Nhóm 05:
Nguyễn Văn Hùng – 16110099
Lê Công Nghĩa – 16110165

Đề 01: Tìm hiểu giải thuật C4.5, C5.0 của Cây Quyết Định (Decision Tree), và tìm hiểu về Rừng Ngẫu Nhiên (Random Forest). Xây dựng ứng dụng demo

- Lưu ý: tham khảo thoải mái nhưng không được đạo văn, không được đạo code có sẵn trên các github, trên internet
- https://en.wikipedia.org/wiki/C4.5_algorithm
- https://www.ibm.com/support/knowledgecenter/en/SS3RA7_15.0.0/com.ibm.spss.modeler.help/c50node_general.htm
- https://www.rulequest.com/see5-unix.html
- http://mercury.webster.edu/aleshunas/Support%20Materials/C4.5/Nguyen-Presentation%20Data%20mining.pdf
- https://www.quora.com/What-are-the-differences-between-ID3-C4-5-and-CART
- https://rulequest.com/see5-comparison.html
- http://article.nadiapub.com/IJDTA/vol11_no1/1.pdf
- http://mercury.webster.edu/aleshunas/Support%20Materials/C4.5/Nguyen-Presentation%20Data%20mining.pdf
- https://www.researchgate.net/publication/265162251_A_comparative_study_of_decision_tree_ID3_and_C45
- https://www.researchgate.net/publication/284082342_Comparison_of_C5_0_CART_classification_algorithms_using_pruning_technique
- https://sinhvientot.net/thuat-toan-cay-quyet-dinh-c45/
- https://ongxuanhong.wordpress.com/2015/09/22/c4-5-hoi-gi-dap-nay/
- http://uet.vnu.edu.vn/~thuyhq/Student_Thesis/K46_Nguyen_Thi_Thuy_Linh_Thesis.pdf
- https://slideplayer.com/slide/11186155/
- http://kcntt.duytan.edu.vn/Home/ArticleDetail/vn/128/3419/so-sanh-thuat-toan-cay-quyet-dinh-id3-va-c45
- https://fr.scribd.com/document/333891349/c4-5-Cay-quy%E1%BA%BFt-%C4%91%E1%BB%8Bnh-Thu%E1%BA%ADt-toan-phan-l%E1%BB%9Bp-c4-5

1. YÊU CẦU:

Báo cáo cần trình bày được TỐI THIỂU các phần sau

a. Lý thuyết
   i. Giới thiệu [algorithm/topic name]
   ii. [algorithm/topic name] dùng để làm gì
   iii. Lý thuyết nền tảng hoặc các bước của [algorithm/topic name]
   iv. Các ứng dụng của [algorithm/topic name]
   v. So sánh ưu/nhược điểm của C4.5 với C5.0 so với các giải thuật Decision Tree khác (ví dụ, ID3, M5…)

b. Code demo cho 1 bài toán dùng[algorithm/topic name] với dataset tùy sinh viên lựa chọn
c. Đánh giá model
(các subsection nhỏ hơn sinh viên tự thêm vào nếu cần)

Sinh viên chú ý khi làm Đồ Án:

- Trên mạng có rất nhiều tài liệu + code demo có sẵn
- Code demo:
  - Không bắt buộc dùng thư viện scikit-learn, các nhóm có thể dùng bất kì thư viện nào
  - Không cần xây dựng giao diện ứng dụng kiểu Window forms hay Web app (tuy nhiên nếu nhóm nào làm thì rất khuyến khích và được tăng điểm)
  - Code demo nên TỰ LÀM, dataset thì có thể lấy trên mạng
    - Không khuyến khích copy hay dùng lại code demo của người khác dù là có trích dẫn nguồn gốc thì cũng không được đánh giá cao

- Không được Copy & Paste từ internet, các tài liệu khác, báo cáo đồ án của các trường ĐH khác, các hóa trước…. vào báo cáo của mình mà KHÔNG TRÍCH DẪN NGUỒN GỐC → **phạm lỗi "Đạo văn" (**)** → tự động Rớt môn học

- Không nên Copy & Paste từ internet vào báo cáo mà KHÔNG HIỂU RÕ ý nghĩa, vì:
  - Khi vấn đáp cuối HK cho đồ án, Giảng viên sẽ hỏi bất kì 1 biến số x,y,z nào đó, công thức nào đó, dòng code nào đó,… trong những cái mà sinh viên có ghi trong báo cáo
  - Câu hỏi sẽ dành cho bất kì 1 bạn nào trong nhóm (đặc biệt là những bạn hay im lặng, ít đóng góp ý kiến, hay nghỉ học, và chưa từng lên trình bày, present)
  - Nếu 1 thành viên không trả lời được (vì chỉ đi copy paste mà không hiểu ý nghĩa) → trừ điểm nhóm

- Các nhóm làm các đề có topic gần tương tự nhau (ví dụ phần "Regression" , "Rừng ngẫu nhiên"…):
  - Được phép thảo luận bàn bạc với nhau để chia sẻ kiến thức
  - Không được phép copy bài làm lẫn nhau.
    - Nếu bị phát hiện → trừ điểm cho tất cả các nhóm copy bài lẫn nhau

- Đề bài có mức độ Khó/Dễ tương đối là NHƯ NHAU. Tuy nhiên vẫn sẽ có những đề khó hơn 1 chút (dù không khác biệt nhiều). Tiêu chí chấm điểm công bằng:
  - Dựa trên độ Khó/Dễ của từng đề.
    - Nhóm nào làm đề khó hơn sẽ được chấm điểm nhẹ tay hơn so với những nhóm làm đề dễ
  - Dựa trên số lượng sinh viên trong 1 nhóm
    - Nhóm nào ít sinh viên hơn sẽ được chấm điểm nhẹ tay hơn so với những nhóm đông
  - Dựa trên mức độ nội dung tìm hiểu + trình bày của nhóm
  - Dựa trên code, dataset
  - Dựa trên phần trả lời vấn đáp cuối HK

- **(**)** "Đạo văn" là gì?
  - https://www.hotcourses.vn/study-abroad-info/choosing-a-university/dao-van-pragiasism-la-gi/
  - https://baomoi.com/chong-dao-van-co-ai-noi-cho-cac-em-dau/c/25227423.epi
  - https://vietpsy.wordpress.com/2013/07/14/tranh-dao-van/

2. BÁO CÁO TIẾN ĐỘ:

- Nộp file World/pdf báo cáo tiến độ và những gì đã làm/tìm hiểu được vào các thời hạn sau:
  - 1/10/2019
  - 15/10/2019
  - 1/11/2019
  - 15/11/2019
  - 1/12/2019
  - 15/12/2019
  - Gợi ý: các nhóm bắt đầu viết báo cáo đồ án từ đầu, và dùng chính bản báo cáo đó làm báo cáo tiến độ, giảng viên sẽ theo dõi tiến độ dựa trên việc so sánh các bản báo cáo sau mỗi 2 tuần
- Nộp qua email, không nhận nộp qua facebook
  - Email: sangtlm@fit.hcmute.edu.vn
  - Title bắt buộc (*giống nhau cho tất cả các lần gửi mail báo cáo tiến độ*) :
    Nhóm 05 – Đề 01 – báo cáo tiến độ
- Hẹn gặp giảng viên hướng dẫn: linh động

3. NỘP BÀI:

- Mỗi nhóm in báo cáo:
  - In 2 mặt
  - Đóng bìa mềm, không cần đóng bìa cứng, không cần bìa kiếng
  - Không cần in màu
  - Trong báo cáo nếu có đồ thị, biểu đồ, hình ảnh nào cần in màu để cho thấy sự phân tách rõ hơn của các nhóm, lớp, cụm…. thì hãy in màu

- Thời hạn nộp báo cáo đồ án: **30/12/2019**
  - Nộp trễ:
    - Từ 0h ngày 31/12/2019 bị tính là nộp trễ
    - Mỗi ngày nộp trễ trừ 1 điểm cho toàn bộ nhóm
  - Nộp sớm: không cộng điểm
  - Report nộp bản giấy
  - Source code + Dataset gửi qua email/ google drive
    - Title email bắt buộc: **Nhóm 05 – Đề 01 – báo cáo đồ án - final**

- Đồ án không thuyết trình/ present nhưng có thi vấn đáp

- Vấn đáp: từ 02/01/2020, thời gian chính xác sẽ thông báo sau
    - Câu hỏi sẽ dành cho bất kì 1 thành viên nào trong nhóm (đặc biệt là những bạn hay im lặng, ít đóng góp ý kiến, hay nghỉ học, và chưa từng lên trình bày, present)