



Universitat d'Alacant
Universidad de Alicante

DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN E
INTELIGENCIA ARTIFICIAL
ESCUELA POLITÉCNICA SUPERIOR

Predicción de asistencia en accidentes de tráfico con un modelo de aprendizaje profundo

LUIS PÉREZ-SALA GARCÍA-PLATA

Tesis presentada para aspirar al grado de

DOCTOR O DOCTORA POR LA UNIVERSIDAD DE ALICANTE
DOCTORADO EN INFORMÁTICA

Dirigida por

Dr. Jose Francisco Vicent Francés
Dr. Manuel Curado Navarro

En caso de financiación: aquí vendría el texto explicativo...

Índice general

0.1. Abstract	5
0.2. Acknowledgements	6
1. Introduction	7
1.1. Motivación	7
2. ¿Cómo medir la gravedad de un accidente?	9
3. ¿Cómo predecir la gravedad de un accidente de tráfico?	11
3.1. Revisión sobre modelos de predicción de gravedad de los accidentes de tráfico	11
3.2. Métodos de optimización de hiperparámetros	12
3.3. Algoritmos Genéticos	13
3.4. Algoritmos de construcción de matrices	15
3.5. Algoritmos de medición de importancia de características	16
3.6. Algoritmos CNN	18
3.7. Modelos estado del arte	22
3.8. Medidas de evaluación de una red neuronal	25
4. Construcción de un modelo de predicción de la gravedad de un accidente de tráfico	29
4.1. Modelo preliminar	30
4.2. Modelo GTAAF	34
4.3. Preprocesamiento	35
4.3.1. Limpieza	37
4.3.2. Discretización	37

4.3.3.	Transformación (Sin/Cos)	38
4.3.4.	Filtrado de Áreas	39
4.3.5.	Normalización	40
4.3.6.	División Train-Val-Test	41
4.3.7.	Resampling	42
4.4.	Postprocesamiento	43
4.4.1.	Construcción de Matrices	44
4.4.2.	Feature Importance Algorithm	46
4.4.3.	Algoritmo Genético	46
4.4.4.	Construcción de Matrices	47
4.4.5.	Diseño del modelo	50
4.5.	Evaluación del modelo: Eficiencia y Robustez	50
5. Experimentos y resultados		51
5.1.	Resultados preliminares - Prototipo	51
5.1.1.	Construcción de matrices	60
5.1.2.	Conclusiones	65
5.2.	Resultados finales	65
5.2.1.	Dataset	65
5.2.2.	Descripción de datos	66
5.2.3.	Limpieza	70
5.2.4.	Filtrado de áreas	71
5.2.5.	Discretización	73
5.2.6.	Transformación (Sin/Cos)	78
5.2.7.	Resampling	78
5.2.8.	Normalización	79
5.2.9.	Categorización	79
5.2.10.	División Train-Val-Test	79
5.2.11.	Cálculo de pesos	80
5.2.12.	Feature Importance Algorithm	80
5.2.13.	Algoritmo Genético	80
5.2.14.	Pesos de categorías	82
5.2.15.	Construcción de matrices	82

0.1. ABSTRACT	5
5.2.16. Métricas de evaluación	82
5.3. Pruebas de estrés	103
6. Conclusiones	107
7. Publications	109
8. Anexos	111

0.1. Abstract

Luis: en principio OK, pero hay que repasarlo, me he liado en algún punto, la intuición más o menos creo que sería esta.

En esta tesis se presenta un nuevo modelo general que predice la necesidad de asistencia médica en accidentes de tráfico de cualquier región en base a la descripción del accidente. Conocer la gravedad del accidente una vez se produce es de vital importancia, ya que permite asignar recursos médicos de forma eficiente una vez se conocen las características del mismo, permitiendo evitar así consecuencias más graves en los afectados a corto y largo plazo al disponer de asistencia médica en un tiempo acorde a la gravedad del mismo. Con el objetivo de implementar un modelo general que pueda ser aplicado en distintas regiones independientemente de los datos disponibles, debido principalmente a las limitaciones socioeconómicas de la región, se presenta una metodología generalizable que permite adaptar cualquier conjunto de datos recibido a la entrada de este nuevo modelo clasificador.

La principal desventaja, para este caso de uso, de los modelos de clasificación, es que las características que requieren para sus predicciones deben ser las mismas a sus entradas respecto a los datos con los que se han entrenado. Por lo que si se pretendiese diseñar un modelo general aplicable a cualquier región con independencia de la información disponible en cada caso no sería posible, y se requeriría de un desarrollo específico para cada población en la que se quisiese aplicar, ya que cada una de estas, por la naturaleza socioeconómica de las poblaciones, puede no recoger ciertos datos que sí están presentes en otras. La metodología diseñada en esta tesis permite solventar este problema mediante un enfoque basado en la categorización de las características de los accidentes, donde en función de la naturaleza de cada dato disponible estos puedan ser asignados a categorías que engloban información a un nivel más alto, permitiendo así que el nuevo modelo propuesto sea independiente a los datos que estén disponibles en la región.

Para validar este enfoque se compararán los resultados de este nuevo modelo con otros 6 modelos del estado del arte que han sido aplicados históricamente

para la predicción de la necesidad de asistencia médica en accidentes a lo largo de 8 regiones distintas en distintos países, donde la información disponible en cada uno de estos conjuntos de datos es distinta por la naturaleza de socioeconómica de regiones. Además se utilizarán técnicas para evaluar la robustez del nuevo modelo mediante pruebas de estrés, donde para cada uno de los conjuntos de datos se iran eliminando características de mayor y menor importancia y se reevaluarán estos resultados commparándolos con a los modelos del estado del arte.

0.2. Acknowledgements

Luis: hay que hacerlo.

blablabla

Capítulo 1

Introduction

Luis: hay que hacerlo.

blablabla

1.1. Motivación

Aquí explicas la motivación de tu tesis (hay una necesidad, saber si un accidente de trafico va a ser grave, y te propones resolverlo mediante un modelo blablabla.

Luis: hay que hacerlo.

Capítulo 2

£Cómo medir la gravedad de un accidente?

Breve estado del arte para explicar como se mide la gravedad de un accidente en la literatura.

Luis: aquí me he liado, por el solapamiento de contenido respecto a la revisión de los modelos que existen de accidentes, tengo que separar estas dos secciones y readactarlas de cero.

La definición de la gravedad de un accidente de tráfico es un componente esencial para crear un modelo predictivo. La gravedad que implica un accidente de tráfico puede ser medido de muchas formas distintas. A lo largo de los años, muchos han sido las investigaciones que han estudiado a distintos niveles el impacto de estos, tanto a nivel económico, físico y social. Es por esto por lo que en función de los criterios que se utilicen para categorizarlos, el valor que puede aportar un modelo predictivo puede ser muy amplio y aportar valor en distintos ámbitos.

Existen estudios donde es común medir la gravedad de los accidentes en función del coste que supone la consecución de los accidentes para las autoridades, estos pueden ser categorizados en tres categorías () .

Otro ejemplo de evaluación en la severidad de los accidentes es en función de la cantidad total de daños a la propiedad, número de víctimas con lesiones y número de víctimas mortales que este ha producido [?], clasificando finalmente estos datos en cuatro clases distintas (leves, generales, graves y muy graves).

como a bajo nivel, donde se mide la gravedad en función d elas consecuencias físicas de la víctima. E

Nosotros nos centraremos exclusivamente en el impacto físico que supone a la víctima del accidente.

10 CAPÍTULO 2. ¿CÓMO MEDIR LA GRAVEDAD DE UN ACCIDENTE?

Aún disponiendo de los enfoques anteriores, en el estado del arte es más común observar la evaluación de la gravedad del accidente de tráfico en función de la gravedad de la lesión de la víctima. Es por este enfoque por el que nos regiremos a partir de ahora.

En función del criterio o las variables que se consideren, pueden tomarse distintos enfoques para implementar un modelo. Por lo que es importante definir de forma coherente los criterios mediante los que serán categorizados. Para implementar un modelo útil y que aporte valor a los organismos de asistencia médica es necesario revisar las distintas formas de interpretarlos en base al estado del arte.

Capítulo 3

£Cómo predecir la gravedad de un accidente de tráfico?

3.1. Revisión sobre modelos de predicción de gravedad de los accidentes de tráfico

La predicción de la severidad de los accidentes de tráfico ha sido un campo ampliamente estudiado a lo largo de los últimos años. En la historia reciente, la tendencia en la aparición de nuevos modelos de Aprendizaje Estadístico e Inteligencia Artificial ha ido aumentando en paralelo con los avances disruptivos en el campo de las Ciencias de la Computación. Tanto es así que la proposición de nuevos métodos en el último año ha sido exponencial. A lo largo del tiempo distintos enfoques han sido aplicados para solventar este problema en distintas poblaciones de todo el mundo, donde la severidad de los accidentes han sido consideradas de distintas formas, como se ha mencionado en el apartado anterior.

El componente de los datos a la hora de presentar estos modelos del estado del arte es crucial, ya que un buen entrenamiento y unos buenos resultados sobre un conjunto de datos de una región no implican que este modelo pueda ser aplicado a otras localizaciones, tanto por la falta de características respecto a otros conjuntos de datos y a las peculiaridades del conjunto de datos en cuestión.

Como principal punto de partida en la historia reciente podemos tomar [?], donde la severidad de los accidentes es considerada en tres clases (accidentes que únicamente daños a la propiedad, aquellos donde se han producido lesiones y accidentes con consecuencias fatales). Este modelo está entrenado en base a

12CAPÍTULO 3. ¿CÓMO PREDECIR LA GRAVEDAD DE UN ACCIDENTE DE TRÁFICO?

los accidentes producidos en la Autopista Norte-Sur, Malasia. Este conjunto de datos dispone de características que describen los accidentes, como las condiciones climáticas, la fecha y hora del accidente, tipo de colisión del vehículo, entre otras. El modelo propuesto está basado en Redes Neuronales Recurrentes (RNNs), donde las características de los datos son insertados a lo largo de dos capas LSTM, con el objetivo de capturar las correlaciones temporales entre las características de los accidentes.

Aplican random forest, ANN y árboles de decisión [?]. . .

Enfoques más recientes contemplan [?] [2018-2020 Sistema Nacional Chino de Investigación en Profundidad de Accidentes Automovilísticos (Chinese National Automobile Accident In-Depth Investigation System (NAIS))]. El conjunto de datos sobre el que trabaja esta investigación contiene 18 características que describen los accidentes, tales como . . . Esta investigación propone un modelo de Árboles de Decisión sobre el que se entrena el conjunto de datos en base a estas características.

Otro enfoque interesante es el aplicado en [?], donde se aplican en conjunto distintos árboles de decisión para crear un ensemble del tipo Random Forest, cuyos hiperparámetros son optimizados mediante Optimización Bayesiana (BO). Esta publicación busca predecir la severidad de los accidentes en US, entre los estados de Montgomery (Alabama) y el estado de Pensilvania. Este dataset dispone de muchísimas características, como tal tal y tal. La clasificación de los accidentes se divide en tres clases (leves, serios y fatales).

Como se puede intuir, existen distintos enfoques aplicados a muchas ciudades distintas. El principal inconveniente de los modelos citados anteriormente es que están muy acoplados a los datos disponibles para cada uno de estos datasets, entrenando modelos que requieren las características explícitas enumeradas en cada uno de ellos. Esto se traduce en una falta de generalización si se quisiese aplicar a otros conjuntos de datos pertenecientes a poblaciones donde estos datos puedan no estar disponibles, ya sea por la dificultad de su recogida o por las condiciones socioeconómicas de la región en concreto.

En general nuestra metodología es común en otros papers... Ya sea con XG-Boost optimizando hiperparámetros,

3.2. Métodos de optimización de hiperparámetros

A lo largo del tiempo distintos han sido los enfoques desarrollados para optimizar hiperparámetros de modelos predictivos.. [?]

3.3. Algoritmos Genéticos

Explicación de los diferentes algoritmos genéticos que hay, formulas, descripción, etc...

Luis: hay que terminarlo, y revisarlo, pero el enfoque propuesto es este.

Los algoritmos genéticos son métodos inspirados en la evolución biológica, que buscan optimizar soluciones a problemas matemáticos mediante la simulación la evolución de una población que genera descendencia a lo largo de generaciones. Estos algoritmos han sido ampliamente utilizados en casos como tal tal y tal. La principal fortaleza de estos algoritmos es que son métodos eficientes y seguros para llegar a soluciones aproximadas a la óptima ideal, reduciendo así el coste computacional en muchos casos exponencial, que supondría la búsqueda de la solución ideal (óptimo global) mediante métodos de combinación a lo largo de todo el conjunto total de posibles soluciones (espacio de búsqueda).

Existen muchos algoritmos genéticos conocidos, tanto para problemas de objetivo único como multi-objetivo (SPEA-II, NSGA-II), etc... [Aquí puedo bajar mucho.. Puedo usar mi TFG.](#)

El funcionamiento de un algoritmo genético consta de una serie de etapas que son repetidas a lo largo de las sucesivas generaciones: *inicialización, evaluación, selección, cruce, mutación y reemplazamiento*.

En la primera de ellas (*etapa de inicialización*), se crea una población original de N individuos aleatorios, donde cada uno de estos representa una posible solución al problema que se quiere optimizar. Estos individuos son evaluados mediante una función heurística, donde a cada uno se le asigna una puntuación de calidad en base a un criterio que mida el rendimiento que ofrece dicha solución al problema planteado (*etapa de evaluación*). Una vez se dispone de las puntuaciones de calidad, aquellos individuos que mejor se adapten al problema (mejor puntaje reciban) serán escogidos para dar lugar a descendencia (*etapa de selección*). La información que contienen los M mejores individuos es combinada entre sí (*etapa de cruce*), simulando el intercambio de información que supondría el intercambio genético en la naturaleza. Una vez se disponen de los nuevos individuos, estos pueden sufrir modificaciones aleatorias sobre su información resultante (*etapa de mutación*). Como en cualquier población biológica, la combinación de la misma información a lo largo de sucesivas generaciones provoca un estancamiento en la sociedad. La falta de diversidad en la población implica que no exista variabilidad en los individuos sucesores y por tanto que se tienda a explotar una zona del espacio de búsqueda provocando el riesgo de caer en un mínimo local del problema, es decir, una solución subóptima al problema respecto al mínimo global de la función buscado por estos algoritmos. Por este motivo es crucial introducir un componente aleatorio que pueda modificar la información de los individuos generados para tender a explorar este espacio de búsqueda. En este punto se evalúan los nuevos individuos y los M miembros

14 CAPÍTULO 3. ¿CÓMO PREDECIR LA GRAVEDAD DE UN ACCIDENTE DE TRÁFICO?

con peor puntuación de la población son eliminados, de esta forma la población en cada generación siempre constará de N individuos. En caso de que existan individuos iguales en la población, estos son eliminados, lo que provoca que se integren en la población el mismo número de los que se han descartado. Estas etapas son repetidas a lo largo de G generaciones hasta llegar a una condición de parada, tras la cual se seleccionará el individuo que mejor puntuación haya obtenido mediante la función heurística.

Hay que decir cómo los individuos se cruzan (punto aleatorio en el vector de la soluciones, porque luego al explicar los parámetros del algoritmo genético se explica el crossover index random) Existen distintas estrategias para de cruce, como intercambiar los siguiendo algún. (etapa de mutación) Una vez se dispone de la descendencia (etapa de reemplazamiento)...

En la figura ?? se puede observar cada una de las etapas que compone un algoritmo genético.

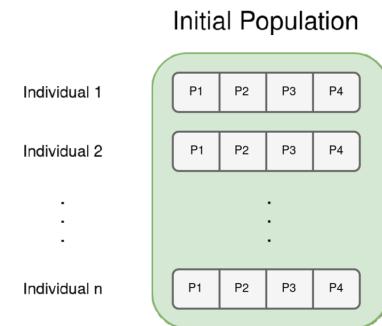


Figura 3.1: TFM.

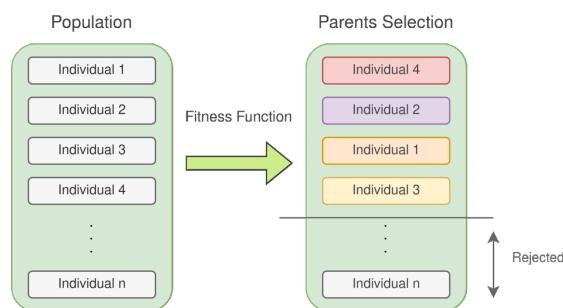


Figura 3.2: TFM.

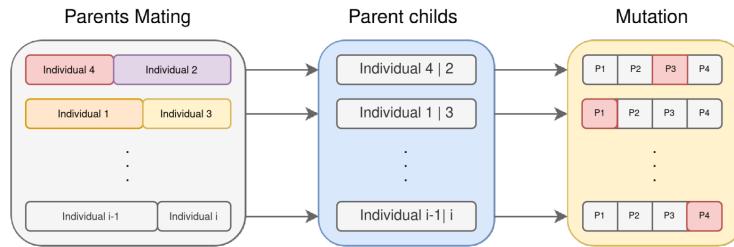


Figura 3.3: TFM.

3.4. Algoritmos de construcción de matrices

Luis: hay que revisar la redacción, pero el enfoque propuesto es este.

La tendencia de utilizar modelos convolucionales a lo largo de la historia reciente ha tenido un notable incremento debido a la eficiencia y rendimiento que estos demuestran a lo largo de múltiples contextos. Estos modelos trabajan sobre datos matriciales bajo los que pueden aplicar convoluciones y encontrar patrones más o menos complejos sobre los que aprenden a clasificar las muestras. La naturaleza de los datos tabulares impide aplicar este tipo de modelos a este tipo de datos, ya que no presentan una estructura matricial. Tanto es el incremento del uso de las convolucionales, que a lo largo de los últimos años se han diseñado técnicas para transformar datos tabulares a matriciales con el objetivo de poder aplicar estos modelos. Este problema no es un problema trivial, la forma en la que los datos son transformados a estas matrices debe generar una estructura que tenga sentido para los modelos convolucionales, maximizando la forma de representar la información para estos modelos.

En los últimos años se han diseñado distintas estrategias para solventar este problema, como OmicsMapNet [?], orientado a construir representaciones matriciales de las características de genes de pacientes que presentan cáncer. Para ello se consideran las descripciones bibliográficas de los genes para posicionar en localizaciones cercanas en la matriz aquellos que mayor semejanza presentan a través de Tree Map. Otro de los enfoques referentes en el estado del arte es Dee-Insight [?], que utiliza vectores de características de los datos originales para proyectarlos en un espacio bidimensional aplicando la visualización estadística T-SNE [?], donde aquellas características más cercanas bajo este espacio son seleccionadas para situarlas en posiciones cercanas de la matriz final construida. Más recientemente, se presentó REFINED (REpresentation of Features as Images with NEighborhood Dependencies) [?], una técnica que busca proyectar las características originales de los datos en un espacio bidimensional utilizando un escalador multidimensional bayesiano, que permite mantener la distribución de

16 CAPÍTULO 3. ¿CÓMO PREDECIR LA GRAVEDAD DE UN ACCIDENTE DE TRÁFICO?

las características en su espacio dimensional original, posteriormente se aplica un algoritmo de Escalada Simple (hill climbing) que optimiza la asignación de las características a los píxeles finales de la imagen.

En lo que respecta a un de los enfoques más recientes en el estado del arte, se presenta Image Generator for Tabular Data (IGTD) [?], una técnica que se aplica sobre descriptores de genes en pacientes que sufren cáncer. Esta técnica asigna las características más correlacionadas entre sí a posiciones cercanas dentro de la matriz, con el objetivo de aplicar una red neuronal convolucional que pueda operar sobre ella. Para lograr esto hace uso de técnicas de minimización de rankings entre pares de características y técnicas de minimización de rankings entre píxeles, donde en cada iteración se reasignan pares de características a la posición de aquellas otras que no han sido consideradas desde hace tiempo. De esta forma se logra una representación final de la matriz que resulta en agrupación de características similares cercanas.

3.5. Algoritmos de medición de importancia de características

Luis: hay que revisar la redacción, pero el enfoque propuesto es este.

En el campo de la Inteligencia Artificial y Análisis de Datos la medición de la importancia de las características dentro de un conjunto de datos toma un papel clave para distintos propósitos. Estas técnicas permiten conocer el peso que tienen cada una de las variables respecto al resto de ellas para un conjunto de datos, ya sea por la relación que presentan entre sí (fácilmente deducible por el ser humano o no), o por la importancia que han tenido a la hora de construir un modelo predictivo. Comúnmente, valores más altos de importancia representan una mayor relevancia de una característica para...

En el estado del arte, existen distintos métodos que tienen como propósito medir el peso de las características en un conjunto de datos, tanto para problemas de regresión como de clasificación. Uno de los enfoques más clásicos dentro del Aprendizaje estadístico para problemas de naturaleza regresiva es la técnica de Regresión Lineal, donde la magnitud de las variables está basada en el valor y la dirección de los coeficientes en base al resultado del aprendizaje del método predictivo. Tomando como referencia esta base, existen enfoques más complejos derivados de esta técnica, como son las Elastic Net Regression, que durante el proceso de aprendizaje del modelo de Regresión Lineal, se utilizan términos de penalización para reducir los coeficientes del predictor. Por otra parte, existen técnicas orientadas exclusivamente a problemas de clasificación que permiten medir la importancia de las características. Un método muy común en este campo es la Regresión Logística, un modelo estadístico que tiene como objetivo

3.5. ALGORITMOS DE MEDICIÓN DE IMPORTANCIA DE CARACTERÍSTICAS17

deducir la probabilidad de que ocurra un evento binario en función de uno o más predictores. Para calcular la importancia de las características, se utilizan las probabilidades logarítmicas para un cambio de una unidad en la variable predictiva. Los valores absolutos más grandes indican una relación más fuerte entre el predictor y la variable objetivo [?].

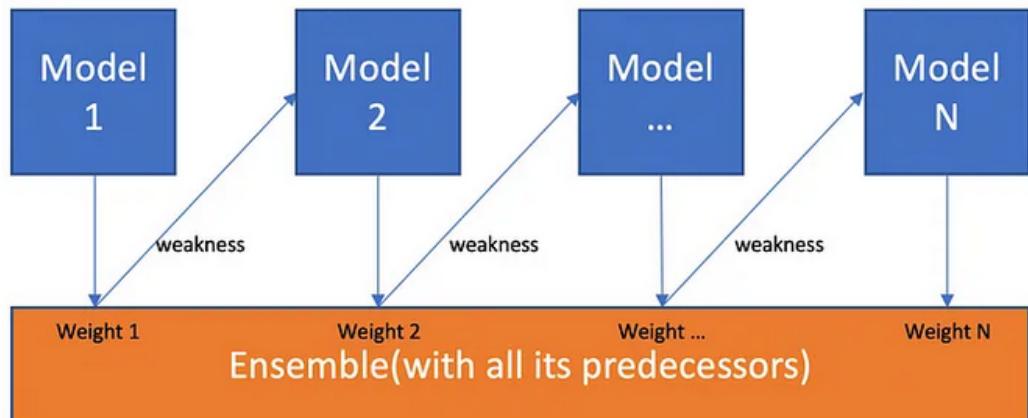
Permutation Feature Importance (<https://academic.oup.com/bioinformatics/article/26/10/1340/193348?login=false>)
Feature Selection with Importance: no he encontrado nada en 2 min.

Linear Regression Feature Importance: <https://machinelearningmastery.com/calculate-feature-importance-with-python/>

Por otra parte, existen otras técnicas que se alejan del aprendizaje estadístico y son métodos ampliamente utilizados para calcular la importancia de las variables, como los algoritmo de aprendizaje supervisado Random Forest. Estos métodos modelos cuyo funcionamiento se basa en la composición de varios modelos para dar una clasificación final (ensembles).

Dentro la filosofía de los modelos ensemble, los algoritmos Random Forest pertenecen a la tipología Bagging, que se basan en el concepto de crear múltiples modelos que son entrenados con distintas técnicas de reemplazo (bootstrap) sobre los datos, y la predicción final del conjunto es la combinación de la salida de cada uno de ellos de manera independiente por votación. Este enfoque de los algoritmos ensemble permite obtener modelos robustos que son menos sensibles al sobreajuste de un modelo general. Random Forest se compone en N árboles de decisión, donde cada uno de ellos es entrenado con un subconjunto de muestras y características de forma independiente para dar lugar a un modelo combinado donde la predicción de nuevas muestras se elige aquella clase más votada de entre todo el conjunto de árboles. La importancia de las características en este algoritmo se calcula mediante el peso que ha tenido cada característica a la hora de construir los árboles, en función del número de muestras que divide en cada nivel.

No obstante, existe otra técnica más potente dentro de los ensembles que tiene como base el funcionamiento de los Random Forest, el algoritmo tipo Boosting XGBoost [?]. Los algoritmo tipo Boosting que se caracterizan por crear modelos secuencialmente donde cada nuevo modelo se enfoca en corregir los errores cometidos por los modelos anteriores. XGBoost construye N árboles de decisión secuenciales, donde cada uno de estos se centra en corregir el error cometido por el error anterior, reduciendo así el sesgo y mejorando la precisión del modelo final al enfocarse en reducir el error de los modelos anteriores. Esta técnica es ampliamente utilizada para problemas de XXXX 3.4 [1].



One is weak, together is strong, learning from past is the best

Figura 3.4: Alguna imagen así hecha por mi.

[1] (<https://towardsdatascience.com/boosting-algorithms-explained-d38f56ef3f30>)

[2] <https://medium.com/geekculture/xgboost-versus-random-forest-898e42870f30>

<https://towardsdatascience.com/best-practice-to-calculate-and-interpret-model-feature-importance-14f0e11ee660>

3.6. Algoritmos CNN

Explicación des diferentes algoritmos CNN con los que luego te comparas, con sus formulas y explicacion

Luis: hay que terminar y revisar la redacción, en la estructura me he liado, seguramente haya que reestructurarlo con el contenido que hay escrito.

Las redes neuronales convolucionales (CNNs) son modelos de Inteligencia Artificial supervisados que principalmente están orientados al reconocimiento de patrones en imágenes. Estos modelos han sido ampliamente utilizados para distintos objetivos, como clasificación de imágenes, detección de elementos de interés dentro de estas o problemas de regresión. La naturaleza de su arquitectura ha permitido que además, estos modelos sean aplicados al campo de la Inteligencia Artificial Generativa, como la creación de imágenes mediante GAN, autoencoders o representación n-dimensional de imágenes a través de embeddings mediante el entrenamiento de redes siamesas, etc.

La principal característica que distingue a estos modelos respecto al resto

de redes neuronales, y los hace especialmente efectivos en problemas basados en imágenes, es que su arquitectura se basa en capas convolucionales. Estas capas están compuestas por filtros, que durante el proceso de entrenamiento aprenden operaciones que se aplican sobre los datos de entrada, permitiendo así generar y reconocer patrones que se encuentren presentes en ellos.

Dentro de las redes neuronales convolucionales existen diferentes tipos, cada uno con sus ventajas y desventajas en función del problema que se quiere resolver. No obstante, existen partes comunes a ellas que es necesario mencionar, las capas de las que normalmente constan estas redes son las siguientes:

1. **Capas Convolucionales:** Estas capas aplican convoluciones sobre las muestras de entrada. Las convoluciones no son más que multiplicaciones sobre posiciones de un vector que calculan la suma ponderada de todos los vecinos de la muestra de entrada para dar lugar a un único resultado en su salida, que será asignado a la salida de la capa convolucional en la misma posición sobre la que se ha aplicado la operación sobre la muestra de entrada. Los valores de ponderación (pesos) de esta suma son aprendidos por la red en su etapa de entrenamiento.
2. **Filtros:** Los filtros son pequeñas matrices de las que están compuestas las capas convolucionales y son utilizadas para realizar las operaciones. Cada uno de estos filtros tiene asociado una serie de pesos en cada posición de la matriz. Estos filtros, al ser aplicados, generan los denominados feature maps, que no son más que mapas de activación sobre los que se aplicarán la función de activación.
3. **Función de Activación:** activation functions in CNNs introduce nonlinearities, enabling the network to learn complex patterns and relationships within the data, typically, an activation function like ReLU (Rectified Linear Unit) is applied element-wise to the feature maps to introduce non-linearity.

$$\text{ReLU}(x) = \max(0, x)$$

4. **Capas Pooling:** estas capas aplican operaciones sobre los mapas de características con el objetivo de simplificar la información y reducir la dimensionalidad, que permite reducir la complejidad computacional de las redes durante su entrenamiento. Estas operaciones tienen una naturaleza de agrupación que son aplicadas en pequeñas zonas de los mapas de características para simplificar áreas y contemplar patrones relevantes en ellas. Estas operaciones pueden ser promediar un conjunto de características, mantener el mínimo de ellas o el máximo entre otras.
5. **Capas Densas:** Las capas densas son capas que interconectan completamente un conjunto de entrada de neuronas con las neuronas especificadas en esta capa. A diferencia de su aplicación en otro tipo de redes neuronales,

20CAPÍTULO 3. ¿CÓMO PREDECIR LA GRAVEDAD DE UN ACCIDENTE DE TRÁFICO?

en las redes convolucionales estas capas toman como entrada el conjunto de características extraídas de los procesos convolucionales para dar lugar a una clasificación final.

$$\begin{aligned}z_i &= \sum_{j=1}^n w_{ij} \cdot x_j + b_i \\y_i &= f(z_i)\end{aligned}$$

CNN-1D

Las redes neuronales convolucionales unidimensionales (CNN-1D) son redes cuya característica principal es que los filtros que aplican en cada una de sus convoluciones son de una dimensión [?].

$$(f * g)[n] = \sum_{m=0}^{M-1} f[m] \cdot g[n - m]$$

Estos modelos son ampliamente utilizados para problemas orientados a detecciones de patrones en señales, donde la naturaleza de los datos es principalmente secuencial. Por ejemplo, algunas de las aplicaciones donde las CNN-1D han demostrado ser efectivas han sido el monitoreo de electrocardiograma en tiempo real [?], detección de daños estructurales basada en vibraciones en infraestructuras civiles [?] o para clasificación de audios musicales [?]. Estas arquitecturas son una buena opción en problemas de este tipo, tanto en la calidad de resultados que presentan en este dominio, como en la rapidez de inferencia que demuestran, permitiendo ser aplicadas en tiempo real en dispositivos que requieren baja demanda de recursos computacionales, como teléfonos móviles. Sin embargo, la principal limitación de estas arquitecturas se

CNN-2D

CNN-3D?

El proceso de aprendizaje de las redes neuronales está dividido en varias fases. Las redes en su etapa de entrenamiento realizan predicciones sobre los datos de entrada, aplicando operaciones matemáticas sobre ellos utilizando los pesos configurados en la red (Forward Propagation).

Forward Pass:

$$Z = W \cdot X + b$$

$$A = \sigma(Z)$$

Posteriormente, en la capa clasificadora, los valores predichos son comparados con el valor real de las muestras que han sido introducidas en esta etapa a la red, de tal forma que el error que han producido sobre estas muestras durante esta fase es medible y calculado mediante una función de pérdida.

Loss Calculation:

$$\text{Loss} = \text{Calculate Loss}(\text{Actual}, \text{Predicted})$$

Gracias a la derivabilidad de las funciones que componen la red, en función de este error los pesos asociados a cada una de las capas de la red son optimizados con el objetivo de minimizar el error en la siguiente etapa de entrenamiento (Back Propagation). Gracias a la repetición de estos procesos la red toma conocimiento sobre los datos.

Backward Pass (Calculating Gradients):

$$\begin{aligned}\frac{\partial \text{Loss}}{\partial Z} &= \frac{\partial \text{Loss}}{\partial A} \times \frac{\partial A}{\partial Z} \\ \frac{\partial \text{Loss}}{\partial W} &= \frac{\partial \text{Loss}}{\partial Z} \cdot X^T \\ \frac{\partial \text{Loss}}{\partial b} &= \text{sum of } \frac{\partial \text{Loss}}{\partial Z} \text{ over the examples}\end{aligned}$$

Updating Weights and Biases:

$$\begin{aligned}W &= W - \alpha \cdot \frac{\partial \text{Loss}}{\partial W} \\ b &= b - \alpha \cdot \frac{\partial \text{Loss}}{\partial b}\end{aligned}$$

Para llevar a cabo este proceso, es necesario especificar la función de pérdida que medirá el error durante la fase de entrenamiento. La función de pérdida más común en problemas de clasificación binaria es la Binary Cross Entropy:

$$\text{Binary Cross Entropy} = \frac{-1}{N} \sum_{i=1}^N (y_i * \log p_i + (1 - y_i) * \log (1 - p_i)).$$

Donde:

- N es el número de muestras totales en el conjunto de datos.
- y_i es la etiqueta de la clase (0 ó 1) de la muestra actual.
- p_i es la probabilidad de que la muestra actual pertenezca a la clase 1.
- \log denotes the natural logarithm.

Esta ecuación penaliza en tiempo de entrenamiento la clasificación errónea de las muestras. El término $(y_i * \log p_i)$ penaliza la probabilidad p_i de pertenencia de la muestra y_i a la clase 0, siempre y cuando el valor verdadero de la muestra sea la clase 1. Por el contrario, el término $y_i * \log p_i + (1 - y_i) * \log (1 - p_i)$ penaliza la probabilidad p_i de la muestra y_i de pertenencia a la clase 1 siempre y cuando el valor real sea la clase 0. Valores de probabilidad altos p_i de la predicción de

22CAPÍTULO 3. ¿CÓMO PREDECIR LA GRAVEDAD DE UN ACCIDENTE DE TRÁFICO?

la red a las clases incorrectas, generan una acumulación del error. El símbolo negativo de la ecuación describe la minimización de esta función de pérdida. Esta función se utilizará para actualizar los pesos de toda la red mediante Back Propagation de cara a minimizar esta función para la siguiente época.

Una vez se define la función de pérdida de la red, la actualización de los pesos internos de las capas de la red y por tanto, el conocimiento de la misma sobre los datos, viene dado por el proceso de Back Propagation. Este proceso hace uso de la regla de la cadena, que permite calcular las derivadas parciales de la función de pérdida con respecto a los pesos de la red neuronal. Esto se aplica mediante el cálculo de las derivadas parciales de las capas superiores para calcular las derivadas de las capas inferiores. Comienza a partir de la capa de salida, retrocediendo a través de las capas ocultas, actualizando los pesos de la red conjuntamente en cada etapa.

3.7. Modelos estado del arte

Modelos del estado del arte contra los que nos comparamos.

Luis: esto lo he metido yo. Según tengo entendido tenemos que poner fórmulas de todos los modelos, incluyendo los 5 ó 6 contra los que nos comparamos no? Simplemente he copiado las fórmulas de ChatGPT

Estadísticos

Naive Bayes

$$P(C_k|x_1, x_2, \dots, x_n) = \frac{P(C_k) \cdot P(x_1, x_2, \dots, x_n|C_k)}{P(x_1, x_2, \dots, x_n)}$$

Where:

- $P(C_k|x_1, x_2, \dots, x_n)$ is the posterior probability of class C_k given the features.
- $P(C_k)$ is the prior probability of class C_k .
- $P(x_1, x_2, \dots, x_n|C_k)$ is the likelihood, the probability of observing the given features given class C_k .
- $P(x_1, x_2, \dots, x_n)$ is the probability of observing the features.

Logistic Regression

Logistic Regression models the probability $P(Y = 1|X)$ of a binary outcome Y given predictors $X = (X_1, X_2, \dots, X_n)$ using the logistic function:

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n)}}$$

Where:

- $P(Y = 1|X)$ is the probability of the positive class given the predictors.
- $\beta_0, \beta_1, \beta_2, \dots, \beta_n$ are the coefficients.
- X_1, X_2, \dots, X_n are the predictor variables.
- e is the base of the natural logarithm.

No supervisados

KNN The k-Nearest Neighbors (KNN) algorithm predicts the class of a data point by considering its k nearest neighbors in the feature space. The predicted class \hat{y} for a new data point x is determined by a majority vote among its k nearest neighbors:

$$\hat{y} = \arg \max_{y_i} \sum_{i=1}^k I(y_i = y)$$

Where:

- \hat{y} is the predicted class for the new data point.
- x is the new data point to be classified.
- k is the number of nearest neighbors to consider.
- y_i represents the classes of the k nearest neighbors.
- $I(y_i = y)$ is an indicator function returning 1 if y_i is equal to the predicted class y , and 0 otherwise.

Supervisados

NNs

MLP

A Multi-Layer Perceptron (MLP) is a type of feedforward neural network composed of multiple layers, including an input layer, hidden layers, and an output layer. Let's denote:

- $x = (x_1, x_2, \dots, x_n)$ as the input vector.
- $h^{(i)} = (h_1^{(i)}, h_2^{(i)}, \dots, h_{m_i}^{(i)})$ as the activations of the i -th hidden layer.
- $W^{(i)}$ as the weight matrix connecting the i -th and $(i + 1)$ -th layers.
- $b^{(i)}$ as the bias vector added to the i -th layer.

24 CAPÍTULO 3. ¿CÓMO PREDECIR LA GRAVEDAD DE UN ACCIDENTE DE TRÁFICO?

- f as the activation function.
- $y = (y_1, y_2, \dots, y_k)$ as the output vector.

The computation in an MLP can be represented as follows:

$$h^{(i)} = f(W^{(i)}h^{(i-1)} + b^{(i)})$$

where $h^{(i-1)}$ is the activation from the previous layer.

The final output of the MLP is computed as:

$$y = f(W^{(n)}h^{(n-1)} + b^{(n)})$$

Here, n represents the number of hidden layers in the MLP.

Random Forest

1. For $t = 1$ to T (the number of trees in the forest):
 - a) Draw a bootstrap sample D_t by sampling with replacement from D .
 - b) Grow a decision tree T_t using D_t :
 - For each node of the tree:
 - 1) Randomly select m features from M .
 - 2) Split the node using the best feature among the m selected features.
2. The final prediction for a new sample is obtained by aggregating predictions from all trees (for classification, typically a majority vote).

SVC

The Support Vector Classifier (SVC) aims to find the optimal hyperplane that best separates data into different classes. Given a training dataset $\{(x_i, y_i)\}$ where $x_i \in R^n$ represents input features and $y_i \in \{-1, 1\}$ represents class labels, the objective of the SVC is to find:

$$\min_{w,b,\xi} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^m \xi_i$$

subject to:

$$y_i(w \cdot x_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0$$

Where:

- w is the weight vector of the hyperplane.
- b is the bias term.
- ξ_i are slack variables allowing for misclassification or data points within the margin.
- C is a regularization parameter controlling the trade-off between maximizing the margin and minimizing the classification error.

The decision function for classifying a new sample x_{new} is given by:

$$\text{Predict}(x_{\text{new}}) = \text{sign}(w \cdot x_{\text{new}} + b)$$

LSTM para [?] LSTM Equations:

$$\begin{aligned} i_t &= \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \\ f_t &= \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \\ g_t &= \tanh(W_{xg}x_t + W_{hg}h_{t-1} + b_g) \\ o_t &= \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \\ c_t &= f_t \odot c_{t-1} + i_t \odot g_t \\ h_t &= o_t \odot \tanh(c_t) \end{aligned}$$

3.8. Medidas de evaluación de una red neuronal

Aquí explicas el F1-score y demás, con sus fórmulas y con detalle.

Luis: esto yo creo que estaría OK.

En este apartado se presentan los indicadores de calidad utilizados para medir el rendimiento y generalización de los modelos expuestos en esta tesis. Uno de los componentes fundamentales en el desarrollo de modelos de Inteligencia Artificial es conocer la capacidad y calidad de los modelos ante la predicción de nuevas muestras que nunca ha visto durante su etapa de entrenamiento, tanto como para poder compararlos como para conocer en profundidad cómo se comportan los modelos ante nuevas situaciones.

Para evaluar los modelos, es común aplicar una fase de validación o de test, donde se utilizan los modelos para realizar predicciones contra muestras de las que se conoce su variable verdadera. De esta forma es posible comparar la calidad de los modelos respecto a muestras que nunca antes han visto y aplicar fórmulas y métricas que nos dan idea del rendimiento de los modelos. Para esto,

26 CAPÍTULO 3. ¿CÓMO PREDECIR LA GRAVEDAD DE UN ACCIDENTE DE TRÁFICO?

es necesario introducir dos conceptos básicos TP, FP, FN y TN, estos datos son calculados para cada una de las clases que puede predecir el modelo.

1. **True Positives:** Los True Positives (TP) representan el número de muestras que han sido correctamente clasificadas por el modelo como positivas. Es decir, el modelo clasifica correctamente la muestra como la clase a la que pertenece.
2. **True Negatives:** Los True Negatives (TN) representan el número de muestras que han sido correctamente clasificadas por el modelo como negativas. Es decir, el modelo
3. **False Positives:** Los False Positives (FP) representan el número de muestras que han sido incorrectamente clasificadas por el modelo como positivas. Es decir, el modelo ha clasificado una muestra que no pertenecía a esa clase como positiva.
4. **False Negatives:** Los False Negatives (FN) representan el número de muestras que han sido incorrectamente clasificadas por el modelo como negativas. Es decir, el modelo ha clasificado una muestra positiva como negativa.

En función del problema que nos encontremos, es posible que sea preferible un modelo que tienda a más a un tipo de clasificación que a otra debido a la criticidad del problema a costa de reducir el número de ocurrencias de otro indicador. No tiene la misma importancia matar a cinco personas por FN que darle un medicamento a 5 que realmente no la necesitan (FP) pero que no sufrirían de consecuencias graves al tomarlo. Es por ello que utilizando estos conceptos básicos es posible crear indicadores de calidad que ofrezcan más información para cada una de las clases predichas. En el estado del arte, se utilizan dos métricas comunes que pueden ser utilizadas para la composición de indicadores aún más complejos, estas métricas son calculadas para cada una de las posibles clases dentro del conjunto de datos.

La primera de ellas es la precisión (Precision), que mide el porcentaje de muestras clasificadas correctamente de una clase, respecto al total de muestras que existen de dicha clase en el conjunto de datos.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

Por otra parte, el Recall representa la proporción de elementos de una clase que el modelo identifica correctamente como esa clase.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

Existe otra métrica que combina los dos indicadores anteriores, considerando la precisión que tiene el modelo a la hora de predecir muestras como una clase y cuántos de los casos positivos fueron captados por el modelo (recuerdo), de tal forma que para cada una de las clases se pueda obtener una evaluación individual, siendo más sencillo en análisis sobre esto.

$$\text{F1 score} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

Capítulo 4

Construcción de un modelo de predicción de la gravedad de un accidente de tráfico

Luis: esta intro y todo el punto del modelo preliminar no me termina de convencer, lo hablamos cuando podais.

En esta tesis se expone una metodología y un modelo general de predicción de gravedad de accidentes de tráfico aplicable a cualquier región. Para llegar a este fin, inicialmente se realizó una investigación y se implementó una primera metodología a modo de prototipo sobre la que aplicar modificaciones hasta llegar al objetivo final, un procedimiento que no fuese sensible a la disponibilidad de datos y fuera independiente de la región sobre la que se aplicase, es decir, un modelo de predicción de la necesidad de asistencia en los accidentes de tráfico general. Por este motivo, este apartado se divide en dos subsecciones. La primera de ella describe la intuición sobre el primer prototipo, describiendo brevemente las fases que lo componen, los objetivos finales de este, incidiendo en las partes que han evolucionado respecto al modelo final. La siguiente sección de este apartado expone la metodología final tras la evolución del prototipo como referencia, justificando las decisiones tomadas en cada caso.

4.1. Modelo preliminar

Explicas el enfoque de paper 1, brevemente, a nivel metodológico.
Luis: a lo mejor estaría bien meter todo lo del congreso

Este primer prototipo se presentó en el artículo [?], se implementó con el objetivo de predecir la gravedad de los accidentes de tráfico en la ciudad de Madrid, concretamente en tres clases distintas disponibles en este conjunto de datos: (1) Leves, (2) Severos y (3) Fatales.

Para llegar al entrenamiento de un modelo predictivo, se diseñó una metodología prototipo que estaba compuesta por cinco fases secuenciales 4.1. Y tenía como objetivo realizar transformaciones y operaciones sobre datos, inicialmente tabulares, para transformarlos en datos matriciales. De esta forma era posible experimentar con dos modelos convolucionales, el primero de ellos unidimensional y el segundo bidimensional, CNN-1D y CNN-2D respectivamente.

Para ello, era necesario definir una categorización de características, las cuales serían utilizadas como apoyo para la construcción de estas matrices que junto a la importancia de cada variable dentro del conjunto de datos, era posible asignarlas a coordenadas dentro de una matriz.

Finalmente la metodología y los modelos convolucionales propuestos eran comparados con otros tres modelos del estado del arte (GNB, SVC y KNN) para evaluar sus rendimientos respecto a la ciudad de Madrid.

A continuación se enumeran las etapas que definen el flujo de la metodología prototipo.

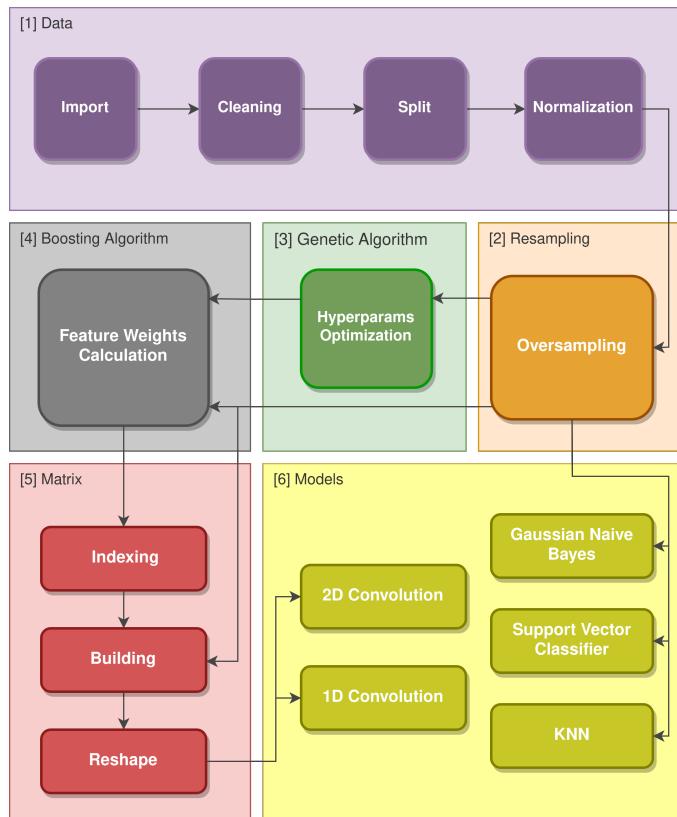


Figura 4.1: Flow chart of the proposed model with its different phases.

[1] Datos

La primera fase de esta metodología prototipo está orientada al tratamiento de los datos. Estos datos originales del dataset eran datos en bruto, donde se podían encontrar errores en los valores, valores atípicos y variables con valores cualitativos que había que discretizar. En esta etapa a los datos se les aplica un proceso de limpieza, discretización y normalización. Con el objetivo de construir un conjunto de datos interpretable por los modelos de Inteligencia Artificial.

[2] Resampling

La segunda fase tenía como objetivo trabajar sobre el desbalanceo de los datos presente el dataset. Debido a la naturaleza de los accidentes de tráfico, gran parte de ellos eran de tipo leve, mientras que el resto de tipos de accidentes (severos y fatales), presentaban una proporción mucho menor respecto a los del primer tipo. Para evitar un sesgo en los modelos, y que tendiesen a predecir cualquier nueva muestra como la clase mayoritaria, se estudiaron distintas técnicas de balanceo de datos, utilizando finalmente la técnica Borderline SMOTE-II para balancear las clases minoritarias, aplicando generación de datos sintéticos hasta

32 CAPÍTULO 4. CONSTRUCCIÓN DE UN MODELO DE PREDICCIÓN DE LA GRAVEDAD DE UN

igualar las clases hasta la mayoritaria.

Luis: Capaz que juntaba los algoritmos genéticos y el Boosting , es necesario saber del Boosting para justificar el uso de algoritmos genéticos

[3-4] Algoritmos Genéticos / Boosting

Para transformar los datos tabulares a datos matriciales interpretables por el modelo convolucional de dos dimensiones, se requería de algún tipo de estrategia para la asignación de cada una de las variables del dataset a coordenadas dentro de una matriz bidimensional, con el objetivo de aplicar los modelos convolucionales propuestos en este prototipo. Para llevar a cabo esto, se tomó una estrategia que requería de conocer la importancia de cada variable dentro del conjunto de datos. Como método para hallar el peso de cada característica dentro del dataset se utilizó un algoritmo tipo Boosting. Los algoritmos tipo boosting son un algoritmos clasificadores que ofrecen la importancia numérica de cada variable en función del peso que han tenido durante su entrenamiento. Estos algoritmos necesitan una configuración de hiperparámetros que se realizó mediante la evolución de un algoritmo genético.

[5] Matrices

Una vez se disponían de los pesos de las características gracias al cálculo del algoritmo tipo Boosting, se categorizaron las variables en distintas características. Para tener una referencia de dos dimensiones sobre las que comenzar a indexar las variables. En primer lugar se calculó el peso total de las categorías, que era la suma de cada una de las características que contenía, como resultado de esto, cada categoría se indexaba a una fila de la matriz, donde aquella que más peso presentaba era asignada a la fila central, la segunda en la posición inmediatamente superior, la siguiente en la inferior y así sucesivamente. Las características que las componían se asociaban a las columnas dentro de su categoría de la misma manera, la de mayor peso en la posición central, la siguiente en su posición inmediatamente a la izquierda, la siguiente a la derecha etc. Como resultado de este proceso, cada registro perteneciente al dataset original era transformado en una matriz de 5x5.

[6] Modelos CNN propuestos

Las arquitecturas prototipo propuestas constaban de cuatro capas convolucionales con tamaños de kernels de 1×3 para la CNN 1D y 3×3 para la CNN-2D respectivamente. Estos núcleos se proyectaban en 256 y 512 canales para formar el filtro convolucional asociado con cada capa. Posteriormente se aplicaba un proceso de normalización de batch a la salida de cada uno de los mapas de características.

El padding de los kernels estableció en 1 para ambos tipos de redes, de modo que las convoluciones se apliquen agregando ceros a los límites de las matrices, y las sa 1 para la CNN 1D y 1, 1 para la CNN 2D. Por lo tanto, el desplazamiento de los núcleos se realiza píxel por píxel en ambas redes convolucionales.

En la salida de cada capa convolucional, se aplica la función de activación

Unidad Lineal Rectificada (ReLU).

La salida de la última capa de convolución transforma la matriz de mapa de características generada de tamaño 5×5 en una capa que aplanará la matriz a un vector unidimensional de 1×25 . A continuación, se aplica una capa densa que conecta cada uno de los 25 nodos de la capa aplanada con los 128 nodos de la capa densa, que genera los logits antes de aplicar la última función de activación Softmax que devuelve la clase predicha. En la figura 4.2 se observa el diseño de la arquitectura de la red propuesta.

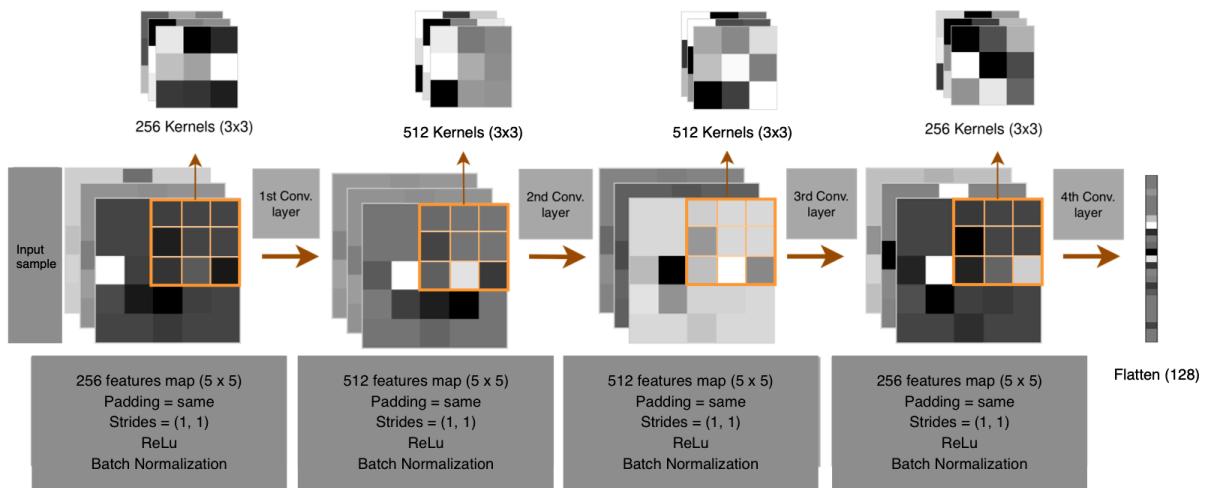


Figura 4.2: Architecture of the 2D-Convolutional neural network.

[7] Comparaciones

Por último se compararon los dos modelos propuestos (CNN-1D y CNN-2D) contra 3 modelos del estado del arte.

Conclusiones.. Analizando los resultados de la metodología y modelo prototípico se observa que la decisión de dividir los accidentes de tráfico en tres clases, ofrece una utilidad bastante pobre. Siendo dos de estas clases minoritarias y aplicando generaciónn de datos sintéticos contra tantas muestras de accidentes leves podría carecer de sentido a nivel práctico. El siguiente paso para resolver esto fue la categorización de estos accidentes en dos clases, además de implementar un proceso de filtrado de áreas que uscaba rebajar el número de accidentes tipo leve de forma natural en el conjunto de datos. Otra de las propuestas de mejora y que tenía todo el sentido del mundo era incrementar el número de características. A medida que se dispone de mayor información en el conjunto de datos

4.2. Modelo GTAAF

Después de analizar los resultados ofrecidos del primer prototipo, encontrar debilidades en los enfoques y decisiones tomadas, se propone una nueva metodología basada en la anterior. Esta nueva metodología denominada GTAAF (General Model for Traffic Accident Assistance Forecasting), busca incrementar el rendimiento de su predecesora, con el principal objetivo de diseñar un procedimiento de predicción de asistencia de accidentes generalizable a cualquier región. El principal problema de los conjuntos de datos de accidentes es que dependiendo de la región y/o gobierno que los ofrezca, estos disponen de información muy dispar entre ellos, debido principalmente al coste que supone obtener ciertos datos y/o la naturaleza social de la población. Ss por esto que en caso de querer aplicar un modelo de predicción de necesidad de asistencia en accidentes, requiere un trabajo de analizar qué categorías están disponibles y cuáles pueden ser influyentes en la necesidad de asistencia de los accidentes. Para solventar esto y conseguir una generalización independiente de los datos disponibles, la metodología GTAAF propuesta se basa en categorización de las características disponibles individuales dependientes de cada conjunto de datos, donde en función de la naturaleza a la que pertenezca cada dato disponible estos puedan ser asignados a una de las categorías propuestas en esta metodología, cuyas propiedades son de fácil adquisición. Esto sorteja las peculiaridades individuales de la disponibilidad de datos de cualquier región. Para evaluar esto, GTAAF es comparado con otros seis modelos del estado del arte a lo largo de 8 regiones distintas en las mismas condiciones.

En esta sección se explicará con detalle cada una de las etapas por las que pasan los datos, la justificación de las decisiones tomadas para la construcción de esta metodología y las principales diferencias entre la versión preliminar y la versión final.

En primer lugar, las fases de la nueva metodología son asignadas a tres etapas claramente diferenciadas: (1) la fase de Preprocesamiento, donde se contemplan procesos de limpieza de datos, transformación y balanceo de datos, (2) la fase de Postprocesado donde se aplican técnicas de transformación para representar los datos de accidentes en formato tabular a formato matricial, y (3) la fase de entrenamiento, donde se entrenará un modelo neuronal convolucional en base a esta representación para predecir la necesidad de asistencia en los accidentes. En la figura 4.3 se muestran, en modo de diagrama, cada una de las fases que componen la metodología GTAAF.

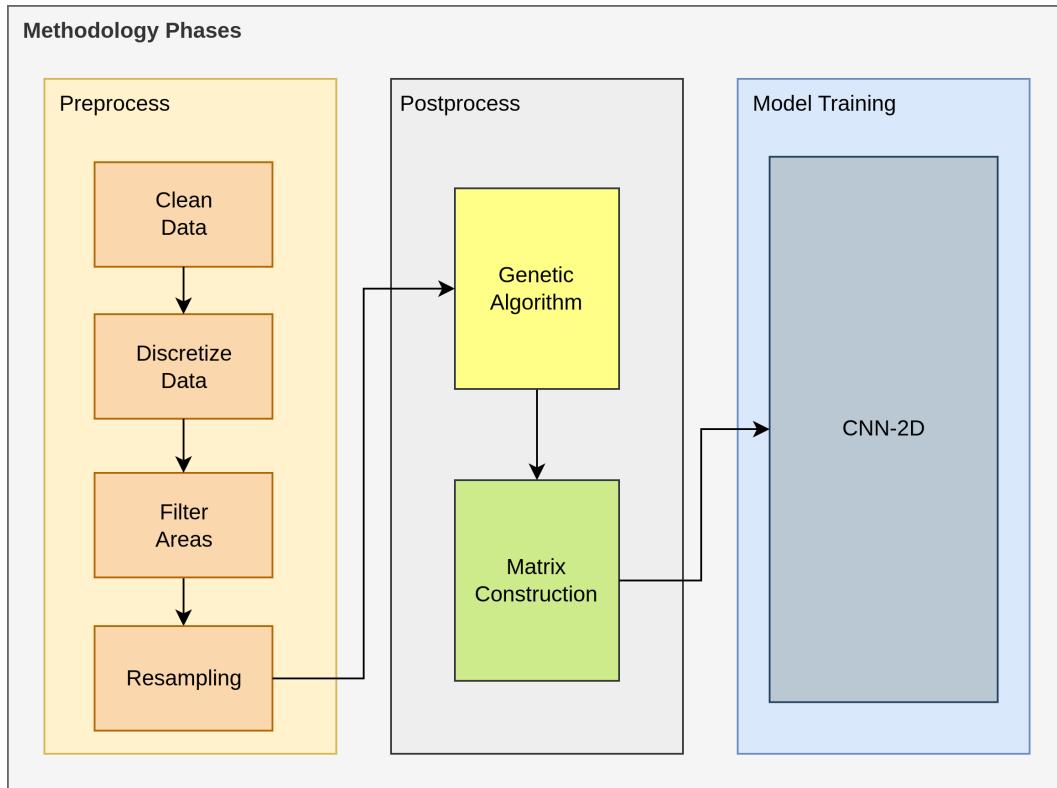


Figura 4.3: Methodology flowchart: data preprocessing, postprocessing and model training.

En segundo lugar, como segunda consideración importante respecto a la metodología prototipo, el concepto de gravedad de los accidentes es reasignado a dos clases (accidentes sin necesidad de asistencia y accidentes con necesidad). Esto se debe a que el valor que aporta distinguir entre la clases Severo y Fatal no es lo suficientemente enriquecedor como para arriesgarse a distinguir entre tres clases, ya que un modelo de clasificación a medida que incrementa el número de clases, tiene más posibilidades de realizar predicciones erróneas, sobre todo si son clases minoritarias como son los accidentes severos y fatales. Por este motivo se distinguen entre accidentes con necesidad de asistencia y aquellos que no.

4.3. Preprocesamiento

Esta sección explica las diferentes etapas que componen la fase de preprocesamiento de la metodología GTAAF propuesta. Esta es la primera de las etapas y es donde a los datos se les aplican transformaciones para dar lugar a

36 CAPÍTULO 4. CONSTRUCCIÓN DE UN MODELO DE PREDICCIÓN DE LA GRAVEDAD DE UN

un conjunto de datos refinado interpretable para cualquier modelo que trabaje con datos tabulares. Esta etapa está compuesta por cuatro fases: (1) proceso de limpieza de datos, donde se identifican, corrigen y se tratan las inconsistencias sobre los datos, (2) la discretización, donde se convierten las variables continuas en variables discretas y se codifican los valores cualitativos de las características, (3) el filtrado de áreas, donde se reduce el desbalanceo de los datos escogiendo subregiones de la ciudad donde se localicen ambos tipos de accidentes, y (4) el remuestreo, donde se generan muestras sintéticas de la clase minoritaria para disponer de un dataset balanceado. En la Figura 4.4 se muestra el flujo sobre el que pasan los datos para cada una de las diferentes fases que componen la etapa de Preprocesamiento. Esta figura será referenciada en las siguientes subsecciones en la explicación de las fases de Preprocesamiento.

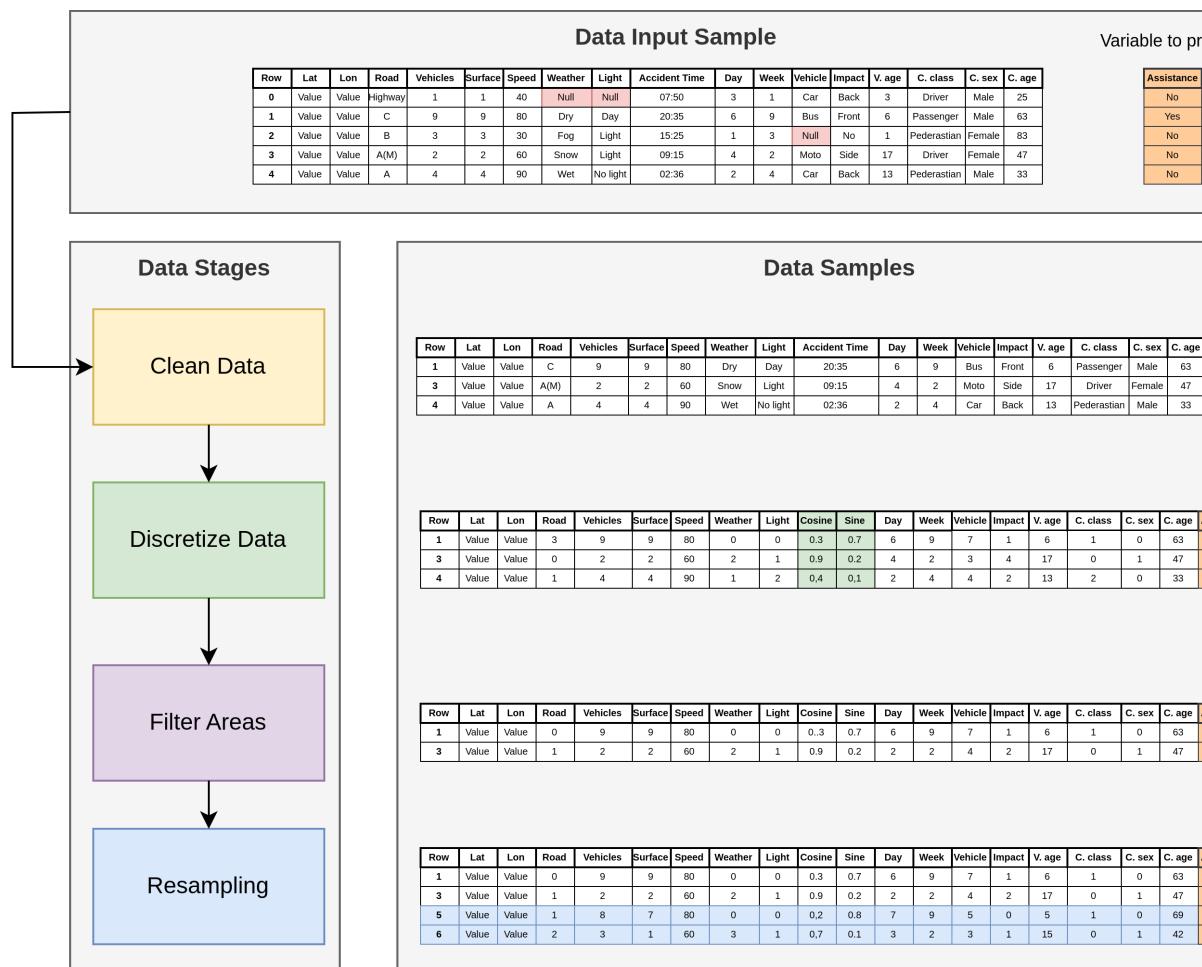


Figura 4.4: Pre-processing data flow.

4.3.1. Limpieza

Luis: es que aquí no hay mucho más que decir...

La limpieza de datos es un proceso esencial en cualquier proyecto de Análisis de datos o Inteligencia Artificial. Esta fase tiene como objetivo tratar los datos de tal forma que el dataset procesado no disponga de valores ausentes, atípicos, presenten inconsistencias o errores. Este proceso asegura que los datos estén listos para análisis y modelado. Un conjunto de datos limpio y refinado es la base para comenzar a trabajar con modelos predictivos, ya que de otra forma los datos no son fiables [?].

La primera fase de la metodología contempla un proceso de limpieza, en el que aquellos registros de los datos que presenten valores nulos o aquellos que se muestren atípicos sobre las variables escogidas serán eliminados del dataset. Esto provoca que haya un porcentaje de los datos que son eliminados. Estos casos se encuentran representados de color rojo en la primera etapa de la figura 4.4, donde los registros de accidentes con identificador 0 y 2 son eliminados del conjunto de datos al presentar valores nulos en alguna de sus características.

4.3.2. Discretización

Luis: con pinzas, no me termina de convencer, realmente no digo nada y no me gusta meter ejemplos de variables del dataset cuando aún no hemos presentado el dataset.

Los modelos predictivos trabajan con datos numéricos, que son los que son capaces de interpretar, sobre los que realizan operaciones matemáticas adquirir conocimiento sobre estos y poder realizar inferencias sobre muestras nunca antes vistas. Es por esto que las características ofrecidas en los conjuntos de datos deben ser transformadas a estos valores sin perder el valor que representa la información. A la hora de describir un accidente, gran parte de la información que se obtiene tiene una naturaleza cualitativa, es decir, son valores que no representan valores numéricos sino descriptivos. Esto se puede ejemplificar de forma clara con la característica 'First Point of Impact', en la que los valores que toma este campo en los conjuntos de datos originales representan una descripción del punto de impacto, como por ejemplo, 'Colisión frontal', 'Colisión lateral', etc. Por este motivo es necesario aplicar un proceso de discretización, este proceso busca transformar estos valores descriptivos a valores numéricos de tal forma que los datos puedan ser interpretados por los modelos, buscando representar de forma jerárquica la importancia de cada uno de los posibles valores descriptivos, teniendo como objetivo que la información descriptiva contenida sea coherente con su representación numérica.

En esta tesis se ha seguido un procedimiento de discretización incremental,

donde a cada posible valor del conjunto de datos se le ha asignado un valor numérico en función de la importancia que se le ha asignado.

4.3.3. Transformación (Sin/Cos)

Luis: esto yo creo que estaría, a la espera de poner más bonito el dibujo.

Como se ha comentado en la sección anterior, los modelos de Inteligencia Artificial y Aprendizaje estadístico interpretan los datos en forma numérica. El valor numérico que se le asigna a cada campo es crítico, ya que será así como el modelo interprete el orden de los valores cualitativos que los humanos somos capaces de comprender. La representación del formato de las horas y minutos del día, por su naturaleza, no es una excepción. El concepto de la hora del día tiene un componente cíclico que es necesario representar para que el modelo comprenda que las once y cincuentainueve de la noche es una hora muy próximas a las doce de la noche. Esto es algo a lo que los seres humanos estamos acostumbrados, pero debe ser indicado de forma coherente para los modelos de IA que interpretarían que estas dos horas muy parejas son valores totalmente opuestos en los posibles valores que puede contener la característica hora con el formato 24 horas que conocemos (23:59, 00:00). Con el objetivo de representar de forma consistente la información de la hora del accidente, es necesario aplicar una transformación que interprete las horas y minutos en formato 24h a un formato cíclico, y para ello se transformará este campo inicialmente de una dimensión, a dos dimensiones sinusoidales. Para realizar este proceso en primer lugar se transforma la hora y el minuto en el que se ha producido cada accidente a segundos. Posteriormente se aplican las siguientes fórmulas sobre los segundos para representar la hora del accidente en dos componentes, el senoidal y el cosenoidal:

$$\begin{aligned} \sin((2 \cdot \pi \cdot DaySeconds) / SecondsInDay) \\ \cos((2 \cdot \pi \cdot DaySeconds) / SecondsInDay) \end{aligned}$$

(Dibujito explicativo de senos y cosenos)[puedo poner las 23:59 de la noche representada en seno y coseno, las 00:00 y las 15:00 para que se vean las diferencias]. En la figura 4.5 se muestra un ejemplo de la naturaleza cíclica de la representación de la variable Hora en forma de seno (eje de ordenadas) y coseno (eje de coordenadas), donde se observa que la hora 00:00 en el espacio bidimensional se encuentra más cercana a la hora 07:58:00 respecto a cualquier otra posible representación unidimensional.

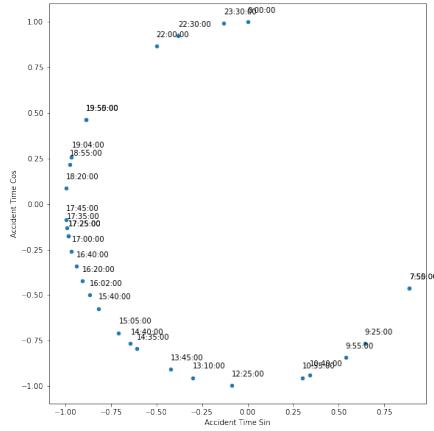


Figura 4.5: Algo así, no es lo definitivo.

En la Figura 4.4 de referencia, se muestra el ejemplo donde el accidente sin necesidad de asistencia con identificador 4 se elimina del conjunto de datos porque no convive con otro accidente de tipo asistencia dentro de su mismo área.

4.3.4. Fitrado de Áreas

Luis: esto yo creo que estaría.

Uno de los retos más comunes en el campo de la Inteligencia Artificial es disponer de un conjunto de datos no balanceado. Este problema implica tener una desproporción del número de muestras en base a la variable a predecir. Esta casuística afecta negativamente al entrenamiento de los modelos de Inteligencia Artificial, ya que estos en su etapa de entrenamiento adquieren el conocimiento prediciendo sobre estas muestras y son penalizados cuando sus predicciones durante esta fase son erróneas. Si la distribución de datos de entrenamiento dispone de muchas más muestras de una clase que de otra, el modelo tenderá a aprender durante su entrenamiento a predecir siempre aquella clase mayoritaria, ya que se le ha penalizado en menos ocasiones durante esta fase, obteniendo así un modelo sesgado que está condicionado por naturaleza a predecir sobre la clase más común.

En lo que respecta la naturaleza de la distribución de datos de accidentes de tráfico, siempre existirán muchos más accidentes que no han necesitado asistencia respecto a los que sí. Por lo que durante esta fase de la metodología se

40 CAPÍTULO 4. CONSTRUCCIÓN DE UN MODELO DE PREDICCIÓN DE LA GRAVEDAD DE UN

busca paliar este efecto tratando de reducir la diferencia entre el número de registros de la clase mayoritaria (sin necesidad de asistencia) y la clase minoritaria (necesidad de asistencia).

Para solventar esto se aplica un filtrado basado en áreas, que buscará balancear los datos escogiendo áreas estratégicas donde coexistan accidentes con ambos tipos de consecuencias. Para cada población se establece una ventana de dimensiones (X,Y) que recorrerá secuencialmente el área total que engloba cada una de las regiones escogidas en esta tesis. Esta ventana buscará si en ese área coexisten accidentes de tipo No-Asistencia y Asistencia, de tal forma que si esto se cumple, dicha subárea se mantendrá en el dataset, y en caso contrario se eliminará. Esto consigue un balanceo de los datos que minimiza el número de accidentes de tipo No-Asistencia en el dataset que no sean estrictamente necesarios. En la figura ?? se muestra un ejemplo del criterio seguido para aplicar este filtrado, donde se seleccionan únicamente aquellas regiones donde coexisten accidentes sin necesidad de asistencia (verde) y con necesidad de asistencia (rojo).

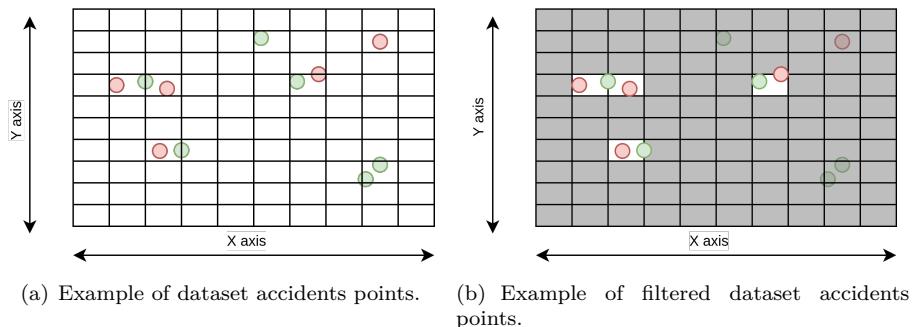


Figura 4.6: Ejemplo de filtrado de áreas. Los puntos verdes representan accidentes con lesiones leves, mientras que los puntos rojos representan accidentes que requirieron asistencia.

En la Figura 4.4 de referencia, se muestra el ejemplo donde el accidente sin necesidad de asistencia con identificador 4 se elimina del conjunto de datos porque no convive con otro accidente de tipo asistencia dentro de su mismo área.

4.3.5. Normalización

Luis: esto yo creo que estaría, no se me ocurre qué más poner, no obstante no me convence poner ejemplos de los datasets cuando aún no los hemos presentado.

En cualquier modelo de Inteligencia Artificial es imprescindible normalizar los datos. Los modelos de Inteligencia Artificial trabajan con valores numéricos realizando operaciones sobre ellos. En los conjuntos de datos suelen coexistir variables cuyos valores se encuentran representados en distintas escalas, es decir, que los valores que pueden tomar ciertas características suelen presentar un rango de valores mucho más amplio que otras dentro del mismo conjunto de datos, haciendo que las características sean incomparables entre sí debido a su magnitud. Un ejemplo de esto puede observarse en la características 'Semana en Año' y 'Sexo', la primera de estas variables puede contener un amplio conjunto de posibles valores (desde el 0 hasta el 51), en función de la semana en la que se ha producido el accidente, mientras que la segunda variable únicamente puede tomar dos valores (0 ó 1). Esta variabilidad numérica en los posibles valores de los datos provoca que las operaciones matemáticas que aplican los modelos durante su fase de entrenamiento sean desproporcionadas en las características con rango de valores más altos, produciendo una desproporción en estas operaciones, haciendo los datos incomparables entre sí, dándole más importancia a unas características que a otras. Es por esto por lo que es necesario un proceso de logre acotar el rango de posibles valores del conjunto total de datos. Existen distintas técnicas para aplicar la normalización en los datos, Existen diferentes técnicas de normalización como Mean Centered (MC), Variable Stability Scaling (VSS) o Min-Max Normalization (MMN), entre otras [?]. En esta tesis, para normalizar los datos y hacerlos comparables entre sí se ha utilizado la técnica de Z-Score (ZSN) debido a que logra representaciones de acuerdo con una distribución normal. Para hacerlo, se utilizan la media y la desviación estándar para reescalar los datos de manera que su distribución esté definida por una media de cero y una desviación estándar unitaria.

$$Z = \frac{(X - \mu)}{\sigma}$$

4.3.6. División Train-Val-Test

Luis: esto yo creo que estaría, lo único hacer inciso en lo último.

Los modelos supervisados de Inteligencia Artificial aprenden patrones sobre datos que son ofrecidos en la etapa de entrenamiento del modelo. Durante esta fase los modelos realizan predicciones sobre de datos y posteriormente se les enseña la clase a la que pertenecía cada uno de los datos que ha predicho, de esta forma se mide el error que han cometido durante este proceso y los pesos de la red son actualizados para minimizar el error en la siguiente fase. Si este aprendizaje se repite durante muchas etapas, el modelo tiende a aprenderse los datos de memoria, lo que se conoce como sobreajuste de la red u overfitting, provocando que la red no sea capaz de generalizar ante nuevas muestras tras su entrenamiento. Por este motivo es importante mantener el control del entrenamiento de la red mediante la evaluación del rendimiento de la red en cada

42CAPÍTULO 4. CONSTRUCCIÓN DE UN MODELO DE PREDICCIÓN DE LA GRAVEDAD DE UN

época mediante un conjunto de datos que nunca ha visto durante sus fases de entrenamiento, este conjunto de datos es conocido como conjunto de validación, y es utilizado para parar el entrenamiento cuando el modelo no sea capaz de generalizar sobre estas muestras. Por otra parte, existe el conjunto un conjunto de datos de test, utilizado para medir el rendimiento del modelo final una vez ha acabado su fase de entrenamiento. Este conjunto pertenece a muestras que la red no ha visto durante su fase de aprendizaje ni ha sido utilizado como validación.

En esta tesis se ha dividido el conjunto de datos original de cada una de las ciudades mediante... (80 % lo normal..)

4.3.7. Resampling

Luis: no me termina de convencer, estoy presentando métodos de resampling para justificar por qué usamos el SMOTE, pero me da la sensación de que colapsaría con el estado del arte si incluyésemos esto ahí.

El conjunto de datos, una vez se han reducido considerablemente el desbalanceo entre las dos clases gracias al proceso de filtrado de áreas, sigue presentando cierto desbalanceo. Por mucho que se haya acotado el problema a regiones individuales, es lógico que se hayan producido más accidentes sin necesidad de asistencia respecto a las que sí. En problemas de Inteligencia Artificial y Aprendizaje Estadístico es común aplicar técnicas que permitan igualar el número de muestras en conjuntos de datos donde se presenta desbalanceo. Existen dos principales corrientes que tienen como objetivo reducir la diferencia del desbalanceo de los datos. La primera de ellas consiste en igualar el número de muestras de la clase minoritaria hasta llegar a la mayoritaria (upsampling) mediante técnicas de reemplazamiento de datos (resampling), tal y tal . La segunda filosofía consiste en eliminar aleatoriamente registros de la clase mayoritaria hasta llegar al número de la minoritaria (undersampling) [?]. De esta forma se consigue un dataset balanceado que no provoque un sesgo en el entrenamiento de la red. En el caso de estudio de esta tesis, aplicar técnicas que eliminan accidentes sin necesidad de asistencia hasta igualar el número de aquellos que sí la requieren es un inconveniente, ya que al disponer de tan pocas muestras de la segunda clase, el conjunto de datos resultante se vería notablemente reducido, lo que afectaría negativamente al entrenamiento de la red, que requiere de un conjunto de datos lo más extenso posible para favorecer la generalización en sus predicciones. Por este motivo, en esta tesis se opta por métodos de aumento de datos (upsampling), que mantienen el valor que aportan las muestras de los accidentes sin necesidad de asistencia, aumentando los datos de aquellos que sí la requieren. Contras del resampling... Por estos motivos en este trabajo se ha optado por una técnica de generación de datos sintética denominada Synthetic Minority OverSampling Technique (SMOTE-II), que busca incrementar el número de clases de

las muestras minoritarias mediante la generación de nuevas muestras artificiales.

Para la generación de nuevas muestras sintética mediante SMOTE-II, se hace uso del espacio de características, para generar nuevas muestras de la clase mayoritaria que se encuentren cercanas al espacio de características que divide ambas clases. Para ello, se proyecta una nueva muestra de la clase minoritaria entre la línea que divide una muestra aleatoria respecto a uno de sus vecinos más cercanos.

Esta técnica permite generar datos sintéticos en base al contexto que conforman las muestras de la clase minoritaria hasta llegar a la mayoritaria.

En la Figura 4.4 de referencia se observa, marcados en azul, cómo los registros con identificadores 5 y 6 han sido generados en base a las modificaciones de los valores de los registros 1 y 3 para balancear el dataset.

4.4. Postprocesamiento

La segunda fase de la metodología implica transformar los datos refinados y balanceados en matrices interpretables por el modelo GTAAF recién propuesto. Este proceso implica mapear los atributos de las muestras tabulares en posiciones dentro de estas matrices. Para realizar esto, se hará uso de un método de transformación que toma en consideración la importancia de cada característica dentro del conjunto de datos. El objetivo es posicionar estratégicamente las características más relevantes en la matriz para maximizar su impacto en el modelo GTAAF, como se ilustra en la Figura 4.7. La determinación de la importancia de las características se basa en un algoritmo tipo boosting, que asigna pesos a las características según su relevancia en la separación de datos durante el entrenamiento. Para garantizar un entrenamiento óptimo del modelo, se realiza una optimización de hiperparámetros utilizando un algoritmo genético. A lo largo de generaciones sucesivas, este algoritmo genético hace evolucionar los hiperparámetros, guiado por la métrica de F1-Score, que actúa como la función heurística.

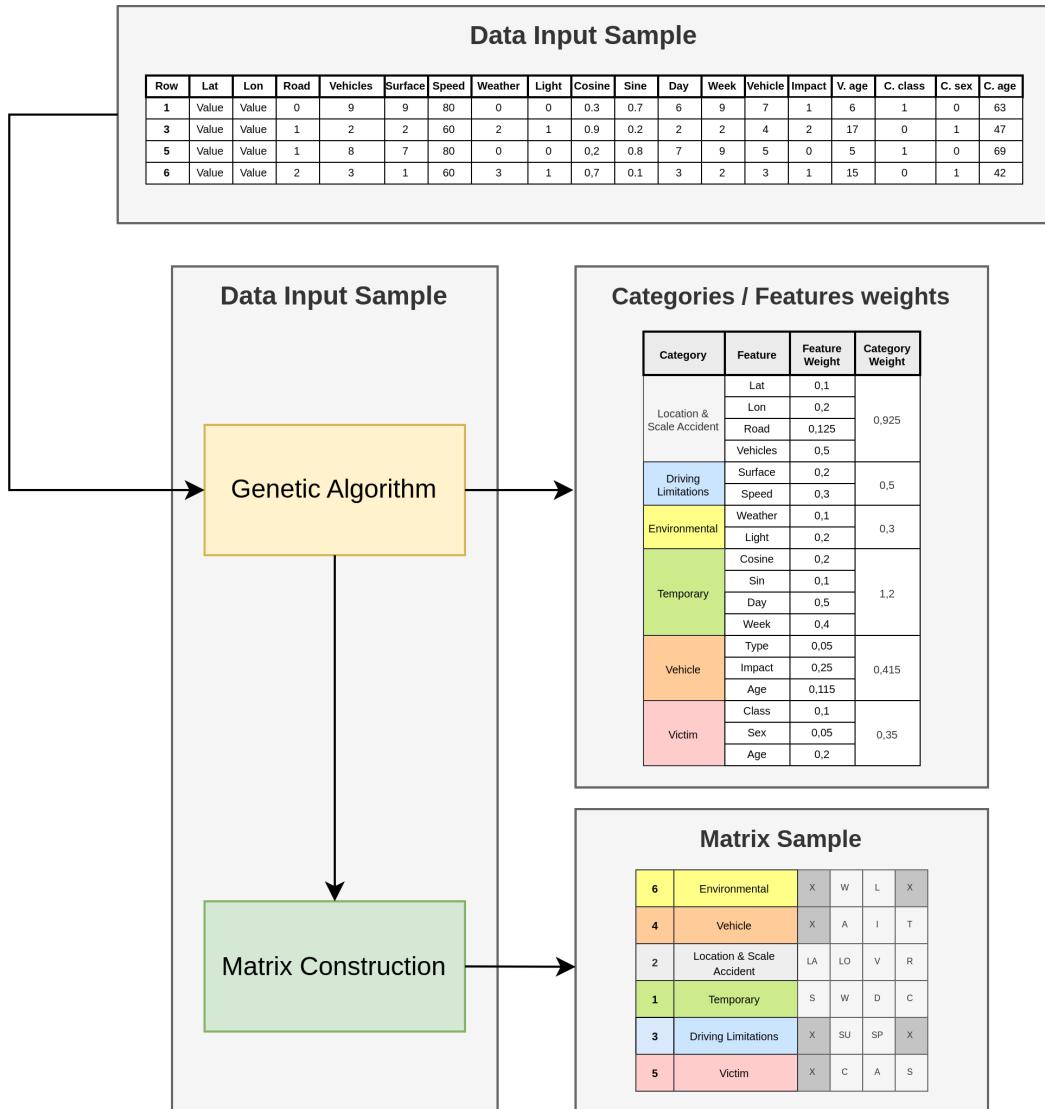


Figura 4.7: Category and feature weights.

4.4.1. Construcción de Matrices

Luis: este es el tinte que le quiero un poco dar, justificar por qué usamos lo que usamos y sobre todo justificar por qué no usamos otros métodos del estado del arte.

En esta tesis, se presenta un método para construir datos inicialmente ta-

bulares a datos matriciales con los que podrá trabajar el modelo convolucional propuesto, esta transformación hace uso de la categorización de las características y la importancia de cada una de ellas individualmente dentro del conjunto de datos a las que pertenecen. En la sección X se explicó el funcionamiento del proceso de categorización propuesto, que buscaba poder aplicar esta metodología a cualquier conjunto de datos de accidentes agrupando las características en conceptos básicos y de fácil categorización. El siguiente paso para lograr la transformación de los datos inicialmente en filas y columnas a datos matriciales es asignar cada una de las características del conjunto de datos a una posición dentro de la matriz, de tal forma que los datos puedan ser interpretados por el modelo convolucional. Para tener un contexto de la importancia en el orden en el que se asignan estas características, se explica brevemente la intuición sobre la que trabajan las redes neuronales convolucionales. Los píxeles que componen una imagen representan patrones que, para los seres humanos, son reconocibles, las redes convolucionales aprenden a reconocer estas variaciones, inicialmente en una escala pequeña (pocos píxeles), y, a medida que aumenta el número de capas, estas redes son capaces de aprender patrones más complejos en base a la composición del reconocimiento de aquellos más simples. Este funcionamiento, por definición, implica que la forma en la que se compone una imagen sea crítica, es decir, que el contenido que representa la imagen debe estar formado de manera coherente para que las redes puedan aprender estos patrones, requiriendo un sentido y/o contexto completo en su composición.

Existen distintos métodos que logran transformar datos tabulares a una representación matricial de los mismos, buscando dar un sentido a la asignación de las características en posiciones de la matriz. En la sección 3.4 se presentaron distintos métodos como REFINED, DeepInsight o IGTD, que buscan optimizar la posición de las características en base a la similaridad que presenten entre ellas, principalmente en datos orientados a la descripción genética. Sin embargo, estas técnicas presentan distintas limitaciones debido a la magnitud de los datos para las que han sido diseñadas (del orden de 2.500 características), esto provoca que estos métodos sean difícilmente aplicables a datos de baja dimensionalidad, como asignar espacios en blanco ante la falta de características o que los métodos no sean capaces de converger al trabajar con tan pocos datos. En el caso de estudio de esta tesis las características disponibles son mucho menores, del orden de 20 variables.

Debido a las limitaciones de los métodos anteriores, en esta tesis se presenta un método de composición de matrices en base a la importancia de las características, que permite asignar cada una de las variables del dataset a posiciones estratégicas dentro de la matriz haciendo uso de dos conceptos fundamentales, los Algoritmos Genéticos y los Algoritmos de Medición de Importancia de Características (feature importance).

4.4.2. Feature Importance Algorithm

Luis: Floja la justificación, pero van por ahí los tiros.

Como se ha presentado en la sección 3.5, existen distintos métodos que permiten evaluar la importancia de las variables en función de distintos criterios, como la correlación que presentan las variables entre sí o el nivel de importancia de cada característica a la hora de entrenar un modelo predictivo, ejemplos como estos son la Regresión Logística, técnicas de ensambles de tipo Bagging como los Random Forest o métodos ensambles tipo Boosting.

En esta tesis se trabaja con un dataset desbalanceado, por lo que a la hora de aplicar algoritmos de medición de características es importante escoger técnicas que sean insensibles a esto. Una de las muchas propiedades que ofrecen de los métodos de ensambles es que se adaptan especialmente bien a conjuntos de datos sesgados. Estos modelos, en sus distintas formas, se benefician de estar compuestos de una combinación de modelos y distintas técnicas de muestreo que reducen considerablemente el sobreajuste que pueda darse con otros métodos.

Dentro de estos modelos, los ensambles tipo Boosting son ampliamente conocidos por adaptarse especialmente bien en estos casos. Estos modelos utilizan técnicas de regularización durante su entrenamiento y se centran en minimizar el error producido cuando clasifican muestras de aquellas clases más conflictivas, que en el caso de un dataset desbalanceado serán las muestras minoritarias. Por otra parte, son modelos muy robustos que generalmente ofrecen un mayor rendimiento respecto a otros tipos de ensambles como los Random Forest, que únicamente ofrece que cada uno de los modelos sea entrenado con un subconjunto de los datos originales.

En esta metodología se utilizará el algoritmo tipo Boosting XGBoost, donde se minimizará la métrica F1-Score resultante de la clasificación de ambas clases de accidentes (Sin necesidad de asistencia y necesidad de asistencia). La principal limitación de este algoritmo para ofrecer un buen rendimiento es que requiere de una optimización de hiperparámetros con los que se entrena los datos.

4.4.3. Algoritmo Genético

Explicar qué métodos existen para optimizar hiperparámetros, por qué no nos sirven (costes computacionales) y por qué optamos por algoritmos genéticos (son muy buenos sin necesidad de hacer fuerza bruta)

Luis: Ni leerlo, tengo que

Existen numerosos métodos para optimizar hiperparámetros...

Es por esto por lo que en esta tesis, los algoritmos genéticos son utilizados para optimizar los hiperparámetros de entrenamiento del algoritmo XGBoost, que ofrecerá la importancia de las características, necesaria para la construcción de las matrices de entrada a la red CNN-2D propuesta. Donde cada uno de los individuos de la población del algoritmo genético representará una posible combinación de hiperparámetros, concretamente los valores de (Max Depth, ETA y N árboles). La función heurística que será optimizada será el F1-Score otorgado sobre los datos de test de cada uno de los conjuntos de datos.

4.4.4. Construcción de Matrices

Una vez se dispone de la categorización de los datos y de los pesos de las características gracias al modelo XGBoost, se aplica el proceso de asignación de cada una de las variables a posiciones de la matriz. Como se ha comentado en secciones anteriores, la forma en la que se compone una matriz sobre la que opera una red convolucional es de vital importancia, los filtros que estas aprenden seguir..... **Aquí otra vez?: Existen diferentes enfoques para construir matrices en base a datos tabulares, pero estos enfoques como se ha comentado en la sección XX sufren de limitaciones aplicados a nuestro caso de eso, (dimensiones de los datos, etc. ensar algo)** . Por este motivo se ha diseñado una estrategia que pretende posicionar las características más relevantes de cualquier conjunto de datos (definidas por el algoritmo XGBoost) a posiciones cercanas al centro de la matriz, que son las que... Este proceso teniendo en cuenta la categorización inicial que permite aplicar esta metodología a cualquier región, siendo tolerante a la falta de características en la disponibilidad de los datos que se ofrezcan.

El método diseñado sigue los siguientes pasos:

1. En primer lugar, las características son asociadas a sus categorías, de tal forma que pueda medirse la importancia de cada categoría dentro del conjunto de datos. Esta es obtenida mediante la suma del peso total de cada característica individual que la contiene
2. El segundo paso es asignar cada categoría con una fila de la matriz según su peso, donde aquella con el mayor peso se posiciona en la fila central, la segunda categoría más importante se asocia a la fila inmediatamente superior, la siguiente a la fila inmediatamente inferior y así sucesivamente (ver Figura 4.8).
3. Una vez que las categorías están asociadas a una fila de la matriz, cada una de las características dentro de su categoría se asocia en cada columna siguiendo el mismo procedimiento definido en el apartado anterior. La característica más importante de una categoría se posiciona en el centro, la segunda característica más importante se sitúa inmediatamente a su izquierda, mientras que la siguiente característica más importante ocupa el lugar a su izquierda y así sucesivamente (ver Figura 4.9).

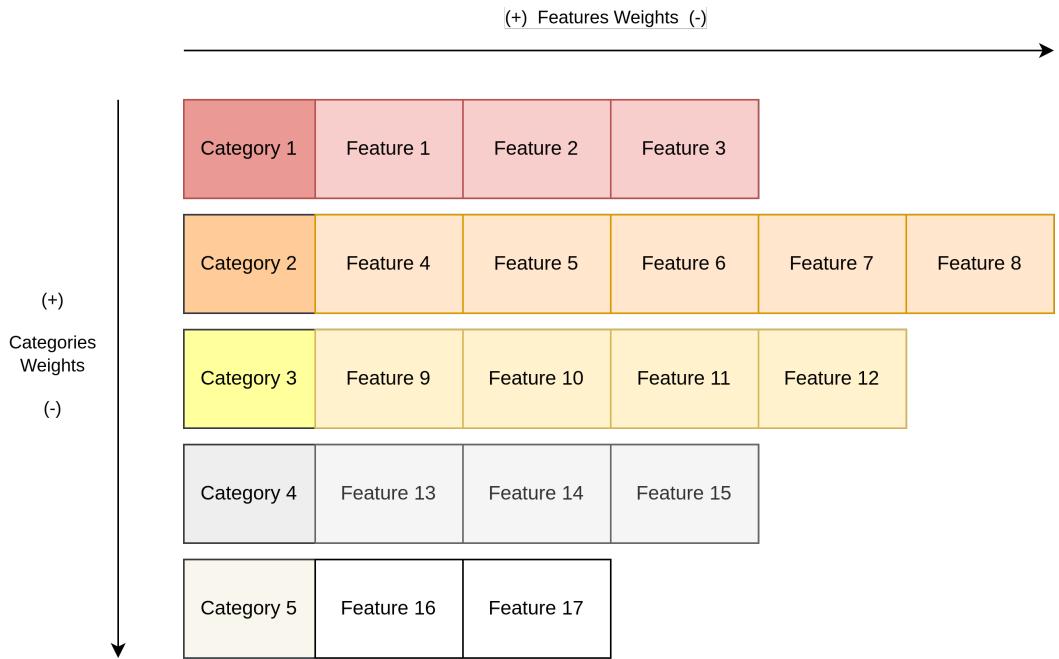


Figura 4.8: Category and feature weights.

El resultado de este proceso es una transformación de datos inicialmente tabulares en una matriz $n \times m$, donde n es el número de categorías disponibles en los datos y m es el número de máximo de características que contienen las categorías. Estas matrices están conformadas siguiendo que las variables más importantes para los datos se encuentran en las posiciones centrales, como se muestra en la Figura 4.9.

Category 4	0	Feature 14	Feature 13	Feature 15	0
Category 2	Feature 7	Feature 5	Feature 4	Feature 6	Feature 8
Category 1	0	Feature 2	Feature 1	Feature 3	0
Category 3	Feature 12	Feature 10	Feature 9	Feature 11	0
Category 5	0	Feature 17	Feature 16	0	0

Figura 4.9: Categories and feature positions.

En la Figura 4.10 se muestra un ejemplo del procedimiento

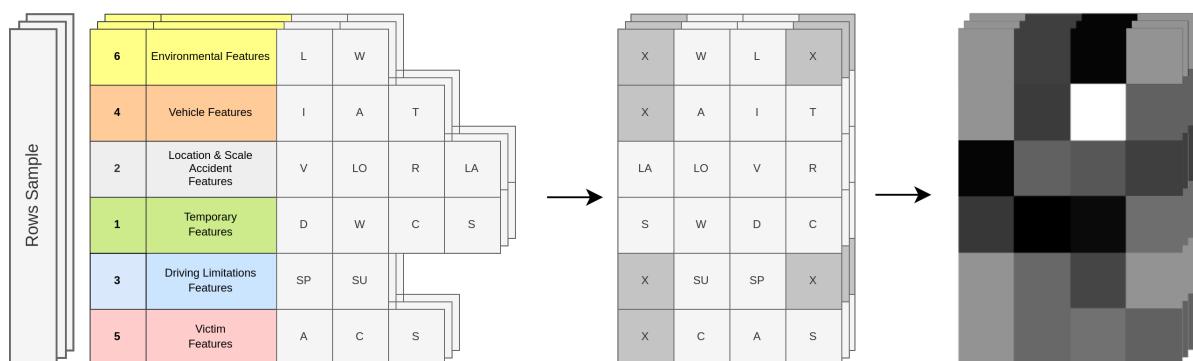


Figura 4.10: Assignment process of features to matrix positions. Categories are arranged based on their weight and assigned to rows of the matrix; subsequently, features within their respective categories are positioned.

4.4.5. Diseño del modelo

Luis: TODO Cambiar los hiperparams por los últimos (esto está copiado del primer paper).

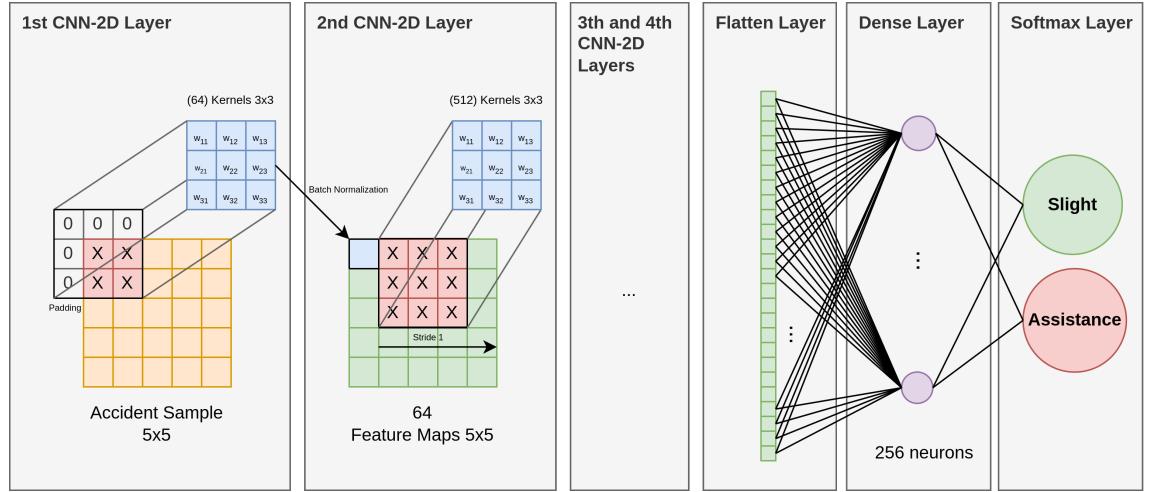


Figura 4.11: Proposed CNN-2D architecture summary.

4.5. Evaluación del modelo: Eficiencia y Robustez

Capítulo 5

Experimentos y resultados

5.1. Resultados preliminares - Prototipo

Aquí pones los resultados del paper 1

Explicar el artículo, indicando los enfoques tomados y las decisiones...

Como etapa previa al modelo final, y a modo de prototipo, se construyó un modelo primigenio que fue evolucionando hasta llegar a la metodología final expuesta en esta tesis. Sobre este primer modelo, se fueron aplicando modificaciones y mejoras en base al análisis de los resultados obtenidos durante su ciclo de vida hasta llegar a la "versión definitiva" de esta tesis. A modo de justificar las decisiones y criterios expuestos en este documento, en esta sección se expondrá el procedimiento inicial, los análisis de resultados y las mejoras propuestas que dan lugar a la versión final.

Este prototipo se presentó en el artículo [?], y se construyó con el objetivo de predecir la gravedad de los accidentes de tráfico en la ciudad de Madrid, dividiendo la severidad de los accidentes en tres clases (Leves, Severos y Fatales).

Descripción de datos

Los datos originales presentados en este prototipo pertenecían a la ciudad de Madrid, que describían ocurrencias de accidentes de tráfico a lo largo de toda la ciudad a través de 18 características entre los años 2019 y 2022, con un total de 60.966 registros. La variable a predecir representaba la lesividad que había sufrido la víctima implicada en el accidente, y que en el conjunto de datos era considerada en 7 clases, que se interpretaron finalmente como 3:

1. Leve: esto varía desde aquellos que no han sido heridos hasta aquellos que han necesitado ser admitidos en un hospital por no más de 24 horas. La cuantificación numérica es:

- Atención de emergencia sin posterior admisión hospitalaria: 1.
 - Admisión hospitalaria menor o igual a 24 horas: 2.
 - Atención médica ambulatoria después del accidente: 5.
 - Atención médica solo en el lugar del accidente: 6.
 - Sin atención médica: 7.
2. Grave: aquellos involucrados que han requerido hospitalización por más de 24 horas. En este caso, la cuantificación numérica es:
- Hospitalización por más de 24 horas: 3.
3. Fatal: fatalidades dentro de las 24 horas posteriores al accidente. La asignación numérica para este campo es:
- Fallecido dentro de las 24 horas: 4.

Limpieza

El resto de características describían información del accidente, como el lugar en el que se había producido, información del vehículo o información sobre la víctima. No obstante, existían conjuntos de variables que presentaban correlaciones entre sí (algo que afecta negativamente al rendimiento de los modelos) y contenían valores atípicos o nulos. Es por esto por lo que en primer lugar era necesario aplicar un proceso de análisis para evaluar el alcance y la calidad los datos aplicar que comenzaba por un proceso de limpieza que pretendía disponer de un dataset refinado e interpretable por distintos métodos, por lo que se eliminaron los registros con valores atípicos y aquellos que presentaban valores nulos, resultando un dataset final con 54.364 registros, un 10.82 % de pérdida de información respecto al original.

Discretización

Luis: Me parece raro presentar los datos ya filtrados y luego después de la tabla explicar que son los resultantes del proceso de eliminación en función del 0,44 de correlación).

En la figura 5.1 se muestra la descripción detallada de cada variable en esta etapa de la metodología.

Luis: **TODO** Esta tabla está copiada y pegada del paper 1).

Atributo	Descripción
ID de Incidente	Identificador del incidente, si varios registros tienen el mismo número de archivo, se consideran el mismo accidente y cada registro representa a cada una de las personas involucradas en él (conductor, pasajero o peatón)
Fecha	Día, mes y año en que ocurrió el incidente
Hora	Hora y minutos en que ocurrió el incidente
Tipo de Carretera	Tipo de carretera donde ocurrió el incidente
Nombre	Nombre de la calle donde ocurrió el incidente
Número de Calle	Número de la calle donde ocurrió el incidente
Distrito	Nombre del distrito donde ocurrió el incidente
Tipo de Accidente	Puede ser: doble colisión, colisión múltiple, alcance, colisión con un obstáculo, atropello, vuelco, caída u otras causas
Condiciones climáticas	Condiciones climáticas en el momento del incidente
Vehículo	Clasificación según tipos de vehículos
Persona	Rol de la persona involucrada: conductor, pasajero o peatón
Edad	Rango de edad de la persona involucrada
Género	Mujer u hombre
Severidad	Consecuencias físicas de la persona involucrada, si han necesitado atención médica, si han sido hospitalizados o si han sido fatales
X	Coordenada X - UTM
Y	Coordenada Y - UTM
Alcohol	Si la persona involucrada ha dado positivo en alcohol (S o N)
Drogas	Si la persona involucrada ha dado positivo en drogas (S o N)

Cuadro 5.1: Variables del conjunto de datos y sus descripciones.

Una vez se disponen de unos datos refinados, era necesario transformarlos para hacerlos interpretables por los modelos. Este proceso se hizo mediante la asignación de valores numéricos a cada una de las variables cualitativas del dataset, en función de la fuerza del significado de los valores de cada característica, en la Figura 5.2 se muestra la discretización de las variables seleccionadas de este conjunto de datos.

Características	Característica	Tipificación
'Gravedad!' Gravedad!Gravedadpt<	ipo de Camino!ipo de Camino!	0: Leve (1, 2, 5, 6, 7) 1: Grave (3) 2: Fatal (4)
iempo!iempo!Tiempopt<		1: Noche (6 PM - 6 AM) 2: Día (6 AM - 6 PM)
istrito!istrito!Distritopt<		Basado en orden de aparición
!!Xpt<		Posición de Coordenada UTM X
!!Ypt<		Posición de Coordenada UTM Y
ipo de Accidente!ipo de Accidente!Tipo de Accidentept<	identept<1: Colisión frontal - tamaño condicione	Meteorológicas!Condiciones Meteorológicas!
	2: Colisión trasera	
	3: Choque lateral	
	4: Colisión con obstáculo fijo	
	5: Choque en cadena	
	6: Atropello a peatón	
	7: Colisión frontal	
	8: Otro	
ehículo!ehículo!Vehícuopt<		
ersona!persona!Personapt<	9: Salida de la carretera	
	10: Vuelco de vehículo	
	11: Atropello a animal	
	12: Caída	
dad!dad!Edadpt<		
ipo de Camino!ipo de Camino!Tipo de Caminopt<	1: Estacionamiento	
	2: Aeropuerto	
	3: Parque	
	4: Túnel	
	5: Zona industrial	
énero!énero!Géneroopt<	6: Pista	
	7: Rotonda	
	8: Glorieta	
lcohol	9: Puerta	Drogas!Alcohol o
		Drogas!Alcohol o

Cuadro 5.2: Asignación numérica de las variables del conjunto de datos.

Para entrenar un modelo de Inteligencia Artificial es necesario analizar la dependencia entre cada par de variables, por esto se analizó la relación entre variables mediante una matriz de correlación. Estas matrices muestran la fuerza mediante la que variable es dependiente respecto al resto de las demás, los coeficientes de correlación varían entre -1 y 1, indicando la magnitud y dirección de esta dependencia. Una vez analizadas estas métricas, se aplicó un límite de correlación entre variables del $\pm 0,44$, lo que quiere decir que aquellas que presentasen un índice que superase este valor se verían excluidas del dataset. En la figura 5.1 se muestra la matriz de correlación resultante tras eliminar las características que superasen este umbral de dependencia entre sí.



Figura 5.1: Correlation matrix between the dataset variables.

Resampling

Una vez aplicado el proceso de limpieza de datos y elección de características del dataset, se analizó la distribución final de los datos en base a la clase a predecir, la severidad de accidente. Atendiendo a los registros resultantes (Leve, Grave y Fatal), se puede observar que el conjunto de datos está claramente desbalanceado. Se disponían de 53,009 accidentes leves, 1,271 graves y 84 fatales. Esto se convierte en un problema para los modelos de clasificación, ya que tienden a predecir las muestras como pertenecientes a la mayoría del conjunto de pruebas. Para paliar este problema se aplicó la técnica de remuestreo Borderline SMOTE-II para generar más muestras de accidentes pertenecientes a clases minoritarias (Grave y Fatal), evitando que el modelo se sobreajuste. Una vez aplicado el algoritmo, se obtienen 42,508 muestras de cada una de las clases de accidentes, es decir, un total de 127,524 registros.

Normalización

En la Tabla ?? se muestra un ejemplo de la aplicación de la normalización de datos en base a la técnica Z-Score, donde en la primera columna se observan los datos previos a esta normalización, mientras que la segunda columna contiene los valores de estas características normalizados.

Luis: Aquí metemos el ejemplo del TFM

Característica	Valor
hora	2
tipo carretera	19
distrito	14
tipo accidente	1
estado meteorológico	1
tipo vehiculo	4
tipo persona	1
rango edad	3
sexo	1
drogas alcohol positivo	2
vehiculos implicados	1
coordenada x utm	438950266
coordenada y utm	4473953232

Figura 5.2 Muestra de accidente tipificada.

Característica	Valor
hora	1.2548
tipo carretera	0.4597
distrito	-0.0297
tipo accidente	-1.4528
estado meteorológico	-0.2508
tipo vehiculo	-0.1621
tipo persona	-0.5316
rango edad	0.2129
sexo	-0.7004
drogas alcohol positivo	0.1488
vehiculos implicados	-1.4591
coordenada x utm	-0.0524
coordenada y utm	0.0081

Figura 5.3 Muestra de accidente normalizada.

^{.4}				
0.0	0.0	-0.1621	0.0	0.0
1.2548	0.0081	-0.0524	-1.4528	-1.4591
0.2129	-0.7004	-0.5316	0.1488	0.0
0.0	-0.0297	0.4597	0.0	0.0
0.0	0.0	-0.2508	0.0	0.0

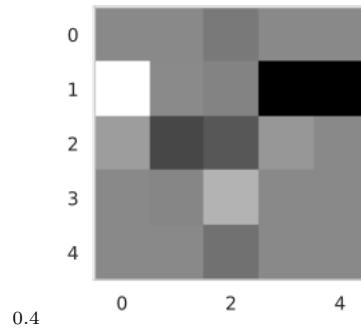
Figura 5.3 Características.**Figura 5.4** Imagen de las características.

Figura 5.5: Proceso que sigue una muestra del conjunto de datos tipificada hasta llegar a una matriz ($a \rightarrow b \rightarrow c \rightarrow d$).

Una vez se disponen de los datos normalizados, estos ya son comparables y por tanto pueden ser utilizados para el entrenamiento de cualquier modelo que acepte valores numéricos como entrada.

División Train-Val-Test

Una vez se disponen de las categorías

Categorización

Como se ha comentado en la sección de metodología, las redes neuronales convolucionales (CNN) aprenden patrones utilizando matrices como datos de entrada, y cuando se trata de datos tabulares, es necesario transformar estos datos a datos matriciales.

Para lograr este objetivo, uno de los requisitos de esta transformación era asignar cada característica a una categoría del dataset. Sobre este conjunto de datos, las variables eran asignadas a 5 categorías: Características del accidente, Condiciones de la carretera, Condiciones meteorológicas, Características del vehículo y Características del conductor. En la Tabla 5.3 se observa la categorización de cada característica a una categoría en función del concepto que describan.

Categoría	Característica
Accidente	X
	Y
	Hora
	Tipo de accidente
Carretera	Severidad
	Tipo de carretera
Distrito	
Clima	Condiciones climáticas
Vehículo	Vehículo
Conductor	Persona
	Género
	Edad
	Alcohol o Drogas

Cuadro 5.3: Clasificación de las Características (variables del conjunto de datos) en Categorías.

Algoritmo Genético

En este punto, se analiza la optimización de los hiperparámetros del algoritmo de Boosting mediante un algoritmo genético. La figura 5.6 muestra la evolución de los tres hiperparámetros a lo largo de las generaciones. Como se puede observar, los hiperparámetros toman distintos valores en función del mejor individuo evaluado en la población en cada etapa, estos hiperparámetros convergen aproximadamente en la iteración 42.

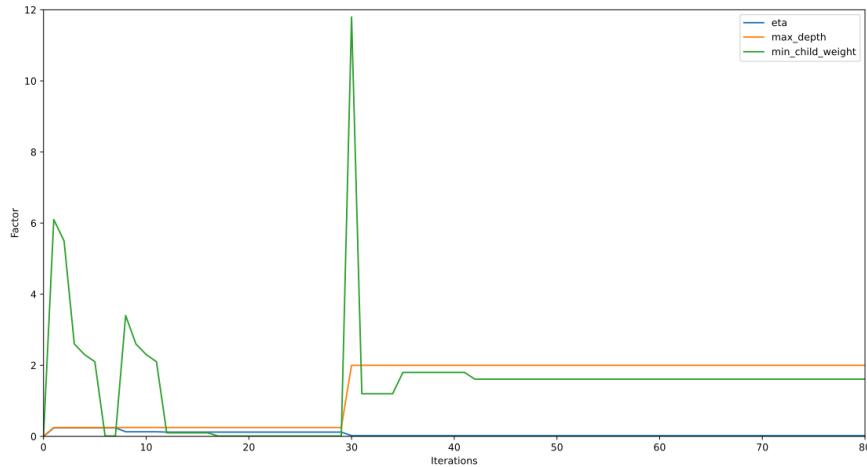


Figura 5.6: Evolution of hyperparameters throughout the iterations.

En la tabla 5.4 se observa el valor tomado el mejor individuo entre las distintas generaciones para cada uno de los hiperparámetros del XGBoost, esta será la configuración con la que se entrenará el algoritmo para obtener el peso de las características del dataset.

Hyperparameters	Value
Max deep	2
Minimum weight of children	1.6
ETA	0.007
Gamma	0.3
Alpha	0
Lambda	1

Cuadro 5.4: Optimized values of the parameters after applying the genetic algorithm.

En la Tabla 5.5 se observa el peso asignado a cada una de las características individuales resultante del entrenamiento XGBoost con los hiperparámetros optimizados. En la columna "Categoría de Peso" se observa el peso de cada categoría, que es la suma del peso de las características individuales que las componen.

Categoría	Categoría de Peso	Característica	Peso de la Característica
Accidente	0.299	- Coordenada X - Coordenada Y - Hora - Tipo de accidente - Severidad	0.071 0.066 0.055 0.051 0.057
Carretera	0.187	- Distrito - Tipo de Carretera	0.059 0.127
Clima	0.050	- Condiciones Climáticas	0.050
Vehículo	0.070	- Vehículo	0.070
Conductor	0.394	- Persona - Género - Edad - Alcohol o Drogas	0.177 0.111 0.050 0.056

Cuadro 5.5: Ejemplo con los pesos de todas las características estudiadas, así como los pesos de las cinco categorías.

5.1.1. Construcción de matrices

Una vez se disponían de las características y categorías evaluadas, se aplicaba el proceso de asignación de posiciones de cada característica a una coordenadas dentro de la matriz, aplicando el algoritmo de construcción de matrices.

En la tabla 5.6 se observa un ejemplo de un registro (originalmente tabular) transformado a formato matricial.

0.0	0.0	0.05	0.0	0.0
0.0	0.059	0.128	0.0	0.0
0.050	0.111	0.177	0.056	0.0
0.055	0.066	0.071	0.057	0.051
0.0	0.0	0.070	0.0	0.0

Cuadro 5.6: A specific accident matrix.

Entrenamientos

Las figuras 5.7 y 5.8 muestran la evolución de la métrica de puntuación F1 a lo largo de las 100 ejecuciones para las redes neuronales convolucionales 1D y 2D. Al visualizar la convolución unidimensional (Figura 5.7), se puede verificar que la puntuación F1 de entrenamiento aumentaba ligeramente a lo largo de las épocas, experimentando altibajos a medida que el modelo se entrena, comenzando inicialmente con un valor de entrenamiento inferior a 0,58 y llegando hasta 0,68, mostrando poca capacidad de aprendizaje y generalización ante nuevas muestras.

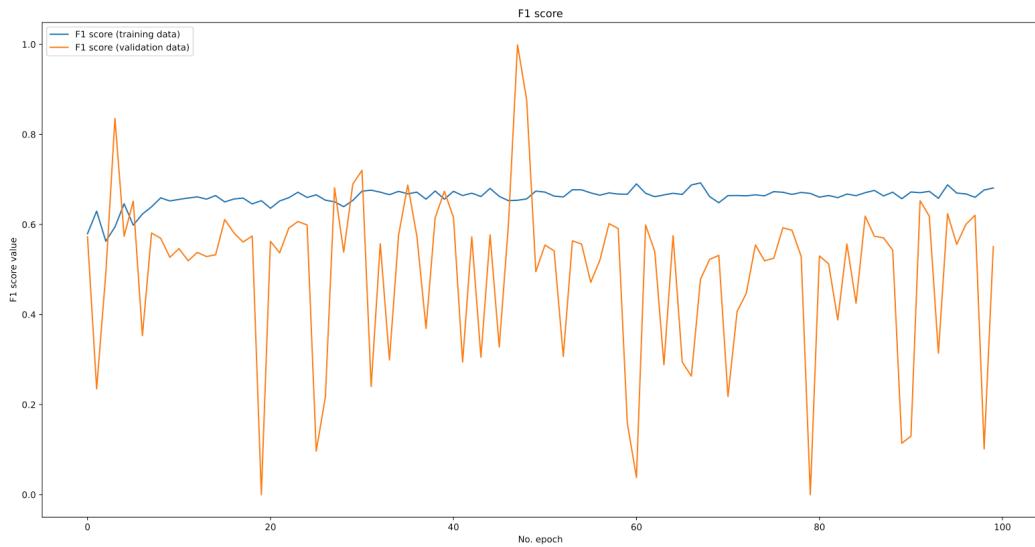


Figura 5.7: Evolution of the F1-score of the 1D-CNN in the training and test set.

Por otro lado, la Figura 5.8 muestra el gráfico de entrenamiento y validación de la red neuronal convolucional bidimensional. Se observó que la tendencia de la función de pérdida en el conjunto de datos de entrenamiento era más estable. Se puede ver cómo la red, en la primera ejecución, comienza con un puntaje F1 de 0,62 hasta alcanzar 0,78 en la iteración 100, por lo que se puede deducir que esta red logró un mejor rendimiento en el conjunto de entrenamiento en comparación con la red convolucional unidimensional, sufriendo menos altibajos respecto en el conjunto de validación.

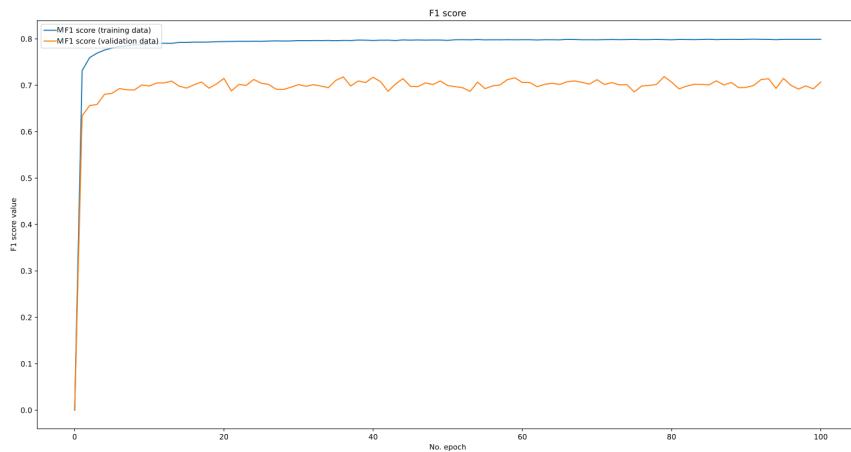


Figura 5.8: Evolution of the F1-score of the 2D-CNN in training and test set.

Resultados

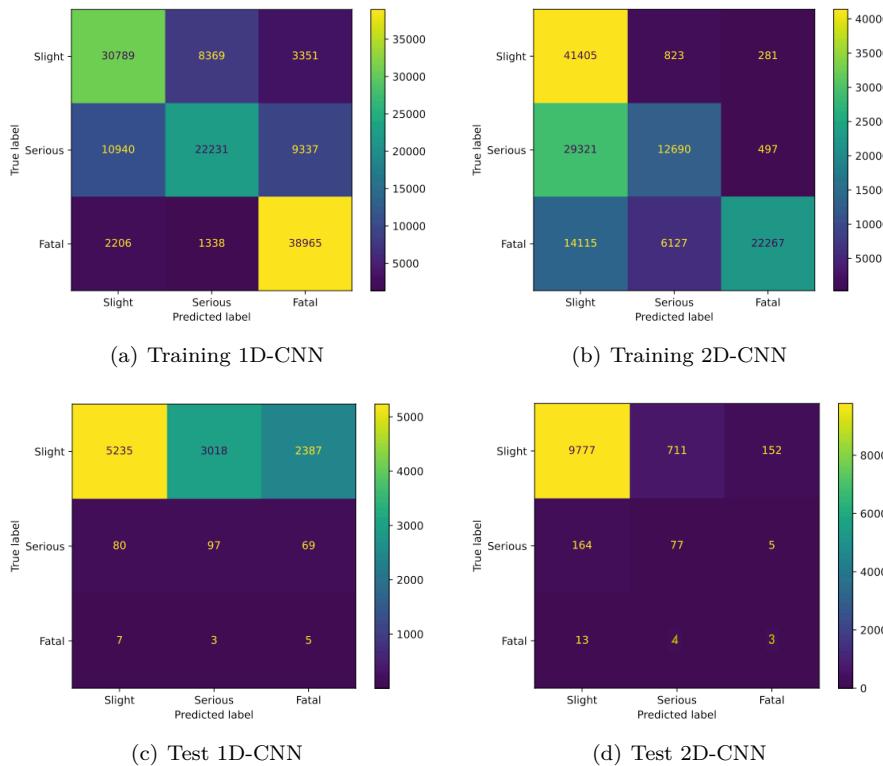


Figura 5.9: Confusion Matrices for Convolutional neural networks.

Resultados

Para evaluar el rendimiento de la metodología y los modelos propuestos, se realizó una comparación con tres modelos del estado del arte, el Gaussian Naive Bayes, Support Vector Classifier y K-Nearest Neighbor.

Los resultados de las métricas de clasificación se muestran en la Tabla 5.9 para cada una de las clases predichas en el conjunto de **pruebas**. Estos informes muestran la información con la que evaluamos los modelos, ya que explica cómo se comportan con respecto a nuevos datos.

Como se puede ver en esta tabla, el modelo KNN obtiene mejores resultados en todas las medidas de todas las clases excepto en Recall en accidentes graves, donde el GNB es un poco mejor.

Si analizamos la métrica de Precisión, se puede observar que el modelo que presenta el mejor promedio para las clases Leves es la Red Neuronal Convolutacional 1D (1D-CNN) con 0,984, seguido por la Red Neuronal Convolutacional 2D (2D-CNN) y el modelo KNN con 0,982. Además, el 2D-CNN también ofrece la mejor métrica para accidentes graves con 0,097, con una gran diferencia respecto

al modelo KNN que le sigue con 0,042. En cuanto a los accidentes fatales, tanto los modelos 1D-CNN como 2D-CNN tienen un valor similar, obteniendo 0,002.

Respecto a la métrica de Recall, el mejor promedio para las clases Leves es la Red Neuronal Convolucional 2D (2D-CNN) con 0,919, seguido por el modelo KNN con 0,689. Además, el modelo GNB ofrece la mejor métrica para accidentes graves con 0,699. En accidentes fatales, el 2D-CNN tiene el mejor valor con 0,1.

Es necesario señalar que el F1-score es una forma de combinar las métricas de Precisión y Recall, y se define como la media armónica de la Precisión y Recall del modelo. Teniendo esto en cuenta, si analizamos el F1-score de los informes, el modelo que presenta el mejor promedio para las clases Leves es la Red Neuronal Convolucional 2D (2D-CNN), alcanzando 0,950, muy por encima del siguiente modelo KNN, que ofrece un valor de 0,810. Además, el 2D-CNN también ofrece la mejor métrica para accidentes graves con 0,148, alcanzando el doble del rendimiento en comparación con el modelo que le sigue, el KNN con 0,076. Respecto a los accidentes fatales, los modelos con la mejor clasificación son tanto la Red Neuronal Convolucional 1D como la 2D, obteniendo 0,004, el doble que KNN, que son los siguientes mejores modelos en esta clase con 0,002.

Podemos concluir que el modelo propuesto, basado en redes neuronales convolucionales, presenta mejores predicciones en cuanto a la métrica F1-score, que es una combinación de Precisión y Recall.

Metric/Severity	1D-CNN			2D-CNN		
	Slight	Serious	Fatal	Slight	Serious	Fatal
Precision	0.701	0.696	0.754	0.488	0.646	0.966
Recall	0.724	0.523	0.917	0.974	0.299	0.524
F1-score	0.712	0.597	0.828	0.650	0.409	0.679

Cuadro 5.7: Training metrics for 1D-CNN and 2D-CNN.

Metric/Severity	1D-CNN			2D-CNN		
	Slight	Serious	Fatal	Slight	Serious	Fatal
Precision	0.984	0.031	0.002	0.982	0.097	0.002
Recall	0.429	0.394	0.333	0.919	0.313	0.1
F1-score	0.596	0.058	0.004	0.950	0.148	0.004

Cuadro 5.8: Test metrics for 1D-CNN and 2D-CNN.

Resultados

Metric/Severity	GNB			SVC			KNN		
	Slight	Serious	Fatal	Slight	Serious	Fatal	Slight	Serious	Fatal
Precision	0.980	0.025	0	0.979	0.029	0	0.982	0.042	0.001
Recall	0.369	0.699	0	0.644	0.411	0	0.689	0.382	0.067
F1-score	0.536	0.048	0	0.777	0.054	0	0.810	0.076	0.002

Cuadro 5.9: Test metrics classification for GNB, SVC and KNN.

5.1.2. Conclusiones

Trabajos a futuro en función de los resultados obtenidos, justificando el por qué de dichos cambios...

Analizando los resultados de este artículo se propusieron una serie de mejoras a implementar para crear un modelo útil en la severidad de los accidentes, y aportar valor a los cuerpos de emergencia. Estos cambios a implementar fueron:

1. Unión de la severidad original de los accidentes de tres clases a dos clases para paliar el efecto de la superposición en la clasificación. Concretamente realizar una agrupación de las clases Accidentes Severos y Accidentes Fatales a Necesidad de Asistencia.
2. Inclusión de nuevas características para enriquecer la información con la que trabaja el modelo, tanto recopilación de nuevos datos como aplicar transformaciones sobre ellos para disponer de mayor información
- 3.

5.2. Resultados finales

Aquí pones los resultados del paper 3

5.2.1. Dataset

Los datos escogidos para esta tesis pertenecen a tres conjuntos de datos distintos, con el objetivo de poder evaluar la metodología y el nuevo modelo propuesto en diferentes contextos. Esta evaluación se basará en dos factores principales; la distinta disponibilidad de información en los datos, y en distintos casos de estudio en función de la densidad de población de las regiones escogidas. Teniendo en cuenta la densidad de población podemos distinguir entre tres casos de estudio claramente diferenciados: (1) alta concentración de población, (2) concentración media y (3) concentración dispersa. Con esta variabilidad en los datos se busca medir la robustez y generalización de la técnica desarrollada.

El primer dataset seleccionado contiene información de accidentes sobre la Comunidad de Madrid, donde se describen los accidentes de tráfico producidos entre 2019 y 2022 a lo largo de toda la comunidad. La alta densidad de población de Madrid convierte este conjunto de datos en un caso de estudio de alta concentración de población. Este conjunto de datos ha sido extraído del Portal de Datos Abiertos del Ayuntamiento de Madrid [?].

El segundo caso de estudio contempla una situación de concentración de población dispersa, concretamente a lo largo del estado de Victoria, Australia, contemplando los accidentes producidos entre el 2000 y el 2005. Este conjunto de datos ha sido obtenido a través del Departamento de Transportes y Planificación del Gobierno de Victoria [?].

El tercer y último conjunto de datos pertenece al Departamento de Transportes de Reino Unido [?], donde se contempla información de los accidentes producidos entre 2005 to 2020 a lo largo de todo el país. Sobre este conjunto de datos se han extraído accidentes pertenecientes a 6 regiones diferentes, concretamente: Southwark, Manchester, Birmingham, Liverpool, Sheffield y Cornwall. Cada una de ellas presenta un caso de uso distinto en función de su densidad de población.

5.2.2. Descripción de datos

En esta sección se explicará la variabilidad en la disponibilidad de los datos entre los conjuntos de datos escogidos. Como se ha comentado anteriormente, cada uno de los datasets contiene distinta información en función de los recursos que disponga cada región, como puede ser la capacidad que se tenga para realizar pruebas de alcoholemia, la recogida de información de las condiciones actuales de la carretera o entre otras, lo que provoca una heterogeneidad que debería ser tratada individualmente para cada población en específico . Para solventar este problema, y con el objetivo de crear una metodología y modelo predictivo generalizables a cualquier población independientemente de las características individuales que esta contenga, se propone agrupar las características disponibles en categorías fácilmente reconocibles. De esta forma, todos aquellos descriptores del accidente serán asignados a un concepto, donde cada uno de estos permite asignar características de muy fácil obtención, permitiendo así utilizar la metodología tanto para conjuntos de datos donde se disponga de información muy específica como para conjuntos de datos donde se contemple información más simplificada. Las categorías propuestas donde serán englobadas las características son las siguientes:

1. Magnitud y ubicación del accidente: enfocado en información relativa a la localización y magnitud del accidente, como datos geográficos.
2. Limitaciones de Conducción: abarcan características que limitan al conductor, como regulaciones reales a los límites de velocidad o condiciones actuales de la carretera.

3. Factores ambientales: condiciones climáticas y de visibilidad.
4. Información Temporal: relacionada con el momento del accidente.
5. Información del Vehículo: características que describan al vehículo objeto del accidente.
6. Información de la Víctima: descriptores que definan a la víctima en el momento del incidente, factores como la edad, sexo, postivo en sustancias estupefacientes, etc.

En la Figura 5.10 se presentan las características disponibles en cada uno de los conjuntos de datos escogidos en esta tesis, con el objetivo de mostrar la variabilidad de información que puede existir entre distintas poblaciones. Los campos marcados en naranja son aquellos que representan información distinta entre los datasets pero que pueden ser incluidos en las categorías correspondientes. Por otra parte, aquellos campos marcados en rojo indican la ausencia de este tipo de características en comparación con el resto de datasets.

	UK	Madrid	Victoria
Location & Scale	Latitude Longitude Road Class Number of Vehicles	Latitude Longitude District Number of Vehicles	Latitude Longitude Type of Accident Place Number of Vehicles
Driving Limitations	Road Surface Speed Limit	X X	Road Surface Speed Limit
Environmental	Weather Conditions Lighting Conditions	Weather Conditions X	Weather Conditions Lighting Conditions
Temporary	Cosine Hour Sine Hour Day on Week Week on Year	Cosine Hour Sine Hour Day on Week Week on Year	Cosine Hour Sine Hour Day on Week Week on Year
Vehicle	Vehicle Type First Point of Impact Age of Vehicle	Vehicle Type First Point of Impact X	Vehicle Type First Point of Impact Age of Vehicle
Victim	Casualty Class Casualty Sex Casualty Age X	Casualty Class Casualty Sex Casualty Age Alcohol/Drugs Positive	Casualty Class Casualty Sex Casualty Age X

Figura 5.10: Classification of variables. Fields shown in yellow represent features of the same nature but differing in data granularity. Additionally, missing features compared to other datasets are highlighted in red.

Partes comunes entre los datos

Normalmente, en cualquier conjunto de datos que describa accidentes, existe información básica y de fácil obtención que suele ser común entre distintas poblaciones.

Estas características comunes suelen ser información espacial, como es la localización del accidente, las condiciones climáticas en el momento del suceso y la hora y fecha en la que se ha producido. Por otra parte, como es lógico, existe información de fácil obtención que puede ser recogida rápidamente 'echando un vistazo rápido' al los vehículos implicados en el accidente, como es el tipo de vehículo colisionado y en cuál ha sido el primer punto de impacto.

Principales diferencias entre los datos

Como se puede observar, cada uno de los conjuntos de datos tiene una naturaleza diferente y ofrecen distinta información.

UK

En el caso de UK se observan ligeras diferencias respecto al resto de conjuntos de datos. Como es el caso de la característica Road class (para la categoría Location & Scale Accident), cuyo significado varía en comparación con el resto de conjuntos de datos, y la ausencia de información sobre controles de estupefacientes a la víctima (categoría Victim). En el caso de Road Class, este campo representa la clasificación de la carretera en la que se ha producido el accidente en base al tráfico que suelen contener. Esta clasificación es responsabilidad del Gobierno de UK y se clasifican las vías en seis tipos diferentes: (1) Motorways: se trata de autopistas de alta velocidad que permiten el movimiento de vehículos entre los principales pueblos y ciudades. (2) A(M): se trata de carreteras principales que interconectan poblaciones y destinos de interés, estas vías pueden contener secciones transformadas en autovía. (3) A: carreteras importantes que conectan grandes densidades de tráfico entre zonas. Generalmente son las más anchas y directas, y son las de mayor importancia para el tráfico que contiene el área, estas carreteras pueden estar abiertas a distintos usuarios, como vianandantes, ciclistas o caballos, aunque normalmente esto está restringido por las autoridades locales competentes. (4) Las carreteras B alimentan el tráfico entre las vías A y las carreteras más pequeñas de la red, siguen siendo de especial importancia para el tráfico, pero menos que las A. (5) Las carreteras tipo C son generalmente más pequeñas e interconectan las vías de tipo A y B. Normalmente unen urbanizaciones con el resto de carreteras de la red, son carreteras de menor importancia que las anteriores pero son de mayor relevancia respecto a las del siguiente tipo. (6) Carreteras no clasificadas, se tratan de vías destinadas al tráfico local, por su naturaleza la mayoría de las vías pertenecen a este tipo, generalmente tienen muy poca importancia y a nivel local [?].

Madrid

En el caso del conjunto de datos de Madrid, las diferencias respecto al resto de datasets es más notable. La información disponible es considerablemente menor en comparación con el dataset de UK y de Victoria. Analizando la Figura se puede observar que hay ciertas características que no están presentes, llegando a dejar incluso una categoría vacía (Driving Limitations) al no disponer de información de este tipo. Por otra parte, tampoco se dispone de la información de Lighting Conditions para la categoría Environmental ni de Age of Vehicle, en la categoría de características del vehículo. No obstante, aún faltando esta información, el resto de características pueden ser asignadas a las categorías definidas, convirtiendo, por tanto, este dataset aplicable a esta metodología.

Sin embargo, el conjunto de datos de Madrid ofrece información sobre si la víctima se encuentra bajo los efectos del alcohol o de sustancias estupefacientes.

Al ser un dato que describe a la víctima del incidente, este será asignado a la categoría Victim.

Por otro lado, en la categoría Location & Scale Accident los datos de Madrid presentan una diferencia en lo que representa la característica District respecto al resto de datasets. Este campo ha sido obtenido mediante expresiones regulares, buscando distintos tipos de vía sobre la columna que ofrece información acerca del nombre de la calle. De tal forma que contiene engloba información del tipo de vía urbana o interurbana sobre la que transitaba el vehículo en el momento en el que se produjo el accidente, como avenidas, boulevares, entre otras. Al ser una característica que ofrece información sobre la localización del accidente, será incluida en la categoría de Location & Scale Accident. En la tabla X de anexos pueden consultarse los distintos valores que esta característica puede tomar.

Victoria

El conjunto de datos de Victoria contempla un caso parecido al de los datos de UK, donde no se disponen de datos que describan si la víctima se encontraba bajo los efectos de estupefacientes o del alcohol, como es el caso del conjunto de datos de Madrid. Por lo que esta característica quedará vacía también en el dataset de Victoria.

Por otra parte, la característica Type Of Accident Place, ofrece información sobre el lugar del accidente, concretamente el lugar donde se ha producido, como autopista, parking, tunel, etc. por lo que irá asignada a la categoría Location & Scale Accident.

5.2.3. Limpieza

En la tabla 5.12 se expone el número total de registros del conjunto de datos original y el número de muestras resultante tras haber aplicado la limpieza de estos datos para cada una de las poblaciones contempladas en esta tesis.

Data Distribution			
UK			
Region	Assistance	Original	Cleaned
Southwark	No	X	X
	Yes	X	X
Manchester	No	X	X
	Yes	X	X
Birmingham	No	X	X
	Yes	X	X
Liverpool	No	X	X
	Yes	X	X
Sheffield	No	X	X
	Yes	X	X
Cornwall	No	X	X
	Yes	X	X
Spain			
Region	Assistance	Original	Filtered
Madrid	No	X	X
	Yes	X	X
Australia			
Region	Assistance	Original	Filtered
Victoria	No	X	X
	Yes	X	X

Cuadro 5.10: XX

Como se puede observar, la población que más valores nulos presenta en las categorías de interés es la ciudad de XX, obteniendo una pérdida del X % respecto al conjunto de datos inicial...

En la Figura X se muestra una representación visual del número de registros previo al proceso de limpieza respecto al resultado de esta etapa. (Histogramas pre y post limpieza).

5.2.4. Filtrado de áreas

Para ilustrar los parámetros con los que se aplica el filtrado de áreas, se presenta la Tabla 5.11, donde se muestra para cada población el número de áreas por las que pasará el filtro, además del tamaño de la ventana X,Y para cada región, donde en función de la extensión y la densidad de población tomará valores más grandes (poblaciones más dispersas) o valores más pequeños (cuando la densidad de población es alta). Los tamaños de ventana para cada

población han sido escogidos mediante un procedimiento experimental, en el que se maximiza el rendimiento final de los modelos.

Areas Split			
UK			
City	Axis	Areas Number	Areas Size
Southwark	X	529	10
	Y	487	20
Manchester	X	791	14
	Y	1069	20
Birmingham	X	3519	12
	Y	1557	17
Liverpool	X	2107	12
	Y	717	21
Sheffield	X	1896	12
	Y	1115	18
Cornwall	X	10090	15
	Y	5597	19
Spain			
City	Axis	Areas Number	Areas Size
Madrid	X	5241	5
	Y	4444	7
Australia			
City	Axis	Areas Number	Areas Size
Victoria	X	4931	145
	Y	5241	97

Cuadro 5.11

Una vez se establecen las dimensiones de las ventanas de tamaño X,Y se aplica el filtrado para cada ciudad, donde el número de muestras de la clase mayoritaria se ve considerablemente rebajado respecto a la minoritaria, con el objetivo de tener un conjunto de datos más balanceado. La tabla 5.12 muestra el número de registros original para cada población y el número de registros resultante tras aplicar el filtrado por áreas.

Data Distribution			
UK			
Region	Assistance	Original	Filtered
Southwark	No	27105	4251
	Yes	3109	1256
Manchester	No	48771	4548
	Yes	4570	1466
Birmingham	No	108723	4092
	Yes	11187	2063
Liverpool	No	49291	3640
	Yes	5161	1192
Sheffield	No	43579	2060
	Yes	5887	1638
Cornwall	No	32994	2191
	Yes	4852	2020
Spain			
Region	Assistance	Original	Filtered
Madrid	No	53218	2601
	Yes	1355	1286
Australia			
Region	Assistance	Original	Filtered
Victoria	No	4857	2065
	Yes	5609	2649

Cuadro 5.12

Luis: Para mí aquí irían ya los mapas del tercer paper.. (Mapas)

En la figura X se puede observar la distribución de clases para cada población (Poner aquí un histograma del desbalanceo de las clases y justo al lado la distribución una vez se aplica el filtrado de áreas.)

5.2.5. Discretización

Tabla de discretización.

UK discretización

Classification	Feature	Typing	Value
Location & Scale Accident	Latitude	Real Number	OSGR East Coordinate
	Longitude	Real Number	OSGR North Coordinate
	Road Class	0	Motorway
		1	A(M)
		2	A
		3	B
		4	C
		5	Unclassified
	Number of Vehicles	0-N	Depending on the number of vehicles involved
Driving Limitations	Road Surface	0	Dry
		1	Wet / Damp
		2	Snow
		3	Frost / Ice
		4	Flood
	Speed Limit	0-70	Depending on the speed limit (mph) of the road
Environmental	Weather Conditions	0	Fine without high winds
		1	Raining without high winds
		2	Snowing without high winds
		3	Fine with high winds
		4	Raining with high winds
		5	Snowing with high winds
		6	Fog or mist
	Lighting Conditions	7	Other
		0	Daylight: street lights present
		1	Darkness: no street lighting
		2	Darkness: street lights present and lit
		3	Darkness: street lights present but unlit
		4	Darkness: street lighting unknown
Temporary	Cosine Hour	Real Number	XX
	Sine Hour	Real Number	XX
	Day on Week	0-6	XX
	Week on Year	0-52	XX
Vehicle	Vehicle Type	0-17	Depending on the weight of the vehicle
	First Point of Impact	0	Did not impact
		1	Front
		2	Back
		3	Offside
		4	Nearside
		5	Unknown (self reported)
	Age of Vehicle	0-N	In order of vehicle age
	Casualty Class	0	Driver/Rider
		1	Passenger
		2	Pedestrian
Victim	Casualty Sex	0	Male
		1	Female
	Casualty Age	0	Younger than 18
		1	Between 18 and 25
		2	Between 25 and 65
		3	Older than 65

Cuadro 5.13: UK classification of variables.

Madrid discretización, in progress...

Classification	Feature	Typing	Value
Location & Scale Accident	Latitude	Real Number	Cartesian coordinate system
	Longitude	Real Number	Cartesian coordinate system
	District	0-X	District number (Anexo 1*)
	Number of Vehicles	0-N	Depending on the number of vehicles involved
Environmental	Weather Conditions	0	Fine without high winds
		1	Raining without high winds
		2	Snowing without high winds
		3	Fine with high winds
		4	Raining with high winds
		5	Snowing with high winds
		6	Fog or mist
		7	Other
Temporary	Cosine Hour	Real Number	XX
	Sine Hour	Real Number	XX
	Day on Week	0-6	XX
	Week on Year	0-52	XX
Vehicle	Vehicle Type	0-17	Depending on the weight of the vehicle
	First Point of Impact	1	Head-on - size collision
		2	Rear-end collision
		3	Side crash
		4	Collision again fixed obstacle
		5	Pile-up
		6	Hitting a pedestrian
		7	Head-on collision
		8	Other
		9	Leaving the road
		10	Vehicle rollover
		11	Hitting an animal
		12	Falling
Victim	Casualty Class	0	Driver/Rider
		1	Passenger
		2	Pedestrian
	Casualty Sex	0	Male
		1	Female
	Casualty Age	0	Younger than 18
		1	Between 18 and 25
		2	Between 25 and 65
		3	Older than 65
	Alcohol/Drugs Positive	0	No
		1	Yes

Cuadro 5.14: Victoria classification of variables.

Victoria discretización, in progres...

Classification	Feature	Typing	Value
Location & Scale Accident	Latitude	Real Number	OSGR East Coordinate
	Longitude	Real Number	OSGR North Coordinate
	Road Class	0	Motorway
		1	A(M)
		2	A
		3	B
		4	C
		5	Unclassified
	Number of Vehicles	0-N	Depending on the number of vehicles involved
Driving Limitations	Road Surface	0	Dry
		1	Wet / Damp
		2	Snow
		3	Frost / Ice
		4	Flood
	Speed Limit	0-70	Depending on the speed limit (mph) of the road
Environmental	Weather Conditions	0	Fine without high winds
		1	Raining without high winds
		2	Snowing without high winds
		3	Fine with high winds
		4	Raining with high winds
		5	Snowing with high winds
		6	Fog or mist
		7	Other
	Lighting Conditions	0	Daylight: street lights present
		1	Darkness: no street lighting
		2	Darkness: street lights present and lit
		3	Darkness: street lights present but unlit
		4	Darkness: street lighting unknown
Temporary	Cosine Hour	Real Number	XX
	Sine Hour	Real Number	XX
	Day on Week	0-6	XX
	Week on Year	0-52	XX
Vehicle	Vehicle Type	0-17	Depending on the weight of the vehicle
	First Point of Impact	0	Did not impact
		1	Front
		2	Back
		3	Offside
		4	Nearside
		5	Unknown (self reported)
	Age of Vehicle	0-N	In order of vehicle age
	Casualty Class	0	Driver/Rider
		1	Passenger
		2	Pedestrian
Victim	Casualty Sex	0	Male
		1	Female
	Casualty Age	0	Younger than 18
		1	Between 18 and 25
		2	Between 25 and 65
		3	Older than 65

Cuadro 5.15: UK classification of variables.

5.2.6. Transformación (Sin/Cos)

5.2.7. Resampling

En la tabla 5.16 se muestran los datos resultantes tras haber aplicado el proceso de resampling mediante la generación de datos sintéticos de SMOTE-II, conformando un dataset balanceado en el que se previene el riesgo de sesgo de datos por parte de la red.

**Luis: REPASAR PORQUE LOS NÚMEROS NO ME CUADRNAN,
HEMOS MANDADO EL PAPER 3 TAMBIÉN ASÍ.**

Data Distribution			
UK			
Region	Assistance	Filtered	Oversampled
Southwark	No	4251	2973
	Yes	1256	2973
Manchester	No	4548	3178
	Yes	1466	3178
Birmingham	No	4092	2838
	Yes	2063	2838
Liverpool	No	3640	2554
	Yes	1192	2554
Sheffield	No	2060	1447
	Yes	1638	1446
Cornwall	No	2191	2191
	Yes	2020	2191
Spain			
Region	Assistance	Filtered	Oversampled
Madrid	No	2601	2070
	Yes	1286	2070
Australia			
Region	Assistance	Filtered	Oversampled
Victoria	No	2065	1844
	Yes	2649	1845

Cuadro 5.16: REPASAR PORQUE LOS NÚMEROS NO ME CUADRNAN, HEMOS MANDADO EL PAPER 3 TAMBIÉN ASÍ.

5.2.8. Normalización

En la Tabla 5.17 se muestra un ejemplo de la aplicación de la normalización de datos en base a la técnica Z-Score, donde en la primera columna se observan los datos previos a esta normalización, mientras que la segunda columna contiene los valores de estas características normalizados.

Category	Feature	Original Value	Normalized Value
Location & Scale Accident	Easting	X	X
	Northing	X	X
	1st Road Class	X	X
	Number of Vehicles	X	X
Driving Limitations	Road Surface	X	X
	Speed Limit	X	X
Environmental	Weather Conditions	X	X
	Lighting Conditions	X	X
Temporary	Cosine Hour	X	X
	Sine Hour	X	X
	Day on Week	X	X
	Week on Year	X	X
Vehicle	Vehicle Type	X	X
	First Point of Impact	X	X
	Age of Vehicle	X	X
Victim	Casualty Class	X	X
	Casualty Sex	X	X
	Casualty Age	X	X

Cuadro 5.17: blabla

Una vez se disponen de los datos normalizados, estos ya son comparables y por tanto pueden ser utilizados para el entrenamiento de cualquier modelo que acepte valores numéricos como entrada.

5.2.9. Categorización

5.2.10. División Train-Val-Test

Sacar la tabla de comparación de registros para entrenamiento y test.

5.2.11. Cálculo de pesos

5.2.12. Feature Importance Algorithm

Tabla donde aparezcan los pesos de cada una de las características?

5.2.13. Algoritmo Genético

En la tabla 5.18 se muestran los hiperparámetros del algoritmo genético utilizados para optimizar el algoritmo XGBoost. Durante cada una de las generaciones, el límite máximo de individuos en la población es de 50 individuos (fila Population). Estos individuos en cada generación son evaluados mediante la función heurística a optimizar, la métrica F1-Score resultante del algoritmo XGBoost sobre el conjunto de test accidentes, es decir, en los accidentes no vistos durante el entrenamiento (fila Fitness Function). Una vez son evaluados, aquellos 10 mejores individuos son seleccionados para intercambiar su información, es decir, los padres que darán lugar a 10 nuevos individuos de cara a la próxima generación (fila Parents Mating). La mezcla de información entre padres se realiza mediante una estrategia de cruce mixta (fila Crossover Index), es decir, para cada par de padres se asigna un índice aleatorio en sobre el que se dividirán ambos individuos para luego combinar esta información en el descendiente resultante **esto se aplica para left join y right join**. Una vez se han dado lugar a los 10 nuevos individuos, el valor de cada una de las componentes que los conforman pueden ser modificados con una probabilidad del 40 % (fila Mutation Probability). Este proceso será repetido a lo largo de todas las 50 generaciones (fila Generations).

Hyperparameter	Value
Population	50
Parents Mating	10
Generations	50
Crossover Index	Random
Mutation Probability	0.4
Fitness Function	Boosting Algorithm F1-Score

Cuadro 5.18: Genetic algorithm hyperparameters setup.

La Tabla 5.19 muestran las variaciones en los valores máximos y mínimos permitidos para cada variable a optimizar mediante el algoritmo genético. La fila *Initial* de cada hiperparámetro muestra el rango de valores que cada individuo puede tomar cuando es inicializado. En la fila *Mutation* se observan, para cada hiperparámetro, los valores límites permitidos sobre los que los componentes de un sujeto pueden modificarse en el proceso de mutación, siempre y cuando dicho componente haya sufrido una mutación.

Hyperparameter	Limit	Min	Max
ETA	Initial	0.01	1
	Mutation	-0.2	0.2
Max Depth	Initial	1	25
	Mutation	-3	3
Min Child Weight	Initial	0.01	20
	Mutation	-4	4

Cuadro 5.19: Boosting models hyperparameters limits.

La configuración de estos parámetros, tanto los inciales como los de mutación se han escogido en base a resultados experimentales, en los que para cada

En la Tabla 5.20 se pueden observar los hiperparámetros óptimos resultantes de la ejecución del algoritmo genético, blabla...

Genetic Algorithm Result Values				
UK				
Region	ETA	Maximum Depth	Minimum Children	Weight
Southwark	0.62	13		0.01
Manchester	0.01	1		0.01
Birmingham	0.43	17		0.01
Liverpool	0.83	12		0.01
Sheffield	0.59	20		0.61
Cornwall	0.85	17		0.01
Spain				
Region	ETA	Maximum Depth	Minimum Children	Weight
Madrid	0.01	1		0.01
Australia				
Region	ETA	Maximum Depth	Minimum Children	Weight
Victoria	0.6	25		0.01

Cuadro 5.20: Resulting boosting model hyperparameters after executing the genetic algorithm.

5.2.14. Pesos de categorías**5.2.15. Construcción de matrices****5.2.16. Métricas de evaluación**

En la figura 5.11 se muestra la distribución de los accidentes original y la distribución resultante tras aplicar el filtrado por áreas, aquellos accidentes que no requieren de asistencia se encuentran representados en verde, mientras que aquellos que sí se encuentran representados en rojo. Como puede observarse en la figura 5.11(a), la concentración de los accidentes se ve distribuida principalmente por aquellas zonas más próximas al núcleo urbano de Madrid, contando además con una amplia concentración en aquellas carreteras que pertenecen a las principales arterias de comunicación de Madrid. La figura 5.11(b) muestra la distribución de accidentes resultante tras haber aplicado el proceso de filtrado de áreas. Este proceso de reducción de datos permite una simplificación de la información sin que esto represente una pérdida en sí misma, de ya que se busca equilibrar el número de accidentes necesarios de asistencia y los que no, manteniendo únicamente la información imprescindible para ello.

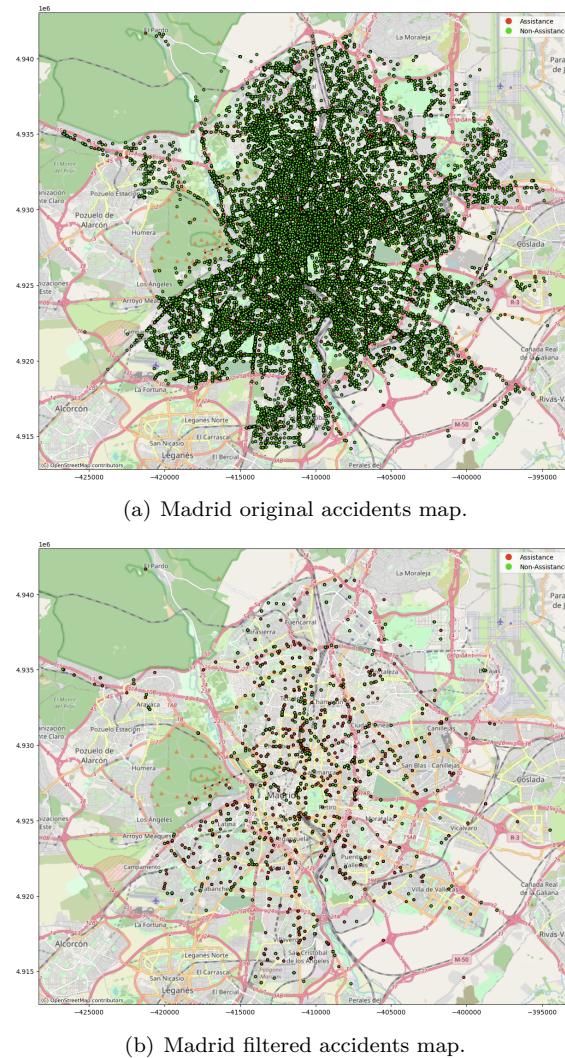


Figura 5.11: Madrid original/filtered accidents map.

En la Figura 5.12 se muestra la evolución de la función a optimizar (F1-Score) a lo largo de las 50 épocas para las que se ha entrenado el modelo GTAAF en la ciudad de Madrid. Se observa cómo el F1-Score para el conjunto de entrenamiento sufre una evolución importante durante las diez primeras épocas, después de las cuales sigue aumentando en menor medida. Por otra parte, la métrica sobre el conjunto de validación sufre una evolución más lenta, hasta aproximadamente la época 30 no se ve una clara evolución en la generalización del modelo sobre datos que nunca ha visto.

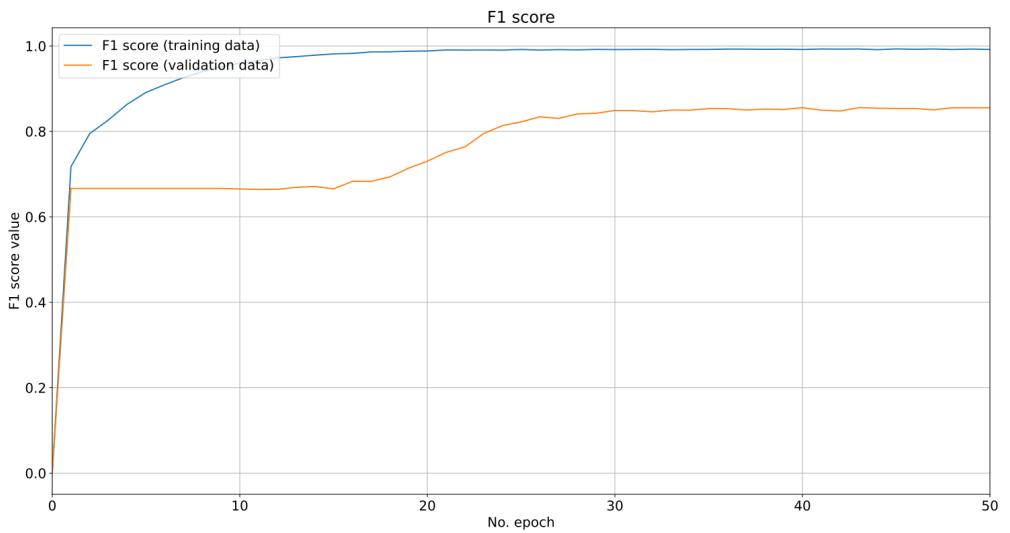


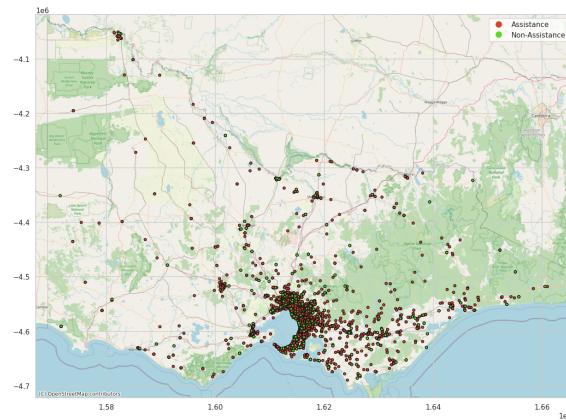
Figura 5.12: Evolution of F1-Score Madrid.

En la tabla 5.21 se observan los resultados de la métrica F1-Score de la predicción de la severidad de los accidentes de cada uno de los modelos sobre el conjunto de test de la ciudad de Madrid. Como se puede comprobar, el valor más alto lo ofrece el nuevo modelo GTAAF propuesto, llegando a mejorar en un 3,9 % al siguiente mejor modelo, el SVC sobre los accidentes Slight, mientras que la mejora sobre los accidentes Assistance se mide en un 5,7 % sobre el siguiente modelo que mejor métricas ofrece, el SVC. Con estos resultados puede interpretar que el nuevo modelo GTAAF propuesto es capaz de generalizar mejor en la predicción de la severidad de nuevos accidentes que no ha visto previamente sobre la ciudad de Madrid.

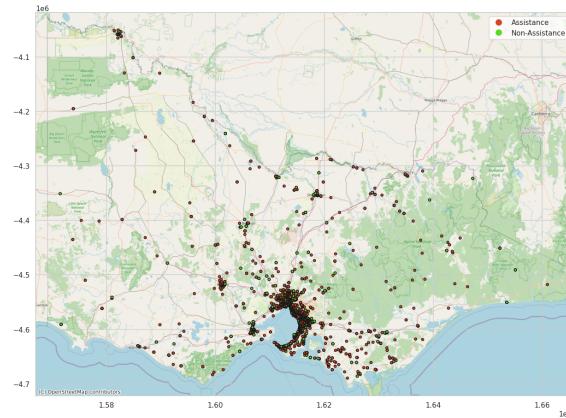
		Spain region F1-Score
Model	Assistance	Madrid
NB	No	0.729
	Yes	0.621
SVC	No	0.862
	Yes	0.748
KNN	No	0.739
	Yes	0.634
RF	No	0.744
	Yes	0.643
LR	No	0.750
	Yes	0.623
MLP	No	0.856
	Yes	0.724
GTAAF	No	0.894
	Yes	0.798

Cuadro 5.21: F1-Scores by Accident Class on Madrid (Spain).

En el segundo caso, tenemos una región dispersa, el estado de Victoria (Australia). Victoria, un estado en Australia, abarca una región diversa con ciudades bulliciosas como Melbourne, situada a lo largo de la costa sureste, conocida por su densidad de población moderada a alta y una mezcla de vitalidad urbana. En la figura 5.13 se muestra la distribución de accidentes sobre la población de Victoria, aquellos Non-Assistance se encuentran marcados en verde mientras que aquellos tipo Assistance se encuentran representados en rojo. Como se puede observar en la figura 5.13(a) gran parte de la concentración de los accidentes se encuentra sobre la ciudad de Melbourne y sus núcleos urbanos próximos (como Ballarat al oeste, Shepparton al norte o Traralgon al este), al igual que en las carreteras que interconectan estas poblaciones. Al ser un estado extenso, el filtrado de áreas es más amplio, lo que resulta una variante respecto a ciudades de mayor concentración. En la Figura 5.13(b) se observa la distribución de accidentes resultante tras aplicar el proceso de filtrado, donde aquellas zonas que presentan más accidentes necesarios de asistencia son en las grandes poblaciones y en las interconexiones entre estas.



(a) Victoria original accidents map.



(b) Victoria filtered accidents map.

Figura 5.13: Victoria original/filtered accidents map.

En la Figura 5.14 se muestran las funciones F1-Score sobre los datos de entrenamiento y validación para la región de Victoria. Esta métrica sobre el conjunto de entrenamiento muestra una curva de aprendizaje más lenta respecto a la ciudad de Madrid, lo cual es comprensible ya que existe más variabilidad de datos en esta región al ser mucho más extensa que la anterior. La función de validación presenta más variaciones a lo largo del aprendizaje, llegando a su máximo aproximadamente en la época 45.

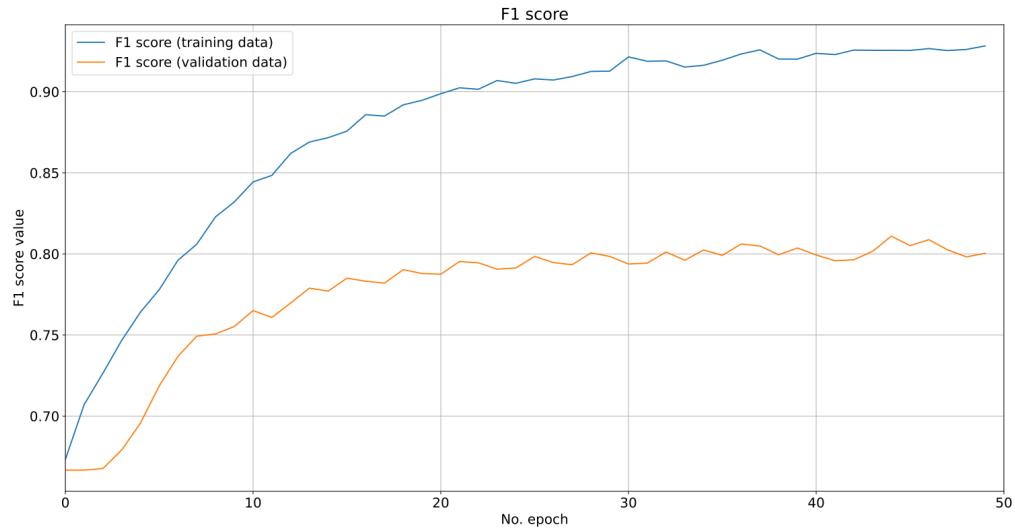


Figura 5.14: Evolution of F1-Score Victoria.

En la Tabla 5.22, se presentan los resultados del F1-Score obtenidos por cada uno de los modelos para ambos tipos de clasificación de accidentes. Específicamente, se observa que para la población de Victoria, el nuevo modelo GTAAF propuesto logra una mejora del 6.5 % en comparación con el siguiente mejor modelo, el SVC, para accidentes de No Asistencia. Por otro lado, en lo que respecta a accidentes de tipo Asistencia, hay una mejora del 9 % en comparación con el MLP. Estos resultados reflejan una mejora significativa en la capacidad de generalización del nuevo modelo propuesto.

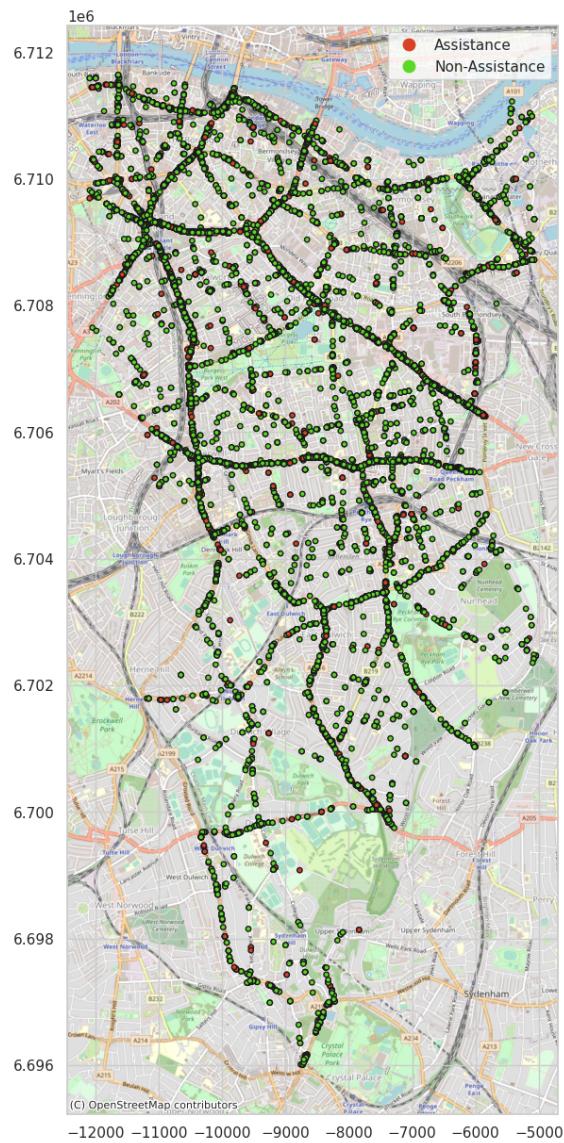
Model	Assistance	Australia region F1-Score
		Victoria
NB	No	0.635
	Yes	0.476
SVC	No	0.662
	Yes	0.679
KNN	No	0.654
	Yes	0.616
RF	No	0.647
	Yes	0.364
LR	No	0.612
	Yes	0.630
MLP	No	0.635
	Yes	0.694
GTAAF	No	0.727
	Yes	0.784

Cuadro 5.22: F1-Scores by Accident Class on Victoria (Australia).

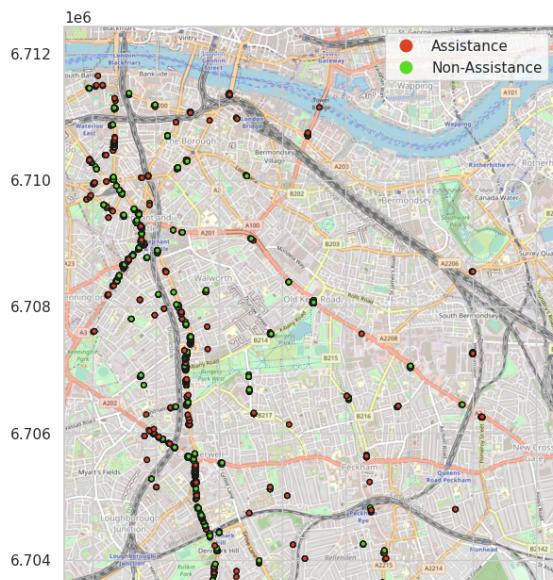
En la Tabla 5.22 se muestran los resultados F1-Score obtenidos de cada uno de los modelos respecto a ambos tipos de clasificación de accidentes. Concretamente se observa que para la ciudad de Victoria el nuevo modelo GTAAF propuesto obtiene una mejora respecto al siguiente mejor modelo, el SVC, para los accidentes Slight del 6,5 %. Por otra parte, en lo que respecta a los accidentes tipo Assistance se obtiene una mejora del 9 % respecto al MLP. Estos resultados reflejan una mejora de generalización significativa del nuevo modelo propuesto.

Southwark

Southwark es un distrito de Londres situado en la orilla sur del río Támesis, con una alta densidad de población. En la Figura 5.15 se observa la distribución de los accidentes a lo largo del municipio de Southwark. Analizando los accidentes del conjunto de datos original, Figura 5.15(a), se observa que estos se producen a lo largo de las distintas vías que conectan el municipio con el centro neurálgico de la ciudad, hecho habitual al conectar zonas menos pobladas con lugares de trabajo y de ocio, mientras que la minoría de ellos se presentan en las calles aledañas. Observando la distribución de accidentes tras el proceso de filtrado por áreas (Figura 5.15(b)) se acentúa este hecho, donde se observan que aquellos principales accidentes Assistance se producen en estas vías. Por otra parte se muestra una concentración minoritaria de este tipo de accidentes en la zona de Dulwich (sur), en la intersección de la circunvalación S Circular red con la carretera que dirige al centro del municipio.



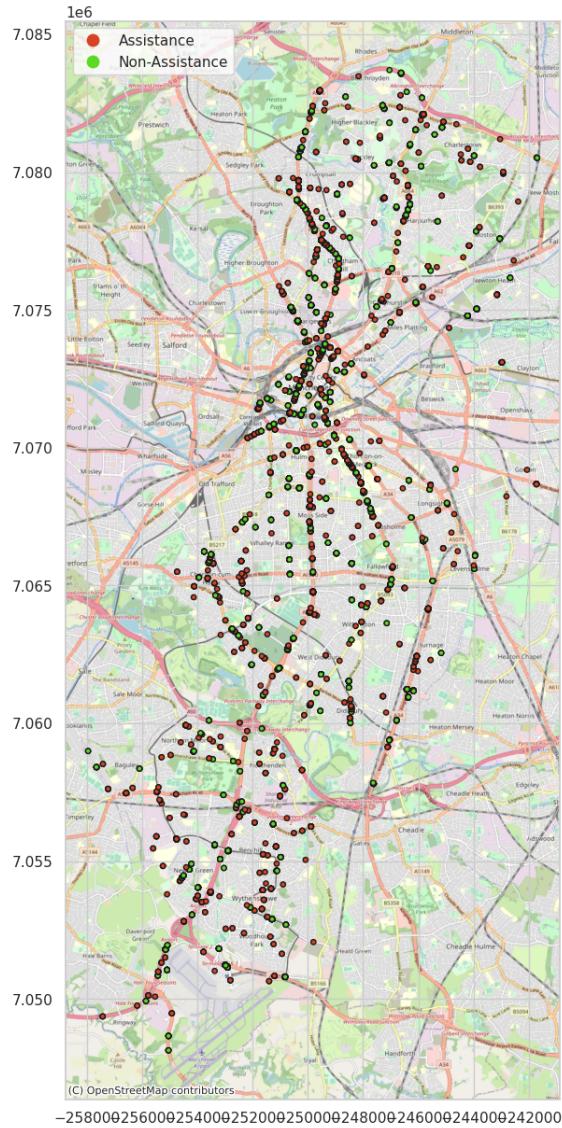
(a) Southwark original accidents map.



Manchester

Manchester, ubicada en el norte de Inglaterra, es una gran ciudad conocida por su legado industrial y su alta densidad de población. En la Figure 5.16 se muestra la distribución de los accidentes de Manchester. Atendiendo a la distribución de accidentes original, en la Figura ??, como es habitual en cualquier población se aprecia una concentración de accidentes importante en la zona central de la ciudad, siendo también considerable en el área de Longsight. Por otra parte, las principales vías que comunican las periferias urbanas (norte) con el centro de la ciudad también presentan una concentración mayor de accidentes, lo que puede deberse a desplazamientos por trabajo. Por otra parte, la carretera de Wythenshawe, cercano a Sale Water Park (Sur), también presenta una concentración elevada de accidentes, motivados por los desplazamientos de ocio y de trabajo. En la figura 5.16(b) se observa la localización de los accidentes una vez se ha aplicado el proceso de filtrado por áreas, donde se vislumbra que gran parte de los accidentes Assistance se distribuyen a lo largo de las carreteras que comunican hacia el centro de la ciudad.

(a) Manchester original accidents map.



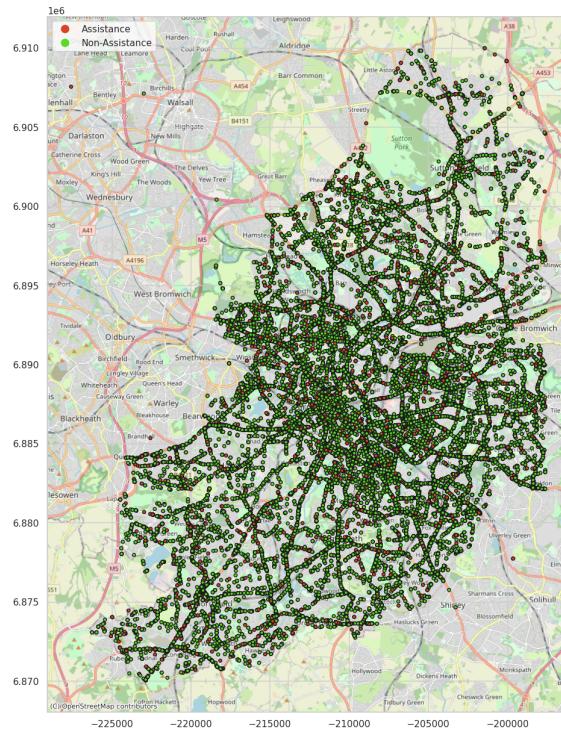
(b) Manchester filtered accidents map.

Figura 5.16: Manchester original/filtered accidents map.

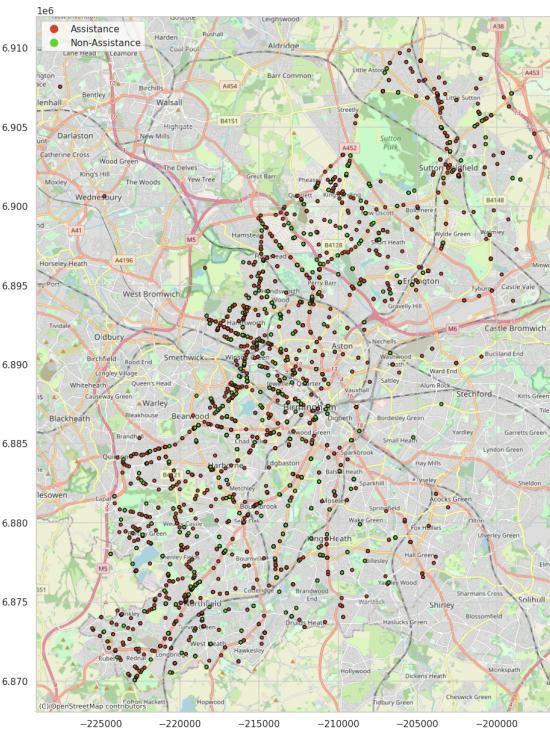
Birmingham

Birmingham, la segunda ciudad más grande de Inglaterra, se extiende por

West Midlands con un paisaje urbano diverso y una densidad de población considerable, famosa por su historia industrial y su vitalidad cultural. En la Figura 5.17 se muestra la distribución de los accidentes de Birmingham, tanto los originales como los resultantes una vez aplicado el proceso de filtrado. Como se puede observar en los accidentes originales en la Figura 5.17(a) se aprecia que gran parte de los accidentes se concentran en la zona centro de la ciudad, una tendencia normal debido a que es el principal foco de actividad de las ciudades. Mientras que los accidentes se van dispersando a medida que distan de este punto. Se aprecian ligeras agrupaciones de accidentes a lo largo de las zonas de incorporaciones a las principales arterias de la ciudad, como es al este, el caos de Handsworth. Por otra parte, en la figura 5.17(b) se muestran los accidentes una vez se ha aplicado el proceso de filtrado por áreas. Como se puede observar, la información ha sido resumida sin dar lugar a pérdidas en el valor de la misma. Se vislumbran ciertas zonas más conflictivas donde se producen accidentes más importantes, como es el caso de la carretera Holyhead Rd de entrada a la ciudad o en Northfield.



(a) Birmingham original accidents map.

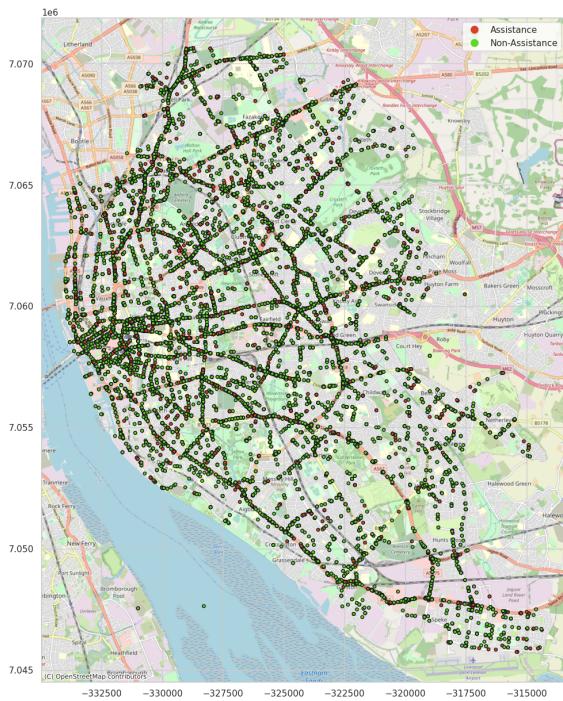


(b) Birmingham filtered accidents map.

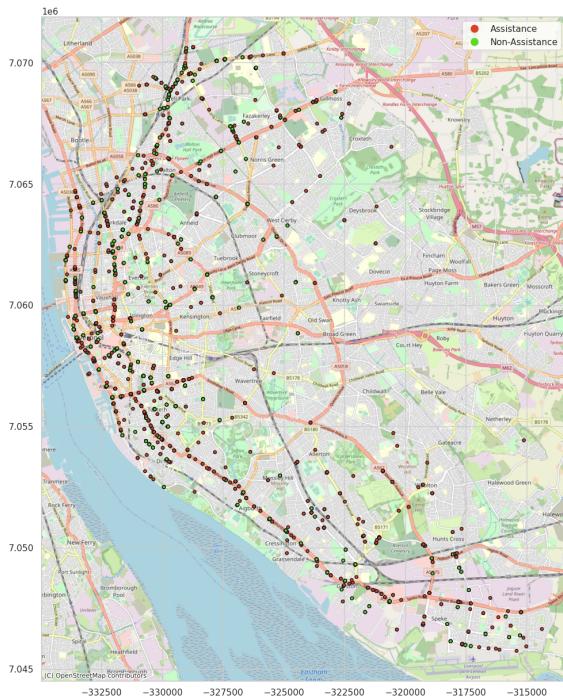
Figura 5.17: Birmingham original/filtered accidents map.

Liverpool

Liverpool, ubicada a lo largo del río Mersey en el noroeste de Inglaterra, prospera como una ciudad marítima con una rica historia, profundidad cultural y una densidad de población significativa, reconocida por su encanto en el frente marítimo y su legado musical. En la Figura 5.18 se muestra la comparativa de la distribución de accidentes originales del dataset y filtrados para la ciudad de Liverpool. En la Figura 5.18(a) se aprecian accidentes concentrados en la zona centro de la ciudad, como viene siendo habitual, además de a lo largo de las circunvalaciones que la rodean. En la Figura 5.18(b), después del proceso de filtrado, se aprecia que gran parte de los accidentes Assistance se producen a lo largo de Strand Street (desde el sur hasta el oeste), convergiendo ambas direcciones en el centro neuralgico. Por otra parte se visualiza otra concentración en la carretera que conecta la localidad de Ormskirk con el centro (noroeste), una de las principales vías de conexión.



(a) Liverpool original accidents map.

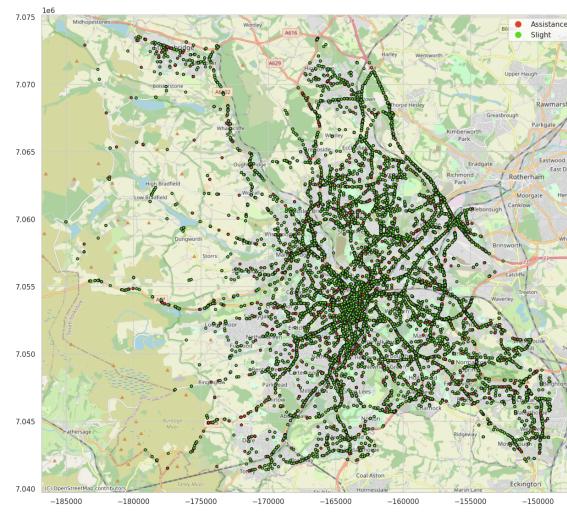


(b) Liverpool filtered accidents map.

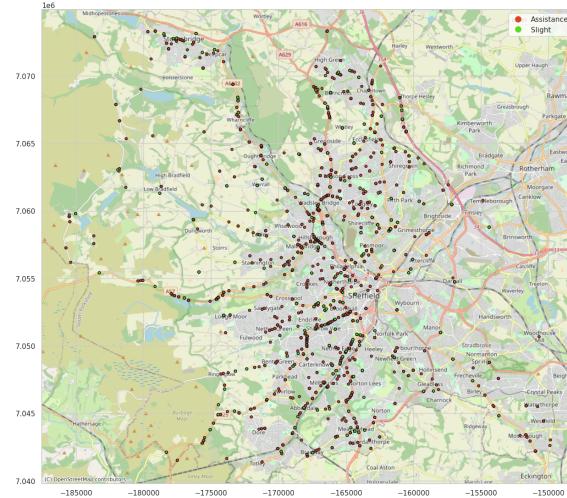
Figura 5.18: Liverpool original/filtered accidents map.

Sheffield

Sheffield, ubicada en South Yorkshire, presume de un patrimonio industrial y paisajes pintorescos, con una densidad de población intermedia. En la Figura 5.19 se muestra la distribución de accidentes para la ciudad de Sheffield, tanto la original como la resultante tras la etapa de filtrado. En la Figura 5.19(a) se pueden apreciar distintas concentraciones en zonas estratégicas. Como suele ser habitual, el núcleo urbano es un centro de mayor densidad de incidentes, mientras que en las intersecciones que conectan la ciudad de Sheffield y la de Rotherham (cruces de Tinsley Viaduct con Meadow Bank Road y la A6178, al noreste de Sheffield). También se aprecian concentraciones en los suburbios de Wadsley Bridge y Malin Bridge, periferias de la ciudad, además de alrededor de todas las vías principales que conectan con el centro. Por otra parte, en la figura 5.19(b) se muestran los accidentes una vez se ha realizado el proceso de filtrado, donde se aprecia que aquellos que han requerido de asistencia normalmente se presentan en las principales arterias, donde se circula a una mayor velocidad.



(a) Sheffield original accidents map.



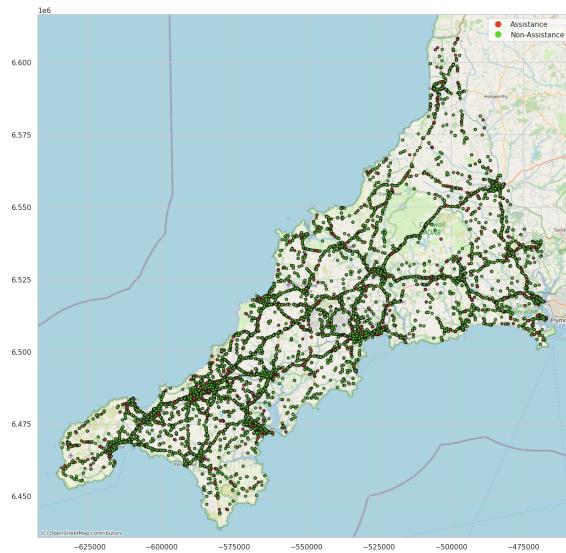
(b) Sheffield filtered accidents map.

Figura 5.19: Sheffield original/filtered accidents map.

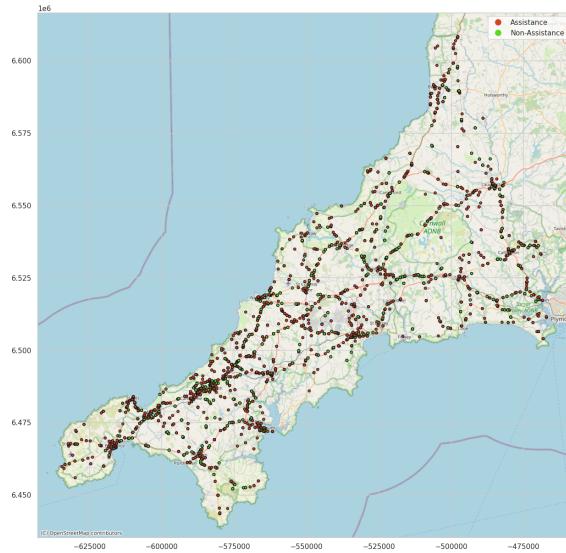
Cornwall

Cornualles, situada en la parte suroeste de Inglaterra con sus apacibles paisajes, encantadores pueblos costeros y extensiones rurales, fomenta un entorno tranquilo alejado de los núcleos de alta densidad de población. En la Figura 5.20 se muestran de nuevo los accidentes originales de dataset y los que resultan tras aplicar el proceso de filtrado sobre el condado de Cornwall. En la

Figura 5.20(a) las principales concentraciones de accidentes se encuentran distribuidas a lo largo de las distintas ciudades del condado. La mayoría de estos se encuentran divididos en dos regiones claramente definidas, la primera de ellas entre las vías que conectan las localidades de Camborne y Redruth (suroeste de Cornwall), y el área comprendida entre St Austell, Duporth, Carlyon Bay y Par, este del condado. No obstante, el resto de regiones también presentan una concentración considerable, como es el caso de la ciudad de Falmouth (sureste), las localidades de Penzance y Hayle (suroeste), en la ciudad de Newquay y sus alrededores (oeste), Bodmin (centro) y Launceston (norte). De nuevo, en este caso, se demuestra que la mayor frecuencia de accidentes se presenta entre los principales núcleos de población y las carreteras que los interconectan, debido a que las grandes ciudades implican más movimientos de vehículos. Atendiendo a la Figura 5.20(b) se observa que la ocurrencia de accidentes necesarios de asistencia, una vez aplicado el proceso de filtrado, tiene la misma tendencia que el expuesto para el conjunto de datos original, distribuyéndose a lo largo de las principales carreteras del condado Cornwall y concentrándose más en los núcleos de población.

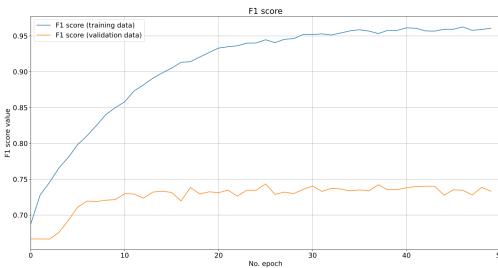


(a) Cornwall original accidents map.

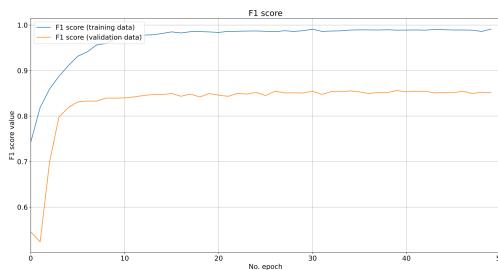


(b) Cornwall filtered accidents map.

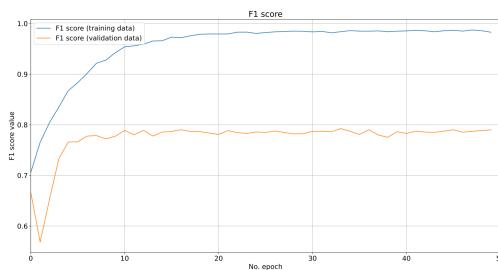
Figura 5.20: Cornwall original/filtered accidents map.



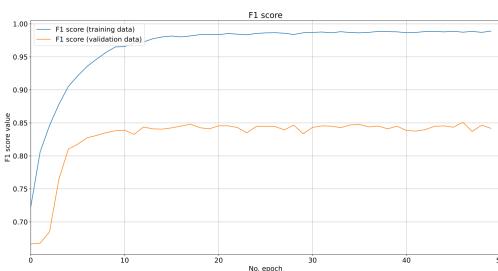
(a) Training 2D-CNN Southwark



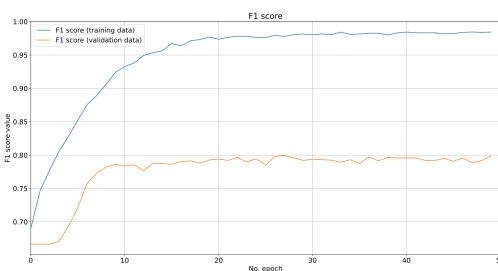
(b) Training 2D-CNN Manchester



(c) Training 2D-CNN Birmingham



(d) Training 2D-CNN Liverpool



(e) Training 2D-CNN Sheffield

En la Tabla 5.23 se muestra el valor F1-Score para cada una de las ciudades de cada modelo sobre el conjunto de test. Como se puede observar, el nuevo modelo propuesto GTAAF es el que mejor métricas ofrece en comparación al resto, obteniendo la mayor diferencia con respecto a su sucesor en los accidentes Slight, el MLP, para la ciudad de Manchester de un 5,33 %, mientras que la mayor diferencia para los accidentes tipo Assistance es de un 13,8 % en la ciudad de Southwark respecto al siguiente mejor modelo, el MLP. La siguiente mayor diferencia se presenta entre el modelo GTAAF se presenta en la ciudad de Southwark para la clase Slight, con un incremento del 4,8 % respecto al siguiente mejor modelo MLP, mientras que para la clase Assistance ésta se presenta en la ciudad de Liverpool con respecto al modelo MLP, llegando a un 13,2 %. Observando los resultados de la tabla, se aprecia que el mayor incremento del rendimiento con respecto al resto de modelos se presenta sobre la clase Assistance, obteniendo de media una mejora de 9,21 % sobre todas las ciudades, mientras que el incremento de rendimiento se acentúa menos en la Slight, ya que el resto de modelos ofrecen unas métricas más altasW, siendo la mejora de un 3,93 % de media. Estos resultados reflejan una mejor generalización del modelo propuesto en comparación al resto de modelos estudiados para cada una de las ciudades de Reino Unido.

		UK areas F1-Score					
Model	Assistance	Southwark	Manchester	Birmingham	Liverpool	Sheffield	Cornwall
NB	No	0.504	0.675	0.567	0.560	0.620	0.653
	Yes	0.400	0.482	0.558	0.417	0.669	0.484
SVC	No	0.826	0.845	0.812	0.865	0.809	0.702
	Yes	0.599	0.624	0.673	0.630	0.773	0.626
KNN	No	0.652	0.723	0.747	0.746	0.754	0.656
	Yes	0.469	0.510	0.609	0.519	0.676	0.559
RF	No	0.561	0.118	0.303	0.742	0.313	0.711
	Yes	0.430	0.379	0.509	0.504	0.585	0.581
LR	No	0.711	0.800	0.761	0.806	0.733	0.630
	Yes	0.415	0.540	0.604	0.530	0.652	0.598
MLP	No	0.916	0.857	0.819	0.910	0.853	0.709
	Yes	0.743	0.632	0.662	0.721	0.810	0.671
GTAAF	No	0.964	0.924	0.858	0.956	0.918	0.722
	Yes	0.881	0.762	0.711	0.853	0.889	0.707

Cuadro 5.23: F1-Scores comparison by traffic accident assistance on six UK areas.

La Figura 5.22 muestra a modo de comparativa el rendimiento del nuevo modelo GTAAF propuesto en los accidentes Slight para cada una de las poblaciones estudiadas respecto al resto de modelos del estado del arte con los que se ha experimentado en esta investigación. Se aprecia un incremento de

rendimiento independientemente de las características individuales en todas las poblaciones respecto al resto de modelos estudiados, siendo el mayor incremento en la población de Victoria con un incremento del 6.5 % respecto al siguiente mejor modelo, el SVC.

F1-Score by region (Non-Assistance accidents)

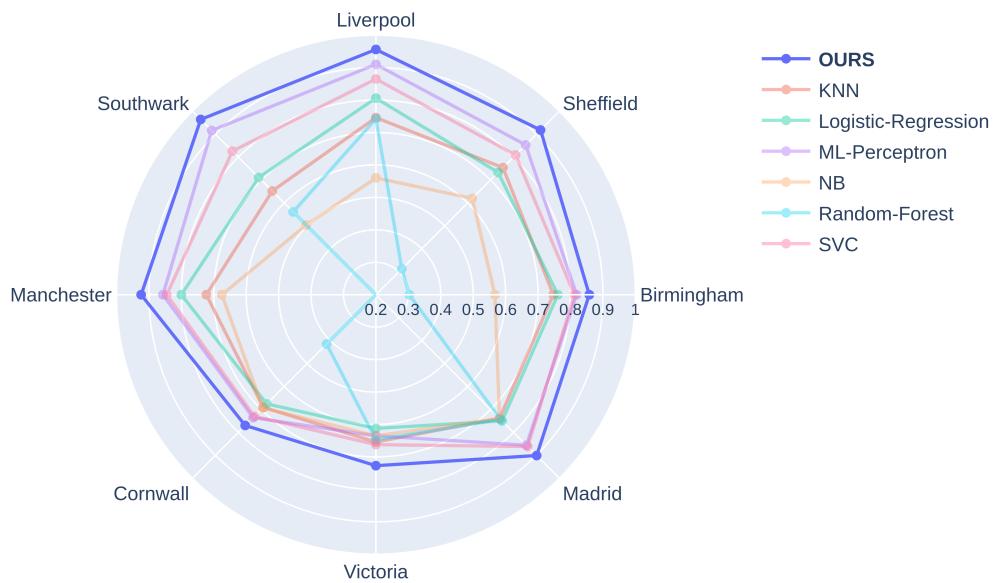
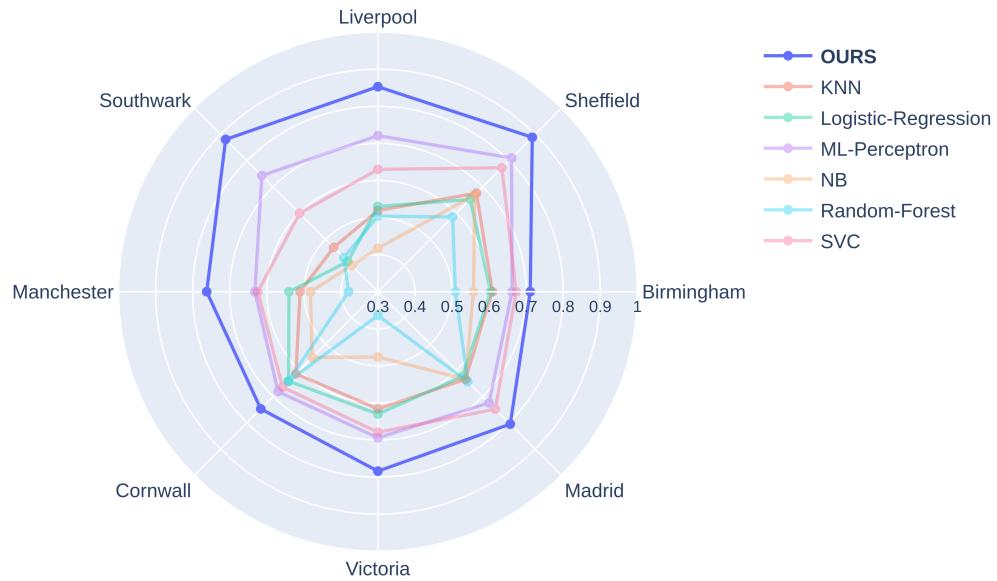


Figura 5.22: F1-Scores Comparison for Non-Assistance Accidents.

La Figura 5.23 muestra la comparativa del rendimiento basado en el F1-Score de los modelos para cada una de las ciudades en los accidentes Assistance. En esta gráfica se puede observar una diferencia considerablemente mayor del nuevo modelo GTAAF propuesto respecto al resto en comparación con los accidentes Slight. La mayor diferencia de mejora de este modelo GTAAF se presenta en la ciudad de Southwark, con un incremento del 13.8 % respecto al siguiente mejor modelo sobre esta población, el MLP.

F1-Score by region (Assistance accidents)

**Figura 5.23:** F1-Scores Comparison for Assistance Accidents.

5.3. Pruebas de estrés

En esta sección se realizarán distintas pruebas de estrés. El objetivo de estas pruebas es medir el rendimiento de la metodología y el modelo propuesto en casos extremos utilizando como base los conjuntos de datos expuestos en esta tesis para tener una aproximación del rendimiento del modelo en otros conjuntos de datos que no dispongan de las características descritas en este documento. Para ello se realizarán tres experimentos para cada conjunto de datos que consistirán en eliminar aquellas características de mayor y menor importancia de forma independiente, y, en un experimento posterior, se eliminarán ambas conjuntamente con el objetivo de medir el rendimiento ante la falta de características más y menos influyentes en futuros conjuntos de datos. La evaluación de la importancia de las características viene dada por el peso asignado a cada una de estas mediante el algoritmo genético.

En la tabla 5.24 ...

Model	Assistance	Cornwall		
		Lowest	Highest	Both
NB	No	0.668	0.597	0.592
	Yes	0.490	0.535	0.519
SVC	No	0.710	0.631	0.646
	Yes	0.628	0.626	0.620
KNN	No	0.671	0.601	0.637
	Yes	0.571	0.532	0.559
RF	No	0.719	0.498	0.514
	Yes	0.603	0.638	0.644
LR	No	0.670	0.575	0.567
	Yes	0.626	0.585	0.580
MLP	No	0.724	0.652	0.680
	Yes	0.695	0.654	0.685
CNN2D	No	0.768	0.736	0.792
	Yes	0.766	0.736	0.787

Cuadro 5.24: F1-Scores comparison with features loss in Madrid dataset. In bold the best result (our model)

Model	Assistance	Victoria		
		Lowest	Highest	Both
NB	No	0.639	0.613	0.607
	Yes	0.465	0.553	0.572
SVC	No	0.653	0.638	0.664
	Yes	0.657	0.650	0.676
KNN	No	0.625	0.627	0.638
	Yes	0.540	0.562	0.566
RF	No	0.630	0.630	0.621
	Yes	0.248	0.161	0.071
LR	No	0.598	0.574	0.599
	Yes	0.609	0.637	0.646
MLP	No	0.635	0.636	0.654
	Yes	0.693	0.686	0.692
CNN2D	No	0.732	0.720	0.778
	Yes	0.780	0.793	0.814

Cuadro 5.25: F1-Scores comparison with features loss in Victoria dataset. In bold the best result (our model)

En este experimento es necesario destacar un resultado: en nuestra propuesta, el modelo GTAAF, existe una gran mejora sobre los resultados del mismo modelo con todas las características. En contraste, hay un gran deterioro en los resultados de los otros modelos con los que se compara. En otras palabras, la diferencia en el puntaje F1 entre GTAAF y los otros modelos aumenta.

Esta circunstancia sugiere que nuestro modelo se ve afectado por las características extremas, donde el modelo de boosting y el algoritmo genético favorecen y desfavorecen las características más y menos relevantes. Este no es el caso para los otros algoritmos, donde todas las características son individualizadas.

En la Figura 5.24 podemos observar la diferencia de los algoritmos con y sin pérdida de características (Tablas 5.23 y 5.22 versus Tablas ?? y 5.25). Mostramos cómo nuestro modelo mejora sus propios resultados en todos los casos.

Por ejemplo, en Cornwall, obtenemos una mejora en nuestro modelo del 4.8% y 5.9% en No Assistance y Assistance (eliminando la peor característica, ver la barra azul en la Figura 5.24-arriba), 1.4% y 2.9% (eliminando la mejor característica, ver la barra verde en la Figura 5.24-arriba) y 7% y 8% (eliminando ambas características, ver la barra gris en la Figura 5.24-arriba), respectivamente.

Evaluando un área más dispersa como Victoria, los resultados son similares pero con una mejora menor: 0.5% y 0.4% (ver la barra azul en la Figura 5.24-abajo), -0.7% y 0.9% (ver la barra verde en la Figura 5.24-abajo), y 5.1% y 3% (ver la barra gris en la Figura 5.24-abajo), respectivamente. Esto indicaría un efecto menor de los valores extremos en la ponderación del algoritmo genético si el área es más dispersa.

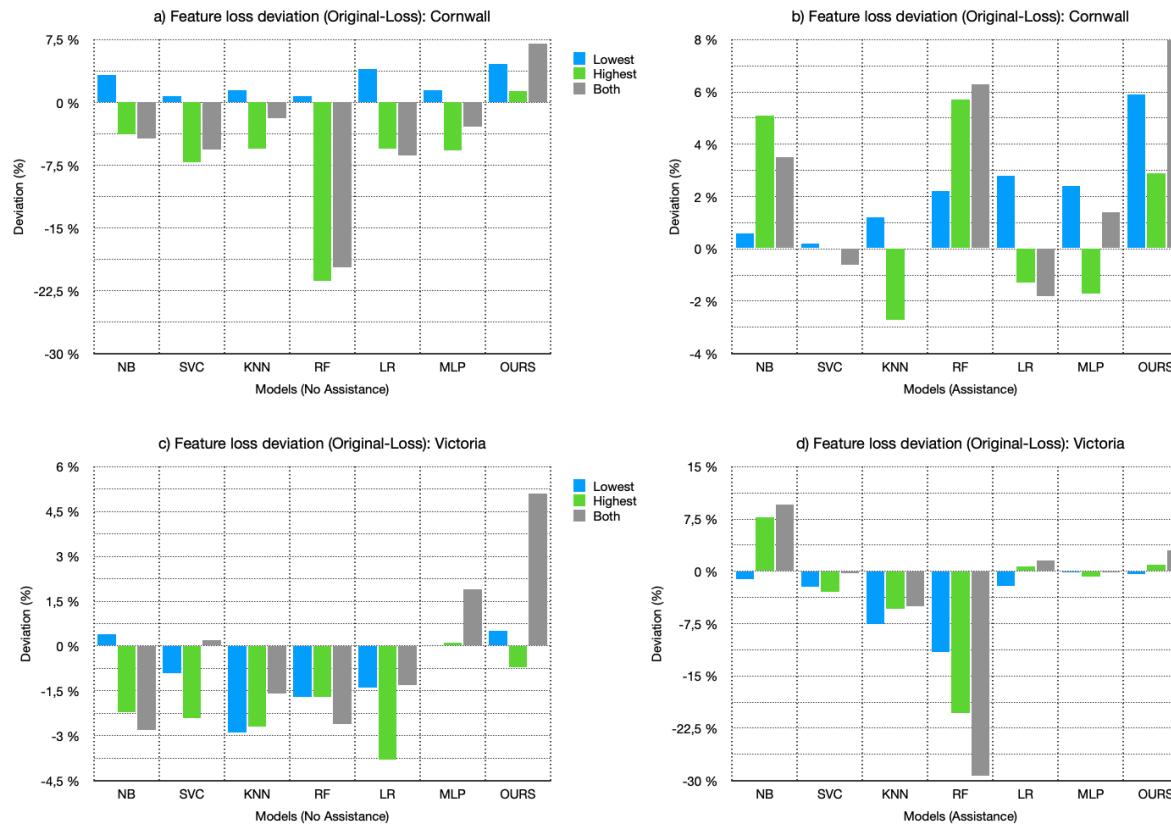


Figura 5.24: Comparación de pérdida de características. Las barras representan la diferencia entre los resultados con todas las características y los resultados sin características extremas: en azul sin la característica más baja, en verde sin la característica más alta y en gris sin ambas características extremas.

Capítulo 6

Conclusiones

En esta tesis se ha propuesto un nuevo modelo que evalúa la necesidad de asistencia médica en los accidentes de tráfico. Esta funcionalidad es extremadamente importante para priorizar la asignación de recursos médicos una vez se conocen las características del accidente, de tal forma que se puedan minimizar las consecuencias físicas a corto y largo plazo de las víctimas. Para ello se ha propuesto una metodología que transforma las características que describen los accidentes, mediante categorizaciones, para alimentar a nuestro modelo convolucional X. Como se ha demostrado en su evaluación, los resultados no solo mejoran ampliamente al estado del arte (con valores de hasta el 13.8 %), sino que la categorización propuesta ha demostrado ser muy robusta respecto a la individualización de características de los demás modelos. Además, nuestro modelo ha mostrado un gran rendimiento en distintos contextos, concretamente en distintos datasets de 8 poblaciones de diferentes densidades de población, siendo relevante este dato en la correlación que tiene con el número de accidentes producidos.

Además, con el objetivo de proponer un modelo general que pueda ser aplicado a nuevas poblaciones que no dispongan de la misma información que los datasets presentados en este artículo, (debido principalmente a la dificultad inherente de recogida de datos específicos, como controles de alcohol y drogas, u otras características cuya obtención esté relacionada con la condición económica de la población), se ha analizado la robustez del modelo eliminando características, excluyendo características de mayor y menor impacto que han resultado del algoritmo genético, obteniendo resultados incluso mejores en nuestro modelo que hace indicar la sensibilidad que tiene éste respecto a estos valores. Como trabajo futuro se analizará cómo reducir esta sensibilidad.

Capítulo 7

Publications

Capítulo 8

Anexos

Madrid paper I discretization:

Features	Feature		Typing	Type of Road!Type of Roadpt<	Roadpt<	Typing
	Type of Road	Type of Road!				
'Severity!Severity!Severitypt<	0: Slight (1, 2, 5, 6, 7)					10: Bridge
	1: Severe (3)					11: Square
	2: Fatal (4)					12: Bouleva
ime!ime!Timept<	1: Night (6 PM - 6 AM)					13: Crossin
	2: Day (6 AM - 6 PM)					14: Roadwa
istrict!district!Districtpt<	Based on order of appearance					15: Road
!!Xpt<	UTM X Coordinate position					16: Avenue
!!Ypt<	UTM Y Coordinate position					17: Highwa
ype of Accident!type of Accident!Type of Accidentpt<	1: Head-on collision					18: Street
	2: Rear-end collision					1: Sunny
	3: Side crash					2: Cloudy
	4: Collision again fixed obstacle					3: Light rai
	5: Pile-up					4: Heavy ra
	6: Hitting a pedestrian					5: Hail
	7: Head-on collision					6: Snowing
	8: Other					7: Unknown
	vehicle!Vehicle!Vehiclep<					Based on o
	9: Leaving the road					
	erson!erson!Personpt<					1: Driver
	10: Vehicle rollover					2: Passenge
	11: Hitting an animal					3: Pedestria
	12: Falling					
	gelge!Agept<					
ype of Road!type of Road!Type of Roadpt<	1: Parking					1: Under 18
	2: Airport					2: From 18
	3: Park					3: From 25
	4: Tunnel					4: Over 65
	5: Industrial state					5: Unknown
	ender!ender!Genderpt<					
	6: Track					1: Male
	7: Round					2: Female
	8: Roundabout					3: Unknown
	lcohol of Drugs!lcohol or Drugs!Alcohol or Drugspt<					
	9: Gate					1: Yes
						2: Not

Cuadro 8.1: Numerical assignment of the dataset variables.