UNIVERSIDAD POLITÉCNICA DE MADRID

ESCUELA TÉCNICA SUPERIOR
DE INGENIEROS DE TELECOMUNICACIÓN

GRADO EN INGENIERÍA DE TECNOLOGÍAS Y
SERVICIOS DE TELECOMUNICACIÓN

TRABAJO FIN DE GRADO

DISEÑO DE UN SISTEMA DE ANÁLISIS DE DATOS PARA EL ESTUDIO DE LA EVOLUCIÓN DE PACIENTES CON TRAUMATISMOS GRAVES

Enrique Salinas Martín

2021

UNIVERSIDAD POLITÉCNICA DE MADRID

ESCUELA TÉCNICA SUPERIOR
DE INGENIEROS DE TELECOMUNICACIÓN

GRADO EN INGENIERÍA DE TECNOLOGÍAS Y SERVICIOS DE TELECOMUNICACIÓN

TRABAJO FIN DE GRADO

DISEÑO DE UN SISTEMA DE ANÁLISIS DE DATOS PARA EL ESTUDIO DE LA EVOLUCIÓN DE PACIENTES CON TRAUMATISMOS GRAVES

Autor
Enrique Salinas Martín

Tutora
Blanca Larraga

2021

Resumen

Los traumatismos suponen un problema de salud pública mundial siendo una de las principales causas de muerte, especialmente en los menores de cuarenta años.

La digitalización de la información y la recolección de registros de trauma han generado grandes bases de datos sobre las que se han realizado análisis para la extracción de conclusiones y la generación de modelos predictivos.

La base de datos TPQ PUF año 2016 contiene todos los registros enviados al Banco Nacional de Datos de Trauma (NTDB) del año 2016 por los centros de traumas, estadounidenses, participantes en el Trauma Quality Improvement Program (TQIP).

En este Trabajo Fin de Grado se realiza un estudio de la base de datos de traumas TQP PUF año 2016. Se analizan las distintas tablas y relaciones de la base de datos para ver cuales son de interés. Seleccionadas las variables más significativas, se realiza una limpieza de las mismas, eliminando valores erróneos y outliers a partir de la representación de su distribución con Histogramas y Gráficos de Caja. Una vez limpiada la base de datos, se estudia la correlación lineal de Pearson entre variables sin encontrar relaciones lineales a parte de las evidentes. También se emplea Análisis de Componentes Principales (PCA) y Random Forest para intentar identificar las variables más importantes y reducir la complejidad. Adicionalmente se entrenan árboles decisores regresivos para predecir el tiempo del paciente en el Departamento de Emergencias obteniendo modelos inservibles con alto overfitting.

Palabras clave: Base de datos, Análisis de datos, Lesiones, Traumatismos, Correlación, PCA, Árboles de decisión.

Abstract

Traumatisms are a global public health issue, being one of the main causes of death especially in those under forty years old.

The digitalization of information and the collection of trauma records have generated large databases from which analysis have been carried out to draw conclusions and generate predictive models.

The TPQ PUF database of 2016 contains all the records submitted to the National Trauma Data Bank (NTDB) of 2016 by trauma centers, from the United States, participating in the Trauma Quality Improvement Program (TQIP).

In this Final Degree Thesis a study of the trauma database TQP PUF of the year 2016 is carried out. The different tables and relationships of the database are analyzed to see which ones are of interest. Once the most significant variables are selected, these are evaluated, eliminating erroneous values and outliers from their represented distribution using Histograms and BoxPlots. Once the database has been cleaned, the Pearson linear correlation between variables is studied without finding linear relationships other than the obvious ones. Principal Component Analysis (PCA) and Random Forest are also used to identify the most important variables and reduce complexity. In addition, regressive decision trees are trained to predict the patient's time in the Emergency Department, obtaining unusable models with high overfitting.

Palabras clave: Database, Data analysis, Injuries, Traumatisms, Correlation, PCA, Decision Trees.

Agradecimientos

Me gustaría agradecer a mis padres por apoyarme y motivarme durante toda mi vida, a mis amigos y compañeros de la Universidad por acompañarme en esta etapa de mi vida y al equipo de Robolabo ya que sin ellos este trabajo no habría sido posible.

Índice general

Resumen v

Abstract vii

Agradecimientos ix

Índice General xi

Índice de Figuras xiii

Índice de Tablas xv

Lista de Acrónimos xvii

1. Introducción 1
- 1.1. Traumatismos . . . 1
- 1.2. Base de datos de Traumatismos . . . 3
- 1.3. Objetivos . . . 3
- 1.4. Estructura . . . 4

2. Estado del arte 5
- 2.1. Aplicaciones del análisis de datos . . . 5
  - 2.1.1. Aplicaciones del análisis de datos al cuidado de la salud . . . 6
    - 2.1.1.1. Aplicaciones del análisis de datos al estudio de traumatismos . . . 9

3. Descripción de la base de datos 13
- 3.1. Descripción de las tablas . . . 13
- 3.2. Relaciones entre tablas . . . 15
- 3.3. Eliminación de columnas y filas . . . 16
- 3.4. Conceptos para la limpieza de datos . . . 17
  - 3.4.1. Outliers . . . 17
  - 3.4.2. Visualización de la distribución de los datos . . . 17
    - 3.4.2.1. Diagramas de Cajas . . . 18
    - 3.4.2.2. Histograma . . . 19
- 3.5. Descripción tablas de interés . . . 19
  - 3.5.1. Tabla de códigos de lesión AIS, AISPCODE . . . 20
  - 3.5.2. Tabla DEMO . . . 22

3.5.3. Tabla ED 23
3.5.4. Tabla PROTDEV 25
3.5.5. Tablas TBI_CM y TBI_GSC_Q 26
3.5.6. Tabla TRANSPORT 26
3.5.7. Tablas COMORBID y COMPLIC 27
3.5.8. Tabla DISCHARGE 28
3.5.9. Utilización de instancias 29

4. Desarrollo **31**
4.1. Herramientas para el desarrollo 31
4.2. Variables categóricas 32
4.3. Correlación de Pearson 33
4.3.1. Matriz de correlación 33
4.4. Análisis de componentes principales (PCA) 34
4.5. Modelos de predicción 35
4.5.1. Árboles de decisión 35
4.5.2. Random Forest 36
4.6. Resultados Correlaciones 37
4.7. Resultados PCA 40
4.8. Resultados Árboles de decisión 42
4.9. Resultados Random Forest 45

5. Conclusiones y líneas futuras **47**
5.1. Conclusiones 47
5.2. Líneas futuras 48

Referencias **49**

A. Aspectos éticos, económicos, sociales y ambientales **55**
A.1. Introducción 55
A.2. Descripción de impactos relevantes relacionados con el proyecto 55
A.3. Conclusiones 56

B. Presupuesto económico **57**

Índice de figuras

2.1. Preprocesamiento de las imágenes MRI en la publicación [1] 7
2.2. *(a)* Precisión de la Red Neuronal Artificial y *(b)* Precisión de la arquitectura VGG-16 para los datos de entrenamiento y para los datos de validación en la publicación [1] 8
2.3. *(a)* Ejemplo de pregunta y *(b)* Ejemplo de diagnóstico de la aplicación Ada. 9
2.4. Representación de la Red Neuronal empleada en la publicación [2] . . 12
2.5. Rendimiento de las regresiones lineales, las redes neuronales artificiales (ANN) y los clínicos prediciendo la supervivencia de 100 pacientes en la publicación [2]. Las barras blancas indican ANN; las negras, clínicos; las grises, regresiones lineales 12

3.1. Diagrama de clases 16
3.2. Zonas de un Diagrama de Caja [3] 18
3.3. Ejemplo de Histograma [4] 19
3.4. AISCODE - Primer dígito [5] 20
3.5. Distribución Severidad - Zona 21
3.6. Número de lesiones por incidente 22
3.7. Distribución de Edades 23
3.8. Diagrama de Caja del peso y la altura antes de la eliminación de entradas. 24
3.9. *(a)* Histograma de pesos antes de eliminación y *(b)* Distribución normal después de eliminación. 24
3.10. Histograma de los dispositivos de protección. 25
3.11. *(a)* Diagrama de caja de `LOSMIN` antes de eliminación y *(b)* Histograma de `LOSMIN` después de eliminación. 29
3.12. Porcentaje de utilización de las instancias de cada tabla. 29

4.1. Ejemplo de PCA con 2 variables [6] 34
4.2. Ejemplo de árbol de decisión de Profundidad 2 36
4.3. Correlación de Pearson de las tablas AISPCODE y DEMO 39
4.4. *(a)* Scree Plot con ISS y *(b)*Proyección sobre PC1 y PC2 con ISS . . . 40
4.5. *(a)* Scree Plot sin ISS y *(b)*Proyección sobre PC1 y PC2 sin ISS 41
4.6. Distribución de `EDDAYS` objetivo y valores obtenidos. *(a)* Unión con AISPCODE *(b)* Unión con las 4 tablas. 44

Índice de tablas

1.1. Intencionalidad de las lesiones de pacientes que fallecieron en Estados Unidos en 2019 [7] 1
1.2. Probabilidad de muerte asociada a cada severidad [8] 2

2.1. Resultados de los clasificadores del artículo [9] 7
2.2. Importancia relativa de cada variable en la severidad del accidente en la publicación [10] 10

3.1. Descripción de las tablas de la BD 14
3.2. 5 filas de la tabla `ED` 15

4.1. Ejemplo de conversión de variables categóricas a numéricas usando One Hot Encoding 33
4.2. Ejemplo de Matriz de Correlación. 33
4.3. Loadings de PC1 al aplicar PCA a la unión de ED y AISPCODE . . . 41
4.4. Árboles de decisión regresivos. 43
4.5. R^2 de entrenamiento, R^2 de validación y puntuación OOB de los diferentes Random Forest 45

B.1. Costes de personal. 57
B.2. Costes de recursos materiales. 57
B.3. Costes totales. 58

Lista de Acrónimos

AIS: Abbreviated Injury Scale.

ED: Emergency Department.

EMS: Emergency Medical Services.

TQIP: Trauma Quality Improvement Program.

GCS: Glasgow Coma Scale.

TQP: Trauma Quality Programs.

PUF: Participant Use File.

ECODE: External Cause of Injury CODE.

DCODE: Diagnosis CODE.

PCODE: Procedure CODE.

ICD-9-CM: International Classification of Diseases, Ninth Revision, Clinical Modification.

ICD-10-CM: International Classification of Diseases, Tenth Revision, Clinical Modification.

BD: Base de Datos.

PCA: Análisis de componentes principales.

ISS: Injury Severity Score.

ECM: Error Cuadrático Medio.

OOB: Out of Bag.

Capítulo 1

Introducción

En este Capítulo se evidencia la importancia de los traumas, se presenta la base de datos a utilizar y se exponen el enfoque y los objetivos de este Trabajo Fin de Grado. Adicionalmente, se expone la estructura de capítulos del documento.

1.1. Traumatismos

Los traumatismos son lesiones que sufre una persona por una fuerza externa en sus órganos o tejidos. En función de la zona afectada se puede distinguir entre[11]:

- Traumatismos auriculares cuando la lesión se produce en articulaciones.
- Traumatismos óseos cuando la lesión afecta a huesos.

Los traumatismos provocan la muerte de más de 5.8 millones de personas anualmente, son por tanto un problema de salud pública mundial[12].

En el año 2019 en Estados Unidos fallecieron 246.041 personas por traumatismos. 169.628 hombres y 76.413 mujeres[7]. En la Tabla 1.1, se clasifican estas muertes en base a la intención cuando se produjo la lesión. En la última columna se expresa el porcentaje de muertes totales por trauma producido con esta intención. De la tabla se concluye que los accidentes involuntarios son la principal causa de muerte por trauma en Estados Unidos, representando más fallecimientos que el resto de intenciones juntas; hay un número elevado de muertes por suicidios y el número de muertes por intervención legal es relativamente bajo.

Intencionalidad	**Número de muertes**	**% de muertes**
Involuntario	173.040	70,3
Suicidio	47.511	19,3
Homicidio	19.141	7,8
Intencionalidad indeterminada	5.683	2,3
Intervención legal	652	0,3

Tabla 1.1: Intencionalidad de las lesiones de pacientes que fallecieron en Estados Unidos en 2019 [7]

Según [13], no existe un método globalmente aceptado para la comparación de traumatismos graves. Los más utilizados son la puntuación de la gravedad de traumatismos y lesiones (TRISS), la puntuación de la gravedad de la lesión (ISS) y la escala abreviada de lesiones (AIS).

En la base de datos que se va a analizar en este Trabajo Fin de Grado se incluyen ISS y AIS.

Injury Severity Score (ISS) es una puntuación médica establecida para evaluar la gravedad de las lesiones asociadas a un incidente. Se profundizará más sobre esta puntuación en la Sección 3.5.3.

Abbreviated Injury Scale (AIS) es un sistema de codificación de lesiones individuales basado en anatomía, que clasifica cada lesión por su región del cuerpo y asigna una severidad[14]. A continuación se adjunta el significado asociado a cada valor de severidad[14]:

1. Menor
2. Moderado
3. Serio
4. Severo
5. Crítico
6. Máximo (actualmente intratable)

Según [15], se puede calcular la probabilidad de muerte de un herido a partir de la severidad con la que se clasifica la lesión. En la Tabla 1.2 se puede ver la probabilidad de muerte asociada a cada severidad.

Severidad	Probabilidad de muerte (%)
1	0
2	1 – 2
3	8 – 10
4	5 – 50
5	5 – 50
6	100

Tabla 1.2: Probabilidad de muerte asociada a cada severidad [8]

1.2. Base de datos de Traumatismos

La primera base de datos de traumas procesada en ordenadores fue creada en 1969 en el hospital de Chicago Cook County. Esta base de datos se convirtió en el prototipo para los registros de traumas en Illinois, logrando en 1971 contar con información de más de 50 hospitales[16].

En 1997, el comité de trauma del Colegio Americano de Cirujanos (ACS) decidió que era importante la recolección de entradas de traumas de todo el país, para mejorar la calidad de la información sobre traumas. Por este motivo, crearon un subcomité para el desarrollo del Banco de Datos de Traumas Nacional (NTDB)[17].

La base de datos TPQ PUF año 2016 contiene todos los registros enviados al Banco Nacional de Datos de Trauma (NTDB) del año 2016. Esta base de datos contiene entradas aportadas por los diferentes centros de traumas participantes en el Trauma Quality Improvement Program (TQIP). A día de hoy hay más de 850 centros participando en Estados Unidos[18] y esta versión de la base de datos cuenta con más de 960000 incidentes.

1.3. Objetivos

El propósito de este Trabajo Fin de Grado consiste en realizar un estudio exhaustivo de la base de datos TPQ PUF año 2016 para intentar facilitar una mejor atención y correcta predicción del desarrollo del traumatismo. Al analizar la base de datos se podrá estudiar la evolución de los casos recogidos en la misma para poder ver si, de alguna manera, se puede predecir lo que cabe esperar en casos futuros de trauma. Para lograr este propósito principal, se tratarán de alcanzar los siguientes objetivos:

- Realizar un estudio del estado del arte de sistemas de predicción de trauma en la industria de la salud.
- Estudiar de la base de datos TPQ PUF year 2016: estructura, relación entre tablas y contenido de las tablas.
- Realizar una limpieza de la base de datos como, por ejemplo, inconsistencias o entradas incompletas.
- Estudio de relaciones entre atributos/tablas de la base de datos.
- Si se encuentran correlaciones importantes, entrenamiento de un modelo predictivo.
- Pruebas y validación de los resultados obtenidos.

1.4. Estructura

En esta sección se expone la estructura de capítulos en la que se divide este documento con una breve descripción del contenido de cada capítulo.

- **Capítulo 1:** En este Capítulo se evidencia la importancia de los traumas, se presenta la base de datos a utilizar y se exponen el enfoque y los objetivos de este Trabajo Fin de Grado. Adicionalmente, se expone la estructura de capítulos del documento.

- **Capítulo 2:** En este Capítulo se proporcionan distintas definiciones de *análisis de datos* y se exponen aplicaciones recientes de análisis de datos y generación de modelos desde una perspectiva general hasta el enfoque del trabajo, los traumatismos.

- **Capítulo 3:** En este Capítulo se describe la base de datos. Se presentan las tablas que la componen, los atributos de cada tabla de interés y las relaciones entre las distintas tablas. Adicionalmente se incluye una sección en la que se especifica cómo se manejarán los valores nulos o erróneos y se detalla, en cada descripción de tabla, el procedimiento seguido para la limpieza de la misma.

- **Capítulo 4:** En este Capítulo se explicarán las herramientas empleadas para el desarrollo de este Trabajo Fin de Grado. Después se explicarán las nociones básicas necesarias para el entendimiento de los análisis realizados y finalmente se mostrarán los resultados y conclusiones obtenidos de los análisis. Para analizar las relaciones entre tablas se utilizará la Correlación Lineal de Pearson y para intentar reducir las variables se utilizará PCA y Random Forest. Adicionalmente se generarán 6 árboles de decisión para intentar predecir el tiempo que pasa un paciente en el Departamento de Emergencias (ED).

- **Capítulo 5:** En este Capítulo se presentan las conclusiones extraídas del desarrollo de este proyecto y se sugieren líneas futuras de investigación.

Capítulo 2

Estado del arte

En este Capítulo se proporcionan distintas definiciones de *análisis de datos* y se exponen aplicaciones recientes de análisis de datos y generación de modelos desde una perspectiva general hasta el enfoque del trabajo, los traumatismos.

Una posible definición de análisis de datos es: ciencia que estudia un conjunto de datos para extraer conclusiones sobre la información de los datos, para a partir de estas tomar decisiones o ampliar los conocimientos sobre esa información[19].

Otra posible definición es: el análisis de datos consiste en el estudio exhaustivo de un conjunto de datos, para que una entidad o una empresa tome una decisión en base a las conclusiones sacadas del estudio[20].

Wikipedia define análisis de datos como: "proceso que consiste en inspeccionar, y transformar datos con el objetivo de resaltar información útil, para sugerir conclusiones y apoyo en la toma de decisiones"[21].

2.1. Aplicaciones del análisis de datos

Con el constante auge de la tecnologías de la información y comunicación (TIC), cada vez resuena más el término *Análisis de datos* y es que, sus aplicaciones parecen ilimitadas.

Entre las aplicaciones del análisis de datos se encuentra el uso para finanzas. Las empresas bancarias han empleado la gran cantidad de datos que recolectaron de sus préstamos, para estudiar perfiles de cliente para analizar el riesgo detrás de cada préstamo pudiendo elegir apropiadamente cuando conceder uno[22].

Otro ejemplo de aplicación, son las búsquedas de Internet; los motores de búsqueda como Google utilizan algoritmos de ciencia de datos para proporcionar el mejor resultado a la búsqueda.

Cuando realizamos una búsqueda en Google, recibimos las páginas más relevantes en menos de un segundo. Para devolvernos este resultado, Google utiliza múltiples algoritmos de búsqueda para encontrar entre su amplia biblioteca de páginas web las que mejor se adaptan a la búsqueda. Estos algoritmos de búsqueda tienen en

cuenta múltiples factores como las palabras que hemos introducido en la búsqueda, la popularidad y relevancia de las páginas web o nuestra ubicación.

Para que los algoritmos de búsqueda de Google funcionen, primero Google debe ser capaz de interpretar la búsqueda realizada para entender nuestras intenciones e intereses al realizar la búsqueda. Como parte del procesamiento de texto de nuestra búsqueda, Google reconoce faltas de ortografía y aplica un sistema de sinónimos. Este sistema de sinónimos tardó más de 5 años en desarrollarse pero mejoró los resultados de búsqueda en más de un 30 % en todos los idiomas. Este sistema permite que; al buscar *cómo cambiar una bombilla*, el motor de búsqueda también considere los resultados de la búsqueda *cómo sustituir una bombilla*, aumentando notablemente el número de resultados[23].

Otro ámbito de aplicación del análisis de datos y de la generación de modelos predictivos es el reconocimiento facial.

En 2014 un equipo del departamento de investigación de Inteligencia Artificial de Facebook desarrolló *DeepFace*, un sistema de reconocimiento facial. El proceso habitual hasta la fecha para el reconocimiento facial consistía en 4 pasos: detección, alineamiento, representación y clasificación. El éxito del equipo de Facebook se debe a su replanteamiento de los pasos 2 y 3; emplearon un modelado de caras 3D para lograr una transformación afín por partes y, mediante el uso de una red neuronal profunda de 9 capas, la obtención de una representación de caras. Los resultados de entrenar esta red neuronal profunda fueron altamente satisfactorios, logrando un 97.35 % de precisión en el conjunto de datos de caras etiquetadas en la naturaleza (LFW)[24].

El análisis de datos y la generación de modelos, tienen por tanto un impacto significativo en la sociedad, teniendo multitud de aplicaciones como el reconocimiento facial utilizado en el ámbito de la seguridad, el reconocimiento de voz en los asistentes personales, la conducción automática o el estudio de precios de casas; y por supuesto, una posible aplicación es el uso del análisis de datos para el cuidado de la salud.

2.1.1. Aplicaciones del análisis de datos al cuidado de la salud

En esta sección se muestran algunas de las múltiples aplicaciones del análisis de datos al cuidado de la salud.

Existen sistemas para facilitar procedimiento médicos como la detección de tumores o la estenosis de arterias. Estos sistemas emplean técnicas de machine learning y análisis de ondas para cumplir su función[22].

Los tumores cerebrales son una enfermedad causada por el crecimiento de células anormales en el cerebro. En los últimos 10 años se han realizado distintos intentos de aplicación de técnicas de aprendizaje automático para la detección de tumores cerebrales.

En 2012 un grupo de investigadores de la Universidad de Ciencia y Tecnología

de Jordania (JUST) publicó el artículo *Machine Learning Approach for Brain Tumor Detection*. En esta publicación se presenta un acercamiento a la detección de tumores sobre una imagen por resonancia magnética (MRI)[1], aplicando aprendizaje automático. Para el desarrollo de este experimento preprocesaron distintas características de las 27 imágenes MRI utilizadas, como la zona de interés o el color. Hecho esto, entrenaron 4 distintos clasificadores para comparar los resultados. Los resultados obtenidos se muestran en la Tabla 2.1. Los investigadores a la vista de los resultados en esa tabla, concluyeron que las redes neuronales eran el mejor clasificador para esas imágenes MRI[9].

Algoritmo	**Tasa de acierto**	**Precisión**
Red Neuronal	66,6 %	0,63
Árbol J48	59,2 %	0,628
Naïve Bayes	59,2 %	0,691
Lazy-IBk	62,9 %	0,652

Tabla 2.1: Resultados de los clasificadores del artículo [9]

En 2020 un grupo de ingenieros del Instituto de Tecnología KLE publicaron el artículo *Brain Tumor Detection Using Deep Learning Models*. En este experimento se emplearon 253 imágenes MRI de pacientes, de las cuales 155 eran imágenes con tumor y el resto no, para comparar los resultados entre redes neuronales convolucionales y una arquitectura VGG-16 construida por ellos, en la detección de tumores cerebrales. Las imágenes utilizadas en el experimento tenían distintos tamaños, por lo que era necesario una etapa de preproceso antes de entrenar los modelos[1].

Se muestra un ejemplo de imagen MRI en el penúltimo paso de preprocesamiento en la Figura 2.1, en la imagen se ve el contorno dibujado en azul y los puntos extremos. A partir de estos puntos se recortaron las imágenes para quedarse con lo importante de la imagen.

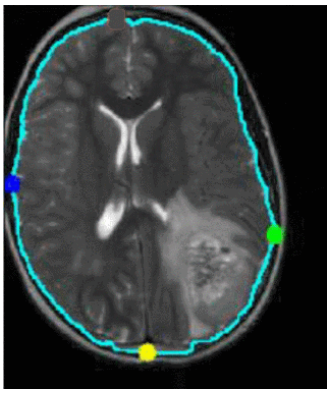


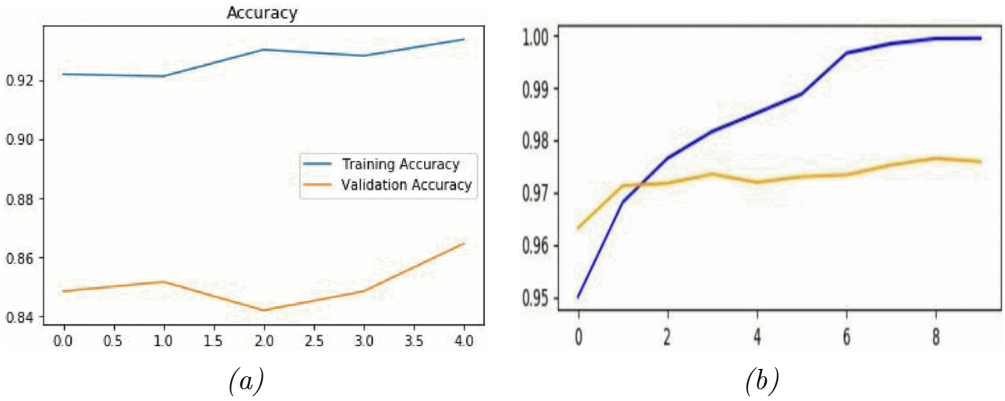
Figura 2.1: Preprocesamiento de las imágenes MRI en la publicación [1]

[1] MRI utiliza campos magnéticos y ondas de radio para producir imágenes detalladas del interior del cuerpo

Entre los pasos de preprocesamiento cabe destacar el aumento de datos de imagen realizado a partir de pequeñas modificaciones sobre las imágenes originales. De este modo, pasaron de tener 253 instancias a tener 2065 instancias.

Para el experimento, en cada algoritmo, utilizaron 1445 instancias para entrenamiento, 310 para test y 310 para validación[1]. Se estudió la precisión de los dos algoritmos durante 25 generaciones evolutivas y se obtuvieron las gráficas de la Figura 2.2.

Del experimento llevado a cabo concluyeron que ambos modelos tenían unos resultados satisfactorios para la identificación de tumores cerebrales. Comparando ambos modelos, afirmaron que la Red Neuronal Artificial tenía peores resultados que VGG-16, pero requería menor capacidad de cómputo[1].



(a) *(b)*

Figura 2.2: *(a)* Precisión de la Red Neuronal Artificial y *(b)* Precisión de la arquitectura VGG-16 para los datos de entrenamiento y para los datos de validación en la publicación [1]

Otra aplicación, de ciencia de datos en el campo del cuidado de la salud, es la personalización de tratamientos basados en la genética y genómica. Las técnicas de ciencia de datos permiten tener un conocimiento más profundo de los problemas genéticos en el uso de medicamentos concretos[22].

Otro ejemplo destacado es el desarrollo de LYNA (Lymph Node Assistant) realizado por Google. Esta aplicación identifica tumores de cáncer de mama que se metastatizan a los ganglios linfáticos cercanos. El ojo humano tiene dificultades para visualizar esto, mientras que LYNA obtuvo una tasa de acierto del 99 % en su primera prueba realizada sobre la combinación de dos bases de datos distintas[25].

Una aplicación interesante es el desarrollo de la aplicación móvil Ada. Esta aplicación móvil realiza preguntas de respuesta cerrada sobre tus síntomas y mediante el uso de Inteligencia Artificial ofrece un diagnóstico[26]. En la Figura 2.3 se muestra un ejemplo de pregunta y un ejemplo de diagnóstico de la aplicación Ada.

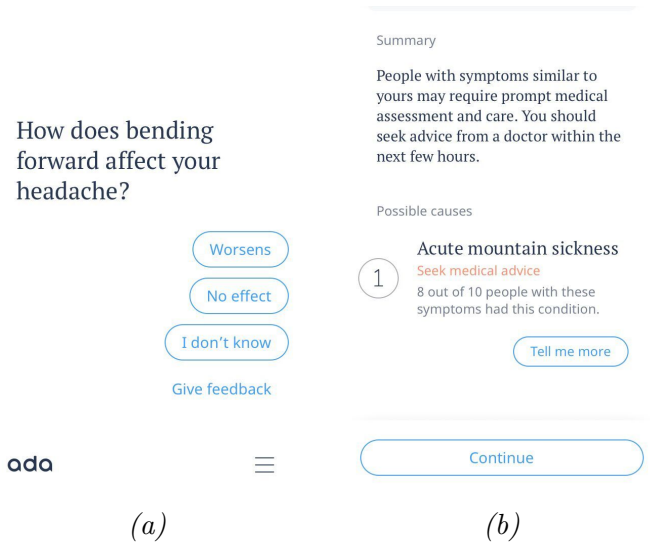


Figura 2.3: *(a)* Ejemplo de pregunta y *(b)* Ejemplo de diagnóstico de la aplicación Ada.

De igual forma que las anteriores aplicaciones, Oncora surge del empleo de análisis de datos en el área del cuidado de la salud. Oncora consiste en un software que emplea técnicas de machine learning para realizar recomendaciones a pacientes con cáncer basado en datos de los 15 años anteriores. El equipo de radiólogos y científicos de datos de este proyecto colaboraron en el análisis de los últimos 15 años, lo que supusieron más de 50000 registros de pacientes con cáncer para el desarrollo de este herramienta capaz de sugerir regímenes personalizados de radiación y quimioterapia[27].

2.1.1.1. Aplicaciones del análisis de datos al estudio de traumatismos

Por último, se exponen aplicaciones del análisis de datos en el área de estudio de este Trabajo Fin de Grado, los traumatismos.

En 2011 se analizaron datos de accidentes de coche producidos en Irán entre 2006 y 2008 con el propósito de identificar los factores más influyentes en la severidad de las lesiones que sufren los conductores en los accidentes de tráficos de las carreteras rurales de Irán; puesto que, el 70 % de las muertes en accidentes de tráficos del país suceden en este tipo de carreteras. La detección de estos factores permitiría su eliminación o control logrando mejorar la seguridad vial del país[10]. Cabe mencionar que en esta publicación la variable severidad puede tomar los siguientes valores:

- Sin lesión
- Lesión
- Fatalidad

Por lo tanto, no tiene el mismo significado que la severidad AIS mencionada en el Capítulo 1.

Para el estudio de la base se emplearon árboles predictores y se realizó una medición de la importancia de cada variable (VIM). En la Tabla 2.2 se adjuntan las variables presentes en el estudio y la importancia relativa de cada una; como se puede observar el cinturón de seguridad es la variable más importante, siendo 90 veces más importante que el tipo de colisión; el motivo del accidente es 16 veces más importante que el tipo de colisión; y el resto de variables no son importantes.

Variable	**Medida de la Importancia de la Variable (VIM)**
Cinturón de seguridad	0.8214
Causa del accidente	0.1484
Tipo de colisión	0.0088
Tipo de vehículo	0.0029
Condiciones climatológicas	0.0023
Edad	0.0023
Tipo de hombro	0.0023
Ancho del hombro	0.0023
Estado de la superficie de la carretera	0.0023
Iluminación	0.0023
Tipo de localización	0.0023
Ocurrencia	0.0023
Género	0.0001

Tabla 2.2: Importancia relativa de cada variable en la severidad del accidente en la publicación [10]

Una vez encontradas las variables más influyentes, realizaron un árbol de clasificación utilizando el 30 % de las instancias para prueba y el resto para entrenamiento; obteniendo unas tasas de acierto de 72,82 % y 72,49 % con los datos de entrenamiento y prueba, respectivamente[10].

Como resultado de este experimento, se descubrió que los principales factores asociados con la severidad de los traumatismos producidos en accidentes de coches consistían en: la realización de adelantamientos imprudentes, el exceso de velocidad y la falta de uso del cinturón de seguridad[10].

Otra publicación de interés es *A neural network approach to the diagnosis of morbidity outcomes in trauma care*. En esta publicación se exponen los resultados del uso de redes artificiales neuronales para evaluar las complicaciones traumáticas. Además de presentar los resultados obtenidos de su prototipo de red neuronal, muestran el significativo beneficio económico que conllevaría la inclusión de redes neuronales para la mejoría de diagnósticos de trauma. En este estudio solo se trata de predecir si se produce la sepsis, el resto de complicaciones no se estudian[28].

Para el desarrollo de este experimento emplearon 912 instancias de traumas. Tras

la limpieza quedaron 515 casos, de los que 13 tuvieron sepsis como complicación. De estas 515 instancias emplearon 256 para entrenamiento y 259 para pruebas. La red neuronal tendría como variable objetivo la sepsis y utilizaría otras 18 variables como son la edad del paciente, la puntuación de la gravedad de lesión (ISS) o la temperatura en el Departamento de Emergencias[28].

Su modelo logra diagnosticar correctamente la complicación de sepsis en víctimas de lesiones de contusión traumática. Para el diseño de la red neuronal se usa como función de transferencia la sigmoide, al someter el modelo a pruebas con el conjunto de datos de test se obtuvo un 100 % de sensibilidad[2] y un 96.5 % de especificidad[3]; por lo que el potencial de este modelo es elevado[28].

Por destacar una última publicación sobre la aplicación de análisis de datos al estudio de traumatismos, vamos a comentar la publicación *Use of an artificial neural network to predict head injury outcome*. En esta publicación los autores ilustran las redes neuronales artificiales como una herramienta innovadora y potente para el modelado, mostrando el potencial que podría tener su aplicación para la predicción en el campo de la neurocirugía. Los autores designaron un modelo para predecir la supervivencia de los pacientes con lesión cerebral traumática[2].

En la Figura 2.4 se muestra la arquitectura de la Red Neuronal diseñada para el experimento, está formada por 11 nodos de entrada interconectados uno a uno a las capas ocultas de la red; cada capa oculta se conecta al único nodo de salida. Los nodos ocultos hacen posibles las interacciones no lineales entre las diferentes variables. Para el entrenamiento de la red se introducen entradas con salidas conocidas y la red ajusta los pesos de cada enlace en función de la realimentación[2].

Para demostrar el potencial de las redes neuronales, los autores comparan los resultados obtenidos de las redes neuronales con las predicciones de 4 médicos del personal de neurocirugía y 5 neurocirujanos, y con las predicciones de 2 regresiones lineales. Para ello, primero entrenaron 30 redes neuronales y se quedaron con las 5 que tenían mejores resultados. En la Figura 2.5 se recogen la exactitud, la sensibilidad y la especificidad de los clínicos, las regresiones lineales y las redes neuronales al intentar predecir 100 instancias.

En esta publicación, tal y como se ve en la Figura 2.5, Las redes neuronal artificiales proporcionan mejores predicciones que los médicos especialistas o los modelos de regresión; siendo más precisas y sensibles[2].

[2] La sensibilidad es el porcentaje de positivos que son correctamente identificados como positivos
[3] La especificidad es el porcentaje de negativos que son correctamente identificados como negativos

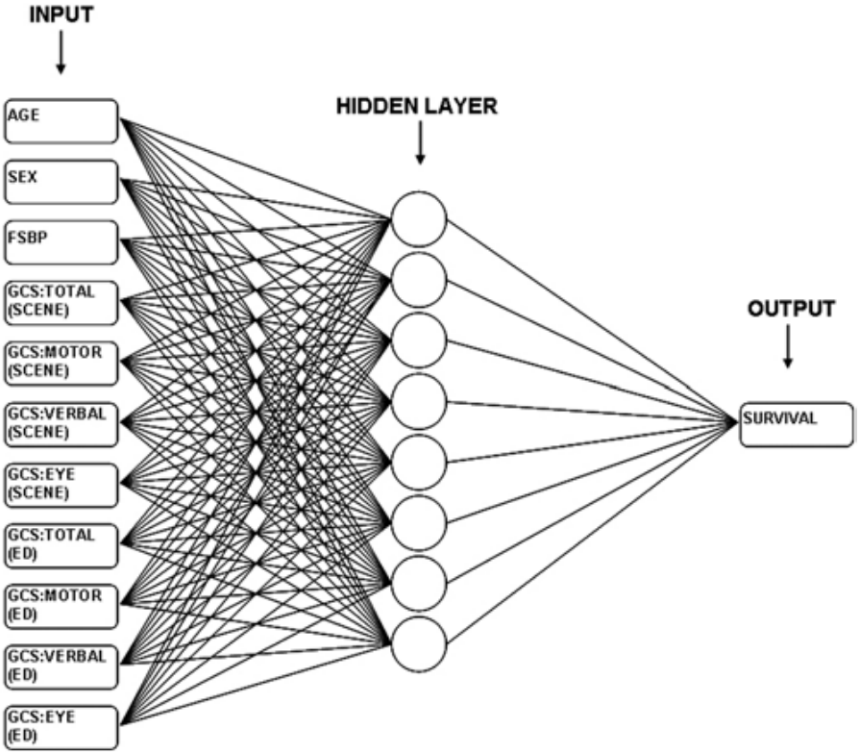


Figura 2.4: Representación de la Red Neuronal empleada en la publicación [2]

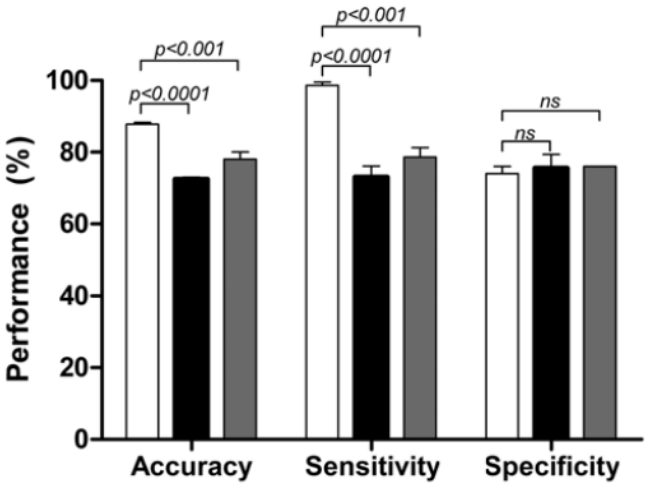


Figura 2.5: Rendimiento de las regresiones lineales, las redes neuronales artificiales (ANN) y los clínicos prediciendo la supervivencia de 100 pacientes en la publicación [2]. Las barras blancas indican ANN; las negras, clínicos; las grises, regresiones lineales

Capítulo 3

Descripción de la base de datos

En este Capítulo se describe la base de datos. Se presentan las tablas que la componen, los atributos de cada tabla de interés y las relaciones entre las distintas tablas. Adicionalmente se incluye una sección en la que se especifica cómo se manejarán los valores nulos o erróneos y se detalla, en cada descripción de tabla, el procedimiento seguido para la limpieza de la misma.

3.1. Descripción de las tablas

La base de datos TPQ PUF ha variado sus tablas con el cambio de versiones. La versión de 2016, que es la que se emplea en este Trabajo Fin de Grado, está compuesta por 29 tablas tal y como se muestra en la Figura 3.1. Los nombres de cada tabla suelen ser descriptivos; en la Tabla 3.1 se puede encontrar una descripción y el número de instancias de 20 de las 29 tablas. Las otras 9 tablas son: `AISDES`, `ECODEDES`, `ICD10_ECODEDES`, `ICD10_PCODEDES`, `DCODEDES`, `ICD10_DCODEDES`, `ICD10_LOCDES`, `PCODEDES` y `AIS05TO98_CROSSWALK`. Las 8 tablas con el sufijo "DES" no se incluyen en la tabla porque se pueden explicar todas simultáneamente. Las 8 tablas con el sufijo "DES" son tablas para describir el significado de los códigos de las tablas a las que se asocian. Por ejemplo, la tabla `DCODEDES` proporciona una descripción de los códigos de diagnóstico ICD-9-CM, DCODE, que parten de la tabla con el mismo nombre, `DCODE`. La tabla `AIS05TO98_CROSSWALK` proporciona una conversión de códigos AIS 05 a AIS 98. Esta tabla es de utilidad si, además de las entradas de la versión de 2005, se dispone de entradas de la versión de 1998; esto permite trabajar con un mayor número de entradas simultáneamente. No obstante, todas las entradas que se tienen en la base de datos TPQ PUF 2016 utilizan AIS 05, por lo tanto, la versión de 98 no se utilizará. Cabe destacar la tabla `ECODE` que aporta información acerca del mecanismo externo de la lesión y de la intencionalidad de la misma[29].

ICD-x-CM es el sistema oficial de asignación de códigos a diagnósticos y procedimientos asociados utilizado en hospitales en Estados Unidos[30]. Como se observa en la tabla 3.1, en esta BD hay entradas de la novena y de la décima versión de estos códigos. Entre una versión y otra hay notables diferencias; hay alrededor de 19 veces más ICD-10-CM PCODEs y alrededor de 5 veces más DCODEs, siendo

los PCODE los códigos de procedimiento y los DCODE los de diagnóstico. Además ICD-10 cuenta con categorías alfanuméricas en vez de numéricas[31]. En ICD-9-CM, tanto los códigos de procedimiento como los de diagnóstico, tienen el formato AB.CD, por ejemplo *99.23* significa *Inyección de esteroides*. Los dos primeros dígitos de los códigos de procedimientos ICD-9-CM especifican la región anatómica y los dos últimos el procedimiento seguido[32]. Los códigos ICD-10-CM son más complejos y tienen el formato ABC.DEFG, por ejemplo *HZ2.ZZZZ* significa *Servicios de Desintoxicación para el tratamiento por abuso de substancias*.

TABLA	ENTRADAS	DESCRIPCIÓN
AISPCODE	3237059	El código AIS calculado a partir de los códigos de diagnóstico ICD-9
COMORBID	1648445	Condiciones comórbidas
COMPLIC	991382	Complicaciones en hospital
DEMO	963134	Información demográfica
DCODE	1441701	Códigos de diagnóstico ICD-9-CM
DISCHARGE	962957	Información de alta y resultado
ECODE	962957	ECODE ICD-9-CM
ED	963134	Información Departamento de Emergencias
ICD10_DCODE	3372858	Códigos de diagnóstico ICD-10-CM
ICD10_ECODE	962957	ECODE ICD-10-CM
ICD10_LOC	963134	Códigos de localización ICD-10-CM
PCODE	5308716	Códigos de procedimiento (tanto ICD-9-CM como ICD-10-CM)
PM	748189	Información sobre los procesos TQIP de elementos de medidas de cuidado
PM_EMBOLIZE_SITE	748390	Información sobre los procesos TQIP de elementos de medidas de cuidado para zona de embolización
PM_TBI_CM	749579	Información sobre los procesos TQIP de elementos de medidas de cuidado para monitorización cerebral
PM_TBI_GSC_Q	761845	Información sobre los procesos TQIP de elementos de medidas de cuidado para calificadores de evaluación GCS.
PROTDEV	1162673	Dispositivos de protección
TRANSPORT	1932697	Información sobre transporte al ED
VITALS	1925914	Constantes vitales de EMS y ED
INCLUSION	963134	Información relativa a la afiliación con TQIP de la institución y sobre el cumpliendo del cumplimiento de inclusión TQIP

Tabla 3.1: Descripción de las tablas de la BD

En la tabla 3.2 se muestran 5 filas de la tabla `ED`, que proporciona información del departamento de emergencia, a modo de ejemplo. Todas las tablas de la BD presentan el formato de esta tabla, cada fila representa una instancia y cada columna representa una variable. Por este motivo, se hablará a lo largo del documento indistintamente de muestra, instancia o fila para referirnos a las filas de la tabla y variable, atributo, característica y columna, para referirnos a las columnas de la tabla.

INC_KEY	TRANSFER	SIGNSOFLIFE	EDMIN	EDDAYS	ISSAIS	HEIGHT	WEIGHT
160021573	Yes	Arrived with signs of life	571.0	1.0	2	168.0	75.0
160021574	No	Arrived with signs of life	64.0	1.0	34	193.0	150.0
160136517	No	Arrived with signs of life	1440.0	1.0	1	175.0	82.0
160021577	Yes	Arrived with signs of life	512.0	1.0	50	165.0	52.0
160136519	No	Arrived with signs of life	148.0	1.0	2	185.0	98.0

Tabla 3.2: 5 filas de la tabla `ED`

3.2. Relaciones entre tablas

Un diagrama UML es una representación gráfica parcial del modelo de un sistema en desarrollo, implementación o existente[33].

Dado que el número de tablas es elevado, se decide hacer un diagrama UML para poder visualizar fácilmente todas las tablas, así como la forma que tienen de relacionarse unas con otras. Este diagrama puede verse en la Figura 3.1. En esta figura se han añadido las tablas que componen la base de datos, así como una tabla que se ha creado para facilitar la visualización de las relaciones entre tablas que es la tabla INC_KEY TABLE.

Cuando dos tablas están relacionadas se representa, además de la unión, el número de instancias que hay en una tabla por cada instancia de la otra en cada extremo de la unión. Cuando la tabla X toma el valor "1" en su unión con la tabla Y, significa que por cada instancia de Y hay una única instancia de X asociada a dicha entrada. Cuando la tabla X toma el valor "1..*" en su unión con la tabla Y, significa que por cada instancia de Y hay al menos una, pudiendo haber más de una, instancia de X asociada a dicha entrada. El caso "1..*" en las tablas asociadas a la tabla INC_KEY TABLE viene dado porque hay múltiples entradas asociadas a un mismo INC_KEY, es decir, hay varias entradas en la tabla que pertenecen al mismo incidente y por tanto tienen el mismo código de incidente (INC_KEY). Por ejemplo, la tabla `COMPLIC` tiene varias instancias con la misma INC_KEY dado que puede haber varias complicaciones en el tratamiento de un paciente.

De esta manera, la Figura 3.1 nos permite ver fácilmente la forma de conectar distintas tablas y las tablas cuyas entradas pueden ser unívocamente identificadas mediante la INC_KEY. INC_KEY es un identificador asociado a cada incidente pero

cada incidente puede tener múltiples entradas en la tabla como acabamos de ver.

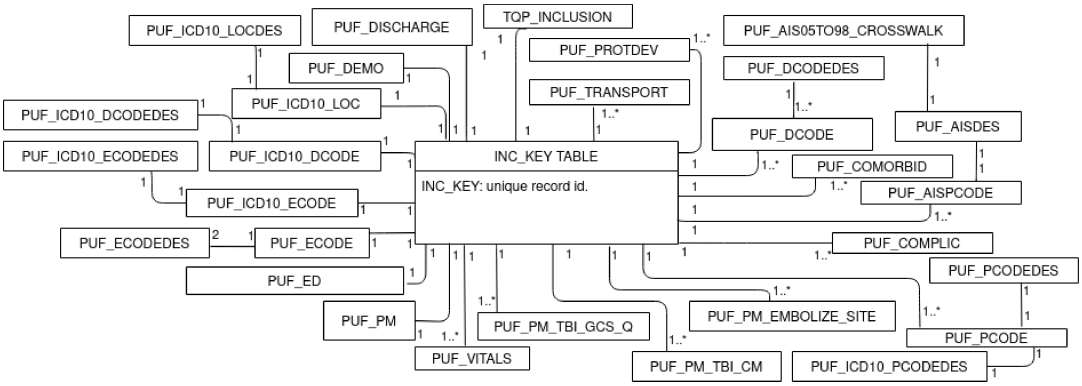


Figura 3.1: Diagrama de clases

3.3. Eliminación de columnas y filas

En esta sección se van a destacar consideraciones importantes a tener en cuenta cuando se opera con una tabla, así como las decisiones que se toman a la hora de eliminar filas y columnas.

Antes de nada, es importante destacar los siguientes puntos:

- Cuanto mayor sea el número de filas, mayor precisión tendrán los resultados obtenidos.
- El aporte de información de una columna aislada de las demás es siempre ≥ 0.
- En una tabla de una BD hay columnas con mayor aporte de información que otras.
- Cuantas más columnas independientes introducimos, mayor es la probabilidad de introducir valores erróneos, provocando que la fila se convierta en errónea.

De estas afirmaciones podemos extraer lo siguiente:

- Hay un compromiso entre el aporte de información de una columna y la cantidad de valores nulos o inválidos que agrega dicha columna a la tabla.
- La adición de columnas puede disminuir el número de filas no erróneas.

Con objeto de obtener resultados precisos, hay interés en tener el mayor número de filas posibles, por ello, cuando se considere que una columna no aportará demasiada información, se eliminará de la tabla para evitar reducir el número de filas no erróneas a causa de esa columna.

Por otro lado, una vez eliminadas las columnas que aportan menor información, seguiremos teniendo valores nulos o erróneos. Una posible solución es completar los valores nulos. El problema de esta solución es con qué valor completamos los valores

nulos. Con cualquier valor que pongamos, estaremos introduciendo un sesgo. Otra solución más sencilla, apropiada para cuando hay gran cantidad de datos como es nuestro caso, es la eliminación de la fila entera cuando hay un valor nulo o erróneo en la fila[34]. Esta última solución es la que emplearemos salvo se indique lo contrario.

3.4. Conceptos para la limpieza de datos

En esta sección se explican 3 conceptos para el entendimiento y la interpretación de la limpieza de datos llevada a cabo.

3.4.1. Outliers

Los outliers o valores atípicos, son muestras que se encuentran distantes del resto. Estos valores alejados pueden tener un impacto elevado en los modelos creados o en las conclusiones extraídas[35]. Por ello es importante tener en cuenta la detección de estos valores y su eliminación cuando sea procedente. Otro tipo de outlier, aparte de los valores alejados, son lo que se conoce como error-outliers. Los error-outliers son valores que se han introducido por error y por tanto, son más sencillos de identificar porque carecen de sentido en la población estudiada [35]. Si por ejemplo, tenemos una variable que es la edad de los heridos y vemos muestras con el valor -5, es evidente que se trata de un error-outlier pues la edad de una persona no puede ser negativa.

Existen diversas formas de detectar outliers y no hay una verdad absoluta sobre cuando una instancia es un outlier. La detección y eliminación de outliers está entonces sujeta a interpretación e interés.

3.4.2. Visualización de la distribución de los datos

Observar la distribución de los datos suele ser de gran utilidad para la limpieza de los mismos. Representar la distribución de las variables permite, entre otras cosas, detectar outliers. Dos ejemplos de visualización de distribución son los Diagramas de Cajas y los Histogramas.

3.4.2.1. Diagramas de Cajas

Los Diagramas de Cajas permiten comparar distribuciones e identificar valores atípicos, representando los datos en cuartiles. En la Figura 3.2 se muestra un ejemplo de Diagrama de Cajas para facilitar la comprensión.

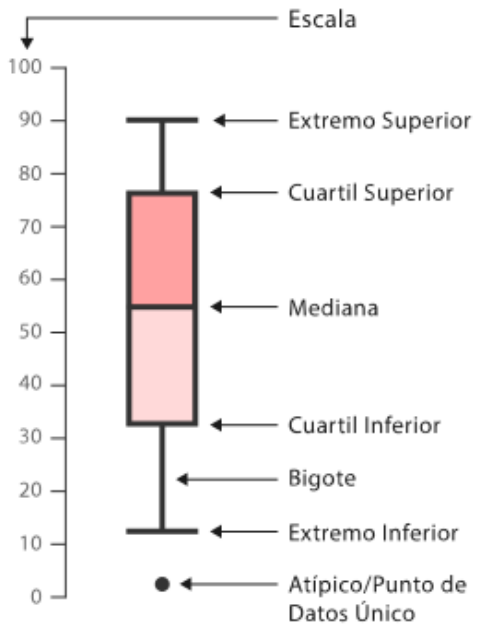


Figura 3.2: Zonas de un Diagrama de Caja [3]

- El 25 % de las instancias son menores o iguales al cuartil inferior.
- Los valores entre el origen y el segundo cuartil o mediana representan el 50 % de las instancias.
- Los valores entre el origen y el tercer cuartil o cuartil superior representan el 75 % de las instancias.

Como se puede deducir, en la caja se encuentran el 50 % de las instancias, la caja será más grande cuanto más dispersos estén los datos.

Los dos segmentos que continúan la caja por los dos extremos son los bigotes. Los bigotes marcan el límite entre valores con una distribución cercana y valores atípicos[36]. Los bigotes tienen una longitud máxima que debe cumplir la siguiente inecuación:

$$L_{bigote} \leq 1{,}5 \cdot RIC \tag{3.1}$$

Donde L_{bigote} es la longitud de cada bigote (ambos tienen que medir lo mismo) y RIC es el Rango Intercuartílico, que se define como la diferencia entre el valor del cuartil superior y del cuartil inferior[37].

3.4.2.2. Histograma

Un histograma es similar a un gráfico de barras, la diferencia principal es que en el histograma las barras son la agrupación de valores en rangos. La altura de un histograma representa el número de instancias que caen en ese rango[38]. Los histogramas nos muestran la distribución de los datos, pudiendo ver tendencias, concentración de datos o identificar valores muy alejados. En la Figura 3.3 se puede ver un ejemplo de histograma. En este histograma vemos que se agrupan los valores en rangos de una unidad.

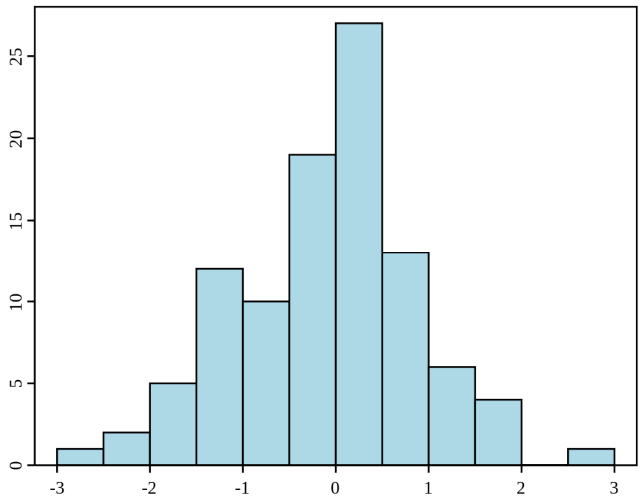


Figura 3.3: Ejemplo de Histograma [4]

3.5. Descripción tablas de interés

Si bien se ha dado una descripción genérica de todas las tablas presentes en la base de datos, en esta sección se detallarán más las tablas con las que se va a trabajar en el capítulo 4. Aunque no se mencione explícitamente, no se eliminará la columna INC_KEY de ninguna de las tablas que contenga dicha columna ya que esta servirá para unir tablas.

Antes de proceder con la descripción de tablas de interés, es importante destacar la existencia de múltiples entradas asociadas a una misma INC_KEY en varias tablas como se puede ver en la Figura 3.1. Para poder hacer uniones entre tablas y operar con varias tablas simultáneamente, es necesario que INC_KEY sea única en nuestras tablas ya que utilizaremos INC_KEY como clave para la unión de casi todas las tablas. Esto nos hace tener que prescindir de múltiples entradas en las tablas con varias entradas asociadas a una misma INC_KEY. En cada descripción de tabla se mencionará qué criterio se ha seguido para la reducción de instancias. Las únicas tablas que no se

unen utilizando la clave INC_KEY son las tablas con el sufijo "DES" que se unirán con el código asociado. Por ejemplo, las tablas `DCODE` y `DCODEDES` se unen mediante el código DCODE y las tablas `PCODE` y `PCODEDES` mediante el código PCODE.

3.5.1. Tabla de códigos de lesión AIS, AISPCODE

AIS (Abbreviated Injury Scale) es un sistema de codificación de lesiones basado en la anatomía. En la tabla `AISPCODE` se presenta cada código AIS descompuesto en las variables `SEVERITY` y `PREDOT`. Cada código AIS está compuesto por siete números, los seis primeros, previos al punto, reflejan las lesiones de los pacientes y el último la severidad. Los dígitos tienen un significado asociado. Siendo la estructura de un AISCODE "123456.7":

- 1 indica la región del cuerpo
- 2 el tipo de estructura anatómica
- 3 y 4 la estructura anatómica específica
- 5 y 6 el nivel de lesión
- 7 indica el nivel de severidad basado en el PREDOT.

Por ejemplo, una fractura de la diáfisis femoral puede ser codificada como 851814.3 siendo la severidad 3. [39]

En la Figura 3.4 se muestra el significado que tiene el primer dígito de los códigos AIS asociado a cada número que puede tomar así como las regiones que comprenden dicha zona.

AIS Numerical Descriptor	AIS Section Descriptor	Body Regions Included
1	Head	Cranium, brain
2	Face	Eye, ear, lips
3	Neck	Neck, throat
4	Thorax	Thoracic contents, including rib-cage
5	Abdomen/Pelvic Contents	Abdominal/pelvic organs
6	Spine	Spinal column/cord
7	Upper extremities	Upper limbs including shoulder
8	Lower extremities	Lower limbs including pelvis
9	External	Integumentary system, including burns

Figura 3.4: AISCODE - Primer dígito [5]

En la Figura 3.5 se representan las severidades encontradas en la tabla para cada zona AIS. Vemos que, a excepción de la cara y las extremidades inferiores, hay al menos una entrada de severidad 6 para cada una de las zonas. La escala AIS es una medida para lesiones individuales, actualmente no hay ninguna forma de combinar códigos AIS que sea popularmente aceptada. Esto supone un problema, pues

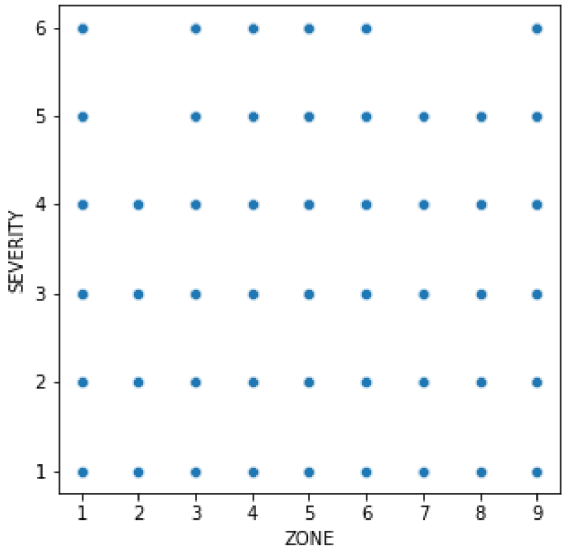


Figura 3.5: Distribución Severidad - Zona

si analizamos la tabla `AISPCODE` hay registros con múltiples entradas para un mismo incidente y no se especifica cómo ha influenciado cada entrada en el resto de tablas. Esto conlleva la siguiente problemática asociada, que es que sabemos los minutos que un paciente ha estado en el Departamento de Emergencias y las complicaciones que ha tenido, pero no sabemos qué lesión ha sido la causante o si ha sido una combinación de todas las lesiones.

En la Figura 3.6 se recogen cuántas lesiones tienen múltiples entradas en la tabla `AISPCODE` después de la eliminación de valores erróneos y nulos. Cómo se puede observar, la mayoría de incidentes son de 1, 2, 3 y 4 lesiones. Se decide que, emplear sólo los incidentes con una lesión no es suficiente ya que solo se estarían empleando el 30 % de los incidentes disponibles. Dada la distribución de la Figura 3.6 se decide incluir la lesión más severa de cada uno de los incidentes con dos lesiones. Esto es, si un incidente tiene dos lesiones, se supondrá que la entrada de mayor severidad es más significativa, asumiendo que es la lesión que causa la mayor parte de los datos asociados a dicho incidente. El resto de lesiones se descartan. Si bien gran parte de los incidentes tienen 3 lesiones, parece erróneo asumir que una lesión es siempre más significativa que las otras dos lesiones de un mismo incidente y dado que no hay forma de cuantificar cuánto influye cada una, se descartan todos los incidentes de 3 o más lesiones.

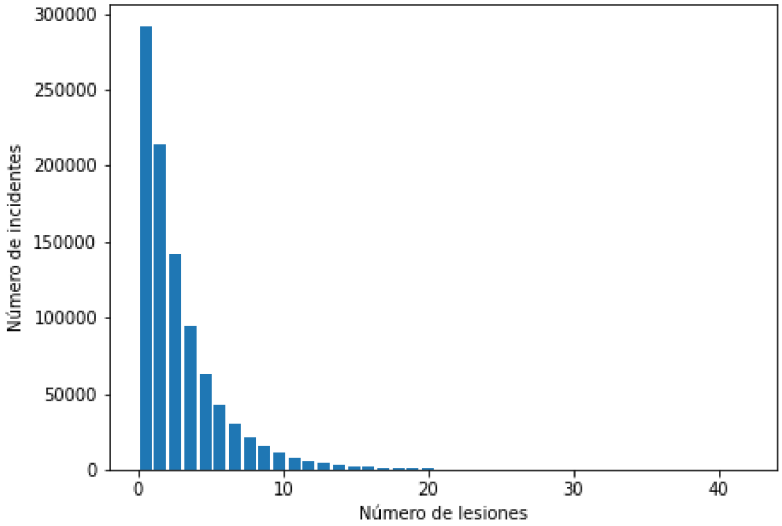


Figura 3.6: Número de lesiones por incidente

3.5.2. Tabla DEMO

La tabla `DEMO` incluye información demográfica del incidente.

Entre las columnas de esta tabla, se considera no influye en el desarrollo del traumatismo ninguno de los siguientes campos: `TEACHINGSTATUS, ETHNIC, RACE1, RACE2, STATEDESIGNATION, PEDIATRICSTATEDESIGNATION` por lo que se eliminan siguiendo el razonamiento explicado en la Sección 3.3.

Se han encontrado distintos resultados de trauma basados en el género. Si bien parte de la causa de estas diferencias encontradas se debe a las diferencias existentes en los patrones de lesión entre hombres y mujeres; varios estudios han confirmado una mayor supervivencia femenina ante traumas. Cabe destacar los cambios hormonales agudos que tienen lugar tras las lesiones y que pueden afectar a los resultados del trauma[40]. Debido a esto, parece conveniente mantener la columna `GENDER`.

El atributo `HOSPITALTYPE` toma los valores *Non-profit* y *For profit* por lo que, a pesar de las diferencias que puedan existir entre ambos tipos de hospital, se descarta al no tener información suficiente para distinguir entre hospitales puesto que no hay especificación de recursos, personal o políticas. Además, las posibles diferencias entre ambos tipos de hospitales no necesariamente son extrapolables a otros hospitales fuera de Estados Unidos.

El atributo `YOBIRTH` se elimina porque `AGE` nos aporta la misma información sin tener que realizar operaciones en la columna para trabajar con `AGE`.

En la columna `AGE`, hay 56720 entradas para el valor -99, de acuerdo con el manual, se pone el valor -99 cuando la edad es menor que 1 o mayor que 89. Estas entradas se eliminan. Quedando entonces la distribución de edades como se muestra en la Figura 3.7. El atributo `BEDSIZE` toma los valores $\leq$*200, 201 - 400,* $\geq$*600, 401 - 600* y se mantiene por si fuera de interés. Quedando las columnas:

INC_KEY	AGE	GENDER	BEDSIZE

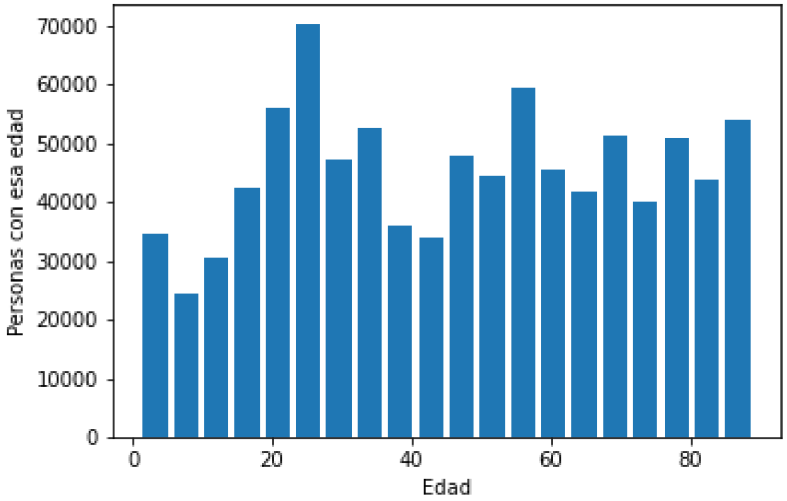


Figura 3.7: Distribución de Edades

3.5.3. Tabla ED

En esta tabla se encuentra información recolectada en el Departamento de Emergencias y contiene información como la edad del paciente, la ocupación, la altura o los minutos que pasa el paciente en el Departamento de Emergencias.

De una rápida búsqueda obtenemos: El promedio de altura y peso de un paciente de trauma en Estados Unidos es de 1.72 metros de alto y pesa 90 kilos; una mujer promedio mide 1.67 metros y pesa 77 kilos[41]. Tanto la altura como el peso parecen campos influyentes en el desarrollo de traumas y en la clasificación de pacientes. Para estudiar la distribución de la altura y del peso, podemos observar la Figura 3.8, donde se muestra un Diagrama de Cajas del peso y la altura. A partir del diagrama, se deciden eliminar las instancias con peso mayor a 170 kilos, 7128 instancias y las que corresponden a personas que tienen un peso inferior a 9 kilos, 5123 instancias.

En la Figura 3.9 observamos un histograma del peso antes de la eliminación de las 12251 entradas y la distribución con la que quedan los pesos tras la eliminación. Observamos a excepción del extremo izquierdo que la distribución tiende a ser normal. Dado que la distribución es casi normal, es adecuado eliminar las entradas cuyo peso esté a 3 desviaciones estándar de la media. Puesto que en una distribución normal el 99.7 % de los valores se encuentran a menos de 3 desviaciones estándar de la media[42]. Es una práctica habitual la eliminación de valores a más de 3 desviaciones estándar de la media cuando la distribución es similar a una normal. Con esto eliminamos otras 6692 filas, quedando 715930 filas pero habiendo reducido notablemente los outliers.

Por otro lado, teniendo en cuenta que la eliminación de outliers de la columna `WEIGHT` eliminó outliers de la columna `HEIGHT` y viendo la distribución, se decide

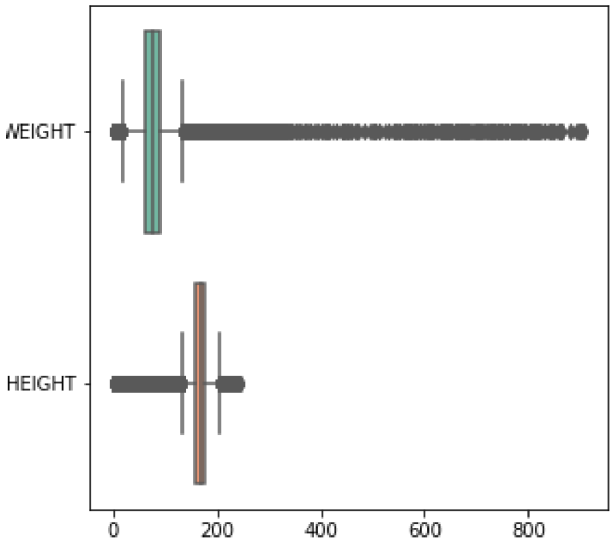


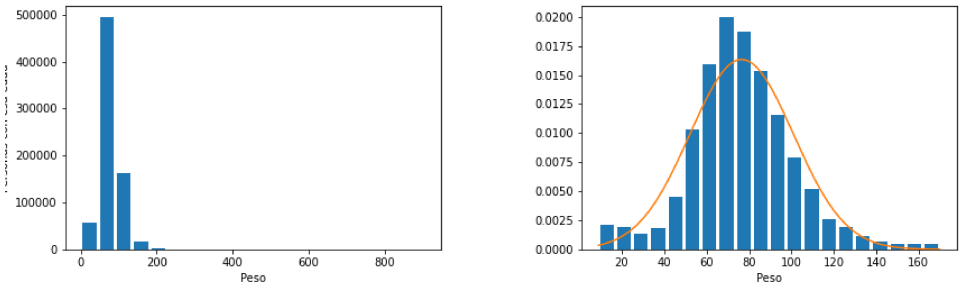
Figura 3.8: Diagrama de Caja del peso y la altura antes de la eliminación de entradas.

eliminar solo 4663 instancias, para alturas menores a 60 cm. Otra columna de interés es `EDDAYS`, que recoge los días que el paciente estuvo en ED. En la limpieza de esta columna, se eliminan todas las entradas con más de 14 días ya que solo se pierden 42 entradas pero nos libramos de outliers muy separados como 366 días o 1096 días.

El campo `ISSAIS` recoge el Injury Severity Score (ISS) asociado a cada incidente. El ISS tomará el máximo valor, 75, si alguna de las lesiones tiene severidad 6 ya que se considera intratable. Si ninguna de las lesiones del incidente tiene severidad 6, el ISS se calcula empleando la siguiente ecuación[43]:

$$IIS = S_1^2 + S_2^2 + S_3^2 \tag{3.2}$$

Donde S_1, S_2 y S_3 son las severidades de las 3 lesiones de mayor severidad asociadas a ese incidente.



(a) *(b)*

Figura 3.9: *(a)* Histograma de pesos antes de eliminación y *(b)* Distribución normal después de eliminación.

3.5.4. Tabla PROTDEV

La tabla `PROTDEV` indica el dispositivo de protección presente en el accidente. Esta información se recoge en la columna `PROTDEV`. Esta tabla está pensada para los accidentes de coche y por lo tanto el resto de incidentes tomarán el valor `None` en la columna `PROTDEV`.

Hay 730934 instancias con `PROTDEV` igual a *None*, es decir, sin protección en el momento del incidente. En la Figura 3.10 se muestra un histograma de los valores que puede tomar la columna `PROTDEV` cuando sí hay dispositivo de protección. Los valores del eje X se han puesto como números siguiendo la leyenda que se adjunta a continuación:

1. Protective Non-Clothing Gear (e.g., shin guard)
2. Airbag Present
3. Lap Belt
4. Helmet (e.g. bicycle, skiing, motorcycle)
5. Child Restraint (booster seat or child car seat)
6. Protective Clothing (e.g., padded leather pants)
7. Eye Protection
8. Shoulder Belt
9. Other
10. Personal Floatation Device

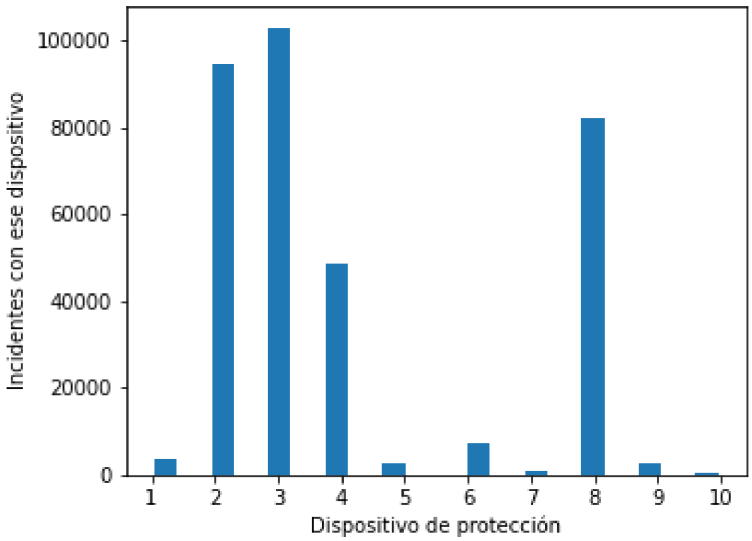


Figura 3.10: Histograma de los dispositivos de protección.

Las otras dos variables presentes en la tabla son CHILDRES, que describe el tipo de silla de coche del niño, y AIRBAG, que describe los airbags que saltaron con el accidente. Las variables CHILDRES y AIRBAG sólo toman valores válidos para las entradas con PROTDEV=*Child Restraint (booster seat or child car seat)* y PROTDEV=Airbag Present por lo que no existen instancias que contengan valores válidos en las 3 columnas simultáneamente. Para solventar la problemática expuesta, se generarán 3 tablas a partir de PROTDEV:

- PROTDEV_DEV contendrá únicamente INC_KEY y PROTDEV, 344852 instancias.
- PROTDEV_AB contendrá únicamente INC_KEY y AIRBAG, 29496 instancias.
- PROTDEV_CH contendrá únicamente INC_KEY y CHILDRES, 2426 instancias.

En las dos últimas tablas se podría mantener la columna PROTDEV pero dado que siempre toma el mismo valor, *Child Restraint (booster seat or child car seat)* y PROTDEV=Airbag Present respectivamente, no tiene sentido trabajar con esa columna.

3.5.5. Tablas TBI_CM y TBI_GSC_Q

Estas dos tablas se comentan juntas por su similitud en la estructura y en el tratamiento de las mismas. Ambas tablas están compuestas por, además de INC_KEY, dos columnas.

La tabla TBI_CM contiene información sobre los dispositivos empleados para la monitorización cerebral del paciente, esta información se almacena en la columna TBI_CEREBRAL_MONITOR_DESC y en la otra columna se incluye un código asociado a esta descripción. La tabla TBI_GSC_Q contiene información sobre medidas de cuidado realizadas al paciente como puede ser sedar o paralizar al paciente, esta información se recoge en la columna TBI_GCS_QUALIFIER_DESC y en la otra columna se incluye un código asociado a esta descripción.

De TBI_CM nos quedamos con TBI_CEREBRAL_MONITOR_DESC y de TBI_GSC_Q nos quedamos con TBI_GCS_QUALIFIER_DESC ya que en ambas tablas no tiene sentido utilizar la columna de códigos, ya que aunque sean números, en realidad son valores categóricos y serían menos descriptivos que las propias descripciones.

3.5.6. Tabla TRANSPORT

Esta tabla recopila información sobre el medio de transporte utilizado para el desplazamiento del paciente. En cuanto a la composición de la tabla, a parte de INC_KEY, contiene dos columnas: TRANTYPE y TMODE. La primera indica si se trata del transporte primario utilizado o de otro transporte. La segunda indica el modo de transporte pudiendo tomar los siguientes valores:

- Helicopter Ambulance

- Ground Ambulance
- Private/Public Vehicle/Walk-in
- Police
- Fixed-wing Ambulance
- Other

Para cada incidente puede haber múltiples entradas tal y como se ha explicado anteriormente. Para solucionar este inconveniente, nos quedaremos únicamente con las entradas cuyo valor de `TRANTYPE` es igual a *Primary* al ser el transporte primario utilizado y por tanto el más importante. Según[40], el transporte inicial de pacientes con traumas a centros no traumatológicos estaba relacionado con un aumento del 30 % en la tasa de mortalidad durante las primeras 48 horas después del incidente. Esta tabla nos permitirá estudiar la influencia del medio de transporte utilizado.

3.5.7. Tablas COMORBID y COMPLIC

La tabla `COMORBID` recoge información sobre las condiciones comórbidas del paciente en la llegada al Hospital. La tabla `COMPLIC` contiene información sobre las complicaciones surgidas durante el tratamiento del paciente.

Estas tablas contienen múltiples entradas por incidente y también tienen una estructura de 3 columnas, siendo la primera `INC_KEY`, la segunda el contenido mencionado en el párrafo anterior y la última códigos asociados a las descripciones de la segunda tabla. Analizando las complicaciones, no se puede realizar un ranking que nos permita priorizar ciertas complicaciones respecto a otras, por ello sólo se emplearán las instancias con una única complicación. En la tabla `COMORBID` tampoco se puede afirmar cuál de las enfermedades coexistentes puede tener más importancia en el desarrollo del trauma, por ello se sigue la misma estrategia que en `COMPLIC`. Las columnas `COMORKEY` y `COMPLKEY` son los códigos asociados a `COMORDES` y `COMPLICDES` respectivamente. Siguiendo el razonamiento de la subsección 3.5.5 se eliminan.

Se adjuntan a continuación 5 de los 25 valores posibles de `COMORDES`:

- Hypertension requiring medication
- Diabetes mellitus
- Drug use disorder
- Current smoker
- Alcohol Use Disorder

Se adjuntan a continuación 5 de las 22 posibles complicaciones recogidas en `COMPLICDES`:

- Unplanned admission to the ICU
- Unplanned intubation

- Drug or alcohol withdrawal syndrome
- Pulmonary embolism
- Cardiac arrest with CPR

No se ha encontrado una clasificación de la influencia que tienen estas complicaciones o condiciones comórbidas en el desarrollo de traumas. La generación de un modelo teniendo en cuenta estos campos podría darnos una idea del peso que tiene cada valor en el desarrollo de traumas.

3.5.8. Tabla DISCHARGE

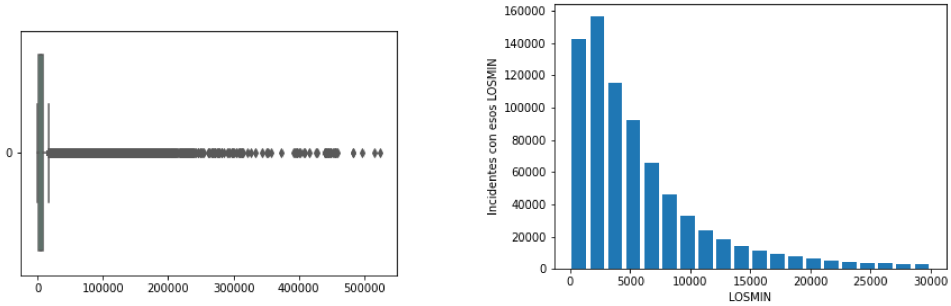
En esta tabla se recoge información de los tiempos que ha pasado el paciente en el hospital, así como el año de alta médica, la transferencia del paciente y el método de pago, indicando si el pago lo cubre un seguro médico, la persona...

En esta tabla podemos prescindir de la columna que refleja el año en el que se da el alta al paciente, `YODISCH`, al tomar siempre el mismo valor, que es 2016, y de la columna `Payment` al no ser relevante para el estudio de traumas.

En esta tabla se incluye el tiempo que pasa el paciente en el hospital expresado en minutos y en días en las variables `LOSMIN` y `LOSDAYS` respectivamente; los días que el paciente ha utilizado un ventilador mecánico en la variable `VENTDAYS` y los días que ha pasado en la unidad de cuidados intensivos `ICUDAYS`. También se incluye en esta tabla la variable `HOSPDISP` que determina la disposición del paciente en el alta hospitalario, esta variable toma, entre otros valores, *Dado de alta a casa, Transferido a un hospital de cuidado, Fallecido.*

Las columnas `LOSMINS, LOSDAYS, ICUDAYS` y `VENTDAYS` están correlacionadas por lo que para los análisis sólo se utilizará `LOSMINS`, evitando perder instancias por la adición de valores nulos de las otras columnas, y en caso de encontrarse otras relaciones con `LOSMINS`, se pasará a confirmar que ocurren con el resto de variables. En la Figura 3.11*(a)* se puede ver un Diagrama de Cajas antes de la limpieza, utilizado para detectar outliers. Solo hay 1424 instancias con más de 100000 minutos, lo que equivale aproximadamente a 69 días, por eso se muestra discontinua esa parte. La eliminación de outliers se realiza mediante prueba y error viendo la distribución y el número de instancias que iba quedando al eliminar instancias, se decide eliminar aquellas entradas cuya duración total de estancia en el hospital es mayor a 30000 minutos, lo que expresado en días sería alrededor de 21 días. En total se eliminaron 25796 instancias, quedando 765906 instancias.

En la Figura 3.11*(b)*, se muestra un histograma de la duración de la estancia de los pacientes en el hospital en minutos. Si bien el número de instancias a partir de 15000 minutos es claramente menor al número de instancias con pocos minutos, no parece razonable eliminar estos datos porque al generar un modelo predictor del tiempo en el hospital, el modelo no sería capaz de predecir correctamente estas instancias.



(a) *(b)*

Figura 3.11: *(a)* Diagrama de caja de `LOSMIN` antes de eliminación y *(b)* Histograma de `LOSMIN` después de eliminación.

3.5.9. Utilización de instancias

A lo largo de esta sección se ha descrito la limpieza realizada en las diferentes tablas a utilizar. Esta limpieza ha provocado la eliminación de múltiples instancias por tabla. A continuación se adjunta una gráfica que muestra el porcentaje efectivo de instancias de cada tabla que vamos a emplear en los análisis. Hay que tener presente este número de instancias corresponde a cada tabla aislada. Evidentemente, el número de instancias se reducirá al combinar distintas tablas.

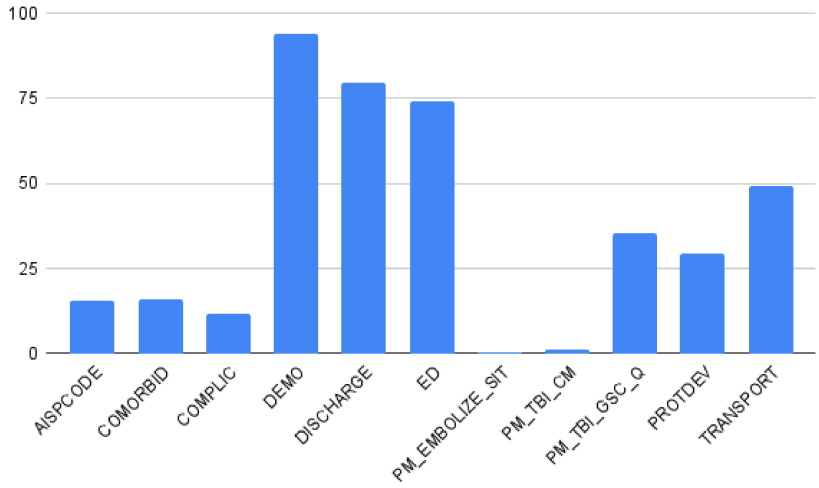


Figura 3.12: Porcentaje de utilización de las instancias de cada tabla.

Capítulo 4

Desarrollo

En este Capítulo se explicarán las herramientas empleadas para el desarrollo de este Trabajo Fin de Grado. Después se explicarán las nociones básicas necesarias para el entendimiento de los análisis realizados y finalmente se mostrarán los resultados y conclusiones obtenidos de los análisis. Para analizar las relaciones entre tablas se utilizará la Correlación Lineal de Pearson y para intentar reducir las variables se utilizará PCA y Random Forest. Adicionalmente se generarán 6 árboles de decisión para intentar predecir el tiempo que pasa un paciente en el Departamento de Emergencias (ED).

4.1. Herramientas para el desarrollo

En esta sección se describen las herramientas software empleadas en este Trabajo Fin de Grado. El tratamiento de datos se realiza mediante cuadernos Jupyter, empleando la interfaz de usuario JupyterLab. JupyterLab es un entorno de desarrollo interactivo basado en tecnología web y es una herramienta popularmente utilizada para ciencia de datos y aprendizaje automático[44]. Nos permite el tratamiento, la representación y el análisis de los datos.

Python es un lenguaje de programación interpretado cuyo código es especialmente legible, como consiguiente, es fácil su aprendizaje, uso y lectura[45]. Para el tratamiento de los datos se ha empleado NumPy, Pandas y Scikit-learn, para la representación de los datos se ha empleado Matplotlib y Seaborn. Pandas es una herramienta flexible, potente y rápida para manipulación y análisis de datos construida sobre Python[46]. NumPy es una biblioteca Python con una gran colección de funciones matemáticas de alto nivel para operar sobre matrices multidimensionales[47]. Scikit-learn es una biblioteca que proporciona herramientas para el análisis predictivo[48]. Seaborn es una biblioteca para la visualización de datos basada en Matplotlib, proporciona una interfaz de alto nivel para dibujar gráficos estadísticos[49]. Matplotlib es una biblioteca que permite la creación de visualizaciones estáticas, animadas e interactivas en Python[50].

4.2. Variables categóricas

En esta sección se menciona el concepto de variables categóricas, se especifica cómo se han tratado las variables categóricas presentes en la BD y por qué es necesaria su conversión a variables numéricas. Las variables categóricas son variables no numéricas, almacenadas típicamente como valores textuales. Dado que muchos algoritmos no son capaces de soportar este tipo de datos, es necesario hacer una conversión de estos atributos a valores numéricos. Para este procedimiento no hay una respuesta única, destacan las siguientes dos formas de conversión:

- **Label Encoding**: Convertir cada valor de la columna categórica a un valor numérico.
- **One Hot Encoding**: Convertir cada valor categórico a una nueva columna y asignar un 1 cuando la fila original contenga ese valor y un 0 cuando no.

Label Encoding tiene la ventaja de mantener los datos de forma compacta y no aumentar el número de columnas, pero los valores numéricos pueden ser malinterpretados por los algoritmos[51]. Tomemos la columna que habla sobre los dispositivos de protección, columna `PROTDEV`, de la tabla que proporciona información sobre estos dispositivos, `PROTDEV`, para explicar esta problemática. A continuación se muestran algunos de los 11 valores que puede tomar la columna `PROTDEV`:

1. Sin protección
2. Presencia de Airbag
3. Cinturón de hombro
4. Protección de ojos

Si convertimos, por ejemplo, cada valor en su posición en la lista estaríamos afirmando que la protección de ojos protege 4 veces más que no tener protección, 2 veces más que la presencia de Airbag y 1.3 veces más que el cinturón. Esto evidentemente no es así y por ello esta forma de conversión no es adecuada para esta columna. En cambio, usando One Hot Encoding al generar columnas independientes no estamos asignando pesos a cada posible valor y aunque se generan más columnas no generaría tanta falsa información como Label Encoding[51].

Label Encoding no resultó apropiado para ninguna de las columnas con valores categóricos estudiadas por no poder establecer pesos coherentes para los valores. Por este motivo, todas las variables categóricas se convirtieron a variables numéricas utilizando One Hot Encoding.

En la Tabla 4.2, se muestra un ejemplo de One Hot Encoding para facilitar el entendimiento. Podemos ver como cada posible valor de la columna `PROTDEV` pasa a ser una nueva columna y la columna `PROTDEV` se elimina de la tabla. [1]

[1] Dado que la variable `PROTDEV` tiene 11 valores distintos, debería haber 11 columnas pero solo se representan 4 por simplicidad.

PROTDEV	Presencia de Airbag	Cinturón de hombro	Protección de ojos	Sin protección
Presencia de Airbag	1	0	0	0
Cinturón de hombro	0	1	0	0
Protección de ojos	0	0	1	0
Sin protección	0	0	0	1
Protección de ojos	0	0	1	0

Tabla 4.1: Ejemplo de conversión de variables categóricas a numéricas usando One Hot Encoding

4.3. Correlación de Pearson

La correlación de Pearson puede aparecer como ρ_{XY} si se ha calculado para toda la población o como r_{XY} si se ha calculado sobre una muestra[52]. A continuación se adjunta la ecuación para la población:

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{Cov(X,Y)}{\sqrt[2]{Var(X)Var(Y)}} \tag{4.1}$$

σ_{XY} es la covarianza de X,Y. σ_X es la desviación estándar de la variable X y σ_Y es la desviación estándar de la variable Y.

El coeficiente de correlación de Pearson toma valores en el intervalo [-1,1]. Cuanto mayor sea el valor absoluto del coeficiente, más fuerte es la correlación. Si el coeficiente toma un valor positivo, existe una relación directa entre las variables, es decir, cuando una aumenta la otra aumenta. Si el coeficiente toma un valor negativo, la relación es inversa y cuando una aumenta la otra disminuye.

4.3.1. Matriz de correlación

Cuando se estudian las correlaciones entre las variables de una tabla, es habitual realizar una matriz de correlación para visualizar todas las correlaciones simultáneamente.

Una matriz de correlación es una tabla de 2 dimensiones que muestra una misma lista de variables horizontal y verticalmente. En esta tipo de tabla, la celda de intersección entre cada columna y fila indica la correlación de las variables al inicio de la columna y fila. Esto hace que siempre se genere una diagonal de unos ya que la correlación de una variable consigo misma es siempre 1.

	A	B	C
A	1	0.2	-0.8
B	0.2	1	0.5
C	-0.8	0.5	1

Tabla 4.2: Ejemplo de Matriz de Correlación.

En la Tabla 4.2 se adjunta un ejemplo de matriz de correlación, vemos en este caso la correlación entre C y A es -0.8, por lo que la relación entre C y B es elevada e inversa.

4.4. Análisis de componentes principales (PCA)

El análisis de componentes principales (PCA) es un procedimiento estadístico que permite comprimir la información de grandes tablas de datos en algo que captura la esencia de los datos originales empleando un número menor de variables, lo que facilita su análisis y visualización. Se basa en la proyección de las variables a un subespacio de menor dimensión. PCA trata de encontrar una forma significativa de reducir la dimensionalidad centrándose en las diferencias entre las instancias[53].

En una muestra con n instancias de s variables PCA permite encontrar z factores subyacentes, siendo $z < s$, capaces de explicar aproximadamente lo mismo que las s variables originales. Cada una de las nuevas z variables se denomina componente principal (PC)[6].

Para explicar cómo funciona PCA se va a emplear un ejemplo. Supongamos que tenemos una tabla con n instancias y cada instancia tiene 2 variables, X_1 y X_2. En la Figura 4.1 se representan todas las instancias en términos de estas dos variables. La línea roja define la primera componente principal (PC1), la primera componente principal marca la dirección en la que los datos muestran mayor varianza[6]. La línea verde define la segunda componente principal, la segunda componente principal marca la dirección en la que los componentes muestran mayor varianza sin estar correlacionada con la primera componente, esto implica que las componentes principales son perpendiculares entre sí[6].

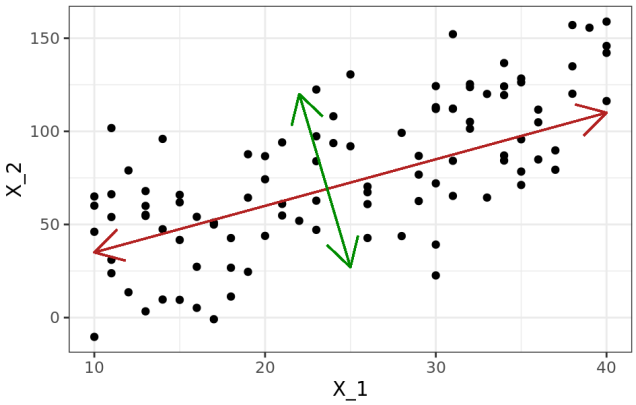


Figura 4.1: Ejemplo de PCA con 2 variables [6]

A continuación se explica cómo se calcula cada PC:

Para un grupo de variables $(X_1, X_2, \ldots, X_p)$ cada PC se obtiene como combinación

lineal de las variables originales empleando la ecuación 4.2

$$PCi = \sum_{j=1}^{p} \phi_{ij} \cdot X_{ij} \tag{4.2}$$

Los términos ϕ_{ij} se denominan loadings y se pueden entender como el peso que tiene cada una de las variables originales en PC. Para el cálculo de PC_1 se resuelve un problema de optimización para identificar los loadings que maximizan la varianza y una vez calculado para el primer componente se repite para el resto siguiendo el mismo procedimiento pero además forzando que se cumpla la condición de perpendicularidad entre el nuevo PC y los demás PCs ya calculados[6].

Es importante destacar que antes de calcular los componentes principales, se deben escalar las variables para que su desviación estándar sea 1 y su media sea 0[53]. De no aplicar este escalado las variables de mayor escala tendrán mayor influencia, dando pie a una falsa interpretación de los resultados.

4.5. Modelos de predicción

En esta sección se describen conceptos importantes sobre modelos de predicción y los dos modelos utilizados. Cuando se elige un modelo de predicción hay que tener en cuenta lo que se quiere predecir para ver si el modelo a utilizar ha de ser de regresión o de clasificación.

Los modelos de clasificación toman como resultado una clase entre un número limitado de clases. Por ejemplo, un modelo capaz de predecir si un cliente devolverá el crédito sólo puede tomar un valor de las clases *Sí* y *No*. En cambio se tratará de regresión cuando el resultado sea un número dentro de un conjunto infinito de posibles resultados[54]. Dado que nosotros queremos entrenar un modelo capaz de decir los días en ED, se utilizarán modelos de regresión.

Dos conceptos importantes son el overfitting y el underfitting. El overfitting ocurre cuando el modelo aprende demasiado los detalles de los datos de entrenamiento, haciendo que obtenga muy malos resultados cuando se introduce en el modelo datos que no había visto antes. El underfitting ocurre cuando el modelo no es capaz de extraer generalizaciones de los datos de entrenamiento ni modelar estos datos[55].

En este Trabajo Fin de Grado se entrenarán árboles de decisión y Random forests.

4.5.1. Árboles de decisión

Para la construcción de árboles de decisión se emplea la biblioteca scikit-learn, esta biblioteca utiliza una versión optimizada del algoritmo CART[56].

Los árboles de decisión ramifican sus nodos a partir de la impureza. Al configurar el árbol se emplea la impureza Gini, que mide las divergencias entre las distribuciones

de los valores del atributo a predecir y divide el nodo en dos nodos empleando la condición que proporcione la menor cantidad de impureza posible[57].

El algoritmo general para la construcción de árboles de decisión es el siguiente:

1. Seleccionar condición.

2. Creación de ramas para cuando se cumple y cuando no se cumple la condición.

3. Repetir 1 y 2 en cada una de las ramas obtenidas con otra condición.

Si todos los elementos en cada rama de un nodo tienen la misma predicción se trata de una rama pura y no se sigue ramificando dicho nodo. También se puede configurar la variable *min_samples_split* para fijar el mínimo de instancias que debe haber en un nodo para ramificar, esta variable se deja en su valor por defecto, 2. Otra forma de limitar la ramificación es mediante el parámetro *max_depth* que fija la máxima profundidad del árbol[58]. La función utilizada para medir la calidad de una ramificación es el Error Cuadrático Medio (ECM).

Siendo Y un vector con todos los valores que deberíamos predecir y Y' es un vector con todas las predicciones realizadas, el error cuadrático medio se calcula empleando la ecuación 4.3.

$$ECM = \frac{1}{n}\sum_{i=1}^{n}(Y_i' - Y_i)^2 \tag{4.3}$$

Donde n es el número total de instancias.

En la Figura 4.2 se muestra un árbol de decisión de Profundidad 2 a modo de ejemplo. MSE son las siglas de Error Cuadrático Medio en inglés.

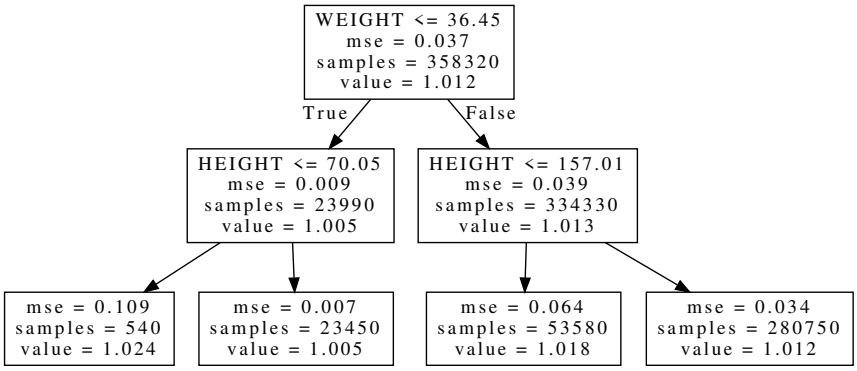


Figura 4.2: Ejemplo de árbol de decisión de Profundidad 2

4.5.2. Random Forest

Los Random Forest o Bosques Aleatorios son una combinación de árboles de decisión predictores. El valor que se devuelve como predicción, es la media de las predicciones de cada árbol. La combinación de árboles en la que se basa Random Forest disminuye el overfitting producido por un único árbol decisor[59].

Para este modelo se hace uso del método *RandomForestRegressor()* de scikit-learn. Este método permite utilizar Bootstrap aggregation y utilizar muestras Out-of-bag para estimar el error[60].

Bootstrap aggregation utiliza muestreo aleatorio con repetición, es decir, de los datos originales se extrae un subconjunto de muestras para generar el árbol de manera aleatoria. Las muestras que se han utilizado para entrenar un árbol se mantienen en el conjunto de datos originales, pudiendo ser utilizadas para entrenar otros árboles[61].

El error Out-of-bag (OOB) es la media de los errores de predicción en cada muestra de entrenamiento al utilizarse para test en los árboles que no utilizaron esa muestra para entrenamiento[62].

4.6. Resultados Correlaciones

En esta sección se recogen las diferentes correlaciones realizadas justificando por qué se hicieron y exponiendo los resultados obtenidos.

Dada la gran cantidad de tablas e interconexiones de tablas, se pueden realizar diversas combinaciones. De cara a obtener resultados útiles, se decide centrar prioritariamente el análisis en la influencia de la severidad de las lesiones en el resto de tablas. Esto es, la mayoría de los análisis incluirán la tabla `AISPCODE` descrita previamente en la subsección 3.5.1.

Otro aspecto importante a tener en cuenta es la problemática explicada en la subsección 3.3 sobre la adición de valores nulos al agregar columnas. Esto es una complicación a tener en cuenta al unir tablas, ya que la unión de dos tablas no deja de ser la adición de las columnas de una tabla a la otra tabla.

Inicialmente se intentaron unir las tablas `AISPCODE`, `TRANSPORT` y `DISCHARGE` para ver cómo la severidad que se obtiene de la tabla `AISPCODE` y el transporte primario utilizado que se extrae de la tabla `TRANSPORT` influyen en el tiempo que un paciente pasaría en el hospital, información que se obtiene de la tabla `DISCHARGE`. Esta unión redujo excesivamente el número de instancias. Por ello se decide que, para estudiar las correlaciones, es mejor estudiar las tablas de dos en dos para tener un mayor número de instancias y que los resultados sean más fiables.

A continuación se exponen las correlaciones realizadas y el motivo por el que se realizan.

- `AISPCODE` y `COMPLIC`: Con la unión de estas dos tablas se pretende estudiar cómo influye la severidad de una lesión en las complicaciones que surgen.
- `COMORBID` y `COMPLIC`: Estas dos tablas se unen para ver la relación que tienen las condiciones comórbidas del paciente y las complicaciones que surgen.
- `AISPCODE` y `DEMO`: Se estudian estas dos tablas para ver si existe relación entre la severidad, el género, la edad y el tamaño de cama asignado.
- `AISPCODE` y `DISCHARGE`: Se estudia la correlación entre las variables de estas dos tablas para ver cómo influye la severidad de la lesión en el tiempo que pasa el

paciente en el hospital y la disposición del paciente al abandonar el hospital, es decir, el campo `HOSPDISP`.

- `AISPCODE` y `ED`: Se estudia la correlación entre estas dos tablas para ver la influencia que tiene la severidad de la lesión en la información recogida en el Departamento de Emergencias.

- `AISPCODE` y `TRANSPORT`: Con el estudio de la relación entre la severidad y el medio de transporte primario utilizado, se intenta descubrir si existe una tendencia a utilizar un medio u otro dependiendo de la severidad de la lesión.

- `AISPCODE` y `PROTDEV`: Se combinan estas dos tablas para intentar extraer conclusiones sobre la influencia de los dispositivos de protección en la severidad de las lesiones en los accidentes de coche. Desgraciadamente en la tabla `PROTDEV` no se recogen otros datos de interés del accidente tales como la velocidad del vehículo del accidente o una clasificación en función del impacto. Tal y como se comentó en la subsección 3.5.4 la tabla `PROTDEV` se descompone en 3 tablas, por lo que se realizan 3 matrices de correlación distintas para estudiar la influencia de los dispositivos de protección, la influencia del airbag y la influencia de la silla de niño en la severidad.

- `AISPCODE` y `EMBOLIZE_SITE`: La unión de estas dos tablas se pretendía realizar para ver si existe algún tipo de relación entre la severidad del paciente y la producción de una embolia en las diferentes zonas recogidas en esta tabla. Cabe destacar los pocos incidentes en los que se produjo una embolia, ya que de los 748189 incidentes recogidos en la tabla `EMBOLIZE_SITE`, solo 2449 personas habían sufrido una embolia por lo que aunque se estudia la correlación entre estas dos tablas, hay que tener presente la baja fiabilidad de los resultados debido al reducido número de instancias.

- `AISPCODE` y `TBI_CM`: Se estudia la relación entre la severidad de la lesión y el dispositivo empleado para la monitorización cerebral, para tratar de ver si la severidad de la lesión condiciona el dispositivo a utilizar para la monitorización cerebral. Encontrar relaciones entre estas dos variables podría ser relevante para decidir a qué hospital debe ir un paciente en función de la severidad de la lesión y la disponibilidad de equipos de los hospitales de las proximidades.

- `AISPCODE` y `TBI_GSC_Q`: Con la unión de estas dos tablas se pretende encontrar una relación entre la severidad de la lesión y las medidas de cuidado a realizar al paciente.

De los estudios realizados sobre las uniones comentadas no se han logrado obtener grandes correlaciones, por lo que podemos afirmar que no existe relación lineal o proporcionalidad entre las variables estadísticas estudiadas a excepción de las evidentes, como son; el peso y la altura, la severidad o la puntuación de la gravedad de lesión (ISS).

A modo de ejemplo se presenta la Figura 4.3 donde se ve un mapa de calor con los resultados de las distintas correlaciones entre variables de las tablas `AISPCODE` y `DEMO`.

Como se puede ver en la barra de la derecha, a medida que aumenta la saturación de color, aumenta el valor absoluto del coeficiente de correlación. En rojo se representan las variables cuya relación es inversa y en azul las variables cuya correlación es directa. Adicionalmente, para facilitar la interpretación del mapa de calor, el tamaño del cuadrado es proporcional al módulo del coeficiente de correlación.

Como se comentó en 4.3.1, la diagonal de unos no tiene significado. Vemos que no hay grandes correlaciones en el resto de la matriz, es decir, el módulo de los coeficientes de correlación es bajo. Las variables `Is_a_Male` y `Is_a_Female` son las dos columnas resultado de la conversión de la variable categórica `GENDER` a variables numéricas, es decir, no tiene sentido interpretar la correlación entre estas dos variables. Lo mismo sucede con las variables con el prefijo "bedsize". La variable n_INSTANCES indica el número de lesiones del paciente.

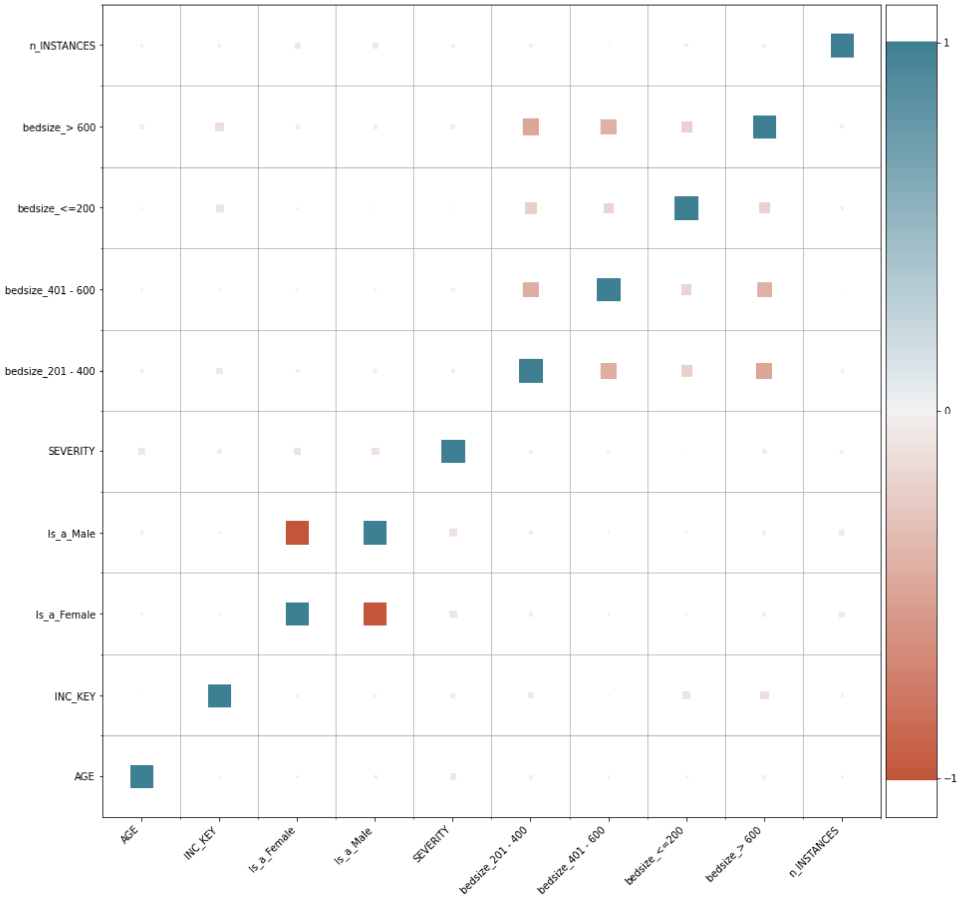


Figura 4.3: Correlación de Pearson de las tablas AISPCODE y DEMO

Para poder continuar con los análisis, se decide emplear PCA para intentar reducir la complejidad expresando los datos en nuevas variables no correlacionadas.

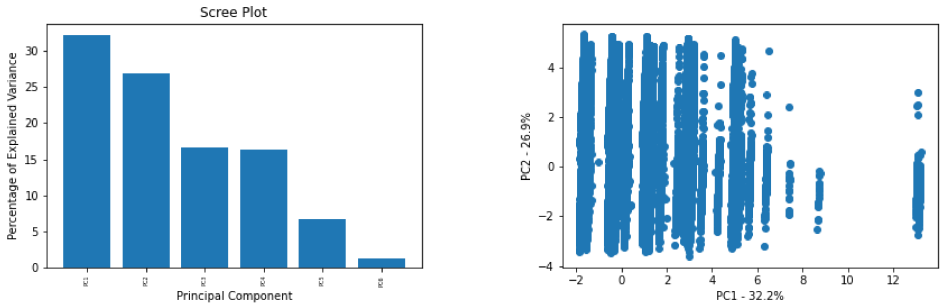
4.7. Resultados PCA

En la BD hay bastantes variables categóricas que hemos tratado de utilizar aplicando One Hot Encoding pero esta conversión genera columnas con datos binarios. PCA trata de minimizar la varianza. El concepto de varianza no tiene demasiado sentido con variables binarias[63]. Por este motivo emplearemos PCA sin tener en cuenta las variables categóricas, evitando que estas nos provoquen sacar conclusiones erróneas.

En todas las tablas estudiadas, salvo en `ED`, hay menos de 2 variables numéricas por lo que carece de sentido aplicar PCA en estas tablas. Se aplica PCA a la unión de `AISPCODE` y `ED`, y se representa el porcentaje de varianza que representa cada PC en la Figura 4.4*(a)*. En esta Figura, se ve que todos los PC salvo los dos últimos tienen un aporte significativo en comparación con el inmediatamente anterior.

En la Figura 4.4*(b)* se representa la proyección de instancias sobre PC1 y PC2, y por lo tanto, estamos representando la variación de los datos con un 58.1 % de la varianza total por lo que no es una gráfica muy representativa pero sí puede ser útil para ver clusters, tendencias o outliers. Destacan las incidencias situadas más a la derecha y podrían ser outliers. Viendo los loadings de PC1 en la Tabla 4.3 podemos deducir que esas instancias son los incidencias con puntuación de la gravedad de lesión `(ISSAIS) = 75`, es decir, las de severidad 6 y por lo tanto no son outliers.

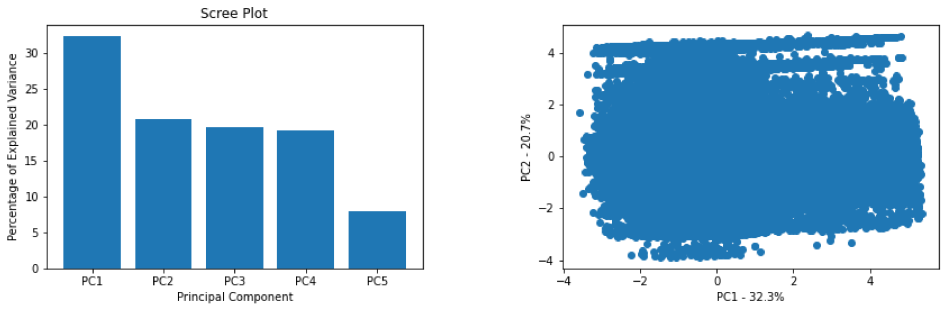
En la misma tabla podemos ver que el peso de `ISSAIS` y `SEVERITY` es muy elevado con respecto a las demás entradas, luego estas variables son las que están marcando la posición de los puntos. `SEVERITY` e `ISSAIS` están altamente correlacionadas (de hecho, como ya hemos visto, `ISSAIS` se calcula a partir de `SEVERITY`), esto causa que cuando una instancia está muy alejada en `SEVERITY` está también muy alejada en `ISSAIS` lo que provoca que la mayor varianza la marquen estas dos variables simultáneamente. Se decide entonces eliminar `ISSAIS`.



(a) *(b)*

Figura 4.4: *(a)* Scree Plot con ISS y *(b)*Proyección sobre PC1 y PC2 con ISS .

Adicionalmente, se sustituye el tiempo en el departamento de emergencias expresado en días (`EDDAYS`) por el tiempo expresado en minutos (`EDMIN`) para ganar precisión. Repetimos entonces PCA y obtenemos los resultados de la Figura 4.5. Como podemos ver, los puntos están todos más concentrados. Esto se debe a que el peso de `WEIGHT` y `HEIGHT` son los que más influyen en la distribución, ya que son los valores con mayor aporte según la Tabla 4.3.



(a) *(b)*

Figura 4.5: *(a)* Scree Plot sin ISS y *(b)*Proyección sobre PC1 y PC2 sin ISS .

No podemos extraer grandes conclusiones de usar PCA, sólo podemos prescindir de `n_INSTANCES`. No descartamos `EDMIN` porque nos interesa generar un modelo capaz de estimar esta variable a partir de las demás.

LOADINGS DE PC1			
CON ISS		SIN ISS	
ISSAIS	0.706352	WEIGHT	-0.698164
SEVERITY	0.703771	HEIGHT	-0.696386
n_INSTANCES	0.068330	n_INSTANCES	-0.151102
WEIGHT	-0.032485	EDMIN	-0.054254
EDDAYS	0.006925	SEVERITY	0.042880
HEIGHT	-0.000963		

Tabla 4.3: Loadings de PC1 al aplicar PCA a la unión de ED y AISPCODE

4.8. Resultados Árboles de decisión

Se quiere predecir el tiempo que pasa el paciente en el servicio de urgencias. Es importante destacar la limitación de equipos, camas y personal existentes en los hospitales. Saber con antelación el tiempo que requerirá cada paciente en el departamento de emergencias (ED), tendría beneficios como la optimizar del personal y los equipos necesarios; o el conocimiento de en qué hospital estará disponible antes un respirador. Dado que no existen relaciones lineales entre la variable que se quiere poder predecir y el resto de columnas, una posible solución es la utilización de un modelo que pueda tratar nativamente relaciones no lineales. Se decide utilizar árboles de decisión.

Se han empleado árboles de decisión, con distintas combinaciones de tablas, para tratar de predecir el tiempo que pasará un herido en el departamento de emergencias. En la Tabla 4.4 se recoge información de interés sobre las 3 uniones realizadas. Como se puede observar, inicialmente solo se une la tabla `ED` con la tabla `AISPCODE`. En esta primera unión se trata de poder predecir la variable del tiempo que pasa el paciente en el departamento de emergencias en días, `EDDAYS`, en función del resto de variables de `ED` y de la variable `SEVERITY` de la tabla `AISPCODE`. El objetivo de los 3 árboles es predecir `EDDAYS` a partir del resto de variables.

Con la primera unión se obtiene un Error Cuadrático Medio (ECM) bastante bajo pero se realiza el modelo con las otras tres tablas para ver si otras variables pueden ayudar en la clasificación logrando disminuir ECM y aumentar la precisión. La razón de limitar la profundidad de las ramas en vez de utilizar la máxima profundidad posible es evitar el overfitting. Si continuamos analizando la Tabla 4.4 podemos destacar las siguientes afirmaciones:

- **La adición de tablas supone la disminución de instancias**. Esto empeora la fiabilidad del modelo pues el peso que tiene cada instancia de entrenamiento es mayor además de que el modelo conoce un menor número de combinación de variables, pudiendo provocar un mayor error al encontrarse con una combinación desconocida.

- **El error cuadrático medio (ECM) disminuye con la adición de nuevas tablas**. Esto viene causado tanto por la reducción de instancias como por la adición de nuevas variables. Si observamos la condición del nodo raíz, vemos que viene determinada por variables presentes en las nuevas tablas (`AGE` pertenece a `DEMO` y `Severe sepsis` a `COMPLIC`). Esto es un claro indicador de que las nuevas variables son útiles para la clasificación de instancias en grupos y para la predicción de `EDDAYS`.

Según [64], La sepsis es la respuesta extrema del cuerpo humano a una infección. La sepsis es una emergencia médica que puede ser mortal. Requiere un tratamiento rápido, pues de no ser así puede provocar daños en los tejidos, problemas en los órganos o la muerte. Por lo que tiene sentido que esta complicación sea relevante en

la división en subgrupos para predecir la duración del paciente en ED.

Recordemos Severe sepsis es una de las posibles complicaciones recogidas en la variable que describe las complicaciones de la tabla `COMPLIC`, `COMPLDES`, y la variable `complic_Severe sepsis` es una de las variables generadas al realizar la conversión de valores categóricos a numéricos de la variable `COMPLDES`, por lo tanto la variable `complic_Severe sepsis` es una variable binaria que tomará un 1 cuando el incidente tenga esa complicación y un 0 cuando no. Por lo que la condición $Severesepsis \leq 0,5$ tiene la siguiente interpretación: si se cumple la condición, la complicación no ha sucedido en ese incidente; si por el contrario no se cumple la condición, la complicación ha sucedido en ese incidente.

ED			
Unido con:	**AISPCODE**	**AISPCODE y DEMO**	**AISPCODE, DEMO, COMORBID y COMPLIC**
Máxima profundidad	38	42	24
Error cuadrático medio a máxima profundidad	0,02	0,001	0
Profundidad elegida	14	15	8
Error cuadrático medio a la profundidad elegida	0,030	0,023	0,021
Instancias	357560	333206	11324
Condición nodo raíz	WEIGHT $\geq$ 36,45	AGE $\geq$ 28,5	Severe sepsis $\leq$ 0,5

Tabla 4.4: Árboles de decisión regresivos.

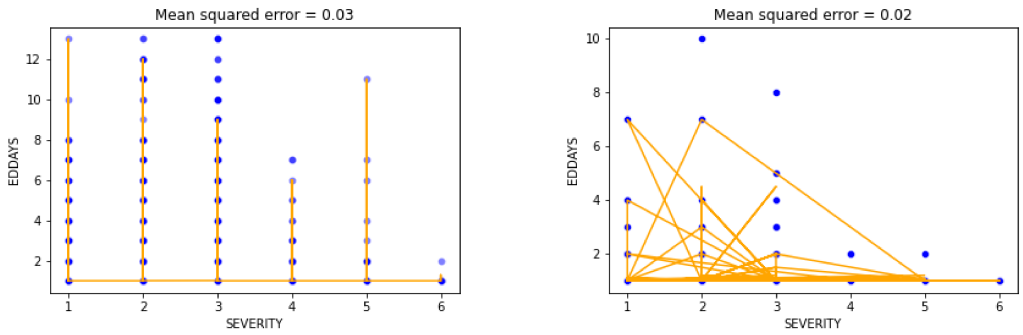
A pesar de las bajos ECM obtenidos no se puede llegar a afirmar que sea un modelo útil, debido a la distribución que representa `EDDAYS`: hay 35.586 instancias con 1 día de ED, 2.170 con 2 días, 219 con 3 días y hay menos de 12 instancias con 9, 10, 11, 12, 13 días).

En la Figura 4.6 se recogen, en azul, los valores de `EDDAYS` que se desean predecir y, en naranja, los rangos que predice el árbol. Realmente `SEVERITY` no es la variable más representativa y sería una mejor representación usar la variable del nodo raíz para estudiar su influencia en el valor predicho. Se decide representar los datos en función de `SEVERITY` por el interés que se tiene en ver la influencia de `SEVERITY` en los `EDDAYS` y porque la representación es simple y fácilmente interpretable, además de que sería difícil comparar ambas gráficas si se utilizan distintos ejes de coordenadas.

La Figura 4.6 ha sido la forma de determinar la profundidad elegida. Se ha realizado un proceso de prueba y error tratando de encontrar la mejor profundidad posible. Sabemos que hay pocas instancias con más de 8 `EDDAYS` por lo que un modelo capaz de predecir adecuadamente todas las instancias con más de 8 `EDDAYS` sería claro

candidato a producir overfitting. En cambio, si disminuimos mucho la profundidad del árbol, por ejemplo a 3, la gran mayoría de valores superiores a 4 `EDDAYS` no son estimados nunca y el modelo al ser tan simple se basa en la estimación de valores próximos a 1 ya que al haber tantas muestras con `EDDAYS` = 1 es la estimación que más reduce ECM.

Por otro lado, si comparamos las Figuras 4.6*(a)* y 4.6*(b)* podemos ver cómo la disminución de instancias ha tenido un impacto significativo en la mejora del ECM. Los modelos *(a)* y *(b)* se han hecho con una profundidad de 14 y 8 respectivamente pero aún así, el modelo *(b)* tiene mejor ECM. La disminución del ECM viene marcada, además de por el aumento de variables, por la disminución de instancias. Como vemos en la Figura 4.6*(b)*, entre las instancias eliminadas se encuentran la gran mayoría de instancias con más de 5 `EDDAYS`. Al ser los valores más alejados, disminuye el ECM la eliminación de los mismos.



(a) *(b)*

Figura 4.6: Distribución de `EDDAYS` objetivo y valores obtenidos. *(a)* Unión con AISPCODE *(b)* Unión con las 4 tablas.

Al ver que los modelos realizados para los días que pasa un paciente en ED no aparentan ser muy útiles, por la distribución de los días, se decide hacer de nuevo los modelos pero esta vez empleando la variable `EDMIN` en vez de `EDDAYS`. Con este cambio se espera ganar precisión en el modelo intentando predecir los minutos que puede pasar un paciente en el departamento de emergencias.

Para intentar predecir los minutos que pasa un paciente en el Departamento de Emergencias se generarán árboles de decisión empleando sólo las instancias de un día. Los árboles tratarán de predecir los minutos que pasará un paciente en el departamento de emergencias en caso de estar solo un día ya que casi todos los casos son para un día. Tras realizar los 3 árboles con la máxima profundidad posible, el menor MSE obtenido se produce para la unión de `AISPCODE, DEMO, COMORBID` y `COMPLIC` y es de 185,24. Dado el alto MSE obtenido para la máxima profundidad y teniendo en cuenta que se han utilizado las instancias de entrenamiento para test, podemos afirmar que estos modelos no sirven.

4.9. Resultados Random Forest

Dado que no se han podido elaborar modelos útiles en la Sección 4.8, se decide emplear Random Forest para descubrir la importancia que tiene cada variable a la hora de predecir el tiempo en el departamento de emergencias del paciente.

Para ver la importancia de cada variable, una vez entrenado un Random Forest y obtenido el coeficiente de determinación (R^2) y la puntuación fuera de bolsa (OOB), una solución es desordenar esa variable, dejando el resto como estaban y reentrenar el modelo para ver la variación de estos valores. Es importante destacar que cuanto más preciso sea el modelo, más confianza tienen los resultados obtenidos sobre la importancia de las variables[65].

El coeficiente de determinación (R^2) se calcula utilizando la ecuación 4.4

$$R^2 = 1 - \frac{u}{v} \tag{4.4}$$

Siendo u el sumatorio de errores cuadráticos de la regresión,
$u = \sum_{i=1}^{n}(Y'_{esperado} - Y_{predicho})^2$, y v la suma de errores cuadráticos de la predicción empleando la media, $v = \sum_{i=1}^{n}(Y'_{esperado} - Y_{media})^2$. El valor máximo que puede tomar R^2 es 1 y puede ser negativo si el modelo es arbitrariamente peor[60].

En la Tabla 4.9 se muestran los diferentes valores obtenidos para R^2 de Entrenamiento y R^2 de Validación además de la puntuación OOB para los diferentes modelos realizados.

	ED					
Unido con:	**AISPCODE**		**AISPCODE y DEMO**		**AISPCODE, DEMO, COMORBID y COMPLIC**	
Prediciendo:	**EDDAYS**	**EDMIN**	**EDDAYS**	**EDMIN**	**EDDAYS**	**EDMIN**
R^2 **Entrenamiento**	0,48	0,49	0,85	0.85	0,84	0,85
Puntuación OOB	-0,23	-0,22	-0,1	-0,1	-0,11	-0,07
R^2 **Validación**	-0,21	-0,22	-0,1	-0,1	-0,24	-0,08

Tabla 4.5: R^2 de entrenamiento, R^2 de validación y puntuación OOB de los diferentes Random Forest

Para la realización de los random forest de la Tabla 4.9 se ha dividido el conjunto inicial, empleando el 80 % de las instancias para entrenamiento y el resto para test, y se han empleado 1000 árboles en cada Random Forest. Como se observa en la tabla, el coeficiente de determinación, R^2, es bueno para los 4 últimos modelos, pero tanto R^2 calculado con las muestras para validación como la puntuación OOB, toman valores negativos. Por este motivo, no tiene sentido emplear estos modelos para averiguar la importancia de cada variable.

Que los valores obtenidos para R^2 de entrenamiento sean mucho mejores que los obtenidos para validación y OOB demuestra que el modelo tiene mucho overfitting y, como consiguiente, no se adapta correctamente a los datos que no ha visto en la fase de entrenamiento.

Capítulo 5

Conclusiones y líneas futuras

En este Capítulo se presentan las conclusiones extraídas del desarrollo de este proyecto y se sugieren líneas futuras de investigación.

5.1. Conclusiones

Este Trabajo Fin de Grado ha requerido una primera fase de familiarización con los traumatismos y con la base de datos, para el entendimiento de las variables almacenadas y las relaciones entre tablas. Este estudio previo ha permitido escoger qué relaciones entre tablas eran interesantes para analizar.

La exhaustiva limpieza llevada a cabo, analizando variable a variable cada una de las tablas presentes en la base de datos, ha permitido la eliminación de valores erróneos, nulos y alejados de la base de datos. Este proceso ha sido lento, pero es muy importante ya que, como se ha comentado, la eliminación de estos valores es fundamental para la generación de modelos útiles.

Adicionalmente, fruto de los análisis llevados a cabo, se ha podido priorizar entre variables de cada tabla, prescindiendo de las menos significativas. Esto es un logro importante, ya que suaviza la problemática asociada a la reducción de instancias por la presencia de valores nulos.

Seleccionadas las variables más importantes y eliminadas los valores erróneos, nulos y alejados, se ha podido estudiar la distribución de valores de las variables más significativas. Estudiar la distribución de las distintas variables, ha permitido sacar conclusiones como las siguientes: en los más de 960.000 incidentes de traumatismos recogidas en el año 2016 no hubo traumatismos faciales de severidad mayor que 4, ni traumatismos en las extremidades inferiores de severidad máxima; la mayoría de incidentes son de 1, 2 o 3 lesiones, en ese orden; la inmensa mayoría de pacientes con lesiones traumáticas pasan 1 día en el Departamento de Emergencias.

Por otro lado, se han estudiado las relaciones entre diferentes variables; los análisis realizados no han permitido encontrar relaciones estadísticas que no fueran evidentes, como la altura y el peso de una persona.

Por último, se estudió cómo se adaptan dos algoritmos a los datos para la predicción del tiempo en el Departamento de Emergencias, estos modelos no resultaron de utilidad por no hacer buenas predicciones. Se considera que los malos

resultados obtenidos con estos modelos, se deben a la distribución de datos y a la complejidad de la predicción.

5.2. Líneas futuras

Como consecuencia de los análisis realizados, se presentan líneas futuras de investigación que podrían ser interesantes de cara a conseguir encontrar relaciones significativas o la generación de modelos predictores:

- Aumentar el número de instancias mediante la inclusión de la base TQP PUF del año 2019, publicada en Junio de 2021. Esto permitiría emplear más variables en los modelos predictores al suavizar los efectos del compromiso existente entre el número de variables y el número de instancias. Dicho de otra forma, al aumentar el número de instancias podríamos permitirnos usar más variables.

- Dada la complejidad de la BD, sería interesante emplear modelos de Deep Learning teniendo presente el overfitting encontrado en los árboles entrenados para no realizar una red neuronal con excesivas capas obteniendo sobreentrenamiento.

- Uso de modelos predictores que puedan tratar nativamente relaciones no lineales entre variables. Se considera lógico seguir esta línea evitando modelos lineales debido a la falta de relaciones lineales entre variables.

- Estudiar el tiempo en el Departamento de Emergencias de los pacientes con lesiones en una zona concreta, por ejemplo la cabeza. Este estudio reduciría notablemente el número de instancias, pero al realizarse un enfoque más específico es posible que los algoritmos encuentren relaciones más acentuadas que cuando se estudian todas las zonas del cuerpo simultáneamente, permitiendo mejorar los resultados de las predicciones.

Referencias

References

[1] Sneha Grampurohit, Venkamma Shalavadi, Vaishnavi R. Dhotargavi, Megha Kudari, and Soumya Jolad. Brain tumor detection using deep learning models. In *2020 IEEE India Council International Subsections Conference (INDISCON)*, pages 129–134, 2020.

[2] Rughani AI, Dumont TM, Lu Z, Bongard J, Horgan MA, Penar PL, and Tranmer BI. Use of an artificial neural network to predict head injury outcome. 09 2010.

[3] Gráfico Blox Plot en JASP. `https://statutorial534846024.wordpress.com/2018/11/25/grafico-blox-plot-en-jasp/`, 2018.

[4] Histograma. `https://es.wikipedia.org/wiki/Histograma`.

[5] Andrew Morris, Jo Barnes, Brian Fildes, Fulvio Bentivegna, and Keith Seyer. Effectiveness of adr 69: a case-control study of crashed vehicles equipped with airbags. 06 2021.

[6] Joaquín Amat Rodrigo. Análisis de Componentes Principales (Principal Component Analysis, PCA) y t-SNE. `https://www.cienciadedatos.net/documentos/35_principal_component_analysis`, 2017.

[7] Centers for Disease Control and Prevention. CAUSES OF INJURY-RELATED DEATH. `https://www.cdc.gov/`.

[8] Severity. `https://en.wikipedia.org/wiki/Abbreviated_Injury_Scale#Severity`.

[9] Mahmoud Al-Ayyoub, Ghaith Husari, Omar Darwish, and Ahmad Alabed. Machine learning approach for brain tumor detection. 04 2012.

[10] Ali Kashani, Afshin Shariat, and Andishe Ranjbari. A data mining approach to identify key factors of traffic injury severity. 23, 01 2011.

[11] Traumatismos. `https://www.topdoctors.es/diccionario-medico/traumatismos#`.

[12] Oscar Sáenz-Morales, Mayra Murillo-Gomez, Jorge Russi-Lozano, Francisco Viera-Molina, Dafna García, Catalina Torres, and Manuel Garay-Fernandez. Coagulopatía inducida por trauma. revisión basada en la evidencia y propuesta de manejo. *Acta Colombiana de Cuidado Intensivo*, 2021.

[13] F. Alberdi, I. García, L. Atutxa, and M. Zabarte. Epidemiología del trauma grave. *Medicina Intensiva*, 38(9):580–588, 2014.

[14] Abbreviated Injury Scale. `https://aci.health.nsw.gov.au/get-involved/institute-of-trauma-and-injury-management/data/injury-scoring/abbreviated_injury_scale`.

[15] M. S. ULMAN and R. L. Stalnaker. EVALUATION OF THE AIS AS A MEASURE OF PROBABILITY OF DEATH. `http://www.ircobi.org/wordpress/downloads/irc1986/pdf_files/1986_8.pdf`.

[16] Beuran M, Stoica B, Negoi I, Tănase I, Gaspar B, Turculeţ C, and Păun S. Trauma registry – a necessity of modern clinical practice. 04 2014.

[17] Zain G. Hashmi, Amy H. Kaji, and Avery B. Nathens. Practical Guide to Surgical Data Sets: National Trauma Data Bank (NTDB). *JAMA Surgery*, 153(9):852–853, 09 2018.

[18] Trauma Quality Improvement Program (TQIP). `https://www.facs.org/quality-programs/trauma/tqp/center-programs/tqip`.

[19] Análisis de datos. `https://www.questionpro.com/es/analisis-de-datos.html`.

[20] Guillermo Westreicher. Análisis de datos. `https://economipedia.com/definiciones/analisis-de-datos.html`, 2020.

[21] Análisis de datos. `https://es.wikipedia.org/wiki/An%C3%A1lisis_de_datos`.

[22] Upasana. Top 10 Data Science Applications. `https://www.edureka.co/blog/data-science-applications/`, 2020.

[23] How Google Search works. `https://www.google.com/search/howsearchworks/`.

[24] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1701–1708, 2014.

[25] Martin Stumpe and Craig Mermel. Applying Deep Learning to Metastatic Breast Cancer Detection. `https://ai.googleblog.com/2018/10/applying-deep-learning-to-metastatic.html`, 2018.

[26] Cuídate con Ada. `https://ada.com/es/app/`.

[27] Mae Rice. 17 Data Science Applications and Examples. https://builtin.com/data-science/data-science-applications-examples, 2019.

[28] Marble RP and Healy JC. A neural network approach to the diagnosis of morbidity outcomes in trauma care. 08 1998.

[29] National Center for Health Statistics. ICD Injury Matrices. https://www.cdc.gov/nchs/injury/injury_matrices.htm, 2015.

[30] National Center for Health Statistics. International Classification of Diseases,Ninth Revision, Clinical Modification (ICD-9-CM). https://www.cdc.gov/nchs/icd/icd9cm.htm, 2015.

[31] National Center for Health Statistics. International Classification of Diseases, (ICD-10-CM/PCS) Transition - Background. https://www.cdc.gov/nchs/icd/icd10cm_pcs_background.htm, 2015.

[32] Marta Terron Cuadrado. ICD-9-CM: International Classification of Diseases, Ninth Revision, Clinical Modification. https://ec.europa.eu/cefdigital/wiki/display/EHSEMANTIC/ICD-9-CM%3A+International+Classification+of+Diseases%2C+Ninth+Revision%2C+Clinical+Modification, 2019.

[33] Kirill Fakhroutdinov. The Unified Modeling Language. https://www.uml-diagrams.org/uml-25-diagrams.html.

[34] Marcelo R. Ferreyra. ¿Qué hacer con los NULLS? http://powerhousedm.blogspot.com/2007/10/qu-hacer-con-los-nulls.html, 2007.

[35] Alicia Horsch. Detecting And Treating Outliers In Python. https://towardsdatascience.com/detecting-and-treating-outliers-in-python-part-1-4ece5098b755, 2020.

[36] Daniel Montes. Diagrama BoxPlot. https://www.pgconocimiento.com/diagrama-boxplot/.

[37] Facultad de Medicina UNNE. GRÁFICO DE CAJA. https://med.unne.edu.ar/sitio/multimedia/imagenes/ckfinder/files/files/aps/GR%C3%81FICO%20DE%20CAJA.pdf, 2018.

[38] Histograms. https://www.mathsisfun.com/data/histograms.html.

[39] Abbreviated Injury Scale predot code AAAM. https://loinc.org/75890-4/.

[40] Kinjal N. Sethuraman, Evie G. Marcolini, Maureen McCunn, Bhakti Hansoti, Federico E. Vaca and Lena M. Napolitano. Gender-specific Issues in Traumatic Injury and Resuscitation: Consensus-based Recommendations for Future Research. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4313572/, 2014.

[41] Estadounidense promedio al borde de la obesidad, indican datos. https://gestion.pe/mundo/eeuu/estadounidense-promedio-borde-obesidad-indican-datos-253616-noticia/, 2016.

[42] Normal Distribution. https://www.mathsisfun.com/data/standard-normal-distribution.html.

[43] Injury Severity Score. https://aci.health.nsw.gov.au/get-involved/institute-of-trauma-and-injury-management/data/injury-scoring/injury_severity_score.

[44] Jupyter. https://jupyter.org/.

[45] Python. https://www.python.org/.

[46] Pandas. https://pandas.pydata.org/.

[47] NumPy. https://numpy.org/.

[48] Scikit-learn. https://scikit-learn.org/stable/.

[49] Seaborn. https://seaborn.pydata.org/.

[50] Matplotlib. https://matplotlib.org/.

[51] Chris Moffitt. Guide to Encoding Categorical Values in Python. https://pbpython.com/categorical-encoding.html, 2017.

[52] Fco.J.Herrero and Marcelino Cuesta. Correlación de Pearson. https://web.archive.org/web/20091215105427/http://www.psico.uniovi.es/Dpto_Psicologia/metodos/tutor.6/fcope.html.

[53] What Is Principal Component Analysis (PCA) and How It Is Used? https://www.sartorius.com/en/knowledge/science-snippets/what-is-principal-component-analysis-pca-and-how-it-is-used-507186.

[54] Jose Martinez Heras. ¿Clasificación o Regresión? https://www.iartificial.net/clasificacion-o-regresion/, 2020.

[55] Jason Brownlee. Overfitting and Underfitting With Machine Learning Algorithms. https://machinelearningmastery.com/overfitting-and-underfitting-with-machine-learning-algorithms/, 2019.

[56] Tree algorithms: ID3, C4.5, C5.0 and CART. https://scikit-learn.org/stable/modules/tree.html#tree-algorithms-id3-c4-5-c5-0-and-cart.

[57] Mukesh Mithrakumar. How to tune a Decision Tree? https://towardsdatascience.com/how-to-tune-a-decision-tree-f03721801680, 2019.

[58] sklearn.tree.DecisionTreeRegressor. https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeRegressor.html.

[59] Random forest. https://en.wikipedia.org/wiki/Random_forest#cite_note-ho1995-1.

[60] sklearn.ensemble.RandomForestRegressor. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html.

[61] Bootstrap aggregating. https://en.wikipedia.org/wiki/Bootstrap_aggregating.

[62] Out-of-bag error. https://en.wikipedia.org/wiki/Out-of-bag_error.

[63] PCA For categorical features? https://stackoverflow.com/questions/40795141/pca-for-categorical-features, 2016.

[64] MedlinePlus, Biblioteca Nacional de Medicina de EE. UU. Sepsis. https://medlineplus.gov/spanish/sepsis.html, 2021.

[65] Eryk Lewinson. Explaining Feature Importance by example of a Random Forest. https://towardsdatascience.com/explaining-feature-importance-by-example-of-a-random-forest-d9166011959e, 2019.

Apéndice A

Aspectos éticos, económicos, sociales y ambientales

En este anexo se analizan los posibles impactos éticos, económicos, sociales y ambientales actuales y futuros como consecuencia de la realización de este Trabajo Fin de Grado.

A.1. Introducción

Tal y como se ha comentado en el Capítulo 2, el análisis de datos está a la orden del día en la sociedad y ha sido empleado, entre otras cosas, para mejorar los diagnósticos y tratamientos médicos. Se ha demostrado que los análisis de datos realizados en el campo de la medicina tiene grandes impactos en la sociedad.

A.2. Descripción de impactos relevantes relacionados con el proyecto

Este Trabajo Fin de Grado tiene un impacto directo en aspectos éticos, sociales y económicos en la actualidad; además, es la base para lograr el desarrollo de un modelo capaz de predecir distintos aspectos relacionados con los traumas, como puede ser el tiempo que pasará un paciente en el departamento de emergencias o la recuperación del paciente en base a la severidad de la lesión y otras variables; lo que tendría enormes impactos a futuro. A continuación se describen los impactos ético, social y económico asociados a este trabajo:

- Impacto ético: La limpieza, ordenación y clasificación de los datos realizada permite su visualización y tratamiento prácticamente inmediato, mejorando la información disponible del personal santario, para evitar errores de tratamiento de traumas. Los errores de tratamiento provocan lesiones más graves o incluso la muerte, reducir estos errores tiene un componente ético muy acentuado.

- Impacto social: El aumento de la información relevante sobre traumatología contribuye a la mejoría de atención médica y por tanto tiene un remarcable impacto social, puesto que la información sobre el tiempo que pasan los pacientes

en el Departamento de Emergencias permite adaptar el personal médico, la formación y los recursos hospitalarios para que el cuidado sea adecuado, de este modo, se reducen los fallecimientos y las discapacidades severas de los pacientes.

- Impacto económico: Como consecuencia de las mejoras ya destacadas en el impacto social, existe un ahorro económico importante a nivel hospilatario, reduciendo el número de días de ingreso y reduciendo el número de días de rehabilitación.

Este trabajo no tiene impactos ambientales destacables pues la realización del mismo no perjudica ni contribuye a la mejoría del medio ambiente.

A.3. Conclusiones

Resumiendo lo comentado en el apartado anterior, este proyecto tiene impactos económicos, sociales y éticos directos por su contribución en el ámbito traumatológico.

Apéndice B

Presupuesto económico

En este apéndice se recoge una estimación del coste necesario, diferenciando entre costes de personal y de recursos materiales, para cubrir el desarrollo de este Trabajo Fin de Grado.

- **Personal**: Para el cálculo de los costes de personal se tiene en cuenta la participación del estudiante y la tutora del proyecto.

	Coste horario (€)	Horas	Total (€)
Directora del trabajo	30	30	900
Estudiante de ingeniería	13	300	3900
TOTAL			**4800**

Tabla B.1: Costes de personal.

- **Costes de recursos materiales**: Para la realización de este trabajo ha sido necesario el uso de un ordenador. Se ha utilizado un ordenador con un procesador i5, 16GB de RAM y 500GB de almacenamiento SSD. El precio de el ordenador es 1000€ y el tiempo de vida estimado es de 5 años. También ha sido necesario el uso de la base de datos de traumas TQP PUF año 2016, que tiene un precio de 500$; realizando el cambio de divisas de dólares estadounidenses a euros a día 1 de Junio, el precio es de 409,84€. Se recogen estos costes en la Tabla B.2

	Tiempo de vida (años)	Uds.	Coste (€)	Amortización (€/mes)	Uso (meses)	Total (€)
Ordenador	5	1	1000	16.67	4	66.67
Base de datos	5	1	409,84	6,83	4	27,32
TOTAL						**93,99**

Tabla B.2: Costes de recursos materiales.

Habiendo mostrado en las tablas anteriores el coste desglosado para el desarrollo de este proyecto, se recoge en la siguiente tabla el coste total del proyecto, IVA incluido.

	Coste
Costes de personal	4800€
Costes de material	93,99€
Subtotal	4893,99€
IVA	1027,74€
Total	5921,73€

Tabla B.3: Costes totales.