



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

1. Introduction	11
1.1. Motivación	11
1.2. Objetivos	11
2. Estado del arte	13
2.1. ¿Cómo medir la gravedad de un accidente?	13
2.2. ¿Cómo predecir la gravedad de un accidente de tráfico?	14
3. Background	17
3.1. Modelos estado del arte	17
3.2. Algoritmos CNN	20
3.3. Algoritmos de construcción de matrices	25
3.4. Métodos de Remuestreo	26
3.5. Métodos de optimización de hiperparámetros	27
3.6. Algoritmos de medición de importancia de características	28
3.7. Algoritmos Genéticos	30
3.8. Medidas de evaluación de una red neuronal	32
4. Construcción de un modelo general de predicción de la gravedad de un accidente de tráfico	35
4.1. Modelo preliminar	35
4.2. Modelo GTAAF	40
4.2.1. Preprocesamiento	42
4.2.1.1. Limpieza	43
4.2.1.2. Discretización	44

4.2.1.3. Transformación (Sin/Cos)	44
4.2.1.4. Fitrado de Áreas	46
4.2.1.5. Normalización	47
4.2.1.6. División Train-Val-Test	48
4.2.1.7. Resampling	49
4.2.2. Postprocesamiento	49
4.2.2.1. Construcción de Matrices	51
4.2.2.2. Feature Importance Algorithm	52
4.2.2.3. Algoritmo Genético	52
4.2.2.4. Construcción de Matrices	53
4.2.2.5. Diseño del modelo	56
4.3. Evaluación del modelo: Eficiencia y Robustez	57
5. Experimentos y resultados	59
5.1. Adaptación de los datos	59
5.2. Prototipo: resultados preliminares	73
5.3. Configuración GTAAF	87
5.3.1. Limpieza	92
5.3.2. Discretización	94
5.3.3. Filtrado de áreas	100
5.3.4. Resampling	102
5.3.5. Normalización	103
5.3.6. Algoritmo Genético	104
5.3.7. Construcción de matrices	106
5.4. Evaluación	106
5.4.1. Comparativas	107
5.4.2. Pruebas de estrés	119
6. Conclusiones	123
7. Publications	125
8. COSAS IMPORTANTES	127
8.1. Cosas que faltan:	127

<i>ÍNDICE GENERAL</i>	5
8.2. Dudas generales:	128
9. Anexos	129

Abstract

En esta tesis se presenta un nuevo modelo general que predice la necesidad de asistencia médica en accidentes de tráfico, independientemente de la ciudad en que se produzca, en base a información sobre 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.

Los modelos de clasificación existentes actualmente para este caso de uso tienen, como principal desventaja, que las características que requieren para sus predicciones deben ser las mismas respecto a los datos con los que se han entrenado. Es decir, requiere de un desarrollo específico para cada dataset en la que se quisiese aplicar, ya que cada una de estos, por la naturaleza socioeconómica de las poblaciones sobre los que se aplica, 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 seis 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 regiones distintas en distintos países, donde la información disponible en cada uno de estos conjuntos de datos es distinta por la naturaleza 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 irán eliminando características de mayor y menor importancia y se reevaluarán estos resultados comparándolos con los modelos del estado del arte.

Acknowledgements

Capítulo 1

Introducción

La Organización Mundial de la Salud (OMS) estima que alrededor de 1,19 millones de personas a lo largo de todo el mundo mueren anualmente como resultado de accidentes de tráfico [?]. Esto supone que los accidentes de tráfico sean considerados como un importante desafío de salud pública, que requiere de esfuerzos coordinados a nivel global para prevenir lesiones y salvar vidas, así como para abordar las repercusiones económicas y sociales que estos suponen. Numerosos estudios avalan que el tiempo en el que responden los servicios de emergencia ante accidentes de gravedad están directamente correlacionados con tasas de mortalidad más altas [?], lo que convierte esto en un factor clave para minimizar las consecuencias de las víctimas.

El hecho de conocer si un accidente puede tener consecuencias graves para alguna de las personas implicadas, una vez se ha producido el accidente es una necesidad básica, ya que permitiría a las administraciones públicas y privadas una asignación de recursos médicos (por ejemplo, una ambulancia) eficiente y prioritaria para minimizar a corto y largo plazo tanto las consecuencias físicas para las víctimas como las de los costes económicos que pudieran suponer los tratamientos posteriores necesarios para su posible recuperación.

La creación de un modelo predictivo general aplicable a cualquier área urbanizada tiene como principal restricción la dependencia en la disponibilidad de la información que ofrece cada administración. Para la creación de un modelo son necesarios datos que describan el accidente, para que este tome conocimiento sobre ellos y pueda realizar predicciones ante nuevas muestras. Cada población cuenta con recursos económicos distintos y condiciones sociales diferentes, por lo que los datos disponibles en cada una de estas suelen variar, haciendo difícil diseñar una metodología general que no sea dependiente de los datos individuales ofrecidos por cada administración.

En base a la necesidad de solventar esta debilidad, el **objetivo principal** de esta tesis es proponer una metodología y un modelo general aplicable a cualquier

área urbana para conocer la gravedad de un accidente con víctimas implicadas en base a la descripción del incidente, que de otra forma sólo sería posible conocer una vez acudiesen las asistencias médicas al lugar del mismo. Para ello, se propone una metodología y un modelo convolucional generalizables que hace uso de la categorización de la información disponible en cualquier conjunto de datos de accidentes, ya que engloba en información básica características individuales presentes en los datos, de tal forma que se pueda aplicar a cualquier sitio, creando una herramienta práctica para cualquier servicios de emergencia a lo largo de todo el mundo. Otro punto importante a destacar es la de dotar al modelo de una robustez frente a pérdida de características, para que sea del todo generalizable y no dependa del país donde se produzca el accidente (por ejemplo, si un país tiene un dataset sin datos de test de alcohol/drogas por falta de desarrollo económico, el modelo siga funcionando correctamente).

El principal avance en la investigación, desarrollada en el marco de este programa de doctorado, es el desarrollo de una metodología general para la predicción de la gravedad de los accidentes de tráfico en base a la descripción de las circunstancias que rodean al mismo. El modelo se basa en la utilización de redes neuronales convolucionales de dos dimensiones y puede ser aplicable a cualquier población del mundo.

Para la creación del modelo se ha tenido que estudiar una serie de elementos que se enumeran a continuación:

1. Técnicas de transformación de datos tabulares a matrices.
2. Algoritmos de medición de importancia de características.
3. Algoritmos evolutivos para optimización de hiperparámetros.
4. Diseño de redes convolucionales aplicadas a datos de baja dimensionalidad.
5. Técnicas de balanceo de datos.
6. Métodos de tipificación y categorización de datos.
7. Estudio de modelos del estado del arte para la comparativa.
8. Técnicas de exploración de datos geográficos espaciales.
9. Análisis de resultados e implementación de mejoras.

Para llevar a cabo los objetivos propuestos, en el Capítulo 2 se estudiará el estado del arte de la predicción de la gravedad de los accidentes de tráfico. La contextualización del marco teórico relativo a las herramientas utilizadas para el desarrollo de este trabajo se tratará en el Capítulo 3. Seguidamente, en el Capítulo 4, se definirá la metodología propuesta para resolver la necesidad de la predicción de la gravedad de los accidentes, tanto un modelo preliminar como un modelo definitivo desarrollado en base a este, aplicando mejoras sobre

el primero con el objetivo de aumentar el rendimiento del primer prototipo y transformandolo en un modelo aplicable a cualquier población (GTAAF). En el Capítulo 5 se expondrán los datos utilizados para evaluar ambas metodologías y los resultados obtenidos en comparación con otros modelos del estado del arte. Por último, en el capítulo 6, se interpretará la utilidad de este trabajo y se expondrán posibles mejoras a futuro para seguir incrementando el rendimiento de la metodología GTAAF final propuesta.

Capítulo 2

Estado del arte

2.1. ¿Cómo medir la gravedad de un accidente?

La definición de la gravedad de un accidente de tráfico es el punto de partida para enfocar cualquier investigación relacionada con el estudio de los accidentes de tráfico. La interpretación de la gravedad que implica un accidente puede ser muy variada, de hecho, a lo largo de los años, muchas han sido las investigaciones que han estudiado desde distintos puntos de vista el impacto que supone su consecuencia, tanto a nivel económico, como físico y/o social. Es por esto por lo que, en función del prisma con el que se mire, los criterios que se utilicen para definirlos pueden ser muy variados, y el valor que pueda aportar un modelo predictivo puede ser de muy distinta índole en función de la definición sobre la que se estudien.

Una vertiente de los estudios se centran en medir la gravedad de los accidentes en función del coste que suponen para las autoridades, junto con el número total de víctimas involucradas en ellos, considerando así la gravedad y agrupándola en distintas clases [?]. En otros estudios, se evalúa la gravedad de los accidentes en función de la cantidad total de daños a la propiedad, número de víctimas con lesiones graves y número de víctimas mortales que se han producido [?], clasificando finalmente estos datos en cuatro clases distintas: **leves, generales, graves y muy graves**.

No obstante, parece más relevante a nivel social un enfoque más orientado al daño físico de la persona accidentada, en detrimento del impacto económico, que puede ser solucionado con dinero. Es más, esta interpretación es la más común en la literatura: consecuencias físicas que supone para cada una de las víctimas individuales implicadas en el accidente. La clasificación más común dentro de estos puede ser la división entre **lesiones fatales, graves y sin lesiones**. Otros estudios (ver [?]) toman como referencia la agrupación de la gravedad de las víctimas hasta en cuatro clases: sin lesiones, lesiones leves, lesiones moderadas

y accidentes fatales.

Estos enfoques donde se clasifican la gravedad en un conjunto de niveles tiene una problemática: la subjetividad y/o solapamiento de niveles. Por ejemplo, un accidente puede situarse de forma difusa entre lesión grave y fatal, e incluso puede ser grave en un momento y convertirse en fatal en función de múltiples factores externos que no es posible controlar. A nivel de predicción, también se puede encontrar el problema de que la falta de datos haga más difícil evitar ese solapamiento y producir errores de predicción graves.

Por ello, es interesante valorar clasificaciones binarias [?, ?], donde la gravedad puede clasificarse como por ejemplo, **accidente fatal/no fatal**, como **lesión/no lesión**, o incluso **accidente con lesiones o solo daños materiales** [?, ?].

En este trabajo, la forma en la que se medirá la gravedad de los accidentes de tráfico se orienta a las consecuencias físicas que suponen para cada una de las víctimas individuales implicadas en ellos, poniendo el foco en **la necesidad de asistencia sanitaria** a las personas en un accidente de tráfico.

2.2. ¿Cómo predecir la gravedad de un accidente de tráfico?

La predicción de la gravedad de los accidentes de tráfico ha sido un campo ampliamente estudiado a lo largo de los últimos años, debido a la importancia que tienen para las autoridades a lo largo de todo el mundo. 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 los últimos años ha sido exponencial. Es por esto que, para solventar este problema, se han aplicados diferentes enfoques a lo largo de distintas poblaciones en todo el mundo, donde la gravedad de los accidentes han sido consideradas de forma diferente, como se ha mencionado en el apartado anterior.

Como principal punto de partida en la historia reciente se puede tomar [?], donde el modelo propuesto está entrenado en base a los accidentes producidos en la autopista Norte-Sur de 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. Por otra parte, en [?] se aplican tres métodos distintos al contexto para predecir la gravedad de los accidentes en la ciudad de Seúl, concretamente Random Forest, Perceptrones Multicapa y árboles de decisión, con el objetivo

2.2. ¿CÓMO PREDECIR LA GRAVEDAD DE UN ACCIDENTE DE TRÁFICO?17

de comparar cuál de ellos generaliza mejor sobre los datos, siendo el Random Forest el que mejores resultados presentaba.

Enfoques más recientes se contemplan en el *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 y propone un modelo de árboles de decisión sobre el que se entrena el conjunto de datos en base a estas características.

Otra consideración interesante es el utilizado en [?], donde se aplican en conjunto distintos árboles de decisión para crear un ensamblaje del tipo Random Forest, cuyos hiperparámetros son optimizados mediante Optimización Bayesiana (BO). Esta publicación busca predecir la gravedad de los accidentes a lo largo de Estados Unidos, concretamente entre los estados de Alabama y el estado de Pensilvania. Para ello, se utiliza un conjunto de datos relativo a accidentes de tráfico producidos entre los años 2016 y 2019.

Otras perspectivas contemplan únicamente la predicción de la gravedad de los accidentes sobre un subconjunto de vehículos, como por ejemplo los vehículos de dos ruedas. Una de estas investigaciones se presenta en [?], donde se predice la gravedad del accidente exclusivamente en conductores de vehículos de dos ruedas en la ciudad de Chennai (India) entre los años 2016 y 2018. Para esto, se desarrollan dos modelos independientes para compararlos, el primero de ellos un modelo Random Forest convencional y el segundo Conditional Inference Forest (CIF). Este algoritmo es similar al Random Forest pero, en lugar de utilizar el índice Gini para separar las muestras en cada nodo, se utilizan métodos estadísticos para determinar la importancia de la separación de los datos en base al p-valor, asegurando así que la división de los datos se realiza en base a las características más significativas. Este enfoque permite además medir la importancia de cada característica disponible en el dataset. Para evaluar el rendimiento de este modelo se compara con el otro método desarrollado (Random Forest) y un método Ordered Probit, ambos siendo superados por CIF. Igualmente, siguiendo con la predicción de la gravedad de accidentes en vehículos de dos ruedas, en [?] se propone un tratamiento orientado a ciclistas. Se quieren predecir las características clave que influyen en la gravedad de los accidentes de ciclistas a lo largo de las carreteras de Italia, utilizando un conjunto de datos que contempla registros de accidentes desde el año 2011 hasta el 2013. Para ello utilizan un árbol de decisión tipo CHAID que evalúa la importancia de las características más relevantes que producen este tipo de accidentes, posteriormente, las ocho más significativas son incluidas en el entrenamiento de un Optimizador Bayesiano clasificador de gravedad.

Otro tipo de investigaciones se han orientado a la predicción de la gravedad de vehículos más pesados, como es el caso de [?]. El objetivo principal de este estudio es predecir las características más influyentes en accidentes donde se encuentran involucrados camiones. Para ello, los autores utilizan un conjunto de datos de Irán, donde se estudian los accidentes a lo largo de ocho provincias entre los años 2011 y 2014. En base a estos datos, se entrena un modelo del tipo

Support Vector Machoíne (SVM) y un tipo de árbol de decisión Random Parameter Binary Logit (RPBL) por separado, para predecir qué variables afectan en mayor medida a la consecución de accidentes graves.

En la historia reciente también es común encontrar investigaciones que combinan distintos modelos para lograr un mejor rendimiento, como es el caso de [?]. En este artículo los investigadores implementan un enfoque de clasificación híbrida basada en modelos machine learning sobre los datos de la autopista de Pakistán N-5 entre los años 2015 y 2019. Para ello se utiliza un algoritmo de selección de características (Boruta Algorithm) para decidir las características son influyentes en la predicción de la gravedad de los accidentes, apoyándose en un clasificador Random Forest. Posteriormente, estas características resultantes se incluyen para el entrenamiento de cuatro modelos clasificadores con el fin de comparar el rendimiento entre ellos, concretamente Naive Bayes (NB), k-Nearest Neighbor (KNN), Binary Logistic Regression (BLR) y XGBoost, siendo este último el que mejor generaliza sobre los datos.

Otra de las vertientes propuestas recientemente es la basada en aplicar distintos algoritmos de aprendizaje automático para preprocessar la entrada a un modelo basado en *Autoencoders*. Esta propuesta utiliza algoritmos para seleccionar características influyentes que afecten a la gravedad de los accidentes de tráfico para utilizar posteriormente algoritmos de agrupación en base a las características geográficas de los datos, con el objetivo final de entrenar un *Stacked Sparse Autoencoder (SSAE)* [?] que predice la severidad de dichos accidentes.

Por otra parte, estudios como [?] ofrecen la comparación de distintos modelos predictivos como Multi Layer Perceptron (MLP), MLP con embeddings y TabNet, utilizando optimizadores bayesianos para optimizar los hiperparámetros de estos modelos.

El componente de los datos a la hora de presentar todos 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, debido tanto a la falta de características comunes respecto a otros conjuntos de datos como a las peculiaridades concretas de cada conjunto de datos en cuestión.

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. Esta dependencia de los datos, es una debilidad que impide el desarrollo de modelos generales y que ha motivado el estudio presentado en esta tesis.

En resumen, se puede observar como los estudios sobre la predicción y aná-

2.2. ¿CÓMO PREDECIR LA GRAVEDAD DE UN ACCIDENTE DE TRÁFICO?19

lisis de accidentes de tráfico son muy específicos, centrándose en un dataset concreto, en un área acotada. Además, la mayoría se enfoca a un perfil de accidente concreto (camiones, bicicletas, etc) y analiza la gravedad en base a un criterio como los citados en la sección anterior. Es por ello que se plantea en esta tesis un modelo general, donde se analice la necesidad de asistencia médica o no de los accidentes de tráfico independientemente de la urbe y del vehículo implicado, entre otras características.

En el siguiente capítulo se hace un estudio base de las técnicas y modelos estudiados en esta tesis, previo a la explicación detallada del modelo propuesto.

Capítulo 3

Background

En este capítulo se expone el marco contextual de las herramientas sobre las que se desarrolla esta tesis, cobrando especial importancia para comprender la metodología de este estudio en su totalidad. Es por esto por lo que en este apartado se introducen conceptos y métodos que se mencionan a lo largo de todo el documento. Las siguientes secciones definen cada una de las herramientas utilizadas en este trabajo.

3.1. Modelos utilizados en la comparativa

En el campo de la inteligencia artificial y, más concretamente, el aprendizaje automático, muchos han sido los modelos y los métodos que se han desarrollado a lo largo de la historia reciente. Parte de ellos han tenido una gran relevancia debido a que han sido ampliamente aplicados en distintos tipos de problemas ofreciendo gran calidad en sus resultados, llegando a una gran capacidad de generalización en distintos contextos. Por este motivo, en este trabajo se tomarán como referencia algunos de estos métodos para la realización de una comparativa de eficiencia a través de distintas métricas. En esta sección, se describen a nivel teórico los modelos de Aprendizaje Automático que son relevantes para la comprensión y desarrollo de esta tesis.

Gaussian Naive Bayes

El algoritmo Naive Bayes es un método de Aprendizaje Supervisado ampliamente extendido en problemas de clasificación de datos, tanto por su simplicidad como por la capacidad de generalización en la calidad de los resultados que ofrece en numerosos problemas de clasificación, por ejemplo, en investigaciones recientes ha sido aplicado en detección de enfermedades cardiovasculares [?], detección de ciberataques en redes inalámbricas [?] o reconocimiento de escenas

a través de imágenes mediante características geométricas presentes en ellas [?]. La denominación de este modelo se debe a que su funcionamiento está basado el teorema de Bayes, ya que este algoritmo calcula la probabilidad de pertenencia de una muestra en base a la suposición de que cada una de las variables (predictoras) presenta una independencia condicional respecto al resto de ellas. Una de las variantes de este algoritmo para trabajar con variables numéricas es el Gaussian Naive Bayes. Este algoritmo trabaja con las probabilidades a priori de pertenencia de las muestras a una clase, calculada sin tener en cuenta las variables predictoras de los datos, junto con las probabilidades a posteriori, aquellas individuales de cada característica una vez se toma conocimiento de los datos de entrenamiento, y sus predicciones se basan en la composición de ambas:

La fórmula que representa la probabilidad a priori $P(y)$ de una clase y viene representada por la siguiente fórmula:

$$P(y) = \frac{\text{Número de muestras de la clase } y}{\text{Número total de muestras}}$$

Por otra parte, cada una de las variables de los datos de entrenamiento son proyectadas en distribuciones gaussianas, para cada una de las clases a predecir siguiendo la siguiente fórmula:

En tiempo de inferencia, la predicción de una nueva muestra se interpreta como aquella probabilidad más alta en base a las clases, haciendo uso de la información que ha obtenido el modelo en base a sus datos de entrenamiento.

Figura 3.1: Ejemplo de distribuciones en base a Gaussian Naive Bayes [?].

Multi Layer Perceptron

El Perceptrón Multicapa o Multi-Layer Perceptron (MLP) es una arquitectura de red neuronal artificial. Es una generalización del Perceptrón simple y surgió como consecuencia de las limitaciones de dicha arquitectura en lo referente al problema de la separabilidad no lineal. El MLP es un aproximador universal, en el sentido de que cualquier función continua en un espacio R^n puede aproximarse con un Perceptrón multicapa, con al menos una capa oculta de neuronas. Esto, sitúa al MLP como un modelo matemático útil a la hora de aproximar o interpolar relaciones no lineales entre datos de entrada y salida.

La arquitectura de Perceptrón multicapa se caracteriza porque tiene sus neuronas agrupadas en capas de diferentes niveles: la capa de entrada, las capas ocultas y la capa de salida. Las neuronas de la capa de entrada se encargan únicamente de recibir las señales del exterior y propagarlas a todas las neuronas de la siguiente capa. La última capa proporciona la respuesta de la red para cada uno de los patrones de entrada.

Este tipo de arquitecturas son ampliamente utilizadas en problemas de aprendizaje supervisado, como clasificación y regresión, y ofrecen un gran rendimiento en contextos muy variables. Por ejemplo, en publicaciones recientes, este tipo de modelos ha sido aplicado a problemas como segmentación de imágenes médicas para la prevención de múltiples enfermedades [?], para la clasificación de nubes de puntos en tres dimensiones [?] o como apoyo para la predicción del tiempo estimado de trayectos en taxi de la ciudad de Nueva York [?]. Al pertenecer a la familia de las redes neuronales, estos modelos aprenden los pesos asignados a cada una de las conexiones entre neuronas de las capas para dar lugar a las fases comunes de optimización de redes neuronales que detallarán en la sección ??.

Figura 3.2: Ejemplo de arquitectura MLP [?].

Regresión Logística

La regresión logística es un modelo de aprendizaje estadístico utilizado históricamente para solventar problemas de clasificación binaria. Este método tiene como objetivo deducir la probabilidad de que ocurra un evento binario en función de uno o más predictores, siendo ampliamente extendido debido a su sencillez y capacidad de generalización. Por ejemplo, en investigaciones recientes, este modelo se ha aplicado de forma exitosa a la predicción de la pérdida de clientes en sectores de telecomunicación [?], para la creación de rápidos protocolos computacionales de seguridad para preservar la privacidad del genoma humano [?], o para la predicción de diabetes [?] en base a descriptores humanos. La regresión logística asigna un coeficiente a cada una de las variables predictoras, y estos son ajustados durante el proceso de aprendizaje para minimizar una función objetivo, normalmente R^2 . Este proceso de ajuste de coeficientes en base a los datos de entrenamiento resulta en una combinación lineal de variables independientes ante nuevas muestras:

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

Donde $\beta_0, \beta_1, \dots, \beta_n$ son los coeficientes asignados a cada una de las variables predictoras, siendo β_0 el término de intercepción, y z es el valor de la combinación de las características multiplicadas por dichos coeficientes.

La regresión logística hace uso de una función sigmoide, que transforma los valores continuos resultantes de la combinación lineal z a una probabilidad de pertenencia a una clase en función de la separabilidad a través de esta dimensión de las muestras de entrenamiento.

$$P(y = 1|x) = \frac{1}{1 + e^{-z}}$$

Donde $P(y = 1|x)$ representa la probabilidad de la muestra x de pertenecer a la clase positiva.

Figura 3.3: Ejemplo de clasificación mediante Regresión Logística [?].

k-Nearest Neighbours

El algoritmo k-Nearest Neighbours (KNN) es un algoritmo de aprendizaje automático, ampliamente utilizado tanto para problemas de clasificación como de regresión. Estudios recientes han aplicado esta técnica para distintos propósitos, como agrupar datos de dimensiones arbitrarias para reconocer picos de densidad de forma eficiente [?], para clasificación de textos [?] o como apoyo para la detección de intrusos en redes inalámbricas [?] en base al número de usuarios por nodo, peticiones recibidas de la red, etc.. Este método se basa en la clasificación de nuevas muestras en base a la distancia de sus características respecto a la proyección de las características de las muestras de entrenamiento sobre un espacio N-dimensional. Cuando una nueva muestra x_0 es clasificada, KNN identifica los K puntos más cercanos almacenados de sus muestras de entrenamiento respecto a las de la observación x_0 (N_0) y genera una probabilidad de pertenencia de la nueva muestra x_0 a la clase j , en función de la fracción de puntos de N_0 que pertenezcan a esa clase. La fórmula para calcular la probabilidad de pertenencia es la siguiente:

$$P(y = j|X = x_0) = \frac{1}{K} \sum_{i \in N_0} I(y_i = j)$$

Figura 3.4: Ejemplo de clasificación del KNN ante una nueva muestra, donde se miden las distancias entre los N puntos más cercanos ante una nueva muestra y se clasifica como aquella que más vecinos cercanos presente [?].

Random Forest

Los algoritmos Random Forest son algoritmos basados en árboles de decisión que pertenecen al conjunto de modelos Ensembles. Los métodos Ensembles son arquitecturas que están conformadas por varios modelos entrenados simultáneamente para dar lugar a un modelo predictivo final. Dentro de la categorización de Ensembles, el método Random Forest, pertenece a la subcategoría Bagging, cuya principal característica es que se apoyan en crear múltiples modelos que son entrenados con distintas técnicas de reemplazo (bootstrap) sobre los datos, conformando un modelo predictivo final como la combinación de la salida de cada uno de ellos de manera independiente por votación. Este tipo de modelos

ha sido aplicado de forma efectiva en contextos muy variados, como para predicción de la contaminación de nitratos de aguas subterráneas [?], predicción de precios de inmuebles en base a sus características [?] o para la medición de características geográficas socioeconómicas importantes que influyeron en los fallecidos por COVID-19 en Estados Unidos [?]. Concretamente, 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 al resto 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. De forma generalizada, los modelos tipo Ensemble son métodos robustos que son menos sensibles al sobreajuste de los datos gracias a sus técnicas de reemplazo durante su etapa de entrenamiento y la composición de un modelo final en base a varios modelos.

Figura 3.5: Ejemplo de arquitectura ensemble Random Forest. El dataset es dividido en subconjuntos mediante Bootstrap y se entrena un árbol de decisión sobre cada uno de ellos. La clasificación final viene dada por la predicción mayoritaria de los modelos [?].

Support Vector Classifier

El Support Vector Classifier (SVC) es un algoritmo de aprendizaje supervisado utilizado comúnmente para problemas de clasificación. Recientemente, este modelo ha sido aplicado a distintos contextos, como para la predicción de la deformación de los túneles durante su excavación en base a las condiciones geológicas [?], como apoyo para la detección temprana de cánceres de piel [?] o incluso para la predicción de defectos en aplicaciones *software* mediante métricas comunes en el desarrollo de los proyectos [?]. Se basa en el concepto de encontrar un hiperplano óptimo que maximice la separabilidad de las clases de entrenamiento en base al espacio generado por la proyección de las características de los datos. El SVC define estos hiperplanos en base a las muestras de distintas clases que más cerca se encuentren a lo largo del espacio N-dimensional. La fórmula que define un hiperplano para el SVC en un problema de clasificación binario se define como:

$$W \cdot X - b = 0$$

En tiempo de predicción, el SVC utiliza los hiperplanos definidos para determinar si una nueva muestra x_0 pertenece a una clase u otra mediante la siguiente ecuación:

$$f(x_0) = \text{sign}(W \cdot x_0 - b)$$

Donde W es el vector de pesos asociado a cada característica, X es el vector

de características de la muestra y b es el término de sesgo.

Figura 3.6: Ejemplo de SVC propuesta en [?].

3.2. Algoritmos CNN

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 dentro de este contexto, como clasificación de imágenes, detección de elementos de interés, o incluso han sido aplicadas a problemas de regresión. Tal es su redimento en estos problemas, que estas arquitecturas han sido extrapoladas al campo de la Inteligencia Artificial Generativa, ofreciendo soluciones en distintos problemas como la generación de imágenes artificiales a través de redes generativas antagónicas (GAN), segmentación de elementos de interés dentro de imágenes, o incluso en la representación de imágenes mediante vectores n-dimensionales para comparar la similitud de imágenes entre sí mediante redes siamesas.

La principal característica que distingue a estos modelos respecto al resto de redes neuronales y que las hace tan efectivas en problemas orientados a imágenes, es que su diseño 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 quiera 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:** las funciones de activación de las CNN introducen no linealidades, lo que permite a la red aprender patrones y relaciones complejas dentro de los datos. Normalmente, se aplica una función de activación como ReLU (Rectified Linear Unit) a los mapas de características para introducir la no linealidad 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, 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.

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

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 en cada etapa de la red, a esta etapa se le conoce como Forward Propagation, y es definida mediante la siguiente fórmula:

$$\mathbf{z}^{[l]} = \mathbf{W}^{[l]} \cdot \mathbf{a}^{[l-1]} + \mathbf{b}^{[l]},$$

donde:

- $\mathbf{z}^{[l]}$ es la activación antes de aplicar la función de activación en la capa l .
- $\mathbf{W}^{[l]}$ es la matriz de pesos.
- $\mathbf{a}^{[l-1]}$ es la salida de la capa anterior.

- $\mathbf{b}^{[l]}$ es el vector de sesgo.

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. Para llevar a cabo este proceso, es necesario introducir el concepto de función de pérdida que medirá el error durante la fase de entrenamiento

Las funciones de pérdida en las redes neuronales miden el error producido por la red al realizar predicciones sobre los datos en su etapa de entrenamiento. Estas funciones tienen com objetivo comparar la calidad de la predicciones de la red respecto a la clase verdadera a la que pertenecen. Un valor alto de esta función indica un error alto en las predicciones, mientras que un valor bajo representa una buena calidad de predicciones:

$$\text{Pérdida} = \text{Cálculo de Pérdida}(\text{Verdadero}, \text{Predicho})$$

La función de pérdida más común en problemas de clasificación binaria es la Binary Cross Entropy cuya fórmula es

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

Con N representando el número de muestras totales en el conjunto de datos, y_i la etiqueta verdadera de la clase (0 ó 1) de la muestra actual y p_i la probabilidad de que la muestra actual pertenezca a la clase 1.

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. Así, valores de probabilidad altos de la predicción de 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.

La actualización de los pesos de la red, viene dada por el proceso de Back Propagation. Este método utiliza la función de pérdida de la época actual para calcular la dirección en la que actualizar la red mediante el descenso por gradiente. Para actualizar los pesos de la red se hace uso de la regla de la cadena para actualizar los pesos de la red, esto 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. La siguiente fórmula representa la actualización de los pesos para una de estas capas:

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:

$$\mathbf{W}^{[l]} = \mathbf{W}^{[l]} - \alpha \frac{\partial J}{\partial \mathbf{W}^{[l]}},$$

con:

- α es la tasa de aprendizaje.
- $\frac{\partial J}{\partial \mathbf{W}^{[l]}}$ es el gradiente de la función de pérdida J con respecto a los pesos.

En este trabajo, se estudiarán en detalle dos tipos de redes de este tipo, las redes neuronales convolucionales unidimensionales, CNN-1D, y las redes neuronales convolucionales bidimensionales, CNN-2D).

CNN-1D

Las redes neuronales convolucionales unidimensionales son redes cuya característica principal es que los filtros que aplican en cada una de sus convoluciones son de una dimensión [?]. 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 debe a que no terminan de adaptarse correctamente a encontrar patrones bidimensionales, debido a su naturaleza.

CNN-2D

Las redes neuronales convolucionales bidimensionales (CNN-2D) son redes ampliamente utilizadas en todo tipo de problemas relativos a imágenes [?]. La principal característica que distingue a estas redes es que están especialmente adaptadas a la naturaleza bidimensional de las imágenes, utilizando filtros de dos dimensiones para capturar los patrones presentes en estas. Muy distintos han sido los casos de estudio en los que se han aplicado estas arquitecturas, como clasificación de imágenes médicas para detección de enfermedades tempranas [?], IA generativa para reidentificación facial [?] mediante generación de representaciones en los patrones detectados o la segmentación de objetos de interés en imágenes [?].

3.3. Algoritmos de construcción de matrices

La tendencia de utilizar modelos neuronales convolucionales en los últimos años ha tenido un considerable incremento debido a la eficiencia y rendimiento que estos demuestran a lo largo de múltiples contextos. La particularidad de estos modelos es que trabajan sobre datos matriciales sobre las que pueden aplicar convoluciones para aprender patrones más o menos complejos. Gran parte de la naturaleza de los problemas que debe resolver el campo de la Inteligencia Artificial, tiene un origen tabular, es decir, los datos que se ofrecen están conformados por registros (bases de datos, hojas de cálculo, etc.). Al no disponer de una tipología matricial o en forma de imagen, este problema impide aplicar modelos convolucionales a este tipo de datos. Para sortear este problema, a lo largo de los últimos años se han ido desarrollando técnicas para transformar datos tabulares a matriciales con el fin de poder aplicar este tipo de modelos. La forma en la que se transforman datos tabulares a datos matriciales no tiene una solución trivial, ya que la composición resultante 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 ajustarse a este requisito, como la propuesta de OmicsMapNet [?], un algoritmo orientado a construir representaciones matriciales de las características de los genes en 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 DeepInsight [?], 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 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.

3.4. Métodos de Remuestreo

En el campo de la Inteligencia Artificial el problema del desbalanceo de los datos ha sido ampliamente estudiado a lo largo de los años debido a los inconvenientes que generan para el entrenamiento de ciertos modelos predictivos y la frecuencia con la que los datos presentan este problema. Un conjunto de datos desbalanceado, respecto a una clase a predecir, se define como aquel que dispone proporcionalmente de muchas más muestras pertenecientes a una clase respecto a las demás. Los modelos predictivos de Inteligencia Artificial aprenden y actualizan su conocimiento en base a los datos, y gran parte de ellos se entrena prediciendo sobre las muestras de entrenamiento en su etapa de aprendizaje para posteriormente actualizar su conocimiento en base a la clase real a la que pertenecían dichas muestras. Si el conjunto de datos sobre el que aprende un modelo de IA presenta una clase mayoritaria, este será penalizado en más ocasiones cuando se equivoque al predecir estas muestras como cualquier otra clase, por lo que el modelo para evitar ser penalizado tenderá a predecir todas las muestras como la clase mayoritaria, de tal forma que se obtiene un modelo sesgado producido por un conjunto de datos desbalanceado.

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, y para ello existen dos principales corrientes que tienen como objetivo reducir la diferencia entre el número de clases de los datos:

1. 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 mediante Random oversampling resampling [?] o Bootstrap [?] entre otros.

2. 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 a costa de perder información sobre la clase mayoritaria.

Por otra parte, existe un enfoque orientado al remuestreo de datos de las clases minoritarias mediante la generación de datos sintéticos. Estas técnicas generan nuevos datos utilizando distintos métodos, como modelos estadísticos o algoritmos para imitar patrones de datos reales. El objetivo de esta corriente es crear datos que se asemejen a los datos reales, tanto en propiedades estadísticas como en relaciones entre variables. Una de las técnicas más conocidas de generación de datos sintéticos es Borderline SMOTE-II [?].

SMOTE-II funciona como generador de datos sintéticos, para ello hace uso del espacio generado al proyectar las características de los datos. En base a esto es posible generar nuevas muestras de las clases minoritarias que se encuentren cercanas al espacio de características que divide dicha clase con el resto que conviven en el conjunto de datos. 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.

3.5. Métodos de optimización de hiperparámetros

En el campo del aprendizaje automático, la optimización de los hiperparámetros (HPO) cobra un papel fundamental en el desarrollo de los modelos. Los hiperparámetros son configuraciones que son establecidas durante la etapa previa a iniciar el proceso de aprendizaje, por lo que afectan de forma significativa a la forma en que el modelo aprende sobre los datos, hace predicciones sobre nuevas muestras y, por ende, a su rendimiento.

Es por esto por lo que una buena configuración de hiperparámetros es de vital importancia, un modelo potencialmente aplicable a un problema puede llegar a quedar inservible por el simple hecho de no tener una configuración eficiente de estos. Estas configuraciones son dependientes del problema, los datos, y el modelo propuesto para resolverlo. Además, el espacio de búsqueda (o combinación) de las configuraciones pueden ser más o menos amplias dependiendo de la naturaleza de cada modelo de aprendizaje y de las posibilidades de parametrización que este ofrezca. Debido a esto, es necesario optimizar los

hiperparámetros, y existen distintos métodos por los que hacerlo.

La principal limitación a la hora de encontrar una configuración de hiperparámetros consistente es el coste que puede suponer en términos de recursos computacionales. Para evaluar una de estas configuraciones se requiere de entrenar el modelo con dicha configuración y evaluar el rendimiento que ofrece sobre datos que nunca ha visto. De esta forma se puede tener una intuición de la calidad de esos hiperparámetros, por lo que es necesario aplicar técnicas que permitan maximizar la calidad de la configuración minimizando el coste que esto supone.

A lo largo del tiempo, distintos han sido los enfoques desarrollados para HPO de modelos predictivos, cada uno con sus fortalezas y debilidades dependiendo del contexto en el que se apliquen [?]. Una de las técnicas clásicas y referentes debido a su precisión es la técnica Grid Search. Esta técnica permite probar y es idónea para modelos ligeros y con pocos datos de entrenamiento, ya que permite probar cada combinación posible. Otra de las técnicas ampliamente reconocidas para el problema HPO y basada en el método Grid Search, es el Random Search [?]. Esta técnica establece también una parrilla donde se especifican los posibles valores que pueden tomar los hiperparámetros para seleccionar combinaciones de estos de manera aleatoria. Esto permite probar un gran número de configuraciones sin tener que pasar por cada una de las individualmente, reduciendo considerablemente el coste computacional permitiendo explorar zonas del espacio de búsqueda muy distintas.

Sin embargo, al ser una selección aleatoria, es posible que se pasen configuraciones que pueden converger.. Por otra parte, existen métodos de HPO que utilizan modelos probabilísticos para calcular el set óptimo de hiperparámetros, como es el Optimizador Bayesiano. Este modelo busca la relación entre los parámetros de entrada y los valores de salida creando un modelo Gausiano probabilístico, reduciendo así el número de evaluaciones necesaria para llegar a una solución óptima.

Otro enfoque interesante para lograr una buena configuración de hiperparámetros es el uso de algoritmos genéticos [?]. La naturaleza de estos algoritmos permite aplicarlos de forma eficiente al problema HPO, ya que son métodos óptimos para problemas de minimización, donde en función de unos parámetros de entrada, estos son capaces de maximizar la calidad de la soluciones minimizando el esfuerzo que implica llegar a ella a lo largo de las generaciones.

3.6. Algoritmos de medición de importancia de características

En el campo de la Inteligencia Artificial la medición de la importancia de las características entre los conjuntos de datos toma un papel fundamental para el análisis y desarrollo de modelos. 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 en el papel que ha interpretado en el entrenamiento de un modelo predictivo, mientras que valores más bajos suelen representar poca relevancia durante su ajuste.

En el estado del arte, existen distintos métodos que tienen como propósito medir el peso de las características en conjuntos 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 (predicción de variables continuas), es la técnica de regresión lineal. Este método mide la magnitud de las variables en base al valor y la dirección de los coeficientes respecto 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. Estos modelos durante el proceso de aprendizaje del modelo de regresión lineal utilizan términos de penalización para reducir los coeficientes del predictor, con el objetivo de introducir un componente de regularización, que favorecerá la generalización del modelo al evitar que pocos predictores sean demasiado influyentes en las predicciones de nuevas muestras. 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, que para calcular la importancia de las variables, 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 [?].

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 son los modelos basados en filosofías de tipo ensemble, ya introducidos en la sección 3.1.

Dentro la filosofía de los modelos *ensemble*, los algoritmos Random Forest han sido históricamente utilizados para calcular la importancia de las características de su entrenamiento. Para este fin, este algoritmo calcula mediante el peso que ha tenido cada característica a la hora de construir los árboles, en función del número de muestras que dividen cada uno de los niveles en los distintos árboles construidos.

No obstante, existe otra técnica más con mayor potencial dentro de los ensambles que tiene como base el funcionamiento de los Random Forest, el algoritmo tipo Boosting XGBoost [?]. Los algoritmos tipo Boosting 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 modelo anterior, reduciendo así el sesgo que pudiera llegar a

producirse por datos desbalanceados y mejorando la precisión del modelo final, además ofrece de términos de penalización para evitar una complejidad del modelo excesiva. Estos algoritmos disponen de una serie de hiperparámetros que deben ser configurados para un buen rendimiento, como limitar la profundidad de los árboles que se generan, la tasa de aprendizaje, que define la contribución de cada árbol al modelo final, o el número de instancias mínimas que debe separar un nodo para poder ser creado, esto permite reducir el sobreajuste de los modelos a casos muy específicos encontrados en los datos de entrenamiento. XGBoost ha sido ampliamente utilizado a lo largo de los últimos años debido a su gran potencial para problemas muy diversos, como para la predicción de empresas que entran en bancarrota en base a sus características [?], para predecir los niveles de agua subterránea en base a descriptores meteorológicos y geográficos [?], o incluso para predecir la fluctuación del valor del oro a lo largo del tiempo [?]

Figura 3.7: Ejemplo de algoritmo XGBoost [?].

3.7. Algoritmos Genéticos

Los algoritmos genéticos son métodos inspirados en la evolución biológica, que buscan optimizar soluciones a problemas matemáticos mediante la simulación de la evolución de una población de individuos que producen descendencia a lo largo de generaciones. Estos algoritmos han sido ampliamente utilizados en casos como la optimización del flujo de tráfico en la red para balancear la carga de los nodos [?], para analizar la capacidad de agua en el suelo mediante imágenes remotas [?] o incluso para simular con el uso de autómatas distintas enfermedades como el COVID-19 [?]. La principal fortaleza de estos algoritmos es que son métodos eficientes y seguros para llegar a soluciones aproximadas a la óptima ideal, reduciendo el coste computacional (en muchos casos exponencial), que supondría la búsqueda del óptimo global mediante métodos de combinación a lo largo de todo el espacio de búsqueda. Existen numerosos algoritmos genéticos conocidos, y muchos han sido aplicados a distintos contextos, tanto para problemas de objetivo único como a problemas multi-objetivo [?].

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

En la primera de ellas (*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 (véase Figura 3.1). 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 (*evaluación*), como se muestra en la Figura

3.2. Una vez se dispone de las puntuaciones de calidad, aquellos individuos que mejor se adapten al problema, es decir que mejor puntuación reciban, serán escogidos para dar lugar a descendencia (*selección*). La información que contienen los M mejores individuos es combinada entre sí (*cruce*), simulando el intercambio de información producido en el intercambio genético en la naturaleza. Una vez se disponen de los nuevos individuos, estos pueden sufrir modificaciones aleatorias sobre su información resultante (*mutación*), como se puede ver en la Figura 3.3. 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 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 varias generaciones hasta llegar a una condición de parada, tras la cual se seleccionará el individuo que mejor puntuación haya obtenido mediante la función heurística.

Cabe mencionar que dentro de la etapa de cruce, existen infinitas estrategias que se pueden aplicar para dar lugar al individuo descendiente. Por ejemplo, una de las más comunes suele consistir en dividir en dos a cada par de individuos que se reproducirán para generar su descendiente en base a composición de estas dos mitades. Otro método común, es seleccionar un punto aleatorio en cada proceso de descendencia del vector solución para dividir ambos progenitores, de tal forma que el individuo generado puede contener más información de un progenitor que de otro.

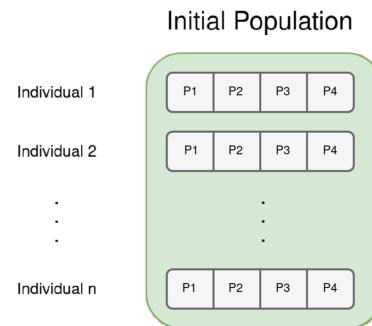


Figura 3.8: Ejemplo de inicialización de individuos de cuatro características en un algoritmo genético.

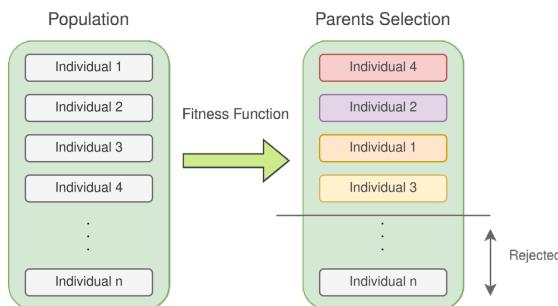


Figura 3.9: Ejemplo del proceso de selección de los cuatro mejores individuos de una generación a través de su calidad en base a la función heurística (*fitness function*).

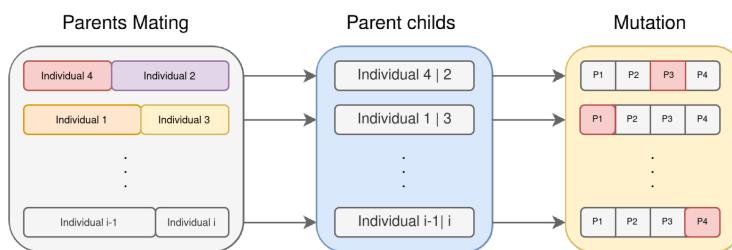


Figura 3.10: Ejemplo del proceso de cruce entre individuos de una población y mutación de una de las características del descendiente resultante.

De esta forma, mediante la mejora continua de individuos a través de las generaciones, seleccionando los mejores y combinando su información entre sí, da lugar a una solución aproximada a la ideal a lo largo de las generaciones.

3.8. Medidas de evaluación de una red neuronal

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, es necesario introducir conceptos básicos que deben ser calculados para cada una de las clases que puede predecir el modelo.

1. **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 (TN):** representan el número de muestras que han sido correctamente clasificadas por el modelo como negativas. Es decir, el modelo
3. **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 (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 encontramos, es posible que sea preferible un modelo que tienda a predecir con más facilidad futuras muestras a un tipo de clase respecto otra. Por ejemplo, es mejor predecir erróneamente que un accidente necesita asistencia (FP sobre la clase asistencia) y que luego no sea necesaria ninguna intervención, que predecir erróneamente uno que no necesita asistencia (FN sobre la clase No Asistencia). Este análisis del balance es posible evaluarlo gracias a los indicadores definidos anteriormente, no obstante, este tipo de decisiones son dependientes de la criticidad del problema que se quiere resolver.

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.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

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

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

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.

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

La métrica F1-Score es particularmente útil en problemas de clasificación binaria, sobre todo cuando existe un desbalanceo en las clases. Particularmente en estos casos, este indicador muestra información muy relevante en comparación a otras por sí solas, ya que toma en cuenta la combinación tanto de los falsos positivos como de los falsos negativos. Debido a la visión general del rendimiento de los modelos que ofrece en este tipo de casos, será la métrica definida para la evaluación de los modelos en esta tesis.

Capítulo 4

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

En este trabajo se expone una metodología y un modelo general de predicción de gravedad de accidentes de tráfico aplicable a cualquier región e independiente a los datos que puedan estar disponibles en cada una de ellas. 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 e hipótesis 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 general 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

En una primera fase de la investigación, se propuso un modelo preliminar de predicción de la gravedad de accidentes de tráfico aplicado a una ciudad [?], teniendo como meta la predicción de la gravedad de los accidentes en base a 3 niveles (leve, severo y fatal). Como el objetivo de este trabajo es diseñar un

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

modelo predictivo general a cualquier ciudad, en este apartado se expondrá brevemente el enfoque del modelo preliminar, a modo de introducción y acentuando los matices que condujeron a la reclasificación posterior de las clases que describían la necesidad de asistencia. El modelo general que se propone se explicará con total detalle en la siguiente sección 4.2.

Este primer prototipo se presentó en [?] y se implementó con el objetivo de predecir la gravedad de los accidentes de tráfico utilizando un dataset de la ciudad de Madrid. Concretamente, se dividió la gravedad de los accidentes entre categorías distintas que estaban en ese conjunto de datos: (1) Leves, (2) Severos y (3) Fatales.

Para llegar al entrenamiento del modelo predictivo, se diseñó una metodología compuesta por cinco fases secuenciales, mostradas en la Figura 4.1. Esta metodología tenía como objetivo realizar transformaciones y operaciones sobre datos, inicialmente tabulares, para transformarlos en datos matriciales sobre los que pudieran operar modelos diseñados para tratar imágenes. 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.

Primeramente fue necesario definir una categorización de características, las cuales fueron utilizadas como apoyo para la construcción de estas matrices. Para la construcción de las matrices que alimentan las CNN se le asignó una importancia a cada variable dentro del conjunto de datos con la finalidad de situarlas en determinados lugares de estas matrices. Es decir, dependiendo de la importancia asignada a cada variable, se le asignaba una posición en la matriz.

Finalmente la metodología y los modelos convolucionales propuestos se compararon con otros tres modelos del estado del arte (Naive Bayes, Support Vector Classification y k-Nearest Neighbor) para evaluar sus rendimientos respectivos en el dataset de la ciudad de Madrid.

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

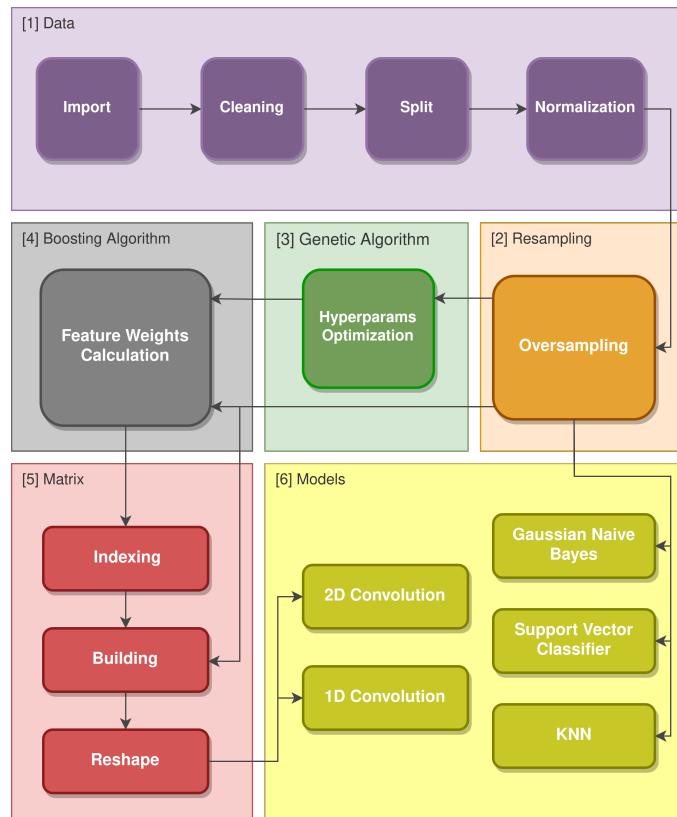


Figura 4.1: Diagrama de flujo de la metodología preliminar propuesta con sus diferentes fases.

La primera fase de esta metodología (1 en la Figura 4.1) prototipo está orientada al tratamiento de los datos. Los 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. Es por esto que en esta etapa se diseñaron métodos para tratar estos datos, concretamente aplicándoles un proceso de limpieza, una discretización para que fuesen interpretables por los modelos, un tratamiento para su normalización y otro para evitar que estuviesen sobredimensionados.

En una segunda fase (2), se aplicó un proceso para trabajar sobre el desbalanceo de los datos presente el dataset ya que, debido a la naturaleza de los accidentes de tráfico, gran parte de ellos eran de tipo leve, mientras que el resto de accidentes (severos y fatales), presentaban una proporción mucho menor. Para evitar un sesgo en los modelos es decir, la tendencia a predecir cualquier nueva muestra como la de clase mayoritaria, se estudiaron distintas técnicas de balanceo de datos. Finalmente, se utilizó la técnica Borderline SMOTE-II para balancear las clases minoritarias, aplicando generación de datos sintéticos hasta

igualar las clases hasta la mayoritaria.

Una tercera fase (3) buscaba transformar los datos tabulares resultantes a datos matriciales interpretables por los modelos convolucionales. Para esto 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 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 en la fase (4).

Una vez se disponían de los pesos de las características gracias al cálculo del algoritmo tipo Boosting (5), se categorizaron las variables en distintas características (6) para tener una referencia bidimensional sobre la que comenzar a indexar las variables. En primer lugar se calculó el peso total de las categorías, que era la suma de los pesos de cada una de las variables 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. Po otro lado, las características que las componían se asociaban a las columnas dentro de su categoría de forma similar, 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 tamaño 5×5 .

Las arquitecturas que se propusiero en 1 prototipo eran dos redes convolucionales, de una y dos dimensiones. Estas 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 kernels 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 aplicaban agregando ceros a los límites de las matrices, de 1 para la CNN-1D y 1, 1 para la CNN-2D. Por lo tanto, el desplazamiento de los núcleos se realizaba elemento a elemento en ambas redes. En la salida de cada capa convolucional, se aplicaba la función de activación Rectified Linear Unit (ReLU). La salida de la última capa de convolución transformaba los mapas de características generados de tamaño 5×5 a un vector unidimensional de 1×25 . A continuación, se aplicaba una capa densa que conectaba cada uno de los 25 nodos de la capa aplanada con los 128 nodos de la capa densa, que generaba los logits antes de aplicar la última función de activación Softmax que devolvía la clase predicha. En la figura 4.2 se observa el diseño de la arquitectura de la red propuesta.

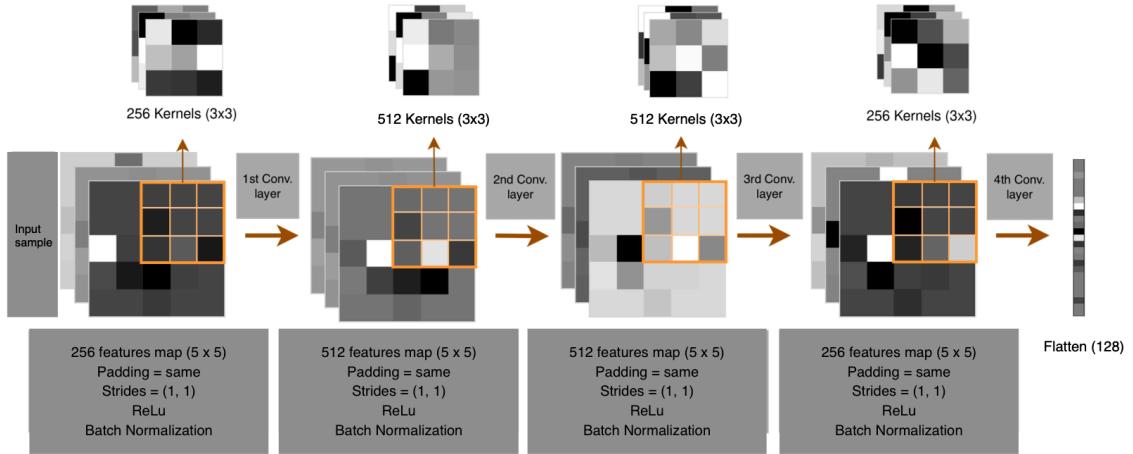


Figura 4.2: Arquitectura de la red neuronal convolucional 2D.

En última instancia (6), los dos modelos propuestos, CNN-1D y CNN-2D se compararon contra tres modelos del estado del arte, utilizando como referencia de generalización el indicador F1-Core sobre conjuntos de datos de test.

Aprendizaje resultante del prototipo

Una vez aplicada la metodología sobre el conjunto de datos da Madrid, se analizaron los resultados que esta ofrecía sobre el conjunto de datos de test. Esto sirvió para observar ciertas debilidades y localizar puntos de mejora que podrían aumentar el rendimiento de cara a implementar una metodología generalizable a cualquier región.

En primer lugar, se observó que la decisión de dividir los accidentes de tráfico en tres clases (leves, severos y fatales) era un condicionante que perjudicaba considerablemente el rendimiento del modelo. Al tener dos clases notablemente desproporcionadas respecto a la mayoritaria, la generación de datos sintéticos podría carecer de sentido a nivel práctico. Esto hizo que se pensara en agrupar los accidentes en dos categorías (leves y resto) con la finalidad de mejorar el rendimiento del modelo a nivel funcional.

Al observarse incongruencias en la interpretación sobre los valores numéricos asignados a determinadas variables, se plantearon cuestiones sobre la discretización de los datos y su inclusión en áreas geolocalizadas. Además, observando las curvas de aprendizaje en las gráficas de entrenamiento de los modelos, se pensó en una propuesta de mejora añadiendo más características a la matriz de entrada a las redes convolucionales, planteándose la inclusión de más variables que pudiesen ser obtenidas en base a transformaciones sobre los datos existentes para aportar más información al modelo.

4.2. Modelo GTAAF

Después de analizar los resultados ofrecidos del primer prototipo, se propusieron una serie de mejoras, en el modelo, enfocadas a mejorar las debilidades observadas. Estos cambios se estructuraron en una nueva metodología denominada GTAAF (General Model for Traffic Accident Assistance Forecasting) que busca incrementar el rendimiento del prototipo y cuyo principal objetivo es diseñar e implementar un procedimiento de predicción de asistencia de accidentes de tráfico generalizable a cualquier dataset y región.

El principal problema observado en los conjuntos de datos de accidentes de tráfico es que, dependiendo de la región y/o gobierno que los ofrezca, estos disponen de información diferente, debido principalmente al coste que supone obtener ciertos datos y a la naturaleza social de la población. Es por esto que, la implementación de un modelo de predicción de necesidad de asistencia en accidentes de tráfico general, requiere un trabajo de análisis de la categorización de las variables disponibles y cuáles pueden ser influyentes en la necesidad de asistencia.

Con la finalidad de solventar estos problemas y presentar una generalización del modelo que sea independiente de los datos disponibles, la metodología GTAAF propuesta se basa en categorización de las características disponibles individuales dependientes de cada conjunto de datos. Así, 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 sortea las peculiaridades individuales de la disponibilidad de datos de cualquier región.

Para evaluar la eficacia del modelo GTAAF, se compara con otros seis modelos del estado del arte a lo largo de ocho regiones distintas en las mismas condiciones.

En esta sección se va a explicar, 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 primera instancia, las fases de la nueva metodología son asignadas a tres etapas claramente diferenciadas: (en naranja en la Figura 4.3) la fase de Preprocesamiento, donde se contemplan procesos de limpieza de datos, transformación y balanceo de datos, (en gris) la fase de Postprocesado donde se aplican técnicas de transformación para representar los datos de accidentes en formato tabular a formato matricial, y (en azul) 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.

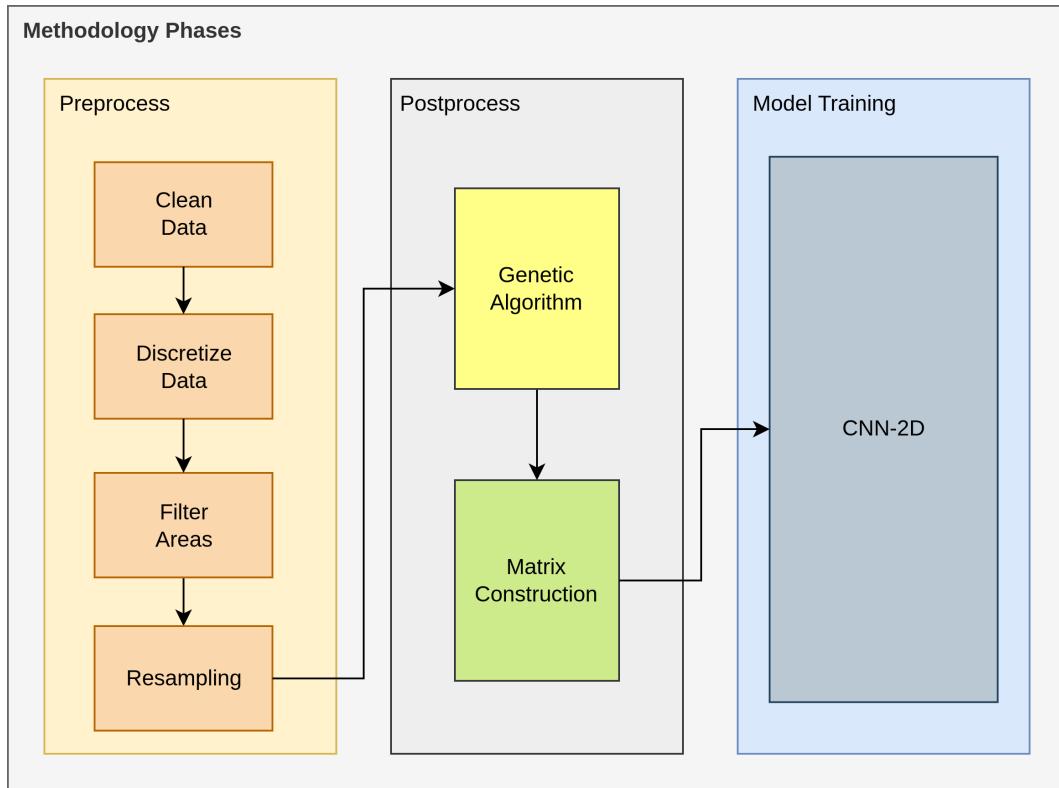


Figura 4.3: Diagrama de flujo de la metodología final con las tres fases simplificadas: preprocessado de datos, postprocesado y entrenamiento del modelo.

En segundo lugar, el concepto de gravedad de los accidentes es reasignado de tres a dos clases, accidentes sin necesidad de asistencia y accidentes con necesidad de asistencia. Esto se debe a que el balance entre lo que aporta y lo que resta el distinguir tres clases de accidentes, se decanta claramente por esto último 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. Por este motivo se distinguen dos clases de accidentes: **con asistencia y sin asistencia**.

Como tercera consideración, para paliar aún más el efecto del desbalanceo presente en los datos se diseña una técnica para balancear el dataset en base a la clase minoritaria aplicando un filtrado de áreas. Esta técnica tiene como objetivo dividir el mapa de la población en celdas, para seleccionar aquellas zonas de la región donde existan ambos tipos de accidentes. Con esto se consigue que la información interpretada por los modelos no se vea condicionada por la naturaleza de los mismos.

Para aportar más información, se añaden transformaciones sobre los datos en

base a las variables ya existentes. Así, para capturar la naturaleza periódica de la hora del accidente se representó mediante dos componentes cíclicas utilizando funciones seno y coseno.

Como quinta variación a considerar, se redefinen las categorías donde son asignadas las características, pasando de cinco a seis categorías finales. Con este cambio, se busca una reorganización que facilite la asignación de características a conceptos más generales que representan estas categorías. Esto implica que las matrices que se construyen pasan de tener dimensión 5×5 a 6×4 , sobre estos conjuntos de datos.

Por otra parte, se han centrado los esfuerzo en desarrollar el modelo convolucional de dos dimensiones CNN-2D. Esto se debió al análisis de entrenamiento de los modelos CNN-1D y CNN-2D, optando por descartar el primero debido a la poca capacidad de generalización sobre el conjunto de validación ofrecido por el modelo unidimensional.

Por último, en un intento de evaluar la generalización del modelo propuesto, se ha probado en distintos datasets y áreas, ampliando los conjuntos de datos sobre los que se aplica la metodología. De esta forma, se incluyendo seis regiones de Reino Unido, una de Australia y Madrid. Además, se ampliaron los modelos del estado del arte contra los que comparar el rendimiento, llegando a seis, SVC, Naive Bayes, Bagging Random Forest, KNN, Regresión Logística y una red neuronal Perceptrón Multicapa.

4.2.1. Preprocesamiento

Esta sección explica las diferentes etapas que componen la fase de preprocesamiento de la metodología GTAAF propuesta. En esta etapa es donde se les aplica transformaciones a los datos para obtener a un conjunto de datos refinado e interpretable para cualquier modelo que trabaje con datos en formato tabular.

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: Diagrama de flujo del preprocesamiento de datos con sus cuatro fases propuestas: (1) limpieza, (2) discretización, (3) filtrado de áreas y (4) resampling.

4.2.1.1. Limpieza

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

50CAPÍTULO 4. CONSTRUCCIÓN DE UN MODELO GENERAL DE PREDICCIÓN DE LA GRAVEDAD

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.

4.2.1.2. Discretización

Los modelos predictivos trabajan con datos numéricos, que son los que son capaces de interpretar, son sobre los que realizan operaciones matemáticas para 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 y descriptiva. Esto se puede intuir de forma clara con el ejemplo de una característica que describa el punto de impacto del accidente, donde los valores que esta variable pudiera tomar se refieren a una descripción cualitativa del punto de impacto del vehículo, como pudieran ser frontal, lateral o por alcance, entre otras. Por este motivo es necesario aplicar un proceso de cuantificación y discretización que busque transformar estos valores descriptivos en valores numéricos, de tal forma que los datos puedan ser interpretados por los modelos. Se busca 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 este trabajo 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.

4.2.1.3. Transformación (Sin/Cos)

Como se ha comentado anteriormente, 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 la 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:

$$\sin((2 \cdot \pi \cdot DaySeconds) / SecondsInDay) \quad (4.1)$$

$$\cos((2 \cdot \pi \cdot DaySeconds) / SecondsInDay) \quad (4.2)$$

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.

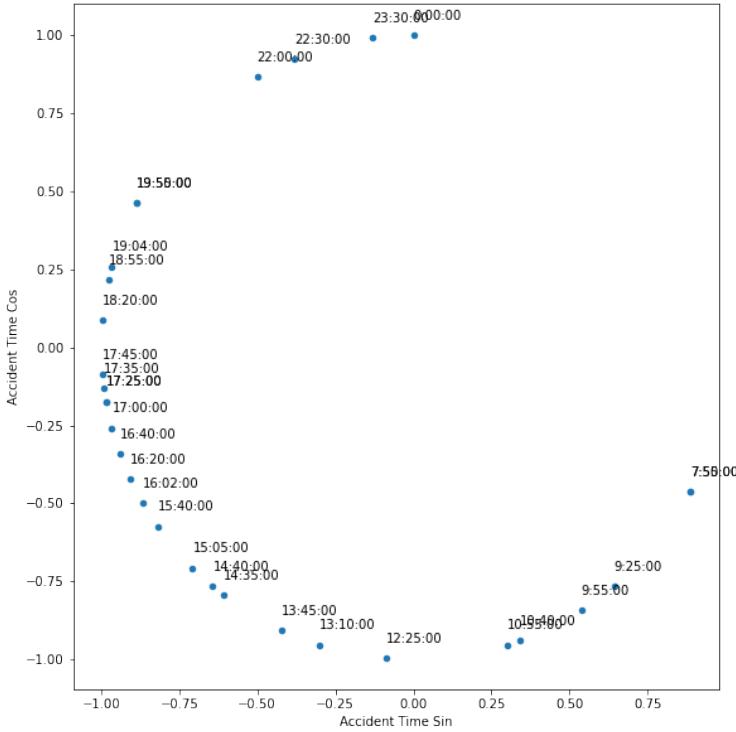


Figura 4.5: Algo así, no es lo definitivo.

4.2.1.4. Filtrado de Áreas

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 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 coexisten 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).

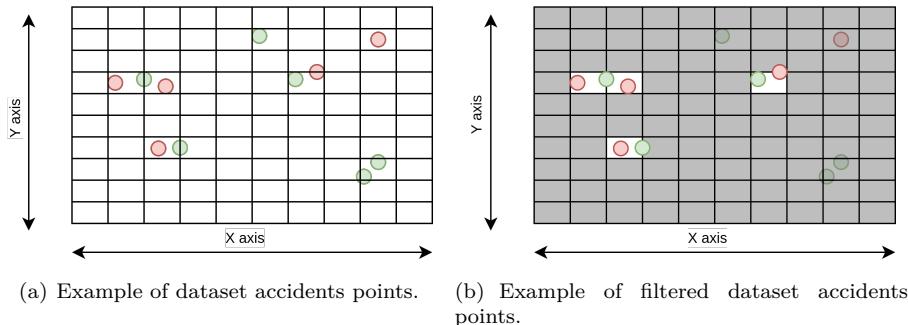


Figura 4.6: Ejemplo de filtrado de áreas. Los puntos verdes representan accidentes que no requieren asistencia, mientras que los puntos rojos representan accidentes que necesitan asistencia.

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.

4.2.1.5. Normalización

En cualquier modelo de inteligencia artificial es imprescindible normalizar los datos. Los modelos predictivos 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 de ellas 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 una característica que pudiera describir la semana dentro del año en el que se ha producido el accidente y, por otra parte, el sexo de la víctima. 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, como Mean Centered (MC), Variable Stability Scaling (VSS) o Min-Max Normalization (MMN), entre otras [?]. En este trabajo, 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.

$$Z = \frac{(X - \mu)}{\sigma} \quad (4.3)$$

4.2.1.6. División Train-Val-Test

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 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 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 este trabajo se ha dividido el conjunto de datos original de cada una de las ciudades mediante el porcentaje 80 % para entrenamiento y 20 % para validación.

4.2.1.7. Resampling

Una vez se disponen de los datos refinados y normalizados, es necesario aplicar algún proceso que logre balancear los datos en función de la clase a la que pertenecen. 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 mas accidentes sin necesidad de asistencia respecto a las que sí.

En el caso de estudio de esta tesis, aplicar técnicas de Undersampling 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, se opta por métodos de aumentado 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. 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.

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.

4.2.2. Postprocesamiento

La segunda fase de la metodología implica transformar los datos refinados y balanceados en matrices interpretables por el modelo GTAAF propuesto. Este proceso implica mapear los atributos de las muestras tabulares en posiciones dentro de estas matrices. Para realizar esto, se hace uso de un método de transformación que toma en consideración la importancia de cada característica dentro del conjunto de datos. El objetivo es posicionarse 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 variables 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 algoritmos evolutivos. 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.

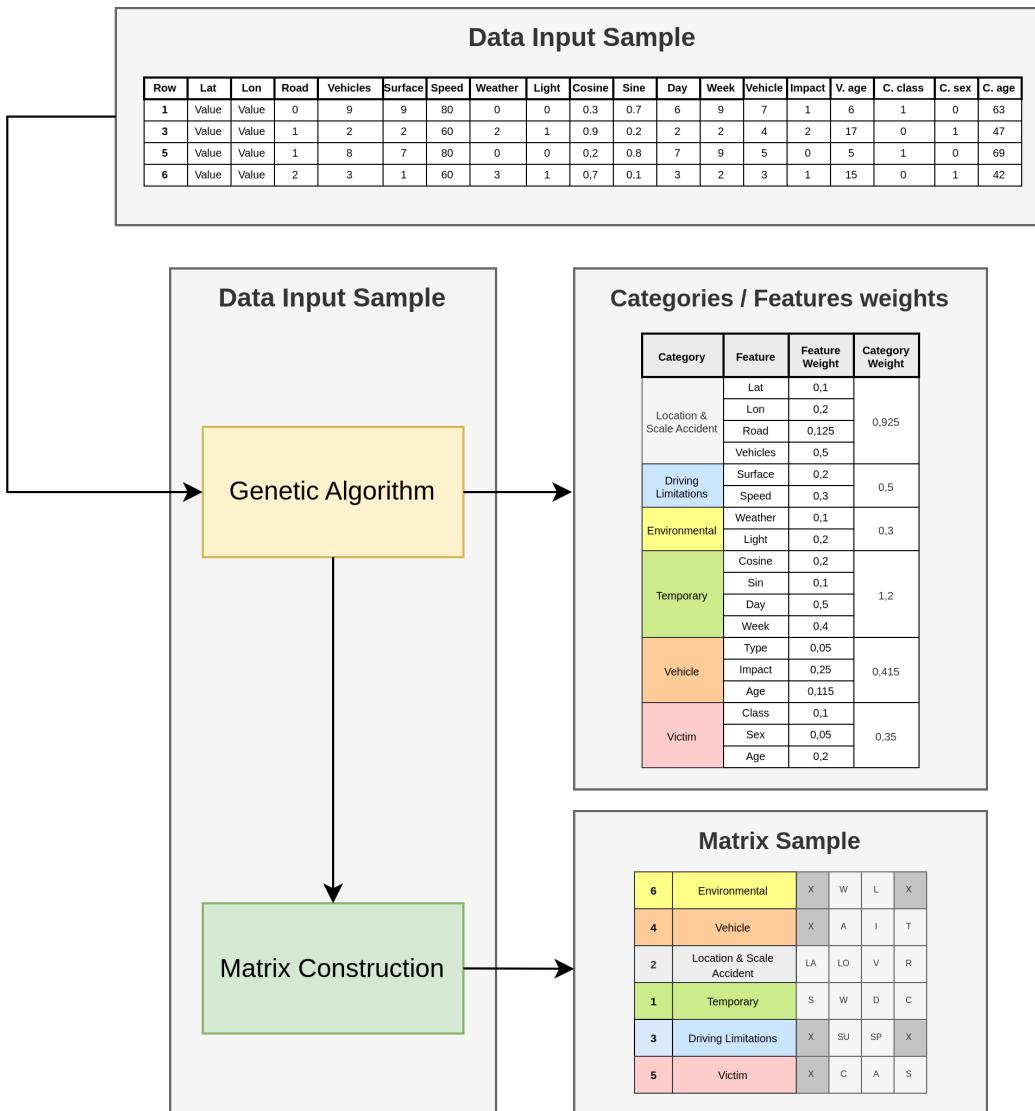


Figura 4.7: Diagrama de flujo del postprocesamiento de datos con las dos fases propuestas: (1) aplicación del algoritmo genético para la optimización de hiperparámetros de XGBoost, y (2) construcción de matrices en base a los pesos de las categorías.

4.2.2.1. Construcción de Matrices

En este trabajo, se presenta un método para transformar datos, inicialmente tabulares, 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 secciones anteriores se ha explicado el funcionamiento del proceso de categorización propuesto, que busca 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 3.3 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 2500 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 este trabajo 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.

4.2.2.2. Algoritmos de medición de importancia de características

Como se ha presentado en la sección 3.6, 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 ensembles de tipo Bagging como los Random Forest o métodos ensembles tipo Boosting.

En esta tesis se trabaja con conjunto de datos desbalanceados, 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 ensembles 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 ensembles 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 *ensembles* 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 el error del modelo mediante la métrica F1-Score resultante de la clasificación de ambas clases de accidentes (sin necesidad de asistencia y con necesidad de asistencia).

Este algoritmo ofrece una serie de hiperparámetros, que permiten configurar el método para maximizar su rendimiento. Del total de hiperparámetros disponibles para su configuración, se escogerán aquellos más relevantes, concretamente la profundidad máxima que puede tomar el árbol, el número de árboles que minimizarán el error de sus predecesores y la tasa de aprendizaje o learning rate. Para ello se aplicarán técnicas de optimización de hiperparámetros basadas en algoritmos evolutivos

4.2.2.3. Algoritmo Genético

Como se ha comentado en la sección 3.5, existen numerosos métodos para optimizar hiperparámetros, cada uno con sus ventajas y desventajas dependiendo del contexto y los datos en el que se apliquen.

Debido a las limitaciones computacionales que supone la combinación de todos los posibles hiperparámetros, el método Grid Search no se adapta adecuadamente al caso de uso contemplado en esta tesis. Por otra parte, siguiendo la línea de probar combinaciones de hiperparámetros sin una evolución en su convergencia, el método Random Search aunque es más eficiente que la búsqueda de cuadrícula, no es idóneo para este caso, ya que no sigue ningún patrón que explote las mejores soluciones que se van obteniendo, siendo estas combinaciones

meramente aleatorias.

Debido a esto, en esta tesis se utilizan algoritmos genéticos para optimizar los hiperparámetros de entrenamiento del algoritmo XGBoost. Este tipo de algoritmos permiten una exploración amplia del espacio de búsqueda de los hiperparámetros óptimos, acentuando además la explotación en soluciones cercanas al óptimo ideal. El algoritmo con los hiperparámetros optimizados XGBoost 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.

4.2.2.4. 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 y por esto es necesario aplicar un método que logre transformar estos datos de manera coherente y eficiente. Existen diferentes enfoques para construir matrices en base a datos tabulares, pero estos enfoques como se ha comentado en la sección 4.2.2.2 sufren de limitaciones aplicados a nuestro caso de uso, ya sea porque necesitan conocimiento del dominio o porque han sido diseñados para un número de características mucho mayor respecto a las disponibles en los conjuntos de datos de accidentes de tráfico. 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 zonas de importancia para los filtros de las redes convolucionales. Este proceso, teniendo en cuenta la categorización inicial que permite aplicar esta metodología a cualquier región, es 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).

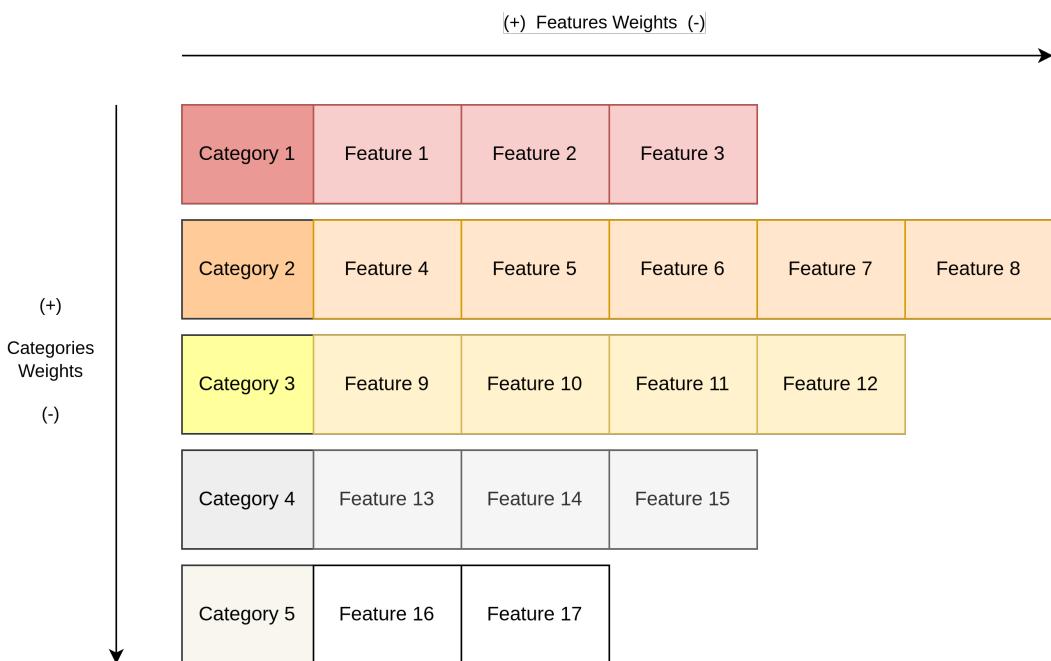


Figura 4.8: Diagrama de la asignación de pesos a las categorías en base a las características.

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 la máxima de que las variables más importantes para los datos se encuentran en las posiciones centrales, como se muestra en la Figura 4.9.

62CAPÍTULO 4. CONSTRUCCIÓN DE UN MODELO GENERAL DE PREDICCIÓN DE LA GRAVEZ

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: Posición de las categorías y características en base a sus pesos.

En la Figura 4.10 se muestra un ejemplo del procedimiento

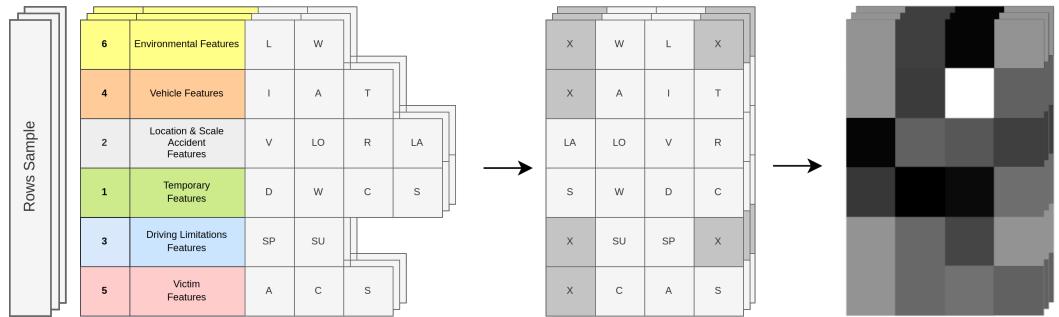


Figura 4.10: Proceso de asignación de características a posiciones de las matrices. Las categorías se asignan a las filas de la matriz en base a su peso; seguidamente, las características de cada categoría se posicionan en las columnas.

4.2.2.5. Diseño del modelo

El nuevo modelo propuesto presenta una arquitectura de cuatro capas convolucionales de dos dimensiones cada una, con un tamaño de kernel de 3×3

y una función de activación ReLU. A la salida de cada capa convolucional se aplica un proceso de *Batch Normalization*.

La primera capa convolucional de la red consta de 64 kernels, la segunda de 512, la tercera de 128 y la cuarta de 256. Estos kernels contienen los pesos que se entrena durante la fase de ajuste del modelo a partir de la salida conocida de los datos etiquetados, aprendiendo qué multiplicaciones en los datos minimizan la función de pérdida definida de la red (entropía cruzada binaria) gracias al proceso de retropropagación. La salida de cada capa convolucional son los mapas de características, que son el resultado de aplicar la multiplicación de estos filtros a su entrada. El paso, o número de unidades que avanzan los kernels, para un mapa de características, es 1. También se aplica relleno en las convoluciones, es decir, si la multiplicación del kernel excede los límites de la matriz, se agregarán ceros a estos límites para realizar la convolución. Los mapas de características resultantes de la última capa pasan a través de una capa de aplanamiento, que transforma los datos a una sola dimensión una vez que han finalizado las convoluciones. Cada uno de estos datos aplanados está interconectado con los 256 nodos definidos de la capa densa (Fully Connected Network). Finalmente, la capa densa está conectada a una capa final con la función de activación Softmax, que da la probabilidad de que cada nueva muestra pertenezca a una de las dos clases. En la figura 4.11 se puede observar, a modo de diagrama, la arquitectura básica de la red.

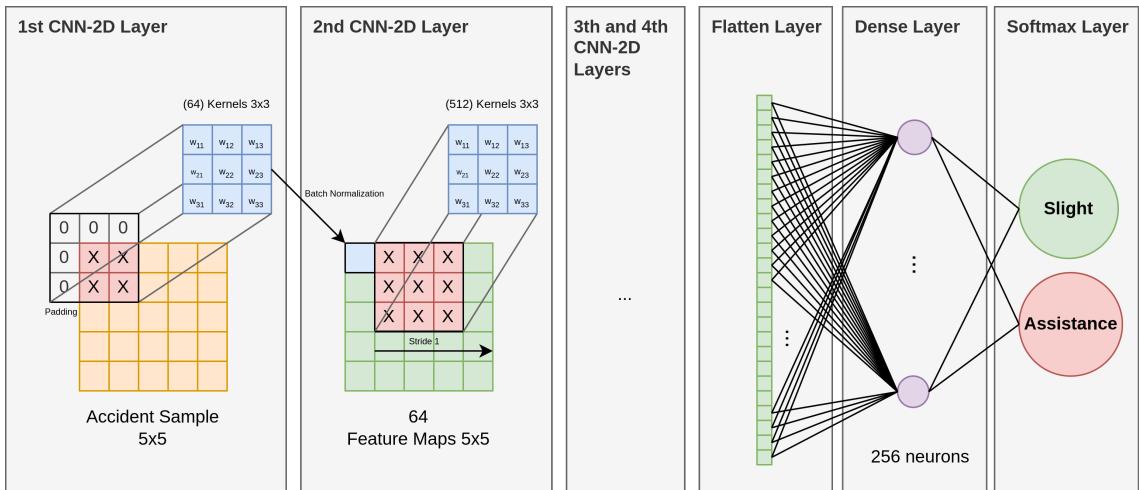


Figura 4.11: Arquitectura del modelo CNN-2D propuesto.

4.3. Evaluación del modelo: Eficiencia y Robustez

Para medir el rendimiento y evaluar la capacidad de generalización de la metodología GTAAF, se compararán los resultados ofrecidos por este respecto a otros seis modelos de clasificación del estado del arte (SVC, Naive Bayes, Random Forest, KNN, Regresión Logística y MLP) a lo largo de datasets de accidentes de tráfico de áreas situadas en distintos lugares del mundo. Concretamente España (Madrid), Reino Unido (Southwark, Manchester, Birmingham, Liverpool, Sheffield y Cornwall), y Australia (estado de Victoria).

Con el objetivo de medir la precisión del modelo en distintos contextos, estas regiones han sido seleccionadas debido a que presentan una alta variabilidad, tanto en los datos que contienen, como la extensión de las regiones y en la densidad de población que presentan. De esta forma es posible evaluar el rendimiento de la metodología distinguiendo entre tres casos de estudio claramente definidos: (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.

Por otra parte, se han realizado ejecutado el modelo GTAAF eliminando características en los datos. En primer lugar, se prescinde aquella que más relevancia tiene para la metodología (mayor peso), en segundo lugar, aquella que menos relevancia tiene (menor peso) y, en un tercer experimento, quitando ambas. Estas pruebas tienen un doble objetivo: evaluar la robustez de la metodología y simular la aplicación del modelo a futuros conjuntos de datos donde no se disponga de valores de todas las características de los conjuntos de datos seleccionados, "eliminando variables significativamente poderosas para los resultados de esta tesis".

Para medir la eficiencia de los modelos se utilizará la métrica F1-Score, ya que es una métrica ampliamente utilizada en problemas de clasificación y que representa una buena aproximación a la generalización del modelo, ya que para su cálculo se tienen en cuenta componentes más básicas como la precisión y el recuerdo, indicadores esenciales para una evaluación robusta de cualquier modelo predictivo.

Capítulo 5

Experimentos y resultados

En esta sección se expone la información técnica de interés relativa a los experimentos y resultados realizados en este trabajo. En primer lugar se presentarán las tecnologías utilizadas durante el transcurso de este trabajo, en segundo lugar se mostrarán los datos sobre los que se han realizado los experimentos. En tercera instancia se exponen los resultados obtenidos por un **prototipo preliminar** tras su ejecución bajo una población específica, Madrid, que sirve como base para los experimentos relativos al cuarto apartado, donde se presentarán los resultados de la metodología final **GTAAF** aplicada a datos de ocho regiones distintas. La finalidad última es validar la generalización de la metodología propuesta a datasets de varias áreas, concretamente Reino Unido (municipio de Southwark, ciudad de Manchester, ciudad de Birmingham, ciudad de Liverpool, ciudad de Sheffield y condado de Cornwall), España (Madrid) y Australia (estado de Victoria).

5.1. Adaptación de los datos

En este punto se presentan los datos sobre los que se han aplicado las metodologías. Los datos escogidos para la validación de la metodología pertenecen a tres regiones distintas, donde cada una de ellas presenta distintas singularidades. El criterio escogido para la selección de estos datos se basa en la búsqueda de la variabilidad a lo largo de distintos contextos y casos de uso con el objetivo de validar que la metodología propuesta es capaz de generalizar a lo largo de circunstancias distintas. Esta evaluación se basará en dos factores principales: la distinta disponibilidad de información en los conjuntos de datos y en función de la densidad de población de las regiones escogidas. A su vez, para este último factor de la densidad de población, distinguiremos 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 metodología GTAAF desarrollada.

Madrid

El primer dataset seleccionado para su adaptación a las metodologías contiene información de accidentes de tráfico sobre la Comunidad de Madrid, comprendidos entre los años 2019 y 2022 a lo largo de toda la ciudad. Estos datos describen los accidentes mediante 18 características para 73 514 registros totales. Este conjunto de datos ha sido extraído del Portal de Datos Abiertos del Ayuntamiento de Madrid [?].

La alta densidad de población de Madrid convierte este conjunto de datos en un caso de estudio de alta concentración de población, en la Tabla 5.1 se pueden observar los datos de geográficos de interés sobre esta ciudad.

Dataset	Km^2	Habitantes	Habitantes por km^2	Ratio de accidentes
Madrid	604,3	3.339.931	5.530	0,006

Cuadro 5.1: Descripción de las propiedades geográficas de la ciudad de Madrid.

Las características originales presentes en este conjunto de datos se encuentran descritas en la Tabla 5.2:

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
Clima	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 5.2: Variables del conjunto de datos y sus descripciones.

En lo que respecta a la variable a predecir, la **Severidad** del accidente, en este conjunto de datos se presentan distintos valores que puede tomar. A continuación se presentan las descripciones que puede tomar esta característica junto con su valor numérico original:

- Atención de emergencia sin posterior admisión hospitalaria: 1.
- Admisión hospitalaria menor o igual a 24 horas: 2.
- Hospitalización por más de 24 horas: 3.
- Fallecido dentro de las 24 horas: 4.
- Atención médica ambulatoria después del accidente: 5.
- Atención médica solo en el lugar del accidente: 6.
- Sin atención médica: 7.

Hay que resaltar que el conjunto de datos de la ciudad de Madrid se utilizará para validar ambas metodologías (preliminar y GTAAF). Por este motivo, la interpretación de los valores que toma la variable de la gravedad del accidente, serán considerados de forma distintas en cada una de ellas, tanto en la metodología preliminar, donde se interpretará la gravedad como tres posibles clases (**leves, severos y fatales**), como en la metodología GTAAF, donde la gravedad es considerada como dos clases (**con necesidad de asistencia y sin necesidad de asistencia**).

Victoria

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 desde el año 2000 y hasta el 2005, contando con un total de 14 123 registros. Este conjunto de datos ha sido obtenido a través del Departamento de Transportes y Planificación del Gobierno del estado de Victoria [?].

Como se puede apreciar en la tabla 5.3, el caso de uso aplicado del estado de Victoria se muestra como una situación de baja concentración de población. Al ser un estado muy extenso, las densidades de población se encuentran en los núcleos urbanos.

Dataset	Km ²	Habitantes	Habitantes por km ²	Ratio de accidentes
Victoria	227.416	5.603.100	25	0,0004

Cuadro 5.3: Descripción de las propiedades geográficas del estado de Victoria.

En este caso, el dataset dispone de 140 características, divididas en varias bases de datos sobre las que ha habido que realizar operaciones de unión para disponer de los datos en su totalidad. Gran parte de estas características contemplan información que no aporta valor descriptivo para los accidentes, ya que se tratan de definiciones de los valores numéricos asignados en otras columnas, identificadores para hacer uniones entre las tablas, etc. Estas tablas contienen información relativa a:

- **Víctimas del accidente** 5.4
- **Descriptores del accidente** 5.5
- **Características de la carretera** 5.6
- **Condiciones atmosféricas** 5.7
- **vehículos implicados** 5.8

- **Evento del accidente 5.9**

- **Localización del accidente 5.10**

- **Características del nodo 5.11**

Luis: Me parece mucho lío meter todas estas tablas aquí, las smetería en anexos.

déjalo, siempre estamos a tiempo de quitarlo

Tabla Víctima	
Atributo	Descripción
ACCIDENT_NO	Número de accidente
PERSON_ID	ID de la persona
VEHICLE_ID	ID del vehículo
SEX	Género
AGE	Edad
AGE_GROUP	Grupo de edad
INJ_LEVEL	Nivel de lesión
INJ_LEVEL_DESC	Descripción del nivel de lesión
SEATING_POSITION	Posición en el asiento
HELMET_BELT_WORN	¿Casco o cinturón de seguridad usado?
ROAD_USER_TYPE	Tipo de usuario de la carretera
ROAD_USER_TYPE_DESC	Descripción del tipo de usuario de la carretera
LICENCE_STATE	Estado de la licencia
PEDEST_MOVEMENT	Movimiento del peatón
POSTCODE	Código postal
TAKEN_HOSPITAL	Hospital al que fue llevado
EJECTED_CODE	Código de expulsión

Cuadro 5.4: Variables del conjunto de datos de Victoria y sus descripciones. Tabla relativa a las víctimas.

Tabla Accidente	
Atributo	Descripción
ACCIDENT_NO	Número de accidente
ACCIDENT_DATE	Fecha del accidente
ACCIDENT_TIME	Hora del accidente
ACCIDENT_TYPE	Tipo de accidente
ACCIDENT_TYPE_DESC	Descripción del tipo de accidente
DAY_OF_WEEK	Día de la semana
DAY_OF_WEEK_DESC	Descripción del día de la semana
DCA_CODE	Código DCA
DCA_DESCRIPTION	Descripción DCA
DIRECTORY	Directorio
EDITION	Edición
PAGE	Página
GRID_REFERENCE_X	Referencia de cuadrícula (X)
GRID_REFERENCE_Y	Referencia de cuadrícula (Y)
LIGHT_CONDITION	Condición de luz
LIGHT_CONDITION_DESC	Descripción de la condición de luz
NODE_ID	ID de nodo
NO_OF_VEHICLES	Número de vehículos
NO_PERSONS	Número de personas
NO_PERSONS_INJ_2	Número de personas heridas levemente
NO_PERSONS_INJ_3	Número de personas heridas gravemente
NO_PERSONS_KILLED	Número de personas fallecidas
NO_PERSONS_NOT_INJ	Número de personas no heridas
POLICE_ATTEND	Asistencia policial presente?
ROAD_GEOMETRY	Geometría de la carretera
ROAD_GEOMETRY_DESC	Descripción de la geometría de la carretera
SEVERITY	Severidad del accidente
SPEED_ZONE	Zona de velocidad

Cuadro 5.5: Variables del conjunto de datos de Victoria y sus descripciones. Tabla Accidente.

Tabla Características de Carretera	
Atributo	Descripción
ACCIDENT_NO	Número de accidente
SURFACE_COND	Condición de la superficie
SURFACE_COND_DESC	Descripción de la condición de la superficie
SURFACE_COND_SEQ	Secuencia de la condición de la superficie

Cuadro 5.6: Variables del conjunto de datos de Victoria y sus descripciones. Tabla Características de Carretera.

Tabla Características Atmosféricas	
Atributo	Descripción
ACCIDENT_NO	Número de accidente
ATMOSPH_COND	Condición atmosférica
ATMOSPH_COND_SEQ	Secuencia de la condición atmosférica
ATMOSPH_COND_DESC	Descripción de la condición atmosférica

Cuadro 5.7: Variables del conjunto de datos de Victoria y sus descripciones. Tabla Características Atmosféricas.

Tabla Vehículo	
Atributo	Descripción
ACCIDENT_NO	Número de accidente
VEHICLE_ID	ID del vehículo
VEHICLE_YEAR_MANUF	Año de fabricación del vehículo
VEHICLE_DCA_CODE	Código DCA del vehículo
INITIAL_DIRECTION	Dirección inicial
ROAD_SURFACE_TYPE	Tipo de superficie de la carretera
ROAD_SURFACE_DESC	Descripción del tipo de superficie de la carretera
REG_STATE	Estado de registro
VEHICLE_BODY_STYLE	Estilo del cuerpo del vehículo
VEHICLE_MAKE	Marca del vehículo
VEHICLE_MODEL	Modelo del vehículo
VEHICLE_POWER	Potencia del vehículo
VEHICLE_TYPE	Tipo de vehículo
VEHICLE_TYPE_DESC	Descripción del tipo de vehículo
VEHICLE_WEIGHT	Peso del vehículo
CONSTRUCTION_TYPE	Tipo de construcción
FUEL_TYPE	Tipo de combustible
NO_OF_WHEELS	Número de ruedas
NO_OF_CYLINDERS	Número de cilindros
SEATING_CAPACITY	Capacidad de asientos
TARE_WEIGHT	Peso tara
TOTAL_NO_OCCUPANTS	Número total de ocupantes
CARRY_CAPACITY	Capacidad de carga
CUBIC_CAPACITY	Capacidad cúbica
FINAL_DIRECTION	Dirección final
DRIVER_INTENT	Intención del conductor
VEHICLE_MOVEMENT	Movimiento del vehículo
TRAILER_TYPE	Tipo de remolque
VEHICLE_COLOUR_1	Color del vehículo 1
VEHICLE_COLOUR_2	Color del vehículo 2
CAUGHT_FIRE	Incendio del vehículo
INITIAL_IMPACT	Impacto inicial
LAMPS	Lámparas
LEVEL_OF_DAMAGE	Nivel de daño
OWNER_POSTCODE	Código postal del propietario
TOWED_AWAY_FLAG	Indicador de remolque
TRAFFIC_CONTROL	Control de tráfico
TRAFFIC_CONTROL_DESC	Descripción del control de tráfico

Cuadro 5.8: Variables del conjunto de datos de Victoria y sus descripciones. Tabla Vehículo.

Tabla Evento Accidente	
Atributo	Descripción
ACCIDENT_NO	Número de accidente
EVENT_SEQ_NO	Número de secuencia del evento
EVENT_TYPE	Tipo de evento
EVENT_TYPE_DESC	Descripción del tipo de evento
VEHICLE_1_ID	ID del vehículo 1
VEHICLE_1_COLL_PT	Punto de colisión del vehículo 1
VEHICLE_1_COLL_PT_DESC	Descripción del punto de colisión del vehículo 1
VEHICLE_2_ID	ID del vehículo 2
VEHICLE_2_COLL_PT	Punto de colisión del vehículo 2
VEHICLE_2_COLL_PT_DESC	Descripción del punto de colisión del vehículo 2
PERSON_ID	ID de la persona
OBJECT_TYPE	Tipo de objeto
OBJECT_TYPE_DESC	Descripción del tipo de objeto

Cuadro 5.9: Variables del conjunto de datos de Victoria y sus descripciones. Tabla Evento Accidentes.

Tabla Localización de Accidente	
Atributo	Descripción
ACCIDENT_NO	Número de accidente
NODE_ID	ID del nodo
ROAD_ROUTE_1	Ruta de la carretera 1
ROAD_NAME	Nombre de la carretera
ROAD_TYPE	Tipo de carretera
ROAD_NAME_INT	Nombre de la carretera (intersección)
ROAD_TYPE_INT	Tipo de carretera (intersección)
DISTANCE_LOCATION	Distancia de la ubicación
DIRECTION_LOCATION	Dirección de la ubicación
NEAREST_KM_POST	Kilómetro de poste más cercano
OFF_ROAD_LOCATION	Ubicación fuera de la carretera

Cuadro 5.10: Variables del conjunto de datos de Victoria y sus descripciones. Tabla Localización de Accidente.

Tabla Nodo	
Atributo	Descripción
ACCIDENT_NO	Número de accidente
NODE_ID	ID del nodo
NODE_TYPE	Tipo de nodo
AMG_X	Coordenada AMG-X
AMG_Y	Coordenada AMG-Y
LGA_NAME	Nombre del área del gobierno local (LGA)
REGION_NAME	Nombre de la región
DEG_URBAN_NAME	Nombre del área urbana
LAT	Latitud
LONG	Longitud
POSTCODE_N0	Código postal

Cuadro 5.11: Variables del conjunto de datos de Victoria y sus descripciones. Tabla de nodos.

En lo que respecta a la variable predictiva que indica la gravedad de las víctimas en los accidentes, este conjunto de datos categoriza las lesiones en cuatro clases:

- Fatales: 1.
- Graves: 2.
- Otro tipo de lesiones: 3.
- Sin lesiones: 4.

Reino Unido

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

Este dataset dispone, en su versión original, con 77 características que describen información acerca de los accidentes, las víctimas y los vehículos implicados en ellos, en bases de datos separadas. Por lo que para obtener la información total desglosada por víctima se realizan operaciones de unión entre ellas. En la tabla 5.12 se puede observar la cantidad de registros resultantes tras realizar estas operaciones en cada una de las regiones seleccionadas.

Distribución de datos de Reino Unido	
Región	Número de muestras
Southwark	30.214
Manchester	53.341
Birmingham	119.910
Liverpool	54.452
Sheffield	49.466
Cornwall	37.846

Cuadro 5.12: Número original de muestras de las regiones del Reino Unido.

En la tabla 5.13 se observan datos geográficos acerca de cada una de las regiones escogidas de Reino Unido, cada una de ellas presentando una densidad de población distinta y un ratio de accidentes por habitante diferente.

Dataset	Km^2	Habitantes	Habitantes por km^2	Ratio de accidentes
Southwark	29	317.256	10.997	0,006
Manchester	116	547.627	4.721	0,006
Birmingham	268	1.144.919	429	0,007
Liverpool	116	500.500	4315	0,007
Sheffield	368	534.500	1.452	0,006
Cornwall	3.563	569.578	160	0,004

Cuadro 5.13: Descripción de las propiedades geográficas las regiones de Reino Unido.

Luis, pregunta: ¿Cómo podemos describir los datos como se ha hecho en el apartado anterior de Madrid, siendo las variables 77? **MANU:** Las pondria todas. Al fin y al cabo, eso te sirve para justificar uno de los puntos fuertes del modelo: la robustez de la categorización en el caso de que falten características

JOSE: opino lo mismo, se ponen todas

Luis: Hecho, pero yo lo metería en el anexo, esto me parece que queda un poco forzado.

En las siguientes tablas se muestran las características originales disponibles en el conjunto de datos de Reino Unido, cada una de estas tablas incluye información descriptiva de los accidentes 5.14, de los vehículos implicados 5.15 y de las víctimas 5.16.

Tabla Accidente	
Atributo	Descripción
accident_index	Índice del accidente
accident_year	Año del accidente
accident_reference	Referencia del accidente
location_easting_osgr	Coordenada este de la ubicación (OSGR)
location_northing_osgr	Coordenada norte de la ubicación (OSGR)
longitude	Longitud
latitude	Latitud
police_force	Fuerza policial
accident_severity	Gravedad del accidente
number_of_vehicles	Número de vehículos
number_of_casualties	Número de víctimas
date	Fecha
day_of_week	Día de la semana
time	Hora
local_authority_district	Distrito de la autoridad local
local_authority_ons_district	Distrito ONS de la autoridad local
local_authority_highway	Carretera de la autoridad local
first_road_class	Clase de la primera carretera
first_road_number	Número de la primera carretera
road_type	Tipo de carretera
speed_limit	Límite de velocidad
junction_detail	Detalle de la intersección
junction_control	Control de la intersección
second_road_class	Clase de la segunda carretera
second_road_number	Número de la segunda carretera
pedestrian_crossing_human_control	Control humano del cruce peatonal
pedestrian_crossing_physical_facilities	Instalaciones físicas del cruce peatonal
light_conditions	Condiciones de iluminación
weather_conditions	Condiciones meteorológicas
road_surface_conditions	Condiciones de la superficie de la carretera
special_conditions_at_site	Condiciones especiales en el sitio
carriageway_hazards	Peligros en la calzada
urban_or_rural_area	Área urbana o rural
did_police_officer_attend_scene_of_accident	Asistió un oficial de policía a la escena del accidente
trunk_road_flag	Indicador de carretera principal
lsoa_of_accident_location	LSOA de la ubicación del accidente

Cuadro 5.14: Variables del conjunto de datos de Reino Unido y sus descripciones.
Tabla Accidente.

Tabla Vehículo	
Atributo	Descripción
accident_index	Índice del accidente
accident_year	Año del accidente
accident_reference	Referencia del accidente
vehicle_reference	Referencia del vehículo
vehicle_type	Tipo de vehículo
towing_and_articulation	Remolque y articulación
vehicle_manoeuvre	Maniobra del vehículo
vehicle_direction_from	Dirección del vehículo desde
vehicle_direction_to	Dirección del vehículo hacia
vehicle_location_restricted_lane	Ubicación del vehículo en carril restringido
junction_location	Ubicación en la intersección
skidding_and_overturning	Derrape y vuelco
hit_object_in_carriageway	Objeto golpeado en la calzada
vehicle_leaving_carriageway	Vehículo abandonando la calzada
hit_object_off_carriageway	Objeto golpeado fuera de la calzada
first_point_of_impact	Primer punto de impacto
vehicle_left_hand_drive	Vehículo de conducción izquierda
journey_purpose_of_driver	Propósito del viaje del conductor
sex_of_driver	Sexo del conductor
age_of_driver	Edad del conductor
age_band_of_driver	Grupo de edad del conductor
engine_capacity_cc	Capacidad del motor (cc)
propulsion_code	Código de propulsión
age_of_vehicle	Edad del vehículo
generic_make_model	Modelo genérico del vehículo
driver_imd_decile	Decil de IMD (Índice de Marginación Deprivación) del conductor
driver_home_area_type	Tipo de área de residencia del conductor

Cuadro 5.15: Variables del conjunto de datos de Reino Unido y sus descripciones.
Tabla Vehículo.

Tabla Víctima	
Atributo	Descripción
accident_index	Índice del accidente
accident_year	Año del accidente
accident_reference	Referencia del accidente
vehicle_reference	Referencia del vehículo
casualty_reference	Referencia de la víctima
casualty_class	Clase de la víctima
sex_of_casualty	Sexo de la víctima
age_of_casualty	Edad de la víctima
age_band_of_casualty	Grupo de edad de la víctima
casualty_severity	Gravedad de la víctima
pedestrian_location	Ubicación del peatón
pedestrian_movement	Movimiento del peatón
car_passenger	Pasajero de automóvil
bus_or_coach_passenger	Pasajero de autobús o autocar
pedestrian_road_maintenance_worker	Trabajador de mantenimiento de carreteras peatonal
casualty_type	Tipo de víctima
casualty_home_area_type	Tipo de área de residencia de la víctima
casualty_imd_decile	Decil de IMD (Índice de Marginación Deprivación) de la víctima

Cuadro 5.16: Variables del conjunto de datos de Reino Unido y sus descripciones.
Tabla Víctima.

Sobre este conjunto de datos, la variable a predecir puede tomar tres valores distintos que contemplan las consecuencias del accidente en las víctimas:

- Fatal: persona fallecida debido a consecuencias del accidente, categorización 1.
- Grave: víctimas que han sufrido consecuencias moderadas como fracturas, cortes profundos o lesiones internas, categorización 2.
- Leve: víctimas que han tenido consecuencias livianas y de fácil recuperación, entre estos casos pueden destacar los esguinces, moratones o shock emocional, categorización 3.

Los detalles de lo que engloban estas clases puede consultarse en la web de Departamento de Transportes de Reino Unido [?].

5.2. Prototipo: resultados preliminares

En esta sección se exponen los resultados sobre la metodología preliminar, aplicándola sobre los datos de la ciudad de Madrid y presentada en el artículo [?]. Como se ha comentado en la sección de metodología 4.1, este desarrollo tenía como objetivo crear un método que transformase datos tabulares en datos matriciales para aplicar dos redes neuronales convolucionales, de una y dos dimensiones. Este método tenía el fin predecir la severidad en los accidentes de tráfico divididos en tres clases claramente diferenciadas (**leves, severos y fatales**).

Para lograr esta clasificación, se realizaron transformaciones sobre los valores de la variable a predecir, la lesividad del accidentado, reasignando las 7 clases originales disponibles en el conjunto de datos a tres, en función de su gravedad. Esta asignación se ha realizado en base al siguiente criterio:

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.

Para describir los resultados, se acompañará cada etapa por la que ha pasado los datos a través de la ejecución de la metodología.

Limpieza

En un primer lugar, se analizaron las características de los datos en bruto que describían información de los accidentes. Gracias a esto, se observó que existían conjuntos de variables que contenían valores atípicos (outliers) y/o valores nulos. Es por esto por lo que se requería de aplicar un proceso de limpieza

que ofreciese un dataset refinado e interpretable por los distintos métodos. Para ello se eliminaron los registros con valores atípicos y aquellos que presentaban valores en blanco, resultando un dataset final con 65 158 registros, un 11,38 % de pérdida de información respecto al original (73 514).

Además, se eliminaron aquellas características que no aportaban valor a los modelos predictivos, como el Identificador del accidente. Por otra parte, se analizó la dependencia entre cada par de variables mediante matrices de correlación que muestran la intensidad con la que una variable es dependiente del resto. Remarcar que los coeficientes de correlación varían entre -1 y 1 , indicando la magnitud y dirección de esta dependencia. Después, se ha aplicado un umbral de correlación entre variables del $\pm 0,44$, lo que quiere decir que aquellas que presentasen una correlación superior a este valor son excluidas. En la figura 5.1 se muestra la matriz de correlación resultante tras eliminar las características que superasen este límite de dependencia entre sí, pasando a ser 14 de las 18 originales.



Figura 5.1: Matriz de correlación entre las variables del conjunto de datos.

Discretización

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 que representaban los valores de cada característica. Por otra parte, las variables originales *Positivo en Drogas* y *Positivo en Alcohol* se unieron en una nueva característica *Alcohol o Drogas*, con el objetivo de recoger esta información en un único campo para no descartar muchos registros que presentaban valores nulos en ambas columnas. En la Tabla 5.17 se muestra la discretización realizada para cada una de las variables.

Si pones dos columnas en vez de cuatro quedará mejor

Característica	Tipificación	Característica	Tipificación
Gravedad	0: Leve (1, 2, 5, 6, 7)	Tiempo	1: Noche (6 PM - 6 AM)
	1: Grave (3)		2: Día (6 AM - 6 PM)
	2: Fatal (4)	Distrito	En base al orden de aparición
X	Posición Coordenada UTM X		1: Colisión frontal
Y	Posición Coordenada UTM Y		2: Colisión trasera
Tipo de Carretera	1: Estacionamiento	Tipo de Accidente	3: Choque lateral
	2: Aeropuerto		4: Colisión contra obstáculo fijo
	3: Parque		5: Choque en cadena
	4: Túnel		6: Atropello a peatón
	5: Zona industrial		7: Colisión frontal
	6: Pista		8: Otro
	7: Rotonda		9: Salida de la carretera
	8: Glorieta		10: Vuelco de vehículo
	9: Puerta		11: Atropello a animal
	10: Puente		12: Caída
Condiciones Meteorológicas	1: Soleado	Vehículo	En base al orden de aparición
	2: Nublado		1: Conductor
	3: Lluvia ligera	Persona	2: Pasajero
	4: Lluvia intensa		3: Peatón
	5: Granizo		
	6: Nevando	Edad	1: Menos de 18 años
	7: Desconocido		2: De 18 a 25 años
Género	1: Masculino		3: De 25 a 65 años
	2: Femenino		4: Más de 65 años
	3: Desconocido		5: Desconocida
		Alcohol o Drogas	1: Sí / 2: No

Cuadro 5.17: Asignación numérica de las variables del conjunto de datos.

División Train-Val-Test

Como es habitual en el diseño de los modelos predictivos, es común dividir los datos en subconjuntos para el aprendizaje de los modelos y su evaluación real sobre registros que nunca ha visto. Del total de muestras resultantes del proceso de filtrado (65 158) se asignan el 80 % de ellas para el conjunto de entrenamiento (54 211), resultando en 53 213 accidentes leves, 984 graves y 50 fatales. El 20 % de los datos restantes se asignaron al conjunto validación o test (10 911), concretamente 10 640 accidentes leves, 256 graves y 15 fatales.

Resampling

Una vez aplicado el proceso de limpieza de datos y elección de características del dataset, se analizó la distribución final de los datos de entrenamiento 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 53 213 accidentes leves, 984 graves y 50 fatales. Esto, como se ha comentado en la sección 3.4 se convierte en un problema para los modelos de clasificación, ya que en estos casos tienden a predecir las nuevas muestras como aquellas que pertenecen a la clase mayoritaria del conjunto de entrenamiento. Para paliar este problema se aplicó la técnica de remuestreo Borderline SMOTE-II, con el objetivo de generar más muestras de accidentes pertenecientes a clases minoritarias (Grave y Fatal) hasta llegar a la mayoritaria (Leves), evitando que el modelo se sobreajuste. Una vez aplicado el algoritmo, se disponen de 53 213 muestras de la clase leve, 53 213 de la clase grave y 53 213 de la clase fatal, haciendo un total de 159 639 registros en el nuevo conjunto de datos balanceado.

Normalización

En la Figura 5.4 se muestra un ejemplo de la aplicación de la normalización de datos en base a la técnica Z-Score, que transformará los datos a una distribución normal con media 0 y desviación 1 de cada valor. En la Tabla 5.2 se observa un registro de datos discretizado del dataset y en la Figura 5.3 se observa esta misma muestra tras haber aplicado el proceso de normalización. Este proceso se aplica para cada una de las muestras de tal forma que la dimensión de los datos quede bajo la misma magnitud, para poder ser interpretarse eficientemente por los modelos.

Característica	Valor
hora	2
tipo carretera	19
distrito	14
tipo accidente	1
estado meteorológico	1
tipo vehículo	4
tipo persona	1
rango edad	3
sexo	1
drogas alcohol positivo	2
vehículos 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 vehículo	-0.1621
tipo persona	-0.5316
rango edad	0.2129
sexo	-0.7004
drogas alcohol positivo	0.1488
vehículos implicados	-1.4591
coordenada x utm	-0.0524
coordenada y utm	0.0081

Figura 5.3 Muestra de accidente tipificada.**Figura 5.4** Ejemplo de normalización de una muestra del dataset.

Una vez se dispone de todos los datos normalizados, pueden ser utilizados para el entrenamiento de cualquier modelo predictivo.

Categorización

Como se ha comentado, las redes neuronales convolucionales (CNN) aprenden patrones utilizando matrices como datos de entrada. Así pues, como disponemos de datos tabulares, se hace necesario transformarlos en matrices.

Uno de los requisitos de esta transformación era asignar cada característica a una categoría del dataset. Sobre este conjunto de datos de Madrid, las variables eran asignadas a 5 categorías en base a la información de la que se

disponía: Características del accidente, Condiciones de la carretera, Condiciones meteorológicas, Características del vehículo y Características del conductor. En la Tabla 5.18 se observa la categorización de cada característica en función de la información que describen.

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

Cuadro 5.18: 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 XGBoost a través del algoritmo genético mediante la maximización de la métrica F1-Score de los datos de entrenamiento. Los hiperparámetros del algoritmo XGBoost a optimizar fueron: profundidad máxima del árbol, peso mínimo de los hijos y el ratio de aprendizaje. El algoritmo genético se configuró para optimizar esta función durante 80 generaciones, con un máximo de 40 individuos en la población y los 10 mejores en cada iteración eran seleccionados para reproducirse mediante una estrategia de cruce aleatorio.

La figura 5.5 muestra la evolución de los tres hiperparámetros del XGBoost 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, donde no se observan modificaciones a partir de esta generación.

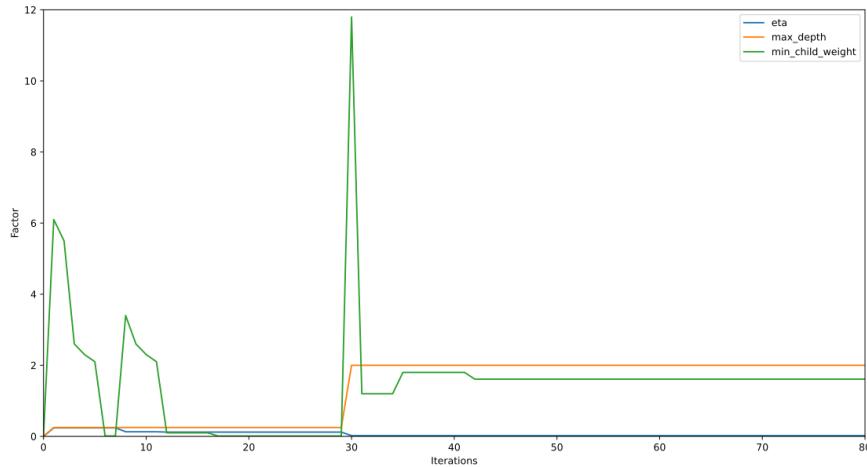


Figura 5.5: Evolución de los hiperparámetros del XGBoost a lo largo de las generaciones.

En la tabla 5.19 se observa el valor tomado por el mejor individuo resultante de la ejecución de las distintas generaciones para cada uno de los hiperparámetros del XGBoost.

Hiperparámetro	Valor
Profundidad Máxima	2
Peso Mínimo de Hijos	1.6
ETA	0.007

Cuadro 5.19: Valores optimizados de los parámetros después de aplicar el algoritmo genético.

Con la configuración de hiperparámetros mostrada en la Tabla 5.19 se entrena el algoritmo XGBoost para obtener el peso de todas las características del dataset. así, en la Tabla 5.20 se muestra el peso asignado a cada una de las características individuales, donde la columna peso de la Categoría muestra el peso de cada categoría como suma de los peso de cada una de las características individuales que la componen.

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

Cuadro 5.20: Ejemplo con los pesos de todas las características estudiadas, así como los pesos de las cinco categorías.

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 Figura 5.6 se observa un ejemplo de un registro transformado a formato matricial una vez aplicado el algoritmo de construcción haciendo uso de la importancia de las características de la tabla 5.20. Igualmente, en 5.7 se observa la representación en imagen de escala de grises de dicha matriz.

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 5.6 Matriz resultante tras la transformación de un registro a formato matricial.

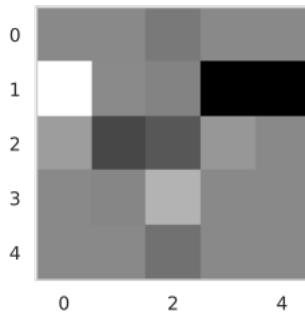


Figura 5.7 Imagen de la matriz en escala de grises.

Figura 5.8: Ejemplo de construcción matricial de una muestra normalizada del dataset.

Entrenamientos

Una vez que las matrices han sido construidas, se describe la red convolucional utilizada en el modelo propuesto. En esta sección se presenta la información resultante de los entrenamientos realizados para los dos modelos convolucionales, de una y dos dimensiones, con los que trabaja esta metodología preliminar. De esta forma se analizará la evolución de la función de pérdida en cada uno de estos modelos.

Las figuras 5.9 y 5.10 muestran la evolución de la métrica de puntuación F1-score a lo largo de las 100 ejecuciones para las redes neuronales convolucionales 1D y 2D. Al visualizar la convolución unidimensional (Figura 5.9), se pudo 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.

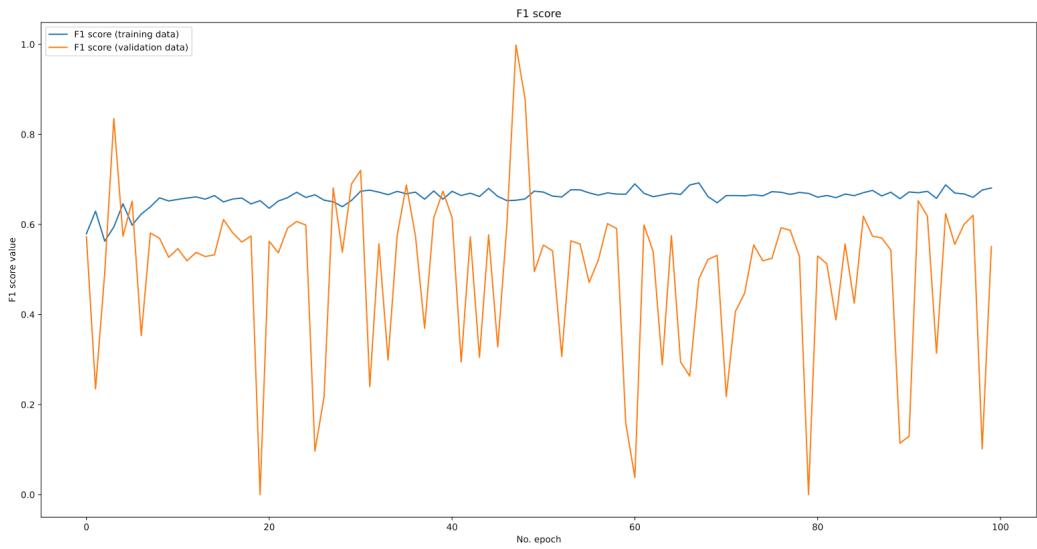


Figura 5.9: Evolución del F1-score de la red neuronal convolucional unidimensional (CNN-1D) en los conjuntos de entrenamiento y test.

Por otro lado, la Figura 5.10 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.

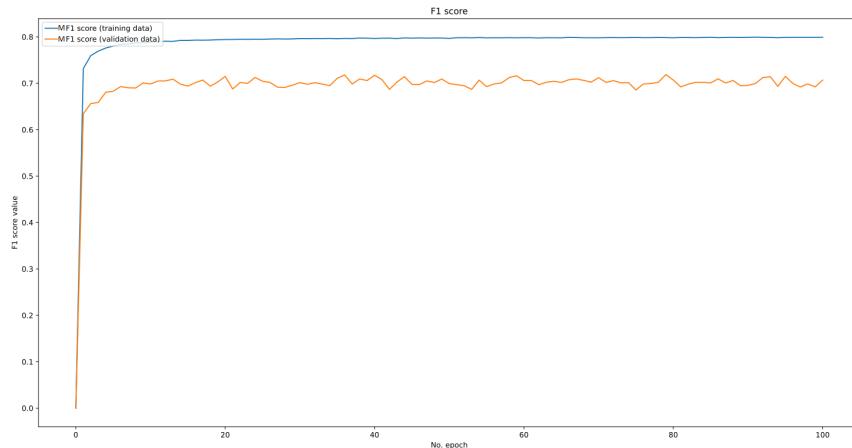


Figura 5.10: Evolución del F1-score de la red neuronal convolucional 2D (CNN-2D) en los conjuntos de entrenamiento y test.

Evaluación

Para evaluar el rendimiento de la metodología preliminar y los modelos propuestos, se realizó una comparación con tres modelos del estado del arte, Gaussian Naive Bayes (GNB), Support Vector Classifier (SVC) y K-Nearest Neighbor (KNN). En esta sección se utilizarán las métricas obtenidas tras la ejecución de los cinco modelos sobre el conjunto de datos de test (CNN-1D, CNN-2D, GNB, SVC y KNN), además se muestran los resultados de predicción basados en matrices de confusión.

Las Tablas 5.21 y 5.22 detallan las métricas resultantes de la clasificación de las redes para los conjuntos de entrenamiento y prueba respectivamente. Se observó que, para el conjunto de entrenamiento, el modelo CNN-1D obtenía un mejor F1-score en la clasificación de todas las clases de accidentes en comparación con la red CNN-2D. Sin embargo, al analizar las métricas de los datos de test, el modelo que presentaba mejor F1-score para accidentes leves y graves es CNN-2D, con 0,950 y 0,148 respectivamente, mientras que en accidentes fatales ambas redes ofrecían el mismo rendimiento con un 0,004 de F1-Score.

Métrica/Gravedad	CNN-1D			CNN-2D		
	Leve	Grave	Fatal	Leve	Grave	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 5.21: Métricas sobre el conjunto de entrenamiento para los modelos CNN-1D y CNN-2D.

Métrica/Gravedad	CNN-1D			CNN-2D		
	Leve	Grave	Fatal	Leve	Grave	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 5.22: Métricas sobre el conjunto de test para los modelos CNN-1D y CNN-2D.

Luis: Percepción me da la sensación de que hay mucho contenido aquí. La forma de exponer los resultados (primero los de las CNNs y luego los modelos del estado del arte) provoca que esto sea muy largo cuando en mi opinión, estos resultados no tienen mucha importancia. Lo reestructuramos o qué pensais?

Jose: Tienes razón, se me está haciendo un poco pesado de leer, lo que significa que deberíamos reducirlo. ¿Quitamos lo referente a las matrices de confusión y lo centramos en los resultados del F1-score?

Luis: Me parece buena idea, he quitado lo relativo a las matrices. Además creo que deberíamos reducir todo lo referente a los resultados preliminares..

La información con la que se evalúan los modelos, es decir, las métricas de clasificación resultantes, para el conjunto de test se muestran en la Tabla 5.23 para cada una de las clases predichas. 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 GNB es un poco mejor.

Si analizamos la métrica de Precisión en los cinco modelos comparados, se puede observar que el modelo que presenta el mejor promedio para las clases Slight es la Red Neuronal Convolutacional 1D (CNN-1D) con 0,984, seguida por la Red Neuronal Convolutacional 2D (CNN-2D) y el modelo KNN con 0,982. Además, la CNN-2D también ofrece la mejor métrica para accidentes graves, 0,097, con una gran diferencia respecto al modelo que le sigue, KNN con 0,042. En cuanto a los accidentes fatales, ambos modelos CNN-1D y CNN-2D tienen un valor similar, obteniendo 0,002.

Con respecto a la métrica de Recall, el mejor promedio para las clases Slight es la Red Neuronal Convolutacional 2D (CNN-2D) con 0,919, seguida por el modelo KNN con 0,689. Además, el modelo GNB ofrece la mejor métrica para accidentes graves, 0,699. En accidentes fatales, CNN-2D 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 el 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 Slight es la Red Neuronal Convolutacional 2D (CNN-2D), alcanzando 0,950, muy por encima del siguiente modelo KNN, que ofrece un valor de 0,810. Además, la CNN-2D también ofrece la mejor métrica para accidentes graves, 0,148, alcanzando el doble de rendimiento en comparación con el modelo que le sigue, KNN con 0,076. En cuanto a los accidentes fatales, los modelos con mejor clasificación son tanto CNN-1D como CNN-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 con respecto a la métrica F1-score, que es una combinación de Precisión y Recall.

Métrica/Gravedad	GNB			SVC			KNN		
	Leve	Graves	Fatal	Leve	Graves	Fatal	Leve	Graves	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 5.23: Métricas de clasificación sobre el conjunto de test de los modelos GNB, SVC y KNN.

Jose: ¿Quitamos lo de las matrices de confusión en este apartado?

Luis: QUITADO

Conclusiones

Luis: esto es realmente lo mismo que la introducción al modelo GTAAF pero más justificado en base al análisis de resultados, no sé si estaría bien repetirse.

Jose: Aquí pondría esto

Analizando los resultados finales de esta metodología preliminar se propusieron una serie de mejoras a desarrollar para crear un modelo general en la severidad de los accidentes de tráfico, aportando así un posible valor a las administraciones públicas para asignar recursos médicos en los accidentes de tráfico.

En primer lugar, se detectó que la decisión de disgregar la clasificación de la gravedad de los accidentes en tres clases provocaba un efecto conflictivo en la clasificación de los modelos debido a que dos de las clases en las que se dividía la Severidad de los accidentes eran minoritarias, lo que penalizaba el aprendizaje del modelo. Por otra lado, al intentar generalizar el modelo, la adaptación de la metodología para poder aplicarla en cualquier conjunto de datos era prioritaria. De esta forma, cualquier información que pudiese ser relevante en la predicción de la gravedad de los accidentes, orientado a la falta de disponibilidad de datos en cada región, debería ser añadida. Finalmente se comprobó que había datos ya existentes de los que se podían sacar más características importantes. Por ejemplo, la hora del accidente no estaba representada debido a su naturaleza cíclica. Todos estas debilidades hicieron necesario la implantación de soluciones que aportaran generalidad al modelo definitivo.

Jose: Los siguientes párrafos los pasaría al modelo definitivo, al principio de la sección siguiente

Como se ha comentado, al analizar los resultados de la metodología preliminar se detectaron una serie de mejoras que era necesario estudiar si se quería crear un modelo general que fuera posible utilizarlo por las administraciones públicas.

En primer lugar, se detectó que la decisión de disgregar la clasificación de la gravedad de los accidentes en tres clases provocaba un efecto conflictivo en la clasificación de los modelos. Al tener dos clases minoritarias con tan pocas muestras, el efecto del *resmapling* datos no presentaba el rendimiento esperado, ya que la diferencia entre dos clases con tan pocas muestras penalizaba el aprendizaje del modelo. Además, el valor que aporta la distinción entre accidentes graves y fatales es insuficiente, ya que en ambos casos la asistencia de los organismos de emergencia es necesaria. Por este motivo, con la finalidad de aumentar la utilidad de la metodología y paliar el efecto de superposición, se agruparon las de severidad graves y fatales en una sola, **necesidad de asistencia**.

Luis: Cuidado con este párrafo que en Madrid nos quedamos con 5x4 y estamos diciendo que tenemos 6 categorías..

Jose: Lo que hay que insistir es en que tenemos siempre matrices 6x4 y si hay alguna categoría de la que no se tienen datos, pues una fila de ceros pero el tamaño de la matriz es siempre la misma

Por otra parte, la adaptación de la metodología para poder aplicarla en cualquier población era prioritaria, y cualquier información que pudiese ser relevante en la predicción de la gravedad de los accidentes, orientado a la falta de disponibilidad de datos en cada región, debería ser añadida. Para llegar a esto en futuros conjuntos de datos, se redefinieron las categorías para que contemplan información que pudieran describir conceptos más genéricos y de fácil asignación. Se propuso un replanteamiento de las categorías, pasando de ser 5 a 6: (1) Información del accidente, (2) limitaciones en la conducción, (3) factores ambientales, (4) información temporal, (5) información del vehículo y (6)

información de la víctima. Esto implica que la dimensionalidad de la matriz de características crezca verticalmente, pudiendo agregar más información para aumentar su dimensión en distintos conjuntos de datos, pasando de ser de 5×5 a 6×5 .

Por otra parte, se planteó aplicar transformaciones sobre los datos ya existentes para aumentar el número de características en base a la información ya seleccionada. Por ejemplo, una de las principales debilidades del modelo anterior es que la hora del accidente no estaba representada en base a su naturaleza cíclica. Para discretizar mejor esta variable se aplicaron transformaciones en forma de senos y cosenos para contemplar esta información, que gracias a la propuesta de la nueva categorización, esta, junto a muchas otras, podría ser incluida.

En base al análisis de resultados, se planteó abordar el problema del desbalanceo de los datos añadiendo un componente más. El número de muestras originales evidenciaban que aún con el método de *resampling* no era posible conseguir una generalización en las predicciones. Para este efecto sobre los datos originales, se propuso un método que filtrase los datos de forma que se redujesen el número de muestras de accidentes leves en comparación con el resto, mediante un sistema de selección de áreas donde coexistiesen todos los tipos de accidente.

Por último, en base a estos resultados, el modelo CNN-2D presentaba mejores métricas sobre el entrenamiento de los datos, evidenciando que esta arquitectura era capaz de capturar patrones más representativos, por lo que se propuso centrar los esfuerzos en el desarrollo de este modelo, descartando así la CNN-1D.

5.3. Configuración GTAAF

En esta sección se presentan los resultados de la metodología GTAAF sobre los tres conjuntos de datos descritos en la sección 5.1. Como se ha comentado en apartados anteriores, esta metodología es una evolución del prototipo anterior, por lo que presentará diferencias respecto a su predecesora. Como consideración importante a tener en cuenta en este punto es que una de las principales diferencias de esta metodología es la clasificación de la gravedad del accidente en **dos clases: sin necesidad de asistencia y con necesidad de asistencia**. En la siguiente tabla 5.24 se muestra la asignación del valor de la gravedad del accidente original de cada uno de los tres conjuntos de datos a estas dos clases.

Luis: Inquietud Viendo lo de Madrid.... Lo hemos hecho mal... Qué hacemos?

Jose: no hacemos nada, callarnos como...

Asignación de valores de Asistencia		
Región	Valor Final	Valor Original
Madrid	Sin Asistencia	Sin atención médica Admisión hospitalaria menor o igual a 24 horas Atención médica solo en el lugar del accidente Atención médica ambulatoria después del accidente Atención de emergencia sin posterior admisión hospitalaria
	Con Asistencia	Hospitalización por más de 24 horas Fallecido dentro de las 24 horas
Reino Unido	Sin Asistencia	Leve
	Con Asistencia	Grave Fatal
Victoria	Sin Asistencia	Sin lesiones
	Con Asistencia	Otro tipo de lesiones Grave Fatal

Cuadro 5.24: Asignación de los valores de la gravedad de los accidentes en los tres conjuntos de datos a las dos clases a predecir (sin necesidad de asistencia y asistencia).

Como se ha comentado en las secciones anteriores, la variación entre la disponibilidad de información es un problema común entre distintos conjuntos de datos. En función de las condiciones sociales y económicas de poblaciones alrededor de todo el mundo, la disponibilidad de la información es variable. El esfuerzo que supone la recogida de ciertos datos puede suponer un coste más alto en comparación a otras poblaciones, lo que se traduce en una heterogeneidad en la información disponible entre distintos conjuntos de datos. Para paliar este problema y con el objetivo de proponer un proceso consistente, escalable, e invariante a las condiciones particulares de cada población, en esta metodología se propone agrupar la información disponible en un máximo de 6 categorías, las cuales engloban conceptos donde es fácilmente asignar los datos asumiendo que estos no tienen por qué estar siempre disponibles ni que deban representar exactamente la misma información entre distintos conjuntos de datos:

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 relativas 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.

Este enfoque permite aplicar esta metodología incluso en casos donde no se dispongan de las seis categorías propuestas, proponiendo un modelo tolerante a información inconsistente entre distintos conjuntos de datos. Para visualizar esta casuística, en la Figura 5.11 se presentan las características disponibles en cada uno de los conjuntos de datos bajo la categorización propuesta con el objetivo de mostrar la variabilidad de información que puede existir entre los datos de distintas regiones. 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
<u>Location & Scale</u>	Latitude Longitude Road Class Number of Vehicles	Latitude Longitude District Number of Vehicles	Latitude Longitude Type of Accident Place Number of Vehicles
<u>Driving Limitations</u>	Road Surface Speed Limit	X X	Road Surface Speed Limit
<u>Environmental</u>	Weather Conditions Lighting Conditions	Weather Conditions X	Weather Conditions Lighting Conditions
<u>Temporary</u>	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
<u>Vehicle</u>	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
<u>Victim</u>	Casualty Class Casualty Sex Casualty Age X	Casualty Class Casualty Sex Casualty Age Alcohol/Drugs Positive	Casualty Class Casualty Sex Casualty Age X

Figura 5.11: Clasificación de variables. Los campos mostrados en amarillo representan características de la misma naturaleza pero difieren en la granularidad de los datos. Además, las características ausentes en comparación con otros conjuntos de datos están resaltadas en rojo.

A continuación se detallarán aquellas partes comunes entre estos tres conjuntos de datos y las principales diferencias entre ellos.

Partes comunes entre los datos en base a la categorización

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 suelen presentarse en forma de información espacial y temporal del accidente, como es la localización, las condiciones climáticas en el momento del suceso o la hora y fecha en la que se ha producido, como es el caso entre estos tres conjuntos de datos. Por otra parte, como es lógico, existe información básica que puede ser recogida rápidamente observando la

escena del accidente, como es la localización del mismo, el número de vehículos implicados, las condiciones climáticas, la hora, el tipo de vehículo, el punto de impacto y las características del accidentado.

Principales diferencias entre los datos

Cada una de las regiones ofrece información distinta en cada conjunto de datos. En este punto se analizarán las principales diferencias entre los datos en base a la categorización propuesta:

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 de 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 Reino Unido y de Victoria. Analizando la Figura 5.11 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 aérea 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, bulevares, entre otras.~~ Este campo contempla el distrito dentro de Madrid en el que se ha producido el accidente, y es interpretado de forma numérica en función del orden de aparición de los distritos en los datos. Al ser una característica que ofrece información sobre la localización del accidente, será incluida en la categoría de Location & Scale Accident.

Como diferencia más notable sobre el conjunto de datos de Madrid es que **dispone de características para contemplar 5 categorías** respecto a las 6 de los otros dos conjuntos de datos.

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.

Respecto a la variable *First Point of Impact*, este campo indica el tipo de colisión del vehículo, es decir, contra qué objeto ha impactado el vehículo, en comparación contra otros conjuntos de datos que indican también de qué parte del mismo ha impactado primero.

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, túnel, etc. por lo que irá asignada a la categoría Location & Scale Accident.

5.3.1. Limpieza

En la tabla 5.30 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 propuestas.

Distribución de Datos			
Reino Unido			
Región	Asistencia	Original	Limpieza
Southwark	No	27.105	11.065
	Sí	3.109	2.703
Manchester	No	48.771	24.110
	Sí	4.570	1.885
Birmingham	No	108.723	64.147
	Sí	11.187	6.191
Liverpool	No	49.291	25.936
	Sí	5.161	2.276
Sheffield	No	43.579	25.622
	Sí	5.887	2.703
Cornwall	No	32.994	20.803
	Sí	4.852	2.842
España			
Región	Asistencia	Original	Limpieza
Madrid	No	72.042	63.853
	Sí	1.472	1.305
Australia			
Región	Asistencia	Original	Limpieza
Victoria	No	7.064	5.556
	Sí	7.059	6.024

Cuadro 5.25: Comparación de la distribución de datos tras el proceso de limpieza.

Como se puede observar, tras el proceso de limpieza, cada uno de los conjuntos de datos se ve afectado por una pérdida de registros significativa. Para Reino Unido, de un total de 301,650 registros, se obtienen finalmente 216,219, lo que supone un 28,3 % de pérdida de información respecto al conjunto de datos inicial. En la ciudad de Madrid, 73,514 registros en bruto iniciales acaban resultando en 65,158 con información interpretable, conllevando un 11,38 % de pérdida de información. Por último, en el estado de Victoria, de un total de 14,123 registros iniciales, resultan 11,580 finales, asumiendo un 18 % de pérdida respecto al conjunto de datos inicial.

5.3.2. Discretización

En esta sección se expone la discretización escogida para las variables de los tres conjuntos de datos seleccionados para una mayor comprensión del tratamiento de las características en cada región.

Luis: Pregunta aquí la idea es poner todas las tablas de discretización, pero no sé qué comentar en cada una de ellas si hay que poner más texto. Tal vez se puede poner en anexos y mencionarlo en este párrafo de arriba que se discretizan y si se quiere más info que se vayan a anexos?

Reino Unido

La tabla 5.26 muestra la asignación de valores aplicada a los datos de las regiones de Reino Unido (Southwark, Manchester, Birmingham, Liverpool, Sheffield y Cornwall).

Categoría	Característica	Valor	Descripción
Localización y Escala del accidente	Latitud	Número Real	Coordenada Este (OSGR)
	Longitud	Número Real	Coordenada Norte (OSGR)
	Clase de Carretera	0	Autopista
		1	A(M)
		2	A
		3	B
		4	C
		5	No clasificada
	Número de Vehículos	0-N	Dependiendo del número de vehículos involucrados
Limitaciones de Conducción	Superficie de la Carretera	0	Seca
		1	Mojada / Húmeda
		2	Nieve
		3	Helada / Hielo
		4	Inundación
	Límite de Velocidad	0-70	Dependiendo del límite de velocidad (mph) de la carretera
Ambiental	Condiciones Meteorológicas	0	Buen tiempo sin viento fuerte
		1	Lluvia sin viento fuerte
		2	Nieve sin viento fuerte
		3	Buen tiempo con viento fuerte
		4	Lluvia con viento fuerte
		5	Nieve con viento fuerte
		6	Niebla o neblina
		7	Otro
	Condiciones de Iluminación	0	Luz del día: luces de la calle presentes
		1	Oscuridad: sin iluminación en la calle
		2	Oscuridad: luces de la calle presentes y encendidas
		3	Oscuridad: luces de la calle presentes pero apagadas
		4	Oscuridad: iluminación de la calle desconocida
Temporal	Hora Coseno	Número Real	Representación cosenaloidal de la hora del accidente
	Hora Seno	Número Real	Representación senoidal de la hora del accidente
	Día de la Semana	0-6	Dependiendo del día de la semana en que ocurrió el accidente
	Semana del Año	0-52	Dependiendo de la semana en que ocurrió el accidente
Vehículo	Tipo de Vehículo	0-17	Dependiendo del peso del vehículo
	Primer Punto de Impacto	0	No impactó
		1	Frente
		2	Parte trasera
		3	Lado contrario
		4	Lado cercano
		5	Desconocido (autoreportado)
	Edad del Vehículo	0-N	En orden de antigüedad del vehículo
	Clase de Víctima	0	Conductor/Motociclista
		1	Pasajero
		2	Peatón
Víctima	Sexo de la Víctima	0	Masculino
		1	Femenino
	Edad de la Víctima	0	Menor de 18 años
		1	Entre 18 y 25 años
		2	Entre 25 y 65 años
		3	Mayor de 65 años

Cuadro 5.26: Discretización propuesta de las variables para el conjunto de datos de Reino Unido.

Madrid

La tabla 5.27 muestra la asignación de valores aplicada a los datos de las regiones de Madrid. Como cuestión a considerar en este caso es que a la variable Distrito se le aplica un valor numérico en función del orden de aparición, sin que este represente ninguna importancia incremental.

Categoría	Característica	Valor	Descripción
Ubicación & Escala del Accidente	Latitud	Número Real	Sistema de coordenadas cartesianas
	Longitud	Número Real	Sistema de coordenadas cartesianas
	Distrito	0-X	Número de distrito (Anexo 1*)
	Número de Vehículos	0-N	Dependiendo del número de vehículos involucrados
Ambiental	Condiciones Meteorológicas	0	Despejado
		1	Nublado
		2	Lluvia ligera
		3	Lluvia intensa
		4	Granizo
		5	Nevando
		6	Desconocido
Temporal	Hora Coseno	Número Real	Representación cosenoidal de la hora del accidente
	Hora Seno	Número Real	Representación senoidal de la hora del accidente
	Día de la Semana	0-6	Dependiendo del día de la semana en que ocurrió el accidente
	Semana del Año	0-52	Dependiendo de la semana en que ocurrió el accidente
Vehículo	Primer Punto de Impacto	0-17	Dependiendo del peso del vehículo
		0	Colisión frontal - tamaño
		1	Colisión trasera
		2	Choque lateral
		3	Colisión contra obstáculo fijo
		4	Choque en cadena
		5	Atropello a peatón
		6	Colisión frontal
		7	Otro
		8	Abandonando la carretera
		9	Vuelco del vehículo
		10	Atropello a animal
		11	Caida
Víctima	Clase de Víctima	0	Conductor/Motociclista
		1	Pasajero
	Sexo de la Víctima	2	Peatón
		0	Masculino
	Edad de la Víctima	1	Femenino
		0	Menor de 18 años
		1	Entre 18 y 25 años
		2	Entre 25 y 65 años
	Alcohol/Drogas Positivo	3	Mayor de 65 años
		0	No
		1	Sí

Cuadro 5.27: Discretización propuesta de las variables para el conjunto de datos de Madrid.

Victoria

La tabla 5.28 muestra la discretización del valor las características que pueden tomar los registros de accidentes en el Estado de Victoria.

Categoría	Característica	Valor	Descripción	
Ubicación & Escala del Accidente	Latitud	Número Real	Coordenada Este OSGR	
	Longitud	Número Real	Coordenada Norte OSGR	
	Clase de Carretera	0-58	En orden de aparición	
	Número de Vehículos	0-N	Dependiendo del número de vehículos involucrados	
Limitaciones de Conducción	Superficie de la Carretera	0	Seco	
		1	Mojado / Húmedo	
		2	Fangosa	
		3	Nevada	
		4	Hielo	
		5	Otro	
Ambiental	Condiciones Meteorológicas	0-70	Dependiendo del límite de velocidad (mph) de acuerdo a la condición	
		0	Despejado	
		1	Lluvia	
		2	Nevando	
		3	Niebla	
		4	Humo en el ambiente	
		5	Polvo en el ambiente	
		6	Fuertes Vientos	
	Condiciones de Iluminación	7	Otro	
		0	Día	
		1	Amanecer	
		2	Oscuridad: luces de la calle encendidas	
		3	Oscuridad: luces de la calle apagadas	
		4	Oscuridad: sin luces de la calle	
Temporal	Temporal	5	Oscuridad: iluminación de la calle descolorida	
		6	Otro	
		7	Otro	
		8	Otro	
Vehículo	Vehículo	Hora Coseno	Número Real	Representación cosenoidal de la hora del día
		Hora Seno	Número Real	Representación senoidal de la hora del día
		Día de la Semana	0-6	Dependiendo del día de la semana en que haya ocurrido
		Semana del Año	0-52	Dependiendo de la semana en que haya ocurrido
	Vehículo	Tipo de Vehículo	0-17	Dependiendo del peso del vehículo
			0	Colisión con vehículo
			1	Atropello a peatón
			2	Atropello a animal
		Primer Punto de Impacto	3	Colisión contra obstáculo fijo
			3	Colisión contra otro objeto
			5	Vuelco del vehículo (sin colisión)
			6	Caída desde de un vehículo en movimiento
			7	Ninguna colisión y ningún objeto golpeado
Víctima	Víctima	Edad del Vehículo	0-N	En orden de antigüedad del vehículo
			0	Peatón
			1	Conductor
			2	Pasajero
			3	Motociclista
			4	Ciclista
			5	Desconocido
	Víctima	Sexo de la Víctima	0	Masculino
			1	Femenino
			0	Menor de 18 años
			1	Entre 18 y 25 años
			2	Entre 25 y 65 años
			3	Mayor de 65 años
		Edad de la Víctima	4	Otro
			5	Otro

Cuadro 5.28: Discretización propuesta de las variables para el conjunto de datos de Victoria.

5.3.3. Filtrado de áreas

Para ilustrar los parámetros con los que se aplica el filtrado de áreas, se presenta la Tabla 5.29, donde se muestra para cada población el número de áreas proyectadas resultante de la elección del tamaño de las ventanas en función de los valores X,Y. La elección de estos tamaños variará función de la extensión y densidad de población, tomando valores de ventana más grandes en poblaciones más dispersas y tamaños más pequeños cuando la densidad de población es más 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.

División de Áreas			
Reino Unido			
Región	Eje	Número de Áreas	Tamaño de Área
Southwark	X	529	10
	Y	487	20
Manchester	X	791	14
	Y	1.069	20
Birmingham	X	3.519	12
	Y	1.557	17
Liverpool	X	2.107	12
	Y	717	21
Sheffield	X	1.896	12
	Y	1.115	18
Cornwall	X	10.090	15
	Y	5.597	19
España			
Región	Eje	Número de Áreas	Tamaño de Área
Madrid	X	5.241	5
	Y	4.444	7
Australia			
Región	Eje	Número de Áreas	Tamaño de Área
Victoria	X	4.931	145
	Y	5.241	97

Cuadro 5.29: Número de áreas resultante tras la definición de su tamaño para cada región.

Una vez se establecen las dimensiones de las ventanas de tamaño X,Y se aplica el filtrado para cada región, donde el número de muestras de la clase mayoritaria se ve considerablemente reducido respecto a la minoritaria, buscando

obtener un conjunto de datos más balanceado. La Tabla 5.30 muestra el número de registros originales para cada población y el número de registros resultante tras aplicar el filtrado.

Distribución de Datos			
Reino Unido			
Región	Asistencia	Limpieza	Filtrado
Southwark	No	11.065	4.251
	Sí	2.703	1.256
Manchester	No	24.110	4.548
	Sí	1.885	1.466
Birmingham	No	64.147	4.092
	Sí	6.191	2.063
Liverpool	No	25.936	3.640
	Sí	2.276	1.192
Sheffield	No	25.622	2.060
	Sí	2.703	1.638
Cornwall	No	20.803	2.191
	Sí	2.842	2.020
España			
Region	Assistance	Cleaned	Filtered
Madrid	No	63.840	2.601
	Sí	1.305	1.286
Australia			
Region	Assistance	Cleaned	Filtered
Victoria	No	5.556	2.065
	Sí	6.024	2.649

Cuadro 5.30: Distribución de datos tras el proceso de filtrado para cada una de las regiones.

Una vez aplicado el filtrado de áreas los conjuntos de datos resultantes presentan un desbalanceo considerablemente menor. Para el dataset de Reino Unido se ha conseguido reducir, de media, una desproporción de las clases del 90,2% y 9,8% de los registros de las clases Sin Asistencia y Con Asistencia respectivamente, al 53,6% y 46,4%. Para la ciudad de Madrid 98% y 2% al 67% y 33%. Para la ciudad de Victoria de un 48% y 52% al 43,8% 56,2%, por lo que sobre esta última, el proceso de remuestreo de clases afectará a la clase Sin necesidad de Asistencia, siendo el único caso en el que esta es mayoritaria.

5.3.4. Resampling

Se ha seleccionado el 80 % de los datos refinados como conjunto de entrenamiento, y 20 % restante como conjunto de validación o test. En la tabla ?? se muestra la distribución de estos datos junto con el número total de muestras de entrenamiento sobre el que se ha aplicado el proceso de aumentado de datos mediante SMOTE-II.

En la tabla 5.31 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.

Distribución de Datos					
Reino Unido					
Región	Asistencia	Filtrado	Entrenamiento	Test	Oversampled (Entrenamiento)
Southwark	No	4.251	3.400	851	3.400
	Sí	1.256	1.004	252	3.400
Manchester	No	4.548	3.638	910	3.638
	Sí	1.466	1.172	294	3.638
Birmingham	No	4.092	3.273	819	3.273
	Sí	2.063	1.650	413	3.273
Liverpool	No	3.640	2.912	728	2.912
	Sí	1.192	953	239	2.912
Sheffield	No	2.060	1.648	412	1.648
	Sí	1.638	1.310	328	1.648
Cornwall	No	2.191	1.752	439	1.752
	Sí	2.020	1.616	404	1.752
España					
Región	Asistencia	Filtrado	Entrenamiento	Test	Oversampled (Entrenamiento)
Madrid	No	2.601	2.080	521	2.080
	Sí	1.286	1.028	258	2.080
Australia					
Región	Asistencia	Filtrado	Entrenamiento	Test	Oversampled (Entrenamiento)
Victoria	No	2.065	1.652	413	2.119
	Sí	2.649	2.119	530	2.119

Cuadro 5.31: Distribución de datos para las ciudades seleccionadas. La columna Assistance representa si el accidente ha requerido de asistencia o no, las dos clases objetivo de este documento. Filtered indica el número de muestras disponibles tras el proceso de filtrado. La columna Train representa el 80 % de las muestras de entrenamiento seleccionadas del total de los datos filtrados. La columna Test muestra el 20 % de los datos utilizados para la futura validación de los modelos. Finalmente Oversampled engloba el número de muestra tras aplicar el aumentado de datos sobre el conjunto de entrenamiento de cada población mediante la técnica SMOTE-II para la clase minoritaria.

5.3.5. Normalización

Luis: Comentario esto directamente es que lo quitaba

En la Tabla 5.32 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 5.32: 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.

5.3.6. Algoritmo Genético

En la tabla 5.33 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, entrenado con dicha configuración de hiperparámetros, 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 mediante cruce

por punto aleatorio de las soluciones. 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).

Hiperparámetro	Valor
Población	50
Padres Emparejados	10
Generaciones	50
Índice de Cruzamiento	Aleatorio
Probabilidad de Mutación	0.4
Función Heurística	Puntuación <i>F1-Score</i> del <i>XGBoost</i>

Cuadro 5.33: Configuración de hiperparámetros del algoritmo genético.

La Tabla 5.34 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 *Mutación* 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.

Hiperparámetro	Límite	Mínimo	Máximo
ETA	Inicial	0.01	1
	Mutación	-0.2	0.2
Profundidad Máxima	Inicial	1	25
	Mutación	-3	3
Peso Mínimo de Hijos	Inicial	0.01	20
	Mutación	-4	4

Cuadro 5.34: Límites de inicialización y mutación de los hiperparámetros del algoritmo genético.

La configuración de estos parámetros, tanto los iniciales como los de mutación se han escogido en base a resultados experimentales, donde se ha priorizado la maximización de la métrica *F1-Score* entre distintos experimentos.

En la Tabla 5.35 se pueden observar los hiperparámetros óptimos resultantes de la ejecución del algoritmo genético, donde para cada región se observa la tasa de aprendizaje de los árboles (*ETA*), la profundidad máxima de los árboles y el peso mínimo de los hijos para decidir la separación a través del nivel de los árboles.

Valores Resultantes del Algoritmo Genético			
Reino Unido			
Región	ETA	Profundidad Máxima	Peso Mínimo de Hijos
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
España			
Región	ETA	Profundidad Máxima	Peso Mínimo de Hijos
Madrid	0.01	1	0.01
Australia			
Región	ETA	Profundidad Máxima	Peso Mínimo de Hijos
Victoria	0.6	25	0.01

Cuadro 5.35: Hiperparámetros resultantes del *XGBoost* después de ejecutar el algoritmo genético.

5.3.7. Construcción de matrices

Una vez obtenidos los pesos de las categorías para cada población mediante el algoritmo XGBoost optimizado con hiperparámetros, los registros tabulares de cada una de las poblaciones son convertidos a matrices, siendo estas interpretables por el modelo convolucional propuesto CNN-2D.

Como los conjuntos de datos utilizados disponen de distintas características, el número de categorías disponibles varía entre ellos y por tanto la dimensionabilidad de las matrices resultantes es variable entre las distintas poblaciones. En el caso del conjunto de datos de Madrid se disponen de 5 de las 6 categorías propuestas al no disponer de información sobre las limitaciones de la conducción, por lo que las matrices de entrada para el modelo convolucional serán de 5x4 mientras que para los datos de Reino Unido y Victoria serán de 6x4 al disponer de características que pueden ser englobadas en las 6 categorías.

Luis: Inquietud aquí no sé qué más poner.

5.4. Evaluación

En esta sección se presentarán los resultados del modelo convolucional tras la ejecución de la metodología GTAAF sobre las ocho poblaciones propuestas en esta tesis. Para evaluar la generalización del modelo en distintos contextos,

los resultados del modelo convolucional se compararán con otros seis modelos del estado del arte. Los resultados se exponen para cada conjunto de datos, donde se comenzará con Madrid, en segundo lugar se analizarán los resultados de Victoria y en último lugar las seis poblaciones de Reino Unido, cada una con una densidad distinta de habitantes. Los dos primeros casos (Madrid y Australia) presentan condiciones de densidad de población opuestas, alta y baja respectivamente.

5.4.1. Comparativas

Madrid

Este primer caso presenta una situación de alta densidad. En la figura 5.12 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 5.12(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 5.12(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.

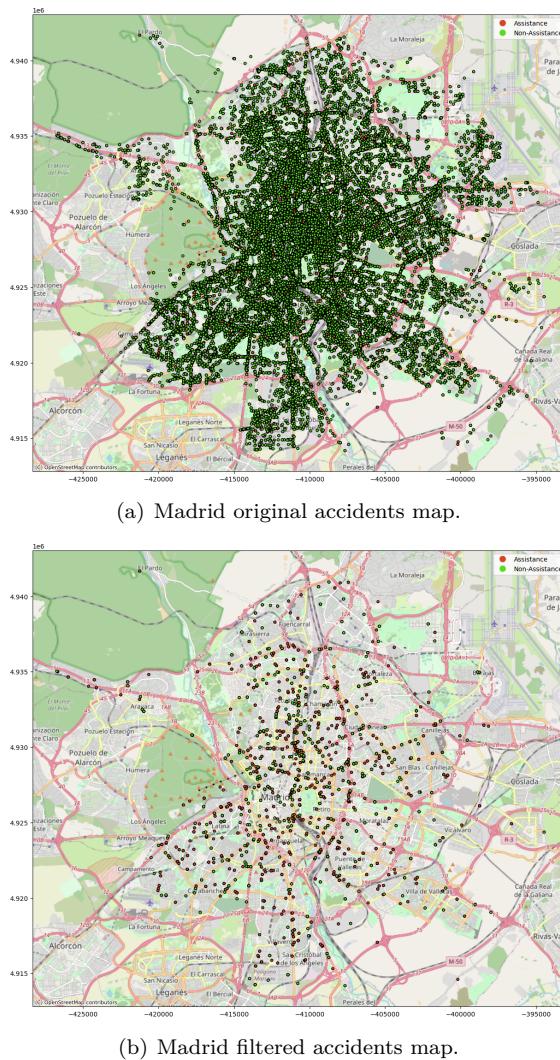


Figura 5.12: Madrid original/filtered accidents map.

En la Figura 5.13 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.

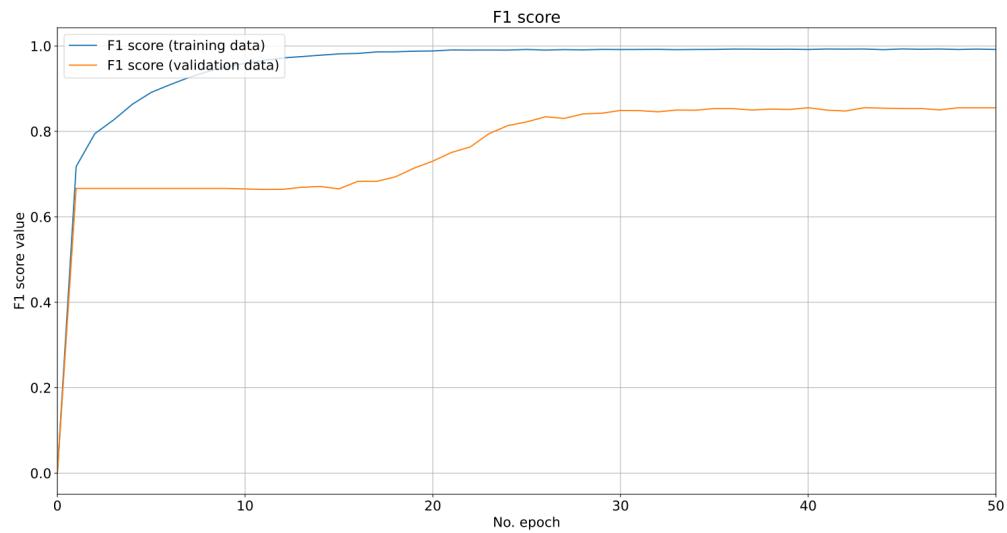


Figura 5.13: Evolution of F1-Score Madrid.

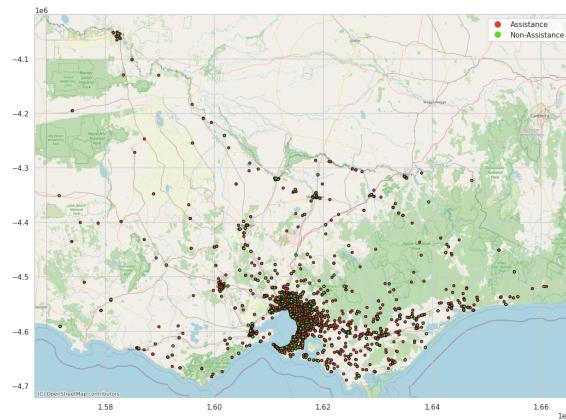
En la tabla 5.36 se observan los resultados de la métrica F1-Score de la predicción de la severidad de los accidentes de cada uno de los modelos sobre el conjunto de test de la ciudad de Madrid. Como se puede comprobar, el valor más alto lo ofrece el nuevo modelo GTAAF propuesto, llegando a mejorar en 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.

Modelo	Asistencia	F1-Score España
		Madrid
NB	No	0.729
	Sí	0.621
SVC	No	0.862
	Sí	0.748
KNN	No	0.739
	Sí	0.634
RF	No	0.744
	Sí	0.643
LR	No	0.750
	Sí	0.623
MLP	No	0.856
	Sí	0.724
GTAAF	No	0.894
	Sí	0.798

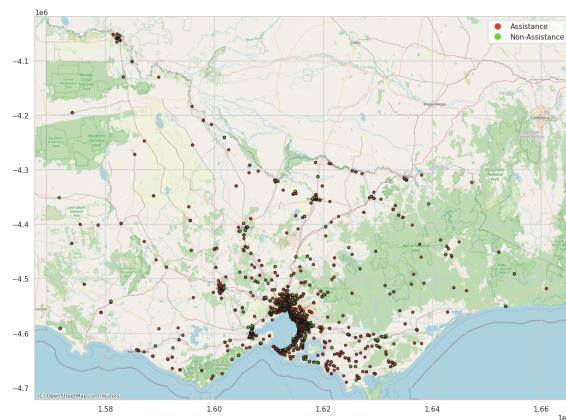
Cuadro 5.36: F1-Score por modelo y clase de accidente en Madrid (España).

Victoria

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 5.14 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 5.14(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 5.14(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 5.14: Victoria original/filtered accidents map.

En la Figura 5.15 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.

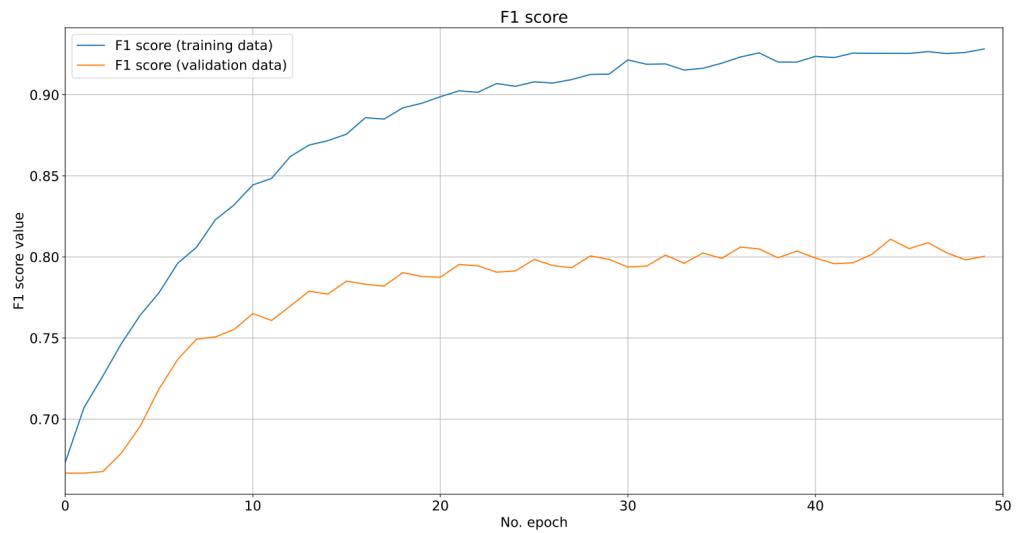


Figura 5.15: Evolution of F1-Score Victoria.

En la Tabla 5.37, 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.

		<i>F1-Score Australia</i>
Modelo	Asistencia	Victoria
NB	No	0.635
	Sí	0.476
SVC	No	0.662
	Sí	0.679
KNN	No	0.654
	Sí	0.616
RF	No	0.647
	Sí	0.364
LR	No	0.612
	Sí	0.630
MLP	No	0.635
	Sí	0.694
GTAAF	No	0.727
	Sí	0.784

Cuadro 5.37: F1-Score por modelo y clase de accidente en Victoria (Australia).

En la Tabla 5.37 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.

Reino Unido

Southwark

Southwark es un distrito de Londres situado en la orilla sur del río Támesis, con una alta densidad de población. En la Figura ?? se observa la distribución de los accidentes a lo largo del municipio de Southwark. Analizando los accidentes del conjunto de datos original, Figura ??, 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 ??) 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.

Manchester

Manchester, ubicada en el noroeste de Inglaterra, es una gran ciudad conocida por su legado industrial y su alta densidad de población. En la Figura ?? 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 ?? 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.

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 ?? 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 ?? 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 ?? 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.

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 ?? se muestra la comparativa de la distribución de accidentes originales del dataset y filtrados para la ciudad de Liverpool. En la Figura ?? 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 ??, 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 neurálgico. 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.

Sheffield

Sheffield, ubicada en South Yorkshire, presume de un patrimonio industrial y paisajes pintorescos, con una densidad de población intermedia. En la Figura ?? 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 ?? 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 ?? 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.

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 ?? 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 ?? 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 ?? 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.

En la Tabla 5.38 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 **sin necesidad de asistencia**, el MLP, para la ciudad de Manchester con un 6,7 %, mientras que la mayor diferencia para los accidentes **con necesidad de asistencia** es de un 13,8 % en la ciudad de Southwark respecto al siguiente mejor modelo, de nuevo el MLP. La siguiente mayor diferencia se presenta entre el modelo GTAAF se presenta en la ciudad de Southwark para la clase sin necesidad de asistencia, con un incremento del 4,8 % respecto al siguiente mejor modelo MLP, mientras que para la clase con necesidad de asistencia 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 con necesidad de asistencia, obteniendo de media una mejora de 9,21 % sobre todas las ciudades, mientras que el incremento de rendimiento se acentúa menos en accidentes sin necesidad de asistencia, ya que el resto de modelos ofrecen unas métricas más altas, 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.

		F1-Score Reino Unido					
Modelo	Asistencia	Southwark	Manchester	Birmingham	Liverpool	Sheffield	Cornwall
NB	No	0.504	0.675	0.567	0.560	0.620	0.653
	Sí	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
	Sí	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
	Sí	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
	Sí	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
	Sí	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
	Sí	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
	Sí	0.881	0.762	0.711	0.853	0.889	0.707

Cuadro 5.38: F1-Score por modelo y clase de accidente para cada una de las poblaciones de Reino Unido.

La Figura 5.16 muestra a modo de comparativa el rendimiento del nuevo modelo GTAAF propuesto en los accidentes **sin necesidad de asistencia** 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 aumento del 6.5 % respecto al siguiente mejor modelo, el SVC.

F1-Score by region (Non-Assistance accidents)

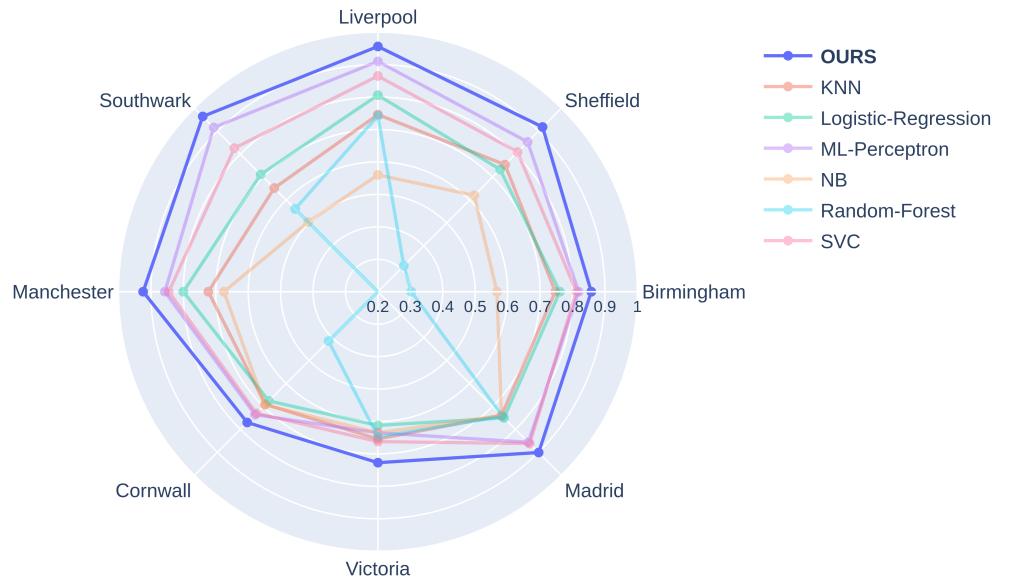


Figura 5.16: F1-Scores Comparison for Non-Assistance Accidents.

La Figura 5.17 muestra la comparativa del rendimiento basado en el F1-Score de los modelos para cada una de las ciudades en los accidentes **con necesidad de asistencia**. En esta gráfica se puede observar una diferencia considerablemente mayor del nuevo modelo GTAAF propuesto . La mayor mejora se presenta en la ciudad de Southwark, con un incremento del F1-Score del 13.8% respecto al siguiente mejor modelo sobre esta población, el MLP.

F1-Score by region (Assistance accidents)



Figura 5.17: F1-Scores Comparison for Assistance Accidents.

5.4.2. 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 futuros conjuntos de datos que no dispongan de la totalidad de las características descritas en este documento. Para ello se realizarán tres experimentos para cada conjunto de datos, estos consistirán en eliminar aquellas variables 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 XGBoost optimizado mediante el algoritmo genético.

En la tabla ?? ...

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

Cuadro 5.39: Comparativa F1-Score con eliminación de características sobre el conjunto de datos de Cornwall. La columna Menor representa los resultados obtenidos ejecutando la metodología y eliminando la característica que menor peso presenta mediante el algoritmo genético, la columna Mayor presenta los resultados eliminando la característica de mayor importancia y Ambas presenta los resultados eliminando las dos simultáneamente. En negrita se muestran los resultados obtenidos por el mejor modelo (GTAAF).

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

Cuadro 5.40: Comparativa F1-Score con eliminación de características sobre el conjunto de datos de Victoria. La columna Menor representa los resultados obtenidos ejecutando la metodología y eliminando la característica que menor peso presenta mediante el algoritmo genético, la columna Mayor presenta los resultados eliminando la característica de mayor importancia y Ambas presenta los resultados eliminando las dos simultáneamente. En negrita se muestran los resultados obtenidos por el mejor modelo (GTAAF).

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 la puntuación *F1-Score* entre GTAAF y los otros modelos aumenta.

Esta circunstancia sugiere que nuestro modelo se ve afectado por las características extremas, donde el modelo de XGBoost 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 5.18 podemos observar la diferencia de los algoritmos con y sin pérdida de características (Tablas 5.38, columna Cornwall, y 5.37 contra las Tablas 5.39 y 5.40 respectivamente). 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 los accidentes **sin necesidad de asistencia** y **con necesidad de asistencia** (eliminando la peor característica, ver la barra azul en la Figura 5.18-arriba), 1.4 % y 2.9 % (eliminando la mejor característica, ver la barra verde

en la Figura 5.18-arriba) y 7% y 8% (eliminando ambas características, ver la barra gris en la Figura 5.18-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 5.18-abajo), -0.7% y 0.9% (ver la barra verde en la Figura 5.18-abajo), y 5.1% y 3% (ver la barra gris en la Figura 5.18-abajo), respectivamente. Esto indicaría un efecto menor de los valores extremos en la ponderación del algoritmo genético si el área es más dispersa.

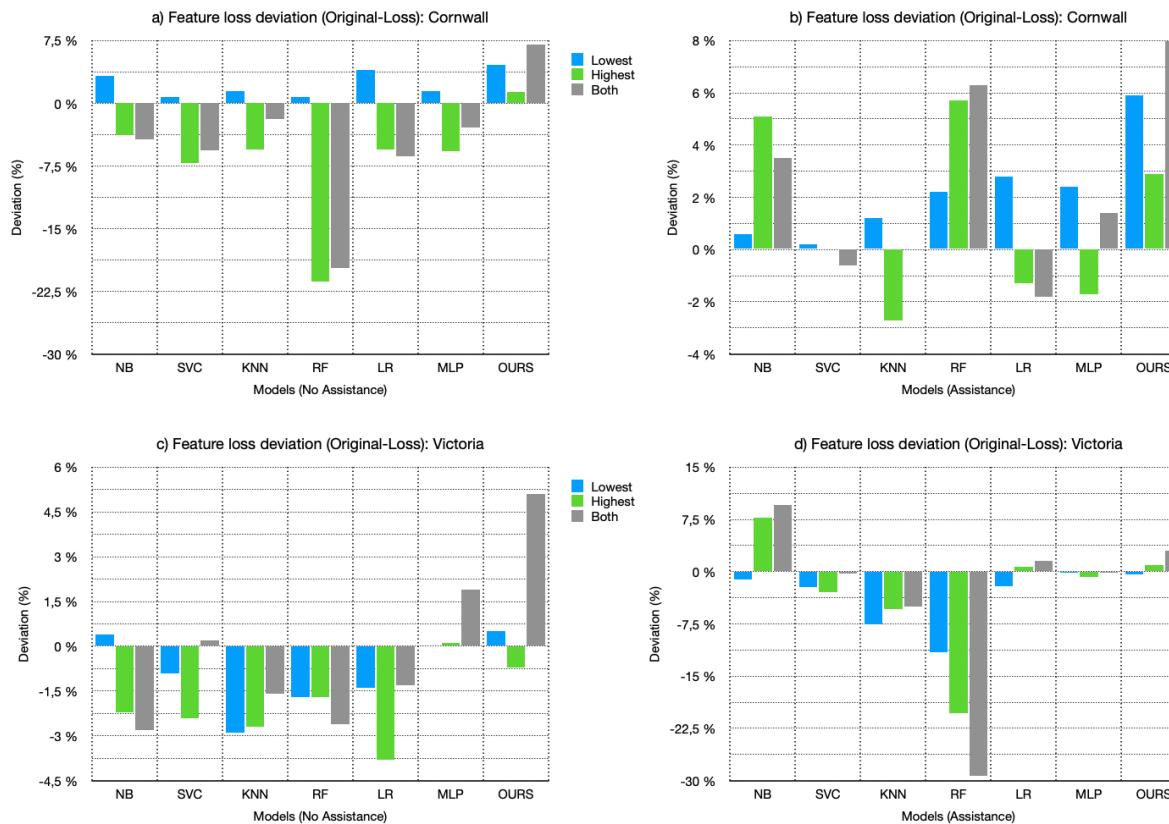


Figura 5.18: Comparación de pérdida de características. Las barras representan la diferencia entre los resultados con todas las características y los resultados sin características extremas: en azul sin la característica más baja, en verde sin la característica más alta y en gris sin ambas características 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 los afectados. Para ello se ha propuesto una metodología que transforma las características que describen los accidentes, mediante categorizaciones, en datos matriciales para alimentar a nuestro modelo convolucional GTAAF. Como se ha demostrado en su evaluación, los resultados no solo mejoran ampliamente a los modelos del estado del arte (con valores de hasta el 13.8 %), sino que la categorización propuesta ha demostrado robustez respecto a la interpretación de las características de forma individual 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.

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 aquellas de mayor y menor impacto que han resultado del *XGBoost* optimizado mediante el algoritmo genético, obteniendo resultados incluso mejores en nuestro modelo. Esto hace indicar la sensibilidad que tiene respecto a estos casos extremos.

Como principal medida medida como trabajo a futuro se propone aplicar este modelo a otros conjuntos de datos de accidentes para evaluar la utilidad que ofrece en comparación con otros modelos, por otra parte se debe analizar cómo reducir la sensibilidad del modelo GTAAF ante características extremas.

Capítulo 7

Publications

Capítulo 8

COSAS IMPORTANTES

8.1. Cosas que faltan:

1. Reforzar mucho que Madrid tiene solo 5 categorías en la sección de evaluación del 3er paper.
2. Pensar qué hacemos con la tabla de asignación de gravedad de accidentes de Madrid (que nos desmonta toda la tesis)
3. Pensar si presentamos la densidad de población en el primer apartado de datos o en la sección final como en el paper 3.
4. Repasar los números de muestras de Madrid para que cuadren
5. Conclusiones del paper 1
6. Contar las características que tiene Victoria
7. Repasar la consistencia en el nombrado de modelos (CNN-1D vs CNN-1D y para 2D igual)
8. Ver qué hacemos con el paper 1. Las erratas de poner la clase severidad en las tablas de categorizaciones y la matriz de confusión CNN-2D
9. La matriz de confusión del CNN-2D del primer paper no tiene el mismo nº de muestras de test que el resto de modelos. Crear de nuevo la matriz de confusión con los registros corregidos. Esto implica todos los cálculos de la métrica CNN-2D (precision, recall, f1, etc.)
10. Traducciones de las tablas?
11. Traducciones de las imágenes? (esto puede ser gordo)

12. Pensar cómo presentar la evaluación de la metodología preliminar. Para mí es como muy larga y no tiene excesiva importancia.
13. Concluir bien de una forma directa para llegar a un objetivo (desde que se empiezan a presentar la evaluación de GTAAF)
14. Cambiar la plantilla para que se vea decente
15. Darle una vuelta importante a cómo llamamos las cosas (severidad, gravedad, etc.)
16. Repasar toda la redacción
17. Descomentar los mapas para que se muestren

8.2. Dudas generales:

1. **aLa categorización de impacto entre Madrid y Victoria tiene valores iguales en distintas posiciones! TAMBIEŃ Clase de Víctima.**
2. La metodología es GTAAF? Es el modelo? O es todo junto?
- Por que los datos de la primera metodología de Madrid no cuadran con los de la segunda?

Capítulo 9

Anexos