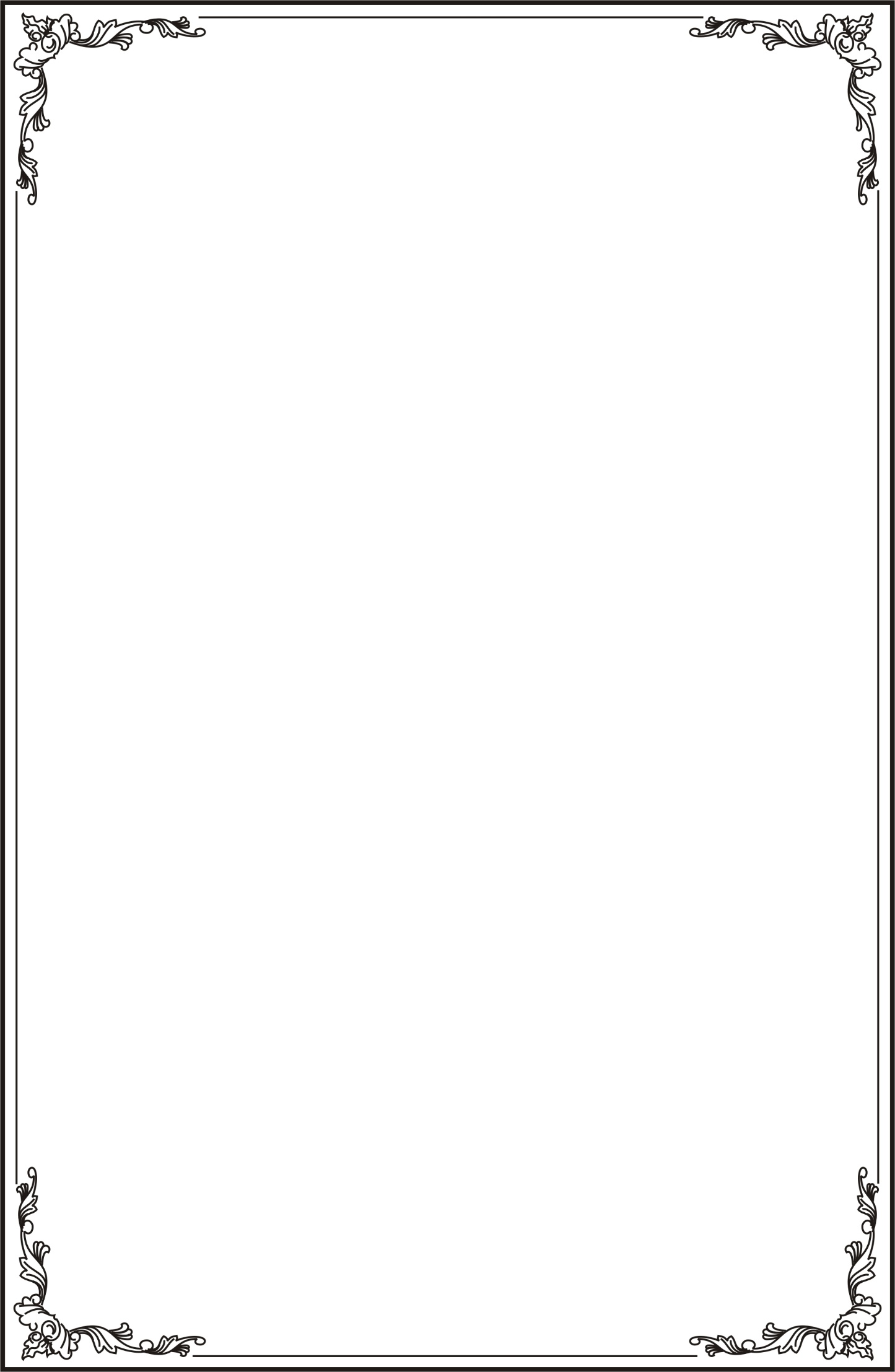
****TRƯỜNG ĐẠI HỌC NHA TRANG

**KHOA CÔNG NGHỆ THÔNG TIN**

-----🙞🙜🕮🙞🙜-----

****

**BÁO CÁO ĐỀ TÀI KHAI PHÁ DỮ LIỆU**

**“ĐỀ TÀI: XÂY DỰNG HỆ HỖ TRỢ KỸ NĂNG CHO HỌC SINH”**

**Giảng viên hướng dẫn :** Nguyễn Khắc Cường

**Sinh viên thực hiện** **:** Hàng Tuấn Kiệt - 63130597

Huỳnh Thanh Luân **-** 63134535

**Lớp :**  62.CNTT-1

**Khánh Hòa 2024 - 2025**

Mục Lục

[LỜI NÓI ĐẦU 3](#_Toc162228967)

[**I.** **TỔNG QUAN VỀ KHAI PHÁ DỮ LIỆU** 4](#_Toc162228968)

[**II.** **QUY TRÌNH KHAI PHÁ DỮ LIỆU** 4](#_Toc162228969)

[**1.** **Phát triển bài toán và xây dựng giả thuyết** 4](#_Toc162228970)

[**2.** **Thu thập dữ liệu** 4](#_Toc162228971)

[**3.** **Tiền xử lý dữ liệu** 4](#_Toc162228972)

[**4.** **Xây dựng mô hình** 5](#_Toc162228973)

[**5.** **Huấn luyện và Đánh giá mô hình** 5](#_Toc162228974)

[**6.** **Tối ưu và Điều chỉnh mô hình** 5](#_Toc162228975)

[**7.** **Triển khai mô hình** 5](#_Toc162228976)

[**III.** **BẢNG SERVEY** 5](#_Toc162228977)

[**IV.** **GIAO DIỆN VÀ HƯỚNG DẪN SỬ DỤNG** 7](#_Toc162228978)

[**1.** **Giao diện người dùng nhập vào:** 7](#_Toc162228979)

[**2.** **Giao diện dự đoán kỹ năng:** 8](#_Toc162228980)

[**3.** **Hướng dẫn sử dụng:** 8](#_Toc162228981)

[**V.** **TỔNG KẾT** 9](#_Toc162228982)

[**1.** **Kết quả đạt được** 9](#_Toc162228983)

[**2.** **Hướng phát triển** 9](#_Toc162228984)

[**3.** **Tổng kết** 9](#_Toc162228985)

[**VI.** **TÀI LIỆU THAM KHẢO** 9](#_Toc162228986)

# LỜI NÓI ĐẦU

Khai phá dữ liệu hay còn gọi với cái tên “Data Mining” là quá trình quan trọng và hấp dẫn trong lĩnh vực khoa học máy tính, trí tuệ nhân tạo. Trong bối cảnh thời đại số hóa ngày nay, dữ liệu ngày càng trở nên phong phú và đa dạng, từ đó tạo ra cơ hội mới để hiểu biết sâu sắc về nhiều lĩnh vực khác nhau.

Đây là một lĩnh vực khoa học tuy không mới, nhưng cho thấy lĩnh vực khai phá dữ liệu đang ngày càng phát triển và có thể tiến xa hơn trong tương lai.Đồng thời, thời điểm này nó được xem là một lĩnh vực “nóng” và dành rất nhiều mối quan tâm để phát triển nó một cách mạnh mẽ, bùng nổ hơn.

Hiện nay, việc quan tâm machine learning càng ngày càng tăng lên là vì nhờ có machine learning giúp gia tăng dung lượng lưu trữ các loại dữ liệu sẵn,việc xử lý tính toán có chi phí thấp và hiệu quả hơn rất nhiều.

Những điều trên được hiểu là nó có thể thực hiện tự động, nhanh chóng để tạo ra những mô hình cho phép phân tích các dữ liệu có quy mô lớn hơn và phức tạp hơn đồng thời đưa ra những kết quả một cách nhanh và chính xác hơn.

Điều quan trọng trong Khai phá dữ liệu là áp dụng các phương pháp và kỹ thuật khai phá dữ liệu một cách chính xác và hiệu quả để có được cái nhìn toàn diện về dữ liệu. Trong báo cáo này, chúng em tập trung vào việc khai phá dữ liệu về biểu hiện của con người về kỹ năng tư duy và tranh luận. Mục tiêu của chúng em là phân tích các biểu hiện này và dự đoán khả năng có kỹ năng tư duy, tranh luận của một cá nhân dựa trên các biểu hiện đó.

Chúng em xin chân thành gửi lời cảm ơn đến thầy Nguyễn Khắc Cường Trường Đại học Nha Trang đã tận tình giảng dạy, truyền đạt cho chúng em những kiến thức cũng như kinh nghiệm quý báu trong suốt quá trình học và đã tận tình theo sát giúp đỡ, trực tiếp chỉ bảo, hướng dẫn trong suốt quá trình nghiên cứu và học tập của chúng em.

Em xin chân thành cảm ơn !

1. **TỔNG QUAN VỀ KHAI PHÁ DỮ LIỆU**

Thuật ngữ Khai phá dữ liệu (KPDL) ra đời vào cuối những năm 80 thế kỷ trước. Có nhiều định nghĩa khác nhau về KPDL, nhưng diễn đạt một cách dễ hiểu thì **Khai phá dữ liệu** là một tập hợp các kỹ thuật được sử dụng để khai thác và tìm ra các mối quan hệ lẫn nhau giữa các dữ liệu trong một tập hợp dữ liệu khổng lồ và phức tạp, đồng thời cũng tìm ra các tri thức, mẫu tiềm ẩn trong tập dữ liệu đó.

Hiện kỹ thuật KPDL đang được áp dụng một cách rộng rãi trong nhiều lĩnh vực kinh doanh và đời sống khác nhau, như: Thương mại (phân tích dữ liệu bán hàng và thị trường, phân tích đầu tư, quyết định cho vay, phát hiện gian lận,...); Thông tin sản xuất (điều khiển và lập kế hoạch, hệ thống quản lý, phân tích kết quả thử nghiệm…); Thông tin khoa học (dự báo thời tiết,...); CSDL sinh học (ngân hàng gen,…); Khoa học địa lý (dự báo động đất,…); Trong y tế, marketing, ngân hàng, viễn thông, du lịch, internet…

Những gì thu được từ KPDL thật đáng giá. Điều đó được chứng minh bằng thực tế: Chẩn đoán bệnh trong y tế dựa trên kết quả xét nghiệm đã giúp cho bảo hiểm y tế phát hiện ra nhiều trường hợp xét nghiệm không hợp lý; Trong dịch vụ viễn thông đã phát hiện ra những nhóm người thường xuyên gọi cho nhau bằng mobile và thu lợi hàng triệu USD; IBM Suft-Aid đã áp dụng KPDL vào phân tích các lần đăng nhập Web vào các trang liên quan đến thị trường để phát hiện sở thích khách hàng, từ đó đánh giá hiệu quả của việc tiếp thị qua Web và cải thiện hoạt động của các Website; trang Web mua bán qua mạng cũng tăng doanh thu nhờ áp dụng KPDL trong việc phân tích sở thích mua bán của khách hàng.

1. **QUY TRÌNH KHAI PHÁ DỮ LIỆU**
2. **Phát triển bài toán và xây dựng giả thuyết**

Trước hết, cần xác định rõ bài toán muốn giải quyết và xác định giả thuyết ban đầu. Trong trường hợp này bài toán là dự đoán khả năng có hoặc không có kỹ năng tư duy, tranh luận của một cá nhân dựa trên các biểu hiện của họ.

1. **Thu thập dữ liệu**

Thu thập dữ liệu liên quan từ các nguồn khác nhau. Trong đề tài này chúng em thu thập dữ liệu từ sách, báo chí, trang mạng, các video nói về biểu hiện của các kỹ năng tư duy và tranh luận.

1. **Tiền xử lý dữ liệu**

Mã hóa dữ liệu: chuyển đổi dữ liệu phân loại thành dạng số để có thể xử lý bằng mô hình RNN. Có thể sử dụng các phương pháp như mã hóa nhị phân (binary encoding) hoặc mã hóa one-hot.

Xử lý dữ liệu thiếu: Kiểm tra và xử lý các dữ liệu thiếu (nếu có) bằng cách sử dụng các phương pháp như điền giá trị trung bình, median, hoặc sử dụng mô hình để dự đoán giá trị thiếu.

Phân chia dữ liệu: Chia dữ liệu thành các tập huấn luyện, validation và kiểm tra để đánh giá hiệu suất của mô hình một cách chính xác.

1. **Xây dựng mô hình**

Xác định kiến trúc mô hình phù hợp, có thể bao gồm các tầng LSTM hoặc GRU để xử lý hiệu quả dữ liệu có tính chuỗi thời gian hoặc tuần tự.

Chọn hàm mất mát và bộ tối ưu phù hợp dựa trên bài toán cụ thể (ví dụ, mất mát cross-entropy cho bài toán phân loại).

1. **Huấn luyện và Đánh giá mô hình**

Huấn luyện mô hình sử dụng tập huấn luyện và theo dõi hiệu suất trên tập validation để điều chỉnh tham số và tránh overfitting.

Đánh giá mô hình cuối cùng trên tập kiểm tra để xác định độ chính xác và hiệu suất tổng thể.

1. **Tối ưu và Điều chỉnh mô hình**

Sử dụng kỹ thuật như tuning tham số, kỹ thuật regularization (ví dụ, dropout) để tối ưu mô hình và cải thiện khả năng khái quát hóa.

1. **Triển khai mô hình**

Phát triển một ứng dụng hoặc hệ thống có thể sử dụng mô hình đã được huấn luyện để dự đoán kỹ năng của học sinh mới dựa trên biểu hiện của họ.

1. **BẢNG SERVEY**

|  |  |  |
| --- | --- | --- |
| **Số Lần Thay đổi Model + Data** | **Mô Tả** | **Kết Quả / Nhận Xét** |
| 1 | Thu thập dữ liệu | Thu thập 14 mẫu, bao gồm dữ liệu: Sách, Báo chí, Tiểu thuyết, Talk Show. |
| Tiền xử lý dữ liệu | Loại bỏ dữ liệu hỏng, mã hóa nhãn, và chuẩn hóa dữ liệu. |
| Chia tập dữ liệu | 80% dữ liệu huấn luyện, 20% dữ liệu kiểm tra. |
| Huấn luyện mô hình | Huấn luyện mô hình **SVM** với test\_size=0.2, random\_state=42 và độ chính xác trên tập huấn luyện là 0% |
| Đánh giá mô hình | Độ chính xác trên tập kiểm tra là 0%. |
| 2 | Thu thập dữ liệu | Thu thập 16 mẫu, bao gồm dữ liệu: Sách, Báo chí, Tiểu thuyết, Talk Show. |
| Tiền xử lý dữ liệu | Loại bỏ dữ liệu hỏng, mã hóa nhãn, và chuẩn hóa dữ liệu. |
| Chia tập dữ liệu | 80% dữ liệu huấn luyện, 20% dữ liệu kiểm tra. |
| Huấn luyện mô hình | Huấn luyện mô hình **Random Forest Classifier** với test\_size=0.2, random\_state=42 và độ chính xác trên tập huấn luyện là 33%. |
| Đánh giá mô hình | Độ chính xác trên tập kiểm tra là 33%. |
| 3 | Thu thập dữ liệu | Thu thập 16 mẫu, bao gồm dữ liệu: Sách, Báo chí, Tiểu thuyết, Talk Show. |
| Tiền xử lý dữ liệu | Loại bỏ dữ liệu hỏng, mã hóa nhãn, và chuẩn hóa dữ liệu. |
| Chia tập dữ liệu | 80% dữ liệu huấn luyện, 20% dữ liệu kiểm tra. |
| Huấn luyện mô hình | Huấn luyện mô hình **kNN** với test\_size=0.2, random\_state=42 và chính xác trên tập huấn luyện là 100% |
| Đánh giá mô hình | Độ chính xác trên tập kiểm tra là 0%. |
| 4 | Thu thập dữ liệu | Thu thập 18 mẫu, bao gồm dữ liệu: Sách, Báo chí, Tiểu thuyết, Talk Show. |
| Tiền xử lý dữ liệu | Loại bỏ dữ liệu hỏng, mã hóa nhãn, và chuẩn hóa dữ liệu. |
| Chia tập dữ liệu | 80% dữ liệu huấn luyện, 20% dữ liệu kiểm tra. |
| Huấn luyện mô hình | Huấn luyện mô hình **CNN** với test\_size=0.2, random\_state=42 ,learning\_rate=0.001, epochs=100, batch\_size=8 và độ chính xác trên tập huấn luyện là 66,67%. |
| Đánh giá mô hình | Độ chính xác trên tập kiểm tra là 0.6666666865348816. |
| 5 | Thu thập dữ liệu | Thu thập 26 mẫu, bao gồm dữ liệu: Sách, Báo chí, Tiểu thuyết, Talk Show. |
| Tiền xử lý dữ liệu | Loại bỏ dữ liệu hỏng, mã hóa nhãn, và chuẩn hóa dữ liệu. |
| Chia tập dữ liệu | 80% dữ liệu huấn luyện, 20% dữ liệu kiểm tra. |
| Huấn luyện mô hình | Huấn luyện mô hình **RNN** với test\_size=0.2, random\_state=42, learning\_rate=0.001, epochs=50, batch\_size=8 và độ chính xác trên tập huấn luyện là 0%. |
| Đánh giá mô hình | Độ chính xác trên tập kiểm tra là 0%. |
| 6 | Thu thập dữ liệu | Thu thập 728 mẫu, bao gồm dữ liệu: Sách, Báo chí, Tiểu thuyết, Talk Show. |
| Tiền xử lý dữ liệu | Loại bỏ dữ liệu hỏng, mã hóa nhãn, và chuẩn hóa dữ liệu. |
| Chia tập dữ liệu | 80% dữ liệu huấn luyện, 20% dữ liệu kiểm tra. |
| Huấn luyện mô hình | Huấn luyện mô hình **ANN** với test\_size=0.2, random\_state=0, learning\_rate=0.0005, epochs=60, batch\_size= 128, độ chính xác trên tập huấn luyện là 71.73%. |
| Đánh giá mô hình | Độ chính xác trên tập kiểm tra là 69.8630154132843% |
| 7 | Thu thập dữ liệu | Thu thập 819 mẫu, bao gồm dữ liệu: Sách, Báo chí, Tiểu thuyết, Talk Show. |
| Tiền xử lý dữ liệu | Loại bỏ dữ liệu hỏng, mã hóa nhãn, và chuẩn hóa dữ liệu. |
| Chia tập dữ liệu | 80% dữ liệu huấn luyện, 20% dữ liệu kiểm tra. |
| Huấn luyện mô hình | Huấn luyện mô hình **DNN** với test\_size=0.2, random\_state=0, learning\_rate=0.0005, epochs=60, batch\_size= 128 và độ chính xác trên tập huấn luyện là 88.40%. |
| Đánh giá mô hình | Độ chính xác trên tập kiểm tra là 79.2682945728302 %. |

1. **GIAO DIỆN VÀ HƯỚNG DẪN SỬ DỤNG**
2. **Giao diện người dùng nhập vào:**

**A screenshot of a computer

Description automatically generated**

1. **Giao diện dự đoán kỹ năng:**

**A screenshot of a message

Description automatically generated**

1. **Hướng dẫn sử dụng:**

* Tải đầy đủ các thư viện trước khi chạy chương trình
* Chạy code huấn luyện Models: *Huan\_Luyen\_Model🡪 Notebooks🡪 DNN.py / Model\_DNN.ipynb*

file *DNN* vs *Model\_DNN* đều chạy giống nhau và có thể chạy trên môi trường PyCharm, Jupyter Notebook

**A screen shot of a computer program

Description automatically generated**

* Chạy giao diện: *Huan\_Luyen\_Model🡪 App🡪 main.py*

**A screen shot of a computer

Description automatically generated**

1. **TỔNG KẾT**
2. **Kết quả đạt được**

Xây dựng được một mô hình với độ chính xác cao để dự đoán khả năng có kỹ năng tư duy, tranh luận của một cá nhân dựa trên các biểu hiện của họ.

Hiểu biết sâu hơn về các kỹ thuật khai phá dữ liệu

Xây dựng giao diện cho người dùng để sử dụng khi người dùng nhập vào các biểu hiện của bản thân về các kỹ năng để có thể biết được bản thân có những kỹ năng nào và chưa có kỹ năng nào giúp họ có thể phát triển những kỹ năng đó.

1. **Hướng phát triển**

Dựa trên kết quả của nghiên cứu, chúng em đề xuất một số hướng phát triển tiếp theo như sau:

* Mở rộng phạm vi và thu thập thêm dữ liệu để cải thiện mô hình.
* Nghiên cứu các phương pháp và kỹ thuật khác tiên tiến hơn trong lĩnh vực học máy và khai phá dữ liệu để tăng độ chính xác của mô hình .
* Áp dụng kết quả khai phá vào thực tiễn, bằng cách phát triển các ứng dụng và công cụ hỗ trợ trong các lĩnh vực như giáo dục.

1. **Tổng kết**

Qua quá trình tìm hiểu và nghiên cứu đề tài. Chúng em đã hoàn thành được nội dung đặt ra với đề tài là khai phá dữ liệu biểu hiện của con người về kỹ năng tư duy và tranh luận.

Do kinh nghiệm bản thân còn hạn chế và thời gian có hạn nên không tránh khỏi sai sót. Chúng em rất mong nhận được sự đóng góp của thầy để mô hình được hoàn thiện hơn nữa.

Em xin chân thành cảm ơn!

1. **TÀI LIỆU THAM KHẢO**
   * Giáo trình KHAI PHÁ DỮ LIỆU trường Đại học Nha Trang – Nguyễn Khắc Cường
   * ["Data Mining: Concepts and Techniques" - Jiawei Han, Micheline Kamber, và Jian Pei.](https://myweb.sabanciuniv.edu/rdehkharghani/files/2016/02/The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf)
   * [Machine Learning cơ bản – Nguyễn Thanh Tuấn](https://nttuan8.com/bai-3-neural-network/)