CC5205 MINERIA DE DATOS

PRESENTACIÓN HITO 3 GRUPO 22

INTEGRANTES:

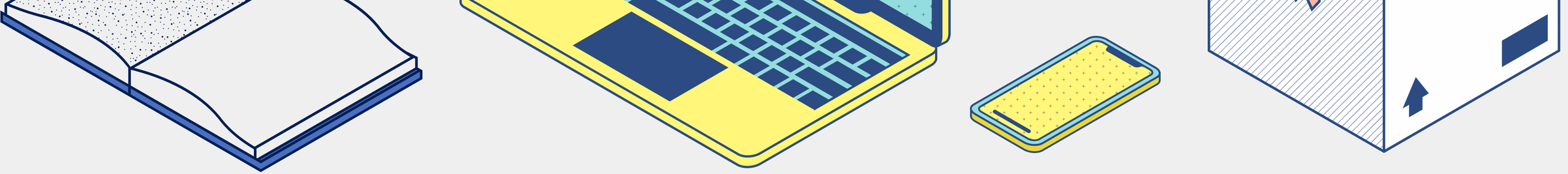
Camila Fuentes

Javier Kauer

Felipe Mellado

Diego Faúndez

Benjamín San Martín

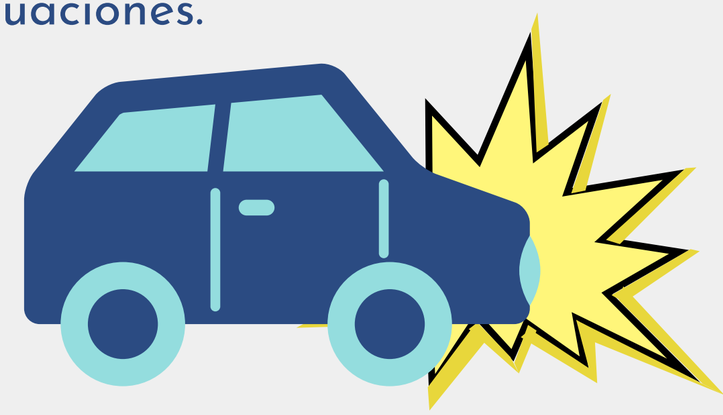


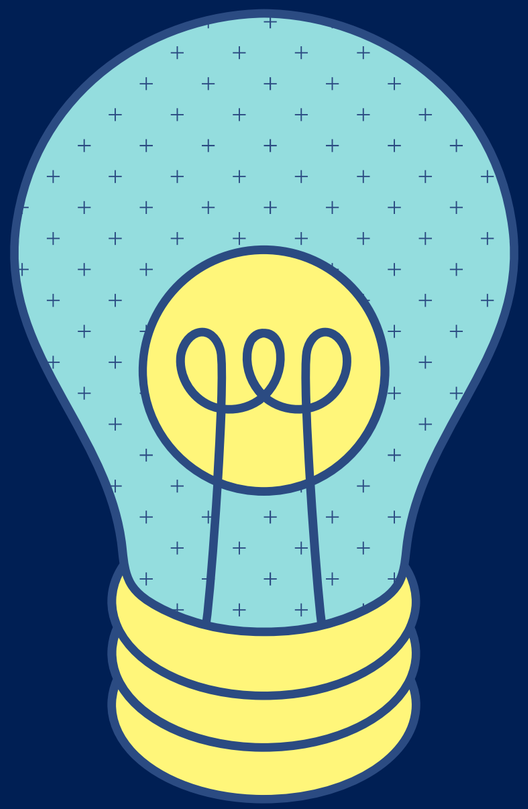
PROBLEMA

- Los accidentes automovilisticos causan daños materiales y perdida de vehículos.
- También interrumpen el tráfico y, lo más lamentable, provocan pérdida de vidas.
- Es importante anticipar los factores del entorno para reducir la gravedad de los accidentes viales.

OBJETIVO

Nuestra misión es analizar una base de datos de accidentes de tráfico para crear un modelo capaz de predecir su gravedad mediante el analisis de diversos factores, con el fin de estar preparados en estas situaciones.





CAMBIOS CON RESPECTO A LOS HITOS ANTERIORES

Se consideró que se podía sacar provecho de atributos eliminados en el hito 1, tales como la fecha.

Se construyó una nueva matriz de correlación.

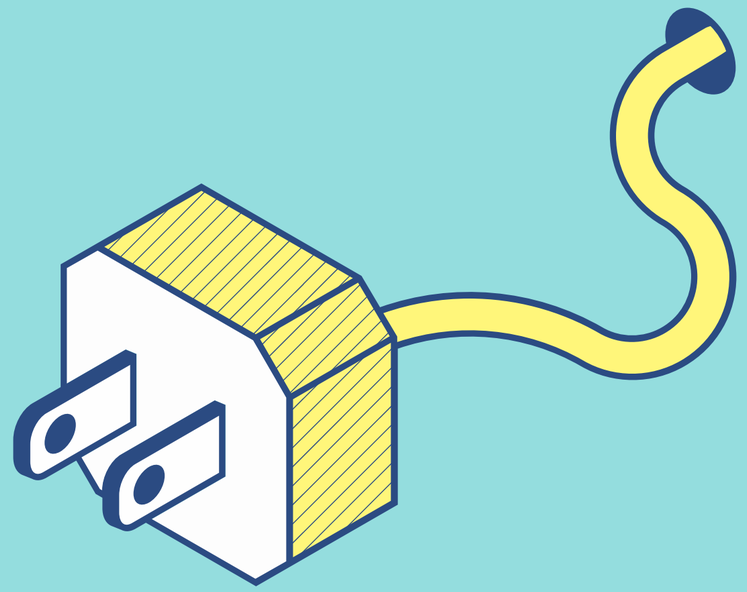
Se cambiaron preguntas que se consideraban poco interesantes.

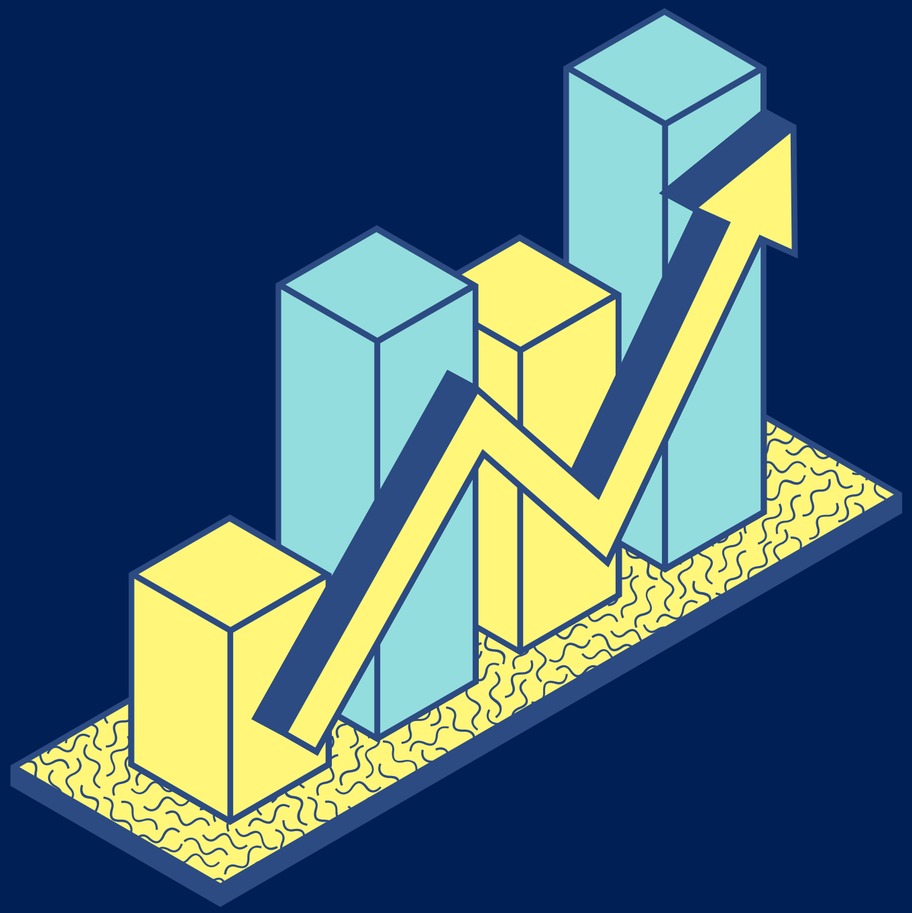
PREGUNTAS Y PROBLEMAS

1. Ocupando los datos balanceados mediante undersampling, ¿Se puede generar un modelo de predicción aceptable de la severidad de un accidente en base a las cualidades del entorno?

2. ¿Qué tanto varían las métricas estadísticas de un modelo desbalanceado a uno balanceado?

3. ¿Existen patrones comunes entre los accidentes al evaluar las condiciones del entorno?





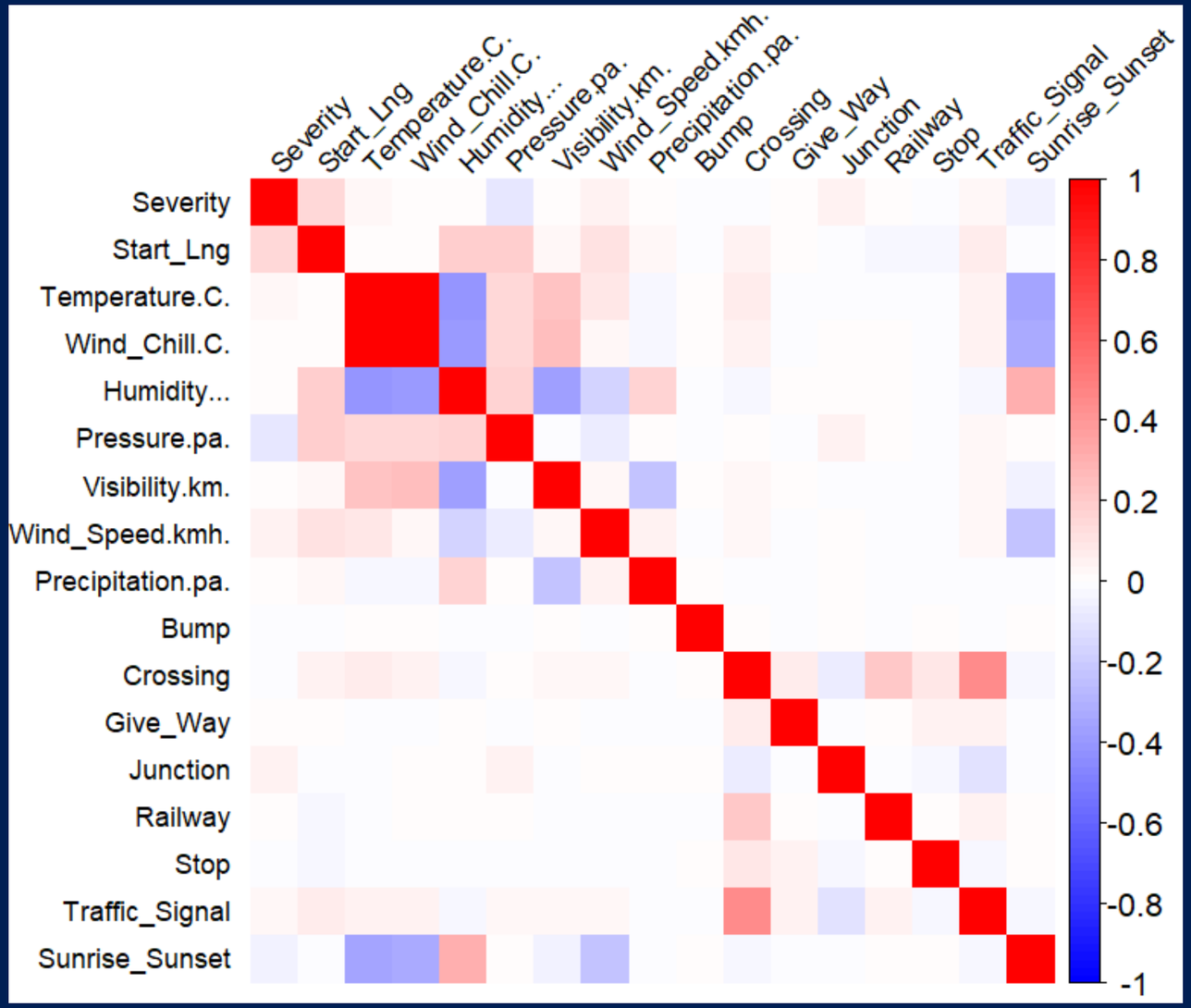
DATOS

Habían muchas columnas que tenían datos faltantes en abundancia (Latitud, Longitud accidente, descripción por mencionar algunos)

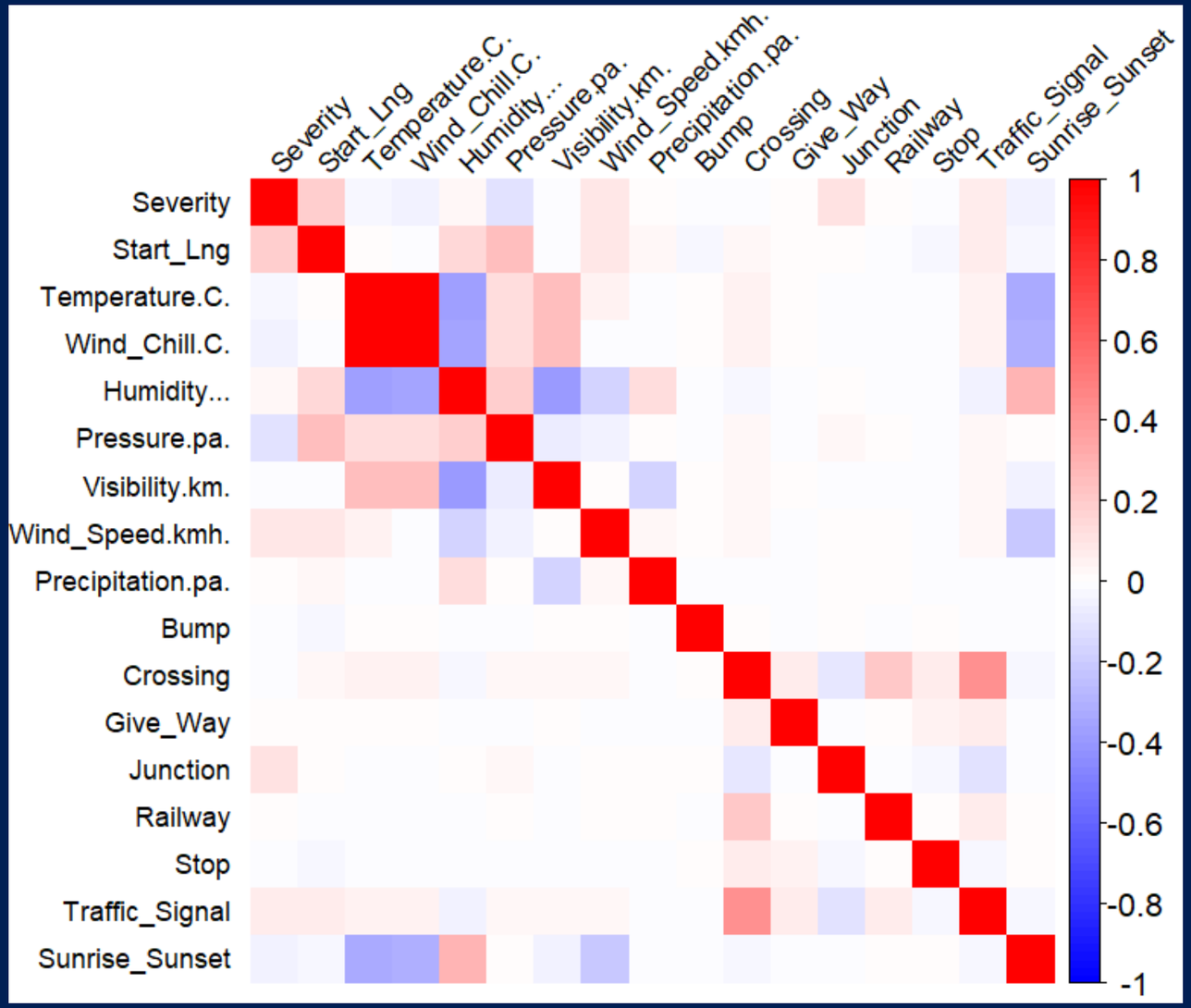
Hubieron algunas que no consideramos por no ser muy relevantes (Distancia, calle, ciudad, condado, etc.)

La gran mayoría de los datos de severidad se agrupaban en la categoría 2.

MATRIZ SIN UNDERSAMPLING



MATRIZ CON UNDERSAMPLING



MÉTODOS EXPERIMENTALES Y RESULTADOS

MÉTODO EXPERIMENTAL PREGUNTA 1

Ocupando los datos balanceados mediante undersampling, ¿Se puede generar un modelo de predicción aceptable de la severidad de un accidente en base a las cualidades del entorno?

1 — 2 — 3 — 4

PASO 1

Se eliminaron columnas de string con poca relevancia y se convirtieron las columnas booleanas a 0's y 1's en las bases de datos.

PASO 2

Se entrenaron 3 diferentes modelos (Decision tree, BD y KNN (nn=10)).

PASO 3

Entrenamiento: Se dividió la base CON undersampling en 70% entreno y 30% testing.

PASO 4

Se evaluaron las métricas para responder la pregunta planteada.

RESULTADOS PREGUNTA 1

Testing usando el dataset con undersampling (70/30)

Base Dummy

Precision prom: 0,50

Recall prom: 0,50

F1-score prom: 0,50

Decision Tree

Precision prom: 0,67

Recall prom: 0,73

F1-score prom: 0,70

KNN

Precision prom: 0,67

Recall prom: 0,56

F1-score prom: 0,61

MÉTODO EXPERIMENTAL PREGUNTA 2

¿Qué tanto varían las métricas estadísticas de un modelo desbalanceado a uno balanceado?

1 PASO

Se eliminaron columnas de string con poca relevancia y se convirtieron las columnas booleanas a 0's y 1's en las bases de datos.

2 PASO

Se entrenaron 3 diferentes modelos (Decision tree, BD y KNN (nn=10)).

3 PASO

Entrenamiento 1: Se dividió la base CON undersampling en 70% entreno y 30% testing.

4 PASO

Entrenamiento 2: Se utilizó el entrenamiento implementado en la pregunta anterior

5 PASO

Se evaluaron las métricas y se eligieron los mejores modelos de cada entrenamiento para responder la pregunta planteada.

RESULTADOS PREGUNTA 2

Testing usando el dataset con undersampling (70/30)

Base Dummy

Precision prom: 0,50

Recall prom: 0,50

F1-score prom: 0,50

Decision Tree

Precision prom: 0,67

Recall prom: 0,73

F1-score prom: 0,70

KNN

Precision prom: 0,67

Recall prom: 0,56

F1-score prom: 0,61

Testing usando el dataset sin undersampling (70/30)

Base Dummy

Precision prom: 0,11

Recall prom: 0,11

F1-score prom: 0,11

Decision Tree

Precision prom: 0,61

Recall prom: 0,18

F1-score prom: 0,28

KNN

Precision prom: 0,58

Recall prom: 0,06

F1-score prom: 0,10

MÉTODO EXPERIMENTAL PREGUNTA 3

¿Existen patrones comunes entre los accidentes al evaluar las condiciones del entorno?

1 PASO

Se ocupa todo los atributos del dataset, y coneritods en valores numéricos

2 PASO

Usando la base de datos sin undersampling se utiliza el método del codo para obtener la cantidad óptima de clusters.

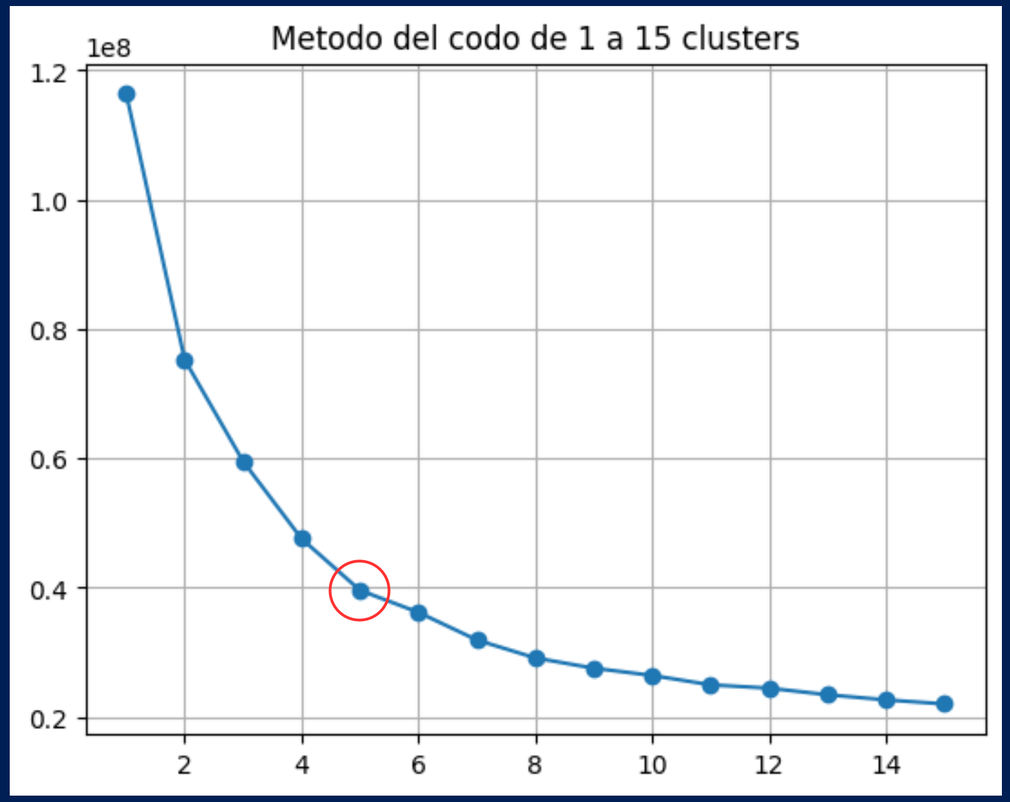
3 PASO

Se utiliza K-Means para hacer la separación de los datos

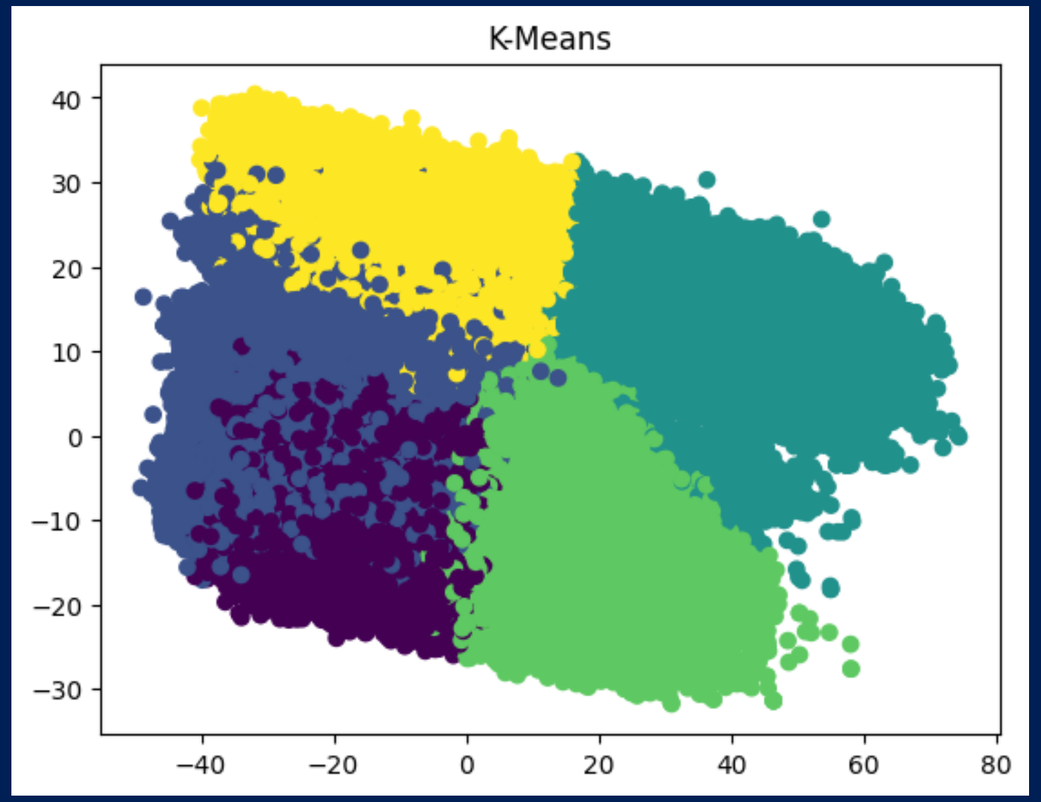
4 PASO

Finalmente se utiliza PCA para poder visualizar los resultados obtenidos

RESULTADOS PREGUNTA 3



Visualización con 5 clusters



FUTURAS DIRECCIONES

1

Utilizar modelos mas avanzados de clasificación con algoritmos mas complejos como random forest o redes neuronales para comparar el rendimiento.

2

Complementar con una base que mantenga mas un patrón y no sea tan aleatorio.

3

Utilizar modelos de regresión para realizar mejores predicciones numericas sobre la severidad.

4

Separar estados, hacer algo mas específico.

5

Utilizar otras tecnicas de balanceos de datos como oversampling o el metodo de generacion de muestras sinteticas (SMOTE).

MUCHAS GRACIAS POR SU ATENCIÓN