Instituto Tecnológico y de Estudios Superiores Monterrey

Inteligencia Artificial para la Ciencia de Datos

Entrega 2: Diseño Metodológico y Preprocesamiento

Equipo 4
Esteban Sierra Baccio | A00836286
Diego Esparza Ruíz | A00837527
Javier Jorge Hernández Verduzco | A01722667
Sergio Omar Flores García | A01285193
Sergio Aarón Hernández Orta | A01613878

24 de agosto del 2025

Resumen Ejecutivo

Decisiones metodológicas clave

Algunas de las decisiones tomadas para la realización del presente trabajo ha sido la creación de features que nos permitan trabajar con más información recopilada para obtener mejores resultados dentro de los próximos análisis. Dentro de los cuáles se aprovecha el modelado para la imputación de datos faltantes o algún otro método de mitigación.

Features más importantes creadas

Durante el preprocesamiento de los datos fue necesaria la creación de variables que permitan leer la información y trabajar con ella de una forma más sencilla. Algunas de las features más importantes son *Family Size*, *Age Group* y *Title.*

Diseño Metodológico

Framework de validación de hipótesis

HIPÓTESIS #1: Los pasajeros con hijos menores de edad tuvieron mayor probabilidad de supervivencia que los pasajeros sin hijos menores de edad.

1. OPERACIONALIZACIÓN:
   - Variable dependiente: Supervivencia.
   - Variables independientes: Tiene hijos menores de edad.
   - Variables de control: El tamaño de la familia.

2. MÉTODO ESTADÍSTICO:
   - Test principal: Test de tendencia para múltiples categorías ordinales (1 hijo, 2 hijos, 3 hijos…)
   - Justificación: Este test nos permite visualizar cómo va variando la probabilidad de supervivencia a partir del tamaño de la familia entre padres e hijos
   - Supuestos a verificar: Verificar el número de hijos o padres y su posibilidad de supervivencia.
   - Tests alternativos: chi-cuadrado

3. CRITERIOS DE DECISIÓN:
   - Nivel de significancia: $\alpha = 0.05$
   - Tamaño del efecto mínimo: Tendría que haber un cambio de mínimo un 10% para poder decir que es relevante a la posibilidad de supervivencia.
   - Corrección por comparaciones múltiples: [Si aplica] No aplica

4. PLAN DE ANÁLISIS:
   - Análisis principal: Visualización con gráfico de barras y Tabla de contingencia 2x2.
   - Análisis de sensibilidad: Diferentes umbrales para definir a un menor (12, 16 y 18 años)
   - Análisis de subgrupos: Estratificación por clase social (1era, 2da, y 3ra)

HIPÓTESIS #2: La probabilidad de supervivencia en primera clase es al menos 50% mejor que la de tercera clase, para todas las edades y géneros.

1. OPERACIONALIZACIÓN:
   - Variable dependiente: Supervivencia
   - Variables independientes: Clase del pasajero, Género y Edad.
   - Variables de control: Género y Edad.

2. MÉTODO ESTADÍSTICO:
   - Test principal: Regresión Logística.
   - Justificación: Es el método más adecuado porque la variable dependiente (survived) es binaria (solo tiene dos resultados posibles: sobrevivir o no). La regresión logística nos permite modelar la probabilidad de este resultado en función de las variables independientes, controlando los efectos de la edad y el género simultáneamente
   - Supuestos a verificar: La relación entre las variables independientes numéricas (en este caso, la edad) y el logaritmo de las probabilidades es lineal.
   - Tests alternativos: Chi cuadrado

3. CRITERIOS DE DECISIÓN:
- Nivel de significancia: $\alpha = 0.05$
- Tamaño del efecto mínimo: Un odds ratio de al menos 1.5 para la variable PClass. Esto significa que la probabilidad de supervivencia en primera clase debe ser 50% mayor que en tercera clase, lo que corresponde directamente al enunciado de la hipótesis.
- Corrección por comparaciones múltiples: No aplica para el análisis.

4. PLAN DE ANÁLISIS:
- Análisis principal: Ajustar el modelo de regresión logística con la clase, género y edad como variables independientes.
- Análisis de sensibilidad: Probar diferentes métodos para manejar los valores faltantes en la edad, puede ser usar la mediana en vez de la media.
- Análisis de subgrupos: Realizar el mismo análisis de regresión logística de forma separada para los subgrupos de hombres y mujeres para verificar la hipótesis en cada género.

HIPÓTESIS #3: La posibilidad de supervivencia disminuye en cuestión de ser de tercera vinculados a varios familiares provenientes de Southampton

1. OPERACIONALIZACIÓN:
- Variable dependiente: **Supervivencia**
- Variables independientes: **Tamaño Familia, Origen Clase 3 y Embarcación Origen Southampton**
- Variables de control: **Sexo, Edad, Ticket, Tarifa y Cabina Conocidad**

2. MÉTODO ESTADÍSTICO:
- Test principal: **Chi-cuadrado sobre el dataframe de Categorías Familia y Supervivencia que se ubican dentro del subgrupo 1.**

- Justificación: **La razón de porque este método es apropiado debido a que se evalúa si el factor de "Tamaño Familia" vinculado a familias de pasajeros pertenecientes a la tercera clase y del puerto de Southampton es negativo para su supervivencia.**

- Supuestos a verificar:
  - **Suficientes observaciones en celdas de interacción:** Al crear varias combinaciones/interacciones con varias columnas, 3 columnas en este caso, fácilmente podemos tener pocas o ninguna observación en ciertas celdas del dataframe vinculado a este subgrupo causado por tres conceptos principales que vienen siendo **explosión combinatoria, valores de baja frecuencia y correlación entre variables.**

  - **Linealidad en el logit para tamaño familia (en el caso de usar la variable de forma numérica):** Para evitar que la estimación esté sesgada, nos aseguramos de que la relación entre Tamaño Familia y el log-odds (logit) de la variable de supervivencia sea lineal.

  - **Independencia y uso de errores cluster por ticket:** Se había encontrado un patrón en donde se vincula tres tickets diferentes con 6 o 7 pasajeros que pertenecían a la misma familia, provenían del puerto de Southampton y eran de tercera clase lo cual nos indica que se está violando la suposición de independencia.

- Tests alternativos:
  - **Si se viola el supuesto de linealidad en el logit para tamaño familia** entonces podemos transformar la variable de "tamaño familia", podemos añadir términos polinómicos para obtener la curvatura, podemos realizar modelos de regresión con splines y, por último, podemos utilizar modelos de árboles de decisión (Random Forest).

  - **Si se viola la independencia y uso de errores cluster por ticket** entonces podemos utilizar una regresión logística con errores estándar robustos o clusters para así ajustar los errores estándar mientras se considera la correlación de cada cluster, podemos emplear el uso de modelos de efectos mixtos o modelos jerárquicos lo cual nos ayuda a modelar la variabilidad de los datos, en este caso el subgrupo 1, a diferentes niveles y, también, podemos integrar modelos GEE (Generalized Estimating Equations) dándonos la oportunidad de modelar la media de la variable dependiente (Supervivencia) sin tener que especificar la distribución completa de los datos para esta hipótesis.

3. CRITERIOS DE DECISIÓN:
  - Nivel de significancia: $\alpha = 0.05$
  - Tamaño del efecto mínimo: **Para aceptar la hipótesis 3, se puede establecer un umbral por medio de una razón de Odds (OR) que sea menor a 0.67.**

(1 - 0.67 = 0.33) - Esto implica una reducción de los odds de 33%.

Lo anterior, nos ayuda a establecer que los odds de supervivencia para que los pasajeros cumplan con los criterios de la hipótesis 3 (pasajeros de tercera clase con familia y provenientes del puerto de Southampton) tienen que ser un mínimo de 33% menores en comparación a otros pasajeros que no cumplan con los criterios.

Adicionalmente, si nos vamos por el lado del concepto de la caída absoluta en la probabilidad, podemos establecer que la caída absoluta con respecto a la probabilidad de supervivencia/tasa de supervivencia sea de al menos 10 puntos porcentuales lo cual se traduce a que al menos la probabilidad de sobrevivir de los pasajeros pertenecientes al subgrupo 1 es menos 10 puntos porcentuales en comparación a los puntos porcentuales del grupo de pasajeros que no cumplan con los criterios del subgrupo 1.

  - Corrección por comparaciones múltiples: Si aplica

4. PLAN DE ANÁLISIS:
  - Análisis principal

Gracias al test principal seleccionado, el modelo que se realizará para evaluar la hipótesis 3 consiste en un modelo de regresión logística con triple interacción.

  - Análisis de sensibilidad

Se utilizaran modelos alternativos para verificar la robustez de los datos y se realizará una comparación con otros modelos.

- Análisis de subgrupos (si es relevante)

Se explorarán diferencias por sexo y por edad dentro del subgrupo 1 para evaluar si el efecto previamente mencionado es consistente.

Estrategia de modelado

La estrategia de modelado que adoptamos define el problema como una tarea de clasificación binaria donde la variable dependiente es la supervivencia de los pasajeros del Titanic. Buscamos desarrollar modelos predictivos que integren variables demográficas (edad, género), socioeconómicas (clase de pasajero, tarifa) y familiares (tamaño de familia, presencia de menores), considerando específicamente las interacciones complejas planteadas en las tres hipótesis de investigación. El abordaje metodológico contempla primordialmente el uso de regresión logística, justificado por la naturaleza binaria de la variable dependiente y la necesidad de interpretar efectos específicos como la triple interacción entre tamaño familiar, clase social y puerto de embarque (Hipótesis #3). Para gestionar el desbalance de clases inherente al dataset (~38% supervivientes vs ~62% no supervivientes), se consideran técnicas como la ponderación de clases inversamente proporcional a su frecuencia, SMOTE para oversampling de la clase minoritaria, y optimización de umbrales de clasificación, lo que permite obtener resultados más equilibrados y validar adecuadamente las hipótesis sobre subgrupos específicos como familias con menores de edad y pasajeros de tercera clase embarcados en Southampton.

Métricas y validación

Para evaluar el desempeño de los modelos, queremos establecer la exactitud como medida principal debido a que proporciona una visión general del rendimiento del modelo en la clasificación de ambas clases (supervivientes y no supervivientes), lo cual es fundamental para contrastar las hipótesis planteadas en diferentes grupos de pasajeros. Como medidas complementarias utilizaremos la sensibilidad para medir qué tan bien el modelo identifica a los supervivientes reales, la especificidad para evaluar su capacidad de reconocer correctamente a quienes no sobrevivieron, el F1-score que balancea la precisión y el recall especialmente útil en casos de desbalance, y el área bajo la curva ROC que permite analizar el comportamiento del modelo bajo distintos puntos de corte. El proceso de validación se estructura mediante la partición de los datos en tres conjuntos: 70% para entrenamiento, 15% para validación y 15% para pruebas finales, proporción que se considera adecuada dado el tamaño relativamente pequeño del dataset del Titanic. Adicionalmente, se aplicará validación cruzada estratificada con 5 divisiones para garantizar que cada subconjunto mantenga la misma proporción de casos de supervivencia que el dataset original, evitando así que los resultados dependan de una partición particular de los datos. Este enfoque es especialmente importante considerando que necesitamos evaluar el comportamiento del modelo en los subgrupos específicos definidos en nuestras hipótesis. Para mantener la integridad de los análisis, se verificarán los supuestos estadísticos como la relación lineal entre variables continuas y el logit, además de considerar la posible dependencia entre observaciones de una misma familia mediante técnicas de errores estándar robustos. Es importante reconocer que

los hallazgos estarán limitados al contexto específico del naufragio del Titanic, con las particularidades sociales y culturales de principios del siglo XX, lo que restringe su aplicabilidad a otros escenarios o épocas.

Ingeniería de Features

Resumen de variables creadas

Para abordar los sesgos identificados en la integridad de los datos (específicamente en la variable 'Cabin'), se aplicó ingeniería de variables con el objetivo de transformar las limitaciones del dataset en información modelable. En lugar de imputar valores, se crearon features que capturan explícitamente los patrones de datos faltantes:

- **Title**: El título de la persona
- **FamilySize**: El tamaño de la familia
- **IsAlone**: Si el tamaño de la familia es de 1
- **AgeGroup**: grupo de edad entre niño, adolecente, adulto y viejo.
- **FarePerPerson**: La tarifa por persona
- **CabinDeck**: CUBIERTA DE LA CABINA
- **CabinKnown**: Valor binario sobre si se conoce la cabina o no
- **TicketFrequency**: Cuenta la frecuencia del ticket
- **NameLength**: Longitud del nombre
- **HasCabinNeighbor**: Si tiene un vecino en la cabina
- TicketPrefix: El prefijo de ticket si es que existe

Justificación de las más importantes

- **Family Size:** Esta variable se encarga de obtener el tamaño de la familia. Para lograr esto es necesario conseguir a las personas que compartan el mismo apellido y compartan el mismo número de padres/hijos o hermanos en el buque. Este procesamiento nos permitirá trabajar con la primera hipótesis.
- **Age Group:** Esta variable permite agrupar por edad a las personas, delimitando la supervivencia por edad en grupos para facilitar su lectura y trabajo. Así mismo, podremos trabajar con la hipótesis #1
- **Cabin Known:** Esta variable permite trabajar directamente con los datos faltantes, específicamente los datos faltantes respecto a la cabina, ya que se ha encontrado una correlación que aumenta la posibilidad de supervivencia si es conocida la cabina del usuario. Esta variable permite que el modelo trabaje con una variable que anteriormente había parecido datos aleatorios faltantes.
- **Title:** El título de la persona permite conocer su rango. Esta variable separa el nombre que pasa desapercibido y permite agrupar a las personas por su rango volviéndose importante respecto a su supervivencia.
- **IsAlone:** Esta variable permite identificar a los pasajeros que viajan sin familia. es un indicador binario que facilita el análisis del efecto de viajar solo en la probabilidad de sobrevivir.
- **FarePerPerson**: esta variable permite ajustar la tarifa total del ticket por el número de personas en él. es un indicador más preciso del estatus socioeconómico de cada pasajero, lo cual influye en la probabilidad de supervivencia.
- **CabinDeck**: Esta variable permite conocer la cubierta de la cabina (a, b, c, etc.). se ha encontrado una correlación entre la cubierta y la posibilidad de supervivencia, ya que estaba relacionada con la cercanía a los botes salvavidas.
- **TickerFrequency**: Esta variable cuenta cuántas personas comparten el mismo número de ticket. es un indicador del tamaño del grupo de viaje que puede influir en la evacuación y, por lo tanto, en la supervivencia.

- **NameLenght**: Esta variable mide el número de caracteres en el nombre del pasajero. Se utiliza como un proxy para el estatus social, ya que los nombres más largos podrían estar asociados con un mayor estilo formal o posición de élite.
- **HasCabinNeighbor**: Esta variable fue creada para identificar a personas con cabina conocida que no viajaban solas. permite analizar si compartir la cabina con otros, además del tamaño de la familia, tuvo un impacto en la supervivencia.

Impacto esperado en el modelo

Esta estrategia resultará en un modelo que es:

- Más robusto: Menos propenso a aprender patrones espurios derivados de los sesgos en los datos.
- Más Interpretable: Las razones detrás de las predicciones (e.g., "no sobrevivió por ser de tercera clase sin cabina registrada") serán claras y se alinearán con el contexto histórico.
- Más Justo: Al reconocer y modelar explícitamente los grupos subrepresentados (e.g., pasajeros de tercera clase sin datos), las predicciones para estos grupos serán más conscientes y menos dependientes de suposiciones erróneas.

Tratamiento de Datos Faltantes

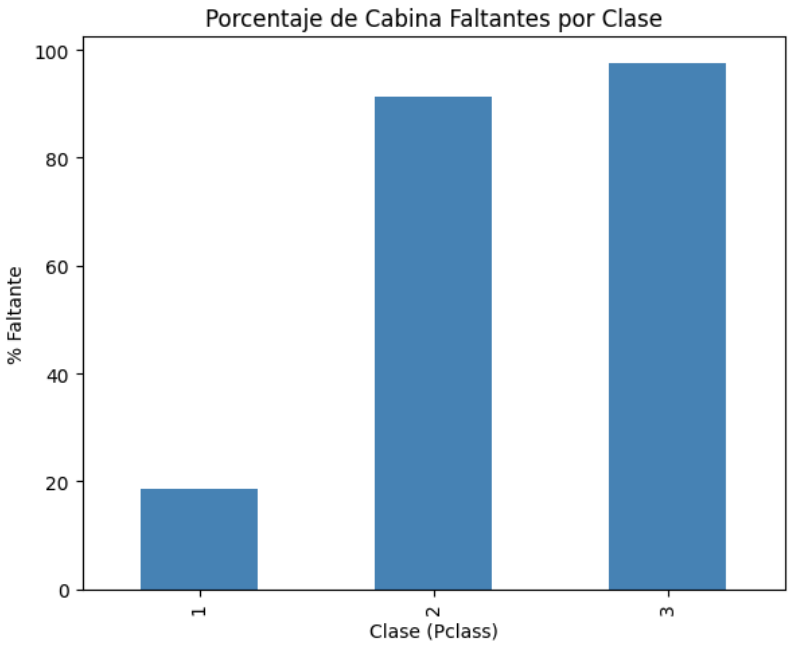
Estrategia general adoptada

Hay solamente 3 columnas con valores faltantes en el dataset:

1. Edad (`Age`)
2. Cabina (`Cabin`)
3. Puerto de embarcación (`Cabin`)

La estrategia general que se adoptó fue la del **método simple.** Otros métodos no presentaron mejoras en los modelos y agregan incertidumbre.

De las tres, la cabina es la columna con más valores vacíos, siendo casi todas las personas de tercera y segunda clase.



En este caso, los datos faltantes no se pueden reemplazar o imputar. Si se necesita la columna de Cabin, se tienen que eliminar las filas con el valor vacío, pues no son valores en una escala que se pueda cuantificar.

Para la edad, se decidió que se tomará la mediana de los datos en la muestra para imputar los datos faltantes, al no encontrarse mejora en los modelos con otros métodos.

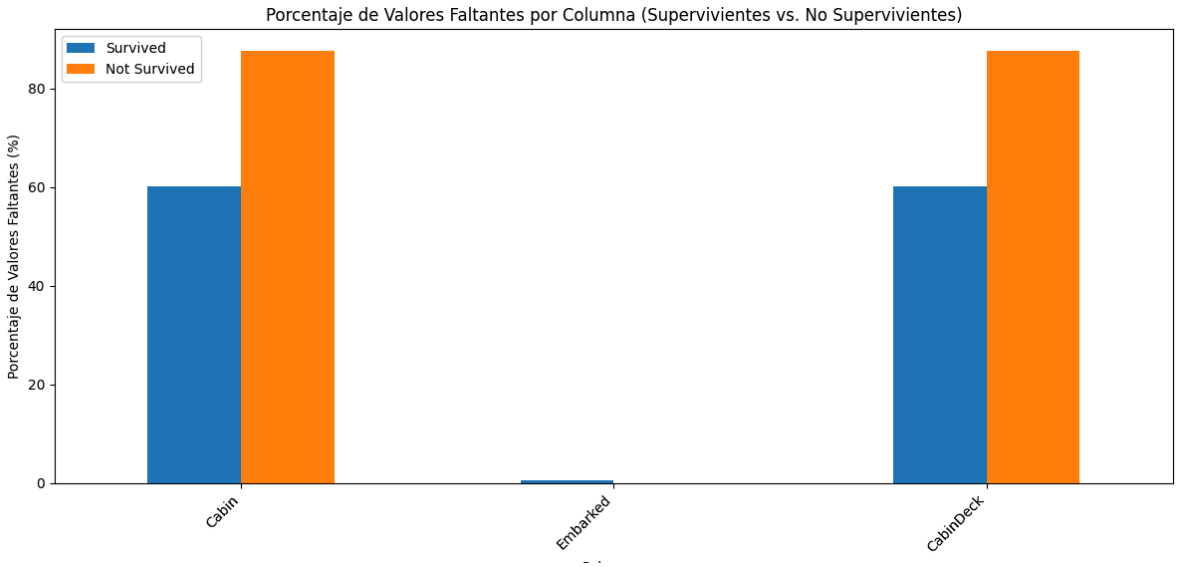
Finalmente, solo hay dos datos faltantes para la columna del puerto de Embarcación; dos mujeres sin relación, que comparten la misma cabina en primera clase. La moda en cada clase y a lo largo de la muestra es Southampton, por bastante, por lo que es el valor indicado sin duda para reemplazar esos dos valores vacíos.

Nuestra conclusión después de probar los múltiples métodos lamentablemente es que este dataset no se presta a métodos muy complejos, se tiene poca información e incluso hay variables que no se pueden cuantificar, por lo que

Consideraciones de Sesgo y Ética

Sesgo de supervivencia

Sistemáticamente hay información faltante respecto a los no supervivientes, esta información aumenta en la cabina de un 60% de los supervivientes a un 85% en los no supervivientes.



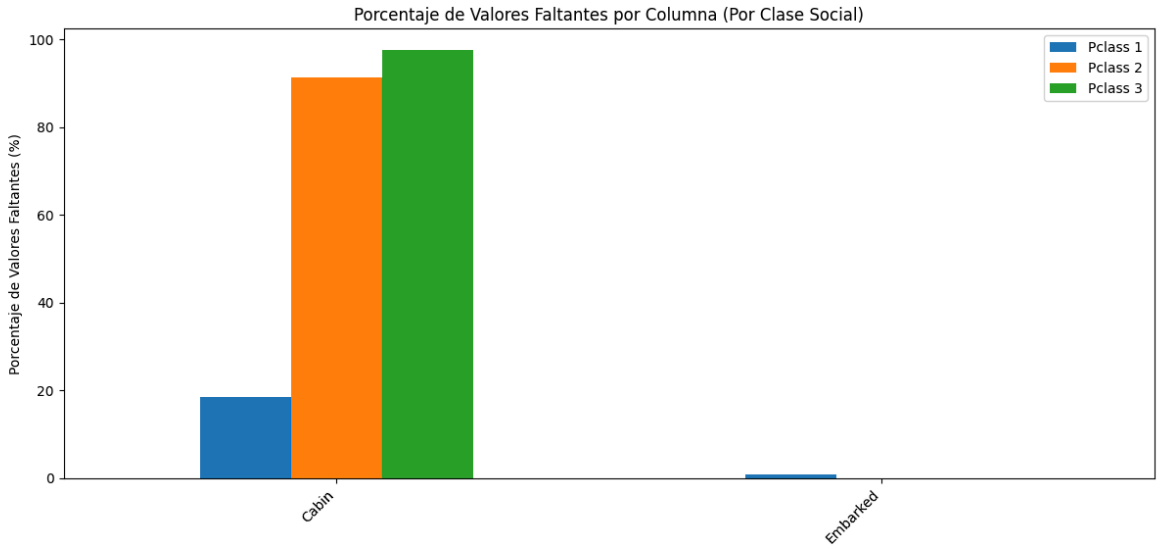
Estrategia de mitigación: Análisis de sensibilidad.

Implementación
Tasa de supervivencia original: 0.3838383838383838
Tasa de supervivencia con imputación pesimista: 0.1526374859708193
Tasa de supervivencia con imputación optimista: 0.9236812570145904

Documentación de limitaciones residuales:
Este análisis determina que los datos faltantes en Cabin NO son aleatorios.

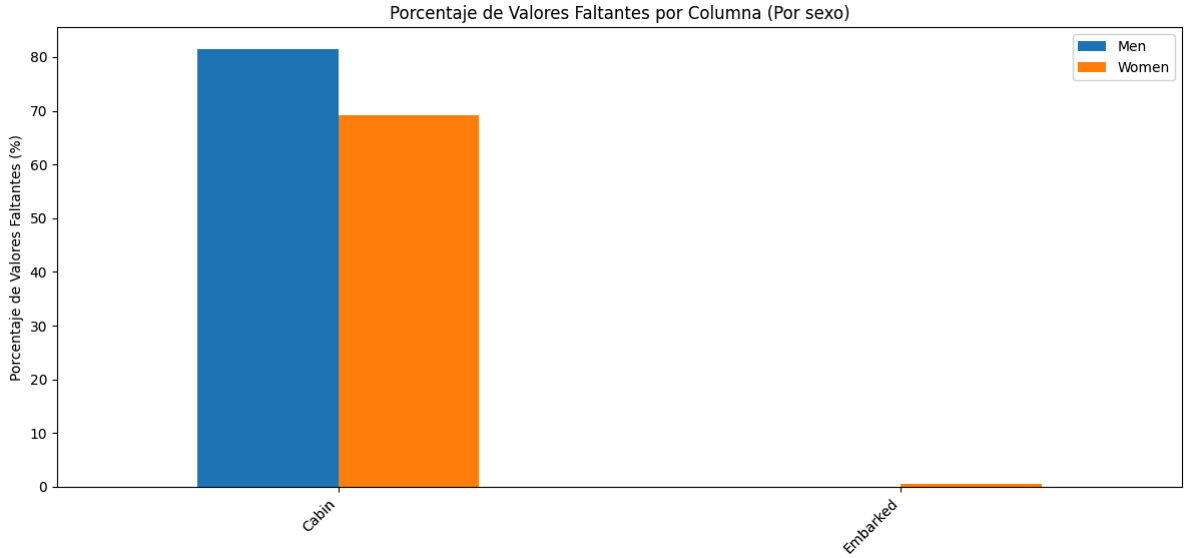
Sesgo de clase social

Sistemáticamente, hay más información faltante respecto a las personas de segunda y tercera clase en la variable de cabina. Pasando de un 20% en primera clase, a 90% en segunda y 95% en tercera clase.



Sesgo de género

Podemos ver que si existe un sesgo de género, pasando del 80% de datos faltantes en la columna “Cabin” en hombres a un 69% en mujeres. Sin embargo, esta información puede estar altamente influenciada por el sesgo de supervivencia, donde la mayoría de las mujeres sobrevivieron al accidente.



Sesgo histórico

Algunos de los grupos se encuentran representados en las secciones anteriores y ya han sido explicadas anteriormente. La clase social, el género y el puerto de embarque son cruciales para determinar la probabilidad de que una persona sobreviva al accidente o no.

Conclusiones y Próximos Pasos

¿Cómo afectan sus decisiones de preprocesamiento a la interpretabilidad del modelo final?
Nuestras decisiones de preprocesamiento mejoran radicalmente la interpretabilidad del modelo al transformar datos crudos y sesgados en características (features) con significado contextual claro.

¿Qué trade-offs enfrentaron entre completitud de datos y calidad?
El principal trade-off fue priorizar la calidad y la honestidad de los datos sobre la completitud artificial.

¿Cómo podrían sus decisiones introducir o amplificar sesgos existentes?
Nuestras decisiones están diseñadas para mitigar sesgos, pero un uso incorrecto podría amplificarlos a través del reforzamiento de estructuras sociales como el valor “title”. O a través de la máscara de sesgo no mitigado que no mejora la información y nos permite pasarla por alto.

¿Qué información adicional del Titanic les hubiera sido más útil?
La información más valiosa sería aquella que ayudará a llenar los vacíos que nuestros features proxy intentan capturar como la ubicación exacta de la cabina. Registros de rescate, ocupación, manifiesto de apellidos y relaciones.

¿Cuáles son los próximos pasos?
Los próximos pasos son crear un modelo de machine learning con este pipeline de procesamiento de datos de tal forma que el modelo pueda predecir la probabilidad de supervivencia de un usuario del titanic de la mejor forma posible.