Taiyō
Machine Learning Assignment
Natural Language Understanding

By Remita Austin

PROBLEM STATEMENT

Use World Bank Projects dataset.

Option 2. Binary Classifier. Using the status variable build a binary classifier to predict the probability whether a project will be “closed” or “canceled/distressed”.

DATASET

The dataset that has been used for the time series analysis has been obtained from World Bank Projects. It has information on all of the World Bank's lending projects from 1947 to the present. The dataset includes basic information such as the project title, task manager, country, project id, sector, themes, commitment amount, product line, and financing. It also provides links to publicly disclosed online documents.

Link to Dataset: https://datacatalog.worldbank.org/search/dataset/0037800

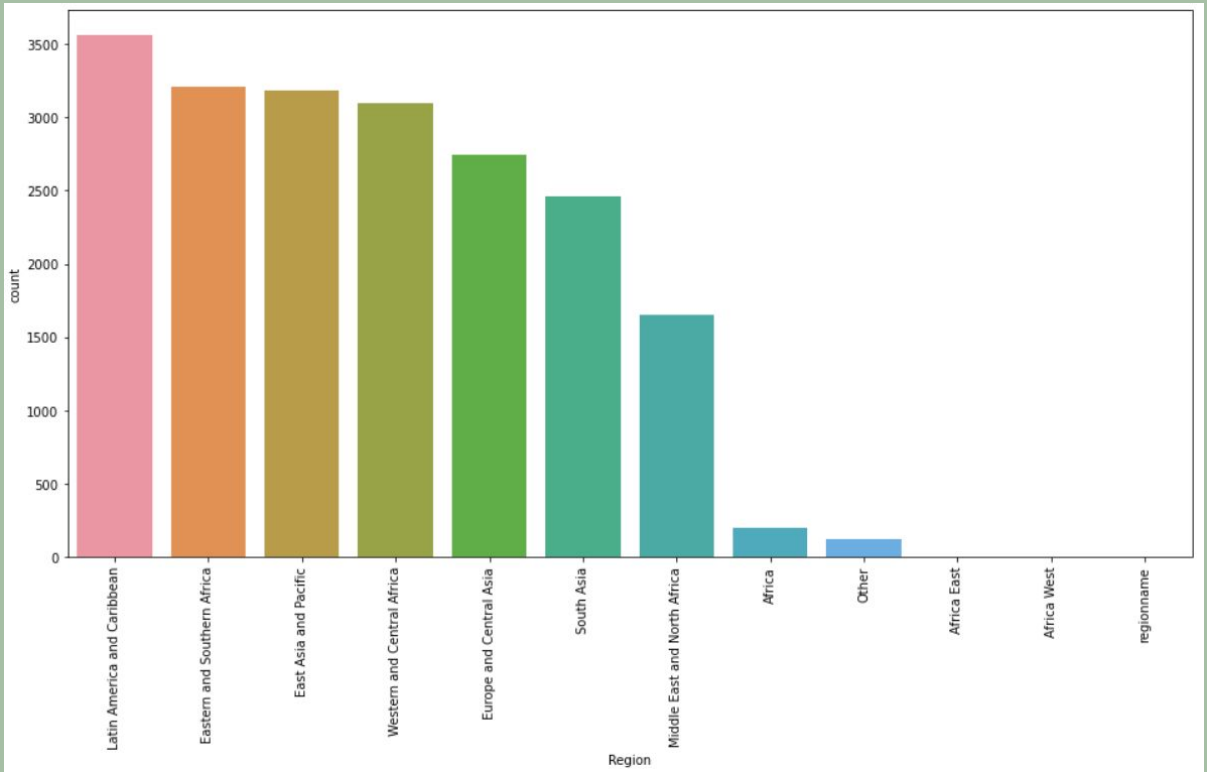
IMPLEMENTATION

1. **Data Pre-processing**

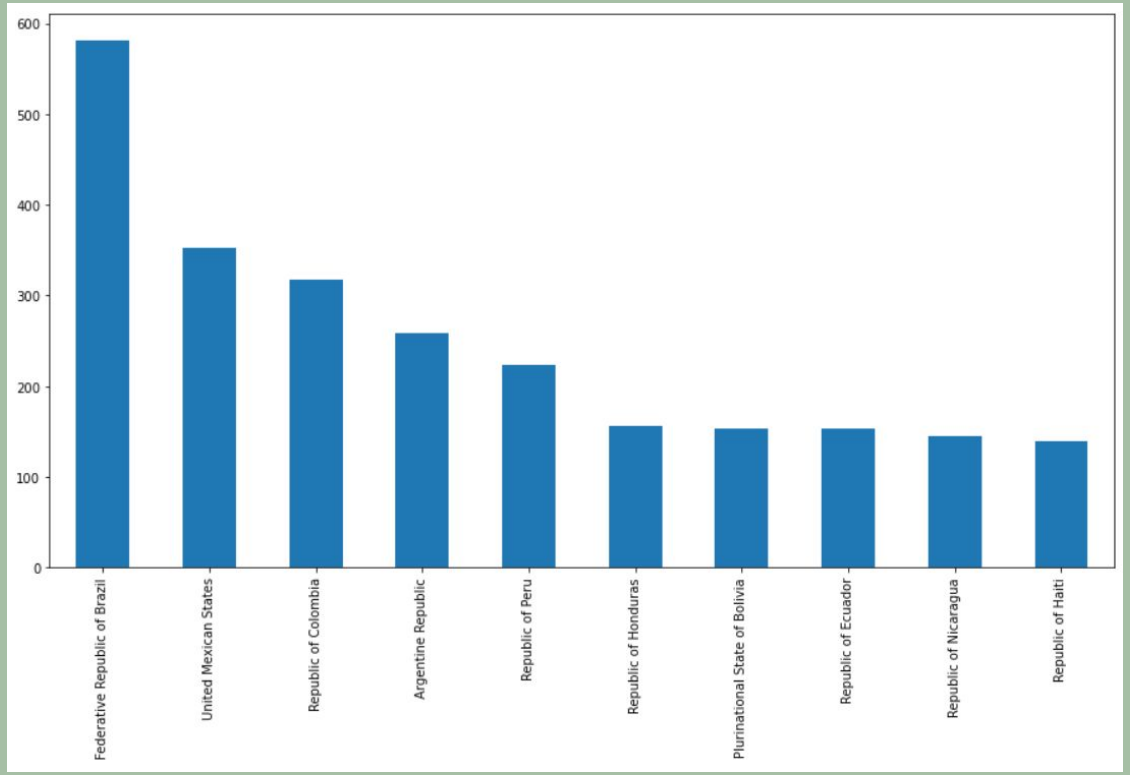
   The Projects dataset was imported from World Bank Projects and loaded. We checked if the dataset contained missing/null values. The dataset was cleaned by removing NA values, and dropping irrelevant fields and records with Project Status=Pipeline since this is not required for the given problem statement.

2. **Exploratory Analysis**

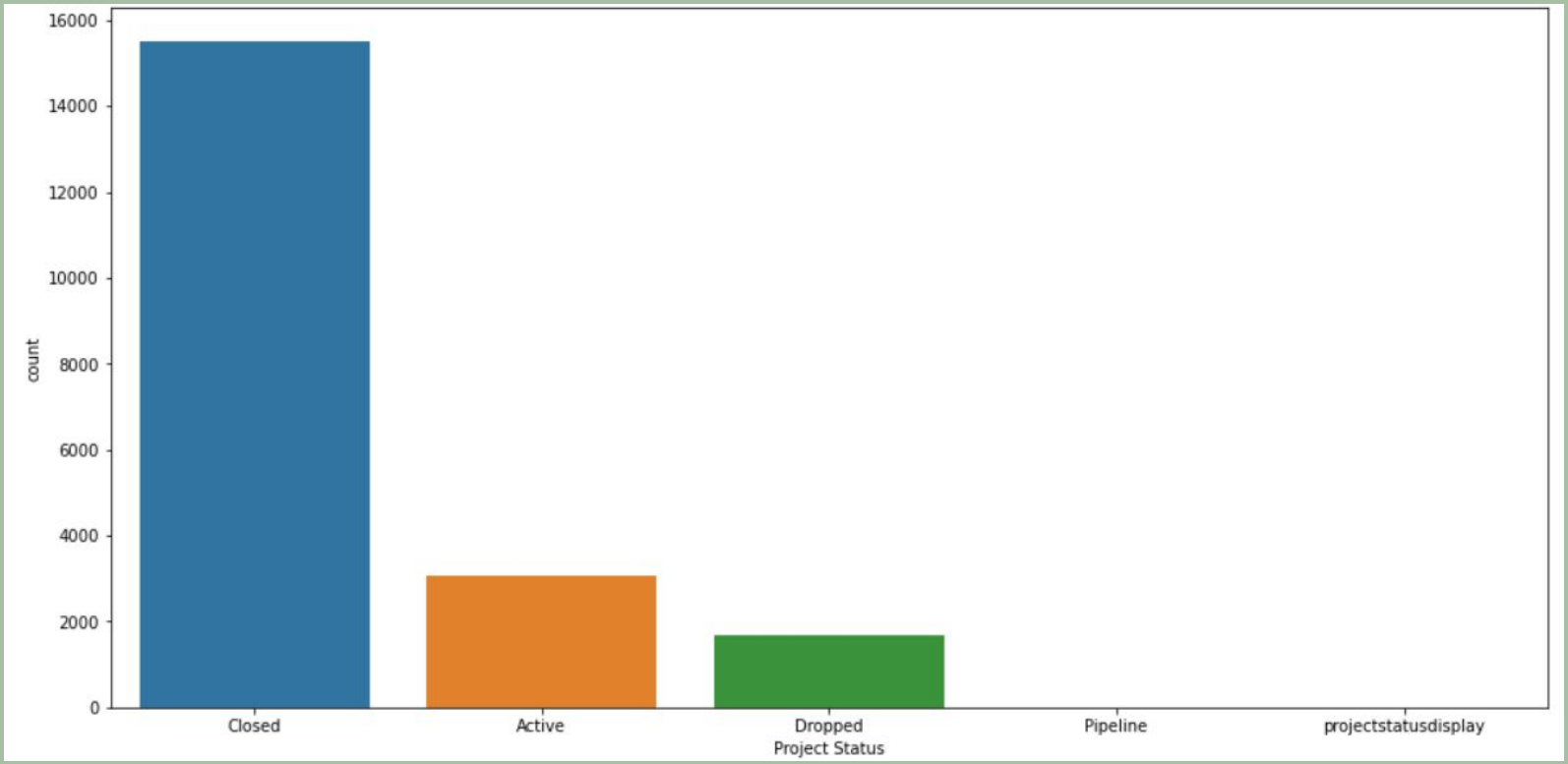
   We need to observe and analyze the data to see what we are working with. For this, the matplotlib and seaborn libraries were used and the data was visualized and inferences were observed.



Most of the projects have been implemented in Latin America and Caribbean region.



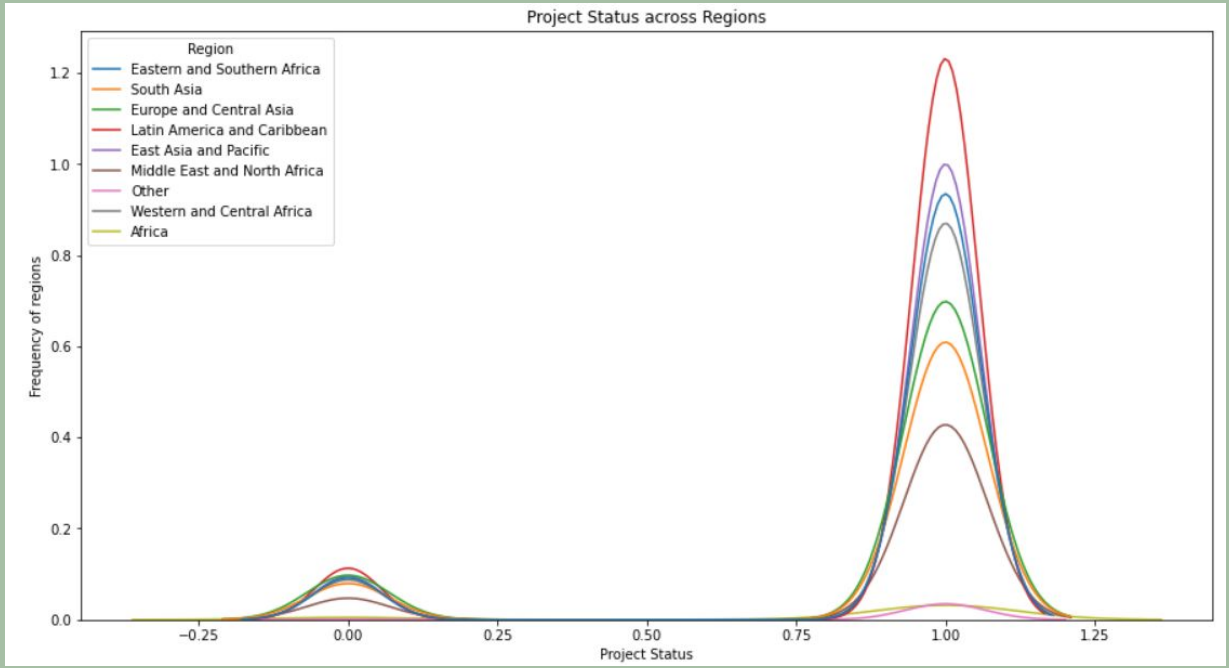
Here the top ten countries of Latin America and Caribbean region with the most number of projects are displayed.



Most of the projects have their status as Closed. Few projects are Active and even fewer projects have been Dropped.

3. Split the data wrt Project Status

The dataset was split such that active projects consisted of those whose Project Status=Active and closed/dropped projects consisted of those whose Project Status!=Active. Then the relevant features were selected and dummy variables were given for categorical data for ease of prediction.

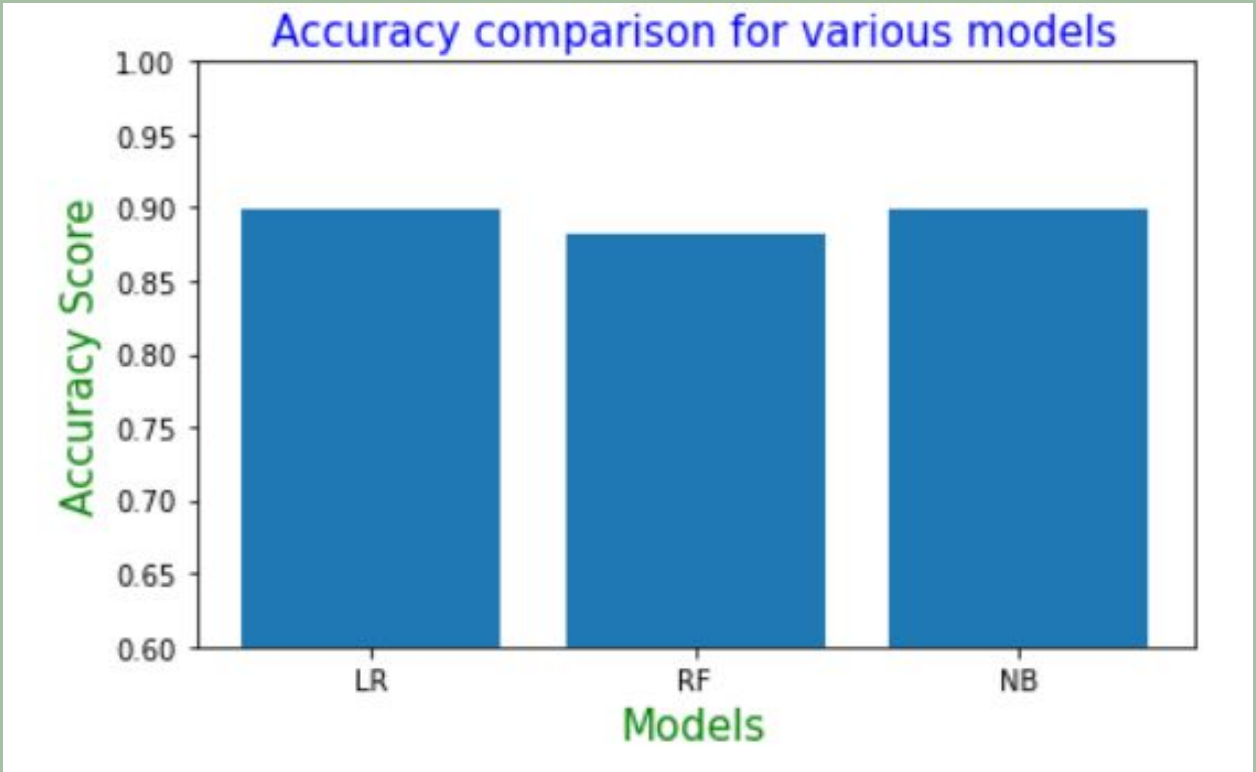


4. Split the Data into Training Data and Test Data

The next step is to split the data into training and testing data to train our machine learning algorithm with the training data and evaluate the model's performance using the testing data.

5. Performance Evaluation

To find the best model for our training data, we've taken three algorithms- Logistic Regression, Random Forest, and Naive Bayes Classifiers. We then computed the accuracy scores of these models and visualized them using a barplot. We inferred that the Logistic Regression and Naive Bayes models gave the highest accuracy of 89.8%. Then we did predictions based on the Logistic Regression model over a sample input data to check if the project status is closed or canceled/distressed.



From the barplot, it is clear that LR and NB shows highest accuracy score, and RF shows least accuracy score.

CONCLUSION

The World Bank Projects data was analyzed, and the data was cleaned and modified in order to train the model using Logistic Regression, Random Forest, and Naive Bayes algorithms. We then computed the accuracy scores of these models and visualized them using a barplot. The LR and NB models gave the highest accuracy. Then the project status of sample input data was predicted using the LR model.