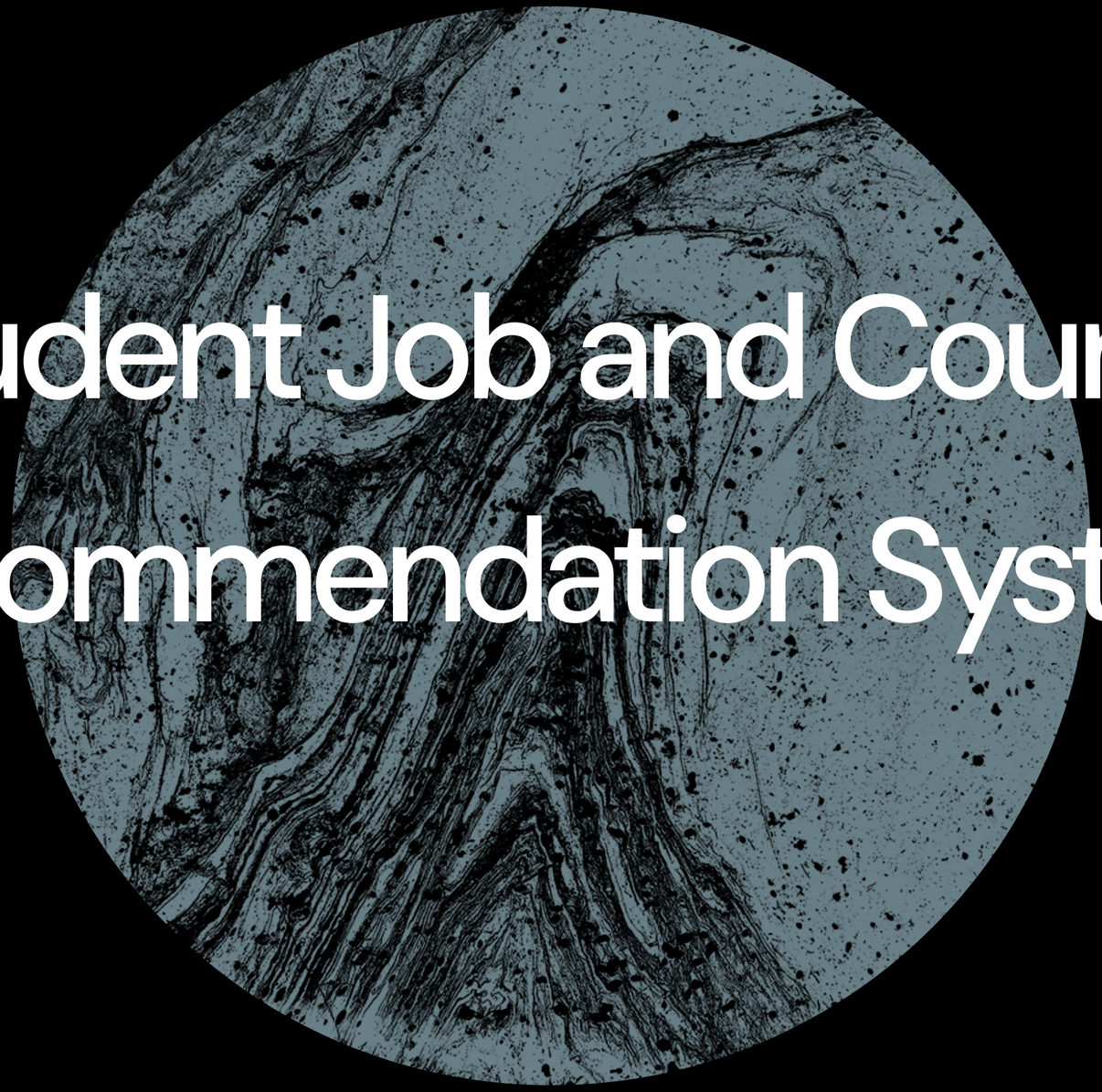


Student Job and Course Recommendation System

Sashank Boppana

Tejesh Boppana

Intro to Data Science

Data and Objective

Choosing the right career path can be overwhelming for students. This project aims to simplify that journey by providing personalized job and course recommendations based on their academic background, skills, and career goals.

Helps students find the right job opportunities by matching their skills and interests with relevant job roles.

Suggests online courses that can help them develop in-demand skills and improve their chances of landing a great job.

Data Used

This project utilizes three key datasets to provide job and course recommendations.

Datasets: (Kaggle)

- **Studentdata.csv** – Contains student academic and career-related information.
- **Online_Courses.csv** – Lists various online courses with relevant details.
- **Job_Postings.csv** – Includes job opportunities along with required skills and company details.

EDA RESULTS

Key Insights from Student Data:

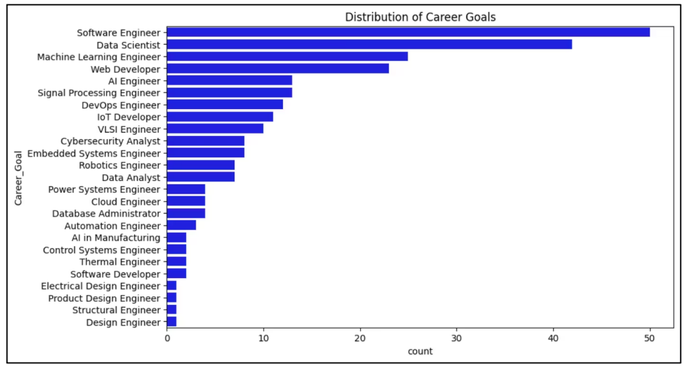
- Most students scored 75–95% in 12th grade.
- Top goals: ML Engineer, Web Developer, Data Scientist.
- Top skills: Python, SQL, HTML, Java, ML.

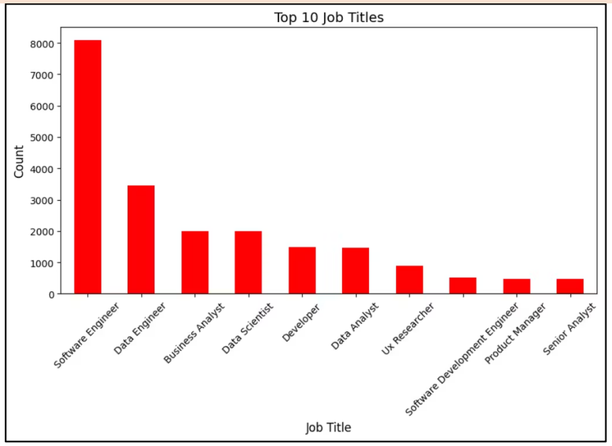
Course Insights:

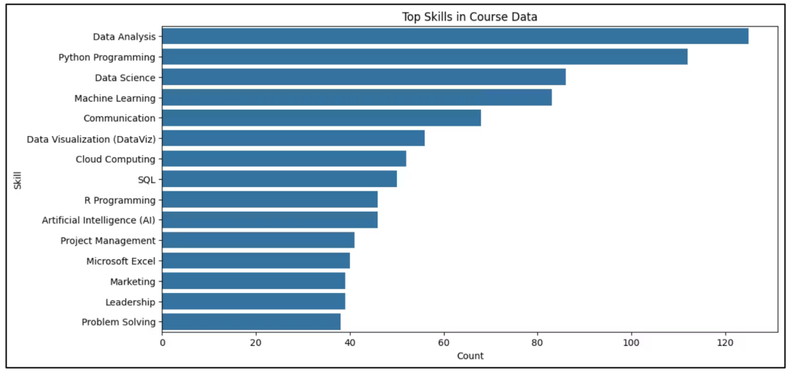
- Popular: Standard courses over specializations.
- Focused on: Python, Data Science, ML, Business.

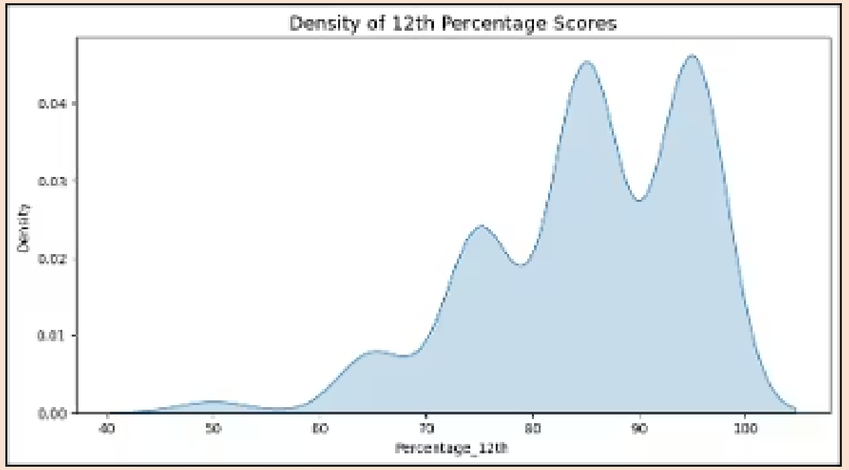
Job Market Highlights:

- Top Roles: Software Engineer, Data Engineer, Analyst.
- Locations: USA major hubs California, NY, Texas.
- Skills mismatch indicates need for targeted upskilling.









Feature Selection

Methods Used:

- **Random Forest** – Measures feature impact on model accuracy
- **Mutual Information** – Captures dependency between features and target

Top Features Identified:

- Course_Type_Encoded, Skills_Count
- Keywords: machine, programming, data, sql, analysis

Insights:

- RF captures **predictive power**
- MI captures **statistical relevance**
- Combined scores for **robust feature ranking**

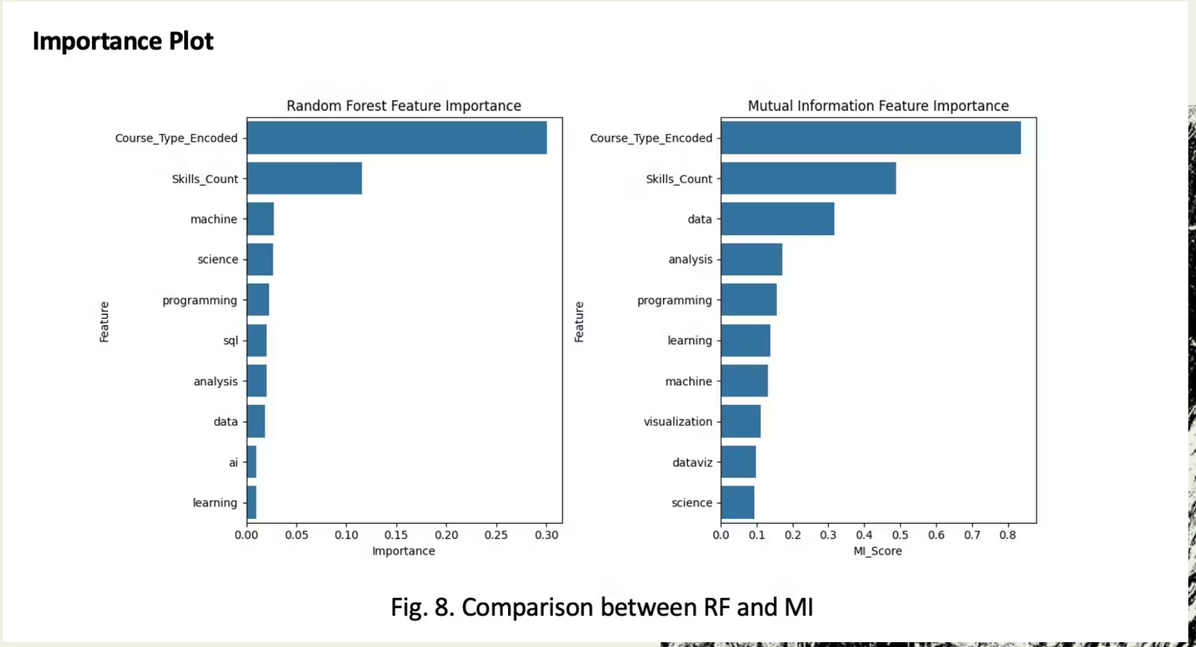
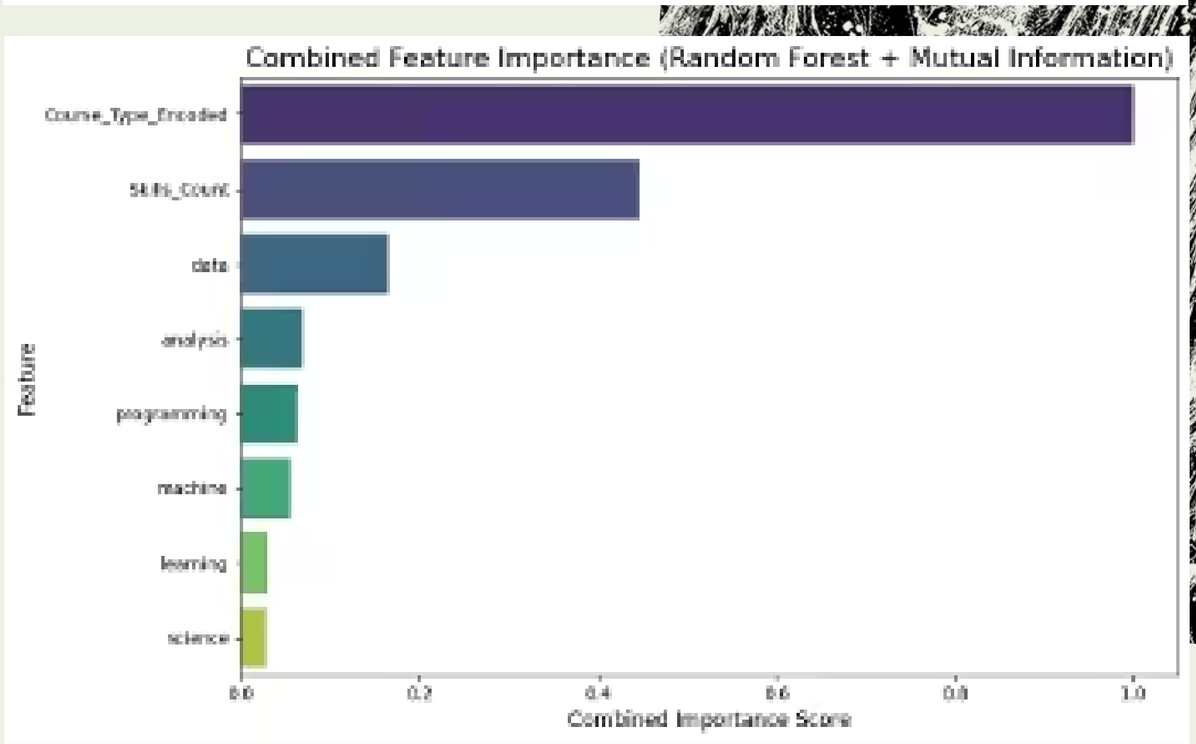


Fig. 8. Comparison between RF and MI



MODEL SELECTION AND EVALUATION

The **Decision Tree model** was selected as the best performer with **94.2% accuracy**, **94.3% F1 score**, and **90.1% cross-validation accuracy**, showing strong and balanced results. Although **Random Forest** had a slightly higher ROC-AUC (90.5%), its lower F1 score made it less ideal. **SVM** and **Logistic Regression** underperformed, especially in recall and handling complex patterns. Thus, the Decision Tree offered the best trade-off between accuracy, precision, recall, and interpretability.

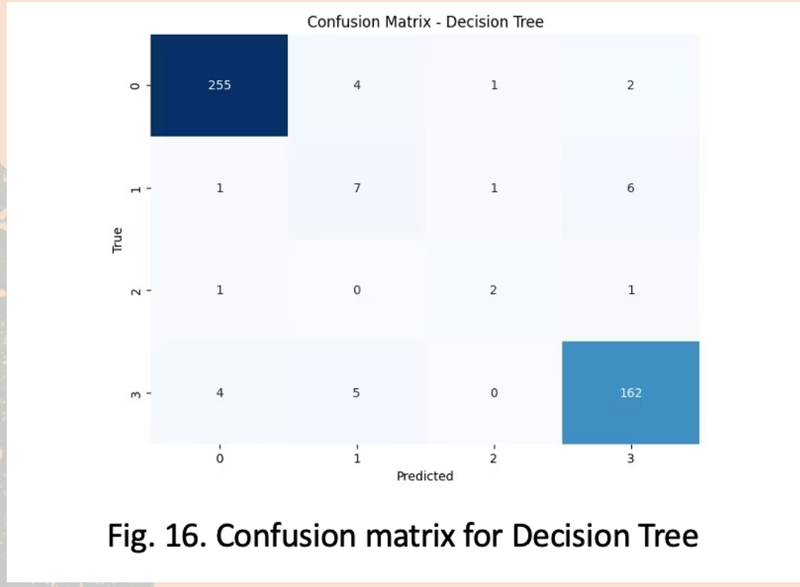


Fig. 16. Confusion matrix for Decision Tree

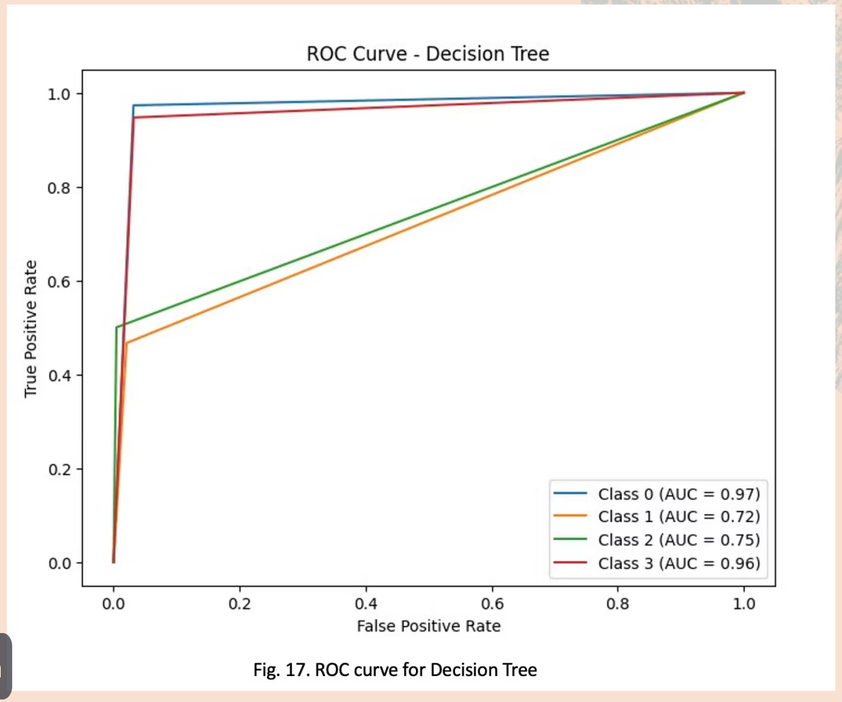
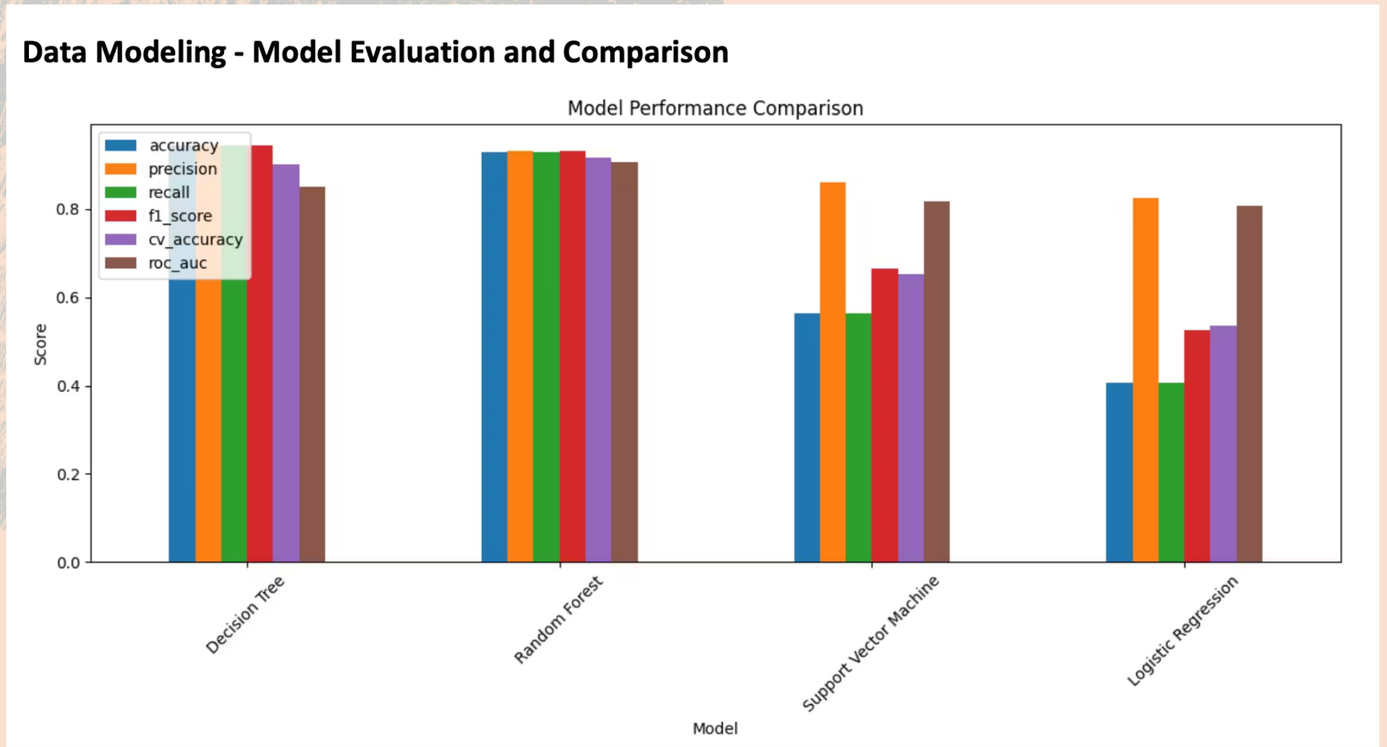


Fig. 17. ROC curve for Decision Tree



Takeaways & Future Work

01 Tree-based models outperformed linear/SVM models

02 Skills and Course type are crucial for predictions

03 Deploy model with interactive Flask/Streamlit frontend

04 Allow real-time input for real-time recommendations