# Technical Report

## LendSmart

Sofia Knutas, A01831481
Kristina Kristiansen Stapnes, A01830792

Aplicación de métodos multivariados en ciencia de datos (Gpo 601)
Group 6

## 1. Objective

The aim of this analysis was to develop and evaluate statistical classification models that accurately predict loan default risk using applicant financial and demographic characteristics, enabling improved risk assessment and reduction of default rates.

## 2. Data Structure and Initial Assessment

The dataset consists of 2,500 observations and 18 variables. The target variable loan_status is binary, indicating default (1) or non-default (0). The dataset contains no missing values, allowing direct modeling without imputation. Summary statistics confirmed appropriate value ranges and no structural inconsistencies. The observed default rate is 28%, meaning the classification task involves moderate imbalance but remains suitable for standard discriminant methods.

## 3. Exploratory Data Analysis

Univariate distribution plots showed that several financial predictors, including loan_amount and savings_ratio, exhibit skewness and heavy tails, indicating departures from normality. Credit score and job stability metrics were more symmetrically distributed. Categorical predictors such as education_level and marital_status displayed measurable variation in default proportions. A correlation heatmap revealed strong relationships between loan_status and variables related to repayment behavior, credit utilization, and financial ratios. These patterns indicate informative predictor structure suitable for classification modeling.

## 4. Data Preparation and Encoding

Categorical predictors were converted using one-hot encoding. The predictor matrix X and target vector y were defined, excluding identifier and date variables. The dataset was split into training and testing subsets using an 80/20 stratified partition to preserve class distribution. Standardization using z-score scaling was applied based on the training subset and transferred to the test subset to maintain consistency across variables and ensure suitability for discriminant analysis.

## 5. Statistical Assumption Considerations

Linear Discriminant Analysis (LDA) assumes multivariate normality within each class and equality of covariance matrices across groups, producing linear separation boundaries. Quadratic Discriminant Analysis (QDA) relaxes the equal-covariance assumption, allowing distinct covariance structures and generating quadratic boundaries. Given the observed skewness in several predictors, the normality assumption is only partially met. This suggests that QDA could theoretically capture more complex separation structures, but empirical validation is required to determine performance differences.

## 6. Linear Discriminant Analysis (LDA) Model Results

The LDA model was trained on the standardized data. Examination of coefficient magnitudes showed that variables related to financial pressure—credit utilization, debt-to-income ratio—and repayment reliability measures contributed most to class discrimination. Model evaluation on the test set demonstrated high accuracy and strong separation capacity, with balanced performance across default and non-default classes.

## 7. Quadratic Discriminant Analysis (QDA) Model Results

A QDA model was fitted using the same scaled predictors. Allowing group-specific covariance matrices did not yield a performance advantage. The model achieved similarly high classification accuracy. The ROC curve and AUC metric were comparable to LDA, indicating that the additional flexibility of quadratic boundaries did not translate into improved predictive ability in this dataset.

## 8. Performance Comparison and Interpretation

Both models achieved high accuracy and strong classification metrics when evaluated on the unseen test data. Confusion matrices showed low false-positive and false-negative rates. ROC curves for both models exhibited strong separation, with AUC scores indicating excellent discriminative power. The lack of improvement from QDA suggests that covariance structures between classes are similar enough for linear boundaries to perform equivalently.

## 9. Technical Conclusion

Both LDA and QDA are effective for predicting loan default within this dataset. However, LDA offers equivalent predictive accuracy with a simpler linear decision boundary, reduced parameter complexity, and greater numerical stability. Therefore, LDA is the more technically appropriate model for operational deployment and further analytical extensions. Future improvements may include dimensionality reduction, feature selection diagnostics, and evaluation using additional classification frameworks such as logistic regression, SVM, or tree-based models.