ĐỀ CƯƠNG
Học AI cho Khoa học Dữ liệu:
Machine Learning

Thời gian: 12 tuần (2 buổi/tuần, mỗi tuần 1 chủ đề)
(Tập trung vào phân tích dự đoán và phân tích đề xuất)

Một nhà khoa học dữ liệu thì không thể không biết về Statistical Machine Learning (Học Máy Thống Kê)!!!*. Trong ngành DS, các kiến thức trong Xác suất Thống kê và Học Máy là những nền tảng cốt lõi để thực hiện tốt các phân tích chuyên sâu, đặc biệt là cho các dạng phân tích cao cấp như dự đoán/dự báo và đề xuất/đề nghị. Khóa học này giúp các học viên củng cố kiến thức nền để phát triển sự nghiệp trong ngành DS một cách vững chắc, đồng thời hiểu được về các phương pháp cốt lõi trong AI. Mục tiêu của khóa học là trang bị cho học viên nền tảng cần thiết để thật sự chuyển mình từ một DA thông thường sang một Data Scientist. Tuy nhiên, các phương pháp Thống kê và Học Máy không đơn giản và cần có thời gian để nắm bắt, luyện tập. Thông qua 1 buổi lý thuyết và 1 buổi thực hành hằng tuần,* ***học viên sẽ được luyện tập cách đưa ra các dự đoán, khuyến nghị về quyết định hoặc kế hoạch (thậm chí chiến lược) dựa trên dữ liệu (data-driven) cho doanh nghiệp thông qua các bài thực hành về phân cụm, phân loại và phân tích dự đoán/đề xuất trong kinh doanh, bán lẻ, tài chính, chứng khoán...*** *sử dụng các phương pháp Học Máy Thống Kê, bao gồm Top 05 giải thuật trong phân tích dữ liệu, theo Edureka: Decision Tree, Random Forest, Association Rule Mining, Linear Regression và K-Means Clustering. Sau khóa học này, học viên có khả năng phân loại các bài toán và đề xuất giải pháp cho các dự án DA/DS dựa trên các cơ sở vững chắc, chọn lựa các tiêu chí đánh giá phù hợp về tiềm năng, kết quả cần đạt, cũng như khả năng truyền tải các thông tin đó đến đồng nghiệp, cấp trên, khách hàng, đối tác,… một cách khoa học và dễ hiểu. Đồng thời, các kiến thức trong khóa học giúp tạo nền tảng vững chắc cho các học viên muốn đi sâu hơn vào AI và đặc biệt là Deep Learning.*

Với các kiến thức, kinh nghiệm từ khóa học này, học viên từ chưa có kinh nghiệm đi làm có thể apply vào các vị trí junior Data Scientist phụ trách phân tích dữ liệu ở các công ty. Với các học viên đã đi làm và có kiến thức/kinh nghiệm chuyên ngành hoàn toàn có thể trở thành senior/leader phụ trách phân tích dữ liệu, thậm chí là chuyên gia cấp cao về phân tích dữ liệu trong chuyên ngành cụ thể (marketing, ngân hàng, chứng khoán, bảo hiểm, bán lẻ, logistics, kinh doanh - thương mại điện tử…).

Yêu cầu đầu vào: *Lập trình Python (nắm vững Numpy và Pandas, trực quan hóa dữ liệu với Matplotlib, Pandas), toán THPT (nắm vững về hàm số và xác suất), tiếng Anh (đọc hiểu cơ bản). Với các học viên đã biết lập trình Python nhưng chưa nắm vững Pandas có thể học bổ sung vài buổi trước khóa học.*

	Nội Dung	Thực Hành & Thảo luận
Tuần 1 (Bảo)	Các loại phân tích thống kê (p1) - Ôn tập Probability/Statistics - Thống kê mô tả (Descriptive Statistics)	- Làm sạch dữ liệu và thực hiện phân tích mô tả - Correlation Coefficient - Central limit theorem & confidence interval
Tuần 2 (Kiên)	Các loại phân tích thống kê (p2) - Thống kê suy luận (Inferential Statistics)	- Estimation vs. Hypothesis Testing - Alternate Hypothesis - Error Type I và Type II - T-Test/Z-Test/F-Test/ ANOVA - A/B Testing
Tuần 3 (Bảo)	Tổng quan về Machine Learning (p1): hệ thống hồi quy (regression) và hệ thống phân lớp (classification)	- Supervised vs. Unsupervised - Linear Regression và Logistic Regression - Ridge, Lasso và ElasticNet Regression - Phân lớp với k-Nearest Neighbor (kNN)
Tuần 4 (Kiên)	Tổng quan về Machine Learning (p2): Đánh giá kết quả hệ thống dự đoán và hệ thống phân lớp	- Error metrics: MAE, MSE, RMSE, R-squared, Adjusted R-squared - Accuracy/Confusion Matrix - True/False Positive/Negative - Precision, Recall, F1-score - ROC và AUC - Bias vs. variance
Tuần 5 (Bảo)	Bài toán giảm chiều dữ liệu (dimensionality reduction)	- Principal Component Analysis - Latent Derichlet Analysis
Tuần 6 (Kiên)	Decision Tree và Random forest	- Áp dụng suy luận Bayes và Naive Bayes để dự đoán khách hàng rời bỏ ngân hàng - Sử dụng cây quyết định để phân tích đơn hàng, dự đoán khoản vay, hoặc khách hàng rời bỏ - BootStrapping
Tuần 7 (Bảo)	Bài toán phân cụm (clustering)	Bài toán phân cụm khách hàng với: - k-Means & mini-batch k-Means - Mean-Shift clustering - Density-Based Spatial Clustering of Applications with Noise (DBSCAN) - Ước đoán số cụm, hiệu suất gom cụm: hệ số Silhouette, phương pháp Elbow
Tuần 8 (Kiên)	Khai phá dữ liệu (data mining) với Association Rule Mining (khai phá luật kết hợp) và Affinity Analysis (phân tích tương đồng)	- Bài toán phân tích giỏ hàng - Thuật toán Apriori tìm tập phổ biến trong database giỏ hàng
Tuần 9 (Bảo)	Bài toán gợi ý (recommendation)	- Content-based Filtering - Collaborative Filtering - Áp dụng cho bài toán gợi ý mua hàng
Tuần 10	Bài toán time series forecasting	- ARIMA/SARIMA

(Kiên)		- Kalman-filter - Dự báo giá chứng khoán
Tuần 11 (Bảo)	Support Vector Machine (SVM)	- Ứng dụng SVM cho bài toán dự đoán và phân lớp - Thực hiện phân tích phản hồi khách hàng (sentiment analysis) với SVM
Tuần 12 (Kiên)	- Giới thiệu Mạng neural nhân tạo (ANN) - Project cuối khóa	- Ứng dụng ANN cho bài toán dự đoán giá nhà, chứng khoán, phân lớp sản phẩm - Ôn tập & Trình bày project cuối khóa