Đại học Quốc gia Thành phố Hồ Chí Minh
Trường Đại học Khoa học Tự nhiên
Khoa Công nghệ Thông tin

ĐỒ ÁN THỰC HÀNH
CSC00004 – NHẬP MÔN CÔNG NGHỆ THÔNG TIN (CHO KHOA HỌC DỮ LIỆU)

1 TỔNG QUAN

Loại bài tập	Đồ án thực hành
Thời gian thực hiện	(Dự kiến) 4 tuần
Deadline nộp bài	(Dự kiến) 15/1/2023 23:59:00
Hình thức thực hiện	Theo nhóm (max 4 thành viên)
Hình thức nộp bài	Nộp thông qua Moodle FIT
GV hướng dẫn	Kiều Vũ Minh Đức Lê Nhựt Nam
Thông tin liên lạc	kvmduc3@gmail.com (KVM Đức) lenam.fithcmus@gmail.com (LN Nam)

2 CHUẨN ĐẦU RA CẦN ĐẠT

Đồ án này nhằm mục tiêu đạt được các chuẩn đầu ra sau:

- Tham gia thảo luận, tranh luận theo nhóm trên chủ đề môn học.
- Phân tích, tổng hợp và viết tài liệu kỹ thuật theo mẫu cho trước theo cá nhân hoặc cộng tác nhóm.
- Liệt kê và giải thích ít nhất <n> thuật ngữ tiếng Anh ngành và chuyên ngành.
- Áp dụng những kiến thức cơ bản của 5 chuyên ngành KHMT, CNTT, CNPM, HTTT và MMT có liên quan được thể hiện trong đề tài. (Tập trung định hướng Khoa học Dữ liệu)
- Tìm, sử dụng, so sánh các công cụ hợp pháp để thực hiện các yêu cầu cụ thể.
- Báo cáo (viết và miệng) quá trình xây dựng và hướng dẫn sử dụng sản phẩm.
-

3 CÔNG NGHỆ SỬ DỤNG

Các công nghệ/ công cụ sử dụng cho đồ án

MongoDB	https://www.mongodb.com/
Mongo Compass	https://www.mongodb.com/products/compass
Mongo Atlas	https://www.mongodb.com/atlas/database
Ngôn ngữ lập trình Python	https://www.python.org/
Jupyter Notebooks	https://jupyter.org/
PyMongo	https://pymongo.readthedocs.io/en/stable/
Sckit-learn	https://scikit-learn.org/

4 MÔ TẢ ĐỒ ÁN

Nghiên cứu và làm việc với cơ sở dữ liệu phi quan hệ phân tán đơn giản.

a) Giới thiệu

"*Distributed programming is the art of solving the same problem that you can solve on a single computer using multiple computers*."

Hai tác vụ cơ bản trong bất kỳ một hệ thống tính toán hiện nay nào là:

- Lưu trữ (storage)
- Tính toán (computation)

Với lượng tài nguyên và thời gian nghiên cứu phát triển hữu hạn, lập trình viên nói chung cần phải đối mặt với nhiều vấn đề như kích thước của bài toán vượt khỏi khả năng tính toán của hệ thống hiện tại. Tính toán phân tán cho phép người ta có thể xử lý bài toán trên một máy đơn bằng việc sử dụng nhiều máy tính khác.

Trong ngành Khoa học Dữ liệu, xử lý và phân tích dữ liệu đóng vai trò quan trọng, một trong những công cụ thường hay xử dụng đó là Jupyter Notebook và Python. Python cho phép xử lý, phân tích số liệu và trực quan hóa với nhiều thư viện có sẵn như NumPy, Pandas, Matplotlib, … Hơn nữa, để có thể mô hình hóa và giải quyết một số bài toán như dự đoán hay phân lớp, Sckit-learn là một trong những thư viện rất dễ sử dụng.

b) Nội dung đồ án

Cho trước tập dữ liệu iris được phân tách sẵn. Tập dữ liệu bao gồm các tập tin như sau:

- iris_train.csv: Tập dữ liệu Iris, gồm 4 cột và 120 dòng. Cột index chỉ thứ tự của các quan sát, cột sepal.length chỉ chiều dài của đài hoa Iris, cột sepal.width chỉ chiều rộng của đài hoa Iris, cột petal.length chỉ chiều dài của cánh hoa Iris, cột petal.width chỉ chiều rộng của cánh hoa Iris, và cột variety chỉ loại hoa Iris (0 - Iris Setosa, 1 - Iris Versicolour, 2- Iris Virginica)
- iris_test_input.csv: Tập dữ liệu kiểm tra Iris, gồm 4 cột và 30 dòng. Cột index chỉ thứ tự của các quan sát, cột sepal.length chỉ chiều dài của đài hoa Iris, cột sepal.width chỉ chiều rộng của đài hoa Iris, cột petal.length chỉ chiều dài của cánh hoa Iris, cột petal.width chỉ chiều rộng của cánh hoa Iris.
- iris_test_label.csv: Tập nhãn Iris đúng để kiểm tra, gòm 2 cột và 30 dòng. Cột index chỉ thứ tự của các quan sát và cột variety chỉ loại hoa Iris (0 - Iris Setosa, 1 - Iris Versicolour, 2- Iris Virginica)

Sinh viên thực hiện tạo cơ sở dữ liệu với ba tập tin cho trước với Mongo Atlas với:

- Database name: kdl_nmcntt_iris
- Collection name lần lượt là: iris_train, iris_test_input và iris_test_label

Sau đó, sinh viên sử dụng PyMongo để truy xuất cơ sở dữ liệu và đọc dữ liệu bằng Pandas. Kế tiếp, sinh viên sử dụng Sckit-learn với các mô hình học máy như sau để huấn luyện, đưa ra kết quả dự đoán được dựa trên dữ liệu iris_test_input, và lưu vào cơ sở dữ liệu với collection name là iris_predicted.

- Logistic Regression
- Support Vector Machine

c) Quy trình thực hiện đồ án

Đồ án được thực hiện theo qui trình Scrum, với các vai trò như sau:

- Product Owner: GVTH đóng vai trò là Product Owner, là người đưa ra các yêu cầu về đồ án mà nhóm cần thực hiện. Ngoài ra, hàng tuần (trong giờ thực hành) Scrum Master đại diện nhóm, báo cáo tiến độ với Product Owner.

- Scrum Master: nhóm trưởng của nhóm đóng vai trò là Scrum Master, là người quản lí nhóm, đưa ra danh sách các công việc, nhận kết quả của từng thành viên và báo cáo hàng tuần với Product Owner.
- Business Analyst: mỗi nhóm cử ra 1-2 thành viên đóng vai trò là Business Analyst, người trực tiếp nghe yêu cầu từ phía khách hàng, truyền đạt lại với các thành viên trong nhóm.
- Designer: đóng vai trò là người đưa ra các thiết kế về xử lí các chức năng mà khách hàng yêu cầu, thiết kế về resource, thiết kế dữ liệu lưu trữ.
- Developer: cài đặt bằng ngôn ngữ theo thiết kế đề ra (trong phạm vi đồ án, sinh viên có thể tìm các ứng dụng đã có sẵn không cần cài đặt).
- Tester: kiểm tra các chức năng đã được hoàn thành chưa, trước khi báo cáo hàng tuần với Product Owner. Nếu chưa hoàn thành, tester cần ghi nhận lại để yêu cầu các thành viên khác trong nhóm chỉnh sửa.
- Do nhóm có 4 thành viên, nên ngoài vai trò Scrum Master (do nhóm trưởng làm), mỗi vai trò khác nên có 2-3 người đảm nhận.

5 HƯỚNG DẪN CHI TIẾT THEO TUẦN

Nội dung thực hiện được viết trên cơ sở 4 tuần thực hiện

Tuần 7	Công bố các đề tài đồ án thực hành
Tuần 8	Báo cáo giai đoạn 1
Tuần 9	Báo cáo giai đoạn 2
Tuần 10	Báo cáo giai đoạn 3
Tuần 11	Vấn đáp đồ án

Nhóm (đứng đầu là trưởng nhóm với vai trò Scrum Master) báo cáo hàng tuần cho Giảng viên hướng dẫn thực hành (vai trò Product Owner) trong giờ thực hành hàng tuần.

Kết quả thảo luận nhóm sẽ phải được trình bày trong meeting minutes. Ghi lại những nội dung

sau: Những vấn đề nhóm đã thảo luận, kết quả thảo luận và tìm cách giải quyết, phân công công việc cho lần gặp tới ra sao.

6 CÁC QUY ĐỊNH CHO ĐỒ ÁN

Các kết quả cần đạt được

- **Project plan.** Kế hoạch thực hiện dự án theo qui trình Scrum.
- **Website** đăng các hướng dẫn của một qui trình với các công cụ. Hướng dẫn các công cụ được trình bày hướng dẫn bằng bài viết và hình ảnh, trong đó có ít nhất một công cụ hướng dẫn có kết hợp video. Khuyến khích các cơ chế trình bày tương tác cao hơn (có kịch bản).
- **Meeting minutes**. Ghi lại nội dung các cuộc họp của nhóm. Các thành phần cơ bản gồm có: Các công việc của tuần trước, tiến độ của các công việc này (đã xong, các vấn đề còn tồn động và các giải pháp đã chọn sau khi thảo luận, các công việc mới, phân công công việc cho tuần mới).
- **Uri**. Lưu lại uri cơ sở dữ liệu sau cùng của nhóm.
- **Source code**. Lưu lại các source code trong quá trình nhóm đã thực hiện.
- **Reflective report**. Báo cáo rút kinh nghiệm, nội dung gồm: Những điểm đã làm tốt, những điểm còn tồn đọng. Mức độ đóng góp của các thành viên.

7 ĐÁNH GIÁ

Việc đánh giá đồ án của sinh viên được dựa trên các phần như sau:

Tiêu chí	**Tỷ lệ**
Báo cáo đầy đủ và chi tiết	20%
Xây dựng được cơ sở dữ liệu như yêu cầu	20%
Sử dụng các thư viện để đọc dữ liệu và mô hình hóa	15%
Đưa ra kết quả dự đoán và báo cáo độ chính xác	15%
Đưa kết quả đã dự đoán được lên cơ sở dữ liệu	10%
Trình bày vấn đáp	10%

Trả lời câu hỏi	10%
Tổng cộng	**100%**

--- Good luck :) ---