Guided Activity on Titanic Survivors Classification

Start Assignment

- Due Sep 24 by 11:59pm
- Points 100
- Submitting a file upload
- File Types pdf
- Available Sep 9 at 12am - Dec 20 at 11:59pm

Objective:

In this activity, you will use three classification techniques—Logistic Regression, Linear Discriminant Analysis (LDA), and Naive Bayes—to analyze the **Titanic dataset** (https://www.kaggle.com/competitions/titanic). The goal is to determine which factors contributed most to survival and compare the performance of each classifier.

1. Understanding the Dataset

The Titanic dataset contains information about passengers, such as their age, gender, ticket class, and whether they survived the Titanic disaster.

- **Target Variable:** `Survived` (1 = Survived, 0 = Did not survive)
- **Predictor Variables** (selected for simplicity, you can use all or do your own selection):
  - `Pclass`: Passenger class (1 = 1st, 2 = 2nd, 3 = 3rd)
  - `Sex`: Gender (Male, Female)
  - `Age`: Age in years
  - `Fare`: Passenger fare
  - `SibSp`: Number of siblings/spouses aboard
  - `Parch`: Number of parents/children aboard

2. Data Preprocessing

Before applying the classifiers, you will need to:

- Handle missing values in the `Age` column (e.g., impute the mean).
- Convert categorical variables (`Sex` and `Pclass`) into numerical format (e.g., one-hot encoding for `Sex`).
- Split the data into training and test sets (e.g., 80% training, 20% test).

3. Logistic Regression

- **Task 1:** Fit a logistic regression model using the predictors `Pclass`, `Sex`, `Age`, `Fare`, `SibSp`, and `Parch`.

- **Task 2:** Interpret the coefficients of the logistic regression model. Which variables have the largest impact on survival?
- **Task 3:** Calculate the accuracy, precision, recall, and F1-score for the logistic regression model on the test set.

4. Linear Discriminant Analysis (LDA) or Quadratic Discriminant Analysis (QDA)

- **Task 4:** Fit an LDA (or QDA) model using the same predictors.
- **Task 5:** Compare the LDA model's accuracy with the logistic regression model. Is there any difference in the results?
- **Task 6:** Examine the decision boundaries formed by LDA/QDA and compare them to the logistic regression decision boundaries.

5. Naive Bayes

- **Task 7:** Fit a Gaussian Naive Bayes model using the same predictors.
- **Task 8:** Evaluate the performance of the Naive Bayes classifier and compare it to the logistic regression and LDA models.
- **Task 9:** Interpret the results of the Naive Bayes model. Does it perform better with certain predictors?

6. Model Comparison

- **Task 10:** Compare the performance metrics (accuracy, precision, recall, F1-score) of all three models (logistic regression, LDA/QDA, and Naive Bayes). Which model performs the best in terms of overall accuracy? Which one performs better on specific classes (e.g., survivors vs. non-survivors)?

7. Conclusion

- **Task 11:** Based on your findings, what variables are most important in predicting survival on the Titanic? Which model provides the most reliable results?