



Escuela
Politécnica
Superior

Título del Trabajo Fin de Grado/Máster



Grado en Ingeniería en Sonido e
Imagen en Telecomunicación

Trabajo Fin de Grado

Autor:

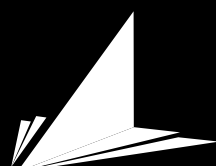
Luis Pérez-Sala García-Plata

Tutores:

José Francisco Vicent Frances

Nombre Apellido1 Apellido2 (tutor2)

Mayo 2022



Escuela
Politécnica
Superior

Título del Trabajo Fin de Grado/Máster

Subtítulo del proyecto

Autor

Luis Pérez-Sala García-Plata

Tutores

José Francisco Vicent Frances

Ciencia de la Computacion e Inteligencia Artificial

Nombre Apellido1 Apellido2 (tutor2)

Departamento del cotutor

Grado en Ingeniería en Sonido e Imagen en Telecomunicación

ALICANTE, Mayo 2022

Preámbulo

Poner aquí un texto breve que debe incluir entre otras:

“las razones que han llevado a la realización del estudio, el tema, la finalidad y el alcance y también los agradecimientos por las ayudas, por ejemplo apoyo económico (becas y subvenciones) y las consultas y discusiones con los tutores y colegas de trabajo. (?)”

Agradecimientos¹

Este trabajo no habría sido posible sin el apoyo y el estímulo de mi colega y amigo, Doctor Rudolf Fliesning, bajo cuya supervisión escogí este tema y comencé la tesis. Sr. Quentin Travers, mi consejero en las etapas finales del trabajo, también ha sido generosamente servicial, y me ha ayudado de numerosos modos, incluyendo el resumen del contenido de los documentos que no estaban disponibles para mi examen, y en particular por permitirme leer, en cuanto estuvieron disponibles, las copias de los recientes extractos de los diarios de campaña del Vigilante Rupert Giles y la actual Cazadora la señorita Buffy Summers, que se encontraron con William the Bloody en 1998, y por facilitarme el pleno acceso a los diarios de anteriores Vigilantes relevantes a la carrera de William the Bloody.

También me gustaría agradecerle al Consejo la concesión de Wyndham-Pryce como Compañero, el cual me ha apoyado durante mis dos años de investigación, y la concesión de dos subvenciones de viajes, una para estudiar documentos en los Archivos de Vigilantes sellados en Munich, y otra para la investigación en campaña en Praga. Me gustaría agradecer a Sr. Travers, otra vez, por facilitarme la acreditación de seguridad para el trabajo en los Archivos de Munich, y al Doctor Fliesning por su apoyo colegial y ayuda en ambos viajes de investigación.

No puedo terminar sin agradecer a mi familia, en cuyo estímulo constante y amor he confiado a lo largo de mis años en la Academia. Estoy agradecida también a los ejemplos de mis difuntos hermano, Desmond Chalmers, Vigilante en Entrenamiento, y padre, Albert Chalmers, Vigilante. Su coraje resuelto y convicción siempre me inspirarán, y espero seguir, a mi propio y pequeño modo, la noble misión por la que dieron sus vidas.

Es a ellos a quien dedico este trabajo.

¹Por si alguien tiene curiosidad, este “simpático” agradecimiento está tomado de la “Tesis de Lydia Chalmers” basada en el universo del programa de televisión Buffy, la Cazadora de Vampiros.<http://www.buffy-cavampiros.com/Spiketesis/tesis.inicio.htm>

*A mi esposa Marganit, y a mis hijos Ella Rose y Daniel Adams,
sin los cuales habría podido acabar este libro dos años antes*²

²Dedicatoria de Joseph J. Roman en "An Introduction to Algebraic Topology"

*Si consigo ver más lejos
es porque he conseguido auparme
a hombros de gigantes*

Isaac Newton.

Índice general

1	Introducción	1
1.1	Motivación	1
1.2	Objetivos	2
1.2.1	Objetivos Generales	2
1.2.2	Objetivos Específicos	2
2	Marco Teórico	3
3	Tecnologías	5
3.1	Modelos	5
3.1.1	Algoritmos Genéticos	5
3.1.2	XGBoost	7
3.1.3	Redes Neuronales	8
3.1.3.1	Convolucionales	12
3.1.3.1.1	Unidimensionales	13
3.1.3.1.2	Bidimensionales	15
3.1.4	KNN	15
3.2	Especificaciones técnicas	15
3.2.1	Herramientas utilizadas	16
3.2.2	Especificaciones del servidor	17
4	Metodología	19
4.1	Diagrama de Gantt	19
4.2	Diagrama de flujo	20
4.3	Proceso	21
4.3.1	Datos	22
4.3.2	Remuestreo	30
4.3.3	Algoritmo Genético	32
4.3.4	XGBoost	34
4.3.5	Imágenes	34
4.3.6	Modelos	36
5	Resultados	39
5.1	Algoritmo Genético	39
5.2	XGBoost	41
5.3	KNN	41
5.4	Entrenamiento de modelos	42
5.4.1	Tiempos de entrenamiento	42
5.4.2	Gráficas de entrenamiento	42

5.4.3	Reportes de clasificación	42
5.4.4	Matrices de confusión	42
5.4.5	Comparativa entre modelos	42
5.5	Predicciones de modelos	44
5.5.1	Reportes de clasificación	44
5.5.2	Matrices de confusión	48
5.5.3	Comparativa entre modelos	49
6	Conclusiones	53
6.1	Trabajos para futuro	53

Índice de figuras

3.1	Inicialización de la población. Cada individuo está constituido por n parámetros del XGBoost a optimizar.	6
3.2	Selección de padres candidatos.	6
3.3	Mutación de hijos.	7
3.4	Proceso de optimización de XGBoost.	8
3.5	https://www.v7labs.com/blog/neural-networks-activation-functions . Arquitectura de una Red Neuronal.	9
3.6	https://towardsdatascience.com/cross-entropy-loss-function-f38c4ec8643e Cross-Entropy entre la función Softmax (izquierda) y el valor de la clase verdadera (derecha).	11
3.7	http://makeyourownneuralnetwork.blogspot.com/2020/02/calculating-output-size-of-convolutions.html	12
3.8	https://indoml.com/2018/03/07/student-notes-convolutional-neural-networks-cnn-introduction/	13
3.9	https://www.researchgate.net/journal/Transport-in-Porous-Media-1573-1634 . Ejemplo de ReLU aplicado a una convolución de dos dimensiones.	14
3.10	https://www.researchgate.net/publication/329201018/figure/fig2/AS:697400008138753@154328452 architecture-of-our-1D-CNN-model-This-model-consists-of-two-convolutional-layers.png	14
3.11	https://peltarion.com/knowledge-center/documentation/modeling-view/build-an-ai-model/blocks/1d-convolution Observamos la entrada de la red (izquierda) a la que se le aplica un kernel (en el centro) que resulta en la convolución de un mapa de características (derecha).	15
3.12	https://forum.huawei.com/enterprise/es/data/attachment/forum/202108/20/150033de3l4nyvmcbo KNN aplicado con distintos valores de k	16
4.1	Diagrama de Gantt de la planificación del proyecto.	19
4.2	Flujo de datos del proceso del proyecto.	21
4.3	Histograma de clases de accidente en el conjunto de datos original.	28
4.4	Matriz de correlación para los predictores escogidos del dataset.	29
4.5	https://towardsdatascience.com/train-test-split-and-cross-validation-in-python-80b61beca4b6 . División de un conjunto de datos en datos de entrenamiento y test.	30
4.6	Histograma de clases de accidente del conjunto de entrenamiento tras aplicar SMOTE-II.	31
4.7	TSNE de 2 y 3 componentes aplicado a los datos originales y a los generados sintéticamente (SMOTE-II).	32
4.8	Representación de las muestras de accidentes en forma de matriz de grises.	36
4.9	https://www.researchgate.net/journal/Transport-in-Porous-Media-1573-1634 . Función ReLU	37

4.10	Arquitectura de la CNN-2D mostrando una kernels aprendidos durante el entrenamiento.	38
5.1	Evolución de hiperparámetros a lo largo de las iteraciones.	39
5.2	Evolución del macro F1 score a lo largo de las iteraciones.	40
5.3	Pesos asignados por XGboost a las características.	41
5.4	Comparación entre los tiempos de entrenamiento de los modelos.	44
5.5	Comparación de TODO(tipo) F1 score sobre el conjunto de entrenamiento y validación CNN-1D.	45
5.6	Comparación de TODO(tipo) F1 score sobre el conjunto de entrenamiento y validación CNN-2D.	46
5.7	Matrices de confusión de los modelos sobre el conjunto de entrenamiento. . .	47
5.8	Comparativa de las métricas de las predicciones sobre el conjunto de entrenamiento de los modelos.	50
5.9	Matrices de confusión de los modelos sobre el conjunto de entrenamiento. . .	51
5.10	Comparativa de las métricas de las predicciones sobre el conjunto de test de los modelos.	52

Índice de tablas

4.1	Descripción de los datos.	23
4.2	Transformaciones aplicadas a los datos.	27
4.3	Límites de inicialización y mutación de los hiperparámetros de los individuos.	33
4.4	Transformaciones aplicadas a los datos.	35
5.1	Mejores parámetros de XGBoost tras aplicar el algoritmo genético.	40
5.2	Mejores parámetros de KNN tras aplicar GridSearch.	41
5.3	Cálculo de pesos de características y categorías mediante XGBoost.	42
5.5	Métricas CNN-1D.	45
5.6	Métricas Naive Bayes.	46
5.7	Métricas SVC.	48
5.8	Métricas SVC.	48
5.9	Métricas CNN-2D.	48

Índice de Códigos

1 Introducción

1.1 Motivación

Los accidentes de tráfico son uno de los principales problemas de Salud Pública en nuestros días. La multicausalidad, la variedad de las fuentes de datos y la poca cantidad de análisis específicos, apuntan a una gran complejidad en su tratamiento. Mas de 3.500 personas mueren, en las carreteras, diariamente en todo el mundo. Esto, significa casi 1,3 millones de muertes y aproximadamente 50 millones de lesiones cada año. Por lo tanto, la predicción de la gravedad de los accidentes de tráfico es un componente muy importante que se debe tener en cuenta, adoptando cualquier mejora en la predicción de la gravedad de los mismos. Además, el impacto económico y social asociado con los accidentes de tráfico lleva a las administraciones a buscar de manera activa mejoras. En el campo de la investigación sobre seguridad vial, el desarrollo de metodologías fiables para predecir y clasificar el nivel de gravedad, de los accidentes de tráfico, en función de diversas variables es un componente clave. La creciente disponibilidad de datos a gran escala de varias fuentes, incluidos los vehículos conectados y autónomos exige una comprensión profunda de las relaciones causales entre la seguridad y factores asociados con la ayuda de nuevas metodologías.

En las últimas dos décadas, los rápidos desarrollos en los métodos de aprendizaje automático (ML) y su regresión precisa y rendimiento de clasificación han atraído mucho la atención de los investigadores. Ha habido un número creciente de aplicaciones de métodos ML en la investigación de la gravedad de los accidentes. Aunque los métodos estadísticos tradicionales tienen formas funcionales rigurosas y claramente definidas, los métodos ML tienen una gran flexibilidad, requieren poca o ninguna suposición previa con respecto a los datos de gravedad de choques y son capaces de manejar valores perdidos, ruidos y valores atípicos.

A pesar de que la utilización de técnicas de inteligencia artificial (ML) en lugar de otras técnicas simples permite identificar las relaciones ocultas entre los factores de accidentes debido a la heterogeneidad de los datos de accidentes de tráfico, el principal inconveniente de estos modelos es que muy pocos estudios utilizan modelos de aprendizaje profundo para mejorar el rendimiento lo que se traduce en un bajo rendimiento en accidentes graves. Además, como los conjuntos de datos de accidentes de tráfico existentes, están muy desequilibrados, es difícil mejorar el rendimiento de las clases menos comunes como los accidentes graves. Basándose en lo expuesto anteriormente, en este trabajo se plantea estudiar y formalizar un modelo de aprendizaje profundo para predicción de la gravedad de los accidentes de tráfico. Con la finalidad de las características de los datos y preparar la entrada para el modelo de aprendizaje profundo, se ha utilizado una técnica para transformar números en matrices bidimensionales (Sharma et al., 2019) consistente en convertir un vector de características del accidente en una matriz de características, teniendo en cuenta que los vectores de características eran las variables de los datos del accidente y la matriz de características tenía las posiciones de las variables. Para fines de comparación, los resultados se han comparado con otros resultados

de modelos de aprendizaje automático.

1.2 Objetivos

1.2.1 Objetivos Generales

El objetivo principal de este trabajo final de Máster es desarrollar un sistema predictivo de la aavedad de accidentes de tráfico utilizando características que se pueden identificar en los lugares del accidente, como el sexo del conductor, el tipo de vehículo, el entorno o los atributos de la carretera.

1.2.2 Objetivos Específicos

Con la finalidad de realizar este objetivo principal, se plantean los siguientes objetivos secundarios:

1. Utilización de técnicas para la transformación de características cualitativas de los accidentes de tráfico en matrices numéricas.
 2. Utilización de algoritmos de aprendizaje automático basado en árboles de decisiones (XGBoost) para inferir pesos de las características de accidentes de tráfico.
 3. Utilizar técnicas de algoritmia evolutiva para optimización de parámetros de entrada del algoritmo XGBoost.
 4. Desarrollar un enfoque basado en el aprendizaje profundo, redes convolucionales, para predecir la gravedad de los accidentes de tráfico.
 5. Comparar los resultados de los modelos de aprendizaje profundo con otros modelos utilizando indicadores de rendimiento.
-

2 Marco Teórico

3 Tecnologías

Una vez definido el problema a tratar, los objetivos generales y específicos del proyecto, la solución propuesta y los requisitos a implementar hay que analizar las herramientas y tecnologías a utilizar. Para esto habrá que plantearse diferentes modelos y tecnologías disponibles para el desarrollo y optar por aquellos que aporten estabilidad y buen rendimiento.

3.1 Modelos

En primer lugar analizaremos los modelos utilizados durante la realización del proyecto.

3.1.1 Algoritmos Genéticos

Un *Algoritmo Genético* ?, es un modelo que imita la evolución de las especies en el ámbito de la biología, con el objetivo de encontrar una solución potencialmente óptima a un problema. El planteamiento de estos modelos consiste en evaluar los individuos pertenecientes a la población de la generación correspondiente para tomar el subconjunto de aquellas soluciones que mejor calidad ofrezcan mediante una función heurística denominada *función fitness* que aproxima soluciones ideales a un problema.

Las funciones heurísticas son funciones aproximadas a una solución ideal del problema, debido a que todas las posibles combinaciones de los parámetros a optimizar genera un espacio de búsqueda exponencial (problema NP-hard), es necesario crear una aproximación mediante la función heurística para acotar éste espacio.

Para comprender el funcionamiento de los algoritmos genéticos es necesario analizar cada una de las fases que lo constituyen:

1. Inicialización de la Población:

En primer lugar es necesario generar los n individuos de la población aleatoriamente. Cada uno de estos individuos está formado por el conjunto de parámetros que son objeto de optimización 3.1.

2. Evaluación de Población:

Es la primera fase del proceso iterativo del algoritmo genético y, en él, cada uno de los individuos pertenecientes a la población son evaluados mediante la función *fitness*.

3. Selección de Padres:

Una vez evaluados los individuos de la población, se procede a seleccionar aquellos que mejor valor han obtenido para crear nuevas soluciones a partir de éstos padres.

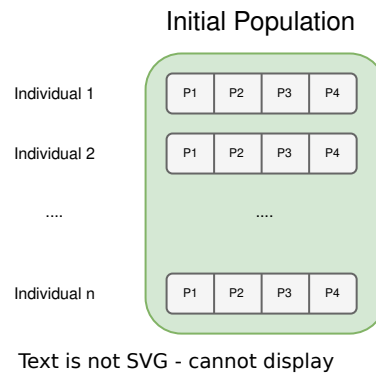


Figura 3.1: Inicialización de la población. Cada individuo está constituido por n parámetros del XGBoost a optimizar.

Existen distintas técnicas de selección, como por ejemplo escoger aquellos n mejores padres de la población o técnicas basadas en métodos probabilísticos, para que aquellos individuos que menos puntuación logren, tengan alguna probabilidad de ser elegidos para la generación de nuevas soluciones. Este tipo de técnicas se utilizan para aumentar la diversidad en la población y evitar caer en soluciones acotadas a mínimos locales del problema. Se puede ver el proceso de cómo son seleccionados los padres en función de su *fitness* en la figura 3.2.

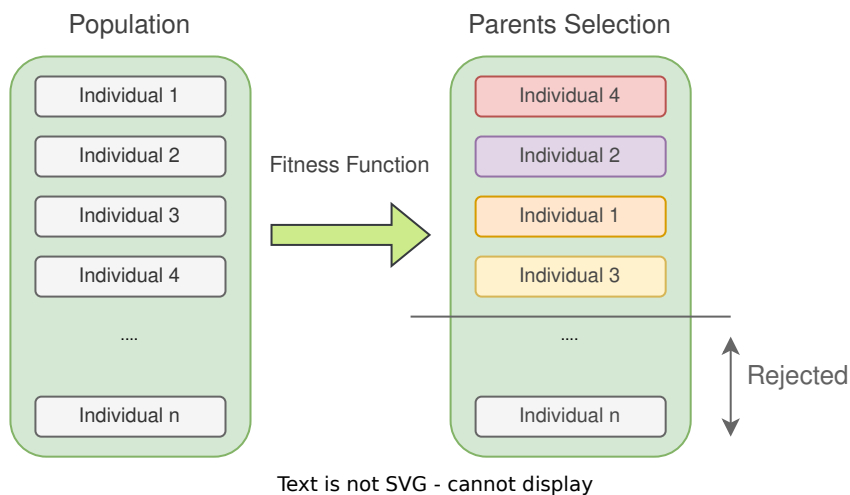
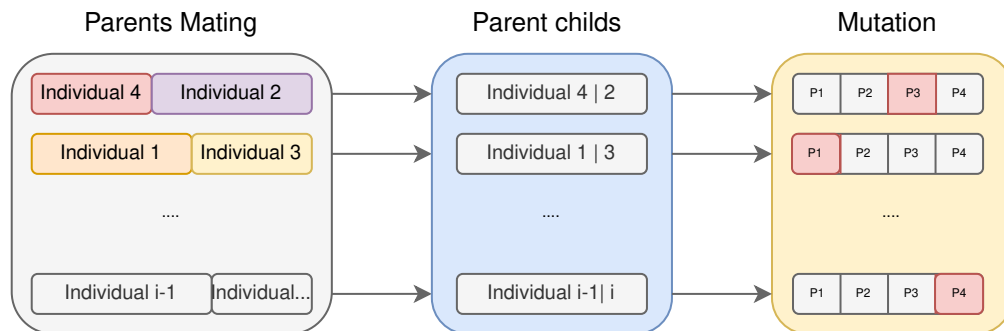


Figura 3.2: Selección de padres candidatos.

4. Cruce:

Una vez seleccionados los padres es necesario que intercambien información entre ellos para dar lugar a nuevos individuos, este proceso de herencia de información es posible realizarlo aplicando distintas filosofías; seleccionar un punto o posición de cruce a lo largo del vector de los padres y combinar ambos padres, o escoger aleatoriamente las posiciones de los parámetros de ambos padres y combinarlos para dar lugar a la solución

generada figura [3.3].



Text is not SVG - cannot display

Figura 3.3: Mutación de hijos.

5. Mutación:

Cuando los padres seleccionados son combinados para dar lugar a los candidatos de la nueva generación, es importante aplicar un proceso de mutación entre éstos debido a la necesidad de generar diversidad en la población. Si se combina la misma información entre generaciones se corre el riesgo de converger prematuramente a un mínimo local del espacio de búsqueda. Por este motivo se introduce el concepto de mutación, que consiste en aplicar una modificación aleatoria sobre alguno de los parámetros de los individuos generados para poder ampliar el espacio de soluciones.

Cada una de las fases enumeradas anteriormente se repite de forma iterativa entre generaciones hasta que se cumpla alguna condición de parada.

A continuación enumeraremos los parámetros con los que configurar un algoritmo genético. Dependiendo del tipo de algoritmo utilizado es posible encontrar distintos parámetros con los que inicializar el algoritmo, sin embargo estos son los más extendidos:

- Probabilidad de cruce:

Probabilidad con la que dos padres intercambian su información entre sí para dar lugar a un nuevo individuo hijo.

- Probabilidad de mutación:

Indica la probabilidad con la que cada hiperparámetro puede variar su valor.

3.1.2 XGBoost

Extreme Gradient Boosting Algoritihm (XGBoost) es una librería open-source que implementa algoritmos *ML Ensembles de tipo Boosting*. Se trata de uno de los algoritmos más importantes de los últimos años, logrando un rendimiento diez veces mayor a otras soluciones populares en una misma máquina y llegando a las primeras posiciones en numerosos retos propuestos en *Kaggle* ?.

XGBoost se basa en los modelos *Ensemble Boosting* de los algoritmos de *ML*, de ahí su nombre *Gradient Boosting*. La técnica *Boosting* se basa en construir un modelo robusto (*strong learner*) combinando una serie de modelos débiles (*weak learners*) ?.

Concretamente *XGBoost* implementa *Gradient Boosting Decision Trees (GBDT)*, que es un algoritmo similar a los *Random Forest*, y es utilizado tanto para clasificación como para regresión.

La diferencia principal entre los algoritmos *Random Forest* y *XGBoost* radica en el hecho de que los primeros utilizan algoritmos *Ensembles* tipo *Bagging* y construyen un modelo final predictivo mediante en el que la salida será la combinación de las predicciones de cada uno de los modelos individuales. En el caso de los modelos *Ensembles Boosting (XGBoost)* se combinan secuencialmente de forma que su objetivo es minimizar el error producido por los árboles generales utilizando para ello el método de descenso por gradiente.

En la figura [3.4] se puede observar cómo se van construyendo los árboles clasificadores y cómo cada árbol minimiza el residuo producido por el árbol anterior.

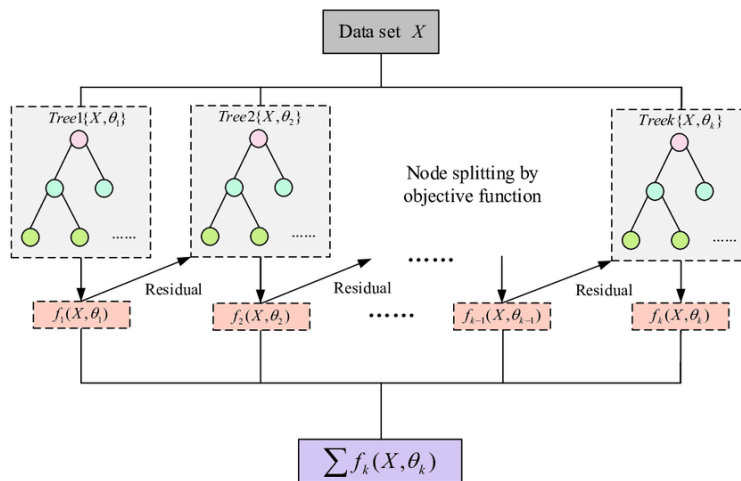


Figura 3.4: Proceso de optimización de XGBoost.

Además, *XGBoost* aplica técnicas de regularización en su función de pérdida, de tal forma que reduce la influencia individual de cada árbol generado y de sus hojas para poder dar lugar a posteriores árboles que consigan mejorar el modelo, evitando de esta manera el sobreajuste u *overfitting*.

3.1.3 Redes Neuronales

Las Redes Neuronales ? son modelos que emulan comportamiento del cerebro a la hora de procesar la información, entrenándose sobre un conjunto de datos de entrenamiento para identificar patrones entre ellos.

A modo de comprensión, se muestra la arquitectura de una *NN*, donde se pueden observar las conexiones entre cada capa de la red y sus salidas en la figura 3.5. Existe una capa de entrada inicial que corresponde a cada una de las características del ejemplo de entrada (en el caso de imágenes se trata de cada uno de los píxeles). Cada una de las características de la capa de entrada se conecta con cada una de las neuronas de la primera capa oculta *hidden layer* a través de conexiones denominadas pesos (*weights*) que son los que serán aprendidos en la red con el objetivo de encontrar patrones entre las características mediante el método de *Descenso por Gradiente* o *Gradient Descent (GD)*.

Esta metodología se aplica para cada una de las *hidden layers* presentes en la arquitectura hasta la última capa, llegando hasta la última función *SoftMax* que dictará la clase predicha en base a la salida de la capa final, u *output layer*.

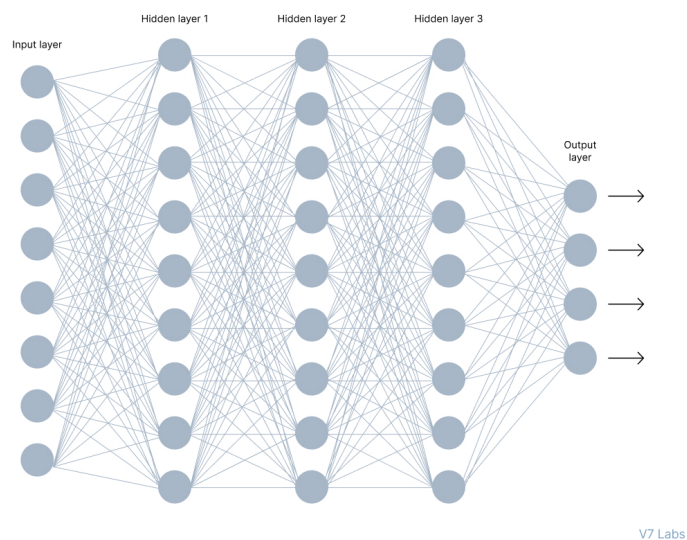


Figura 3.5: <https://www.v7labs.com/blog/neural-networks-activation-functions>. Arquitectura de una Red Neuronal.

Existen una serie de conceptos comunes a todas las *NN* que es conveniente nombrar para su comprensión:

- **Función de activación:**

Las funciones de activación son las encargadas de decidir si una neurona es activada o no en función de la entrada a esta función. Esta función suele situarse en la salida de cada capa de la red para decidir si una determinada neurona de la siguiente capa se activa o no.

Existen distintos enfoques en lo que respecta a las funciones de activación, como la función logística, tangente hiperbólica y *softmax* entre otras.

- **Entropía:**

La entropía sobre una variable aleatoria X es el grado de incertidumbre producido en base a los posibles valores que puede tomar como respuesta. Se calcula en base a la siguiente fórmula:

$$H(X) = - \sum_{i=1}^n p(x_i) \log_2 p(x_i)$$

Donde x_i es cada uno de los posibles valores que puede tomar la variable como respuesta y $p(x_i)$ es la probabilidad de obtener dicho valor.

- Softmax:

Se trata de la generalización en forma multidimensional de la función logística o *logistic function*. *Softmax* es una función que permite la transformación de un vector de números reales a una distribución de probabilidad. Esta función se suele aplicar como capa de activación en el último nivel en las *NN*, representando la probabilidad de que la salida de la red pertenezca a cada una de las K clases posibles en el modelo ?.

Esto es necesario debido a que normalmente los componentes de la salida de la última capa de una red neuronal (*logits*) son el resultado de la predicción en forma de números reales no normalizados, por lo tanto pueden ser mayores que 1 e incluso negativos. La función permite calcular la distribución de probabilidad en base a estos *logits* y calcular la probabilidad de pertenencia de la muestra a cada una de las K clases que componen el modelo mediante la siguiente fórmula:

$$\sigma(z_i) = \frac{e^{z_i}}{\sum_{j=1}^K e^{z_j}} \quad \text{for } i = 1, 2, \dots, K$$

Esta función es continua y diferenciable, por lo que es posible aplicar el método *GD* para actualizar los valores de los pesos de la red neuronal en las capas anteriores gracias a estas propiedades que permiten la derivabilidad de la función.

- Cross-Entropy:

El objetivo de cross-entropy es minimizar la función de pérdida (*log loss*) comparando la probabilidad de la clase predicha con respecto a su valor verdadero. Al aplicar una penalización logarítmica, las probabilidades de las clases que más disten respecto a su valor verdadero se verán más acentuadas ?.

$$L_{CE} = - \sum_{i=1}^n y_i \log_2(p_i)$$

Donde n es el número de clases a predecir, el valor y_i es la etiqueta de la clase verdadera, y p_i es la probabilidad de que la muestra actual pertenezca a la clase i .

- Back Propagation:

En el proceso de entrenamiento de la *NN* es necesario actualizar los pesos de las capas ocultas con el objetivo de minimizar la cross-entropy total del modelo. Para minimizar este valor se utiliza un algoritmo de optimización que maximiza el valor de una función *GD*.

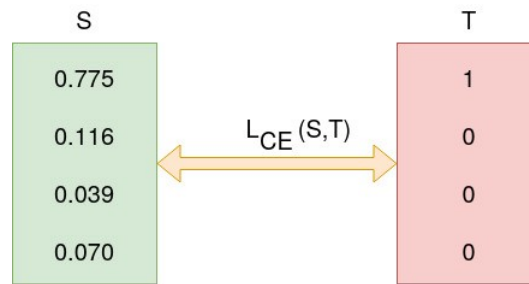


Figura 3.6: <https://towardsdatascience.com/cross-entropy-loss-function-f38c4ec8643e> Cross-Entropy entre la función Softmax (izquierda) y el valor de la clase verdadera (derecha).

Gracias a las propiedades de la función *Softmax* es posible utilizar descenso por gradiente para calcular la derivada de la función de pérdida respecto de cada uno de los pesos de las capas intermedias de la red y actualizar cada uno de ellos con el objetivo de minimizar la función de pérdida.

- Learning Rate:

La *tasa de aprendizaje*, o *learning rate* define el tamaño de paso que se aplicará a a la hora de optimizar los pesos de la red. El valor del *learning rate* se comprende en el intervalo $[0, 1]$, de tal forma que a menor valor de la tasa de aprendizaje menor será el cambio que se produce en los pesos de las neuronas del modelo, el caso contrario ocurre cuando adopta un valor alto.

Este parámetro no es sencillo de especificar, el *learning rate* es un regularizador que con valores altos permite que la red no caiga en *overfitting*, caso que ocurriría si el valor fuera demasiado pequeño y la minimización de la función de pérdida cayese en un mínimo local debido a aprender demasiado de los datos de entrenamiento. Si el valor es muy alto los pesos de los parámetros de la red variarán excesivamente impidiendo la convergencia en un mínimo de la función.

Esta disyuntiva ha sido ampliamente estudiada en los últimos tiempos, acerca de cuál es el valor ideal del *learning rate* en las *NN* ?, aunque comunmente se configura con el valor 0.1.

- Batch Normalization:

El proceso de entrenamiento de una red neuronal puede ser costoso, ya que existen una gran cantidad de parámetros para cada capa que son optimizados durante el proceso de entrenamiento. Esto hace que este proceso sea demasiado sensible a los parámetros inicializados en el inicio del entrenamiento de la red, como es el ejemplo de establecer una tasa de aprendizaje *learning rate* muy baja para poder converger. Esto hace que el entrenamiento sea muy lento, por lo que es necesario normalizar las entradas para cada mini-batch (subconjuntos de datos de entrenamiento con el que se entrena la red en cada época).

El método *Batch Normalization* ? normaliza los datos de tal forma que permite utilizar tasas de aprendizaje mayores además de actuar como regularizador, reduciendo así el número de capas *Dropout* que pudieran existir en la arquitectura.

3.1.3.1 Convolucionales

A contiución definiremos los parámetros específicos aplicados a las *CNNs*:

- Kernels:

Son las estructuras que se encargan de realizar la convolución. Los kernels son matrices de pesos unidimensionales ($1 \times n$) o bidimensionales ($n \times m$) dependiendo del tipo de convolución que se esté aplicando (Convolución 1-D o Convolución 2-D), y se encargan de inferir patrones en la matriz de entrada gracias a la actualización de sus pesos. En función de la profundidad de la capa, los kernels podrán encontrar patrones más simples o más complejos debido a que la entrada a cada capa es una convolución (aplicación del filtro) de las capas anteriores.

Un kernel realiza el producto escalar de sus pesos con respecto a las posiciones de la matriz de entrada (imagen o *feature map*) que colisionen en ese momento con el kernel, generando un nuevo *feature map* de menor dimensión, a menos que se utilice la técnica de *Padding* ?.

- Filtros:

Un filtro no es más que una agrupación de kernels, siempre tendrán una dimensión mayor que la definida para el kernel. Por ejemplo, si disponemos de un kernel unidimensional ($1 \times n$), el filtro contendrá k kernels de ($1 \times n$) en la capa convolucional definida ?

- Strides:

Son los saltos producidos por un kernel en cada etapa de la convolución, dependiendo si la convolución es dimensional o bidimensional, el stride constará de una dimensión (i), es decir, los pasos hacia la derecha en la imagen, o de dos dimensiones (i, j), los pasos hacia la derecha y hacia abajo en la imagen.

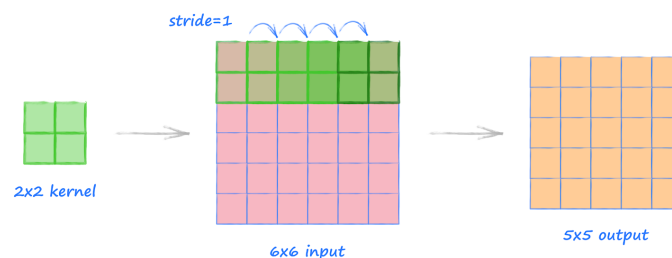


Figura 3.7: <http://makeyourownneuralnetwork.blogspot.com/2020/02/calculating-output-size-of-convolutions.html>

- **Padding:**

El padding es un método utilizado para incrementar la dimensionalidad de la entrada de la capa debido el desvanecimiento de los valores posicionados en zonas comprometedoras una vez aplicada la convolución.

El *padding* es el proceso de generar celdas artificiales inicializadas con valor 0 que permiten que el filtro convolucional sea aplicado en su totalidad en la posición de estas zonas comprometedoras, manteniendo la información de los límites de la imagen. De otra forma los valores de estos extremos se desvanecerían a medida se incrementa la profundidad de la red (número de capas) con respecto al resto de valores de la matriz.

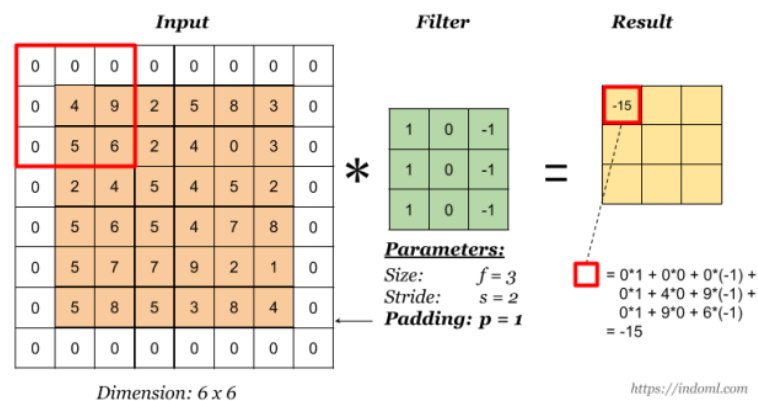


Figura 3.8: <https://indoml.com/2018/03/07/student-notes-convolutional-neural-networks-cnn-introduction/>

- **Función de activación:**

En el caso de las redes neuronales convolucionales, la función de activación se aplica a cada casilla proyectada por cada kernel sobre la matriz de entrada, de tal forma que cambia el valor a 0 sobre aquellos valores que resulten negativos de la convolución.

En la figura 3.9 se puede apreciar cómo se aplica la función de activación *ReLU* para cada casilla de la proyección de un *kernel* bidimensional sobre una *red convolucional de dos dimensiones*.

En el proceso de entrenamiento de las *CNNs* cada uno de los valores de los kernels son optimizados mediante el método *GD*, de tal forma que los pesos de cada capa son actualizados gracias a la derivabilidad de la función de pérdida.

3.1.3.1.1 Unidimensionales Una *Red Neuronal Convolucional de 1 Dimensión*, o *Convolutional Neural Network 1-Dimensional (CNN-1D)*, es un tipo de red convolucional que utiliza kernels (*filters*) que son aplicados a las imágenes de entrada con el objetivo encontrar ciertas

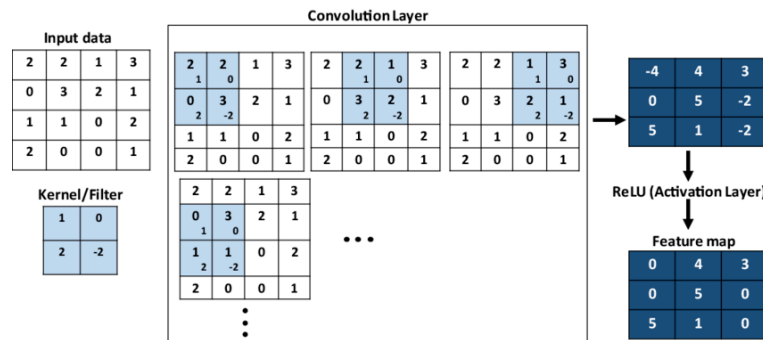


Figura 3.9: <https://www.researchgate.net/journal/Transport-in-Porous-Media-1573-1634>. Ejemplo de ReLU aplicado a una convolución de dos dimensiones.

características en ellas.

Las complejidad computacional de las *CNN-1D* es mucho más simple que las *CNN-2D*, además, estas redes se han demostrado ser más efectivas en distintos campos con respecto a las *CNN-2D*, como en técnicas de análisis de señales. La baja carga computacional de esta arquitectura la hace atractiva para aplicarla en tiempo real en dispositivos con pocos recursos como teléfonos móviles ?.

El resultado de este proceso es la proyección del filtro sobre un espacio dimensional denominado mapa de características (*feature maps*). Este filtro se utiliza para convolucionar los feature maps de la capa anterior ? o de la imagen de entrada si se trata de la primera capa de la red.

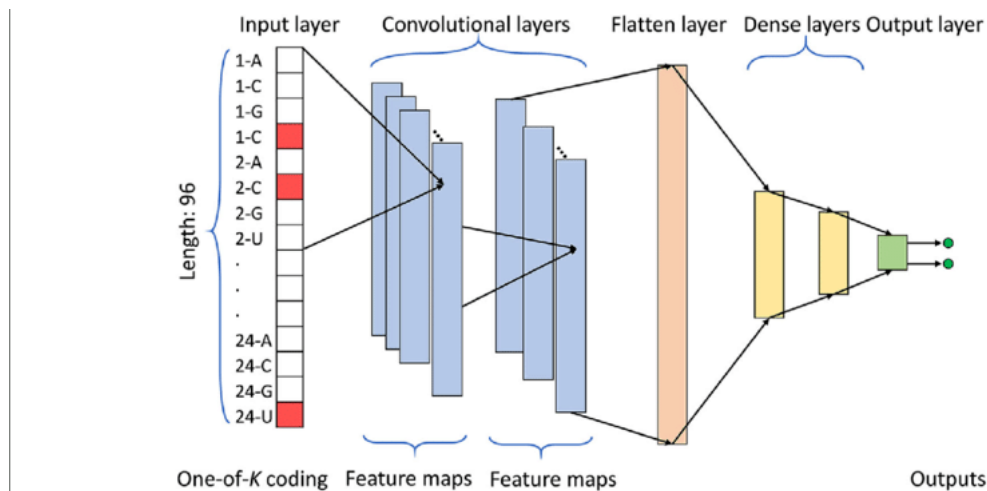


Figura 3.10: <https://www.researchgate.net/publication/329201018/figure/fig2/AS:697400008138753@1543284527161/1-architecture-of-our-1D-CNN-model-This-model-consists-of-two-convolutional-layers.png>

Al ser una arquitectura unidimensional, el *kernel* tendrá que tener un tamaño $(1 \times n)$.

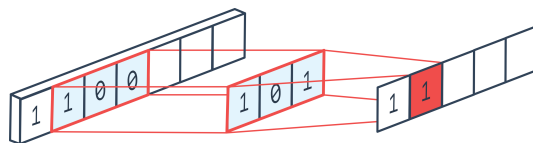


Figura 3.11: <https://peltarion.com/knowledge-center/documentation/modeling-view/build-an-ai-model/blocks/1d-convolution> Observamos la entrada de la red (izquierda) a la que se le aplica un kernel (en el centro) que resulta en la convolución de un mapa de características (derecha).

3.1.3.1.2 Bidimensionales Otra de las técnicas convolucionales implementadas en este proyecto son las *Redes Neuronales Convolucionales de 2 Dimensiones*, o *Convolutional Neural Networks 2-Dimensional (CNN-2D)*. Al igual que en el caso de las *CNN-1D*, el filtro se aplica sobre el input de la capa, resultando en *feature maps*, sin embargo, en este caso *kernel* será de bidimensional al ser una convolución de dos dimensiones, por lo que es necesario especificar el tamaño de la matriz que recorrerá la matriz pasada como input a la capa correspondiente.

3.1.4 KNN

El algoritmo *K vecinos más próximos*, o K-Nearest Neighbors (KNN) ?, es una técnica de aprendizaje supervisado ampliamente utilizada para tareas de clasificación y regresión, es uno de los algoritmos más básicos de *ML* y tiene una gran importancia en el campo de la *Ciencia de Datos* debido a su simplicidad de implementación y a su interpretabilidad.

El algoritmo *KNN* asume que muestras con características parecidas deben pertenecer a la misma clase. O lo que es lo mismo, muestras parecidas deben estar cercanas en el espacio dimensional.

Esta técnica se basa en representar las muestras de un conjunto de datos en base a sus características en un espacio n -dimensional y clasificar la muestra correspondiente como la clase mayoritaria al calcular la distancia de dicha muestra a los k vecinos más próximos. Normalmente la distancia empleada es la *Euclídea*, aunque es posible calcular otras distancias como *Manhattan* o *Chebyshev*.

A medida que se incrementan las dimensiones del espacio resulta más costoso computacionalmente calcular distancias entre los puntos.

Por lo tanto, el único parámetro a configurar es el número de vecinos más cercanos que se calculará en base al punto que queramos clasificar, en la figura 3.12 se muestra un ejemplo de un problema de clasificación aplicando distintos valores al parámetros k .

3.2 Especificaciones técnicas

En esta sección detallaremos las herramientas que han sido utilizadas para poder realizar este proyecto.

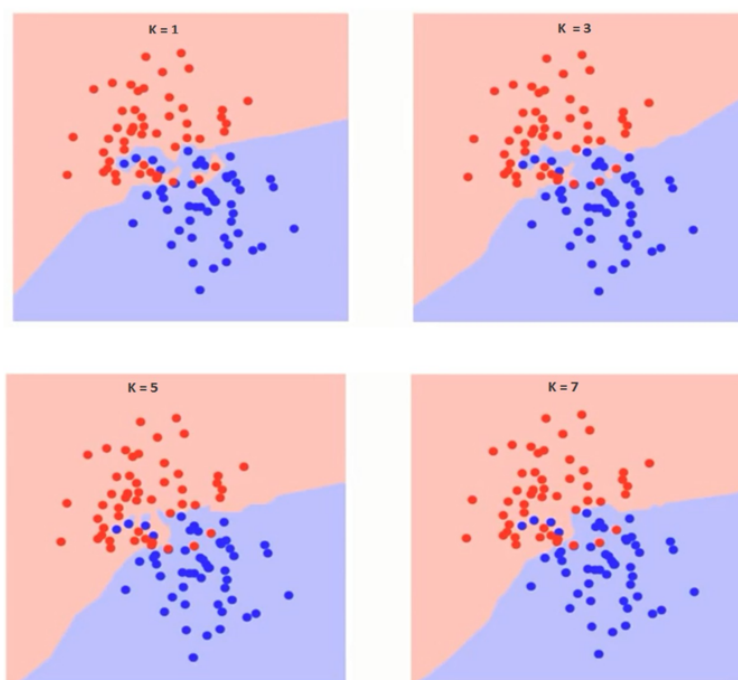


Figura 3.12: <https://forum.huawei.com/enterprise/es/data/attachment/forum/202108/20/150033de3l4nyvmcbozh14.p>
KNN aplicado con distintos valores de k

3.2.1 Herramientas utilizadas

En primer lugar se ha utilizado de forma intensiva *Python*, que es un lenguaje de programación interpretado de nivel alto multiparadigma. La versión empleada para el desarrollo del proyecto ha sido 3.9.11. ?

Las librerías más utilizadas han sido:

Pandas: librería que provee de herramientas que permiten el análisis y la manipulación de datos. La versión utilizada para este proyecto es la 1.3.5 ?.

Tensorflow: librería utilizada para la implementación de redes neuronales permitiendo su ejecución. Este proyecto se basa en la versión 2.8.0 ?.

Sklearn: librería que contiene múltiples modelos predictivos implementados, basado en NumPy, SciPy y Matplotlib. La versión configurada para este proyecto es 1.0.2 ?.

XGboost: paquete que contiene la implementación del algoritmo XGBoost, además de múltiples configuraciones como la ejecución en *CPU*, *GPU* y *GPU paralelizada*. La importación de esta librería ha sido en base a la versión 1.5.0 ?.

Respecto a las herramientas *Software* utilizadas destacan:

CUDA: plataforma de computación paralelizada que permite la ejecución de código en *GPU*, esto permite que las redes neuronales puedan entrenarse con mayor rapidez que en *CPU*

debido a la velocidad con la que se realizan las operaciones orientadas a datos en las tarjetas gráficas. Se ha utilizado la versión 11.6 ?.

Anaconda: distribución open-source que ofrece la flexibilidad de mantener varios entornos con distintas configuraciones y versiones de distintas librerías, facilitando además la migración entre equipos. La versión utilizada ha sido la 4.12.0 ?.

Jupyter Notebook: entorno interactivo que permite la creación, edición y ejecución de notebooks de forma local y remota. La versión utilizada para el desarrollo de este proyecto ha sido la 6.4.10 ?.

Jupyter Lab: interfaz de nueva generación que convive con el entorno Jupyter Notebook, ofrece numerosas funcionalidades como es la navegación entre distintos repositorios dentro de la interfaz. La versión instalada para la realización del proyecto ha sido la 3.3.2 ?.

DiagramsNet: plataforma utilizada para la confección de figuras mostradas en este documento ?.

Google Meets: plataforma utilizada para realizar reuniones semanales con el tutor ?.

Como sistema de control de versiones se ha utilizado Github ?, repositorio donde es posible subir versiones evolutivas del proyecto imprescindible en los proyectos de desarrollo ?.

3.2.2 Especificaciones del servidor

La mayor parte de los experimentos realizados en este trabajo se han ejecutado bajo un servidor con CPU *Dual AMD Rome 7742* (128 cores) y contando con una GPU *DGX NVIDIA A100* de 40 GigaBytes (*GB*).

4 Metodología

El desarrollo de este proyecto respeta tanto el proceso de desarrollo *Software* como el ciclo de vida de proyectos de *Ciencia de Datos*, por lo que se ha dividido en varias fases acontecidas que definiremos a continuación en base a la planificación mediante un *Diagrama de Gantt*:

4.1 Diagrama de Gantt

Un diagrama de Gantt es una herramienta que ayuda a visualizar las fases de evolución de un proyecto. Genera una vista global de cada una de las fases desarrolladas y su estimación temporal. A continuación se muestra el diagrama de Gantt realizado respecto a cada una de las fases de este proyecto 4.1.

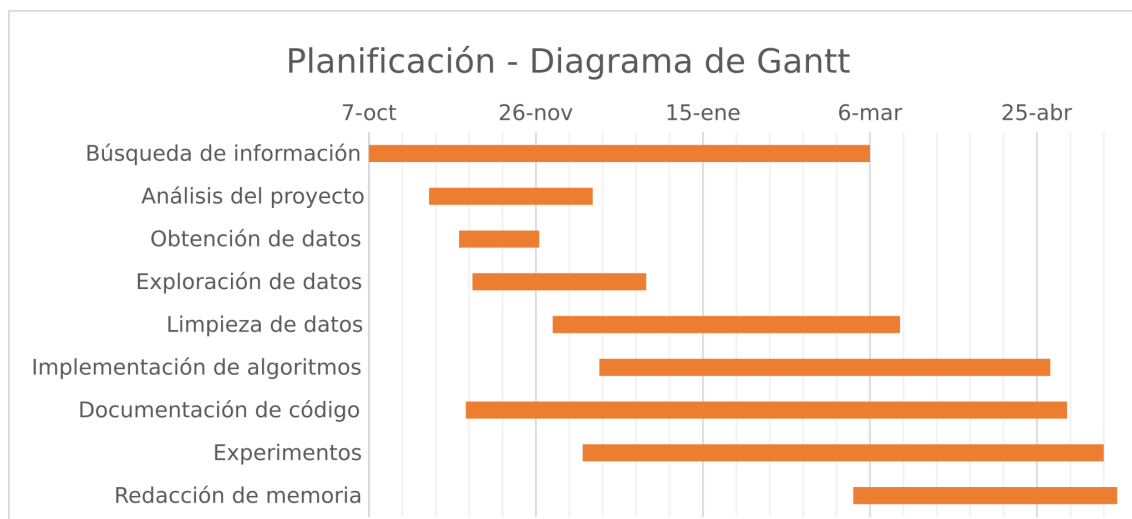


Figura 4.1: Diagrama de Gantt de la planificación del proyecto.

La búsqueda de información incluye la investigación de todo lo referente al proyecto, desde qué es una red neuronal hasta cómo diseñar sus arquitecturas. Por este motivo esta fase es la primera en el desarrollo y ha sido un proceso que se ha prolongado durante casi toda la evolución del trabajo.

El análisis del proyecto consiste en estudiar la viabilidad del mismo, buscando fuentes de información desde las que poder obtener los datos además de explorarlos. Es una de las fases más críticas en un proyecto *Data Science* ya que si las conclusiones del estudio de viabilidad no son sólidas, se podría cometer el riesgo de empezar un proyecto que no es factible para el cubrir los objetivos.

La obtención de datos consiste en la búsqueda de repositorios donde encontrar datos que

contengan la información requerida para abarcar los objetivos del proyecto, se ha realizado visitando distintos portales de datos abiertos nacionales hasta encontrar un conjunto de datos que encajase con los objetivos, en esta fase también se involucra la exploración inicial de los datos.

La exploración de datos abarca la fase de análisis de datos, esto se realiza utilizando métodos visuales para entender las características propias del dataset, además de hacer una valoración de la calidad de los mismos. Las primeras etapas de esta fase se complementan con el análisis del proyecto.

El siguiente paso ha sido la limpieza de datos, gracias a la exploración se comprenden la naturaleza de los datos y se decide qué técnicas usar para el tratamiento de valores atípicos del conjunto de datos para su posterior transformación.

Una de las tareas más importantes en el ciclo de vida de desarrollo Software que garantizan la calidad del producto es la documentación de código. Tiene como fin clarificar conceptos y explicar el funcionamiento de los métodos que componen el desarrollo. Es una fase que se prolonga desde el inicio de la codificación en la limpieza de datos hasta poco tiempo después de terminar de implementar los modelos.

La última fase de este proyecto a nivel funcional de este proyecto ha sido la realización de experimentos, comienza poco antes que la implementación de los modelos y consiste en testar las distintas implementaciones, tanto de limpieza de datos como el funcionamiento de los modelos.

La última etapa consiste en la redacción de este documento, plasmando toda la información obtenida durante la evolución del proyecto en esta memoria.

4.2 Diagrama de flujo

En primer lugar se define el flujo de ejecución del proyecto como se muestra en la figura 4.2, donde se han dividido las etapas del proyecto en seis bloques claramente diferenciados. Cada uno de estos está orientado a una disciplina distinta en proyectos de *Ciencia de Datos* y por lo tanto serán detallados en cada una de las subsecciones posteriores.

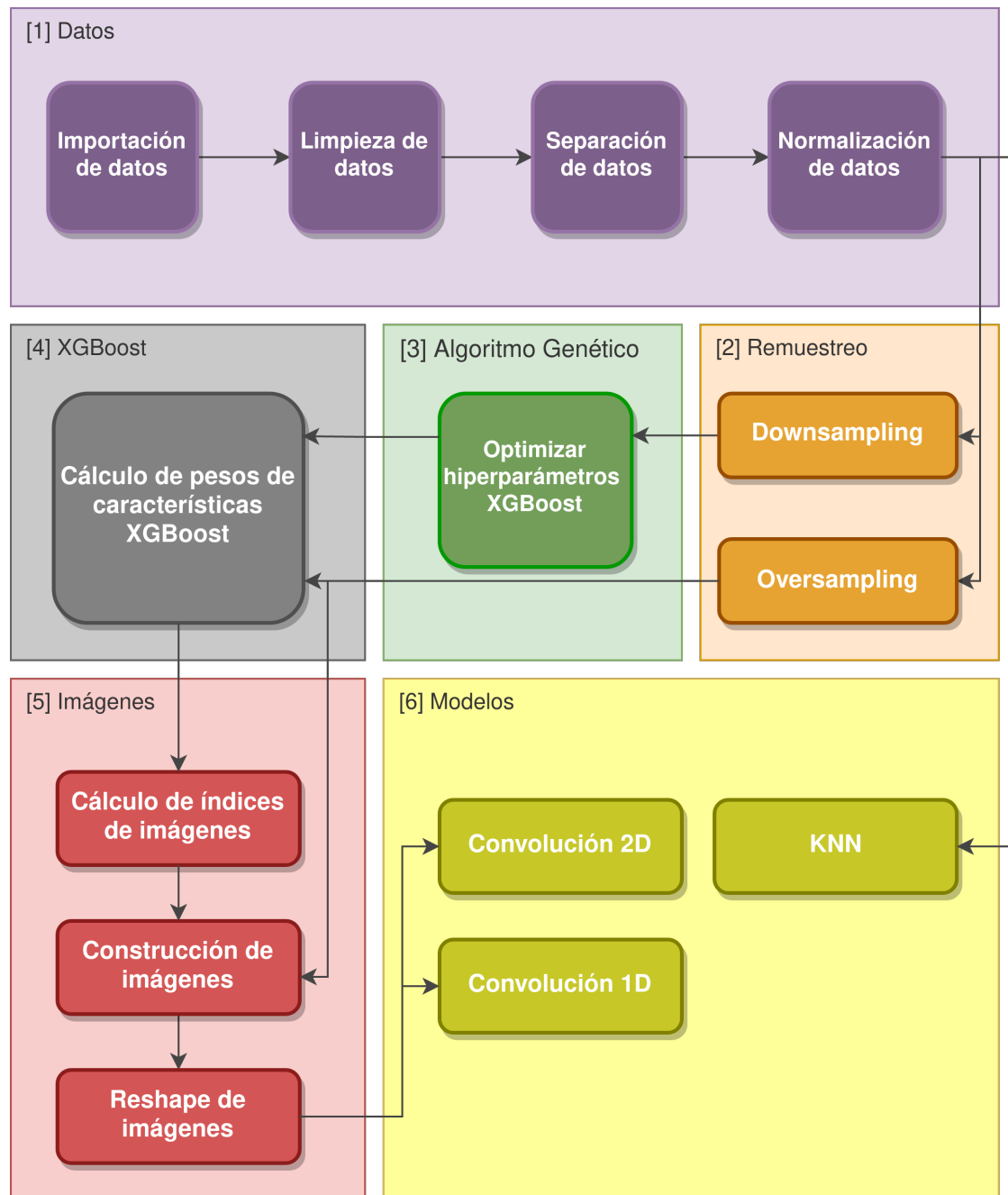


Figura 4.2: Flujo de datos del proceso del proyecto.

4.3 Proceso

En esta sección describiremos cada uno de los pasos que ha seguido el desarrollo del proyecto, desde la búsqueda inicial de datos hasta la implementación final de los algoritmos:

4.3.1 Datos

Como en cualquier proyecto de *Ciencia de Datos* el primer paso contempla la obtención, preparación y estudio de los datos. Este es el proceso en el que más tiempo se invierte en el ciclo de vida de un proyecto, algunos estudios apuntan que alrededor del 80 por ciento del tiempo en un proyecto se destina a esta tarea [?], por lo que es extremadamente importante aplicar buenas prácticas a la hora de llevar a cabo esta fase.

El conjunto de datos con el que se ha trabajado en este proyecto describe accidentes de tráfico de la ciudad de Madrid en un periodo concreto, desde el año 2019 hasta el 2022. Este dataset ha sido obtenido desde [?], repositorio donde es posible encontrar distintos datasets relativos a la Comunidad de Madrid.

El número total de registros en este periodo de tiempo es de 60.966 instancias, cada una de las cuales consta de 17 características que se describirán en la tabla 4.1:

Atributo	Descripción
Número de expediente	Identificador del incidente, si varios registros tienen el mismo un número de expediente se consideran como un mismo accidente y cada registro representa cada una de las distintas personas involucradas en él (Conductor, Pasajero o Peatón).
Fecha	Día mes y año en el que se ha producido el incidente.
Hora	Hora y minuto del día.
Localización	Nombre de la calle (si procede).
Número	Número de la calle donde ha ocurrido el incidente (si procede).
Distrito	Nombre del distrito de Madrid donde ha ocurrido el incidente.
Tipo de accidente	Tipología de accidente, puede ser de diversos tipos (Colisión doble, Colisión múltiple, Alcance, Choque contra obstáculo, Atropello, Vuelco, Caída, Otras causas).
Estado meteorológico	Condiciones climatológicas en el momento del incidente (Despejado, Nublado, Lluvia débil, Lluvia intensa, Granizado o Nevando)
Tipo de vehículo	Clasificación en función del tipo de vehículos, p.e. motocicleta, turismo, cuadriciclo, etc.
Tipo de persona	Rol de la persona involucrada (Conductor, Pasajero o Peatón)
Rango de edad	Intervalo de edad la persona implicada.
Sexo	Sexo de la persona implicada (Hombre o Mujer).
Lesividad	Consecuencias físicas de la persona implicada, si ha necesitado asistencia sanitaria, si ha sido ingresada o si ha sido mortal.
Coordenada X	Coordenada X del accidente en formato <i>UTM</i> .
Coordenada Y	Coordenada Y del accidente en formato <i>UTM</i> .
Positivo en alcohol	Si la persona implicada ha dado positivo en el control de alcoholemia (S o N).
Positivo en drogas	Si la persona implicada ha dado positivo en el control de estupefacientes (S o N).

Tabla 4.1: Descripción de los datos.

La variable respuesta de este proyecto es la lesividad, siguiendo la descripción del conjunto de datos existen una serie de valores asignados a ésta categoría. Cada uno de ellos pertenece al menos a uno de los siguientes casos:

- Consecuencias leves: comprende desde aquellas personas que no han resultado heridas hasta aquellas que han necesitado ingresar en un centro hospitalario no más de 24 horas, el tipado numérico de estas casuísticas es el siguiente:
 - 1: Atención en urgencias sin posterior ingreso.
 - 2: Ingreso inferior o igual a 24 horas.
 - 5: Asistencia sanitaria ambulatoria con posterioridad.
 - 7: Asistencia sanitaria sólo en el lugar del accidente.
 - 14: Sin asistencia sanitaria.
- Consecuencias severas: implicados que han requerido un ingreso hospitalario superior a 24 horas, el tipado numérico del conjunto de datos corresponde al siguiente caso:
 - 3: Ingreso superior a 24 horas.
- Consecuencias fatales: víctimas mortales dentro del margen de las 24 horas posteriores al accidente, la asignación numérica a este campo es:
 - 4: Fallecido 24 horas.

Se pueden consultar más detalles de las características del dataset descripción del conjunto de datos del portal *Open Data* de Madrid ?.

1. Limpieza de datos

Una vez importados los datos en el proyecto, es requisito hacer una limpieza de éstos (*Data Cleaning*), eligiendo las características que serán utilizadas como variables explicativas en las predicciones atendiendo además a los valores de las instancias, ya que pueden contener valores nulos, valores atípicos (*outliers*) o contener valores erróneos. Esta problemática requiere de la aplicación de distintas fases de análisis.

En el caso del dataset de accidentes de tráfico de este proyecto se han escogido las siguientes características como variables explicativas *hora*, *distrito*, *tipo accidente*, *estado meteorológico*, *tipo vehiculo*, *tipo persona*, *rango edad*, *sexo*, *positivo alcohol*, *positivo drogas*, *vehiculos implicados*, *coordenada x utm*, *coordenada y utm*.

Una vez que se han obtenido los predictores con los que trabajarán los modelos, se han eliminado aquellos registros duplicados y aquellos que tuvieran algún valor nulo en alguno de sus predictores. Por lo tanto, las dimensiones finales del conjunto de datos pasarán a ser 12 variables y 54.364 registros con respecto a las 17 características y 60966 filas originales.

2. Transformaciones de datos

En esta sección detallaremos las transformaciones sobre los datos que han sido necesarias para que los modelos trabajen con un conjunto de datos bien definido y consistente. Gran parte de los modelos de ML necesitan de datos de entrada numéricos para poder comprenderlos, además existe la necesidad de normalizar estos valores para que estén en la misma escala y el entrenamiento sea exitoso, por lo tanto se requiere de una serie de transformaciones que conviertan las variables categóricas a variables numéricas.

En primer lugar ha sido necesario transformar las columnas *coordenada x utm* y *coordenada y utm* a números enteros, ya que estas variables eran inicialmente de tipo *String*. Además, se da el caso de que el rango de estas variables es del orden de 7 a 10 dígitos, evitando cualquier formato de decimales estandarizado (utilizando puntos y comas indistintamente en distintas posiciones de los mismos), debido a esto ha sido necesario crear un proceso que analizase cada casuística y los tradujese a un formato estandarizado.

Las columnas *positiva alcohol* y *positiva drogas* se han unido en una nueva columna debido a la cantidad de valores nulos que existía en la segunda variable, por lo que se crea una nueva columna que hace referencia a la intoxicación etílica o de estupefacientes.

La variable *localización* ha sido transformada a *tipo de carretera*, con el objetivo de definir un criterio para clasificar el tipo de vía donde se ha producido el incidente, como por ejemplo una carretera, una autovía, glorieta, etc. Al no existir un servicio que ofrezca la tipología de la calle dado el nombre de la vía o las coordenadas UTM, ha sido necesario tratar este problema desde otro punto de vista mediante expresiones regulares.

El término *expresión regular* debe su origen a la teoría de las matemáticas y ciencias de computación, reflejan la propiedad de *regularidad* que caracterizan a las expresiones matemáticas. Una expresión regular es un tipo de patrón utilizado para encontrar una combinación de caracteres en una cadena de texto. Para clasificar el tipo de carreteras en este proyecto se han diseñado una serie de expresiones regulares que coinciden con ciertos criterios respecto a los valores de la variable *localización* con el fin de distinguir distintos tipos de vía.

NOTA: Ponemos una tabla con todas las RegEx que se han utilizado? Es sin duda la peor parte y menos científica del proyecto, y es lo único que no me gustaría mostrar, aunque ha llevado mucho tiempo me daría un poco de vergüenza poner esto en un proyecto de Ciencia de datos la verdad

Las tipologías de clases de carreteras son *Parking*, *Aeropuerto*, *Cuesta*, *Paseo*, *Parque*, *Túnel*, *Polígono*, *Camino*, *Glorieta*, *Puerta*, *Puente*, *Plaza*, *Bulevar*, *Travesía*, *Calzada*, *Carretera*, *Avenida*, *Autovía* y *Calle*

No obstante, la aplicación de las expresiones regulares generalmente no son una solución empírica a un problema, ya que hay casos que se pueden ser no contemplados al diseñar estos patrones manualmente, y en el caso de este proyecto su aplicación conlleva una serie de inconvenientes. Debido a las necesidades del problema están diseñadas jerárquicamente, de tal forma que aquellos valores de la variable *localización* que coincidan con el primer patrón definido de la expresión regular serán asignados a esta categoría. Esto supone un impedimento a tener en cuenta, ya que muchos de los accidentes se producen en intersecciones, carriles de aceleración que conectan dos vías, vías cuyo nombre contiene el tipo de otra vía, etc. Si se da el caso de que haya algún patrón que no se ha definido como expresión regular, provocará que estas localizaciones se asignen con respecto al primer patrón definido encontrado, llegando a clasificar erróneamente el tipo de vía. Como consecuencia, el orden de la definición de las expresiones es un punto crítico porque se busca maximizar la clasificación de aquellos tipos de vía con

menor número de apariciones, como es el caso de parking o aeropuerto, mientras se busca minimizar el error en la clasificación.

Para solucionar esto herramientas como Open Refine ? que utilizan técnicas de clustering nominal para detectar semejanzas entre nombres podría ser una solución.

Por otra parte hay que tratar aquellos valores que no hayan coincidido con los patrones definidos (TODO: poner el número), sin embargo, al ser una cantidad de registros se han podido revisar manualmente coincidiendo todas con nombres de calles al que no se les ha asignado la palabra "calle".

En la variable *hora*, al ser un campo continuo ha sido necesario discretizarla según intervalos, distinguiendo entre noche y día en función de la hora en la que se ha producido el incidente.

En la tabla 4.2 se define el resto de codificaciones aplicadas sobre el resto de variables del modelo para transformar estos campos a formato numérico.

Característica	Tipado
Lesividad	0: Accidentes leves (1, 2, 5, 6, 7, 14). 1: Accidentes severos (3). 2: Accidentes fatales (4).
Hora	1: Noche (6 PM - 6 AM). 2: Día (6 AM - 6 PM).
Distrito	Numeración en función de orden de aparición.
Tipo Accidente	1: Colisión fronto-lateral. 2: Alcance. 3: Colisión lateral. 4: Choque contra obstáculo fijo. 5: Colisión múltiple. 6: Caída. 7: Atropello a persona. 8: Colisión frontal. 9: Otro. 10: Solo salida de la vía. 11: Vuelco. 12: Atropello a animal. 13: Despeñamiento.
Estado Meteorológico	1: Despejado. 2: Nublado. 3: Lluvia débil. 4: Lluvia intensa.

	5: Granizando. 6: Nevando. 7: Se desconoce.
Tipo Vehículo	Numeración en función de orden de aparición.
Tipo Persona	1: Conductor. 2: Pasajero. 3: Peatón.
Rango Edad	1: Menores de 18 años. 2: De 18 a 25 años. 3: De 25 a 65 años. 4: Mayores de 65 años. 5: Edad desconocida.
Sexo	1: Hombre. 2: Mujer. 3: Desconocido.
Positivo	1: Sí. 2: No.

Tabla 4.2: Transformaciones aplicadas a los datos.

3. Análisis de datos

Una de las fases más importantes antes de comenzar el modelado en cualquier proyecto *Data Science* es el análisis de datos. Este proceso tiene como objetivo la descripción de los datos, identificación de *outliers*, valores erróneos y tendencias que se puedan dar en ellos.

Comenzaremos analizando los datos que correspondan a cada una de las tres posibles clases (leves, severos y fatales) realizando un histograma 4.3. Como podemos observar, nos encontramos ante un dataset claramente desbalanceado con respecto a la variable respuesta, contando con 53.009 accidentes leves, 1.271 severos y 84 fatales.

Un conjunto de datos desbalanceado se define como aquel que contiene un número de instancias mucho mayor de determinadas clases con respecto al resto ?. Esto se convierte en un problema para los modelos de clasificación ya que estos modelos tenderán a predecir las muestras como aquellas que pertenecen a las mayoritarias sobre el conjunto de test debido a que las reglas de clasificación de las clases menos numerosas tienden a ser ignoradas.

El desbalanceo de datos es un problema ampliamente estudiado a lo largo de los años y existen numerosos métodos orientados a solventarlo mediante distintas técnicas de muestreo ?, por lo que será necesario aplicar distintas técnicas sobre el conjunto de datos.

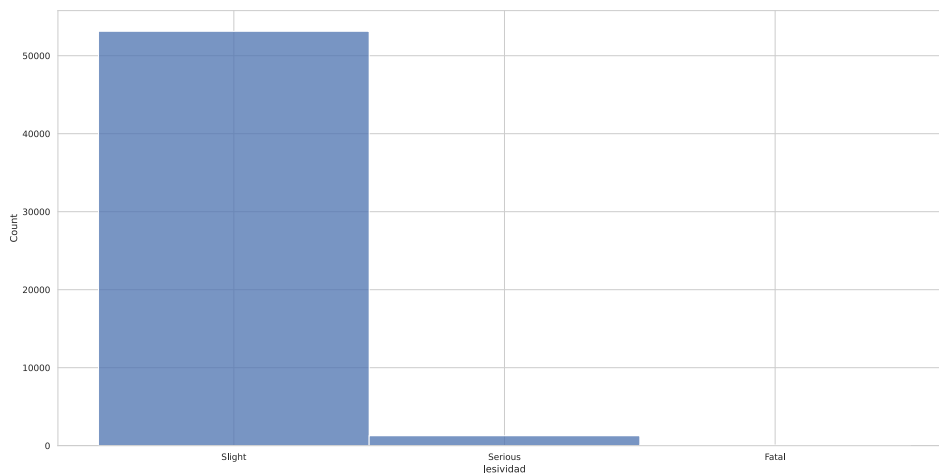


Figura 4.3: Histograma de clases de accidente en el conjunto de datos original.

En la figura 4.4 podemos observar la matriz de correlación de las variables explicativas del conjunto de datos del proyecto. Esta matriz muestra las relaciones existentes entre las variables indicando el grado de correlación entre ellas. Estos valores están acotados entre $[-1, 1]$ en base al coeficiente de correlación de *Pearson* ?, que indica la relación entre pares de campos, de tal forma es posible saber tanto el grado como la proporcionalidad de esta variación.

Analizando la matriz se puede comprobar que las únicas variables que están medianamente correlacionadas son *coordenada x utm* y *distrito*, con una correlación positiva de 0.44, sin embargo no es un valor tan alto como para eliminar alguna de las dos las características.

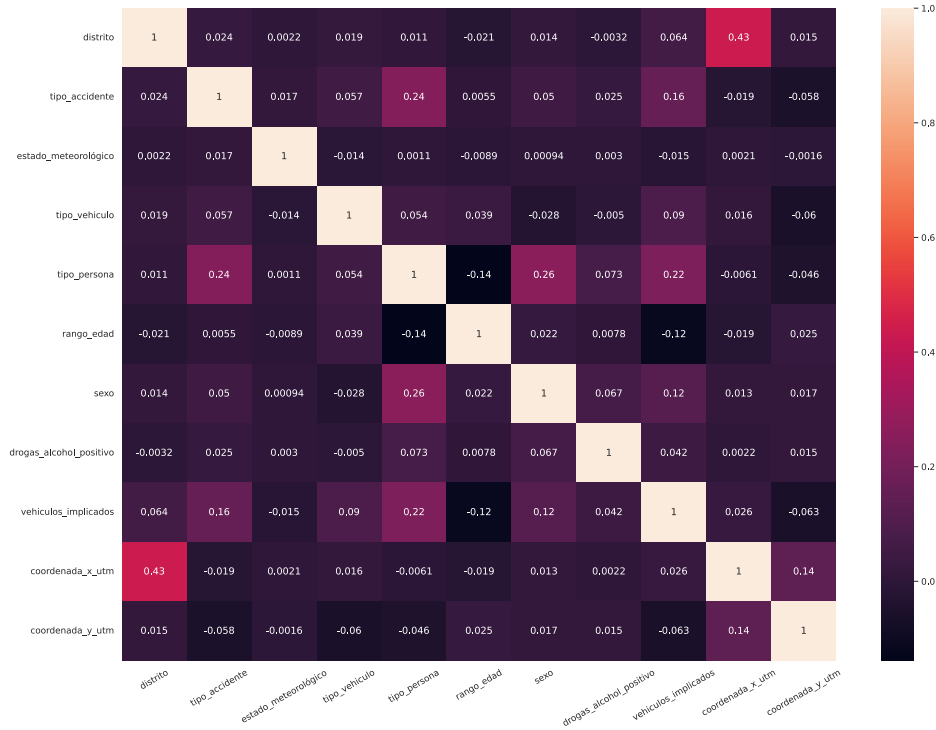


Figura 4.4: Matriz de correlación para los predictores escogidos del dataset.

4. Normalización de datos

La normalización de datos es un proceso necesario a la hora de obtener buenos resultados en los modelos predictivos de ML. Cuando un modelo es entrenado, es bastante probable existan características representadas en distintas escalas. En consecuencia, las características que contengan un rango de valores numérico más alto, ya sea por la naturaleza de la variable o porque se encuentren en otro rango, dominarán a aquellas características que contengan un rango menor, influenciando negativamente a los modelos ML, que normalmente son muy sensibles a la desnormalización de datos, como por ejemplo las CNN o KNN ?.

El proceso de normalización tiene como objetivo minimizar el *bias* de aquellas características cuya contribución sea mayor a la hora de encontrar patrones entre los datos. Existen distintas técnicas de normalización como por ejemplo Mean Centered (MC), Variable Stability Scaling (VSS) o Min-Max Normalization (MMN) entre otras ?.

En este proyecto se ha utilizado la normalización Z-Score Normalization (ZSN) debido a las propiedades que ofrece, ZSN que utiliza la media y la desviación típica para rees-

calar los datos de tal forma que la distribución de ellos esté definida por una media de cero y una desviación típica unitaria, consiguiendo representaciones de acuerdo a una distribución normal.

Los resultados obtenidos después de aplicar esta técnica se pueden interpretar como la distancia de cada valor con respecto a la media.

$$x^* = \frac{x-u}{\sigma}$$

Donde x^* es una muestra de una característica de los datos, u es la media total de dicha característica y σ es la desviación típica total de los valores de la característica.

5. Separación de datos

La siguiente fase en cualquier modelo de ML es la separación de datos (*split*). Esta fase consiste en dividir el conjunto total de registros en al menos dos subconjuntos, uno de entrenamiento y otro de test. El modelo se entrenará en base al conjunto de entrenamiento y se evaluará con el de test, los resultados sobre este conjunto permitirán comparar los modelos en base a las predicciones sobre las muestras que nunca han visto.

Comúnmente la proporción de datos de entrenamiento y test está establecida en 0.8 y 0.2 respectivamente, en la figura 4.5 se muestra una representación visual de esta división.

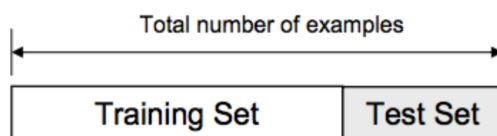


Figura 4.5: <https://towardsdatascience.com/train-test-split-and-cross-validation-in-python-80b61beca4b6>. División de un conjunto de datos en datos de entrenamiento y test.

4.3.2 Remuestreo

Son técnicas permiten operar sobre el conjunto de datos para balancearlo, con el objetivo de que el modelo no se vea muy afectado debido a la diferencia de las clases mayoritarias respecto a las minoritarias. Estas técnicas de forma general se aplican únicamente al conjunto de entrenamiento, ya que la intención es influir sobre el aprendizaje del modelo para aumentar el rendimiento de las predicciones sobre el conjunto de test, si aplicásemos remuestreo sobre este último conjunto estaríamos falseando los resultados de las predicciones.

En este proyecto se han hecho uso de las siguientes técnicas de remuestreo:

1. Downsampling: tiene como objetivo balancear el conjunto de datos para que todas las clases tengan el mismo número de muestras igualando el número de muestras de

las clases mayoritarias a aquellas clases que contienen menos muestras, descartando aquellas que sobrepasen este límite ?. **NOTA: Aún no se sabe si se va a aplicar Downsampling o no, sigo con los experimentos, de momento lo dejaré así a la espera de ver más resultados, queda pendiente para revisión**

2. Generación de datos sintéticos: técnica que permite generar datos artificiales en base a los límites que separan unas clases de otras. Se ha hecho uso de la técnica Borderline Synthetic Minority Over-sampling Technique 2 (SMOTE-II) ?. Esta técnica selecciona los vecinos más cercanos de la misma clase y genera nuevas muestras en base al espacio entre la clase minoritaria y sus vecinos más cercanos.

SMOTE-II ha sido utilizado para generar más muestras artificiales de los accidentes pertenecientes a las clases minoritarias (*severos* y *graves*). Este nuevo conjunto de datos servirá como entrenamiento para los modelos, evitando que tiendan al sobreajuste. En la figura 4.6 se aprecia el histograma tras haber aplicado SMOTE-II, contando ahora con 42508 muestras de cada una de las clases de accidentes.

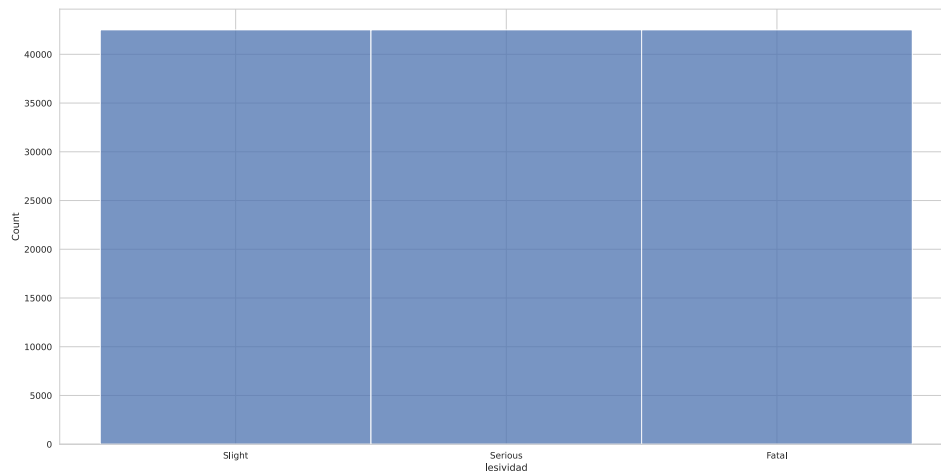


Figura 4.6: Histograma de clases de accidente del conjunto de entrenamiento tras aplicar SMOTE-II.

Una vez aplicadas las técnicas de remuestreo, es útil observar cuál es la representación espacial de los datos sintéticos generados por SMOTE-II en deferencia a los datos originales mediante el método t-Distributed Stochastic Neighbor Embedding (T-SNE) ?. T-SNE es un algoritmo que permite visualizar las proyecciones de datos multidimensionales en espacios bidimensionales o tridimensionales. Este modelo debe su origen a su antecesor Stochastic Neighbor Embedding (SNE), siendo una versión optimizada de éste.

Mientras que SNE proyecta los datos convirtiendo las distancias Euclídeas entre las muestras a probabilidades condicionales sean similares entre sí siguiendo probabilidades normales, el T-SNE modifica la función de coste para proyectarlos mediante una distribución t-Student,

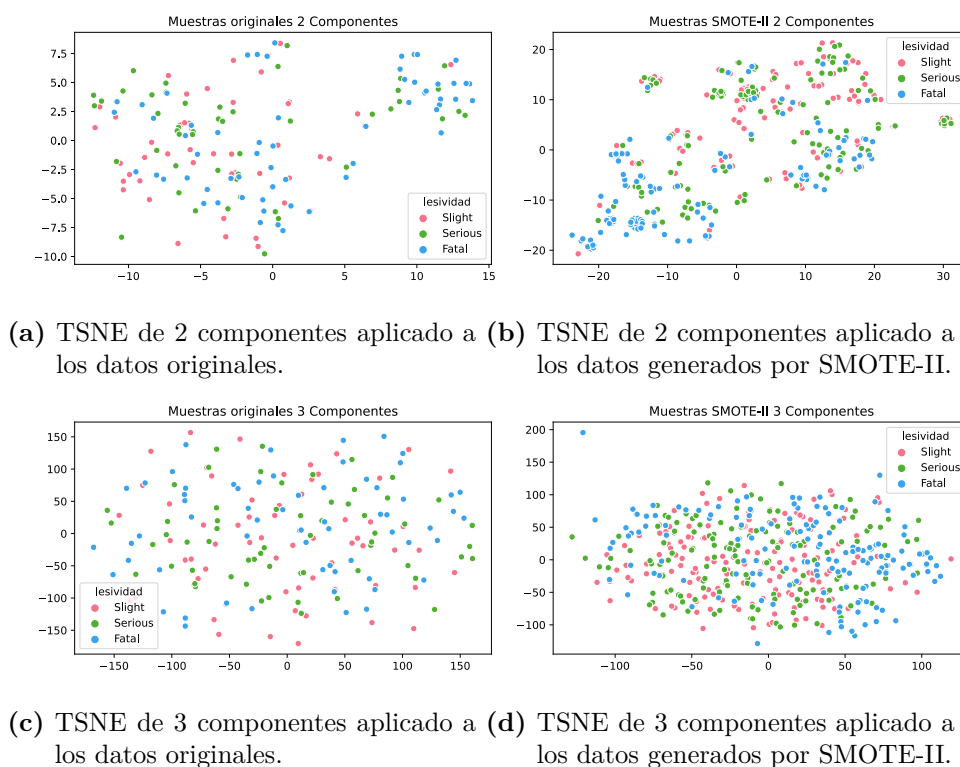


Figura 4.7: TSNE de 2 y 3 componentes aplicado a los datos originales y a los generados sintéticamente (SMOTE-II).

permitiendo además trabajar con gradientes simplificados.

En la figura 4.7 se muestran las proyecciones de T-SNE (tanto la representación bidimensional como la tridimensional) sobre el conjunto de datos de entrenamiento original y aquellos que han sido generados artificialmente por SMOTE-II.

El T-SNE de 2D aplicado sobre los datos generados por SMOTE-II 4.7b nos permite observar cómo se han generado las muestras con respecto a los datos originales 4.7a en una proyección bidimensional. Se aprecia cómo se han producido las nuevas muestras de los accidentes graves (azul) calculando la cercanía de los puntos en base a las fronteras de división. **Nota: Realmente yo no veo nada.**

4.3.3 Algoritmo Genético

Para un ideal entrenamiento de un modelo predictivo es conveniente optimizar los hiperparámetros con los que el modelo será entrenado, en este proyecto se aplicarán los algoritmos genéticos para optimizar hiperparámetros del algoritmo XGBoost.

Para ello se ha hecho uso de algoritmos genéticos [?], donde cada individuo perteneciente a la población de una generación está formado por una configuración de hiperparámetros específica, de tal forma que a lo largo de las iteraciones los individuos evolucionarán mediante el cruce y la mutación para dar lugar a nuevas configuraciones de hiperparámetros optimizadas

?

Debido al coste computacional que tendría optimizar todos los parámetros por la magnitud del espacio de búsqueda y al no ser necesario tener en cuenta todos ellos, se ha seleccionado un subconjunto de aquellos que más influencia tienen en el entrenamiento del modelo. Los hiperparámetros de los que consta cada individuo de la población son:

1. Profundidad máxima: es la máxima altura que puede tomar el árbol. Si el árbol de decisión alcanza demasiada profundidad tenderá al *overfitting* ya que aprenderá relaciones complejas entre los datos que pueden deberse a ruido en los datos de entrenamiento.
2. Peso mínimo de los hijos: es el mínimo peso que se establece a la hora de crear un nuevo nodo en el árbol. Cuando se entrena un árbol de decisión éste genera nuevos nodos en base a máxima separabilidad de los datos de entrenamiento en cada nivel. Con el límite de peso de los hijos establecemos un umbral mínimo de muestras que deben pertenecer a un nodo para realizar la separación. Un valor bajo en este parámetro permitirá crear nodos con menos muestras y por lo tanto el modelo tenderá al *overfitting*.
3. ETA: tamaño de paso utilizado para aplicar descenso por gradiente para minimizar la pérdida de los árboles anteriores.

La inicialización y mutación de los valores de los individuos viene dada por una limitación mínima y máxima que pueden, ya que, si no se contemplase, los hiperparámetros podrían tomar valores extremos, reduciendo así el rendimiento de entrenamiento y el de las predicciones. Por lo tanto los individuos variarán sus parámetros tomando un valor aleatorio dentro de estos rangos definidos. Los parámetros de las nuevas soluciones mutarán dentro de unos límites específicos para cada parámetro establecidos como máximos y mínimos para que no tomen valores extremos. La tabla ? muestra los límites de los hiperparámetros.

Hiperparámetro	Inicialización	Mutación
Profundidad Máxima	[1, 25]	[-6, 6]
Peso mínimo de los hijos	[0.01, 20.0]	[-7, 7]
ETA	[0.01, 1]	[-0.3, 0.3]

Tabla 4.3: Límites de inicialización y mutación de los hiperparámetros de los individuos.

Una vez inicializados aleatoriamente los TODO:X individuos especificados en la población, éstos se evaluarán instanciando TODO:X modelos XGBoost con los valores de cada individuo en la población. El conjunto de datos sobre el que entrenará cada instancia XGBoost será el conjunto de entrenamiento TODO:XX. La función *fitness* que evaluará cada individuo será la métrica TODO:F1-X aplicada sobre el conjunto de test obtenido en el proceso de separación de datos.

TODO: No se redactará más hasta tener claro cuál es la mejor forma de entrenar el XGBoost (función de validación y datos de entrenamiento)

4.3.4 XGBoost

TODO: Decir si se entrena con downsampled o smote-ii en función de los experimentos cuando acabemos

Una vez se han optimizados los hiperparámetros del XGBoost, se entrena un nuevo modelo con los `pq4m354ow` calculados sobre el conjunto de entrenamiento remuestreado. Con este nuevo modelo XGBoost entrenado se obtiene la importancia de cada clase mediante la ponderación de sus pesos. ?. Estos pesos son calculados por XGBoost, e indican el grado de importancia que ha tenido cada característica del conjunto de datos a la hora de entrenar el árbol, por lo tanto, aquellas características a las que se les asigne más peso habrán sido aquellas que han jugado un papel clave en las decisiones de los árboles.

Los pesos son asignados para cada característica del conjunto de datos con el que ha sido entrenado XGBoost de tal forma que permite realizar un análisis comparativo de los atributos y saber qué predictores son más importantes.

4.3.5 Imágenes

Las CNN aprenden patrones sobre la entrada de datos como imágenes, esto implica que éstas estén construidas de tal forma que se maximice la representación de la información, es decir, es necesario aplicar técnicas que posicionen cada característica en un pixel de la matriz maximizando la representación de los accidentes (TODO: **relamente no entiendo muy bien la motivación que ha seguido la gente de TASPCNN a la hora de construir las imágenes, falta entender por qué lo hacen así, no lo he encontrado en el paper.**)

Por lo tanto, una vez se tienen los datos normalizados y muestreados, el siguiente paso será transferir cada una de las muestras tabulares de los accidentes a una imagen en escala de grises, donde a cada característica se le asignará una posición de la matriz (5×5), de tal forma que éstas sean la entrada a las CNN.

Para este objetivo se aplicará el algoritmo FV2GI propuesto en el artículo ?, que asigna las características de una muestra en representación tabular en función de la importancia que éstas presenten en una jerarquía.

Tal y como se propone en ? las características que provocan un accidente de tráfico pueden englobarse en una serie de elementos o categorías principales, concretamente: *características del conductor, el estado de la carretera, características propias del vehículo y condiciones del ambiente*. Por este motivo se realizará una asignación de cada característica del dataset en una de estas categorías, se muestra en la tabla 4.4 las asignaciones de las variables explicativas del conjunto de datos en función de esta jerarquía.

Categoría	Variable
Accidente	Coordenada X. Coordenada Y. Hora. Vehículos implicados. Tipo de accidente.
Ambiente	Estado meteorológico.
Carretera	Tipo Carretera. Distrito.
Conductor	Tipo Persona. Sexo. Rango Edad. Positivo.

Tabla 4.4: Transformaciones aplicadas a los datos.

Una vez definida la jerarquía de características y los pesos asociados a cada una de ellas gracias al algoritmo XGBoost, se construyen las imágenes de acuerdo al criterio *FV2GI*. Este proceso consta de los siguientes pasos:

1. Generación de n imágenes inicializadas a 0, donde n es el número de muestras tabulares en el dataset.
2. Asignación de una fila a cada padre en función de su peso.
3. Asignación de la la posición de cada característica hija en función de su peso dentro de la fila de su padre.

Se deben tomar las siguientes consideraciones a la hora de aplicar *FV2GI*:

1. La importancia de un padre viene dada por la suma de los pesos de las características hijas, en la tabla 5.3 se observa el cálculo de cada característica para las categorías principales.
2. La asignación de las filas de los padres se realiza de forma intercalada en función de su peso. Aquel padre que más importancia tenga irá posicionado en la fila central de la matriz, el segundo padre irá posicionado por encima de éste y el tercero por debajo y así sucesivamente, de tal forma que se irá creando una estructura en la que dichos padres se interpolan entre las filas en función de su peso.

- Una vez se han asignado las categorías padre a las filas se realiza el mismo procedimiento con las características hijas a nivel de columna. Estas características se asignarán en una posición dentro de la fila de su categoría padre, donde aquella que más importancia tenga irá posicionada en el centro, la segunda irá posicionada a su izquierda, la tercera a la derecha y así sucesivamente.

Nota: No estoy contento con la explicación, creo que estaría bien meter una fórmula matemática o pseudocódigo y cambiar la redacción

Como resultado de este proceso se obtienen las imágenes con un único canal gris 4.8, que serán la entrada de las CNN en forma de representaciones matriciales.

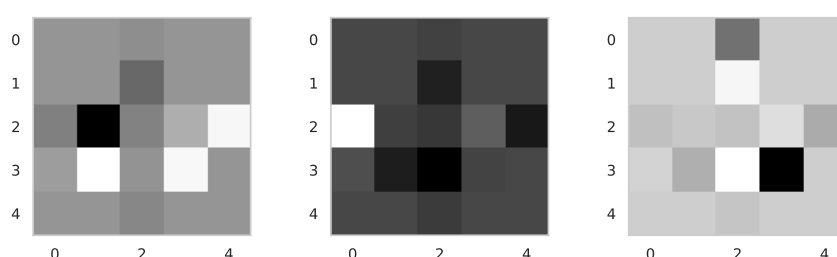


Figura 4.8: Representación de las muestras de accidentes en forma de matriz de grises.

4.3.6 Modelos

En esta sección analizaremos los modelos utilizados en este proyecto con el objetivo de realizar un estudio comparativo entre las arquitecturas.

1. KNN

El método KNN servirá como referencia para testar el rendimiento del resto de modelos. Al ser un método que se aplica sobre un conjunto de datos tabular original, no hará uso de las imágenes generadas.

Es necesario optimizar los parámetros de este algoritmo para conseguir un buen rendimiento, por lo que aplicaremos la técnica *Grid Search* ?. Esta técnica permite la optimización de los valores de los hiperparámetros mediante una búsqueda exhaustiva en un espacio de búsqueda definido por el usuario, probando distintas combinaciones hasta cubrirlo por completo.

KNN construirá el espacio de proyecciones mediante los datos de entrenamiento remuestreados por XGBoost, y clasificará las muestras del conjunto de test en base a la cercanía de los vecinos proyectados del conjunto de entrenamiento.

2. CNN

Una vez definidos los pasos que sigue el entrenamiento de una *NN* y las características propias de las *CNNs*, podemos analizar las arquitecturas de las *CNNs* implementadas que se han aplicado en este proyecto. Cabe mencionar que el funcionamiento de ambas *CNNs* únicamente difiere en el tamaño del *kernel* (unidimensionales o bidimensionales

según sea el caso correspondiente) y la forma en la que éste se desplaza debido a su dimensionalidad, por lo tanto detallaremos ambas arquitecturas de la misma forma.

La arquitectura consta de cuatro capas convolucionales con tamaños de kernel (1×3) en caso de las CNN-1D y de tamaño (3×3) para las CNN-2D. Estos kernels se proyectarán en 256 canales para formar el filtro convolucional asociado a cada capa. A la salida de cada uno de los mapas de características se aplica un proceso de BN.

El *padding* del kernel se ha establecido en 1 para ambos tipos de redes, de tal forma que las convoluciones se aplicarán añadiendo ceros en los límites de las imágenes, y los *strides* en (1) para las CNN-1D y $(1, 1)$ para las CNN-2D, por lo tanto el desplazamiento de los kernels se hará píxel a píxel tanto en las CNN-1D como en las CNN-2D.

A la salida de cada capa convolucional se aplica la función de activación Rectified Linear Unit (ReLU), que se comporta devolviendo el valor 0 para aquellas entradas que sean negativas y el valor original para aquellas que sean positivas, se puede apreciar el comportamiento en la figura 4.9. El rendimiento de esta función de activación en el campo de las imágenes está ampliamente extendido, además, evita la probabilidad de aparición del gradiente evanescente ?.

$$f(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ x & \text{if } x > 0 \end{cases}$$

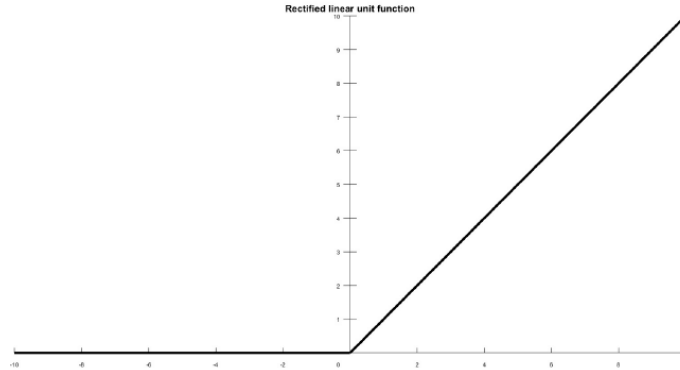


Figura 4.9: <https://www.researchgate.net/journal/Transport-in-Porous-Media-1573-1634>. Función ReLU

La salida de la última capa de la convolución transformará el mapa de características de tamaño (5×5) generada a una capa Flatten, que aplanará la matriz de tal forma que la convertirá en un vector unidimensional de (1×25) . A continuación se aplicará una capa densa que conectará cada uno de los 25 nodos de la capa Flatten con los 128 nodos de la capa densa, que generará los logits antes de aplicar la última función de activación *Softmax* que devolverá la clase predicha.

Para ejemplificar la arquitectura de las CNN propuestas se muestra el caso de la CNN-2D en la figura 4.10

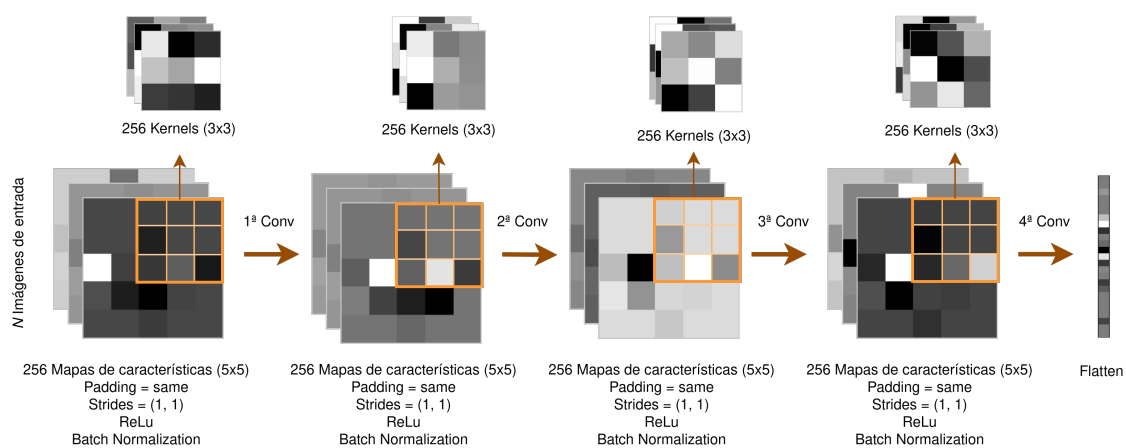


Figura 4.10: Arquitectura de la CNN-2D mostrando una kernels aprendidos durante el entrenamiento.

?

5 Resultados

En esta sección se detallarán los resultados de los experimentos y se realizará un análisis comparativo entre ellos.

En primer lugar....

5.1 Algoritmo Genético

En la figura [5.1] se pueden visualizar la evolución de los parámetros del mejor individuo en cada generación.

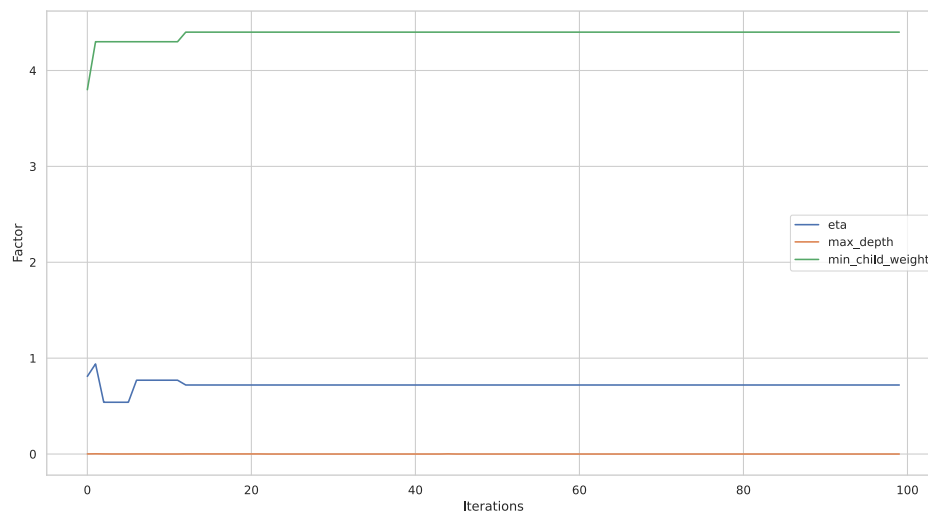


Figura 5.1: Evolución de hiperparámetros a lo largo de las iteraciones.

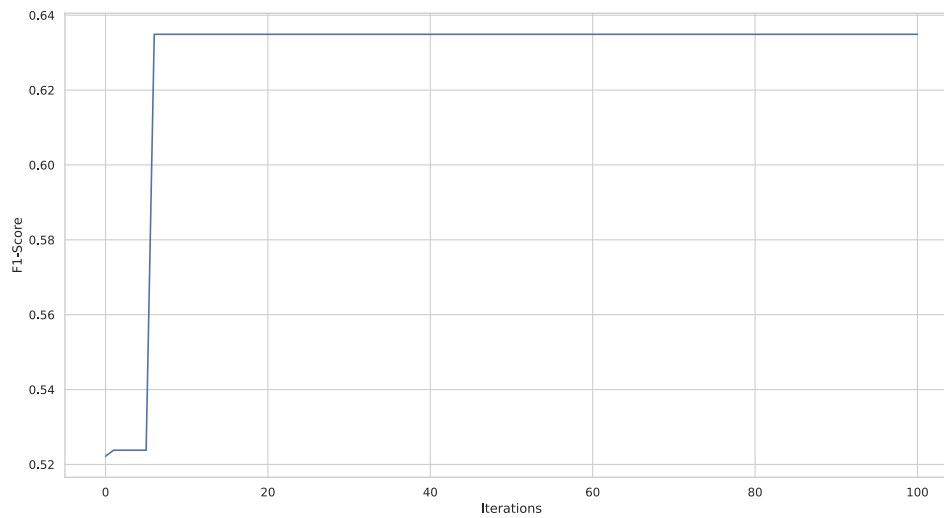


Figura 5.2: Evolución del macro F1 score a lo largo de las iteraciones.

En la tabla 5.1 se pueden observar el mejor individuo obtenido después de *TODO*: *X* generaciones.

Hiperparámetro	Valor
Profundidad Máxima	TODO
Peso mínimo de los hijos	TODO
ETA	TODO

Tabla 5.1: Mejores parámetros de XGBoost tras aplicar el algoritmo genético.

5.2 XGBoost

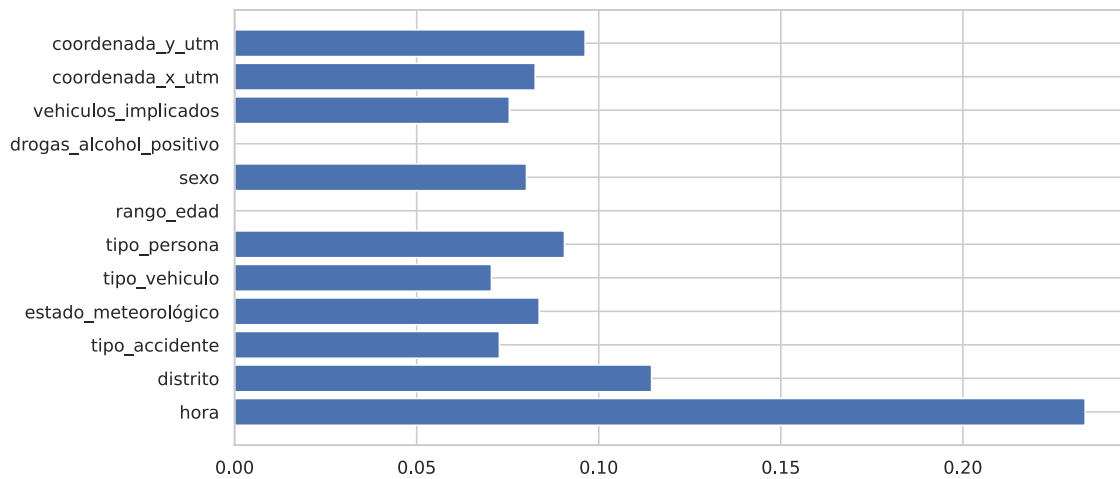


Figura 5.3: Pesos asignados por XGboost a las características.

5.3 KNN

Se puede observar en la tabla 5.2 los mejores parámetros para *KNN* después de haber ejecutado *Grid Search* sobre el conjunto de entrenamiento.

Atributo	Valor
Número de vecinos	91
Tipo de distancia	Minkowski

Tabla 5.2: Mejores parámetros de KNN tras aplicar GridSearch.

Categoría	Peso Categoría	Característica	Peso Característica
Accidente	TODO: 1	Coordenada X	TODO: 1
		Coordenada Y	TODO: 1
		Hora	TODO: 1
		Vehículos implicados	TODO: 1
Ambiente	TODO: 1	Estado meteorológico	TODO: 1
Carretera	TODO: 1	Tipo de accidente	TODO: 1
		Distrito	TODO: 1
Conductor	TODO: 1	Tipo Persona	TODO: 1
		Sexo	TODO: 1
		Rango Edad	TODO: 1
		Positivo	TODO: 1

Tabla 5.3: Cálculo de pesos de características y categorías mediante XGBoost.

El resultado de aplicar ambas técnicas da lugar a la representación en forma matricial de los accidentes, como se puede observar en la figura 4.8.

5.4 Entrenamiento de modelos

5.4.1 Tiempos de entrenamiento

El KNN no necesita entrenamiento, el tiempo obtenido es el de la optimización de hiperparámetros mediante Grid Search, por lo que no incluiría el tiempo en la gráfica.

5.4.2 Gráficas de entrenamiento

Explicar las gráficas de evolución F1-score de las redes neuronales.

5.4.3 Reportes de clasificación

Explicar los reportes de clasificación con datos de **ENTRENAMIENTO**.

5.4.4 Matrices de confusión

Explicar las matrices de confusión con datos de **ENTRENAMIENTO**.

5.4.5 Comparativa entre modelos

Explicar la comparativa de modelos con datos de **ENTRENAMIENTO**.

	precision	recall	f1-score	support
Slight	0.984	0.184	0.31	42509.0
Serious	0.526	0.919	0.669	42509.0
Fatal	0.867	0.923	0.894	42508.0
accuracy	0.675	0.675	0.675	0.675
macro avg	0.792	0.675	0.624	127526.0
weighted avg	0.792	0.675	0.624	127526.0

(a) Métricas de clasificación del conjunto de entrenamiento CNN-1D.

	precision	recall	f1-score	support
Slight	0.733	0.458	0.564	42509.0
Serious	0.675	0.452	0.541	42509.0
Fatal	0.565	0.964	0.712	42508.0
accuracy	0.625	0.625	0.625	0.625
macro avg	0.658	0.625	0.606	127526.0
weighted avg	0.658	0.625	0.606	127526.0

(b) Métricas clasificación del conjunto de entrenamiento Naive Bayes.

	precision	recall	f1-score	support
Slight	0.834	0.792	0.812	42509.0
Serious	0.833	0.768	0.799	42509.0
Fatal	0.852	0.961	0.903	42508.0
accuracy	0.84	0.84	0.84	0.84
macro avg	0.839	0.84	0.838	127526.0
weighted avg	0.839	0.84	0.838	127526.0

(c) Métricas clasificación del conjunto de entrenamiento SVC.

	precision	recall	f1-score	support
Slight	0.52	0.86	0.648	42509.0
Serious	0.839	0.596	0.697	42509.0
Fatal	0.875	0.557	0.681	42508.0
accuracy	0.671	0.671	0.671	0.671
macro avg	0.745	0.671	0.675	127526.0
weighted avg	0.745	0.671	0.675	127526.0

(d) Métricas clasificación del conjunto de entrenamiento KNN.

	precision	recall	f1-score	support
Slight	0.989	0.088	0.161	10640.0
Serious	0.024	0.899	0.046	247.0
Fatal	0.002	0.071	0.003	14.0
accuracy	0.106	0.106	0.106	0.106
macro avg	0.338	0.353	0.07	10901.0
weighted avg	0.966	0.106	0.158	10901.0

(e) Métricas clasificación del conjunto de entrenamiento CNN-2D.

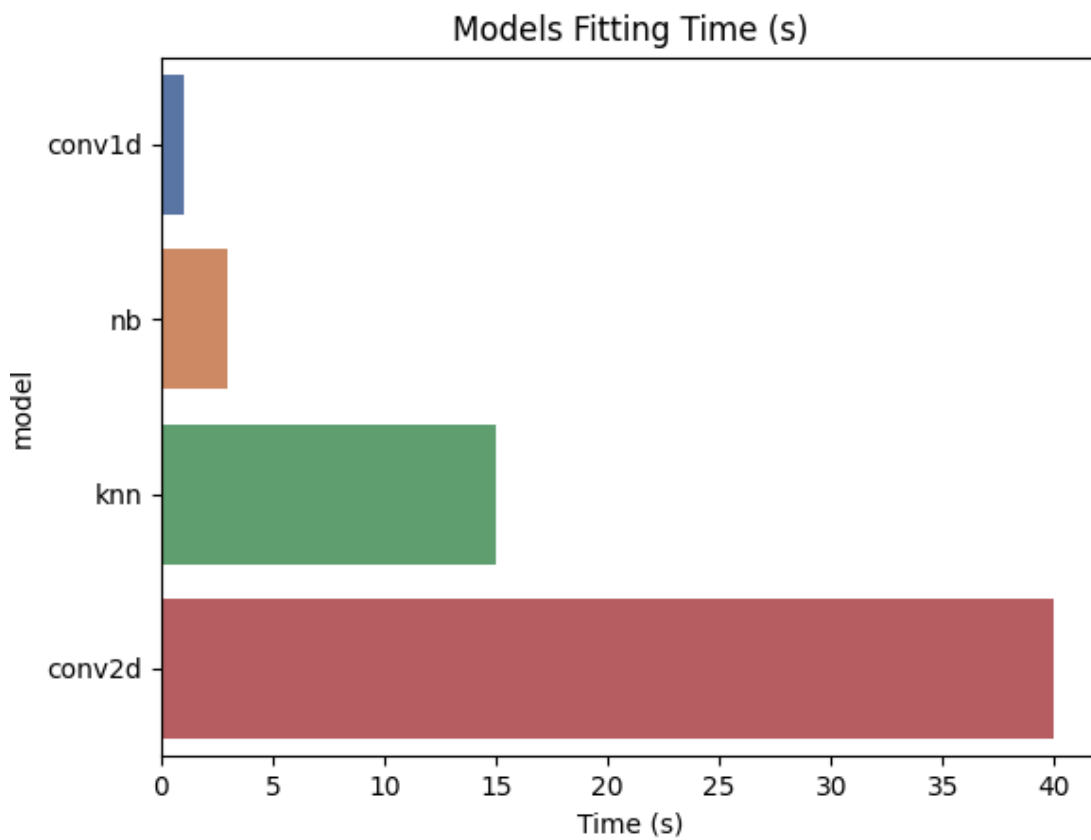


Figura 5.4: Comparación entre los tiempos de entrenamiento de los modelos.

5.5 Predicciones de modelos

5.5.1 Reportes de clasificación

Explicar los reportes de clasificación con datos de **TEST**.

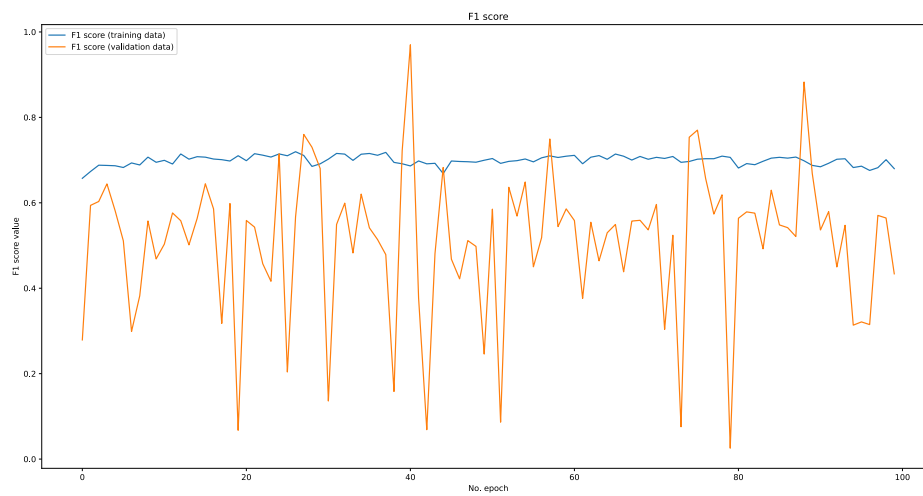


Figura 5.5: Comparación de TODO(tipo) F1 score sobre el conjunto de entrenamiento y validación CNN-1D.

	precision	recall	f1-score	support
Slight	0.984	0.184	0.31	42509.0
Serious	0.526	0.919	0.669	42509.0
Fatal	0.867	0.923	0.894	42508.0
accuracy	0.675	0.675	0.675	0.675
macro avg	0.792	0.675	0.624	127526.0
weighted avg	0.792	0.675	0.624	127526.0

Tabla 5.5: Métricas CNN-1D.

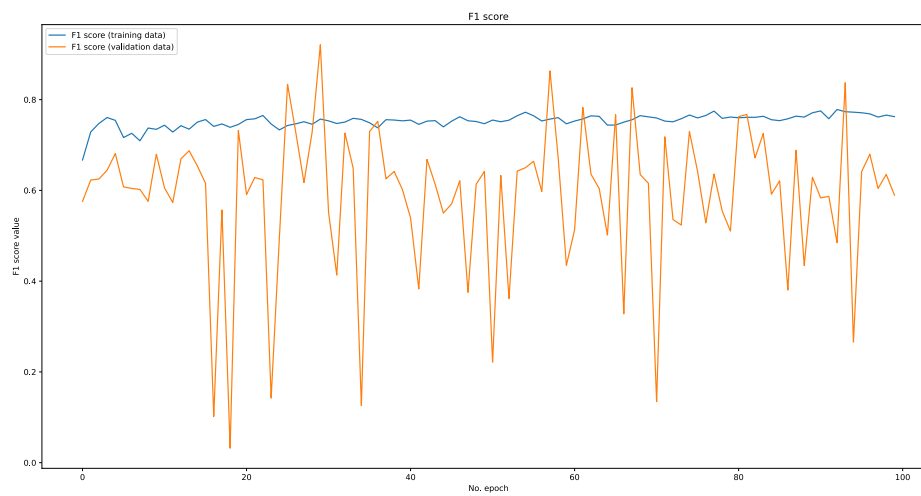
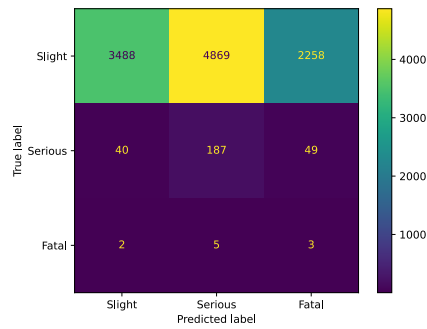


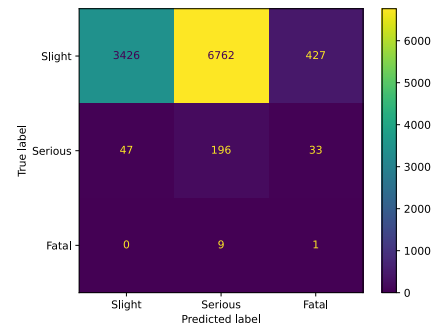
Figura 5.6: Comparación de TODO(tipo) F1 score sobre el conjunto de entrenamiento y validación CNN-2D.

	precision	recall	f1-score	support
Slight	0.98	0.448	0.615	10640.0
Serious	0.026	0.636	0.05	247.0
Fatal	0.0	0.0	0.0	14.0
accuracy	0.452	0.452	0.452	0.452
macro avg	0.335	0.361	0.222	10901.0
weighted avg	0.957	0.452	0.601	10901.0

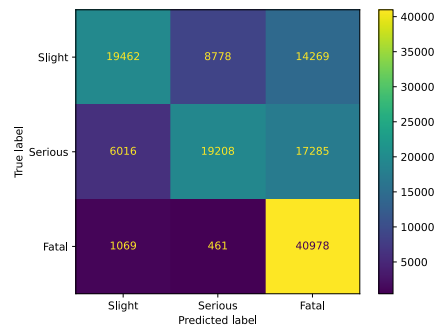
Tabla 5.6: Métricas Naive Bayes.



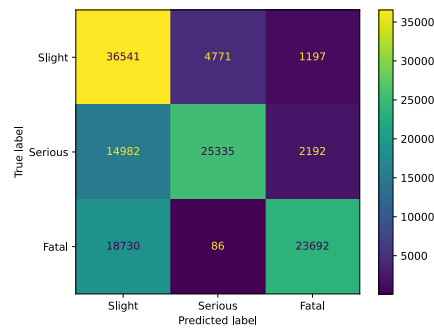
(a) Matriz de confusión sobre el conjunto de datos de entrenamiento CNN-1D.



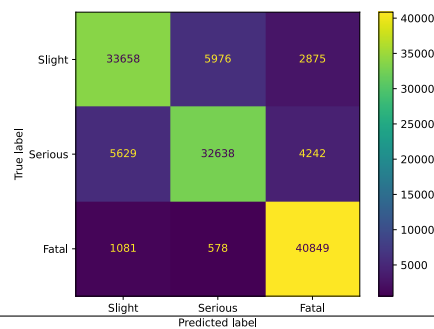
(b) Matriz de confusión sobre el conjunto de datos de entrenamiento CNN-2D.



(c) Matriz de confusión sobre el conjunto de datos de entrenamiento NB.



(d) Matriz de confusión sobre el conjunto de datos de entrenamiento KNN.



(e) Matriz de confusión sobre el conjunto de datos de entrenamiento SVC.

Figura 5.7: Matrices de confusión de los modelos sobre el conjunto de entrenamiento.

	precision	recall	f1-score	support
Slight	0.981	0.678	0.802	10640.0
Serious	0.035	0.393	0.065	247.0
Fatal	0.0	0.0	0.0	14.0
accuracy	0.67	0.67	0.67	0.67
macro avg	0.339	0.357	0.289	10901.0
weighted avg	0.959	0.67	0.784	10901.0

Tabla 5.7: Métricas SVC.

	precision	recall	f1-score	support
Slight	0.981	0.771	0.864	10640.0
Serious	0.041	0.348	0.073	247.0
Fatal	0.002	0.071	0.004	14.0
accuracy	0.761	0.761	0.761	0.761
macro avg	0.342	0.397	0.314	10901.0
weighted avg	0.959	0.761	0.845	10901.0

Tabla 5.8: Métricas SVC.

	precision	recall	f1-score	support
Slight	0.989	0.088	0.161	10640.0
Serious	0.024	0.899	0.046	247.0
Fatal	0.002	0.071	0.003	14.0
accuracy	0.106	0.106	0.106	0.106
macro avg	0.338	0.353	0.07	10901.0
weighted avg	0.966	0.106	0.158	10901.0

Tabla 5.9: Métricas CNN-2D.

5.5.2 Matrices de confusión

Explicar las matrices de confusión con datos de **TEST**.

5.5.3 Comparativa entre modelos

Explicar la comparativa de modelos con datos de **ENTRENAMIENTO**.

Como podemos observar, los modelos entrenados predicen bien sus muestras de entrenamiento sin embargo ninguno de los modelos aplicados es capaz de generalizar bien sobre el conjunto de datos de test sin smote esto es debido a la naturaleza de los datos, de la información que tenemos y de el tipo de limpieza que se ha hecho. Se podría hacer un estudio del valor numérico que se le asigna a cada variable en la limpieza, ya que seguramente al estar nosotros haciendolo aleatoriamente seguramente se esté creando algún patrón de jerarquía dentro de una columna o se le da mas importancia a un todoterreno que a un avion porque el todoterreno al tener un valor asignado más alto sera mas negro y por lo tanto la red predicira lo que sea

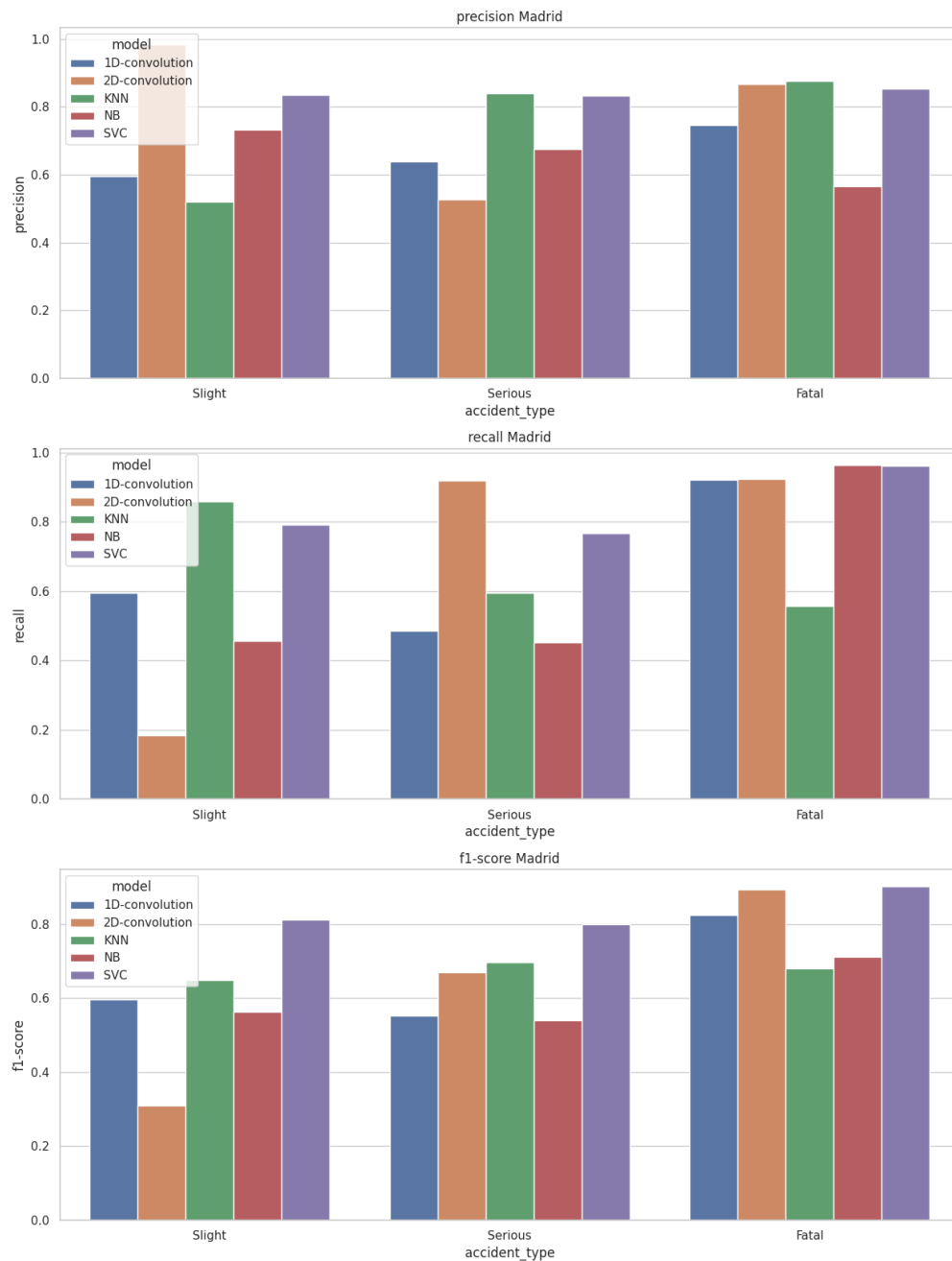
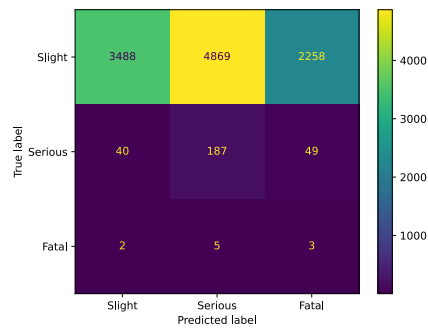
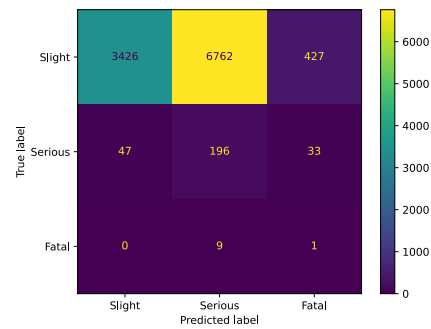


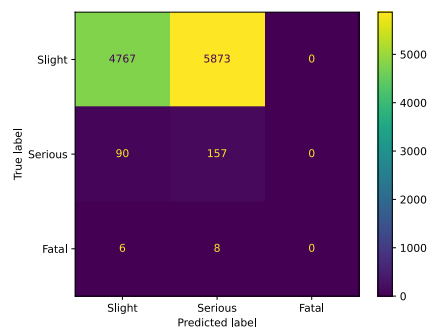
Figura 5.8: Comparativa de las métricas de las predicciones sobre el conjunto de entrenamiento de los modelos.



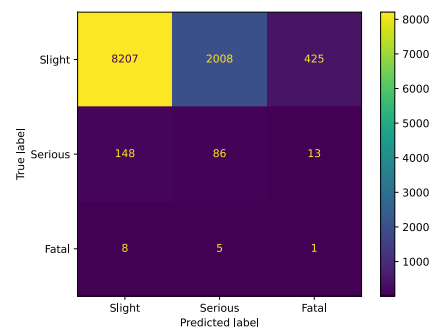
(a) Matriz de confusión CNN-1D.



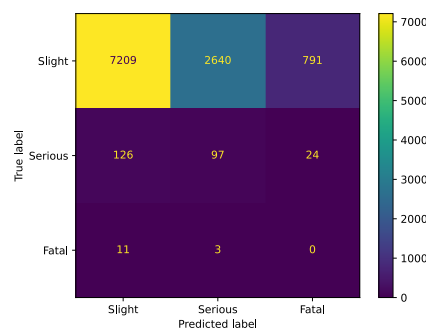
(b) Matriz de confusión CNN-2D.



(c) Matriz de confusión NB.



(d) Matriz de confusión KNN.



(e) Matriz de confusión SVC.

Figura 5.9: Matrices de confusión de los modelos sobre el conjunto de entrenamiento.

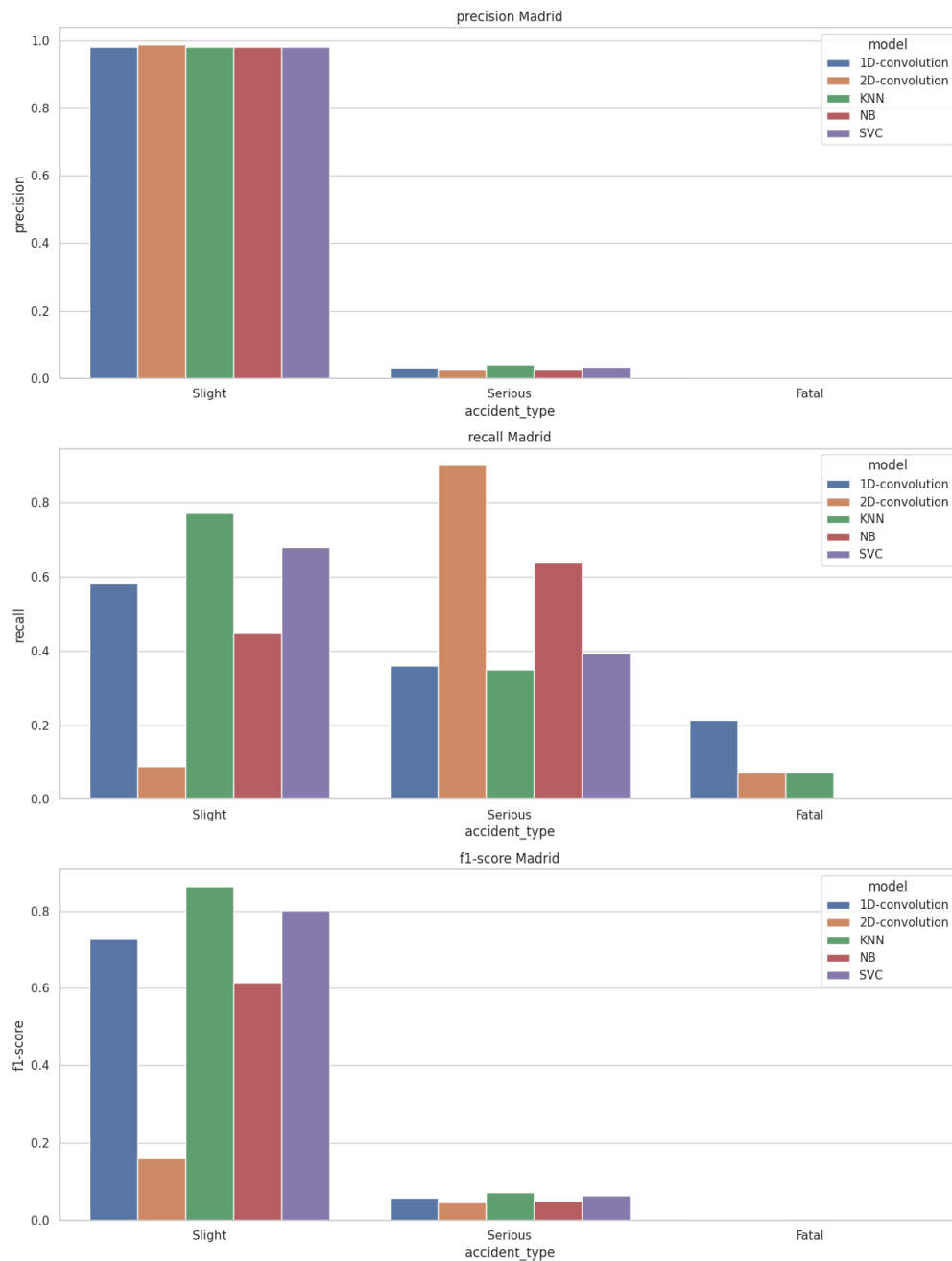


Figura 5.10: Comparativa de las métricas de las predicciones sobre el conjunto de test de los modelos.

6 Conclusiones

6.1 Trabajos para futuro

1. Utilizar NAS para encontrar la mejor estructura de la red para un problema, debe ser más **simple** de las que ha construido Autokeras.
2. Utilizar metaaprendizaje para entrenar modelos que averiguen los mejores hiperparámetros para otros modelos
3. Añadir más características al dataset como la velocidad para aumentar el rendimiento, esto se puede hacer con otro dataset de Madrid que hay velocidades medias por tramos.
4. Usar OpenRefine para clusterizar nombres de tipos de via
5. Las transformaciones sobre las variables del dataset son un punto crítico, por ejemplo la variable día/noche podría tener otro rango de horas, que afectaría notablemente a los árboles de decisión de XGBoost.
6. Otra opción es aumentar la complejidad del dataset en base a los datos que hay, como por ejemplo añadir la estación del año, añadir la temperatura ambiente (tal vez las temperaturas más frías de la carretera minimicen el agarre de los neumáticos)
7. Probar a hacer resampling de otra forma (reduciendo el número de muestras artificiales por ejemplo)

