BỘ GIÁO DỤC VÀ ĐÀO TẠO



TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

KHOA CÔNG NGHỆ THÔNG TIN

NHẬP MÔN HỌC MÁY

**BÁO CÁO BÀI TẬP MÔN HỌC**

**NỘI DUNG: PRACTICE COMPETITION**

**PHASE 01- SCIKIT LEARN**

**Giảng viên hướng dẫn** :Nguyễn Thanh Tình

**Lớp** :CQ2022/24

**Sinh viên thực hiện** :Đinh Viết Lợi-22120188

Nguyễn Trần Lợi- 22120190

Nguyễn Nhật Long-22120194

*Hồ Chí Minh, ngày 23 tháng 5 năm 2025*

# MỤC LỤC

[MỤC LỤC 1](#_Toc198762855)

[PHẦN 1: TỔNG QUAN ĐỒ ÁN 2](#_Toc198762856)

[I. Tổng quan đồ án 2](#_Toc198762857)

[1. Giới thiệu đồ án 2](#_Toc198762858)

[2. Yêu cầu đồ án 2](#_Toc198762859)

[II. Kết quả đạt được 3](#_Toc198762860)

[PHẦN 2: BÁO CÁO NHÓM 3](#_Toc198762861)

[I. Thông tin nhóm 3](#_Toc198762862)

[II. Quy trình thực hiện 3](#_Toc198762863)

[PHẦN 3: MODEL DEVELOPMENT 4](#_Toc198762864)

[I. Phân tích bộ dữ liệu 4](#_Toc198762865)

[II. Chi tiết quá trình xây dựng mô hình 4](#_Toc198762866)

[III. Các phương pháp khác 4](#_Toc198762867)

[PHẦN 4: MODEL EVALUATION 4](#_Toc198762868)

[I. Các bộ chỉ số đánh giá 4](#_Toc198762869)

[II. Đánh giá và cải thiện bộ chỉ số 4](#_Toc198762870)

[III. Một số mô hình khác 4](#_Toc198762871)

[PHẦN 5: ĐÁNH GIÁ PHƯƠNG PHÁP 4](#_Toc198762872)

[I. Phương pháp tiền xử lý dữ liệu 4](#_Toc198762873)

[II. Phương pháp xây dựng mô hình 4](#_Toc198762874)

[PHẦN 6: TÀI LIỆU THAM KHẢO 4](#_Toc198762875)

[I. Tài liệu nhóm 4](#_Toc198762876)

[II. Tài liệu tham khảo 4](#_Toc198762877)

# PHẦN 1: TỔNG QUAN ĐỒ ÁN

## Tổng quan đồ án

### **Giới thiệu đồ án**

* + Đồ án Practice Competition là đồ án kết thúc môn học của bộ môn Nhập môn Học máy của lớp CQ2022/24 Trường Đại học Khoa học Tự nhiên học kỳ II năm học 2024-2025.
  + Đồ án tập trung vào thực hành quy trình xây dựng các mô hình học máy phục với 3 giai đoạn chính: Exploratory Data, Model Development, and Model Evaluation. Trọng tâm quy trình sẽ nằm ở hai giai đoạn sau.
  + Hình thức thực hiện đồ án được tổ chức dưới dạng một cuộc thi trên Kaggle với tư cách cá nhân hoặc theo nhóm. Các thông tin về cuộc thi được trình bày tại Kaggle competition.
  + Bài toán được đặt ra bởi cuộc thi sẽ được yêu cầu giải quyết trong vòng 3 tuần với 3 giai đoạn (phase) khác nhau. Tuần đầu của cuộc thi được yêu cầu sử dụng Scikit-learn.

### **Yêu cầu đồ án**

* + Các sinh viên được mong đợi tuân thủ một quy trình phát triển cho dự án Học máy thực tế, được cấu trúc với ba bước chính: Phân tích Dữ liệu Khám phá (Exploratory Data Analysis), Phát triển Mô hình (Model Development) và Đánh giá Mô hình (Model Evaluation).
  + Bài tập được thực hiện dưới hình thức một cuộc thi Kaggle, sinh viên cần tham gia theo thể thức cá nhân hoặc theo đội nhóm.
  + Dự án được lên kế hoạch diễn ra trong tổng cộng 3 tuần, được cấu trúc thành 3 giai đoạn riêng biệt. Mỗi giai đoạn kéo dài một tuần và bao gồm toàn bộ chu trình của hai bước chính: Phát triển Mô hình, và Đánh giá và Phân tích Lỗi.
  + Mỗi sinh viên/ nhóm được yêu cầu nộp quá trình thực hiện hàng tuần bao gồm mã nguồn Jupyter notebook và báo cáo cho từng Phase.
  + Mã nguồn Jupyter Notebook cần có: thông tin sinh viên/ nhóm sinh viên, giải thích chi tiết từng bước bằng hình ảnh minh hoạ, biểu đồ, công thức, chú thích (comment) từng bước xử lý và trực quan, notebook được format tốt.
  + Báo cáo bao gồm: thông tin sinh viên/ nhóm sinh viên, tự đánh giá kết quả, ưu và nhược điểm của các phương pháp xử lý, mô hình và frameworks khác nhau. Thông tin chi tiết mỗi giai đoạn cần đảm bảo:
    - **Model development**: Trình bày việc lựa chọn mô hình, thiết kế kiến trúc và thiết lập huấn luyện cùng với lý do đằng sau các lựa chọn của bạn.
    - **Đánh giá Mô hình:** Báo cáo về hiệu suất mô hình bằng cách sử dụng các chỉ số phù hợp, đồng thời bao gồm các quan sát về thời gian chạy, mức sử dụng tài nguyên tính toán. Cung cấp phân tích chi tiết về bất kỳ loại lỗi nào mà sinh viên và/hoặc nhóm đã xác định là điểm yếu của mô hình.

## Kết quả đạt được

* + Quá trình xây dựng mô hình của nhóm được áp dụng mạnh mẽ và đa dạng các thư viện, framwork được hỗ trợ sẵn.
  + Nhóm áp dụng nhiều các phương pháp xử lý khác nhau nhằm chuẩn bị được bộ dữ liệu tốt nhất, phân tích các đặc điểm của dữ liệu hỗ trợ, gây khó khăn cho quá trình huấn luyện.
  + Áp dụng rộng rãi các các mô hình học máy khác nhau, đánh giá ưu nhược điểm từng mô hình, thiết lập bộ trọng số tốt nhất cho mô hình thông qua các phương pháp Grid Search, Random Search…
  + Hầu hết mô hình được nhóm lựa chọn sau cùng đều cho kết quả ấn tượng, áp dụng tốt cho bộ dữ liệu test.
  + Thông tin cụ thể về quá trình làm cũng như động lực được trình bày rõ ràng trong notebook và báo cáo.

# PHẦN 2: BÁO CÁO NHÓM

## Thông tin nhóm

|  |  |
| --- | --- |
| **Thành viên** | **Vai trò** |
| Đinh Viết Lợi | Phương pháp SMOTE, tách frame, viết báo cáo tổng hợp. |
| Nguyễn Trần Lợi | Phương pháp mean, feature engineering. |
| Nguyễn Nhật Long | Lựa chọn mô hình, tìm kiếm bộ siêu tham số phù hợp, lựa chọn tham số đánh giá. |

## Quy trình thực hiện

* + Tận dụng số lượng submit có hạn hằng ngày trong thời gian ngắn, nhóm liên tục cập nhật source code hằng ngày để kiểm tra mức độ cải thiện qua từng phiên bản:
    - Ngày 1**:** phương pháp mean lấy trung bình các vector audio\_embedding, đánh dấu trực tiếp nhãn, mô hình Random Forest.
    - Ngày 2: Sử dụng Grid Search tìm kiếm siêu tham số tối ưu cho mô hình Random Forest.
    - Ngày 3: Thử nghiệm nhiều mô hình hơn cho bài toán, phát hiện SVM có hiệu suất tốt nhất trong các mô hình, tiến hành tìm bộ tham số tối ưu cho mô hình SVM.
    - Ngày 4:
    - Ngày 5:
    - Ngày 6:

# PHẦN 3: MODEL DEVELOPMENT

## Phân tích bộ dữ liệu

## Chi tiết quá trình xây dựng mô hình

## Các phương pháp khác

# PHẦN 4: MODEL EVALUATION

## Các bộ chỉ số đánh giá

## Đánh giá và cải thiện bộ chỉ số

## Một số mô hình khác

# PHẦN 5: ĐÁNH GIÁ PHƯƠNG PHÁP

## Phương pháp tiền xử lý dữ liệu

## Phương pháp xây dựng mô hình

# PHẦN 6: TÀI LIỆU THAM KHẢO

## Tài liệu nhóm

## Tài liệu tham khảo