COMS 4995 Applied Machine Learning Assignment 1: From Dirty Data to Predictive Models

Federico Giorgi fg2617

1 Introduction

The sinking of the Titanic is among the most famous maritime disasters in history. On April 15, 1912, during her inaugural voyage, the widely considered "unsinkable" RMS Titanic struck an iceberg and sank in the North Atlantic. Due to an insufficient number of lifeboats to accommodate all passengers and crew members, 1502 of the 2224 individuals aboard perished. While there was some element of chance involved in survival, historical analyses indicate that certain groups, including women, children, and first class passengers, had significantly higher survival rates. This tragedy, besides its historical resonance, offers a compelling case study for statistical analysis and predictive modelling.

Goal of the project The aim of this assignment is to construct an end-to-end machine learning pipeline that predicts passenger survival. Specifically, the project will:

1. clean and transform the raw dataset, handling missing and noisy values;
2. engineer features that better capture socio-demographic patterns, including family size, passenger titles and transformation of fare prices;
3. train and compare models, including a generative approach, Naïve Bayes, and discriminative approaches, linear regression and its regularised variants;
4. evaluate performance using standard metrics, accuracy, precision, recall, and F1 score, together with visualizations such as confusion matrices and roc curves.

This structure not only highlights the predictive power of different modelling philosophies but also offers insights into the social factors that shaped survival outcomes on the Titanic.

2 Exploratory data analysis

2.1 Univariate data analysis

Survival distribution The training set contains 891 passengers, of whom about 38% survived and 62% did not. The bar chart in Figure 1 clearly shows the class imbalance. The imbalance highlights why accuracy alone is not an appropriate metric: a trivial model that predicts "not survived" for everyone would achieve around 62% accuracy. This motivates the use of complementary metrics such as precision, recall, and F1-score when evaluating model performance.

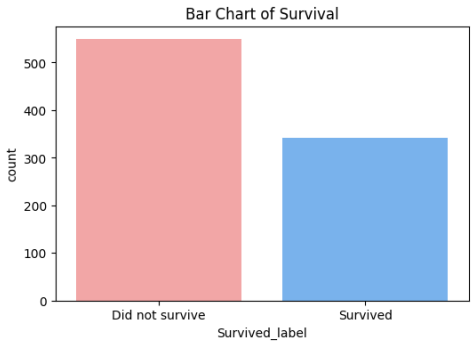

Figure 1: *Distribution of the target variable* `Survived` *across the training set.*

Demographic variables The majority of passengers belonged to third class (55.1%), with far fewer in first (24.2%) or second class (20.7%). Approximately two-thirds (64.8%) of passengers were male and one-third (35.2%) female. `Age` was right-skewed, with most passengers between 20 and 40 years old and fewer elderly travelers. These patterns suggest that socio-economic status and gender might strongly influence survival.
Both `SibSp`, the number of siblings or spouses aboard, and `Parch`, the number of of parents or children aboard, are heavily skewed toward zero: most people traveled alone or with one family member, while large families were rare.

Because the two variables represent similar concepts, it is natural to combine them into a single family size feature later on.
The `ticket` number variable obviously does not provide meaningful information. Its distribution is highly fragmented, with most values unique or appearing only a handful of times, making it effectively an identifier rather than a useful predictive feature.
Ticket `fare` is highly skewed, with a long tail of very expensive tickets. Applying a logarithmic transformation improves symmetry and facilitates modelling. The `cabin` variable has a very high proportion of missing values, with over 77% of entries unavailable. Among the non-missing cases, the distribution across decks is uneven, with decks C and B most represented Port of embarkation is dominated by Southampton (about 72%), with smaller fractions from Cherbourg and Queenstown, indicating another potential proxy for socio-economic status.

2.2 Multivariate Data Analysis

Figure 2 shows the correlation matrix among the main numerical variables. The heatmap provides a compact overview of how socio-economic indicators, demographic features, and survival are interrelated, serving as a useful guide for subsequent feature engineering and model design.

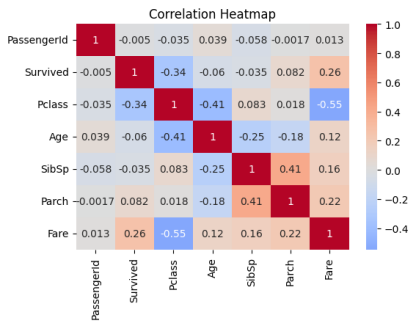

Figure 2: *Correlation heatmap of numerical features.*

Survival vs. predictors Survival varies substantially across different groups. First-class passengers enjoyed survival rates over 60%, second-class around 50% and third-class under 30%. Women survived at much higher rates than men; more than 70% of females lived whereas only about 20% of males did. Children under ten had better survival than adults, while the probability of survival decreased for older age groups (Figure 3).

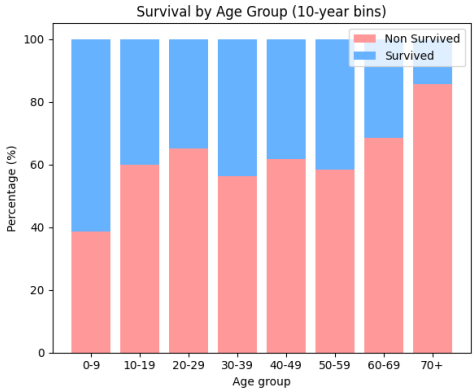

Figure 3: *Age-survival relationship: survival percentages within 10-year age bins.*

Passengers traveling alone or with very large families had lower survival, whereas those with one or two family members fared better. Ticket fare also played a role: survivors generally paid higher fares on average, as shown in Figure 4, consistent with the advantage of passengers in higher socio-economic classes.

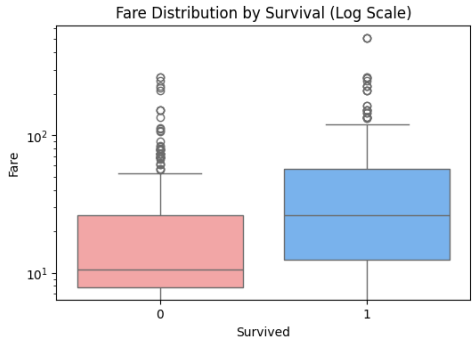

Figure 4: *Distribution of ticket fare by survival status, displayed on a logarithmic scale.*

Larger scope analysis Beyond the effect of individual predictors, it is crucial to study how variables interact with one another and jointly shape survival outcomes. Socio-economic status, for instance, is reflected both in passenger class and in ticket fare (Figure 5), while demographic factors such as gender and age strongly mediate survival chances.

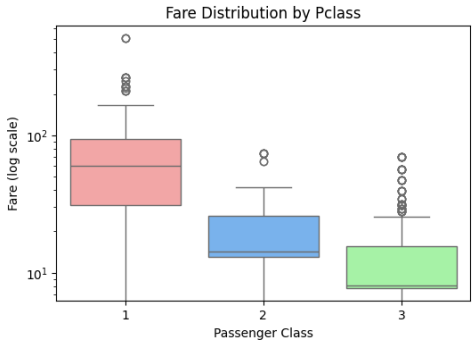

Figure 5: *Relationship between ticket fare and passenger class, highlighting higher fares in first class.*

When combined, these dimensions reveal systematic patterns: women in higher classes enjoyed almost universal survival, whereas men in third class had very low probabilities of survival (Figure 6).

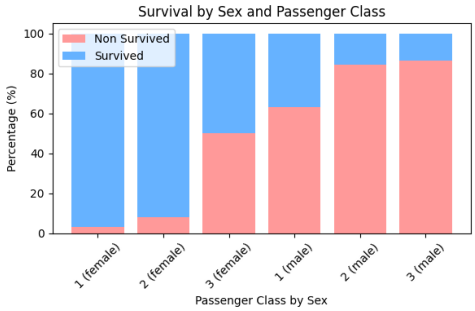

Figure 6: *Survival distribution jointly by passenger class and sex.*

This highlights how multiple features intertwine, showing that survival cannot be explained by a single attribute but rather by the interaction of socio-economic position, gender, and other demographic variables.

3 Data cleaning

Missing values Missing values are common in the Titanic data and must be handled carefully to avoid introducing bias. The main pre-processing steps were as follows:

- **Cabin**: about 77% of values were missing. Imputing such a large proportion would have introduced excessive noise and unreliable information, so the column was dropped entirely.
- **Age**: moderately affected by missingness. Instead of using a single global median, ages were imputed with the median within each (`Pclass`, `Sex`) group, reflecting the observed pattern that age decreases with passenger class and is generally lower for females.
- **Embarked**: only two missing values, which were replaced with the mode (Southampton), consistent with the majority of passengers.
- **Fare**: in the test set, a few missing values were replaced by the median fare.

After these imputations, the dataset contained no missing values.

Outliers Continuous features were screened for extreme values using the IQR rule. The distribution of `Age` shows only a small number of high values (roughly 65–80 years), which are historically plausible and were therefore retained. By contrast, `Fare` is strongly right-skewed and yields many flagged cases; multivariate checks against `Pclass` indicate these cluster in first class, consistent with luxury tickets rather than data errors. Accordingly, no observations were removed, and any undue influence is handled at the modelling stage if needed.

4 Feature engineering

Effective modelling requires transforming raw variables into more informative features.

Family group We defined *family size* as `SibSp + Parch`, representing the total number of relatives aboard. To capture its non-linear effects, it was categorized into four ordered groups: Alone (0), Small (1–3), Medium (4–6), and Large (7–11). Survival follows a distinct non-linear trend, with passengers traveling with a small family showing the highest chances of survival, while in large groups fared to zero (Figure 7).

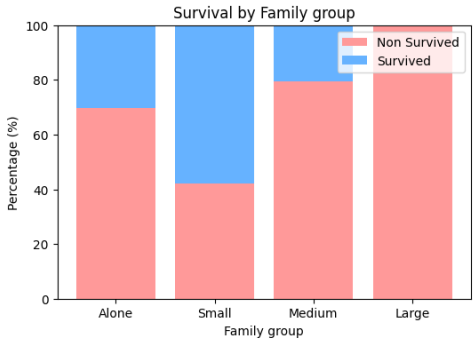

Figure 7: *Survival by family group.*

Standardized log-fare Since the ticket fare distribution was highly skewed, we applied a logarithmic transformation, $\log(\texttt{Fare})$, followed by standardization; reducing the influence of extreme values. As previously observed in Figure 4, survivors generally paid higher fares on average; the transformed variable thus preserves this relationship while mitigating skewness.

Age Bin Given the strong age–survival relationship observed in Figure 3, the binned age variable (`AgeBin`) was retained to preserve this informative structure. This transformation simplifies interpretation by grouping passengers into ordered age categories, enabling clearer comparisons across demographic segments while maintaining the predictive signal observed during exploration.

Title status Passenger names contain titles such as "Mr", "Mrs", "Miss", and "Master", which provide social and demographic information. We extracted these titles and grouped all rare ones under "Other." Titles capture meaningful patterns linked to age and gender: for instance, "Mrs" typically denotes adult married women, whereas "Master" indicates young boys. The survival pattern across titles (Figure 8) reflects known demographic trends: females ("Mrs", "Miss") survived far more often than males ("Mr"), while "Master" shows higher survival among young boys, consistent with the "women and children first" rule. Hence, the title variable effectively acts as an implicit interaction of gender, age, and marital status, enriching the socio-demographic representation.

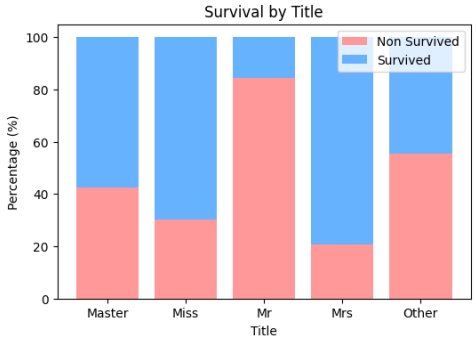

Figure 8: *Survival by title.*

Sex–class interaction Motivated by the clear stratification visible in Figure 6, we introduced an interaction term, `Sex_Pclass`, combining passenger sex and class into a single categorical variable. This allows even simple models to exploit the non-additive relationship between gender and socio-economic status.

5 Model training

The Linear Regression and Naïve Bayes models embody complementary philosophies: Linear Regression captures discriminative decision boundaries, whereas Naïve Bayes models the underlying class-conditional distributions in a generative fashion. They are trained on two distinct representations of the dataset, each tailored to the assumptions and strengths of the respective algorithm.

- For the **Linear Regression** models, categorical variables (`Sex`, `Embarked`, `Title`, `Pclass`, `AgeBin`, and `Family group`) were one-hot encoded with one reference category dropped to avoid perfect collinearity. Additionally, the interaction between `Sex_Pclass` was included, and one-hot encoded, to capture non-linear, non-additive effects that a purely linear model would otherwise fail to represent. Regularised variants (L1 and L2) were later introduced to mitigate overfitting and enhance generalisation.
- For the **Naïve Bayes** models, all one-hot encoded categories were retained, i.e., no baseline level was dropped, since the model treats each feature independently rather than as part of a linear combination. Two variants were trained on different transformations of the same dataset: a **BernoulliNB** model, where all features were binarised to represent the presence or absence of traits, and a **GaussianNB** model, where all features were kept continuous and standardised. Interaction terms were not included, as they contradict the conditional independence assumption underlying Naïve Bayes.

Each model was evaluated through 5-fold cross-validation, computing accuracy, precision, recall, F1 score, and AUC.

5.1 Linear regressions

Although binary outcomes are typically modelled using logistic regression, in this assignment we employed **linear regression** as a baseline classifier, converting continuous predictions into binary outcomes by thresholding at 0.5. The model specification, reported in Equation 1, includes all engineered numerical and categorical predictors discussed earlier.

$$\begin{aligned}\widehat{\text{Survived}}_i = {} & \beta_0 + \beta_1 \,\texttt{LogFare}_i + \beta_2 \,\texttt{Sex\_male}_i \\ & + \sum_{e \in \{Q,S\}} \beta_e \,\texttt{Embarked}_{e,i} \\ & + \sum_{t \in \{Miss,Mr,Mrs,Other\}} \beta_t \,\texttt{Title}_{t,i} \\ & + \sum_{p \in \{2,3\}} \beta_p \,\texttt{Pclass}_{p,i} + \sum_a \beta_a \,\texttt{AgeBin}_{a,i} \\ & + \sum_f \beta_f \,\texttt{FamilyGroup}_{f,i} + \sum_s \beta_s \,\texttt{Sex\_Pclass}_{s,i} + \varepsilon_i \end{aligned} \tag{1}$$

To improve generalisation and mitigate potential multicollinearity, two regularised variants were also trained: **Lasso regression** (L1 penalty) and **Ridge regression** (L2 penalty). Both models were implemented within a `Pipeline` combining feature standardisation and regression. Standardisation is crucial because the magnitude of the L1 and L2 penalties directly depends on the scale of the coefficients; scaling therefore guarantees that all predictors contribute comparably to the regularisation term. The L1 regulariser encourages sparsity, driving some coefficients exactly to zero and thus performing implicit feature selection, whereas the L2 regulariser distributes shrinkage more smoothly across correlated predictors, retaining all features but reducing their influence. For both regularised models, a grid of values of the regularisation strength α was explored, and the configuration yielding the highest F1 score was selected.

5.2 Naïve Bayes

Two Naïve Bayes classifiers were implemented to explore generative approaches under different feature assumptions. In contrast to linear models, Naïve Bayes estimates class-conditional feature distributions and applies Bayes' theorem to infer posterior probabilities. This approach relies on the strong assumption of conditional independence between predictors given the target class, which often does not hold in socio-demographic data but provides a useful baseline for comparison.

Bernoulli Naïve Bayes. Given that most engineered variables are binary (one-hot encodings of sex, class, title, age bins, family group, and embarkation port), BernoulliNB provides a natural fit. Continuous variables such as `LogFare` were discretised into two quantile-based bins using a `KBinsDiscretizer`, ensuring compatibility with binary features. The model was trained through a `Pipeline` combining this discretisation step with the Naïve Bayes estimator. To account for unseen feature–class combinations, several degrees of Laplace smoothing were tested.

Gaussian Naïve Bayes. A GaussianNB classifier was trained on continuous, standardised features to model class-conditional Gaussian densities. Categorical predictors were converted into continuous numerical representations using a `TargetEncoder`, which replaces each category with the smoothed mean survival rate observed in the training data. The resulting dataset was standardised within a `Pipeline` and passed to the Gaussian model, ensuring that all feature distributions were on comparable scales. Although this representation allowed GaussianNB to exploit continuous variation, its independence and normality assumptions remained severely violated.

6 Model comparison and evaluation

6.1 Overview of results

Table 1 reports the main evaluation metrics for the final models trained and validated on the held-out subset. Across all configurations, performance differences were moderate, with regularised linear models consistently outperforming Naïve Bayes classifiers.

Model	Accuracy	Precision	Recall	F1	AUC
Linear Regression	0.835	0.824	0.725	0.771	0.878
L1 Regression	0.835	0.806	0.751	0.778	0.867
L2 Regression	0.835	0.822	0.728	0.772	0.877
Bernoulli NB	0.799	0.742	0.731	0.736	0.846
Gaussian NB	0.696	0.670	0.409	0.508	0.744

Table 1: *Model performance comparison on the validation data.*

Accuracy Accuracy measures the proportion of correctly classified passengers over the total. Given the mild imbalance between survivors (38%) and non-survivors (62%), accuracy alone can be misleading, since a trivial classifier predicting all passengers as "not survived" would already achieve 0.62. All linear models reached 0.835 accuracy, Bernoulli NB followed with 0.799, while Gaussian NB lagged behind at 0.696.

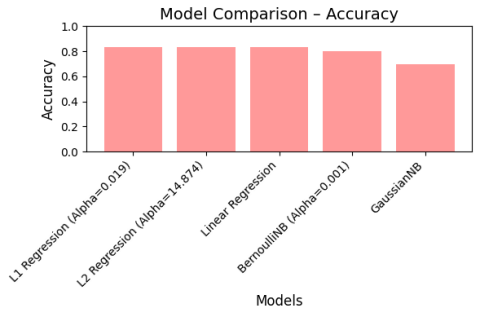

Figure 9: *Model comparison in terms of accuracy.*

Precision Precision quantifies the proportion of predicted survivors who actually survived. Linear Regression achieved the highest precision (0.824), followed closely by Ridge (0.822) and Lasso (0.806) regressions, Bernoulli NB obtained a slightly lower precision (0.742), while Gaussian NB fell to 0.670.

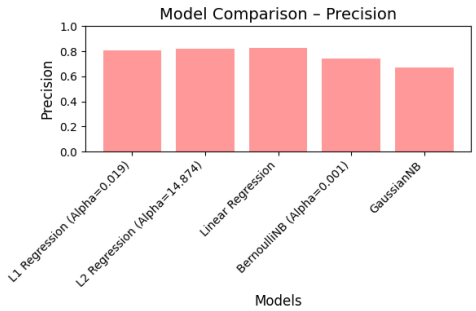

Figure 10: *Model comparison in terms of precision.*

Recall Recall measures the proportion of actual survivors correctly identified and captures model sensitivity to the minority class. Lasso regression achieved the best recall (0.751), followed by Ridge (0.728) and Linear regression (0.725). Again, Bernoulli NB performed similarly (0.731), whereas Gaussian NB achieved merely 0.409.

Figure 11: *Model comparison in terms of recall.*

F1 score The F1 score captures the trade-off between false positives and false negatives and it is particularly suitable for datasets with modest class imbalance like this one. Lasso regression achieved the highest F1 (0.778), followed closely by Ridge (0.772) and Linear regression (0.771), Bernoulli NB obtained 0.736, whereas Gaussian NB dropped to 0.508.

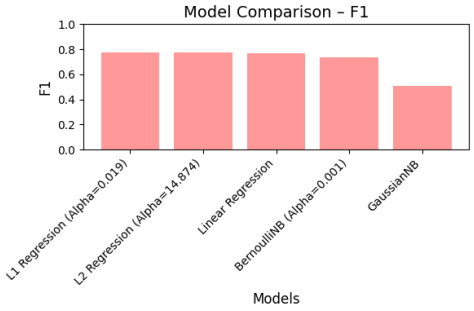

Figure 12: *Model comparison in terms of F1 score.*

AUC The Area Under the ROC Curve evaluates the model's ranking ability across all thresholds, independent of any fixed decision rule. AUC confirms the strong discriminative capacity of linear models.

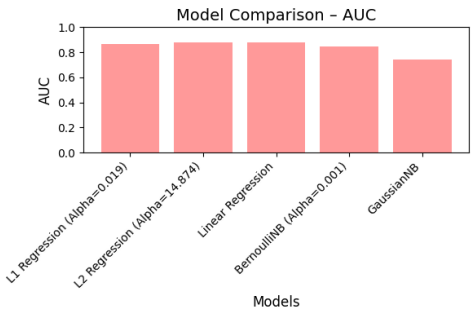

Figure 13: *Model comparison in terms of AUC.*

Confusion matrices Confusion matrices show the distribution of true positives, false positives, true negatives, and false negatives, enabling direct inspection of each model's misclassification pattern. Linear models displayed a balanced distribution between false positives and false negatives, while Gaussian NB exhibited a marked bias toward predicting the positive class. The detailed confusion matrices are provided in Appendix.

ROC curves Receiver Operating Characteristic (ROC) plots confirm the numerical AUC results: the linear models dominate the upper-left region, Bernoulli NB follows closely, and Gaussian NB lies near the diagonal, indicating near-random discrimination. All ROC curves are reported in the Appendix.

6.2 Interpretation of coefficients

The baseline **linear regression** model, performing on par with its regularised counterparts, provides interpretable coefficients. Only the *signs* are discussed, as magnitudes vary with feature scales, but their directions align with expected socio-demographic survival trends. The complete list of estimated coefficients appear in the Appendix.

- `Sex_male` shows a large negative coefficient, confirming that being male significantly reduced survival probability.
- Titles such as `Miss` and `Mrs` have strong positive effects, consistent with the prioritisation of women and children during evacuation.
- `Pclass_2` and `Pclass_3` coefficients are negative, while `LogFare` is positive, wealthier, first-class passengers enjoyed higher survival chances.
- Moderate family groups (`Small`, `Medium`) contribute positively, whereas large families reduce survival odds.
- Passengers embarking from Cherbourg (`Embarked_C`) show a mild positive effect, reflecting that port's concentration of first-class travellers.

Regularisation strength. Figure 14 shows the cross-validated F1 scores for L1 and L2 regularisation across a logarithmic grid of α values. Both regressions remained stable over a wide range of regularisation strengths, with performance deteriorating only under excessively large penalties. This behaviour suggests that the extensive feature engineering and careful encoding performed earlier effectively captured the main predictive patterns, leaving little need for strong regularisation.

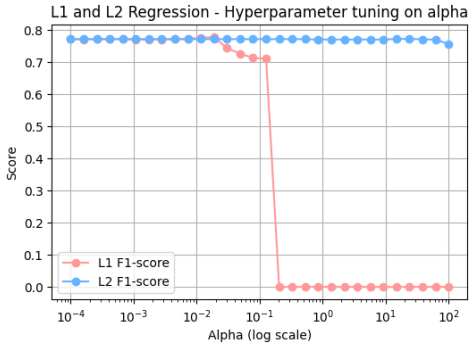

Figure 14: *Cross-validated F1 scores for L1 (Lasso) and L2 (Ridge) regression across different regularisation strengths α.*

6.3 Effect of Smoothing

Figure 15 shows that the F1-score is remarkably stable across several orders of magnitude of the smoothing parameter α. Values between 10^{-3} and 10^{-1} yield almost identical performance and only for extreme regularization ($\alpha \gg 1$) does the score slightly deteriorate, as probability estimates become overly uniform. Overall, smoothing has a negligible impact on performance but improves robustness to rare feature combinations.

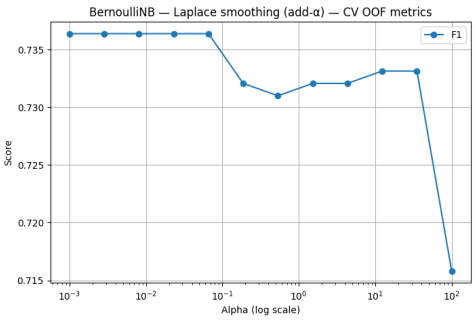

Figure 15: *Effect of Laplace smoothing (`add-α`) on Bernoulli Naïve Bayes cross-validated F1-score.*

7 Discussion

The Titanic dataset highlights how the structure and interdependence of socio-demographic variables strongly influence model choice and performance. Two main conclusions emerge from the analysis.

Strong feature correlations shape survival prediction. The main predictors, `Sex`, `Age`, `Fare`, and `Pclass`, are deeply interconnected. Female passengers were on average younger and more likely to travel in higher classes, which in turn corresponded to higher ticket fares and better access to lifeboats. Similarly, `Pclass` and `Fare` jointly capture socio-economic status, while titles extracted from names correlate with both gender and age. These structural dependencies create clusters of correlated variables that together drive survival outcomes. As a result, the dataset violates the conditional-independence assumption required by Naïve Bayes.

Linear models outperform generative ones under correlated features. Because linear regression directly learns discriminative decision boundaries, it can exploit correlations among predictors. Regularised variants such as ridge and lasso regression further stabilise coefficient estimates, achieving superior performances without relying on unrealistic independence assumptions. In contrast, Naïve Bayes treats each feature as statistically independent, leading to oversimplified posterior probabilities and reduced predictive reliability when predictors are highly collinear.

8 AI tool usage disclosure

Artificial intelligence played an extensive yet carefully supervised role throughout the development of this project. I used ChatGPT to accelerate workflow, verify understanding, and improve the quality of both code and written explanations. The tool was particularly helpful for recalling syntactic details of Python and LaTeX, for instance, specific plotting functions or formatting commands, that would otherwise require lengthy documentation searches. This allowed me to focus my effort on conceptual reasoning rather than on mechanical coding recall.
I also employed the AI to clarify some mathematical assumptions, to test alternative ways of presenting statistical ideas, and to refine the narrative structure of the report. **All generated material was carefully reviewed, verified, and edited**: every formula, paragraph, and code block was checked for correctness and rewritten where necessary.

9 Appendix

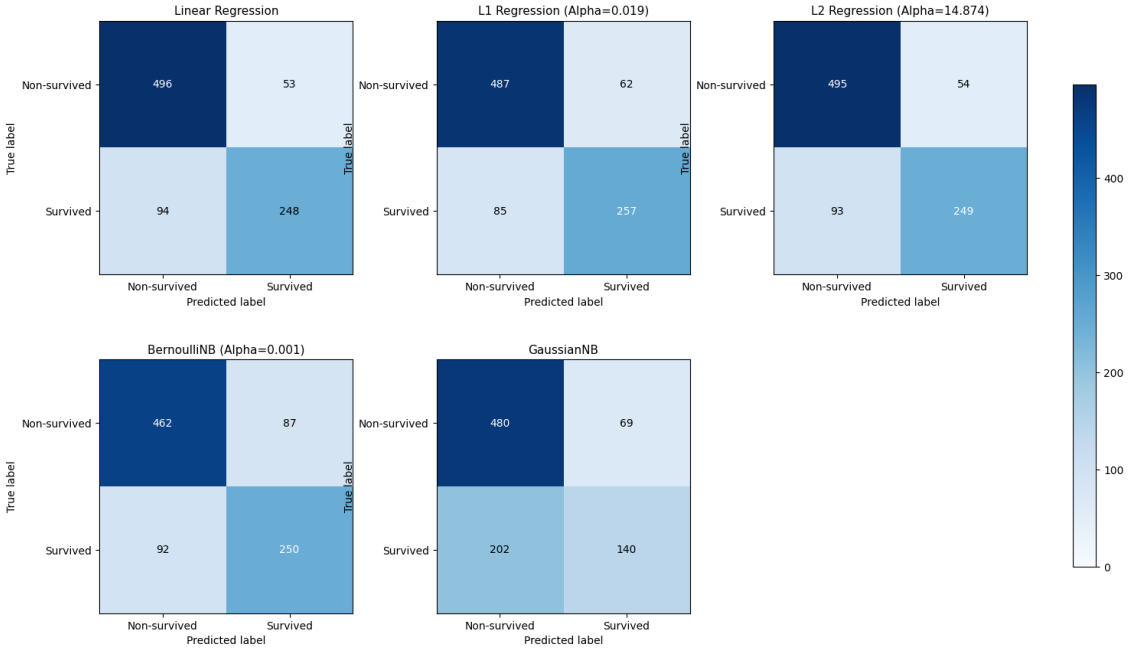

Figure 16: *Confusion matrices for all models on the validation data.*

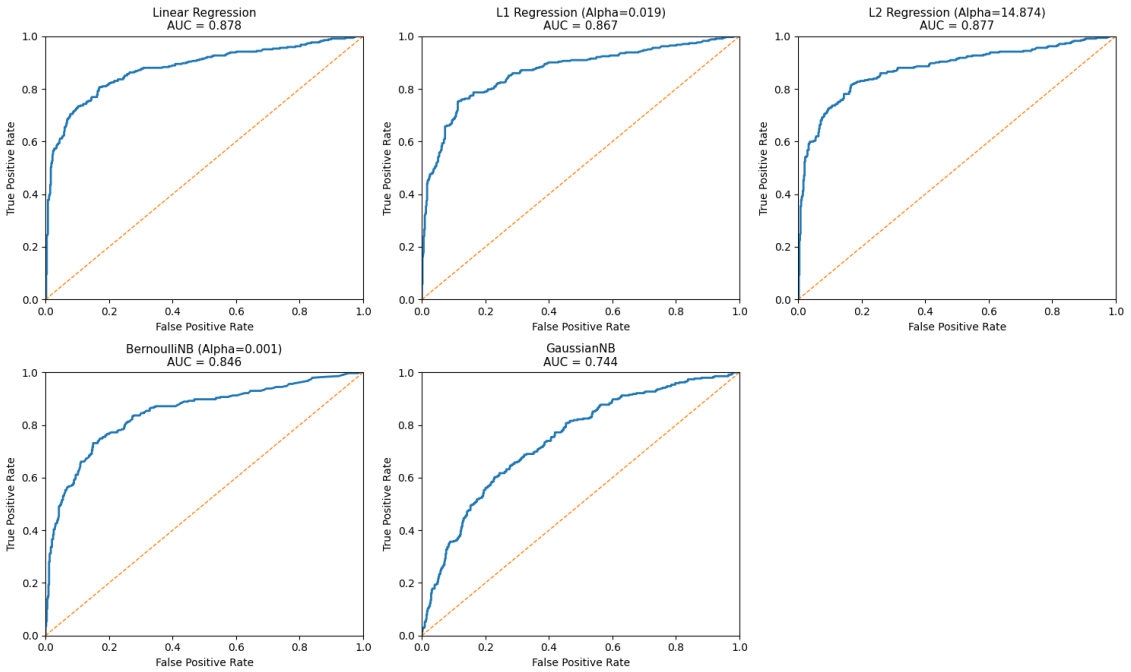

Figure 17: *Receiver Operating Characteristic (ROC) curves for all models.*

Feature	Coef_mean	Coef_std
LogFare	0.043	0.008
Sex_male	-0.446	0.040
Embarked_Q	0.008	0.025
Embarked_S	-0.052	0.011
Title_Miss	-0.552	0.077
Title_Mr	-0.547	0.031
Title_Mrs	-0.512	0.072
Title_Other	-0.509	0.075
Pclass_2	-0.112	0.006
Pclass_3	-0.223	0.010
AgeBin_10–19	-0.081	0.019
AgeBin_20–29	-0.104	0.033
AgeBin_30–39	-0.074	0.040
AgeBin_40–49	-0.179	0.023
AgeBin_50–59	-0.214	0.054
AgeBin_60–69	-0.194	0.036
AgeBin_70+	-0.242	0.082
Family group_Small	-0.040	0.012
Family group_Medium	-0.477	0.023
Family group_Large	-0.544	0.034
Sex_Pclass_female_2	0.123	0.008
Sex_Pclass_female_3	-0.119	0.020
Sex_Pclass_male_1	-0.106	0.022
Sex_Pclass_male_2	-0.235	0.009
Sex_Pclass_male_3	-0.104	0.021

Table 2: *Estimated coefficients from the linear regression model with cross-validated mean and standard deviation.*