

SYLLABUS

Certificate Program in Artificial Intelligence and Machine Learning A Practitioner's Approach (CAIML)



Jointly Organized by



E&ICT Academy

National Institute
of
Technology, Warangal



Special Interest **G**roup
BIG DATA
ANALYTICS

ABOUT THE PROGRAM



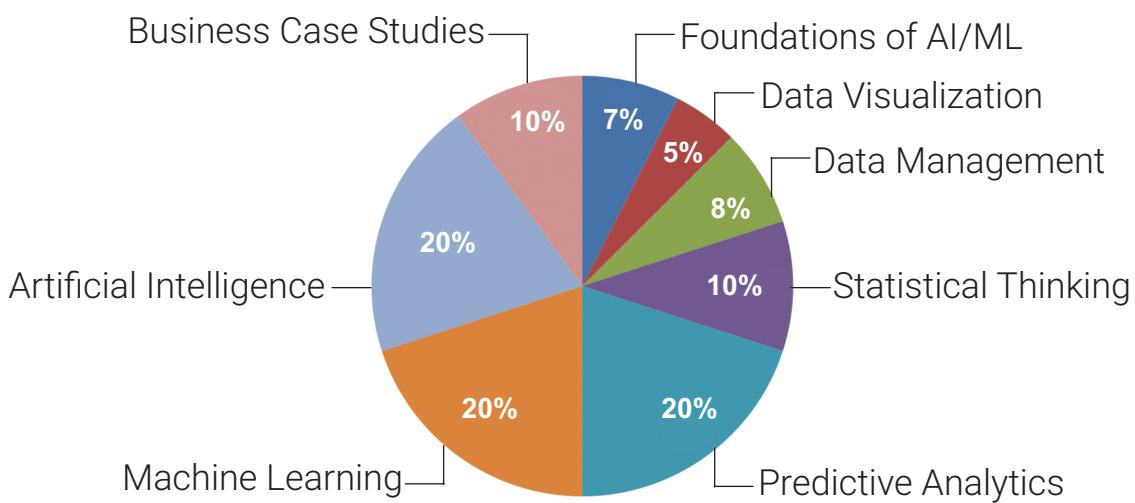
PRACTITIONER'S APPROACH TO ARTIFICIAL INTELLIGENCE & MACHINE LEARNING

CAIML is an intensive application oriented, real-world scenario based program in AI & ML. CAIML is a 6 Months (Weekends), intensive skill oriented, practical training program required for building business models for analytics. It is designed to give the participant enough exposure to the variety of applications that can be built using techniques covered under this program. This course is designed for the experienced professionals from variety of IT backgrounds. No prior knowledge of statistics or modeling is assumed.

OBJECTIVES

- Acquire advanced Data Analysis skills.
- Stay Industry relevant and grow in your career.
- Create AI/ML solutions for various business problems.
- Build and deploy production grade AI/ML applications.
- Apply AI/ML methods, techniques and tools immediate

COVERAGE and DURATION



MODULE 1: FOUNDATIONS OF AI AND ML



AI a multidisciplinary field that requires a range of skills in statistics, mathematics, predictive modeling and business analysis. An AI professional should feel at ease to build the algorithms necessary, work with various data sources (often in disparate forms) and an innate ability to ask the right questions and find the right answer. This module helps layout the canvas on which the rest of the modules are built.

Unit 1: Introduction to Data Science and AI & ML

- Data Science, AI & ML
- Use Cases in Business and Scope
- Scientific Method
- Modeling Concepts
- CRISP-DM Method



Unit 2: R Essentials (Tutorial)

Programming

- Commands and Syntax
- Packages and Libraries
- Introduction to Data Types
- Data Structures in R - Vectors, Matrices, Arrays, Lists, Factors, Data Frames
- Importing and Exporting Data.
- Control structures and Functions

Descriptive Statistics

- Data exploration (histograms, bar chart, box plot, line graph, scatter plot)
- Qualitative and Quantitative Data
- Measure of Central Tendency (Mean, Median and Mode),
- Measure of Positions (Quartiles, Deciles, Percentiles and Quantiles),
- Measure of Dispersion (Range, Median, Absolute deviation about median, Variance and Standard deviation), Anscombe's quartet
- Other Measures: Quartile and Percentile, Interquartile Range

Unit 3: Statistical Analysis

Initial Data Analysis

- Relationship between attributes: Covariance, Correlation Coefficient, Chi Square
- Measure of Distribution (Skewness and Kurtosis), Box and Whisker Plot (Box Plot and its parts, Using Box Plots to compare distribution) and other statistical graphs

Probability

- Probability (Joint, marginal and conditional probabilities)
- Probability distributions (Continuous and Discrete)
- Density Functions and Cumulative functions

MODULE 2: DATA MANAGEMENT



This is foundational to Data Scientists. This requires a nontrivial understanding of the real-world problems. It involves judgments such as those about the relevance and representativeness of the data. This module helps participants to have a good understanding of the methods, methodologies and techniques from the basics of statistics and probability obtain supporting evidence through data, isolate or identify factors to construct models that can uncover relationships and variation in processes.

Unit 3: Data Acquisition

- Gather information from different sources.
- Internal systems and External systems.
- Web APIs, Open Data Sources, Data APIs, Web Scrapping
- Relational Database access (queries) to process/access data.

Unit 4: Data Pre-processing and Preparation

- Data Munging, Wrangling
- Plyr packages
- Cast/Melt

Unit 5: Data Quality and Transformation

- Data imputation
- Data Transformation (minmax, log transform, z-score transform etc.,).
- Binning, Classing and Standardization.
- Outlier/Noise& Anomalies

Unit 6: Handling Text Data

- Bag-of-words
- Regular Expressions
- Sentence Splitting and Tokenization
- Punctuations and Stop words, Incorrect spellings
- Properties of words and Word cloud
- Lemmatization and Term-Document TxD computation
- Sentiment Analysis (Case Study)

Unit 7: Principles of Big Data

- Introduction to Big Data
- Challenges of processing Big Data (Volume, Velocity and Variety perspective)
- Use Cases

Unit 8: Big Data Frameworks – Hadoop, Spark and NoSQL

- Processing, Storage and Programming Framework
- Hadoop eco-system Components and their functions
- Essential Algorithms (Word count, Page Rank, IT-IDF)
- Spark: RDDs, Streaming and Spark ML
- NoSQL concepts (CAP, ACID, NoSQL types)



MODULE 3: STATISTICAL DECISION MAKING



This is foundational to Data Scientists. This requires a nontrivial understanding of the real-world problems. It involves judgments such as those about the relevance and representativeness of the data. This module helps participants to have a good understanding of the methods, methodologies and techniques from the basics of statistics and probability obtain supporting evidence through data, isolate or identify factors to construct models that can uncover relationships and variation in processes.

Unit 9: Data Visualization

Visualizing and Communicating clearly and effectively about the patterns we find in data is a key skill for a successful data professional. This module focuses on the design and implementation of complementary visual and verbal representations of patterns and analyses in order to convey findings, answer questions, drive decisions, and provide persuasive evidence supported by data.

- Science of Visualization
- Visualization Periodic Table
- Aesthetics and Story telling
- Concepts of measurement - scales of measurement
- Design of data collection formats with illustration
- Principles of data visualization - different methods of presenting data in business analytics.
- Concepts of Size, Shape, Color
- Various Visualization types
- Bubble charts
- Geo-maps (Chloropeths)
- Gauge charts
- Tree map
- Heat map
- Motion charts
- Force Directed Charts etc.,

Unit 10: Sampling and Estimation

- Sample versus population
- Sample techniques (simple, stratified, clustered, random)
- Sampling Distributions
- Parameter Estimation
- Unbalanced data treatment

Unit 11: Inferential Statistics

- Develop an intuition how to understand the data, attributes, distributions
- Procedure for statistical testing, etc.
- Test of Hypothesis (Concept of Hypothesis testing, Null Hypothesis and Alternative Hypothesis)
- Cross Tabulations (Contingency table and their use, Chi-Square test, Fisher's exact test),
- One Sample t test (Concept, Assumptions, Hypothesis, Verification of assumptions, Performing the test and interpretation of results)
- Independent Samples t test
- Paired Samples t test
- One way ANOVA (Post hoc tests: Fisher's LSD, Tukey's HSD).
- z-test and F-test



MODULE 4: PREDICTIVE ANALYTICS



Predictive analytics is an area of statistics that deals with extracting information from data and using it to predict trends and behavior patterns. Predicting an outcome, predicting counts, predicting a value - all these have immeasurable use cases in CRM, Fraud detection, Portfolio Management, Sales and Marketing. Predictive Analytics is approached from Regression (glm) and Time Series models in this module.

Unit 12: Linear Regression

- Regression basics: Relationship between attributes using Covariance and Correlation
- Relationship between multiple variables: Regression (Linear, Multivariate) in prediction.
- Residual Analysis
- Identifying significant features, feature reduction using AIC, multi-collinearity
- Non-normality and Heteroscedasticity
- Hypothesis testing of Regression Model
- Confidence intervals of Slope
- R-square and goodness of fit
- Influential Observations – Leverage

Unit 13: Multiple Linear Regression

- Polynomial Regression
- Regularization methods
- Lasso, Ridge and Elastic nets
- Categorical Variables in Regression



Unit 14: Non-Linear Regression

- Logit function and interpretation
- Types of error measures (ROCR)
- Logistic Regression in classification

Unit 15: Forecasting models

- Trend analysis
- Cyclical and Seasonal analysis
- Smoothing; Moving averages; Box-Jenkins, Holt-winters, Auto-correlation; ARIMA
- Examples: Applications of Time Series in financial markets

MODULE 5: MACHINE LEARNING

Predictive analytics is an area of statistics that deals with extracting information from data and using it to predict trends and behavior patterns. Predicting an outcome, predicting counts, predicting a value - all these have imnumerable use cases in CRM, Fraud detection, Portfolio Management, Sales and Marketing. Predictive Analytics is approached from Regression (glm) and Time Series models in this module.

Unit 16: Foundations for ML

- ML Techniques overview
- Validation Techniques (Cross-Validations)
- Feature Reduction/Dimensionality reduction
- Principal components analysis (Eigen values, Eigen vectors, Orthogonality)

Unit 17: Clustering

- Distance measures
- Different clustering methods (Distance, Density, Hierarchical)
- Iterative distance-based clustering;
- Dealing with continuous, categorical values in K-Means
- Constructing a hierarchical cluster
- K-Medoids, k-Mode and density-based clustering
- Measures of quality of clustering

Unit 18: Classification

Naïve Bayes Classifier

- Model Assumptions, Probability estimation
- Required data processing
- M-estimates, Feature selection: Mutual information
- Classifier

K-Nearest Neighbors

- Computational geometry; Voronoi Diagrams; Delaunay Triangulations
- K-Nearest Neighbor algorithm; Wilson editing and triangulations
- Aspects to consider while designing K-Nearest Neighbor

Support Vector Machines

- Linear learning machines and Kernel space, Making Kernels and working in feature space
- SVM for classification and regression problems.

Decision Trees

- ID4, C4.5, CART

Ensembles methods

- Bagging & boosting and its impact on bias and variance
- C5.0 boosting
- Random forest
- Gradient Boosting Machines and XGBoost

Unit 19: Association Rule mining

- The applications of Association Rule Mining: Market Basket, Recommendation Engines, etc.
- A mathematical model for association analysis; Large item sets; Association Rules
- Apriori: Constructs large item sets with mini sup by iterations; Interestingness of discovered association rules;
- Application examples; Association analysis vs. classification
- FP-trees

MODULE 6: ARTIFICIAL INTELLIGENCE



Artificial Intelligence is utilized heavily in computizing cognitive functions such as speech and Vision. Often these functions are achieved through the use of Neural networks. In this module, we will study very popular NN architectures for achieving various cognitive functions such as Object recognition, natural language processing besides explore reinforcement learning. We will study and practice various use cases ranging from text generation, object tagging to fraud detection and learning games such as 2048.

Unit 20: Foundations for AI

- AI: Application areas
- AI Basics (Divide and Conquer, Greedy, Branch and Bound, Gradient Descent)
- NN basics (Perceptron and MLP, FFN, Backpropagation)

Unit 21: Convolution Neural Networks [2]

- Image classification
- Text classification
- Image classification and hyper-parameter tuning
- Emerging NN architectures

Unit 22: Recurrent Neural Networks [2]

- Building recurrent NN
- Long Short-Term Memory
- Time Series Forecasting

Unit 23: Deep Learning [6]

- Auto-encoders and unsupervised learning
- Stacked auto-encoders and semi-supervised learning
- Regularization - Dropout and Batch normalization

MODULE 7: BUSINESS USE CASES* (TENTATIVE)



Case Study 1: Churn Analysis and Prediction (Survival Modelling)

- Cox-proportional models
- Churn Prediction

Case Study 2: Credit card Fraud Analysis

- Imbalanced Data
- Neural Network

Case study 3: Sentiment Analysis or Topic Mining from New York Times

- Similarity measures (Cosine Similarity, Chi-Square, N Grams)
- Part-of-Speech Tagging
- Stemming and Chunking

Case Study 4: Sales Funnel Analysis

- A/B testing
- Campaign effectiveness, Web page layout effectiveness
- Scoring and Ranking

Case Study 5: Recommendation Systems and Collaborative filtering

- User based
- Item Based
- Singular value decomposition-based recommenders

Case Study 6: Customer Segmentation and Value

- Segmentation Strategies
- Lifetime Value

Case Study 7: Portfolio Risk Conformance

- Risk Profiling
- Portfolio Optimization

Case Study 8: Uber Alternative Routing

- Graph Construction
- Route Optimization

ORGANIZERS



E & ICT Academy

NIT-Warangal is a premier institute known for imparting technical education of high standards. Ministry of Electronics and IT, GoI has established E&ICT academy to providing specialized training in emerging disciplines like AI & IoT. Continuing education for working professionals is a priority for this academy.



Computer Society of India (CSI) is a premier professional society established in 1965. The purpose of the society is to advance the theory and practice of Computer Science and Information Technology



SIG-BDA is the special interest group of CSI focused on the advancement of Big Data Analytics. It promotes education of this discipline through evangelism and publication of journal - **Visleshana** is its flag-ship publication.

