ĐẠI HỌC KHOA HỌC TỰ NHIÊN
TRUNG TÂM TIN HỌC

Địa chỉ: 227 Nguyễn Văn Cừ, Q.5, TP.HCM
Tel: (028) 38351056
Email: tuvan@csc.hcmus.edu.vn
Website: https://csc.edu.vn/

Machine Learning with Python

(MÁY HỌC VỚI PYTHON)

Thời gian: 5 tuần

Thời lượng: 48 giờ (64 tiết)

Học phí: 6.000.000 đ

1. MỤC TIÊU

- Khóa học cung cấp cho học viên các kiến thức quan trọng và cần thiết về Machine Learning, một phần rất "qua trọng" của Khoa học dữ liệu (Data Science).
- Trang bị các kiến thức cơ bản về Toán được ứng dụng trong các thuật toán của Machine Learning.
- Trang bị kiến thức và kỹ năng vận dụng các thuật toán thuộc nhóm Supervised Learning (Classification, Regression), Unsupervised Learning (Clustering, Dimensionality Reduction) thông qua việc sử dụng các bộ thư viện, công cụ mạnh mẽ, mã nguồn mở như Python, Jupyter Notebooks, Numpy, Pandas, Matplotlib, Seaborn, sklearn…
- Xây dựng nền tảng vững chắc về Machine Learning với Python, tạo tiền đề cho việc tìm hiểu kiến thức về Deep Learning.
- Là khóa học thứ năm trong chương trình "Data Science and Machine Learning Certificate"

2. ĐỐI TƯỢNG HỌC

- HV đã tham gia khóa Data Preprocessing & Data Analysis hoặc có kiến thức tương đương
- Sinh viên các trường Đại học, Cao đẳng
- HV có định hướng sẽ làm việc trong lĩnh vực Machine Learning/ Data Science

3. KẾT QUẢ ĐẠT ĐƯỢC

Sau khi hoàn thành khóa học, học viên có thể đạt được các kiến thức và kỹ năng:

- Hiểu các khái niệm cơ bản của Toán được vận dụng trong Machine Learning, bao gồm Đại số ma trận; Tối ưu hàm số với Gradient Descent; Xác suất và Thống kê.
- Hiểu, áp dụng và triển khai các thuật toán trong nhóm Supervised Learning như Logistic Regression, Linear Regression, Naïve Bayes, K-Nearest Neighbors (KNN), Decision Tree, Random Forest, Support Vector Machine (SVM), Mạng Nơron; Boosting và AdaBoost với Python.
- Hiểu, áp dụng và triển khai các thuật toán trong nhóm Unsupervised Learning như K-Means clustering, Hierarchical Clustering, Gaussian Mixture Models (GMM)
- Dimensionality Reduction với Principal Component Analysis (PCA) và Linear Discriminant Analysis (LDA).
- Time Series Analysis với ARIMA

- Vận dụng các thuật toán Machine Learning trong việc giải quyết các vấn đề thực tế, cụ thể
- Triển khai project theo Data Science process
- Xây dựng nền tảng vững chắc về Machine Learning với Python

4. NỘI DUNG KHOÁ HỌC

- **Giới thiệu Machine Learning**
  - Giới thiệu
  - Tổng quan về Machine Learning
    - Định nghĩa
    - Lịch sử phát triển
    - Tầm quan trọng của ML trong thế giới hiện nay
  - Phân loại
    - Học có giám sát
    - Học không giám sát
    - Học tăng cường
  - Quy trình phát triển ML project (ôn lại)
    - Các bước thực hiện
    - Đánh giá mô hình
  - Hỏi đáp
- **KNN**
  - Vector - Ma trận
  - Các thao tác cơ bản với vector – ma trận
  - Thuật toán kNN
    - Khoảng các giữa 2 vector
    - Cài đặt thuật KNN **(ChatGPT hỗ trợ)**
    - Sử dụng thư viện sklearn
- **Tối ưu hàm số với Gradient Descent**
  - Giới thiệu đạo hàm
  - Một số đạo hàm cơ bản
  - Đạo hàm đa biến
  - Tim cực trị hàm số với phương pháp Gradient Descent
- **Mô hình hồi quy tuyến tính (Linear Regression)**
  - Hàm mất mát (Loss function / Cost function) MSE
  - Tối ưu hàm mất mát MSE
    - Ứng dụng phân rã SVD tìm ma trận giả đảo
    - Ứng dụng thuật toán Gradient Descent tối thiểu hóa MSE
  - Sử dụng thư viện sklearn
  - Vấn đề đa công tuyến trong hồi quy đa biến.
  - Chọn lựa thuộc tính với SelectKBest
- **Mô hình hồi quy Logistic (Logistic Regression)**

- Giới thiệu
- Binary classification
- Hàm Sigmoid
- Cài đặt Logistic Regression với Gradient Descent
- Multi class classification với Logistic Regression
    - Khái niệm OVR và OVO
    - Sử dụng thư viện sklearn trong bài toán hồi quy logistic đa biến.
- Chọn lựa thuộc tính với SelectKBest

➢ **Naive Bayes**

- Lý thuyết xác suất
- Công thức Bayes
- Phân phối dữ liệu và hàm phân phối xác suất
    - Phân phối dữ liệu (Distribution)
    - Hàm phân phối xác suất và hàm mật độ xác suất
    - Phân phối chuẩn (Standard distribution)
    - Phân phối nhị thức (Binomial distribution)
- Ứng dụng công thức Bayes trong bài toán phân lớp (classification)
- Minh họa thuật toán với Excel/Python
- Thực hành cài đặt mô hình với thư viện sklearn

➢ **Đánh giá và lựa chọn mô hình (Cross validation)**

➢ **Cây quyết định (Tree Decision) - Rừng (Random Forest)**

- Giới thiệu về cây quyết định
- Hoạt đông tổ chức cây
    - Entropy, Information gain
    - Inpurity Gini
- Ưu & Khuyết điểm
- Cài đặt ứng dụng với thư viện sklearn
- Thuật toán Random Forest
    - Giới thiệu
    - Ensemble models
    - Ưu & khuyết điểm
    - Thực hành cài đặt mô hình với thư viện sklearn
- Chọn lựa thuộc tính với FeaturesImportance

➢ **Mạng nơron (Neuron network)**

- Tổng quan về mạng nơron
- Các thành phần của một mạng noron đơn giản
    - Layers
    - Activation functions
    - Feed forward process

        - Back propagation
    - Biến thể của Gradient Descent
        - Stochastic
        - Mini-batch
    - Cài đặt một mạng noron đơn giản
    - Giới thiệu Deep Learning

- **SVM**
    - Cơ bản về SVM
        - Tối đa hóa lề phân tách
        - Khái niệm support vector
    - Kernel Trick và bài toán phân tách không tuyến tính (non-linear separate)
        - Giới thiệu các Kernel trick
        - Giới thiệu các siêu tham số (C, gamma) trong SVM
    - Thực hành cài đặt mô hình với sklearn
- **Các kỹ thuật khác**
    - Pipeline
    - Lưu trữ và tái sử dụng model
    - Tối ưu các giá trị tham số (Parameters tuning)
        - Grid search
        - Random search
- **Ensemble Techniques: Boosting**
    - Tổng quan về Boosting
    - Các thuật toán Boosting phổ biến
        - AdaBoost
        - GradientBoost
        - XGBoost
    - Thực hành cài đặt mô hình
- **Các thuật toán giảm chiều dữ liệu**
    - Tổng quan giảm chiều dữ liệu
    - Các kiến thức Toán liên quan
        - Eigen values và Eigen Vectors
        - Eigen decomposition
        - Phân tích thành phần chính (PCA)
    - Các ứng dụng của PCA
        - Trực quan hóa
        - Giảm chiều dữ liệu
    - Thuật toán LDA
        - Giới thiệu

        - Các ứng dụng
- **Phân cụm dữ liệu**
    - KMeans
        - Hoạt động của thuật toán
        - Chọn số nhóm (Elbow method, Sihoute)
        - Ưu và khuyết điểm
        - Các biến thể của KMeans
        - Thực hành cài đặt mô hình.
    - Hierachical clustering
        - Hoạt động của thuật toán
        - Dendogram
        - Ưu và khuyết điểm
        - Thực hành cài đặt mô hình.
    - Gausian Mixture Model (GMM)
        - Phân phối chuẩn
        - Thuật toán GMM
        - Giới thiệu EM
        - Cài đặt ứng dụng GMM
- **Time series analysis: ARIMA**
    - Tổng quan
    - Các thành phần
        - Trend, seasonality, cyclic patterns, and noise.
    - Phân loại
        - Univariate vs. multivariate time series.
    - Stationarity.
        - Dickey-Fuller test.
    - Time Series Decomposition
    - Thuật toán ARIMA
- **Đồ án cuối môn**