



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. Resultados preliminares - Prototipo	59
5.2. Resultados finales	73
5.2.1. Dataset	73
5.2.2. Descripción de datos	74
5.2.2.1. Partes comunes entre los datos	76
5.2.2.2. Principales diferencias entre los datos	77
5.2.3. Limpieza	78
5.2.4. Filtrado de áreas	79
5.2.5. Discretización	81
5.2.6. Transformación (Sin/Cos)	86
5.2.7. Resampling	86
5.2.8. Normalización	87
5.2.9. Categorización	87
5.2.10. División Train-Val-Test	87
5.2.11. Cálculo de pesos	88
5.2.12. Feature Importance Algorithm	88
5.2.13. Algoritmo Genético	88
5.2.14. Pesos de categorías	90
5.2.15. Construcción de matrices	90

<i>ÍNDICE GENERAL</i>	5
5.2.16. Métricas de evaluación	90
5.3. Pruebas de estrés	111
6. Conclusiones	115
7. Publications	117
8. Anexos	119

Abstract

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

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

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

Para validar este enfoque se compararán los resultados de este nuevo modelo con otros 6 modelos del estado del arte que han sido aplicados históricamente para la predicción de la necesidad de asistencia médica en accidentes a lo largo de

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

Acknowledgements

Luis: hay que hacerlo.

blablabla

Capítulo 1

Introduction

Luis: hay que hacerlo.

blablabla

1.1. Motivación

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

Luis: hay que hacerlo.

1.2. Objetivos

Luis: hay que hacerlo.

Capítulo 2

Estado del arte

2.1. ¿Cómo medir la gravedad de un accidente?

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

La definición de la gravedad de un accidente de tráfico es la base fundamental para enfocar cualquier investigación. 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 de estudio sobre 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 toman como referencia la agrupación de la severidad 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 borrosamente 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 también puede clasificarse como por ejemplo, **accidente fatal/no fatal**, como **lesión/no lesión**, o incluso **accidente con lesiones o solo daños materiales** [?, ?].

En esta tesis, la forma en la que se medirá la gravedad de los accidentes 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 (o no)** a las personas en un accidente de tráfico.

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

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

La predicción de la severidad de los accidentes de tráfico ha sido un campo ampliamente estudiado a lo largo de los últimos años, 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 el último año ha sido exponencial. Es por esto que distintos enfoques han sido aplicados para solventar este problema a lo largo de distintas poblaciones en todo el mundo, donde la severidad de los accidentes han sido consideradas de distintas formas, como se ha mencionado en el apartado anterior.

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

2.2. ¿CÓMO PREDECIR LA GRAVEDAD DE UN ACCIDENTE DE TRÁFICO?15

Por otra parte, en [?] se aplican tres métodos distintos al contexto para predecir la severidad de los accidentes en la ciudad de Seúl, concretamente Random Forest, Perceptrones Multicapa y árboles de decisión, con el objetivo de comparar cuál de ellos generaliza mejor sobre los datos, siendo el Random Forest el que mejores resultados presentaba.

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

Otro enfoque interesante es el aplicado en [?], donde se aplican en conjunto distintos árboles de decisión para crear un ensemble del tipo Random Forest, cuyos hiperparámetros son optimizados mediante Optimización Bayesiana (BO). Esta publicación busca predecir la severidad de los accidentes a lo largo de Estados Unidos, concretamente entre los estados de Montgomery (Alabama) y el estado de Pensilvania, para ello se utiliza un conjunto de dataset relativo a accidentes de tráfico producidos entre el año 2016 y 2019.

Otros enfoques 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, es comparado con el otro método desarrollado Random Forest y un método Ordered Probit, ambos siendo superados por CIF.

Siguiendo con la predicción de la gravedad de accidentes en vehículos de dos ruedas, se propone un enfoque orientado a ciclistas [?], donde se quieren predecir las características clave que influyen en la severidad 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, aquellas 8 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 grandes camiones. Para ello, se utiliza un conjunto de datos de Irán, donde se estudian los accidentes a lo largo de 8 provincias entre los años 2011 y 2014. En base a estos datos, se enrena un modelo 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 encontrarse con investigaciones que combinen distintos modelos para lograr un mejor rendimiento, como es el caso de . Donde la investigación implementa 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 qué características son más 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 NB, KNN, BLR, XGBoost, siendo este último el que mejor generaliza sobre los datos.

Por otra parte, estudios como [?] ofrecen la comparación de distintos modelos predictivos (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 estos modelos del estado del arte es crucial, ya que un buen entrenamiento y unos buenos resultados sobre un conjunto de datos de una región no implican que este modelo pueda ser aplicado a otras localizaciones, tanto por la falta de características respecto a otros conjuntos de datos y a las peculiaridades del conjunto de datos en cuestión.

Como 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.

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 que son utilizadas en este trabajo.

3.1. Modelos estado del arte

Luis: No me termina de convencer, tengo que darle otra vuelta. Me da la sensación de que se explican las fórmulas y ya, no se interpretan. Lo dejo para más adelante sabiendo esto.

Modelos del estado del arte contra los que nos comparamos.

En el campo de la Inteligencia Artificial y 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 como comparativa como punto de partida. 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 este ofrece en numerosos problemas de clasificación. Su denominación 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:

Luis: buscar las fórmulas a posteriori

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.

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. 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 R². 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..

k-Nearest Neighbors

El algoritmo k-Nearest Neighbors (KNN) es un algoritmo de aprendizaje automático, ampliamente utilizado tanto para problemas de clasificación como de regresión. 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:

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

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. 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 Enbemble 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.

Support Vector Classifier

El Support Vector Classifier (SVC) es un algoritmo de aprendizaje supervisado utilizado comúnmente para problemas de clasificación. 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.

MLP

Luis: si explico el MLP en detalle, entraría en conflicto con la siguiente sección de modelos CNN, tendría que explicar de nuevo el Bakpropagation, gradientes, etc..

El Perceptrón Multicapa o Multi-Layer Perceptron (MLP) es una arquitectura de red neuronal artificial. 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. Al pertenecer a la familia de las redes neuronales, estos modelos aprende los pesos asignados a cada una de las conexiones entre neuronas de las capas para dar lugar a

3.2. Algoritmos CNN

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

Luis: Yo creo que OK, hay que repasar el hilo que conecta los párrafos.

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 redimiento 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 GAN, segmentación de elementos de interés dentro de imágenes, o incluso en la representación de imágenes mediante vectores n-dimensionales para comparar la similaridad 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:** activation functions in CNNs introduce nonlinearities, enabling the network to learn complex patterns and relationships within the data, typically, an activation function like ReLU (Rectified

Linear Unit) is applied element-wise to the feature maps to introduce non-linearity.

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

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

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

Proceso de entrenamiento de una red convolucional

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

Función de pérdida

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 como objetivo comparar la calidad de las 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:

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

Donde N representa el número de muestras totales en el conjunto de datos, y_i es la etiqueta verdadera de la clase (0 ó 1) de la muestra actual y p_i es 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. Valores de probabilidad altos p_i 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.

Backpropagation

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.

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

Donde:

- 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 esta tesis, se estudiarán en detalle dos tipos de redes de este tipo, las redes neuronales convolucionales unidimensionales (CNNs 1D) y las redes neuronales convolucionales bidimensionales (CNNs 2D) y se expondrán en los siguientes puntos:

CNN-1D

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

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

A lo largo de la historia reciente, 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, excels, 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

Luis: repasar la redacción del último párrafo.

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:

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.

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 (hyper-parameter optimization o 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

Luis TODO: Falta explicar el XGBoost al final

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

3.6. ALGORITMOS DE MEDICIÓN DE IMPORTANCIA DE CARACTERÍSTICAS29

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 potente dentro de los ensembles 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. Esta técnica es ampliamente utilizada para problemas de **TODO: estado del arte del XGBoost.... ??** [1].

Luis TODO: Falta explicar el XGBoost aquí

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 de la solución ideal (óptimo global) mediante métodos de combinación a lo largo de todo el conjunto total de posibles soluciones (espacio de búsqueda). Existen numerosos algoritmos genéticos conocidos, y muchos han sido aplicados a distintos contextos, tanto para problemas de objetivo único (single objective) como a problemas multi-objetivo (multi-objective), algunos algoritmos de este segundo caso son: SPEA-II, NSGA-II o AMGA [?].

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 (*etapa de inicialización*), se crea una población original de N individuos aleatorios, donde cada uno de estos representa una posible solución al problema que se quiere optimizar, ver 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 (*etapa de evaluación*), Figura 3.2. Una vez se dispone de las puntuaciones de calidad, aquellos individuos que mejor se adapten al problema (mejor puntaje reciban) serán escogidos para dar lugar a descendencia (*etapa de selección*). La información que contienen los M mejores individuos es combinada entre sí (*etapa de cruce*), simulando el intercambio de información que supondría el intercambio genético en la naturaleza. Una vez se disponen de los nuevos individuos, estos pueden sufrir modificaciones aleatorias sobre su información resultante (*etapa de mutación*), 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 G generaciones hasta llegar a una condición de parada, tras la cual se seleccionará el individuo que mejor puntuación haya obtenido mediante la función heurística.

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.

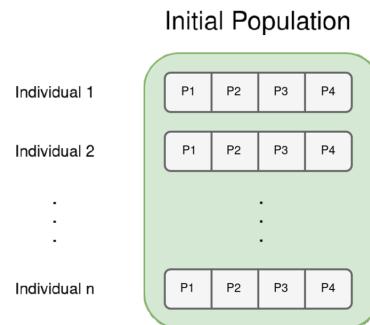


Figura 3.1: TFM.

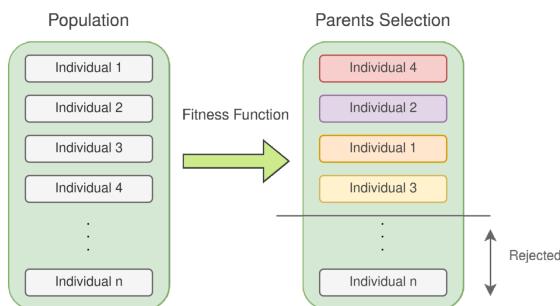


Figura 3.2: TFM.

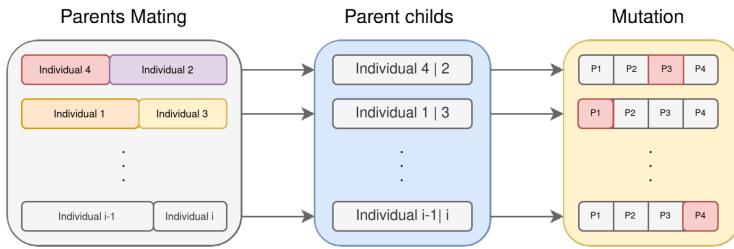


Figura 3.3: TFM.

De esta forma, mediante la mejora continua de individuos a través de las generaciones, seleccionando aquellos mejores y combinando su información entre sí, se 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 dos conceptos básicos TP, FP, FN y TN, estos datos son calculados para cada una de las clases que puede predecir el modelo.

1. **True Positives (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 sea necesaria ninguna intervención, que predecir erróneamente uno que no necesita asistencia y luego muera el accidentado (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.

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

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.

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

Capítulo 4

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

En esta tesis 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 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

~~MANU: En una primera fase de la tesis doctoral, se propuso un modelo preliminar aplicado a un dataset concreto~~

Luis: En una primera fase de la tesis doctoral, se propuso un modelo preliminar de predicción de la gravedad de accidentes de tráfico aplicado a una ciudad

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

en concreto [?], teniendo como meta la predicción de la gravedad en base a 3 niveles (leve, severo y fatal). Esta tesis tiene como objetivo diseñar un modelo predictivo general a cualquier ciudad, por lo que en este apartado se expondrá brevemente el enfoque del modelo preliminar, a modo de introducción y acentrando 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 el artículo , se implementó con el objetivo de predecir la gravedad de los accidentes de tráfico en la ciudad de Madrid, concretamente en tres clases distintas disponibles en este conjunto de datos: (1) Leves, (2) Severos y (3) Fatales.

Para llegar al entrenamiento de un modelo predictivo, se diseñó una metodología prototípica que estaba compuesta por cinco fases secuenciales, mostradas en la Figura 4.1. Y tenía como objetivo realizar transformaciones y operaciones sobre datos, inicialmente tabulares, para transformarlos en datos matriciales sobre los que puedieran 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.

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

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

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

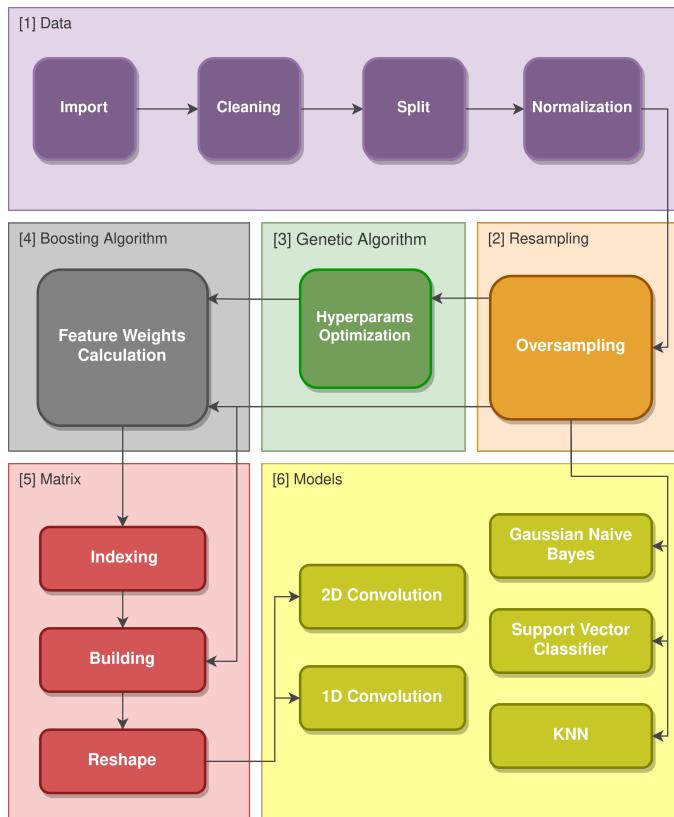


Figura 4.1: Flow chart of the proposed model with its different phases.

La primera fase de esta metodología prototipo está orientada al tratamiento de los datos (1). 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, y un tratamiento para normalizarlos y que no estuviesen sobredimensionados. En una segunda fase (2), se aplicó un proceso para trabajar sobre el desbalanceo de los datos presente el dataset. Debido a la naturaleza de los accidentes de tráfico, gran parte de ellos eran de tipo leve, mientras que el resto de tipos de accidentes (severos y fatales), presentaban una proporción mucho menor respecto a los del primer tipo. Para evitar un sesgo en los modelos, y que tendiesen a predecir cualquier nueva muestra como la clase mayoritaria, se estudiaron distintas técnicas de balanceo de datos, utilizando finalmente la técnica Borderline SMOTE-II para balancear las clases minoritarias, aplicando generación de datos sintéticos hasta 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 un algoritmos clasificadores que ofrecen la importancia numérica de cada variable en función del peso que han tenido durante su entrenamiento. Estos algoritmos necesitan una configuración de hiperparámetros que se realizó mediante la evolución de un algoritmo genético (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 de dos dimensiones sobre las que comenzar a indexar las variables. En primer lugar se calculó el peso total de las categorías, que era la suma de cada una de las características que contenía, como resultado de esto, cada categoría se indexaba a una fila de la matriz, donde aquella que más peso presentaba era asignada a la fila central, la segunda en la posición inmediatamente superior, la siguiente en la inferior y así sucesivamente. Las características que las componían se asociaban a las columnas dentro de su categoría de la misma manera, la de mayor peso en la posición central, la siguiente en su posición inmediatamente a la izquierda, la siguiente a la derecha etc. Como resultado de este proceso, cada registro perteneciente al dataset original era transformado en una matriz de 5x5.

Las arquitecturas propuestas 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 píxel por píxel 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.

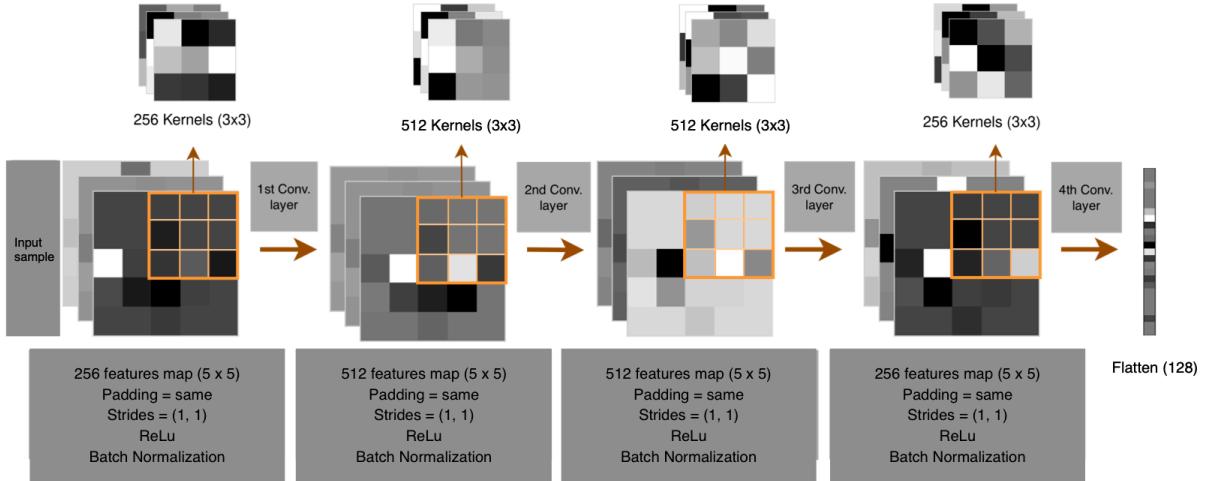


Figura 4.2: Architecture of the 2D-Convolutional neural network.

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.

Conclusiones

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, Graves y Fatales) era un condicionante que perjudicaba considerablemente el rendimiento del modelo. Al tener dos clases notablemente desproporcionadas respecto a la mayoritaria (Graves y Fatales), el modelo ser siendo dos de estas clases minoritarias, aún aplicando generación de datos sintéticos contra tantas muestras de accidentes leves podría carecer de sentido a nivel práctico, ya que la diferencia entre distinguir entre las clases Grave y Fatal respecto a agruparlas en una, a nivel funcional era **algo que tenía sentido y mejoraría el rendimiento del modelo... (algo así)**.

Por otra parte, se plantearon cuestiones sobre la discretización de los datos.. se observaron incongruencias en la interpretación sobre los valores numéricos asignados....

Por otra parte, observando las curvas de aprendizaje en las gráficas de entrenamiento de los modelos, una propuesta de mejora interesante era añadir

más características a la matriz de entrada a las redes convolucionales, por lo que trabajo a futuro se planteó 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

Luis: aquí de alguna forma te tienes que traer la categorización de la sección de resultados, para decir que la metodología generalizable se usa en base a esto.

Luis: este párrafo inicial entra en colapso con la sección de Resultados 1.1.1 Descripción de datos, se dicen cosas muy parecidas. Hay que analizar cuál nos gusta más y reestructurarlo

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

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

En primera instancia, las fases de la nueva metodología son asignadas a tres etapas claramente diferenciadas: (1) la fase de Preprocesamiento, donde se

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

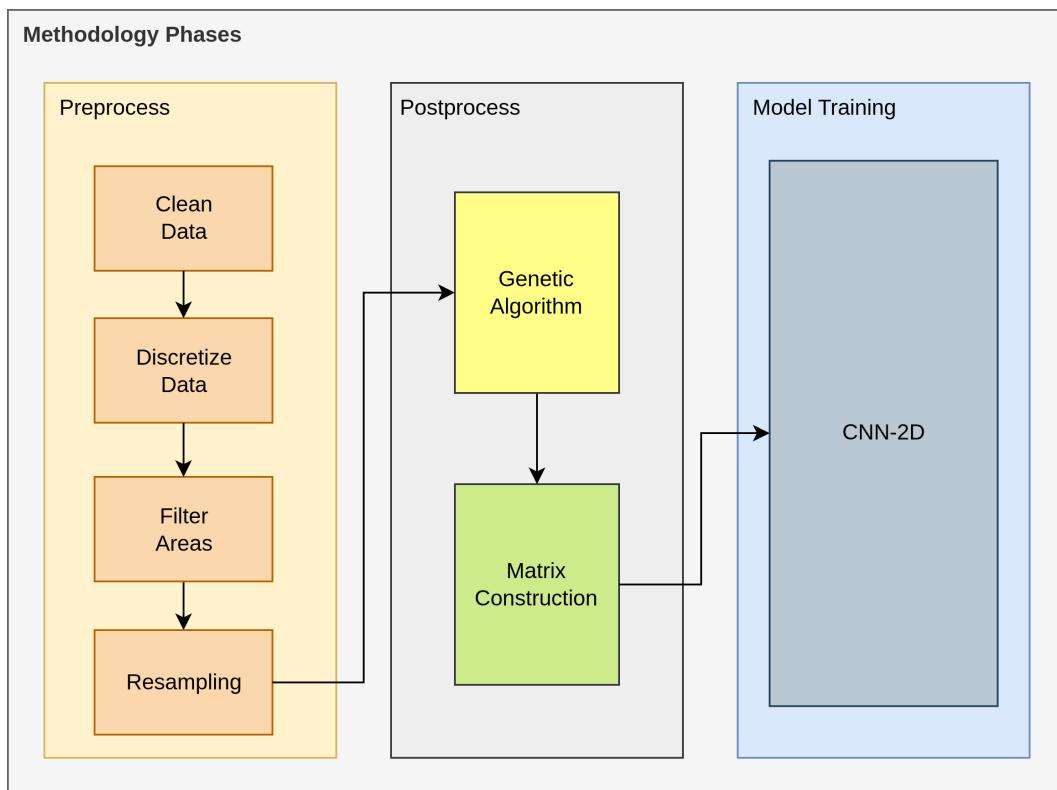


Figura 4.3: Methodology flowchart: data preprocessing, postprocessing and model training.

En segundo lugar, el concepto de gravedad de los accidentes es reasignado de tres a dos clases (accidentes sin necesidad de asistencia y accidentes con necesidad). Esto se debe a que el valor que aporta distinguir entre la clases Severo y Fatal no es lo suficientemente enriquecedor como para arriesgarse a distinguir entre tres clases, ya que un modelo de clasificación a medida que incrementa el número de clases, tiene más posibilidades de realizar predicciones erróneas, sobre todo si son clases minoritarias como son los accidentes severos y fatales. Por este motivo se distinguen entre dos clases de accidentes: **con necesidad de asistencia y sin necesidad de 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 como objetivo dividir el mapa de la población en celdas, para seleccionar aquellas zonas de la región donde existan ambos tipos de accidentes para que la información que interpretan los modelos obtenida aporte no se vea condicionada por la naturaleza.

Como cuarta modificación, se añaden transformaciones sobre los datos con el fin de aportar más información en base a las variables ya existentes. Para ello se representó la variable de la hora del accidente a dos componentes cíclicas mediante senos y cosenos, para capturar la naturaleza periódica de la hora del accidente.

**Luis: no dices nada de la categorización en el modelo preliminar
creo, tienes que mencionarlo**

Como quinta variación a considerar, se redefinen las categorías donde son asignadas las características, pasando de ser originalmente 5 a 6 categorías finales, buscando una reorganización que facilita la asignación de características a conceptos más generales que representan estas categorías. Esto implica que las matrices que se construyen pasan de ser de dimensiones 5x5 a 6x4 sobre estos conjuntos de datos.

Por otra parte, se centraron los esfuerzo en desarrollar el modelo convolucional de dos dimensiones. En base al análisis de entrenamiento del modelo unidimensional y bidimensional, se optó por descartar el primero debido a la poca capacidad de generalización sobre el conjunto de validación ofrecido por el modelo unidimensional.

Por último, se ampliaron los conjuntos de datos sobre los que se amplió el conjunto de datos sobre los que aplicar la metodología, incluyendo seis regiones de Reino Unido y una de Australia, con el objetivo de evaluar la generalización de la metodología en distintos contextos. Además, se ampliaron los modelos del estado del arte contra los que comparar el rendimiento, llegando a ser seis, X de aprendizaje supervisado (SVC, clasificador probabilístico Naive Bayes, clasificador Bagging Random Forest, KNN y Regresión Logística), y una red neuronal Perceptrón Multicapa (Multilayer Perceptron o MLP).

4.2.1. Preprocesamiento

Esta sección explica las diferentes etapas que componen la fase de preprocesamiento de la metodología GTAAF propuesta. Esta es la primera de las etapas y es donde a los datos se les aplican transformaciones para dar lugar a un conjunto de datos refinado interpretable para cualquier modelo que trabaje con datos tabulares. Esta etapa está compuesta por cuatro fases: (1) proceso de limpieza de datos, donde se identifican, corrigen y se tratan las inconsistencias sobre los datos, (2) la discretización, donde se convierten las variables continuas en variables discretas y se codifican los valores cualitativos de las características,

(3) el filtrado de áreas, donde se reduce el desbalanceo de los datos escogiendo subregiones de la ciudad donde se localicen ambos tipos de accidentes, y (4) el remuestreo, donde se generan muestras sintéticas de la clase minoritaria para disponer de un dataset balanceado. En la Figura 4.4 se muestra el flujo sobre el que pasan los datos para cada una de las diferentes fases que componen la etapa de Preprocesamiento. Esta figura será referenciada en las siguientes subsecciones en la explicación de las fases de Preprocesamiento.

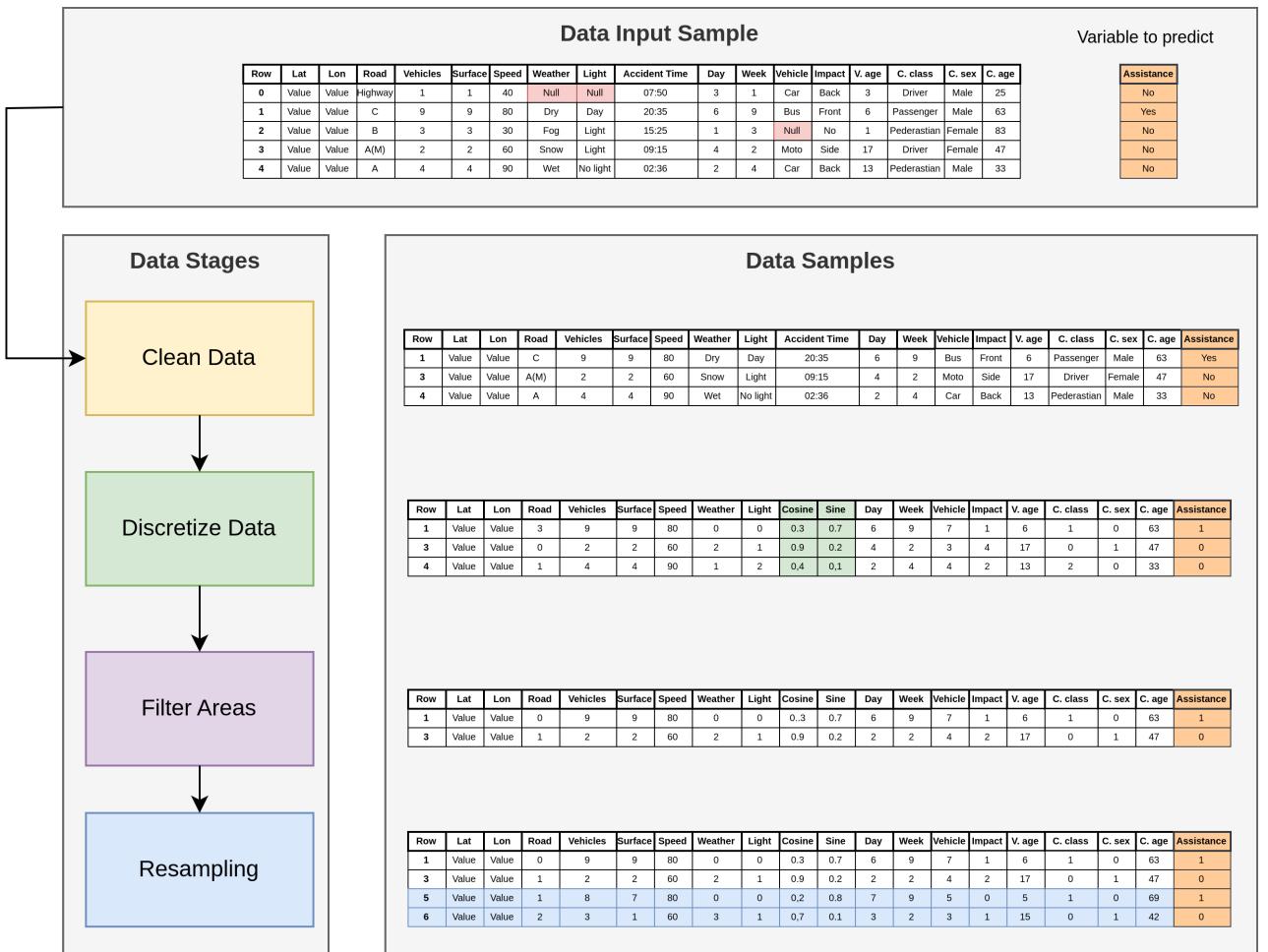


Figura 4.4: Pre-processing data flow.

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

En esta tesis se ha seguido un procedimiento de discretización incremental, donde a cada posible valor del conjunto de datos se le ha asignado un valor numérico en función de la importancia que se le ha asignado.

4.2.1.3. Transformación (Sin/Cos)

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

Como se ha comentado en la sección anterior, los modelos de Inteligencia

Artificial y Aprendizaje estadístico interpretan los datos en forma numérica. El valor numérico que se le asigna a cada campo es crítico, ya que será así como el modelo interprete el orden de los valores cualitativos que los humanos somos capaces de comprender. La representación del formato de 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:

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

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

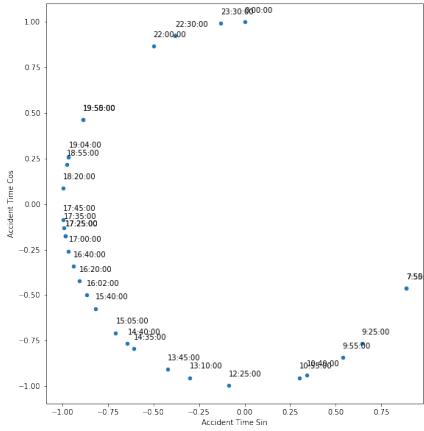


Figura 4.5: Algo así, no es lo definitivo.

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

4.2.1.4. Fitrado 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).

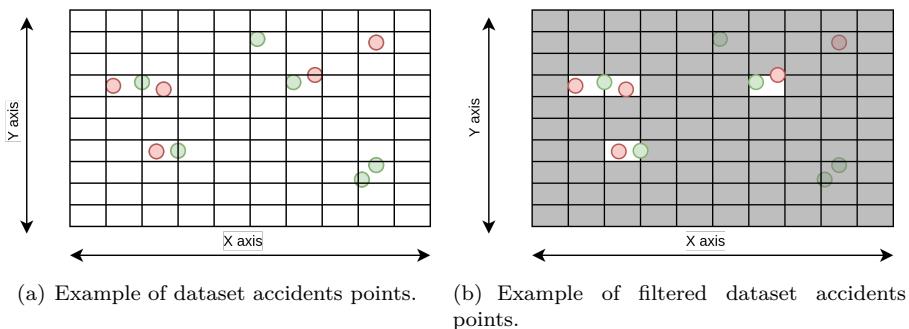


Figura 4.6: Ejemplo de filtrado de áreas. Los puntos verdes representan accidentes con lesiones leves, mientras que los puntos rojos representan accidentes que requirieron asistencia.

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

4.2.1.5. 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. Existen diferentes técnicas de normalización como Mean Centered (MC), Variable Stability Scaling (VSS) o Min-Max Normalization (MMN), entre otras [?]. En esta tesis, para normalizar los datos y hacerlos comparables entre sí se ha utilizado la técnica de Z-Score (ZSN) debido a que logra representaciones de acuerdo con una distribución normal. Para hacerlo, se utilizan la media y la desviación estándar para reescalar los datos de manera que su distribución esté definida por una media de cero y una desviación estándar unitaria.

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

4.2.1.6. División Train-Val-Test

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

Los modelos supervisados de Inteligencia Artificial aprenden patrones sobre datos que son ofrecidos en la etapa de entrenamiento del modelo. Durante esta fase los modelos realizan predicciones sobre de datos y posteriormente se les enseña la clase a la que pertenecía cada uno de los datos que ha predicho, de esta forma se mide el error que han cometido durante este proceso y los pesos de la red son actualizados para minimizar el error en la siguiente fase. Si este aprendizaje se repite durante muchas etapas, el modelo tiende a aprenderse los datos de memoria, lo que se conoce como sobreajuste de la red u overfitting, provocando que la red no sea capaz de generalizar ante nuevas muestras tras su entrenamiento. Por este motivo es importante mantener el control del entrenamiento de la red mediante la evaluación del rendimiento de la red en cada época mediante un conjunto de datos que nunca ha visto durante sus fases de entrenamiento, este conjunto de datos es conocido como conjunto de validación, y es utilizado para parar el entrenamiento cuando el modelo no sea capaz de generalizar sobre estas muestras. Por otra parte, existe el conjunto un conjunto de datos de test, utilizado para medir el rendimiento del modelo final una vez ha acabado su fase de entrenamiento. Este conjunto pertenece a muestras que la red no ha visto durante su fase de aprendizaje ni ha sido utilizado como validación.

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

4.2.1.7. 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 pertenezcan. El conjunto de datos, una vez se han reducido considerablemente el desbalanceo entre las dos clases gracias al proceso de filtrado de áreas, sigue presentando cierto desbalanceo. Por mucho que se haya acotado el problema a regiones individuales, es lógico que se hayan producido más accidentes sin necesidad de asistencia respecto a las que sí.

En 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, en esta tesis 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. Contras del resampling... Por estos motivos en este trabajo se ha optado por una técnica de generación de datos sintética denominada Synthetic Minority Oversampling Technique (SMOTE-II), que busca incrementar el número de clases de las muestras minoritarias mediante la generación de nuevas muestras artificiales.

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 recién propuesto. Este proceso implica mapear los atributos de las muestras tabulares en posiciones dentro de estas matrices. Para realizar esto, se hará uso de un método de transformación que toma en consideración la importancia de cada característica dentro del conjunto de datos. El objetivo es posicionar estratégicamente las características más relevantes en la matriz para maximizar su impacto en el modelo GTAAF, como se ilustra en la Figura 4.7. La determinación de la importancia de las características se basa en un algoritmo tipo boosting, que asigna pesos a las características según su relevancia en la separación de datos durante

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

el entrenamiento. Para garantizar un entrenamiento óptimo del modelo, se realiza una optimización de hiperparámetros utilizando un algoritmo genético. A lo largo de generaciones sucesivas, este algoritmo genético hace evolucionar los hiperparámetros, guiado por la métrica de F1-Score, que actúa como la función heurística.

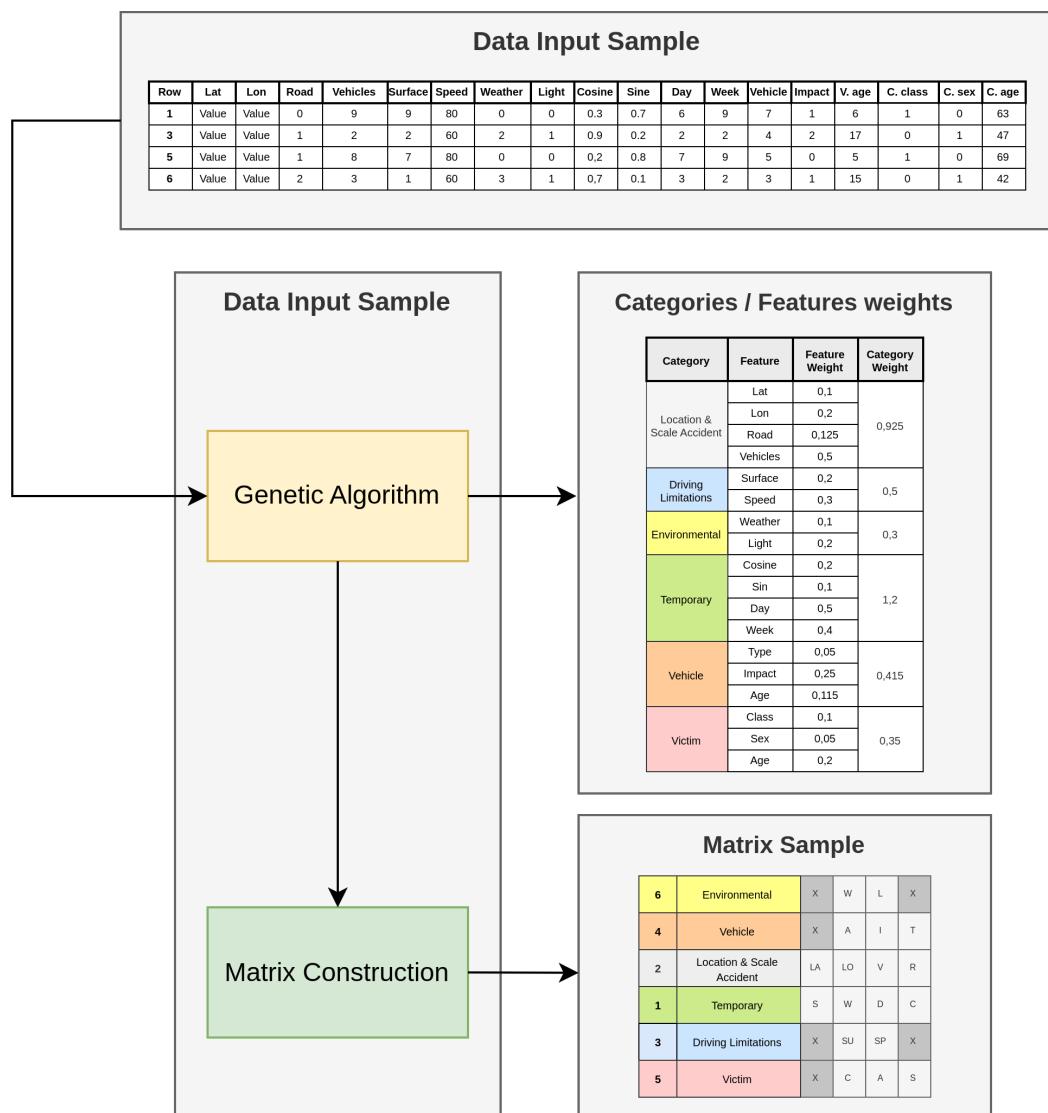


Figura 4.7: Category and feature weights.

4.2.2.1. Construcción de Matrices

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

Existen distintos métodos que logran transformar datos tabulares a una representación matricial de los mismos, buscando dar un sentido a la asignación de las características en posiciones de la matriz. En la sección 3.3 se presentaron distintos métodos como REFINED, DeepInsight o IGTID, que buscan optimizar la posición de las características en base a la similaridad que presenten entre ellas, principalmente en datos orientados a la descripción genética. Sin embargo, estas técnicas presentan distintas limitaciones debido a la magnitud de los datos para las que han sido diseñadas (del orden de 2.500 características), esto provoca que estos métodos sean difícilmente aplicables a datos de baja dimensionalidad, como asignar espacios en blanco ante la falta de características o que los métodos no sean capaces de converger al trabajar con tan pocos datos. En el caso de estudio de esta tesis las características disponibles son mucho menores, del orden de 20 variables.

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

4.2.2.2. Feature Importance Algorithm

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

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 ensambles de tipo Bagging como los Random Forest o métodos ensembles tipo Boosting.

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

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

En esta metodología se utilizará el algoritmo tipo Boosting XGBoost, donde se minimizará el error del modelo mediante la métrica F1-Score resultante de la clasificación de ambas clases de accidentes (Sin necesidad de asistencia y 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.

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.

Es por esto por lo que en esta tesis, los algoritmos genéticos son utilizados para optimizar los hiperparámetros de entrenamiento del algoritmo XGBoost, que permiten una exploración amplia del espacio de búsqueda de los hiperparámetros 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 que... Este proceso teniendo en cuenta la categorización inicial que permite aplicar esta metodología a cualquier región, siendo tolerante a la falta de características en la disponibilidad de los datos que se ofrezcan.

El método diseñado sigue los siguientes pasos:

1. En primer lugar, las características son asociadas a sus categorías, de tal forma que pueda medirse la importancia de cada categoría dentro del conjunto de datos. Esta es obtenida mediante la suma del peso total de cada característica individual que la contiene

2. El segundo paso es asignar cada categoría con una fila de la matriz según su peso, donde aquella con el mayor peso se posiciona en la fila central, la segunda categoría más importante se asocia a la fila inmediatamente superior, la siguiente a la fila inmediatamente inferior y así sucesivamente (ver Figura 4.8).
3. Una vez que las categorías están asociadas a una fila de la matriz, cada una de las características dentro de su categoría se asocia en cada columna siguiendo el mismo procedimiento definido en el apartado anterior. La característica más importante de una categoría se posiciona en el centro, la segunda característica más importante se sitúa inmediatamente a su izquierda, mientras que la siguiente característica más importante ocupa el lugar a su izquierda y así sucesivamente (ver Figura 4.9).

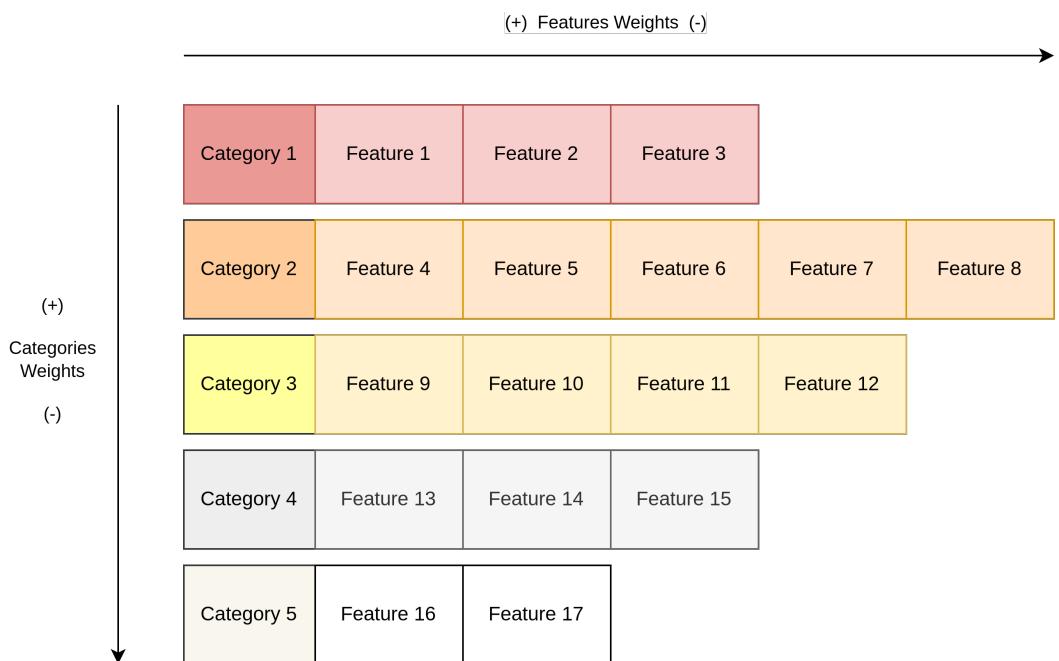


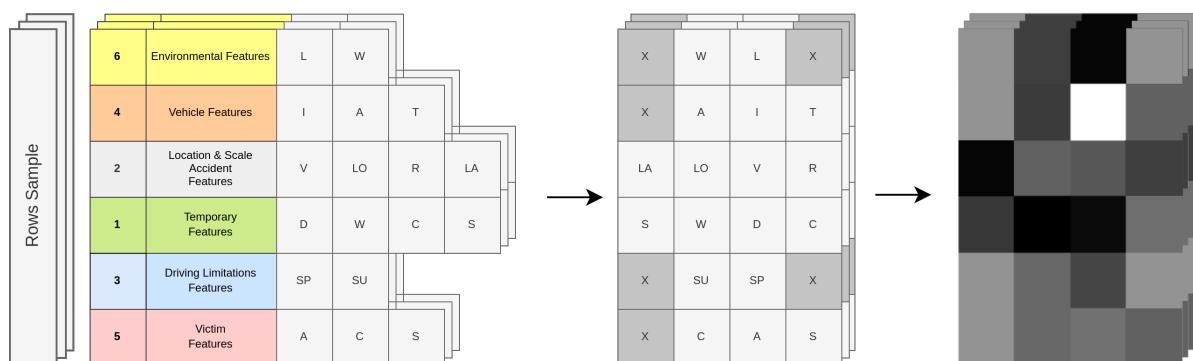
Figura 4.8: Category and feature weights.

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

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

Figura 4.9: Categories and feature positions.

En la Figura 4.10 se muestra un ejemplo del procedimiento

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

4.2.2.5. 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 densa 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 intuición de la arquitectura de la red.

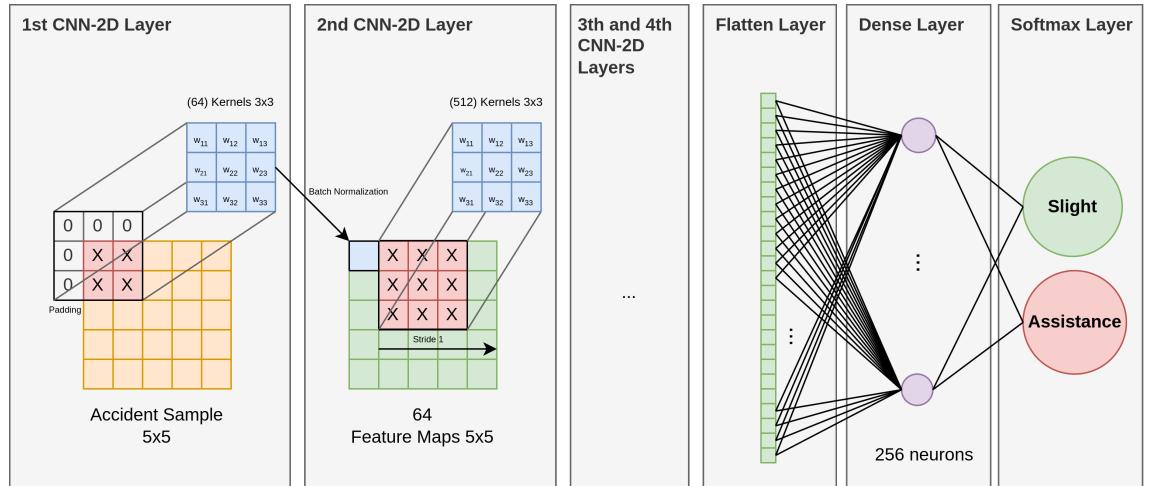


Figura 4.11: Proposed CNN-2D architecture summary.

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 ocho regiones 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, en un apartado posterior, se realizarán pruebas de la ejecución de la metodología eliminando características en los datos. En primer lugar, se prescinde aquella que más relevancia tiene para la metodología, en segundo lugar, aquella que menos relevancia tiene y, en un tercer experimento, ambas conjuntamente. 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 toda la información presentada en 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 .

Capítulo 5

Experimentos y resultados

5.1. Resultados preliminares - Prototipo

Aquí pones los resultados del paper 1

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

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

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

Descripción de datos

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

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

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

Limpieza

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

Discretización

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

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

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

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

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

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

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

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

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



Figura 5.1: Correlation matrix between the dataset variables.

Resampling

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

Normalización

En la Figura 5.4 se muestra un ejemplo de la aplicación de la normalización de datos en base a la técnica Z-Score, donde en la primera Tabla 5.2 se observa un registro de datos discretizado del dataset y el 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

estén bajo la misma magnitud, para poder ser eficientemente interpretables 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 disponen de los datos normalizados, estos ya se encuentran en las mismas magnitudes y por tanto pueden ser utilizados para el entrenamiento de cualquier modelo predictivo.

División Train-Val-Test

El siguiente paso, necesario en cualquier procedimiento de entrenamiento de un modelo predictivo, fue la división entre datos de entrenamiento y datos de validación. Esta división se realiza asignando un 80 % de los datos del dataset para asignarlos al entrenamiento y un 20 % a validación o test.

Categorización

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

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

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

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

Algoritmo Genético

En este punto, se analiza la optimización de los hiperparámetros del algoritmo de Boosting mediante un algoritmo genético. La figura 5.5 muestra la evolución de los tres hiperparámetros a lo largo de las generaciones. Como se puede observar, los hiperparámetros toman distintos valores en función del mejor individuo evaluado en la población en cada etapa, estos hiperparámetros convergen aproximadamente en la iteración 42, donde no se observan modificaciones a partir de esta generación.

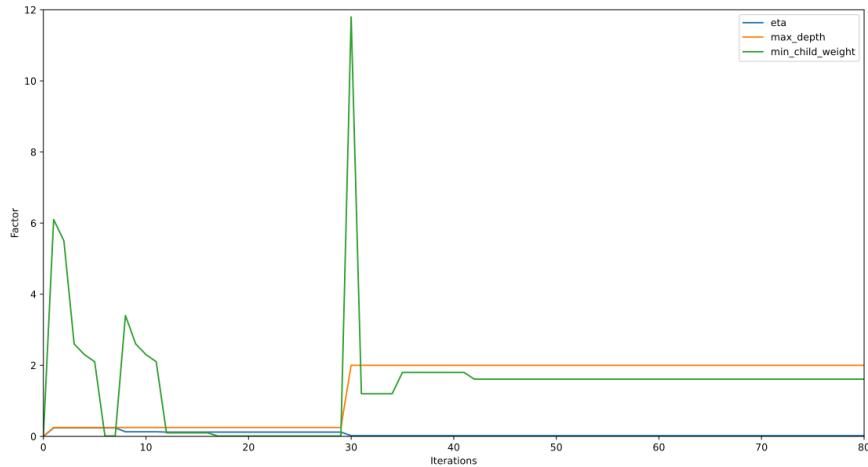


Figura 5.5: Evolution of hyperparameters throughout the iterations.

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

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

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

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

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

Cuadro 5.5: 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 5.5, y en 5.7 se observa la representación en imagen 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.

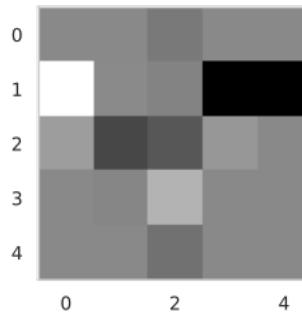


Figura 5.7 Imagen de las características.

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

Entrenamientos

Las figuras 5.9 y 5.10 muestran la evolución de la métrica de puntuación F1 a lo largo de las 100 ejecuciones para las redes neuronales convolucionales 1D y 2D. Al visualizar la convolución unidimensional (Figura 5.9), se puede verificar que la puntuación F1 de entrenamiento aumentaba ligeramente a lo largo de las épocas, experimentando altibajos a medida que el modelo se entrena, comenzando inicialmente con un valor de entrenamiento inferior a 0,58 y llegando hasta 0,68, mostrando poca capacidad de aprendizaje y generalización ante nuevas muestras.

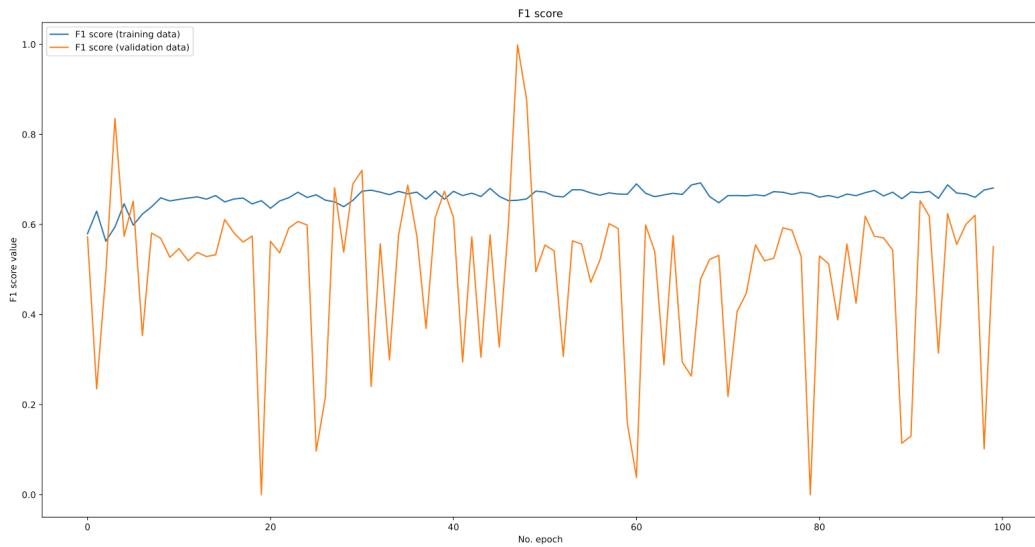


Figura 5.9: Evolution of the F1-score of the 1D-CNN in the training and test set.

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.

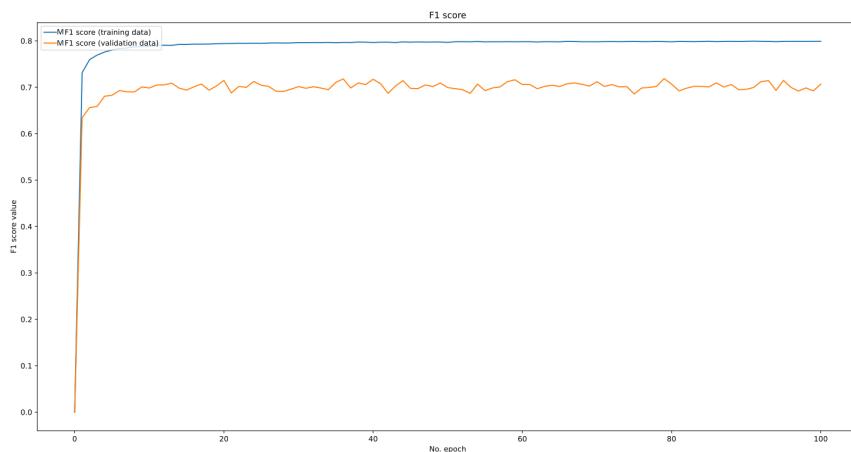


Figura 5.10: Evolution of the F1-score of the 2D-CNN in training and test set.

Resultados

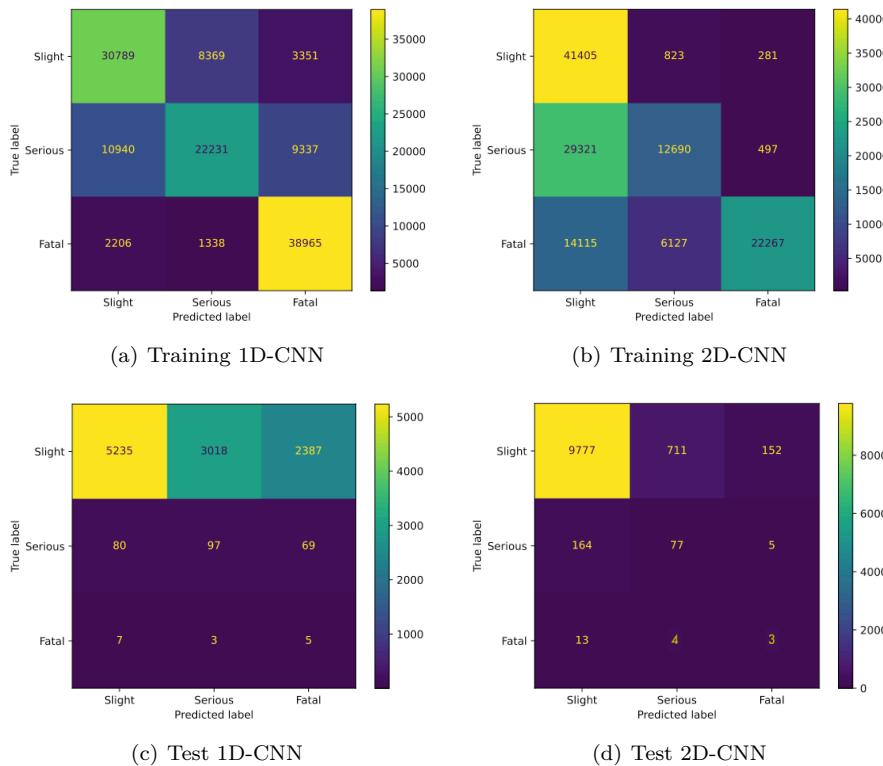


Figura 5.11: Confusion Matrices for Convolutional neural networks.

Resultados

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

Los resultados de las métricas de clasificación se muestran en la Tabla 5.8 para cada una de las clases predichas en el conjunto de **pruebas**. Estos informes muestran la información con la que evaluamos los modelos, ya que explica cómo se comportan con respecto a nuevos datos.

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

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

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

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

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

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

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

Cuadro 5.6: Training metrics for 1D-CNN and 2D-CNN.

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

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

Resultados

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

Cuadro 5.8: Test metrics classification for GNB, SVC and KNN.

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

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

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

5.2. Resultados finales

Aquí pones los resultados del paper 3

En esta sección se expondrán los resultados ofrecidos por la metodología GTAAF, como se ha comentado anteriormente, esta metodología es una evolución del prototipo anterior. Para ello se analizarán los datos con los que se ha ejecutado este método, y los resultados de cada una de las etapas que lo componen aplicadas a estos datos.

5.2.1. Dataset

Los datos escogidos para esta tesis pertenecen a tres conjuntos de datos distintos, con el objetivo de poder evaluar la metodología y el nuevo modelo propuesto en diferentes contextos. Esta evaluación se basará en dos factores principales; la distinta disponibilidad de información en los datos, y en distintos casos de estudio en función de la densidad de población de las regiones escogidas. Teniendo en cuenta la densidad de población podemos distinguir entre tres casos

de estudio claramente diferenciados: (1) alta concentración de población, (2) concentración media y (3) concentración dispersa. Con esta variabilidad en los datos se busca medir la robustez y generalización de la técnica desarrollada.

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

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

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

5.2.2. Descripción de datos

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

1. Magnitud y ubicación del accidente: enfocado en información relativa a la localización y magnitud del accidente, como datos geográficos.

2. Limitaciones de Conducción: abarcan características que limitan al conductor, como regulaciones 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.
5. Información del Vehículo: características que describan al vehículo objeto del accidente.
6. Información de la Víctima: descriptores que definan a la víctima en el momento del incidente, factores como la edad, sexo, positivo en sustancias estupefacientes, etc.

En la Figura 5.12 se presentan las características disponibles en cada uno de los conjuntos de datos escogidos en esta tesis, con el objetivo de mostrar la variabilidad de información que puede existir entre distintas poblaciones. Los campos marcados en naranja son aquellos que representan información distinta entre los datasets pero que pueden ser incluidos en las categorías correspondientes. Por otra parte, aquellos campos marcados en rojo indican la ausencia de este tipo de características en comparación con el resto de datasets.

	UK	Madrid	Victoria
<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 5.12: Classification of variables. Fields shown in yellow represent features of the same nature but differing in data granularity. Additionally, missing features compared to other datasets are highlighted in red.

5.2.2.1. Partes comunes entre los datos

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

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

5.2.2.2. Principales diferencias entre los datos

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

UK

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

Madrid

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

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

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

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

Victoria

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

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

5.2.3. Limpieza

En la tabla 5.11 se expone el número total de registros del conjunto de datos original y el número de muestras resultante tras haber aplicado la limpieza de estos datos para cada una de las poblaciones contempladas en esta tesis.

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

Cuadro 5.9: XX

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

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

5.2.4. Filtrado de áreas

Para ilustrar los parámetros con los que se aplica el filtrado de áreas, se presenta la Tabla 5.10, donde se muestra para cada población el número de áreas por las que pasará el filtro, además del tamaño de la ventana X,Y para cada región, donde en función de la extensión y la densidad de población tomará valores más grandes (poblaciones más dispersas) o valores más pequeños (cuando la densidad de población es alta). Los tamaños de ventana para cada pobla-

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

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

Cuadro 5.10

Una vez se establecen las dimensiones de las ventanas de tamaño X,Y se aplica el filtrado para cada ciudad, donde el número de muestras de la clase mayoritaria se ve considerablemente rebajado respecto a la minoritaria, con el objetivo de tener un conjunto de datos más balanceado. La tabla 5.11 muestra el número de registros original para cada población y el número de registros resultante tras aplicar el filtrado por áreas.

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

Cuadro 5.11

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

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

5.2.5. Discretización

Tabla de discretización.

UK UK discretización

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

Cuadro 5.12: UK classification of variables.

Madrid Madrid discretización, in progress...

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

Cuadro 5.13: Victoria classification of variables.

Victoria

Victoria discretización, in progres...

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

Cuadro 5.14: UK classification of variables.

5.2.6. Transformación (Sin/Cos)

5.2.7. Resampling

En la tabla 5.15 se muestran los datos resultantes tras haber aplicado el proceso de resampling mediante la generación de datos sintéticos de SMOTE-II, conformando un dataset balanceado en el que se previene el riesgo de sesgo de datos por parte de la red.

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

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

Cuadro 5.15: REPASAR PORQUE LOS NÚMEROS NO ME CUADRNAN, HEMOS MANDADO EL PAPER 3 TAMBIÉN ASÍ.

5.2.8. Normalización

En la Tabla 5.16 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 5.16: blabla

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

5.2.9. Categorización

5.2.10. División Train-Val-Test

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

5.2.11. Cálculo de pesos

5.2.12. Feature Importance Algorithm

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

5.2.13. Algoritmo Genético

En la tabla 5.17 se muestran los hiperparámetros del algoritmo genético utilizados para optimizar el algoritmo XGBoost. Durante cada una de las generaciones, el límite máximo de individuos en la población es de 50 individuos (fila Population). Estos individuos en cada generación son evaluados mediante la función heurística a optimizar, la métrica F1-Score resultante del algoritmo XGBoost sobre el conjunto de test accidentes, es decir, en los accidentes no vistos durante el entrenamiento (fila Fitness Function). Una vez son evaluados, aquellos 10 mejores individuos son seleccionados para intercambiar su información, es decir, los padres que darán lugar a 10 nuevos individuos de cara a la próxima generación (fila Parents Mating). La mezcla de información entre padres se realiza mediante una estrategia de cruce mixta (fila Crossover Index), es decir, para cada par de padres se asigna un índice aleatorio en sobre el que se dividirán ambos individuos para luego combinar esta información en el descendiente resultante **esto se aplica para left join y right join**. Una vez se han dado lugar a los 10 nuevos individuos, el valor de cada una de las componentes que los conforman pueden ser modificados con una probabilidad del 40 % (fila Mutation Probability). Este proceso será repetido a lo largo de todas las 50 generaciones (fila Generations).

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

Cuadro 5.17: Genetic algorithm hyperparameters setup.

La Tabla 5.18 muestran las variaciones en los valores máximos y mínimos permitidos para cada variable a optimizar mediante el algoritmo genético. La fila *Initial* de cada hiperparámetro muestra el rango de valores que cada individuo puede tomar cuando es inicializado. En la fila *Mutation* se observan, para cada hiperparámetro, los valores límites permitidos sobre los que los componentes de un sujeto pueden modificarse en el proceso de mutación, siempre y cuando dicho componente haya sufrido una mutación.

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

Cuadro 5.18: Boosting models hyperparameters limits.

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

En la Tabla 5.19 se pueden observar los hiperparámetros óptimos resultantes de la ejecución del algoritmo genético, blabla...

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

Cuadro 5.19: Resulting boosting model hyperparameters after executing the genetic algorithm.

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

En la figura 5.13 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 5.13(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 5.13(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.

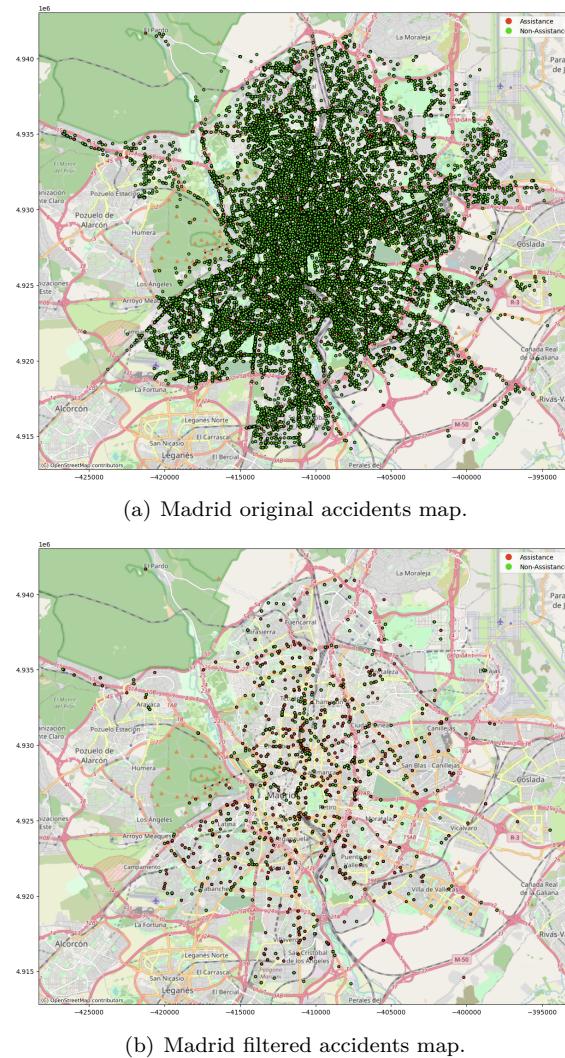


Figura 5.13: Madrid original/filtered accidents map.

En la Figura 5.14 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.

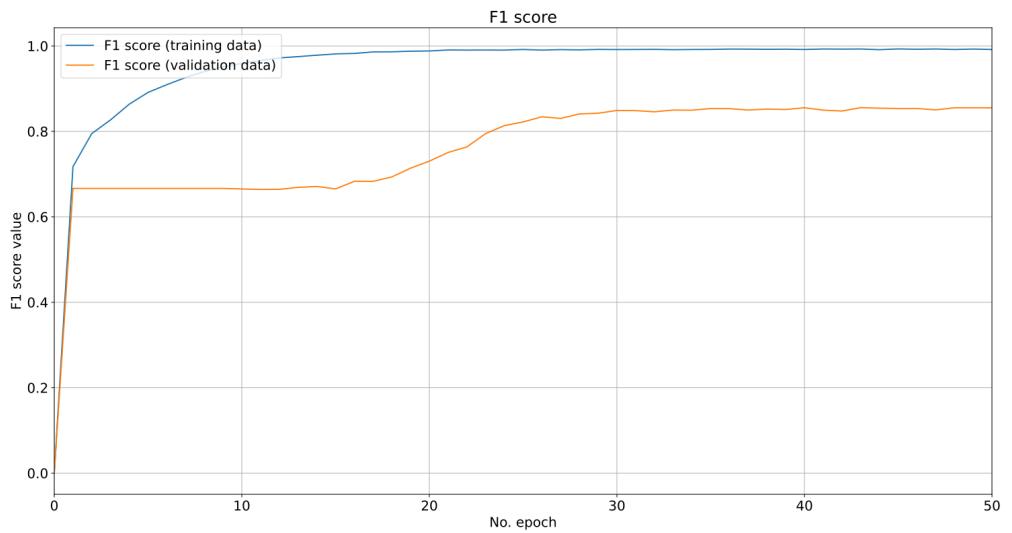


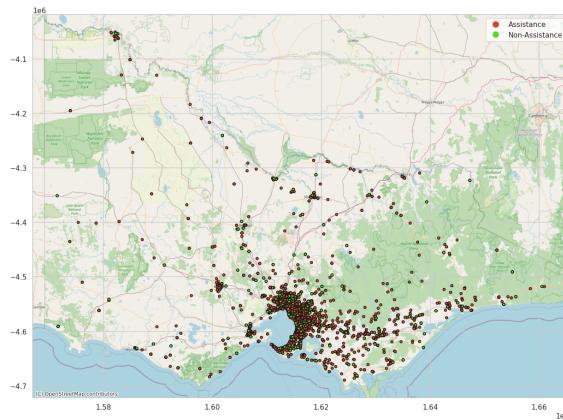
Figura 5.14: Evolution of F1-Score Madrid.

En la tabla 5.20 se observan los resultados de la métrica F1-Score de la predicción de la severidad de los accidentes de cada uno de los modelos sobre el conjunto de test de la ciudad de Madrid. Como se puede comprobar, el valor más alto lo ofrece el nuevo modelo GTAAF propuesto, llegando a mejorar en un 3,9 % al siguiente mejor modelo, el SVC sobre los accidentes Slight, mientras que la mejora sobre los accidentes Assistance se mide en un 5,7 % sobre el siguiente modelo que mejor métricas ofrece, el SVC. Con estos resultados puede interpretar que el nuevo modelo GTAAF propuesto es capaz de generalizar mejor en la predicción de la severidad de nuevos accidentes que no ha visto previamente sobre la ciudad de Madrid.

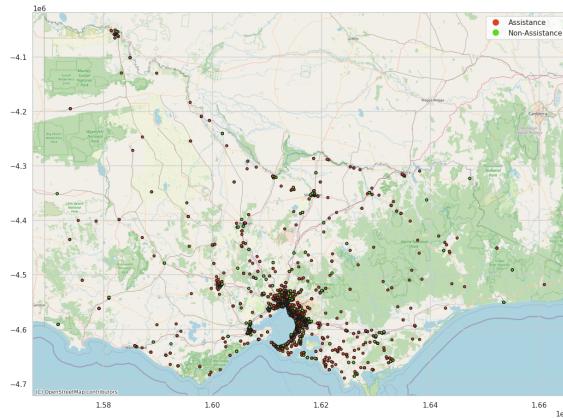
		Spain region F1-Score
Model	Assistance	Madrid
NB	No	0.729
	Yes	0.621
SVC	No	0.862
	Yes	0.748
KNN	No	0.739
	Yes	0.634
RF	No	0.744
	Yes	0.643
LR	No	0.750
	Yes	0.623
MLP	No	0.856
	Yes	0.724
GTAAF	No	0.894
	Yes	0.798

Cuadro 5.20: F1-Scores by Accident Class on Madrid (Spain).

En el segundo caso, tenemos una región dispersa, el estado de Victoria (Australia). Victoria, un estado en Australia, abarca una región diversa con ciudades bulliciosas como Melbourne, situada a lo largo de la costa sureste, conocida por su densidad de población moderada a alta y una mezcla de vitalidad urbana. En la figura 5.15 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 5.15(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 5.15(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 5.15: Victoria original/filtered accidents map.

En la Figura 5.16 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.

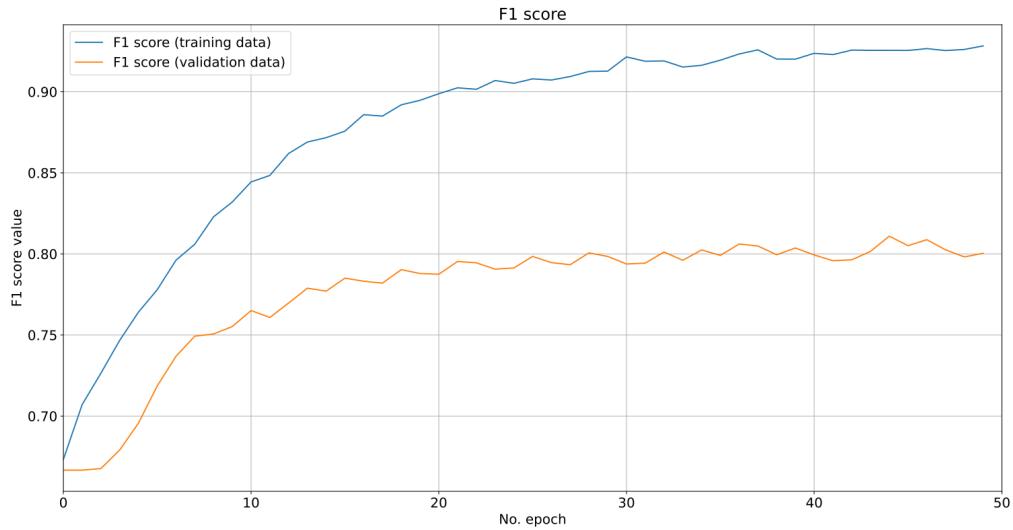


Figura 5.16: Evolution of F1-Score Victoria.

En la Tabla 5.21, se presentan los resultados del F1-Score obtenidos por cada uno de los modelos para ambos tipos de clasificación de accidentes. Específicamente, se observa que para la población de Victoria, el nuevo modelo GTAAF propuesto logra una mejora del 6.5 % en comparación con el siguiente mejor modelo, el SVC, para accidentes de No Asistencia. Por otro lado, en lo que respecta a accidentes de tipo Asistencia, hay una mejora del 9 % en comparación con el MLP. Estos resultados reflejan una mejora significativa en la capacidad de generalización del nuevo modelo propuesto.

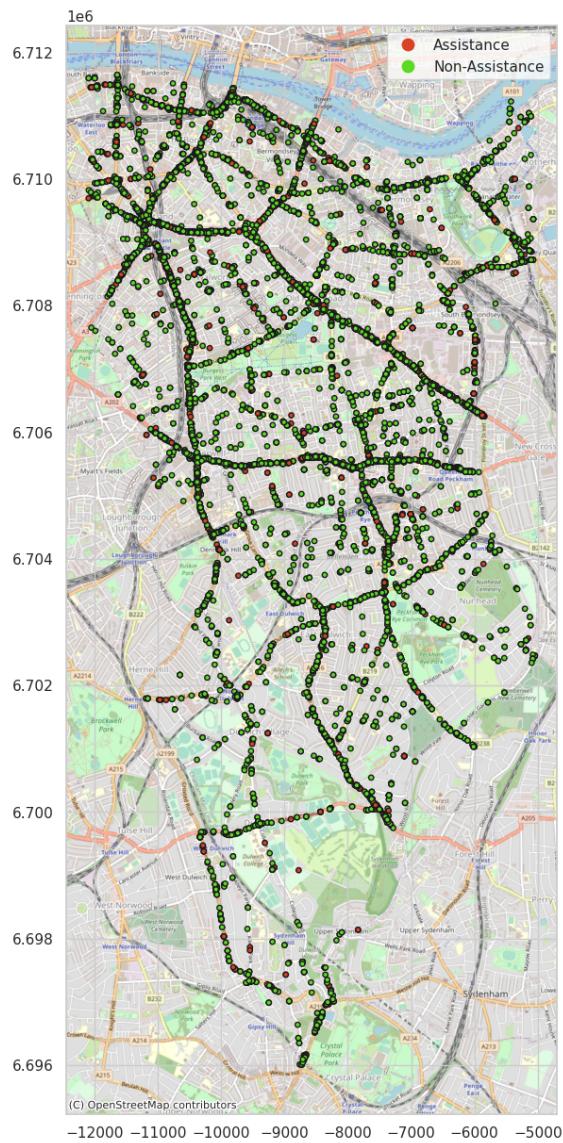
Model	Assistance	Australia region F1-Score
		Victoria
NB	No	0.635
	Yes	0.476
SVC	No	0.662
	Yes	0.679
KNN	No	0.654
	Yes	0.616
RF	No	0.647
	Yes	0.364
LR	No	0.612
	Yes	0.630
MLP	No	0.635
	Yes	0.694
GTAAF	No	0.727
	Yes	0.784

Cuadro 5.21: F1-Scores by Accident Class on Victoria (Australia).

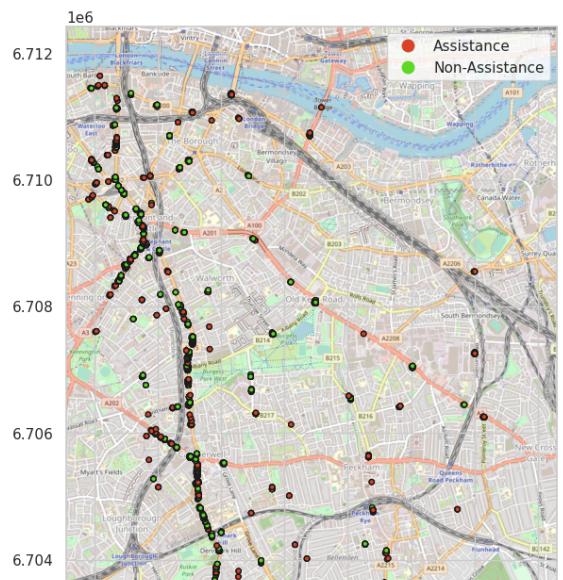
En la Tabla 5.21 se muestran los resultados F1-Score obtenidos de cada uno de los modelos respecto a ambos tipos de clasificación de accidentes. Concretamente se observa que para la ciudad de Victoria el nuevo modelo GTAAF propuesto obtiene una mejora respecto al siguiente mejor modelo, el SVC, para los accidentes Slight del 6,5 %. Por otra parte, en lo que respecta a los accidentes tipo Assistance se obtiene una mejora del 9 % respecto al MLP. Estos resultados reflejan una mejora de generalización significativa del nuevo modelo propuesto.

Southwark

Southwark es un distrito de Londres situado en la orilla sur del río Támesis, con una alta densidad de población. En la Figura 5.17 se observa la distribución de los accidentes a lo largo del municipio de Southwark. Analizando los accidentes del conjunto de datos original, Figura 5.17(a), se observa que estos se producen a lo largo de las distintas vías que conectan el municipio con el centro neurálgico de la ciudad, hecho habitual al conectar zonas menos pobladas con lugares de trabajo y de ocio, mientras que la minoría de ellos se presentan en las calles aledañas. Observando la distribución de accidentes tras el proceso de filtrado por áreas (Figura 5.17(b)) se acentúa este hecho, donde se observan que aquellos principales accidentes Assistance se producen en estas vías. Por otra parte se muestra una concentración minoritaria de este tipo de accidentes en la zona de Dulwich (sur), en la intersección de la circunvalación S Circular red con la carretera que dirige al centro del municipio.



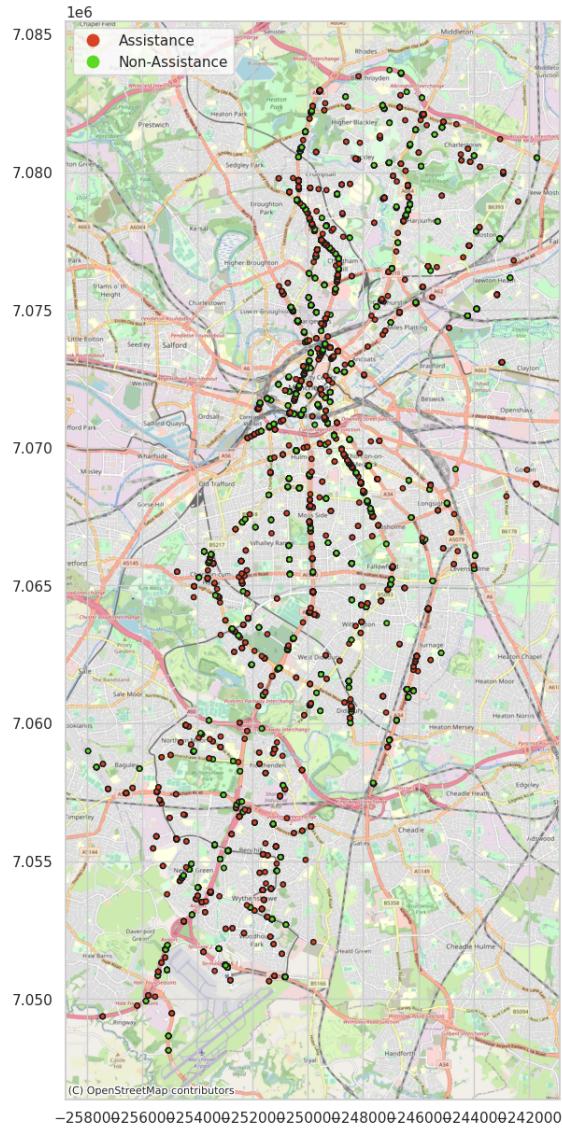
(a) Southwark original accidents map.



Manchester

Manchester, ubicada en el noroeste de Inglaterra, es una gran ciudad conocida por su legado industrial y su alta densidad de población. En la Figure 5.18 se muestra la distribución de los accidentes de Manchester. Atendiendo a la distribución de accidentes original, en la Figura ??, como es habitual en cualquier población se aprecia una concentración de accidentes importante en la zona central de la ciudad, siendo también considerable en el área de Longsight. Por otra parte, las principales vías que comunican las periferias urbanas (norte) con el centro de la ciudad también presentan una concentración mayor de accidentes, lo que puede deberse a desplazamientos por trabajo. Por otra parte, la carretera de Wythenshawe, cercano a Sale Water Park (Sur), también presenta una concentración elevada de accidentes, motivados por los desplazamientos de ocio y de trabajo. En la figura 5.18(b) se observa la localización de los accidentes una vez se ha aplicado el proceso de filtrado por áreas, donde se vislumbra que gran parte de los accidentes Assistance se distribuyen a lo largo de las carreteras que comunican hacia el centro de la ciudad.

(a) Manchester original accidents map.



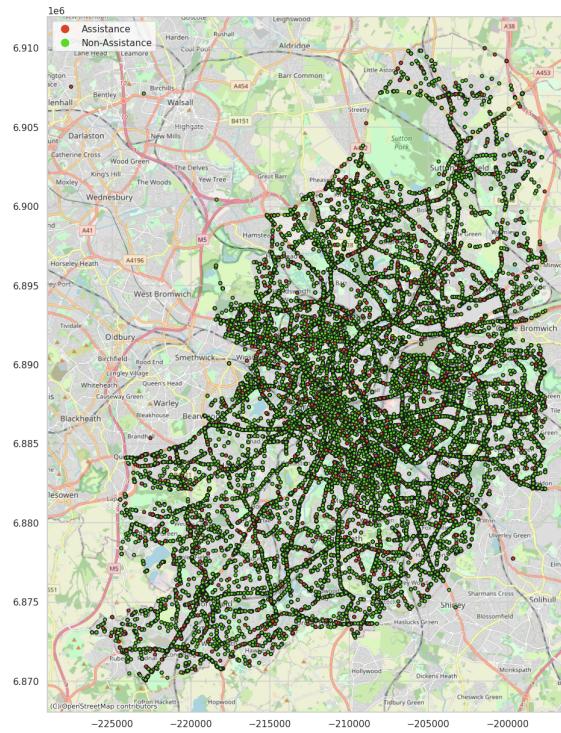
(b) Manchester filtered accidents map.

Figura 5.18: Manchester original/filtered accidents map.

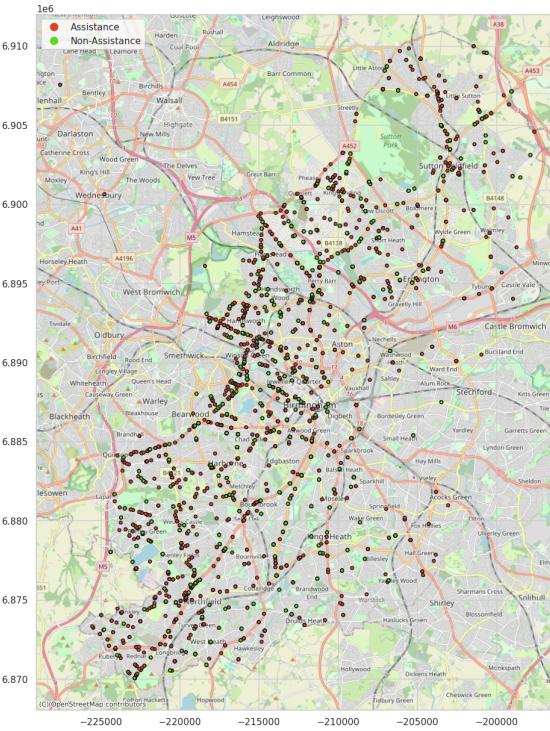
Birmingham

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

West Midlands con un paisaje urbano diverso y una densidad de población considerable, famosa por su historia industrial y su vitalidad cultural. En la Figura 5.19 se muestra la distribución de los accidentes de Birmingham, tanto los originales como los resultantes una vez aplicado el proceso de filtrado. Como se puede observar en los accidentes originales en la Figura 5.19(a) se aprecia que gran parte de los accidentes se concentran en la zona centro de la ciudad, una tendencia normal debido a que es el principal foco de actividad de las ciudades. Mientras que los accidentes se van dispersando a medida que distan de este punto. Se aprecian ligeras agrupaciones de accidentes a lo largo de las zonas de incorporaciones a las principales arterias de la ciudad, como es al este, el caos de Handsworth. Por otra parte, en la figura 5.19(b) se muestran los accidentes una vez se ha aplicado el proceso de filtrado por áreas. Como se puede observar, la información ha sido resumida sin dar lugar a pérdidas en el valor de la misma. Se vislumbran ciertas zonas más conflictivas donde se producen accidentes más importantes, como es el caso de la carretera Holyhead Rd de entrada a la ciudad o en Northfield.



(a) Birmingham original accidents map.

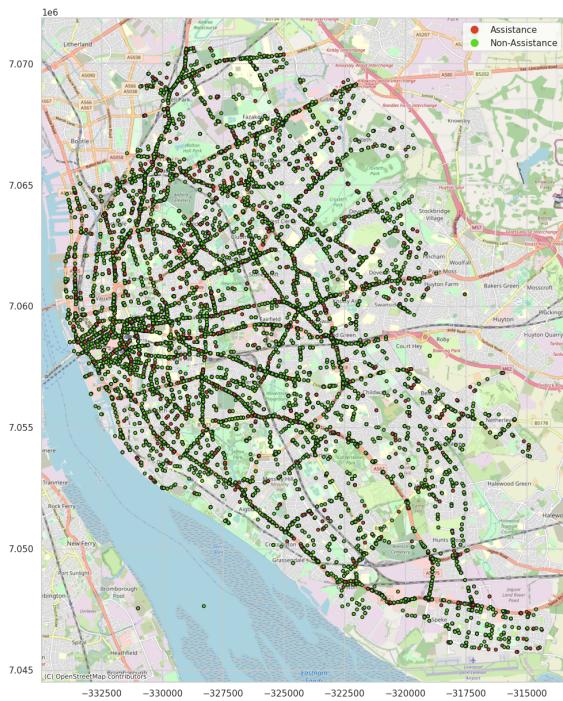


(b) Birmingham filtered accidents map.

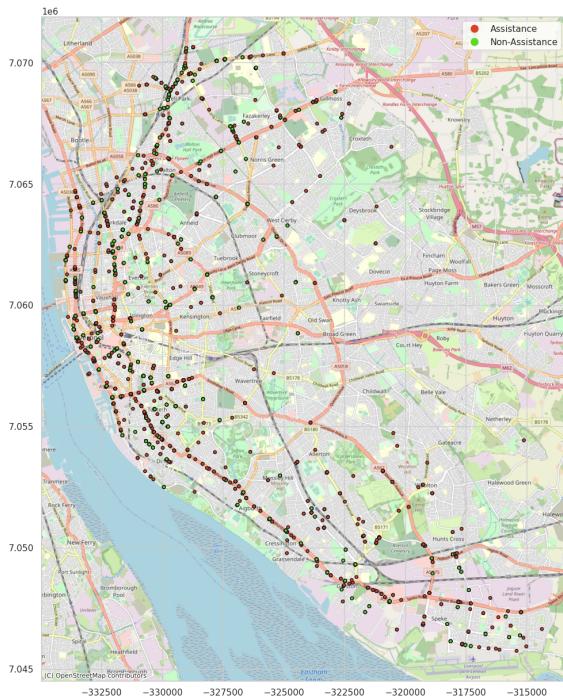
Figura 5.19: Birmingham original/filtered accidents map.

Liverpool

Liverpool, ubicada a lo largo del río Mersey en el noroeste de Inglaterra, prospera como una ciudad marítima con una rica historia, profundidad cultural y una densidad de población significativa, reconocida por su encanto en el frente marítimo y su legado musical. En la Figura 5.20 se muestra la comparativa de la distribución de accidentes originales del dataset y filtrados para la ciudad de Liverpool. En la Figura 5.20(a) se aprecian accidentes concentrados en la zona centro de la ciudad, como viene siendo habitual, además de a lo largo de las circunvalaciones que la rodean. En la Figura 5.20(b), después del proceso de filtrado, se aprecia que gran parte de los accidentes Assistance se producen a lo largo de Strand Street (desde el sur hasta el oeste), convergiendo ambas direcciones en el centro 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.



(a) Liverpool original accidents map.

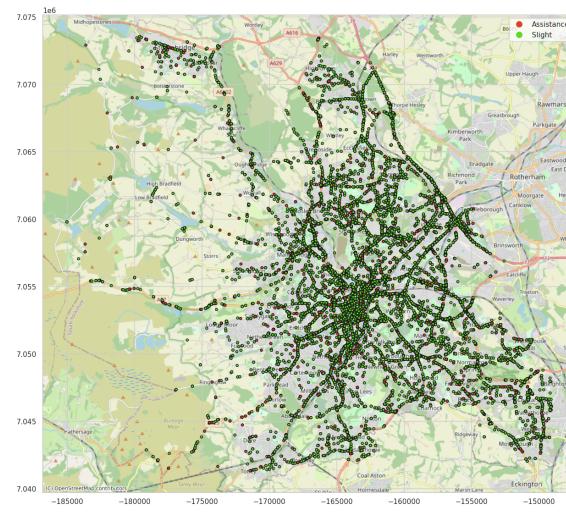


(b) Liverpool filtered accidents map.

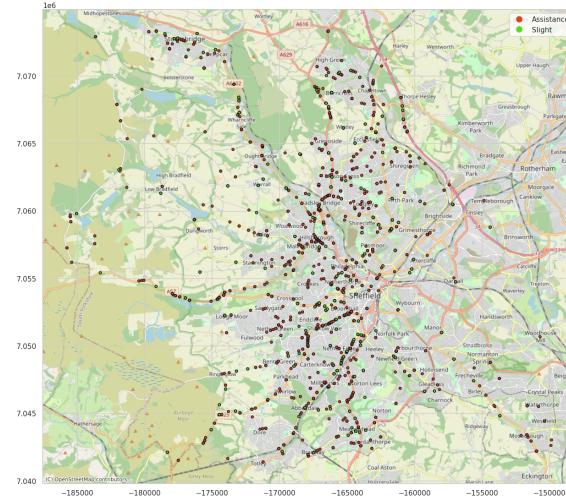
Figura 5.20: Liverpool original/filtered accidents map.

Sheffield

Sheffield, ubicada en South Yorkshire, presume de un patrimonio industrial y paisajes pintorescos, con una densidad de población intermedia. En la Figura 5.21 se muestra la distribución de accidentes para la ciudad de Sheffield, tanto la original como la resultante tras la etapa de filtrado. En la Figura 5.21(a) se pueden apreciar distintas concentraciones en zonas estratégicas. Como suele ser habitual, el núcleo urbano es un centro de mayor densidad de incidentes, mientras que en las intersecciones que conectan la ciudad de Sheffield y la de Rotherham (cruces de Tinsley Viaduct con Meadow Bank Road y la A6178, al noreste de Sheffield). También se aprecian concentraciones en los suburbios de Wadsley Bridge y Malin Bridge, periferias de la ciudad, además de alrededor de todas las vías principales que conectan con el centro. Por otra parte, en la figura 5.21(b) se muestran los accidentes una vez se ha realizado el proceso de filtrado, donde se aprecia que aquellos que han requerido de asistencia normalmente se presentan en las principales arterias, donde se circula a una mayor velocidad.



(a) Sheffield original accidents map.



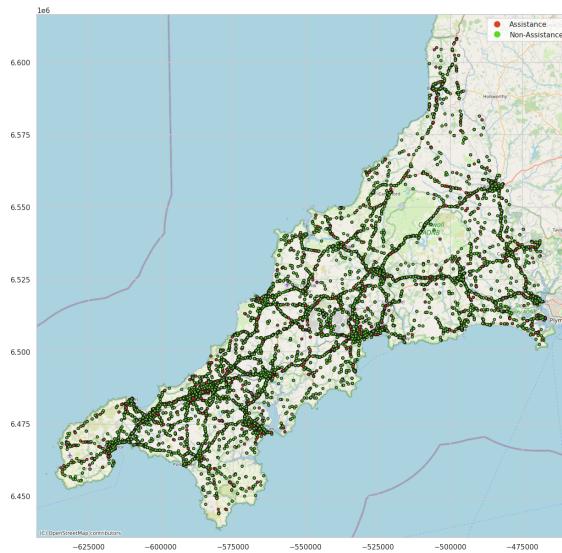
(b) Sheffield filtered accidents map.

Figura 5.21: Sheffield original/filtered accidents map.

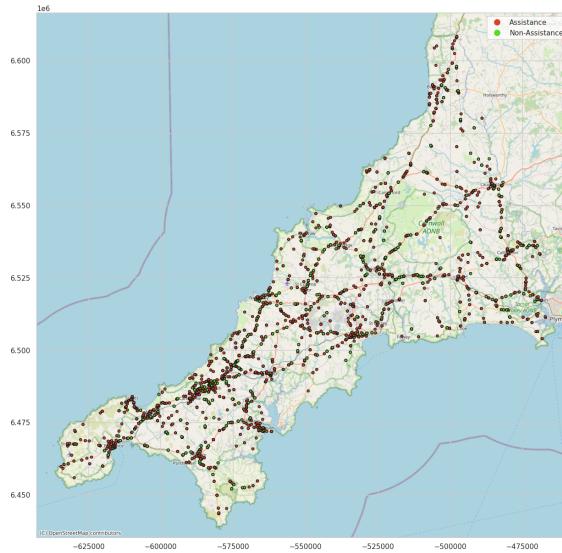
Cornwall

Cornualles, situada en la parte suroeste de Inglaterra con sus apacibles paisajes, encantadores pueblos costeros y extensiones rurales, fomenta un entorno tranquilo alejado de los núcleos de alta densidad de población. En la Figura 5.22 se muestran de nuevo los accidentes originales de dataset y los que resultan tras aplicar el proceso de filtrado sobre el condado de Cornwall. En la

Figura 5.22(a) las principales concentraciones de accidentes se encuentran distribuidas a lo largo de las distintas ciudades del condado. La mayoría de estos se encuentran divididos en dos regiones claramente definidas, la primera de ellas entre las vías que conectan las localidades de Camborne y Redruth (suroeste de Cornwall), y el área comprendida entre St Austell, Duporth, Carlyon Bay y Par, este del condado. No obstante, el resto de regiones también presentan una concentración considerable, como es el caso de la ciudad de Falmouth (sureste), las localidades de Penzance y Hayle (suroeste), en la ciudad de Newquay y sus alrededores (oeste), Bodmin (centro) y Launceston (norte). De nuevo, en este caso, se demuestra que la mayor frecuencia de accidentes se presenta entre los principales núcleos de población y las carreteras que los interconectan, debido a que las grandes ciudades implican más movimientos de vehículos. Atendiendo a la Figura 5.22(b) se observa que la ocurrencia de accidentes necesarios de asistencia, una vez aplicado el proceso de filtrado, tiene la misma tendencia que el expuesto para el conjunto de datos original, distribuyéndose a lo largo de las principales carreteras del condado Cornwall y concentrándose más en los núcleos de población.

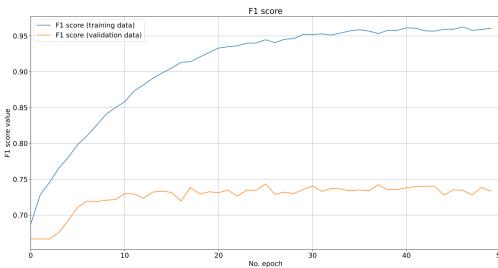


(a) Cornwall original accidents map.

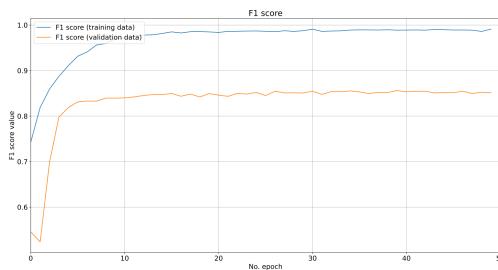


(b) Cornwall filtered accidents map.

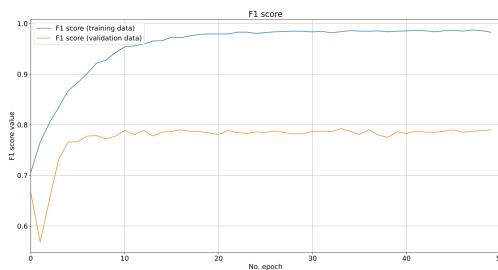
Figura 5.22: Cornwall original/filtered accidents map.



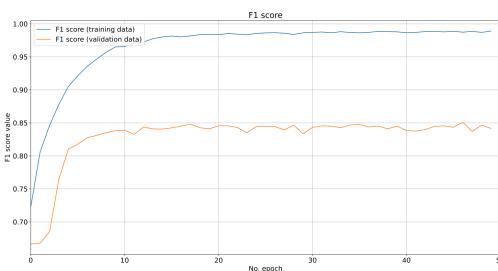
(a) Training 2D-CNN Southwark



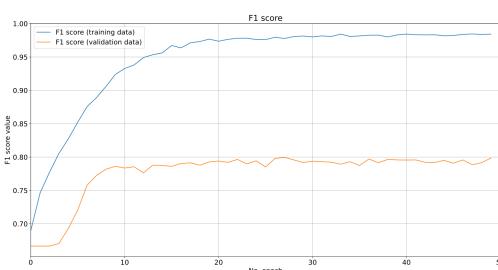
(b) Training 2D-CNN Manchester



(c) Training 2D-CNN Birmingham



(d) Training 2D-CNN Liverpool



(e) Training 2D-CNN Sheffield

En la Tabla 5.22 se muestra el valor F1-Score para cada una de las ciudades de cada modelo sobre el conjunto de test. Como se puede observar, el nuevo modelo propuesto GTAAF es el que mejor métricas ofrece en comparación al resto, obteniendo la mayor diferencia con respecto a su sucesor en los accidentes Slight, el MLP, para la ciudad de Manchester de un 5,33 %, mientras que la mayor diferencia para los accidentes tipo Assistance es de un 13,8 % en la ciudad de Southwark respecto al siguiente mejor modelo, el MLP. La siguiente mayor diferencia se presenta entre el modelo GTAAF se presenta en la ciudad de Southwark para la clase Slight, con un incremento del 4,8 % respecto al siguiente mejor modelo MLP, mientras que para la clase Assistance ésta se presenta en la ciudad de Liverpool con respecto al modelo MLP, llegando a un 13,2 %. Observando los resultados de la tabla, se aprecia que el mayor incremento del rendimiento con respecto al resto de modelos se presenta sobre la clase Assistance, obteniendo de media una mejora de 9,21 % sobre todas las ciudades, mientras que el incremento de rendimiento se acentúa menos en la Slight, ya que el resto de modelos ofrecen unas métricas más altasW, siendo la mejora de un 3,93 % de media. Estos resultados reflejan una mejor generalización del modelo propuesto en comparación al resto de modelos estudiados para cada una de las ciudades de Reino Unido.

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

Cuadro 5.22: F1-Scores comparison by traffic accident assistance on six UK areas.

La Figura 5.24 muestra a modo de comparativa el rendimiento del nuevo modelo GTAAF propuesto en los accidentes Slight para cada una de las poblaciones estudiadas respecto al resto de modelos del estado del arte con los que se ha experimentado en esta investigación. Se aprecia un incremento de

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

F1-Score by region (Non-Assistance accidents)

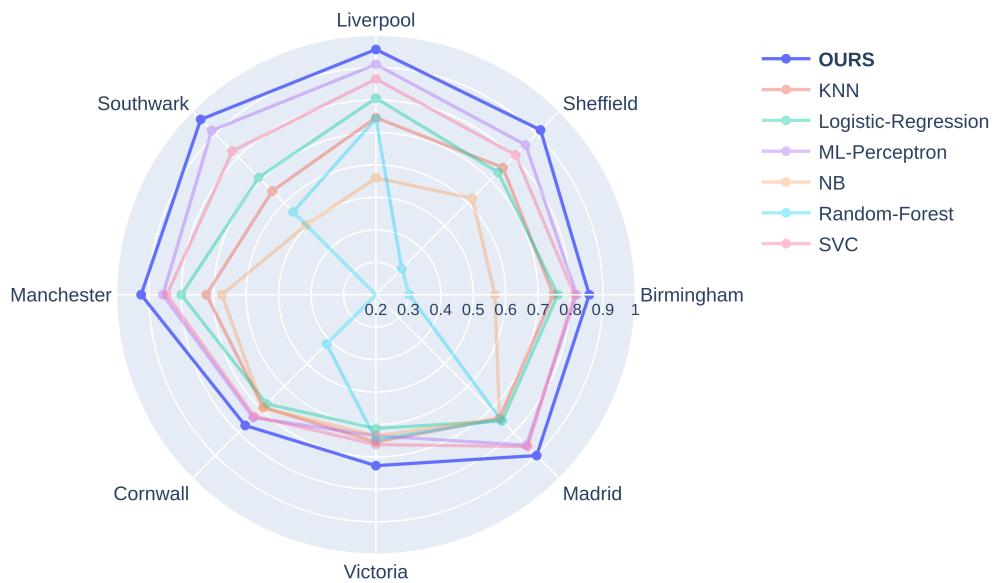
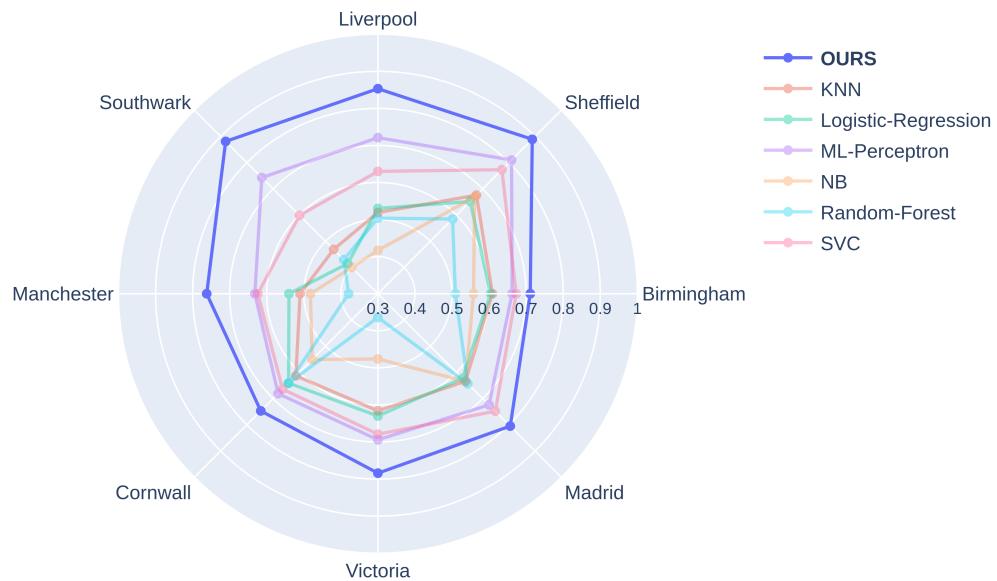


Figura 5.24: F1-Scores Comparison for Non-Assistance Accidents.

La Figura 5.25 muestra la comparativa del rendimiento basado en el F1-Score de los modelos para cada una de las ciudades en los accidentes Assistance. En esta gráfica se puede observar una diferencia considerablemente mayor del nuevo modelo GTAAF propuesto respecto al resto en comparación con los accidentes Slight. La mayor diferencia de mejora de este modelo GTAAF se presenta en la ciudad de Southwark, con un incremento del 13.8 % respecto al siguiente mejor modelo sobre esta población, el MLP.

F1-Score by region (Assistance accidents)

**Figura 5.25:** F1-Scores Comparison for Assistance Accidents.

5.3. Pruebas de estrés

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

En la tabla 5.23 ...

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

Cuadro 5.23: F1-Scores comparison with features loss in Madrid dataset. In bold the best result (our model)

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

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

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

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

En la Figura 5.26 podemos observar la diferencia de los algoritmos con y sin pérdida de características (Tablas 5.22 y 5.21 versus Tablas ?? y 5.24). Mostramos cómo nuestro modelo mejora sus propios resultados en todos los casos.

Por ejemplo, en Cornwall, obtenemos una mejora en nuestro modelo del 4.8% y 5.9% en No Assistance y Assistance (eliminando la peor característica, ver la barra azul en la Figura 5.26-arriba), 1.4% y 2.9% (eliminando la mejor característica, ver la barra verde en la Figura 5.26-arriba) y 7% y 8% (eliminando ambas características, ver la barra gris en la Figura 5.26-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 5.26-abajo), -0.7% y 0.9% (ver la barra verde en la Figura 5.26-abajo), y 5.1% y 3% (ver la barra gris en la Figura 5.26-abajo), respectivamente. Esto indicaría un efecto menor de los valores extremos en la ponderación del algoritmo genético si el área es más dispersa.

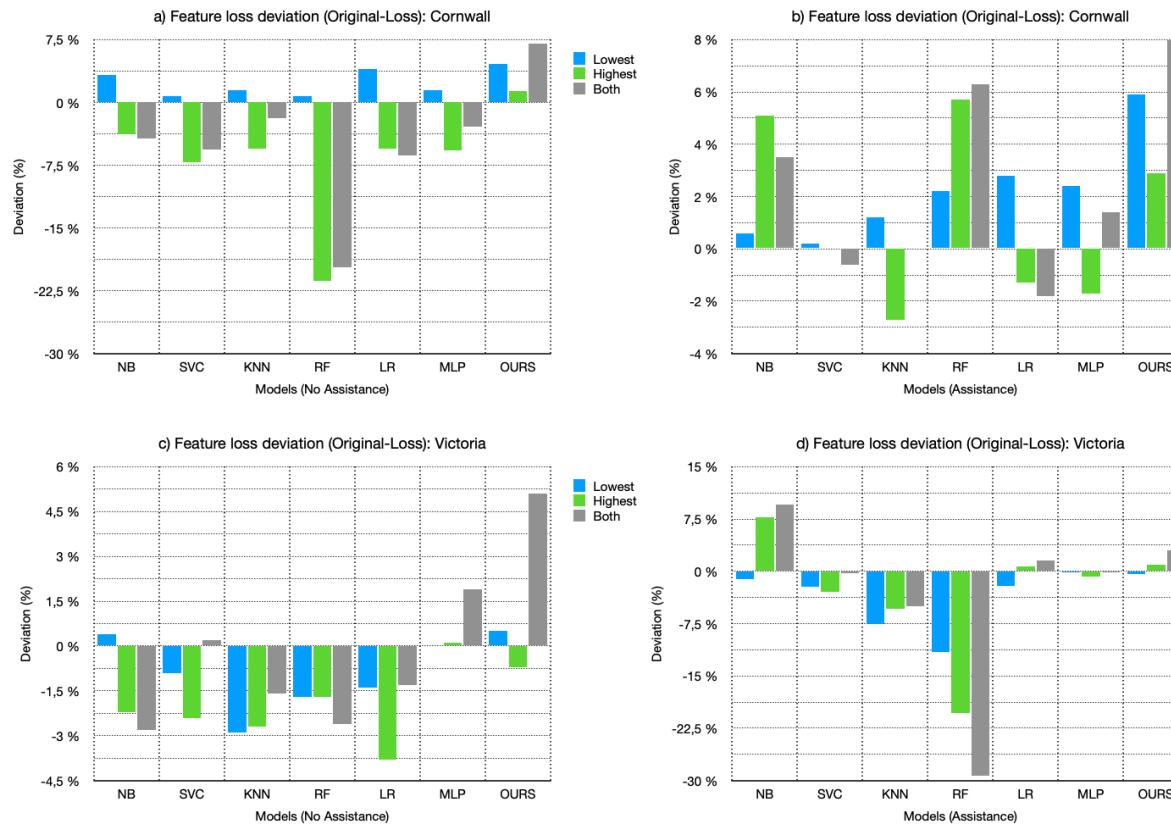


Figura 5.26: Comparación de pérdida de características. Las barras representan la diferencia entre los resultados con todas las características y los resultados sin características extremas: en azul sin la característica más baja, en verde sin la característica más alta y en gris sin ambas características extremas.

Capítulo 6

Conclusiones

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

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

Capítulo 7

Publications

Capítulo 8

Anexos

Madrid paper I discretization:

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

Cuadro 8.1: Numerical assignment of the dataset variables.