📊 Final Technical Report - Titanic Survival Prediction

📋 Project Information

Team: SG1_Team3_ML
Members: Andrés López, Héctor Eguiarte, Yahwthani Morales, Omar Vidaña
Course: COM 139 - Simulation & Visualization
University: Universidad Panamericana
Date: May 2025

🎯 Executive Summary

Achieved Objective

We developed a machine learning model that predicts RMS Titanic passenger survival with **84.4% accuracy**, exceeding the academic goal of 80% and revealing significant historical patterns about history's most famous maritime tragedy.

Key Results

- 🏆 **Best Model**: Support Vector Machine (SVM) with RBF kernel
- 📊 **Performance**: 84.4% accuracy, F1=0.78, AUC=0.86
- 🎯 **Historical Insights**: Quantitative validation of "women and children first" protocol
- ⚖️ **Key Factor**: Gender-class intersection determined survival

---

📊 Methodology and Data

Dataset

- **Source**: Kaggle Titanic Dataset
- **Dimensions**: 891 passengers × 12 original features
- **Target Variable**: Survival (38.4% historical rate)
- **Quality**: Missing values in Age (19.9%), Cabin (77.1%), Embarked (0.2%)

CRISP-DM Process

1. **Data Understanding**: Complete EDA identified gender-class patterns
2. **Data Preparation**: Smart imputation, advanced feature engineering
3. **Modeling**: 4 algorithms compared (Logistic Regression, Random Forest, SVM, Naive Bayes)
4. **Evaluation**: Cross-validation, hyperparameter tuning, error analysis
5. **Deployment**: Production model with historical interpretation

---

🔧 Feature Engineering

Created Variables (9 → 29 features)

Successful Derived Variables

1. **FamilySize** (SibSp + Parch + 1): Families of 2-4 people → 55-72% survival
2. **IsAlone**: Accompanied (50.6%) vs Alone (30.4%) survival
3. **AgeGroup**: Child (57.4%) > Young_Adult (33.7%) > Senior (26.9%)
4. **FareBin**: Premium (58.1%) > High (45.5%) > Medium (30.4%) > Low (19.7%)
5. **Title** (extracted from names): Mrs (79.4%) > Miss (70.1%) > Master (57.5%) > Mr (15.7%)

Critical Interaction Variables

- **Sex_Pclass**: female_Class1 (96.8%) vs male_Class3 (13.5%)
- **Age_Sex**: Adult_Female (75.3%) > Young (54.0%) > Adult_Male (16.6%)

Top 3 Most Predictive Features

1. **AgeSex_Adult_Female** (r=0.486): Adult women
2. **SexPclass_female_Class1** (r=0.413): First-class women
3. **Title_Mrs** (r=0.342): Female social status

🤖 Modeling and Results

Algorithm Comparison

Model	Accuracy	F1-Score	AUC-ROC	Overfitting	Interpretability
SVM	**84.4%**	**0.781**	**0.859**	✅ No	Low
Logistic Regression	84.3%	0.781	**0.910**	✅ No	**High**
Random Forest	83.2%	0.783	0.879	⚠️ Yes	Medium
Naive Bayes	82.0%	0.750	0.876	✅ No	Medium

Final Model: Optimized SVM

- **Hyperparameters**: C=1, kernel='rbf', gamma='auto'
- **Validation**: Consistent 5-fold cross-validation
- **Generalization**: Stable performance across train/val/test

Detailed Metrics

- **Precision**: 84.7% (reliability in positive predictions)
- **Recall**: 72.5% (ability to find survivors)
- **F1-Score**: 78.1% (precision-recall balance)
- **AUC-ROC**: 85.9% (excellent discriminative capability)

🔍 Error Analysis

16.2% Total Errors (144 cases)

False Positives (52 cases)

- **Profile**: Mainly young 3rd class women
- **Interpretation**: Model optimistic about female survival
- **Historical Context**: Exceptional cases where location/circumstances prevented evacuation

False Negatives (92 cases)

- **Profile**: Mostly men (86/92) who survived unexpectedly
- **Interpretation**: Cases of heroism, luck, or evacuation assistance
- **Human Value**: Reveal extraordinary survival stories

🏛️ Historical Validation

Confirmed Hypotheses

H1 - "Women and Children First" Protocol ✅

- **Evidence**: Women 74.2% vs Men 18.9% survival
- **Ratio**: 3.9x higher female survival
- **Context**: Birkenhead protocol faithfully applied

H2 - Social Class Determinant ✅

- **Evidence**: 1st class (63.0%) > 2nd class (47.3%) > 3rd class (24.2%)
- **Interpretation**: Privileged access to lifeboats
- **Social Implication**: 1912 social structure reflected in survival

H4 - Gender-Class Intersection ✅

- **Extreme Evidence**: 1st class women (96.8%) vs 3rd class men (13.5%)
- **Multiplicative Factor**: Gender + Social class = More precise prediction
- **Top Features**: The 3 most predictive variables involve this intersection

Validation with Historical Data

- **Lifeboat Capacity**: 1,178 people (53% of 2,224 onboard)
- **Actual Survivors**: ~710 people (32% historical vs 38.4% dataset)
- **Protocol**: Our model faithfully captures "women and children first"

⚖️ Bias and Fairness Analysis

Performance by Subgroups

By Gender

- **Women**: Recall=97.4% (excellent at detecting female survivors)
- **Men**: Precision=100% (conservative, only predicts survival when very confident)

By Social Class

- **Class 1**: Precision=97% (very accurate for first class)
- **Class 2**: Accuracy=92.4% (best overall performance)
- **Class 3**: Recall=62% (difficulty predicting third-class survival)

Ethical Reflection

The model reflects **real historical biases** of the era (1912), not algorithmic biases. Performance differences across subgroups reflect the social inequalities that determined survival access.

🧠 Lessons for Modern Protocols

Applicable Insights

1. **Evacuation Planning**: Consider all social classes equitably
2. **Clear Protocols**: Define transparent and fair priorities
3. **Sufficient Capacity**: Ensure emergency resources for 100% occupancy
4. **Predictive Analysis**: Use ML to optimize evacuation plans

Value of Historical Analysis

- **Quantification**: Convert historical narratives into measurable data
- **Validation**: Confirm or refute historical theories with statistical evidence
- **Learning**: Extract lessons applicable to modern situations

---

📈 Robustness and Limitations

Model Strengths

- ✅ **No Overfitting**: Consistent performance across sets
- ✅ **Generalization**: Stable AUC in cross-validation (0.856 ± 0.031)
- ✅ **Interpretability**: Results aligned with historical knowledge
- ✅ **Reproducibility**: Fully documented and replicable pipeline

Identified Limitations

- ⚠️ **Dataset Size**: 891 samples for 29 features (30:1 ratio)
- ⚠️ **Missing Data**: 19.9% Age imputation may introduce bias
- ⚠️ **Biased Survivors**: Only passengers in lifeboats, not water victims
- ⚠️ **SVM Interpretability**: RBF kernel complicates direct explanation

Potential Improvements

- **Ensemble Methods**: Combine multiple models for greater robustness
- **External Data**: Incorporate cabin location information
- **Temporal Analysis**: Consider exact sinking timeline
- **Cross-Validation**: Temporal if data from other shipwrecks becomes available

---

🎯 Conclusions

Academic Objectives ✅

- **Performance**: 84.4% accuracy > 80% target
- **Methodology**: Complete CRISP-DM application
- **Algorithms**: Systematic comparison of 4 ML techniques
- **Documentation**: Complete development log with challenges

Historical Contribution 🏛️

- **Quantitative Validation**: "Women and children first" protocol statistically confirmed
- **Intersectionality**: Gender + Social class as critical determining factor
- **Exceptional Cases**: Model errors reveal extraordinary human stories
- **Modern Lessons**: Insights applicable to current emergency protocols

Technical Impact 🔬

- **Feature Engineering**: 29 created variables capture complex historical patterns
- **Error Analysis**: 16.2% errors provide insights about the disaster's chaotic nature
- **Reproducibility**: Complete pipeline available for future research
- **Methodology**: Framework applicable to analysis of other historical events

---

📚 References and Resources

Data Sources

- Kaggle Titanic Dataset
- Encyclopedia Titanica (historical validation)
- British Board of Trade Report (1912)

Technical Methodology

- James, G. et al. "An Introduction to Statistical Learning"
- Castillo, G. "ML-Practical.pdf" (course document)
- Scikit-learn documentation

Historical Context

- Lord, Walter. "A Night to Remember" (1955)
- Official British disaster investigation (1912)
- Naval historical archives

Final Technical Report
Project SG1_Team3_ML
Universidad Panamericana - May 2025