TRƯỜNG ĐẠI HỌC VĂN LANG
**ĐƠN VỊ: KHOA KỸ THUẬT CƠ – ĐIỆN VÀ MÁY TÍNH**

**ĐỀ THI/ĐỀ BÀI, RUBRIC VÀ THANG ĐIỂM**
**THI KẾT THÚC HỌC PHẦN**
**Học kỳ 2, năm học 2023-2024**

## I. Thông tin chung

| | | | |
|---|---|---|---|
| Tên học phần: | Phân tích dữ liệu lớn | | |
| Mã học phần: | 71DSBD40014 | Số tín chỉ: | 3 |
| Mã nhóm lớp học phần: | 232_71DSBD40014_01 | | |
| Hình thức thi: **Tiểu luận** | | Thời gian làm bài: **7** | ngày |
| ☐ Cá nhân | ☒ Nhóm | | |
| ***Quy cách đặt tên file*** | ***Mã SV_Ho va ten SV_Nhom*** | | |

### 1. Format đề thi

- Font: Times New Roman
- Size: 13
- Quy ước đặt tên file đề thi/đề bài:

+ Mã học phần_Tên học phần_Mã nhóm học phần_TIEUL_De 1

### 2. Giao nhận đề thi

Sau khi kiểm duyệt đề thi, đáp án/rubric. **Trưởng Khoa/Bộ môn** gửi đề thi, đáp án/rubric về Trung tâm Khảo thí qua email: khaothivanlang@gmail.com bao gồm file word và file pdf (***nén lại và đặt mật khẩu file nén***) và nhắn tin + họ tên người gửi qua số điện thoại **0918.01.03.09** (Phan Nhất Linh).

## II. Các yêu cầu của đề thi nhằm đáp ứng CLO

*(Phần này phải phối hợp với thông tin từ đề cương chi tiết của học phần)*

| Ký hiệu CLO | Nội dung CLO | Hình thức đánh giá | Trọng số CLO trong thành phần đánh giá (%) | Câu hỏi thi số | Điểm số tối đa | Lấy dữ liệu đo lường mức đạt PLO/PI |
|---|---|---|---|---|---|---|
| (1) | (2) | (3) | (4) | (5) | (6) | (7) |
| **CLO1** | Nắm vững các khái niệm cơ bản về dữ liệu lớn, xác suất và thống kê | Tiểu luận | 15% | 1,2,3 | 2,3,5 | PI1.1 |
| **CLO2** | Tận dụng hiểu biết nâng cao về dữ liệu lớn và lập trình Python để vận hành hệ thống quản lý, xử lý và phân tích dữ liệu lớn | Tiểu luận | 30% | 2,3 | 3,5 | PI2.1 |
| **CLO3** | Nắm vững kiến thức liên quan của các thuật toán Máy học | Tiểu luận | 20% | 2,3 | 3,5 | PI3.1 |
| **CLO4** | Vận dụng linh hoạt các kỹ thuật học máy trong phân tích dữ liệu lớn | Tiểu luận | 20% | 1,2,3 | 2,3,5 | PI4.1 |
| **CLO5** | Có ý thức tự tìm hiểu, học hỏi, áp dụng các kỹ thuật mới trong quản trị và phân tích dữ liệu lớn | Tiểu luận | 15% | 1,2,3 | 2,3,5 | PI10.1 |

## III. Nội dung đề bài

### 1. Đề bài

**Câu 1 (2 điểm):** Sinh viên chọn 1 trong các chủ đề bên dưới và trình bày các kiến thức liên quan đến chủ đề gồm: Khái niệm, Ứng dụng trong thực tế và liên hệ với lĩnh vực đang học.

- Chủ đề 1: Hệ sinh thái Hadoop và MapReduce.
- Chủ đề 2: Hadoop Distributed File System (HDFS) và Apache Spark.

**Câu 2 (3 điểm):** Sinh viên sử dụng nền tảng Cloudera trên hệ điều hành Cent OS để thực hiện chương trình WordCount.

**Lưu ý:**

- Sinh viên tự chọn 1 tài liệu tiếng Anh khoảng 2000 chữ.
- Bài báo cáo phải có hình ảnh minh họa chi tiết quá trình thực hiện chương trình WordCount.

**Câu 3 (5 điểm):**

Một nhà phát hành game cho ứng dụng điện thoại thông minh cần phân tích các giao dịch phát sinh của người dùng. Trong đó, việc giao dịch mua/bán các sản phẩm trong game là

nguồn thu chính của công ty. Các giao dịch này được lưu trong tập dữ liệu ***game_data.csv*** (đính kèm). Công ty cần xác định được người dùng nào có khả năng cao sẽ thực hiện các giao dịch trong quá trình chơi game. Với vai trò là kỹ sư Trí tuệ nhân tạo của công ty, sinh viên hãy thực hiện các yêu cầu bên dưới:

- Sử dụng **phần mềm KNIME** và **ngôn ngữ lập trình Python trên nền tảng Apache Spark** để xây dựng mô hình Machine Learning nhằm phân loại khách hàng dựa trên hành vi mua/bán sản phẩm của họ.
- So sánh kết quả trên phần mềm KNIME và trên nền tảng Apache Spark.

**Lưu ý:**

- Bài báo cáo cần mô tả rõ các nội dung chính như sau: Chuẩn bị dữ liệu, Phân chia tập dữ liệu, Đánh giá, Phân tích kết quả.
- Mô hình cần phân loại nhóm khách hàng mua những sản phẩm giá trị cao và nhóm khách hàng mua những sản phẩm giá trị thấp.

## 2. Hướng dẫn thể thức trình bày đề bài

**A. Bài làm tối thiểu phải có các nội dung sau:**

1. **Tóm tắt** (*chỉ áp dụng cho **Câu 3***):
- Giới thiệu tổng quát bài toán
- Các kỹ thuật/phương pháp liên quan
- Các ưu/khuyết điểm của các kỹ thuật/phương pháp (nếu có).
- Trình bày các nội dung thực hiện.
- Kỹ thuật/phương pháp thực hiện.

2. **Mô tả dữ liệu** (*chỉ áp dụng cho **Câu2 và Câu 3***)

3. **Kỹ thuật/phương pháp**: giải thích chi tiết kỹ thuật/phương pháp sử dụng

4. **Kết quả**
- Hình ảnh: đặt tên theo thứ tự, ảnh không bị vỡ, chữ ghi chú trong ảnh không quá nhỏ so với font size trong bài, ghi rõ nguồn hình ảnh (khuyến khích tự vẽ hình), giải thích ý nghĩa của hình ảnh.
- Bảng (mô tả dữ liệu/kết quả kèm giải thích…)

5. **Thảo luận** (*chỉ áp dụng cho **Câu 3***)

6. **Kết luận** (*chỉ áp dụng cho **Câu 3***):
- Tóm tắt nội dung, phương pháp đã làm, kết quả và hướng triển khai tiếp theo (nếu có)

**B. Hình thức trình bày**

**1. Trang bìa: Học phần, Tên đề tài, Họ tên sinh viên, Giảng viên**
**2. Mục lục**
**3. Header: Tên môn học**
**4. Footer: Tên sinh viên, Mã sinh viên, đánh số trang/Tổng số trang**
**BM-006**
**5. Độ dài: 10 – 20 trang**
**6. Canh lề:**
Lề trên: Cách mép trên từ 20 – 25mm (2cm – 2.5cm). Lề dưới: Cách mép dưới từ 20 – 25mm (2cm – 2.5cm). Lề trái: Cách mép trái từ 30 – 35 mm (3cm – 3.5cm). Lề phải: Cách mép phải từ 15 – 20 mm (1.5cm – 2cm).
**7. Font: Time New Roman, Font size: 13**

## 3. Rubric và thang điểm

| Tiêu chí | Trọng số (%) | Tốt 100% | Khá 75% | Trung bình 50% | Yếu 25% | Kém 0% |
|---|---|---|---|---|---|---|
| Hình thức trình bày | 10% | Căn chỉnh hợp lý | 1 – 3 đoạn căn chỉnh không hợp lý | 4 – 5 đoạn căn chỉnh không hợp lý | 6 – 8 đoạn căn chỉnh không hợp lý | Hơn 8 đoạn căn chỉnh không hợp lý |
| Nội dung lý thuyết | 30% | Trình bày đầy đủ các nội dung lý thuyết được sử dụng | Thiếu 1 nội dung | Thiếu 2 nội dung | Thiếu 3 nội dung | Thiếu hơn 3 nội dung |
| Giới thiệu về bộ dữ liệu | 10% | Giới thiệu rõ: nguồn và thời gian thu thập, tên và thang đo của các biến | Không giới thiệu 1 nội dung | Không giới thiệu 2 nội dung | Không giới thiệu 3 nội dung | Không giới thiệu cả 4 nội dung |
| Kết quả phân tích | 30% | Chính xác | 1 – 3 kết quả sai | 4 – 5 kết quả sai | 6 – 8 kết quả sai | Hơn 8 kết quả sai |
| Nhận xét kết quả | 20% | Khớp với kết quả phân tích | Có 1 nhận xét không khớp | Có 2 nhận xét không khớp | Có 3 nhận xét không khớp | Có hơn 3 nhận xét không khớp |

*TP. Hồ Chí Minh, ngày ..... tháng ..... năm 2024*

**Người duyệt đề** **Giảng viên ra đề**

**TS. Nguyễn Quốc Dũng** **TS. Trương Quốc Trí**