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/

Mới

Data Pre-processing and Analysis

(Tiền xử lý dữ liệu và Phân tích dữ liệu)

Thời gian: 5 tuần

Thời lượng: 40 giờ (53 tiết)

Học phí: 5.500.000 đ

1. MỤC TIÊU

- Khóa học cung cấp cho học viên (HV) các kiến thức và kỹ năng cần thiết khi thực hiện việc tiền xử lý và phân tích dữ liệu
- Trang bị cho HV những kỹ thuật khai thác dữ liệu, chuyển đổi dữ liệu thô thành dữ liệu có định dạng dễ hiểu
- Hướng dẫn HV cách chuẩn bị dữ liệu để phân tích, thực hiện phân tích thống kê, tạo trực quan hóa dữ liệu có ý nghĩa
- Cung cấp cho HV các thư viện tiền xử lý và phân tích dữ liệu mạnh mẽ và ưu việt của Python như Numpy, Scipy, Pandas, Matplotlib, Seaborn, pandas profiling, dataprep…
- Hướng dẫn cách tiền xử lý dữ liệu tiếng Anh, tiếng Việt
- Cung cấp cho HV quy trình quản lý phân tích dữ liệu hiệu quả
- Hướng dẫn HV sử dụng thư viện mã nguồn mở sklearn để triển khai một số thuật toán Machine Learning giúp xây dựng các mô hình dự đoán thông minh
- Là khóa học thứ tư trong chương trình “Data Science and Machine Learning Certificate”

2. ĐỐI TƯỢNG HỌC

- Sinh viên các trường Đại học, Cao đẳng
- HV có định hướng sẽ làm việc trong lĩnh vực Data Science, Machine Learning
- Người làm việc trong lĩnh vực dữ liệu, phân tích dữ liệu và khoa học dữ liệu muốn nâng cao kỹ năng phân tích và tiền xử lý dữ liệu.

3. KẾT QUẢ ĐẠT ĐƯỢC

Sau khi hoàn thành khóa học, học viên sẽ đạt được các kỹ năng:

- Hiểu và vận dụng các bước trong quy trình tiền xử lý dữ liệu (Data Pre-processing) khi triển khai dự án Data Science
- Nắm được quy trình và kỹ thuật phân tích dữ liệu (Data Analysis)

- Phân tích dữ liệu khám phá (Exploratory Data Analysis - EDA) để có cái nhìn ban đầu về dữ liệu, xác định các yếu tố quan trọng trong bộ dữ liệu
- Thực hiện các thao tác làm sạch dữ liệu (Data Cleaning)
- Áp dụng linh hoạt các kỹ thuật chuẩn hóa dữ liệu (Data Standardization) khác nhau tùy vào các bộ dữ liệu và yêu cầu của dự án
- Tạo các tính năng cần thiết (Feature Engineering)
- Tiền xử lý dữ liệu tiếng Anh, tiếng Việt
- Phân tích, triển khai và đánh giá mô hình phân tích dữ liệu
- Giải thích kết quả từ các phân tích dữ liệu
- Quy trình quản lý phân tích dữ liệu hiệu quả
- Sử dụng thư viện mã nguồn mở sklearn để triển khai một số thuật toán Machine Learning
- Kết hợp trực quan hóa dữ liệu, kết quả thống kê để tạo các báo cáo phân tích dữ liệu mạch lạc, thuyết phục

4. NỘI DUNG KHOÁ HỌC

- Giới thiệu Data Analysis
    - Giới thiệu
    - Quy trình:
        - Define the Problem and Objectives
        - Data Collection
        - Data Cleaning
        - EDA
        - Feature Engineering
        - Model Selection and Building
        - Model Evaluation
        - Communication
        - Deployment
- Giới thiệu Data Pre-processing
    - Giới thiệu
    - Quy trình:
        - Data Collection
        - Data Integration
        - Data Cleaning
        - EDA
        - Feature Engineering
        - Handling Text Data
- Data Pre-processing
    - Giới thiệu Data Pre-processing
    - Giới thiệu các loại dữ liệu
    - Làm sạch dữ liệu (Data Cleaning):
        - Removing Duplicates
        - Handling Missing Values
        - Correcting Inconsistent Data
        - Standardizing Data Formats
        - [Outliers]

- Phân tích EDA (Exploratory Data Analysis)
    - Giới thiệu
    - Các khái niệm xác suất thống kê cơ bản
    - Các phương pháp phân tích:
        - Summary Statistics
        - Data Visualization
        - Hypothesis testing
    - Phân tích 1 biến:
        - Category
        - Continuous
    - Phân tích 2 biến:
        - Category and Category
        - Contiuous and Continuous
        - Category and Continous
    - Phát hiện và xử lý outliers:
        - Quantiles method
        - Z-score method
    - Một số công cụ phân tích:
        - Dataprep
        - ttth-mds5-analyzer
- Feature Engineering
    - Giới thiệu
    - Feature Scaling:
        - Log normalization
        - Standard Scaler
        - Min-max Scaler
        - Robust Scaler
        - Binarizer
    - Data Transformation:
        - Pivot and UnPivot
        - Category Encoder: Label encoder and One-Hot encoder
- NLP (Natural Language Processing) cơ bản
    - Giới thiệu
    - Text Data Pre-processing
    - Text Data Transformation:
        - Count Vectorizer
        - Tf-Idf Vectorizer
- Mô hình hồi quy tuyến tính (Linear Regression)
    - Giới thiệu
    - Simple Linear Regression
    - Multiple Linear Regression
    - Evaluation Model

- Mô hình hồi quy Logistic (Logistic Regression)
    - Giới thiệu
    - Logistic Regression
- Xử lý tập dữ liệu mất cân bằng (Imbalanced Dataset)
    - Giới thiệu
    - Strategies to handling Imbalanced Dataset:
        - Resampling: Oversampling and Undersampling
        - Synthetic Data Generation: SMOTE
    - Performance Metric Classification:
        - Accuracy
        - Confusion Matrix
        - ROC Curves (ROC – AUC)