**TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP TP HỒ CHÍ MINH**

**KHOA CÔNG NGHỆ THÔNG TIN**

****

**ĐỒ ÁN CUỐI KÌ**

**NHẬP MÔN KHOA HỌC DỮ LIỆU**

*Người thực hiện*: **Nguyễn Văn Nam – MSHV**

**Nguyễn Tấn Đạt – MSHV**

Lớp **: 10050301**

Khoá  **: 17**

*Người hướng dẫn*: **TS BÙI THANH HÙNG**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2023**

**TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP TP HỒ CHÍ MINH**

**KHOA CÔNG NGHỆ THÔNG TIN**

****

**ĐỒ ÁN CUỐI KÌ**

**NHẬP MÔN KHOA HỌC DỮ LIỆU**

Người thực hiện: **Nguyễn Văn Nam - MSSV**

**Nguyễn Tấn Đạt - MSSV**

Lớp **: 10050301**

Khoá  **: 17**

Người hướng dẫn: **TS. BÙI THANH HÙNG**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2023**

**LỜI CẢM ƠN**

Chúng em xin gửi lời cảm ơn đến TS.Bùi Thanh Hùng thầy là người hướng dẫn định hướng, và chỉ điểm chúng em để chúng em có thể hoàn thành đồ án này, chúng em chân thành cảm ơn thầy đã giúp đỡ và chia sẻ những kiến thức bổ ích để chúng em có thể thuận lợi hoàn thành đồ án kết thúc môn học này.

Một lần nữa chúng em chân thành cảm ơn thầy TS.Bùi Thanh Hùng rất nhiều.

Trân trọng,

**ĐỒ ÁN ĐƯỢC HOÀN THÀNH**

**TẠI TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP TP HỒ CHÍ MINH**

Tôi xin cam đoan đây là sản phẩm đồ án của riêng chúng tôi và được sự hướng dẫn của TS. Bùi Thanh Hùng. Các nội dung nghiên cứu, kết quả trong đề tài này là trung thực và chưa công bố dưới bất kỳ hình thức nào trước đây. Những số liệu trong các bảng biểu phục vụ cho việc phân tích, nhận xét, đánh giá được chính tác giả thu thập từ các nguồn khác nhau có ghi rõ trong phần tài liệu tham khảo.

Ngoài ra, trong đồ án còn sử dụng một số nhận xét, đánh giá cũng như số liệu của các tác giả khác, cơ quan tổ chức khác đều có trích dẫn và chú thích nguồn gốc.

**Nếu phát hiện có bất kỳ sự gian lận nào tôi xin hoàn toàn chịu trách nhiệm về nội dung đồ án của mình.** Trường đại học Công nghiệp TP Hồ Chí Minh không liên quan đến những vi phạm tác quyền, bản quyền do tôi gây ra trong quá trình thực hiện (nếu có).

*TP. Hồ Chí Minh, ngày tháng năm*

*Tác giả*

*(ký tên và ghi rõ họ tên)*

*Nguyễn Văn Nam*

*Nguyễn Tấn Đạt*

**PHẦN ĐÁNH GIÁ CỦA GIẢNG VIÊN**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Tp. Hồ Chí Minh, ngày tháng năm

(kí và ghi họ tên)

**TÓM TẮT**

Trình bày tóm tắt vấn đề nghiên cứu, các hướng tiếp cận, cách giải quyết vấn đề và một số kết quả đạt được, những phát hiện cơ bản trong vòng 1 -2 trang của **từng Bài toán, mỗi bài toán là 1 đoạn văn bản khác nhau**.

**MỤC LỤC**

[LỜI CẢM ƠN i](#_heading=h.gjdgxs)

[PHẦN XÁC NHẬN VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN i](#_heading=h.30j0zll)

[TÓM TẮT i](#_heading=h.1fob9te)

[MỤC LỤC 1](#_heading=h.3znysh7)

[DANH MỤC CHỮ VIẾT TẮT 1](#_heading=h.2et92p0)

[DANH MỤC CÁC BẢNG BIỂU, HÌNH VẼ, ĐỒ THỊ 1](#_heading=h.2et92p0)

[CHƯƠNG 1 – BÀI 1 1](#_heading=h.tyjcwt)

[1.1 Tiểu mục cấp 1 1](#_heading=h.2s8eyo1)

[1.1.1 Tiểu mục cấp 2 1](#_heading=h.3dy6vkm)

[1.1.1.1 Tiểu mục cấp 3 1](#_heading=h.17dp8vu)

[1.1.1.2 Tiểu mục cấp 3 tiếp theo. 1](#_heading=h.3rdcrjn)

[1.1.2 Tiểu mục cấp 2 tiếp theo 1](#_heading=h.26in1rg)

[1.2 Nội dung của chương này 1](#_heading=h.1t3h5sf)

[CHƯƠNG 2 – BÀI 2 1](#_heading=h.lnxbz9)

2[.1 Tiểu mục 1 1](#_heading=h.35nkun2)

2[.2 Tiểu mục 2 1](#_heading=h.1ksv4uv)

[CHƯƠNG 3 – BÀI 3 1](#_heading=h.lnxbz9)

3[.1 Tiểu mục 1 1](#_heading=h.35nkun2)

3[.2 Tiểu mục 2 1](#_heading=h.1ksv4uv)

LÀM VIỆC NHÓM 2

TỰ ĐÁNH GIÁ

6

**DANH MỤC KÍ HIỆU VÀ CHỮ VIẾT TẮT**

**CÁC KÝ HIỆU**

*f Tần số của dòng điện và điện áp (Hz)*

*p Mật độ điện tích khối (C/m3)*

**CÁC CHỮ VIẾT TẮT**

CSTD Công suất tác dụng

MF Máy phát điện

BER Tỷ lệ bít lỗi

**DANH MỤC CÁC HÌNH VẼ**

[Hình 2.1: Kiến trúc FTP 1](#_heading=h.44sinio)

Hình 2.2: Kiến trúc TP 1

**DANH MỤC CÁC BẢNG**

[Bảng 3.1 Ví dụ cho chèn bảng 1](#_heading=h.2jxsxqh)

[Bảng 3.2 Kết quả 1](#_heading=h.2jxsxqh)

**CHƯƠNG 1**

**THU THẬP DỮ LIỆU**

**1.1 Yêu cầu thu thập dữ liệu**

Yêu cầu thu thập dữ liệu là thu thập ý kiến của sinh viên đại học IUH về 5 chủ đề khác nhau. Mỗi chủ đề gồm 12 ý kiến, bao gồm 4 ý kiến tích cực (positive), 4 ý kiến tiêu cực (negative) và 4 ý kiến trung tính (neutral).

**1.2 Cách thức tiến hành thu thập dữ liệu**

Để thu thập dữ liệu, chúng ta có thể sử dụng các phương pháp như đăng ký trực tuyến và thu thập dữ liệu thông qua mẫu câu hỏi hoặc phỏng vấn trực tiếp các sinh viên. Tuy nhiên, trong trường hợp này, chúng ta sẽ tạo ra một bản khảo sát online bằng google form và gửi nó đến những sinh viên IUH để khảo sát.

**1.3 Kết quả**

Kết quả đạt được:

Sau khi thu thập xong chúng ta sẽ có bộ dữ liệu có 60 dòng dữ liệu gồm 5 chủ đề mỗi chủ đề CSVC, GV, CTDT, CTSV, TVTS, mỗi chủ đề có 12 dòng dữ liệu bao gồm 4 dòng dữ liệu đánh giá tích cực (được đánh số là 1), 4 dòng dữ liệu đánh giá t

Cách đánh giá tính chính xác của bộ dữ liệu :

Để đánh giá bằng cách đọc từng dòng dữ liệu, trước hết ta cần xác định các tiêu chí đánh giá để có thể đánh giá độ chính xác và tính đúng đắn của dữ liệu. Các tiêu chí đánh giá có thể bao gồm:

1. Sự phù hợp của đánh giá với nội dung ý kiến: điểm đánh giá phải phản ánh đúng ý nghĩa và tình huống được mô tả trong ý kiến.
2. Sự thống nhất của đánh giá trong các ý kiến liên quan: các ý kiến liên quan đến cùng một chủ đề và có nội dung tương tự nên có điểm đánh giá tương đồng với nhau.
3. Độ chính xác của đánh giá: điểm đánh giá phải phù hợp với thực tế và đánh giá phản ánh đúng quan điểm của người dùng.

Sau khi xác định các tiêu chí đánh giá, ta có thể thực hiện đánh giá từng dòng dữ liệu bằng cách đọc và hiểu nội dung của từng ý kiến, sau đó đưa ra một điểm đánh giá theo ba mức độ -1,0,1. Với mỗi ý kiến, ta cần đánh giá xem điểm đánh giá có phù hợp với nội dung ý kiến hay không, và có thể so sánh với các ý kiến khác để đánh giá tính thống nhất của dữ liệu.

Tuy nhiên, phương pháp này có thể gặp khó khăn khi đánh giá các ý kiến phức tạp hoặc khó hiểu, đặc biệt là trong trường hợp có sự không đồng nhất giữa các ý kiến liên quan đến cùng một chủ đề.

**CHƯƠNG 2**

**ĐẶT TÊN NGẮN GỌN LIÊN QUAN ĐẾN CÂU HỎI ĐẶT RA**

**(Ví dụ Phân tích các ý kiến phản hồi về Chương trình Đào tạo ngành Khoa học dữ liệu ở Đại học Công nghiệp TP Hồ Chí Minh)**

**2.1 Câu hỏi đặt ra**

Trình bày ít nhất 2 câu hỏi có liên quan đến 5 chủ đề dưới đây, lý giải tại sao lại đặt ra câu hỏi này, ý nghĩa của câu hỏi đặt ra.

| STT | Mã chủ đề | Diễn giải |
| --- | --- | --- |
| 1 | CSVC | Cơ sở vật chất |
| 2 | GV | Đội ngũ giảng viên |
| 3 | CTDT | Chương trình đào tạo |
| 4 | CTSV | Công tác sinh viên |
| 5 | TVTS | Tư vấn tuyển sinh |

**2.2 Xây dựng câu hỏi khảo sát**

Trình bày cách thức xây dựng ít nhất 10 câu hỏi câu hỏi khảo sát để trả lời cho 2 câu hỏi đặt ra ở 2.1

Liệt kê ít nhất 10 câu hỏi câu hỏi khảo sát và giải thích chi tiết các câu hỏi khảo sát này

**2.3 Phân tích**

**2.3.1 Kết quả khảo sát**

Trình bày kết quả khi tiến hành khảo sát ít nhất 10 câu hỏi khảo sát trên với ít nhất 50 bạn trong trường về các nội dung như: Cách thức tiến hành khảo sát, thời gian, số lượng, đối tượng, kết quả thu được,…

**2.3.2 Phân tích dữ liệu**

Trình bày cách tiền xử lý dữ lý dữ liệu và

Kết quả phân tích 10 câu hỏi khảo sát ở trên

**Chú ý tập trung vào các vấn đề đề bài gợi ý:**

Mỗi phân tích đều bao gồm các nội dung theo yêu cầu sau:

* Mô tả về dữ liệu
* Thực hiện các thống kê căn bản
* Tìm mối tương quan giữa các câu hỏi khảo sát và kết quả
* Xác định các yếu tố quan trọng ảnh hưởng đến kết quả
* Trực quan hóa dữ liệu và kết quả

**2.4 Kết luận**

Trình bày kết luận sau khi thực hiện bài tập này

Những điểm còn hạn chế và hướng phát triển trong tương lai

**CHƯƠNG 3**

**KHAI PHÁ DỮ LIỆU TỪ CÂU NÓI CỦA NGƯỜI NỔI TIẾNG**

**3.1 Thu thập dữ liệu**

Trình bày cách thu thập dữ liệu và kết quả dữ liệu thu thập được

**3.2 Khai phá dữ liệu**

**3.2.1. Xử lý dữ liệu- Data Imputation**

Trình bày về yêu cầu, cách tiếp cận giải quyết, kết quả dạt được, đánh giá, thảo luận, v.v.phù hợp để bạn hoàn thành yêu cầu này một cách hiệu quả.

***Yêu cầu***:

- Bạn hãy thêm vào Trường STT và điền tự động dữ liệu của trường này?

- Một số giá trị của dữ liệu Trường ngày sinh chưa có, bạn hãy đề xuất cách điền?

- Bạn hãy thêm vào Trường Tuoi (Tuổi) và đề xuất cách điền tuổi của các tác giả?

**3.2.2. Khám phá dữ liệu- Data Exploration**

Trình bày về yêu cầu, cách tiếp cận giải quyết, kết quả dạt được, đánh giá, thảo luận, v.v*.* phù hợp để bạn hoàn thành yêu cầu này một cách hiệu quả.

***Yêu cầu***: Bạn cần khám phá dữ liệu để hiển thị một số thông tin thống kê và phân tích của tập dữ liệu đã cho. Chẳng hạn như:

- Thống kê về tác giả và câu nói nổi tiếng có trong bộ dữ liệu,

- Thống kê về năm sinh và độ tuổi của các tác giả,

- Thống kê về các câu nói nổi tiếng như: câu dài nhất, ngắn nhất, số từ, …

- Thống kê về các từ được sử dụng trong các câu nói,

- Phân tích, trực quan mối quan hệ giữa giữa tác giả và câu nói nổi tiếng,

- Phân tích, trực quan mối quan hệ giữa các tác giả với nhau,…

Trên đây chỉ là một số gợi ý, bạn có thể đề xuất thêm các phân tích, thống kê khác.

**3.2.3. Trích xuất đặc trưng- Feature Extraction**

Trình bày về yêu cầu, cách tiếp cận giải quyết, kết quả dạt được, đánh giá, thảo luận, v.v*.* phù hợp để bạn hoàn thành yêu cầu này một cách hiệu quả.

***Yêu cầu***: Hãy đề xuất cách trích xuất đặc trưng từ bộ dữ liệu đã cho, cung cấp lý do và giải thích cách làm của bạn.

**3.2.4. Suy luận**

Trình bày về yêu cầu, cách tiếp cận giải quyết, kết quả dạt được, đánh giá, thảo luận, v.v*.* phù hợp để bạn hoàn thành yêu cầu này một cách hiệu quả.

***Yêu cầu***: Bạn được yêu cầu phân loại câu nói theo tên người nổi tiếng và tính độ tương đồng phong cách nói giữa các tác giả theo 2 yêu cầu sau:

- Hãy dự đoán tên của người nổi tiếng theo câu nói dựa trên các đặc trưng bạn trích xuất ở trên và đánh giá trên bộ dữ liệu đã cho với tỉ lệ Train/Test và các độ đo phù hợp?

- Hãy đề xuất cách tính độ tương đồng phong cách nói giữa các tác giả và tìm ra các tác giả có phong cách nói tương đồng nhau nhất?

**LÀM VIỆC NHÓM**

Trình bày tóm tắt cách thức làm việc nhóm

Phân chia công việc của các thành viên trong nhóm

Tổng số lần gặp nhau (tính theo buổi)

Tổng thời gian gặp nhau (tính theo giờ)

**TÀI LIỆU THAM KHẢO**

**Chú ý các tài liệu tham khảo phải được chỉ mục vào trong các Chương nếu có tham khảo, phải có ít nhất từ 5-15 tài liệu tham khảo chính thống: Sách, bài báo Hội nghị, tạp chí - journal, … và tài liệu này phải được đánh chỉ mục từ nhỏ tới lớn và gắn vào trong báo cáo.**

**Các hình ảnh tham khảo hay các phần tham khảo ít thì có thể đánh thành footnote ngay trong trang tham khảo**

1. Quách Ngọc Ân (1992), “Nhìn lại hai năm phát triển lúa lai”, *Di tuyền học ứng dụng*, 98(1), tr. 10-16.
2. Bộ nông nghiệp & PTNT (1996), *Báo cáo tổng kết 5 năm (1992-1996) phát triển lúa lai,* Hà Nội.
3. Nguyễn Hữu Đống, Đào Thanh Bằng, Lâm Quang Dụ, Phan Đức Trực (1997), *Đột biến –* *Cơ sở lý luận và ứng dụng,* Nhà xuất bản nông nghiệp, Viện khoa học kỹ thuật nông nghiệp Việt Nam, Hà Nội.
4. Nguyễn Thị Gấm (1996), *Phát hiện và đánh giá một số dòng bất dục đực cảm ứng nhiệt* *độ,* Luận văn thạc sĩ khoa học nông nghiệp, Viện khoa học kỹ thuật nông nghiệp Việt Nam, Hà Nội.

……….

1. Anderson J.E. (1985), The Relative Inefficiency of Quota, The Cheese Case, *American* *Economic Review*, 75(1), pp. 178-90.
2. Borkakati R. P.,Virmani S. S. (1997), Genetics of thermosensitive genic male sterility in Rice, *Euphytica* 88, pp. 1-7.
3. Boulding K.E. (1955), *Economics Analysis*, Hamish Hamilton, London.
4. Burton G. W. (1988), “Cytoplasmic male-sterility in pearl millet (penni-setum glaucum L.)”, *Agronomic Journal* 50, pp. 230-231.
5. Central Statistical Oraganisation (1995), *Statistical Year Book*, Beijing.
6. FAO (1971), *Agricultural Commodity Projections (1970-1980)*, Vol. II. Rome.
7. Institute of Economics (1988), *Analysis of Expenditure Pattern of Urban Households in* *Vietnam,* Departement pf Economics, Economic Research Report, Hanoi.

**PHỤ LỤC**

Phần này bao gồm những nội dung cần thiết nhằm minh họa hoặc hỗ trợ cho nội dung đồ án như số liệu, biểu mẫu, tranh ảnh. . . . nếu sử dụng những câu trả lời cho một *bảng câu hỏi thì bảng câu hỏi mẫu này phải được đưa vào phần Phụ lục ở dạng nguyên bản* đã dùng để điều tra, thăm dò ý kiến; **không được tóm tắt hoặc sửa đổi**. Các tính toán mẫu trình bày tóm tắt trong các biểu mẫu cũng cần nêu trong Phụ lục của luận văn. Phụ lục không được dày hơn phần chính của đồ án

**MỘT SỐ CHÚ Ý KHI VIẾT BÁO CÁO**

1. Thống nhất kích cỡ chữ, kiểu chữ trong toàn bộ báo cáo. Không tô màu chữ, chỉ dùng màu đen
2. Các công thức phải tự gõ và đánh số theo Chương, ví dụ 1.1, 2.1, 2.2, 2.3
3. Các hình và Bảng phải đánh số theo chương, ví dụ Hình 1.1, Hình 2.1, Bảng 3.1, Bảng 3.2
4. Các hình nếu lấy ở ngoài phải đề footnote chú thích nguồn ở dưới
5. Hình mô hình tổng quát phải tự vẽ bằng Word, không dán hình
6. Các tài liệu tham khảo phải đính vào luận văn theo thứ tự từ nhỏ tới lớn, bắt đầu từ 1, ít nhất phải từ 5-15 tài liệu tham khảo, lựa chọn các tài liệu tham khảo mới
7. Tóm tắt trình bày được các nội dung sau: giới thiệu, phương pháp làm, kết quả, nhận xét (không dùng hình, bảng ở mục này)

**TỰ ĐÁNH GIÁ**

| Câu | Nội dung | Điểm chuẩn | Tự chấm | Ghi chú |
| --- | --- | --- | --- | --- |
| Chương 1  (15 điểm) | Thu thập dữ liệu | 15 |  |  |
| Chương 2  (35 điểm) | **2.1 Câu hỏi đặt ra** | 5 |  |  |
| 2.2 **Xây dựng câu hỏi khảo sát** | 5 |  |  |
| **2.3 Phân tích** |  |  |  |
| **2.3.1 Kết quả khảo sát** | 10 |  |  |
| **2.3.2 Phân tích dữ liệu** | 15 |  |  |
| Chương 3  (50 điểm) | **3.1 Thu thập dữ liệu** | 15 |  |  |
| **3.2 Khai phá dữ liệu** |  |  |  |
| **3.2.1. Xử lý dữ liệu** | 3 |  |  |
| **3.2.2. Khám phá dữ liệu** | 12 |  |  |
| **3.2.3. Trích xuất đặc trưng** | 5 |  |  |
| **3.2.4. Suy luận** | 15 |  |  |
| **Báo cáo** | **(**chú ý các chú ý 2,3,4,6 ở trang trước, nếu sai sẽ bị trừ điểm nặng**)** | 10đ |  |  |
| **Làm việc nhóm** | Chú ý trả lời đúng 4 yêu cầu trong phần làm việc nhóm | 10đ |  |  |
| **Tổng điểm (120)** | | |  |  |
| **Đôi qua thang điểm 10 (120=10 điểm)** | | |  |  |