SRI VENKATESWARA UNIVERSITY DEGREE COURSE IN BACHELOR OF COMPUTER APPLICATIONS (BCA) UNDER CBCS W.E.F.2021-22

BIG DATA & MACHINE LEARNING

IV-SEMESTER

S.No	Paper Code	Subject	Hours/ Week	No of Credits	Max.Marks Internal assessment	Max. Marks University Exam	Total Marks
1.	C10	Data Warehousing & Data Mining	4	3	25	75	100
2.	C10-P	DM & DW Lab	2	2	-0-	50	50
3.	C11	Introduction To Big Data	4	3	25	75	100
4.	C12	Cloud Computing	4	3	25	75	100
5.	C13	NO SQL Databases	4	3	25	75	100
6.	C14	Big Data with Hadoop	4	3	25	75	100
7	C14-P	Big Data with Hadoop lab	2	2	-0-	50	50
8	C-15	Predictive Modeling And Analytics	4	3			
Total			28	25	125	475	600

C10-DATA WAREHOUSING AND DATA MINING

OBJECTIVES:

- Study data warehouse principles and its working
- Learn Data mining concepts and understand Association Rule Mining
- Study Classification Algorithms 4. Gain knowledge of how data is grouped using clustering techniques.

OUTCOMES:

- Comparison of functional differences between data warehouse and database systems.
- Ability to perform the pre-processing of data and apply mining techniques on it.
- Capability to identify the association rules, classification and clusters in large data sets.
- Skills to solve real world problems in business and scientific information using data mining.

UNIT-I Data warehouse:

Introduction to Data warehouse, Difference between operational database systems and data warehouses, Data warehouse Characteristics, Data warehouse Architecture and its Components, Extraction-Transformation-Loading, Logical(Multi-Dimensional), Data Modeling, Schema Design, Star and Snow-Flake Schema, Fact Constellation, Fact Table, Fully Addictive, Semi-Addictive, Non Addictive Measures; Fact-Less-Facts, Dimension Table Characteristics; OLAP Cube, OLAP Operations, OLAP Server Architecture-ROLAP, MOLAP and HOLAP.

UNIT-II Introduction:

Fundamentals of data mining, Data Mining Functionalities, Classification of Data Mining systems, Data Mining Task Primitives, Integration of a Data Mining System with a Database or Data Warehouse System, Major issues in Data Mining. Data Preprocessing: Need for Preprocessing the Data, Data Cleaning, Data Integration &Transformation, Data Reduction, Discretization and Concept Hierarchy Generation.

UNIT-III Association Rules:

Problem Definition, Frequent Item Set Generation, The APRIORI Principle, Support and Confidence Measures, Association Rule Generation; APRIOIRI Algorithm, The Partition Algorithms, FP-Growth Algorithms, Compact Representation of Frequent Item Set- Maximal Frequent Item Set, Closed Frequent Item Set.

UNIT-IV Classification:

Problem Definition, General Approaches to solving a classification problem, Evaluation of Classifiers, Classification techniques, Decision Trees-Decision tree Construction, Methods for Expressing attribute test conditions, Measures for Selecting the Best Split, Algorithm for Decision tree Induction; Naive-Bayes Classifier, Bayesian Belief Networks; K- Nearest neighbor classification-Algorithm and Characteristics. Prediction: Accuracy and Error measures, Evaluating the accuracy of classifier or a predictor, Ensemble methods

UNIT-V Clustering:

Clustering Overview, A Categorization of Major Clustering Methods, Partitioning Methods, Hierarchical Methods, Partitioning Clustering-K-Means Algorithm, PAM Algorithm; Hierarchical Clustering-Agglomerative Methods and divisive methods, Basic Agglomerative Hierarchical Clustering Algorithm, Key Issues in Hierarchical Clustering, Strengths and Weakness, Outlier Detection.

TEXT BOOKS:

1) Data Mining- Concepts and -1.chniques- Jiawei Han, Micheline Kamber, Morgan Kaufmann Publishers, Elsevier, 2 Edition, 2006.

2) Introduction to Data Mining, Psng-Ning Tan, Vipin Kumar, Michael Steinbanch, Pearson Educatior.

REFERENCE BOOKS:

1) Data Mining Techniques, Arun KPujari, 3rd Edition, Universities Press.

2) Data Warehousing Fundament's, Pualraj Ponnaiah, Wiley Student Edition.

3) The Data Warehouse Life CycleToolkit — Ralph Kimball, Wiley Student Edition.

4) Data Mining, Vikaram Pudi, P Rddha Krishna, Oxford University Press

C 11- INTRODUCTION TO BIG DATA

COURSE OBJECTIVES

- The course gives an overview of the Big Data phenomenon, focusing then on extracting value from the Big Data using predictive analytics techniques
- Students will gain knowledge on analyzing Big Data. It serves as an introductory course for graduate students who are expecting to face Big Data storage, processing, analysis, visualization, and application issues on both workplaces and research environments.
- Gain knowledge on this fast-changing technological direction. Big Data Analytics is probably the fastest evolving issue in the IT world now.
- Get insight on what tools, algorithms, and platforms to use on which types of real world use cases.

COURSE OUTCOMES

Upon completion of this course, the students will be able to

- Outline the importance of Big Data Analytics
- Apply statistical techniques for Big data Analytics.
- Analyze problems appropriate to mining data streams.
- Apply the knowledge of clustering techniques in data mining.
- Use Graph Analytics for Big Data and provide solutions

UNIT1: INTRODUCTION TO BIG DATA

Evolution of Big data - Best Practices for Big data Analytics - Big data characteristics - Validating - The Promotion of the Value of Big Data - Big Data Use Cases- Characteristics of Big Data Applications - Perception and Quantification of Value -Understanding Big Data Storage - Evolution Of Analytic Scalability - Analytic Processes and Tools - Analysis vs Reporting - Modern Data Analytic Tools - Statistical Concepts: Sampling Distributions - Re-Sampling - Statistical Inference - Prediction Error.

UNIT2: DATA ANALYSIS, CLUSTERING AND CLASSIFICATION

Regression Modeling - Multivariate Analysis - Bayesian Modeling - Support Vector and Kernel Methods - Analysis of Time Series: Linear Systems Analysis - Nonlinear Dynamics - Rule Induction. Overview of Clustering - K-means - Use Cases - Overview of the Method - Determining the Number of Clusters - Diagnostics - Reasons to Choose and Cautions .- Classification: Decision Trees - Overview of a Decision Tree - The General Algorithm - Decision Tree Algorithms - Evaluating a Decision Tree - Decision Trees in R - Naïve Bayes - Bayes‘ Theorem - Naïve Bayes Classifier.

UNIT3: STREAM MEMORY

Introduction To Streams Concepts – Stream Data Model and Architecture - Stream Computing - Sampling Data in a Stream – Filtering Streams – Counting Distinct Elements in a Stream – Estimating Moments – Counting Oneness in a Window – Decaying Window - Real time Analytics Platform(RTAP) Applications - Case Studies - Real Time Sentiment Analysis, Stock Market Predictions.

UNIT4: ASSOCIATION AND GRAPH MEMORY

Advanced Analytical Theory and Methods: Association Rules - Overview - Apriori Algorithm - Evaluation of Candidate Rules - Applications of Association Rules - Finding Association& finding similarity - Graph Analytics for Big Data:

UNIT5: GRAPH ANALYTICS

The Graph Model - Representation as Triples - Graphs and Network Organization - Choosing Graph Analytics - Graph Analytics Use Cases - Graph Analytics Algorithms and Solution Approaches - Technical Complexity of Analyzing Graphs- Features of a Graph Analytics Platform.

TEXT BOOKS

1.David Loshin, "Big Data Analytics: From Strategic Planning to Enterprise Integration with Tools, Techniques, NoSQL, and Graph", 2013. 2. AnandRajaraman and Jeffrey David Ullman, “Mining of Massive Datasets”, Cambridge University Press, 2012 3. Michael Berthold, David J. Hand, “Intelligent Data Analysis”, Springer, 2007.

REFERENCE BOOKS

1. EMC Education Services, "Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data", Wiley publishers, 2015.
2. Bart Baesens, "Analytics in a Big Data World: The Essential Guide to Data Science and its Applications", Wiley Publishers, 2015.
3. Kim H. Pries and Robert Dunnigan, "Big Data Analytics: A Practical Guide for Managers " CRC Press, 2015

C 12-CLOUD COMPUTING

OBJECTIVES:

- Cloud computing is a colloquial expression used to describe a variety of different computing concepts that involve a large number of computers involves a large number of computers that are connected through a real-time communication network.
- In science, cloud computing is a synonym for distributed computing over a network and means the ability to run a program on many connected computers at the same time.
- This course covers basic concepts of cloud types, services and security etc.

OUTCOMES:

- Learners will develop knowledge about Software Development Life Cycle, which is very essential for any Software Developer to design and develop any application or software.
- This course also includes UNITs on Software testing which forms an essential part of SDLC
- Students are ability to apply knowledge of mathematics, science, and engineering
- Students are ability to design and conduct experiments, as well as to analyze and interpret data.
- Students are ability to function on multi-disciplinary teams.
- Students are ability to analyze, design, verify, validate, implement, apply, and maintain software system.

At the end of the course, the students will be able to:

- Learn the underlying principles of Cloud Technology and various types of cloud computing architecture and types.
- Evaluate between different cloud solutions offered by various providers based on their merits and demerits.
- Understand the Cloud Cost Management and Selection of Cloud Provider.
- Understand the IT governance in cloud computing.
- Track the Ten cloud do an do nots.:

UNIT I: Introduction

Introduction to Cloud Computing, History and Evolution of Cloud Computing, Types of clouds, Private Public and hybrid clouds, Cloud Computing architecture, Cloud computing infrastructure, Merits of Cloud computing, , Cloud computing delivery models and services (IaaS, PaaS, SaaS), obstacles for cloud technology, Cloud vulnerabilities, Cloud challenges, Practical applications of cloud computing.

UNIT II: Cloud Computing Companies and Migrating to Cloud

Web-based business services, Delivering Business Processes from the Cloud: Business process examples, Broad Approaches to Migrating into the Cloud, The Seven-Step Model of Migration into a Cloud, Efficient Steps for migrating to cloud., Risks: Measuring and assessment of risks, Company concerns Risk Mitigation methodology for Cloud computing, Case Studies

UNIT III: Cloud Cost Management and Selection of Cloud Provider

Assessing the Cloud: software Evaluation, System Testing, Seasonal or peak loading, Cost cutting and cost-benefit analysis, selecting the right scalable application. Considerations for selecting cloud solution. Understanding Best Practices used in selection of Cloud service and providers, Clouding the Standards and Best Practices Issue: Interoperability, Portability, Integration, Security, Standards Organizations and Groups associated with Cloud Computing, Commercial and Business Consideration

UNIT IV: Governance in the Cloud

Industry Standards Organizations and Groups associated with Cloud Computing, Need for IT governance in cloud computing, Cloud Governance Solution: Access Controls, Financial Controls, Key Management and Encryption, Logging and Auditing, API integration. Legal Issues: Data Privacy and Security Issues, Cloud Contracting models, Jurisdictional Issues Raised by Virtualization and Data Location, Legal issues in Commercial and Business Considerations

UNIT V: Ten cloud do an do nots

Don't be reactive, do consider the cloud a financial issue, don't go alone, do think about your architecture, don't neglect governance, don't forget about business purpose, do make security the centerpiece of your strategy, don't apply the cloud to everything, don't forget about Service Management, do start with a pilot project.

TEXT BOOKS:

1. Cloud Computing: Principles and Paradigms, Rajkumar Buyya, James Broberg, Andrzej M. Goscinski,, John Wiley and Sons Publications, 2011

REFERENCES:

1. Brief Guide to Cloud Computing, Christopher Barnett, Constable & Robinson Limited, 2010
2. Handbook on Cloud Computing, Borivoje Furht, Armando Escalante, Springer, 2010

LIST OF PROGRAMS: Study the basic cloud architecture and represent it using a case study

1. Enlist Major difference between SAAS PAAS & Iaas also submit a research done on various companies in cloud business and the corresponding services provided by them , tag them under SAAS , Paas & Iaas.
2. Study and present a report on Jolly cloud.
3. Present a report on obstacles and vulnerabilities in cloud computing on generic level
4. Present a report on Amazon cloud services.
5. Present a report on Microsoft cloud services.
6. Present a report on cost management on cloud
7. Enlist and explain legal issues involved in the cloud with the lelp of a case study
8. Explain the process of migrating to cloud with a case study.
9. Present a report on google cloud and cloud services
10. Create a scenario based on real time domain

C 13-NOSQL DATABASES

OBJECTIVES: The course is aimed at:

- Provide students an exposure to unstructured data.
- Work with query unstructured database.
-

OUTCOMES: At the end of the course, the students will be able to:

- Identify the use of unstructured data.
- Know the knowledge of features of NO SQl Data Base.
- Know the Key-Value Databases, Document Databases.
- Learn various concepts of Graph Databases .
- Analyze the advantage & disadvantages of Relational database

UNIT I: INTRODUCING NOSQL

The value of Relational Databases, Disadvantages of Relational Databases, A Brief History of NoSQL, Features of NoSQL : Features of NoSQL, ACID vs. BASE, Managing Different Data Types

UNIT II: DATA MODELS

Aggregates, key-value and document data models, Column-Family Stores, relationships, graph databases, schema-less databases, materialized views. Distribution models: Single Server, sharding, master-slave replication, peer-peer replication, sharding and replication

UNIT III: UPDATE AND READ CONSISTENCY

Update Consistency, Read Consistency. Relaxing Consistency: Relaxing Consistency, Relaxing Durability

UNIT IV: NOSQLDATABASES

Key-Value Databases, Document Databases, Column-Family Stores

UNIT V: GRAPH DATABASES

Graph Databases, Beyond NoSQL.

TEXT BOOKS:

1. P. J. Sadalage and M. Fowler, "NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot Persistence",Copyright © 2013 Pearson Education, Inc. 2012.
2. NoSQL For Dummies®, 2015 by John Wiley & Sons, Inc
3. Professional NoSQL, Shashank Tiwari, Wrox
4. E. Capriolo, D. Wampler, and J. Rutherglen, "Programming Hive", O'Reilley, 2012.

REFERENCES:

1. Lars George, "HBase: The Definitive Guide", O'Reilley, 2011.
2. Eben Hewitt, "Cassandra: The Definitive Guide", O'Reilley, 2010.
3. “MongoDB: The Definitive Guide” by Kristina Chodorow

C14 -BIG DATA WITH HADOOP

COURSE OBJECTIVES :

- Understand the Big Data Platform and its Use cases
- Provide an overview of Apache Hadoop
- The HDFS file system, MapReduce frameworks are studied in detail
- Apply analytics on Structured, Unstructured Data,
- Introduction to YARN and MapReduce

COURSE OUTCOMES: The students will be able to:

- Identify Big Data and its Business Implications.
- List the components of Hadoop and Hadoop Eco-System
- Access and Process Data on Distributed File System
- Manage Job Execution in Hadoop Environment
- Develop Big Data Solutions using Hadoop Eco System
- Understand the use of predictive analytics on big data

UNIT I Introduction to Big Data, Characteristics and its Use Case

Introduction – Why Big data - What is big data – Facts about Big Data - importance of Big Data Evaluation of Big Data – Market Trends – Sources of Data Explosion – Types of Data – Case Study for Netflix and the house of card. Need of Big Data – Big Data and its sources – Characteristics of Big Data – Difference between Traditional IT Approach and Big Data Technology – Capabilities of Big Data – Handling Limitations of Big Data - Technologies Supporting Big Data - Big Data Use Cases.

UNIT II Introduction to Hadoop

Introduction – Why Hadoop – What is Hadoop – History and Milestone of Hadoop – Core Components of Hadoop – Difference between Regular File System and HDFS – Common Hadoop Shell Commands – Hadoop Configuration.

UNIT III Hadoop Distributed File System (HDFS)

Concepts and Architecture - Data Flow (File Read, File Write) - Fault Tolerance - Java Base API - Different Daemons in Hadoop cluster (NameNode, Secondary NameNode, Job Tracker, Task Tracker and DataNode) - Loading a dataset into the HDFS.

UNIT IV Introduction to YARN and MapReduce

What is YARN – YARN Infrastructure - Introduction of MapReduce – Analogy of MapReduce – MapReduce Architecture - Example of MapReduce –Sorting, Shuffling – Reducing – Combiner – Partitioner – Creating MapReduce program by using Eclipse.

UNIT V Introduction to Big Data Streaming

Real time Big Data Streaming, Big data streaming framework, data streaming process, tools for big data streaming, industry use cases for big data streaming.

TEXT BOOKS:

1. Seema Acharya (Author), Subhashini Chellappan, Big Data and Analytics (2015). Wiley Publication.
2. Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data (2015), EMC Education Services

REFERENCES:

1. Big Data, Black Book: Covers Hadoop 2, MapReduce, Hive, YARN, Pig, R and Data Visualization (2016), DT Editorial Services 2. Tom White, Hadoop: The Definitive Guide, 4th Edition (2015)

C15- PREDICTIVE MODELING AND ANALYTICS

COURSE OBJECTIVES:

- Formulate complex decision-making problems with data for predictive analysis in business context.
- Analyze and evaluate predictive model outcomes for informing decision-making.
- Ability to apply specific statistical and regression analysis methods applicable to predictive analytics to identify new trends and patterns, uncover relationships, create forecasts, predict likelihoods, and test predictive hypotheses.
- Ability to develop and use various quantitative and classification predictive models based on various regression and decision tree methods.

OUTCOMES : Upon completion of this course, the students will be able to

- Understand the basics of predictive analytics and summarize Data, Categorize Models, and techniques
- Apply Decision tree, Support Vector Machine for Data Classification
- Apply Methods such as Naïve Bayes Markov Model, Linear Regression, Neural Networks to Boost Prediction Accuracy for Data Classification.
- Develop predictive models for various Real-Time Applications.
- Analyze and Visualize predictive Model's results using Data Visualization tools

UNIT1: DATA PREPARTION

Introduction – Predictive Analytics in the Wild – Exploring Data types and associated Techniques - Complexities of data - Applying Models: Models and simulation, Categorizing Models, Describing, summarizing data, and decisions – Identify similarities in Data: Data Clustering, converting Raw Data into a Matrix, Identify K-groups in Data.

UNIT2: DATA CLASSIFICATION – PART I

Background – Exploring Data classification process - Using Data Classification to predict the future: Decision tree, Algorithm for generating Decision Trees, Support Vector Machine.

UNIT3: DATA CLASSIFICATION – PART II

Ensemble Methods to Boost Prediction Accuracy: Naïve Bayes Classification Algorithm, The Markov Model, Linear Regression, Neural Networks – Deep learning.

UNIT4: DATA PREDICTION

Adopt predictive analytics - Processing data: identifying, cleaning, generating, reducing dimensionality of data – Structuring Data – Build predictive model: develop and test the model.

UNIT5: DATA VISUALIZATION

Introduction to visualization tool – Evaluate the data – visualize Model's Analytical Results: hidden grouping, data classification results, outliers, decision trees, prediction – Novel visualization in Predictive Analytics.

TEXT BOOKS

1. Anasse Bari, Mohamed Chaouchi, Tommy Jung, “Predictive Analytics For Dummies”, Wiley Publisher, 2nd Edition, 2016.

REFERENCE BOOKS

1. Bertt Lantz, Machine Learning with R: Expert techniques for predictive modeling to solve all your data analysis problems, Pack Publisher, 2nd Edition, 2015.
2. Aurelien,”Hands-On Machine Learning with Scikit-Learn & TensorFlow”, O’Reilly Publisher, 5th Edition, 2017.
3. Max Kuhn, Kjell Johnson, “ Applied Predictive Modeling” Springer, 2013.

SRI VENKATESWARA UNIVERSITY

BCA DEGREE COURSE IN BIGDATA & MACHINE LEARNING

SECOND YEAR – FOURTH SEMESTER

(Syllabus under CBCS w.e.f. 2021-22)

MODEL QUESTION PAPER (for all papers)

Time: 3 hours Marks: 75 marks

Note: This question paper contains two parts A and B.

Part A is compulsory which carries 25 marks. Answer any five of the following questions in Part A.
Part B consists of 5 Units. Answer any one full question from each unit. Each question carries 10 marks

PART – A

Answer any *Five* of the following question. **(5X5=25M)**

1	
2	
3	
4	
5	
6	
7	
8	

PART – B

Answer All The Questions. Each question carries 10 marks **(5X10= 50M)**

9	(A) OR (B)
10	(A) OR (B)
11	(A) OR (B)
12	(A) OR (B)
13	(A) OR (B)