CENTRAL BANK OF EGYPT
Egyptian Banking Institute

Markov ACADEMY

Data Analysis and Data Science Track

106 Hours

Virtual

Center of **Excellence** and **Knowledge Hub**
www.ebi.gov.eg

Course Description:

The ultimate guide to career transformation—go from novice to professional! In this class, you will learn the different techniques of machine learning and how to apply the data science life cycle on data sets. You will also get the practical knowledge of implementing the most effective methods by yourself. Moreover, you will get both the theoretical underpinnings of learning along with the practical know-how needed to strongly apply these methodologies and techniques to new problems.

Target Audience:

This course is designed for professionals and learners who want to build a strong foundation in data science and machine learning. It is ideal for individuals in fields such as business, economics, finance, engineering, and healthcare who wish to leverage data-driven decision-making. No prior experience in data science or programming is required.

Course Objectives:

- Build essential statistical knowledge and how to apply it on different business questions.
- Will be able to analyse your data quickly and visualize your results using SPSS and Python.
- You will learn how to use Python to analyse and visualize your data using different libraries.
- You'll explore the four crucial steps for any data analysis project. Reading, describing, cleaning, and visualizing data.
- You will work with the most common and popular tools that data analysts use every day.
- You will be able to confidently extract knowledge and answers from data.
- Build your expertise in the most widely used AI & ML tools and technologies.
- Understand the complete Data Science life cycle and its application on different Data Sets
- Acquire the ability as a data scientist to independently solve business problems using machine learning.
- Learn how data scientists exactly work in completing an end-to-end project. Starting from the understanding of a business problem then, data visualization, data preparation and moving through the whole data pipeline till deployment and reporting the results in the most effective way.
- Apply the whole data science life cycle on different real world use cases (end to end)

Course Outline:

1- Data Analysis and Visualization (30hrs)

Module 1: introduction to basics of Statistics.

- Introduction.
- Types of Variables.
- Sampling Techniques.
- Sample size.

Module 2: Data preparation and Descriptive Statistics using SPSS-software

- Introduce SPSS.
- Data coding, and data entry.
- Importing data to SPSS from excel.

- Data manipulation in SPSS.
- Data checking and editing in SPSS Introduction to descriptive statistics using SPSS.
- Measures of central tendency for different types of data.
- Measures of Variability for different types of data.
- One-way Tabulation distinct types of variables.
- Two-way and 3-way tabulations for distinct types of variables.

Module 3: Data Analysis & Visualization using Python Introduction to Python.

- Descriptive Statistics
- Python Programming
- Data Types and Operators
- Data Structures
- Control Flow
- Functions
- Scripting
- Working with Data in Python
- Data Analysis
- Pandas and NumPy
- Data Analysis Process
- Data Mining
- Data Wrangling
- Assessing and Cleaning Data
- Exploratory Data Analysis
- Anomaly Detection

Data Visualization

- Introduction to Data Visualization Tools
- Basic and Specialized Visualization
- Matplotlib and Seaborn
- Different Data Charts
- Advanced Visualization Tools
- Heat Maps Plotting
- Word Cloud
- Folium Maps
- Intensity Maps
- Geospatial Maps in Python
- Choropleth Maps

Data Prediction Models

- Linear Regression
- Logistic Regression

Use Case-1 Explore US Bikeshare Data Use python to understand US bikeshare data. Calculate statistics and build an interactive environment where a user can choose the data and filter for a dataset to show.
Use Case-2 House Sales in King County, USA Analyze and predict housing prices using attributes or features such as square footage, number of bedrooms, number of floors and so on
Use Case-3 Audience Interest in Data Science Topics
Use python to generate visualization plots to summarize the results of a survey that was conducted to gauge an audience interest in different data science topics
Use Case-4 San Francisco Incidents Distribution Use python Folium maps to generate a Choropleth map of the crime rate in San Francisco, based on data for one year, and show distribution of different crimes' type ities

2- Data Science and Python (36hrs)
Module 1: Statistics, Linear Algebra & Probability for Data Science

- Descriptive Statistics
  - Measures of Central Tendency
  - Measures of Variability
  - Skewness and Outliers
- Probability
  - Introduction to Probability
  - Probability Laws
  - Bayesian Theorem
  - Probability Distribution
  - Gaussian Distribution
  - Sampling Distribution
  - Central Limit Theorem
- Normalization.
  - Z-score
  - Min-Max Method
  - Decimal Scaling Method
- Inferential Statistics
  - 1. T-Test and ANOVA
  - 2. Chi-Square Test
  - 3. Spearman Correlation Coefficient
  - 4. Pearson Correlation Coefficient
  - 5. Regression Analysis
- Linear Algebra Review
  - Review on Matrices
  - Operations on Matrices
  - Eigen Values and Eigen Vectors.
  - Dimensionality Reduction (Principal Component Analysis - PCA)

Module 2: Data Science and Machine Learning

- Introduction to Artificial Intelligence
- Introduction to Data Science
- Data Science life cycle.
- Introduction to Machine Learning & Data Mining
- Machine Learning
- Data Mining
- Supervised and Unsupervised Learning
- Types of Data
- Data Preprocessing
- Frequent Item Sets
- Association Rules & Apriori Algorithm
- Regression for Data Science
  - Linear Regression
  - Polynomial Regression
- Bias and Variance
- Base Classifiers for Data Science
  - Logistic Regression
  - Decision Tree based Methods
  - K-Nearest Neighbor
  - Neural Networks
  - Naïve Bayes
  - Support Vector Machines
- Clustering
  - K-mean Clustering
  - Hierarchical Clustering
  - Cluster Evaluation
- Evaluation of Learning Models for a Data Scientist
  - F1-Score
  - ROC
  - Lift Curves

Module 3: Python for Data Science & Machine Learning

- Python Basics
  - General Syntax
  - Data Types
- Python Data Structures
  - Lists
  - Tuples
  - Sets
  - Dictionaries

- Python Programming Fundamentals
  - Functions
  - Methods
  - Loops
  - Conditional Statements
  - Classes and Objects
- Data Science Libraries
  - Numpy
  - Pandas
  - Matplotlib
  - Seaborn
  - Sklearn

Module 4: Class Projects

- Project 1: Market Basket Analysis (Apriori)
  - Importing Packages
  - Loading the Data
  - Data Preprocessing
  - Build Frequent Item Set
  - Crating Association Rules
- Project 2: Automotive Price Prediction (Linear Regression Algorithim)
  - Importing packages
  - Loading the data
  - Date Preprocessing
  - Creating Linear Model
  - Evaluating the Model
- Project 3: Fraud Detection (Logistic Regression Algorithm)
  - Importing Packages
  - Loading the Data
  - Data Exploration
  - Date Preprocessing
  - Split the Data (train & Test)
  - Train Logistic Regression Algorithm
  - Test the Trained Model
  - Evaluating the Model
- Project 4: Customer Churn Prediction (Neural Networks Algorithm)
  - Importing Packages
  - Loading the Data
  - Date Preprocessing
  - Split the Data (Train & Test)
  - Train NN Algorithm.
  - Test the trained Model
  - Evaluating the Model

- Project 5: Customer Segmentation (K-Means Clustering)
  - Importing Packages
  - Loading the Data.
  - Date Preprocessing
  - Choose the Optimum Number of Clusters
  - Apply K-Means
  - Visualize the Output

3- Applied Data Science (40hrs)

Module 1

- Data Science and AI Methodology Pipeline
- Understanding Business Problem
- Analytic Approach
- Data Collection and Preparation
- Modelling
- Deployment
- Submitted Project Document
- Cloud-Based Platforms (Google Colab - Amazon SageMaker - IBM Watson - MS Azure ML)

Module 2

Geographical Visualization of different clusters empowers data analysis and gives wonderful and unpredicted insights to geospatial data which enables business decision makers to take the rightest decisions with minimum risk.

Project 1: Clustering San Francisco Police Department Incidents.

- Collect incidents data for one year
- Segment incidents based on type, address and location
- Python Folium maps
- Geographical visualization of crime distribution
- Intensity maps of crime rate

Project 2: Analysing International Immigration Flows to Canada.

- Collect world immigration data for 20 years
- Python Choropleth maps
- Geographical features of non-spatial attributes
- GeoJSON files
- Geographic variation of immigrants' density

Module 3

Un-supervised learning is a key type of machine learning algorithms used to draw inferences from unlabeled datasets where cluster analysis techniques can be used for exploratory data analysis to find hidden patterns or groupings.

Project 3: Franchise Decision to Invest in Prominent Locations

- Scrap neighbourhoods' data of different cities
- Python Geopy client
- Locate neighbourhoods' coordinates
- Python FourSquare API
- Explore particular venues of industries
- K-Means un-supervised algorithm
- Elbow curve evaluation

Module 4:

Ensemble techniques are always present with supervised classification algorithms in machine learning to boost prediction accuracy through combining several base models in order to produce one optimal predictive model.

Project 4: Predicting Individuals' Income for Non-profit Organizations

- Exploratory data analysis
- Data pre-processing
- Training and prediction Pipeline
- Logistic Regression, KNN, SVM, Decision Tree, Random Forest
- Algorithms hyper parameters
- Grid Search
- Model evaluation
- Ensemble Learning
- Optimize best candidate algorithm

Module 5:

Deep learning algorithms have always been the best choice to handle challenging cases of analysing visual imagery through applying deep neural network and transfer learning to use the power of pre-trained model and apply it to a related problem.

Project 5: Images Classification using Deep Neural Network

- Images' features extraction
- Transfer learning
- Python PyTorch
- Torch Vision models (VGG and DenseNet)
- Images pre-processing
- Neural Network model classification layer
- Build classifier with different model architectures
- Python Argument Parser
- Control user arguments to the model
- Different computing platforms (CPU - GPU)

Module 6:

Computer Vision is one of the most important widely used applications of computer science and AI where computers are trained to interpret and understand the visual world. Using digital images from cameras and videos and deep learning models, machines can accurately identify and classify objects, and then react to what they see.

Project 6: Text Localization, Detection and Recognition

- Text detection and localization
- Optical character recognition
- Compute bounding boxes
- Tesseract OCR
- Python OpenCV
- Python Pytesseract
- Confidence of detected text
- Command line arguments

Module 7:

Recommendation systems are a collection of algorithms used to recommend items to users based on information taken from the user. They have become ubiquitous, and can be commonly seen in online stores, movies databases and job finders.

Project 7: E-Commerce Company Recommender System

- Pre-processing users' data
- Pre-processing items' data
- Content-Based filtering
- Collaborative filtering
- Pearson correlation function
- Similarity score
- Predict best match item

4- Final project

Prerequisites:

None

HOTLINE
15200
One number to better serve you!

Headquarters – Nasr City
22 A, Dr. Anwar El Mofty St., Tiba 2000
P.O.Box 8164 Nasr City, Cairo, Egypt
Tel.: (+2) 02 24054472
Fax: (+2) 02 24054471

Working hours: 9:00 am - 5:00 pm
www.ebi.gov.eg

Like us on
facebook.com/EgyptianBankingInstitute

Follow us on
twitter.com/EBItweets

Join us on
linkedin.com/company/egyptian-banking-institute

Watch us on
YouTube Channel: Egyptian Banking Institute (EBI)