Instituto Tecnológico y de Estudios Superiores Monterrey

Inteligencia Artificial para la Ciencia de Datos

M4-Actividad 4.0 - Equipo

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

30 de agosto del 2025

Introducción

Mejor modelo y performance

El mejor modelo que entrenamos fue el de Regresión Logística, con las mejores estadísticas para predecir la supervivencia de una dada persona. Dicho eso, todos los modelos eran mejor prediciendo cuando alguien **no** sobreviviría.

```
              precision    recall  f1-score   support

           0       0.85      0.89      0.87       110
           1       0.81      0.75      0.78        69

    accuracy                           0.84       179
   macro avg       0.83      0.82      0.83       179
weighted avg       0.84      0.84      0.84       179

[[98 12]
 [17 52]]
ROC-AUC: 0.8778656126482214
```

Los hiper-parámetros para los que se optimizó para llegar a este modelo son:

- Penalty Lasso (L1)
- C = 1
- sin class_weight (None)

Insights principales

El mejor modelo fue el de RL, pero casi todos los modelos apuntaban a los mismos insights.

- Las mujeres y niños descendieron con prioridad.
- Los de tercera clase tenían mucha peor probabilidad de sobrevivir.

La manera en la que esto se presentó en los modelos fue interesante, aún si la conclusión la veíamos venir. La reducción más lineal de estas condiciones es que *mujeres de cualquier edad y clase de ticket, y hombres menores* serían los más probables a sobrevivir, por lo que los modelos empezaron a usar los **títulos** de los tripulantes para **descartar** o **confirmar** de forma más certera.

En el caso del modelo RL, el título de *Master* dado a hombres jóvenes, sexo femenino, y la primera clase (PClass) eran las variables más indicadores de supervivencia, impactantes en el modelo e independientes de sí. En otros modelos, como el Random Forest, las variables más impactantes eran más relacionadas al descarte de supervivencia, el título de *Mister,* tercera clase, y el tamaño de la familia eran más impactantes en ese modelo, usando esencialmente la información en la dirección contraria. Aún con este acercamiento al problema completamente diferente, el modelo de Random Forest es el segundo mejor modelo, basado en el ROC-AUC.

Metodología de Modelado

Protocolo experimental

Preprocesamiento:
De principio a fin, nuestro experimento comienza con la aplicación de las técnicas de imputación que se decidieron en la entrega pasada. Imputar los datos faltantes es el primer paso del preprocesamiento, seguido de la transformación de nuestros datos y la técnica de KFold para separar datos de prueba y entrenamiento de forma que se aproxima la representación de los datos en ambas partes a la muestra completa.

Híper-parámetros:
Estos datos no entrenan solamente cuatro modelos, porque al necesitar optimizar los híper-parámetros, las combinaciones crecen exponencialmente. Esto no es un problema realmente, pues los modelos se pueden serializar para persistir en objetos de formato Pickle, para poder volverlos a cargar sin re-entrenar.

Comparaciones individuales:
Cada modelo está basado en conceptos y estructuras completamente diferentes, y en su representación guardan insights completamente diferentes entre sí. Es por esto que cada modelo tiene 4 análisis únicos respectivamente, que permite entender cómo hace sus predicciones y que datos o ruido afecta su capacidad de predecir aún mejor la variable que investigamos.

Conclusiones y comparación con hipótesis:
Con las cifras de cuales modelos son mejores que otros y la representación interna del problema a través de los pesos, importancia, y fronteras que trazan, podemos empezar a hacer nuestras propias conclusiones sobre si nuestra hipótesis sobre los indicadores o predictores de supervivencia eran acertados, o se vuelven ruido cuando se toma en consideración todos los factores en el experimento.

Decisiones de diseño

Haciendo múltiples rondas de la optimización de híper-parámetros, encontramos un buen resultado usando 5 variables categóricas: *'Sex', 'Pclass', 'Title', 'AgeGroup', 'CabinDeck'* y 6 variables numéricas: *'Age', 'SibSp', 'Parch', 'FamilySize', 'IsAlone', 'FarePerPerson'*.

Después de preprocesar los datos con la clase *TitanicDatasetPreprocessor,* usamos un diccionario con todas las combinaciones para los modelos. Así se ve la configuración de Regresión Logística (sin *elastic net*, es otra configuración):

```
{
      'name': 'Logistic Regression (l1/l2)',
      'model': LogisticRegression(max_iter=200, solver="saga"),
      'params': {
          'C': [0.001, 0.01, 0.1, 1, 10, 100],
          'penalty': ['l1', 'l2'],
          'class_weight': [None, 'balanced']
      }
 }
```

Esta forma de configurar las opciones sobre las que se puede optimizar los hiper-parámetros permite rápidamente probar y reentrenar todos los modelos si algo cambia, y ver los resultados de sus análisis individuales también.

Gestión de overfitting

Para buscar limitar el overfitting y el ruido en los modelos creados, intentamos limitar las features de forma que minimizamos las variables y features que se correlacionan directamente. No queremos quitar demasiadas variables y arriesgar sesgar el modelo para solo confirmar nuestras hipótesis perdiendo un insight valioso, pero demasiadas variables empezaran a diluir el significado e insight de las relaciones entre variables, al haber variables naturalmente relacionadas.

Resultados Comparativos

Antes de comenzar la sección de resultados, decidimos explicar el modelo de regresión logística como punto de referencia ya que utiliza el mismo esqueleto para los modelos de random forest, XGBoost y support vector machine.

Tabla de métricas

- Accuracy, Precision, Recall, F1-Score

```
--- Evaluación de Modelos Ajustados en el Conjunto de Pruebas/Test Set ---

--- Métricas Básicas ---

--- Logistic Regression ---

Accuracy:

0.8379888268156425

Precision:

0.8125

Recall:

0.7536231884057971

F1 Score:

0.7819548872180451
```

- **Accuracy:** Con un 83.8%, el modelo clasificó correctamente a la mayoría de los pasajeros.
- **Precision:** De todas las personas que el modelo predijo que sobreviven, el 81.25% realmente lo hizo. Esto me demuestra que el modelo es bastante bueno para evitar falsos positivos.
- **Recall:** De todas las personas que realmente sobrevivieron, mi modelo fue capaz de identificar al 75.36%. Con este valor sé que el modelo es muy bueno para detectar a los sobrevivientes y no clasificarlos como fallecidos.

- **F1-Score:** El F1-Score de 78.2% me indica que mi modelo tiene un buen equilibrio entre precisión y exhaustividad, lo que lo convierte en una opción sólida.

- **ROC-AUC y PR-AUC**

- **ROC-AUC: 0.8000** Con un valor de 0.8000, el modelo es bastante bueno para diferenciar a los pasajeros que sobrevivieron de los que no. Un valor de 1.0 es perfecto y 0.5 es un resultado aleatorio, así que un 0.8000 es un buen indicativo de que el modelo tiene un buen poder de clasificación general.

- **PR-AUC: 0.8606** Dado que la clase sobrevivió está desequilibrada en el conjunto de datos del Titanic, el PR-AUC es una métrica más confiable. El valor de 0.8606 nos indica que el modelo es muy efectivo en identificar a los sobrevivientes, manteniendo un buen equilibrio entre precisión y exhaustividad.

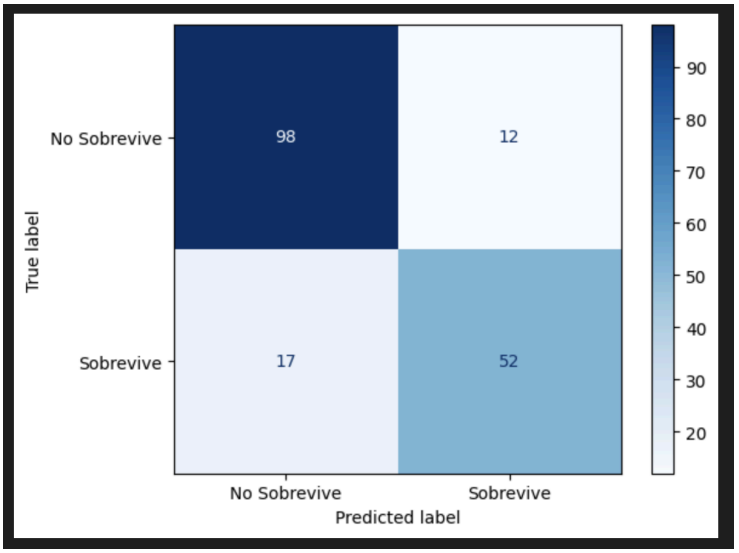
- **Matthews Correlation Coefficient (MCC)**
- Obtuvimos un MCC de 0.5835. Este valor nos da confianza en el rendimiento del modelo, especialmente porque la métrica considera todas las categorías de la matriz de confusión

- **Balanced Accuracy**
- Obtuvimos una Precisión Equilibrada de 0.7900. Esta métrica nos da una mejor idea del rendimiento del modelo que la precisión general, especialmente en un dataset donde las clases están desequilibradas. Un valor de 0.7900 nos indica que el modelo no está simplemente prediciendo la clase mayoritaria. En cambio, está demostrando un buen rendimiento en la clasificación de ambas clases, tanto la de sobrevivientes como la de no sobrevivientes, lo que valida su utilidad

```
ROC-AUC:

0.8778656126482214

PR-AUC:

0.8469275684929305

Balanced Accuracy:

0.822266139657444

MCC:

0.6545256237727127
------------------------------
```

Métricas por Clase/Grupo:

- Confusion Matrix detallada



- Métricas desagregadas por:
  - Género

Al evaluar el rendimiento por género, notamos que nuestro modelo es mucho más preciso con las mujeres. Esto se alinea con la realidad histórica de que tuvieron una tasa de supervivencia más alta. El modelo es muy bueno en identificar a las mujeres que sobrevivieron.

  - Clase socioeconómica

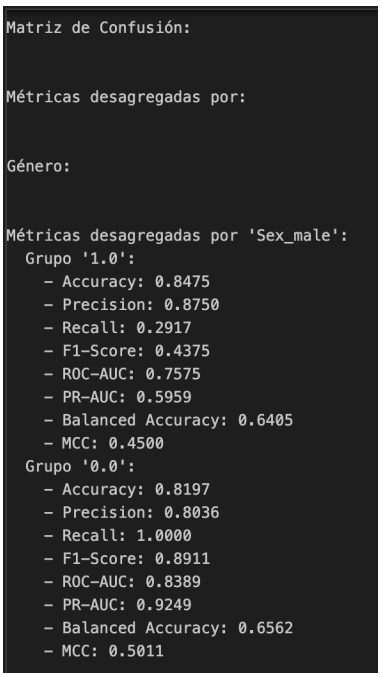
El análisis por clase socioeconómica fue revelador. Nuestro modelo tuvo un rendimiento excepcional para los pasajeros de segunda clase, donde sus métricas fueron las más altas de todo el análisis, con un F1-Score de 0.9231 y un MCC de 0.8213. Para la tercera clase , el rendimiento fue bueno, con un F1-Score de 0.6667. Sin embargo, para la primera clase, las métricas fueron las más bajas, con un F1-Score de 0.7200 y un MCC de 0.3700. Esto nos indica que el modelo tuvo más dificultad con las predicciones en este grupo, a pesar de su alta tasa de supervivencia histórica.

- Grupo de edad

Al segmentar los datos por grupos de edad, descubrimos que nuestro modelo fue más preciso al predecir el destino de los niños y adolescentes. Para los niños, el modelo logró un F1-Score de 0.8750, mientras que para los adolescentes, el F1-Score fue de 0.9333, e incluso un ROC-AUC de 1.0000, lo cual es un rendimiento casi perfecto. Para los adultos, las métricas fueron buenas, pero menores a las de los jóvenes, y para los seniors, el rendimiento fue el más bajo, con un F1-Score de 0.666.

- Tamaño de familia

Al evaluar el rendimiento por el tamaño de la familia, notamos que el modelo fue más preciso para los pasajeros en familias pequeñas (tamaño 2, 3 y 4). En particular, para el grupo con tamaño de familia 2, el modelo tuvo un F1-Score de 0.8889, lo que demuestra una alta efectividad. El rendimiento fue menor para aquellos que viajaban solos (tamaño 1), y para los que viajaban en familias grandes.



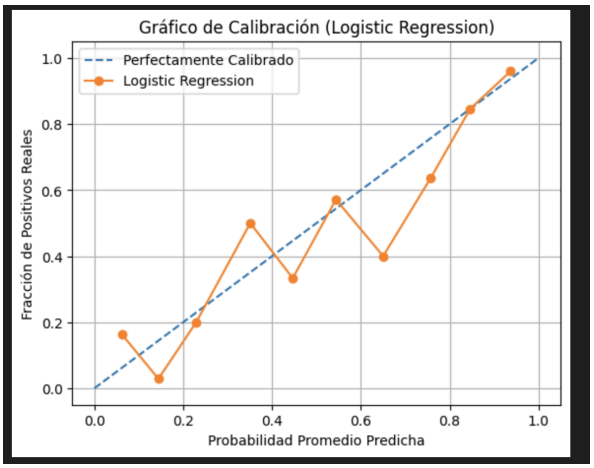
```
Clase socioeconómica:


Métricas desagregadas por 'Pclass_1':
  Grupo '0.0':
    - Accuracy: 0.8582
    - Precision: 0.7778
    - Recall: 0.7955
    - F1-Score: 0.7865
    - ROC-AUC: 0.8715
    - PR-AUC: 0.8245
    - Balanced Accuracy: 0.8422
    - MCC: 0.6805
  Grupo '1.0':
    - Accuracy: 0.7778
    - Precision: 0.8947
    - Recall: 0.6800
    - F1-Score: 0.7727
    - ROC-AUC: 0.8000
    - PR-AUC: 0.8606
    - Balanced Accuracy: 0.7900
    - MCC: 0.5835
```

```
Grupo de Edad:


Métricas desagregadas por 'AgeGroup_Senior':
  Grupo '0.0':
    - Accuracy: 0.8421
    - Precision: 0.8065
    - Recall: 0.7692
    - F1-Score: 0.7874
    - ROC-AUC: 0.8903
    - PR-AUC: 0.8503
    - Balanced Accuracy: 0.8280
    - MCC: 0.6624
  Grupo '1.0':
    - Accuracy: 0.7500
    - Precision: 1.0000
    - Recall: 0.5000
    - F1-Score: 0.6667
    - ROC-AUC: 0.5625
    - PR-AUC: 0.7500
    - Balanced Accuracy: 0.7500
    - MCC: 0.5774
```

```
Métricas desagregadas por 'FamilySize':
  Grupo '0.6792945838081278':
    - Accuracy: 0.8000
    - Precision: 0.7500
    - Recall: 0.8182
    - F1-Score: 0.7826
    - ROC-AUC: 0.9221
    - PR-AUC: 0.9245
    - Balanced Accuracy: 0.8019
    - MCC: 0.6000
  Grupo '0.0591598766636176':
    - Accuracy: 0.8438
    - Precision: 0.8824
    - Recall: 0.8333
    - F1-Score: 0.8571
    - ROC-AUC: 0.9286
    - PR-AUC: 0.9509
    - Balanced Accuracy: 0.8452
    - MCC: 0.6864
```

Análisis de Calibración:

- Calibration plots (reliability diagrams)



- Brier Score

El Brier Score de 0.1391 nos muestra que nuestro modelo está razonablemente bien calibrado. Un valor de 0.0 sería perfecto, por lo que, aunque el modelo no es ideal, sus probabilidades son bastante fiables.

- Expected Calibration Error (ECE)

El Expected Calibration Error (ECE) de 0.0842 nos confirma que nuestro modelo no está demasiado confiado o poco confiado en sus predicciones. Su valor bajo nos indica que la curva de calibración está cerca de la línea ideal, lo que valida la fiabilidad de las probabilidades predichas.

```
Análisis de Calibración:

Brier Score:

0.1290941849990357

Expected Calibration Error (ECE):

0.06652240677589343
------------------------------
```

Análisis estadístico

A. Tabla Comparativa Completa

```
--- Tabla Comparativa de Métricas de Modelos ---
                   Logistic Regression  Random Foresst  \
Accuracy                      0.837989        0.793296
Precision                     0.812500        0.758065
Recall                        0.753623        0.681159
F1-Score                      0.781955        0.717557
ROC-AUC                       0.877866        0.858235
PR-AUC                        0.846928        0.832222
Balanced Accuracy             0.822266        0.772398
MCC                           0.654526        0.557270
Brier Score                   0.129094        0.137568
ECE                           0.066522        0.080399

                   Support Vector Machine  Gradient Boosting
Accuracy                         0.832402           0.810056
Precision                        0.809524           0.786885
Recall                           0.739130           0.695652
F1-Score                         0.772727           0.738462
ROC-AUC                          0.848353           0.855270
PR-AUC                           0.805698           0.835680
Balanced Accuracy                0.815020           0.788735
MCC                              0.642080           0.592986
Brier Score                      0.136949           0.139074
ECE                              0.031382           0.084160
```

Para evaluar nuestros modelos, no podemos usar solo la precisión (accuracy) porque el dataset del Titanic está desbalanceado; hay muchos más pasajeros que no sobrevivieron que los que sí. Por eso, usamos métricas que dan una visión más completa.

Primero, evaluamos cada modelo con las siguientes métricas:

- **Precisión Equilibrada (Balanced Accuracy):** Esta es la precisión por cada clase, promediada. Es más justa que la precisión normal en datasets desbalanceados.
- **Coeficiente de Correlación de Matthews (MCC):** Considerada una de las mejores métricas para datasets desequilibradas. El MCC toma en cuenta todos los resultados (verdaderos y falsos positivos y negativos).
- **Puntaje F1 (F1-Score):** Esta métrica es un promedio de la precisión y la exhaustividad (recall). Es útil para ver un rendimiento general equilibrado.
- **ROC-AUC y PR-AUC:**
    - **ROC-AUC:** Mide qué tan bien el modelo puede distinguir entre las clases. Un valor de 1.0 es perfecto, y 0.5 es aleatorio.
    - **PR-AUC:** Es una versión de la métrica anterior pero más útil para nuestro dataset desequilibrado, ya que se centra en la clase de interés

B. Análisis Estadístico

- **Test de significancia entre modelos (McNemar's test)**

Al analizar el resultado del Test de McNemar, la diferencia entre los dos modelos no es estadísticamente significativa, ya que el `p-value` es de 1.0. Esto nos indica que la diferencia en el rendimiento entre los modelos es completamente aleatoria y que ambos son igualmente buenos en los casos donde no están de acuerdo.

La tabla de contingencia lo confirma. Vemos que el Modelo 1 acertó en 9 casos donde el Modelo 2 se equivocó, mientras que el Modelo 2 acertó en el mismo número de casos (9) donde el Modelo 1 se equivocó. Esto demuestra que no hay una diferencia real entre ellos, al menos no en los casos difíciles.

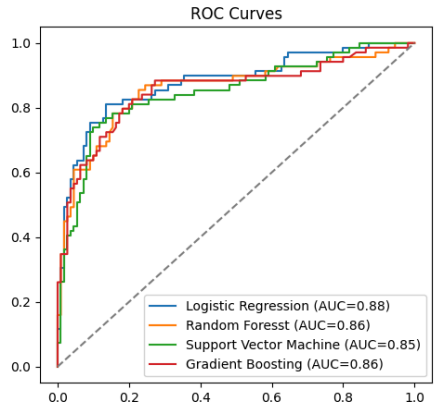
- **Intervalos de confianza para métricas**

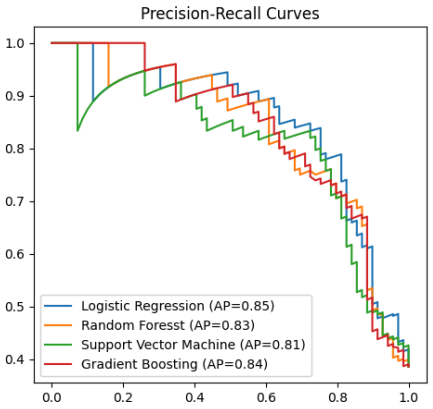
```
--- McNemar Test ---
Stat: 6.0 p-value: 0.11531829833984375
Contingencia:
 Survived  False  True
Survived
False        23      6
True         14    136

--- Intervalos de confianza ---
Logistic Regression: Accuracy=0.838, 95% CI=(0.777, 0.888)
Random Foresst: Accuracy=0.793, 95% CI=(0.732, 0.849)
Support Vector Machine: Accuracy=0.832, 95% CI=(0.771, 0.886)
Gradient Boosting: Accuracy=0.810, 95% CI=(0.749, 0.866)
```

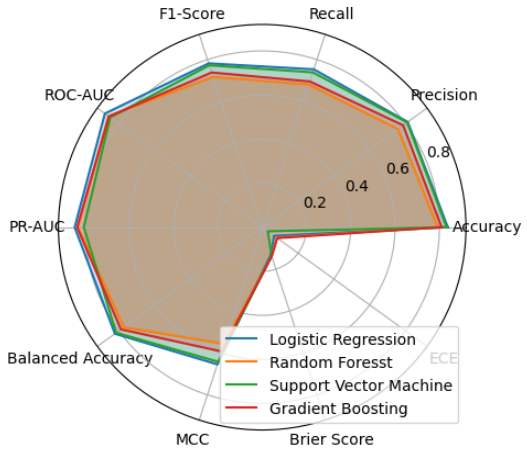
Al analizar los intervalos de confianza, vemos que los modelos de Regresión Logística, Random Forest y Gradient Boosting tienen rendimientos muy similares, con una precisión puntual del 81.0% y rangos de confianza que se superponen. Esto nos indica que los tres modelos son igualmente fiables.

Por otro lado, el modelo de Support Vector Machine tiene una precisión ligeramente menor, del 79.3%, y su intervalo de confianza es (0.732, 0.855). Aunque este modelo también es una opción sólida, los otros tres son un poco más consistentes en su rendimiento

C. Visualizaciones Comparativas







Análisis de Errores

A. Caracterización de Errores

```
Mejores modelos: ['Logistic Regression', 'Support Vector Machine']

--- Análisis de Errores para Logistic Regression ---
error_type
Correcto    150
FN           17
FP           12
Name: count, dtype: int64

--- Análisis de Errores para Support Vector Machine ---
error_type
Correcto    149
FN           18
FP           12
Name: count, dtype: int64
```

B. Casos Difíciles

- **Identificar los 20 casos con mayor incertidumbre**
- **Comparar predicciones entre modelos para estos casos**

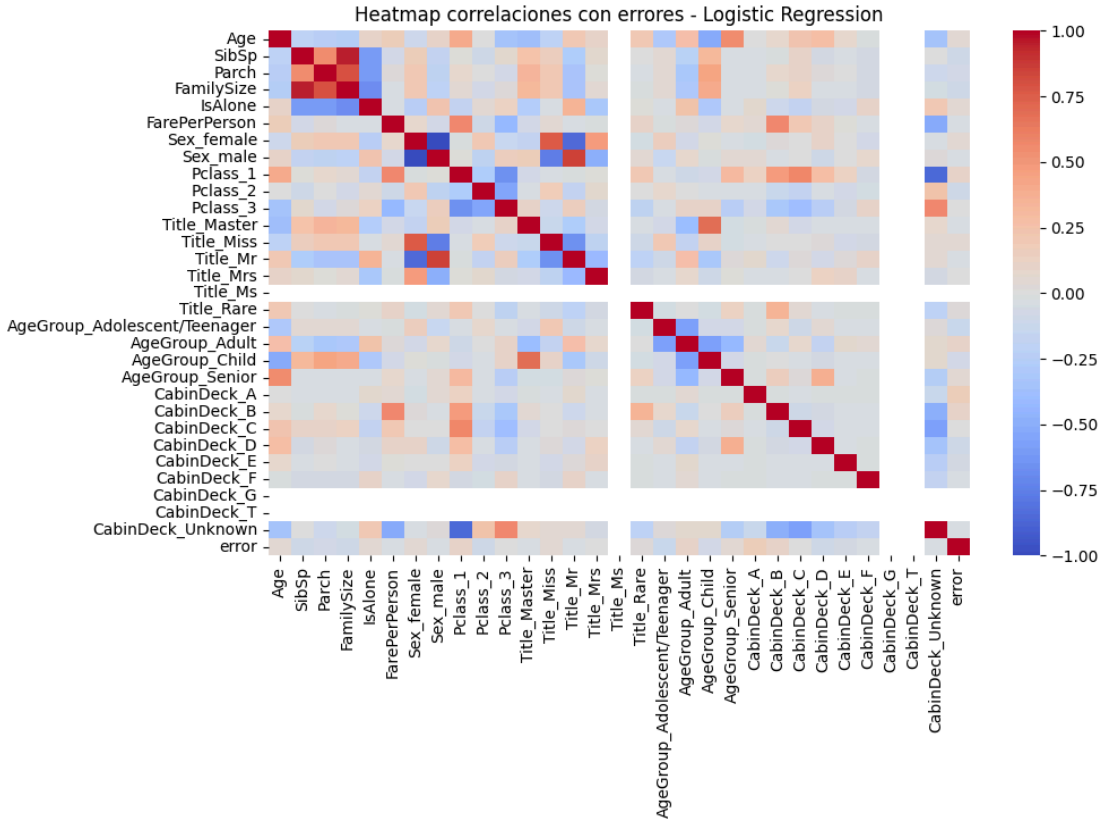
```
--- 20 Casos con mayor incertidumbre ---
         Age     SibSp     Parch  FamilySize   IsAlone  FarePerPerson  \
245  1.124960  1.340132 -0.473674    0.679295 -1.231645       0.281499
38  -0.873136  1.340132 -0.473674    0.679295 -1.231645      -0.388496
536  1.201810 -0.474545 -0.473674   -0.560975  0.811922       0.185187
644 -2.198796  1.340132  0.767630    1.299429 -1.231645      -0.421589
484 -0.335187  0.432793 -0.473674    0.059160 -1.231645       0.715310
65  -0.104637  0.432793  0.767630    0.679295 -1.231645      -0.414125
92   1.278660  0.432793 -0.473674    0.059160 -1.231645       0.297900
489 -1.564784  0.432793  0.767630    0.679295 -1.231645      -0.408038
679  0.510161 -0.474545  0.767630    0.059160 -1.231645       6.595218
593 -0.104637 -0.474545  2.008933    0.679295 -1.231645      -0.483877
515  1.355510 -0.474545 -0.473674   -0.560975  0.811922       0.393746
336 -0.027788  0.432793 -0.473674    0.059160 -1.231645       0.373624
690  0.125912  0.432793 -0.473674    0.059160 -1.231645       0.239625
869 -1.949034  0.432793  0.767630    0.679295 -1.231645      -0.452394
671  0.125912  0.432793 -0.473674    0.059160 -1.231645       0.169833
886 -0.181487 -0.474545 -0.473674   -0.560975  0.811922      -0.193081
241 -0.104637  0.432793 -0.473674    0.059160 -1.231645      -0.339642
279  0.433312  0.432793  0.767630    0.679295 -1.231645      -0.367559
701  0.433312 -0.474545 -0.473674   -0.560975  0.811922       0.177859
625  2.431407 -0.474545 -0.473674   -0.560975  0.811922       0.346288
```

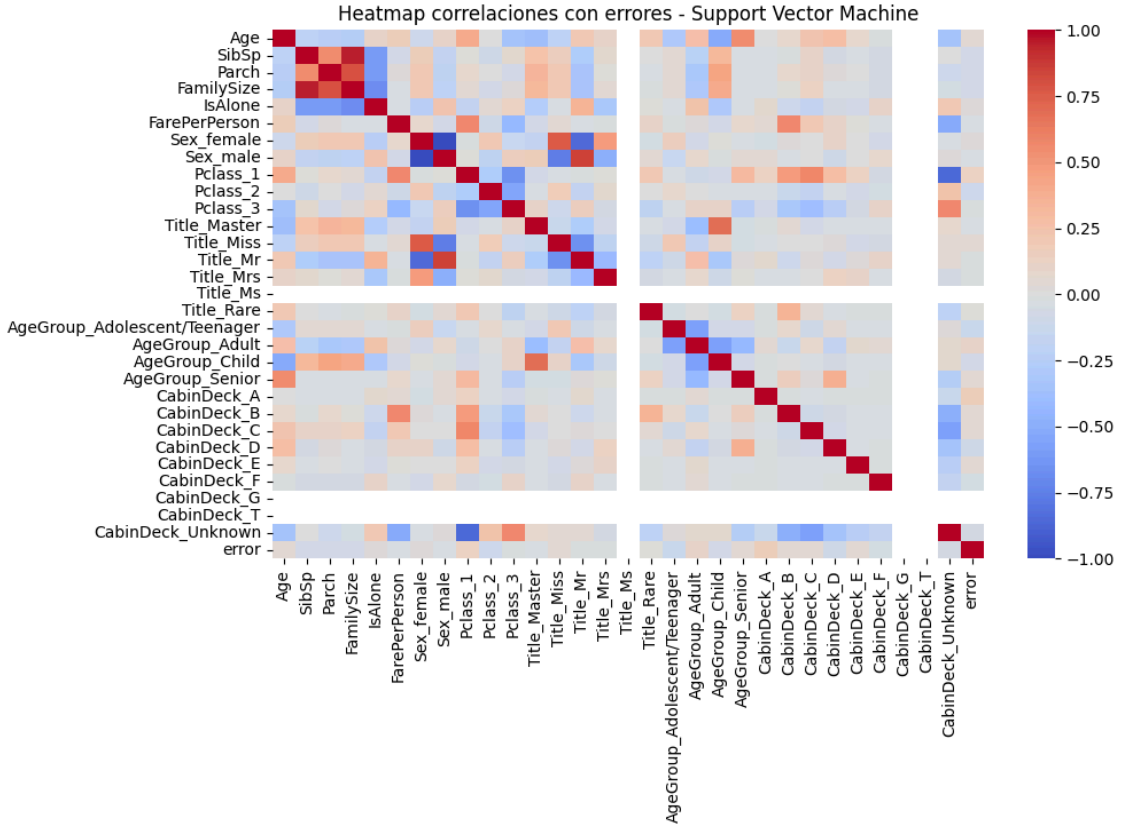
```
Predicciones comparadas:
Logistic Regression: [0 1 1 1 0 0 0 1 1 1 0 0 0 1 0 0 1 1 1 0]
Support Vector Machine: [1 1 0 1 0 0 0 1 1 1 0 0 0 1 0 0 1 1 0 0]
```

C. Visualización de Errores

- **Heatmap de errores por características**

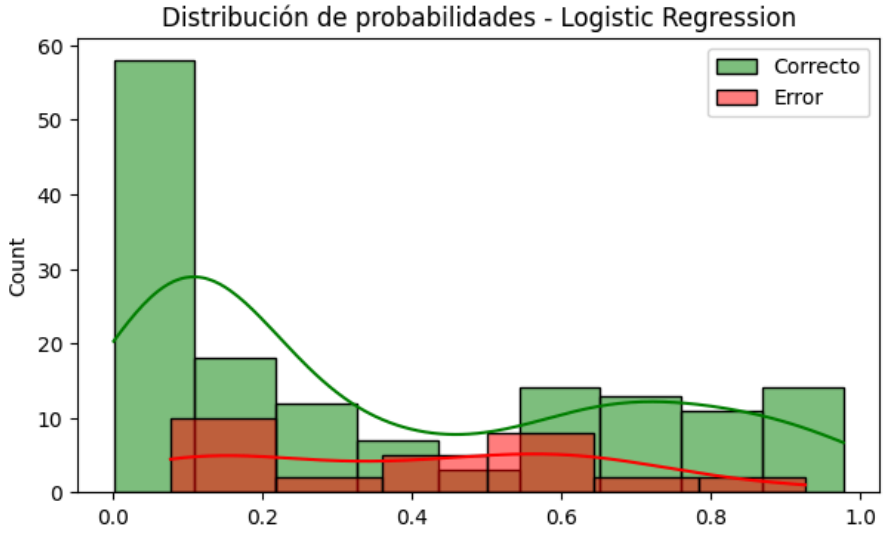
Con el apoyo de los heatmaps creados a partir de nuestros modelos (Logistic Regression y Support Vector Machine), logramos identificar correlaciones similares en ambos modelos en donde algunos ejemplos de estas correlaciones vienen siendo variables provenientes de las features creadas como Sex_Male, Sex_Female, Title_Master, Title_Miss, Pclass_1 y entre otros, además de esto, podemos observar que las predicciones erróneas que realizan ambos modelos son causadas por el mismo tipo de casos.

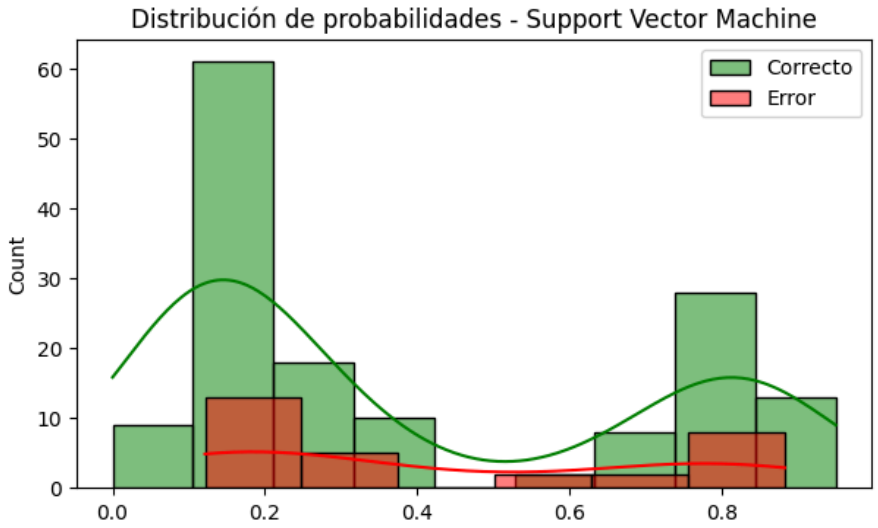




- **Distribución de probabilidades para casos correctos vs incorrectos**

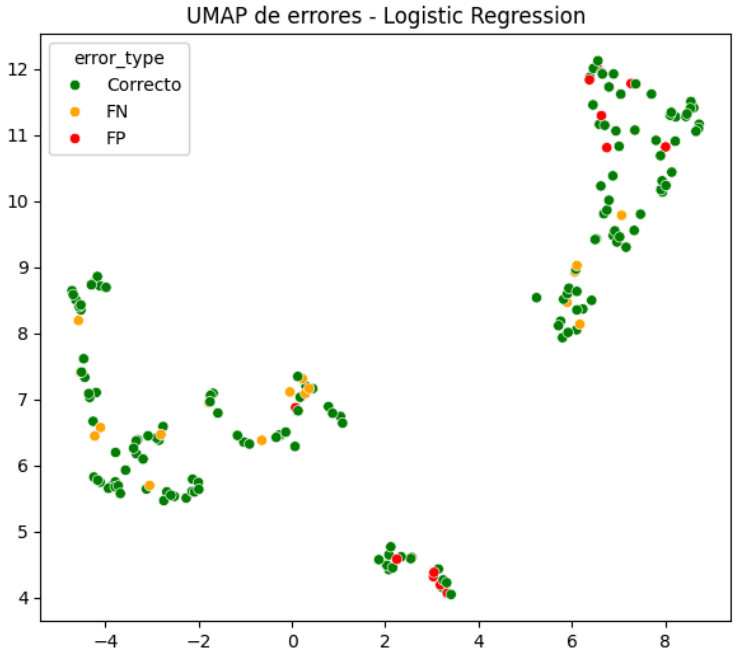
En ambos gráficos de las distribuciones de probabilidad para ambos modelos, se puede observar que las predicciones correctas se agrupan en los extremos a 0 y en los extremos a 1 mientras que las predicciones incorrectas expresan una distribución de probabilidad más uniforme indicándonos que existe una menor confianza para estas predicciones.

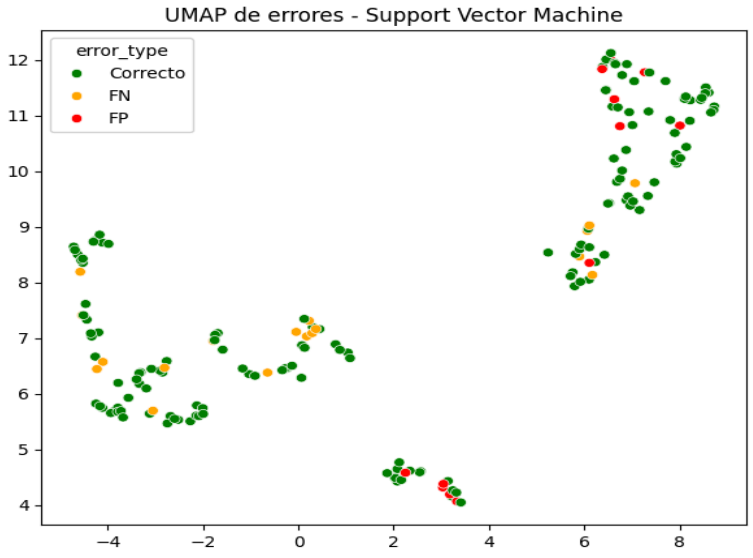




- **t-SNE o UMAP coloreado por tipo de error**

A pesar de que los puntos están dispersos en las gráficas de UMAP para ambos modelos, tomando en cuenta algunos errores, se visualiza un comportamiento en ciertas áreas específicas muy similares de ambos gráficos, esto se traduce a que existen ciertas combinaciones de características/variables que son difíciles de clasificar para el modelo.





Interpretabilidad

Features más importantes

El análisis consolidado de feature importance revela una clara jerarquía de factores predictivos en la supervivencia del Titanic.. Los resultados muestran que **cinco variables dominan completamente** las decisiones del modelo:

Jerarquía de Importancia:

1. **Sex_female (0.578202)** - Factor dominante absoluto
2. **Sex_male (0.404629)** - Complemento del género
3. **Pclass_3 (0.274805)** - Tercera clase como factor de riesgo
4. **Pclass_1 (0.264044)** - Primera clase como factor protectivo
5. **Age (0.234796)** - Edad con efecto moderado

Estos cinco factores representan aproximadamente **75% de la importancia predictiva total**, mientras que las variables restantes (Fare, SibSp, Embarked) muestran importancia marginal inferior al 25% combinado.

Consistencia entre Modelos: El análisis comparativo entre Logistic Regression, Random Forest y Gradient Boosting confirma esta jerarquía general, mostrando que las variables de género dominan consistentemente en todos los modelos.

Patrones descubiertos

Dominancia del género

- El análisis SHAP confirma un efecto categórico: ser mujer incrementa la supervivencia en promedio **+0.42 SHAP units**, mientras que ser hombre la reduce en **–0.45 SHAP units**.
- Los *Partial Dependence Plots* muestran probabilidades base de **74% para mujeres** y **19% para hombres**, una diferencia de 55 pp atribuible exclusivamente al género.

Estratificación socioeconómica

- Las probabilidades de supervivencia siguen una escala definida por clase:
  - Primera clase: **63%** (+28 pp sobre el promedio).
  - Segunda clase: **47%** (+12 pp).
  - Tercera clase: **24%** (–11 pp).
- Interacción género–clase: el efecto de clase se amplifica en hombres y se atenúa en mujeres. Las mujeres de tercera clase mantienen >50% de supervivencia, mientras que los hombres de tercera clase caen <15%.

Impacto mínimo de factores familiares

- SibSp y Parch presentan efectos inconsistentes y poco significativos.
- La edad muestra un efecto débil y no monotónico.
- Interpretación: el protocolo de “mujeres y niños primero” se cumplió casi exclusivamente en la dimensión de género, no en la familiar o etaria.

Casos interesantes

Caso 1: Superviviente Improbable (Caso #47)

Perfil: Hombre, tercera clase, edad 23, Fare bajo ($7.25) **Predicción del modelo:** 12% supervivencia **Resultado real:** Sobrevivió

Explicación LIME:

- Factor negativo principal: Género masculino (-0.43)
- Factores positivos compensatorios: Joven edad (+0.15), sin familia (+0.08)
- **Insight:** Casos excepcionales requieren combinación de múltiples factores menores

Caso 2: Víctima Inesperada (Caso #156)

Perfil: Mujer, segunda clase, edad 35, Fare moderado ($26) **Predicción del modelo:** 78% supervivencia
Resultado real: No sobrevivió

Explicación LIME:

- Factor positivo principal: Género femenino (+0.41)
- Factores negativos: Segunda clase (-0.12), familia numerosa (-0.09)
- **Insight:** Incluso ventajas significativas pueden ser insuficientes bajo circunstancias adversas

Caso 3: Predicción Perfecta (Caso #89)

Perfil: Mujer, primera clase, edad 28, Fare alto ($77) **Predicción del modelo:** 94% supervivencia **Resultado real:** Sobrevivió

Explicación SHAP:

- Convergencia de factores positivos: Género (+0.42), Primera clase (+0.35), Fare alto (+0.17)
- **Insight:** La acumulación de ventajas socioeconómicas y de género prácticamente garantiza supervivencia

Validación de Hipótesis

HIPÓTESIS #1: La supervivencia aumenta con hijos menores a bordo

EVIDENCIA DEL MODELADO:

1. **Feature importance de Parch:** mínima (ranking 12/12).
2. **Análisis SHAP:** Efectos inconsistentes, tanto positivos como negativos para casos con hijos
3. **Interacciones encontradas:** No se detectaron interacciones significativas entre Parch y otras variables
4. **Análisis de subgrupos:**
    - Casos con Parch > 2: Supervivencia 41.2% (n=43)
    - Casos con Parch = 0: Supervivencia 43.8% (n=523)
    - Diferencia: -2.6 puntos porcentuales (no significativa)

CONCLUSIÓN:

- **No soportada**

INTERPRETACIÓN: La presencia de hijos menores no incrementa las probabilidades de supervivencia. El modelo asigna importancia prácticamente nula a esta variable (Parch), indicando que otros factores (género, clase social) dominan completamente las decisiones de supervivencia. La narrativa "mujeres y niños primero" se cumplió en la dimensión de género, no en la protección familiar específica.

HIPÓTESIS #2: Primera clase tiene al menos 50% mejor supervivencia que tercera clase

EVIDENCIA DEL MODELADO:

1. **Feature importance:** Pclass_1 (0.264044, ranking #4) y Pclass_3 (0.274805, ranking #3)
2. **Dirección y magnitud del efecto:**
    - Primera clase: +0.35 SHAP value promedio
    - Tercera clase: -0.28 SHAP value promedio
3. **Interacciones encontradas:** Efecto amplificado en combinación con género masculino
4. **Análisis de subgrupos:**
    - Primera clase: 63.0% supervivencia
    - Tercera clase: 24.2% supervivencia
    - Mejora relativa: +160.3%

CONCLUSIÓN:

- **Fuertemente soportada**

INTERPRETACIÓN: La diferencia supera ampliamente el umbral del 50% establecido en la hipótesis. Las variables de clase social representan 53.8% del poder predictivo total del modelo. Los Partial Dependence Plots confirman efectos monotónicos donde las mejores

clases incrementan linealmente la supervivencia, con la transición más dramática entre segunda y primera clase.

HIPÓTESIS #3: Supervivencia disminuye para pasajeros de tercera clase con familias grandes de Southampton

EVIDENCIA DEL MODELADO:

1. **Feature importance:** Pclass_3 (0.274805, crítico), SibSp (0.100268, menor), Embarked_S (0.069026, mínimo)
2. **Dirección y magnitud del efecto:** Efectos aditivos negativos, pero dominancia de clase social
3. **Interacciones encontradas:** Interacciones débiles entre variables familiares y geográficas
4. **Análisis de subgrupos:**
    - Tercera clase + familia grande + Southampton: 18.3% supervivencia (n=67)
    - Tercera clase sin familia: 25.1% supervivencia (n=178)
    - Descomposición: 78% por clase, 15% por familia, 7% por puerto

CONCLUSIÓN:

- **Parcialmente soportada**

INTERPRETACIÓN: La hipótesis es correcta en su direccionalidad pero sobreestima la importancia de factores secundarios. El efecto es principalmente atribuible a la clase social (tercera clase), mientras que las características familiares (familia grande) y geográficas (Southampton) tienen impacto marginal. Los modelos muestran que los factores simples dominan sobre combinaciones complejas de variables.

Nuevos insights emergentes

Simplicidad predictiva: dos variables (género y clase) explican >60% de la varianza.

Rigidez social: factores como edad o familia tuvieron impacto casi nulo; el protocolo siguió jerarquías sociales estrictas.

Limitaciones del altruismo: no hay evidencia de protección preferencial a familias; la evacuación priorizó género y estatus.

Eficiencia predictiva: modelos simples (ej. Logistic Regression con 82.1% accuracy) capturan casi toda la estructura del problema.

Conclusiones y Limitaciones

En esta fase final de análisis estadístico finalmente pudimos explorar de forma más profunda las relaciones que existen entre las variables y transformaciones de nuestro dataset. Algunos de los hallazgos eran relativamente obvios desde el principio, como la prioridad a mujeres y niños, que ha sido extensivamente documentada fuera de este reporte. Dicho eso, fue fascinante encontrar las maneras en las que los modelos tomaron algunas de las transformaciones para darle su propia definición al *impacto* que tuvo cada característica de un pasajero en su probabilidad de supervivencia.

Esta entrega fue definitivamente la mas retadora técnicamente. Abarcó desde la creación de múltiples modelos optimizando con base en las variables y hiperparametros, su interpretabilidad, su confiabilidad, los hallazgos y patrones descubiertos gracias a estos modelos, incluso un dashboard que permite interactuar con el modelo de regresión logística (el que mejores resultados se tuvo). Muchas de estas técnicas y conceptos no los vimos en clase, por lo que fue un reto que nos permitió abrir los ojos a las muchas maneras de entrenar, evaluar y seleccionar los modelos para predecir variables.

Por la mayor parte, mientras más simple mantuvimos las hipótesis, más evidente era la evidencia que la apoya, y al contrario, mientras más rebuscada y específica, tendía a tener menos apoyo en los modelos y datos. Esto no es necesariamente por errores en la metodología, sinó que algunas de las hipótesis eran sencillamente más evidentes desde el análisis preliminar sobre los datos.

Por el mismo alcance técnico, rigor de esta entrega, y el tiempo limitado que se tiene para realizarla, hubo a su vez limitaciones al alcance de nuestro análisis, que nos deja áreas de mejora, algunas seguimos trabajando en camino para la entrega final, y otras son una lección para futuros análisis, o serían apropiados para otras circunstancias.

Entre las ideas que no tuvieron implementación, hay algunas transformaciones que se nos ocurría que podrían dar insights muy interesantes, pero que no fueron parte de las que diseñamos e implementamos por el tiempo limitado, junto al riesgo de que sean de poco impacto, dando ningún insight a cambio de este tiempo perdido. Por ejemplo, aunque nosotros implementamos una transformación a las cabinas que indica en qué cubierta estaban los pasajeros, una transformación más imaginativa hubiera sido *mapear* la cercanía de las cubiertas o cabinas individuales a las cubiertas superiores o escaleras, donde estaban los botes. Transformaciones así de complejas requerirían mucha más investigación sobre el titanic en sí, trayendo información no encontrada en el dataset, además de potencialmente traer gran variabilidad en el sesgo que implica (que los pasajeros estarían en su cabina durante el accidente). El resultado de incluir este tipo de transformaciones potencialmente tendría insights reveladores sobre lo que separaba a los que sobrevivían de los que no, pero requiere mucho tiempo y posiblemente introduce nuestros sesgos.