PREDICTIVE SURVIVAL ANALYSIS ON THE TITANIC: APPLICATION OF MACHINE LEARNING ALGORITHMS FOR PASSENGER CLASSIFICATION

Juan J. Bonilla, Ricardo Muñoz, Valentina Isaza, Nelcy L. Zapata,

Master's students in Artificial Intelligence and Data Science
Universidad Autónoma de Occidente, Cll 25 # 115-85 Km 2 Vía Cali - Jamundí · Cali, Colombia,

***Abstract*-- This paper addresses the problem of predicting the survival of Titanic passengers using machine learning techniques. Based on variables such as age, gender, social class and family relationships, supervised models were trained to identify patterns associated with survival.**

The results confirm that being a woman, traveling first class and paying a high fare significantly increased the probability of survival, reflecting social decisions and structural inequalities of the time. The models also show the disadvantage of men and third-class passengers, possibly due to their lower access to lifeboats.

Among the algorithms evaluated, XGBoost and SVM offered the best balance between accuracy and sensitivity, while Logistic Regression, although simpler, showed competitive performance and high interpretability, ideal for a first approach to the problem.

I. INTRODUCTION

Predictive analytics based on historical data has emerged as a fundamental tool in the field of data science and artificial intelligence. A representative case widely used for educational purposes is the 1912 RMS Titanic disaster, where more than 1,500 people lost their lives after the sinking of the ocean liner. This event, while historical, provides an ideal context for applying and demonstrating modern machine learning techniques in a controlled environment. The present study addresses the problem of predicting the probability of survival of Titanic passengers, using as a basis the dataset provided by the Kaggle platform. Based on variables such as age, gender, socioeconomic class (Pclass), number of family members on board and port of embarkation, the aim is to build a supervised model capable of identifying hidden patterns and making accurate predictions about the survival of individuals.

This problem, beyond its historical relevance, allows researchers and students to practice key concepts of the data science workflow, including exploratory analysis, missing value imputation, feature engineering, supervised modeling and model cross-validation. It also explores the use of various classification algorithms, such as logistic regression, Random Forest and the extreme gradient method (XGBoost), evaluating their performance using metrics such as precision, recall and F1-score.

II. OBJECTIVE

The purpose of this work is to develop and implement a supervised machine learning model to predict the probability of survival of the RMS Titanic passengers from the variables provided in the original dataset distributed by Kaggle. To this end, a systematic approach will be employed that includes selection and transformation of relevant features, imputation of missing values, and validation of multiple classification algorithms. The model will be trained on a subset of the dataset (titanic_training.csv) and evaluated using reserved data, with the objective of maximizing its predictive capability through performance metrics such as accuracy, recall and F1 score. The proposal seeks to ensure the replicability of the process and provide a solid foundation for future improvements in educational and professional environments.

III. PROBLEM TO BE ADDRESSED

This project aims to address the following question: what kind of people were most likely to survive the sinking of the Titanic? To answer this question, we propose the construction of a supervised machine learning model to predict, based on historical data, whether or not a passenger survived the maritime disaster. The prediction is based on relevant variables such as age, gender, socioeconomic class (Pclass) and family ties on board (SibSp and Parch). The problem is framed within a binary classification task, where the target variable is passenger survival. The solution involves identifying patterns in the data that explain the relationship between individual attributes and the observed outcome.

IV. BACKGROUND

The sinking of the RMS Titanic in 1912 has been the subject of numerous studies in the field of machine learning, due to the availability of detailed data on the passengers and the circumstances of the disaster. Several investigations have explored the application of classification algorithms to predict passenger survival, considering variables such as age, gender, ticket class, and boarding point.

Sherlock et al [1] used data mining tools, specifically Weka, to analyze the relationship between passenger characteristics and their probability of survival. The study identified that cabin class, age, and port of embarkation were significant factors influencing survival rates.

Liang [2] applied decision tree and random forest algorithms to predict passenger survival. The results showed that age, gender, and ticket class were the variables most correlated with survival. Furthermore, it was observed that the decision tree algorithm outperformed the random forest algorithm in terms of accuracy in this specific context.

Amalia and Rahayu [3] compared multiple classification models, including logistic regression, random forests, and XGBoost, to predict the survival of Titanic passengers. The study concluded that the random forests model achieved the highest accuracy, highlighting the effectiveness of ensemble methods on data sets with mixed variables.

Liao et al [4] proposed a hard voting model combining logistic regression, random forests and decision trees to predict passenger survival. This integrative approach achieved an accuracy of 87.64%, evidencing that combining classifiers can improve the robustness of predictions in structured data sets.

These studies provide a solid foundation for the development of predictive models in the Titanic context, and serve as a reference for the implementation and evaluation of machine learning algorithms in binary classification problems.

V. EXPLORATORY DATA ANALYSIS

Dataset Preparation and Exploration

A. Importing Libraries

For exploratory analysis and data preprocessing, widely used libraries in the Python data science ecosystem were employed: pandas and numpy for data structure manipulation, matplotlib and seaborn for graphical visualization, and scikit-learn for predictive modeling and variable transformation. Also included were scipy.stats for statistical testing and google.colab.drive for integration with Google Drive, which served as a data source.

B. Loading the Dataset

The training set was obtained from the specified path in Google Drive, corresponding to the file titanic_training.csv, which contains 999 records with passenger information on board the Titanic.

C. Dataset Context

The dataset is framed by the historical event of the sinking of the RMS Titanic that occurred between April 14 and 15, 1912, during its maiden voyage between Southampton (UK) and Nova Scotia (Canada). After a collision with an iceberg, the ship sank in less than three hours, causing the death of 1502 people out of a total of 2224 passengers and crew. The insufficient number of lifeboats and the boarding conditions influenced the survival rate.

The dataset contains relevant passenger information, with both demographic and operational variables, which can be determinant in predicting survival. Table I summarizes the characteristics included in the dataset:

Variable	*Descripción*	*Valores*
survival	Sobrevivió?	0 = No, 1 = Sí
pclass	Clase del boleto de embarque	1 = primera, 2 = segunda, 3 = tercera
sex	Sexo	male = Masculino, female = Femenino
age	Edad en años	
sibsp	Cantidad de hermanos o esposas del sujeto a bordo	
parch	Cantidad de padres o hijos del sujeto a bordo Titanic	
ticket	Número de pase de abordaje	
fare	Tarifa de embarque	
cabin	Número de habitación	
embarked	Puerto de embarque	C = Cherbourg, Q = Queenstown, S = Southampton

Table Number 1 - Dataset Context

D. Initial Exploration of the Dataset

Preliminary analysis of the DataFrame revealed a total of 999 entries and 10 columns. Summary statistics showed the presence of null values in multiple variables:

Nº	*Columna*	*Valores no nulos*	*Tipo de dato*	*Observaciones*
0	survived	999	float64	Variable objetivo (0 = No, 1 = Sí)
1	pclass	999	float64	Clase del boleto (1, 2, 3)
2	sex	999	object	Género (male, female)
3	age	804	float64	Edad (valores faltantes)
4	sibsp	999	float64	Nº de hermanos/cónyuge a bordo
5	parch	999	float64	Nº de padres/hijos a bordo
6	ticket	999	object	Número de boleto
7	fare	998	float64	Tarifa pagada
8	cabin	227	object	Cabina asignada (muchos nulos)
9	embarked	997	object	Puerto de embarque (C, Q, S)

Table Number 2 - Exploration Dataset

It was observed that the variables age, fare, cabin and embarked had missing values in varying proportions. Particularly, cabin had a high proportion of missing data (~77.3%), so its use was subsequently reconsidered. In addition, one record was identified with a null value in the survived variable, representing the target label. This record was discarded from the analysis, since its absence prevents supervised training of the model.

E. Ordinal Coding of Categorical Variables

Since machine learning algorithms require numerical data, the categorical variables sex, embarked and cabin were coded using Ordinal Encoder, thus obtaining Coded Titanic DataFrame and then generating the descriptive statistics of the transformed DataFrame.

Estadística	*survived*	*pclass*	*age*	*sibsp*	*parch*	*fare*	*sex_encoded*	*embarked_encoded*	*cabin_encoded*
count	999.0	999.0	804.0	999.0	999.0	998.0	999.0	997.0	227.0
mean	0.386	2.287	30.26	0.504	0.409	34.622	0.651	1.518	78.194
std	0.487	0.844	14.63	1.058	0.897	54.390	0.477	0.801	43.059
min	0.0	1.0	0.167	0.0	0.0	0.0	0.0	0.0	0.0
25%	0.0	1.0	21.0	0.0	0.0	7.896	0.0	1.0	42.5
50%	0.0	3.0	28.0	0.0	0.0	14.458	1.0	2.0	78.0
75%	1.0	3.0	39.0	1.0	0.0	32.116	1.0	2.0	113.5
max	1.0	3.0	80.0	8.0	9.0	512.329	1.0	2.0	152.0

Table Number 3 - Dataset Coding

Correlation Analysis with the Target Variable

In order to identify linear associations between the survived target variable (survival) and other continuous or high cardinality numerical variables, the point-biserial correlation coefficient was used, which is suitable for measuring the relationship between a binary variable and a continuous variable. This coefficient takes values between -1 and 1, where values close to the extremes indicate a high association (positive or negative), and values close to zero indicate a weak or null association.

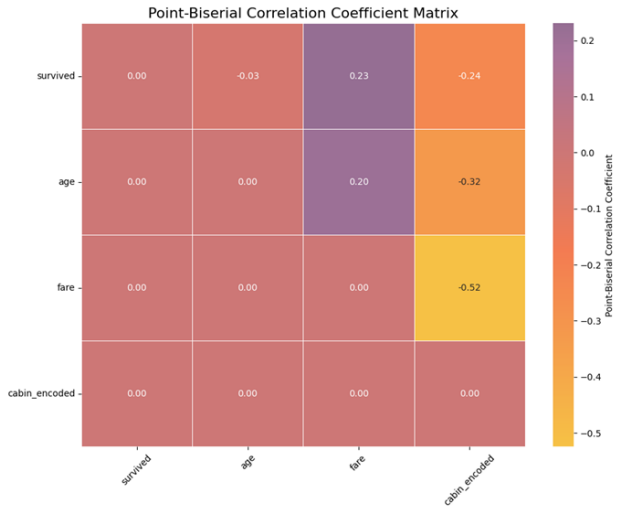
A. Handling Missing Values

Since the calculation of correlations using scipy.stats.pointbiserialr does not support null values, a new DataFrame was generated with specific imputations:

- age and embarked: imputed with the mode, since there is no presence of decimals in age.
- fare: Imputed with the mean, since it is a continuous monetary variable.
- cabin_encoded: Assigned a new value called "Unknown", which can represent shared or unassigned cabins, common among third class passengers.

B. Point-Biserial Correlation Matrix

In addition, the point-biserial correlation between survived and the following variables: age, fare, and cabin_encoded was calculated. The correlation coefficients are presented below.



Matrix Number 1 - Point-Biserial Correlation Matrix

The following hypotheses are presented:

- ***fare***

Correlation with survived: +0.231

Interpretation: There is a moderate positive correlation between the value of the fare and the probability of surviving. This suggests that passengers who paid more for their ticket (possibly from higher classes) were more likely to survive. This is consistent with historical facts: first class passengers had priority in the lifeboats.

- ***cabin_encoded***

Correlation with survived: -0.241

Interpretation: This moderate negative correlation indicates that as cabin number/coding increases (or when no cabin is assigned, such as in low class), the probability of survival decreases. That is, passengers with simpler cabins or no cabin tended to have lower survival rates.

- ***age***

Correlation with survived: -0.033

Interpretation: This correlation is very weak and negative, indicating that age had very little impact on survival based on these data. Although one would expect children to have higher priority, the signal is so small that it is not considered significant.

- ***fare and cabin_encoded***

Correlation: -0.525

Interpretation: they also have a strong negative correlation with each other, which may indicate that more “coded” (or absent) cabins were associated with cheaper tickets. This makes sense, since third class passengers generally did not have their own cabins.

- ***age and fare***

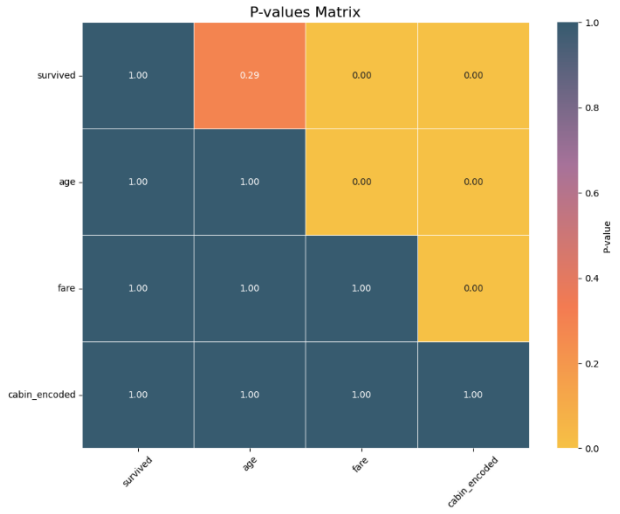
Correlation: 0.204

Interpretation: they have a low positive correlation (0.204), which could suggest that older people paid slightly more expensive fares (possibly because they traveled in better conditions), although the relationship is weak.

C. Matrix of P - Values

Keeping in mind the p-values, we take into consideration that a p-value indicates the probability that an observed correlation occurs by chance.

- If $p < 0.05$, it is considered statistically significant (i.e., we are confident that the correlation is real).
- If $p > 0.05$, the correlation is considered non-significant (it could have occurred by chance).



Matriz Número 2 – Matriz de P-valores

- ***fare vs survived***

p = 1.62e-13: Highly significant

Interpretación: The positive correlation between passage value and probability of survival is statistically reliable.

- ***cabin_encoded vs survived***

p = 1.04e-14: Highly significant

Interpretación: We can also rely on the negative correlation between cabin type/level and survival.

- **age vs survived**

p = 0.292: Not significant

Interpretación: There is insufficient statistical evidence to claim that age influences the probability of survival in this data set. It may have been a random effect or too weak.

Next, we will examine how certain individual characteristics of Titanic passengers were related to their probability of surviving the disaster. Correlation results are detailed for fare paid (Fare), cabin number (Cabin_encoded) and age (Age). For each variable, the correlation coefficient, the associated p-value, and an interpretation of the findings are presented, along with possible contextual explanations based on the social and physical structure of the ship.

1. Fare

- There is a moderate positive correlation between the fare paid and the probability of survival.
- This is statistically significant, indicating that this pattern is not a product of chance.
- **Possible explanation:** Passengers who paid higher fares tended to travel in first class, where they were located in areas closer to the lifeboats and received priority attention during the evacuation. In addition, Titanic personnel were conditioned to protect upper-class passengers first, reflecting a hierarchical social structure typical of the era.

2. Cabin_encoded

- There is a moderate negative correlation between cabin number and survival.
- This suggests that people with high cabin_encoded values, which may correspond to high numbered, shared or even unknown cabins, were less likely to survive.
- **Possible explanation:** Many third-class passengers did not have an individual cabin, and if they did, it was not identified on the manifest. These passengers were located on the lower levels of the ship, away from the exits and boats. This reflects a structural factor of exclusion and unequal access during the emergency.

3. Age

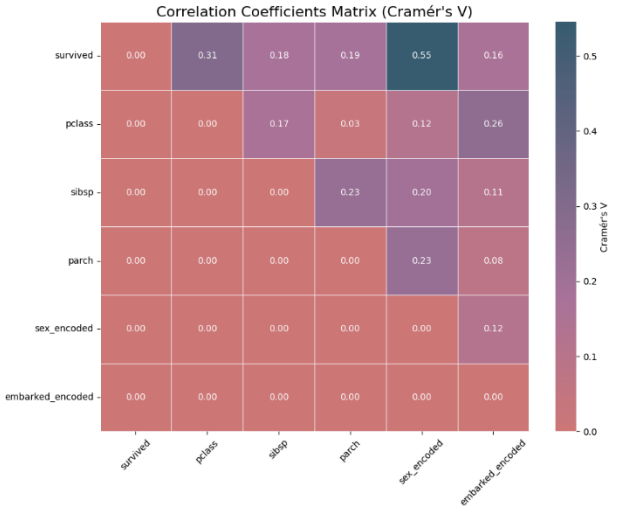
- Age has a very weak and non-significant correlation with survival. This means that, on average, age had no clear or systematic effect on who did or did not survive.
- **Possible explanation:** Although there was a "women and children first" slogan, in practice it was not always consistently followed. In addition, the age variable alone does not distinguish between young children, adolescents, or young adults, which may be diluting the effect.

VI. CATEGORICAL VARIABLES ANALYSIS

In order to understand the strength of the relationship between the categorical variables in the Titanic data set and the target variable (survived), Cramér's V coefficient of association was calculated. This coefficient measures the strength of association between two categorical variables, taking values between 0 (no association) and 1 (perfect association). The analysis was complemented with the corresponding p-values to evaluate the statistical significance of each relationship.

A. Cramér's V Coefficient Matrix

To understand the relationship between categorical variables and the probability of survival on the Titanic, the Cramér's V coefficient was used, which quantifies the strength of association between pairs of categorical variables. This analysis allows us to identify which characteristics have the greatest explanatory power on the outcome of the passengers (survived or not).



Matrix Number 3 - Cramér's V Coefficients

Relationship with the Target Variable (survived):

- **sex_encoded (passenger sex):**

Cramér's V = 0.545

This is the strongest association observed. This reflects a clear and marked difference between males and females in the probability of survival. The high value suggests that sex was a determining factor during evacuation, reinforcing the well-known historical "women and children

- **pclass:**

Cramér's V = 0.306

It shows a moderate association with survival. First class passengers had greater access to lifeboats and better location within the ship, which influenced their likelihood of survival. This finding is consistent with patterns of social inequality during the shipwreck.

- **parch:**

Cramér's V = 0.187

Although weak, it indicates that the presence of close relatives could have affected behavior in emergency situations, e.g., the desire to stay together or mutual support.

- **sibsp:**

Cramér's V = 0.177

Like parch, it suggests a slight relationship with survival. It could influence both positively (family cooperation) and negatively (difficulty in group evacuation).

- **embarked_encoded:**

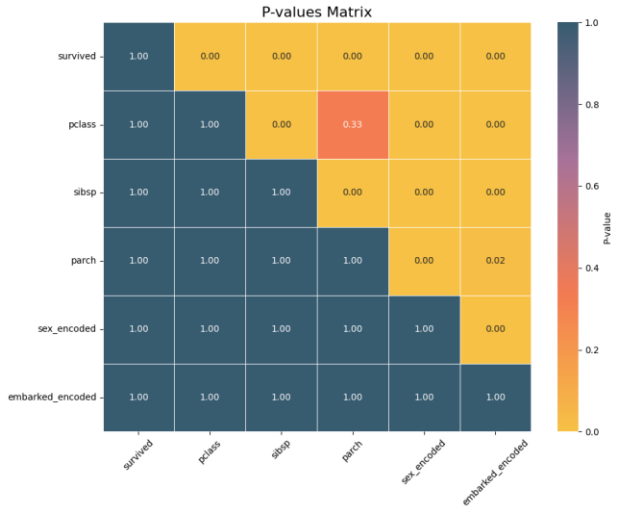
Cramér's V = 0.164

Weak association. It does not seem to have a determinant weight, although it may reflect indirect differences in the socioeconomic profile of passengers according to boarding location.

This analysis reinforces the importance of sex and pclass as key variables for predicting survival in classification models. Their inclusion is essential for both predictive performance and historical interpretability of the phenomenon.

B. Matrix of P-values for Cramér's V

After calculating the Cramér's V coefficients that estimate the strength of association between categorical variables, it is necessary to verify whether these associations are statistically significant. For this purpose, p-values were used, which indicate the probability of observing the detected relationship (or a more extreme one) if an association did not actually exist in the population.



Matrix Number 4 - P-values for Cramér's V

- **sex_encoded: p ≈ 9.26e-67:**

The association between sex and survival is extremely significant. This value quantitatively reinforces what the history and Cramér's V coefficient already suggested: sex had a deterministic impact on the odds of surviving the shipwreck.

- **Pclass: p ≈ 1.68e-21:**

Also highly significant. Indicates that passenger class strongly influenced the outcome. Boat access and location on the ship, associated with class, likely account for this influence.

- **parch, sibsp, embarked_encoded: p < 0.00001:**

Although their strength of association was low (see Cramér's V analysis), the p-values show that these relationships are not random. That is, having relatives on board or place of embarkation may have a subtle but real influence on survival.

- **pclass y parch: p ≈ 0.33:**

This is the only non-significant relationship in the entire matrix. It indicates that class is not statistically associated with the number of parents/children on board. Which makes sense: passengers of any class could travel with or without family.

- **Generalidades**

all other combinations between variables also show extremely low p-values, which supports the idea that the categorical variables used contain relevant, non-random signals about the structure of the data.

VII. MODELING

A. Evaluation of Predictive Models

In order to evaluate the performance of different classification approaches for predicting Titanic passenger survival, six supervised models were implemented and compared: Logistic Regression, Random Forest, XGBoost, Support Vector Machines (SVM), K-Nearest Neighbors (KNN) and Naive Bayes. In addition to considering traditional metrics such as accuracy, the analysis focused especially on the performance of each model against the minority class (survivors), using measures such as recall, precision and F1-score.

Modelo	*Exactitud*	*Recall (Clase 1)*	*Precision (Clase 1)*	*F1-score (Clase 1)*
Regresión Logística	0.795	0.704	0.714	0.709
Random Forest	0.790	0.718	0.699	0.708
XGBoost	0.805	**0.761**	0.711	**0.735**
SVM	**0.810**	0.676	**0.762**	0.716
KNN	0.785	0.676	0.706	0.691
Gaussian Naive Bayes	0.780	0.747	0.671	0.707

Table Number 4 - Predictive Modelss

This approach is crucial in problems with unbalanced classes, where the cost of classification errors varies by class. Detailed observations on the performance of each model are presented below, with emphasis on their ability to correctly identify survivors (class 1), as well as their potential advantages and contextual limitations.

- **Logistic Regression**

This model is presented as the interpretable basis of the analysis. Its overall performance is solid, with acceptable accuracy and easily explainable coefficients. However, its ability to correctly detect survivors is limited: approximately 30% of them are not identified (low recall), which could be problematic in scenarios where false negatives have a high cost.

- **Random Forest**

Offers similar performance to logistic regression, but shows greater robustness to nonlinear relationships between variables. It shows a slight improvement in class 1 recall, indicating better recognition of survivors, although still only moderately.

- **XGBoost**

Stands out as the model with the best balance between accuracy and recall. Its architecture allows capturing complex patterns in the data, which translates into a more efficient identification of survivors without sacrificing too much accuracy. This balance makes it a particularly suitable option in contexts where minimizing false negatives is a priority.

- **Support Vector Machine (SVM)**

The SVM model demonstrates a higher accuracy in predicting survivors, suggesting a low false positive rate. However, its recall is the lowest among the models evaluated, indicating a limited ability to detect all individuals in the positive class. This characteristic makes it more appropriate in scenarios where false positives are more costly than false negatives.

- **K-Nearest Neighbors (KNN)**

Although simple and highly dependent on the local structure of the data, the KNN model offers reasonable performance. Despite its sensitivity to noise and its overall lower generalization capability, it achieves performance comparable to other more complex models, although it does not reach the levels of accuracy or recall observed in XGBoost or SVM.

- **Naive Bayes**

Despite its simplicity and strong assumption of conditional independence between features, Naive Bayes proves competitively effective. The model succeeds in identifying a significant proportion of survivors (good recall), although it does so at the cost of lower precision, which implies a higher number of false positives. This trade-off may be acceptable depending on the application context.

B. Hyperparameter Optimization

Logistic Regression, XGBoost and Support Vector Machine (SVM) models were selected for the hyperparameter optimization stage due to their theoretical relevance, generalizability and their performance observed in the initial tests:

- **Logistic Regression**

Although it is a simple model, Logistic Regression showed a competitive performance (F1-score = 0.709) comparable to that of more complex models. It stands out for its interpretability and low number of hyperparameters, which allows a clear evaluation of the effect of its optimization.

- **XGBoost**

This model presented one of the best overall performances, with a high accuracy (0.805) and the highest recall (0.761), indicating its ability to correctly identify survivors (class 1). This makes it an ideal candidate for optimization, as small adjustments to its multiple hyperparameters (such as number of trees, depth, learning rate) can further maximize its performance.

- **Support Vector Machine (SVM)**

SVM obtained the highest accuracy (0.762) for class 1, indicating that when it predicts a survivor, it is generally correct. Although its recall was lower (0.676), this feature makes it valuable in contexts where false positives are more costly than false negatives. Since SVM is very sensitive to hyperparameter settings such as kernel, C and gamma, its inclusion allows exploring its potential for improvement through appropriate optimization.

After applying GridSearchCV to tune the hyperparameters of three selected models: Logistic Regression, XGBoost and SVM, it was observed that performance improvements were marginal or non-existent.

	Modelo	*Exactitud (Test)*	*Recall (Clase 1)*	*Precision (Clase 1)*	*F1-score (Clase 1)*	*Recall (Clase 0)*	*Precision (Clase 0)*	*F1-score (Clase 0)*
0	Regresión Logística (GridSearchCV)	0.795	0.6901	0.7206	0.705	0.8527	0.8333	0.8429
1	XGBoost (GridSearchCV)	0.805	0.6479	0.7667	0.7023	0.8915	0.8214	0.855
2	SVM (GridSearchCV)	0.81	0.662	0.7705	0.7121	0.8915	0.8273	0.8582

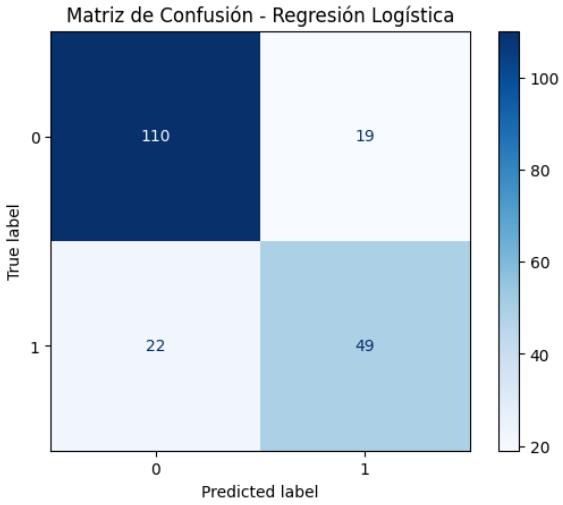
Table 5 - Optimization

- Logistic Regression showed almost identical performance after optimization, suggesting that it was already operating close to its optimal point or that its capacity for improvement by hyperparameters is limited given the linear structure of the model.
- In the case of XGBoost, while accuracy improved in the positive class (survivors), there was a considerable decrease in recall, implying that fewer survivors were correctly identified. This loss of sensitivity reduces its usefulness in scenarios where false negatives are costly, as is the case in the present problem.
- SVM, on the other hand, maintained practically the same level of performance as before optimization, with slight variations in precision and recall, but without significant impact on the overall metrics.

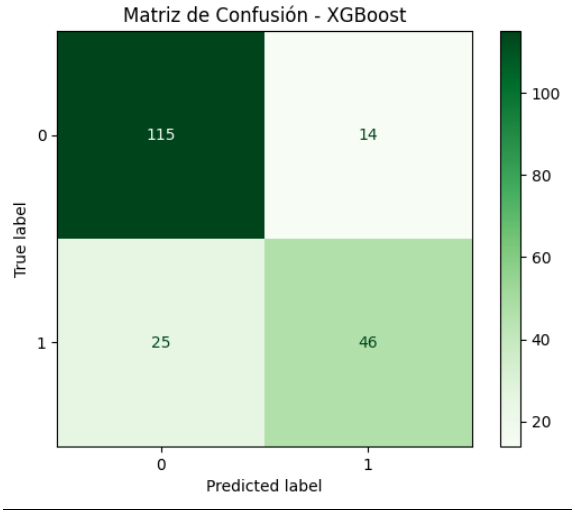
VIII. VISUALIZATION AND INTERPRETABILITY

A. Confusion Matrix

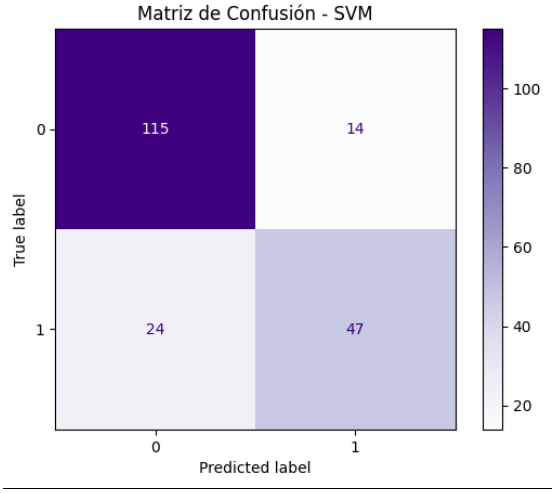
Confusion matrices allow to clearly observe type I and type II errors (false positives and false negatives), which is crucial in contexts such as survival prediction, where not detecting a survivor may be more serious than predicting one incorrectly.



Matrix Number5 - Confusion Matrix - Logistic Regression



Matrix Number 6 - Confusion Matrix - XGBoost



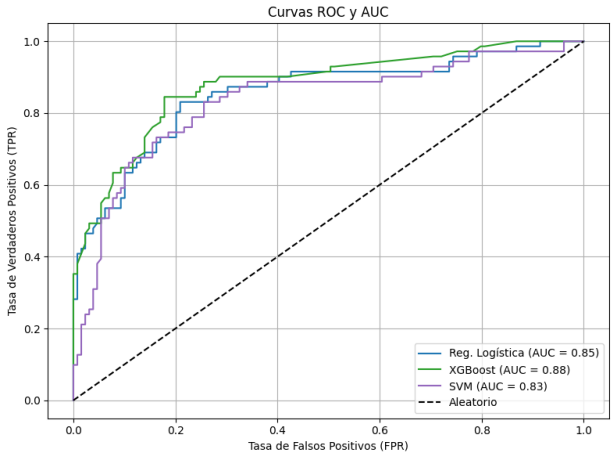
Matrix Number 7 - Confusion Matrix - SVM

Logistic Regression manages to identify a larger number of survivors (49), which is relevant given that, in the Titanic context, survivors were a minority and their identification is key to understanding survival patterns (e.g., sex, class or age).

XGBoost and SVM identify fewer survivors correctly, but are more conservative in labeling someone as a survivor (lower false positive rate).

B. ROC curves and AUC

ROC curves and AUC indicate how well a model distinguishes between survivors and non-survivors, without the need to set a specific threshold.



Graph Number 1 - ROC and AUC curves

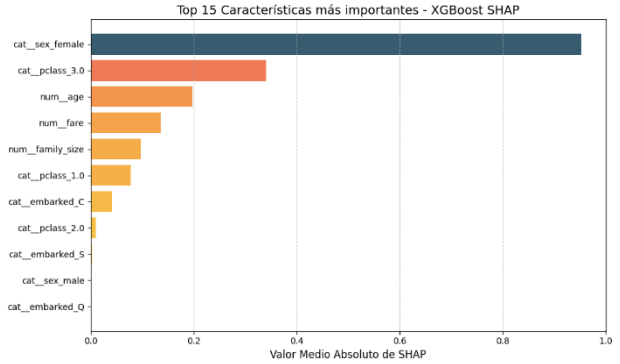
XGBoost achieves the best AUC, suggesting that it has a high ability to correctly detect both survivors and non-survivors, even with different thresholds. This is important if one wishes to adjust the model for different contexts (e.g., historical simulations or hypothetical decisions about evacuation priorities).

Logistic regression, although simpler, also offers high discrimination, reinforcing its usefulness as an interpretable model for understanding what historical factors influenced survival.

SVM, although it has a lower AUC, still performs reasonably well, and suggests that the maximum margin it draws between classes is still useful, but perhaps less well adjusted to the social and demographic pattern of Titanic passengers.

C. Significance of Characteristics with SHAP

SHAP analysis reveals which variables were most determinative of model decisions. On the Titanic, this helps confirm or disprove historical hypotheses about who was most likely to survive.



Graph Number 2 - SHAP

Characteristic: sex_female
Mean SHAP: 0.953
Interpretation: Being female was the most decisive factor. This confirms the "women and children first" policy, which prevailed during the evacuation.

Characteristic: pclass_3.0
Average SHAP: 0.341
Interpretation: Traveling in third class was associated with a lower probability of survival, possibly due to location on the ship and limited access to lifeboats.

Characteristic: age
Mean SHAP: 0.198
Interpretation: Age is also influential, as children had higher priority.

Characteristic: fare
Mean SHAP: 0.136
Interpretation: Higher fares (indicative of first class) correlate with higher survival.

Characteristic: family_size
Mean SHAP: 0.097
Interpretation: Passengers with larger families may have faced more difficulty evacuating together.

Characteristic: pclass_1.0
Mean SHAP: 0.078
Interpretation: Reinforces that the first class had a better chance of survival.

Visual analysis and interpretability allow verification that the models not only predict with good accuracy, but also consistently capture the historical factors that determined survival on the Titanic. This approach adds narrative and explanatory value to the models, facilitating their use as a tool for historical and sociological analysis.

IX. ANALYSIS OF RESULTS AND CONCLUSION

1. Key survival factors: Through predictive models and interpretability with SHAP, it is confirmed that gender (being female), passenger class (first class) and, to a lesser extent, age and fare paid were the most influential factors in the probability of survival. This reflects the social and structural decisions of the time, in particular the prioritization of women and upper-class passengers during evacuation.

2. Evidence of structural inequality: The results reinforce a critical finding: third-class passengers and men were significantly less likely to survive, suggesting unequal access to lifeboats. The models capture these historical disparities clearly and consistently, highlighting how decisions made during the disaster were influenced by socioeconomic status and gender.

3. XGBoost and SVM: the most balanced models: After optimization, the XGBoost and SVM models offered a superior balance between accuracy, precision and sensitivity, correctly predicting a higher proportion of survivors without significantly compromising the false positive rate. This is especially relevant in contexts where correctly identifying those in need of assistance is a priority, such as in evacuation simulations.

4. Logistic Regression: simple, interpretable and competitive: Despite its simplicity, Logistic Regression showed competitive performance and valuable interpretability. Although its ability to detect survivors was slightly lower, it is an excellent starting point because of its clarity in the relationship between variables and outcome. It is useful for constructing a clear narrative of the phenomenon.

5. Importance of interpretability in historical events: Interpretability techniques (SHAP, confusion matrices, ROC curves) not only helped to understand how models work, but also made it possible to connect numerical patterns with real historical human decisions. This traceability between data, model and context reinforces the educational and analytical value of the approach.

X. REFERENCES

[1] J. Sherlock, M. Muniswamaiah, L. Clarke, y S. Cicoria, "Classification of Titanic Passenger Data and Chances of Surviving the Disaster," arXiv preprint arXiv:1810.09851, 2018. [En línea]. Disponible en: https://arxiv.org/abs/1810.09851

[2] W. Liang, "Titanic Disaster Prediction Based on Machine Learning Algorithms," BC Publication, 2023. [En línea]. Disponible en: https://bcpublication.org/index.php/BM/article/view/4860

[3] A. E. Amalia y C. Rahayu, "Comparison of Machine Learning Classification Models in Predicting The Titanic Survival Rate," International Journal of Computer Science and Human-Aided Intelligence, vol. 3, no. 2, pp. 52–58, 2023. [En línea]. Disponible en: https://journal.binus.ac.id/index.php/ijcshai/article/view/12163

[4] Y. Liao, S. Zhang, y Z. Zhang, "Research on Titanic Survival Prediction Based on Machine Learning Method," Proceedings of the Asia Education Management and Public Service Conference, 2023. [En línea]. Disponible en: https://www.ewadirect.com/proceedings/aemps/article/view/19451

[5] Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. Springer.

[6] Powers, D. M. W. (2011). *Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation*. *Journal of Machine Learning Technologies*, 2(1), 37–63.

[7] Lundberg, S. M., & Lee, S. I. (2017). *A Unified Approach to Interpreting Model Predictions*. In *Advances in Neural Information Processing Systems* (NeurIPS 2017).

[8] Chen, T., & Guestrin, C. (2016). *XGBoost: A Scalable Tree Boosting System*. In *Proceedings of the 22nd ACM SIGKDD* (pp. 785–794).