

Machine Learning para la seguridad y eficiencia en el entorno naval:

Análisis predictivo de incidentes navales en EEUU, 2002 - 2015

Oscar Antón Pérez

Memoria del Trabajo Fin de Máster en
*Big Data y Data Science aplicados a
la economía y a la administración
y dirección de empresas (UNED)*

Diciembre 2023



Resumen

Este trabajo de fin de máster se centra en el análisis predictivo de accidentes navales. Parte de una base de datos de accidentes registrados por la Guardia Costera de Estados Unidos (USCG) durante los años 2002 y 2015, junto con datos meteorológicos proporcionados por la Oficina Nacional de Administración Oceánica y Atmosférica (NOAA). Esta base de datos contiene características de los incidentes, así como de las embarcaciones implicadas y meteorología registrada.

Para ello, se van a aplicar diferentes técnicas enmarcadas en la ciencia de datos. En concreto, a través de la comparativa entre diversos algoritmos de clasificación, se pretende aportar información y herramientas estadísticas útiles que permitan avanzar en la prevención de riesgos vitales, medioambientales y financieros en el entorno marino.

Abstract

This master's thesis focuses on the predictive analysis of naval accidents. It is based on a database of accidents recorded by the United States Coast Guard (USCG) from 2002 to 2015, along with meteorological data provided by the National Oceanic and Atmospheric Administration (NOAA). This database contains incident characteristics, as well as information about the involved vessels and recorded meteorological conditions.

To achieve this, various techniques within the field of data science will be employed. Specifically, by comparing different classification algorithms, the study aims to provide information and useful statistical tools that contribute to advancing the prevention of life, environmental, and financial risks in the maritime environment.



*Esta obra está sujeta a una licencia Creative Commons BY-NC-ND 4.0 DEED
Atribución-NoComercial-SinDerivadas 4.0 Internacional*
<https://creativecommons.org/licenses/by-nc-nd/4.0/>

Índice

1. Introducción	1
2. Objetivos.....	2
3. Bases de datos	3
3.1. USCG: Marine Casualty & Pollution Data for Researchers	3
3.2. NOAA: Global Marine data	5
3.3. NOAA: Daily Summaries	6
3.4. Fusión de datos: obtención de los datasets.....	6
4. Análisis Exploratorio de Datos (EDA).....	12
4.1. Dataset VesselBalancedSample	12
4.1.1. Características de los barcos	13
4.1.2. Aspectos legales y normativos	15
4.1.3. Características de los incidentes.....	16
4.1.4. Correlaciones.....	17
4.2. Dataset MergedActivity.....	18
4.2.1. Localización	19
4.2.2. Cronología	21
4.2.3. Meteorología.....	22
4.2.4. Características de los barcos	23
4.2.5. Correlaciones.....	24
5. Aplicación de modelos de Machine Learning	25
5.0. Marco teórico	27
5.0.1. Modelos de machine learning	27
5.0.2. Métricas	30
5.1. Posibilidad de incidente: VesselBalancedSample.....	32
5.1.1. Comparativa de los modelos “Posibilidad de incidente”	32
5.1.2. Explicabilidad de los modelos “Posibilidad de incidente”	36
5.2. Tipo de incidente: MergedActivity.....	38
5.2.0. Limpieza de datos: Equilibrado y tratamiento de valores ausentes	38
5.2.1. Comparativa de los modelos “Tipo de incidente”	41
5.2.2. Explicabilidad de los modelos “Tipo de incidente”	44
6. Conclusiones	45
7. Bibliografía.....	46

1. Introducción

La industria marítima ha desempeñado desde siempre un papel vital en la economía a nivel global. Debido al medio en que se desarrolla su actividad, la seguridad ha sido siempre uno de sus principales desafíos.

Los incidentes navales conllevan consecuencias potencialmente catastróficas a nivel vital, medioambiental y financiero. Es por esto que su prevención ha sido fuente de estudio desde antiguas civilizaciones como Antiguo Egipto, el Imperio Romano, China, Antigua Grecia o los pueblos vikingos.

Desde entonces, se han utilizado registros navales a los que se han ido aplicando sucesivos avances matemáticos, estadísticos y científicos para prevenir incidentes. Siguiendo esta senda, hoy en día se puede contar con la computación. Y llegando a la punta de este iceberg, está el big data y la ciencia de datos. Por lo que, paradójicamente, este trabajo pretende resolver una cuestión muy antigua con las metodologías y técnicas más desarrolladas hasta ahora.

Al aprovechar la riqueza de datos recopilados por la Guardia Costera de Estados Unidos y complementarlos con información meteorológica proporcionada por la Oficina Nacional de Administración Oceánica y Atmosférica, este estudio busca no solo analizar incidentes pasados, sino también anticipar y mitigar riesgos futuros

A lo largo de estas páginas, se pondrá de manifiesto la importancia de recopilar y procesar datos relevantes de incidentes pasados, así como la implementación de algoritmos avanzados de aprendizaje automático y análisis de datos para pronosticar posibles incidentes en el futuro.

El análisis de datos se va a desarrollar con la ayuda del lenguaje R dentro del entorno de Rstudio. La elección de esta herramienta se sustenta en su capacidad especializada para el análisis de datos y la implementación de algoritmos de aprendizaje automático que son desarrollados y testados por una activa comunidad científica. Además, su capacidad para generar documentación reproducible mediante R Markdown, hacen que sea una buena elección para dar a conocer los resultados de una manera muy clara.

La transmisión de información obtenida en el análisis, se instrumentaliza mediante la presente memoria, así como sus correspondientes anexos que contienen el código utilizado en RStudio. Los cuales se detallan a continuación:

- Anexo 3.1. Preprocesado All Casualty & Pollution
- Anexo 3.2. Preprocesado Weather Ocean
- Anexo 3.3. Preprocesado Weather River
- Anexo 3.4. Preprocesado Merged Activity y VesselBalancedSample

- Anexo 4.1. EDA VesselBalancedSample
- Anexo 4.2. EDA MergedActivity

- Anexo 5.0. Funciones de R

- Anexo 5.1. Modelado VesselBalancedSample
- Anexo 5.2. Modelado MergedActivity

La numeración corresponde a los apartados de esta memoria y serán referenciados cuando corresponda para facilitar su comprensión.

2. Objetivos

En este trabajo se van a perseguir objetivos generales tanto en la materia de estudio (seguridad en el mar), como en el ámbito de la ciencia de datos.

Por un lado, se pretende aportar información precisa para la seguridad y eficiencia tanto de tripulaciones, como de los activos utilizados en el entorno marítimo. En concreto, se quiere identificar las condiciones en las cuales se han dado más incidentes durante el periodo del estudio, de tal manera que se pueda extrapolar al presente y a otras zonas del mundo mediante modelos predictivos. Principalmente, se pretende dar respuesta a estas dos cuestiones para un determinado barco:

1. ¿Puede verse envuelto en algún accidente?
2. Y en ese caso, ¿en qué tipo de accidente?

Las herramientas para dar respuesta a esas preguntas van a ser modelos predictivos de clasificación. Para la primera cuestión, se va a utilizar una variable binaria que clasifica si un barco está implicado en un incidente o no. Y para la segunda cuestión, será a través de una variable que recoge el tipo de accidente.

Para este análisis, se ha considerado un alto volumen de datos. De este modo, se van a poner a prueba la efectividad de planteamientos estadísticos y probabilísticos como los modelos bayesianos, las redes neuronales, Gradient Boosting, etc. que más adelante serán expuestos. Se pretende que estos modelos sean capaces de predecir los posibles incidentes con la entrada de nuevos datos reales, por lo que se maximizará el ajuste de cada modelo con los parámetros más adecuados.

En cuanto a los objetivos operativos para la consecución de este trabajo, podemos señalar los siguientes puntos:

- Obtención de datos. Los datos sobre los incidentes navales, se encuentran a disposición pública en la web de Guardia Costera de Estados Unidos (USCG) y los datos meteorológicos están igualmente disponibles en la web de la oficina estadounidense de administración oceánica y atmosférica (NOAA).
- Preprocesado. Como se verá más adelante, hay una importante labor en este apartado, ya que se han tenido que unir datos de tres fuentes independientes, así como realizar una limpieza y filtrado exhaustivo para obtener las bases de datos.
- Exploración estadística. Se trata una parte importante para cualquier análisis predictivo. Pero en este caso, se trata de uno de los principales objetivos del análisis, ya que la información de datos históricos es especialmente valiosa en la prevención de eventos futuros.
- Modelado. Es uno de los motivos centrales de este trabajo, representando el grueso de información expuesta. Se han implementado varios algoritmos, comparándose entre sí mediante las métricas de desempeño más habituales en este tipo de análisis. Para comprender cómo han trabajado, en este apartado se tratará su explicabilidad.
- Conclusiones. Como consecuencia de los anteriores aspectos y teniendo en cuenta que esta información debe ser útil para personas sin conocimientos profundos en ciencia de datos, se ha sintetizado la información obtenida en el modelado.

3. Bases de datos

Las bases de datos que se han utilizado para este análisis proceden de tres fuentes diferentes, proporcionadas por organismos gubernamentales de Estados Unidos:

- USCG: Marine Casualty & Pollution Data for Researchers ([link](#))
- NOAA: Global Marine data ([link](#))
- NOAA: Daily Summaries ([link](#))

La elección de estos datos obedece a criterios de disponibilidad. En primera instancia, se intentó conseguir los datos de seguridad marítima en Europa. Pero la autoridad que los custodia, EMSA, se negó a compartirlos con fines académicos alegando cuestiones de protección de datos. Tras una búsqueda a “puerta fría”, en segundo lugar, se evaluaron los datos recogidos por el Transportation Safety Board of Canadá para su país. Sin embargo, al ser evaluados, se comprobó que la calidad de sus datos no era satisfactoria. Finalmente, fueron los datos publicados por la Guardia Costera de Estados Unidos, los que cumplieron las condiciones para satisfacer los objetivos marcados.

Estos conjuntos de datos han sido objeto de una reducción de información antes de ser combinadas entre sí en los dos datasets utilizados para responder a cada una de las preguntas anteriores:

1. VesselBalancedSample: Contiene características de barcos involucrados en accidentes, así como de los no involucrados. De esta manera se pretende extraer las características de los barcos accidentados.
2. MergedActivity: Contiene información tanto de los incidentes como lo de los barcos involucrados en los mismos. Con este conjunto de datos se pretende extraer las características de cada tipo de accidente

Para la reducción de información antes mencionada, por un lado, se ha acotado geográficamente la información para obviar aquellos incidentes fuera de jurisdicción estadounidense, ya que no son representativos dentro del conjunto y podrían distorsionar el análisis. En este sentido, hay que tener en cuenta que van a evaluar los incidentes acontecidos dentro del cuadro delimitador (*o bounding box*) con las siguientes coordenadas: Latitud de 15°N a 70°N, Longitud de -180°E a -45°E.

Por otro lado, se ha realizado una preselección de variables para evitar el manejo de variables poco relevantes por su descripción o variabilidad. El proceso seguido, queda recogido en un diagrama que será presentado al final de este apartado en forma de resumen.

Todo el código utilizado en estas tareas, está disponible en los anexos de esta memoria. A continuación, se explicarán origen, contenidos y preselección de datos de estas tres fuentes de datos.

3.1. USCG: Marine Casualty & Pollution Data for Researchers

Los datos principales del análisis se han obtenido de la web de Guardia Costera de Estados Unidos (USCG). En concreto, tienen un apartado dedicado a la publicación de datos para la investigación: Marine Casualty & Pollution Data for Researchers. En este apartado, han publicado dos datasets: uno hasta diciembre de 2001 y otro desde enero de 2002 hasta julio de 2015. Para el estudio, se ha elegido este último, entendiéndose que los datos más cercanos tendrán un mayor valor a la hora de extrapolarse al presente. Desde inicios del siglo XX, se han producido diversas actualizaciones en la

reglamentación sobre la construcción, operación y medio ambiente. Por citar algún ejemplo, tenemos el anexo IV del convenio MARPOL entró en vigor en septiembre de 2003, marcando las pautas de descarga de aguas sucias o la obligatoriedad del doble casco para buques petroleros.

Este conjunto de datos, contiene 10 tablas en formato de texto plano (.txt), presentadas en un contenedor .zip de 82,5Mb, que al descomprimirse ocupan un volumen de 623 Mb. El código de R utilizado para su preprocesado se puede encontrar en el Anexo 3.1. Con el objetivo de centrar este análisis en los datos de incidentes, se han seleccionado las siguientes cuatro tablas:

· **MisleVslEvents:** Contiene variables que describen los incidentes navales que son objeto de este análisis. Se han preseleccionado 14 variables:

- | | |
|----------------|--------------|
| - activity_id | - waterway |
| - vessel_id | - event_type |
| - vin | - damage_sta |
| - vessel_name | - latitude |
| - vessel_class | - longitude |
| - flag_desc | - date |
| - vsl_act_role | - hour |

· **MisleVessel:** Contiene variables referentes a las características técnicas y operativas de las embarcaciones implicadas en los incidentes registrados. Respecto a la preselección de variables, se han elegido las siguientes 25 variables:

- | | | |
|--------------------------|-------------------------|----------------------|
| - vessel_id | - solas_desc | - build_shipyard |
| - vessel_name | - vessel_class | - build_year |
| - gross_ton | - propulsion_type | - horsepower_ahead |
| - length | - hull_material | - horsepower_astern |
| - breadth | - hull_design_type | - forebody_type_desc |
| - depth | - hull_double_bottom_t | - hull_configuration |
| - dead_weight_ton | - hull_double_side_type | - hull_shape |
| - flag_abbr | - primary_vin | |
| - classification_society | - imo_number | |

· **MisleVslPoll:** Contiene variables sobre los incidentes con consecuencias sobre el medio ambiente (polución). En este caso, se han preseleccionado las siguientes 14 variables:

- | | |
|----------------|-----------------|
| - activity_id | - waterway |
| - vessel_id | - chris_cd |
| - vin | - latitude |
| - vessel_name | - longitude |
| - vessel_class | - discharge_ |
| - flag_desc | - damage_status |
| - vsl_act_role | |

· **MisleInjury:** Contiene las variables de los incidentes con consecuencias sobre seres humanos (heridos y decesos). Por último, las variables preseleccionadas han sido estas 13:

- | | |
|----------------|-----------------|
| - activity_id | - relationship |
| - vessel_id | - waterway |
| - vin | - accident_type |
| - vessel_name | - casualty_type |
| - vessel_class | - latitude |
| - flag_desc | - longitude |
| - vsl_act_role | |

3.2. NOAA: Global Marine data

Al evaluar la información contenida en el dataset anteriormente descrito, se echó en falta los datos referentes a las condiciones meteorológicas durante cada incidente registrado. En el entorno naval es conocido que malas condiciones climatológicas pueden causar catástrofes. Tres ejemplos clásicos, pueden ser:

- El hundimiento del RMS Titanic en 1912: Este trágico suceso es uno de los naufragios más famosos. El Titanic chocó con un iceberg en el Atlántico Norte durante una noche de niebla, lo que llevó al hundimiento del lujoso trasatlántico y a la pérdida de más de 1500 vidas.
- La Armada Española de 1588 (La Armada Invencible): La Armada Española, una flota enviada por el rey Felipe II de España con el objetivo de invadir Inglaterra, se vio gravemente afectada por una serie de tormentas en las aguas alrededor de las Islas Británicas. Las tormentas causaron la pérdida de numerosas naves y vidas, lo que influyó significativamente en el fracaso de la invasión.
- La tormenta de 1703 en el Canal de la Mancha: Esta tormenta fue una de las más intensas en afectar el sur de Inglaterra. Conocida como "La Gran Tempestad", provocó la pérdida de cientos de barcos y miles de vidas.

Debido a que la información meteorológica es especialmente relevante en la navegación, se ha decidido incluirla para este análisis. En concreto, se han incluido los datos de las balizas y sondas marinas gestionadas por la National Oceanic and Atmospheric Administration (NOAA). Estas sondas recogen datos como temperatura, altura de ola, presión atmosférica, precipitación, dirección e intensidad de viento, etc. ya sea en localizaciones fijas en el mar (boyas) o en itinerancia si están montadas en barcos (estaciones móviles).

La publicación de la NOAA para esta cuestión, consiste en archivos .csv para cada estación, agrupados y comprimidos mensualmente en .tar.gz. En este caso, se han considerado los 168 archivos que comprenden los años de 2002 a 2015, con un peso conjunto de 4,42Gb que, al ser descomprimidos, originan un total de 79.677 archivos .csv con un peso total de 56,84Gb en 97.958.906 observaciones. El código utilizado para el preprocesado de estas tablas se puede encontrar en el Anexo 3.2.

Sobre este conjunto de datos, se han preseleccionado las siguientes 9 variables:

- | | |
|-------------|--------------|
| - STATION | - WIND_SPEED |
| - DATE | - VISIBILITY |
| - LATITUDE | - AIR_TEMP |
| - LONGITUDE | - WAVE_HGT |
| - PAST_WX | |

Ya que las estaciones registran datos varias veces al día y muchos de los incidentes carecen de registro de hora concreta, para simplificar la gestión del conjunto, se ha optado por resumir en un solo registro diario, calculando la media de valores durante del día para variables continuas y la moda para variables discretas.

Más adelante, se comprobará que mientras algunas variables aportan información relevante, el análisis estadístico establecerá que para otras no es tanto así.

3.3. NOAA: Daily Summaries

Los incidentes navales registrados por la USCG abarcan tanto los acontecidos en el mar como en líneas fluviales (ríos). El Global Marine Data descrito anteriormente no contiene datos para ríos. Así que, para cubrir este segundo ámbito, se han incluido los datos meteorológicos de estaciones terrestres a través de la publicación Daily Summaries de la NOAA.

En concreto, las áreas fluviales que se van a explorar son: Mississippi y Ohio. Estas áreas cuentan con unas características diferentes a las encontradas en el mar. Por ejemplo, son conocidos los bancos de arena del río Mississippi que provocan encallamiento o hundimientos de barcos. Además, la ausencia de oleaje y salinidad, influyen en los requerimientos de construcción de los barcos. Los materiales utilizados, diseño y tamaño de casco de embarcaciones destinadas a ríos son, en general, diferentes a los barcos destinados a operar en mar. Por este motivo, los incidentes marinos y fluviales van a ser analizados por separado.

Los datos de Daily Summaries proporcionados por las estaciones terrestres de la NOAA, incluyen datos similares a los anteriores, exceptuando la altura de ola que, en este caso, no tendría sentido.

El formato presentado por la NOAA son archivos .csv.gz de conjuntos anuales. Se han recopilado 14 archivos (del año 2002 a 2015), con un peso total de 2,17Gb. El código utilizado para su preprocesado está disponible en el anexo 3.3. Al ser descomprimidos, se han obtenido 18,76Gb de datos: 154.188.126 observaciones con estas 8 variables preseleccionadas:

- | | |
|-----------|-------------|
| - STATION | - PRCP |
| - DATE | - AWND |
| - TMAX | - LONGITUDE |
| - TMIN | - LATITUDE |

La acotación geográfica para Mississippi, ha sido: Latitud de 30.5°N a 45°N, Longitud de -88°E a -94°E. Para Ohio, corresponden los siguientes valores: Latitud de 37.5°N a 41°N y Longitud de -79.5°E a -88°E, se ha reducido a 12.281.982 observaciones.

3.4. Fusión de datos: obtención de los datasets

Para obtener los dos datasets que sirven de base para el análisis, se han combinado los datos de las tres fuentes descritas anteriormente, utilizando como base la tabla MisleVsIEvents. El código de R utilizado para esta labor se encuentra disponible en el anexo 3.4. A este punto, sería recomendable tener presente el diagrama de preprocesado que se presenta al final de este apartado para ver las relaciones entre tablas, así como el descarte de variables realizado en cada paso.

Con el objetivo de facilitar el procesado de datos de una manera más ágil, se ha creado una tabla intermedia “Events” que recoge los datos filtrados de eventos sin repetir y excluyendo aquellos eventos que no se han producido en las principales zonas de este análisis: zonas oceánicas de EEUU y cuenca fluvial del río Mississippi. Se ha utilizado la librería *maturalearth* para delimitar esas zonas y que coincidan con los datos meteorológicos.

Tanto los incidentes navales como los datos meteorológicos están geolocalizados por coordenadas. Estas coordenadas no coinciden exactamente, por lo que, para asignar

condiciones meteorológicas a cada incidente, se ha calculado la distancia a la estación más cercana mediante la siguiente aproximación:

$$Distancia = \sqrt{(Lat_1 - Lat_2)^2 + (Long_1 - Long_2)^2}$$

Una vez realizadas las acotaciones geográficas, la preselección de variables y la agregación vertical de las observaciones, se han obtenido los dos conjuntos de datos señalados anteriormente:

1. **VesselBalancedSample:** Son 109.836 observaciones con datos de 54.198 barcos implicados en incidentes y otros 54.198 barcos sin incidentes registrados. Los datos de los 54.198 barcos involucrados en incidentes proceden de registros únicos de vessel_id en la tabla Events. Sobre la tabla Vessel, se ha realizado una diferencia exclusiva para obtener los vessel_id que no son comunes entre ambas tablas, entendiendo que los barcos que sí aparezcan en Vessel pero no aparezcan en Events, están exentos de incidentes. Para no sesgar la muestra, se ha realizado una selección aleatoria de los barcos sin incidente del mismo tamaño que el total de los envueltos en incidentes. El resultado ha sido una muestra con 50% de barcos con incidente + 50% de barcos sin incidente. Contiene 12 variables que, en esta ocasión, van a ser descritas y sometidas a un análisis estadístico descriptivo en el siguiente apartado de este documento:

- vessel_id: Identificación unitaria del barco, asignada por el USCG. Es utilizada a efectos de esta base de datos, siendo este dato único para cada barco. Este dato no es utilizado fuera de este ámbito por lo que es útil para el análisis, pero no para exportar información de este.

- imo_number: Identificación unitaria del barco, asignada para el IMO (International Maritime Organization). Esta identificación tiene carácter universal y es frecuentemente utilizada por diversos organismos y empresas para identificar el nombre. Cabe mencionar que este dato no varía durante la vida del barco, pero no todos los barcos tienen obligación de tenerlo. Esto provoca ausencias en la base de datos con relativa frecuencia. En general, los barcos obligados a tener número IMO (Organización Marítima Internacional) son aquellos que cumplen alguno de los siguientes criterios:

- Buques mercantes de 300 toneladas de registro bruto (GT) o más.
- Buques de pasajeros de 100 GT o más.
- Buques de carga de más de 500 GT que transportan sustancias peligrosas o materiales nocivos.
- Plataformas móviles de perforación offshore.

Barcos de guerra, buques de pesca y pequeñas embarcaciones de recreo generalmente no están obligados a llevar el número IMO.

- vessel_name: Nombre del barco. Este dato facilita una identificación rápida, pero al poder ser cambiado durante la vida del mismo, no es mejor criterio edificativo. Incluso podrían aparecer nombres muy similares entre sí por conveniencia, así que este dato es meramente informativo.

- vessel_class: Tipo de barco. Se han diferenciado los siguientes tipos:

- | | | |
|---------------------|----------------------|-----------------|
| · Barge | · Passenger | · School Ship |
| · Bulk Carrier | · Recreational | · Tank Ship |
| · Fishing Vessel | · Refrigerated Cargo | · Towing Vessel |
| · General Dry Cargo | · Research Ship | · UNSPECIFIED |
| · Miscellaneou | · Ro-Ro Cargo Ship | · Warship |

- Offshore
- Passenger
- School Ship

- **build_year**: Año de construcción. Con esta variable se podrá calcular la antigüedad al momento del incidente.

- **gross_ton**: Gross Tonnage (GT) o registro bruto, mide el volumen total de un buque, incluyendo espacios internos. Es una métrica para evaluar su tamaño, pero también determina otros parámetros como legislación aplicable, tamaño de tripulación, etc.

- **length**: Eslora o largo del buque medida en pies. Otra medida de tamaño que describe la embarcación. Se ha optado por obviar la manga (o ancho) ya que existe una relación directamente proporcional entre ambas magnitudes.

- **flag_abbr**: Flag Abbreviation. Se trata de la bandera bajo la que opera. La bandera determina las leyes aplicables, regulaciones y protecciones a cada barco. Por lo tanto, se puede evaluar cómo afectan legislaciones más o menos restrictivas a la salud de cada embarcación.

- **class_soc**: Classification Society. Las sociedades de clasificación proporcionan servicios de certificación para garantizar que los buques cumplan con estándares de seguridad y calidad. Estos certificados son cruciales para la aceptación global y la operación segura de las embarcaciones.

- **solas_desc**: Adhesión a SOLAS, convenio Internacional para la Seguridad de la Vida Humana en el Mar, establece normas de seguridad marítima para prevenir accidentes y salvaguardar vidas en buques internacionales. Se entiende que, si un buque está bajo este convenio, sus requerimientos de seguridad cumplen los estándares universalmente aceptados.

- **event_type**: Tipo de evento. Se trata de la variable objetivo y, por lo tanto, la más importante de este conjunto de datos. Toma valor "No event" para los barcos sin incidentes registrados y para los accidentados los siguientes valores:

- | | | |
|---------------------|------------------------|---------------------|
| · Grounding | · Sinking | · Loss of Stability |
| · Environment | · Emergency | · Damage to Cargo |
| · Failure (Vessels) | · Loss of Power | · Explosion |
| · Allision | · Evasive Maneuvers | · Failure (Diving) |
| · Set Adrift | · Failure (Nonvessels) | · Blowout |
| · Maneuverability | · Fouling | · Falls into Water |
| · Collision | · Abandonment | · Implosion |
| · Flooding | · Capsize | · Loss of Stability |
| · Fire | · Sinking | · Damage to Cargo |

- **damage_status**: Indica si el barco ha sido dañado o no. Son frecuentes los incidentes en los que un barco se ve implicado y no tienen como consecuencia daños propios, por lo que se entiende relevante que este aspecto se evaluado.

2. **MergedActivity**: Se trata de 68.000 observaciones con información detallada de los incidentes registrados. Esta tabla condensa los datos de características de los barcos junto la descripción de los incidentes y la información meteorológica. En este sentido, la mejor opción hubiese sido contar con información detallada de eventos "sin incidencia" para establecer un nivel "sin incidente". Sin embargo, de los datos de barcos sin incidencia, no existen registros de posición y meteorología asequibles. En cierta fase de este análisis, se planteó generar datos sintéticos, pero fue descartado porque estos datos corresponderían al

100% de una clase entera de la variable predictiva, impidiendo el objetivo de traslado y aplicación real del análisis. Este dataset contiene 28 variables, que al igual que la tabla anterior, van a ser descritas y analizadas estadísticamente más adelante:

- **activity_id**: Identificador otorgado por la USCG para cada incidente registrado. Se incluye esta variable como nexo de unión entre los **vessel_id** implicados en un mismo incidente.

- **date**: Fecha del incidente. La fecha del año determina la climatología y actividad comercial. En este caso, se ha simplificado al mes cuando se produjo el hecho para reducir variabilidad.

- **hour**: Hora de incidente. Como en el caso anterior, la hora determina mayor o menor exposición a actividad. Para reducir variabilidad, se ha prescindido de los minutos.

- **region**: Región donde se ha producido el incidente. Calculada en función de las coordenadas de geoposicionamiento, se distinguen las siguientes:

· Alaska	· East Coast	· Mississipi
· Canadá	· Gulf of Mexico	· West Coast

- **longitude**: coordenada de longitud, medida en grados.

- **latitude**: coordenada de latitud, medida en grados.

- **watertype**: Tipo de medio acuático: River o Ocean. Variable calculada gracias a la librería *rnaturalearth*. Para ello, se han obtenido los datos del polígono formado por EEUU y sus países colindantes. Los puntos de longitud y latitud dentro del polígono, se han considerado river; Ocean es todos los demás.

- **event_type**: Tipo de incidente registrado. En este caso, al no haber observaciones sin incidentes, no hay el valor "No event".

- **damage_status**: *(ver descripción anterior)*

- **vessel_id**: *(ver descripción anterior)*

- **imo_numer**: *(ver descripción anterior)*

- **vessel_name**: *(ver descripción anterior)*

- **vessel_class**: *(ver descripción anterior)*

- **build_year**: *(ver descripción anterior)*

- **age**: Antigüedad de la embarcación en el momento del incidente, calculada como fecha del incidente menos fecha de construcción (**build_year**).

- **gross_ton**: *(ver descripción anterior)*

- **length**: *(ver descripción anterior)*

- **flag_abbr**: *(ver descripción anterior)*

- **classif_society**: *(ver descripción anterior)*

- **solas_desc**: *(ver descripción anterior)*

- **air_temp**: Temperatura ambiental registrada durante el incidente, medida en grados Fahrenheit (°F). Temperaturas bajas dan una idea de condiciones más adversas.

- wind_speed: Velocidad del viento, medida en nudos. Igualmente, valores altos de esta variable, implicarían peores condiciones en el mar, pero también en río.
- wave_hgt: Altura de ola, medida en pies. Este valor solo se ha recogido para los incidentes en mar, ya que en río se entiende que no es significativo.
- visibility: Visibilidad según la escala WMO (Organización Meteorológica Mundial). Con una escala de 90 a 99, mide la distancia que es visible a simple vista, siendo 90 el equivalente a 0.05km y con 99, sería hasta más de 50km.
- damage_assessment: Valor económico en dólares de los daños causados por cada accidente. La información es recogida por las comisiones de investigación a partir de datos proporcionados por los seguros implicados.
- casualty: Descripción de daños personales. Principalmente este valor puede indentificarse con muerte, herido o pérdida,
- pollution: Tipo de sustancia derramada al agua.
- event_class: Clase de incidente producido: Critical Events / Onboard Emergencies / Maritime Accidents / Material Issues / Third-party Damages. Se trata de una convergencia de la variable event_type ya que contiene hasta 26 niveles que estaban dificultado el procesado de datos. Para llegar a esta convergencia, se ha establecido la siguiente regla de carácter conceptual:
 1. Critical Events:
Sinking, Implosion, Capsize, Loss of Stability, Vessel Maneuverability, Set Adrift, Abandonment.
 2. Onboard Emergencies:
Loss of Electrical Power, Fire, Emergency Response, Explosion, Flooding, Personnel Casualties, Falls into Water.
 3. Maritime Accidents:
Grounding, Allision, Collision.
 4. Material Issues:
Material Failure Vessels, Material Failure Non-vessels, Material Failure Diving, Blowout.
 5. Third-party Damages:
Damage to the Environment, Damage to Cargo, Fouling, Evasive Maneuvers, UNSPECIFIED

A continuación, como resumen de este aparatado, se presenta un diagrama en el que se exponen las características y dependencias de las tablas utilizadas para llegar a los dos conjuntos de datos en los que se basa el análisis:

Diagrama de preprocesado

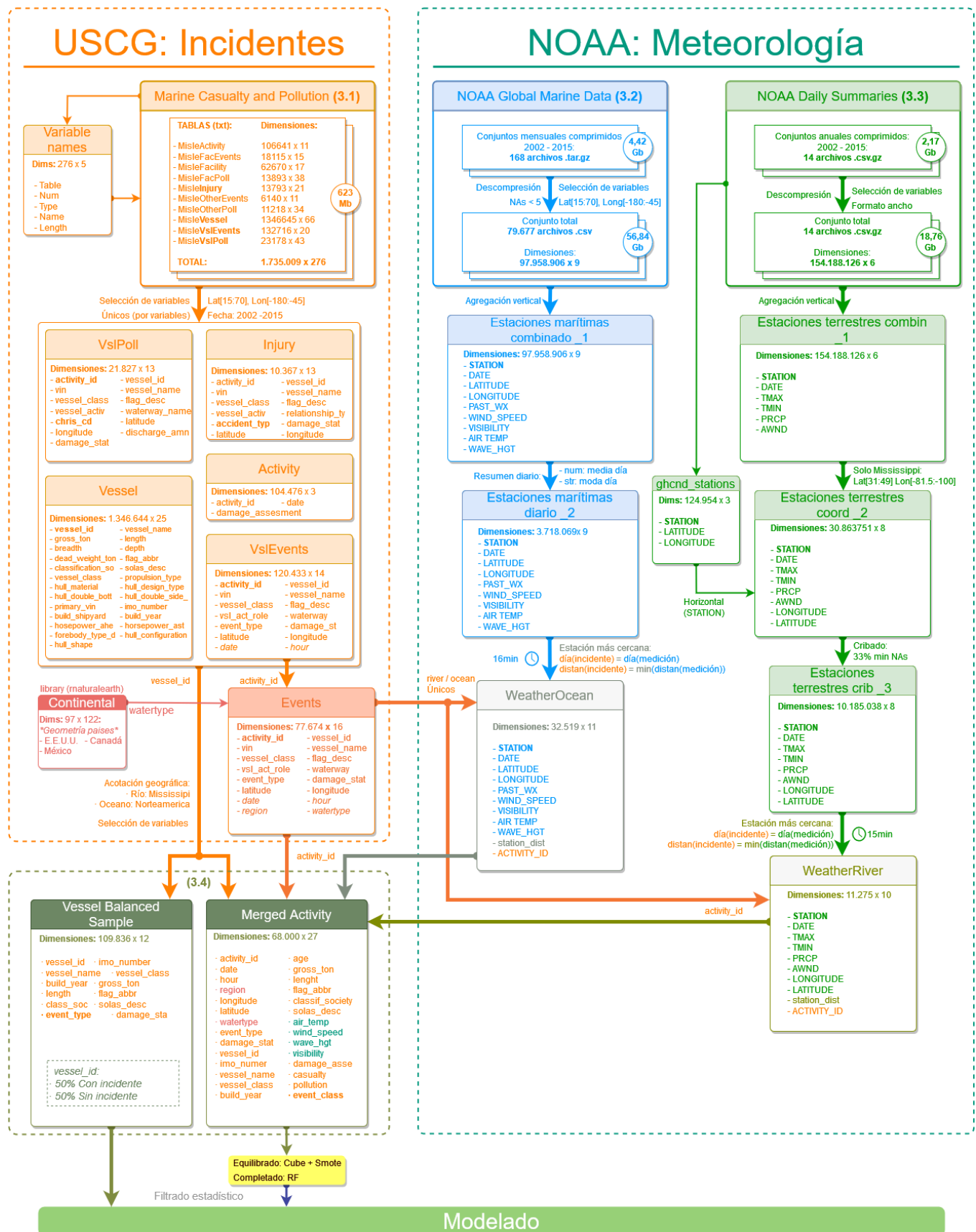


Fig.3.4. Diagrama Preprocesado

4. Análisis Exploratorio de Datos (EDA)

Para hacer un retrato lo más fiel posible de los datos analizados, se van a presentar diferentes aspectos estadísticos de los conjuntos de datos. Estos aspectos van a ser instrumentalizados a través de gráficas y tablas obtenidas en RStudio. El código utilizado está disponible en los anexos finales de esta memoria.

4.1. Dataset VesselBalancedSample

Este conjunto de datos está formado por 109.836 observaciones con datos de 54.198 barcos implicados en incidentes y otros 54.198 barcos sin incidentes registrados. Con estos datos, se pretende establecer un modelo que prediga si un determinado barco va a estar implicado en un incidente. A continuación, se expondrán los resultados más destacables de la exploración estadística de sus variables, pero el código completo se podrá encontrar el anexo 4.1.

Sumario estadístico (librería skimr)




Data summary										
Name	VesselBalancedSample									
Number of rows	109836									
Number of columns	12									
Key	NULL									
Column type frequency:										
character	9									
numeric	3									
Group variables										
None										
Variable type: character										
skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace			
imo_number	0	1	0	7	90591	9728				
vessel_name	0	1	0	50	5	62340				
vessel_class	0	1	5	23	0	16				
build_year	0	1	4	4	0	131				
flag_abbr	0	1	0	2	50	152				
classification_society	0	1	6	58	0	41				
solas_desc	0	1	9	16	0	3				
event_type	0	1	4	30	0	27				
damage_status	0	1	7	35	0	5				
Variable type: numeric										
skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
vessel_id	0	1	340600.76	299164.28	18.0	119634.2	241684	459737	1348019.0	
gross_ton	0	1	2946.08	11476.77	1.0	16.0	67	734	234627.0	
length	0	1	136.59	174.86	6.8	36.3	60	195	1203.8	

Fig. 4.1. VBS Summary

Uno de los datos destacables es que, en este caso, todas las variables tienen datos completos, lo cual resulta especialmente favorable para la aplicación de los algoritmos de aprendizaje automático.

4.1.1. Características de los barcos

Tipo de barco (vessel_class)

¿Qué tipos de barco se han contemplado en el análisis? A continuación, un gráfico de barras con tipo de embarcación, ordenadas de mayor a menor frecuencia y distinguiendo aquellas embarcaciones que han sido objeto de incidentes:

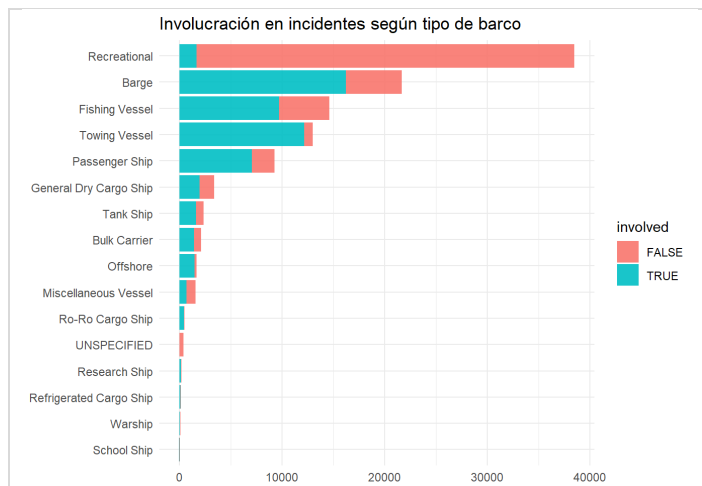


Fig. 4.1.1.1.: VBS vessel_class

Se puede ver que el tipo de barco más frecuente es el recreacional, con casi 4000 barcos registrado.

Sin embargo, son los menos afectados por incidentes

Año de botadura (build_year)

¿Cuál es la antigüedad de estos barcos? Se puede verificar con el siguiente gráfico de barras en el que se ha distinguido la involucración en accidentes

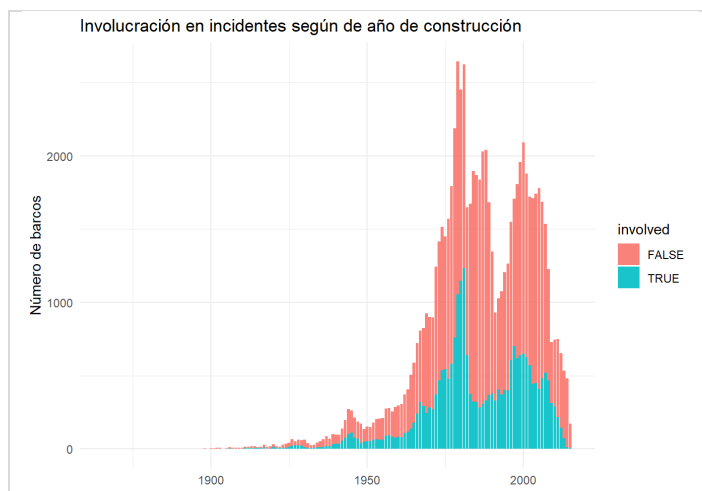


Fig. 4.1.1.2.: VBS build_year

En este caso, se puede ver que la mayoría de embarcaciones se concentran en torno a los inicios de los años 80 y tras una caída, vuelven a crecer en torno al año 2000.

La distribución de afectados por incidentes / no afectados, tiene una tendencia similar

Tamaño: Gross Tonnage (Gross Ton) y eslora (length)

En cuanto a las dimensiones, por un lado, vamos a ver el Gross Tonnage (o Tonelaje bruto), que se refiere al volumen del barco, dividido en cuatro tramos ya que hay muchos barcos pequeños que impiden valorar correctamente la imagen de conjunto completo

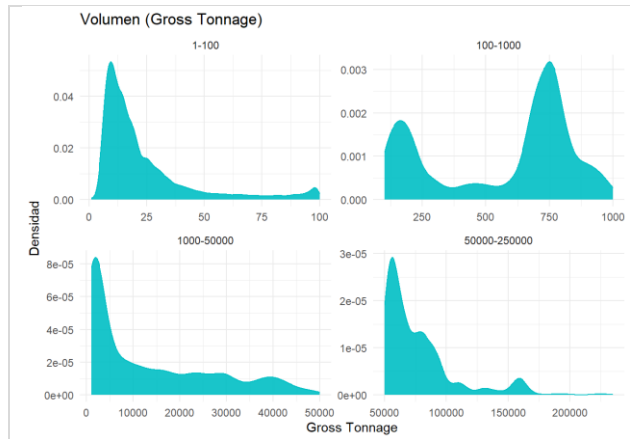


Fig. 4.1.1.3.: VBS Gross Tonnage

Aquí la mayoría se concentran en las 0 y las 25 GT, habiendo un notable pico en 750GT

En cuanto a la eslora, vamos a presentar dos gráficos de densidad con dos tramos para resaltar los valores más altos

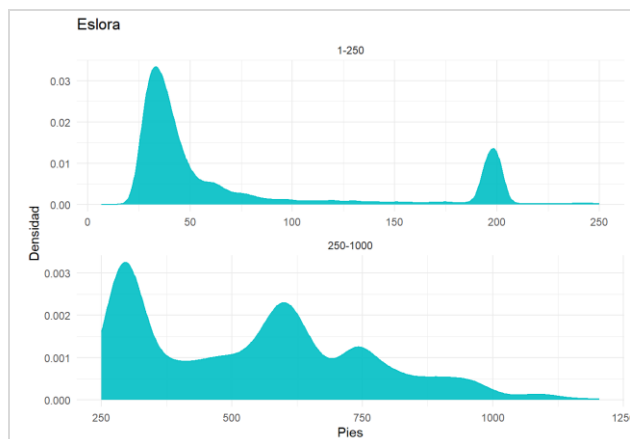


Fig. 4.1.1.4.: VBS Eslora

Se aprecia como la mayoría de embarcaciones se concentran en torno a los 30-40 pies de eslora.

Para evaluar los valores más frecuentes de las variables anteriores, se van a presentar unos gráficos de caja:

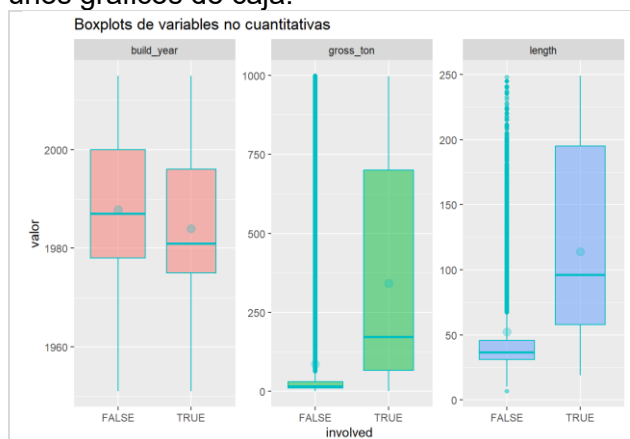


Fig. 4.1.1.5.: VBS Box-Plots Cuanti

En estas variables, se aprecian diferencias de medias según si la embarcación está involucrada en accidentes, por lo que fueron incluidas en el análisis

4.1.2. Aspectos legales y normativos

Bandera (flag_abbr)

En cuanto a la bandera, se presenta un gráfico que recuenta barcos bajo cada una de las banderas más frecuentes:

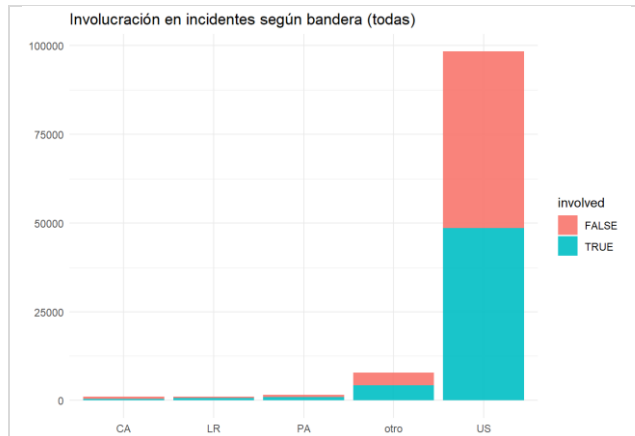


Fig. 4.1.2.1: VBS Bandera

Al centrarse el análisis en E.E.U.U., la gran mayoría de banderas son locales por asuntos legislativos, pero en el resto del mundo no se da esta situación. En cualquier caso, otras banderas destacadas son:

- PA - Panamá
- CA - Canadá
- LR - Liberia

No se aprecian diferencias para la involucración en incidentes.

Sociedad de clasificación (classification_society)

En este caso, también se ha añadido porcentajes de distribución, ya que se aprecian diferencias para la afección a incidentes

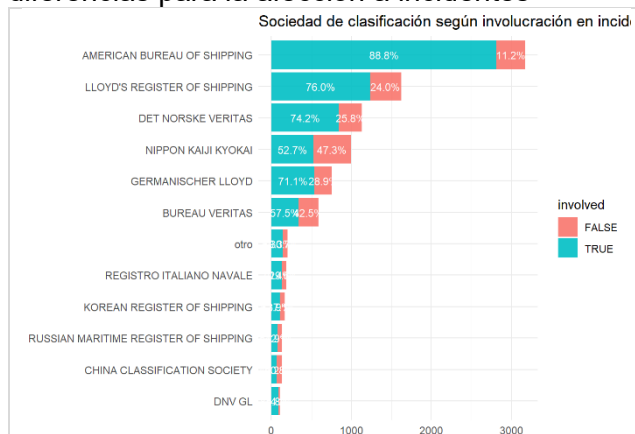


Fig. 4.1.2.2.: VBS Clasificación

En este caso, ABS, se lleva la mayoría de clasificaciones. Sin embargo, se pueden apreciar diferencias para la involucración en incidentes

Adhesión a SOLAS (solas_desc)

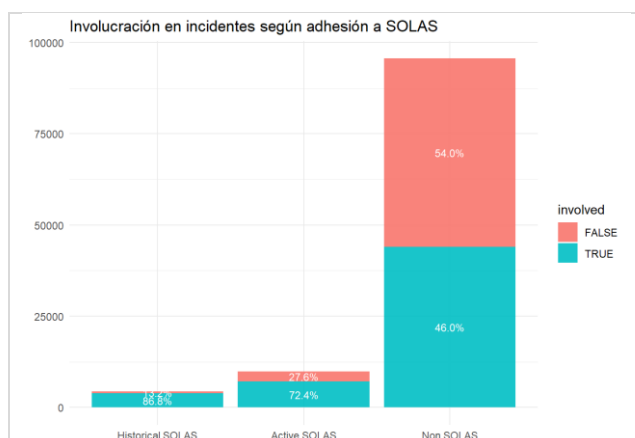


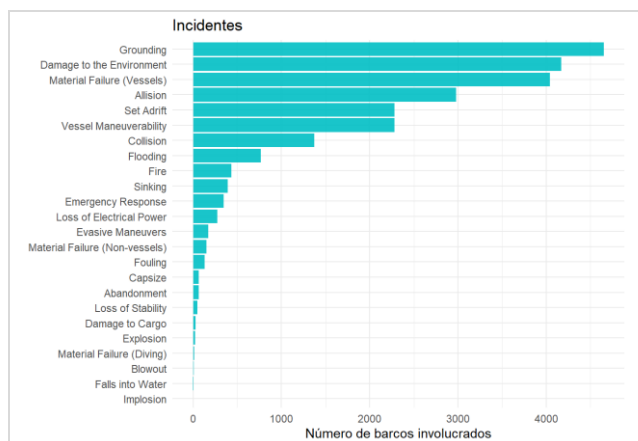
Fig. 4.1.2.3.: VBS Solas

Aquí se puede ver cómo hay una mayoría que no está obligado a cumplir con SOLAS. Sin embargo, para aquellos que sí lo están (Active Solas), hay una paradójica diferencia entre la involucración de incidentes.

4.1.3. Características de los incidentes

Tipo de incidente (event_type)

¿Cuáles son los tipos de accidente más habituales? En el siguiente gráfico de barras se pueden ver los incidentes, ordenados de mayor a menor frecuencia y distinguiendo aquellas embarcaciones que han sido objeto de incidentes:

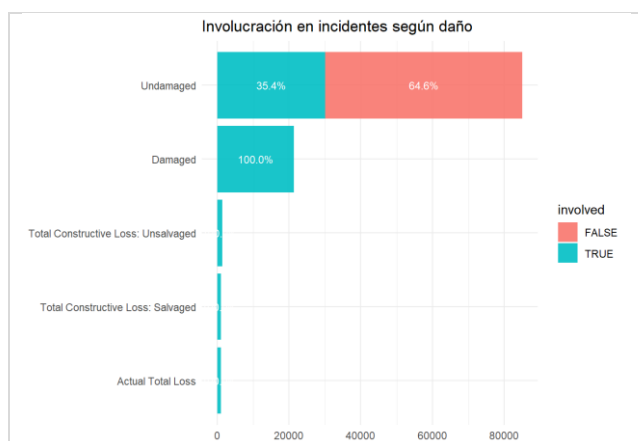


La mayoría de incidentes corresponden a *Grounding* (encallamiento), seguido de *Damage to the Environment* (daño al medio ambiente) y *Material Failure* (Fallo de material)

Fig. 4.1.3.1.: VBS Incidentes

Daños (damage_status)

Tal como se ha visto en el gráfico anterior, una gran parte de estos incidentes son daños al medio ambiente. Puede haber casos en los que estos hechos se produzcan sin causar daños materiales al barco. Por ejemplo, si se abre por error ciertas válvulas, puede haber un vertido nocivo, pero no causa daños. Veamos este dato más en detalle:



En este caso, se puede apreciar cómo un 35.4% de los barcos involucrados en accidentes, no han registrado daños. Por otro lado, lógicamente, los barcos con daños registrados han sido todos involucrados en incidentes.

Fig. 4.1.3.2.: VBS Daños

Nota: dentro de este conjunto de datos se incluyen otras variables como nombre del barco o numero IMO que no tienen valor para este análisis, pero son conservadas como información para posibles extrapolaciones de datos.

4.1.4. Correlaciones

La correcta comprensión de las correlaciones en un conjunto de datos es fundamental antes de aplicar modelos de aprendizaje automático. Este paso es esencial para identificar patrones, relaciones y dependencias entre variables, proporcionando información crucial para el diseño efectivo de modelos predictivos. Analizar correlaciones permite detectar posibles sesgos, evitar multicolinealidad y mejorar la interpretación de los resultados. Con este proceso, se garantiza una toma de decisiones informada y la optimización del rendimiento del modelo.

En este caso, se ha utilizado la librería PerformanceAnalytics que aporta datos adicionales a las correlaciones como significatividad de la correlación e histogramas. El resultado es el siguiente cuadro:

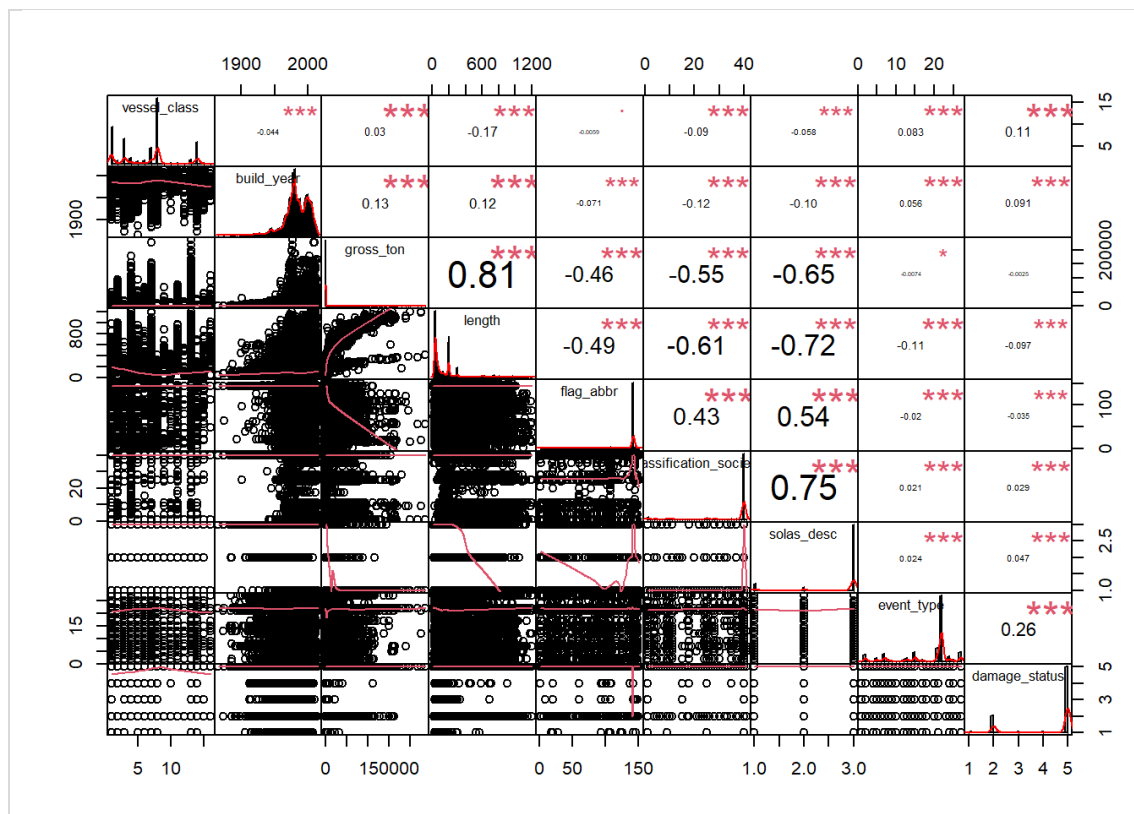


Fig. 4.1.4. VBS Correlaciones

Se aprecian correlaciones altas y significativas entre:

- `length` y `gross_ton`: Según lo esperado, una mayor eslora, implica un mayor volumen, ya que estas magnitudes guardan proporcionalidad por el propio diseño. En cualquier caso, se va a optar por mantener ambas ya al no incluir manga ni calado ni francobordo, se deja ciertos grados de libertad para las características físicas de las embarcaciones.
- `classification_Society` y `solas_desc`: Normalmente, los barcos con mayor volumen están obligados a atenerse a ambas cuestiones, por lo que esta relación está dentro de lo esperable. Del mismo modo, para el análisis, se van conservar para evaluar si las diferentes normas de las sociedades de clasificación tienen influencia en la mejora de la seguridad en el mar.

4.2. Dataset MergedActivity

Este otro conjunto de datos, está formado por un total de 68.000 observaciones y 26 variables que describen con detalles los incidentes. Al focalizarse en la descripción de los incidentes, no hay datos de barcos sin incidentes. Con estos datos, se pretende predecir qué tipo de incidente se producirá cuando se reúnen ciertas circunstancias o características. Tiene en común diversas variables con el conjunto de datos anterior, por lo que no se van a repetir sus descripciones. Sin embargo, todas ellas han sido analizadas y el código empleado está disponible en el anexo 4.2.

Sumario estadístico (skimr)

Data summary

Name	MergedActivity
Number of rows	68000
Number of columns	27
Key	NULL

Column type frequency:

character	15
Date	1
numeric	11

Group variables	None
-----------------	------

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
hour	0	1.00	5	5	0	1438	0
region	0	1.00	6	14	0	6	0
watertype	0	1.00	5	5	0	2	0
event_type	0	1.00	4	30	0	26	0
damage_status	0	1.00	7	35	0	5	0
imo_number	0	1.00	0	7	46583	6013	0
vessel_name	0	1.00	1	50	0	23854	0
vessel_class	0	1.00	5	23	0	16	0
build_year	0	1.00	4	4	0	122	0
flag_abbr	0	1.00	0	2	24	106	0
classification_society	0	1.00	6	58	0	36	0
solas_desc	0	1.00	9	16	0	3	0
casualty	65628	0.03	4	11	0	4	0
pollution	55049	0.19	0	3	73	131	0
event_class	0	1.00	15	19	0	5	0

skim_variable	n_missing	complete_rate	mean	sd	p0	p25
activity_id	0	1.00	3271440.45	942973.00	1838.02	535937.00
latitude	0	1.00	37.62	7.94	15.0	32.26
longitude	0	1.00	-94.41	21.47	-179.8	-94.87
vessel_id	0	1.00	290399.44	268497.49	18.0	90388.00
gross_ton	0	1.00	4391.51	13526.64	1.0	95.00
length	0	1.00	196.07	197.01	18.7	69.70
air_temp	5342	0.92	149.76	95.00	-230.5	82.00
wind_speed	25494	0.63	50.58	30.65	0.0	29.00
wave_hgt	55164	0.19	2.35	2.28	0.0	1.00
visibility	63031	0.07	96.90	1.44	90.0	96.50
damage_assessment	56	1.00	122959.80	3887042.37	0.0	0.00

4.2.1. Localización

Una de las informaciones más características de los incidentes es dónde se han producido. Para ello, como en este caso, se suelen utilizar datos de geoposición con latitud y longitud. Estos valores numéricos proporcionan una referencia única para cualquier punto en la superficie terrestre. La latitud mide la distancia al norte o al sur del ecuador, mientras que la longitud indica la posición al este o al oeste del meridiano de Greenwich. Juntas, estas coordenadas forman un código geográfico universal que nos permite navegar el mundo con precisión y ubicar cualquier lugar con exactitud, facilitando la cartografía, la navegación y la comunicación.

En cuanto al cribado de observaciones, cabe destacar que, durante la exploración estadística, se detectaron algunos incidentes registrados en sitios muy remotos e incluso inverosímiles como en mitad de una llanura. Con la ayuda de la librería *rnatureearth*, se ha podido diferenciar los puntos correspondientes a la zona continental y marítima. Gracias a esta limpieza de datos, solo se van a considerar las regiones de: Alaska, Canadá, Costa Este, Golfo de México, Mississippi (única área fluvial) y Costa Oeste.

Para la representación de mapas, se ha utilizado la librería *leaflet*, que ofrece la posibilidad de crear mapas interactivos. Al ser éste un documento estático, se van a utilizar capturas de los mapas, perdiéndose de esta manera los detalles en forma de pop-up para cada punto.

Regiones (region)

Las características del medio acuático son muy diferentes en las zonas de Alaska de las de Golfo de México o el río Mississippi. Por esta razón, se optó por crear una variable que englobe medios marinos similares. En este caso, se han determinado las siguientes:

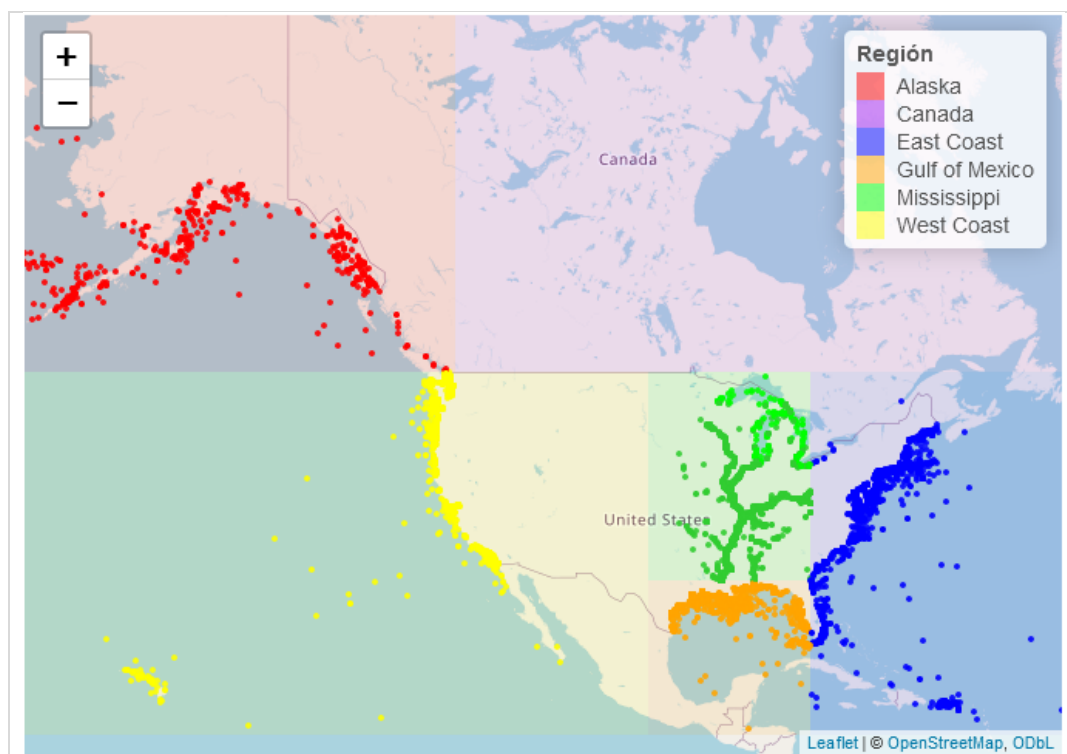


Fig. 4.2.1.1 MA Regiones

En cuanto a la distribución del número de accidentes, debido a la superposición de puntos y la muestra aleatoria empleada para aligerar la carga, no es posible determinar

la cantidad de incidentes sobre el mapa. Sin embargo, se puede comprobar mediante el siguiente gráfico de barras:

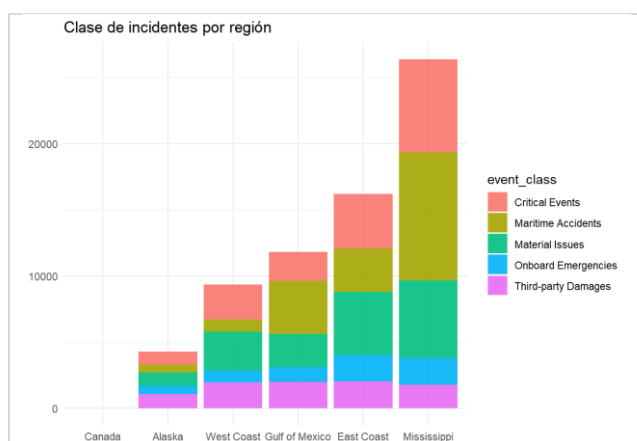


Fig. 4.2.1.2.: MA Incidentes Regiones

La zona con mas incidentes registrados es Mississippi con más de 25.000. En cuanto a la distribución de la clase de incidentes, se puede destacar que los datos a terceros mantienen proporciones diferentes según las regiones.

La misma información, pero sobre un mapa:

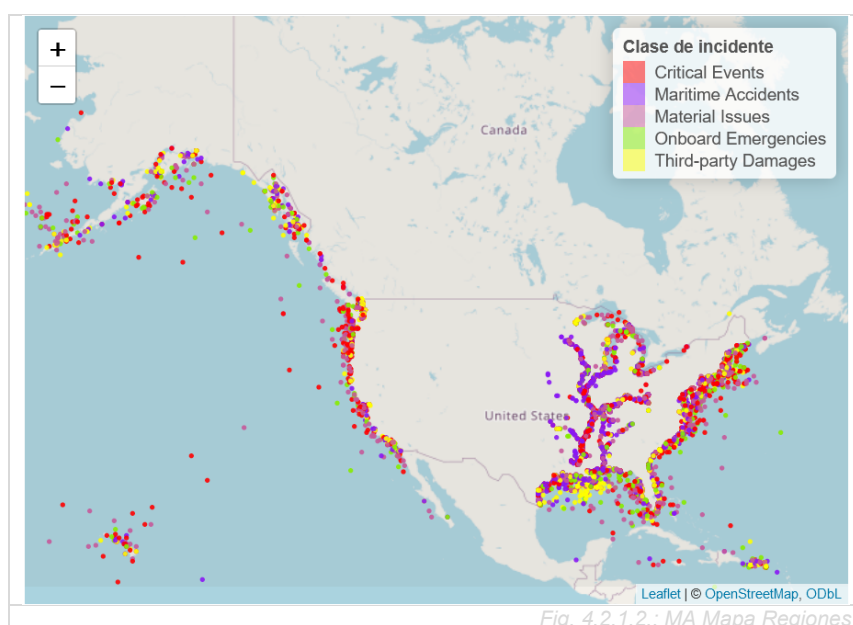


Fig. 4.2.1.2.: MA Mapa Regiones

Evento más común en cada región (event_type)

¿Cuál es el incidente más común en cada región? A continuación, se presenta una tabla con los sucesos más frecuentes

region	suceso_mas_frecuente	num_sucesos_por_region
Mississippi	Grounding	11374
East Coast	Damage to the Environment	9630
Gulf of Mexico	Grounding	7163
West Coast	Vessel Maneuverability	5900
Alaska	Damage to the Environment	2692
Canada	Material Failure (Vessels)	1

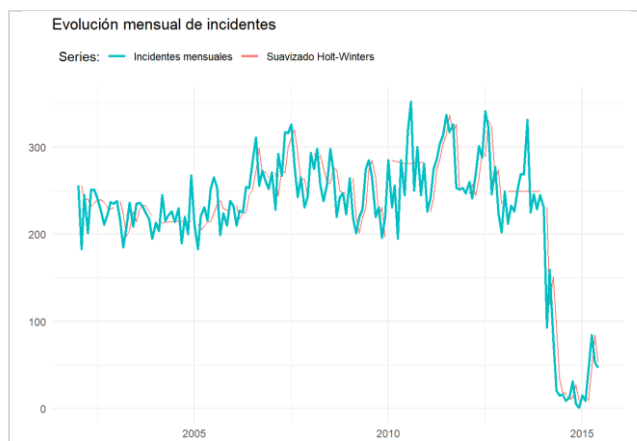
Fig. 4.2.1.3.: MA Incidentes Frecuentes

Los encallamientos son habituales en Mississippi debido a su bajo calado

4.2.2. Cronología

Serie temporal (date)

En este conjunto de datos, se cuenta con información de fecha y hora para cada incidente, por lo que se puede evaluar cómo es la evolución temporal a través de la siguiente serie a la que se añadió un suavizado tipo Holt-Winters. El suavizado Holt-Winters es un método de pronóstico para series temporales que considera tendencia y estacionalidad. Utiliza promedios ponderados y ajusta factores de suavizado para predecir futuros valores con precisión. En este caso, el suavizado no se ajusta completamente, por lo que se puede deducir cierta aleatoriedad.

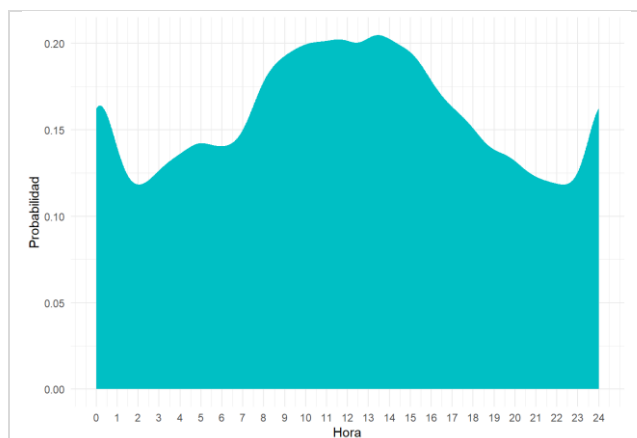


Lo más destacable es una notable caída de incidentes a partir de 2014

Fig. 4.2.2.1.: MA Serie temporal

Perfil horario (hour)

Respecto a la hora en la que se producen los incidentes, saber cuando son más frecuentes puede ayudar a adoptar medidas para evitarlos. En la siguiente gráfica se expone el perfil horario que siguen



Tal como se sospecha, los incidentes son más frecuentes en las horas centrales del día

Fig. 4.2.2.2.: MA Perfil horario

4.2.3. Meteorología

Tras el procesado y unión de datos meteorológicos obtenidos de la NOAA, estos pueden ser incluidos para el análisis, enriqueciendo la información original. A pesar de contar con datos de diversos parámetros, se ha optado por simplificar la cantidad de variables. Las incluidas en este conjunto de datos son: temperatura (air_temp), viento (wind_speed), altura de ola (wave_hgt) y visibilidad (visibility). En el siguiente gráfico se pueden ver sus valores más frecuentes, según clase de evento:

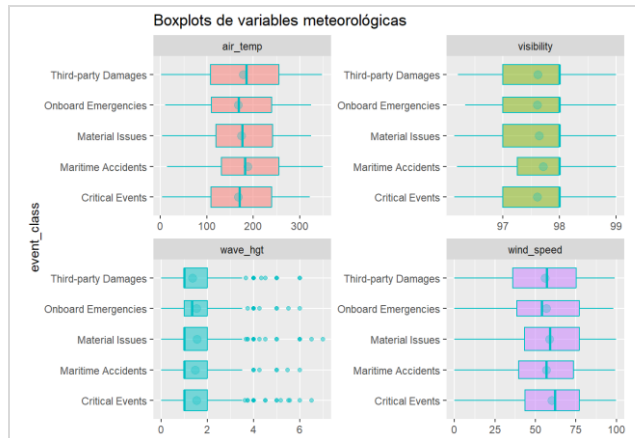


Fig. 4.2.3.1.: MA Box-Plot Meteo

Según parece las variables de visibilidad y altura de ola no tienen valores significativamente diferentes según la clase de incidente, por lo que son candidatas a ser excluidas en los modelos de aprendizaje automático.

Temperatura (air_temp)

Si se considera la temperatura por separado, observando su distribución geográfica y su distribución temporal, que no ofrecen ninguna sorpresa:

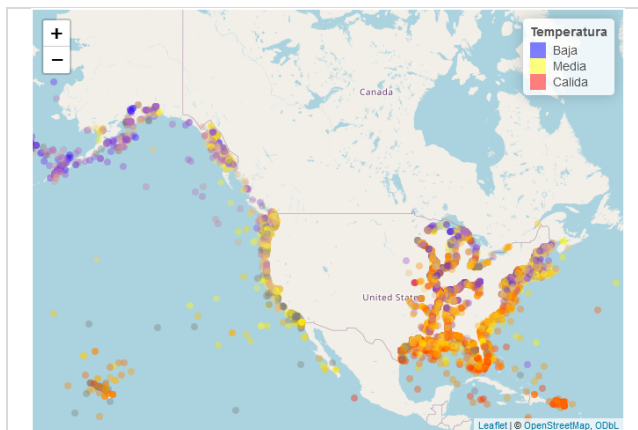


Fig. 4.2.3.2.: MA Serie temporal

Según lo esperado, las temperaturas más bajas, se sitúan a mayor latitud, en Canadá y las altas, más al sur en el Golfo de México

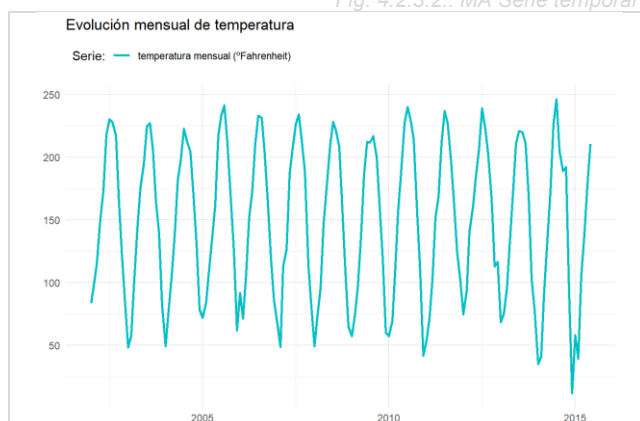


Fig. 4.2.3.3.: MA Serie temperatura

La temperatura media mensual sigue ciclos anuales muy similares entre sí, por lo que se pueden hacer estimaciones de este parámetro

4.2.4. Características de los barcos

Aunque se han revisado la mayoría de sus variables en el apartado 4.1, al incluirse nuevas variables en este conjunto de datos, se pueden evaluar teniendo en cuenta estas.

Antigüedad (date – build_year)

Por ejemplo, con la fecha de construcción y la fecha del incidente, se puede obtener la antigüedad

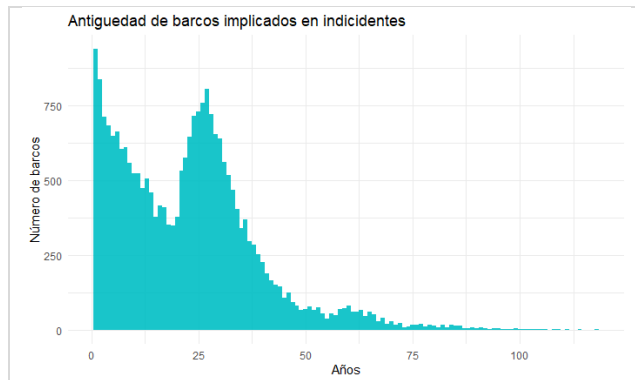


Fig. 4.2.4.1.: MA Antigüedad

Se aprecian dos picos de frecuencia: barcos muy nuevos y barcos con alrededor de 30 años

Veamos cómo se distribuyen según la clase de incidente registrado, junto con las otras dos variables de características de diseño: gross_ton (volumen) y length (eslora)

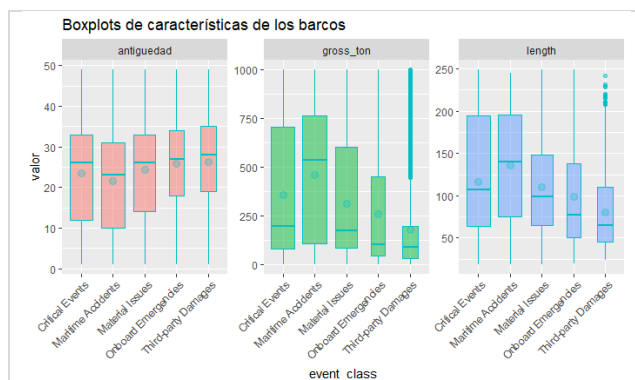
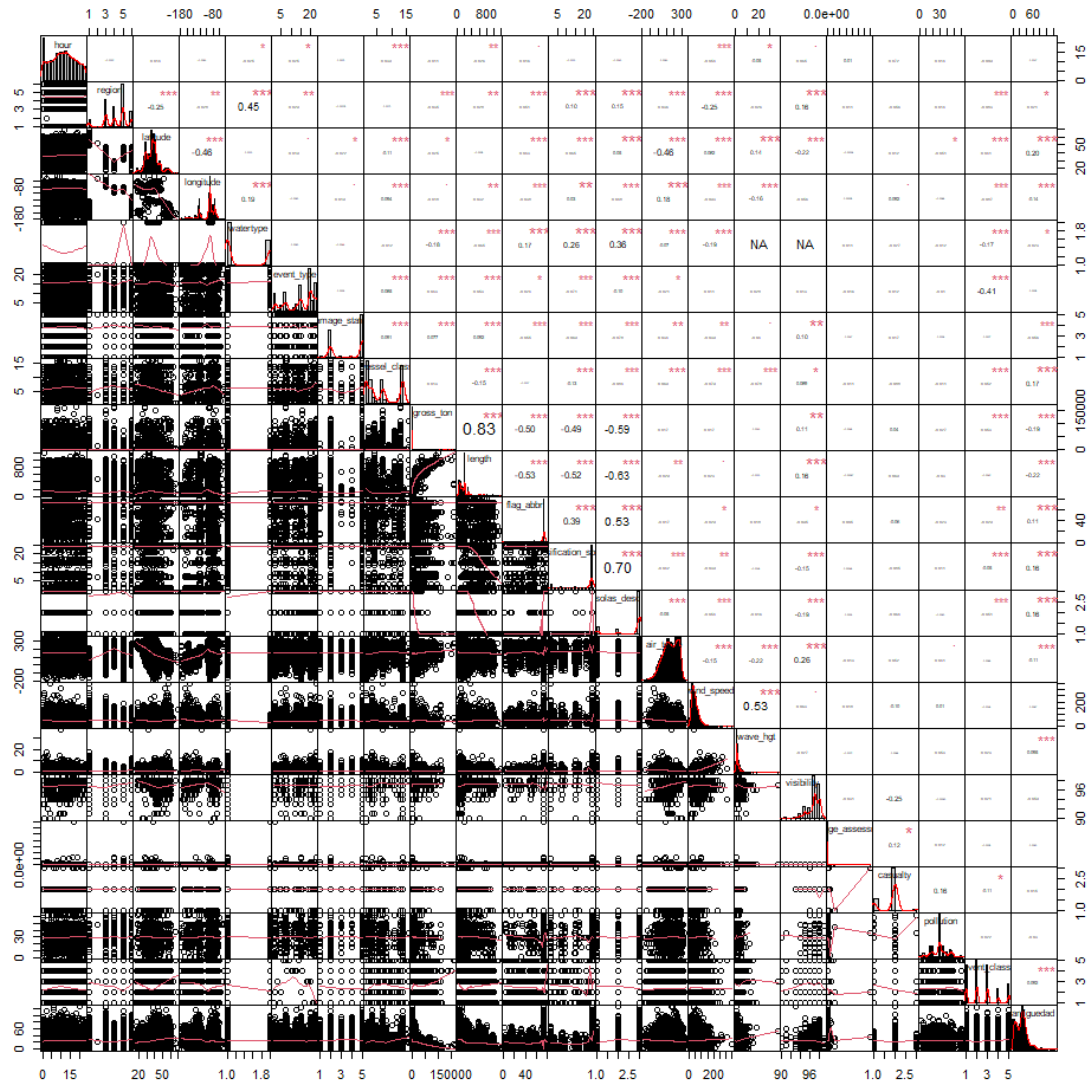


Fig. 4.2.4.2.: MA Box-Plots Barcos

En las tres variables se aprecian diferencias de sus valores más frecuentes, por lo que no serán descartadas para el procesado

4.2.5. Correlaciones

Tal como se ha establecido anteriormente, la relación de las variables es un aspecto fundamental para la aplicación de los algoritmos que se expondrán en lo siguiente apartados.



Para este conjunto de datos, las relaciones entre las variables de características de los barcos presentan los mismos patrones que en el conjunto anterior. En cuanto a las variables meteorológicas, se ve cierta correlación entre altura de ola (wave_hgt) y viento (wind_speed), tal como suele ser habitual en todo el mundo.

5. Aplicación de modelos de Machine Learning

En este apartado se van a exponer los resultados de la aplicación de algoritmos de aprendizaje automático, así como una comparación de rendimiento en base a las métricas más habituales en este ámbito.

Los modelos de machine learning, o aprendizaje automático, son algoritmos y sistemas informáticos diseñados para aprender patrones y realizar tareas específicas sin ser programados explícitamente. En lugar de depender de reglas de programación estáticas, estos modelos utilizan datos para aprender y mejorar su rendimiento con el tiempo.

Estos modelos tienen aplicaciones en una amplia variedad de campos, desde reconocimiento de voz, reconocimiento de imágenes, recomendaciones personalizadas, vehículos autónomos o análisis predictivo como en el caso de este análisis. El machine learning está transformando la manera en que interactuamos con la tecnología y cómo abordamos problemas complejos en diversas disciplinas.

El proceso de aprendizaje implica alimentar al modelo con un conjunto de datos de entrada, permitiéndole identificar patrones y realizar predicciones o tomar decisiones. Existen diversos enfoques de machine learning, como el aprendizaje supervisado, donde el modelo se entrena con ejemplos etiquetados; el aprendizaje no supervisado, donde el modelo encuentra patrones sin etiquetas; y el aprendizaje por refuerzo, donde el modelo toma decisiones basadas en la retroalimentación recibida.

En el proceso de desarrollo de modelos de machine learning, se utiliza un conjunto de datos de entrenamiento y otro de prueba para evaluar y validar el rendimiento del modelo. El conjunto de datos de entrenamiento se emplea para enseñar al modelo patrones y relaciones entre las variables, permitiéndole aprender del comportamiento de los datos. Una vez entrenado, el modelo se evalúa utilizando el conjunto de datos de prueba, que no ha sido visto durante la fase de entrenamiento. Este conjunto independiente sirve para comprobar la capacidad del modelo para generalizar y hacer predicciones precisas en datos no familiares. La división entre entrenamiento y prueba ayuda a identificar posibles problemas de sobreajuste, donde el modelo se adapta demasiado a los datos de entrenamiento, pero no generaliza bien a nuevos datos. Este enfoque de validación cruzada mejora la confianza en la capacidad predictiva del modelo y contribuye a su robustez en situaciones del mundo real.

Según el tipo de problema al que se enfrentan, estos algoritmos pueden ser de regresión o clasificación como los de este caso. La principal diferencia entre un problema de regresión y clasificación radica en la naturaleza de la tarea que se está abordando en el ámbito del machine learning. En un problema de regresión, el objetivo es predecir un valor numérico continuo, como el precio de una casa o la temperatura. En este contexto, los modelos se entrenan para entender la relación entre las variables de entrada y la variable objetivo, buscando generar predicciones que se acerquen lo más posible a los valores reales.

Por otro lado, en un problema de clasificación, el objetivo es asignar una etiqueta o categoría a una instancia dada. Por ejemplo, clasificar correos electrónicos como spam o no spam, predecir si se producirá o no un accidente, etc. Aquí, los modelos aprenden patrones que permiten asignar la entrada a una de varias categorías predefinidas. Es decir, mientras que la regresión se enfoca en predecir cantidades numéricas, la clasificación se centra en asignar etiquetas a datos discretos.

Modelado para este análisis

Problema	Posibilidad de incidente	Clase de incidente
Cuestión	¿Habrá incidente?	¿Qué tipo de incidente?
Conjunto de datos	VesselBalancedSample (50% Incidentes, 50% No)	MergedActivity (Incidentes)
Respuesta	· involved: Yes / No	· event_class: Critical Events / Onboard Emergencies / Maritime Accidents / Material Issues / Third-party Damages
Tipo de respuesta	Clasificación binaria	Clasificación multinivel (5)
Características (Variables)	· vessel_class · build_year · gross_ton · vessel_length · flag_abbr · classification_soc · solas_desc	· activity_id · hour · region · longitude · water_type · damage_status · vessel_class · age · gross_ton · vessel_length · air_temp · wind_speed
Modelos (Algoritmos)	1. Naïve Bayes 2. Tree Augmented Naive Bayes Classifier (TAN) 3. Tree Augmented Naive Bayes Classifier Structure Learner Wrapper (TAN Search) 4. Hill Climbing Tree Augmented Naive Bayes Classifier (TAN Hill Climbling) 5. Aggregating One-Dependence Estimators Naive Bayes Classifier (AODE) 6. Gradient Boosting 7. Extreme Gradient Boosting (XGBTree) 8. Random Forest 9. Máquinas de vector soporte (svmRadial) 10. Perceptrón multicapa (MLP) 11. C5.0 Tree 12. Regresión logística	1. Naïve Bayes 2. Gradient Boosting 3. Random Forest 4. Perceptrón multicapa 5. C5 Tree 6. Keras Sequential Model (Red densamente poblada) 7. Extra: H2o Auto Machine Learning
Evaluación (Métricas)	1. Accuracy (Exactitud) 2. Kappa 3. Sensibilidad y Especificidad 4. ROC y AUC	1. Accuracy (Exactitud) 2. Kappa 3. Sensibilidad y Especificidad 4. ROC y AUC 5. Log Loss

Fig. 5. Resumen Modelado

5.0. Marco teórico

5.0.1. Modelos de machine learning

Para poner en contexto y explicar su funcionamiento, a continuación, se va a presentar de la manera más resumida posible el funcionamiento que sustenta los modelos y sus técnicas. En el siguiente apartado se verán las métricas utilizadas para compararse entre sí. En este análisis, se han utilizado los siguientes modelos:

1. Naïve Bayes

Naïve Bayes es un algoritmo de clasificación probabilística basado en el teorema de Bayes. Supone independencia condicional entre las características dadas las clases, lo que simplifica el cálculo de las probabilidades. Utiliza la regla de Bayes para calcular la probabilidad de pertenencia a una clase dada una observación. El cálculo de la probabilidad condicional de pertenencia a una clase C dado un conjunto de características X , se formula de con la siguiente fórmula:

$$P(C|X) = \frac{P(X|C) \cdot P(C)}{P(X)}$$

Donde $P(X|C)$ es la probabilidad de observar X dado que pertenece a la clase C , $P(C)$ es la probabilidad previa de la clase C , y $P(X)$ es la probabilidad marginal de X .

2. Tree Augmented Naive Bayes Classifier (TAN):

TAN mejora Naïve Bayes al permitir dependencias entre las características, representadas mediante un árbol. La estructura del árbol se aprende durante el entrenamiento, capturando las relaciones de dependencia condicional entre las características y mejorando la precisión en casos donde Naïve Bayes asume independencia. En este caso, la probabilidad conjunta se expresa como:

$$P(X, C) = P(C) \cdot \prod_{i=1}^n P(X_i|C, P_a(X_i))$$

Donde $P_a(X_i)$ son las características padre de X_i en el árbol.

3. Tree Augmented Naive Bayes Classifier Structure Learner Wrapper (TAN S):

Es una variante de TAN que utiliza técnicas de búsqueda para encontrar la mejor estructura del árbol, mejorando la eficiencia del aprendizaje y la calidad del modelo resultante, optimizando la probabilidad conjunta $P(X, C)$

4. Hill Climbing Tree Augmented Naive Bayes Classifier (TAN Hill Climbing):

Similar a TAN, pero utiliza el algoritmo de “escalada de colinas” para buscar la estructura del árbol que maximiza la probabilidad de los datos observados. Busca la mejor solución localmente, explorando iterativamente vecinos cercanos y moviéndose hacia la dirección que mejora el valor objetivo, mejorando la estructura del árbol de manera iterativa.

5. Aggregating One-Dependence Estimators Naive Bayes Classifier (AODE):

AODE es una extensión de Naïve Bayes que permite dependencias entre características, pero asume que estas dependencias pueden agregarse de manera independiente para simplificar el cálculo, pero asume que estas dependencias pueden agregarse de manera independiente:

$$P(X, C) = P(C) \cdot \prod_{i=1}^n P(X_i|C)$$

6. Gradient Boosting:

Gradient Boosting engloba un conjunto de algoritmos que construye modelos predictivos en forma de árboles de decisión de manera secuencial, cada uno corrigiendo los errores del modelo anterior. Se optimiza minimizando una función de pérdida, minimizando la función de pérdida L iterativamente ajustando el modelo F de manera que:

$$F_m(x) = F_{m-1}(x) + p \cdot h_m(x)$$

Donde $h_m(x)$ es el modelo débil ajustado en la m -ésima iteración, y p es la tasa de aprendizaje.

7. Extreme Gradient Boosting (XGBTree):

XGBoost es una implementación eficiente de Gradient Boosting. XGBTree utiliza árboles de decisión como modelos base, incorporando regularización y técnicas de poda para prevenir el sobreajuste. La función objetivo a minimizar es:

$$\sum_{i=1}^n L(y_i, F(x_i)) + \sum_{k=1}^K \Omega(f_k)$$

Donde L es la función de pérdida y $\Omega(f_k)$ es la penalización por complejidad del árbol k .

8. Random Forest:

El algoritmo Random Forest construye múltiples árboles de decisión durante el entrenamiento y combina sus predicciones para obtener una salida más robusta y precisa. En lugar de depender de un solo árbol, Random Forest crea un conjunto de árboles, cada uno entrenado con una muestra aleatoria de datos y características. Luego, combina las predicciones de estos árboles para mejorar la generalización y reducir el sobreajuste, proporcionando un modelo más sólido y eficaz.

9. Máquinas de vector soporte (svmRadial):

Las Máquinas de Vector Soporte son algoritmos de clasificación que buscan encontrar el hiperplano que mejor separa las clases en el espacio de características. svmRadial utiliza un kernel radial para manejar conjuntos de datos no lineales. El kernel radial (Gaussiano) en el algoritmo svmRadial se define como:

$$K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)$$

Donde σ es un parámetro de ancho del kernel.

10. Perceptrón multicapa:

El Perceptrón multicapa es un tipo de red neuronal con una capa oculta. Las redes neuronales consisten en capas de nodos (neuronas) conectadas entre sí, donde cada conexión tiene un peso. El entrenamiento implica ajustar estos pesos para minimizar la diferencia entre las predicciones y los objetivos reales. En una red neuronal, la salida de la capa j se calcula como:

$$y_j = f\left(\sum_{i=1}^n w_{ij} \cdot x_i + b_j\right)$$

Donde w_{ij} son los pesos, x_i son las entradas, b_j es el sesgo y f es la función de activación.

11. **C5.0 Tree:**

C5.0 es un algoritmo de construcción de árboles de decisión que utiliza la ganancia de información y técnicas de poda para crear un modelo predictivo. Es eficiente y efectivo para conjuntos de datos grandes. La ganancia de información se calcula como la reducción en la entropía.

12. **Regresión logística:**

La regresión logística es un modelo de regresión utilizado para problemas de clasificación binaria. Utiliza la función logística para modelar la probabilidad de pertenencia a una clase. Esta función se puede formular de la siguiente manera:

$$P(Y = 1|X) = \frac{1}{1 + \exp(-\beta_0 - \beta_1 \cdot X_1 - \dots - \beta_n \cdot X_n)}$$

Donde $\beta_0, \beta_1, \dots, \beta_n$ son los coeficientes a aprender.

13. **Red densamente poblada**

Una red neuronal densamente poblada, o red neuronal completamente conectada, es un tipo de arquitectura de red neuronal en la que cada neurona en una capa está conectada a todas las neuronas de la capa siguiente. Estas redes son comúnmente utilizadas en tareas de aprendizaje profundo. La salida de una capa se calcula tal como se ha expuesto para la red neuronal. En el caso de este análisis se va utilizar la implementación de keras tensorflow.

5.0.2. Métricas

Las métricas de evaluación en machine learning son herramientas esenciales para medir y comprender el rendimiento de los modelos de clasificación. Cada una de las métricas proporciona información sobre diferentes aspectos del modelo. Esto permite comparar el rendimiento de los modelos y poder elegir las mejores. A continuación, se explican las más utilizadas para los modelos de clasificación como los utilizados en este análisis.

1. Accuracy (Exactitud)

Medir la proporción total de instancias clasificadas correctamente por el modelo. Ayuda a seleccionar modelos con un equilibrio adecuado entre sensibilidad y especificidad y a comparar la calidad del clasificador en diferentes contextos. Se puede calcular mediante la siguiente fórmula:

$$Accuracy = \frac{Verdaderos\ Positivos + Verdaderos\ Negativos}{Total\ de\ Instancias}$$

2. Kappa

La estadística kappa ajusta la exactitud observada por la exactitud esperada al azar y mide la concordancia entre las predicciones y las observaciones. Es útil cuando el desequilibrio de clases afecta la interpretación de la exactitud. Se calcula de este modo:

$$k = \frac{Exactitud\ Observada - Exactitud\ Esperada\ al\ Azar}{1 - Exactitud\ Esperada\ al\ Azar}$$

Un valor de kappa de 1 indica una concordancia perfecta, 0 indica concordancia aleatoria y valores negativos sugieren concordancia peor que aleatoria.

3. Sensibilidad y Especificidad

La sensibilidad mide la proporción de instancias positivas que son correctamente identificadas por el modelo. Es decir, una tasa de verdaderos positivos. Por su parte, la especificidad mide la proporción de instancias negativas que son correctamente identificadas por el modelo. Es decir, una tasa de verdaderos negativos. Son útiles cuando hay asimetría en la importancia de los errores tipo I (falsos positivos) y tipo II (falsos negativos). Se pueden calcular así:

$$Sensibilidad = \frac{Verdaderos\ Positivos}{Positivos\ Reales} ; Especificidad = \frac{Verdaderos\ Negativos}{Negativos\ Reales}$$

4. ROC y AUC

La curva ROC (Receiver Operator Curve) es una representación gráfica de la sensibilidad frente a la tasa de falsos positivos para diferentes umbrales de clasificación. Proporciona una visión completa del rendimiento de un modelo en diferentes niveles de sensibilidad y especificidad.

El AUC (Area Under Curve) mide la capacidad discriminativa global del modelo y se calcula como el área bajo la curva ROC. Un AUC cercano a 1 indica un buen rendimiento, mientras que un AUC de 0.5 indica un rendimiento aleatorio. Se puede calcular como:

$$AUC = \int_0^1 TPR(FPR)dx$$

Donde TPR es la tasa de verdaderos positivos y FPR es la tasa de falsos positivos en cada punto. Ambas métricas ayudan a seleccionar modelos con un

equilibrio adecuado entre sensibilidad y especificidad y a comparar la calidad del clasificador en diferentes contextos.

5. Log Loss

Esta métrica mide cómo un modelo de clasificación es capaz de asignar probabilidades a las clases correctas. Penalizar las predicciones incorrectas de manera más fuerte, asignando una penalización mayor cuando la probabilidad asignada a la clase correcta es baja. La función logarítmica asegura que la penalización crezca exponencialmente a medida que la probabilidad predicha se aleja de la clase correcta. Su cálculo es el siguiente:

$$\text{Log Loss} = -\frac{1}{N} \sum_{i=1}^N \sum_{j=1}^M y_{ij} \log(p_{ij})$$

Donde N es el total de ejemplos, M es número de clases, y_{ij} indica si pertenece a la clase j , p_{ij} es la probabilidad predicha de pertenecer a la clase j

Por otro lado, hay que tener en cuenta que además de estas métricas de evaluación, es crucial considerar la confiabilidad, interpretabilidad, tiempo de entrenamiento y predicción, escalabilidad, manejo de datos faltantes, costo computacional, tolerancia a errores, aprendizaje continuo, ética y sesgo, y documentación/mantenimiento al evaluar modelos de aprendizaje automático. Pero estos otros aspectos proporcionan una visión más completa del rendimiento y la aplicabilidad del modelo en el mundo real.

A continuación, se van a exponer los resultados del entrenamiento y evaluación de cada uno de los modelos. Por un lado, los referentes al problema de clasificación de este análisis “¿Habrà o no incidente”? y por separado los correspondientes a ¿Qué tipo de incidente?

5.1. Posibilidad de incidente: VesselBalancedSample

Para el modelado de este conjunto se ha utilizado el código contenido en el anexo 5.1., no siendo necesario equilibrar ni completar la muestra para aplicar los modelos. Los parámetros de entrenamiento han sido los siguientes:

Método utilizado: para el entrenamiento de los modelos de este problema de clasificación binaria, se ha utilizado la librería caret.

Particionado de datos: 80% de observaciones para entrenamiento (train) y 20% para test. Se ha dividido con una selección al azar.

Validación cruzada: “repeatedcv n=8, rep=2”. Es decir, Validación cruzada repetida. Es una técnica de evaluación del rendimiento del modelo que combina la validación cruzada (cv) con repeticiones. En este caso, el conjunto de datos se divide en 8 pliegues (folds) y se realizará la validación cruzada 2 veces. La repetición puede ayudar a reducir la varianza del rendimiento estimado

Métrica de evaluación: ROC (Receiver Operating Characteristic)

En cuanto a los resultados, esta memoria solo va a destacar la comparativa de modelos, ya que el detalle de los parámetros utilizados en cada modelo y sus outputs concretos se podrán encontrar en el mencionado anexo.

5.1.1. Comparativa de los modelos “Posibilidad de incidente”

Comparación con la Muestra de Entrenamiento

	AUC	Accuracy	Aciertos Clase SI	Aciertos Clase NOK	Kappa	Sensitivity	Specificity
RF	0.955	0.933	0.956	0.911	0.865	0.906	0.953
RL	0.948	0.886	0.920	0.858	0.772	0.846	0.926
XGB	0.947	0.881	0.920	0.849	0.762	0.835	0.927
C5	0.941	0.881	0.915	0.853	0.763	0.841	0.922
GBM	0.933	0.865	0.904	0.832	0.729	0.816	0.913
MLP	0.926	0.856	0.899	0.821	0.712	0.802	0.910
TAN	0.924	0.855	0.879	0.833	0.709	0.822	0.887
AODE	0.924	0.849	0.871	0.830	0.698	0.819	0.879
SVM	0.915	0.835	0.925	0.777	0.671	0.730	0.941
NB	0.894	0.825	0.847	0.805	0.649	0.793	0.856
TANSE	0.894	0.825	0.846	0.806	0.649	0.794	0.856
TANHC	0.894	0.825	0.846	0.806	0.649	0.794	0.856

Comparación con la Muestra de Validación

	AUC	Accuracy	Aciertos Clase SI	Aciertos Clase NO	Kappa	Sensitivity	Specificity
RF	0.957	0.937	0.961	0.916	0.875	0.912	0.963
XGB	0.949	0.883	0.923	0.850	0.766	0.835	0.931
GBM	0.935	0.865	0.907	0.830	0.729	0.813	0.917
C5	0.933	0.868	0.903	0.839	0.737	0.825	0.912
TAN	0.926	0.855	0.880	0.833	0.710	0.822	0.888
AODE	0.926	0.853	0.878	0.831	0.706	0.820	0.886
MLP	0.926	0.851	0.897	0.815	0.703	0.793	0.909
RL	0.923	0.866	0.902	0.836	0.733	0.822	0.911
SVM	0.915	0.832	0.926	0.771	0.663	0.721	0.943
NB	0.896	0.828	0.852	0.808	0.657	0.795	0.862
TANSE	0.896	0.829	0.852	0.808	0.657	0.796	0.861
TANHc	0.896	0.829	0.852	0.808	0.657	0.796	0.861

Los 12 modelos utilizados en este análisis, tienen buenos rendimientos tanto para las muestras de test como en las de validación, por lo que se podrían utilizar para predecir la posibilidad de incidente en el ámbito naval. Gráficamente en conjunto:

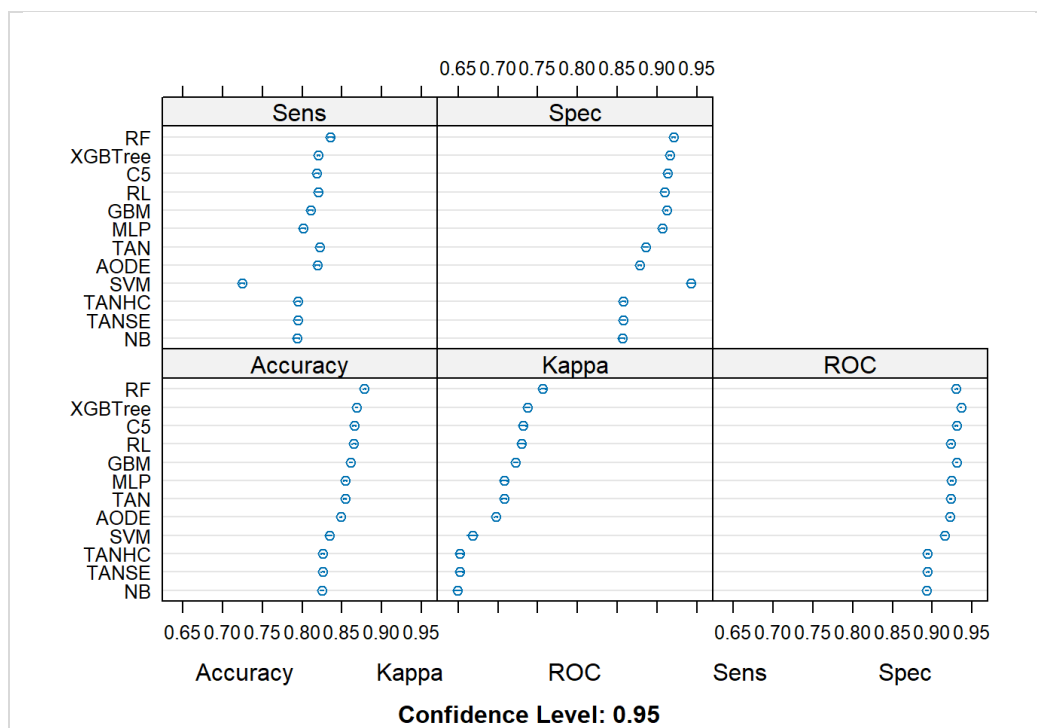


Fig. 5.1.1.1. Comparativa métricas

Y revisando una comparativa de la curva ROC para los principales modelos:

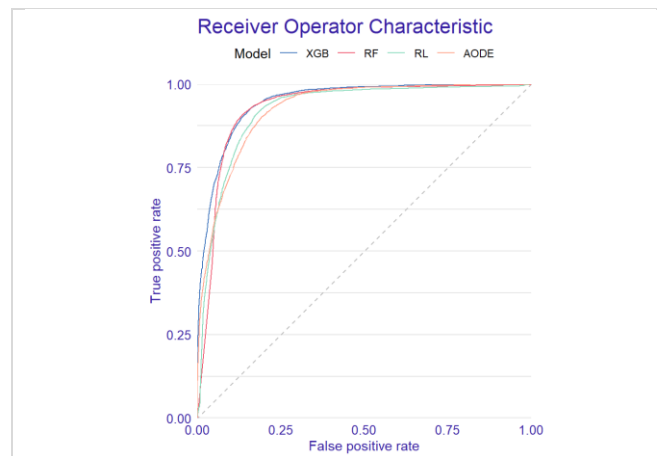


Fig. 5.1.1.2. Comparativa RCO

Se puede ver que siguen una distribución muy similar. Respecto a las métricas de cada modelo, se pueden considerar las siguientes afirmaciones particulares:

- **Random Forest (RF):**

Destaca con la AUC más alta (0.957) y una accuracy impresionante del 93.7%. Además, presenta un buen equilibrio entre sensibilidad (91.2%) y especificidad (96.3%). El coeficiente Kappa de 0.875 indica un acuerdo sustancial entre las predicciones del modelo y las clasificaciones reales.

- **Gradient Boosting (GBM):**

Aunque ligeramente por debajo de Random Forest, GBM muestra un rendimiento sólido con una AUC de 0.935 y una accuracy del 86.5%. La sensibilidad (81.3%) y especificidad (91.7%) sugieren que es capaz de manejar ambas clases de manera efectiva.

- **Decision Tree (C5):**

Similar a GBM en muchos aspectos, el árbol de decisión alcanza una AUC de 0.933 y una accuracy del 86.8%. La sensibilidad (82.5%) y la especificidad (91.2%) son consistentes con un buen rendimiento en la clasificación.

- **Support Vector Machine (SVM):**

Aunque SVM muestra la especificidad más alta (94.3%), su accuracy general es más baja (83.2%). Esto sugiere que es más cauteloso al clasificar la clase positiva, lo cual puede ser beneficioso dependiendo del contexto.

- **Multilayer Perceptron (MLP):**

A pesar de tener una accuracy ligeramente más baja (85.1%), MLP demuestra una alta sensibilidad del 79.3%, lo que indica una capacidad robusta para identificar la clase positiva. La AUC de 0.926 sugiere un buen rendimiento en la discriminación de clases.

- **Naive Bayes (NB):**

Muestra un rendimiento decente con una AUC de 0.896 y una accuracy del 82.8%. Sin embargo, la sensibilidad (79.5%) es relativamente baja en comparación con otros modelos, lo que sugiere que puede tener dificultades para identificar la clase positiva.

En el contexto de un análisis de incidentes en el ámbito marino, la importancia de la sensibilidad frente a la especificidad en los modelos de machine learning se relaciona directamente con la naturaleza crítica y las posibles consecuencias graves de los incidentes marinos. Se pueden considerar los siguientes aspectos:

- **Naturaleza Crítica de los Incidentes Marinos:**

Los incidentes en el ámbito marino, como colisiones, encallamientos o derrames de sustancias peligrosas, pueden tener impactos significativos en la seguridad de la navegación, el medio ambiente marino y la vida humana. La capacidad del modelo para detectar correctamente estos eventos es crucial.

- **Sensibilidad**

La sensibilidad se refiere a la capacidad del modelo para identificar correctamente los casos positivos, es decir, los incidentes reales. En el contexto marino, una alta sensibilidad implica que el modelo es efectivo para detectar la mayoría de los incidentes, lo cual es esencial para garantizar una respuesta rápida y eficaz.

- **Especificidad**

La especificidad se refiere a la capacidad del modelo para identificar correctamente los casos negativos, es decir, situaciones normales sin incidentes. Aunque es importante, en el ámbito marino, la especificidad podría ser menos crítica en comparación con la sensibilidad. La razón es que los falsos negativos (incidentes no detectados) podrían tener consecuencias más graves que los falsos positivos (alarmas incorrectas).

- **Costos Asociados con Falsos Negativos:**

En el ámbito marino, perder un incidente real (falso negativo) puede tener consecuencias desastrosas en términos de seguridad y medio ambiente. Por ejemplo, no detectar un derrame de sustancias peligrosas a tiempo podría llevar a respuestas de emergencia tardías, exacerbando el impacto ambiental.

- **Preferencia por una Sensibilidad Alta:**

Dado que la prioridad en el ámbito marino suele ser la detección temprana y precisa de incidentes, es probable que se prefiera un modelo con una sensibilidad más alta, incluso si esto conlleva una especificidad ligeramente más baja. La capacidad de identificar la gran mayoría de incidentes puede ayudar a minimizar el riesgo y las consecuencias asociadas.

- **Adaptación a las Condiciones Marinas Variables:**

Las condiciones marinas pueden ser variables y desafiantes, lo que aumenta la importancia de tener un modelo robusto que pueda adaptarse a diversas situaciones y detectar incidentes en diferentes escenarios.

En el análisis de incidentes en el ámbito marino, la sensibilidad tiende a ser más crítica que la especificidad. Un modelo que maximice la sensibilidad, aunque a expensas de la especificidad, podría ser preferido para garantizar la detección efectiva de incidentes y, por lo tanto, contribuir a la seguridad marina y la protección del medio ambiente.

Hay varios modelos que se sitúan en esta situación, como es el caso de Random Forest, o Extreme Gradient Boosting que ofrecen buen rendimiento. El caso es que, Random Forest tiene métricas de evaluación ligeramente superiores, está considerado como más robusto y resistente al sobre ajuste. Además, su interpretabilidad es más sencilla y es muy eficiente en términos de tiempo de entrenamiento, por lo que se podría considerar la mejor opción para la resolución de este problema de clasificación.

5.1.2. Explicabilidad de los modelos “Posibilidad de incidente”

Una de las cuestiones más importantes para evaluar un modelo es conocer cómo ha trabajado y en este sentido, saber qué variables han pesado más a la hora de resolver el problema de clasificación. Para el análisis de este apartado, se han elegido los seis algoritmos que han presentado mejores métricas:

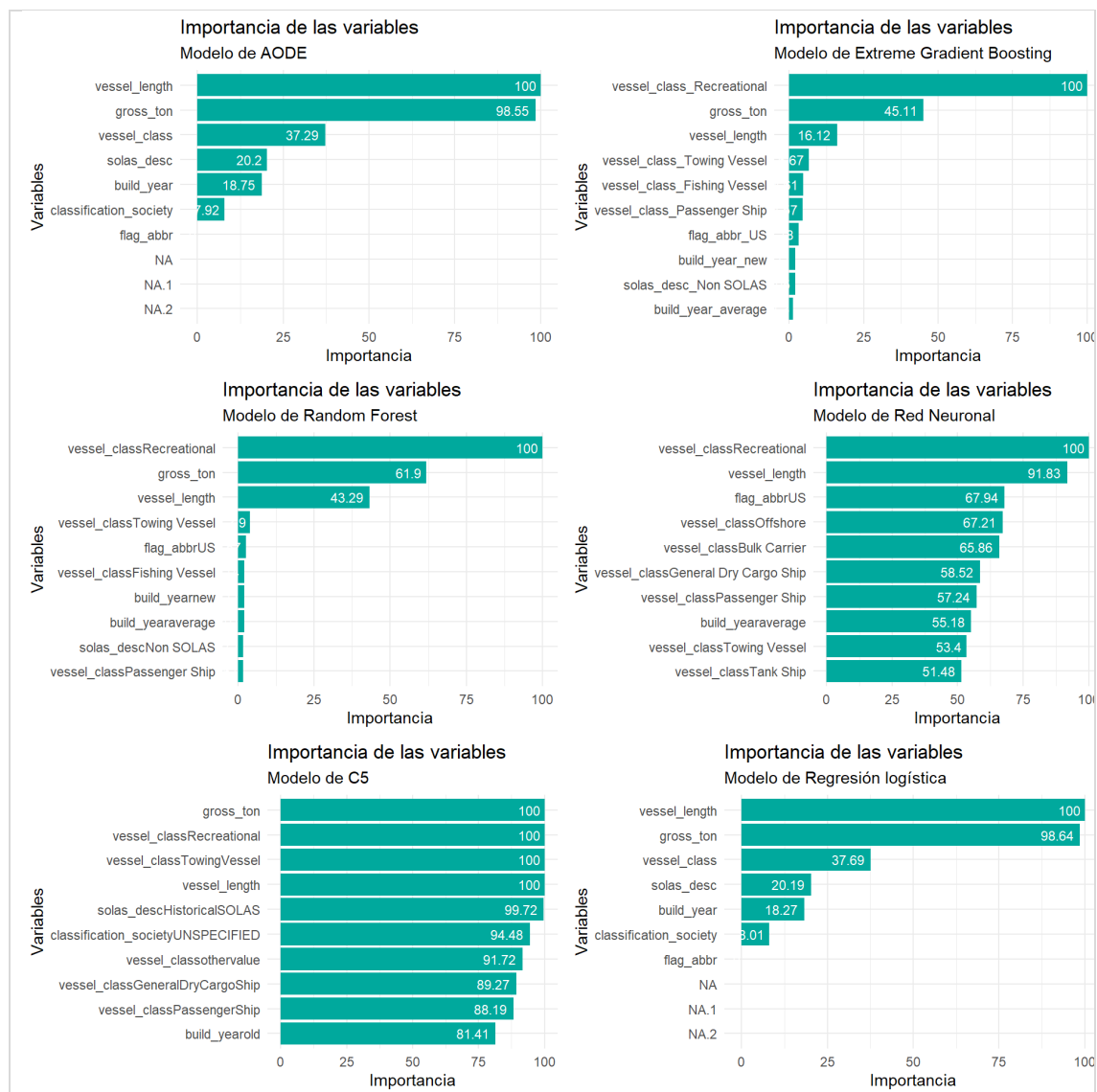


Fig. 5.1.2.1. Importancia variables

Las variables con mayor protagonismo en estos modelos corresponden a características de los barcos. En concreto, gross_ton (volumen) vessel_lenght (eslora) y la clase Recreational de vessel_class.

Random Forest, que ha sido el mejor considerado en base a sus métricas, basa su predicción en esas tres variables. En modelos como C5 o la red neuronal la aportación de las variables al resultado está más repartida, pero la clasificación de variables sigue teniendo las mismas protagonistas.

Por ello, se puede deducir que los mejores modelos predictivos para este problema, deberían contener al menos las tres variables señaladas.

Para ver más en detalle este aspecto, la librería modelStudio, ofrece un widget interactivo donde se puede evaluar la contribución y residuos de una variable concreta. En la siguiente captura hay un ejemplo para la variable más representativa del modelo Random Forest:

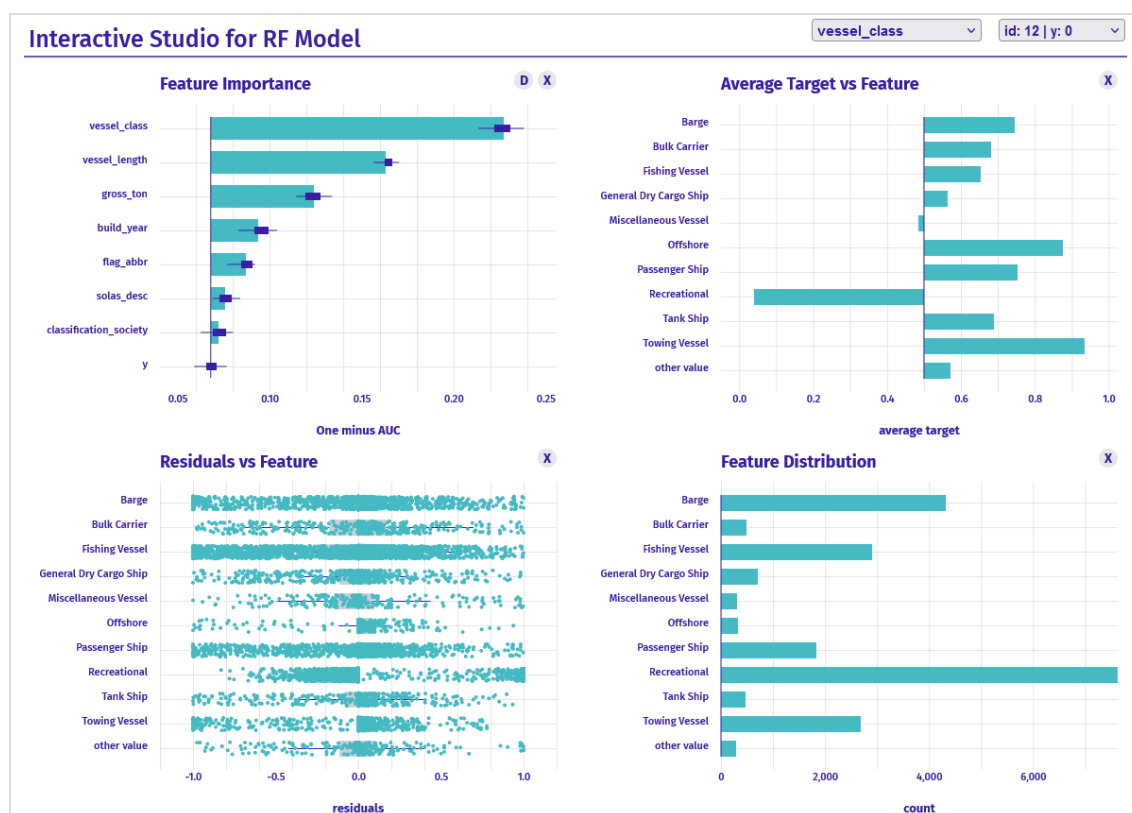


Fig. 5.1.2.2. Interactive Studio

La significancia de "vessel_class - Recreational" se refleja en diversas métricas de evaluación, donde su inclusión ha tenido un impacto significativo en la calidad general del modelo. La atención especial hacia esta variable resalta su poder discriminatorio y su capacidad para aportar información valiosa en estos modelos de machine learning. Este hallazgo no solo contribuye a la comprensión del modelo, sino que también puede tener implicaciones prácticas.

En el anexo 5.1. se pueden encontrar más gráficos sobre la dependencia acumulada de los modelos más importantes gracias a los explicadores de la librería DALEX, así como la contribución de las variables más importantes. Aquí, un ejemplo:

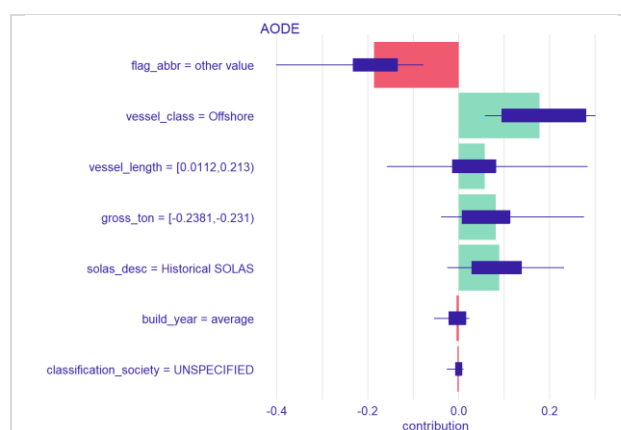


Fig. 5.1.2.3. Dalex: Aportación variable

5.2. Tipo de incidente: MergedActivity

Para el modelado en este conjunto de datos se ha utilizado el código del anexo 5.2. En este caso, sí ha sido necesario equilibrar la muestra y completar los valores faltantes, según los siguientes métodos:

5.2.0. Limpieza de datos: Equilibrado y tratamiento de valores ausentes

La variable respuesta, *event_class*, presenta antes de aplicar los modelos predictivos el siguiente desequilibrio en sus valores:

Variable:	<i>Critical Events</i>	<i>Maritime Accidents</i>	<i>Material Issues</i>	<i>Onboard Emergencies</i>	<i>Third-party Damages</i>
Observaciones:	16938	18467	17158	6630	8807

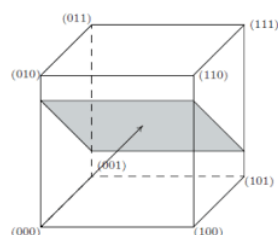
Puesto que los algoritmos necesitan una variable equilibrada para optimizar su funcionamiento, se ha optado por determinar 9000 observaciones por nivel de la variable para conformar la muestra final. Según esta cifra, se deberían generar un 26% de observaciones sintéticas para la clase menos representada (Onboard Emergencies)

Para el submuestreo se ha elegido el **método del Cubo** a Critical-Events, Maritime.Accidents y Material.Issues. Se trata de una técnica utilizada en el ámbito de la minería de datos y el procesamiento de señales para reducir la cantidad de datos en un conjunto de manera que se mantenga una representación significativa de la información original. Para ello, este método selecciona muestras de manera que los estimadores de Horvitz-Thompson sean iguales a los totales de las variables auxiliares conocidas, lo que contribuye a equilibrar la muestra.

Consiste en dos fases, según el método propuesto por Deville y Tillé (2004):

1. Fase de vuelo:

Se puede conceptualizar como un camino aleatorio que inicia en un vector de probabilidad de inclusión π y se mueve a través de un espacio de características. Este camino aleatorio permanece en la intersección entre el cubo y el subespacio definido por las ecuaciones de equilibrio Q .



2. Fase de aterrizaje:

Al finalizar la fase de vuelo, si un vértice (muestra) no ha sido seleccionado, se activa la fase de aterrizaje. Hay tres posibles soluciones para la fase de aterrizaje:

a. Eliminar Progresivamente Variables de Equilibrio: Se eliminan progresivamente las variables de equilibrio y se vuelve a aplicar la fase de vuelo. Es necesario suprimir las variables en orden de menor a mayor importancia.

b. Programación Lineal para Muestra Aproximadamente Equilibrada. Se utiliza la programación lineal para calcular la mejor muestra aproximadamente equilibrada. El objetivo es minimizar la diferencia en equilibrio entre las categorías.

c. Escoger el Vértice más Cercano al Vector de Probabilidades Se selecciona el vértice más cercano al vector de probabilidades obtenido en la fase de vuelo. Se redondean las probabilidades de inclusión que aún no son iguales a 0 o 1.

Implementación Eficiente

Deville y Tillé desarrollaron una implementación más rápida de la fase de vuelo, detallada en "Fast SAS Macros for balancing simples user's guide." Esta implementación contribuye a la eficiencia del método y ofrece ventajas, como no tener restricciones en el tamaño de la población. El tiempo de ejecución depende linealmente del tamaño de la población, lo que mejora la escalabilidad del método.

Este método es comúnmente aplicado en situaciones en las que el conjunto de datos es demasiado grande para procesar eficientemente o cuando se busca equilibrar clases desproporcionadas, tal como sucede en esta ocasión

Para el sobremuestreo, se aplicó el método **SMOTE** a Onboard.Emergencies, Third-party.Damages. Existen otros métodos como ROS y ADASYN, pero para este análisis se ha optado por este por su buen desempeño y amplia aceptación generalizada para esta tarea.

SMOTE (Synthetic Minority Over-sampling Technique) tiene como objetivo abordar el problema de clases desequilibradas mediante la generación de instancias sintéticas de las clases minoritarias. Se pueden distinguir dos fases principales:

1. Selección de Instancias:

Para las instancias a sobremuestrear, son seleccionados k vecinos cercanos (por lo general, se elige $k=5$). Los vecinos se seleccionan utilizando alguna métrica de distancia, como la distancia euclidiana.

2. Proceso de Generación:

Para cada característica de la instancia, SMOTE calcula la diferencia entre el valor de la característica de la instancia original y el valor correspondiente en uno de los vecinos. Multiplica esta diferencia por un número aleatorio entre 0 y 1 y lo suma al valor original de la característica para generar el valor de la nueva instancia sintética.

Este método opera en el espacio de características, creando instancias sintéticas que forman una línea en el espacio de características entre la instancia original y sus vecinos. Esto ayuda a expandir el espacio ocupado por la clase minoritaria y mejora la capacidad del modelo para generalizar correctamente.

SMOTE tiene varias variantes, como SMOTE-NC (para conjuntos de datos con características categóricas), Borderline-SMOTE (se centra en instancias cercanas al límite entre las clases)

Tras la aplicación de ambas técnicas, se ha obtenido una muestra ya balanceada. Sin embargo, sí hay valores ausentes de los que, a continuación, se explica su tratamiento.

Respecto al tratamiento de valores ausentes, en este conjunto de datos, se han detectado tres variables con este problema en mayor o menor medida:

Variable:	<i>air_temp</i>	<i>wind_speed</i>	<i>damage_assessment</i>
Valores ausentes:	2030	5513	15

Los valores ausentes pueden inducir a introducir sesgos en la estimación de modelos ya que distorsionan la variabilidad y generalización. Por lo que evitar un tratamiento de valores ausentes, impacta directamente en la precisión y desempeño. E incluso muchos algoritmos de aprendizaje automático ni siquiera comiencen la fase de entrenamiento si se declara este problema.

En este análisis, se han evaluado dos métodos con ayuda de la librería *mice* para verificar cuál de los dos es más eficiente para este conjunto de datos ya balanceado respecto a su variable objetivo.

Por un lado, la aplicación de un algoritmo CART. (Classification and Regression Trees) es un algoritmo de aprendizaje automático que se utiliza comúnmente para tareas de clasificación y regresión, como los anteriormente explicados. Aunque su aplicación principal es la construcción de árboles de decisión, también puede ser utilizado para la imputación de valores ausentes en conjuntos de datos. La imputación de valores ausentes se realiza mediante arboles de decisión.

En el otro lado de la comparativa, se ha aplicado Random Forest. Este algoritmo, de por sí, tiene tolerancia a los valores ausentes para su aplicación en la construcción de modelos predictivos. Pero también es reconocida su estabilidad y robustez en la imputación de valores ausentes, por lo que se estimó conveniente incluirlo en la comparativa.

Para realizar la comparativa, se ha evaluado la diferencia de medias entre el conjunto de datos original y los obtenidos con ambos métodos:

Comparación de medias

Dataset	mean_air_temp	mean_wind_speed	mean_damage_assessment	dif_total
MergedActivity Balanced (Original)	151.2707	50.51718	104365.3	0.00000
MergedActivity BalancedCart	151.6104	49.00846	104383.4	16.92924
MergedActivity BalancedRF	151.7052	49.09728	104420.8	54.50630

Según el resultado de esta comparativa, el método que menor diferencia de medias ha obtenido es CART, con una diferencia total de 16,92. De modo que se ha optado por este algoritmo para, ahora con los datos balanceados y sin valores ausentes, empezar a trabajar con los modelos predictivos.

En cuanto a la configuración para el entrenamiento de los modelos:

Método utilizado: para el entrenamiento de los modelos de este problema de clasificación binaria, se han utilizado las librerías caret, keras tensorflow y h2o.

Particionado de datos: 80% de observaciones para entrenamiento (train) y 20% para test. Se ha dividido con una selección al azar.

Validación cruzada: “repeatedcv n=8, rep=2”. Es decir, Validación cruzada repetida. Es una técnica de evaluación del rendimiento del modelo que combina la validación cruzada (cv) con repeticiones. En este caso, el conjunto de datos se divide en 8 pliegues (folds) y se realizará la validación cruzada 2 veces. La repetición puede ayudar a reducir la varianza del rendimiento estimado.

Métrica de evaluación: En esta ocasión, al trabajar con un predictor multiclase, se ha elegido “logLoss”. Log Loss penaliza las predicciones incorrectas asignando una mayor penalización cuando la probabilidad asignada a la clase correcta es baja. Esto refleja mejor la confianza del modelo en sus predicciones.

En cuanto a los resultados, al igual que el caso anterior, en esta memoria solo va a destacar la comparativa de modelos, ya que el detalle de los parámetros utilizados en cada modelo y sus outputs concretos se podrán encontrar en el anexo mencionado. En este apartado se han considerado un menor número de modelos, aunque teniendo como objetivo aportar diversidad de herramientas, se han utilizado modelos procesados con tres librerías, siendo una labor sencilla extrapolar las configuraciones a otros modelos.

5.2.1. Comparativa de los modelos “Tipo de incidente”

En este caso, se van a comparar los modelos calculados con Caret y Keras tensorflow conjuntamente y la opción de H2o como extra ya que, por su naturaleza, las métricas ofrecen dificultades para ser comparadas con los demás modelos. En cualquier caso, se dará una comparativa para los 7 modelos más adelante

Comparación con la Muestra de Entrenamiento

	AUC	Accuracy	Kappa	Sensitivity	Specificity
RF	0.969	0.800	0.750	0.800	0.950
C5	0.874	0.598	0.497	0.598	0.899
GBM	0.820	0.524	0.406	0.524	0.881
MLP	0.730	0.415	0.269	0.415	0.854
NB	0.725	0.417	0.271	0.417	0.854
Keras	0.629	0.446	0.308	0.458	0.863

Comparación con la Muestra de Validación

	AUC	Accuracy	Kappa	Sensitivity	Specificity
GBM	0.759	0.438	0.298	0.438	0.860
NB	0.726	0.420	0.275	0.420	0.855
MLP	0.724	0.409	0.261	0.409	0.852
C5	0.714	0.413	0.266	0.413	0.853
RF	0.712	0.380	0.225	0.380	0.845
Keras	0.625	0.427	0.283	0.438	0.857

Ahora los rendimientos alcanzados por los modelos predictivos ni son tan buenos ni son tan similares entre sí, como en la cuestión anterior. Para el conjunto de entrenamiento, no se dan malas métricas en Random Forest, pero en el caso del conjunto de validación, podrían considerarse insuficientes en todos los casos. El desempeño de cada modelo sería el siguiente:

• **Gradient Boosting (GBM):**

GBM tiene un AUC respetable y una alta especificidad, lo que indica su capacidad para identificar casos negativos. Sin embargo, la baja sensibilidad sugiere que puede tener dificultades para detectar la mayoría de los incidentes, lo cual es crítico en el contexto marítimo.

• **Naive Bayes (NB):**

NB muestra resultados moderados en AUC y especificidad. La sensibilidad es relativamente baja, lo que sugiere que podría tener dificultades para identificar eficazmente los incidentes marinos.

• **Multilayer Perceptron (MLP):**

MLP presenta resultados similares a NB con AUC y especificidad moderadas. La sensibilidad es una preocupación, lo que indica la posibilidad de perder incidentes importantes.

• **Decision Tree (C5):**

Similar a MLP y NB en términos de rendimiento. La sensibilidad sigue siendo baja, indicando limitaciones en la identificación de incidentes. Además, como se verá a continuación tiene un Log Loss prácticamente inaceptable.

• **Random Forest (RF):**

RF muestra una especificidad sólida, pero la sensibilidad es baja. Puede ser eficaz para identificar situaciones normales, pero podría perder incidentes marinos importantes.

• **Keras (Red Neuronal Densamente poblada):**

Keras tiene una sensibilidad relativamente alta, lo cual es positivo en el contexto marítimo. Sin embargo, la baja AUC sugiere que puede haber margen de mejora en la discriminación de clases.

Tras diferentes configuraciones de hiperparámetros y selección de características, no se han apreciado grandes mejoras en los resultados, así que se ha incluido en el análisis el modelo AutoML (Auto Machine Learning) que implementa el framework h2o. AutoML explora automáticamente diferentes algoritmos y configuraciones de hiperparámetros, realiza optimización automática, utiliza validación cruzada y combina varios modelos para mejorar la precisión general. Además, maneja automáticamente datos faltantes y variables categóricas.

En este caso, se puede comparar con los demás modelos a través de la métrica Log Loss:

	logloss
Keras	1.330660
GBM	1.338944
MLP	1.425373
AutoML	1.495675
NB	1.552391
RF	1.635759
C5	5.821327

Tomando en cuenta esta métrica, AutoML tiene un desempeño similar a los anteriores modelos por lo que no se puede considerar una opción a tener en cuenta. Este modelo es mucho más complejo y no ofrece una explicabilidad tan buena como Random Forest.

Centrándonos en Random Forest, podemos decir que en general, el modelo parece tener un rendimiento bastante bueno en la muestra de entrenamiento, pero hay señales de posible sobreajuste, ya que el rendimiento en la muestra de validación es considerablemente inferior. Sería procedente explorar estrategias para mejorar la generalización del modelo, como otros ajustes de hiperparámetros o reconsiderar una selección y/o tratamiento diferente de características.

5.2.2. Explicabilidad de los modelos “Tipo de incidente”

En este caso, vamos a centrar el análisis en la importancia de las variables de los cuatro modelos que mejor resultados han obtenido: Random Forest, C5, Gradient Boosting y el Perceptrón Multicapa:

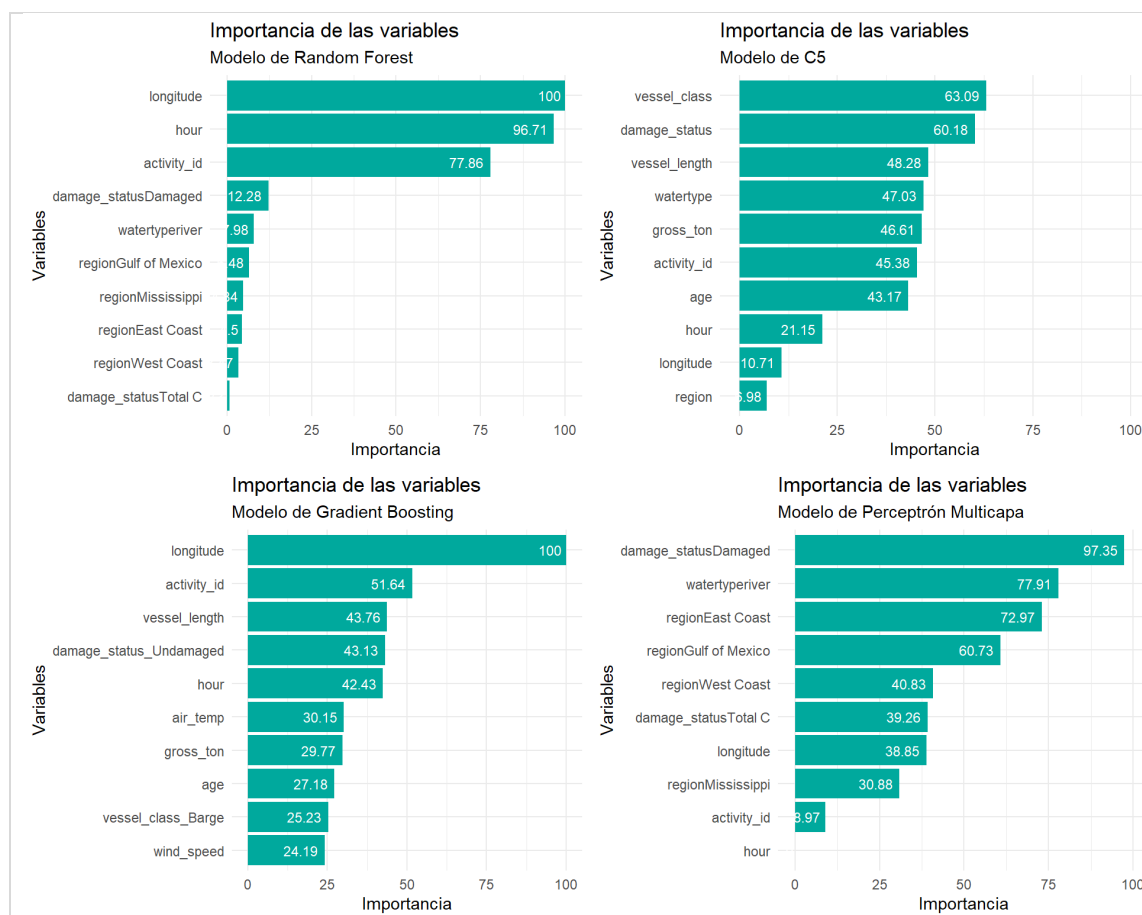


Fig. 5.2.2.1. Importancia variables

En este caso, también se han obtenido resultados dispares entre los modelos. Tanto en los modelos de Random Forest como Gradient Boosting, la variable longitude (es decir, geoposición en sentido horizontal) es la variable más destacada. Estos algoritmos pueden capturar interacciones no lineales entre variables de manera efectiva. Teniendo en cuenta la geografía en la que se desarrolla este análisis, podemos concluir que uno de los aspectos más importantes para determinar la importancia de un incidente según estos modelos es la proximidad a la costa.

De mismo modo, el nivel Golfo de México de la variable region, también aparece en varias ocasiones. Con esto, los modelos nos dicen que, si un barco está ahí, tendrá mayor posibilidad de verse involucrado en determinados incidente.

Otra de las variables que aparece en estas gráficas es damage_status. Esta variable determina si se han producido daños. De una manera lógica, la gravedad de un incidente, da a entender que cuanto mayor es su importancia, mayores daños se producen.

6. Conclusiones

El presente estudio revela la utilidad del modelo Random Forest en la predicción de incidentes marinos, destacando su capacidad para abordar conjuntos de datos complejos y capturar patrones no lineales de manera efectiva. Uno de los factores críticos para el éxito de este análisis ha sido el exhaustivo preprocesamiento de datos, que incluyó la unión de información procedente de diversas fuentes. Esta integración de datos de múltiples sitios no solo enriqueció el conjunto de datos, sino que también proporcionó una representación más completa de las condiciones marinas, mejorando así la generalización del modelo a situaciones del mundo real.

Tal como se ha explicado, una de las claves para la elección de un modelo de aprendizaje automático en el ámbito de la prevención de incidentes marinos es priorizar la sensibilidad frente a la especificidad. En este sentido, la habilidad del modelo propuesto para gestionar datos provenientes de diferentes sitios resalta su robustez y adaptabilidad, aspectos esenciales en el contexto marino, donde la variabilidad geográfica y las condiciones específicas de cada región desempeñan un papel crucial en la frecuencia de incidentes.

La generalización del modelo a situaciones del mundo real se ve respaldada por la diversidad de datos considerados, permitiendo que el modelo aprenda patrones más representativos y, por lo tanto, aumentando su utilidad en la predicción de incidentes en nuevas circunstancias.

A la vista de los resultados, se podría establecer que este análisis puede aportar un modelo que responda a la primera de las cuestiones planteadas: ¿Se verá involucrado un determinado barco en un incidente? Sin embargo, la segunda pregunta en la que se discierne qué tipo de incidente sería, por el momento, no se puede hacer una afirmación tan clara. Queda para el futuro, plantear otros modelos y características que ofrezcan mayores ajustes en este apartado.

También quedaría para el futuro otra interesante cuestión que se ha quedado en el tintero: ¿Cuánto coste tendría un incidente para un determinado barco? En este caso, estaríamos ante un problema de regresión para la variable `damage_assessment`, que cuenta con una aparente buena calidad de datos en el dataset `MergedActivity`.

Sin embargo, la aplicación práctica de estos resultados en operaciones marítimas puede ser inmediata, ya que la capacidad del modelo para prever incidentes puede influir significativamente en la seguridad y la toma de decisiones operativas. La información anticipada proporcionada por el modelo permite una respuesta proactiva a posibles riesgos, lo que puede traducirse en una reducción significativa de incidentes y sus consecuencias.

Es fundamental reconocer que este análisis es un proceso iterativo, y se espera que el modelo se beneficie de mejoras continuas. La retroalimentación constante y la actualización del modelo con nuevos datos contribuirán a su mejora y adaptación a cambios en las condiciones marítimas a lo largo del tiempo.

Por último, es imperativo subrayar la importancia de consideraciones éticas y de privacidad al implementar un modelo predictivo en el entorno marítimo. Garantizar que la recopilación y el uso de datos se realicen de manera ética y respetuosa con la privacidad de las partes involucradas es esencial para el éxito sostenido y la aceptación de la aplicación del modelo en entornos operativos.

7. Bibliografía

Referencias principales:

Clases, guías y material propio impartido durante el Máster en Big Data y Data Science aplicados a la economía y a la administración y dirección de empresas en UNED, curso 2023. En particular, el material correspondiente a los módulos de análisis multivariante y minería de datos. Las funciones en R para la evaluación de los modelos, son una adaptación de las contenidas en el archivo *FUNCIONES_MODULO_7.R* proporcionado durante la autoevaluación del módulo Minería de Datos II.

Recursos web consultados:

- Marine Casualty & Pollution Data for Researchers (USCG)
<https://www.dco.uscg.mil/Our-Organization/Assistant-Commandant-for-Prevention-Policy-CG-5P/Inspections-Compliance-CG-5PC-/Office-of-Investigations-Casualty-Analysis/Marine-Casualty-and-Pollution-Data-for-Researchers/>
- Climate Data Online: Dataset Discovery (NOAA)
<https://www.ncei.noaa.gov/cdo-web/datasets>
- European Maritime Safety Agency (EMSA)
<https://www.emsa.europa.eu/es/>
- Marine occurrence and vessel data from January 2004
<https://www.bst-tsb.gc.ca/eng/stats/marine/data-2.html>

Referencias bibliográficas:

- Adin Urtasun, Aritz (2012): *El método del cubo: Aplicaciones del muestreo equilibrado en la organización estadística vasca*. Euskal Estatistika Erakundea – Instituto Vasco de Estadística.
- Aldás, Joaquin y Uriel, Ezequiel (2017): *Análisis multivariante aplicado con R*, Ediciones Paraninfo, S.A. ISBN: 978-84-283-2969-9
- Bruce, Peter y Bruce, Andrew y Gedeck Peter (2020): *Practical Statistics for Data Scientists*, O'Reilly Media, Inc. ISBN: 978-1-492-07294-2
- Lantz, Brentt (2013): *Machine Learning with R*, Packt Publishing Ltd. ISBN: 978-1-78216-214-8
- Chollet, François y Kalinowski, Tomasz y Allaire, J.J. (2022): *Deep Learning with R*, Manning Publications, Co. ISBN: 9781638350781



Esta obra está sujeta a una licencia Creative Commons BY-NC-ND 4.0 DEED
Atribución-NoComercial-SinDerivadas 4.0 Internacional
<https://creativecommons.org/licenses/by-nc-nd/4.0/>