**Project 3 Loan Prediction - Predict if a loan will get approved or not**

https://www.kaggle.com/datasets/altruistdelhite04/loan-prediction-problem-dataset

In the banking system, banks have many products to sell but the main source of income of any banks is on its credit line. So, they can earn from interest of those loans which they credit. The bank's profit or a loss depends largely on loans, i.e., whether the customers are paying back the loan or defaulting. By predicting the loan defaulters, the bank can reduce its Non- Performing Assets. This makes the study of this phenomenon very important. Previous research in this era has shown that there are so many methods to study the problem of controlling loan defaults. But as the right predictions are very important for the maximization of profits, it is essential to study the nature of the different methods and their comparison. A very important approach in predictive analytics is used to study the problem of predicting loan defaulters: The Logistic regression model. The data is collected from the Kaggle for studying and prediction. Logistic Regression models have been performed and the different measures of performances are computed. The models are compared on the basis of performance measures such as sensitivity and specificity. The results have shown that the model produces different results. Model is marginally better because it includes variables (personal attributes of customer like age, purpose, credit history, credit amount, credit duration, etc.) other than checking account information (which shows wealth of a customer) that should be taken into account to calculate the probability of default on loan correctly. Therefore, by using a logistic regression approach, the right customers to be targeted for granting loan can be easily detected by evaluating their likelihood of default on loan. The model concludes that a bank should not only target the rich customers for granting loans, but it should assess the other attributes of a customer as well which play a very important part in credit granting decisions and predicting the loan defaulters. The results of the different machine learning algorithms are based on accuracy, precision, recall, and F1-score. The ROC curve needs to be plotted based on the confusion matrix.

**Team Members:**
Name: **** (Role: Team Leader)
Name: **** (Role: Feature and Label Analysis, Preprocessing)
Name: **** (Role: Algorithm Design)
Name: **** (Role: Implementation)
Name: **** (Role: Documentation)

**Additional Requirements:**

In addition to report the above specific steps of the machine learning pipeline, the team project also needs to specify which machine learning techniques or deep learning method is used or developed to predict the loan’s approval or decline decisions with high accuracy, precision, recall and AUC scores? What are the state-of-the-art

machine learning based techniques and algorithms? How can you develop a new method or algorithm to improve the prediction performance or make it efficient? How do you quickly detect these characteristics in your model so that you can ensure a high prediction performance? In order to answer these technical questions, you need to

(1) Briefly summarize the existing breakthrough in this topic and summarize the challenges facing in existing studies.
(2) Draw an overall flowchart to show how your model works on such kind of data.
(3) Report the average performance over different fold in the test sets using cross validation.
(4) Report the confusion matrix and AUC curves and scores with different fold of cross validation.
(5) Draw the three comparative curves of training, validation, and test with different iterations or epochs, and show how do you address underfitting and overfitting problems.
(6) Report the running time of your model or algorithm.
(7) Summarize how your team worked together and make this team project complete, what are the efforts of each teammate.
(8) Summarize your contributions to existing studies.