TECHNO INDIA UNIVERSITY

W E S T B E N G A L

EM 4, Sector V, Salt Lake, Kolkata-700091, West Bengal, India
Phone: +91 9836544416/17/18/19, Fax: +91 33 2357 1097

3-Year Master of Computer Application (MCA) Curriculum and Syllabus

Sixth Semester

Course Code	Course Title	Contact Hrs. / Week			Credit
		L	T	P	
Theory					
TIU-UTR-T300	**Career Advancement & Skill Development- For IMCA (Pl/Sql) & For MCA(SAP-ABAP)**	2	0	0	2
TIU-PCA-T312	**Advanced Data Science through R**	3	1	0	4
TIU-PCA-T310	**Big Data Analytics**	3	1	0	4
TIU-PCA-T308	**AI & Machine Learning**	2	1	0	3
Practical					
TIU-PCA-L312	**Advanced Data Science through R programming Lab**	0	0	3	2
TIU-PCA-L308	**Machine Learning with Python**	0	0	3	2
Sessional					
TIU-PES-S398	**Entrepreneurship Skill Development**	0	0	3	3
TIU-PCA-P396	**Major Project using J2EE**	0	0	3	10
TIU-PCA-G398	**Grand Viva**	0	0	0	2
Total Credits					32

Approved by:

External Expert-1   (Prof. Subhadip Basu, J.U.)

External Expert-2   (Prof. Amlan Chakraborty, C.U.)

HOD -   (Prof. A.B. Chaudhuri)

Detailed Syllabus

6th Semester

Career Advancement & Skill Development

TIU-UTR-T300

L-T-P: 2-0-0   Credit 2

SAP – ABAP(iMCA) & Pl/Sql(MCA)

Advanced Data Science through R
TIU-PCA-T312

L-T-P: 3-1-0   Credit: 4

Course Description: Data Science is the study of the generalizable extraction of knowledge from data. Being a data scientist requires an integrated skill set spanning mathematics, statistics, machine learning, databases and other branches of computer science along with a good understanding of the craft of problem formulation to engineer effective solutions. This course will introduce students to this rapidly growing field and equip them with some of its basic principles and tools as well as its general mindset. Students will learn concepts, techniques and tools they need to deal with various facets of data science practice, including data collection and integration, exploratory data analysis, predictive modeling, descriptive modeling, data product creation, evaluation, and effective communication. The focus in the treatment of these topics will be on breadth, rather than depth, and emphasis will be placed on integration and synthesis of concepts and their application to solving problems. To make the learning contextual, real datasets from a variety of disciplines will be used.

Learning Outcomes: At the conclusion of the course, students should be able to:

- Describe what Data Science is and the skill sets needed to be a data scientist.

- Explain in basic terms what Statistical Inference means. Identify probability distributions commonly used as foundations for statistical modeling. Fit a model to data.

Approved by:
        External Expert-1 (Prof. Subhadip Basu, J.U.)
        External Expert-2 (Prof. Amlan Chakraborty, C.U.)
        HOD -   (Prof. A.B. Chaudhuri)

- Use R to carry out basic statistical modeling and analysis.

- Explain the significance of exploratory data analysis (EDA) in data science. Apply basic tools (plots, graphs, summary statistics) to carry out EDA.

- Describe the Data Science Process and how its components interact.

- Use APIs and other tools to scrap the Web and collect data.

- Apply EDA and the Data Science process in a case study. 1

- Apply basic machine learning algorithms (Linear Regression, k-Nearest Neighbors (k-NN), k-means, Naive Bayes) for predictive modeling. Explain why Linear Regression and k-NN are poor choices for Filtering Spam. Explain why Naive Bayes is a better alternative.

- Identify common approaches used for Feature Generation. Identify basic Feature Selection algorithms (Filters, Wrappers, Decision Trees, Random Forests) and use in applications.

- Identify and explain fundamental mathematical and algorithmic ingredients that constitute a Recommendation Engine (dimensionality reduction, singular value decomposition, principal component analysis). Build their own recommendation system using existing components.

- Create effective visualization of given data (to communicate or persuade).

- Work effectively (and synergically) in teams on data science projects.

Detailed Topics:

Unit 1:   **Introduction to Business Analytics:** Overview, Business Decisions and Analytics, Types of Business Analytics, Applications of Business Analytics, Data Science Overview.

Unit 2:   **Review or R programming:** General Programming logics, Data Structures, Graphical presentation, File and Database handling.

Unit 3:   **Data Collection and Data Blending:** Basic Concepts, Case Study 1- Collecting Data from Twitter , Business Intelligence and Data Warehousing , Data Science Project   Life Cycle , Data Visualization, Case Study 2 - Analyzing data from MovieLens,

Unit 4:   **Statistics for Data Science:** Overview, Introduction to Hypothesis, Types of Hypothesis, Data Sampling,   Confidence and Significance Levels, Hypothesis Testing,   Parametric Test,   Non-Parametric Test,   Hypothesis Tests about Population Means, Hypothesis Tests about Population Variance, Hypothesis Tests about Population Proportions.

Unit 5:   **Regression Analysis:** Overview, Introduction to Regression Analysis, Types of Regression Analysis Models, Linear Regression, Demo: Simple Linear Regression, Non-Linear Regression, Demo:

Approved by:
         External Expert-1   (Prof. Subhadip Basu, J.U.)
         External Expert-2   (Prof. Amlan Chakraborty, C.U.)
         HOD -   (Prof. A.B. Chaudhuri)

TECHNO INDIA UNIVERSITY
W E S T B E N G A L

EM 4, Sector V, Salt Lake, Kolkata-700091, West Bengal, India
Phone: +91 9836544416/17/18/19, Fax: +91 33 2357 1097

Regression Analysis with Multiple Variables, Cross Validation, NAIVE BAYES AND LOGISTIC REGRESSION, Non-Linear to Linear Models, Principal Component Analysis, Factor Analysis.

Unit 6: **Machine Learning for Data Science: Clustering:** Overview, Introduction to Clustering, Clustering Example, Clustering Methods: Prototype Based Clustering, Demo: K-means Clustering, Clustering Methods: Hierarchical Clustering, Demo: Hierarchical Clustering, Clustering ,Methods: DBSCAN, Data Preprocessing, Model, Evaluation and Ensembles; Principal Component Analysis, **Association:** Overview, Association Rule, Apriori Algorithm, Demo: Apriori Algorithm; **Data Mining:** Dimensionality Reduction, Clustering, Association Rules, Anomaly Detection, Network Analysis and Recommender Systems; Case Study 3 – Creating a Data Product using some Machine Learning Algorithm; **Big Data** Analytics.

Unit 7: **Specialty Topics:** Data Engineering, Natural Language Processing, and Web Applications.

Unit 8: **Deep Learning:**

Big Data Analytics
TIU-PCA-T310

L-T-P: 3-1-0   **Credit 4**

1 INTRODUCTION TO BIG DATA: Introduction– distributed file system–Big Data and its importance, Four Vs, Drivers for Big data, Big data analytics, Big data applications. Algorithms using map reduce .

2 INTRODUCTION TO HADOOP AND HADOOP ARCHITECTURE: Big Data – Apache Hadoop & Hadoop EcoSystem, Moving Data in and out of Hadoop – Understanding inputs and outputs of MapReduce -, Data Serialization.

3 HDFS, HIVE AND HIVEQL, HBASE HDFS: Overview, Installation and Shell, Java API; Hive Architecture and Installation, Comparison with Traditional Database, HiveQL Querying Data, Sorting And Aggregating, Map Reduce Scripts, Joins & Sub queries, HBase concepts, Advanced Usage, Schema Design, Advance Indexing, PIG, Zookeeper , how it helps in monitoring a cluster, HBase uses Zookeeper and how to Build Applications with Zookeeper.

4 SPARK: Introduction to Data Analysis with Spark, Downloading Spark and Getting Started, Programming with RDDs, Machine Learning with MLib

5. SQOOP, Zookeeper.

6. NoSQL What is it?, Where It is Used Types of NoSQL databases, Why NoSQL?, Advantages of NoSQL, Use of NoSQL in Industry, SQL vs NoSQL, NewSQL .

Approved by:
   External Expert-1 (Prof. Subhadip Basu, J.U.)
   External Expert-2 (Prof. Amlan Chakraborty, C.U.)
   HOD -   (Prof. A.B. Chaudhuri)

TECHNO INDIA UNIVERSITY

W E S T B E N G A L

EM 4, Sector V, Salt Lake, Kolkata-700091, West Bengal, India
Phone: +91 9836544416/17/18/19, Fax: +91 33 2357 1097

7. Data Base for the Modern Web Introduction to MongoDB key features, Core Server tools, MongoDB through the JavaScript's Shell, Creating and Querying through Indexes, Document-Oriented, principles of schema design, Constructing queries on Databases, collections and Documents , MongoDB Query Language

Books for Main Reading:

1. Chris Eaton,Dirk derooset al. , "Understanding Big data ", McGraw Hill, 2012.

 2. BIG Data and Analytics , Sima Acharya, Subhashini Chhellappan, Willey

3. MongoDB in Action, Kyle Banker,Piter Bakkum , Shaun Verch, Dreamtech Press

4. Tom White, "HADOOP: The definitive Guide", O Reilly 2012.

AI & Machine Learning
TIU-PCA-T308

L-T-P: 2-1-0   **Credit 3**

1. **Introduction:**

   Definition of learning systems. Goals and applications of machine learning. Aspects of developing a learning system: training data, concept representation, function approximation.

2. **Inductive Classification:**

   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.

3. **Decision Tree Learning:**

   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. Occam's razor. Over fitting, noisy data, and pruning.

4. **Ensemble Learning:**

   Using committees of multiple hypotheses. Bagging, boosting, and DECORATE. Active learning with ensembles.

5. **Experimental Evaluation of Learning Algorithms:**

Approved by:

      External Expert-1 **(Prof. Subhadip Basu, J.U.)**

      External Expert-2 **(Prof. Amlan Chakraborty, C.U.)**

      HOD -   **(Prof. A.B. Chaudhuri)**

Measuring the accuracy of learned hypotheses.
Comparing learning algorithms: cross-validation, learning curves, and statistical hypothesis testing.

6. **Computational Learning Theory:**

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. PAC results for learning conjunctions, kDNF, and kCNF. Sample complexity for infinite hypothesis spaces, Vapnik-Chervonenkis dimension.

7. **Rule Learning: Propositional and First-Order:**

Translating decision trees into rules. Heuristic rule induction using separate and conquer and information gain. First-order Horn-clause induction (Inductive Logic Programming) and Foil. Learning recursive rules. Inverse resolution, Golem, and Progol.

8. **Artificial Neural Networks:**

Neurons and biological motivation. Linear threshold units. Perceptrons: representational limitation and gradient descent training. Multilayer networks and backpropagation. Hidden layers and constructing intermediate, distributed representations. Over fitting, learning network structure, recurrent networks.

9. **Support Vector Machines:**

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

10. **Bayesian Learning:**

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

11. **Instance-Based Learning:**

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

12. **Text Classification:**

Bag of words representation. Vector space model and cosine similarity. Relevance feedback and Rocchio algorithm. Versions of nearest neighbor and Naive Bayes for text.

13. **Clustering and Unsupervised Learning:**

Approved by:
External Expert-1 (Prof. Subhadip Basu, J.U.)
External Expert-2 (Prof. Amlan Chakraborty, C.U.)
HOD - (Prof. A.B. Chaudhuri)

TECHNO INDIA UNIVERSITY

W E S T B E N G A L

EM 4, Sector V, Salt Lake, Kolkata-700091, West Bengal, India
Phone: +91 9836544416/17/18/19, Fax: +91 33 2357 1097

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

14. **Language Learning :**

Classification problems in language: word-sense disambiguation, sequence labeling. Hidden Markov models (HMM's). Veterbi algorithm for determining most-probable state sequences. Forward-backward EM algorithm for training the parameters of HMM's. Use of HMM's for speech recognition, part-of-speech tagging, and information extraction. Conditional random fields (CRF's). Probabilistic context-free grammars (PCFG). Parsing and learning with PCFGs. Lexicalized PCFGs.

Book for main reading:

1. Machine Learning, Tom Mitchell, McGraw Hill, 1997.

2. Bishop, C. (2006). Pattern Recognition and Machine Learning. Berlin: Springer-Verlag.

Book for Supplementary reading:

1. Machine Learning: A Probabilistic Perspective, By Kevin P. Murphy, MIT press.

2. Research Paper: Ensemble Learning : Thomas G. Dietterich Department of Computer Science Oregon State University Corvallis, Oregon 97331-3202 USA tgd s.orst.edu September 4, 2002

Advanced R Programming Lab
TIU-PCA-L312

L-T-P: 0-0-3   **Credit 2**

Approved by:

External Expert-1   (Prof. Subhadip Basu, J.U.)

External Expert-2   (Prof. Amlan Chakraborty, C.U.)

HOD -   (Prof. A.B. Chaudhuri)