BVM ENGINEERING COLLEGE (AN AUTONOMOUS INSTITUTION)

INFORMATION TECHNOLOGY DEPARTMENT

B. TECH. (INFORMATION TECHNOLOGY) HONOURS DEGREE*

Sr. No.	Course Code & Course Title	L	T	P	H	C
1	**HIT01 : INTRODUCTION TO DATA SCIENCE**	3	0	2	5	4
2	**HIT02 : DATA STRUCTURES AND ALGORITHMS USING PYTHON**	2	0	2	4	3
3	Program Elective I	2	0	2	4	3
4	Program Elective II	3	0	2	5	4
5	**HIT91: PROJECT**	0	0	12	12	6
	Total	**10**	**0**	**20**	**30**	**20**
	Program Elective - I					
1	**HIT11: DATA SCIENCE FOR ENGINEERS**	2	0	2	4	3
2	**HIT12 : INTRODUCTION TO DATA ANALYTICS**	2	0	2	4	3
3	**HIT13: PROBABILITY FOR COMPUTER SCIENCE**	2	0	2	4	3
4	**HIT16: SCALABLE DATA SCIENCE**	2	0	2	4	3
	Program Elective - II					
1	**HIT14: BUSINESS ANALYTICS AND DATA MINING MODELING USING R**	3	0	2	5	4
2	**HIT15 : DATA ANALYTICS WITH PYTHON**	3	0	2	5	4
3	**HIT17: REINFORCEMENT LEARNING**	3	0	2	5	4
4	**HIT18: NATURAL LANGUAGE PROCESSING**	3	0	2	5	4

* A student of B. Tech. Information Technology will be eligible to get B. Tech. Degree with Honours, if he/she gets additional Credits as per above structure.

L=Lecture Hrs./wk; T=Tutorial Hrs./wk; P=Practical Hrs./wk; H=Total Contact Hrs./wk; C=Credits of Course

HIT01 : INTRODUCTION TO DATA SCIENCE
CREDITS – 4 (LTP: 3,0,1)

Course Objective:
To understand and analyze the data for making quicker and better decisions.

Teaching and Assessment Scheme:

Teaching Scheme (Hours per Week)			Credits	Assessment Scheme				Total Marks
				Theory Marks		Practical Marks		
L	T	P	C	ESE	CE	ESE	CE	
3	0	2	4	60	40	20	30	150

Course Contents:

Unit No.	Topics	Teaching Hours
1	**Introduction:** Introduction to Data Science, Data Sources, Challenges, Applications, Introduction to Data Modeling, Statistical Data Modeling, Computational Data Modeling, Statistical Limits on Data: Bonferroni's Principle, Case Studies.	6
2	**Data Gathering and Preprocessing:** Structured and Unstructured data for Data Gathering and Preprocessing, Types, Attributes, Data Cleaning, Data Integration, Data Reduction, Transformation, and Discretization.	6
3	**Exploratory Data Analysis:** Descriptive And Inferential Statistics, Chart Types, Single Variable: Dot Plot, Jitter Plot, Error Bar Plot, Box-And-Whisker Plot, Histogram, Kernel Density Estimate, Cumulative Distribution Function, Two Variable: Bar Chart, Scatter Plot, Line Plot, Log-Log Plot, More than Two Variables: Stacked Plots, Parallel Coordinate Plot, Mean, Variance.	8
4	**Data Modeling:** Basic of Data Model, Function Approximation, Hypothesis Representation, Objective / Loss Function, Linear Regression, Logistic Regression, Gradient Descent.	10
5	**Similarity Measures, Distance Measures and Frequent Item sets:** Feature Extraction: TF, IDF, TF-IDF, Hash Functions, Similarity Measuring Techniques: Shingling, Min-Hashing, Locality Sensitive Hashing, Distance Measures: Triangle Inequality, Euclidean Distance, Cosine Distance, Jaccard Distance, Edit Distance Measures, Frequent Itemsets, The Marketbasket Model, Association Rules, A-Priori Algorithm, PCY (Park-Chen-Yu) Algorithm.	10
6	**Data Streams:** Stream Data Model, Stream Sources, Stream Queries, Issues In Stream Processing, Sampling Data In A Stream, Stream Filtering- Bloom Filter.	5
	Total	**45**

List of References:

1. Jure Leskovec, Anand Rajaraman, Jeffery David Ullman, *"Mining of Massive Datasets"*, Second edition, Cambridge University Press.
2. Avrim Blum, John Hopcroft, Ravindran Kannan, *"Foundations of Data Science"*, Hindustan Book Agency.
3. Ethem Alpaydin*, "Introduction to Machine Learning*", Third edition, PHI Learning Pvt Ltd.

Course Outcomes (COs):

At the end of this course students will be able to …

1. Describe the various areas where data science is applied.
2. Classify and recognize different types of data.
3. Analyze problems in structured framework.
4. Determine appropriate data analysis techniques for problem at hand.
5. Identify visualization for the data analysis problem.
6. Analyze different types of data for inferring meaning.

HIT02 : DATA STRUCTURES AND ALGORITHMS USING PYTHON
CREDITS – 3 (LTP: 2,0,1)

Course Objective:

To understand, design and analyze efficient algorithm for various applications using Python programming.

Teaching and Assessment Scheme:

Teaching Scheme (Hours per Week)			Credits	Assessment Scheme				Total Marks
				Theory Marks		Practical Marks		
L	T	P	C	ESE	CE	ESE	CE	
2	0	2	3	60	40	20	30	150

Course Contents:

Unit No.	Topics	Teaching Hours
1	**Introduction:** Algorithms and Programming: Simple gcd, Improving Naïve gcd, Euclid's Algorithm for gcd, Downloading and Installing Python.	2
2	**Basics of Python:** Types, Expressions, Strings, Lists, Control Flow, Tuples, Python Memory Model: Names, Mutable and Immutable Values, List Operations: Slices etc., Binary Search, Inductive Function Definitions: Numerical and Structural Induction, Elementary Inductive Sorting: Selection and Insertion Sort, In-Place Sorting.	5
3	**Analysis of Algorithms:** Basics of Algorithmic Analysis: Input Size, Asymptotic, Complexity, O() notation, Arrays vs. lists, Merge Sort, Quick Sort, Stable Sorting.	4

Unit No.	Topics	Teaching Hours
4	**Functions:** Dictionaries, More on Python Functions: Optional Arguments, Default values, Passing Functions as Arguments, Higher Order Functions on Lists: map, filter, list comprehension.	3
5	**Exception Handling, File Handling, Input/Output, String Processing :** Exception Handling, Standard Input/Output, Handling Files, String Functions, Formatting printed output, pass, del() and None.	4
6	**Backtracking, Scope, Data structures:** Backtracking: N Queens, Scope in Python: local, global, Nested functions, Generating Permutations, Sets, Stacks, Queues, Priority Queues, Heaps.	5
7	**Classes, objects and user defined data types:** Abstract Data types, Classes and Objects in Python, User Defined Lists, Search Trees.	3
8	**Dynamic programming, wrap-up:** Memoization and Dynamic Programming, Grid Paths, Longest Common Subsequence, Matrix Multiplication, Wrap-Up, Python vs. other languages.	4
	Total	**30**

List of References:

1. Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest and Clifford Stein, *"Introduction to Algorithms",* Third Edition, PHI Publication.
2. Gills Brassard, Paul Bratley, *"Fundamental of Algorithms",* Second Edition, PHI Publication
3. Dave and Dave, *"Design and Analysis of Algorithms"*, Second Edition, Pearson Publication.
4. Anany Levitin, *"Introduction to Design and Analysis of Algorithms",* Third Edition, Pearson Publication.
5. R. Nageswara Rao, *"Core Python Programming",* Paperback, Third Edition, Dreamtech Press.
6. Yashavant Kanetkar, Aditya Kanetkar, *"Let Us Python"*, Third Edition, BPB Publication.

Course Outcomes (COs):

At the end of this course students will be able to …

1. Evaluate the asymptotic performance of sorting and searching algorithms.
2. Understand the basic operations, lists and object-oriented concepts of python programming.
3. Apply the functions of exception handling, file handling, input-output operations and string processing.
4. Solve the problems using dynamic programming techniques.
5. Apply python programming to various data structures.

HIT11: DATA SCIENCE FOR ENGINEERS
CREDITS – 3 (LTP: 2,0,1)

Course Objective:

To introduce R as programming language, mathematical foundation required for data science and first level data science algorithms.

Teaching and Assessment Scheme:

Teaching Scheme (Hours per Week)			Credits	Assessment Scheme				Total Marks
				Theory Marks		Practical Marks		
L	T	P	C	ESE	CE	ESE	CE	150
2	0	2	3	60	40	20	30	

Course Contents:

Unit No.	Topics	Teaching Hours
1	**Introduction to R:** Introduction to R, Variables and Data types in R, Data Frames, Recasting and Joining of Data Frames, Arithmetic, Logical and Matrix Operations in R, Advanced Programming in R: Functions, Control Structures, Data Visualization in R, Basic Graphics.	5
2	**Linear Algebra for Data Science:** Solving Linear Equations, Linear Algebra: Distance, Hyperplanes, and Halfspaces, Eigen values, Eigen vectors.	3
3	**Statistics:** Statistical Modelling, Notion of Probability, Distributions, Mean, Variance, Covariance, Covariance Matrix, Univariate and Multivariate Normal Distributions, Introduction to Hypothesis Testing, Confidence Interval For Estimates.	4
4	**Optimization for Data Science:** Concepts of Optimization, Unconstrained Multivariate Optimization, Gradient (Steepest) Descent (OR) Learning rule.	3
5	**Typology of Data Science Problems and A Solution Framework:** Multivariate Optimization with Equality Constraints, Multivariate Optimization with Inequality Constraints, Introduction to Data Science, Solving Data Analytics Problems.	4
6	**Linear Regression:** Predictive Modelling, Simple Linear Regression and Verifying Assumptions Used in Linear Regression, Model Assessment, Diagnostics to Improve Model Fit, Simple Linear Regression Model Building, Multivariate Linear Regression, Assessing Importance of different variables, Subset Selection.	5
7	**Classification using Logistic Regression:** Cross Validation, Multiple Linear Regression: Model Building and Selection, Classification, Logistic Regression, Performance measures, Logistic Regression implementation in R.	4
8	**Classification using KNN and K-means Clustering:** K-Nearest Neighbors (kNN), K-Means clustering.	2
	Total	**30**

List of References:

1. Gilbert Strang, *"Introduction to Linear Algebra"*, Fifth Edition.
2. Douglas C. Montgomery, George C. Runger *"Applied Statistics and Probability for Engineers"*, Sixth Edition, John Wiley & sons.

Course Outcomes (COs):

At the end of this course students will be able to …

1. Understand R Programming for applications development.
2. Solve the problems related to linear algebra and statistical methods.
3. Explain the optimization techniques and constraints used in data science.
4. Understand the models of linear regression and logistic regression.
5. Understand the algorithms of classification and clustering.

HIT12 : INTRODUCTION TO DATA ANALYTICS
CREDITS – 3 (LTP: 2,0,1)

Course Objective:

To understand the science of analyzing data to convert information to useful knowledge.

Teaching and Assessment Scheme:

Teaching Scheme (Hours per Week)			Credits	Assessment Scheme				Total Marks
				Theory Marks		Practical Marks		
L	T	P	C	ESE	CE	ESE	CE	
2	0	2	3	60	40	20	30	150

Course Contents:

Unit No.	Topics	Teaching Hours
1	**Descriptive Statistics:** Introduction, Descriptive Statistics, Probability Distributions.	4
2	**Inferential Statistics:** Inferential statistics through hypothesis tests, Permutation & Randomization test.	3
3	**Regression & ANOVA:** Type 1 and Type 2 errors, Confidence Intervals, ANOVA (Analysis of Variance), Test of Independence, Regression.	3
4	**Machine Learning : Introduction and Concepts** Introduction to Machine learning, Supervised learning, Unsupervised learning, Differentiating algorithmic and model based frameworks, Regression - Ordinary Least Squares, Ridge Regression, Lasso Regression, Regularization/Coefficients Shrinkage, Data modelling and algorithmic modelling approaches, K Nearest Neighbors Regression & Classification.	5
5	**Supervised Learning with Regression and Classification techniques-1:** Model Validation approaches, Logistic Regression, Linear Discriminant Analysis, Quadratic Discriminant Analysis, Regression & Classification trees, Bias-Variance Dichotomy, Support Vector Machines.	4
6	**Supervised Learning with Regression and Classification techniques-2:** Ensemble Methods and Random Forest, Artificial Neural Networks, Deep Learning.	3
7	**Unsupervised Learning and Challenges for Big Data Analytics:** Clustering, Associative Rule Mining, Hadoop, HDFS Eco System, Spark, NoSQL, Challenges for Big data analytics.	4

Unit No.	Topics	Teaching Hours
8	**Prescriptive Analytics:** Creating data for Analytics through Designed experiments, Active learning and Reinforcement Learning.	4
	Total	**30**

List of References:

1. Hastie, Trevor, et. al; *"The Elements of Statistical Learning",* Second Edition, Springer.
2. Douglas C. Montgomery, George C. Runger, *"Applied Statistics and Probability for Engineers",* Seventh Edition, John Wiley & sons.

Course Outcomes (COs):

At the end of this course students will be able to …

1. Understand the techniques related to differential statistics and inferential statistics.
2. Understand the concepts of machine learning.
3. Analyze the algorithms of supervised learning and unsupervised learning.
4. Apply the methods of regression and classification to solve real time problems.
5. Learn about creating the dataset using prescriptive analysis.

HIT13: PROBABILITY FOR COMPUTER SCIENCE
CREDITS – 3 (LTP: 2,0,1)

Course Objective:

To introduce the concepts of probability and exhibit its applications in computer science and algorithms.

Teaching and Assessment Scheme:

Teaching Scheme (Hours per Week)			Credits	Assessment Scheme				Total Marks
				Theory Marks		Practical Marks		
L	T	P	C	ESE	CE	ESE	CE	150
2	0	2	3	60	40	20	30	

Course Contents:

Unit No.	Topics	Teaching Hours
1	**Introduction to Probability:** Introductory examples of Probability, Probability over Discrete space, Inclusion-Exclusion Principle.	3
2	**Sigma Algebra and Conditional Probability:** Probability over Infinite Space, Conditional Probability, Partition Formula, Independent Events, Bayes Theorem, Fallacies, Random Variables.	3
3	**Famous Random Variables:** Random Variables and Expectations, Independent Random Variables, Conditional Distribution and Expectations, Partition Formula, Linearity of Expectation, Discrete Random Variables: Bernoulli Random Variable, Binomial Random Variable, Geometric Random Variable, Negative	5

Unit No.	Topics	Teaching Hours
	Binomial Random Variable, Continuous Random Variables: Exponential Random Variable, Normal/Gaussian Random Variable, Central Limit Theorem.	
4	**Concentration Inequalities, Boosting by Chernoff:** Equality Checking Protocol, Poison Random Variable, Concentration Inequalities, Markov Inequality, Variance, Weak Linearity of Variance, Law of Large Numbers, Chertoff's Bound, K-wise Independence.	4
5	**Stochastic Process:** Stochastic Process, Markov Chains, Drunkard's Walk, Evolution of Markov Chains.	4
6	**Stationary Distribution:** Perron-Frobenius Theorem, Page Rank Algorithm, Ergodicity, Cell Genetics, Random Sampling.	4
7	**Probabilistic Methods:** Biased-Coin and Hashing, Introduction to Probabilistic Methods, Ramsey Numbers, Large cuts in Graphs, Sum Free Subsets, Discrepancy.	4
8	**Streaming Algorithms:** Extremal Set Families, Super Concentrators, Streaming Algorithms.	3
	Total	**30**

List of References:

1. James L. Johnson, "*Probability and Statistics for Computer Science",* Wiley Publication.
2. P. G. Hoel, S. C. Port, C. J. Stone, "*Introduction to Probability Theory*", Universal Book Stall.
3. S. Ross, *"A First Course in Probability"*, Sixth Edition, Pearson Education India.
4. W. Feller, *"An Introduction to Probability*" Theory and its Applications, Vol. 1, Wiley.
5. D. C. Montgomery, G. C. Runger, *"Applied Statistics and Probability for Engineers*", Wiley.
6. J. L. Devore, "*Probability and Statistics for Engineering and the Sciences"*, Cengage Learning.

Course Outcomes (COs):

At the end of this course students will be able to …

1. Understand the probability concepts related to the computer science and the scientific programming.
2. Apply the methods of discrete and continuous random variables for solving problems related to probability distribution functions.
3. Apply the methods to check the inequalities and boost them.
4. Apply stationary distribution techniques and probabilistic methods to solve real life problems.
5. Understand the working of various streaming algorithms.

HIT14: BUSINESS ANALYTICS AND DATA MINING MODELING USING R
CREDITS – 4 (LTP: 3,0,1)

Course Objective:

To impart knowledge on use of data mining techniques for deriving business intelligence to achieve organizational goals using R programming.

Teaching and Assessment Scheme:

Teaching Scheme (Hours per Week)			Credits	Assessment Scheme				Total Marks
				Theory Marks		Practical Marks		
L	T	P	C	ESE	CE	ESE	CE	150
3	0	2	4	60	40	20	30	

Course Contents:

Unit No.	Topics	Teaching Hours
1	**General Overview of Data Mining and its Components:** Introduction, Data mining process, Introduction to R, Basic statistics: Hypothesis Testing and its techniques, Confidence Interval, Type I and Type II Errors, Power of a test, ANOVA.	4
2	**Data Preparation and Visualization Techniques:** Partitioning, Types of Data sets, Model Building. Visualization techniques: Data Exploration and Conditioning, Line Chart or Graphs, Bar Charts, Scatter Plot, Basic Charts, Distribution Plots, Boxplot, Histogram, Heatmaps, Multidimensional Visualization, In Plot Labels, Specialized Visualization Techniques: Network Graph, Treemaps, Map Charts.	5
3	**Dimension Reduction and Principal Component Analysis:** Introduction to Dimension Reduction, Dimension Reduction Techniques: Domain Knowledge, Data Exploration Techniques, Data Conversion Techniques, Automated Reduction Techniques, Data Mining Techniques, Principal Component Analysis.	5
4	**Performance Metrics and Assessment of Performance Metrics for Prediction and Classification:** Introduction to Performance Metrics, Performance Metrics Based on Naïve Rule, Cutoff Probability Value, Classification Performance, ROC, Cumulative Lift Curve, Asymmetric Misclassification Cost, Oversampling, Rare Class Scenario, Prediction Performance: Prediction Error, Predictive Accuracy Measures.	5
5	**Multiple Linear Regression:** Introduction to Multiple Linear Regression, Objectives, Explanatory Modeling vs. Predictive Modeling, Estimation Technique: Ordinary Least Squares, Variable Selection, Bias-Variance trade-off, Exhaustive Search, Partial-iterative Search, Machine Learning Techniques: k-NN, Naive bayes, Case study.	7
6	**Classification & Regression:** CART, Classification Trees, Recursive Partitioning, Impurity Measures, Gini Index, Entropy Measure, Tree Structure, Classification Trees, Pruning process, Regression trees, Case study.	5
7	**Logistic Regression:** Introduction to Logistic Regression, Logistic Regression Model, Logit, Odds and odds ratio, Logistic Regression for Profiling Task, Case study.	5

Unit No.	Topics	Teaching Hours
8	**Artificial Neural Network**: Introduction to Artificial Neural Networks, Neural Network Architectures, Multilayer Feedforward Networks, Neural Network Training Process, Computing Output Values at Nodes of each Layer type, Linear and Logistic Regression as special cases, Normalization, Estimation Method, Back Propagation, Methods for updating weight and bias-values, Case updating vs. Batch Updating, Stopping criteria for updating, Overfitting issues in Neural Network, Experimenting with Neural Network Models.	6
9	**Discriminant Analysis:** Introduction, Statistical Techniques, Linear Classification Functions, Assumptions and other issues.	3
	Total	**45**

List of References:

1. *"Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data",* EMC Education Services.
2. Shmueli, G., Patel, N. R., Bruce, P. C., *"Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner",* Second Edition, Wiley.

Course Outcomes (COs):

At the end of this course students will be able to …

1. Understand the concepts of data mining and statistical analysis through R programming.
2. Apply the concepts of creating visualizations using R programming language.
3. Evaluate performance metrics for classification techniques.
4. Understand the concepts of regression and its variants.
5. Apply training methods of neural networks for building models.
6. Build, assess and compare models based on real datasets and cases.

HIT15 : DATA ANALYTICS WITH PYTHON
CREDITS – 4(LTP: 3,0,1)

Course Objective:

To understand the use of data analytics and models.

Teaching and Assessment Scheme:

Teaching Scheme (Hours per Week)			Credits	Assessment Scheme				Total Marks
				Theory Marks		Practical Marks		
L	T	P	C	ESE	CE	ESE	CE	150
3	0	2	4	60	40	20	30	

Course Contents:

Unit No.	Topics	Teaching Hours
1	**Introduction to Data Analytics and Python fundamentals:** Introduction to Data Analytics, Python Fundamentals, Central Tendency and Dispersion.	**5**
2	**Introduction to Probability:** Introduction to Probability, Probability Distributions.	**3**
3	**Sampling and Sampling Distributions:** Python Demo for Distributions, Sampling and Sampling Distribution, Distribution of Sample Means, Population and Variance, Confidence Interval Estimation: Single population.	**4**
4	**Hypothesis Testing:** Introducing Hypothesis Testing, Errors in Hypothesis Testing, Hypothesis Testing: Two sample test	**4**
5	**Two sample test and introduction to ANOVA:** ANOVA, Post Hoc Analysis (Tukeya€^{TM}s test), Randomize Block Design (RBD), Two Way ANOVA.	**4**
6	**Linear Regression and Multiple Regression:** Estimation, Prediction of Regression Model Residual Analysis, Multiple Regression Model, Categorical Variable Regression	**5**
7	**Concepts of MLE and Logistic Regression:** Maximum Likelihood Estimation, Logistic Regression, Linear Regression Model vs. Logistic Regression Model.	**5**
8	**ROC and Regression Analysis Model Building:** Confusion Matrix and ROC, Performance of Logistic Model, Regression Analysis Model Building.	**4**
9	**C^2 Test:** Chi-Square Test of Independence, Chi-Square Goodness of Fit Test.	**3**
10	**Clustering analysis:** Introduction to Clustering Analysis, K-Means Clustering, Hierarchical method of Clustering.	**3**
11	**Classification and Regression Trees (CART)** Classification and Regression Trees (CART), Measures of Attribute Selection, Attribute Selection Measures in CART.	**5**
	Total	**45**

List of References:

1. McKinney, W., *"Python for Data Analysis: Data wrangling with Pandas, NumPy, and Python"*, Second Edition, O'Reilly publication.
2. Swaroop, C. H., *"A Byte of Python"*.
3. Ken Black, *"Business Statistics for Contemporary Decision Making",* Sixth Edition, John Wiley & Sons.
4. Anderson Sweeney Williams, *"Statistics for Business and Economics"*, Cengage Learning.
5. Douglas C. Montgomery, George C. Runger, *"Applied Statistics & Probability for Engineering"*, John Wiley & Sons.
6. Jay L. Devore, *"Probability and Statistics for Engineering and the Sciences"*, Cengage Learning.

7. David W. Hosmer, Stanley Lemeshow, *"Applied Logistic Regression (Wiley Series in probability and statistics)"*, Wiley-Inter science Publication.
8. Jiawei Han, Micheline Kamber, *"Data Mining: Concepts and Techniques"*, Third Edition, Morgan Kaufmann Publishers.
9. Leonard Kaufman, Peter J. Rousseeuw, *"Finding Groups in Data: An Introduction to Cluster Analysis"*, John Wiley & Sons.

Course Outcomes (COs):

At the end of this course students will be able to …

1. Understand the basics of data analytics.
2. Apply python programming for solving real life problems.
3. Apply probability and sampling methods to different data sets.
4. Understand the models of linear regression and logistic regression.
5. Apply the techniques of classification and clustering to the probabilistic model.

HIT91: PROJECT
CREDITS - 6 (LTP: 0,0,6)

Course Objective:

To provide exposure in the field of Software/Hardware development to develop real life applications.

Teaching and Assessment Scheme:

Teaching Scheme (Hours per Week)			Credits	Assessment Scheme				Total Marks
				Theory Marks		Practical Marks		
L	T	P	C	ESE	CE	ESE	CE	
0	0	12	6	0	0	120	180	300

Course Conduction Guidelines:

- The project shall be based on Data science/analytics and need of industry/society.
- Students can opt for Industry defined project or User defined project.
- Selection of project definition is subject to approval of departmental academic committee.
- In case of IDP, selection of industry need to be approved by departmental academic committee.
- In case of IDP, the students may be sent to the industry / premier organization for their project during allocated days in respective timetable.
- After approval of project definition, students are required to report their project work on weekly basis to the respective internal guide and/or industry guide.
- A project should incorporate all phases of software development, like requirement analysis, feasibility study, project design, implementation, testing and validation.
- Project will be evaluated in laboratory hours during the semester and final submission will be taken at the end of the semester as a part of continuous evaluation.
- Students have to submit project in CD/DVD with following listed documents at the time of final submission.
    - Project Synopsis
    - Software Requirement Specification
    - Final Project Report
    - Project Setup file with Source code
    - Project Presentation (PPT)

Course Outcomes (COs):

After successful completion of this course student will be able to:

1. Analyze existing systems, thereby select and justify parameters to be improved.
2. Work on proposed engineering solution as per industry / research / societal need.
3. Customize various tools and techniques needed for project development.
4. Understand significance of safe and ethical practices during project.
5. Develop skill to present project related activities effectively to peers, mentors and society.

HIT16: SCALABLE DATA SCIENCE
CREDITS – 3 (LTP: 2,0,1)

Course Objective:

To understand the algorithmic techniques and software paradigms to develop scalable algorithms and systems for the common data science tasks.

Teaching and Assessment Scheme:

Teaching Scheme (Hours per Week)			Credits	Assessment Scheme				Total Marks
				Theory Marks		Practical Marks		
L	T	P	C	ESE	CE	ESE	CE	
2	0	2	3	60	40	20	30	150

Details of Assessment Instruments under CE Practical Component:

30	
Term work	15
Assignment / Quiz / Project	15

Course Contents:

Unit No.	Topics	Teaching Hours
1	**Introduction:** Probability: Concentration inequalities, Linear algebra: PCA, SVD Optimization: Basics, Convex, GD, Machine Learning: Supervised, generalization, feature learning, clustering.	5
2	**Memory Efficient Data Structures:** Memory-efficient Data Structures: Hash functions, Universal / Perfect Hash Families, Bloom filters, Sketches for distinct count, Misra-Gries sketch. Count Sketch, Count-Min Sketch.	6
3	**Approximate Near Neighbors Search:** Introduction, kd-trees, LSH families, MinHash for Jaccard, SimHash for L2, Extensions, Multi-Probe, b-bit hashing, Data dependent variants.	6
4	**Randomized Numerical Linear Algebra:** Random projection, CUR Decomposition, Sparse RP, Subspace RP, Kitchen Sink.	7

Unit No.	Topics	Teaching Hours
5	**Map-Reduce Paradigms:** Map-reduce and related paradigms, Map reduce Programming examples: Page Rank, K-Means, Matrix Multiplication, Big data: Computation goes to Data, Hadoop ecosystem, Scala, Spark.	6
6	**Distributed Machine Learning and Optimization:** Introduction, SGD with Proof, DMM and applications, Clustering.	6
	Total	**36**

List of References:

1. J. Leskovec, A. Rajaraman, JD Ullman, *"Mining of Massive Datasets*", Cambridge University Press, 2nd Edition.
2. Muthukrishnan, S. (2005), *"Data streams: Algorithms and applications",* Foundations and Trends in Theoretical Computer Science, 1(2), 117-236.
3. Woodruff, David P., *"Sketching as a tool for numerical linear algebra"*, Foundations and Trends in Theoretical Computer Science 10.1–2 (2014): 1-157.
4. Mahoney, Michael W., *"Randomized algorithms for matrices and data"*, Foundations and Trends in Machine Learning 3.2 (2011): 123-224.

Course Outcomes (COs):

At the end of this course students will be able to …

1. Understand probabilistic and statistic concepts and procedures.
2. Understand and apply various memory-optimal data structures.
3. Design and apply different Approximate Nearest Neighbor techniques.
4. Solve Linear Algebra using Randomized algorithms.
5. Process Big-Data with Map-Reduce Algorithms.
6. Learn and apply distributed machine learning techniques.

HIT17: REINFORCEMENT LEARNING
CREDITS – 4 (LTP: 3,0,1)

Course Objective:

To understand the concepts of reinforcement learning and various algorithms to solve real world problems.

Teaching and Assessment Scheme:

Teaching Scheme (Hours per Week)			Credits	Assessment Scheme				Total Marks
				Theory Marks		Practical Marks		
L	T	P	C	ESE	CE	ESE	CE	
3	0	2	4	60	40	20	30	150

Details of Assessment Instruments under CE Practical Component:

30	
Term work	15
Assignment / Quiz / Project	15

Course Contents:

Unit No.	Topics	Teaching Hours
1	**Introduction:** Introduction to RL, RL Framework and applications, Introduction to Immediate RL, Bandit Optimality, Value function based methods.	4
2	**Bandit Algorithms:** UCB 1, Concentration Bounds, UCB 1 Theorem, PAC Bounds, Median Elimination, Thompson Sampling.	4
3	**Policy Gradient Methods & Introduction to Full RL:** Policy Search, REINFORCE, Contextual Bandits, Full RL Introduction, Returns, Value Functions and MDPs.	5
4	**MDP Formulation, Bellman Equations & Optimality Proofs**: MDP Modelling, Bellman Equation, Bellman Optimality Equation, Cauchy Sequence and Green's Equation, Banach Fixed Point Theorem, Convergence Proof.	5
5	**Dynamic Programming & Monte Carlo Methods:** Lpi Convergence, Value Iteration, Policy Iteration, Dynamic Programming, Monte Carlo, Control in Monte Carlo.	5
6	**Monte Carlo & Temporal Difference Methods:** Off Policy MC, UCT, TD(0), TD(0) Control, Q-Learning, Aftertaste.	4
7	**Eligibility Traces:** Eligibility Traces, Backward View of Eligibility Traces, Eligibility Trace Control, Thompson Sampling Recap.	4
8	**Function Approximation:** Function Approximation, Linear Parameterization, State Aggregation Methods, Function Approximation and Eligibility Traces, LSTD and LSTDQ, LSPI and Fitted Q.	4
9	**DQN, Fitted Q & Policy Gradient Approaches:** DQN and Fitted Q-Iteration, Policy Gradient Approach, Actor Critic and REINFORCE, Policy Gradient with Function Approximation.	4
10	**Hierarchical Reinforcement Learning and MAXQ:** Hierarchical Reinforcement Learning, Types of Optimality, Semi Markov Decision Processes, Options, Learning with Options, Hierarchical Abstract Machines, MAXQ, MAXQ Value Function Decomposition, Option Discovery.	4
11	**POMDPs:** POMDP Introduction, Solving POMDP.	2
	Total	**45**

List of References:

1. Richard S. Sutton and Andrew G. Barto, *"Reinforcement Learning: An Introduction"*, Second edition, MIT Press.
2. Daniel Jurafsky & James H Martin, *"Speech and Natural Language Processing"*, Pearson Publications.
3. Alberto Leon-Garcia, *"Probability, Statistics, and Random Processes for Electrical*

Engineering", Third edition, PHI Learning.

Course Outcomes (COs):

At the end of this course students will be able to …

1. Understand the concepts of Reinforcement Learning to solve real world problems.
2. Solve problems using Finite Markov Decision process and dynamic program.
3. Apply Monte Carlo, Temporal Difference methods for policy evaluation and prediction.
4. Analyze the Tabular Methods and On-policy Prediction with Approximation.
5. Solve problems using deep reinforcement learning.
6. Recognize current advanced techniques and applications using RL.

HIT18: NATURAL LANGUAGE PROCESSING
CREDITS – 4 (LTP: 3,0,1)

Course Objective:

To understand significance of natural language processing in solving real-world problems.

Teaching and Assessment Scheme:

Teaching Scheme (Hours per Week)			Credits	Assessment Scheme				Total Marks
				Theory Marks		Practical Marks		
L	T	P	C	ESE	CE	ESE	CE	
3	0	2	4	60	40	20	30	150

Details of Assessment Instruments under CE Practical Component:

30	
Term work	15
Assignment / Quiz / Project	15

Course Contents:

Unit No.	Topics	Teaching Hours
1	**Introduction:** Introduction, Uses of NLP, Reason of NLP hard, Empirical Laws, Text Processing Basics.	5
2	**Spelling Correction and Language Modeling:** Spelling Correction: Edit Distance, Weighted Edit Distance, Other Variations, Noisy Channel Model for Spelling Correction. Language Modeling: N-Gram Language Models, Evaluation of Language Models, Basic Smoothing, Advanced Smoothing Models, Computational Morphology, Finite - State Methods for Morphology.	7
3	**POS tagging:**	6

Unit No.	Topics	Teaching Hours
	Introduction to POS Tagging, Hidden Markov Models for POS Tagging, Viterbi Decoding for HMM, Parameter Learning, Baum Welch Algorithm, Maximum Entropy Models, Conditional Random Fields.	
4	**Constituency and Dependency Parsing:** Introduction, Parsing I, CKY, PCFGs, Inside-Outside Probabilities, Introduction to Dependency Grammars and Parsing, Transition Based Parsing, MST-Based Dependency Parsing.	6
5	**Distributional Semantics:** Introduction, Distributional Models of Semantics, Applications and Structured Models of Distributional Semantics, Word Embeddings.	5
6	**Lexical Semantics:** Introduction, Wordnet, Word Sense Disambiguation, Novel Word Sense detection, Introduction to Topic Models, Formulation of Latent Dirichlet Allocation, Gibbs Sampling for LDA and Applications, LDA Variants and Applications.	6
7	**Applications of text mining:** Entity Linking, Introduction to Information Extraction, Relation Extraction, Distant Supervision, Text Summarization, LEXRANK, Optimization based Approaches for Summarization, Summarization Evaluation, Text Classification, Introduction to Sentiment Analysis, Affective Lexicons, Learning and Computing with Affective Lexicons, Aspect - Based Sentiment Analysis.	10
	Total	**45**

List of References:

1. Dan Jurafsky, James Martin. *"Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition",* Prentice Hall, Second Edition, 2009.
2. Chris Manning, Hinrich Schütze. "*Foundations of Statistical Natural Language Processing*", MIT Press, Cambridge, MA: May 1999.
3. Charu C. Aggarwal, *"Machine Learning for Text",* Springer, 2018 edition.
4. Steven Bird, Ewan Klein and Edward Loper, *"Natural Language Processing with Python",* O'Reilly Media. Edition, 2009.
5. Roland R. Hausser, *"Foundations of Computational Linguistics: Human Computer Communication in Natural Language"*, Paperback, MIT press, 2011.

Course Outcomes (COs):

At the end of this course students will be able to …

1. Understand basic concepts of Natural Language Processing.
2. Analyze the spelling correction techniques and language modeling.
3. Explain techniques of constituency parsing and dependency parsing.
4. Apply the models for speech tagging.
5. Use lexical semantics models for various applications.
6. Understand the applications of text mining.