Nombre: Oscar Sebastián Pulido Pinzón

# Problema de negocio

Una empresa aseguradora quiere estudiar el riesgo asociado a la ocurrencia de incidentes entre aviones comerciales y vida salvaje en diferentes regiones geográficas. Para esto cuenta con una colección completa de incidentes de ataques con aviones militares, comerciales o civiles entre 1990 y 2023. Los incidentes se obtuvieron de la base de datos de ataques contra la vida silvestre de la Administración Federal de Aviación (FAA). El conjunto de datos cubre una variedad de información relacionada con estos incidentes, incluidos detalles sobre las aeronaves involucradas, especies de vida silvestre, ubicaciones geográficas y diversos factores que contribuyeron a los sucesos, entre otros.

Las características claves incluyen:

* INCIDENT\_DATE: Fecha del incidente.
* AIRPORT: Aeropuerto inplicado en el accidente.
* AIRCRAFT: Información de la aeronave, incluyendo tipo y registro.
* SPECIES: Especie salvaje implicada en el incidente.
* LOCATION: Localización geografica del incidente.
* PHASE\_OF\_FLIGHT: Fase del vuelo durante la cual ocurrió el incidente.
* DAMAGE\_LEVEL: Nivel de daño resultante del incidente.
* WEATHER: Información acerca de las condiciones de vuelo del incidente.

Se requiere el diseño de una estrategia de análisis predictivo para:

* Entender tendencias temporales y espaciales en la ocurrencia de los incidentes.
* Predecir la severidad de los incidentes para una región espacial y en un momento de tiempo particular.

**Diccionario de datos**

INDEX NR Individual record number

OPID Airline operator code

OPERATOR A three letter International Civil Aviation Organization code for aircraft operators. (BUS = business, PVT = private aircraft other than business, GOV = government aircraft, MIL - military aircraft.)

ATYPE Aircraft

AMA International Civil Aviation Organization code for Aircraft Make

AMO International Civil Aviation Organization code for Aircraft Model

EMA Engine Make Code (see Engine Codes tab below)

EMO Engine Model Code (see Engine Codes tab below)

AC\_CLASS Type of aircraft (see Aircraft Type tab below)

AC\_MASS 1 = 2,250 kg or less: 2 = ,2251-5700 kg: 3 = 5,701-27,000 kg: 4 = 27,001-272,000 kg: 5 = above 272,000 kg

NUM\_ENGS Number of engines

TYPE\_ENG Type of power A = reciprocating engine (piston): B = Turbojet: C = Turboprop: D = Turbofan: E = None (glider): F = Turboshaft (helicopter): Y = Other

ENG\_1\_POS Where engine # 1 is mounted on aircraft (see Engine Position tab below)

ENG\_2\_POS Where engine # 2 is mounted on aircraft (see Engine Position tab below)

ENG\_3\_POS Where engine # 3 is mounted on aircraft (see Engine Position tab below)

ENG\_4\_POS Where engine # 4 is mounted on aircraft (see Engine Position tab below)

REG Aircraft registration

FLT Flight number

REMAINS\_COLLECTED Indicates if bird or wildlife remains were found and collected

REMAINS\_SENT Indicates if remains were sent to the Smithsonian Institution for identifcation

INCIDENT\_DATE Date strike occurred

INCIDENT\_MONTH Month strike occurred

INCIDENT\_YEAR Year strike occurred

TIME\_OF\_DAY Light conditions

TIME Hour and minute in local time

AIRPORT\_ID International Civil Aviation Organization airport identifier for location of strike whether it was on or off airport

AIRPORT Name of airport

STATE State

FAAREGION FAA Region where airport is located

ENROUTE If strike did not occur on approach, climb, landing roll, taxi or take-off, aircraft was enroute. This shows location.

RUNWAY Runway

LOCATION Various information about aircraft location if enroute or airport where strike evidence was found. Some locations show the two airports for the flight departure and arrival if pilot was unaware of the strike.

HEIGHT Feet Above Ground Level

SPEED Knots (indicated air speed)

DISTANCE Nautical miles from airport

PHASE\_OF\_FLT Phase of flight during which strike occurred

DAMAGE Level of damage selected by the Database Manager. See below for ICAO definitions taken from ICAO IBIS Manual Fourth Edition-2001

Blank Unknown

N = None No damage was reported.

M = Minor When the aircraft can be rendered airworthy by simple repairs or replacements and an extensive inspection is not necessary.

M? = Undetermined level The aircraft was damaged, but details as to the extent of the damage are lacking.

S = Substantial When the aircraft incurs damage or structural failure which adversely affects the structure strength, performance or flight characteristics of the aircraft and which would normally require major repair or replacement of the affected component. Bent fairings or cowlings; small dents or puncture holes in the skin;  damage to wing tips, antennae, tires or brakes; and engine blade damage not requiring blade replacement are specifically excluded.

D = Destroyed When the damage sustained makes it inadvisable to restore the aircraft to an airworthy condition.

STR\_RAD Struck radome

DAM\_RAD Damaged radome

STR\_WINDSHLD Struck windshield

DAM\_WINDSHLD Damaged windshield

STR\_NOSE Struck nose

DAM\_NOSE Damaged nose

STR\_ENG1 Struck Engine 1

DAM\_ENG1 Damaged Engine 1

ING\_ENG1 Ingested Engine 1

STR\_ENG2 Struck Engine 2

DAM\_ENG2 Damaged Engine 2

ING\_ENG2 Ingested Engine 2

STR\_ENG3 Struck Engine 3

DAM\_ENG3 Damaged Engine 3

ING\_ENG3 Ingested Engine 3

STR\_ENG4 Struck Engine 4

DAM\_ENG4 Damaged Engine 4

ING\_ENG4 Ingested Engine 4

INGESTED\_OTHER Wildlife ingested in a location other than an engine, effective 3/29/2021 (ALL ingestions for strikes submitted prior to 3/29/2021 are shown here)

STR\_PROP Struck Propeller

DAM\_PROP Damaged Propeller

STR\_WING\_ROT Struck Wing or Rotor

DAM\_WING\_ROT Damaged Wing or Rotor

STR\_FUSE Struck Fuselage

DAM\_FUSE Damaged Fuselage

STR\_LG Struck Landing Gear

DAM\_LG Damaged Landing Gear

STR\_TAIL Struck Tail

DAM\_TAIL Damaged Tail

STR\_LGHTS Struck Lights

DAM\_LGHTS Damaged Lights

STR\_OTHER Struck Other than parts shown above

DAM\_OTHER Damaged Other than parts shown above

OTHER\_SPECIFY What part was struck other than those listed above

EFFECT Effect on flight

EFFECT\_OTHER Effect on flight other than those listed on the form

SKY Type of cloud cover, if any

PRECIP Precipitation

BIRD\_BAND\_NUMBER Bird Band Number associated with the wildlife struck

SPECIES\_ID International Civil Aviation Organization code for type of bird or other wildlife

SPECIES Common name for bird or other wildlife

BIRDS\_SEEN Number of birds/wildlife seen by pilot

BIRDS\_STRUCK Number of birds/wildlife struck

SIZE Size of bird as reported by pilot is a relative scale. Entry should reflect the perceived size as opposed to a scientifically determined value. If more than one species was struck, larger bird is entered.

WARNED Pilot warned of birds/wildlife

COMMENTS As entered by database manager. Can include name of aircraft owner, types of reports received, updates, etc.

REMARKS Most of remarks are from the form but some are data entry notes and are usually in parentheses.

AOS Time aircraft was out of service in hours. If unknown, it is blank.

COST\_REPAIRS Estimated cost of repairs of replacement in dollars (USD)

COST\_OTHER Estimated other costs, other than those in previous field in dollars (USD). May include loss of revenue, hotel expenses due to flight cancellation, costs of fuel dumped, etc.

COST\_REPAIRS\_INFL\_ADJ Costs adjusted to the most recent year based on Consumer Price Index, U.S. Department of Labor.  Inflation-adjusted costs are updated annually.

COST\_OTHER\_INFL\_ADJ Costs adjusted to the most recent year based on Consumer Price Index, U.S. Department of Labor.  Inflation-adjusted costs are updated annually.

REPORTED\_NAME Name(s) of person(s) filing report

REPORTED\_TITLE Title(s) of person(s) filing report

REPORTED\_DATE Date report was written

SOURCE Type of report. Note: for multiple types of reports this will be indicated as Multiple. See "Comments" field for details

PERSON Only one selection allowed. For multiple reports, see field "Reported Title"

NR\_INJURIES Number of people injured

NR\_FATALITIES Number of human fatalities

LUPDATE Last time record was updated

TRANSFER Unused field at this time

INDICATED\_DAMAGE Indicates whether or not aircraft was damaged

# Tarea 1

1. Respecto a la tarea de entender las tendencias temporales y espaciales en los accidentes, describa un problema de analítica predictiva que pueda formular. Especifique el tipo de problema propuesto, describa las variables de entrada y salida a considerar en el modelo.

**Respuesta:** Se puede formular un problema predictivo con clasificación que use como insumos los datos históricos de accidentalidad, tales como la fecha de incidente, el aeropuerto, la localización, la fase del vuelo y el nivel de los daños, los motores dañados, entre otros para agrupar categorías y generar patrones de accidentalidad. En ese orden de ideas, las variables de salida podrían ser aquellas que permitan determinar aquellos lugares, temporadas y horas específicas en la que un accidente se más común a manera de clusters. Con esto se pueden replanificar algunas rutas y horas en función de disminuir la accidentalidad y teniendo patrones de clasificación que faciliten la identificación de estos insights.

1. Respecto a la tarea de predecir la severidad de ocurrencia de los accidentes para una región particular, describa que problema de analítica predictiva que puede formular. Especifique el tipo de problema propuesto, describa las variables de entrada y salida a considerar en el modelo.

**Respuesta:** Para determinar la severidad de los accidentes jugaría un papel clave un modelo de regresión que permita predecir la severidad como variable de salida en función de algunas variables relevantes, tales como el nivel de daño, el número de personas fallecidas, variables relevantes de costos en reparaciones y el nivel del daño. Antes de esto es necesario determinar mediante un filtro los datos a usar para una región en específico, esto se puede hacer usando la variable de localización y puede sugerir que se pueden hacer varias regresiones en simultaneo con la totalidad de los datos. Con esto se pueden estimar unos coeficientes que permitan predecir la severidad de una región e particular dependiendo de la data que se decida.

# Tarea 2

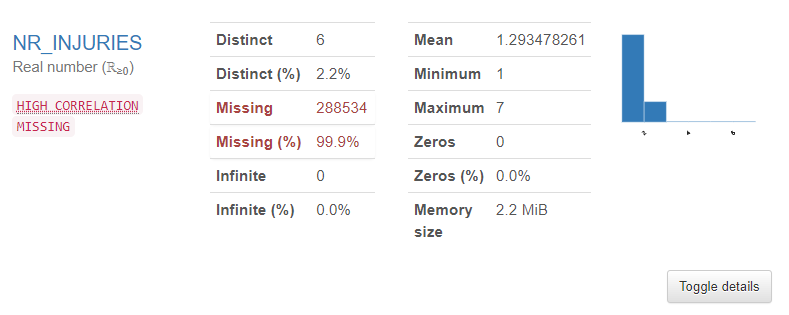
1. Su asistente prepara el siguiente reporte descriptivo sobre los datos disponibles:



Describa un posible problema con los datos a partir de la información de este reporte.

**Respuesta:** Se puede apreciar que la proporción de datos faltantes es significativa, representando casi un 25% del total de celdas del conjunto de datos. Independientemente del tipo de problema es deseable contar con completitud de datos. En esta fase es importante determinar la naturaleza de los faltantes, dado que puede ser esperado que no se tenga cierta información, sin embargo, en este caso parece necesario usar técnicas de imputación de datos o posibles reformulaciones y transformaciones del conjunto de datos para que la modelación no se vea muy impactada.

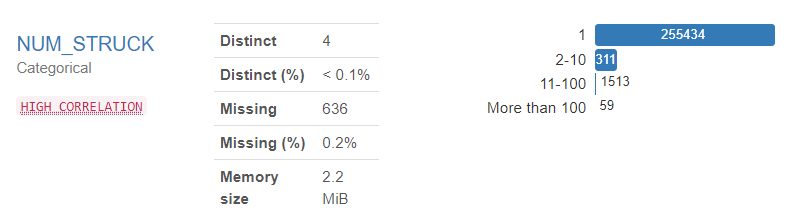
1. La variable NR\_INJURIES presenta el siguiente comportamiento\_



¿Cómo puede explicarse el gran número de datos no encontrados para esta variable?. Sugiera una posible solución.

**Respuesta:** La explicación podría estar en que la naturaleza de la variable parece indicar el número de personas lesiondas dentro del conjunto de datos. Es de esperar que las personas lesionadas en muchos casos sean cero porque se pueden contemplar accidentes “simples” relacionados a aterrizajes de emergencia o regresos al aeropuerto por falta de combustible, pero el valor mínimo es 1, por lo tanto, puede que el repositorio de los datos asuma que los datos entrados como cero sean asumidos como nulos y a eso se deben los posibles nulos. Una posible forma de corregir esta situación es transformar la variable en dos partes usando la naturaleza de duplicado de la variable. Entonces, en la primera parte se tendría la construcción de una variable tipo flag que determine si hubo lesión o no, mediante los valores de 1 y 0. En la segunda parte se podría construir una variable que determine el número de lesiones por individuo, de manera que su valor máximo es 7 y si no hay lesión toma el valor de 0.

1. La variable N\_STRUCK presenta el siguiente comportamiento:

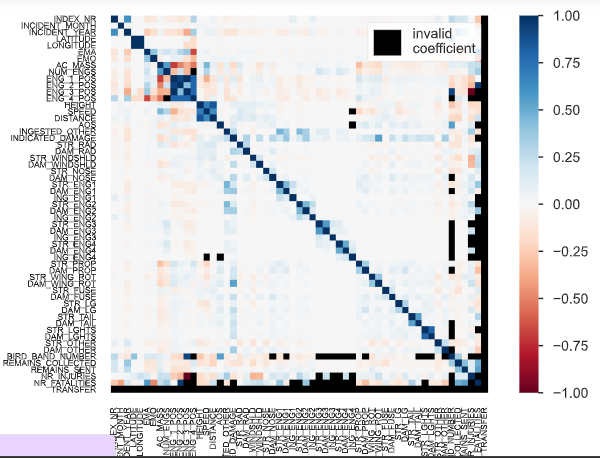


Sugiera tres métodos para completar o imputar los valores no presentes.

**Respuesta:**

* Siempre es una opción pensar en la metodología de imputación de la media como un proceso útil para tratar con datos faltantes.
* (No encontré esta variable en el diccionario) Dependiendo de la naturaleza de la variable se pueden determinar como cero, ya que si parecen indicar número de golpes, pero en realidad no hubo, y los sistemas asumen que un cero equivale a un valor nulo, entonces parece razonable imputar por valores de cero
* Existen otras técnicas para completar los datos como lo es la estandarización Z-Score que permite dar un tratamiento estricto a esos faltantes.

1. Un análisis bivariado sugiere el siguiente patrón de correlación entre las variables bajo estudio:



¿Cómo puede explicarse la alta correlación entre ENG\_1\_POS, ENG\_2\_POS, ENG\_3\_POS y ENG\_4\_POS?

**Respuesta:** Inevitablemente, y debido a la naturaleza de lo aviones, es posible esperar que la ubicación de los 4 motores presente una correlación muy alta, pues ya por temas de aerodinámica se espera que exista un posicionamiento estándar y que eso permita pensar que hay correlación entre estas, por lo tanto se puede pensar en relaciones lineales en dicho sentido.

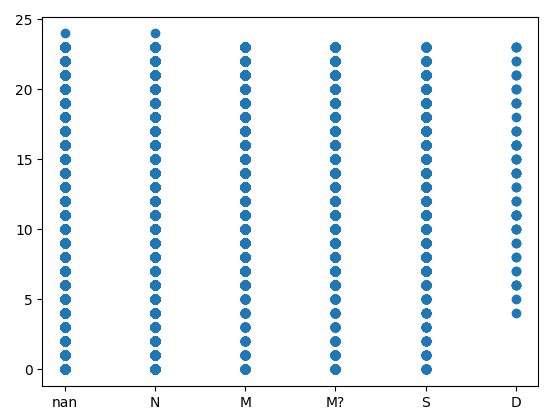
1. ¿Dados estos altos valores de correlación como trabajaría usted con estas cuatro variables en la fase de modelado?

**Respuesta:** Como puede que no aporten tanto al modelo, optaríamos por eliminarlas de la modelación, o por lo menos no usarlas simultáneamente, ya que pueden sesgar los resultados.

1. Su asistente sugiere calcular el coeficiente de correlación de Pearson entre la variable DAMAGE y la variable DISTANCE. ¿Qué puede decir acerca de esta elección de este método de correlación?

**Respuesta:** Que aunque el gráfico de correlaciones no muestre mucho, vale la pena determinar si existe una dependencia lineal entre ambas variables, ya que el daño y la magnitud de esto puede estar explicado en cierta manera por la distancia al aeropuerto más cercano, se esperaría que a mayores distancias mayores daños, una correlación directa, sin embargo de cara a la modelación por regresión puede contribuir a valores de bondad de ajuste inesperadamente elevados que pueden sesgar el análisis.

1. Usted quiere estudiar la relación entre la hora del día y la severidad del daño. Su asistente construye el siguiente gráfico considerando las variables TIME y DAMAGE\_LEVEL:



¿Qué comentarios puede darle a su asistente respecto a esta gráfica?. Sugiera una estrategia mejor para ilustrar la relación entre estas dos variables.

**Respuesta:** Podría pensarse en una transformación del gráfico que permita ver numéricamente la hora y minuto de los accidentes como variable independiente e ilustrada en el contexto de una serie de tiempo. La variable de nivel de daño es categórica, por lo tanto se puede realizar una agrupación que permita identificar los daños altos, medios y bajos y contabilizarlos de acuerdo a la hora y fecha de ocurrencia, se puede pensar en una gráfica de barras con todas las horas del día y en el eje y tendríamos la magnitud de los daños de acuerdo a cantidades de ocurrencia, es decir, una barra segmentada en el tipo de daño e indicado accidentes totales y proporción del tipo de daño dentro de ese total.

1. Su asistente identifica tres posibles variables que identifican la localización del accidente: AIRPORT, STATE, FAAREGION, LATITUDE, LONGITUDE. ¿Qué información sugiere utilizar usted para identificar la localización del incidente?, explique. ¿Cómo puede representar esta información?

**Respuesta:** Las variables elegidas parecen adecuadas, además de esas, yo usaría la variable location.

1. Su asistente sugiere utilizar la siguiente información para describir el momento de ocurrencia del incidente: INCIDENT\_DATE, INCIDENT\_MONTH, INCIDENT\_YEAR, TIME\_OF\_DAY, TIME. ¿Qué opinión tiene usted respecto a esta elección?

**Respuesta:** Estoy de acuerdo

# Tarea 3

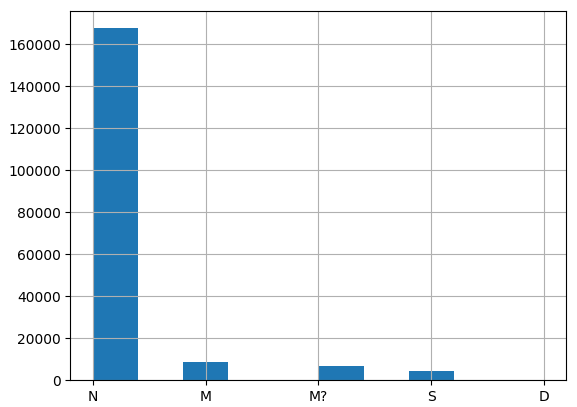
1. Después de discutir con su enlace en la organización se decide que la variable de interés a predecir es el nivel de daño es la variable DAMAGE\_LEVEL. Su asistente sugiere utilizar regresión para abordar el problema. ¿Por qué esta estrategia predictiva puede ser inadecuada para abordar el problema?, ¿Qué estrategia utilizaría?, explique.

**Respuesta:** Porque la regresión se usa con datos numéricos o binarios, dada la naturaleza del conjunto de datos, existen muchas variables que deberían ser eliminadas pero que pueden contribuir a la explicación de la variable de respuesta. Yo usaría un ejercicio de clasificación o de reducción de dimensionalidad e incluso una regresión de Lasso para escoger las variables más relevantes y para luego si usar regresión

1. Describa una medida de desempeño adecuada para evaluar la capacidad de predicción de la variable de interés.

**Respuesta:** Error cuadrático medio o error ajustado porcentual medio.

1. Su asistente despliega el siguiente histograma de la variable DAMAGE\_LEVEL una vez los datos nulos fueron removidos:



En que consiste el problema de imbalance de clases. ¿Existe algún riesgo de imbalance de clases en este caso?, explique.

**Respuesta:** Se ve que en la mayoría de los casos el nivel de daño reportado es nulo, lo cual es esperado. Por lo tanto, las clases no están balanceadas y se corre ese riesgo, al parecer es algo imposible de evitara a menos que el análisis decida enfocarse en aquellos niveles de daño moderado o superior.

1. Describa dos estrategias para abordar el problema de imbalance de clases.

**Respuesta:** Concentrar el análisis en los grupos balanceados. Finalmente, se pueden usar métodos apropiados según sea el caso como el over sampling o el under sampling aleatorio para duplicar o eliminar las muestras con el fin de lograr un equilibrio.

1. Sugiera una estrategia para validar el desempeño de generalización del modelo propuesto, explique.

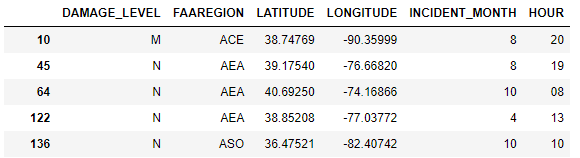
**Respuesta:**

1. ¿Explique en que consiste el *course de la dimensionalidad,* es posible que este fenómeno se presente en este problema?

**Respuesta:** Consiste en el desmejoramiento de un determinado modelo ante la adición de supuestas variables explicativas que no aportan al desempeño del modelo, por lo cual es importante evitar agregar variables que no contribuyan al buen ajuste. Sí se puede presentar en el modelo dada la gran cantidad de variables que puede que no aporten de la mejor manera.

# Tarea 4

1. Después de la fase de entendimiento y construcción del conjunto de datos para el modelo se propone construir un modelo predictivo considerando las siguientes variables:



Su asistente propone utilizar maquinas con regresión lineal, vectores de soporte (SVM), redes neuronales (NN) y árboles de decisión, para abordar el problema predictivo, ¿Qué procesamiento previo debe realizarse a estos datos para que puedan ser utilizados por los modelos propuestos?

**Respuesta:** Podrían hacerse los pasos estándares como el re escalamiento de variables, codificación de variables categóricas y determinación de variables finales de los datos. Una vez hecho eso se segmentan los datos en porción de prueba, validación y entrenamiento para iniciar el proceso de modelado.

1. Una vez evaluados los modelos se obtuvieron los siguientes desempeños en entrenamiento, validación y prueba con respecto al F1-score.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Entrenamiento | Validación | Prueba |
| Regresión lineal | 0.99 | 0.6 | 0.5 |
| SVM | 0.6 | 0.61 | 0.5 |
| NN | 0.6 | 0.8 | 0.5 |
| Arboles de decisión | 0.8 | 0.85 | 0.85 |

* ¿Comente respecto al bias y variance de cada modelo?

**Respuesta:**

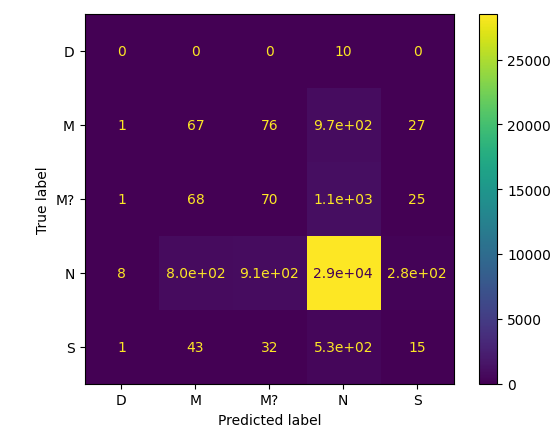
* ¿Cuál modelo escogería y por qué?

**Respuesta:**

1. Su asistente le informa que el modelo de árboles de decisión en test tuvo un accuracy de 0.85. Comente sobre este resultado.

**Respuesta:**

1. Para profundizar en el entendimiento de los resultados su asistente construyó la siguiente matriz de confusión:



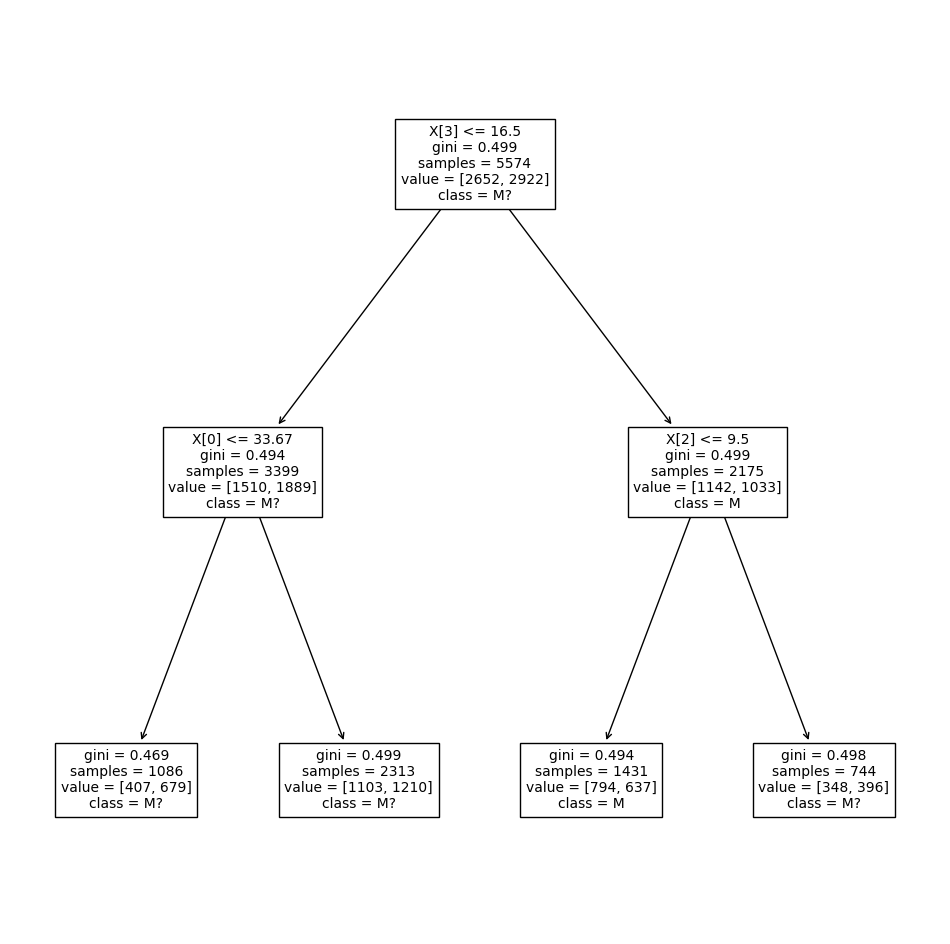
* Explique los resultados presentados.

**Respuesta:**

1. Como parte del análisis se considera incluir un modelo de regresión tipo Lasso. Describa el proceso para buscar el parámetro optimo del valor.

**Respuesta:**

1. El encargado de seguridad pregunta sobre la posibilidad de clasificar daños menores (M) y daños no determinados (M?). Para responder esta pregunta su asistente construye un árbol de decisión considerando las siguientes características ["LATITUDE", "LONGITUDE","INCIDENT\_MONTH","HOUR"]. El árbol de decisión tiene la siguiente forma:



Explique cómo el árbol de decisión toma la decisión de asignar la clase M al incidente.

**Respuesta:**

1. Como complemento al análisis propuesto se decide correr una regresión logística. Describa la función a optimizar en este problema predictivo.

**Respuesta:**

1. Como resultado de este experimento se obtienen los siguientes parámetros, y el correspondiente intercepto:

[[-0.01609793 0.00027143 0.02465072 -0.02042095]] [0.83270049]

¿Qué puede concluir a partir de estos valores?

**Respuesta:**

1. Su asistente menciona que usar la latitud y la longitud para cada llamada de servicio permitiría asignar cada llamada a un código postal. Al utilizar información del censo disponible públicamente, los datos por código postal podrían combinarse con información como la edad promedio, la raza predominante y el ingreso familiar promedio. Evalúe la recomendación de su asistente para detectar posibles inquietudes legales o éticas en este proyecto.

**Respuesta:**