|  |
| --- |
| Trường Đại học Khoa học tự nhiên – Khoa Công nghệ thông tin. |
| Đồ án thực hành cuối kỳ. |
| Programming for Data Science – Lập trình cho Khoa học dữ liệu. |

|  |
| --- |
| Nhóm 3TT  Tháng 12, 2024. |

**TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN**

**KHOA CÔNG NGHỆ THÔNG TIN**

**📖👨‍💻📖**

**A blue and white logo

Description automatically generated**

**ĐỒ ÁN THỰC HÀNH CUỐI KÌ**

**Bộ môn:** Lập trình cho khoa học dữ liệu.

**Tên đề tài:**

*“Phân tích, đánh giá các yếu tố gây bệnh đau tim”.*

**Tên nhóm:** 3TT.

**Thành viên:**

1. 22120384 – Nguyễn Đình Trí.
2. 22120398 – Vũ Hoàng Nhật Trường.
3. 22120412 – Nguyễn Anh Tường.

# **Thông tin chung:**

1. **Bộ môn:** Lập trình cho khoa học dữ liệu.
2. **Giảng viên lý thuyết:** Thầy Phạm Trọng Nghĩa.
3. **Giảng viên thực hành:** Thầy Lê Nhựt Nam.
4. **Mã lớp:** 22\_21.
5. **Tên nhóm:** 3TT.
6. **Danh sách thành viên:**
   1. 22120384 – Nguyễn Đình Trí.
   2. 22120398 – Vũ Hoàng Nhật Trường.
   3. 22120412 – Nguyễn Anh Tường.
7. **Link github repository:** [“Click here to go to our github repository.”](https://github.com/richardnguyen0715/Pro4DS-FinalProject-HandleKaggleDatasetProblems)

Contents

[**ĐỒ ÁN THỰC HÀNH CUỐI KÌ** 2](#_Toc185067394)

[**Thông tin chung:** 3](#_Toc185067395)

[**Section 0:** *Bảng phân công công việc.* 6](#_Toc185067396)

[**I.** **Bảng phân công công việc:** 7](#_Toc185067397)

[**Section 1:** *Data Collection.* 8](#_Toc185067398)

[**I.** **Giới thiệu về tập dữ liệu:** 9](#_Toc185067399)

[**II.** **Một số thông tin khác:** 9](#_Toc185067400)

[**Section 2:** *Data Preprocessing.* 10](#_Toc185067401)

[**I.** **Ý nghĩa các cột trong tập dữ liệu:** 11](#_Toc185067402)

[**II.** **Xét các dòng trùng lặp:** 13](#_Toc185067403)

[**III.** **Kiểu dữ liệu của các cột trong tập dữ liệu:** 14](#_Toc185067404)

[**IV.** **Kiểm tra các giá trị thiếu khuyết của các cột:** 14](#_Toc185067405)

[**V.** **Các kết quả chỉ số liên quan đến thống kê:** 15](#_Toc185067406)

[**VI.** **Kiểm tra điều kiện tồn tại của các giá trị:** 15](#_Toc185067407)

[**VII.** **Xét các phân bố của các giá trị trong cột:** 16](#_Toc185067408)

[**Section 3:** *Data Exploration.* 19](#_Toc185067409)

[**I.** **Góc nhìn sức khỏe cá nhân** 20](#_Toc185067410)

[**1.1** **Chỉ số BMI có liên quan đến tỉ lệ mắc bệnh tim không?** 20](#_Toc185067411)

[**1.2** **Số giờ ngủ có ảnh hưởng đến tình trạng sức khỏe tổng quát như thế nào?** 21](#_Toc185067412)

[**1.3** **Mối liên hệ giữa BMI, thói quen hút thuốc và bệnh tim?** 23](#_Toc185067413)

[**1.4** **Số giờ ngủ và sức khỏe tinh thần có quan hệ gì đối với bệnh tiểu đường không?** 25](#_Toc185067414)

[**1.5** **Mối quan hệ giữa vận động thể chất và bệnh tim?** 27](#_Toc185067415)

[**1.6** **Tình trạng sức khỏe tổng quát có ảnh hưởng đến khó khăn khi đi lại không?** 28](#_Toc185067416)

[**1.7** **So sánh số ngày sức khỏe tinh thần không tốt giữa người uống và không uống rượu** 29](#_Toc185067417)

[**1.8** **Mối quan hệ giữa thời gian ngủ và người hút thuốc?** 31](#_Toc185067418)

[**1.9** **Mối quan hệ giữa chỉ số BMI và tình trạng ung thư da?** 32](#_Toc185067419)

[**1.10** **So sánh tỉ lệ mắc bệnh tim của những người có tiền sử bị và không bị đột quỵ?** 33](#_Toc185067420)

[**II. Góc nhìn cộng đồng và yếu tố nhân khẩu học** 34](#_Toc185067421)

[**2.1** **Nam và nữ có nguy cơ mắc bệnh tim như thế nào?** 34](#_Toc185067422)

[**2.2** **Liệu tuổi tác ảnh hưởng như thế nào đến sức khỏe tinh thần?** 35](#_Toc185067423)

[**2.3** **Tỷ lệ mắc bệnh tiểu đường ở các chủng tộc khác nhau như thế nào?** 37](#_Toc185067424)

[**2.4** **Tỷ lệ mắc bệnh tim ở các chủng tộc khác nhau như thế nào?** 38](#_Toc185067425)

[**2.5** **Việc sử dụng đồ uống có cồn có ảnh hưởng tới thận ở các nhóm tuổi khác nhau không?** 39](#_Toc185067426)

[**2.6** **So sánh khả năng bị hen suyễn giữa nam và nữ?** 40](#_Toc185067427)

[**2.7** **Tuổi tác có phải là yếu tố làm gia tăng nguy cơ mắc bệnh tim không?** 41](#_Toc185067428)

[**2.8** **Thời lượng ngủ trung bình thay đổi như thế nào giữa các nhóm tuổi khác nhau?** 42](#_Toc185067429)

[**2.9** **Sức khỏe tổng quan giữa các chủng tộc có sự khác biệt gì nổi bật?** 43](#_Toc185067430)

[**2.10** **Có mối liên hệ nào giữa giới tính và tỷ lệ đột quỵ không?** 44](#_Toc185067431)

[**III.** **Góc nhìn về các yếu tố hành vi và thói quen sống** 45](#_Toc185067432)

[**3.1** **Mối liên hệ giữa thói quen hút thuốc và bệnh tim?** 45](#_Toc185067433)

[**3.2** **Tác động của việc vận động thể chất đến tình trạng sức khỏe tổng quát** 46](#_Toc185067434)

[**3.3** **Mối quan hệ giữa giấc ngủ, sức khỏe tâm lý và bệnh tim là gì?** 47](#_Toc185067435)

[**3.4** **Tác động của rượu bia đến thời lượng ngủ hàng ngày?** 48](#_Toc185067436)

[**3.5** **Thời lượng ngủ có liên quan như thế nào tới đột quỵ?** 49](#_Toc185067437)

[**3.6** **Liệu những người hút thuốc có thói quen uống rượu bia không?** 50](#_Toc185067438)

# **Section 0:***A blue line drawing of a server Description automatically generated**Bảng phân công công việc.*

1. **Bảng phân công công việc:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Công việc** | **Người thực hiện** | **Tiến độ** | **Đóng góp** | **Note** |
| Data Collection | Nguyễn Anh Tường | 100% | 100% |  |
|  | | | | |
| Data Preprocessing | Nguyễn Anh Tường | 100% | 90% | Trí thêm phần ý nghĩa của các cột. |
| Nguyễn Đình Trí | 10% |
|  | | | | |
| Data Exploration | Nguyễn Đình Trí | 100% | 50% |  |
| Vũ Hoàng Nhật Trường | 50% |

**100%**

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

# **Section 1:** *Khó khăn trong lúc làm việc.*

***A blue line drawing of a server

Description automatically generated***

1. **Nguyễn Anh Tường:**

* Khó khăn trong lúc chọn lựa dataset.
* Về mặt ý nghĩa thì tập dữ liệu này tương đối hàn lâm nhưng nhóm em đã có kinh nghiệm xử lý dataset theo chủ đề y học ở học phần trực quan hóa dữ liệu nên vì thế nhóm chọn dataset này – Heart Disease.

1. **Nguyễn Đình Trí:**
2. **Vũ Hoàng Nhật Trường:**

# **Section 2:** *Học được gì?.*

***A blue line drawing of a server

Description automatically generated***

1. **Nguyễn Anh Tường:**

* Cách quản lý đội nhóm.

1. **Nguyễn Đình Trí:**
2. **Vũ Hoàng Nhật Trường:**

# **Section 3:** *Nếu có thêm thời gian thì làm gì?.*

***A blue line drawing of a server

Description automatically generated***

1. **Xây dựng mô hình:**

* Có thể là một mô hình hồi quy phi tuyến dự đoán khả năng một người mắc bệnh tim.
* Có thể là một mô hình phân loại người bệnh tim theo mức độ - thông qua các thuật toán phân cụm, gôm cụm ( ví dụ: K-Mean Clustering ).

1. **Đánh giá mô hình:**

* Đánh giá tính chính xác của các dự đoán.
* Đánh giá phân loại của cả bệnh nhân ( bao nhiêu nhóm, tình trạng như thế nào,…)