|  |  |  |
| --- | --- | --- |
|  | **TRƯỜNG ĐẠI HỌC BÁCH KHOA**  **KHOA CÔNG NGHỆ THÔNG TIN** |  |

TIỂU LUẬN CUỐI KỲ

HỌC PHẦN: KHOA HỌC DỮ LIỆU

**TÊN ĐỀ TÀI**

***Predicting Movie Ratings***

|  |  |
| --- | --- |
| Nhóm | 5 |
| Họ Và Tên Sinh Viên | Lớp Học Phần |
| Phan Khánh Ngân | 20.10 |
| Lê Văn Thịnh |
| Nguyễn Phan Nhật Hoàng |

ĐÀ NẴNG, 06/2023

**TÓM TẮT**

Tiểu luận này tập trung vào việc giải quyết bài toán dự đoán ratings, dựa vào pre-released features của các bộ phim trước khi phát hành. Thông qua sử dụng các phương pháp, kỹ thuật khai phá dữ liệu và học máy bao gồm thu thập, tiền xử lý dữ liệu, xây dựng và tinh chỉnh mô hình dự đoán, có thể xây dựng hệ thống dự đoán có độ chính xác cao. Có nhiều mô hình hồi quy có thể ứng dụng cho bào toán này, trong đó có Random Forest và Support Vector Machine. Với kết quả thu được, SVM dự đoán cho độ chính xác tốt hơn RF với MAE, RMSE thấp hơn và R2 cao hơn. Kết quả này khẳng định có thể sử dụng mô hình SVM vào hệ thống dự đoán movie ratings thực tế, giúp người sử dụng có thể đưa ra quyết định có hay không nên sản xuất một bộ phim.

**BẢNG PHÂN CÔNG NHIỆM VỤ**

|  |  |  |
| --- | --- | --- |
| Sinh viên thực hiện | Các nhiệm vụ | Tự đánh giá theo 3 mức  (Đã hoàn thành/Chưa hoàn thành/Không triển khai) |
| Nguyễn Phan Nhật Hoàng | -…   * … | -…  -… |
| Phan Khánh Ngân | * Xây dựng hai mô hình hồi quy (RF và SVM) * Triển khai thuật toán lựa chọn tham số tối ưu * Đánh giá, so sánh hai mô hình hồi quy (RF và SVM) | Đã hoàn thành |
| Lê Văn Thịnh |  |  |

**MỤC LỤC**

**CÁC YÊU CẦU VIẾT BÁO CÁO**

**VỀ HÌNH THỨC:**

* Định dạng: căn chỉnh lề kiểu Justify; font chữ 12, times new roman; giãn dòng 1.5, cách trước 6pt; hình ảnh phải đánh số, có chú thích bên dưới hình; bảng phải đánh số, có chú thích bên trên bảng.
* Báo cáo có độ dài từ 15 đến 20 trang và không đóng bìa gương (chỉ cần bìa mềm có màu) gồm: trang bìa (trang 1 của file hướng dẫn này), tóm tắt, bảng phân công nhiệm vụ,mục lục, và các trang nội dung.

**VỀ NỘI DUNG:**

Báo cáo gồm 06 phần sau.

# 1. Giới thiệu

Trong phần này, sinh viên giới thiệu vấn đề cần giải quyết và đề xuất giải pháp tổng quan cần thực hiện.

# 2. Thu thập và mô tả dữ liệu

**2.1. Thu thập dữ liệu**

SV mô tả mô tả giải pháp thu thập dữ liệu gồm nguồn dữ liệu, công cụ thu thập, cách thức sử dụng công cụ, đầu vào và đầu ra của quá trình thu thập, và cho ví dụ minh hoạ.

Yêu cầu:

* SV phải TỰ thu thập (crawl hoặc record) tập dữ liệu.
* Số lượng: >1000 mẫu.

**2.2. Mô tả dữ liệu**

SV mô tả các thống kê tổng quan về tập dữ liệu (số mẫu, số đặc trưng của 1 mẫu, kiểu dữ liệu của mỗi đặc trưng, số mẫu dữ liệu trống của mỗi đặc trưng,...) và xuất ra các thống kê mô tả trực quan về các đặc trưng (ví dụ: dùng boxplot, histogram, scatter plot về độ tương quan,…).

# 3. Trích xuất đặc trưng

SV trình bày việc lựa chọn đặc trưng, làm sạch và chuẩn hóa dữ liệu, giảm chiều dữ liệu, …; và trực quan hoá kết quả/tác dụng của các quá trình trên (trước và sau khi xử lý).

# 4. Mô hình hóa dữ liệu

SV trình bày quá trình xây dựng, xác thực và kiểm thử mô hình gồm:

- Chọn các mô hình phù hợp (ít nhất 2 mô hình) và mô tả ngắn gọn cơ sở lý thuyết của các mô hình đã lựa chọn, mô tả bộ tham số của mỗi mô hình (ví dụ: số lớp của mạng nơ-ron, số node của mỗi lớp, hàm activation được sử dụng,…) thông qua API sử dụng để xây dựng mô hình.

- Chia dữ liệu thành các tập Huấn luyện/Xác thực/Kiểm thử theo tỉ lệ phù hợp.

- Trình bày các tham số của quá trình huấn luyện mô hình (ví dụ: loss function, learning rate, number of epochs,…).

- Trình bày các đồ thị thể hiện hiệu suất (đánh giá bằng Loss hoặc Accuracy) của các mô hình trên các tập Huấn luyện/Xác thực/Kiểm thử.

- Đưa ra các độ đo (metrics) dùng để đánh giá các mô hình (ví dụ: RMSE hoặc MAPE), và so sánh hiệu quả của các mô hình bằng bảng số liệu (hoặc đồ thị) dựa trên các độ đo.

# 5. Kết luận

Sinh viên tổng kết các việc đã làm và các kết quả đạt được. Nêu hướng phát triển để cải thiện kết quả hiện có.

# 6. Tài liệu tham khảo

[Bachelor Degree Project in Informatics G2E, ECTS Spring term 2015] Karl Persson, PREDICTING MOVIE RATINGS - A comparative study on random forests and support vector machines

[1-12018] Yueming Zhang, Predict IMDB score with data mining algorithms