|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **TRƯỜNG ĐẠI HỌC ĐIỆN LỰC**  **KHOA CÔNG NGHỆ THÔNG TIN**    **BÁO CÁO THỰC TẬP TỐT NGHIỆP**  **MÔN**  **ĐỀ TÀI:**   |  |  | | --- | --- | | **Sinh viên thực hiện** | **: TRẦN TUẤN ANH** | | **Giảng viên hướng dẫn** | **:** | | **Ngành** | **: CÔNG NGHỆ THÔNG TIN** | | **Chuyên ngành** | **: CÔNG NGHỆ PHẦN MỀM** | | **Lớp** | **: D14CNPM1** | | **Khóa** | **: 2019-2024** |   ***Hà Nội, tháng 3 năm 2023*** |

1. **Mô tả tóm tắt đề tài**

* Xây dựng mô hình Naive Bayes.

1. **Nội dung thực hiện**

* Chương 1: Tổng quan đề tài
* Chương 2: Tổng quan về khai phá dữ liệu
* Chương 3: Kĩ thuật phân lớp và mô hình Naive Bayes
* Chương 4: Thực nghiệm và kết luận

1. **Kết quả đạt được**

* Hoàn thành đề cương chuyên đề học phần môn: “Khai phá dữ liệu”.
* Xây dựng được mô hình Naive Bayes đơn giản để thực hiện dự đoán dữ liệu.

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

**PHIẾU CHẤM ĐIỂM**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **STT** | **Họ và tên sinh viên** | **Nội dung thực hiện** | **Điểm** | **Chữ ký** |
| 1 | TRẦN TUẤN ANH  (19810310026) |  |  |  |
| 2 |  |  |  |  |
| 3 |  |  |  |  |

|  |  |  |
| --- | --- | --- |
| **Họ và tên giảng viên** | **Chữ ký** | **Ghi chú** |
| Giảng viên chấm 1: |  |  |
| Giảng viên chấm 2: |  |  |

**MỤC LỤC**

[LỜI CẢM ƠN 1](#_Toc106107144)

[TÓM TẮT ĐỀ TÀI 2](#_Toc106107145)

[CHƯƠNG 1: TỔNG QUAN ĐỀ TÀI 3](#_Toc106107146)

[**1.1. Đặt vấn đề 3**](#_Toc106107147)

[**1.2. Mục tiêu đề tài 3**](#_Toc106107148)

[**1.3. Ý nghĩa đề tài 3**](#_Toc106107149)

[**1.3.1. Ý nghĩa chung 3**](#_Toc106107150)

[**1.3.2. Ý nghĩa thực tiễn 4**](#_Toc106107151)

[CHƯƠNG 2: TỔNG QUAN VỀ KHAI PHÁ DỮ LIỆU 5](#_Toc106107152)

[**2.1. Tổng quan về kỹ thuật Khai phá dữ liệu (Data mining) 5**](#_Toc106107153)

[**2.1.1. Khái niệm về khai phá dữ liệu 5**](#_Toc106107154)

[**2.1.2. Quy trình khai phá dữ liệu 5**](#_Toc106107155)

[**2.1.3. Ứng dụng của khai phá dữ liệu 6**](#_Toc106107156)

[**2.2. Classification 6**](#_Toc106107157)

[**2.2.1. Khái niệm 6**](#_Toc106107158)

[2.2.2. Các loại kỹ thuật phân lớp trong khai phá dữ liệu 7](#_Toc106107159)

[**2.3. Navie Bayes 8**](#_Toc106107160)

[**2.3.1. Ưu điểm của Navie Bayes 8**](#_Toc106107161)

[**2.3.2. Nhược điểm của Naive Bayes 8**](#_Toc106107162)

[**2.3.3. Các loại bộ phân loại Naive Bayes 9**](#_Toc106107163)

[**2.3. Ứng dụng của Navie Bayes 9**](#_Toc106107164)

[**2.4. Cơ sở dữ liệu về giáo dục trực tuyến 10**](#_Toc106107165)

[CHƯƠNG 3: KỸ THUẬT PHÂN LỚP VÀ MÔ HÌNH NAVIE BAYES 11](#_Toc106107166)

[**3.1. Khái niệm cơ bản về phân lớp (Khai phá dữ liệu) 11**](#_Toc106107167)

[**3.2. Mô hình phân lớp Navie Bayes 13**](#_Toc106107168)

[**3.3. Một số kiểu mô hình Naive Bayes 15**](#_Toc106107169)

[**3.3.1. Gaussian Naive Bayes 15**](#_Toc106107170)

[**3.3.2. Bernoulli Naive Bayes 15**](#_Toc106107171)

[**3.3.3. Multinomial Naive Bayes 16**](#_Toc106107172)

[CHƯƠNG 4: LẬP TRÌNH 17](#_Toc106107173)

[**4.1. Lập trình 17**](#_Toc106107174)

[**4.1.1. Khai báo thư viện 17**](#_Toc106107175)

[**4.1.2. Đọc và in file CSV 17**](#_Toc106107176)

[**4.1.3. Chuyển dữ liệu sang dạng số để phù hợp với đầu vào của model 18**](#_Toc106107177)

[**4.1.4. Trích chọn các cột để làm dữ liệu huấn luyện và nhãn 18**](#_Toc106107178)

[**4.1.3. Chia dữ liệu train/test theo tỉ lệ 80/20 19**](#_Toc106107179)

[**4.1.4. Số lượng của từng label trong tập dữ liệu train 19**](#_Toc106107180)

[**4.1.5. Huấn luyện bằng mô hình GaussianNB và in ra kết quả huấn luyện 20**](#_Toc106107181)

[**4.1.6. Huấn luyện bằng mô hình BernoulliNB và in ra kết quả huấn luyện 21**](#_Toc106107182)

[**4.1.7. Huấn luyện bằng mô hình MultinomialNB và in ra kết quả huấn luyện 23**](#_Toc106107183)

[**4.2. Kết luận 25**](#_Toc106107184)

[**4.2.1. So sánh kết quả các mô hình 25**](#_Toc106107185)

[**4.2.1.1. Accuracy 25**](#_Toc106107186)

[**4.2.1.2. Precision – recall- f1 score 26**](#_Toc106107187)

[TÀI LIỆU THAM KHẢO 29](#_Toc106107188)

**DANH SÁCH CÁC HÌNH ẢNH**

[**Hình 3.1: Tương quan của khai phá dữ liệu 11**](#_Toc106108541)

[**Hình 3.2: Dữ liệu thử nghiệm sử dụng để ước lượng độ chính xác của quy tắc phân loại 12**](#_Toc106108542)

[**Hình 3.3: Dữ liệu thử nghiệm được sử dụng để ước tính độ chính xác của quy tắc phân lớp 13**](#_Toc106108543)

[**Hình 3.4: Mô tả thuật toán Gaussian Naive Bayes 15**](#_Toc106108544)

[**Hình 3.5: Mô tả thuật toán Bernoulli Naive Bayes 15**](#_Toc106108545)

[**Hình 3.6: Mô tả thuật toán Multinomial Naive Bayes 16**](#_Toc106108546)

[**Hình 4.1: Cấu trúc thư mục của dự án 17**](#_Toc106108547)

[**Hình 4.2: Kết quả đọc file CSV 17**](#_Toc106108548)

[**Hình 4.3: Kết quả sau khi chuyển đổi dữ liệu sang dạng số 18**](#_Toc106108549)

[**Hình 4.4: Kết quả trích chọn cột dùng để làm dữ liệu huấn luyện và nhãn 18**](#_Toc106108550)

[**Hình 4.5: Kết quả sau khi chia tỉ lệ giữa dữ liệu để huấn luyện và dữ liệu để học 19**](#_Toc106108551)

[**Hình 4.6: Số các nhãn được đếm trong tập huấn luyện 19**](#_Toc106108552)

[**Hình 4.7: Các nhãn dự đoán được sau khi dùng mô hình GaussianNB 20**](#_Toc106108553)

[**Hình 4.8: Tỉ lệ chính xác khi đoán nhãn khi dùng mô hình GaussianNB + Báo cáo của việc huấn luyện: 20**](#_Toc106108554)

[**Hình 4.9: Báo cáo sau khi dùng mô hình GaussianNB để huấn luyện 20**](#_Toc106108555)

[**Hình 4.10: Ma trận nhầm lẫn sau khi dùng mô hình GaussianNB để huấn luyện 21**](#_Toc106108556)

[**Hình 4.11: Các nhãn dự đoán được sau khi dùng mô hình BernoulliNB 22**](#_Toc106108557)

[**Hình 4.12: Tỉ lệ chính xác khi đoán nhãn khi dùng mô hình BernoulliNB 22**](#_Toc106108558)

[**Hình 4.13: Báo cáo sau khi dùng mô hình BernoulliNB để huấn luyện 22**](#_Toc106108559)

[**Hình 4.14: Ma trận nhầm lẫn sau khi dùng mô hình BernoulliNB để huấn luyện 23**](#_Toc106108560)

[**Hình 4.15: Các nhãn dự đoán được sau khi dùng mô hình MultinomialNB 24**](#_Toc106108561)

[**Hình 4.16: Tỉ lệ chính xác khi đoán nhãn khi dùng mô hình MultinomialNB 24**](#_Toc106108562)

[**Hình 4.17: Báo cáo sau khi dùng mô hình MultinomialNB để huấn luyện 24**](#_Toc106108563)

[**Hình 4.18: Ma trận nhầm lẫn sau khi dùng mô hình MultinomialNB để huấn luyện 25**](#_Toc106108564)

**DANH SÁCH CÁC BẢNG**

[Bảng 4.1: Bảng so sánh Accuracy 25](#_Toc106108592)

[Bảng 4.2: So sánh Precision của các mô hình 26](#_Toc106108593)

[Bảng 4.3: So sánh Recall của các mô hình 27](#_Toc106108594)

[Bảng 4.4: So sánh f1-score của các mô hình 27](#_Toc106108595)

# LỜI CẢM ƠN

Đầu tiên, chúng em xin gửi lời cảm ơn chân thành đến Trường Đại học Điện Lực đã đưa môn học khai phá dữ liệu vào trương trình giảng dạy. Đặc biệt, chúng em xin gửi lời cảm ơn sâu sắc đến giảng viên bộ môn – thầy Vũ Văn Định đã dạy dỗ, truyền đạt những kiến thức quý báu cho em trong suốt thời gian học tập vừa qua. Trong thời gian tham gia lớp học của thầy, chúng em đã có thêm cho mình nhiều kiến thức bổ ích, tinh thần học tập hiệu quả, nghiêm túc. Đây chắc chắn sẽ là những kiến thức nền rất hữu hiệu và là sự trải nghiệm để giúp chúng em vững bước sau này.

Bộ môn khai phá dữ liệu là môn học thú vị, vô cùng bổ ích và có tính thực tế cao. Đảm bảo cung cấp đủ kiến thức, gắn liền với nhu cầu thực tiễn của sinh viên. Tuy nhiên, do vốn kiến thức còn nhiều hạn chế và khả năng tiếp thu thực tế còn nhiều bỡ ngỡ. Mặc dù chúng em đã cố gắng hết sức nhưng chắc chắn bài tiểu luận khó có thể tránh khỏi những thiếu sót và nhiều chỗ còn chưa chính xác, kính mong thầy xem xét và góp ý để bài tiểu luận của chúng em được hoàn thiện hơn.

Chúng em xin chân thành cảm ơn!

# TÓM TẮT ĐỀ TÀI

Ngành giáo dục luôn là vấn đề sống còn của bất kỳ quốc gia nào trên thế giới. Trong những năm gần đây, chính phủ Việt nam đặc biệt đầu tư cho ngành mũi nhọn này thông qua các chính sách , nguồn vốn dành cho trang thiết bị hạ tầng và nghiên cứu khoa học. Trong lĩnh vực giáo dục, càng ngày càng có nhiều công trình khoa học trong giáo dục. Tuy nhiên các nghiên cứu khoa học về ứng dụng công nghệ thông tin để giải quyết bài toán về giáo dục là không nhiều. Do sự chuyển biến và tình hình lây lan diễn biến phức tạm của đại dịch Covid-19 xảy ra trên toàn thế giới, vậy nên chúng em làm đề tài sử dụng môn học khai phá dữ liệu để xác định đánh gía sự thích ứng học tập trên nhiều học sinh để cho thấy sự ảnh hưởng phần nào của dịch bệnh tới tình hình học tập trực tuyến.

Nghiên cứu tiến hành theo 3 bước chính:

1. Thu nhập thông tin liên quan và tiền xử lý dữ liệu .
2. Tìm hiểu bài toán Phân lớp nhiều lớp trong khai phá dữ liệu, lựa chọn kỹ thuật phân lớp Navie Bayes với dữ liệu thu nhập được.
3. Hiện thực chương trình máy tính và đánh giá ý nghĩa thực tiễn.

# CHƯƠNG 1: TỔNG QUAN ĐỀ TÀI

# TÀI LIỆU THAM KHẢO

1. Blog Machine Learning cơ bản – Vũ Hữu Tiệp: https://machinelearningcoban.com/2017/08/08/nbc/

2. Slide bài giảng Khai Phá Dữ Liệu- Nguyễn Nhật Quang, Trường Đại Học Bách Khoa Hà Nội.

3. Trang chứa dữ liệu: <https://www.kaggle.com/datasets/mdmahmudulhasansuzan/students-adaptability-level-in-online-education>

4. Thuật toán phân lớp Navie Bayes: <https://viblo.asia/p/thuat-toan-phan-lop-naive-bayes-924lJWPm5PM>

5. Tổng quan về thuật toán phân lớp Navie Bayes: http://hoctructuyen123.net/tong-quan-ve-thuat-toan-phan-lop-naive-bayes-classification-nbc/