School of Information Sciences, Manipal University

Master of Engineering – ME (Big Data and Data Analytics)

Program Structure (August 2016 Onwards)

ME (Big Data and Data Analytics) – Semester I										
Subject Code	**Subject Name**	**Common With**	**No of Hrs./week**				**Duration of Exam in Hrs**	**Maximum Marks**		
			Lecture	**Tutorial**	**Practical**	**Credit**		**Internal 50**	**External 50**	**Total 100**
BDA 601	Algorithms and Data Structures for Big Data		3	-	-	3	3	50	50	100
BDA 603	Large Scale Distributed Computing Systems		3	-	-	3	3	50	50	100
BDA 605	Probability and Statistical Inferences		3	-	-	3	3	50	50	100
BDA 607	Modern Database Management Systems		3	-	-	3	3	50	50	100
	Elective – I		3	-	-	3	3	50	50	100
BDA 651	Algorithms and Data Structures for Big Data Lab		-	-	3	1	3	50	50	100
BDA 653	Large Scale Distributed Computing Systems Lab		-	-	3	1	3	50	50	100
BDA 655	Probability and Statistical Inferences Lab		-	-	3	1	3	50	50	100
BDA 657	Modern Database Management Systems Lab		-	-	3	1	3	50	50	100
BDA 659	Elective – I Lab		-	-	3	1	3	50	50	100
BDA 695	Mini Project – I		-	-	-	4	-	100	-	100
BDA 697	Seminar – I		-	-	-	1	-	100	-	100
Total			**15**	**-**	**15**	**25**				

ME (Big Data and Data Analytics) – Semester II

Subject Code	Subject Name	Common With	No of Hrs. / week				Duration of Exam in Hrs	Maximum Marks		
			Lecture	Tutorial	Practical	Credit		Internal 50	External 50	Total 100
BDA 602	Machine Learning		3	-	-	3	3	50	50	100
BDA 604	Architecture of Big Data Systems		3	-	-	3	3	50	50	100
BDA 606	Multiple Linear Regression and Logistic Regression		3	-	-	3	3	50	50	100
BDA 608	Healthcare Informatics		3	-	-	3	3	50	50	100
	Elective – II		3	-	-	3	3	50	50	100
BDA 652	Machine Learning Lab		-	-	3	1	3	50	50	100
BDA 654	Architecture of Big Data Systems Lab		-	-	3	1	3	50	50	100
BDA 656	Multiple Linear Regression and Logistic Regression Lab		-	-	3	1	3	50	50	100
BDA 658	Healthcare Informatics Lab		-	-	3	1	3	50	50	100
BDA 660	Elective – II Lab		-	-	3	1	3	50	50	100
BDA 696	Mini Project – II		-	-	-	4	-	100	-	100
BDA 698	Seminar – II		-	-	-	1	-	100	-	100
TOTAL			**15**	**-**	**15**	**25**				

ME (Big Data and Data Analytics) – Semester III & IV

BDA 799	Project Work	-	-	-	25				
Total Number of Credits to Award Degree							75		

The List of Electives Offered in the First and Second Semesters

Elective I			Elective II		
Sub Code	**Subject**	**Common With**	**Sub Code**	**Subject**	**Common With**
BDA 615.1	DevOps for Big Data Systems		BDA 616.1	Text Retrieval and Search Engines	
BDA 615.2	Mobile Web Application Development		BDA 616.2	Applied Multivariate Analysis	

Semester I

BDA 601: Algorithms and Data Structures for Big Data

[L-T-P-C: 3-0-0-3]

Module Duration: 36 Hours

Algorithm specification and analysis techniques 3 hours

Analysis of recursive programs.
Solving recurrence equations.
General solution for a large class of recurrences.

Elementary data structures 4 hours

Implementation of lists, stacks, queues.

Sorting and Searching Techniques 5 hours

Quick sort, heap sort, merge sort.
Linear search and binary search.

Hashing and Dictionaries 3 hours

Binary search trees 2 hours

Construction.
Inorder, preorder and postorder traversals.

Graphs 6 hours

Representation of graphs. Depth First Searching. Breadth First Searching.
Minimum cost spanning tree.
Single source shortest paths and all-pairs shortest path.

String and text processing techniques 5 hours

Pattern-Matching Algorithms.
Text Compression.
Tries.

Data stream algorithms 8 hours

Sampling, Random Projections, Basic Algorithmic Techniques
Group Testing, Tree Method and Graph sketching.

REFERENCES

1. Introduction to Algorithms - Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest. MIT Press.
2. Data Structures and Algorithms - Aho, Hopcroft and Ulmann. Pearson Publishers.

3. Data Structures and Algorithms in Python - Michael T. Goodrich, Roberto Tamassia, and Michael H. Goldwasser. John Wiley & Sons.
4. Data Streams: Algorithms and Applications - S. Muthukrishnan. Foundations and Trends in Theoretical Computer Science archive, Volume 1 Issue 2, August 2005, Pages 117 – 236.

BDA 603: Large Scale Distributed Computing Systems

[L-T-P-C: 3-0-0-3]

Module Duration: 36 Hours

Introduction to fundamental concepts 2 hours

- Models of distributed computations.
- Global state of a distributed system.
- Modern cloud-based distributed systems.

Architectures 3 hours

- Centralized, decentralized and hybrid architectures.
- Architectures versus middleware.
- Self-management in distributed systems.
- Contemporary cloud-based architectural concepts.

Processes 3 hours

- Clients and server processes.
- Role of virtualization in distributed systems.
- Process migration.

Communication 4 hours

- Remote Procedure Calls.
- Message-oriented communication.
- Stream-oriented communication.

Synchronization - election algorithms 3 hours

- Clock synchronization and logical clocks.
- Distributed mutual exclusion algorithms.
- Elections in wireless and large scale systems.

Consistency, replication and recovery 7 hours

- Replication as a scaling technique.
- Data-centric consistency models. Client-centric consistency models.
- Content distribution and replication.
- Consistency protocols.
  - Continuous consistency. Replicated-write protocols.
  - Cache-coherence protocols. Client-centric consistency.
- CAP theorem.

Security 6 hours

Security Threats, Policies, and Mechanisms.
Authentication. Message Integrity and Confidentiality.
Secure group communication.
Access control - general issues in access control.
Firewalls. Secure mobile mode. Denial of Service attacks.
Security management.
Key Management. Secure Group Management.
Authorization Management
Kerberos

Distributed Web-based systems 8 hours

Client-server architectures.
Processes and communication – HTTP 2.0 and REST.
Distributed Naming Services.
Consistency and replication.
Fault tolerance.
Mobile devices and high-performance browser networking.
Current methods and practices in enforcing Web security.

REFERENCES

1. Distributed Systems: Principles and Paradigms - Andrew S Tanenbaum and Maarten Van Steen. Cambridge University Press.
2. Distributed Computing: Principles, Algorithms, and Systems - Ajay D. Kshemkalyani, Mukesh Singhal. Cambridge University Press.
3. Distributed Systems: Concepts and Design (Fifth edition) - George Coulouris, Jean Dollimore, Tim Kindberg, and Gordon Blair. Pearson Publishers.
4. High Performance Browser Networking: Ilya Grigorik. O'Reilly Atlas. 2013.

BDA 605: Probability and Statistical Inferences

[L-T-P-C: 3-0-0-3]

Module Duration: 36 Hours

Applied probability 6 hours

- Random experiments, events, sample space
- Probability approaches in definition, properties of probability
- Addition and multiplication theorems of probability
- Conditional probability and Bayes theorem.

Random variables and distributions 10 hours

- Random variables, probability mass function
- Probability density function
- Expectation and its properties
- Joint distributions, marginal distributions and conditional distribution
- Discrete distributions
  - Binomial and Poisson distribution
  - Continuous distributions
    - Normal,
    - Student distribution
    - F distribution and Chi-square distribution.

Point and interval estimation 6 hours

- Parameter, statistic, estimate and estimator
- Unbiased estimator, sampling distribution, standard error
- Distribution of sample mean, Central Limit Theorem (only statement)
- Sampling distribution of mean and proportion

Hypothesis testing 14 hours

- Introduction - types of errors, level of significance
- Power, p value- interpretation, NP lemma
- Hypothesis testing
  - Independent sample test
  - Paired test
- One way ANOVA and post hoc tests.
- Introduction to multiple testing problem
  - Family wise error and false discovery rate.
- Introduction to non-parametric methods
  - Mann Whitney test
  - Wilcoxon signed rank test
  - Kruskal Wallis test and Chi square test.

REFERENCES

1. The theory of probability - B. Gnedenko. Mir Publishers, 1978.
2. Modern probability theory: an introductory text book - B. R. Bhat. Wiley Eastern Limited, 1989.
3. Univariate Discrete Distributions - John N L, Kotz S and Kemp AW. John Wiley & Sons, 1992.
4. Continuous Univariate Distributions I & II - John N L, Kotz S and Balakrishnan N. John Wiley & Sons, 1991.
5. Probability and statistical inference - Robert V Hogg and Elliot A Tanis. Macmillan Publishing Company.
6. Statistical Inference - George Cesella & Roger L Berger. Duxbury Thomson Learning.
7. Introduction to Probability Models - M. Ross. Ninth edition; Elsevier Inc, 2007
8. An Introduction to Statistical Learning with Applications in R - Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. Springer, 2013

BDA 607: Modern Database Management Systems

[L-T-P-C: 3-0-0-3]

Module Duration: 36 Hours

Introduction to the growth of traditional and modern database management systems. 1 hour

Foundations of relational data management 4 hours

- Relational operators
- Relational algebra
- Tuple and domain relational calculus.

SQL - syntax and semantics 3 hours

Designing relational data - normal forms 2 hours

Transaction processing 4 hours

- Rollback and compensating transactions.
- ACID properties.
- Locking
    - Deadlocks. Performance. Hotspots. Query-Update problems.
    - B-Tree locking.
    - Locking nested transactions.

Two-Phase commit protocol. 3 hours

- Failure handling.
- Optimizations and variations.

Semi-structured data management 3 hours

- XML, JSON and Web documents.
- XPath and XQuery – an XML Query language

NoSQL – origins, growth and applications. 6 hours

- Schemaless databases and materialized views.
- Varieties of NoSQL stores
- Document, key-value, object-based, tuple store, triple-quad store.
- Graph databases.
- Multi-model and correlation data stores. Ad-hoc query processing.

Distribution models for scalability	4 hours
Horizontal partitioning.	
Data sharding.	
Master-slave replication. Peer-to-peer replication.	
Version stamps – business and system transactions.	
MapReduce	3 hours
Basic MapReduce Partitioning and combining.	
Composing MapReduce calculations.	
Two-stage map-reduce example. Incremental MapReduce.	
UnQL: a query language and algebra for semi-structured data based on structural recursion.	2 hour
NewSQL	1 hour
Difficulties in maintaining transaction consistency in NoSQL data stores.	
The Google Spanner system – background and its implementation.	

REFERENCES

1. Database System Concepts (Sixth Edition) - Avi Silberschatz, Henry F. Korth, and S. Sudarshan. McGraw Hill.
2. Principles of Transaction Processing (Second Edition) – Philip A Bernstein and Eric Newcomer. Morgan Kaufmann Publishers.
3. NoSQL Distilled – Pramod J Sadalage and Martin Fowler. Addison-Wesley Publisher.

BDA 615.1: DevOps for Big Data Systems

[L-T-P-C: 3-0-0-3]

Module Duration: 36 Hours

Agile Infrastructure management and DevOps 1 hours

- History and its importance in managing large scale information systems.
- Workflow Automation
  - Agile methodology and Continuous Delivery model.
  - Configuration management tools.
  - Version control systems.
  - Virtualization and Cloud infrastructure management.

Scripting with Python and Linux Shell 6 hours

- Python for managing computing systems and resources.
- BASH Shell scripting.
- Make, Configure and related build automation tools.

SSH and centralized authentication/authorization 2 hours

Version control systems and Configuration Management Systems 5 hours

- Git and GitHub
- Continuous integration with Github
- Ansible for configuration management.

The Hadoop environment 13 hours

- HDFS
- MapReduce with YARN
- Cluster setup and configuration
- Hive – data summarization, query and analysis
- Spark – high volume in-memory data querying and analysis
- HBase distributed database

HTML5 and JavaScript for DevOps 4 hours

- Web application architecture and their performance behaviours.
- Essential JavaScript for DevOps.

Web Applications, testing and performance analysis 5 hours

- Using Web Browser developer tools for troubleshooting Web applications.
- Web application analytics.
- Headless Website testing with PhantomJS
- Automating Web application testing with Selenium.
- Web application metrics collection, performance monitoring, and alerting.

1. DevOps for Developers - Michael Httermann. APRESS, September 18, 2012.
2. Effective DevOps - Building a Culture of Collaboration, Affinity, and Tooling at Scale. O’Reilly Publisher.
3. Atlassian Git Tutorials - https://www.atlassian.com/git/tutorials/
4. AWS OpsWorks documentation - https://aws.amazon.com/documentation/opsworks/
5. Hadoop: The Definitive Guide. 4th Edition - Storage and Analysis at Internet Scale. Tom White. O'Reilly Media. March 2015.

BDA 615.2: Mobile Web Application Development

[L-T-P-C: 3-0-0-3]

Module Duration: 36 Hours

Challenges of mobile Web application development 2 hours

- The limitations of mobile networks.
- Reducing the page weight - the amount of markup and external elements.
- Avoiding useless network usage.
- Understanding the “mobile-first” design principles.
- Limitations imposed by battery life.

Setting up a personal Web site 4 hours

- Setting free VMs - micro-instances - on AWS.
- Installing and configuring NGINX on AWS micro instances
- Working with routing and reverse proxies
- HTTP and REST APIs

HTML5 and CSS for mobile devices. 4 hours

- Media queries for handling mobile form-factors.
- Principles and practice of responsive design.
- Mobile UX, Viewport, Fluid design and responsive images

Programming with JavaScript and DOM APIs 5 hours

- Accessing document fragments
- Using jQuery and other light-weight libraries
- AJAX and asynchronous programming

Architecture of Android applications 6 hours

Programming for technologies available on smart phones 7 hours

- Introduction to PhoneGap
- Handling Touch events.
- Making use of the accelerometer and the Location APIs.
- Accessing camera and media devices.

Developing offline facilities in mobile web applications 2 hours

- Localstorage and IndexDB APIs

Designing and developing secure mobile web applications 6 hours

- Understanding the single-origin policy
- Dangers of Cross-site scripting
- Principles of the secure socket layer and HTTPS
- Practical encryption for client-server communication in Web applications.
- Best practices in developing secure client-side code

REFERENCES

1. Learning Web App Development (Build Quickly with Proven JavaScript Techniques) - Semmy Purewal. O'Reilly Media. 2014.

2. The Browser Security Handbook. Michal Zalewski. https://code.google.com/p/browsersec/wiki/Main

3. High Performance Responsive Design - Tom Barker. O'Reilly publisher. 2014.

4. Apple UI Design Basics. https://developer.apple.com/library/ios/documentation/UserExperience/Conceptual/MobileHIG/index.html

5. Android Design Principles. https://developer.android.com/design/index.html

6. Android Application Development Reference. https://developer.android.com/develop/index.html

BDA 651: Algorithms and Data Structures for Big Data Lab

[L-T-P-C: 0-0-3-1]

Lab exercises on the subject studied in BDA 601: Algorithms and Data Structures for Big Data

BDA 653: Large Scale Distributed Computing Systems Lab

[L-T-P-C: 0-0-3-1]

Lab exercises on the subject studied in BDA 603: Large Scale Distributed Computing Systems

BDA 655: Probability and Statistical Inferences Lab

[L-T-P-C: 0-0-3-1]

Lab exercises on the subject studied in BDA 605: Probability and Bayesian Methods Using R

BDA 657: Modern Database Management Systems Lab

[L-T-P-C: 0-0-3-1]

Lab exercises on the subject studied in BDA 607: Modern Database Management Systems

BDA 659: Elective – I Lab

[L-T-P-C: 0-0-3-1]

Lab exercises on the subject studied in elective-1

BDA 695: Mini Project – I

[L-T-P-C: 0-0-0-4]

A single semester mini-project. This will be offered by the faculty in consultation with industry partners and mentors. Students have to survey the required background and develop software meeting the requirements.

BDA 697: Seminar – I

[L-T-P-C: 0-0-0-1]

Students select a topic of interest relevant to the field of study. They prepare a short report (not exceeding five pages) and present their understanding in a classroom.

Semester II

BDA 602: Machine Learning

[L-T-P-C: 3-0-0-3]

Module Duration: 36 Hours

Introduction 1 hour

Definition of learning systems.

Goals and applications of machine learning.

Aspects of developing a learning system

Training data, concept representation, and function approximation.

Inductive Classification 3 hours

The concept learning task.

Concept learning as search through a hypothesis space.

General-to-specific ordering of hypotheses.

Finding maximally specific hypotheses.

Version spaces and the candidate elimination algorithm.

Learning conjunctive concepts. The importance of inductive bias.

Predictive analytics – Supervised learning

Decision Tree learning 3 hour

Representing concepts as decision trees.

Recursive induction of decision trees.

Picking the best splitting attribute

Entropy and information gain

Searching for simple trees and computational complexity

Ensemble methods (bagging and boosting) 3 hours

Using committees of multiple hypotheses.

Bagging, boosting, and DECORATE.

Active learning with ensembles.

Computational learning theory. 3 hour

Models of learnability: learning in the limit

Probably approximately correct (PAC) learning.

Sample complexity: quantifying the number of examples needed to PAC learn.
Computational complexity of training. Sample complexity for finite hypothesis spaces.

Bayesian learning 5 hours

Probability theory and Bayes rule.
Naive Bayes learning algorithm.
Parameter smoothing.
Generative vs. discriminative training.
Logistic regression.
Bayes nets and Markov nets for representing dependencies.

Instance-based learning. 3 hour

Constructing explicit generalizations versus comparing to past specific examples.
K-Nearest Neighbour algorithm.
Case-based learning.

Support Vector Machine (SMV) 4 hours

Maximum margin linear separators.
Quadractic programming solution to finding maximum margin separators.
Kernels for learning non-linear functions.

Descriptive analytics – unsupervised learning

Artificial Neural Networks 6 hours

Neurons and biological motivation.
Linear threshold units.
Perceptrons: representational limitation and gradient descent training.
Multilayer networks and back propagation.
Hidden layers and constructing intermediate, distributed representations.
Overfitting,

Clustering 5 hours

Learning from unclassified data.
Clustering. Hierarchical Aglomerative Clustering.
Non-Hierarchical Clustering - k-means partitional clustering.
Expectation maximization (EM) for soft clustering.
Semi-supervised learning with EM using labeled and unlabled data.

REFERENCES

1. Pattern Recognition and Machine Learning - Christopher M. Bishop. Springer.
2. Machine Learning - Tom Mitchell. McGraw Hill.
3. An introduction to support vector machines - Cristianini, N. and J. Shawe-Taylor. Cambridge University Press.
4. Machine Learning: The Art and Science of Algorithms that Make Sense of Data - Flach, Peter. Cambridge University Press.
5. Artificial Intelligence: A Modern Approach (Third Edition) - Russell, Stuart and Peter Norvig. Prentice Hall.
6. Pattern Classification (Second Edition) - Duda, R., P. Hart, and D. Stork. Wiley Publishers.
7. A Course in Machine Learning - Hal Daumé III (http://ciml.info/)
8. Analytics in a Big Data World - Bart Baesens. Wiley.
9. Ensemble Learning - Thomas G. Dietteri in The Handbook of Brain Theory and Neural Networks, Second edition, (M.A. Arbib, Editor), Cambridge, MA: The MIT Press, 2002.
10. Generative and discriminative classifiers: naïve Bayes and logistic regression. http://www.cs.cmu.edu/~tom/mlbook/NBayesLogReg.pdf

BDA 604: Architecture of Big Data Systems

[L-T-P-C: 3-0-0-3]

Module Duration: 36 hours

Classifying big data characteristics 3 hours

- Analysis type - real time or batched for later analysis.
- Processing methodology - predictive, analytical, ad-hoc query, and reporting.
- Data frequency and size
  - On demand, as with social media data
  - Continuous feed, real-time - weather data, transactional data
  - Time series - time-based data
- Data type - transactional, historical, master data and metadata.
- Content formats - structured, unstructured, semi-structured
- Data sources - Web and social media, humans, machines, transaction data and biometric data.
- Data consumers - Enterprise applications, business users and business processes

Big Data processing - the Lambda architecture 5 hours

- Append-only, immutable data
- Batch layer
- Serving layer
- Speed layer
- Case study: Druid - A Real-time Analytical Data Store.

Data storage on the batch layer 5 hours

- Choosing a storage solution for the batch layer
- Distributed file systems
- Vertical partitioning
- Case study: The Hadoop Distributed File System

Computing on the batch layer 5 hours

- Recomputation algorithms vs. incremental algorithms
- Scalability in the batch layer
- MapReduce: a paradigm for Big Data computing
- Case study: Summingbird library for distributed MapReduce platforms.

Serving layer 5 hours

- Performance metrics for the serving layer
- The serving layer solution to the normalization/denormalization problem
- Requirements for a serving layer database
- Case study: ElephantDB

Case study: Apache HBase

Speed layer 6 hours

Computing real time views
Storing real time views
Challenges of incremental computation
Asynchronous versus synchronous updates
Case study: Cassandra's data model.

Alternatives to MapReduce 7 hours

Multipass applications - low-latency data sharing across multiple parallel operations.
Interactive data mining, iterative machine learning algorithms and streaming applications.
Resilient Distributed Datasets (RDDs)
Efficient SQL engines
Discretized streams.
Fault-tolerant stream processing

REFERENCES

1. Big Data: Principles and best practices of scalable real-time data systems - Nathan Marz and James Warren. Manning Publisher.
2. Resilient Distributed Datasets: A Fault-Tolerant Abstraction for In-Memory Cluster Computing. Matei Zaharia, Mosharaf Chowdhury, Tathagata Das, Ankur Dave, Justin Ma, Murphy McCauley, Michael J. Franklin, Scott Shenker, Ion Stoica. NSDI 2012. April 2012.
3. Druid - A Real-time Analytical Data Store - Fangjin Yang, Eric Tschetter, Xavier Léauté, Nelson Ray, Gian Merlino, and Deep Ganguli. SIGMOD'14, June 22–27, 2014, Snowbird, UT, USA.
4. http://static.druid.io/docs/druid.pdf, http://druid.io/docs/0.8.0/design/design.html
5. Big data architecture and patterns - IBM developerWorks. http://www.ibm.com/developerworks/library/bd-archpatterns1/
6. Big Data and Analytics -IBM developerWorks. http://www.ibm.com/developerworks/analytics/
7. http://lambda-architecture.net/
8. Apache HBase - http://hbase.apache.org/
9. Apache Spark Streaming - https://spark.apache.org/streaming/
10. Summingbird MapReduce library - https://github.com/twitter/summingbird

BDA 606: Multiple Linear Regression and Logistic Regression

[L-T-P-C: 3-0-0-3]

Module Duration: 36 hours

Linear regression 18 hours

Pearson's correlation coefficient
Simple linear regression
ANOVA approach to regression, residuals
Multiple linear regression – assumptions, estimation of coefficients,
Coefficient of determination and adjusted coefficient of determination
Multicollinearity, methods to deal with multicollinearty, detection of multicollinearity
    VIF, dummy variables, model building strategies,
Model validation
Residual analysis and sensitivity analysis.

Logistic regression 18 hours

Introduction to logistic regression models
Link functions
Binary and multinomial logistic regression
    Estimation of coefficients
    Interpretation of coefficients
Model building strategies in logistic regression analysis
    Selection of independent variables by forward, backward and stepwise procedures.
Hosmer Lemshow test and Area under the ROC curve
    Assessing goodness of fit of the model and sensitivity analysis.

REFERENCES

1. Applied Linear Statistical Models - John Neter, Michael H Kutner, William Wasserman, Christopher J Nchtsheim.
2. Introduction to Linear Regression Analysis - Douglas C Montgomery, Elizabeth A Peck & G Geoffrey Vining. John Wiley & Sons, Inc.
3. Generalized Linear Models - McCullagh, P. and J.A. Nelder. Chapman and Hall: London, 1989.
4. Applied logistic Regression - David.W.Hosmer and Stanley Lemeshow. Wiley publications

BDA 608: Healthcare Informatics

[L-T-P-C: 3-0-0-3]

Module Duration: 36 hours

Theoretical foundations of health informatics.	1 hours
Evidence-based practice and informatics.	1 hours
Electronic health records and managing patient care.	2 hours
Telehealth and applications for delivering care at a distance.	2 hours
Imaging Technologies and their Applications in Biomedicine.	4 hours
Clinical decision support systems in healthcare.	3 hours
Public health informatics.	3 hours
Participatory healthcare informatics	2 hours
Privacy, confidentiality, security and data integrity.	4 hours
Patient Safety and Quality Initiatives in Informatics	4 hours
Improving user experience for health information technology products.	4 hours
Standards in Healthcare industry and practice.	2 hours
Distance Education: Applications, Techniques and Issues.	2 hours
Information Systems and Technical Tools in Healthcare Education	2 hours

REFERENCES

1. Health Informatics: An Interprofessional Approach - Ramona Nelson and Nancy Staggers.
2. Healthcare Informatics - WIlliam Hanson - McGraw-Hill Education.
3. Handbook of Research on Informatics in Healthcare and Biomedicine - Athina A. Lazakidou.

BDA 616.1: Text Retrieval and Search Engines

[L-T-P-C: 3-0-0-3]

Module Duration: 36 hours

Text and Web search basics	2 hours
Background and history Storing the Documents, detecting duplicates, removing noise Web characteristics, the web graph Advertising as the economic model The search user experience, User query needs, Index size and estimation	
Architecture of a Search Engine	5 hours
Basic building blocks, Text Acquisition, Text Transformation Index Creation, User Interaction, Ranking Blocked sort-based indexing Single-pass in-memory indexing Distributed indexing Dynamic indexing Other types of indexes	
Processing Text	4 hours
Lexical Analysis of the Text Elimination of Stopwords, stemming, Index Terms Selection, Thesaurus Link analysis, anchor text, information extraction Document Clustering, Text Compression	
Queries and Interfaces	3 hours
Query Transformation and Refinement Context and Personalization, Result Pages and Snippets Clustering the Results	
Boolean Retrieval	4 hours
An example information retrieval problem A first take at building an inverted index Processing Boolean queries The extended Boolean model versus ranked retrieval	
Scoring, term weighting and the Vector Space Model	4 hours
Parametric and zone indexes Term frequency and weighting The vector space model for scoring	

Probabilistic Models 5 hours

- Information Retrieval as Classification
- The Probability Ranking Principle
- The Binary Independence Model
- Tree-structured dependencies between terms
- Okapi BM25: a non-binary model

Text classification & Naive Bayes 5 hours

- The text classification problem
- Naive Bayes text classification
- Feature selection

Social Search 4 hours

- User Tags and Manual Indexing
- Searching With Communities
- Filtering and Recommending
- Peer-to-Peer and Metasearch
  - Distributed search
  - P2P Networks

REFERENCES

1. Search Engines: Information Retrieval in Practice - Bruce Croft, Donald Metzler, and Trevor Strohman. Addison-Wesley. 2010.
2. Introduction to Information Retrieval - Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. Cambridge University Press. 2008.
3. Understanding Search Engines: Mathematical Modeling and Text Retrieval (Second Edition) - Micheal W. Berry and Murray Browne. Society for Industrial and Applied Mathematics. 2005.
4. Modern Information Retrieval - Ricardo Baeza-Yates and Berthier Ribeiro-Neto. Addison Wesley. 1999.

BDA 616.2: Applied Multivariate Analysis

[L-T-P-C: 3-0-0-3]

Module Duration: 36 hours

Multivariate data and multivariate normal distribution 6 hours

Introduction to multivariate analysis
Data structure, mean vector, covariance & correlation matrices
Distance between vectors, multivariate normal distribution
properties of multivariate normal random variables.

Multivariate test on mean vectors and discriminant analysis 10 hours

Multivariate one-way analysis of variance model (MANOVA) Wilks' test statistic
Roy's test, Pillai and Lawley–Hotelling tests
The Discriminant Function for Two Groups
Tests for the Two-Group Case, Discriminant Analysis for Several Groups,
Tests for the Several-Group Case, Standardized Coefficients
Classification analysis
Classification into two and several groups
Linear classification function
Quadratic classification functions
Estimating misclassification rate, improved estimates of error rates
Partitioning the sample and Holdout method.

Principal Component Analysis and factor analysis 10 hours

Geometric and Algebraic Bases of Principal Components
Principal Components and Perpendicular Regression
Plotting of Principal Components
Principal Components from the Correlation Matrix
Deciding How Many Components to Retain
Information in the Last Few Principal Components
Interpretation of Principal Components, Factor analysis and factor rotation.

Cluster Analysis 10 hours

Measures of Similarity or Dissimilarity
Hierarchical Clustering- Single Linkage (Nearest Neighbour)
Complete Linkage (Farthest Neighbour), Average Linkage
Centroid, Median, Ward's Method, Divisive Methods
Non-hierarchical Methods – Partitioning - K means
Choosing the Number of Clusters, Cluster Validity, Clustering Variables
Introduction to multidimensional scaling.

REFERENCES

1. Methods of multivariate analysis - Alvin C Rencher. 2nd ed. USA: Wiley interscience; 2002.
2. Applied multivariate statistical analysis - Johnson R A and Wichern D W. Person education
3. Computer–aided multivariate analysis - Abdelmonem Afifi, Virginia A and Susanne May. 4th edition, Chapman & Hall/CRC, USA.
4. Multivariate analysis: methods and applications - Dillon William R and Goldstein Matthew. Wiley Series in Probability and Mathematical Statistics.

BDA 652: Machine Learning Lab

[L-T-P-C: 0-0-3-1]

Lab exercises on the subject studied in BDA 602: Machine Learning

BDA 654: Architecture of Big Data Systems Lab

[L-T-P-C: 0-0-3-1]

Lab exercises on the subject studied in BDA 604: Architecture of Big Data Systems

BDA 656: Multiple Linear Regression and Logistic Regression Lab

[L-T-P-C: 0-0-3-1]

Lab exercises on the subject studied in BDA 606: Statistical Inference and Modeling

BDA 658: Healthcare Informatics Lab

[L-T-P-C: 0-0-3-1]

Lab exercises on the subject studied in BDA 608: Designing with Microcontrollers

BDA 660: Elective – II Lab

[L-T-P-C: 0-0-3-1]

Lab exercises on the subject studied in elective-II

BDA 696: Mini Project – II

[L-T-P-C: 0-0-0-4]

A single semester mini-project. This will be offered by the faculty in consultation with industry partners and mentors. Students have to survey the required background and develop software meeting the requirements.

BDA 698: Seminar – II

[L-T-P-C: 0-0-0-1]

Students select a topic of interest relevant to the field of study. They prepare a short report (not exceeding five pages) and present their understanding in a classroom.