M. Tech (Data Science)

(Two year Programme)

Centurion University of Technology & Management

2022

Data Science

Programme Objectives

- To construct the means for extracting business-focused insights from data. This requires an understanding of how value and information flows in a business, and the ability to use that understanding to identify business opportunities.
- To focus on extracting knowledge from data sets, which are typically large (see big data). The field encompasses analysis, preparing data for analysis, and presenting findings to inform high-level decisions in an organization.

Eligibility Criteria

Bachelor's degree in Engineering / Technology or equivalent degree in Computer Science & Engineering with minimum CGPA of 6.5 or 60% of marks or First Class in the qualifying degree.

Selection Process

The selection processes is through central counseling on the basis of merit in qualifying CUEE or PGAT or GATE score. GATE qualified candidates are eligible for scholarship through AICTE.

Award of degree

After successful completion of degree, student will be awarded with Master of Technology in Data Science Engineering by Centurion University of Technology and Management.

Course Structure

This is a 2-year full-time post graduate program which involves first year (Semester- I & II) of intense coursework and second year (Semester- III & IV) Internship and project at data science companies or Consultancy firm for hands on experience.

Total Credit: 92

Domain Focus: Data Science

Course Structure

1st year-1stsemester

Sl No	Code	Subject name	T	P	P	Credits
1	MTDS1101	Introduction to Data Science	2	2	0	4
2	MTDS1102	Mathematical Foundation for Data Science	2	2	0	4
3	MTDS1104	Data Mining and Data Warehousing	2	2	0	4
4	MTDS1105	Statistical Methods	2	2	0	4
		Practice				
5	MTDS1103	Machine Learning Using Python	0	4	0	4
		TOTAL				**20**

Syllabus

Semester -1

Introduction to Data Science

Subject Name	Code	T-P-P	(Credit)
Introduction to Data Science	MTDS1101	2-2-0	4

Course Objectives:

- To try and discover hidden patterns in raw data.
- Be able to list the steps involved in data science, from data acquisition to insight, and describe the role of each step.
- Distinguish different ways of collecting data, and their impact on the conclusions that can be drawn from the data.
- Manage, summarize, and visualize data using the Python programming language and Jupyter notebooks.
- Explain the basic concepts of statistical inference and implement simulation-based inference methods.
- Apply machine learning methods and assess the quality of predictions.

Module I (10 Hours):

This course will give an introduction to the basic data science techniques including programming in Python, SQL/SPARQL and Map-Reduce for small and big data manipulation and analytics.

Module II (10 Hours):

Data collection, data preparation, data querying, data analytics including pattern mining, classification, clustering, data visualization, and parallel computing platforms.

Module III (10 Hours):

Advanced data analytics including NLP, knowledge extraction, graph analytics, graph querying, knowledge bases and crowd sourcing.

Module IV (10 Hours):

Introduce key application areas of data science including business intelligence, social media, biomedical informatics, computational ecology and e-discovery.

Text Books:

1. Data Science from Scratch (DSS), Joel Grus, O’Reilly Media Inc., http://shop.oreilly.com/product/0636920033400.
2. Python for Data Analysis (PDA), Wes McKinney, O’Reilly Media Inc., http://proquest.safaribooksonline.com/9781449323592
3. Mining of massive datasets (MMD), A. Rajaraman and J.D. Ullman, Cambridge University Press, 2011. ISBN-10: 1107015359, ISBN-13: 978-1107015357, http://www.mmds.org/ (public online access).
4. Natural Language Processing with Python (NLTK Book): http://www.nltk.org/book/ (public online access).

Reference Books:

1. Learning scikit-learn: Machine Learning in Python (MLP), Guillermo Moncecchi and Raul Garreta
2. Doing Data Science (DDS), Cathy O'Neil and Rachel Schutt, O’Reilly Media Inc., http://proquest.safaribooksonline.com/9781449363871

Mathematical Foundation for Data Science

Subject Name	Code	T-P-P	(Credit)
Mathematical Foundation for Data Science	MTDS1102	2-2-0	4

Course Objectives:

- Demonstrate understanding of basic mathematical concepts in data science, relating to linear algebra, probability, and calculus.
- Employ methods related to these concepts in a variety of **data science** applications.
- Apply logical thinking to problem-solving in context.
- Demonstrate skills in writing **mathematics**.

Module I (10 Hours):

Basics of Data Science: Introduction; Typology of problems; Importance of linear algebra, statistics and optimization from a data science perspective; structured thinking for solving data science problems.

Module II (10 Hours):

Linear Algebra: Matrices and their properties (determinants, traces, rank, nullity, etc.); Eigen values and eigenvectors; Matrix factorizations; Inner products; Distance measures; Projections; Notion of hyperplane; half-planes.

Module III (10 Hours):

Probability, Statistics and Random Processes: Probability theory and axioms; Random variables; Probability distributions and density functions (univariate and multivariate); Expectations and moments; Covariance and correlation; Statistics and sampling distributions; Hypothesis testing of means, proportions, variances and correlations; Confidence (statistical) intervals; Correlation functions; White-noise process.

Module IV (10 Hours):

Optimization: Unconstrained optimization; Necessary and sufficiency conditions for optima; Gradient descent methods; Constrained optimization, KKT conditions; Introduction to non-gradient techniques; Introduction to least squares optimization; Optimization view of machine learning.5. Introduction to Data Science Methods: Linear regression as an exemplar function approximation problem; linear classification problems.

Text Books:

1. G. Strang (2016). Introduction to Linear Algebra, Wellesley-Cambridge Press, Fifth edition, USA.
2. Bendat, J. S. and A. G. Piersol (2010). Random Data: Analysis and Measurement Procedures. 4th Edition. John Wiley & Sons, Inc., NY, USA:
3. Montgomery, D. C. and G. C. Runger (2011). Applied Statistics and Probability for Engineers. 5th Edition. John Wiley & Sons, Inc., NY, USA:
4. David G. Luenberger (1969). Optimization by Vector Space Methods, John Wiley & Sons (NY)

Reference Books:

1. Cathy O’Neil and Rachel Schutt (2013). Doing Data Science, O’Reilly Media

Data Mining and Data Warehousing

Subject Name	Code	T-P-P	(Credit)
Data Mining and Data Warehousing	MTDS1104	2-2-0	4

Course Objectives:

- To develop the abilities of critical analysis to data mining systems and applications.
- To implement practical and theoretical understanding of the technologies for data mining
- To understand the strengths and limitations of various data mining models.

Module I (10 Hours):

Data mining tasks – mining frequent patterns, associations and correlations, classification and regression for predictive analysis, cluster analysis , outlier analysis; advanced pattern mining in multilevel, multidimensional space – mining multilevel associations, mining multidimensional associations, mining quantitative association rules, mining rare patterns and negative patterns.

Module II (9 Hours):

Classification by back propagation, support vector machines, classification using frequent patterns, other classification methods – genetic algorithms, roughest approach, fuzz>set approach.

Module III (8 Hours):

Density – based methods –DBSCAN, OPTICS, DENCLUE; Grid-Based methods – STING, CLIQUE; Exception – maximization algorithm; clustering High- Dimensional Data; Clustering Graph and Network Data.

Module IV (13 Hours):

Introduction, web mining, web content mining, web structure mining, we usage mining, Text mining – unstructured text, episode rule discovery for texts, hierarchy of categories, text clustering. **Temporal and Spatial Data Mining:** Introduction; Temporal Data Mining – Temporal Association Rules, Sequence Mining, GSP algorithm, SPADE, SPIRIT Episode Discovery, Time Series Analysis, Spatial Mining – Spatial Mining Tasks, Spatial Clustering. Data Mining Applications.

Text Books:

1. Data Mining Concepts and Techniques, Jiawei Hang Micheline Kamber, Jian pei, Morgan Kaufmannn.
2. Data Mining Techniques – Arun K pujari, Universities Press.

Reference Books:

1. Introduction to Data Mining – Pang-Ning Tan, Vipin kumar, Michael Steinbach, Pearson.
2. Data Mining Principles & Applications – T.V Sveresh Kumar, B.Esware Reddy, Jagadish S Kalimani, Elsevier.

Statistical Methods

Subject Name	Code	T-P-P	(Credit)
Statistical Methods	MTDS1105	2-2-0	4

Course Objectives:

- To understand the role of statistic and probability in the spatial data analysis and design.
- Solve the problems using the advanced statistical approaches.
- Identify the statistical methods for solving geospatial problems, apply the advanced statistical methods for image processing and to use geostatistics for studying spatially varying phenomena.

Module I (10 Hours):

Basic Statistics: Sources of Data, Organization of Data, The Histogram, Measures of central tendency, Mean Deviation, Standard Deviation, Correlation, Coefficient of correlation, Rank correlation, Regression.

Module II (12 Hours):

Probability: equally likely, mutually exclusive events, definitions of probability, additions & multiplication theorems of probability and problems based on them. Bayesian approach, distributions; Poisson, normal, Erlang, Gamma and Weibull probability distributions. Multivariate Data:Random Vectors and Matrices, sample estimate of centroid, standard deviation, SSCP, dispersion, variance, covariance, correlation matrices.

Module III (10 Hours):

Multivariate Regression Models, Multiple linear Regression: Multiple parameter estimation by method of least squares, tests of significance use of dummy variables, problems associated with multi colinearity, heteroscadasticity.

Module IV (8 Hours):

Pattern Analysis, Measures of Arrangements & dispersion, Auto Correlation, Semiveriogram, Kriging

Text Books:

1. Gupta, S.C. and Kapoor, V.K., “Fundamentals of Mathematics Statistics”, Sultan Chand and Sons, 2001.
2. Johnson, R.J., “Miller and Freund’s Probability and Statistics for Engineers” 6th Edition, Prentice Hall of India, 2002.

Reference Books:

1. Jay L. Devore, “Probability and statistics for Engineering and the Sciences”, Thomson and Duxbbury, 2002.
2. Sarma, D.D. “Geostatistics with Applications in Earth Sciences”, Capital Publishing Company, 2002.
3. Cooley W. W and Lohnes P.R .- Multivariate Data Analysis, John Wiley and Sons,1971.

Machine Learning Using Python

Subject Name	Code	T-P-P	(Credit)
Machine Learning Using Python	MTDS1103	0-4-0	4

Course Objectives:

- Apply multilayer perceptron using simple machine learning techniques.
- Use decision trees and statistics models.
- Use data analysis for machine learning.
- UseGeneticalgorithmandreinforcedlearningforappropriateapplications.
- Use the Python programming for machine learning.

Module I (10 Hours):

Learning - Types of machine learning - Supervised learning - The brain and the neurons, Linear Discriminants -Perceptron - Linear Separability -Linear Regression - Multilayer perceptron – Examples of using MLP - Back propagation of error.

Module II (12 Hours):

Decision trees- Constructing decision trees-Classification of regression trees- Regression example - Probability and Learning: Turning data into probabilities - Some basic statistics - Gaussian mixture models - Nearest Neighbor methods.

Module III (10 Hours):

The k-Means algorithm - Vector Quantization’s - Linear Discriminant Analysis - Principal component analysis - Factor Analysis - Independent component analysis - Locally Linear embedding – Isomap - Least squares optimization - Simulated annealing. The Genetic algorithm - Genetic operators - Genetic programming - Combining sampling with genetic programming - Markov Decision Process - Markov Chain Monte Carlo methods: sampling - Monte carlo - Proposal distribution.

Module IV (8 Hours):

Bayesian Networks - Markov Random Fields – Hidden Markov Models -Tracking methods. Python: Installation –Python for MAT LAB and R users-Code Basics –Using NumPy and MatPlotLiB.

Text Books:

1. Kevin P. Murphy, “Machine Learning – A probabilistic Perspective”, MIT Press, 2016.
2. Randal S, “Python Machine Learning, PACKT Publishing, 2016.

Reference Books:

1. Ethem Alpaydin, "Machine Learning: The New AI”, MIT Press, 2016.
2. Shai Shalev-Shwartz, Shai Ben-David, "Understanding Machine Learning: From Theory to Algorithms”, Cambridge University Press, 2014.
3. Sebastian Raschka, “Python Machine Learning”, Packt Publishing Ltd, 2015.

Syllabus

Course Structure

1st year-2ndsemester

Sl No	Code	Subject name	T	P	P	Credits
1	MTDS1201	Design and Analysis of Algorithms	2	2	0	4
2	MTDS1202	Big Data Systems	2	2	0	4
3	MTDS1204	Information Retrieval	2	2	0	4
4	MTDS1205	Computational Intelligence	2	2	0	4
		Practice				
5	MTDS1203	Digital Image processing	0	4	0	4
		TOTAL				**20**

Semester -II

Design and Analysis of Algorithms

Subject Name	Code	T-P-P	(Credit)
Design and Analysis of Algorithms	MTDS1201	2-2-0	4

Course Objectives:

- Focus on the design of algorithms in various domains
- Provide a foundation for designing efficient algorithms.
- Provide familiarity with main thrusts of working algorithms-sufficient to gives context for formulating and seeking known solutions to an algorithmic problem.

Module I (7 Hours):

The role of Algorithms in Computing, Insertion Sort, Analyzing algorithms, Designing algorithms, Asymptotic notations

Module II (13 Hours):

Divide and Conquer Technique, heapsort, merge sort, quick sort and their time complexity, Red-Black Trees
Dynamic Programming: (Matrix-chain multiplication, Longest common subsequences), Greedy Technique: An activity selection problem , Elements of greedy strategy, Huffman codes

Module III (10 Hours):

Single –Source Shortest Paths: The Bellman-Ford algorithm, Single-source shortest paths in directed acyclic graphs, Dijkstra's algorithm.
String Matching: The naïve string matching algorithm, The Rabin Karp algorithm

Module IV (10 Hours):

NP completeness, Reductions, coping with NP completeness, Approximation algorithms: The vertex cover problem, the travelling salesman problem, The set covering problem, The Subset-sum problem. Graph colouring.

Text Books:

5. T. H. Cormen, C. E. Leiserson, R. L. Rivest, Clifford Stein. "Introduction to Algorithms," Third edition ,Prentice Hall India, 2011

Reference Books:

3. Sara. Basse, Allen Van Gelder, "Computer Algorithms: Introduction to Design and Analysis", Pearson, 2000.
4. R. Motwani and P. Raghavan, "Randomized Algorithms," Cambridge University Press, 1995.
5. Dexter C .Kozen, "The Design and Analysis of Algorithms," Springer, 1992.

Big Data Systems

Subject Name	Code	T-P-P	(Credit)
Big Data Systems	MTDS1202	2-2-0	4

Course Objectives:

- To understand the need of Big Data, challenges and different analytical architectures
- Installation and understanding of Hadoop Architecture and its ecosystems
- Processing of Big Data with Advanced architectures like Spark.
- Describe graphs and streaming data in Spark

Module I (10 Hours):

Data Storage and Analysis - Characteristics of Big Data – Big Data Analytics - Typical Analytical Architecture – Requirement for new analytical architecture – Challenges in Big Data Analytics – Need of big data frameworks

Module II (10 Hours):

Hadoop – Requirement of Hadoop Framework - Design principle of Hadoop –Comparison with other system - Hadoop Components – Hadoop 1 vs Hadoop 2 – Hadoop Daemon's – HDFS Commands – Map Reduce Programming: I/O formats, Map side join, Reduce Side Join, Secondary sorting, Pipelining MapReduce jobs

Module III (10 Hours):

Introduction to Hadoop ecosystem technologies: Serialization: AVRO, Co-ordination: Zookeeper, Databases: HBase, Hive, Scripting language: Pig, Streaming: Flink, Storm

Module IV (10 Hours):

Introduction to GPU Computing, CUDA Programming Model, CUDA API, Simple Matrix, Multiplication in CUDA, CUDA Memory Model, Shared Memory Matrix Multiplication, Additional CUDA API Features.

Text Books:

5. Mohammed Guller, Big Data Analytics with Spark, Apress,2015 Reference Books:

Reference Bookss

2. Mike Frampton, "Mastering Apache Spark", Packt Publishing, 2015.
3. TomWhite,"Hadoop:TheDefinitiveGuide",O'Reilly,4thEdition,2015.

3. NickPentreath,MachineLearningwithSpark,PacktPublishing,2015.
4. Donald Miner, Adam Shook, “Map Reduce Design Pattern”, O’Reilly, 2012

Information Retrieval

Subject Name	Code	T-P-P	(Credit)
Information Retrieval	MTDS1204	2-2-0	4

Course Objectives:

- To provide an overview of Information Retrieval.
- Introduce students about insights of the several topics of Information retrieval such as – Boolean retrieval model, Vector space model, indexing.
- Provide comprehensive details about various Evaluation methods.
- Provide implementational insight about the topics covered in the course.

Module I (10 Hours):

Introduction to Retrieval and IR Model: Information, Information Need and Relevance; The IR System; Boolean Retrieval; Term Vocabulary and Postings list; Index Construction; Ranked and other alternative Retrieval Models.

Module II (9 Hours):

Dictionary and Tolerant Retrieval: Tokenization, Stop words, Stemming, Inverted index, Wild card queries, Jaccard coefficient, Understand the importance of Scoring, term weighting.

Module III (8 Hours):

Retrieval Evaluation: Precision, Recall, F-measure, E-measure, Normalized recall, Evaluation problems, Document Processing: Representation; Vector Space Model; Feature Selection; Stop Words; Stemming; Notion of Document Similarity; Standard Datasets.

Module IV (13 Hours):

Classification and Clustering: Notion of Supervised and Unsupervised Algorithms; Naive Bayes, Nearest Neighbour; Clustering Methods such as K-Means, Introduction to recommendation system Collaborative , Content based recommendation.

Text Books:

1. Introduction to Information Retrieval , Christopher D. Manning and Prabhakar Raghavan and Hinrich Schütze, Cambridge University Press, 2008.

Reference Books:

1. Information Storage and Retrieval Systems: Theory and Implementation, Kowalski, Gerald, Mark T Maybury, Springer.
2. Modern Information Retrieval, Ricardo Baeza-Yates, Pearson Education, 2007.
3. Information Retrieval: Algorithms and Heuristics, David A Grossman and Ophir Frieder, 2nd Edition, Springer, 2004.
4. Information Retrieval Data Structures and Algorithms, William B Frakes, Ricardo BaezaYates, Pearson Education, 1992.

Computational Intelligence

Subject Name	Code	T-P-P	(Credit)
Computational Intelligence	MTDS1205	2-2-0	4

Course Objectives:

- Ability to design and implement various machine learning algorithms in a range of real-world applications.
- Enable computers to perform such intellectual tasks as decision making, problem solving, perception and understanding human communication etc.

Module - I (10 Hrs.)

Artificial Neural Network(ANN): Fundamentals of ANN, Basic Models of an artificial Neuron, Neural Network Architecture, Learning methods, Terminologies of ANN, Supervised Learning Networks: Perceptron, MLP, Architecture of a Back propagation Network : back propagation

Module –II (10Hrs)

FUZZY LOGIC

Fuzzy set theory: crisp sets, fuzzy sets, crisp relations, fuzzy relations, Fuzzy Systems: Crisp logic predicate logic, fuzzy logic, fuzzy Rule based system, Defuzzification Methods, Fuzzy rule based reasoning

Module –III (10Hrs)

GENETIC ALGORITHMS

Fundamentals of genetic algorithms: Encoding, Fitness functions, Reproduction. Genetic Modeling: Cross cover, Inversion and deletion, Mutation operator, Bit-wise operators, Bitwise operators used in GA. Convergence of Genetic algorithm. Applications, Real life Problems.

Module – IV (6 Hrs.)

Hybrid Soft Computing Techniques Hybrid system, neural Networks, fuzzy logic and Genetic algorithms hybrids. Genetic Algorithm based Back propagation Networks: GA based weight determination applications: Fuzzy logic controlled genetic Algorithms soft computing tools, Applications.

Text Book :

Principles of Soft Computing- S.N.Sivanandan and S.N.Deepa, Wiley India, 2ndEdition,2011

Reference Book:

1. Neuro Fuzzy and Soft Computing, J. S. R. JANG,C.T. Sun, E. Mitzutani, PHI

2. Neural Networks, Fuzzy Logic, and Genetic Algorithm (synthesis and Application) S.Rajasekaran, G.A. VijayalakshmiPai, PHI

Digital Image processing

Subject Name	Code	T-P-P	(Credit)
Digital Image processing	MTDS1203	0-4-0	4

Course Objectives:

- Learn the fundamental concept of digital image processing
- Study image processing operations
- Understand image processing algorithms
- Expose students to current applications in the field of digital image processing

Module I (10 Hours):

Basic concepts of digital image processing, Steps in Digital image processing, components of an image processing system. Histogram processing, spatial filtering, smoothing spatial filters, sharpening spatial filters.

Module II (12 Hours):

Noise models, Restoration in the presence of noise, periodic noise reduction by frequency domain filtering, linear, position-invariant degradations, estimating the degradation function. Inverse filtering, minimum mean square error filtering, constrained least square filtering, geometric mean filter, image reconstruction from projections.

Module III (10 Hours):

Color models, Pseudo color image processing, basics of full color image processing, color transformations, smoothing and sharpening, image segmentation based on color, color image compression.

Module IV (8 Hours):

Image segmentation, point, line and edge detection, thresholding, region based segmentation, segmentation using morphological watersheds, the use of motion in segmentation

Text Books:

3. Digital Image Processing, Rafeal C.Gonzalez, Richard E.Woods, Second Edition, Pearson Education/PHI.

Reference Books:

4. Image Processing, Analysis, and Machine Vision, Milan Sonka, Vaclav Hlavac and Roger Boyle, Second Edition, Thomson Learning.
5. Introduction to Digital Image Processing with Matlab, Alasdair McAndrew, Thomson Course Technology
6. Computer Vision and Image Processing, Adrian Low, Second Edition, B.S.Publications.
7. Digital Image Processing using Matlab, Rafeal C.Gonzalez, Richard E.Woods, Steven L. Eddins, Pearson Education.

Syllabus

Course Structure (Thesis)

2nd year-3rd semester and 4th semester

Thesis Part I

MTIP 2101 Industry Internship & Project I (16 credits) 0-16-0 (PR)

Thesis Part II

MTIP 2201 Industry Internship & Project II (16 credits) 0-16-0 (PR)

Evaluation Criteria

Theory			Practice
Internal	(30 X3) Average of best of two will be taken into consideration		14 sets of experiment to be identified, out of those 10 will be conducted in the lab having each carries 10marks
Assignment	5Marks		
Attendance	5Marks		
External	60 Marks		

Exceptionally for Research Methodology for Engineers:

Internal Theory Examination	60 Marks
External Theory Examination	40 Marks

Industry Internship Project I: Student will be performing internship for 3-months in Industry and submitted a report and duly certified certificate from the competent Engineer in charge

Industry Internship Project II: Student will be doing live project for 4 months in an Industry and submit a Project Report to concerned guide certified by the Industry Engineer