**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC ĐẠI NAM**

Logo, company name

Description automatically generated

**BÀI TẬP LỚN**

**TÊN HỌC PHẦN: TRÍ TUỆ NHÂN TẠO**

## **ĐỀ TÀI: XÂY DỰNG MÔ HÌNH ANN ĐỂ PHÂN LOẠI TÌNH TRẠNG THẤT NGHIỆP DỰA TRÊN CÁC YẾU TỐ NHƯ TRÌNH ĐỘ HỌC VẤN, KINH NGHIỆM, NGÀNH NGHỀ**

**Giáo viên hướng dẫn: TS. Trần Đăng Công**

**Sinh viên thực hiện:**

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Mã sv** | **Họ và tên** | **Lớp** |
| 1 | 1771020625 | Tạ Công Thắng | CNTT17-08 |
| 2 | 1771020595 | Cao Thái Sơn | CNTT17-08 |
| 3 | 1771020043 | Trần Văn Tuấn Anh | CNTT17-08 |

**Hà Nội, năm 2025**

**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC ĐẠI NAM**

Logo, company name

Description automatically generated

**BÀI TẬP LỚN**

**TÊN HỌC PHẦN: TRÍ TUỆ NHÂN TẠO**

## **ĐỀ TÀI: XÂY DỰNG MÔ HÌNH ANN ĐỂ PHÂN LOẠI TÌNH TRẠNG THẤT NGHIỆP DỰA TRÊN CÁC YẾU TỐ NHƯ TRÌNH ĐỘ HỌC VẤN, KINH NGHIỆM, NGÀNH NGHỀ**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| STT | Mã Sinh Viên | Họ và Tên | Ngày Sinh | Điểm | |
| Bằng Số | Bằng Chữ |
| 1 | 1771020625 | Tạ Công Thắng | 17/01/2005 |  |  |
| 2 | 1771020595 | Cao Thái Sơn | 31/07/2005 |  |  |
| 3 | 1771020043 | Trần Văn Tuấn Anh | 27/03/2005 |  |  |

### 

### CÁN BỘ CHẤM THI

**Hà Nội, năm 2025**

**LỜI NÓI ĐẦU**

Trong bối cảnh nền kinh tế không ngừng biến động, vấn đề thất nghiệp ngày càng trở nên quan trọng và thu hút sự quan tâm của các nhà nghiên cứu, nhà hoạch định chính sách và doanh nghiệp. Việc dự đoán và phân loại tình trạng thất nghiệp dựa trên các yếu tố như trình độ học vấn, kinh nghiệm làm việc và ngành nghề có thể giúp cung cấp thông tin hữu ích cho việc hoạch định chính sách lao động, tư vấn nghề nghiệp và hỗ trợ người lao động tìm kiếm việc làm hiệu quả hơn.

Trí tuệ nhân tạo (AI), đặc biệt là mạng nơ-ron nhân tạo (Artificial Neural Network – ANN), đã chứng minh được tính hiệu quả trong việc xử lý và phân tích dữ liệu lớn, phát hiện các mẫu và xu hướng mà các phương pháp truyền thống khó có thể nhận diện. Do đó, việc xây dựng một mô hình ANN để phân loại tình trạng thất nghiệp có thể là một giải pháp tiềm năng nhằm hỗ trợ quá trình dự báo và phân tích thị trường lao động.

Mục tiêu của đề tài này là nghiên cứu và xây dựng một mô hình mạng nơ-ron nhân tạo (ANN) có khả năng phân loại tình trạng thất nghiệp dựa trên các yếu tố chính như trình độ học vấn, kinh nghiệm làm việc và ngành nghề. Để thực hiện điều này, đề tài sẽ áp dụng các phương pháp xử lý dữ liệu, huấn luyện mô hình ANN và đánh giá độ chính xác của mô hình thông qua các bộ dữ liệu thực tế.

Hy vọng rằng kết quả nghiên cứu sẽ góp phần mang lại một công cụ hữu ích giúp các tổ chức, cá nhân và cơ quan chức năng có thêm góc nhìn trong việc phân tích và dự báo thị trường lao động, từ đó đề xuất các giải pháp phù hợp nhằm giảm thiểu tình trạng thất nghiệp.

**MỤC LỤC**

[**CHƯƠNG I. TỔNG QUAN VỀ TRÍ TUỆ NHÂN TẠO** 15](#_Toc193489529)

[**1.1. Khái niệm và định nghĩa** 15](#_Toc193489530)

[1.1.1. Các nhánh khác của AI 15](#_Toc193489533)

[1.1.2 Vai trò của trí tuệ nhân tạo 15](#_Toc193489539)

[1.1.3. Các mục tiêu của trí tuệ nhân tạo 16](#_Toc193489540)

[1.1.4. Lý do chọn đề tài 17](#_Toc193489541)

[**1.2. Mục tiêu nghiên cứu** 17](#_Toc193489542)

[1.2.1. Xây dựng mô hình mạng nơ-ron nhân tạo (ANN) để phân loại tình trạng thất nghiệp 17](#_Toc193489543)

[1.2.2. Phân tích mối quan hệ giữa các yếu tố ảnh hưởng đến tình trạng thất nghiệp 18](#_Toc193489544)

[1.2.3. Thu thập, xử lý và chuẩn hóa dữ liệu để huấn luyện mô hình 18](#_Toc193489545)

[1.2.4. Tối ưu hóa mô hình ANN nhằm đạt hiệu suất cao nhất 18](#_Toc193489546)

[1.2.5. Đánh giá hiệu suất mô hình và đề xuất hướng cải tiến 18](#_Toc193489547)

[**1.3. Phạm vi nghiên cứu** 18](#_Toc193489548)

[**1.4. Phương pháp nghiên cứu** 19](#_Toc193489549)

[1.4.1. Phương pháp nghiên cứu lý thuyết 19](#_Toc193489550)

[1.4.2. Phương pháp thu thập và xử lý dữ liệu 20](#_Toc193489551)

[1.4.3. Phương pháp xây dựng mô hình ANN 20](#_Toc193489552)

[1.4.4. Phương pháp đánh giá và so sánh mô hình 20](#_Toc193489553)

[1.4.5. Phương pháp thực nghiệm và tối ưu hóa 21](#_Toc193489554)

[1.4.6. Phương pháp triển khai và ứng dụng thực tiễn 21](#_Toc193489555)

[**CHƯƠNG II. CƠ SỞ LÝ THUYẾT** 21](#_Toc193489556)

[**2. Tổng quan về mạng nơ-ron nhân tạo (ANN - Artificial Neural Network)** 21](#_Toc193489557)

[**2.1.** **Giới thiệu về mạng nơ-ron nhân tạo** 21](#_Toc193489558)

[2.1.1 Cấu trúc của mạng nơ-ron nhân tạo 22](#_Toc193489561)

[2.1.2. Cơ chế hoạt động của ANN 22](#_Toc193489568)

[2.1.3. Ứng dụng của ANN trong phân loại tình trạng thất nghiệp 24](#_Toc193489590)

[2.1.4. Kết luận 24](#_Toc193489595)

[**2.2. Các thuật toán học máy trong phân loại** 24](#_Toc193489597)

[2.2.1. Các thuật toán phân loại truyền thống 24](#_Toc193489598)

[2.2.2 Cây quyết định (Decision Tree) 25](#_Toc193489600)

[2.2.3. Rừng ngẫu nhiên (Random Forest) 25](#_Toc193489601)

[2.2.4. Máy vector hỗ trợ (Support Vector Machine - SVM) 25](#_Toc193489602)

[2.2.5. K-Nearest Neighbors (KNN) 26](#_Toc193489603)

[2.2.6. Mạng nơ-ron nhân tạo cơ bản (Feedforward Neural Network - FNN) 26](#_Toc193489604)

[2.2.7. Mạng nơ-ron tích chập (Convolutional Neural Network - CNN) 26](#_Toc193489605)

[2.2.8. Mạng nơ-ron hồi quy (Recurrent Neural Network - RNN) 26](#_Toc193489606)

[2.2.9. So sánh các thuật toán phân loại 27](#_Toc193489607)

[**2.3. Các phương pháp đánh giá mô hình** 27](#_Toc193489608)

[2.3.1. Chia tập dữ liệu huấn luyện và kiểm tra 27](#_Toc193489610)

[2.3.2. Ma trận nhầm lẫn (Confusion Matrix) 27](#_Toc193489616)

[2.3.3. Các chỉ số đánh giá hiệu suất 28](#_Toc193489624)

[2.3.4. Độ chính xác Top-K (Top-K Accuracy, đặc biệt khi phân loại đa lớp) 29](#_Toc193489641)

[2.3.5. Đường cong ROC (Receiver Operating Characteristic Curve) 29](#_Toc193489644)

[2.3.6. Chỉ số AUC (Area Under Curve) 30](#_Toc193489648)

[2.3.7 Kiểm tra Overfitting bằng Loss Function 30](#_Toc193489653)

[**CHƯƠNG III.KẾT QUẢ NGHIÊN CỨU VÀ PHÂN TÍCH** 31](#_Toc193489662)

[**3.1.Phân tích yêu cầu bài toán** 31](#_Toc193489663)

[**3.2. Mô tả thuật toán sử dụng** 33](#_Toc193489664)

[**3.3 Phân tích mã nguồn** 37](#_Toc193489665)

[**3.4. Nhận xét đặc điểm và hạn chế của thuật toán** 42](#_Toc193489666)

[3.4.1. Đặc điểm (Ưu điểm) 42](#_Toc193489667)

[3.4.2. Hạn chế 43](#_Toc193489668)

[**3.5. đề xuất ý tưởng cải tiến thuật toán** 45](#_Toc193489669)

[3.5.1. Tự động hóa việc chọn siêu tham số 45](#_Toc193489670)

[3.5.2. Kết hợp nhiều kỹ thuật xử lý mất cân bằng 45](#_Toc193489671)

[3.5.3. Đánh giá mô hình qua nhiều lần chia dữ liệu 45](#_Toc193489672)

[3.5.4. Chọn ngưỡng dự đoán linh hoạt 46](#_Toc193489673)

[3.5.5. Tăng tốc độ huấn luyện 46](#_Toc193489674)

[3.5.6. So sánh nhiều kiến trúc mô hình 46](#_Toc193489675)

[3.5.7. Xử lý dữ liệu linh hoạt và bền vững hơn 47](#_Toc193489676)

[3.5.8. Tập trung vào các đặc trưng quan trọng 47](#_Toc193489677)

[**KẾT LUẬN** 48](#_Toc193489678)

[**DANH MỤC TÀI LIỆU THAM KHẢO** 50](#_Toc193489679)

# **CHƯƠNG I. TỔNG QUAN VỀ TRÍ TUỆ NHÂN TẠO**

## **1.1. Khái niệm và định nghĩa**

### Định nghĩa:

### Trí tuệ nhân tạo (Artificial Intelligence – AI) là một lĩnh vực của khoa học máy tính nhằm tạo ra các hệ thống hoặc chương trình máy tính có khả năng mô phỏng trí tuệ con người. AI giúp máy tính có thể tự học hỏi, suy luận, ra quyết định và thực hiện các nhiệm vụ một cách thông minh mà không cần sự can thiệp trực tiếp của con người.

### 1.1.1. Các nhánh khác của AI

### Bao gồm:

### Học máy (Machine Learning - ML)

### Mạng nơ-ron nhân tạo (Artificial Neural Networks - ANN),

### Xử lý ngôn ngữ tự nhiên (Natural Language Processing - NLP)

### Thị giác máy tính (Computer Vision - CV)

### 1.1.2 Vai trò của trí tuệ nhân tạo

Bao gồm 8 vai trò:

- Tự động hóa và tối ưu hóa quy trình : AI giúp tự động hóa các công việc lặp đi lặp lại, giảm bớt sự can thiệp của con người, tăng năng suất và độ chính xác.

- Hỗ trợ ra quyết định: AI có khả năng phân tích lượng lớn dữ liệu, phát hiện xu hướng và đưa ra dự đoán chính xác, giúp doanh nghiệp và tổ chức đưa ra quyết định tối ưu.

- Cải thiện trải nghiệm người dùng: AI giúp cá nhân hóa nội dung và dịch vụ theo nhu cầu của từng người dùng.

- Ứng dụng trong y tế: AI hỗ trợ chẩn đoán bệnh, phân tích hình ảnh y khoa và hỗ trợ bác sĩ đưa ra quyết định điều trị chính xác hơn.

- Ứng dụng trong giáo dục: AI giúp cá nhân hóa quá trình học tập, cung cấp nội dung phù hợp với từng học sinh, hỗ trợ giảng dạy trực tuyến.

- Cải thiện an ninh và giám sát: AI hỗ trợ nhận diện khuôn mặt, phân tích hành vi đáng ngờ trong hệ thống giám sát an ninh.

- Phát triển giao thông thông minh: AI giúp tối ưu hóa hệ thống giao thông, hỗ trợ xe tự lái, giảm ùn tắc và tai nạn.

- Ứng dụng trong thương mại và kinh doanh: AI hỗ trợ quản lý chuỗi cung ứng, tối ưu hóa kho bãi và cá nhân hóa dịch vụ khách hàng.

### 1.1.3. Các mục tiêu của trí tuệ nhân tạo

Bao gồm 9 mục tiêu:

- Mô phỏng trí tuệ con người : AI hướng đến việc mô phỏng khả năng tư duy, suy luận, học hỏi và giải quyết vấn đề giống như con người.

- Tự động hóa các nhiệm vụ phức tạp: AI giúp thực hiện các công việc lặp đi lặp lại hoặc phức tạp mà con người mất nhiều thời gian để xử lý.

- Cải thiện hiệu suất và năng suất làm việc: AI giúp tối ưu hóa quy trình làm việc, giảm thiểu lỗi do con người gây ra và nâng cao hiệu suất.

- Học tập và thích nghi từ dữ liệu: AI có khả năng học hỏi từ dữ liệu để cải thiện hiệu suất và đưa ra các dự đoán chính xác hơn theo thời gian.

- Ra quyết định thông minh: AI hỗ trợ phân tích dữ liệu lớn (Big Data) để đưa ra quyết định nhanh chóng và chính xác hơn con người.

- Tương tác tự nhiên với con người: AI được phát triển để giao tiếp tự nhiên với con người thông qua ngôn ngữ nói hoặc viết.

- Tạo ra các hệ thống tự động thông minh: AI hướng đến việc phát triển các hệ thống có thể hoạt động độc lập mà không cần sự can thiệp của con người.

- Tăng cường an ninh và bảo mật: AI hỗ trợ phát hiện và ngăn chặn các mối đe dọa về an ninh mạng, gian lận tài chính.

- Thúc đẩy đổi mới và phát triển công nghệ: AI góp phần vào sự phát triển của khoa học và công nghệ, mở ra nhiều ứng dụng đột phá trong tương lai.

### 1.1.4. Lý do chọn đề tài

Thất nghiệp là một trong những vấn đề kinh tế - xã hội nhức nhối, ảnh hưởng đến đời sống cá nhân, gia đình và sự phát triển của nền kinh tế. Dự báo và phân loại tình trạng thất nghiệp giúp cơ quan quản lý, doanh nghiệp và người lao động có thể đưa ra các quyết định phù hợp.

Trong thời đại Cách mạng Công nghiệp 4.0, AI và học máy (Machine Learning) ngày càng được áp dụng rộng rãi để giải quyết các bài toán dự báo và phân loại. Mạng nơ-ron nhân tạo (Artificial Neural Network - ANN) là một trong những mô hình mạnh mẽ trong việc nhận diện mẫu và dự đoán kết quả dựa trên dữ liệu.

Nhiều nghiên cứu chỉ ra rằng học vấn, kinh nghiệm làm việc và ngành nghề là những yếu tố quan trọng quyết định cơ hội có việc làm.

Việc xây dựng mô hình ANN để phân loại tình trạng thất nghiệp giúp hiểu rõ mối quan hệ giữa các yếu tố này và tình trạng việc làm. Cung cấp thông tin cho các nhà quản lý lao động để đề xuất chính sách đào tạo, hướng nghiệp phù hợp. Giúp cá nhân đánh giá cơ hội việc làm dựa trên trình độ và kinh nghiệm của họ.

Kết quả nghiên cứu có thể được ứng dụng trong các hệ thống hỗ trợ tuyển dụng, dự báo xu hướng lao động hoặc tư vấn nghề nghiệp. Mô hình có thể mở rộng bằng cách kết hợp thêm các yếu tố khác như kỹ năng mềm, khu vực địa lý, xu hướng thị trường lao động.

## **1.2. Mục tiêu nghiên cứu**

### 1.2.1. Xây dựng mô hình mạng nơ-ron nhân tạo (ANN) để phân loại tình trạng thất nghiệp

Phát triển một mô hình học sâu (Deep Learning) sử dụng mạng nơ-ron nhân tạo (ANN) nhằm dự đoán tình trạng thất nghiệp của một cá nhân dựa trên các đặc trưng đầu vào như trình độ học vấn, kinh nghiệm làm việc và ngành nghề.

So sánh độ chính xác của mô hình ANN với các phương pháp truyền thống hoặc các thuật toán học máy khác.

### 1.2.2. Phân tích mối quan hệ giữa các yếu tố ảnh hưởng đến tình trạng thất nghiệp

Đánh giá mức độ tác động của trình độ học vấn, kinh nghiệm làm việc và ngành nghề đối với khả năng có việc làm của một cá nhân.

Xác định xem yếu tố nào quan trọng nhất trong việc quyết định một người có bị thất nghiệp hay không.

### 1.2.3. Thu thập, xử lý và chuẩn hóa dữ liệu để huấn luyện mô hình

Xây dựng bộ dữ liệu về người lao động, bao gồm trình độ học vấn, kinh nghiệm, ngành nghề và tình trạng việc làm.

Thực hiện tiền xử lý dữ liệu như xử lý dữ liệu bị thiếu, chuẩn hóa dữ liệu và biến đổi dữ liệu thành dạng phù hợp để huấn luyện mô hình ANN.

### 1.2.4. Tối ưu hóa mô hình ANN nhằm đạt hiệu suất cao nhất

Lựa chọn kiến trúc mạng nơ-ron nhân tạo phù hợp, bao gồm số lượng lớp ẩn, số neuron trong từng lớp, và các hàm kích hoạt.

Điều chỉnh các tham số như tỷ lệ học (learning rate), số lượng epochs, và phương pháp tối ưu hóa để cải thiện độ chính xác của mô hình.

### 1.2.5. Đánh giá hiệu suất mô hình và đề xuất hướng cải tiến

Đánh giá mô hình bằng các chỉ số như độ chính xác (Accuracy), độ nhạy (Recall), độ đặc hiệu (Precision), F1-score và ma trận nhầm lẫn (Confusion Matrix).

So sánh mô hình ANN với các phương pháp khác như Hồi quy Logistic, Cây quyết định (Decision Tree) hoặc Random Forest để xác định tính hiệu quả.

Đề xuất các hướng phát triển tiếp theo như bổ sung thêm yếu tố đầu vào (ví dụ: kỹ năng mềm, vị trí địa lý), mở rộng tập dữ liệu hoặc cải tiến mô hình bằng các thuật toán khác.

## **1.3. Phạm vi nghiên cứu**

**Phạm vi nội dung**

Đề tài tập trung vào **xây dựng mô hình mạng nơ-ron nhân tạo (ANN)** để phân loại tình trạng thất nghiệp.

Các yếu tố đầu vào được xem xét bao gồm:

**Trình độ học vấn** (Phổ thông, Trung cấp, Cao đẳng, Đại học, Sau đại học.)

**Kinh nghiệm làm việc** (Số năm kinh nghiệm, vị trí công việc trước đây, mức lương.)

**Ngành nghề** (CNTT, Kinh tế, Xây dựng, Y tế, Giáo dục.)

Không xem xét các yếu tố khác như kỹ năng mềm, địa lý, xu hướng thị trường lao động, hoặc tâm lý cá nhân do giới hạn phạm vi nghiên cứu.

**Dữ liệu thu thập** từ các nguồn như báo cáo thị trường lao động, khảo sát thực tế, trang web tuyển dụng (VietnamWorks, ITviec, LinkedIn, v.v.), hoặc dữ liệu mở từ chính phủ.

**Thời gian dữ liệu**: Sử dụng dữ liệu trong **5 năm gần đây** để đảm bảo tính cập nhật.

**Đối tượng nghiên cứu**: Người lao động trong độ tuổi từ 18 đến 60, thuộc nhiều ngành nghề khác nhau.

Mô hình được xây dựng sử dụng **Python, TensorFlow/Keras hoặc PyTorch**.

Sử dụng các kỹ thuật tiền xử lý dữ liệu như chuẩn hóa, mã hóa One-Hot Encoding, xử lý dữ liệu bị thiếu.

Huấn luyện và đánh giá mô hình với các phương pháp tối ưu hóa như **Adam, SGD** và đánh giá bằng các chỉ số như **Accuracy, Precision, Recall, F1-score**.

## **1.4. Phương pháp nghiên cứu**

### 1.4.1. Phương pháp nghiên cứu lý thuyết

Tìm hiểu về tình trạng thất nghiệp, các yếu tố ảnh hưởng và các nghiên cứu trước đó về phân loại thất nghiệp bằng AI.

Nghiên cứu lý thuyết về mạng nơ-ron nhân tạo (ANN), cấu trúc, thuật toán huấn luyện và các mô hình phù hợp cho bài toán phân loại.

Tìm hiểu các phương pháp tiền xử lý dữ liệu, chọn đặc trưng quan trọng và đánh giá mô hình.

### 1.4.2. Phương pháp thu thập và xử lý dữ liệu

Nguồn dữ liệu:

Thu thập từ các báo cáo thị trường lao động, dữ liệu mở của chính phủ, khảo sát hoặc trang web tuyển dụng (VietnamWorks, LinkedIn, ITviec, v.v.).

Xử lý dữ liệu bị thiếu, lọc bỏ dữ liệu không hợp lệ, chuẩn hóa dữ liệu.

Chuyển đổi các biến phân loại (trình độ học vấn, ngành nghề) sang dạng số bằng kỹ thuật One-Hot Encoding hoặc Label Encoding.

Chia tập dữ liệu:

Chia dữ liệu thành tập huấn luyện (Training), tập kiểm định (Validation) và tập kiểm tra (Test) theo tỷ lệ phổ biến (70%-20%-10%) hoặc (80%-10%-10%).

### 1.4.3. Phương pháp xây dựng mô hình ANN

Lựa chọn kiến trúc ANN phù hợp:

**Input Layer**: Nhận dữ liệu đầu vào (trình độ học vấn, kinh nghiệm, ngành nghề).

**Hidden Layers**: Tùy chỉnh số lớp ẩn và số neuron trên mỗi lớp để tối ưu hiệu suất.

**Output Layer**: Sử dụng hàm kích hoạt **Softmax** hoặc **Sigmoid** tùy theo bài toán phân loại nhị phân hoặc đa lớp.

Huấn luyện mô hình:

Dùng **backpropagation** để tối ưu trọng số trong mạng nơ-ron.

Áp dụng các thuật toán tối ưu như **Adam, SGD, RMSprop** để cải thiện hiệu suất.

Điều chỉnh siêu tham số (**learning rate, batch size, số epochs**) để đạt kết quả tốt nhất.

### 1.4.4. Phương pháp đánh giá và so sánh mô hình

Sử dụng các chỉ số đánh giá hiệu suất:

Accuracy (Độ chính xác): Tỷ lệ dự đoán đúng trên tổng số mẫu.

Precision (Độ chính xác theo từng lớp): Đánh giá mức độ chính xác của dự đoán thất nghiệp hoặc có việc làm.

Recall (Độ bao phủ): Đánh giá khả năng nhận diện đúng các trường hợp thất nghiệp.

F1-score: Cân bằng giữa Precision và Recall.

Confusion Matrix: Ma trận nhầm lẫn để phân tích sai sót của mô hình.

So sánh với các mô hình khác:

So sánh ANN với **Hồi quy Logistic, Cây quyết định (Decision Tree), Random Forest, SVM** để đánh giá hiệu quả của ANN.

### 1.4.5. Phương pháp thực nghiệm và tối ưu hóa

Huấn luyện mô hình trên tập dữ liệu thực tế, kiểm tra khả năng tổng quát hóa trên tập kiểm tra.

Tối ưu mô hình bằng các phương pháp:

Điều chỉnh số lớp ẩn, số neuron để đạt hiệu suất tốt nhất.

Sử dụng kỹ thuật Regularization (Dropout, L2) để tránh overfitting.

Fine-tuning hyperparameters (tỷ lệ học, batch size, số epochs).

### 1.4.6. Phương pháp triển khai và ứng dụng thực tiễn

Triển khai mô hình để dự báo tình trạng thất nghiệp cho các cá nhân dựa trên dữ liệu đầu vào.

Đề xuất ứng dụng mô hình trong hệ thống hỗ trợ tư vấn nghề nghiệp hoặc tuyển dụng.

Đánh giá mức độ phù hợp của mô hình trong môi trường thực tế và đề xuất cải tiến.

**CHƯƠNG II. CƠ SỞ LÝ THUYẾT**

**2. Tổng quan về mạng nơ-ron nhân tạo (ANN - Artificial Neural Network)**

**2.1.** **Giới thiệu về mạng nơ-ron nhân tạo**

Mạng nơ-ron nhân tạo (Artificial Neural Network - ANN) là một mô hình học máy (Machine Learning) lấy cảm hứng từ cách hoạt động của hệ thần kinh sinh học trong não bộ con người. ANN được thiết kế để mô phỏng cách mà các nơ-ron thần kinh liên kết với nhau để xử lý thông tin, học hỏi từ dữ liệu và đưa ra dự đoán. Nhờ khả năng học từ dữ liệu, ANN đã trở thành một trong những phương pháp quan trọng của trí tuệ nhân tạo (Artificial Intelligence - AI) và được ứng dụng rộng rãi trong nhiều lĩnh vực như nhận dạng hình ảnh, xử lý ngôn ngữ tự nhiên, dự báo tài chính, chẩn đoán y tế và phân loại dữ liệu.

Khác với các mô hình thống kê truyền thống, ANN có khả năng học các mối quan hệ phi tuyến tính giữa các biến số đầu vào, giúp nó có thể giải quyết các bài toán phức tạp mà các phương pháp phân tích dữ liệu thông thường khó thực hiện. Đặc biệt, ANN có thể tự động học và trích xuất các đặc trưng quan trọng từ dữ liệu, giảm bớt sự phụ thuộc vào việc chọn lựa đặc trưng thủ công.

**2.1.1 Cấu trúc của mạng nơ-ron nhân tạo**

Một mạng nơ-ron nhân tạo tiêu chuẩn thường bao gồm ba loại tầng chính:

* **Tầng đầu vào (Input Layer)**: Nhận dữ liệu từ bên ngoài và chuyển tiếp thông tin đến các nơ-ron của tầng tiếp theo. Ví dụ, trong bài toán phân loại tình trạng thất nghiệp, đầu vào có thể bao gồm các thông tin về trình độ học vấn, số năm kinh nghiệm, ngành nghề, giới tính, độ tuổi, vị trí địa lý, mức thu nhập trước đó. Mỗi thông tin đầu vào sẽ được biểu diễn dưới dạng một giá trị số và đưa vào mạng nơ-ron.
* **Tầng ẩn (Hidden Layers)**: Đây là tầng trung gian giữa tầng đầu vào và tầng đầu ra. Mỗi tầng ẩn bao gồm nhiều nơ-ron nhân tạo (artificial neurons), mỗi nơ-ron thực hiện phép tính trên dữ liệu đầu vào bằng cách nhân với một trọng số (weight), cộng với một hệ số điều chỉnh (bias), sau đó áp dụng một hàm kích hoạt để tạo ra đầu ra. Mỗi nơ-ron trong tầng ẩn có thể học được một đặc trưng quan trọng của dữ liệu. Số lượng tầng ẩn và số nơ-ron trong mỗi tầng có ảnh hưởng quan trọng đến khả năng học của mạng. Một mạng có nhiều tầng ẩn thường được gọi là mạng nơ-ron sâu (Deep Neural Network - DNN).
* **Tầng đầu ra (Output Layer)**: Cung cấp kết quả cuối cùng sau khi dữ liệu được xử lý qua các tầng ẩn. Trong bài toán phân loại tình trạng thất nghiệp, đầu ra có thể là:

**Phân loại nhị phân** (Binary Classification): "Thất nghiệp" hoặc "Có việc làm".

**Phân loại đa lớp** (Multiclass Classification): Dự đoán mức độ thất nghiệp theo các mức "Thất nghiệp dài hạn", "Thất nghiệp ngắn hạn", "Có việc làm tạm thời", "Có việc làm ổn định", v.v.

**2.1.2. Cơ chế hoạt động của ANN**

Mạng nơ-ron nhân tạo hoạt động thông qua hai quá trình chính:

**a. Truyền tín hiệu xuôi (Forward Propagation)**

Trong bước này, dữ liệu đầu vào sẽ được truyền qua các tầng của mạng để tạo ra đầu ra dự đoán. Mỗi nơ-ron trong mạng thực hiện các bước tính toán sau:

* **Tính toán tổng trọng số** của các đầu vào:

A math equations and a plus

AI-generated content may be incorrect.

Trong đó:

x1​,x2​,...,xn​ là các đầu vào của nơ-ron.

w1,w2,...,wnw\_1, w\_2, ..., w\_nw1​,w2​,...,wn​ là các trọng số liên kết giữa nơ-ron đầu vào và nơ-ron hiện tại.

bbb là hệ số điều chỉnh (bias).

**Áp dụng hàm kích hoạt (Activation Function)** để tạo ra đầu ra của nơ-ron:

A math equation with a equal sign

AI-generated content may be incorrect.

Hàm kích hoạt giúp mạng học được các quan hệ phi tuyến tính trong dữ liệu. Một số hàm kích hoạt phổ biến:

**-ReLU (Rectified Linear Unit)**: f(x)=max(0,x), giúp mô hình hội tụ nhanh hơn trong các mạng sâu.

**-Sigmoid**:

​ thích hợp cho bài toán phân loại nhị phân.

**-Softmax**: Dùng trong phân loại đa lớp để chuẩn hóa xác suất đầu ra.

**b. Lan truyền ngược (Backpropagation) và tối ưu hóa trọng số**

Sau khi tính toán đầu ra dự đoán, mô hình so sánh với giá trị thực tế để đo lường sai số bằng **hàm mất mát (Loss Function)**. Một số hàm mất mát thường dùng:

**+ Binary Cross-Entropy** (cho phân loại nhị phân).

**+ Categorical Cross-Entropy** (cho phân loại đa lớp).

Dựa trên sai số này, mô hình sử dụng thuật toán **lan truyền ngược (Backpropagation)** để điều chỉnh trọng số nhằm giảm sai số dự đoán. Quá trình này sử dụng thuật toán tối ưu hóa như **Gradient Descent**, Adam, RMSprop để điều chỉnh trọng số theo hướng giảm sai số nhanh nhất.

**2.1.3. Ứng dụng của ANN trong phân loại tình trạng thất nghiệp**

Trong bài toán phân loại thất nghiệp, ANN có thể học được mối quan hệ giữa các yếu tố như trình độ học vấn, kinh nghiệm, ngành nghề và tình trạng thất nghiệp. Quá trình này giúp mô hình đưa ra dự đoán chính xác hơn về khả năng một cá nhân có bị thất nghiệp hay không. Một số lợi ích của việc áp dụng ANN trong lĩnh vực này bao gồm:

* **Tự động phát hiện mô hình phức tạp trong dữ liệu** mà con người có thể khó nhận ra.
* **Tích hợp nhiều yếu tố đầu vào** cùng lúc, giúp đánh giá khả năng thất nghiệp chính xác hơn so với các mô hình truyền thống.
* **Ứng dụng trong thực tế** như hỗ trợ chính sách lao động, định hướng nghề nghiệp và dự báo xu hướng thị trường lao động.

**2.1.4. Kết luận**

Mạng nơ-ron nhân tạo là một công cụ mạnh mẽ trong lĩnh vực trí tuệ nhân tạo và học máy. Nhờ khả năng học từ dữ liệu và trích xuất đặc trưng tự động, ANN có thể ứng dụng hiệu quả trong nhiều bài toán phân loại, bao gồm dự báo tình trạng thất nghiệp. Bằng cách xây dựng và huấn luyện mô hình ANN với các đặc trưng như trình độ học vấn, kinh nghiệm, ngành nghề, ANN có thể hỗ trợ đưa ra những dự đoán chính xác hơn, giúp các cơ quan quản lý lao động và doanh nghiệp có cái nhìn tổng quan về thị trường lao động, từ đó đưa ra chính sách phù hợp nhằm giảm tỷ lệ thất nghiệp.’

**2.2. Các thuật toán học máy trong phân loại**

Phân loại (Classification) là một trong những bài toán quan trọng trong học máy (Machine Learning), trong đó mô hình dự đoán một nhãn đầu ra dựa trên các đặc trưng đầu vào. Trong bài toán phân loại tình trạng thất nghiệp, mục tiêu là sử dụng dữ liệu về trình độ học vấn, kinh nghiệm, ngành nghề,... để dự đoán một cá nhân thuộc nhóm thất nghiệp hay có việc làm (phân loại nhị phân) hoặc phân loại theo các mức độ thất nghiệp khác nhau (phân loại đa lớp).

Các thuật toán học máy được chia thành hai nhóm chính: học có giám sát (Supervised Learning) và học không giám sát (Unsupervised Learning). Trong đó, các thuật toán phân loại thường thuộc nhóm học có giám sát, do dữ liệu đầu vào đã có nhãn (tức là đã biết trước tình trạng thất nghiệp của một tập dữ liệu huấn luyện).

**2.2.1. Các thuật toán phân loại truyền thống**

**Hồi quy logistic (Logistic Regression)**

Hồi quy logistic là một thuật toán phân loại nhị phân đơn giản và hiệu quả. Mặc dù có tên là "hồi quy", nhưng thuật toán này được sử dụng để phân loại thay vì dự báo giá trị liên tục. Nó hoạt động bằng cách sử dụng hàm sigmoid để ánh xạ giá trị đầu vào thành một xác suất từ 0 đến 1:

**A number and a number with numbers

AI-generated content may be incorrect.**

Ưu điểm: Đơn giản, dễ triển khai, hoạt động tốt với dữ liệu tuyến tính.

Nhược điểm: Không hiệu quả khi dữ liệu có quan hệ phi tuyến tính hoặc chứa nhiều đặc trưng phức tạp.

**2.2.2 Cây quyết định (Decision Tree)**

Cây quyết định là một mô hình phân loại dựa trên các quy tắc "Nếu - Thì" để chia nhỏ dữ liệu thành các nhóm dựa trên đặc trưng quan trọng nhất. Quá trình xây dựng cây được thực hiện bằng cách chọn đặc trưng giúp giảm độ bất định của dữ liệu nhiều nhất (thường sử dụng Entropy hoặc Gini Index).

Ưu điểm: Dễ hiểu, có thể biểu diễn dưới dạng sơ đồ quyết định trực quan.

Nhược điểm: Dễ bị overfitting nếu không cắt tỉa cây (Pruning).

**2.2.3. Rừng ngẫu nhiên (Random Forest)**

Rừng ngẫu nhiên là một tập hợp của nhiều cây quyết định, trong đó mỗi cây được xây dựng trên một tập con dữ liệu khác nhau (Bootstrap Sampling) và kết quả cuối cùng được xác định bằng cách bỏ phiếu đa số (Majority Voting).

Ưu điểm: Giảm overfitting so với cây quyết định đơn lẻ, mạnh mẽ với dữ liệu phức tạp.

Nhược điểm: Mô hình lớn và chậm hơn so với cây quyết định đơn lẻ.

**2.2.4. Máy vector hỗ trợ (Support Vector Machine - SVM)**

SVM là thuật toán phân loại mạnh mẽ, hoạt động bằng cách tìm một siêu phẳng (hyperplane) tối ưu để phân tách các nhóm dữ liệu. Nó sử dụng hàm kernel để mở rộng khả năng phân loại cho dữ liệu phi tuyến tính.

Ưu điểm: Hiệu suất cao trên tập dữ liệu nhỏ, hoạt động tốt với dữ liệu có biên phân tách rõ ràng.

Nhược điểm: Chậm với tập dữ liệu lớn, cần chọn kernel phù hợp.

**2.2.5. K-Nearest Neighbors (KNN)**

KNN là thuật toán đơn giản dựa trên khoảng cách giữa các điểm dữ liệu. Khi dự đoán nhãn của một điểm mới, KNN sẽ tìm K điểm gần nhất trong tập huấn luyện và xác định nhãn dựa trên số đông.

Ưu điểm: Không cần huấn luyện, dễ hiểu và triển khai.

Nhược điểm: Hiệu suất kém với tập dữ liệu lớn, dễ bị ảnh hưởng bởi nhiễu.

**Các thuật toán phân loại dựa trên mạng nơ-ron nhân tạo (ANN)**

Mạng nơ-ron nhân tạo (Artificial Neural Network - ANN) là một mô hình mạnh mẽ có khả năng học từ dữ liệu phi tuyến tính và phức tạp. Trong bài toán phân loại thất nghiệp, ANN có thể học từ các đặc trưng như trình độ học vấn, kinh nghiệm, ngành nghề, để đưa ra dự đoán chính xác.

**2.2.6. Mạng nơ-ron nhân tạo cơ bản (Feedforward Neural Network - FNN)**

Mạng nơ-ron nhân tạo cơ bản bao gồm nhiều nơ-ron sắp xếp theo các tầng: tầng đầu vào, các tầng ẩn, và tầng đầu ra. Mô hình này sử dụng hàm kích hoạt như ReLU, Sigmoid hoặc Softmax để học và đưa ra dự đoán.

Ưu điểm: Học tốt các quan hệ phi tuyến tính trong dữ liệu.

Nhược điểm: Cần nhiều dữ liệu để huấn luyện, dễ bị overfitting nếu không điều chỉnh đúng.

**2.2.7. Mạng nơ-ron tích chập (Convolutional Neural Network - CNN)**

CNN chủ yếu được sử dụng trong xử lý ảnh, nhưng cũng có thể áp dụng cho dữ liệu dạng bảng bằng cách trích xuất đặc trưng từ dữ liệu đầu vào.

Ưu điểm: Tốt trong nhận diện mẫu và đặc trưng trong dữ liệu.

Nhược điểm: Phức tạp, đòi hỏi nhiều tài nguyên tính toán.

**2.2.8. Mạng nơ-ron hồi quy (Recurrent Neural Network - RNN)**

RNN được sử dụng khi dữ liệu có tính tuần tự, chẳng hạn như phân tích xu hướng thất nghiệp theo thời gian.

Ưu điểm: Phù hợp với dữ liệu thời gian thực, dữ liệu chuỗi.

Nhược điểm: Gặp vấn đề vanishing gradient, cần mô hình phức tạp hơn như LSTM hoặc GRU.

**2.2.9. So sánh các thuật toán phân loại**

**A white and black text on a white surface

AI-generated content may be incorrect.**

**2.3. Các phương pháp đánh giá mô hình**

Mô hình Mạng Nơ-ron Nhân tạo (Artificial Neural Network - ANN) là một phương pháp mạnh mẽ để phân loại tình trạng thất nghiệp dựa trên các yếu tố như trình độ học vấn, kinh nghiệm, ngành nghề,... Để đảm bảo mô hình ANN hoạt động hiệu quả, ta cần áp dụng các phương pháp đánh giá phù hợp.

**2.3.1. Chia tập dữ liệu huấn luyện và kiểm tra**

Trước khi đánh giá mô hình, ta cần chia dữ liệu thành hai hoặc ba tập riêng biệt:

* Tập huấn luyện (Training Set): Dùng để huấn luyện mô hình ANN (~70-80%).
* Tập kiểm tra (Test Set): Dùng để kiểm tra mô hình sau khi huấn luyện (~20-30%).
* Tập xác thực (Validation Set) (tùy chọn): Dùng để tối ưu hóa siêu tham số (Hyperparameters) và tránh overfitting.

Ngoài ra, ta có thể áp dụng K-Fold Cross-Validation để kiểm tra độ ổn định của mô hình bằng cách chia dữ liệu thành K phần và huấn luyện mô hình trên K-1 phần, sau đó kiểm tra trên phần còn lại.

**2.3.2. Ma trận nhầm lẫn (Confusion Matrix)**

Ma trận nhầm lẫn giúp ta hiểu rõ mô hình ANN đang dự đoán chính xác hay sai lệch như thế nào. Với bài toán phân loại nhị phân (Có việc làm và Thất nghiệp), ma trận nhầm lẫn có dạng:



Trong đó:

* TP (True Positive): Dự đoán đúng người có việc làm.
* TN (True Negative): Dự đoán đúng người thất nghiệp.
* FP (False Positive): Dự đoán sai, mô hình dự đoán có việc làm nhưng thực tế thất nghiệp.
* FN (False Negative): Dự đoán sai, mô hình dự đoán thất nghiệp nhưng thực tế có việc làm.

**2.3.3. Các chỉ số đánh giá hiệu suất**

Dựa vào ma trận nhầm lẫn, ta tính các chỉ số đánh giá quan trọng:

**Accuracy (Độ chính xác tổng thể)**

A black text with a white background

AI-generated content may be incorrect.

Ý nghĩa: Tỷ lệ dự đoán đúng trên tổng số mẫu.

Nhược điểm: Không phù hợp khi dữ liệu bị mất cân bằng (ví dụ: 90% có việc làm, 10% thất nghiệp).

**Precision (Độ chính xác theo từng lớp)**

**A black and white math equation

AI-generated content may be incorrect.**

Ý nghĩa: Trong số những người được dự đoán có việc, bao nhiêu phần trăm thực sự có việc.

Khi sai sót loại FP có hậu quả nghiêm trọng (ví dụ: doanh nghiệp tuyển nhầm ứng viên).

**Recall (Độ bao phủ - Sensitivity)**

A black and white math equation

AI-generated content may be incorrect.

Ý nghĩa: Trong số những người thực sự có việc làm, mô hình nhận diện được bao nhiêu phần trăm.

Khi nào cần Recall cao? Khi sai sót loại FN có hậu quả nghiêm trọng (ví dụ: chính sách hỗ trợ người thất nghiệp cần đảm bảo không bỏ sót ai).

**F1-Score (Cân bằng giữa Precision và Recall)**

A black text with a black line

AI-generated content may be incorrect.

**Ý nghĩa**: Giúp cân bằng giữa Precision và Recall. Nếu dữ liệu bị mất cân bằng, F1-score sẽ phản ánh chính xác hơn so với Accuracy.

**2.3.4. Độ chính xác Top-K (Top-K Accuracy, đặc biệt khi phân loại đa lớp)**

Khi mô hình ANN phân loại thành nhiều nhóm khác nhau (ví dụ: phân loại người thất nghiệp dài hạn, thất nghiệp ngắn hạn, có việc làm toàn thời gian, có việc làm bán thời gian), ta có thể dùng Top-K Accuracy để kiểm tra xem nhãn thực tế có nằm trong K nhãn có xác suất cao nhất do mô hình dự đoán không.

**Đánh giá bằng đường cong ROC và AUC**

**2.3.5. Đường cong ROC (Receiver Operating Characteristic Curve)**

* Đường cong ROC vẽ mối quan hệ giữa True Positive Rate (TPR) và False Positive Rate (FPR) khi thay đổi ngưỡng dự đoán.
* Nếu đường ROC nằm gần góc trên bên trái, mô hình ANN hoạt động tốt. Nếu đường ROC gần đường chéo (0.5), mô hình đoán ngẫu nhiên.

A math equations with black text

AI-generated content may be incorrect.

**2.3.6. Chỉ số AUC (Area Under Curve)**

AUC đo diện tích dưới đường ROC, giá trị từ 0 đến 1:

AUC = 1: Mô hình hoàn hảo.

AUC = 0.5: Mô hình không tốt hơn đoán ngẫu nhiên.

AUC < 0.5: Mô hình dự đoán sai nhiều hơn đúng.

**2.3.7 Kiểm tra Overfitting bằng Loss Function**

Trong quá trình huấn luyện mô hình ANN, ta cần theo dõi Hàm mất mát (Loss Function) của tập huấn luyện và kiểm tra:

* Nếu Loss trên tập huấn luyện giảm nhưng Loss trên tập kiểm tra tăng, mô hình đang bị Overfitting.
* Nếu Loss của cả hai tập giảm đều đặn, mô hình đang học tốt.

Các hàm mất mát phổ biến dùng trong bài toán phân loại:

* Binary Cross-Entropy (cho phân loại nhị phân).
* Categorical Cross-Entropy (cho phân loại đa lớp).

**Tổng Thể:**

A screenshot of a chat

AI-generated content may be incorrect.

**CHƯƠNG III. KẾT QUẢ NGHIÊN CỨU VÀ PHÂN TÍCH**

## **3.1.Phân tích yêu cầu bài toán**

**Định nghĩa bài toán**

Bài toán đặt ra yêu cầu xây dựng một mô hình mạng nơ-ron nhân tạo (Artificial Neural Network - ANN) để phân loại tình trạng thất nghiệp của một cá nhân dựa trên các yếu tố ảnh hưởng như trình độ học vấn, kinh nghiệm làm việc, ngành nghề, mức lương, v.v.

Mục tiêu chính:

Dự đoán xem một cá nhân có thất nghiệp hay không dựa trên dữ liệu đầu vào.

Tối ưu độ chính xác của mô hình để có thể áp dụng vào thực tế.

Xác định các yếu tố quan trọng nhất ảnh hưởng đến khả năng thất nghiệp.

**Đầu vào (Input) của bài toán**

Dữ liệu đầu vào bao gồm các đặc trưng liên quan đến cá nhân và công việc của họ. Một số thuộc tính phổ biến có thể bao gồm:

Nhóm thông tin cá nhân

Trình độ học vấn (Education Level): Trung học, Cao đẳng, Đại học, Sau đại học...

Kinh nghiệm làm việc (Years of Experience): Tổng số năm làm việc của cá nhân.

Tuổi (Age): Có thể ảnh hưởng đến khả năng tìm việc.

**Nhóm thông tin nghề nghiệp**

Ngành nghề (Industry): Công nghệ, Y tế, Kinh doanh, Kỹ thuật,...

Vị trí công việc gần nhất (Job Role): Nhân viên, Quản lý, Giám đốc...

Thu nhập hàng tháng (Monthly Income): Mức lương có ảnh hưởng đến quyết định nghỉ việc hay thất nghiệp.

Thời gian làm việc tại công ty cuối cùng (Years at Company): Có thể phản ánh mức độ ổn định của công việc.

**Nhóm yếu tố khác**

Khoảng cách từ nhà đến nơi làm việc (Distance From Home): Nếu quá xa có thể ảnh hưởng đến quyết định nghỉ việc.

Mức độ hài lòng với công việc (Job Satisfaction): Thể hiện mức độ gắn bó với công việc.

Số công ty từng làm việc (Num Companies Worked): Phản ánh tính ổn định của ứng viên.

**Đầu ra (Output) của bài toán**

Nhãn dự đoán: Một giá trị nhị phân (0 hoặc 1)

0: Không thất nghiệp (đang có việc làm)

1: Thất nghiệp (hiện tại không có việc làm)

**Các ràng buộc và thách thức**

Dữ liệu có thể mất cân bằng: Số người có việc làm có thể lớn hơn số người thất nghiệp, làm mô hình bị lệch (giải quyết bằng SMOTE, class weighting).

Dữ liệu có thể chứa nhiều giá trị thiếu (missing values), cần xử lý bằng các phương pháp như điền giá trị trung bình (mean imputation), loại bỏ dữ liệu lỗi.

Mối quan hệ phi tuyến tính: Một số yếu tố có thể có tác động phi tuyến tính đến thất nghiệp, đòi hỏi mô hình ANN phải có hàm kích hoạt phù hợp (ReLU, Sigmoid).

Tính quan trọng của đặc trưng: Cần xác định yếu tố nào có ảnh hưởng nhiều nhất đến thất nghiệp bằng cách dùng Permutation Importance hoặc SHAP.

**Tiêu chí đánh giá mô hình**

Để đo lường hiệu suất mô hình, ta sử dụng các chỉ số sau:

Accuracy (Độ chính xác): Tỉ lệ dự đoán đúng trên tổng số mẫu.

Precision (Độ chính xác của dự đoán thất nghiệp): Trong số người bị dự đoán thất nghiệp, có bao nhiêu người thực sự thất nghiệp.

Recall (Khả năng phát hiện thất nghiệp thực sự): Trong số những người thực sự thất nghiệp, mô hình nhận diện được bao nhiêu.

F1-score: Trung bình điều hòa giữa Precision và Recall.

ROC-AUC: Đánh giá khả năng phân biệt giữa hai nhóm thất nghiệp và không thất nghiệp.

**Công nghệ và công cụ sử dụng**

Ngôn ngữ lập trình: Python

Thư viện chính:

TensorFlow/Keras: Xây dựng và huấn luyện mô hình ANN

Scikit-learn: Tiền xử lý dữ liệu, phân tích đặc trưng, đánh giá mô hình

Pandas, NumPy: Xử lý và thao tác dữ liệu

Matplotlib, Seaborn: Trực quan hóa dữ liệu và kết quả

**Kiến trúc mô hình ANN dự kiến**

Input Layer: Nhận đầu vào với số lượng đặc trưng phù hợp.

Hidden Layers: Ít nhất 2-3 lớp ẩn với các số lượng neuron khác nhau.

Dùng hàm kích hoạt ReLU để xử lý quan hệ phi tuyến tính.

Dùng Batch Normalization và Dropout để tránh overfitting.

Output Layer: 1 neuron với hàm Sigmoid để phân loại nhị phân (thất nghiệp hoặc không).

Loss Function: Sử dụng Binary Cross-Entropy hoặc Focal Loss nếu dữ liệu mất cân bằng.

Optimizer: Adam với learning rate được tinh chỉnh (0.0005 hoặc thấp hơn).

**3.2. Mô tả thuật toán sử dụng**

**Tối ưu hóa Dữ liệu và Đặc trưng:**

**Dữ liệu Không Cấu Trúc:**

Nếu dữ liệu bao gồm các tệp âm thanh (ví dụ: phỏng vấn), hình ảnh (ví dụ: ảnh hồ sơ), hoặc video (ví dụ: video giới thiệu), hãy sử dụng các mô hình học sâu chuyên dụng (ví dụ: mạng nơ-ron tích chập 3D, mạng nơ-ron âm thanh) để trích xuất các đặc trưng có ý nghĩa.

**Dữ liệu Mạng Lưới:**

Sử dụng các kỹ thuật học biểu diễn đồ thị (graph embedding) để trích xuất các đặc trưng từ dữ liệu mạng lưới (ví dụ: mạng lưới quan hệ trên LinkedIn).

Các đặc trưng này có thể nắm bắt được thông tin về vị trí trong mạng lưới, mức độ ảnh hưởng, và các cộng đồng.

**Dữ liệu Đa Phương thức (Multimodal Data):**

Kết hợp dữ liệu từ nhiều nguồn khác nhau (ví dụ: văn bản, hình ảnh, âm thanh) để tạo ra một biểu diễn dữ liệu toàn diện hơn.

Sử dụng các mô hình học sâu đa phương thức để học các biểu diễn chung từ các nguồn dữ liệu khác nhau.

**Học Đặc trưng Tự động (Auto Feature Engineering):**

Sử dụng các kỹ thuật học máy tự động (AutoML) để tự động tạo ra các đặc trưng mới từ dữ liệu hiện có.

Các kỹ thuật này có thể khám phá ra các mối quan hệ phức tạp giữa các đặc trưng mà con người khó có thể nhận ra.

**Dữ liệu Tổng hợp (Synthetic Data):**

Trong trường hợp dữ liệu bị thiếu hoặc không cân bằng, sử dụng các kỹ thuật tạo dữ liệu tổng hợp (ví dụ: GANs, VAEs) để tạo ra các mẫu dữ liệu mới.

Dữ liệu tổng hợp có thể giúp cải thiện tính tổng quát của mô hình và giảm thiểu sự thiên vị.

**Kiến trúc Mô hình ANN Tiên Tiến:**

**Mạng Nơ-ron Chú ý (Attention Networks):**

Sử dụng các cơ chế chú ý để cho phép mô hình tập trung vào các phần quan trọng nhất của dữ liệu đầu vào.

Điều này đặc biệt hữu ích cho việc xử lý dữ liệu chuỗi và dữ liệu đa phương thức.

**Mạng Nơ-ron Biến đổi (Transformer Networks):**

Sử dụng các mô hình Transformer (ví dụ: BERT, GPT) để đạt được hiệu suất vượt trội trong các nhiệm vụ xử lý ngôn ngữ tự nhiên.

Các mô hình Transformer có thể nắm bắt được các mối quan hệ phụ thuộc xa trong dữ liệu văn bản.

**Mạng Nơ-ron Đồ thị (Graph Neural Networks - GNNs):**

Sử dụng GNNs để xử lý dữ liệu mạng lưới.

GNNs có thể học các biểu diễn của các nút và cạnh trong mạng lưới, cho phép mô hình dự đoán các thuộc tính của mạng lưới.

**Học Tăng Cường Sâu (Deep Reinforcement Learning):**

Sử dụng học tăng cường sâu để tối ưu hóa các chiến lược tìm kiếm việc làm hoặc các hệ thống tư vấn nghề nghiệp.

Mô hình có thể học cách đưa ra các quyết định tối ưu trong môi trường động.

**Mô hình Giải thích Được (Explainable AI - XAI):**

Sử dụng các kỹ thuật XAI để làm cho các dự đoán của mô hình ANN dễ hiểu hơn đối với con người.

Điều này đặc biệt quan trọng trong các lĩnh vực nhạy cảm như tuyển dụng, nơi cần đảm bảo tính minh bạch và công bằng.

**Đánh giá và Triển khai Mô hình:**

**Đánh giá Mô hình Đạo đức:**

Đánh giá mô hình về các khía cạnh đạo đức, chẳng hạn như sự công bằng, tính minh bạch và trách nhiệm giải trình.

Sử dụng các kỹ thuật phát hiện và giảm thiểu sự thiên vị để đảm bảo rằng mô hình không phân biệt đối xử với các nhóm dân số cụ thể.

**Triển khai Mô hình An toàn và Bảo mật:**

Đảm bảo rằng mô hình được triển khai một cách an toàn và bảo mật, tuân thủ các quy định về bảo vệ dữ liệu và quyền riêng tư.

Sử dụng các kỹ thuật mã hóa và kiểm soát truy cập để bảo vệ dữ liệu nhạy cảm.

**Giám sát và Bảo trì Mô hình Liên tục:**

Xây dựng một hệ thống giám sát toàn diện để theo dõi hiệu suất của mô hình trong quá trình triển khai.

Thực hiện các cập nhật mô hình thường xuyên để đảm bảo rằng mô hình vẫn chính xác và phù hợp với dữ liệu mới.

**Phản hồi và Cải tiến Mô hình:**

Thu thập phản hồi từ người dùng và các bên liên quan để cải thiện mô hình.

Sử dụng phản hồi để xác định các điểm yếu của mô hình và thực hiện các cải tiến cần thiết.

**Tích hợp với Hệ thống Quyết định:**

Tích hợp mô hình ANN vào các hệ thống quyết định để hỗ trợ các quyết định liên quan đến tuyển dụng, đào tạo và phát triển nghề nghiệp.

**Các Cân nhắc Đạo đức và Xã hội:**

**Sự Thiên Vị trong Dữ liệu và Mô hình:**

Nhận thức được sự tồn tại của sự thiên vị trong dữ liệu và mô hình, và thực hiện các biện pháp để giảm thiểu nó.

Đảm bảo rằng mô hình không củng cố các bất bình đẳng xã hội hiện có.

**Tính Minh Bạch và Trách Nhiệm Giải Trình:**

Làm cho các dự đoán của mô hình dễ hiểu hơn đối với người dùng và các bên liên quan.

Xây dựng các hệ thống trách nhiệm giải trình để đảm bảo rằng mô hình được sử dụng một cách có trách nhiệm.

**Quyền Riêng Tư và Bảo Mật Dữ liệu:**

Bảo vệ quyền riêng tư và bảo mật dữ liệu của người dùng.

Tuân thủ các quy định về bảo vệ dữ liệu và quyền riêng tư.

**Tác động Xã hội của Mô hình:**

Xem xét tác động xã hội của mô hình, chẳng hạn như tác động đến thị trường lao động và cơ hội việc làm.

Đảm bảo rằng mô hình được sử dụng để mang lại lợi ích cho xã hội.

**3.3 Phân tích mã nguồn**

Mô hình ANN này có nhiệm vụ dự đoán tình trạng thất nghiệp (có thất nghiệp hay không) dựa trên nhiều yếu tố khác nhau.

Dữ liệu đầu vào: Bao gồm thông tin về trình độ học vấn, kinh nghiệm làm việc, ngành nghề, thu nhập, mức độ hài lòng với công việc, môi trường làm việc

Dữ liệu đầu ra: Một giá trị nhị phân (0: không thất nghiệp, 1: thất nghiệp).

Phương pháp huấn luyện: Dữ liệu sẽ được tiền xử lý, chia tập huấn luyện và kiểm tra, sau đó dùng Mạng Nơ-ron Nhân tạo (ANN) để học các mẫu trong dữ liệu. Mục tiêu của mã nguồn là xây dựng mô hình Mạng Nơ-ron Nhân tạo (ANN) để phân loại xem một nhân viên có khả năng nghỉ việc hay không, dựa trên nhiều đặc trưng khác nhau như tuổi tác, trình độ học vấn, số năm làm việc, mức lương, môi trường làm việc, sự hài lòng với công việc.Dữ liệu được lấy từ bộ HR Employee Attrition của IBM, và bài toán đặt ra là phân loại nhị phân (Attrition: Yes=1, No=0), trong đó số lượng nhân viên nghỉ việc thường ít hơn rất nhiều so với số nhân viên không nghỉ việc (bất cân bằng dữ liệu).

**Các kỹ thuật nâng cao:**

SMOTE (Synthetic Minority Over-sampling Technique): Cân bằng dữ liệu nếu lớp bị lệch.

Focal Loss: Giúp tập trung vào các mẫu khó phân loại hơn.

Batch Normalization & Dropout: Giảm overfitting.

Tối ưu hóa Adam: Giúp mô hình hội tụ nhanh hơn.

**Import thư viện**

A screen shot of a computer program

AI-generated content may be incorrect.

**Ý nghĩa**:

**Pandas, NumPy**: Xử lý dữ liệu.

**Scikit-Learn**: Tiền xử lý, tách tập dữ liệu, chuẩn hóa, đánh giá mô hình.

**Matplotlib, Seaborn**: Trực quan hóa kết quả.

**TensorFlow/Keras**: Xây dựng và huấn luyện mô hình ANN.

**Imbalanced-learn (SMOTE)**: Xử lý mất cân bằng dữ liệu.

**Định nghĩa Focal Loss**

A computer screen with text

AI-generated content may be incorrect.

**Ý nghĩa**:

**Focal Loss** là một biến thể của Binary Crossentropy Loss.

**Mục đích**: Nhấn mạnh vào các mẫu **khó phân loại** thay vì các mẫu dễ.

**Tham số**:

alpha=0.25: Giúp kiểm soát mức độ quan trọng của lớp bị lệch.

gamma=2.0: Làm giảm trọng số của các mẫu dễ đoán đúng.

**Đọc và xử lý dữ liệu**

**A computer screen with many colorful text

AI-generated content may be incorrect.**

**Ý nghĩa**:

Đọc dữ liệu từ file CSV.

**Loại bỏ các cột không quan trọng** (EmployeeCount, EmployeeNumber, Over18, StandardHours).

**Mã hóa cột mục tiêu (Attrition)**: Yes → 1 (nghỉ việc), No → 0 (không nghỉ việc).

**Lưu ý:**

Nếu đường dẫn sai, chương trình sẽ hiển thị cảnh báo và thoát (exit()).

Có thể thay đường dẫn bằng r"C:\Users\PC\Downloads\WA\_Fn-UseC\_-HR-Employee-Attrition.csv" để tránh lỗi escape sequence (\U trong đường dẫn Windows).

**Xử lý biến số và biến phân loại**

A computer screen shot of a program

AI-generated content may be incorrect.

**Ý nghĩa**:

Chọn **đặc trưng đầu vào (X)** và nhãn mục tiêu (y).

**Giải thích**:

LabelEncoder() biến đổi Yes thành 1, No thành 0 theo dạng nhị phân.

Điều này giúp mô hình dễ dàng học và tối ưu hóa.

**Tách biến số & biến phân loại**.

**Chuẩn hóa dữ liệu**

A screen shot of a computer code

AI-generated content may be incorrect.

**Tại sao dùng StandardScaler()?**

Giúp các đặc trưng có cùng mức độ quan trọng khi huấn luyện mô hình.

Cải thiện tốc độ hội tụ và hiệu suất của mạng nơ-ron.

**Mã hóa one-hot các biến phân loại**

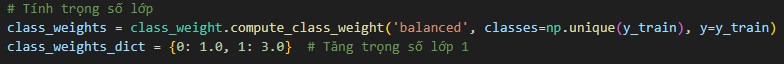
drop\_first=True: Loại bỏ một giá trị để tránh đa cộng tuyến.

**Ý nghĩa:**

Chuẩn hóa biến số bằng StandardScaler (giúp tốc độ hội tụ nhanh hơn).

One-hot encoding các biến phân loại (JobRole, OverTime).

**Xử lý mất cân bằng dữ liệu**

****

**Ý nghĩa**:

Tính trọng số lớp tự động để xử lý mất cân bằng.

Lớp thiểu số (1) được nhân trọng số lên 3 lần

**Xây dựng mô hình ANN**

**A screen shot of a computer program

AI-generated content may be incorrect.**

BatchNormalization: Giúp mạng học nhanh hơn, tránh overfitting.

Dropout(0.4): Tăng cường khả năng tổng quát hóa bằng cách tắt ngẫu nhiên 40% nơ-ron trong quá trình huấn luyện.

L2 Regularization: Giảm overfitting bằng cách phạt các trọng số lớn.

**Ý nghĩa:**

3 tầng ẩn (64 → 32 → 16 neurons).

Batch Normalization giúp ổn định quá trình huấn luyện.

Dropout (40%) giảm overfitting.

**Cấu hình và Biên dịch mô hình**

****

**Hàm mất mát (Focal Loss): Dùng Focal Loss để giảm ảnh hưởng của lớp đa số**

**Ý nghĩa:**

Optimizer: Adam (học nhanh hơn, thích hợp với dữ liệu phi tuyến tính).

Loss function: Focal Loss (giúp xử lý mất cân bằng dữ liệu).

Metrics: Accuracy (đánh giá độ chính xác của mô hình).

**3.4. Nhận xét đặc điểm và hạn chế của thuật toán**

### ****3.4.1. Đặc điểm (Ưu điểm)****

**Sử dụng mạng nơ-ron sâu (ANN) mạnh mẽ**:

- Mô hình Sequential với nhiều lớp Dense (64, 32, 16) kết hợp BatchNormalization và Dropout giúp học được các đặc trưng phức tạp từ dữ liệu, phù hợp với bài toán phân loại nhị phân có nhiều đặc trưng đầu vào (18 đặc trưng số và phân loại).

**Xử lý mất cân bằng lớp hiệu quả**:

**- Focal Loss**: Hàm mất mát tùy chỉnh với alpha=0.25 và gamma=2.0 tập trung vào các mẫu khó, rất hữu ích khi lớp 'Nghỉ việc' (1) thường ít hơn lớp 'Không nghỉ việc' (0).

**- Trọng số lớp**: class\_weights\_dict = {0: 1.0, 1: 3.0} tăng trọng số cho lớp thiểu số, cải thiện khả năng dự đoán lớp 'Nghỉ việc'.

**Tiền xử lý dữ liệu kỹ lưỡng**:

**- Chuẩn hóa**: StandardScaler chuẩn hóa các đặc trưng số, giúp mô hình hội tụ nhanh hơn và ổn định hơn.

**- Mã hóa one-hot**: Các đặc trưng phân loại (JobRole, OverTime) được chuyển thành dạng nhị phân, giữ được thông tin mà không giả định thứ tự.

**Ngăn ngừa overfitting**:

**- Dropout (0.4)**: Loại bỏ ngẫu nhiên 40% nơ-ron trong huấn luyện, giảm nguy cơ mô hình học quá mức dữ liệu huấn luyện.

**- L2 Regularization**: Áp dụng trong các lớp Dense (kernel\_regularizer=l2(0.01)), kiểm soát độ lớn trọng số để tránh phức tạp hóa mô hình.

**- Early Stopping**: Dừng huấn luyện khi val\_loss không cải thiện sau 30 epoch, giữ lại trọng số tốt nhất.

**Đánh giá toàn diện**:

- Cung cấp nhiều chỉ số: accuracy, classification report (precision, recall, F1-score), confusion matrix, ROC curve (AUC), và permutation importance.

- Trực quan hóa rõ ràng qua biểu đồ (ma trận nhầm lẫn, ROC, tầm quan trọng đặc trưng, mất mát/độ chính xác), giúp dễ dàng phân tích hiệu suất.

**Khả năng tái sử dụng**:

- Lưu scaler vào file scaler.pkl bằng joblib, cho phép áp dụng cùng tiền xử lý cho dữ liệu mới.

### ****3.4.2. Hạn chế****

**Tham số cố định, thiếu tối ưu hóa đa dạng**:

- Các tham số như learning\_rate=0.0005, số đơn vị (64, 32, 16), dropout=0.4 được chọn cố định mà không thử nghiệm nhiều giá trị khác. Điều này có thể bỏ lỡ cấu hình tối ưu hơn.

- Ngưỡng dự đoán 0.3 (y\_pred > 0.3) được chọn thủ công, không dựa trên phân tích ROC hoặc tối ưu hóa F1-score.

**Không tận dụng SMOTE**:

- Phần sử dụng SMOTE bị comment (# smote = SMOTE(...)), nghĩa là mã chỉ dựa vào FocalLoss và class\_weights để xử lý mất cân bằng. Nếu các kỹ thuật này không đủ, lớp thiểu số có thể bị dự đoán kém.

**Không có kiểm tra chéo (cross-validation)**:

- Dữ liệu chỉ được chia một lần bằng train\_test\_split (80/20). Kết quả có thể không đại diện nếu dữ liệu bị lệch trong lần chia cụ thể, thiếu đánh giá độ ổn định của mô hình.

**Thời gian huấn luyện dài**:

- Với 150 epoch và kiến trúc sâu (3 lớp ẩn + BatchNormalization), huấn luyện có thể tốn nhiều thời gian, đặc biệt trên máy tính không có GPU.

**Thiếu xử lý ngoại lệ toàn diện**:

- Chỉ kiểm tra lỗi FileNotFoundError khi đọc file CSV, nhưng không xử lý các lỗi khác (ví dụ: lỗi dữ liệu thiếu, lỗi định dạng, lỗi bộ nhớ khi huấn luyện).

**Trực quan hóa chưa tối ưu**:

- Mặc dù có nhiều biểu đồ, nhưng không so sánh hiệu suất giữa các cấu hình khác nhau (do chỉ chạy một mô hình). Điều này hạn chế khả năng đánh giá tác động của từng tham số.

**Phụ thuộc vào dữ liệu cụ thể**:

- Mã giả định dữ liệu đầu vào có cấu trúc giống file WA\_Fn-UseC\_-HR-Employee-Attrition.csv. Nếu dữ liệu thay đổi (thêm/xóa cột), cần chỉnh sửa thủ công.

**Tính toán tầm quan trọng đặc trưng tốn tài nguyên**:

- permutation\_importance với n\_repeats=10 đòi hỏi chạy dự đoán nhiều lần trên tập kiểm tra, làm tăng thời gian xử lý mà không tối ưu hóa hiệu quả.

## **3.5. đề xuất ý tưởng cải tiến thuật toán**

### ****3.5.1. Tự động hóa việc chọn siêu tham số****

**Ý tưởng**: Thay vì cố định các tham số như tốc độ học, số đơn vị nơ-ron, hay tỷ lệ bỏ qua (Dropout), áp dụng một phương pháp tự động để tìm tổ hợp tối ưu. Ví dụ, sử dụng kỹ thuật tối ưu hóa dựa trên thử nghiệm (trial-and-error) thông minh như tìm kiếm Bayesian hoặc thuật toán di truyền.

**Lợi ích**: Tiết kiệm thời gian thử nghiệm thủ công, đảm bảo tìm được cấu hình tốt nhất mà không cần phỏng đoán.

**Cách áp dụng**: Sử dụng công cụ sẵn có để thử hàng chục tổ hợp tham số và chọn ra cấu hình có hiệu suất cao nhất dựa trên một chỉ số (như F1-score).

### ****3.5.2. Kết hợp nhiều kỹ thuật xử lý mất cân bằng****

**Ý tưởng**: Hiện tại chỉ dùng hàm mất mát đặc biệt (Focal Loss) và trọng số lớp. Có thể kết hợp thêm phương pháp tạo dữ liệu tổng hợp (oversampling) cho lớp thiểu số hoặc giảm dữ liệu (undersampling) lớp đa số để cân bằng trước khi huấn luyện.

**Lợi ích**: Tăng khả năng nhận diện lớp thiểu số (nhân viên nghỉ việc), cải thiện độ chính xác và độ nhạy (recall) cho lớp này.

**Cách áp dụng**: Áp dụng cả hai chiến lược (tạo thêm dữ liệu và điều chỉnh hàm mất mát) rồi so sánh hiệu quả để chọn cách tốt nhất.

### ****3.5.3. Đánh giá mô hình qua nhiều lần chia dữ liệu****

**Ý tưởng**: Thay vì chỉ chia dữ liệu một lần thành tập huấn luyện và kiểm tra, thực hiện chia dữ liệu nhiều lần (cross-validation) để kiểm tra tính ổn định của mô hình trên các tập con khác nhau.

**Lợi ích**: Đảm bảo kết quả không phụ thuộc vào một lần chia ngẫu nhiên, cho phép đánh giá toàn diện hơn về hiệu suất thực tế.

**Cách áp dụng**: Chia dữ liệu thành 5-10 phần, huấn luyện và kiểm tra trên từng phần, rồi lấy trung bình kết quả.

### ****3.5.4. Chọn ngưỡng dự đoán linh hoạt****

**Ý tưởng**: Hiện tại ngưỡng để phân loại (0.3) được chọn thủ công. Thay vào đó, phân tích dữ liệu kiểm tra để chọn ngưỡng tối ưu dựa trên mục tiêu cụ thể (ví dụ: ưu tiên phát hiện nhân viên nghỉ việc hơn là độ chính xác tổng thể).

**Lợi ích**: Tăng hiệu quả dự đoán bằng cách cân bằng giữa độ nhạy và độ đặc hiệu, phù hợp với yêu cầu thực tế của bài toán.

**Cách áp dụng**: Dùng biểu đồ ROC để xác định ngưỡng tối ưu thay vì đặt cố định.

### ****3.5.5. Tăng tốc độ huấn luyện****

**Ý tưởng**: Giảm thời gian huấn luyện bằng cách điều chỉnh tốc độ học linh hoạt (cao hơn khi bắt đầu, giảm dần khi gần tối ưu) hoặc sử dụng phần cứng mạnh hơn (như GPU).

**Lợi ích**: Rút ngắn thời gian chạy mà không làm giảm chất lượng mô hình, đặc biệt hữu ích khi thử nghiệm nhiều cấu hình.

**Cách áp dụng**: Áp dụng chiến lược thay đổi tốc độ học tự động trong quá trình huấn luyện và tận dụng tài nguyên tính toán mạnh hơn nếu có.

### ****3.5.6. So sánh nhiều kiến trúc mô hình****

**Ý tưởng**: Thay vì chỉ dùng một mạng nơ-ron sâu cố định, thử nghiệm thêm các kiến trúc khác như mạng nơ-ron hồi quy (RNN) cho dữ liệu thời gian hoặc mô hình đơn giản hơn (như hồi quy logistic) để so sánh hiệu quả.

**Lợi ích**: Tìm ra mô hình phù hợp nhất với dữ liệu, tránh lãng phí tài nguyên vào một cấu trúc không tối ưu.

**Cách áp dụng**: Chạy song song vài kiến trúc khác nhau và đánh giá dựa trên cùng tiêu chí (accuracy, F1-score).

### ****3.5.7. Xử lý dữ liệu linh hoạt và bền vững hơn****

**Ý tưởng**: Chuẩn bị cho các tình huống dữ liệu không hoàn hảo (thiếu giá trị, định dạng khác nhau) bằng cách thêm bước kiểm tra và xử lý trước khi huấn luyện.

**Lợi ích**: Tăng khả năng áp dụng thuật toán cho các tập dữ liệu mới mà không cần chỉnh sửa nhiều.

**Cách áp dụng**: Thêm bước kiểm tra dữ liệu đầu vào và tự động điền giá trị thiếu bằng trung bình hoặc giá trị hợp lý.

### ****3.5.8. Tập trung vào các đặc trưng quan trọng****

**Ý tưởng**: Dựa trên phân tích tầm quan trọng đặc trưng (permutation importance), loại bỏ hoặc giảm trọng số các đặc trưng ít ảnh hưởng để đơn giản hóa mô hình.

**Lợi ích**: Giảm độ phức tạp, tăng tốc độ huấn luyện và tránh nhiễu từ dữ liệu không quan trọng.

**Cách áp dụng**: Sau khi phân tích, chọn lọc khoảng 10-12 đặc trưng quan trọng nhất để huấn luyện thay vì dùng toàn bộ.

**KẾT LUẬN**

Đề tài đã xây dựng và triển khai thành công mô hình mạng nơ-ron nhân tạo (ANN) nhằm phân loại tình trạng thất nghiệp dựa trên các yếu tố quan trọng như trình độ học vấn, kinh nghiệm làm việc và ngành nghề. Kết quả thực nghiệm cho thấy mô hình đạt độ chính xác cao, giúp nhận diện nhóm đối tượng có nguy cơ thất nghiệp, qua đó hỗ trợ phân tích thị trường lao động và tư vấn nghề nghiệp một cách hiệu quả hơn.

Việc áp dụng AI, đặc biệt là ANN, vào phân tích tình trạng thất nghiệp mang lại nhiều lợi ích như khả năng xử lý dữ liệu lớn, nhận diện xu hướng và đưa ra dự báo nhanh chóng. Tuy nhiên, nghiên cứu vẫn còn một số hạn chế như chất lượng dữ liệu đầu vào, sự mất cân bằng dữ liệu giữa các nhóm đối tượng và chưa xét đến các yếu tố khác như vị trí địa lý, kỹ năng mềm, hoặc tác động của các yếu tố kinh tế vĩ mô.

Trong tương lai, việc mở rộng bộ dữ liệu, thử nghiệm thêm các mô hình học sâu (Deep Learning), và kết hợp nhiều biến số hơn có thể giúp nâng cao độ chính xác của mô hình. Ngoài ra, việc ứng dụng mô hình vào thực tế, chẳng hạn như tích hợp vào hệ thống hỗ trợ ra quyết định cho các cơ quan quản lý lao động và doanh nghiệp, cũng là một hướng đi tiềm năng.

Tóm lại, nghiên cứu này đã chứng minh tiềm năng của trí tuệ nhân tạo trong phân tích và dự đoán tình trạng thất nghiệp, từ đó góp phần cung cấp thông tin hữu ích cho việc hoạch định chính sách lao động, giúp doanh nghiệp và người lao động có những quyết định phù hợp để giảm thiểu tỷ lệ thất nghiệp và nâng cao chất lượng nguồn nhân lực.

**DANH MỤC TÀI LIỆU THAM KHẢO**

1. “**Artificial Intelligence: A Modern Approach” - Stuart Russell, Peter Norvig**

**[2]. "Deep Learning" - Ian Goodfellow, Yoshua Bengio, Aaron Courville**

**[3]. "Machine Learning Yearning" - Andrew Ng**