**1. Introducción**

La seguridad vial constituye uno de los principales retos de las administraciones y organismos públicos. El análisis de accidentes de tráfico permite identificar patrones y factores de riesgo asociados a la siniestralidad vial. En este proyecto, se propone utilizar técnicas de Machine Learning para analizar y predecir la gravedad de los accidentes de tráfico en España, basándonos en los datos abiertos proporcionados por la Dirección General de Tráfico (DGT) correspondientes al año 2022.

El objetivo es predecir la gravedad de un accidente (definido como leve, grave o mortal) en función de distintas variables relacionadas con el entorno, la infraestructura y las condiciones del accidente. Esta información podría resultar clave para la toma de decisiones en políticas de prevención y seguridad vial.

**2. Variables a Emplear**

El dataset proporcionado por la DGT contiene información detallada de todos los accidentes de tráfico registrados en España durante el año 2022. A partir de este conjunto de datos, se han seleccionado las siguientes **variables predictoras** que, según estudios previos y conocimiento del dominio, tienen un impacto significativo en la ocurrencia y gravedad de los accidentes:

| **Nombre de la variable** | **Descripción** |
| --- | --- |

|  |  |
| --- | --- |
| MES | Mes del año en que ocurrió el accidente (enero, febrero, etc.). |

|  |  |
| --- | --- |
| DIA\_SEMANA | Día de la semana del accidente (lunes, martes, etc.). |

|  |  |
| --- | --- |
| HORA | Hora del día en la que ocurrió el accidente (0 a 23). |

|  |  |
| --- | --- |
| COD\_PROVINCIA | Código numérico de la provincia donde ocurrió el accidente. |

|  |  |
| --- | --- |
| ZONA | Tipo de zona: urbana, interurbana, travesía, etc. |

|  |  |
| --- | --- |
| TIPO\_VIA | Tipo de vía: autopista, carretera nacional, vía urbana, etc. |

|  |  |
| --- | --- |
| CONDICION\_ILUMINACION | Estado de iluminación: día, noche con iluminación, noche sin iluminación. |

|  |  |
| --- | --- |
| CONDICION\_METEO | Condiciones meteorológicas: lluvia, nieve, niebla, etc. |

|  |  |
| --- | --- |
| CONDICION\_FIRME | Estado del pavimento: seco, mojado, en obras, etc. |

|  |  |
| --- | --- |
| CONDICION\_NIVEL\_CIRCULA | Nivel de circulación: denso, fluido, congestionado, etc. |

|  |  |
| --- | --- |
| VISIB\_RESTRINGIDA\_POR | Causa de visibilidad reducida: curva, niebla, vehículos estacionados, etc. |

|  |  |
| --- | --- |
| TRAZADO\_PLANTA | Diseño de la carretera en planta: recta, curva, cambio rasante, etc. |

|  |  |
| --- | --- |
| PRIORI\_SEMAFORO | Existencia de semáforo en el lugar (1: sí, 0: no). |

|  |  |
| --- | --- |
| PRIORI\_VERT\_STOP | Existencia de señal vertical de STOP (1: sí, 0: no). |

|  |  |
| --- | --- |
| PRIORI\_HORIZ\_CEDA | Existencia de señal horizontal de Ceda el Paso (1: sí, 0: no). |

|  |  |
| --- | --- |
| TOTAL\_VEHICULOS | Número total de vehículos implicados en el accidente. |

|  |  |
| --- | --- |
| ACERA | Si el accidente ocurrió sobre la acera o no. |

|  |  |
| --- | --- |
| TOTAL\_VICTIMAS\_24H | Total de personas heridas o fallecidas en las 24 horas posteriores al accidente. |

|  |  |
| --- | --- |
| TOTAL\_MU24H | Número de fallecidos en las 24 horas posteriores al accidente. |

|  |  |
| --- | --- |
| TOTAL\_HG24H | Número de heridos graves en las 24 horas posteriores al accidente. |

**Variable objetivo (target):** Se definirá una variable categórica que clasifique la gravedad del accidente:

* Leve (sin heridos graves ni muertos)
* Grave (al menos un herido grave)
* Mortal (al menos un fallecido)

**3. Modelos de Machine Learning a Emplear**

En este proyecto, el objetivo es clasificar el tipo de accidente de tráfico en función de diversas variables predictoras, con el fin de detectar patrones que permitan comprender los factores que influyen en la gravedad de los accidentes. Al no contar con una variable objetivo directamente etiquetada como "leve", "grave" o "mortal", se plantea la construcción de una nueva variable target (por ejemplo, en base al número de fallecidos y heridos graves), lo que nos sitúa en un contexto de aprendizaje **supervisado**.

**1. Regresión Logística**

Es un modelo de base ideal para clasificación binaria o multiclase. Su sencillez permite establecer una línea base y entender rápidamente qué variables tienen más peso en la predicción.

* **Ventajas**: Fácil de interpretar, rápido de entrenar.
* **Justificación**: Sirve como modelo de referencia para comparar el rendimiento de técnicas más complejas.

**2. Árboles de Decisión**

Este modelo divide el espacio de datos en función de los valores de las variables predictoras, generando reglas de decisión claras.

* **Ventajas**: Interpretabilidad, manejo de variables categóricas sin necesidad de codificación, no requiere escalado.
* **Justificación**: Permite entender cómo interactúan las variables y qué combinaciones generan mayor riesgo.

**3. Random Forest**

Es un conjunto de árboles de decisión entrenados con distintas muestras del dataset. Mejora la precisión y evita el sobreajuste.

* **Ventajas**: Alta precisión, robustez frente al overfitting, manejo de grandes volúmenes de datos.
* **Justificación**: Muy adecuado para datasets con muchas observaciones (como el de tráfico), y permite evaluar la importancia de cada variable.

**4. XGBoost o LightGBM (si hay tiempo)**

Modelos de boosting que suelen tener un rendimiento muy alto en tareas de clasificación.

* **Ventajas**: Precisión elevada, buen manejo de valores faltantes, optimización eficiente.
* **Justificación**: Ideal para presentar un modelo final con rendimiento optimizado, siempre que se disponga de tiempo suficiente para ajustar hiperparámetros.

El modelo será evaluado usando métricas como:

* Precisión (Accuracy)
* Matriz de confusión
* F1-score
* ROC-AUC

Se dividirá el dataset en conjunto de entrenamiento y test (80/20) y se aplicará validación cruzada para evitar overfitting.

**4. Objetivos Finales del Proyecto**

* Identificar los principales factores que influyen en la gravedad de los accidentes.
* Entrenar un modelo predictivo que permita categorizar la gravedad de futuros accidentes.
* Realizar un dashboard o visualización de los patrones detectados.
* Generar conclusiones que puedan servir como base para futuras intervenciones en seguridad vial.

**5. Principales Desafíos del Proyecto**

El desarrollo de un modelo predictivo aplicado a los accidentes de tráfico en España plantea diversos retos que deben ser considerados a lo largo de todas las fases del proyecto, desde la limpieza y transformación de los datos hasta la interpretación de los resultados obtenidos por los algoritmos de Machine Learning. A continuación, se destacan los desafíos más relevantes:

**1. Calidad y completitud de los datos**

Uno de los principales problemas al trabajar con datos reales es la presencia de valores nulos, inconsistencias o registros incompletos. En el conjunto de datos proporcionado por la Dirección General de Tráfico (DGT), pueden existir variables con registros faltantes o erróneos, especialmente en campos relacionados con las condiciones del entorno o el número de víctimas. Será necesario realizar un análisis exhaustivo y una limpieza cuidadosa para garantizar la calidad de los datos utilizados en el modelo.

**2. Desbalanceo de clases**

En los datos históricos de accidentes, es habitual encontrar un gran número de accidentes leves frente a una proporción mucho menor de accidentes graves o mortales. Esta desproporción puede afectar negativamente al rendimiento de los modelos de clasificación, que tenderán a predecir siempre la clase mayoritaria. Para mitigar este problema se valorará el uso de técnicas como *oversampling* (SMOTE), *undersampling* o modificación de pesos en el modelo.

**3. Selección de variables relevantes**

Aunque el conjunto de datos cuenta con numerosas variables, no todas tienen el mismo peso o correlación con el nivel de gravedad del accidente. Será necesario aplicar técnicas de análisis exploratorio de datos (EDA) y selección de características (feature selection) para identificar aquellas variables más influyentes, con el objetivo de mejorar la interpretabilidad y el rendimiento del modelo.

**4. Naturaleza multivariable y categórica de los datos**

La coexistencia de variables categóricas y numéricas obliga a realizar una correcta transformación y codificación previa (por ejemplo, mediante *One-Hot Encoding*, *Label Encoding* o escalado), lo que puede aumentar la dimensionalidad del problema y requerir una gestión eficiente de los recursos computacionales, especialmente si se trabaja con algoritmos más complejos como Random Forest o XGBoost.

**5. Interpretabilidad del modelo**

Dado que los resultados del modelo pueden tener implicaciones sociales y políticas (por ejemplo, identificar zonas de alto riesgo o factores comunes en accidentes graves), será importante priorizar la interpretación de las predicciones. Esto podría inclinar la balanza hacia modelos más interpretables como los árboles de decisión, aunque se probarán modelos más complejos para evaluar el trade-off entre precisión y comprensibilidad.