CICE
LA ESCUELA PROFESIONAL
DE NUEVAS TECNOLOGÍAS

LAKE

Trabajo de fin de Máster

Escuela profesional de nuevas tecnologías.

Big Data

Álvaro Martínez, Guillermo Herranz, Marta Pérez, Rubén Márquez y Pablo Andreu

INDICE

01. Caso de uso y planteamiento de la empresa

Elección idea

Transporte Público

Twitch

Accidentes tráfico

+ facilidad acceso fuentes información
+ cantidad de información
+ calidad de información

01. Caso de uso y planteamiento de la empresa

Visión del proyecto

Madrid más seguro

Análisis de información

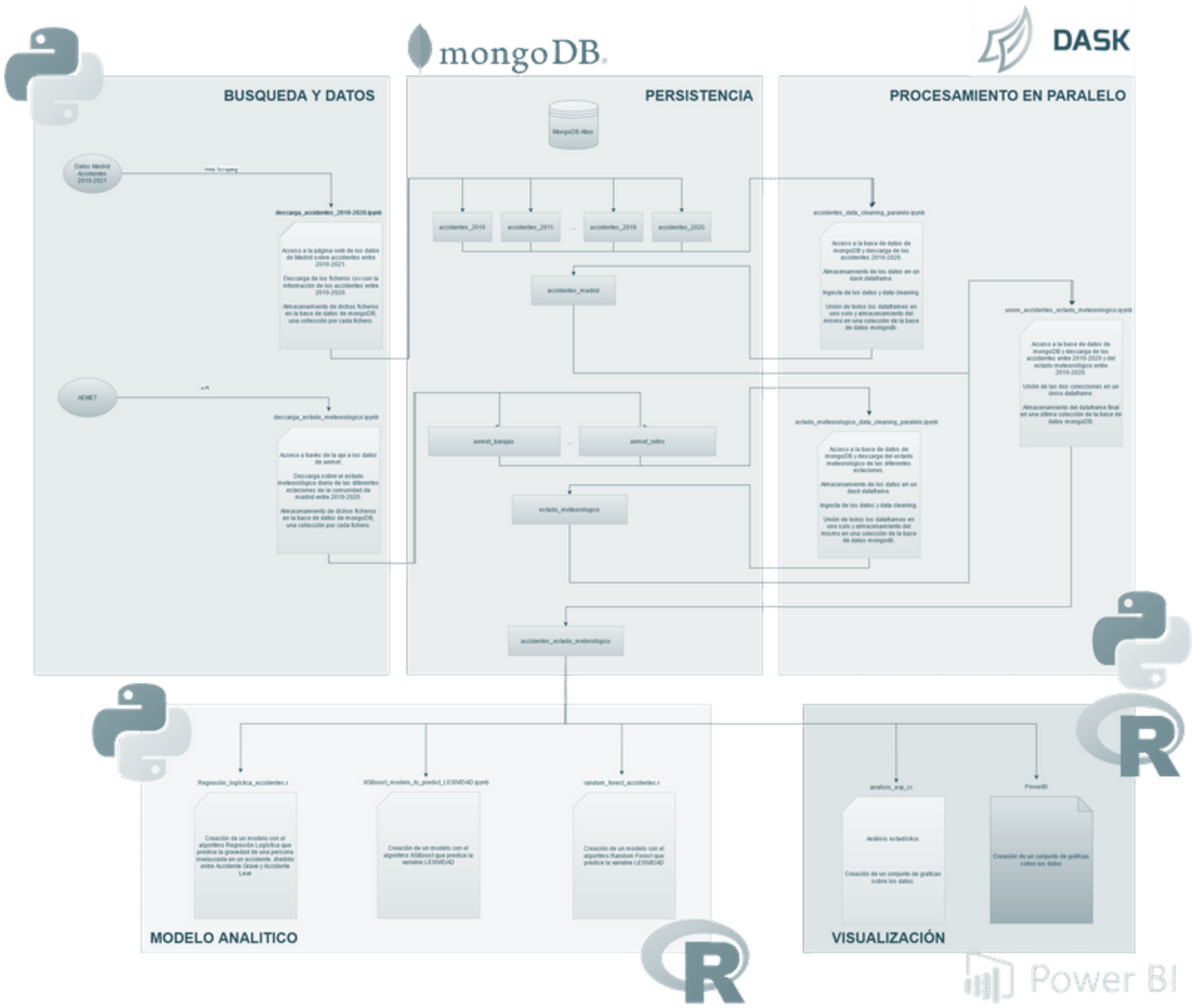
Medidas de desempeño

Desarrollo del proyecto
Rendimiento del proyecto
Suposiciones
Restricciones
Entorno tecnológico
Hitos principales
Evaluación
Alineamiento estratégico
Análisis impacto de la organización

01. Caso de uso y planteamiento de la empresa

Servicios analíticos y estratégicos basados en el tratamiento de grandes cantidades de información.

02. Arquitectura del proyecto



03. Búsqueda y datos

Portal de datos abiertos de la comunidad de Madrid

Acceso a datos de

Accidentes de tráfico
https://datos.madrid.es/portal/site/egob

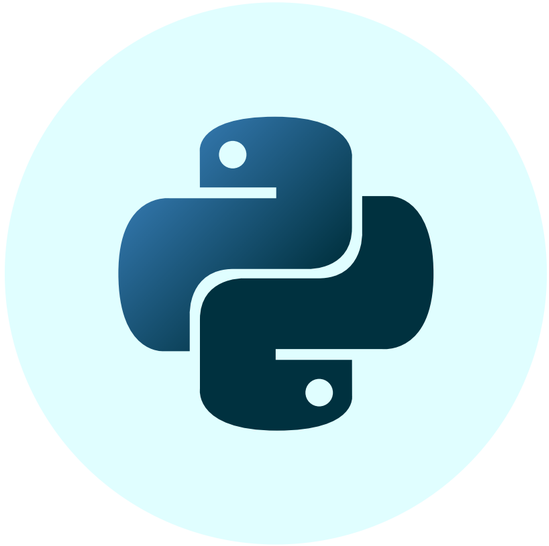
Estados meteorológicos
http://www.aemet.es/es/datos_abiertos

04. Ingesta con persistencia y data cleaning

Almacenamiento de la información

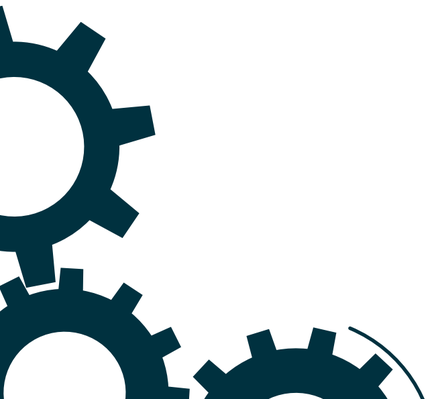
Datacleaning

05. Procesamiento en paralelo



Python

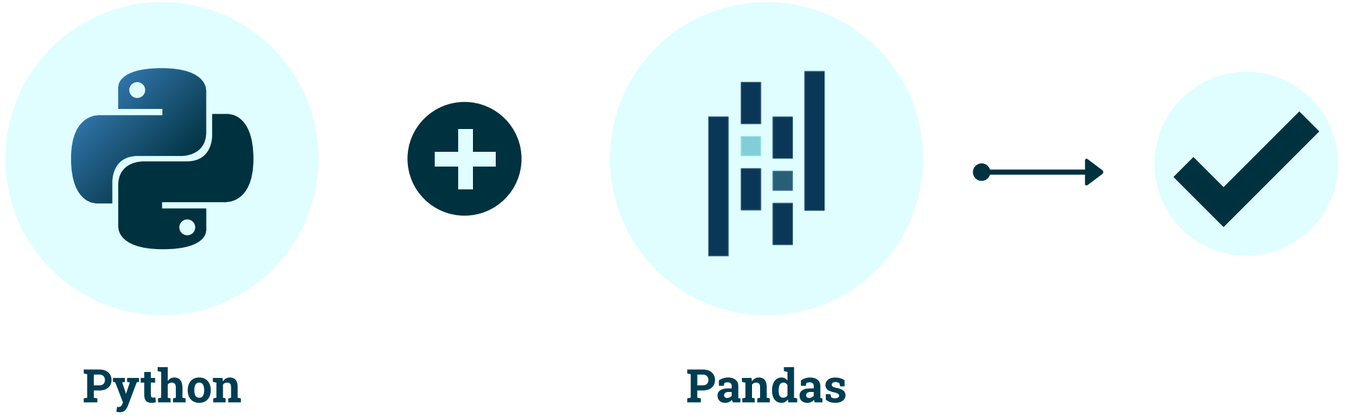
Dask



05. Procesamiento en paralelo

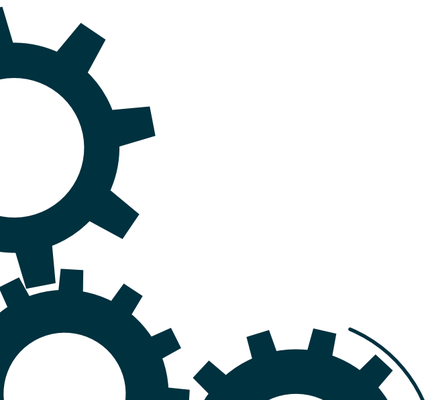
Python + Dask → ✗

No es eficiente para este tipo de muestra



Python + Pandas → ✓

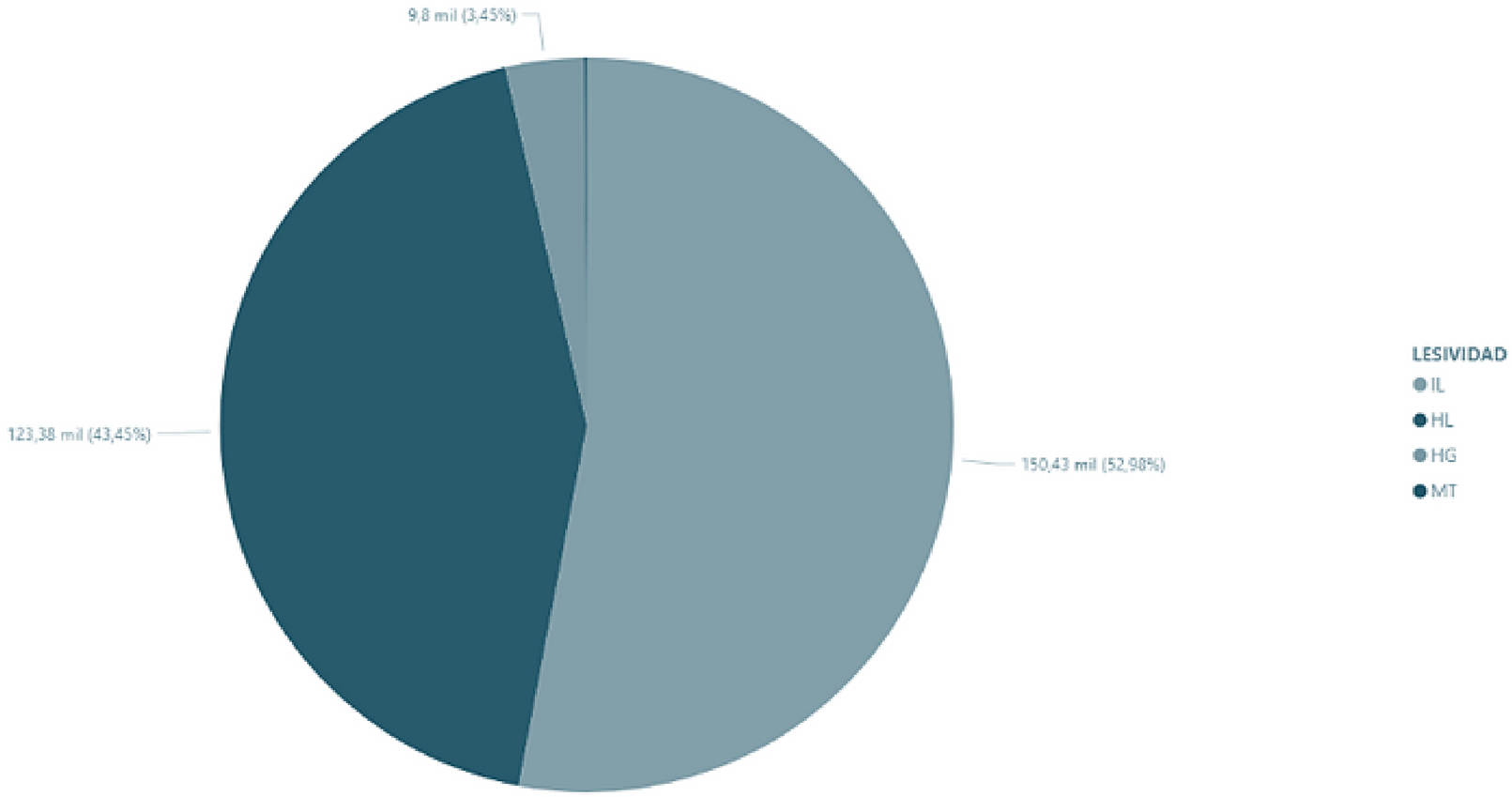
Es eficiente para este tipo de muestra



06. Visualización (Análisis muestra)

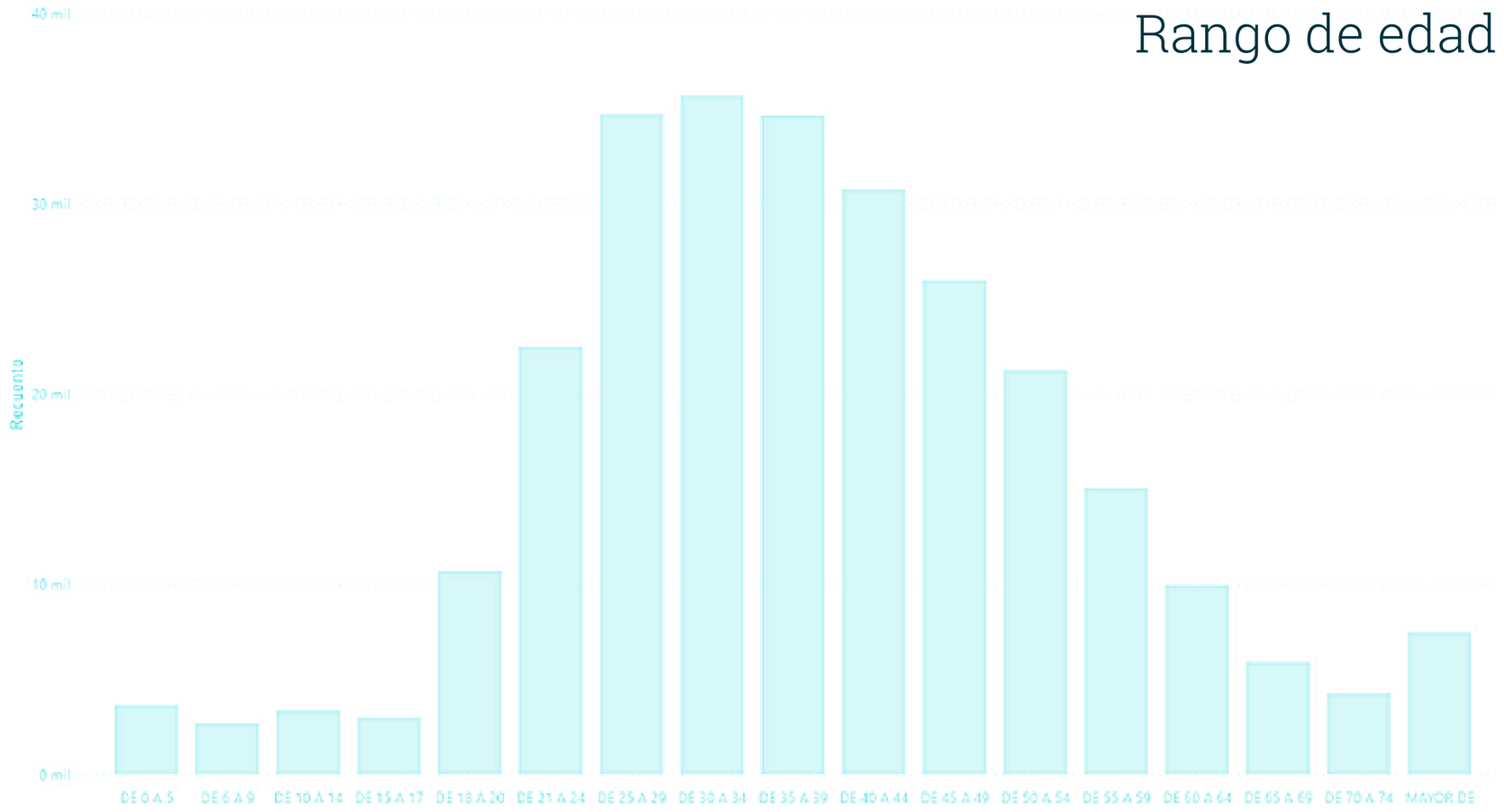
PowerBI

06. Visualización (Análisis muestra)



Tipo de Lesividad

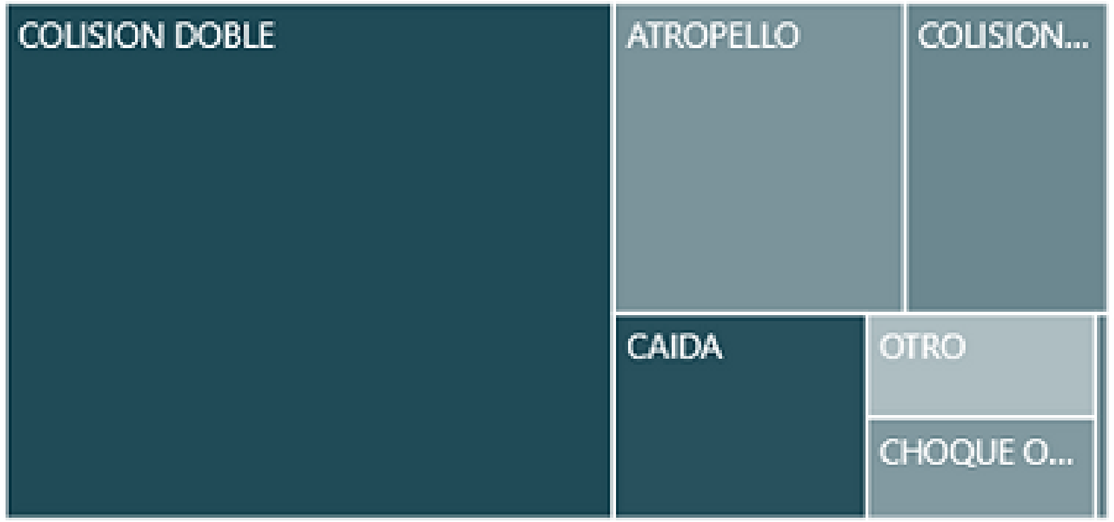
06. Visualización (Análisis muestra)



06. Visualización (Análisis muestra)

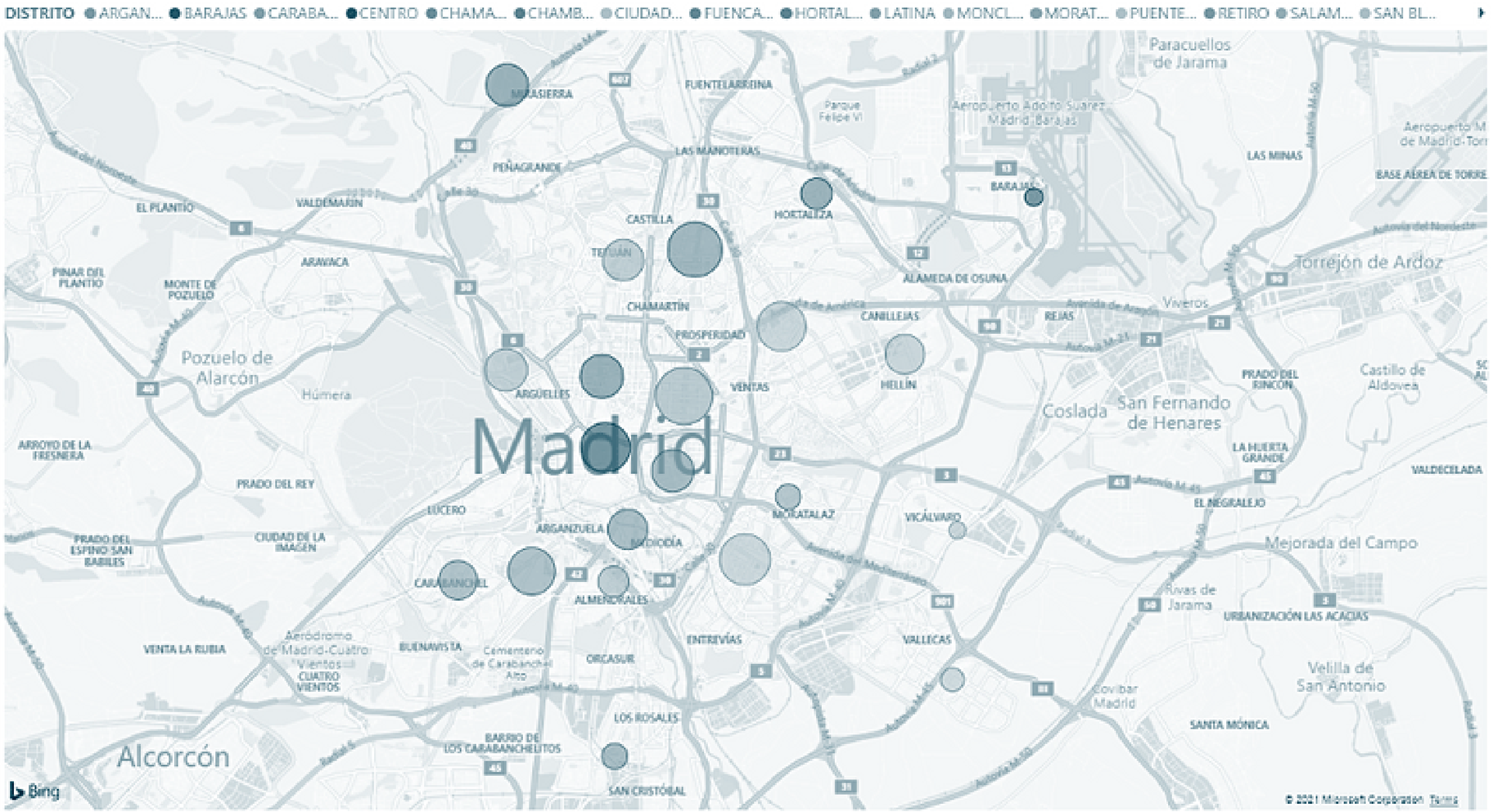
TIPO ACCIDENTE	HG	HL	IL	MT	Total
COLISION DOBLE	4086	67520	84991	79	**156676**
ATROPELLO	3295	14650	27145	171	**45261**
COLISION MULTIPLE	419	12381	18607	15	**31422**
CAIDA	1114	15729	9017	32	**25892**
OTRO	132	5504	6254	5	**11895**
CHOQUE OBSTACULO FIJO	676	6784	4064	40	**11564**
VUELCO	76	806	339	2	**1223**
DESCONOCIDO	2	4	8		**14**
Total	**9800**	**123378**	**150425**	**344**	**283947**

Recuento por TIPO ACCIDENTE



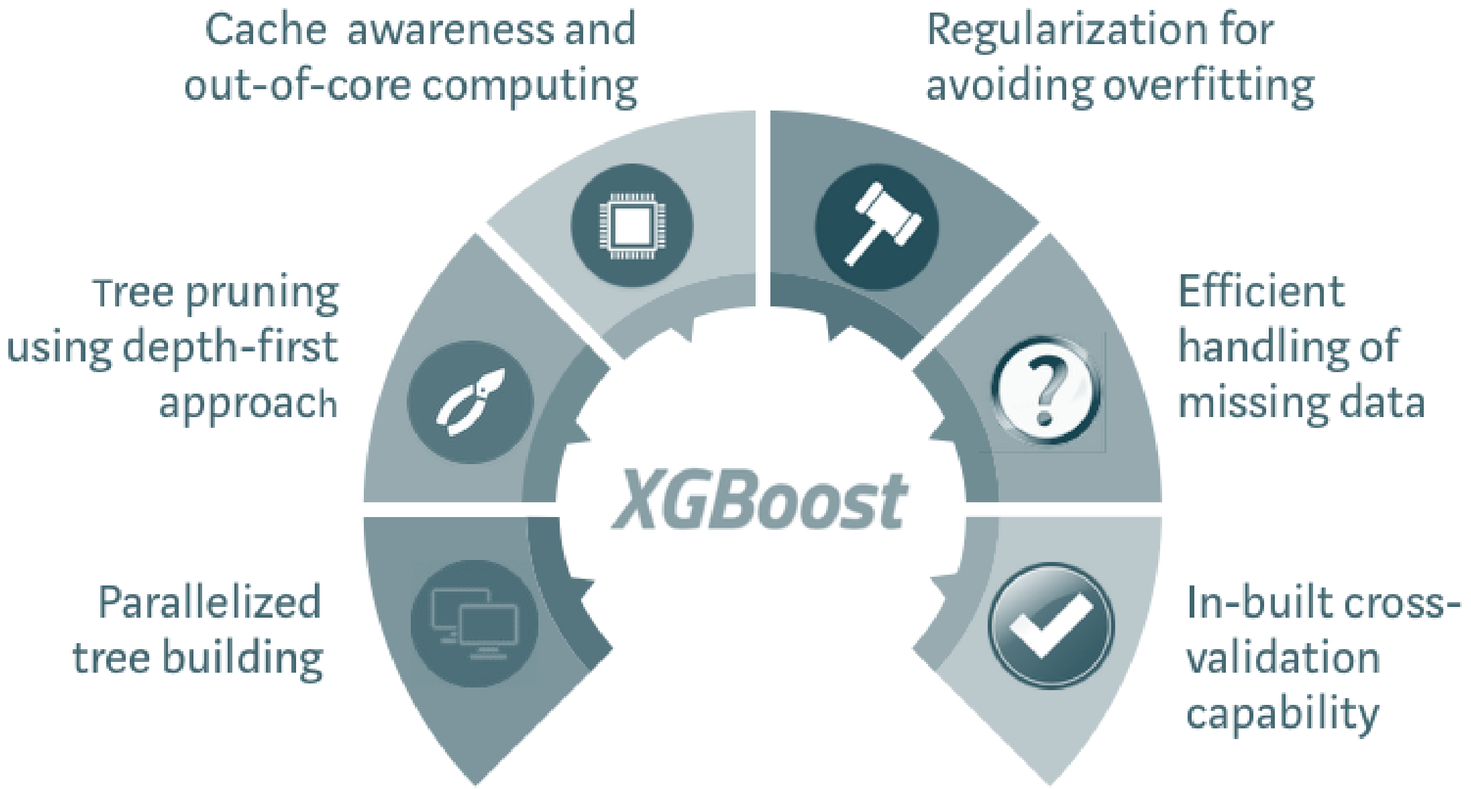
06. Visualización (Análisis muestra)

Mapa DISTRITO



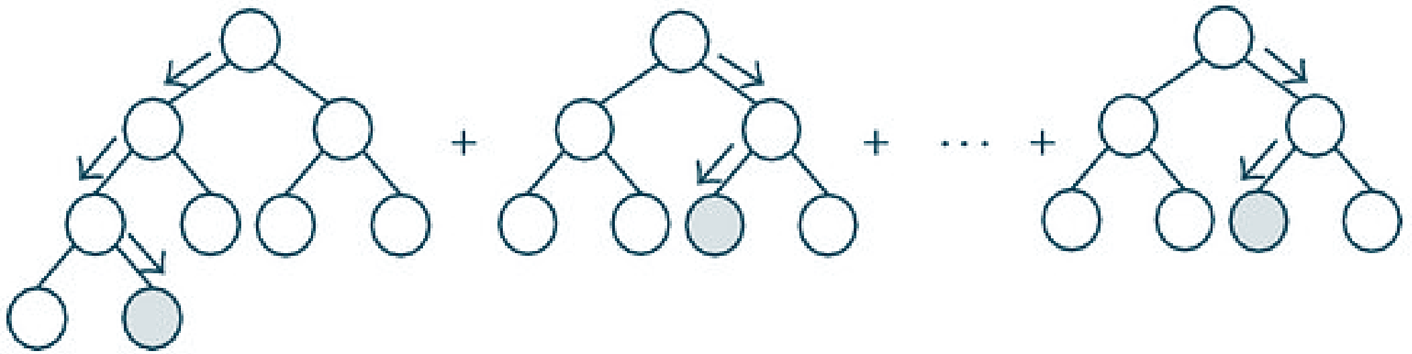
07. Modelos analíticos

XGBoost



07. Modelos analíticos

XGBoost



Basado en Árboles de decisión

Multitud de árboles de decisión secuenciales

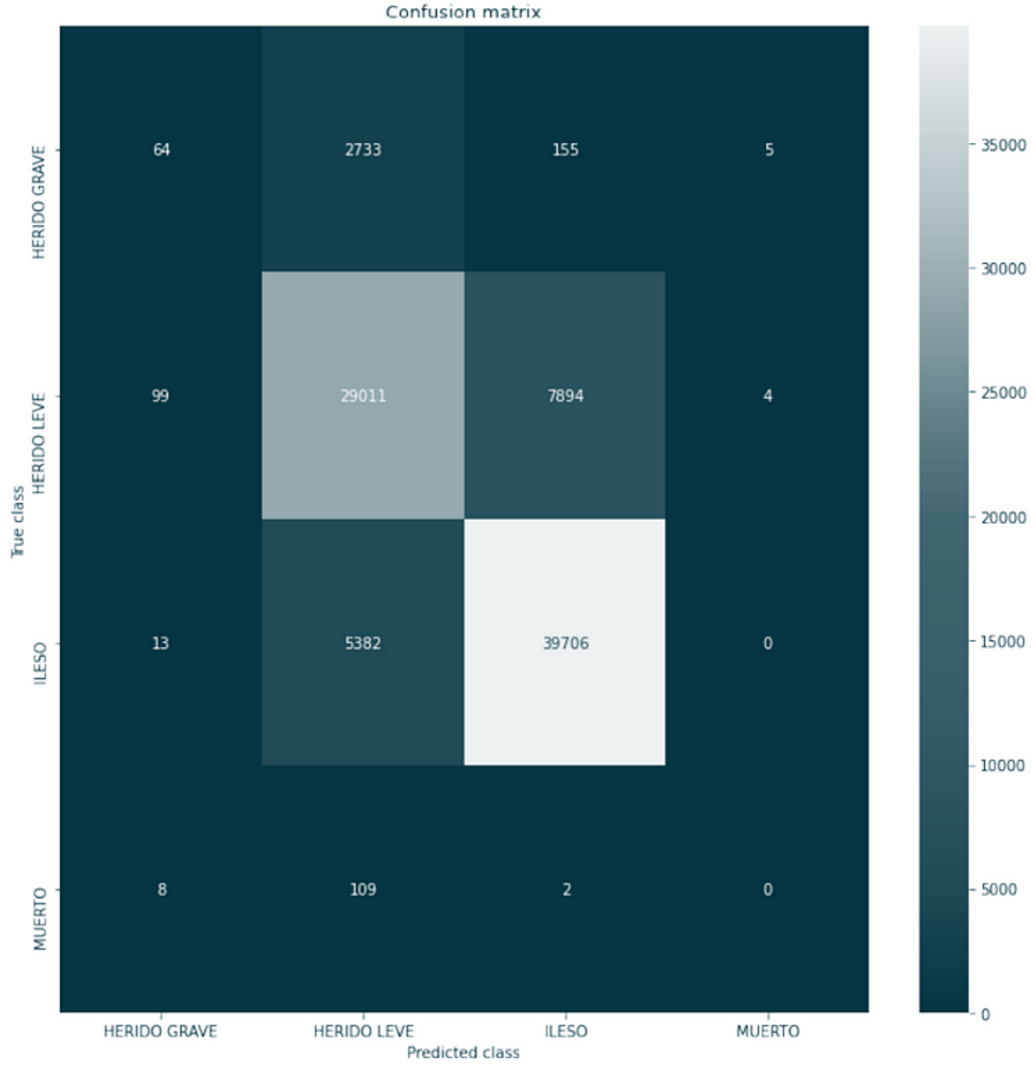
Árboles de decisión cada vez mas profundos

Algoritmo supervisado de ML

Eficiencia computacional

07. Modelos analíticos

XGBoost



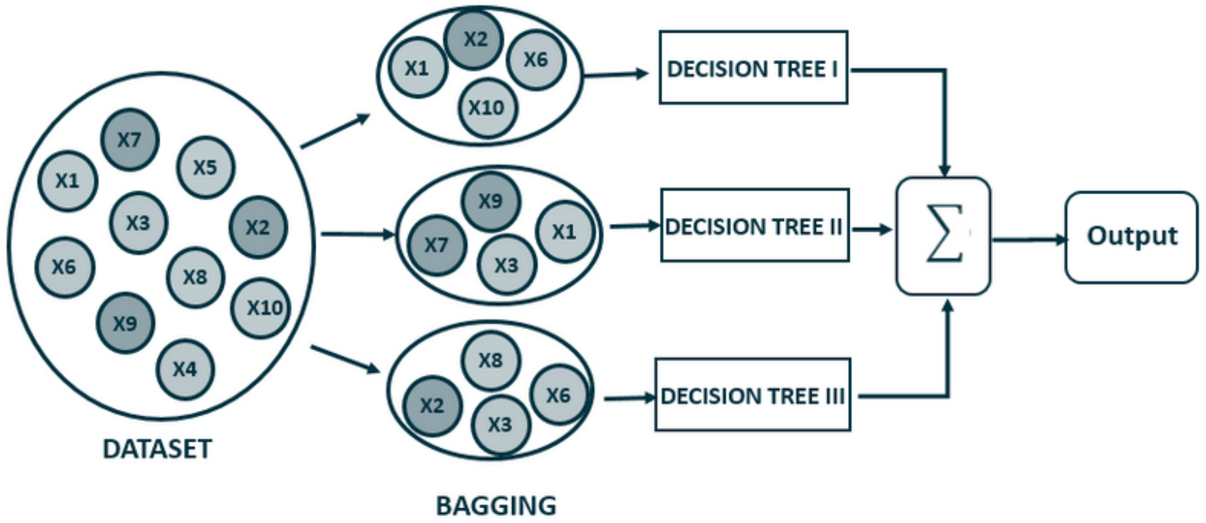
El algoritmo predice la variable objetivo 'LESIVIDAD' con una precisión del 80,74%.

La cantidad de información de la categoría 'MUERTO' y 'HERIDO GRAVE' resulta muy escasa para poder hacer una estimación lo suficientemente precisa.

Si se amplia la información resultaría en un modelo robusto.

07. Modelos analíticos

Random Forest



Equilibrio entre sesgo y varianza.

Buen modelo para la introducción de multitud de variables, discriminando las menos relevantes.

Incorpora métodos efectivos para estimar valores faltantes.

07. Modelos analíticos

Random Forest

RStudio

Accuracy

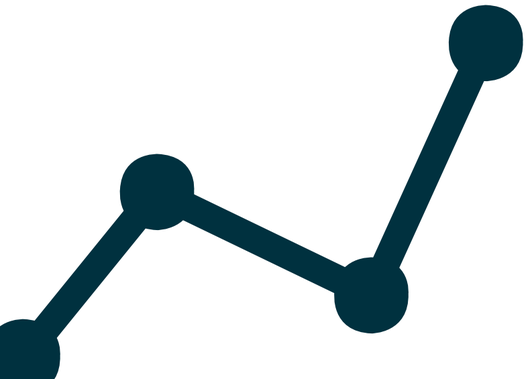
El modelo tiene una precisión del 78% para poder predecir la cantidad de accidentes que se van a producir.

Recall

Heridas graves (HG): 48% de sensibilidad
Heridas leves (HL): 77% de sensibilidad
Ingresos leves (IL): 78% de sensibilidad
Fallecidos (MT): 1% de sensibilidad

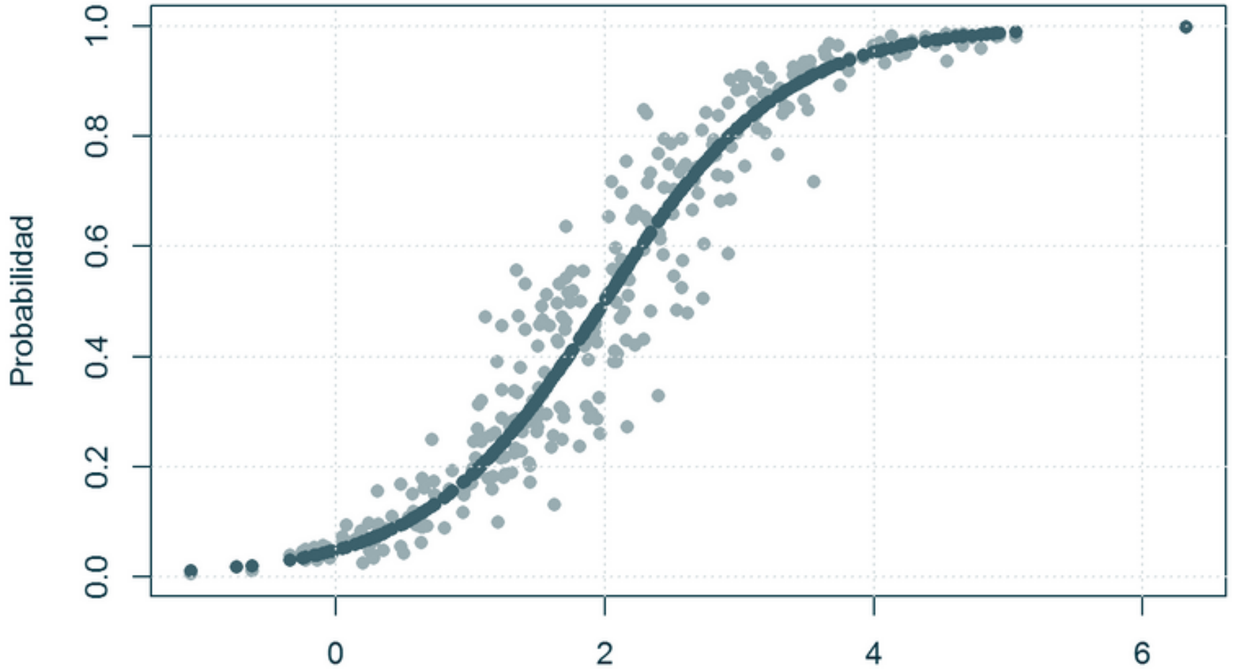
Especificidad

Heridas graves (HG): 96% de sensibilidad
Heridas leves (HL): 79% de sensibilidad
Ingresos leves (IL): 85% de sensibilidad
Fallecidos (MT): 4% de sensibilidad



07. Modelos analíticos

Regresión Logística



Simplicidad.

Resultados facilmente interpretables.

Es extraño que exista sobreajuste

07. Modelos analíticos

Regresión Logística

Accuracy El modelo tiene una precisión del 96% para poder predecir la cantidad de accidentes que se van a producir.

Recall La sensibilidad del modelo es de un 99%

Especificidad La especificidad del modelo es de un 7%