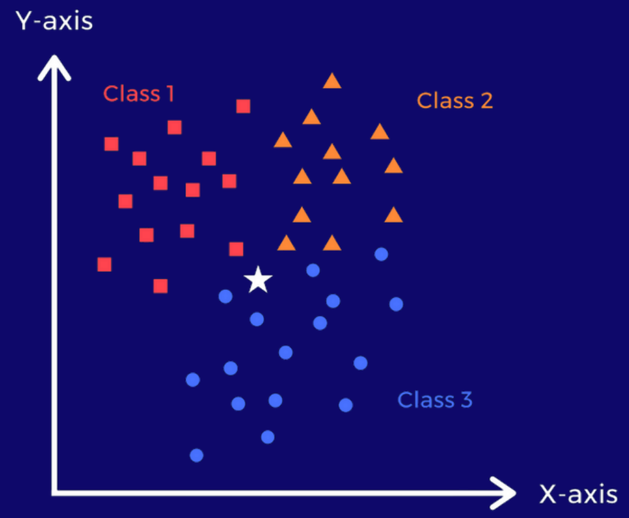
Dataframes con **sólo dos clases.**

Modeling

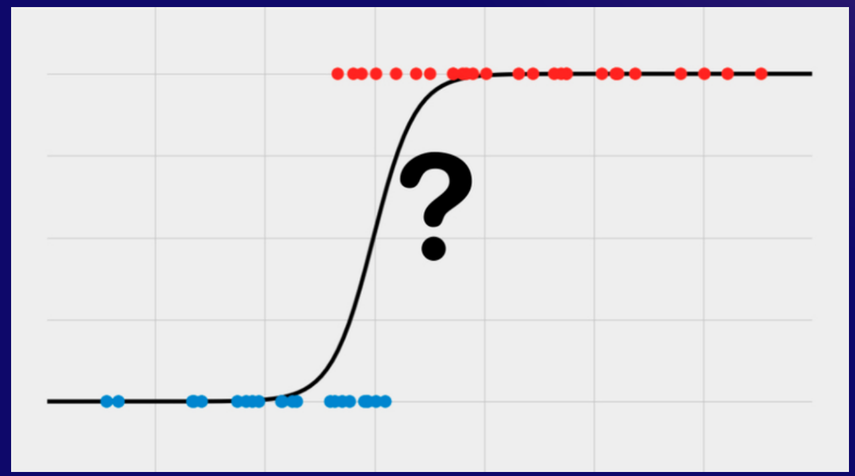
What modeling techniques should we apply?



Classification



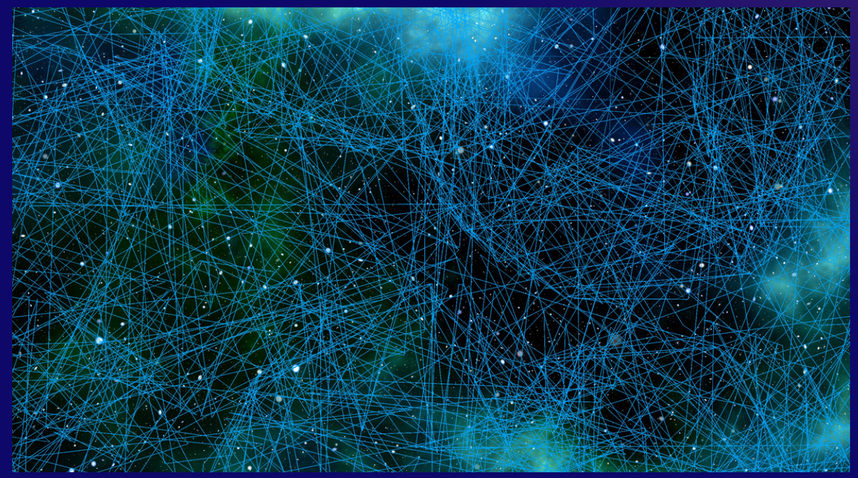
K-Nearest Neighbors, (KNN algorithm)



Logistic Regression Algorithm

Modeling

What modeling techniques should we apply?



Classification

Random Forest Algorithm

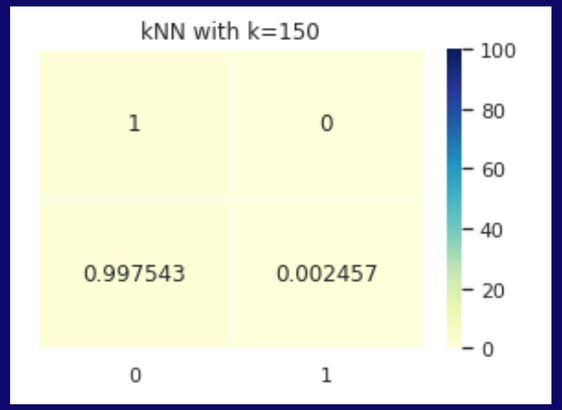
Neural Network Algorithm

Evaluation

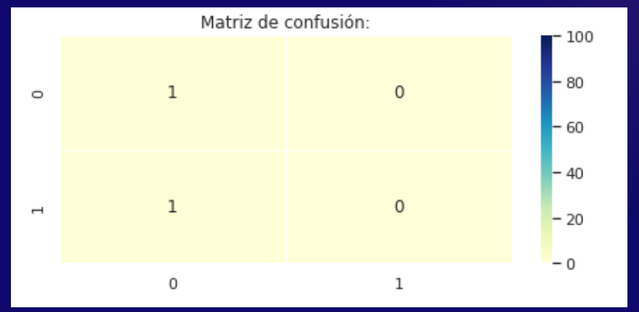
Which model best meets the business objectives?

Matrices de confusión

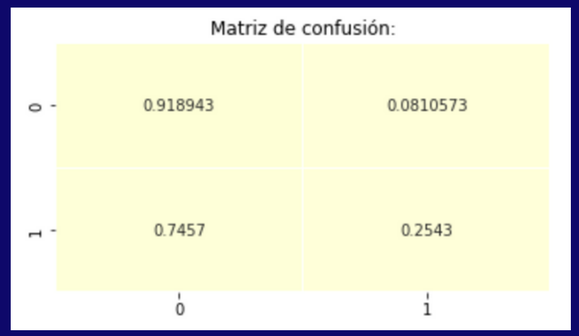
KNN Algorithm



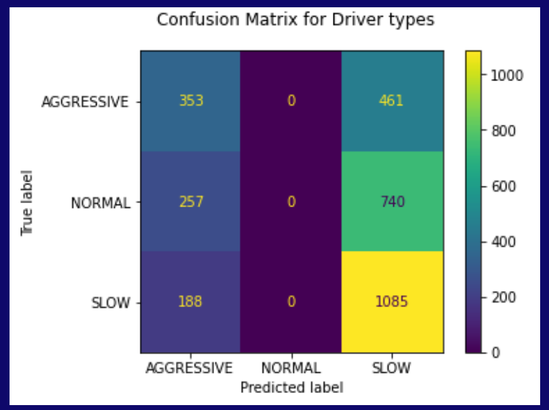
Logistic Regression Algorithm



Random Forest Algorithm



Neural Network Algorithm



Evaluation

Which model best meets the business objectives?

Reportes de clasificación

KNN Algorithm

Reporte de clasificación:

	precision	recall	f1-score	support
0	0.74	1.00	0.85	2270
1	1.00	0.00	0.00	814
accuracy			0.74	3084
macro avg	0.87	0.50	0.43	3084
weighted avg	0.81	0.74	0.63	3084

Logistic Regression Algorithm

Reporte de clasificación:

	precision	recall	f1-score	support
0	0.74	1.00	0.85	2270
1	0.00	0.00	0.00	814
accuracy			0.74	3084
macro avg	0.37	0.50	0.42	3084
weighted avg	0.54	0.74	0.62	3084

Random Forest Algorithm

Reporte de clasificación:

	precision	recall	f1-score	support
0	0.77	0.92	0.84	2270
1	0.53	0.25	0.34	814
accuracy			0.74	3084
macro avg	0.65	0.59	0.59	3084
weighted avg	0.71	0.74	0.71	3084

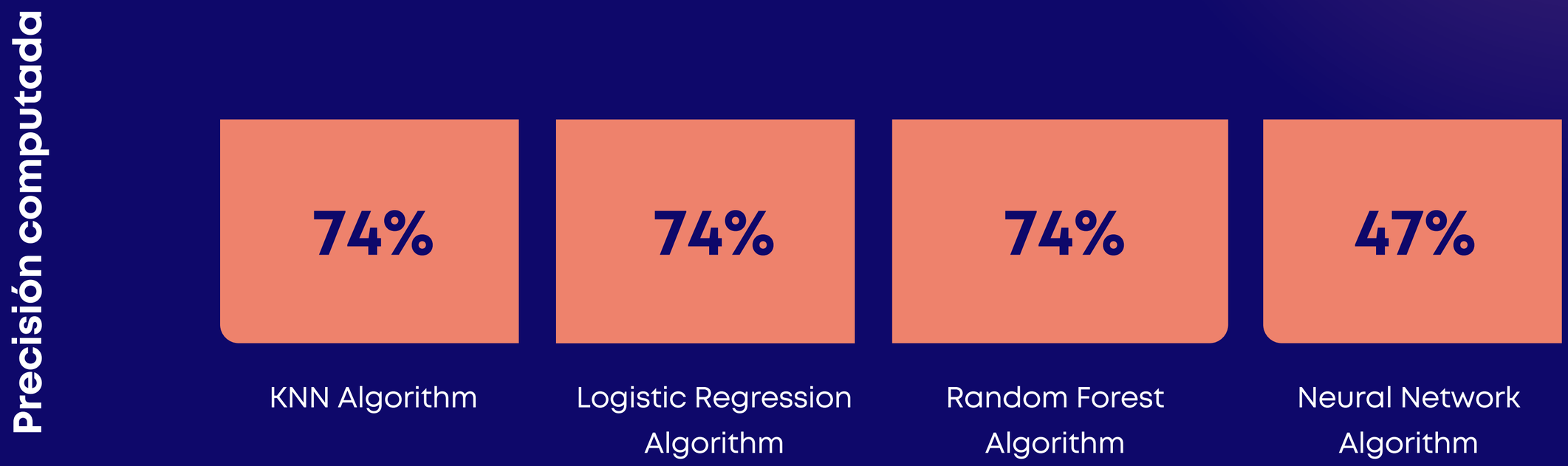
Neural Network Algorithm

Reporte de clasificación:

	precision	recall	f1-score	support
AGGRESSIVE	0.44	0.43	0.44	814
NORMAL	0.00	0.00	0.00	997
SLOW	0.47	0.85	0.61	1273
accuracy			0.47	3084
macro avg	0.31	0.43	0.35	3084
weighted avg	0.31	0.47	0.37	3084

Evaluation

Which model best meets the business objectives?



Deployment

How do stakeholders access the results?

kaggle

Subida de resultados a Kaggle

Documento de LaTeX y presentación de la solución.

Conclusiones

TC3006C.101

Driving Behavior Challenge

Diana

Hay que tomar los resultados con pinzas, al estar sesgado el experimento.

Herón

Es importante conocer todos los algoritmos de machine learning para poder saber elegir el modelo óptimo.

David

Las muestras tomadas son lo más importante al momento de realizar modelos de ML, para una mejor precisión en las predicciones.

Daniel

Para un mejor modelo es importante tener todos los conocimientos teoricos para poder implementarlos en la práctica al resolver una problemática de predicción.