**SYLLABUS**

**FACULTY OF ENGINEERING AND TECHNOLOGY**

**Master of Science (Data Science & Big Data Analytics)**

**M.Sc. (DS & BDA)**

**SCHOOL OF ENGINEERING AND TECHNOLOGY**

**DEPARTMENT OF COMPUTER SCIENCE & APPLICATION**

**Batch: 2023-2025**

**W.E.F A.Y 2023-2024**

**COURSE STRUCTURE**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Course Code** |  | | | |
| **Course Category** | **Program Core** | | | |
| **Course Title** | **Data Warehousing & Data Mining** | | | |
| **Teaching Scheme and Credits**  **Weekly load hrs.** | **L** | **T** | **Laboratory** | **Credits** |
| **3** | **1** | **0** | **4** |
| **Pre-requisites**:   1. Understanding of Relational database normalization techniques, Physical design of a database, Concepts of algorithm design and analysis. 2. Basic understanding of: Software engineering principles and techniques, Probability, and Statistics – Bayesian theory, Regression, Hypothesis Testing | | | | |
| **Course Objectives:**   1. To understand the structure of Data Warehouse 2. To understand different data pre-processing techniques. 3. To understand basic descriptive and predictive data mining techniques. 4. To use data mining tool on different data sets 5. To understand basics of Classification algorithms 6. To understand basics of Prediction algorithms. 7. To understand basics of Clustering algorithms. | | | | |
| **Course Outcomes:**  On completion of the course:   1. The student will get knowledge of data processing and data quality. 2. The student will be able to model and design data warehouses. 3. The student will get knowledge of basic and advanced concepts of algorithms for data mining. 4. The student will be able to understand the complete life-cycle of data mining process and apply by using data mining tools used in industry. | | | | |
| **Course Contents:**   1. **Introduction to Data Mining**   Basic concepts of data mining, Types of Data to be mined.   1. **Introduction to Data Warehouse**   Data Warehouse and DBMS, Architecture of Data Warehouse   1. **Data pre-processing**   Need Data pre-processing, Attributes and Data types   1. **Data Mining Techniques: Association Rule Mining**   Basic idea: item sets, Frequent Item-sets   1. **Data Mining Techniques: Classification**   Definition of Classification, Decision tree Induction: Information gain, gain ratio, Gini Index   1. **Data Mining Techniques: Prediction and Clustering**   Regression and Introduction to clustering | | | | |
| **Learning Resources:**  **Textbooks:**   1. Kamber, M., Han, J., Pei, J. (2011). Data Mining: Concepts and Techniques, Elsevier Science. 2. Margaret H. Dunham, S. Sridhar, Data Mining: Introductory and Advanced Topics. (2006). India: Pearson Education. 3. Data warehousing: fundamentals fot IT professionals 3rd edition, Kimball, Wiley Publication   **Reference Books:**   1. Ian H.Witten, Eibe Frank Data Mining: Practical Machine Learning Tools and Techniques, Elsevier/(Morgan Kauffman). 2. Introduction to Data Mining (2005) By Pang-Ning Tan, Michael Steinbach, Vipin Kumar Addison Wesley 3. [Research-Papers]: Some of the relevant research papers that contain recent results and developments in the data mining field.   **Additional Reading:**  **Websites :** https: //www.tutorialspoint.com  **MOOCS:** Coursera, Udemy, Udacity, Simplilearn, Edx | | | | |
| **Pedagogy:**  Participative learning, discussions, algorithm, demo using latest tools,experiential learning through practical problem solving, assignment, PowerPoint presentations and Case Study. | | | | |
| **Assessment Scheme:**  Class Continuous Assessment (CCA) 60 marks   |  |  |  |  | | --- | --- | --- | --- | | Mid Term Examination  (MCQ/Online Test /Direct Internal Examination) | FAT 1 (Formative Assessment Test 1)  (Assignment, MCQ, Class  Test, Presentation,  Seminar) | FAT 2 (Formative Assessment Test 2)  (Assignment, MCQ, Class  Test, Presentation,  Seminar) | Total | | 20 Marks | 20 Marks | 20 Marks | 60 Marks |   **Term End Examination: 40 Marks** | | | | |

**Syllabus:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Module | Contents | Work load in hrs. | | |
| Theory | Lab | Access |
| 1 | **Introduction to Data Mining:** Basic concepts of data mining, Types of Data to be mined, Stages of the Data Mining Process, Data Mining Techniques, Knowledge Discovery in Databases, Data Mining Issues, Applications of Data. | 8 | - | **14**  **Marks** |
| 2 | **Introduction to Data Warehouse:** Data Warehouse and DBMS Architecture of Data Warehouse, Multidimensional data model, Concepts of OLAP and Data Cube, OLAP operations, Dimensional Data Modelling- Star, Snowflake schemas. | 10 | - | **16**  **Marks** |
| 3 | **Data pre-processing:** Need Data pre-processing, Attributes and Data types, Statistical descriptions of Data, handling missing Data, Data sampling, Data cleaning, Data Integration and transformation, Data reduction-– Curse of Dimensionality, Feature Selection and Feature Engineering, Principal Component Analysis (PCA), Discretization and generating concept hierarchies. | 10 | - | **16**  **Marks** |
| 4 | **Data Mining Techniques: Association Rule Mining:** Basic idea: item sets, Frequent Item-sets, Association Rule Mining, Generating item sets and rules efficiently, FP growth algorithm. | 8 | - | **14**  **Marks** |
| 5 | **Data Mining Techniques: Classification:** Definition of Classification, Decision tree Induction: Information gain, gain ratio, Gini Index, Issues: Over-fitting, tree pruning methods, missing values, continuous classes, Bayesian Classification: Bayes Theorem, Naïve Bayes classifier. | 12 | - | **20**  **Marks** |
| 6 | **Data Mining Techniques: Prediction**: Definition of Prediction, Parametric and Non-Parametric algorithms, Linear Regression Algorithm, Linear Regression Model, OLS, Derivation of Beta coefficients for OLS, OLS Cost function, RMSE, R-Squared Error, Linear Regression Assumptions. Non-linear regression, logistic regression.  **Data Mining Techniques: Introduction to Clustering:** Definition of Clustering, Partitioning Methods, Hierarchical Methods, Distance Measures in Algorithmic Methods, Density Based Clustering | 12 | - | **20**  **Marks** |

**COURSE STRUCTURE**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Course Code** |  | | | |
| **Course Category** | **Program Core** | | | |
| **Course Title** | **Mathematics for Machine Learning** | | | |
| **Teaching Scheme and Credits**  **Weekly load hrs.** | **L** | **T** | **Laboratory** | **Credits** |
| **3** | **0** | **0** | **3** |
| **Pre-requisites**:   1. Basic problem-solving skills 2. Basic Mathematics | | | | |
| **Course Objectives:**   1. To enable students to represent real life problems using mathematical notations 2. To provide a strong foundation for Machine Learning subjects 3. To enable students to understand the notations used in Machine Learning / Deep Learning algorithms/techniques | | | | |
| **Course Outcomes:**  On completion of the course,   1. The students will be able to understand the concept of linear algebraic operations, vectors and matrices and how to apply them in machine learning. 2. The students will have a basic understanding of Calculus | | | | |
| **Course Contents:**   1. **Introduction to Linear Algebra** 2. **Linear Maps and Matrix Representation of Systems** 3. **Linear Dependence and Norms** 4. **Matrix Decomposition** 5. **Introduction to Calculus** **and Multivariate calculus** 6. **Principal Component Analysis.** | | | | |
| **Learning Resources:**  **Textbooks:**   1. Deisenroth, Marc P., Faisal, A. Aldo and Ong, Cheng Soon. Mathematics for Machine Learning. : Cambridge University Press, 2020.   **Reference Books:**   1. Linear Algebra and Optimization for Machine Learning, [Charu C. Aggarwal](https://www.google.co.in/search?tbo=p&tbm=bks&q=inauthor:%22Charu+C.+Aggarwal%22&source=gbs_metadata_r&cad=7), Springer Nature, 2020 2. Trapp, R. (2019). Multivariable Calculus. United Kingdom: OUP Oxford.   **Additional Reading:**  **Websites/ MOOCs:**   1. <https://www.coursera.org/specializations/mathematics-machine-learning> 2. <https://www.coursera.org/learn/multivariate-calculus-machine-learning> 3. <https://ocw.mit.edu/courses/mathematics/18-100c-real-analysis-fall-2012/> 4. https://tutorial.math.lamar.edu/Classes/CalcIII/IteratedIntegrals.aspx   **Pedagogy:** Participative learning, discussions, algorithm, Experiential learning through practical problem solving, Assignments, PowerPoint presentation. | | | | |
| **Assessment Scheme:**  Class Continuous Assessment (CCA) 60 marks   |  |  |  |  | | --- | --- | --- | --- | | Mid Term Examination  (MCQ/Online Test /Direct Internal Examination) | FAT 1 (Formative Assessment Test 1)  (Assignment, MCQ, Class  Test, Presentation,  Seminar) | FAT 2 (Formative Assessment Test 2)  (Assignment, MCQ, Class  Test, Presentation,  Seminar) | Total | | 20 Marks | 20 Marks | 20 Marks | 60 Marks |   **Term End Examination : 40 Marks** | | | | |

**Syllabus:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Module | Contents | Work load in hrs. | | |
| Theory | Lab | Assess |
| 1 | **Introduction to Linear Algebra**  Motivation for Linear Algebra, Significance of Linear Algebra in ML, Mathematical object types - scalars, vectors, matrices and tensors.  Matrix Types - Identity matrix, Inverse matrix, Singular matrix, Triangular matrix, Basic Matrix Operations, Matrix transpose, Trace of a matrix, Determinant and it properties | 6 | **-** | **15**  **Marks** |
| 2 | **Linear Maps and Matrix Representation of Systems**  Linear maps, Dot product and Cross Product of vectors and their properties, Scalar and Vector Projections, Matrix-vector and matrix-matrix multiplication, Hadamard Product. Represent a system of linear equations using matrix notation. | 6 | **-** | **15**  **Marks** |
| 3 | **Linear Span, Linear Dependence, Vector Spaces and Norms**  Linear combination, Span, Linear Dependence and Linear Independence of vectors, Row Figure, Column Figure, Vector Spaces and sub-spaces, Vector norms, role of norms in machine learning, Different kind of norms (L0, L1, L2, Linf) , Examples of different kinds of norms. | 8 | **-** | **20**  **Marks** |
| 4 | **Matrix Decomposition**  Linear transformation of vectors, Eigen vectors, Eigen values, Eigen Decomposition of matrix, Spectral Theorem, Singular Value Decomposition. | 10 | **-** | **20**  **Marks** |
| 5 | **Introduction to Calculus and Multivariate calculus**  Calculus, Significance of calculus in Machine Learning, Functions, Derivative of a function, graphical representation of a function and its derivative, Examples. Multivariate calculus, The Jacobian, The Hessian. Multivariate chain rule, convex functions in multidimensional space, Limits, Partial Derivatives, Interpretations of partial derivatives, High order partial derivatives. | 8 | **-** | **16**  **Marks** |
| 6 | **Principal Component Analysis**  Dimensionality Reduction, Deriving Orthogonal Projections, Lower Dimensional, Subspaces, Characterization through Singular Value Decomposition and Eigenvalue Analysis, Rayleigh Quotient, Kernel PCA, Functional PCA. | 7 | **-** | **14**  **Marks** |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  |  |  |  |

**COURSE STRUCTURE**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Course Code** |  | | | |
| **Course Category** | **Program Core** | | | |
| **Course Title** | **Statistics & Optimization for Machine Learning** | | | |
| **Teaching Scheme and Credits**  **Weekly load hrs** | **L** | **T** | **Laboratory** | **Credits** |
| **3** | **0** | **0** | **3** |
| **Pre-requisites**: Basics of Probability and Mathematics, Knowledge of Python | | | | |
| **Course Objectives:**   1. To teach the concepts of Statistics relevant to Machine Learning 2. To teach different probability distributions 3. To teach different optimization techniques 4. To teach time series analysis of data | | | | |
| **Course Outcomes:**  On completion of the course, student will get knowledge of:   1. The students will be able to understand and relate statistical concepts like probability theory, random variables, regression, and bootstrapping to various machine learning techniques. 2. The students will understand different loss functions and optimization techniques necessary to improve the accuracy of a model. | | | | |
| **Course Contents**   1. Introduction to Statistics 2. Random Variable 3. Probability Distribution 4. Bootstrapping 5. Optimization | | | | |
| **Learning Resources** :  **Textbook:**   1. Wasserman, L. (2013). All of Statistics: A Concise Course in Statistical Inference. Ukraine: Springer New York.   **Reference Books:**   1. Hastie, T., Tibshirani, R., Friedman, J. (2013). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Germany: Springer New York. 2. Fundamentals of Mathematical Statistics. (2020). United Kingdom: Sultan Chand & Sons.   **Additional Reading:**  **Websites:** https: //www.tutorialspoint.com  **MOOCS:** Coursera, Udemy, Udacity, Simplilearn, Edx | | | | |
| **Pedagogy:**  Participative learning, discussions, algorithm, demo using latest tools,experiential learning through practical problem solving, assignment, PowerPoint presentations and Case Study | | | | |
| **Assessment Scheme:**  Class Continuous Assessment (CCA) 60 marks   |  |  |  |  | | --- | --- | --- | --- | | Mid Term Examination  (MCQ/Online Test /Direct Internal Examination) | FAT 1 (Formative Assessment Test 1)  (Assignment, MCQ, Class  Test, Presentation,  Seminar) | FAT 2 (Formative Assessment Test 2)  (Assignment, MCQ, Class  Test, Presentation,  Seminar) | Total | | 20 Marks | 20 Marks | 20 Marks | 60 Marks |   **Term End Examination: 40 Marks** | | | | |

**Syllabus:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Module  No. | Contents | Workload in Hrs | | |
| Theory | Lab | Assess |
| **1** | **Introduction:**  Basic concepts of Statistics, Measures of central tendency, Relative frequency, Class frequency tables, Frequency histogram  Basics of Probability, Complementary events, independent events, Conditional probability, Bayesian Inference - Frequentist vs Bayesian approach. Example for comparing both the approaches, Bayes Theorem, Measures of Variation – Quartiles and Percentiles, Moments - Skewness and Kurtosis, Correlation. | 10 |  | **20**  **Marks** |
| **2** | **Random Variables:**  Definition of random variable, Continuous and Discrete Random Variables, Distribution Function of a Random Variable,  Probability Mass Functions and Probability Density Functions – cdf, pdf, Characteristic Functions, Central Limit Theorems. | 8 |  | **16**  **Marks** |
| **3** | **Probability Distributions:**  Recurrence Relationships, Moment Generating Functions, Cumulant Generating Functions, Continuous Probability Distributions, Rectangular Distribution, Bernoulli Distribution, Binomial Distribution, Poisson Distribution, Continuous Probability Distributions, Uniform Distribution, Normal Distribution, Exponential Distribution, Geometric Distribution | 10 |  | **20**  **Marks** |
| **4** | **Bootstrapping:**  Law of large numbers, bootstrap sample, bootstrap distribution, standard error, margin of error, confidence interval, bootstrap for machine learning. | 7 |  | **16**  **Marks** |
| **5** | **Optimization:**  Linear Optimization Problems, Introduction to Linear, Concave and Convex optimizations. Different Loss functions Optimization: Maxima and Minima, LPP. | 10 |  | **20**  **Marks** |

**COURSE STRUCTURE**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Course Code** |  | | | |
| **Course Category** | **Program Core** | | | |
| **Course Title** | **Big Data Architecture and Ecosystem** | | | |
| **Teaching Scheme and Credits**  **Weekly load hrs** | **L** | **T** | **Laboratory** | **Credits** |
| **3** | **0** | **4** | **5** |
| **Pre-requisites**: Basics of Operating System, Algorithms, Databases | | | | |
| **Course Objectives:**   1. To teach the fundamentals of Big Data Analytics 2. To teach the architecture of distributed Operating system for Hadoop 3. To teach Hadoop framework and components 4. To teach Hadoop ecosystem tools | | | | |
| **Course Outcomes:**  On completion of the course:   1. The students will learn Hadoop history, ecosystem, important features of Big Data and its architecture. 2. The students will learn distributed systems, different components of Hadoop ecosystem and Apache Hadoop 3. The students will be able to apply Big Data concepts in Real World | | | | |
| **Course Contents**   1. Introduction to the course 2. Characteristics of Big Data 3. Introduction to Hadoop 4. Hadoop Framework 5. Hadoop Ecosystem | | | | |
| **Learning Resources**:  **Textbooks:**   1. Hadoop, The Definitive Guide, Tom White, 4th edition, O’Reilly   **Reference Books:**   1. Data Analytics with Hadoop, by Benjamin Bengfort and Jenny Kim   **Additional Reading:**  **Websites :** https: //www.tutorialspoint.com  **MOOCS:** Coursera, Udemy, Udacity, Simplilearn, Edx | | | | |
| **Pedagogy:**  Participative learning, discussions, algorithm, demo using latest tools,experiential learning through practical problem solving, assignment, PowerPoint presentations and Case Study. | | | | |
| **Assessment Scheme:**  Class Continuous Assessment (CCA) 40 marks and Laboratory Continuous Assessment (LCA) 20 marks   |  |  |  |  | | --- | --- | --- | --- | | Mid Term Examination  (MCQ/Online Test /Direct Internal Examination) | FAT 1 (Formative Assessment Test 1)  (Assignment, MCQ, Class  Test, Presentation,  Seminar) | LCA | Total | | 20 Marks | 20 Marks | 20 Marks | 60 Marks |   **Term End Examination: 40 Marks** | | | | |

**Theory Syllabus:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Module  No. | Contents | Workload in Hrs | | |
| Theory | Lab | Assess |
| **1** | **Introduction:** What is Big Data? Different types of data generated. Examples of giant companies generating big data. Applications of Big data. What is Data Science? | 7 | - | **14**  **Marks** |
| **2** | **Characteristics of Big Data**: Volume, Velocity and Variety. Other V’s like variability, veracity, valence and value  Big Data Analytics, Case Studies related to Big Data Architecture | 8 | - | **16**  **Marks** |
| **3** | **Introduction to Hadoop:** Distributed File System, Google File System, History of Hadoop, concept of commodity server, cluster, Hadoop Vs RDBMS, Data integrity in Hadoop,  Basic overview of Hadoop Installation, Hadoop Shell Commands | 10 | 20 | **20**  **Marks** |
| **4** | **Hadoop Framework:**  HDFS: - Block Size, Architecture - Namenode, Datanode, Secondary namenode, Federation, Anatomy of File Read and Write  MapReduce: MapReduce programming model, Mapper and Reducer, Example of a map reduce job, Matrix multiplication using MapReduce.  YARN Architecture: Resource Manager, Node Manager, Application Master, Container, Anatomy of MapReduce Job run in YARN | 10 | 20 | **25**  **Marks** |
| **5** | **Hadoop Ecosystem:**  Overview and comparison of different ecosystem tools like Apache Flume, Pig, Scoop, Hive, HBase, Storm, Zookeeper, Oozie | 10 | 20 | **25**  **Marks** |

**Laboratory Syllabus:**

|  |  |  |
| --- | --- | --- |
| Practical Session | Lab Work | Contact Hours |
| 1 | Download and install VMWare Player / Virtual Box | 8 |
| 2 | Open/Play VM with VMWare player / Virtual Box | 4 |
| 3 | Installation Linux and Hadoop and Hadoop tools | 4 |
| 4 | Download and install Cloudera/Hortonworks VM | 4 |
| 5 | HDFS File Manipulation Commands, HDFS Admin Commands | 4 |
| 6 | HDFS Admin Commands | 8 |
| 7 | Apache Hive with six datasets | 4 |
| 8 | Apache PIG with six datasets: | 4 |
| 9 | Apache Sqoop with MySQL | 8 |
| 10 | Apache HBase commands | 8 |
| * **4 hours allocated for surprise test, viva and submission** | | |