Accidentes en turbinas eólicas: Un estudio de minería de datos

Por
Pablo Jiménez Cruz

UNIVERSIDAD COMPLUTENSE
MADRID

Grado en Ingeniería de Software
Facultad de Informática

Matilde Santos Peñas Ravi Pandit

Accidentes en turbinas eólicas: Un estudio de minería de datos

Degree in Software Engineering
Faculty of Computer Science
Complutense University of Madrid, 2021–2022

Agradecimientos

Me gustaría agradecer a mis tutores Matilde Santos Peñas y Ravi Pandit por confiar en mí y ayudarme en el desarrollo de este proyecto. Sin vosotros, este TFG no habría sido posible.

Agradezco a todos los profesores del grado de Ingeniería del Software, a todos los compañeros que he tenido todos estos años y en especial a mis compañeros del equipo de Física del Rugby, que ha sido corto pero intenso.

Por último agradezco a mi familia y a mis seres queridos, ellos saben mejor que nadie lo bueno y lo malo que ha sido esta etapa de mi vida.

Resumen

Accidentes en turbinas eólicas : Un estudio de minería de datos

En la lucha contra el cambio climático la energía eólica está jugando un papel crucial en la sustitución de energías fósiles. Como aún se trata de una tecnología en fase de expansión y desarrollo los accidentes son eventos que suceden de manera recurrente y en este trabajo nos planteamos realizar un análisis de estos. Este trabajo presenta los resultados obtenidos tras el estudio de 273 accidentes en aerogeneradores por todo el mundo.

Para ello procederemos a un análisis estadístico con la finalidad de ver la cuales son los elementos mas relevantes a tener en cuenta en el accidente y como se relacionan estos con la posibilidad de que el accidente resulte en muertes o lesiones. Para este cometido emplearemos diversas herramientas estadísticas para abordar el análisis desde diversos enfoques.
También procederemos a emplear métodos de selección y ranking de atributos, así como un análisis exploratorio de datos. La finalidad de estos procesos es ampliar el análisis a un estudio no solo de la relación entre los atributos, sino también a un estudio de los valores de esos atributos.

Finalmente, crearemos modelos predictivos utilizando varios algoritmos de clasificación para, en futuros casos, poder prever y evitar accidentes basándonos en los atributos estudiados. Los algoritmos empleados son de dos tipos, de aprendizaje supervisado como puede ser el algoritmo de 'random forest' o 'k-nearest neighbor' y algoritmos de aprendizaje no supervisados como pueden los algoritmos de 'k-means' o 'affinity propagation'. En este segundo grupo también se estudiarán un par de arquitecturas de redes neuronales.

Palabras clave:
Aerogeneradores, Accidente en aerogeneradores, Análisis de datos, Energía eólica, Inteligencia artificial, Algoritmos de clasificación, Redes neuronales,

Abstract

Wind Turbine Accidents : A Data Mining Study

In the fight against climate change, wind energy is playing a crucial role in the substitution of fossil fuels. As it is still a technology in a phase of expansion and development, accidents are recurrent events and in this paper we propose to analyze them. This paper presents the results obtained after the study of 273 wind turbine accidents around the world.

We will proceed to a statistical analysis in order to see which are the most relevant elements to take into account in the accident and how they are related to the possibility of the accident resulting in deaths or injuries. For this purpose we will use various statistical tools to approach the analysis from different approaches.

We will also perform attribute selection and ranking procedures, as well as exploratory data analysis. The purpose of these processes is to extend the analysis to a study not only of the relationship between attributes, but also to a study of the values of those attributes.

Finally, we will create predictive models using various classification algorithms in order to, in future cases, be able to predict and avoid accidents based on the attributes studied. The algorithms used are of two types, supervised learning algorithms such as the 'random forest' or 'k-nearest neighbor' algorithm and unsupervised learning algorithms such as the 'k-means' or 'affinity propagation' algorithms. In this second group, a couple of neural network architectures will also be studied.

Keywords:
Wind turbines, Wind turbine accident, Data analysis, Wind energy, Artificial Intelligence, Classification algorithms, Neural networks, Statistical analysis

Índice general

Índice de figuras

Capítulo 1

Introduction

Se prevé que la demanda mundial de energía crezca en más de dos tercios durante el periodo 2011-2035. Esta demanda se cubrirá con una mezcla de fuentes de energía no renovables (carbón, combustibles fósiles, nuclear) y renovables (eólica, hidráulica, solar, biomasa, biocombustible, geotérmica). Se prevé que la proporción de fuentes de energía renovables en la generación total de electricidad aumente del 20 % en 2011 al 31 % en 2035, y que las renovables acaben superando al gas y al carbón para convertirse en la principal fuente de energía del mundo. Esta tendencia mundial debida al mayor uso de las energías renovables está impulsada principalmente por el indeseado cambio climático global debido a las emisiones de carbono, así como por el agotamiento de los combustibles fósiles. Además, la noción de sostenibilidad de las fuentes de energía renovables está impulsando a los gobiernos a introducir legislación que promueva el uso de las energías renovables.[1]

La energía eólica tiene una larga historia y actualmente se encuentra entre las principales fuentes de energía renovable en términos de capacidad de producción. Según las estadísticas de mercado de 2013 publicadas por el Consejo Mundial de la Energía Eólica, la capacidad acumulada de energía eólica se ha triplicado con creces en nueve años.

Para la realización de este análisis nos basamos en la información de 273 accidentes de aerogeneradores ocurridos en todo el mundo y que hemos recogido de diversas fuentes como se explicará en próximos capítulos. En este estudio trabajamos con dos conceptos principales que constituyen la base de nuestro estudio estadístico.[2]

En primer lugar, la etapa del ciclo de vida del aerogenerador en la que se produjo el accidente, en la imagen anterior podemos ver las posibles etapas en las que se puede producir un accidente, a saber, durante el transporte, la construcción, el funcionamiento y el mantenimiento. En segundo lugar, la causa del accidente del aerogenerador, es decir, la naturaleza, el sistema/equipo y el ser humano. En la figura 1.1 se puede observar estos conceptos.

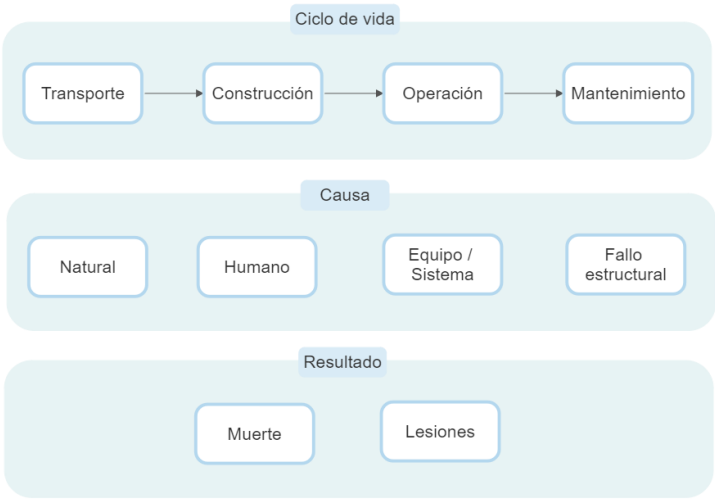


Figura 1.1: Relación causa-efecto y etapas en las que se produce un accidente.

Se investiga la asociación entre estas dos categorías de factores y dos efectos principales (resultados), es decir, la muerte y las lesiones. Así pues, las hipótesis principales de este trabajo son las siguientes:

- Hipótesis 1. Existe una asociación entre las muertes y los atributos predictores (Modelo 1).

- Hipótesis 2. Existe una asociación entre las lesiones y los atributos predictores (Modelo 2).

1.1. Motivacion

Los estudios estadísticos y los algoritmos predictivos tienen una gran capacidad para analizar grandes cantidades de datos y sacar conclusiones o reconocer patrones. Por ello, estas técnicas se utilizan cada vez más para la optimización de la toma de decisiones o para la automatización de las mismas.

Por ello, en este proyecto nos proponemos utilizar estas técnicas para que los resultados obtenidos puedan ser utilizados en futuros estudios y reducir la siniestralidad.

1.2. Objetivos

El objetivo de este proyecto es estudiar diversos accidentes de aerogeneradores utilizando herramientas estadísticas formales a partir de las cuales deducir ciertas conclusiones y obtener modelos predictivos que puedan predecir futuros accidentes de la forma más eficiente posible.

Mas concretamente, los objetivos del proyecto se exponen a continuación:

- Estudio de la literatura sobre los accidentes en aerogeneradores.
- Selección, análisis , ranking y preprocesamiento de los datos de varios accidentes.
- Comparación de modelos predictivos.
- Presentación de los resultados obtenidos en cada fase del proyecto.

1.3. Plan de trabajo

Tarea	Comienzo	Final	Dias
Trabajo de fin de grado	16/10/2021	28/5/2022	210
Investigación	16/10/2021	30/11/2021	44
Cursos herramientas básicas	1/12/2021	31/12/2021	31
Codificar Procesamiento de datos	01/01/2022	31/01/2022	31
Codificar Analisis estadistico y exploratorio	01/02/2022	15/03/2022	44
Codificar Tecnicas del clasificación	16/03/2022	30/04/2022	45
Capítulo 1 : Introduction	01/05/2022	10/05/2022	10
Capítulo 2 : Materials and methods used	11/05/2022	20/05/2022	9
Capítulo 3 : Procesamiento de datos	11/05/2022	20/05/2022	9
Capítulo 4 : Aplicación de técnicas	11/05/2022	20/05/2022	9
Capítulo 5 : Conclusiones y trabajo futuro	20/05/2022	25/05/2022	5
Capítulo 6 : Introduction, conclusion and future work	20/05/2022	20/05/2022	5

1.4. Repositorio

El código generado durante el desarrollo del proyecto se encuentra en el siguiente repositorio público de Github al que se puede acceder en el siguiente URL:

https://github.com/PabloJimenez98/TFG/tree/main

1.5. Asignaturas relacionadas

Las asignaturas cuyo contenido han ayudado a la realización de este proyecto son las siguientes:

- Algebra lineal y calculo.
- Estadística Aplicada.
- Fundamentos de Algoritmia y estructura de datos.

- Ingeniería del Conocimiento.

- Investigación operativa.

1.6. Estructura del proyecto

Para describir el trabajo realizado en cada fase del proyecto, hemos dividido el informe en los siguientes capítulos:

• Capítulo 1: Introducción. Se trata de una introducción en la que exponemos la motivación de esta investigación, los objetivos iniciales y el plan de trabajo para alcanzarlos.

• Capítulo 2: Materiales y métodos utilizados. Recoge toda la información de los datos usados en el estudio y las herramientas utilizadas en este.

• Capítulo 3: Procesamiento de datos.Esta sección contiene una descripción de los datos empleados, las transformaciones que ejecutamos sobre estos, un análisis estadístico y una análisis exploratorio de datos.

• Capítulo 4: Aplicación de técnicas y predicción de accidentes. Esta sección contiene los datos obtenidos con las técnicas de clasificación empleadas.

• Capítulo 5: Conclusiones y trabajo futuro. Esta sección contiene las conclusiones obtenidas durante los diferentes apartados del estudio y algunas medidas que se tomaran en futuros estudios.

• Capítulo 6: Introduction, conclusion and future work.

Capítulo 2

Materiales y métodos utilizados

2.1. Materiales

2.1.1. Bases de datos utilizadas y datos utilizados

Los datos utilizados en el estudio los obtuve del estudio 'Wind Turbine Accidents: A Data Mining Study'.[1] Estos fueron obtenidos en diversas bases de datos. La mayoría de los casos (aproximadamente 200 de ellos) procedían del estudio de s. Braam and Rademakers [3] los cuales se centran en accidentes ocurridos en Dinamarca, Alemania y los Países Bajos. Otra base de datos empleada fue la empleada en el estudio "Handboek Risicozonering Windturbines"[4]. Este manual se elaboró originalmente por encargo de la Agencia Neerlandesa de Energía y Medio Ambiente. Su objetivo es presentar los procedimientos para la evaluación de riesgos de los aerogeneradores. La última base de datos empleada fue la del estudio presentado por Yasuda et al.[5] Los autores se centran en los incidentes de las palas de los aerogeneradores y presentan una nueva clasificación de los mismos. Los autores también clasifican los daños causados por el fuego y sus posibles causas, además de recomendar algunas contramedidas.

Los datos de estos 273 accidentes se estructuraron como una tabla de base de datos, que contiene los atributos que se explicarán en los próximos capítulos. A continuación, en la figura 2.1, expongo una pequeña muestra de los datos en crudo, en la cual podemos observar los atributos de los diferentes accidentes y los cinco primeros casos.

	Accident No	Accident type	Site/area\n	Country\n	Details\n	Web reference/link\n	Year	Month	Day	Turbine Type	Manufacturer	Power of Turbine (KW)	PowerOfWindFarm (KW)	Offshore
0	1.0	Wind	Tjaereborg	Denmark	Damage to all 3 blades	http://www.modernpowersystems.com/story.asp?st...	2002	11	4	Vestas 2000 KW	Vestas	2000	19000	1
1	2.0	Wind	Jiangsu	China	5 died & 4 injured workers	http://www.chinadaily.com.cn/china/2012-03/11/...	2012	3	10	UNKNOWN	UNKNOWN	UNKNOWN	5000	1
2	3.0	Wind	Ireland	UK	The turbine came loose after bolts attaching i...	http://www.modernpowersystems.com/story.asp?st...	2009	1	UNKNOWN	Enercon 2000 KW	Enercon	2000	4000	0
3	4.0	Wind	Berwickshire, Scotland	UK	A 30m turbine near Coldingham was deliberately...	http://www.windbyte.co.uk/safety.html	2011	12	7	Hannevind 22 KW	Hannevind	22	UNKNOWN	0
4	5.0	Wind	Ohio	USA	The fallen wind turbine was supplying power to...	http://www.vindy.com/news/2011/apr/11/wind-tur...	2011	4	10	Proven 15 KW	Proven	15	45	0

During Construction	FULL-DATE	Outcome	Death	Fire	Injury	Mechanical	Structural Break	Transport Accident	Component	Cause	CauseCategory	Source Database	Source of the News	Title of the article
0	April 11, 2002	Structural break	0.0	0.0	0.0	0.0	1.0	0.0	Blade	Human (interference in control systems)	Human	LexisNexis	Major World Publications	Bending with the wind
1	October 3, 2012	Structural break	1.0	0.0	1.0	0.0	1.0	0.0	UNKNOWN	Mechanical (platform collapse at construction ...	Mechanical	LexisNexis	Modern Power System	Death toll in China construction site accident...
0	UNKNOWN	Structural break	0.0	0.0	0.0	1.0	1.0	0.0	Bolt	Mechanical (material fatigue)	Mechanical	LexisNexis	The Sunday Telegraph	ET 'not to blame'
0	December 7, 2011	Mechanical	0.0	0.0	0.0	1.0	0.0	0.0	Tower	Mechanical (blade system failure)	Mechanical	Google	windbyte.co.uk	Wind turbine safety
0	April 11, 2011	Structural break	0.0	0.0	0.0	0.0	1.0	0.0	Tower	UNKNOWN	UNKNOWN	LexisNexis	vindy.com	Wind turbine failure western reserve high scho...

Figura 2.1: Muestra de datos empleados

2.2. Software utilizado

2.2.1. Python

Figura 2.2: Logotipo de Python y Jupyter notebook

Para la codificación del proyecto he utilizado el lenguaje de programación Python y PyCharm como IDE.
Python es actualmente el lenguaje más utilizado en el análisis estadístico, la ciencia de datos y especialmente en el aprendizaje automático ya que dispone de un gran número de librerías y utilidades, muchas de ellas gratuitas, lo que lo convierte en el lenguaje perfecto para este estudio.

Jupyter Notebook es un entorno de desarrollo que nos permite visualizar la salida de código a partir de bloques de código independientes ya sean estos datos obtenidos de diferentes pruebas o gráficos obtenidos a partir de estos. Estas características lo convierten en una herramienta conveniente para realizar flujos de trabajo de ciencia de datos de principio a fin, que pueden ser utilizados para la limpieza de datos, el modelado estadístico, la creación y el entrenamiento de modelos de aprendizaje automático, la visualización de datos y muchos otros propósitos.[6]

2.2.2. Pandas y Numpy

Figura 2.3: Logotipo de Pandas y NumPy

Pandas es una herramienta de análisis y manipulación de datos de código abierto que nos permitirá realizar las transformaciones de datos necesarias para este proyecto.[7]

Numpy es una biblioteca para el lenguaje de programación Python que permite crear vectores y grandes matrices multidimensionales. Para este estudio la utilizaremos para realizar ciertas operaciones con nuestros datos, como una matriz de correlación de atributos.[8]

2.2.3. Scipy y SciKit Learn

Figura 2.4: Logotipo de SciPy y SciKit Learn

SciPy es una biblioteca gratuita y de código abierto que consta de varios algoritmos y pruebas utilizadas en estadística. La utilizaremos para realizar diversas pruebas en el análisis estadístico de datos.[9]

SciKit learn es una biblioteca gratuita de Python que consta de varios algoritmos para el análisis de datos y varias funciones de apoyo para ellos. En el estudio la utilizaremos para la sección de algoritmos de clasificación.[10]

2.3. Métodos

2.3.1. Métodos estadisticos

• Divergencia de Kullback-Leiber: es una medida no simétrica de la similitud o diferencia entre dos funciones de distribución de probabilidad p y q. Para este estudio la utilizamos para comparar cada atributo (p) con nuestras etiquetas (q), que dependerán del modelo (Muerte / Lesiones). [11]

$$D(p||q) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)}.$$

Figura 2.5: Formula de la divergencia de Kullback-Leiber

• Kolmogorov Smirnov: es una prueba no paramétrica que determina la bondad de ajuste de dos distribuciones de probabilidad entre sí.[12]

$$KS = max_x |F_1(x) - F_2(x)|$$

Figura 2.6: Formula de Kolmogorov Smirnovr

• Chi-cuadrado: es la distribución de la suma del cuadrado de k variables aleatorias independientes con distribución normal estándar.[13]

$$X = Z_1^2 + Z_2^2 + \cdots + Z_k^2 = \sum_{i=1}^{k} Z_i^2$$

Figura 2.7: Formula de Chi-square

• T-Test (prueba t de Student): es cualquier prueba en la que el estadístico utilizado tiene una distribución t de Student si la hipótesis nula es verdadera, en este estudio hemos utilizado la prueba t incluida en Scipy que se calcula para las medias de dos muestras independientes de puntuaciones.[14]

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_{X_1 X_2} \cdot \sqrt{\frac{2}{n}}}$$

Figura 2.8: Formula de T-Test 1

$$S_{X_1 X_2} = \sqrt{\frac{1}{2}(S_{X_1}^2 + S_{X_2}^2)},$$

Figura 2.9: Formula de T-Test 2

• Kruskal Wallis: es un método no paramétrico para comprobar si un grupo de datos procede de la misma población.[15]

$$K = (N-1)\frac{\sum_{i=1}^{g} n_i(\bar{r}_{i\cdot} - \bar{r})^2}{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(r_{ij} - \bar{r})^2},$$

Figura 2.10: Formula de Kruskal Wallis

• Shapiro: se utiliza para comprobar la normalidad de un conjunto de datos. [16]

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

Figura 2.11: Formula de Shapiro Wallis

2.3.2. Técnicas

• SVM lineal: es un algoritmo de aprendizaje automático supervisado que puede utilizarse para problemas de clasificación o regresión. Utiliza una técnica llamada el truco del kernel para transformar los datos , obtener los vectores de apoyo y luego, basándose en estas transformaciones, encuentra un límite óptimo entre las posibles salidas. Seguidamente averigua cómo separar sus datos en función de las etiquetas o salidas que ha definido.[17] En la figura 2.12 se puede observar como separa los datos con los vectores de apoyo calculando los hiperplanos.[18]

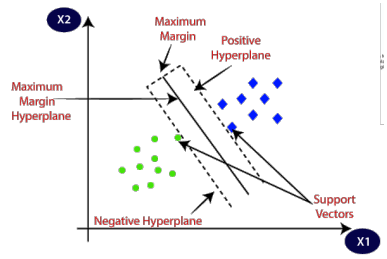


Figura 2.12: Esquema SVM Lineal

• Random Forest: es un popular algoritmo de aprendizaje automático que pertenece a la técnica de aprendizaje supervisado. Puede utilizarse tanto para problemas de clasificación como de regresión en el campo del aprendizaje automático. Se basa en el concepto de aprendizaje de conjunto, que es un proceso de combinación de múltiples clasificadores,mediante 'majority voting' para resolver un problema complejo y mejorar el rendimiento del modelo. La implementación de SciKit Learn utiliza árboles ID3, C4.5 y CART.[19] En la figura 2.13 se puede apreciar un esquema del funcinamiento del algoritmo.[20]

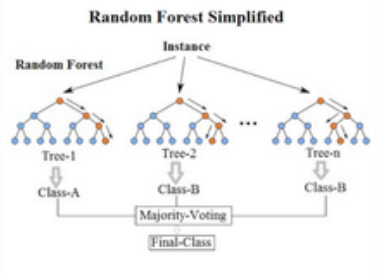


Figura 2.13: Esquema de Random Forest

• Regresión logística: es un popular algoritmo de aprendizaje automático que pertenece a la técnica de aprendizaje supervisado. Puede utilizarse tanto para problemas de clasificación como de regresión en el campo del aprendizaje automático. Se basa en el concepto matemático de regresión y es útil en problemas en los que las variables a predecir no son accesibles.[21] En la figura 2.14 se puede apreciar la funcion empleada en el clasificador.[22]

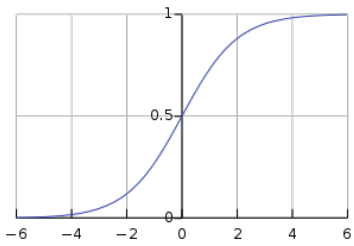


Figura 2.14: Esquema de la función empleada en la Regresión logistica [log(p/p-1)]

• K-nearest neighbor (KNN): es un popular algoritmo de aprendizaje automático que pertenece a la técnica de aprendizaje supervisado. Puede utilizarse tanto para problemas de clasificación como de regresión en el campo del aprendizaje automático. KNN funciona encontrando las distancias entre una consulta y todos los ejemplos de los datos, seleccionando el número especificado de ejemplos (K) más cercanos a la consulta, y luego vota por la etiqueta más frecuente (en el caso de la clasificación) o promedia las etiquetas (en el caso de la regresión).[23] En la figura 2.15 Se puede observar la clasificación de una nueva muestra usando este algoritmo.[20]

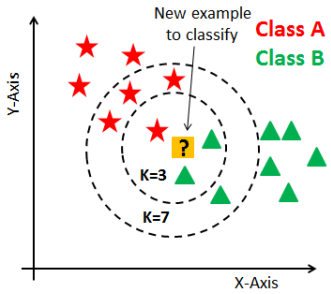


Figura 2.15: Esquema de K-Nearest Neighbot

• Árbol de CART: es un popular algoritmo de aprendizaje automático que pertenece a la técnica de aprendizaje supervisado. Puede utilizarse tanto para problemas de clasificación como de regresión en el aprendizaje automático. Las predicciones se realizan con CART recorriendo el árbol binario dado un nuevo registro de entrada. El árbol se aprende utilizando un algoritmo codicioso sobre los datos de entrenamiento para elegir divisiones en el árbol(GINI).[24]
• K-Means: La agrupación K-means utiliza 'centroides', K puntos diferentes en los datos, asignando cada punto de datos al centroide más cercano. Una vez asignado cada punto, el centroide se desplaza a la media de todos los puntos asignados. A continuación, se repite el proceso: cada punto se asigna a su centroide más cercano, los centroides se desplazan a la media de los puntos asignados. El algoritmo termina cuando ningún punto cambia su centroide asignado.[25]

• Mean Shift: es un algoritmo de clustering no supervisado que pretende descubrir puntos en una densidad suave de muestras. Es un algoritmo basado en el centroide que funciona actualizando los candidatos a centroide para que sean la media de los puntos dentro de

una región determinada (también llamada ancho de banda). Estos candidatos se filtran en un paso de posprocesamiento para eliminar los casi duplicados y formar el conjunto final de centroides. Por tanto, a diferencia de KMeans, no tenemos que elegir nosotros mismos el número de clusters.[26]

• Affinity propagation: es un algoritmo que identifica especímenes entre los puntos de datos y forma clusters de puntos de datos alrededor de estos especímenes. Funciona considerando simultáneamente todos los puntos de datos como posibles especímenes e intercambiando mensajes entre los puntos de datos hasta que surja un buen conjunto de especímenes y agrupaciones.[27]

• Las redes neuronales están formadas por agrupaciones de unidades llamadas neuronas conectadas entre sí para transmitirse información. Las neuronas modifican la información que pasa por ellas sometiéndola a diversas funciones (funciones de activación, protocolos de regresión, etc.). Las neuronas suelen agruparse en capas que tienen una capa de entrada, una o varias capas intermedias y una capa de salida. En este estudio no estudiaremos las distintas arquitecturas de las redes, sino que compararemos dos métodos de construcción de las mismas. En la figura 2.16 se puede observar un esquema sobre la estructura mas comun en redes neuronales.

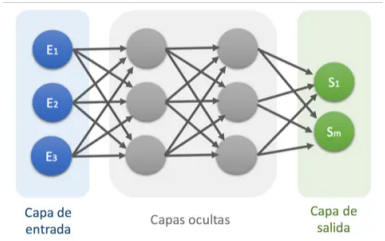


Figura 2.16: Estructura básica de una red neuronal

- Red neuronal secuencial: En el modelo secuencial se define un modelo vacío al que se le va añadiendo capa a capa la arquitectura deseada. No se pueden tener entradas de distintos inputs y no permite la regresión de capas.[28]

- Red neuronal funcional: En el modelo funcional se crea una capa inicial a la que se añade otra definiendo su estructura y su relación con la capa anterior. Para seguir construyendo la red neuronal vas añadiendo capas en la posición que te interesa sin necesidad de tener una estructura lineal. Este método permite tener varias entradas y salidas y también permite funcionalidades más complejas como la regresión entre capas.[29]

Capítulo 3

Procesamiento de datos

3.1. Descripcion de datos

Los datos obtenidos constan de 273 casos de accidentes en aerogeneradores con 28 atributos cada uno. Estos datos pueden tener ciertas restricciones funcionales entre ellos asi como cierta información redundante los cuales se trataran mas adelante. A continiación describo individualmente los atributos obtenidos.

• Lugar/área: Región donde tuvo lugar el accidente (Por ejemplo: Tjaereborg , Ohio)

• País: País donde tuvo lugar el accidente (Por ejemplo: Dinamarca, España ...).

• Modelo de turbina: Modelo del aerogenerador que sufrió el accidente, incluyendo el nombre del fabricante y el nombre/código/número del modelo, y la potencia (Por ejemplo: 'Vestas, V80-2,0 MW').

• Fabricante: La empresa que ha fabricado el aerogenerador (Por ejemplo: Vestas).

• Potencia de la turbina (kW): Potencia del aerogenerador en kW (Por ejemplo, 2000), donde 1 MW = 1000 kW. • Potencia del parque eólico (kW): Potencia total del parque eólico en el que se encuentra el aerogenerador.(Por ejemplo, 20000)

• Muerte: Indica si se ha producido una muerte humana como consecuencia del accidente; toma valores binarios. Toma el valor de 1 cuando se ha producido la muerte.

• Lesión: Indica si se han producido lesiones humanas como consecuencia del accidente; toma valores binarios . Toma el valor de 1 cuando se produce una lesión.

• Incendio: Indica si se ha producido un incendio como consecuencia del accidente; toma valores binarios . Toma el valor de 1 cuando se produce un incendio.

• Mecánico: Indica si se han producido daños mecánicos como consecuencia del accidente; toma valores binarios . Toma el valor de 1 cuando se han producido daños mecánicos.

• Rotura estructural: Indica si se ha producido una rotura estructural debido al accidente; toma valores binarios . En toma el valor de 1 cuando se ha producido una rotura estructural.

• Personas afectadas: Indica si el accidente ha afectado a personas en forma de muerte o lesiones. El valor de este atributo se calcula como el máximo de los valores de los atributos Muerte y Lesiones.

• Sistema/equipo afectado: Indica si el accidente ha afectado al sistema o al equipo de la turbina. El valor de este atributo se calcula como el máximo de los valores de los atributos Incendio, Mecánico, Avería estructural y Accidente de transporte.

• Componente afectado: Todos los componentes principales afectados como consecuencia del accidente, resumidos en forma de cadena (por ejemplo, 'Pala'). Esta cadena puede contener más de un elemento, como 'Torre, Pala'.

• Causa: Indica la causa específica del accidente [por ejemplo, 'Humano (interferencia con los sistemas de control)']. • Categoría de la causa: Indica la categoría general de la causa del accidente. Toma uno de los siguientes valores: 'Humano', 'Naturaleza', 'Sistema/Equipo'.

• Onshore/Offshore: Indica si el aerogenerador está situado en tierra (interior) o en alta mar (en el mar). Toma uno de los valores 'OnShore' u 'OffShore'.

• Ocurrencia del evento: El estado del aerogenerador cuando se produjo el accidente. Toma uno de los siguientes valores 'durante la construcción', 'durante el mantenimiento', 'durante el funcionamiento' y 'durante el transporte'.

• Año del accidente: Año en el que se produjo el accidente (por ejemplo, 2002).

• Mes del accidente: Mes en el que se produjo el accidente (por ejemplo, 11).

• Día del accidente: Día en que se produjo el accidente (por ejemplo, 4).

• Fecha completa: Campo que indica el año, el mes y el día en que se produjo el accidente (por ejemplo, 11 de abril de 2002).

• Ocurrencia del evento: Campo que indica el suceso que produjo el accidente (por ejemplo, Durante la construcción).

3.2. Transformaciones

En la figura 3.1 se muestra una breve descripción de los atributos del conjunto de datos. Podemos ver el nombre de los atributos, el número de valores no nulos (observamos que hay atributos con valores dispares) y el tipo de datos (object y float64).

```
: raw_dataset.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 279 entries, 0 to 278
Data columns (total 29 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   Accident No            273 non-null    float64
 1   Accident type          273 non-null    object
 2   Site/area
            273 non-null    object
 3   Country
               273 non-null    object
 4   Details
               273 non-null    object
 5   Web reference/link
    273 non-null    object
 6   Year                   273 non-null    object
 7   Month                  273 non-null    object
 8   Day                    273 non-null    object
 9   Turbine Type           273 non-null    object
 10  Manufacturer           273 non-null    object
 11  Power of Turbine (KW)  273 non-null    object
 12  PowerOfWindFarm (KW)   273 non-null    object
 13  Offshore               83 non-null     object
 14  During Construction    83 non-null     object
 15  FULL-DATE              273 non-null    object
 16  Outcome                273 non-null    object
 17  Death                  273 non-null    float64
 18  Fire                   273 non-null    float64
 19  Injury                 273 non-null    float64
 20  Mechanical             273 non-null    float64
 21  Structural Break       273 non-null    float64
 22  Transport Accident     273 non-null    float64
 23  Component              273 non-null    object
 24  Cause                  273 non-null    object
 25  CauseCategory          273 non-null    object
 26  Source Database        273 non-null    object
 27  Source of the News     271 non-null    object
 28  Title of the article   273 non-null    object
dtypes: float64(7), object(22)
memory usage: 63.3+ KB
```

Figura 3.1: Formato de datos en crudo

Para el estudio que queremos llevar a cabo, necesitamos que los datos cumplan ciertos criterios para que puedan realizar las pruebas que realizaremos posteriormente. Las transformaciones que he realizado en los atributos para facilitar su análisis y clasificación son las siguientes:

Creación de nuevos atributos:

- Crear un nuevo atributo denominado 'Personas afectadas' que toma el valor 1 si ha habido heridos o muertos.

- Crear un nuevo atributo llamado 'Sistema/Equipo Afectado' que toma el valor 1 si ha habido fuego, daños estructurales o accidentes durante el transporte.

Modificación de atributos:

- Modificar el atributo ' Offshore' cambiando la nomenclatura booleana por la binaria.

- Adaptar el nombre de algunos atributos para que tengan el formato adecuado.

• Separar el conjunto de datos en un conjunto X con todos los atributos y dos listas yMuertes e yHeridas con las etiquetas de los casos que se van a analizar.

Eliminación de atributos:

• Eliminar las columnas 'Accident No', 'Accident type', 'Year', 'Month', 'Day','FULL-DATE','Web reference/link', 'Outcome', 'Structural Break', 'Cause', 'Source Database', 'Source of the News', 'Title of the article'.

La información básica de los atributos resultantes se muestra a continuación, en la figura 3.2 y podemos ver que todos tienen el formato de título deseado, no tienen valores nulos y todos tienen el mismo tipo de datos (float64).

```
df.info()
<class 'pandas.core.frame.DataFrame'>
Int64Index: 273 entries, 0 to 272
Data columns (total 17 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Site/area                 273 non-null    object
 1   Country                   273 non-null    object
 2   Turbine Type              273 non-null    object
 3   Manufacturer              273 non-null    object
 4   Power of Turbine (KW)     273 non-null    object
 5   PowerOfWindFarm (KW)      273 non-null    object
 6   Onshore/Offshore          273 non-null    object
 7   During Construction       273 non-null    object
 8   Fire                      273 non-null    float64
 9   Mechanical                273 non-null    float64
 10  Structural Break          273 non-null    float64
 11  Transport Accident        273 non-null    float64
 12  Component                 273 non-null    object
 13  CauseCategory             273 non-null    object
 14  Afectted Humans           273 non-null    float64
 15  Affected System/Equipment 273 non-null    float64
 16  Event Ocurrence           273 non-null    object
dtypes: float64(6), object(11)
memory usage: 46.5+ KB
```

Figura 3.2: Formato de datos procesados

3.3. Analisis estadistico

Para nuestro análisis estadístico vamos a estudiar la relación entre pares de atributos. Para ello, calculamos la matriz correlacional que se muestra a continuación, en la figura 3.3. En ella podemos ver que no hay patrones claros entre los atributos ni con las etiquetas de los modelos (Muerte o Lesión), lo que indica que los datos son heterogéneos y será necesario analizarlos más a fondo.

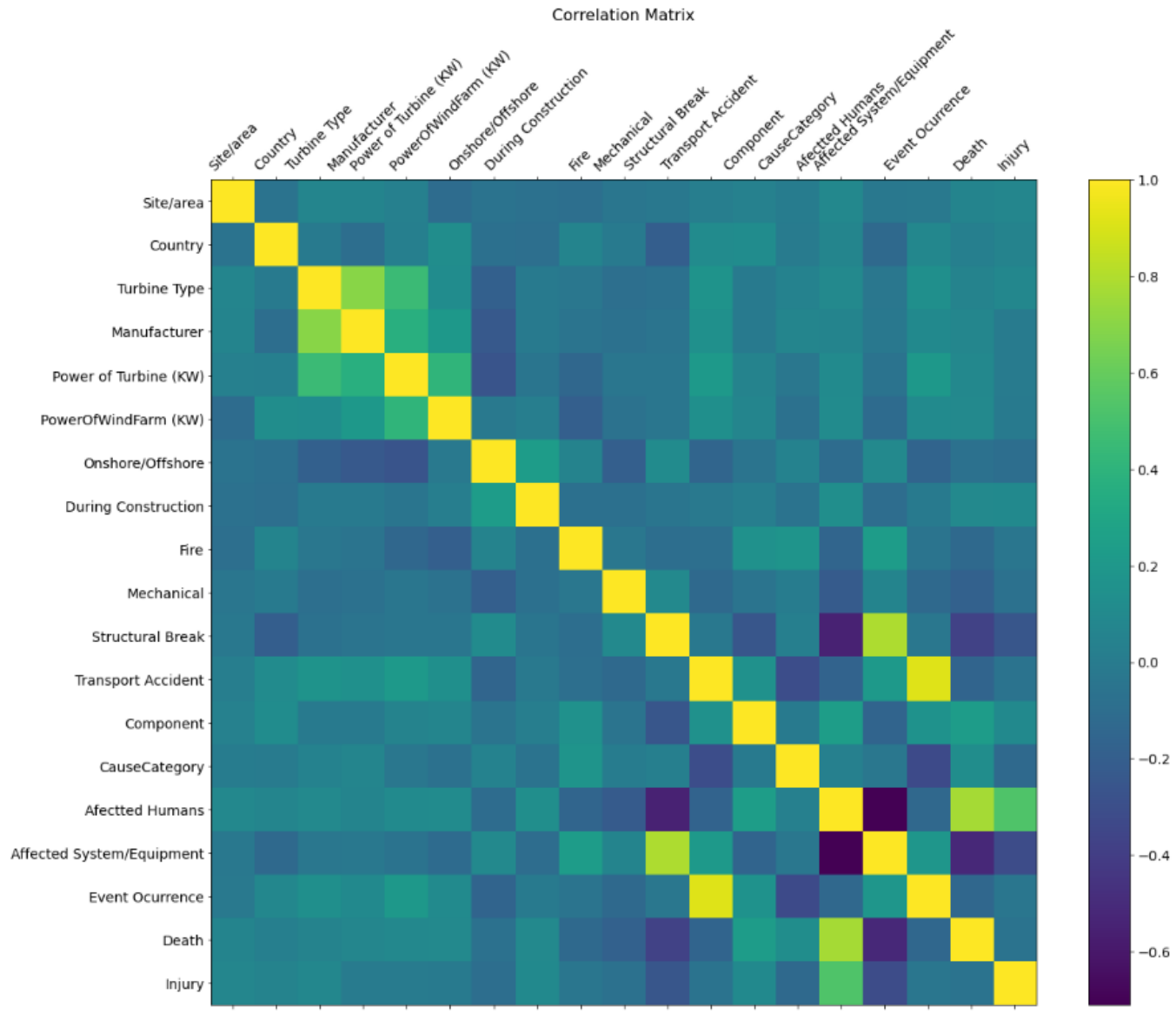


Figura 3.3: Matriz correlacional de atributos

Con los datos obtenidos seleccionamos los pares de atributos con una relación significativa (inferior a -0,2 o superior a 0,2). Eliminamos los pares que tienen un atributo con valor binario, ya que con ellos no se podían realizar las pruebas.

Para cada par de atributos se aplicó una prueba diferente. Si ambos atributos eran categóricos se utilizó la prueba de kolmogorov smirnov, si ambos eran numéricos se utilizó la prueba de chi-cuadrado y si uno era categórico y el otro numérico se comprobó si seguía una distribución normal, si era así se utilizó la prueba t, si no se utilizó la prueba de Kruskal Wallis.

La siguiente tabla (figura 3.4) muestra los resultados obtenidos para los pares de atributos seleccionados. Cada fila muestra la información de cada prueba de la siguiente manera la primera columna (Test No) muestra el número de prueba que se asigna secuencialmente a medida que se ejecutan las pruebas, las dos columnas siguientes (Attribute 1 y Attribute

2) son los nombres de los atributos utilizados en la prueba, la siguiente columna (Correlation) muestra la correlación entre los dos atributos obtenida en la matriz de correlación, la siguiente columna (Correlación) muestra la correlación entre los dos atributos obtenida en la matriz de correlación, la siguiente (Prueba realizada) muestra la prueba seleccionada para este par, la penúltima columna (Valor p) muestra el valor p obtenido al ejecutar la prueba y la última columna (Resultado) muestra si la prueba ha sido estadísticamente relevante (si su valor p es inferior a 0. 05).

	Test No	Attribute1	Attribute2	Correlation	Test Performed	p-Value	Result
4	4	Power of Turbine (KW)	PowerOfWindFarm (KW)	0.000000	Kolmogorov-Smirnov Test	9.431755e-01	-
0	0	Turbine Type	Manufacturer	0.000000	Kolmogorov-Smirnov Test	1.213906e-01	-
16	16	Affected System/Equipment	Transport Accident	0.215729	Chi Square Pearson Test	8.243929e-04	+
11	11	Afectted Humans	Mechanical	-0.221415	Chi Square Pearson Test	5.341658e-04	+
14	14	Affected System/Equipment	Fire	0.236627	Chi Square Pearson Test	2.113838e-04	+
22	22	Structural Break	Injury	0.000000	Chi Square Pearson Test	8.002098e-05	+
24	24	Injury	Affected System/Equipment	-0.311003	Chi Square Pearson Test	7.852081e-07	+
18	18	Structural Break	Death	0.000000	Chi Square Pearson Test	1.537190e-09	+
5	5	Onshore/Offshore	Manufacturer	-0.231374	Kolmogorov-Smirnov Test	2.862711e-10	+
7	7	Onshore/Offshore	During Construction	0.000000	Kolmogorov-Smirnov Test	1.598551e-10	+
21	21	Affected System/Equipment	Death	0.000000	Chi Square Pearson Test	6.079039e-17	+
23	23	Afectted Humans	Injury	0.000000	Chi Square Pearson Test	1.198460e-17	+
12	12	Afectted Humans	Structural Break	-0.548861	Chi Square Pearson Test	3.947759e-19	+
8	8	Power of Turbine (KW)	Transport Accident	0.000000	Kruskal-Wallis Test	1.234211e-28	+
9	9	Structural Break	Component	0.000000	Kruskal-Wallis Test	2.882527e-29	-
13	13	Afectted Humans	Component	0.240552	Kruskal-Wallis Test	1.707477e-29	-
10	10	CauseCategory	Transport Accident	-0.305080	Kruskal-Wallis Test	6.982690e-30	+
19	19	Death	Component	0.232637	Kruskal-Wallis Test	4.001792e-30	-
17	17	Afectted Humans	Affected System/Equipment	0.000000	Chi Square Pearson Test	4.100546e-31	+
1	1	Power of Turbine (KW)	Turbine Type	0.452508	Kolmogorov-Smirnov Test	6.921148e-33	+
20	20	Afectted Humans	Death	0.000000	Chi Square Pearson Test	2.212245e-36	+
15	15	Structural Break	Affected System/Equipment	0.000000	Chi Square Pearson Test	4.745764e-38	+
2	2	Power of Turbine (KW)	Manufacturer	0.370522	Kolmogorov-Smirnov Test	3.084387e-41	+
3	3	PowerOfWindFarm (KW)	Manufacturer	0.204375	Kolmogorov-Smirnov Test	3.070468e-60	+
6	6	Power of Turbine (KW)	Onshore/Offshore	0.000000	Kolmogorov-Smirnov Test	3.821295e-63	+

Figura 3.4: Statistical tests performed and the resulting p-values

Con los datos del análisis estadístico podemos llegar a varias conclusiones.

En relación con el primer modelo podemos observar que las muertes están fuertemente relacionadas con los fallos estructurales (Prueba nº 18). También se puede observar que las víctimas mortales están estadísticamente más afectadas si el equipo está afectado (Prueba nº 21) que si lo están las personas (que incluye a las víctimas mortales y a los heridos) (Prueba nº 20). También es relevante comentar la relación entre las víctimas

mortales y el componente afectado en el accidente (Prueba nº 19). Como hemos visto en el apartado anterior, el componente que más accidentes genera para el modelo 1 son las torres y las palas.

En el segundo modelo, al igual que en el primero, la causa principal de los accidentes es el fallo estructural (Prueba nº 22). También como en el otro modelo, los equipos afectados (Prueba nº 24) tienen una mayor incidencia estadística que las personas afectadas (Prueba nº 23). A diferencia del modelo anterior, el componente no tiene una significación estadística muy grande.

Otras conclusiones que se pueden extraer de este estudio es que la mayoría de los accidentes en los que se declaran afectados los equipos son en accidentes de transporte (Prueba nº 16). Existe una fuerte relación entre las personas afectadas y los equipos afectados, de lo que se puede concluir que probablemente si en un accidente se dañan los equipos también es probable que haya personas afectadas (Prueba nº 17). También se observa que la causa más relevante de los accidentes es la que se produce durante el transporte (Prueba nº 10).

3.4. Selecion de atributos y ranking

La elaboración de este ranking es relevante porque nos ayuda a priorizar la información disponible para su posterior análisis, viendo cuáles son los valores más relevantes de cada atributo.

Para crear un ranking de atributos he utilizado dos métodos diferentes, por un lado uso la medida de ganancia de información (divergencia de Kullback-Leibler) y por otro lado, utilizo el clasificador 'random forest' y compruebo qué atributos son los más utilizados.

La ganancia de información de un atributo 'A' es la información obtenida sobre una respuesta X a partir de la observación de los valores que toma 'A'. El concepto de ganancia de información se utiliza en las ciencias de la información para obtener una clasificación entre atributos, según su ayuda en la predicción de los valores del atributo respuesta. [30]

La figura 3.5 muestra las 11 primeras posiciones de la clasificación.

	Attribute	Values	Information Gain
13	Outcome	48	0.649163
22	Affected System/Equipment	52	0.469388
16	Structural Break	79	0.373016
2	Details	1	0.278826
15	Mechanical	177	0.265560
19	Cause	43	0.263250
17	Transport Accident	182	0.260163
14	Fire	181	0.255144
24	Injury	185	0.241803
10	Onshore/Offshore	157	0.237456
11	During Construction	201	0.225265

	Attribute	Values	Information Gain
18	Cause	22	0.316509
21	Affected System/Equipment	68	0.252747
11	FULL-DATE	7	0.222014
19	CauseCategory	55	0.219063
15	Structural Break	93	0.211864
4	Day	15	0.211581
3	Month	30	0.180183
0	Site/area	4	0.179807
12	Outcome	22	0.155684
7	Power of Turbine (KW)	24	0.143329
9	Onshore/Offshore	169	0.142189

Figura 3.5: Ranking empleando la divergencia de Kullback-Leibler en el modelo 1 y 2

En primer lugar, el ranking obtenido por la prueba de Kullback-Leibler. La clasificación del modelo 1 muestra que las causas más relevantes son el atributo 'Outcome' (que no indica si hubo o no víctimas mortales o heridos, sino el estado de la turbina tras el accidente), el atributo 'Sistema/Equipo afectado' y 'Rotura estructural'. Podemos observar que 'Outcome' tiene una relevancia significativamente mayor que 'Sistema/Equipo afectado', teniendo una 'ganancia de información' mucho mayor que cualquier otro ranking, quizás demasiado alta para ser considerada una medida fiable.('Outcome' no será tomado en cuenta en futuras tecnicas asi que no supone un problema para el estudio)

El ranking del modelo 2 muestra que las causas más relevantes son el atributo 'Outcome' (igual que en modelo anterior), el atributo 'Sistema/Equipo afectado' y 'FECHA COMPLETA'(No se tendra en cuanta en futuras tecnicas). En este caso, la relevancia de los atributos está más equilibrada y parece un resultado más fiable. Observamos que en ambas el atributo 'Sistema/Equipo afectado' está en segunda posición, lo que indica que es un atributo a tener en cuenta en futuros análisis, sin embargo el atributo 'Personas afectadas' no aparece en ninguna de las ranking lo que también es extraño ya que es un atributo que se calcula directamente a partir de los casos de muerte o lesión (véase el capítulo sobre carga y limpieza de datos).
La segunda clasificación se basa en el concepto de 'importancia' dentro del clasificador. La importancia de una característica se calcula como la reducción total (normalizada) del criterio que aporta esa característica. También se conoce como importancia de Gini.

Este método de clasificación es menos consistente si los atributos tienen una cardinalidad alta, es decir, tienen muchos valores únicos. Este no es el caso de este estudio, ya que una gran parte de nuestros atributos toman valores binarios y el resto no alcanza una cardinalidad muy alta.

La figura 3.6 muestra las 10 primeras posiciones de la clasificación.

	Feature	Rank
13	Outcome	0.229626
21	Afectted Humans	0.112967
2	Details	0.100791
5	Day	0.071199
22	Affected System/Equipment	0.063794
3	Year	0.058399
12	FULL-DATE	0.055855
9	PowerOfWindFarm (KW)	0.044013
19	Cause	0.035829
18	Component	0.035060

	Feature	Rank
0	Site/area	0.144986
20	Afectted Humans	0.113944
12	Outcome	0.103066
2	Year	0.083135
21	Affected System/Equipment	0.081120
11	FULL-DATE	0.058456
7	Power of Turbine (KW)	0.056236
6	Manufacturer	0.049946
1	Country	0.043895
18	Cause	0.042492

Figura 3.6: Ranking empleando la el algoritmo random forest en el modelo 1 y 2

En segundo lugar, tenemos las clasificaciones obtenidas mediante el algoritmo de clasificación random forest. Al igual que en el caso anterior, las tablas de arriba muestran los atributos ordenados por su ganancia de información y se comentarán los resultados obtenidos de forma independiente. En el modelo 1 podemos observar que los atributos más relevantes son 'Outcome', 'Affected Humanos' y 'Details'. Como en el caso de los rankings obtenidos con Kullback-Leibler, Outcome' sigue siendo, con diferencia, el atributo más relevante, pero con un valor de 'Ganancia de información' que no es tan alto como en el caso anterior, lo que indica que es una medida mas fiable.

En el modelo 2 podemos ver que los atributos más relevantes son 'Sitio/área', 'Humanos afectados y 'Outcome'. En contraste con las clasificaciones de Kullback-Leibler, el atributo 'Causa' ha ha perdido casi toda su relevancia y el atributo 'Sitio/Área' la ha ganado. Los valores de ganancia de información' están dentro de lo esperado, pero no hay nada que destaque especialmente. El atributo 'Personas afectadas' ocupa la segunda posición en ambas clasificaciones con una ganancia de información', lo que indica que se trata de una medida bastante correcta, ya que los casos de víctimas mortales y heridos son bastante similares
Para la exploración de los datos he decidido utilizar los datos obtenidos de las clasificaciones obtenidos por el clasificador de bosque aleatorio ya que la información obtenida parece ser más sólida (omito los valores binarios como 'Humanos afectados').

Las siguientes figuras, 3.7 y 3.8, muestran los valores más relevantes de cada atributo en los casos en los que se ha producido una muerte para el modelo 1 o una lesión para el modelo 2. La información de cada atributo (columna) es independiente de las demás, las relaciones entre atributos se verán más adelante. Como podemos ver en varios atributos el valor 'UNKNOWN' tiene gran relevancia pero no se tendrá en cuenta porque estos valores (al igual que los NaN y los valores indefinidos) serán eliminados para el análisis posterior. Los atributos de valor binario se han omitido del análisis porque no aportan mucha información

	Component	PowerOfWindFarm (KW)	Site/area	Turbine Type	CauseCategory	Power of Turbine (KW)	Country	Manufacturer
0	Tower	UNKNOWN	UNKNOWN	UNKNOWN	UNKNOWN	UNKNOWN	USA	UNKNOWN
1	UNKNOWN	37500	Palm Springs, California	NEG Micon 1500 KW	Human	2500	Germany	Vestas
2	Blade	44000	Minnesota	Vestas 2000 KW	Mechanical	1500	China	Siemens
3	Crane	49500 KW	Oregon	Nordex 2500 KW	Nature	2000	UK	Sinovel
4	Nacelle	5000	Iowa	Sinovel	Structural	1750	UNKNOWN	Clipper

Figura 3.7: Modelo 1 atributos mas relevantes

	Component	PowerOfWindFarm (KW)	Site/area	Turbine Type	CauseCategory	Power of Turbine (KW)	Country	Manufacturer
0	Tower	UNKNOWN	Texas	UNKNOWN	Human	UNKNOWN	USA	UNKNOWN
1	Blade	60000	Pennsylvania	Vestas 3000 KW	Mechanical	2000	UK	Vestas
2	UNKNOWN	420000	Winnebago County, Illinois	Siemens 2300 KW	UNKNOWN	1500	Australia	Gamesa
3	Crane	420000 KW	Oregon	BWC Excel 10 KW	Nature	2300	China	Siemens
4	Nacelle	20000	Bakersfield	Kenetech 100 KW	Structural	10 KW	Greece	BWC

Figura 3.8: Modelo 2 atributos mas relevantes

Además, para mostrar más información relevante para el análisis, mostramos las causas más relevantes para ambos modelos, las cuales se pueden apreciar en la figura 3.9. Como en el caso anterior no se tendrá en cuenta el valor UNKNOWN. Podemos observar que la causa más relevante en ambos modelos son los casos provocados por causas humanas, seguidos de las causas mecánicas (por colisiones, negligencias, etc.). Observamos que las causas naturales no son muy relevantes en el estudio.

	Cause
0	UNKNOWN
1	Human
2	Mechanical (due to collision)
3	Mechanical
4	Human (transportation)
5	Human (negligance)
6	Mechanical (platform collapse at construction ...
7	Structural (dismantling)
8	Mechanical (falling section of tower)
9	Mechanical (not properly secured foundation bo...
10	Nature (lightening strike)
11	Nature
12	Human (plane crash)
13	Mechanical (sudden drop in air pressure)

	Cause
0	Human
1	UNKNOWN
2	Mechanical
3	Human (negligance)
4	Mechanical (due to collision)
5	Mechanical (platform collapse at construction ...
6	Structural (falling object)
7	Mechanical (not properly secured foundation bo...
8	Nature
9	Mechanical (electrical discharge)
10	Nature (lightening strike)
11	Mechanical (hazardous wind turbine electric ho...
12	Human (wrong action)
13	Mechanical (electrical flash)

Figura 3.9: Main Causes of accidents in model 1 and 2

3.5. Analisis exploratorio

Los dos gráficos siguientes, en la figura 3.10, muestran la relación entre la 'Ocurrencia del suceso' y la 'Categoría de la causa' de los accidentes. Como en el gráfico de barras no se puede apreciar bien la información, ya que en el 'Humano' los accidentes durante el transporte oscurecen la visión de los otros, he añadido un gráfico de mosaico en el que se puede apreciar mejor la información.

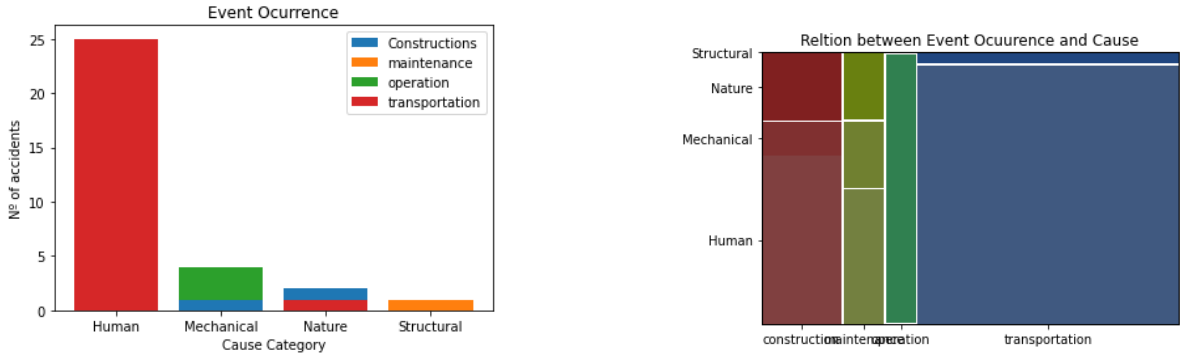


Figura 3.10: Gráfico de barras y gráfico de mosaico entre la ocurrencia del evento y la causa

- Está claro que el mayor número de accidentes se produce durante el transporte debido a errores humanos.

- La mayoría de los accidentes debidos a los equipos (mecánicos y estructurales) se producen durante mantenimiento y transporte.

- No se detectaron fallos estructurales o mecánicos durante el transporte, lo que tiene sentido porque durante el transporte el aerogenerador no está en funcionamiento

A continuación muestro la relación de cada uno de los atributos de forma independiente con los casos de muertes y lesiones.

La figura 3.11 muestra el número de muertos y heridos en la 'categoría de causas'.

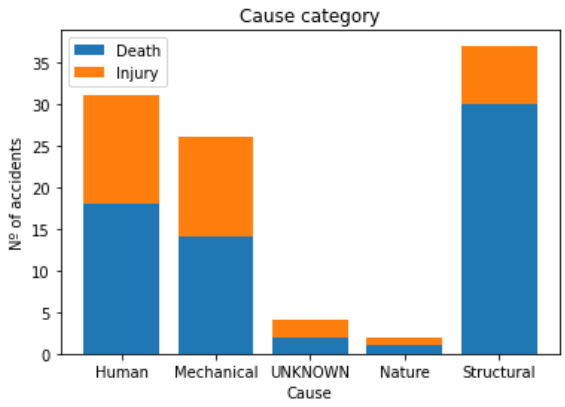


Figura 3.11: Gráfico de mosaico de Causas en el modelo 1 y 2

Con esta información podemos sacar algunas conclusiones.

- La mayoría de los accidentes están causados por fallos estructurales, seguidos de fallos mecánicos y errores humanos.

- El número de víctimas mortales por causas naturales es muy bajo.

- El número de heridos es bajo en relación con el número de muertos.

La figura 3.12 muestra el número de muertos y heridos en la 'Ocurrencia de eventos'.

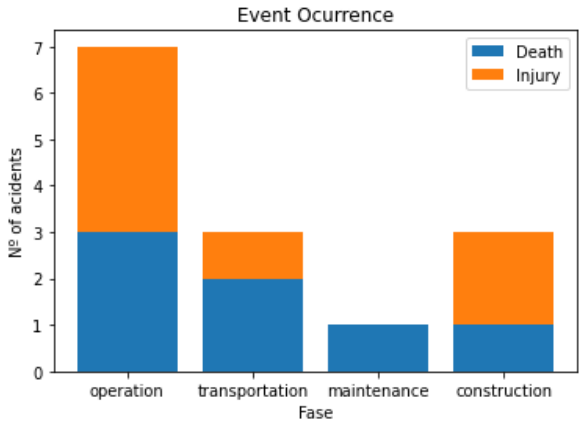


Figura 3.12: Gráfico de mosaico de Eventos en el modelo 1 y 2

Con esta información podemos sacar algunas conclusiones.

- La cantidad de datos de que disponemos es bastante escasa, por lo que el análisis puede no ser fiable. ble.

- El número de muertes se mantiene con valores similares en todos los atributos.

- No se detectaron lesiones durante el mantenimiento. El mayor número de lesiones se detectó durante la construcción.

Capítulo 4

Aplicación de tecnicas y predicción de accidentes

El objetivo en el análisis de clasificación es predecir la clase (si podrian darse casos de muerte o lesiones) a la que podrían pertenecer futuros casos de accidentes basándonos en los atributos predictores de los que disponemos. Para ello, los datos se dividen sistemáticamente en conjuntos de datos de entrenamiento y de prueba, el conjunto de datos de entrenamiento se utiliza para 'enseñar' los algoritmos de clasificación en los datos, y el rendimiento de los algoritmos de clasificación se comprueba con el conjunto de datos de prueba.

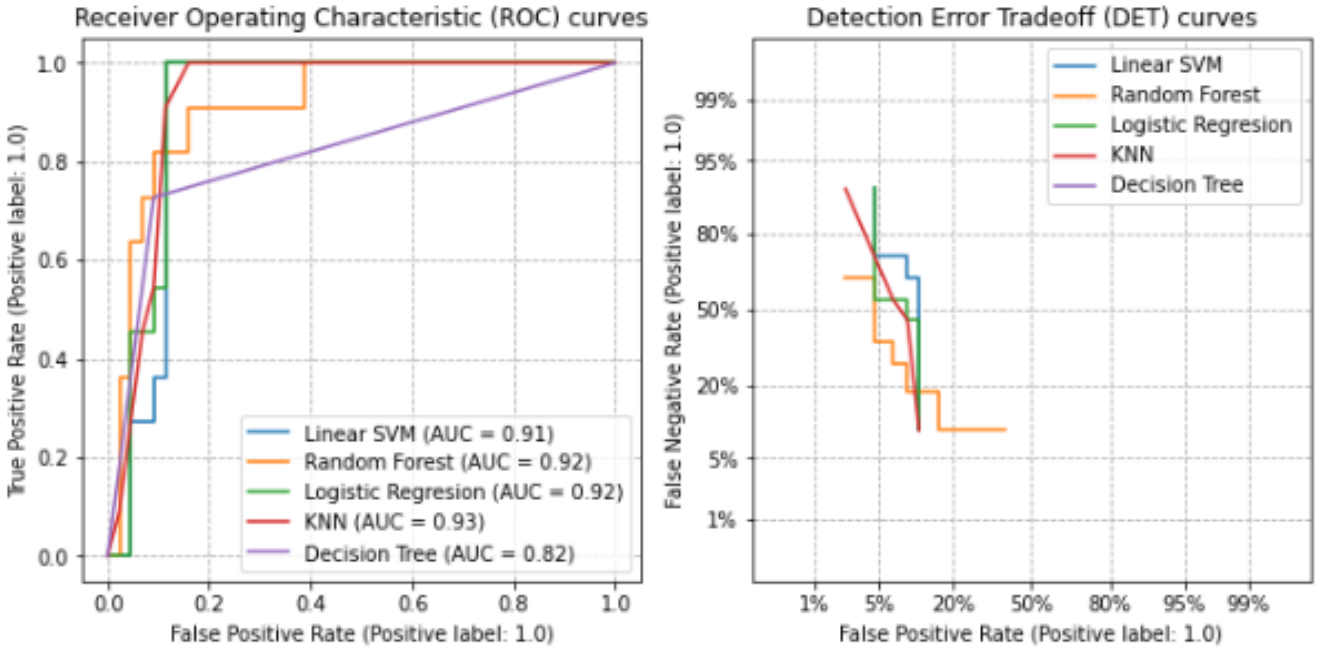
4.1. Algoritmos de clasificación supervisados

Los resultados que he obtenido con la ejecución de los algoritmos de clasificación son los siguientes:

El primer gráfico (Receiver Operating Charasteristics ROC curves) muestra la evolución de la relación entre verdaderos positivos y falsos positivos que indica el porcentaje de éxito del clasificador. En este gráfico distinguimos la tasa de aciertos (CA) y el área bajo la curva que describe la fiabilidad del algoritmo de clasificación.

El segundo gráfico (Curvas DET del error de detección) muestra la evolución de la relación entre verdaderos negativos y falsos negativos, lo que indica la tasa de éxito del clasificador. En él podemos ver que los clasificadores cuya curva DET es más parecida a una 'L' tienen una tasa de error menos relevante y, por tanto, el algoritmo clasificador es más fiable.

Finalmente en la tabla tenemos todos los valores obtenidos al ejecutar los algoritmos de clasificación ordenados por su tasa de aciertos donde tambien se puede apreciar su tasa de acierto.

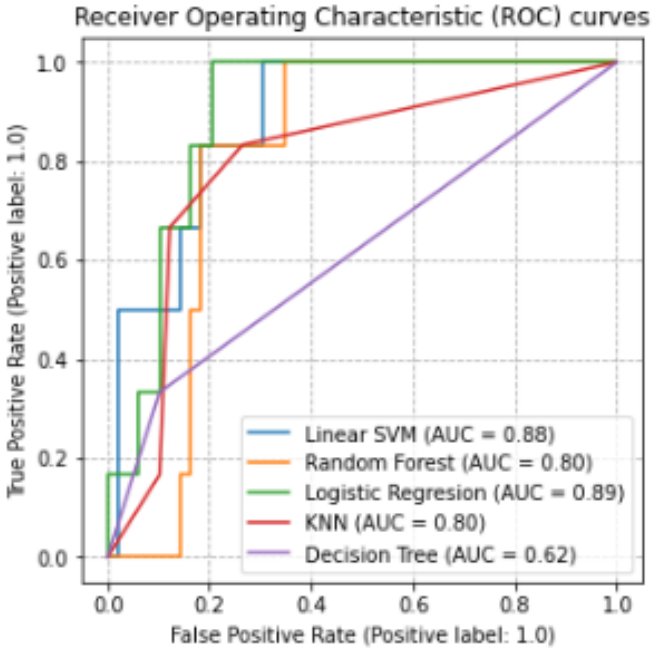


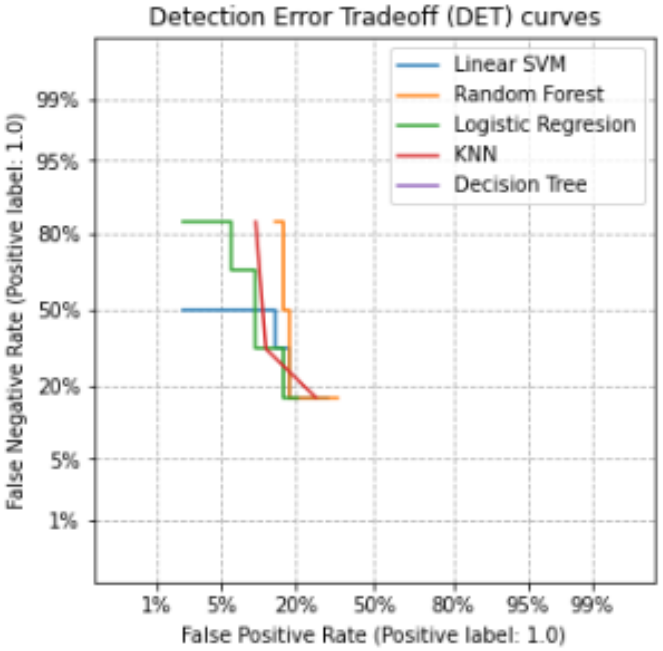
	Classifier	CA	AUC
0	Linear SVM	0.909091	0.943182
2	Logistic Regresion	0.872727	0.852273
4	Decision Tree	0.872727	0.818182
3	KNN	0.836364	0.727273
1	Random Forest	0.854545	0.670455

Figura 4.1: Curva AUC, curva DET y tabla de aciertos en el modelo 1

Como se puede ver en los gráficos anteriores contenidos en la figura 4.1, el clasificador con mejores resultados es el SVM lineal con un porcentaje de éxito del 94,5 %, seguido del clasificador de regresión lineal con un 93,4 %, que empata con el árbol de decisión CART con el mismo porcentaje. Estos tres modelos tienen una curva DET aceptable y un valor AUC que también es bastante bueno, aunque no es fácil de apreciar en el gráfico ROC. Los algoritmos k-Nearest neighbor y random forest no han obtenido un porcentaje de aciertos aceptable y si observamos sus curvas DET podemos ver que su ratio de falsos negativos también es muy alto, por lo que podemos concluir que estos dos modelos no son muy fiables.

A continuación puedes ver las curvas ROC de los clasificadores ejecutados, la curva DET y una tabla donde se ordenan los clasificadores por su AUC para el modelo 2.





	Classifier	CA	AUC
0	Linear SVM	0.927273	0.739796
4	Decision Tree	0.836364	0.615646
2	Logistic Regresion	0.909091	0.583333
3	KNN	0.818182	0.532313
1	Random Forest	0.890909	0.500000

Figura 4.2: Curva AUC, curva DET y tabla de aciertos en el modelo 2

Como se puede ver en los gráficos anteriores contenidos en la figura 4.2, el clasificador con mejores resultados es el árbol de decisión CART con un 95,6 % de éxito, seguido del clasificador k-Nearest neighbor con un 91,2 %, que empata con el clasificador lineal SVM con el mismo porcentaje, seguido de cerca por el clasificador de regresión logística con un 90,1 % de éxito. Estos cuatro modelos tienen una curva TED aceptable y un valor AUC que también es bastante bueno, aunque no es fácil de apreciar en el gráfico ROC. El algoritmo de bosque aleatorio no ha obtenido una tasa de aciertos aceptable y si observamos sus curvas DET podemos ver que su ratio de falsos negativos también es muy alto, por lo que podemos concluir que este modelo no es muy fiable.

4.2. Algoritmos de clasificación no supervisados

Utilizo tres clasificadores: K-means, Mean Shift y Affinity propagation. Para todos ellos utilizo la técnica PCA para reducir la dimensionalidad de los datos a 2 dimensiones para dar más consistencia a los datos y facilitar su visualización.

A continuación muestro en la figura 4.3 los gráficos obtenidos al ejecutar los algoritmos de clustering (son los mismos en ambos modelos ya que en ambos uso los mismos datos (X) y los algoritmos de clustering no utilizan las etiquetas del modelo (y) por lo que genera los mismos gráficos).

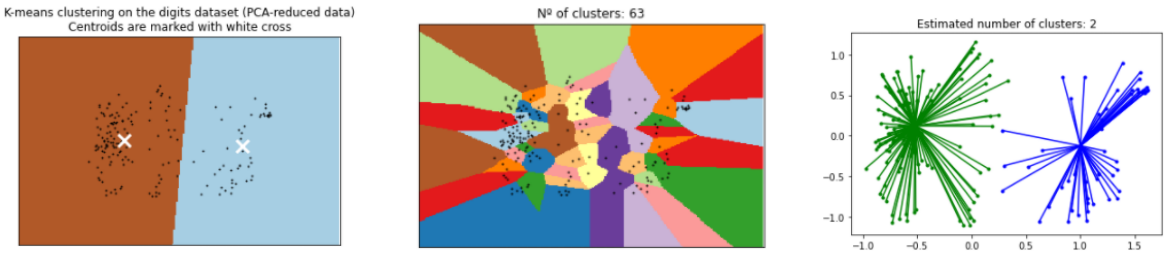


Figura 4.3: K-means, Mean Shift y Affinity Propagation

A continuación muestro los resultados de diferentes pruebas que indican:

- Fowlkes Mallow: Puntuación comparada con el conjunto de pruebas (equivalente al CA en los clasificadores).

- Calinski Harabasz: Densidad de clusters

- Silueta: Superposición

- Davies Bouldin: Puntuación óptima de los clusters (cuanto más baja, mejor)

También añado un informe en el que se indica

- Precisión : Indica el porcentaje de aciertos para cada valor de etiqueta.

- Recall score : Indica la relación entre verdaderos positivos y falsos negativos según la

función siguiente: tp/(tp+fn), donde tp son los verdaderos positivos y fn los falsos negativos.

- Puntuación F1 : La puntuación F1 puede interpretarse como una media armónica de la precisión y la recuperación, donde la puntuación F1 alcanza su mejor valor en 1 y la peor puntuación en 0 según la siguiente fórmula : F1 = 2* (precisión * recuperación) / (precisión + recuperación).

- Soporte : Es el número de veces que el valor predicho coincide con el valor real.

	Cluster	Fowlkes Mallows	Calinski Harabasz	Silhouette	Davies Bouldin
0	K-means	0.793070	233.972318	0.552504	0.725918
1	Means Shift	0.181967	563.294372	0.363000	0.455130
2	Affinity Propagation	0.797480	231.838908	0.543586	0.743614

Figura 4.4: Resultados de las pruebas de los algoritmos de agrupación en el modelo 1

K-Means info:

	precision	recall	f1-score	support
0	0.92	0.88	0.90	138
1	0.67	0.77	0.72	44
accuracy			0.85	182
macro avg	0.80	0.82	0.81	182
weighted avg	0.86	0.85	0.86	182

Mean Shift info:

	precision	recall	f1-score	support
0	0.00	0.00	0.00	138.0
1	0.00	0.00	0.00	44.0
micro avg	0.00	0.00	0.00	182.0
macro avg	0.00	0.00	0.00	182.0
weighted avg	0.00	0.00	0.00	182.0

Affinity propagation info:

	precision	recall	f1-score	support
0	0.33	0.13	0.19	138
1	0.06	0.18	0.09	44
accuracy			0.14	182
macro avg	0.20	0.16	0.14	182
weighted avg	0.27	0.14	0.16	182

Figura 4.5: Reporte F1 en el modelo 1

Como podemos ver en la figura 4.4 Mean Shift es, con diferencia, el algoritmo menos óptimo porque al estimar de forma autónoma el número de clusters obtiene valores alejados de lo esperado, aunque en el test de Calinski-Harabasz obtiene la mejor puntuación porque tiene más clusters y éstos son más densos.

Los algoritmos K.Means y de propagación de afinidad tienen métricas muy similares pero en la prueba Fowles-Mallows que incide en la tasa de aciertos el algoritmo de propagación de afinidad tiene una ligera ventaja sobre K-Means, sin embargo si observamos el informe en la figura 4.5 podemos ver que la exactitud de la medida y el recall son mucho mejores en K-Means lo que indica que en términos generales el algoritmo K-Means ha obtenido un mejor resultado para este modelo.

	Cluster	Fowlkes Mallows	Calinski Harabasz	Silhouette	Davies Bouldin
0	K-means	0.746580	233.972318	0.552504	0.725918
1	Means Shift	0.164387	563.294372	0.363000	0.455130
2	Affinity Propagation	0.740729	231.838908	0.543586	0.743614

Figura 4.6: Resultados de las pruebas de los algoritmos de agrupación en el modelo 2

```
K-Means info:
              precision    recall  f1-score   support

           0       0.95      0.78      0.86       159
           1       0.31      0.70      0.43        23

    accuracy                           0.77       182
   macro avg       0.63      0.74      0.64       182
weighted avg       0.87      0.77      0.80       182

Mean Shift info:
              precision    recall  f1-score   support

           0       1.00      0.08      0.15       159
           1       0.00      0.00      0.00        23

   micro avg       0.57      0.07      0.13       182
   macro avg       0.50      0.04      0.08       182
weighted avg       0.87      0.07      0.13       182

Affinity propagation info:
              precision    recall  f1-score   support

           0       0.69      0.23      0.35       159
           1       0.05      0.26      0.08        23

    accuracy                           0.24       182
   macro avg       0.37      0.25      0.21       182
weighted avg       0.60      0.24      0.31       182
```

Figura 4.7: Reporte F1 en el modelo 2

Como podemos ver en la figura 4.6 Mean Shift es, con diferencia, el algoritmo menos óptimo porque al estimar de forma autónoma el número de clusters obtiene valores alejados de lo esperado, aunque en el test de Calinski-Harabasz obtiene la mejor puntuación porque tiene más clusters y éstos son más densos.

Los algoritmos de K.Means y de Propagación de Afinidad tienen métricas muy similares, pero en el test de Fowles-Mallows que incide en la tasa de aciertos, el algoritmo de K-Means tiene una ligera ventaja sobre el de Propagación de Afinidad. En este modelo las métricas obtenidas en el informe en la figura 4.7 no son tan dispares como en el modelo anterior aunque siguen siendo significativas.

4.3. Redes neuronales

Para cada modelo he implementado dos arquitecturas de redes neuronales en función de la construcción (secuencial o funcional) y de las que obtengo dos métricas, la tasa de aciertos y la pérdida de información (la información que se muestra sobre los gráficos son los valores medios de las métricas).

En la figura 4.8 se muestra la evolucion del indice de aciertos en el primer modelo para cada arquitectura.

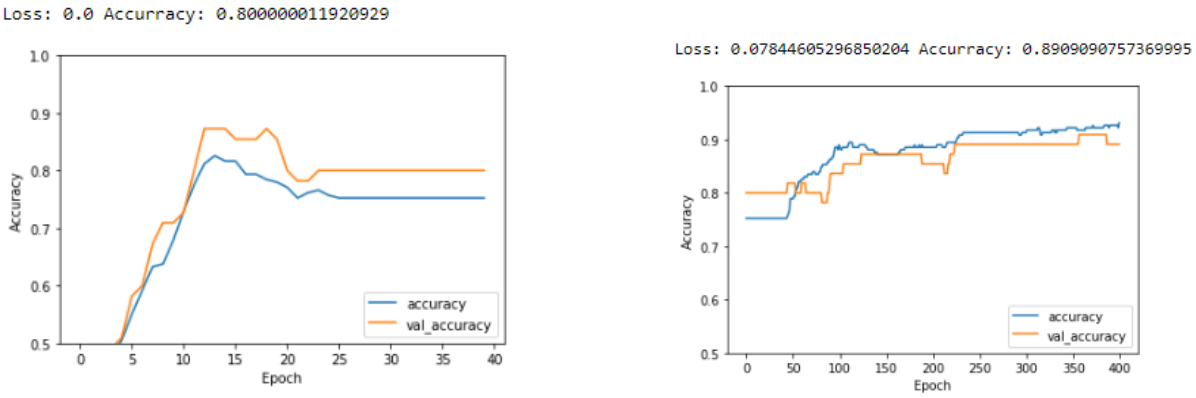


Figura 4.8: Gráfico de la evolución de la precisión en el modelo secuencial (izquierda) y en el modelo funcional (derecha) en el modelo 1.

En los gráficos anteriores podemos ver la función descrita por la tasa de aciertos y la función de pérdida de nuestras redes neuronales. En la arquitectura secuencial podemos ver que el progreso se estanca rápidamente y no consigue un buen porcentaje de aciertos aunque su función de pérdida se mantiene en 0. En la arquitectura funcional podemos ver un progreso continuo en el porcentaje de aciertos que está estrechamente relacionado con la función de pérdida.

Comparando ambas arquitecturas podemos ver que la arquitectura funcional da mejores resultados aunque su función de pérdida es mayor.

La figura 4.9 muestra la tasa de aciertos final de ambas arquitecturas y su pérdida total. Es evidente que la arquitectura funcional obtiene mejores resultados.

	Neural Network	Accuracy	loss
0	Sequential	0.769231	0.000000
1	Functional	0.934066	0.041822

Figura 4.9: Aciertos y pérdidas en el modelo 1

Al igual que en el primer modelo, se observa un rápido estancamiento en la arquitectura secuencial, pero esta vez sí genera alguna pérdida. La arquitectura funcional, a diferencia del modelo anterior, parte de una mejor tasa de éxito pero no avanza demasiado durante

los primeros 150 periodos de entrenamiento, a partir de ahí tiene un progreso continuo.

En la figura 4.10 se muestra la evolucion del indice de aciertos en el segundo modelo para cada arquitectura.

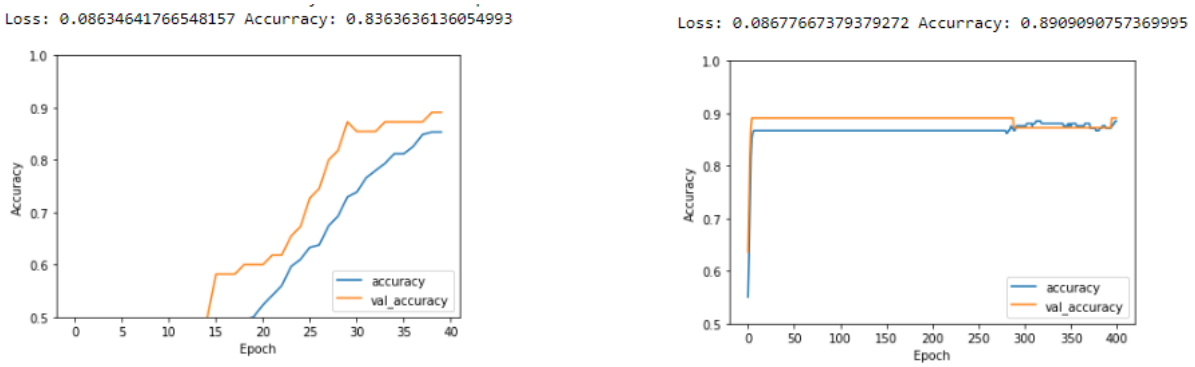


Figura 4.10: Gráfico de la evolución de la precisión en el modelo secuencial (izquierda) y en el modelo funcional (derecha) en el modelo 2.

La figura 4.11 muestra la tasa de aciertos final de ambas arquitecturas y su pérdida total. Es evidente que la arquitectura funcional obtiene mejores resultados.

	Neural Network	Accuracy	loss
0	Sequential	0.769231	0.000000
1	Functional	0.934066	0.041822

Figura 4.11: Aciertos y pérdidas en el modelo 2

4.4. Ranking de técnica

Para concluir, comparamos todas las técnicas desarrolladas y las clasificamos según su precisión. En la figura 4.12 podemos observar la tasa de aciertos de cada algoritmo utilizado para este estudio.

	Techniques	Name	Accuracy
9	Neural Network	NN Functional	0.909091
0	Classifier	Linear SVM	0.909091
2	Classifier	Logistic Regresion	0.872727
4	Classifier	Decision Tree	0.872727
1	Classifier	Random Forest	0.854545
3	Classifier	KNN	0.836364
8	Neural Network	NN Sequential	0.836364
5	Clusters	K-means	0.779084
7	Clusters	Affinity Propagation	0.752045
6	Clusters	Means Shift	0.173424

	Techniques	Name	Accuracy
0	Classifier	Linear SVM	0.927273
10	Neural Network	NN Functional	0.909091
2	Classifier	Logistic Regresion	0.909091
1	Classifier	Random Forest	0.890909
4	Classifier	Decision Tree	0.836364
9	Neural Network	NN Sequential	0.836364
3	Classifier	KNN	0.818182
5	Clusters	K-means	0.779084
7	Clusters	Affinity Propagation	0.752045
8	Clusters	Affinity Propagation	0.718602
6	Clusters	Means Shift	0.173424

Figura 4.12: Clasificación de los atributos en el modelo 1 (izquierda) y en el modelo 2 (derecha)

En ambos modelos podemos apreciar que los algoritmos de clasificación no supervisada han obtenido índices de aciertos muy bajos, probablemente debido a la poca uniformidad de los datos y a la escasa correlación entre ellos, como observamos en el apartado 3.3, por lo que al calcular los clústeres obtenemos resultados tan poco acertados. Los algoritmos de clasificación supervisados ofrecen resultados mas uniformes y en general mas acertados. Destacar especialmente los algoritmos de regresión lineal y el SVM lineal que han obtenido los mejores resultados, debido probablemente a que son algoritmos que funcionan bien datos de alta cardinalidad. Los resultados obtenidos por las redes neuronales son en conjunto prometedores en sus resultados y dispares en los dos modelos. En el modelo 1 la red neuronal funcional obtiene la mejor clasificación, mientras que para el modelo 2 se queda en segunda posición detrás del SVM lineal.

Capítulo 5

Conclusiones y trabajo futuro

5.1. Conclusions

Durante el desarrollo de este proyecto hemos obtenido varias conclusiones basadas en los resultados del análisis exploratorio de datos reales y el análisis estadístico de varios accidentes de aerogeneradores.

Se ha hecho especial hincapié en los atributos más característicos de estos accidentes y su relación con las muertes y lesiones que pueden producirse en los mismos. Estos atributos son el ciclo de vida del aerogenerador, si el accidente se produjo durante su construcción, transporte, operación o mantenimiento, y la causa del accidente, si fue provocado por la naturaleza, por un defecto del equipo o por un error humano.

Las conclusiones más relevantes obtenidas en este estudio son las siguientes:

- Los defectos de los equipos son la principal causa de muertes y lesiones en los accidentes de aerogeneradores. - Los componentes que generan más fallos son la torre del aerogenerador y las palas, que son más susceptibles de causar accidentes en la fase de transporte. - Los países con más probabilidades de sufrir un accidente con víctimas mortales son Estados Unidos, Alemania y China, por este orden. Los países con mayor probabilidad de sufrir un accidente con heridos son EE.UU., Reino Unido y Australia, por este orden. - La gran mayoría de accidentes suceden durante el transporte y son casi en su totalidad por errores humanos.
Uno de los mayores problemas a los que nos hemos enfrentado en el estudio es la cantidad de datos, ya que los datos se obtuvieron escaneando un motor de búsqueda público y las noticias públicas con información relevante son escasas. Por lo tanto, nuestros datos no son completos y son sólo una muestra. Como en cualquier estudio en el que se toma una muestra de la población, existe el riesgo de que nuestra muestra no sea en realidad una muestra aleatoria que represente a la verdadera población.

Además, hemos construido varios modelos de predicción utilizando diversas estrategias. Hemos explorado modelos supervisados y no supervisados, obteniendo ambos márgenes de error aceptables. También hemos trabajado brevemente con redes neuronales sencillas que han arrojado datos muy prometedores.

Para el primer modelo, el que estudia las muertes en accidentes, el modelo que ha obtenido los mejores resultados es el SVM lineal, seguido del árbol de decisión CART y la red neuronal con estructura funcional. Los algoritmos de clustering se han quedado muy atrás. En el segundo modelo, el que estudia las lesiones, el modelo con mejores resultados es el SVM Lineal seguido de la red neuronal con estructura funcional y el árbol de decisión CART. Podemos observar que estas técnicas son superiores en ambos modelos. El árbol de decisión CART y la SVM lineal funcionan bien con cantidades limitadas de datos, por lo que se esperaba su eficacia, pero la red neuronal funcional destaca sobre todo porque en teoría funciona mejor con grandes cantidades de datos, por lo que se puede deducir que en el futuro al aumentar la cantidad de datos para entrenar los modelos, esta técnica podría mejorar aún más sus resultados.

Finalmente, y más allá de los resultados y modelos obtenidos, se ha desarrollado un proyecto siguiendo una metodología adecuada para el estudio y centrándose, sobre todo, en el proceso de aprendizaje.

5.2. Trabajo futuro

Para futuros estudios priorizaría los siguientes puntos para obtener resultados más satisfactorios.

- Disponer de una mayor y más variada cantidad de datos que pueda dar lugar a un análisis más profundo y variado.

- Ampliar el estudio estadístico a conjuntos más amplios de atributos, ya que en este estudio sólo se estudia su relación en pares de atributos.

- Ampliar el estudio exploratorio de datos relacionando los valores más relevantes de los diferentes atributos con las muertes y lesiones.

- Debido a la falta de tiempo, los algoritmos de clasificación y las redes neuronales no se han ajustado adecuadamente hasta un punto que se considere óptimo.

- Estudiar el coste computacional de los distintos algoritmos para optimizar su tiempo de ejecución.

- Ampliar el número de técnicas utilizadas para crear modelos predictivos, especialmente para estudiar más modelos de redes neuronales.

Dado que el sector de los aerogeneradores está creciendo, creemos que las organizaciones gubernamentales y la comunidad académica deberían hacer más hincapié en la recogida y el análisis de datos sobre los accidentes de aerogeneradores. Este estudio ha aportado multitud de ideas y también ha esbozado algunas posibles sugerencias en relación con los accidentes de aerogeneradores. Estas ideas pueden servir de guía para el desarrollo de

diversos estudios en la industria de los aerogeneradores.

Capítulo 6

Introduction, conclusions and future work

6.1. Introduction

World energy demand is expected to grow by more than two-thirds during the period 2011-2035. This demand will be met by a mix of non-renewable (coal, fossil fuels, nuclear) and renewable (wind, hydro, solar, biomass, biofuel, geothermal) energy sources. The share of renewable energy sources in total electricity generation is expected to increase from 20 % in 2011 to 31 % in 2035, with renewables eventually overtaking gas and coal to become the world's leading energy source. This global trend due to the increased use of renewable energies is mainly driven by undesired global climate change due to carbon emissions as well as the depletion of fossil fuels. In addition, the notion of sustainability of renewable energy sources of renewable energy sources is driving governments to introduce legislation that promotes the use of renewable energy.

Wind energy has a long history and is currently among the leading renewable energy sources in terms of production capacity. According to 2013 market statistics published by the World Wind Energy Council, cumulative cumulative wind power capacity has more than tripled in six years.

For the realization of this analysis we rely on the information of 273 wind turbine accidents that have happened around the world and that we have collected from various sources as will be explained in future chapters.

In this study we work with two main concepts that form the basis of our statistical study.

Firstly, the stage of the life cycle of the wind turbine in which the accident occurred, in the image above we can see the possible stages in which an accident can occur, namely during transport, construction, operation and maintenance. Secondly, the cause of the wind turbine accident, i.e. nature, system/equipment and human. The association between these two categories of factors and two main effects (outcomes), i.e., death and injury, is investigated.

Thus, the main hypotheses of this work are as follows:

• Hypothesis 1. There is an association between deaths and predictor attributes (Model 1).

• Hypothesis 2. There is an association between injuries and predictor attributes (Model 2).

6.2. Conclusions and future work

6.2.1. Conclusions

During the development of this project we have obtained several conclusions based on the results of the exploratory analysis of real data and the statistical analysis of several wind turbine accidents.

Special emphasis was placed on the most characteristic attributes of these accidents and their relationship with the deaths and injuries that can occur in these accidents. These attributes are the life cycle of the wind turbine, whether the accident occurred during its construction, transport, operation or maintenance, and the cause of the accident, whether it was caused by nature, by an equipment defect or by human error.

The most relevant conclusions obtained in this study are the following:

• Equipment defects are the main cause of deaths and injuries in wind turbine accidents.

• The components that generate the most failures are the wind turbine tower and blades, which are more susceptible to cause accidents in the transport phase.

• The countries most likely to suffer an accident with fatalities are USA, Germany and China in that order. The countries most likely to have an accident with injuries are USA, UK and Australia in that order.

• Wind turbine farms of 37500Kw are more likely to have an accident than others of higher capacity.

One of the biggest problems we have faced in the study is the amount of data, as the data was obtained by scanning a public search engine and public news with relevant information is scarce. Therefore, our data are not complete and are only a sample. As in any study in which a population is sampled, there is a risk that our sample is not in fact a random sample that represents the true population.

In addition, we have built several predictive models using a variety of strategies. We have explored both supervised and unsupervised models with both obtaining acceptable margins of error. We have also worked briefly with simple neural networks which have

yielded very promising data.

For the first model, the one that studies deaths in accidents, the model that has obtained the best results is the Linear SVM followed by the CART decision tree and the neural network with functional structure. Clustering algorithms have lagged far behind. In the second model, the one that studies lesions, the model with the best results is Linear SVM followed by the neural network with functional structure and the CART decision tree. We can observe that these techniques are superior in both models. The CART decision tree and the linear SVM work well with limited amounts of data, so their effectiveness was expected, but the functional neural network stands out especially since in theory it works better with large amounts of data, so it can be deduced that in the future the amount of data to train the models will increase, this technique could improve its results even more.

Finally, and beyond the results and models obtained, a project has been developed following an appropriate methodology for the study and focusing, above all, on the learning process.

6.2.2. Future work

For future studies I would prioritize the following points in order to obtain more satisfactory results.

- To have a larger and more varied amount of data that could lead to a deeper and more varied analysis.

- To extend the statistical study to larger sets of attributes, since in this study we are only studying their relationship in pairs of attributes.

- To extend the exploratory data study by relating the most relevant values of the different attributes with deaths and injuries.

- Due to lack of time, the classification algorithms and neural networks have not been properly tuned to a point that is considered optimal.

- Study the computational cost of the various algorithms in order to optimize their execution time.

- To expand the number of techniques used to create predictive models, especially to study more neural network models.

As the wind turbine industry is growing, we believe that governmental organizations and the academic community should place more emphasis on collecting and analyzing data on wind turbine accidents. This study has provided a multitude of insights and has also outlined some possible suggestions regarding wind turbine accidents. These ideas can

serve as a guide for the development of various studies in the wind turbine industry.

Bibliografía

[1] Cagri Haksoz Sena Pakter Sobhan Asian, Gurdal Ertek and Soner Ulun. Wind turbine accidents: A data mining study.

[2] Allan N. Zhang Sobhan Asian Gurdal Ertek, Xu Chi. Text mining analysis of wind turbine accidents.

[3] Braam, h., rademakers, l. w. m. m. (2004). guidelines on the environmental risk of wind turbines in the netherlands (no. ecn-rx–04-013). energy research centre of the netherlands ecn.

[4] Faasen, c. j., franck, p. a. l., taris, a. m. h. w. (2014). handboek risicozonering windturbines. rijksdienst voor ondernemend nederland.

[5] Yasuda, y., fujii, t., yamamoto, k., honjo, n., yokoyama, s. (2014, october). classification of wind turbine blade incidents regarding lightning risk management. in 2014 international conference on lightning protection (iclp) (pp. 986-991). ieee.

[6] Python Documentation. https://docs.python.org/3/.

[7] Pandas data analysis library. https://pandas.pydata.org/docs/.

[8] Numpy vector support library. https://numpy.org/doc/stable/.

[9] SciPy statistical tools library. https://docs.scipy.org/doc/scipy/.

[10] SciKit Learn classifiers library. https://scikit-learn.org/stable/modules/classes.html.

[11] Kullback Leibler Divergency. https://es.wikipedia.org/wiki/divergencia$_d e_k ullback - leibler$.

[12] Kolmogorov–Smirnov test. https://en.wikipedia.org/wiki/kolmogorov %e2 %80 %93smirnov$_t est$.

[13] Chi squared test. https://en.wikipedia.org/wiki/chi-squared$_t est$.

[14] Student's t test. https://en.wikipedia.org/wiki/student %27s$_t - test$.

[15] Kruskal–Wallis test. https://en.wikipedia.org/wiki/kruskal %e2 %80 %93wallis$_o ne - way_a nalysis_o f_v ariance$.

[16] Shapiro–Wilk test. https://en.wikipedia.org/wiki/shapiro

[17] Linear SVM. https://scikit-learn.org/stable/modules/generated/sklearn.svm.linearsvc.htmlsklearn.svm.linearsvc.

[18] https://www.analyticsvidhya.com/blog/2021/10/support-vector-machinessvm-a-complete-guide-for-beginners/.

[19] Random Forest Classifiers. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.randomforestclassifier.html ?highlight=random %20forestsklearn.ensemble.randomforestclassifier.

[20] https://en.wikipedia.org/wiki/random$_{f}orest$.

[21] Logistic Regresion Classifier. https://scikit-learn.org/stable/modules/generated /sklearn.linear$_{m}odel.logisticregression.html?highlight$ = $logisticsklearn.linear_{m}odel.logisticregression$.

[22] https://es.wikipedia.org/wiki/regresi %c3 %b3n$_{l}og$ %$c3$ %$adstica$.

[23] K-Nearest Neighbor. https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.nearestneighbors.html? highlight=k %20nearestsklearn.neighbors.nearestneighbors.kneighbors.

[24] CART Decision tree. https://scikit-learn.org/stable/modules/generated/sklearn.tree.decisiontreeclassifier.html? highlight=treesklearn.tree.decisiontreeclassifier.

[25] K-Means clustering. https://scikit-learn.org/stable/modules/generated /sklearn.cluster.k$_{m}eans.html?highlight = k$ %$20meanssklearn.cluster.k_{m}eans$.

[26] Mean Shift clustering. https://scikit-learn.org/stable/modules/generated/sklearn.cluster.meanshift.html?highlight=mean sklearn.cluster.meanshift.

[27] Affinity Propagation. https://scikit-learn.org/stable/modules/generated/sklearn.cluster.affinitypropagation.html? highlight=affinitysklearn.cluster.affinitypropagation.

[28] Neural network Sequencial structure. https://www.tensorflow.org/api$_{d}ocs/python/tf/keras/sequential$.

[29] Neural network Functional structure. https://www.tensorflow.org/guide/keras/functional.

[30] Burnham, k. p., anderson, d. r. (2001). kullback-leibler information as a basis for strong inference in ecological studies. wildlife research, 28(2), 111-119.

UNIVERSIDAD COMPLUTENSE MADRID

AUTORIZACIÓN PARA LA DIFUSIÓN DEL TRABAJO FIN DE GRADO Y SU DEPÓSITO EN EL REPOSITORIO INSTITUCIONAL E-PRINTS COMPLUTENSE

Los abajo firmantes, alumno/a y tutores del Trabajo Fin de Grado (TFG) de la Facultad de Informática, autorizan a la Universidad Complutense de Madrid (UCM) a difundir y utilizar con fines académicos, no comerciales y mencionando expresamente a su autor el Trabajo Fin de Grado (TFG) cuyos datos se detallan a continuación. Así mismo, autorizan a la Universidad Complutense de Madrid a que sea depositado en acceso abierto en el repositorio institucional con el objeto de incrementar la difusión, uso e impacto del TFG en Internet y garantizar su preservación y acceso a largo plazo.

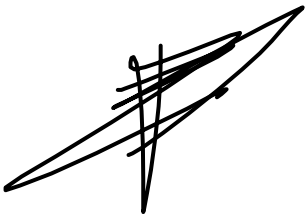
Periodo de embargo (opcional): ☐ 6 meses ☐ 12 meses

Título del TFG: **Accidentes en turbinas eólicas: Un estudio de minería de datos**
Curso académico: 21/22

Alumno/a
Pablo Jiménez Cruz

Tutores y su departamento
Matilde Santos Peñas (DACyA)
Ravi Pandit

Fecha y firma alumno/a



Fecha y firma tutores