GROUP PROJECT

by Borcelle Group

HỆ THỐNG DỰ BÁO TUYỂN DỤNG & MỨC LƯƠNG KHỞI ĐIỂM

Tiêu đề phụ: Ứng dụng Học máy (Machine Learning) trong Định hướng nghề nghiệp

- Người thực hiện: Nguyễn Hoàng Vũ
- Giảng viên hướng dẫn: TS. Nguyễn Sĩ Thìn

- Ngày báo cáo: 22/11/2025

NỘI DUNG TRÌNH BÀY (AGENDA)

1. Tổng quan & Bài toán
2. Phân tích dữ liệu (EDA)
3. Phương pháp luận & Xử lý dữ liệu
4. Huấn luyện & Đánh giá mô hình
5. Demo ứng dụng
6. Kết luận & Hướng phát triển

I. TỔNG QUAN & BÀI TOÁN

Bối cảnh:

- Thị trường lao động cạnh tranh, sinh viên thiếu thông tin định lượng về năng lực bản thân.
- Nhu cầu chuyển dịch từ tuyển dụng cảm tính sang Data-Driven Recruitment.

Mục tiêu dự án:

- Bài toán A (Phân loại): Dự đoán sinh viên Đậu hay Rớt?
  - Input: GPA, Kinh nghiệm, Ngành học...
  - Output: Placed / Not Placed.
- Bài toán B (Hồi quy): Dự đoán Mức lương khởi điểm là bao nhiêu?
  - Điều kiện: Chỉ dự báo cho nhóm đã trúng tuyển.

2. PHÂN TÍCH DỮ LIỆU (EDA)

Tổng quan bộ dữ liệu:

- Nguồn: Hồ sơ sinh viên khối kỹ thuật (Engineering).
- Đặc trưng chính (Features):
  - Định lượng: GPA, Tuổi, Số năm kinh nghiệm.
  - Định tính: Giới tính, Chuyên ngành, Tên trường ĐH.

2. PHÂN TÍCH DỮ LIỆU (EDA)

Phát hiện quan trọng (Key Insights):

- Mất cân bằng dữ liệu: Tỷ lệ sinh viên Đậu/Rớt không đồng đều $\rightarrow$ Cần xử lý kỹ thuật.
- Tương quan mạnh: Kiểm định T-test cho thấy nhóm Đậu có GPA trung bình cao hơn rõ rệt so với nhóm Rớt.
- Ngoại lai (Outliers): Mức lương có phân phối lệch phải (một số ít lương rất cao).

3. PHƯƠNG PHÁP LUẬN & TIỀN XỬ LÝ

Quy trình xử lý (Pipeline):

1

Làm sạch (Cleaning):

- Xử lý dữ liệu thiếu: Điền Mode cho cột Kinh nghiệm.
- Loại bỏ cột thừa: ID, Name, Degree (Variance = 0).

2

Mã hóa (Encoding) - Điểm nhấn kỹ thuật:

- Gender, Stream: Sử dụng One-Hot Encoding.
- College_Name: Sử dụng Frequency Encoding (Thay tên trường bằng tần suất xuất hiện) để xử lý vấn đề số lượng trường quá lớn (High Cardinality).

3

Chuẩn hóa (Scaling):

- Sử dụng StandardScaler cho các biến số (Age, GPA, Exp).

4

Cân bằng dữ liệu:

- Áp dụng SMOTE (Synthetic Minority Over-sampling Technique) trên tập Train để tránh mô hình thiên vị nhóm đa số.

4. HUẤN LUYỆN MÔ HÌNH (MODELING)

1

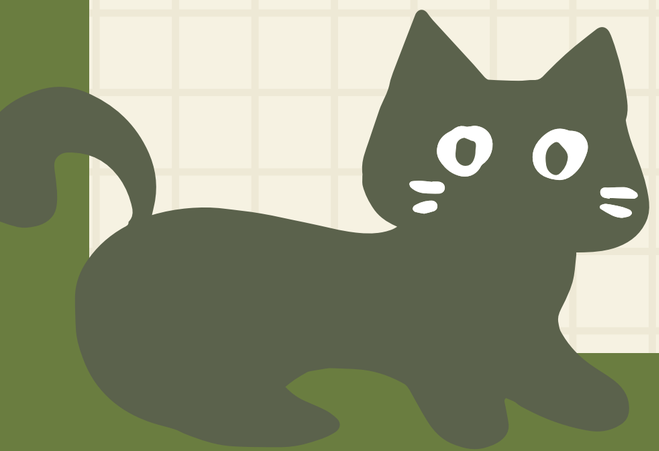
Chiến lược:

- Split ratio: 80% Train - 20% Test.
- Tối ưu hóa: GridSearchCV.

2

Tại sao chọn Random Forest?

- Xử lý tốt dữ liệu hỗn hợp (số & chữ).
- Ít bị ảnh hưởng bởi ngoại lai hơn Linear Regression.
- Nắm bắt được các tương tác phi tuyến (VD: GPA thấp nhưng Kinh nghiệm cao vẫn đậu).



Bài toán	Thuật toán tốt nhất	Metric chính	Kết quả
Phân loại	**Random Forest Classifier**	**F1-Score**	**~ 0.93**
Hồi quy	Random Forest Regressor	RMSE / R^2	R^2 ~ 0.82

ĐÁNH GIÁ CHI TIẾT & GIẢI THÍCH

Hiệu suất phân loại:

- Confusion Matrix: Tỷ lệ nhận diện đúng cả 2 lớp (Đậu/Rớt) đều cao.
- ROC-AUC: Đạt tiệm cận 0.9x → Mô hình tin cậy.

Giải thích mô hình (Feature Importance & SHAP):

1. GPA: Yếu tố quan trọng số 1 (~35%).
2. Kinh nghiệm: Yếu tố quan trọng số 2 (~25%).
3. Trường ĐH: Có ảnh hưởng đáng kể đến mức lương.

→ Kết luận: Năng lực học thuật và kinh nghiệm thực tế là yếu tố quyết định, giới tính ít ảnh hưởng.

5. DEMO ỨNG DỤNG (DEPLOYMENT)

Công nghệ:

- Framework: Streamlit (Python).
- Backend: Scikit-learn, Joblib.

Luồng hoạt động:

1. Người dùng nhập thông tin qua giao diện Web.
2. Hệ thống tự động tiền xử lý (Encoding/Scaling).
3. Mô hình Phân loại chạy dự báo → Nếu Đậu → Chạy tiếp mô hình Lương.
4. Hiển thị kết quả thời gian thực.

6. KẾT LUẬN & HƯỚNG PHÁT TRIỂN

Kết luận:

- Đã xây dựng thành công quy trình End-to-End từ dữ liệu thô đến ứng dụng Web.
- Mô hình Random Forest kết hợp SMOTE giải quyết tốt bài toán tuyển dụng.

Hạn chế:

- Dữ liệu chưa bao gồm kỹ năng mềm (Soft Skills).
- Mô hình chưa có khả năng tự học (Online Learning).

Hướng phát triển:

- Tích hợp NLP để phân tích CV (Resume Parsing).
- Mở rộng dữ liệu sang các khối ngành Kinh tế/Xã hội.

THANK YOU

Q&A