International School of Engineering

upon recommendation of the faculty and
by authority of the Board of Directors,
hereby confers upon

Babji Manohar Erle

THE POSTGRADUATE CERTIFICATE IN BIG DATA ANALYTICS AND OPTIMIZATION

on successful completion of all the requirements of the
376-hour program conducted between
January 05 and June 30, 2019.

This program is certified for quality of content, assessment and pedagogy by the Language Technologies Institute (LTI) of Carnegie Mellon University (CMU). The program curriculum has been developed in collaboration with LTI.

INSOFE

Carnegie Mellon University
Language Technologies Institute

Dated this eleventh day of September, two thousand and nineteen.

Dr. Dakshinamurthy V Kolluru
President

Dr. Sridhar Pappu
Executive VP-Academics

Topics Covered

Essential Engineering Skills in Big Data Analytics Using R and Python
Basics of R and Python languages for data analytics
Data structures, Objects, Control structures, Data I/O
Regular expressions
Data manipulation using advanced commands
Basic visualization in R and Python
Data cleansing and pre-processing of structured data

Foundations of Probability and Statistics for Data Science
Motivation: Importance, scope and challenges in statistical study
Understanding data: Central tendencies and measures of variability
Probability and its relationship to statistics
Bayes theorem
Confusion Matrix and associated evaluation metrics
Probability distributions, Sampling distributions and Central limit theorem
Inferential statistics: Confidence intervals and Hypothesis testing
Statistical tests: z, t, chi-square, F, ANOVA

Statistics and Probability in Decision Modelling
Simple and Multiple Linear regression
Logistic regression
Bias-Variance tradeoff
Regularization methods: Ridge, LASSO and Elastic nets regression
Naive Bayes classifier
Principal components analysis
Time series forecasting

Methods and Algorithms in Machine Learning
Apriori algorithm
Decision trees
k-Nearest Neighbours and Collaborative filtering
Support vector machines
Ensemble methods: Stacking, Random Forest, GBM, XGBoost
Hierarchical, k-Means and k-Medoids Clustering
Planning, thinking and architecting machine learning solutions

Foundations of Text Mining and Search
Pre-processing unstructured text data
Vector space models
Natural language processing
Search: Matrix factorization methods and Singular value decomposition
Application of text classification and sentiment analysis

AI and Decision Sciences
Artificial neural networks
Deep neural nets
Word2Vec, Convolution neural nets
Recurrent neural nets and LSTMs
Planning and architecting AI solutions
Linear Programming: assignment, transportation, integer problems
Monte Carlo simulations and Evolutionary search methods

The Art and Science of Storytelling with Data Visualizations
Communicating with data: Issues and guiding principles
Primary ingredients of data visualization
Visual encodings
Types of charts and which charts to use when
Case: Transition from a simple chart to a powerful visualization
Tools: R-ggplot and Tableau

Applying ML to Big Data Using Hadoop and Spark Ecosystem
Evolution and developments in Big Data applications
Linux and SQL refresher
Distributed and parallel frameworks
HDFS; HDP2.x, NoSQL; GFS
Hadoop ecosystem: Pig, Hive, HBase, Sqoop, Mahout, Flume, Chukwa, Avro, Hue, Oozie, Zookeeper, Kafka
Hadoop Streaming with R and Python
Spark-SQL, Spark ML, Spark Streaming
Security tools: Sentry, Ranger, Kerberous, Knox
Hands-on implementation of various state-of-the-art tools using Hadoop ecosystem
Hadoop cluster exposure: Review the business case, plan and architect solution
Batch and real-time processing of data
Deal with structured and unstructured data
Apply machine learning methods using Spark ML to solve a business problem

Communication, Ethical and IP Challenges for Analytics Professionals (Video Module)
Importance of communication and effective ways of communication
Issues and Challenges: Mix of stakeholders, Explicability of results, Visualization
Guiding Principles: Clarity, Transparency, Integrity, Humility
Framework for Effective Presentations
Data protection, Intellectual property rights, Confidentiality, Contractual liability, Competition law, Licensing of Open Source software and Open Data
How to handle legal, ethical and IP issues at an organization and an individual level