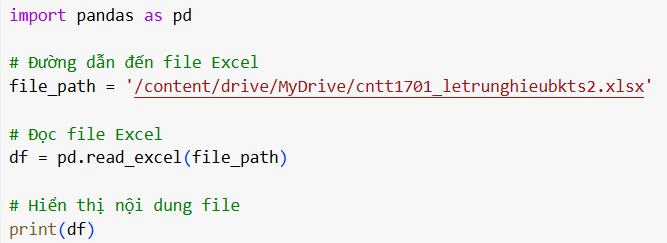
lấy file dữ liệu trên chatgpt

Lưu file vào Excel rồi tải lên driver



Biến file\_path lưu đường dẫn đến file mà mình muốn đọc

pd.read\_excel(file\_path) là hàm dùng để đọc dữ liệu từ file Excel và lưu vào một biến df (DataFrame).

df là một dataframe, tức là một bảng dữ liệu 2D giống như bảng tính trong Excel.

Print (df) in ra nội dung của dataframe

Ảnh có chứa văn bản, Phông chữ

Nội dung do AI tạo ra có thể không chính xác.

Chương trình này thực hiện các bước sau:

1. Nhập thư viện Pandas để xử lý dữ liệu.
2. Định nghĩa đường dẫn file Excel cần đọc.
3. Đọc dữ liệu từ file Excel vào một DataFrame (df).
4. In ra nội dung của DataFrame để xem dữ liệu đã đọc.
5. Tính toán và hiển thị các chỉ số thống kê cơ bản của dữ liệu.

Ảnh có chứa văn bản, ảnh chụp màn hình, Phông chữ

Nội dung do AI tạo ra có thể không chính xác.

df.describe() tính toán các chỉ số thống kê cơ bản:

**count**: Số lượng giá trị không rỗng.

**mean**: Trung bình.

**std**: Độ lệch chuẩn.

**min**: Giá trị nhỏ nhất.

**25%, 50%, 75%**: Các phân vị.

**max**: Giá trị lớn nhất.

print(stats) in kết quả ra màn hình.

plt.subplot(1, 2, 2): Chọn vị trí thứ hai trong vùng vẽ.

sns.histplot(df["Tỷ lệ tốt nghiệp (%)"], kde=True, bins=10): Vẽ biểu đồ phân phối dữ liệu với KDE (đường mật độ).

plt.title(...): Đặt tiêu đề biểu đồ.

Ảnh có chứa văn bản, ảnh chụp màn hình, Phông chữ

Nội dung do AI tạo ra có thể không chính xác.

numpy: Hỗ trợ tính toán số học và xử lý mảng dữ liệu.

matplotlib.pyplot: Dùng để vẽ biểu đồ.

seaborn: Thư viện hỗ trợ vẽ biểu đồ đẹp hơn với Matplotlib.

sklearn.model\_selection.train\_test\_split: Chia dữ liệu thành tập huấn luyện và kiểm tra.

sklearn.linear\_model.LinearRegression: Xây dựng mô hình hồi quy tuyến tính.

sklearn.metrics: Đánh giá chất lượng mô hình.

X chứa dữ liệu về năm, là biến độc lập.

y chứa dữ liệu về tỷ lệ tốt nghiệp, là biến phụ thuộc

train\_test\_split chia dữ liệu thành 80% huấn luyện, 20% kiểm tra.

random\_state=42 giúp đảm bảo việc chia dữ liệu có thể tái lặp.

LinearRegression() tạo một mô hình hồi quy tuyến tính.

model.fit(X\_train, y\_train) huấn luyện mô hình với dữ liệu huấn luyện.

model.predict(X\_test) dự đoán giá trị tỷ lệ tốt nghiệp dựa trên dữ liệu năm trong tập kiểm tra.

mean\_squared\_error(y\_test, y\_pred): Tính toán lỗi trung bình bình phương (MSE), thể hiện mức độ sai số của mô hình.

r2\_score(y\_test, y\_pred): Đánh giá mức độ phù hợp của mô hình với dữ liệu (R² score, càng gần 1 càng tốt).

plt.figure(figsize=(10, 5)): Tạo biểu đồ có kích thước 10x5 inch.

plt.scatter(X\_test, y\_test, color="blue", label="Thực tế"): Vẽ dữ liệu thực tế.

plt.plot(X\_test, y\_pred, color="red", linewidth=2, label="Dự đoán"): Vẽ đường hồi quy tuyến tính.

plt.xlabel("Năm"), plt.ylabel("Tỷ lệ tốt nghiệp (%)"): Đặt nhãn cho trục.

plt.title(...): Tiêu đề biểu đồ.

plt.legend(): Thêm chú thích.

plt.show(): Hiển thị biểu đồ.

**Kết Luận**

**Chất lượng dữ liệu**

Dữ liệu đã được đọc từ file Excel, kiểm tra, làm sạch và chuyển đổi đúng định dạng.

Việc xử lý dữ liệu bị thiếu, sai định dạng giúp cải thiện độ chính xác của mô hình.

**Phân tích dữ liệu**

Các chỉ số thống kê cơ bản giúp hiểu rõ hơn về dữ liệu.

Việc trực quan hóa dữ liệu bằng biểu đồ cột và biểu đồ phân phối giúp thấy rõ xu hướng và sự phân bố của tỷ lệ tốt nghiệp qua các năm.

**Hiệu suất mô hình**

Mô hình hồi quy tuyến tính đã được huấn luyện và kiểm tra với tập dữ liệu đã chia.

Kết quả đánh giá mô hình thông qua MSE và R² Score cho thấy mức độ phù hợp của mô hình với dữ liệu thực tế.

**Hướng cải thiện mô hình**

**Thử nghiệm các mô hình khác** như hồi quy bậc hai, Random Forest, hoặc mạng nơ-ron nhân tạo (ANN).

**Thu thập thêm dữ liệu** từ nhiều năm hơn hoặc bổ sung các yếu tố khác.

**Xử lý đặc trưng dữ liệu tốt hơn** bằng cách tạo thêm biến mới hoặc điều chỉnh mô hình.

Nhìn chung, mô hình hồi quy tuyến tính cung cấp một cách tiếp cận cơ bản để dự đoán tỷ lệ tốt nghiệp theo năm. Tuy nhiên, cần tiếp tục tối ưu hóa để cải thiện độ chính xác.

Link colab: <https://colab.research.google.com/drive/1SEgP2prodE5TeT1pAg9gd15tRyF6NyvQ?hl=vi#scrollTo=wDFwQivQwj9Z>

Link github: