### **Thông tin sinh viên**

* **Họ tên:** Đặng Lê Nin
* **MSSV:** 2202088

### **Tên đề tài: Phân tích và Trực quan hóa dữ liệu để Dự đoán Cơ hội Trúng tuyển Chương trình Cao học**

### **Mô tả dữ liệu**

Bộ dữ liệu [**Admission\_Predict**](https://www.kaggle.com/datasets/mohansacharya/graduate-admissions/data?select=Admission_Predict.csv) bao gồm thông tin hồ sơ của ứng viên đăng ký chương trình cao học, chủ yếu dựa vào các yếu tố như:

| **Thuộc tính** | **Mô tả** |
| --- | --- |
| **GRE Score** | Điểm thi GRE của ứng viên |
| **TOEFL Score** | Điểm TOEFL của ứng viên |
| **University Rating** | Đánh giá xếp hạng của trường đại học ứng viên đăng ký (thang 1–5) |
| **SOP** | Mức độ chất lượng bài tự luận cá nhân (Statement of Purpose) |
| **LOR** | Đánh giá thư giới thiệu (Letter of Recommendation) |
| **CGPA** | Điểm trung bình tích lũy đại học |
| **Research** | Kinh nghiệm nghiên cứu (0 = Không có, 1 = Có) |
| **Chance of Admit** | Xác suất trúng tuyển (giá trị liên tục từ 0 đến 1) |

### **Mục tiêu dự án**

* **Phân tích và trực quan hóa dữ liệu** nhằm hiểu rõ các yếu tố ảnh hưởng đến khả năng trúng tuyển vào **Chương trình Cao học**
* **Xây dựng mô hình học máy** để dự đoán xác suất trúng tuyển của một ứng viên dựa trên hồ sơ đầu vào.

### **Phân tích và Trực quan hóa Dữ liệu (Data Visualization)**

**Các bước thực hiện:**

* **Bước** **1:** Kiểm tra dữ liệu thiếu và xử lý giá trị khuyết.
* **Bước 2:** Thống kê mô tả (trung bình, min, max, độ lệch chuẩn).
* **Bước 3:** Trực quan hóa phân phối các biến đầu vào (histogram, KDE plot).
* **Bước 4:** Vẽ scatter plot giữa GRE, TOEFL, CGPA với Chance of Admit.
* **Bước 5:** Phân tích tương quan bằng biểu đồ heatmap.
* **Bước 6:** So sánh khả năng trúng tuyển giữa nhóm có và không có kinh nghiệm nghiên cứu (boxplot).
* **Bước 7:** Đánh giá ảnh hưởng của University Rating, SOP và LOR.

### **Xây dựng Mô hình Dự đoán**

* **Mục tiêu:** Dự đoán Chance of Admit (hồi quy liên tục).
* **Mô hình sử dụng:** • Linear Regression  
   • Random Forest Regressor  
   • Gradient Boosting Regressor
* **Tiền xử lý:** • Chuẩn hóa đặc trưng (StandardScaler hoặc MinMaxScaler)  
   • Chia tập dữ liệu: Train/Test 80%–20%
* **Đánh giá mô hình:** • MAE – Mean Absolute Error  
   • MSE – Mean Squared Error  
   • R² Score