**Middle Exam Project**

**Course: Introduction to Machine Learning**

**Guidelines for completing the assignment:**

* 2 people per group (in special cases, 3 people).
* All members of the group must work on all tasks. Do not split the tasks where each person does separate parts, and one person does not know what the other has done.
* One file PDF describing the problem and solution for all questions.
* Code files and the data file for all questions.
* One slide file for Presentation.
* Save all results in one folder, compress the folder, and submit the compressed file. The folder name should be a combination of the FirstName\_StudentID of the members.
* All members must submit the assignment.

**Requirement:**

Find a dataset that can be modeled for both classification and regression tasks in machine learning. Use data from UCI datasets.

Use data with attributes that include both numerical and categorical types.

**Question 1 (5 points):**

Use the machine learning methods you have learned (you may also use additional machine learning methods) to solve this problem for both classification and regression approaches. Perform the following steps:

* Provide a statistical summary of the data using data visualization.
* Preprocess the data: type conversion and data normalization.
* Split the data into training and evaluation sets.
* Perform classification (and regression).
* Evaluate and compare the results of the methods: compare classification methods with each other; compare regression methods with each other. Use charts to visually display the comparison.

**Question 2 (3 points):**

* Research and present the concept of overfitting in general, and solutions to address overfitting for the methods in Question 1 (note: you can choose one or more methods to solve overfitting).
* Implement code that includes the solution to overfitting into the methods in Question 1; compare the results with and without overfitting.

**Question 3 (2 points):**

* Research and present feature selection using correlation analysis. Apply this to the problem for the methods in Question 1.

Vietnamese translation:

**Yêu cầu:**

Tìm một tập dữ liệu mà có thể mô hình hoá thành được cả bài toán phân loại hoặc bài toán hồi quy trong học máy. Sử dụng dữ liệu từ UCI datasets.

Sử dụng dữ liệu với attribute bao gồm cả 2 loại: numerical và categorical

**Câu 1:**

Sử dụng các phương pháp học máy đã học (có thể sử dụng thêm các phương pháp học máy khác) để giải bài toán này theo cả phương pháp classification và regression. Thực hiện các bước sau:

* Thống kê đặc điểm của dữ liệu bằng data visualization
* Tiền xử lý dữ liệu: type conversation và data normalization
* Phân chia dữ liệu thành tâp train và tập evaluation
* Thực hiện phân loại (và hồi quy)
* Đánh giá kết quả và so sánh các phương pháp: so sánh các phương pháp phân loại với nhau; so sánh các phương pháp hồi quy với nhau. Sử dụng biểu đồ để hiển thị sự so sánh cho trực quan.

**Câu 2:**

* Tự tìm hiểu và trình bày về vấn đề overfitting nói chung, và giải pháp giải quyết overfitting cho các phương pháp học ở Câu 1 (lưu ý: có thể chọn 1 hoặc nhiều phương pháp để giải quyết overfitting);
* Thực hiện code bổ sung giải pháp overfitting vào các phương pháp ở Câu 1; so sánh kết quả với việc không sử dụng overfitting.

**Câu 3:**

* Tìm hiểu và trình bày Feature selection using correlation analysis. Áp dụng vào bài toán trên đối với các phương pháp ở Câu 1.

**Hướng dẫn làm bài:**

* 2 người 1 nhóm (trường hợp đặc biệt là 3 người)
* Tất cả các thành viên trong nhóm đều cùng làm tất cả các bài. Không làm bài theo kiểu chia mỗi người một bài, và người này không biết bài người kia làm thế nào.
* Nộp bản pdf mô tả bài toán, giải pháp
* Nộp code và data chạy
* Trình bày tóm tắt bằng slide
* Tất cả kết quả lưu vào 1 thư mục, nén lại và submit file nén. Tên thư mục là gộp FirstName\_StudentID của các thành viên
* Tất cả các thành viên đều phải nộp