## Dự án: Dự đoán Đánh giá Yelp bằng MLP

### Giới thiệu:

Dự án này nhằm dự đoán điểm đánh giá Yelp của một nhà hàng bằng mạng nơ-ron đa tầng (MLP). Mô hình được huấn luyện trên một tập dữ liệu bao gồm các đặc trưng khác nhau như cảm xúc, chỉ số nhị phân và giá trị số. Kiến trúc MLP cho phép bắt được mối quan hệ phi tuyến giữa các đặc trưng và biến mục tiêu.

### Bộ dữ liệu:

Bộ dữ liệu được sử dụng cho quá trình huấn luyện và đánh giá là sự kết hợp của nhiều tệp dữ liệu Yelp, bao gồm `yelp\_business.json`, `yelp\_review.json`, `yelp\_user.json`, `yelp\_checkin.json`, `yelp\_tip.json` và `yelp\_photo.json`. Những tệp này cung cấp thông tin về các doanh nghiệp, đánh giá, người dùng, check-in, gợi ý và ảnh trên Yelp. Dữ liệu được tải và hợp nhất thành một DataFrame duy nhất để tiếp tục xử lý.

### Tiền xử lý dữ liệu:

1. Hợp nhất các DataFrame: Các DataFrame riêng lẻ thu được từ các tệp dữ liệu được hợp nhất dựa trên cột `business\_id` chung bằng cách sử dụng hàm `merge()` trong thư viện Pandas.

2. Lựa chọn đặc trưng: Một số cột không liên quan hoặc không hữu ích cho việc dự đoán được loại bỏ khỏi DataFrame bằng cách sử dụng hàm `drop()`. Danh sách `features\_to\_remove` xác định các cột cần loại bỏ.

3. Xử lý giá trị thiếu: Các giá trị thiếu trong DataFrame được điền bằng giá trị phù hợp sử dụng hàm `fillna()`. Các cột cụ thể được điền bằng số không để xử lý giá trị thiếu.

### Kiến trúc mô hình:

Mô hình MLP được triển khai bằng Keras, một thư viện học sâu cấp cao. Kiến trúc mô hình bao gồm nhiều lớp mạng đầy đủ với các hàm kích hoạt, cho phép xử lý biến đổi phi tuyến giữa các đặc trưng và biến mục tiêu. L

ớp đầu vào có số chiều phù hợp với số đặc trưng trong tập dữ liệu. Các lớp ẩn có thể tùy chỉnh bằng cách điều chỉnh số neuron và các hàm kích hoạt. Lớp đầu ra có một neuron, dự đoán điểm đánh giá Yelp.

### Huấn luyện mô hình:

1. Chia dữ liệu: Tập dữ liệu được chia thành các tập huấn luyện và kiểm tra bằng tỷ lệ xác định (ví dụ: 80% huấn luyện, 20% kiểm tra). Điều này được thực hiện bằng hàm `train\_test\_split()` trong module `sklearn.model\_selection`.

2. Biên dịch mô hình: Mô hình MLP được biên dịch với bộ tối ưu hóa, hàm mất mát và chỉ số đánh giá cụ thể. Bộ tối ưu hóa xác định thuật toán được sử dụng cho gradient descent, trong khi hàm mất mát đo lường sai số giữa điểm đánh giá dự đoán và thực tế.

3. Huấn luyện mô hình: Mô hình được huấn luyện trên dữ liệu huấn luyện bằng hàm `fit()`. Số lượng epochs (vòng lặp) và kích thước lô (batch size) có thể điều chỉnh để kiểm soát quá trình huấn luyện.

4. Đánh giá mô hình: Sau khi huấn luyện, mô hình được đánh giá trên dữ liệu kiểm tra bằng hàm `evaluate()`. Điều này cung cấp các độ đo như hàm mất mát và độ chính xác để đánh giá hiệu suất của mô hình.

### Dự đoán mô hình:

Để thực hiện dự đoán bằng mô hình đã huấn luyện, các giá trị đầu vào mong muốn được cung cấp dưới dạng một mảng NumPy. Các giá trị nên có cùng thứ tự và định dạng với các đặc trưng được sử dụng trong quá trình huấn luyện. Hàm `predict()` được sử dụng để thu được dự đoán về điểm đánh giá Yelp.

### Điều chỉnh và Tinh chỉnh:

Mã nguồn cho phép tinh chỉnh và điều chỉnh mô hình. Kiến trúc, hàm kích hoạt, bộ tối ưu hóa, hàm mất mát và các siêu tham số khác có thể được điều chỉnh để thử nghiệm với các c

ấu hình khác nhau và cải thiện hiệu suất của mô hình.

Đây là một phiên bản cơ bản của mã nguồn, và bạn có thể tùy chỉnh và tinh chỉnh nó dựa trên yêu cầu của mình. Hãy sử dụng mã nguồn này như một cơ sở để nghiên cứu và phát triển mô hình dự đoán đánh giá Yelp của riêng bạn.