Project 3

Team 4:

Mohanned Alsahaf

Abdullah Alhuthaily

Razan Alhussainan

Walaa Almajnuni

Abdulrahman Al-Abbas

Project Idea

Dataset Information

Before

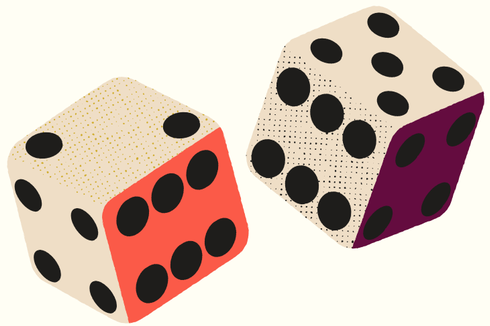
- Name: Measuring Student Persistence and Completion Rate
- Files: Registration – Train – Test – Sample Submission
- Total Records (Train & Test): 7368 Rows
- Total Columns: 24

After

- Name: Measuring Student Persistence and Completion Rate
- Files: Registration – Train – Test – Sample Submission
- Total Records (Train & Test): 6664 Rows
- Total Columns: 24

EDA Summary

- Renamed columns for misspelling
- Unified input values (e.g., "ابها" to "أبها")
- Standardized language to English
- Removed duplicate
- Drop Colunm “Education Speciality“
- Added a new column to calculate age range for each program “Age Range by Program“

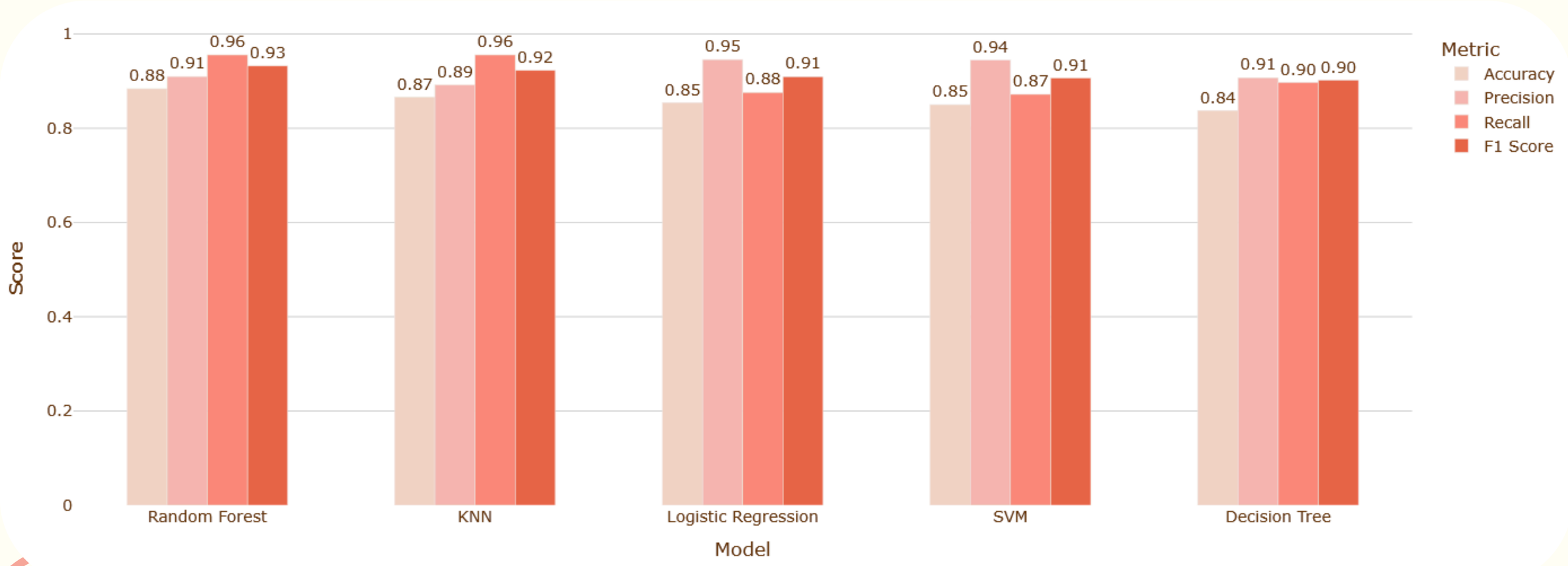


- Converted dates to Datetime format
- Addressed NaN values as follows:
  - Still Working: 4535
  - Job Type: 4535
  - College: 3862
  - Technology Type: 2958
  - Program Skill Level: 1645
  - Program Sub Category Code: 920
  - Employment Status: 557
  - Age: 87

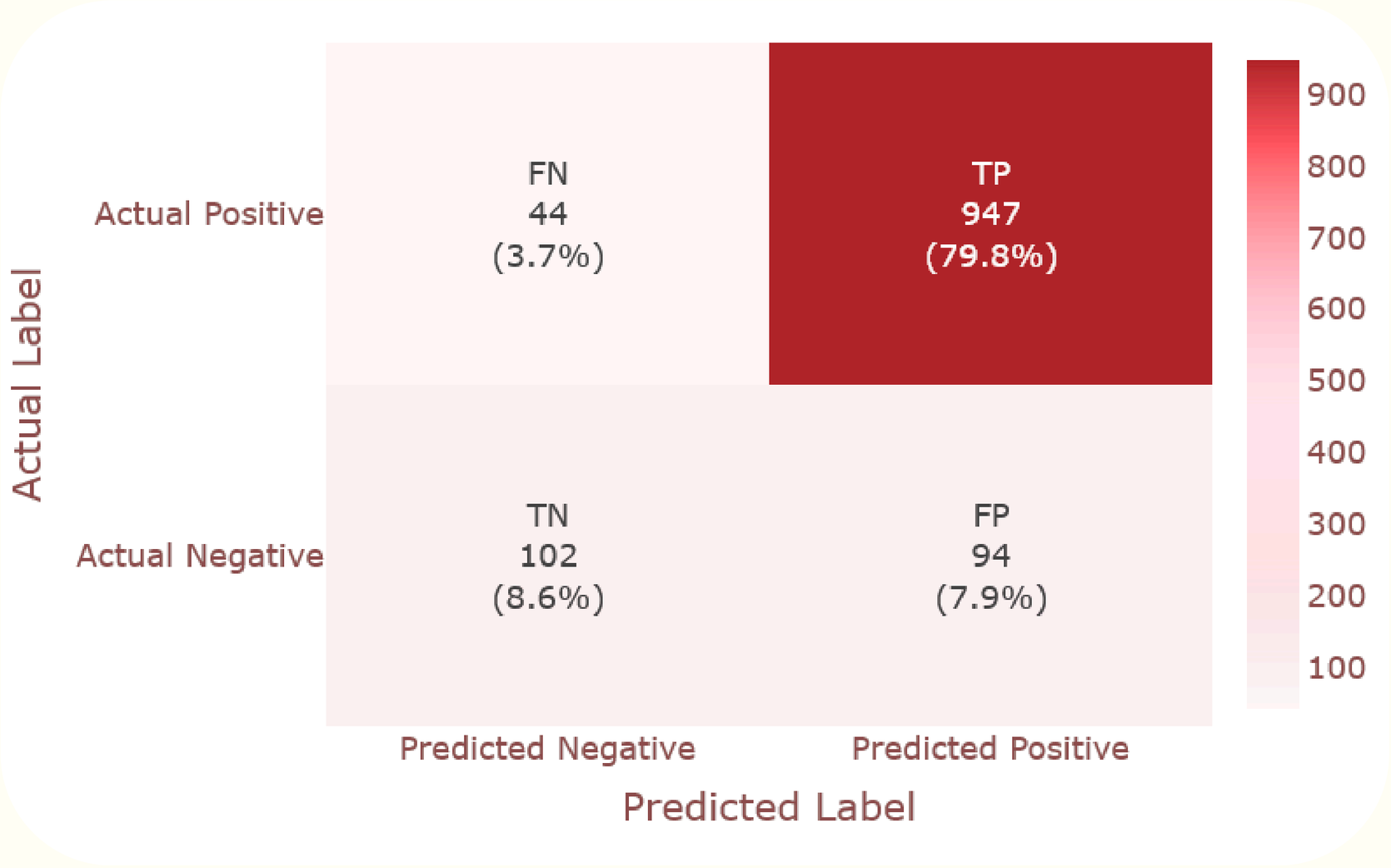
Models used for comparison

- Logistic Regression
- SVM
- Random Forest
- Decision Tree
- KNN

performance comparison



confusion Matrix for Random Forest



Model Improvement

- Five models were evaluated; Random Forest was selected based on performance.

- To optimize model performance, GridSearchCV was applied to each model to fine-tune hyperparameters. The improvement across models was slight.
- New features were added based on the program's age requirements, including:
  Age In Range, Age Gap To Min, Age Gap To Max, and Extreme Age to enhance the model's accuracy in predicting student program completion.

Model Performance

Random Forest Classifier	
Accuracy	0.88
Precision	0.91
Recall	0.95
F1 Score	0.93

Thank You