Sona College of Technology (Autonomous), Salem – 636 005

Continuing Education Centre

Department of Master of Computer Applications

Advanced Diploma in DATA ANALYTICS

CURRICULUM and SYLLABUS

Academic year – 2021-2022

I Year / I Semester

S. No.	Course Code	Course Title	L	T	P	C
Theory & Laboratory						
1.	MCAADDAXX	Introduction to Data Science	3	0	0	3
2.	MCAADDAXX	Python for Data Analytics	3	0	0	3
3.	MCAADDAXX	SQL for Data Analytics	2	0	2	3
4.	MCAADDAXX	Python Programming Laboratory	0	0	2	1
					Total Credits	10

I Year / II Semester

S. No.	Course Code	Course Title	L	T	P	C
Theory & Laboratory						
1.	MCAADDAXX	Statistics and Probability	3	0	0	3
2.	MCAADDAXX	R for Data Analytics	3	0	0	3
3.	MCAADDAXX	R Programming Laboratory	0	0	2	1
Project Work						
4.	MCAADDAXX	Project Work - 1	0	0	6	3
					Total Credits	10

II Year / III Semester

S. No.	Course Code	Course Title	L	T	P	C
Theory & Laboratory						
1.	MCAADDAXX	Big Data and Hadoop	3	0	0	3
2.	MCAADDAXX	Machine Learning	3	0	0	3
3.	MCAADDAXX	Big Data and Hadoop Laboratory	0	0	4	2
4.	MCAADDAXX	Machine Learning with Python Laboratory	0	0	4	2
					Total Credits	10

II Year / IV Semester

S. No.	Course Code	Course Title	L	T	P	C
Theory & Laboratory						
1.	MCAADDAXX	Data Visualization using Tableau & Power BI	2	0	2	3
2.	MCAADDAXX	Big Data Analytics Through Spark	2	0	2	3
Project Work						
3.	MCAADDAXX	Project Work – 2	0	0	8	4
					Total Credits	10

Chairperson/DCC/MCA

Semester I

MCAADDAXX - INTRODUCTION TO DATA SCIENCE

L	T	P	C	M
3	0	0	3	100

COURSE OBJECTIVES:

This course will enable the student to:

- Explain data science and its applications.
- Understand the strategies of data collection and pre-processing.
- Describe EDA.
- Apply statistics methods to develop models.
- Learn the evaluation metrics and techniques.

UNIT I – INTRODUCTION 9

Introduction to Data Science – Evolution of Data Science – Data Science Roles – Stages in a Data Science Project – Applications of Data Science in various fields – Data Security Issues.

UNIT II – DATA COLLECTION AND PRE-PROCESSING 9

Data Collection Strategies – Data Pre-Processing Overview – Data Cleaning – Data Integration and Transformation – Data Reduction – Data Discretization.

UNIT III – EXPLORATORY DATA ANALYTICS 9

Descriptive Statistics – Mean, Standard Deviation, Skewness and Kurtosis – Box Plots – Pivot Table – Heat Map – Correlation Statistics – ANOVA.

UNIT IV – MODEL DEVELOPMENT 9

Simple and Multiple Regression – Model Evaluation using Visualization – Residual Plot – Distribution Plot – Polynomial Regression and Pipelines – Measures for In-sample Evaluation – Prediction and Decision Making.

UNIT V – MODEL EVALUATION 9

Generalization Error – Out-of-Sample Evaluation Metrics – Cross Validation – Overfitting – Under Fitting and Model Selection – Prediction by using Ridge Regression – Testing Multiple Parameters by using Grid Search.

TOTAL = 45 Hours

COURSE OUTCOMES:

This course will enable the student to:

- Present an overview data science and applications.
- Plan the methods of data collection.
- Describe the statistical methods in EDA.

- Apply statistical methods to develop and evaluate the models.
- Becoming an expert in decision making for complex projects.

REFERENCES

1. Jojo Moolayil, “Smarter Decisions: The Intersection of IoT and Data Science”, PACKT, 2016.

2. Cathy O’Neil and Rachel Schutt , “Doing Data Science”, O'Reilly, 2015.

3. David Dietrich, Barry Heller, Beibei Yang, “Data Science and Big data Analytics”, EMC 2013

4. Raj, Pethuru, “Handbook of Research on Cloud Infrastructures for Big Data Analytics”, IGI Global.

MCAADDAXX - PYTHON FOR DATA ANALYTICS

L	T	P	C	M
3	0	0	3	100

COURSE OBJECTIVES:

This course will enable the student to:

- Represent compound data in Python data structures – lists, tuples, and dictionaries.
- Write Python programs with conditionals, loops and functions.
- Handle input/output operations in files.
- Use the aggregations and group operations for data analysis in python.
- Describe visualization methods in python.

UNIT I – PYTHON CONCEPTS AND DATA STRUCTURES 9

Intro to Python: Jupyter Environment – Pseudocode - Interpreter – Program Execution – Statements – Expressions – Using Print () - Wrong usage of print() – Variables - Creating a variable - Reassign a variable - Multiple variable assignment - Flow Controls: conditional and loop statements – Functions – Numeric Data Types – Data type conversion (Implicit) - Data type conversion (Explicit) –Arithmetic, Boolean Operations – Strings – Sequences – Tuples – Lists – Dictionaries.

UNIT II – OOP IN PYTHON 9

Class Definition – Constructors – Object Creation – Inheritance – Overloading – Text Files and Binary Files – Reading and Writing – Exception Handling.

UNIT III – DATA WRANGLING 9

Combining and Merging Data Sets – Reshaping and Pivoting – Data Transformation – String manipulations – Regular Expressions.

UNIT IV – DATA AGGREGATION AND GROUP OPERATIONS 9

GroupBy Mechanics – Data Aggregation – GroupWise Operations – Transformations – Pivot Tables – Cross Tabulations – Date and Time data types.

UNIT V – VISUALIZATION IN PYTHON 9

Matplotlib Package – Plotting Graph - Controlling Graphs – Adding Text – More Graph Types – Getting and Setting Values – Patches.

TOTAL = 45 Hours

COURSE OUTCOMES:

This course will enable the student to:

- Devise Python programs into functions with conditional and loops statements.
- Develop python based application using OOPs concepts and apply file I/O operations.
- Apply string manipulation in python programs.

- Analyze the data by aggregations and grouping operations.
- Develop python application with visualization effects.

REFERENCES

1. Mark Lutz, “Programming Python”, O'Reilly Media, 4th edition, 2010.
2. Joel Grus, “Data Science from scratch”, O'Reilly, 2015.
3. ReemaThareja, “Python Programming using Problem Solving Approach”, Oxford University Press, First edition, 2017
4. Tim Hall and J-P Stacey, “Python 3 for Absolute Beginners”, Apress, 1st edition, 2009
5. Allen B. Downey, “Think Python: How to Think Like a Computer Scientist”, Second Edition, Shroff,O‘Reilly Publishers, 2016 (http://greenteapress.com/wp/thinkpython/)
6. Magnus Lie Hetland, “Beginning Python: From Novice to Professional”, Apress, Second Edition, 2005.
7. Shai Vaingast, “Beginning Python Visualization Crafting Visual Transformation Scripts”, Apress, 2nd edition, 2014.
8. Wes Mc Kinney, “Python for Data Analysis”, O'Reilly Media, 2012.
9. Timothy A. Budd, “Exploring Python”, Mc-Graw Hill Education (India) Private Ltd.,First edition,2011.

MCAADDAXX – SQL FOR DATA ANALYTICS

L	T	P	C	M
2	0	2	3	100

COURSE OBJECTIVES:

This course will enable the student to:

- Understand the basis of relational databases and learn how to retrieve and manipulate data from one or more tables.
- Manipulate data with subqueries and aggregate functions.
- Apply views and joins to manage database.
- Demonstrate stored procedures and triggers.
- To learn emerging databases such as XML, NoSQL.

UNIT I – DESIGNING AND CONSTRUCTING A DATABASE 9

Database design: Database structure, Design process, Pre-design phase of design, Organizing your data, Functional dependency and candidate keys, Entity-relational modeling, Normalization. Creating databases: Creating a database, choosing which database to access, creating a table, Relational data types, Specifying keys, Column constraints, Default values, Design of the movie info database, Indexes. Creating, changing and removing records: Preparing data, INSERT statement, SELECT and INSERT together, DELETE statement, UPDATE statement, TRUNCATE statement, DROP statement, ALTER statement.

UNIT II – RETRIEVING DATA FROM A DATABASE 9

SELECT statement: Anatomy of a SELECT statement, Specifying columns to retrieve, Performing calculations on selected data, Using AS to name columns and expressions, Filtering query results using the WHERE clause, Dealing with null values, Sorting query results, How the equality of string is determined. Slicing Data - Using WHERE Clause: using logical operators in the WHERE clause, the IN clause, The BETWEEN Clause, Matching parts of strings using LIKE, Wildcard characters - Useful functions for WHERE clauses. Aggregating query results: Selecting unique values using DISTINCT, Aggregate functions, COUNT() function, SUM() and AVG() function, Dividing aggregates into categories – GROUP BY, Filtering query results using HAVING. Combining Tables using Joins: Joins and Normalization, What is a Join, Using Joins, Types of joins, Joining More than two tables, Self Joins, Outer Joins, UNION Joins, SQL -92 Join Syntax. Data Wrangling - SQL Data Cleaning - Subqueries: What is a subquery, Types of subqueries, Subqueries that return a list of values, Subqueries that return a single value, Writing complex queries, Using subqueries in UPDATE and DELETE statement, Using subqueries with INSERT.

UNIT III – DATABASE MANAGEMENT 9

Using Views: Creating Views, Advantages of Using Views, Creating Column Aliases, Single - Table Views, Views that Use Joins, Creating Views with Subqueries, Using Other Join Operations in views, Nesting Views, Updating Views, tasks you can accomplish with views. The SQL Security Model: Overview of Database Security, Creating Database Users, Database elements, Using GRANT and REVOKE, Security Roles, Views and Database Security. Real-World Issues Handling Specific Types of Data: Numeric Data Types, String Data Types, dealing with Dates, Converting Data Between Types. Database Performance and Integrity: Improving Database Performance, Performance Measurement Tools, Indexes, The Query Optimizer, Data Integrity, Integrity Versus Performance. Transactions and Cursors: Transactions, Using Transactions in Oracle, Using Transact – SQL, Using Cursors in Oracle PL/SQL.

UNIT IV – STORED PROCEDURE 9

Writing Stored Procedures: Writing a Stored Procedure, working with Variables, Defining Blocks of Code, Conditional Statements Using IF, using Loops, Loop over a Cursor, Triggers. More on Transact – SQL Stored Procedures: General Transact-SQL Programming Information, Global Variables, Using RETURN to Leave Stored Procedures, Handling Errors, Using Temporary Objects, WAITFOR, Advanced Trigger-Writing Techniques. Writing Oracle PL/SQL stored procedures: The Declaration Section, The Executable Section, Exception Handling, Writing Stored Procedures, Creating and Using Custom Functions, Bundling Procedures and Functions in Packages, Debugging PL/SQL Queries, Triggers.

UNIT V – Advanced Databases 9

Window Functions – Pivoting Data in SQL – Rows into Column – Column into Rows- Performance Tuning. Emerging Databases: NoSQL – CAP Theorem – Sharding - Document based – MongoDB Operation: Insert, Update, Delete, Query, Indexing, Application, Replication, Sharding, Deployment – Using MongoDB with PHP / JAVA – Advanced MongoDB Features – Cassandra: Data Model, Key Space, Table Operations, CRUD Operations, CQL Types – HIVE: Data types, Database Operations, Partitioning – HiveQL – OrientDB Graph database – OrientDB Features - XML Database: XML – XML Schema – XML DOM and SAX Parsers – XSL – XSLT – XPath and XQuery.

List of Experiments:

1. Consider the following Order Table:

SALESMAN (Salesman_id, Name, City, Commission, Customer_name, Customer_City, Purchase_amt, Purchase_date)

Perform the following DDL commands in SQL:

i. Creating a database

ii. Viewing all tables in a database

iii. Creating tables (with and without constraints)

iv. Altering tables (with ADD/MODIFY keywords)

v. Dropping a table/database

vi. Truncating a table/database

vii. Renaming a table/database

2. Consider the following College database:

STUDENT (Stud_id, Name, Age, Address, Phone_no, Email_ID);

SEMESTER (Stud_id, Sem_id, Degree, Year_of_Adm)

SUBJECT (Stud_id, Subject_code, Sub_title,Semester,Credits)

Perform the following DCL and TCL commands in SQL:

i. Commit

ii. Rollback

iii. Save Point

iv. Grant

v. Revoke

3. Consider the following Book Table:

BOOK (Book_id, Title, Author_Name, Publisher_Name, Pub_Year, No_of_Copies)

Perform the following DML commands in SQL:

i. Inserting records

ii. Updating the existing records

iii. Deleting the specific records

iv. Selecting records from the existing table

4. For a given set of relational database, create tables and perform the following SQL queries:

SALESMAN (Salesman_id, Name, City, Commission)

CUSTOMER (Customer_id, Cust_Name, City, Grade, Salesman_id)

ORDERS (Ord_No, Purchase_Amt, Ord_Date, Customer_id, Salesman_id)

i. Simple Queries with Select - Where – Between – Like – Distinct clauses.

ii. Simple Queries with Aggregate functions

iii. Queries with group by and having clause

iv. Queries involving - Date Functions, String Functions, Math Functions

v. Sort records with Order by clause.

5. For a given set of relational database for Movie information, create tables and perform the following SQL queries:

ACTOR (Act_id, Act_Name, Act_Gender)

DIRECTOR (Dir_id, Dir_Name, Dir_Phone)

MOVIES (Mov_id, Mov_Title, Mov_Year, Mov_Lang, Act_id, Dir_id)

MOVIE_CAST (Act_id, Actor_name, Mov_id, Role)

RATING (Mov_id, Rev_Stars).

i. Subqueries- With ANY, SOME and ALL clauses.

ii. Join Queries- Inner Join, Outer Join, Left Join, Right Join, Self-Join Subqueries- With IN and NOT IN clause, With EXISTS and NOT EXISTS clause.

iii. Create a simple view that shows all movie records of a particular director.

iv. Create a complex view that shows all movie records of a particular film with highest rating and actor name.

v. Update/Drop the existing views.

6. Create a PL/SQL Cursor to generate student grade calculation.

7. Consider the following Bank table:

Bank (cid, cname, add, accno, acctype, bankname, dep_amt, bal_amt).

i. Write a PL/SQL procedure which accept the account number of a customer and retrieve the balance.

ii. Write a PL/SQL updated trigger on Bank table. The system should keep track of the records that are being updated.

TOTAL = 45 Hours

COURSE OUTCOMES:

This course will enable the student to:

- Design a simple database with DDL and DML commands.
- Write subqueries and join operations for retrieving data from various tables.
- Enforce the security features in multiuser database environment.
- Use NoSQL database systems and manipulate the data associated with it.
- Create PL/SQL blocks for cursors, triggers and stored procedures.

REFERENCES

1. Steve Tale, “SQL: The Ultimate Beginners Guide: Learn SQL Today”, CreateSpace Independent Publishing Platform, 2016.
2. Abraham Silberschatz, Henry F. Korth, S. Sudharshan, “Database System Concepts”, 6th edition, Tata McGraw Hill, 2011
3. A. Silberschatz, H. Korth, S. Sudarshan, Database system concepts, 5/e, McGraw Hill, 2008.
4. Michael McLaughlin, John Harper, “Oracle Database 11g PL/SQL Programming Workbook, ISBN 9780070702264, TMH.
5. Satish Asnani ,”Oracle database 11g : hands on SQL/PL SQL”,EEE edition, PHI.
6. Brad Dayley, “Teach Yourself NoSQL with MongoDB in 24 Hours”, Sams Publishing, First Edition, 2014.
7. ShashankTiwari, “Professional NoSQL”, O’Reilly Media, First Edition, 2011.
8. Vijay Kumar, “Mobile Database Systems”, John Wiley & Sons, First Edition, 2006.

MCAADDAXX - PYTHON PROGRAMMING LABORATORY

L	T	P	C	M
0	0	2	1	100

COURSE OBJECTIVES:

This course will enable the student to:

- Understand programming skills in core Python.
- Define Object Oriented concepts in Python.
- Learn the skill of designing Graphical user Interfaces in Python.
- Write database applications in Python.
- Plan the operation required in data analytics.

List of Experiments

Experiment 1 - Basics

a) Running instructions in Interactive interpreter and a Python Script

b) Write a program to purposefully raise Indentation Error and Correct it

Experiment 2 - Operations

b) Write a program add.py that takes 2 numbers as command line arguments and prints its sum.

Experiment - 3 Control Flow

a) Write a Program for checking whether the given number is an even number or not.

b) Write a program using for loop that loops over a sequence.

c) Find the sum of all the primes below two million. Each new term in the Fibonacci sequence is generated by adding the previous two terms. By starting with 1 and 2, the first 10 terms will be: 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, ...

Experiment 4 - DS - DICTIONARY

a) Write a program to count the numbers of characters in the string and store them in a dictionary data structure

b) Write a program to use split and join methods in the string and trace a birthday with a dictionary data structure.

Experiment - 5 DS – Lists

a) Write a program combine_lists that combines these lists into a dictionary.

b) Write a program to count frequency of characters in a given file. Can you use character frequency to tell whether the given file is a Python program file, C program file or a text file?

Experiment - 6 Files

a) Write a program to print each line of a file in reverse order.

b) Write a program to compute the number of characters, words and lines in a file.

Experiment - 7 Functions

a) Find mean, median, mode for the given set of numbers in a list.

b) Write a function ball_collide that takes two balls as parameters and computes if they are colliding. Your function should return a Boolean representing whether or not the balls are colliding. Hint: Represent a ball on a plane as a tuple of (x, y, r), r being the radius. If (distance between two balls centers) <= (sum of their radii) then (they are colliding)

Experiment - 8 Functions – Strings

a) Write a function nearly_equal to test whether two strings are nearly equal. Two strings a and b are nearly equal when a can be generated by a single mutation on b.

b) Write a function dups to find all duplicates in the list.

c) Write a function unique to find all the unique elements of a list.

Experiment - 9 - Functions - Problem Solving

a) Write a function cumulative_product to compute cumulative product of a list of numbers.

b) Write a function reverse to reverse a list. Without using the reverse function.

c) Write function to compute gcd, lcm of two numbers. Each function shouldn't exceed one line.

Experiment 10 - Multi-D Lists

a) Write a program that defines a matrix and prints

b) Write a program to perform addition of two square matrices

c) Write a program to perform multiplication of two square matrices

Experiment 11 - Modules

a) Install packages requests, flask and explore them using (pip)

b) Write a script that imports requests and fetch content from the page.

c) Write a simple script that serves a simple HTTP Response and a simple HTML Page

Experiment 12- OOP

a) Class variables and instance variable and illustration of the self-variable i) Robot ii) ATM Machine

b) Apply Inheritance and overloading concepts.

Experiment 13 - GUI, Graphics

a) Write a GUI for an Expression Calculator using tk

b) Write a program to visualize the graphs and charts.

Experiment 14 - Testing

a) Write a test-case to check the function even _numbers which return True on passing a list of all even numbers

b) Write a test-case to check the function reverse_string which returns the reversed string

Experiment 15 – Connect with Database

a) Data Aggregation and GroupWise Operations

COURSE OUTCOMES:

This course will enable the student to:

- Understand and comprehend the basics of python programming.
- Demonstrate the principles of structured programming and be able to describe, design, implement, and test structured programs using currently accepted methodology.
- Explain the use of the built-in data structures list, sets, tuples and dictionary.
- Make use of functions and its applications.
- Identify real-world applications using oops, files and exception handling provided by python

Semester II

MCAADDAXX - PROBABILITY AND STATISTICS

L	T	P	C	M
3	**0**	**0**	**3**	**100**

COURSE OBJECTIVES:

This course will enable the student to:

- Understand the basic concepts in probability theory and statistical analysis.
- Learn the fundamental theory of distribution of random variables, the basic theory and techniques of parameter estimation and tests of hypotheses.
- Apply calculators and tables to perform simple statistical analyses for small samples
- Use popular statistics packages, such as SPSS, S-Plus, R or Mat Lab, to perform simple and sophisticated analyses for large samples.

UNIT I – INTRODUCTION TO PROBABILITIES 9

About Data: Data definition - Raw and Processed data; Data Types (NOIR) - Descriptive Stats: Measure of Central Tendency - Measure of Dispersion - Measure of Association - Sample space and events, Probability – Basic terminology - Rules and Events The axioms of probability - Some Elementary theorems - Conditional probability, Baye's theorem, Random variables, Discrete and continuous distributions - Distribution function.

UNIT II – DATA DISTRIBUTION 9

Data Distribution: Skewness - t-Distribution – Distribution Functions: Uniform Distribution Binomial, Poisson, normal distribution, Geometric Distribution, Gaussian Distribution - related properties, Standard Normal Distribution - Moment generating function, Moments of standard distributions – properties - Central Limit Theorem.

UNIT III – SAMPLES AND TECHNIQUES 9

Population and samples, Sampling distribution of mean (with known and unknown variance), proportion, variances, Sampling distribution of sums and differences, Point and interval estimators for means, variances, proportions - Sampling techniques: Random Sampling; Stratified Sampling.

9

UNIT IV – HYPOTHESIS AND STATISTICAL TESTING

Inferential Stats: Estimation technique - Hypothesis Testing (t-statistic calculations) - Statistical Hypothesis – Errors of Type I and Type II errors and calculation, One tail, two-tail tests, Testing hypothesis concerning means, proportions and their differences using Z-test, Tests of hypothesis using Student's t-test, F-test and test. Test of independence of attributes, ANOVA for one-way and two-way classified data - Chi-Square.

UNIT V – ADVANCED STATISTICAL METHODS 9

Statistical Quality Control methods, Methods for preparing control charts, Problems using x-bar, p, R charts and attribute charts, Simple Correlation and Regression, Queuing Theory: Pure Birth and Death Process M/M/1 Model and Simple Problems.

TOTAL = 45 Hours

COURSE OUTCOMES:

This course will enable the student to:

- Demonstrate the basic knowledge on fundamental probability concepts, including random variable, probability of an event, additive rules and conditional probability.
- Derive the probability density function of transformations of random variables and use these techniques to generate data from various distributions
- Demonstrate the basic statistical concepts and measures
- Discuss several well-known distributions, including Binomial, Geometrical, Negative Binomial, Normal and Exponential Distribution
- Prove the model by hypotheses testing.

REFERENCES

1. Probability and Statistics for Engineers, Miller and John E. Freund, Prentice Hall of India
2. Probability and Statistics, D. K. Murugeson & P. Guru Swamy, Anuradha Publishers
3. Probability, Statistics and Random processes. T. Veerrajan, Tata Mc.Graw Hill, India
4. Probability, Statistics and Queuing theory applications for Computer Sciences, 2 Ed, Trivedi, John Wiley

MCAADDAXX - R FOR DATA ANALYTICS

L	T	P	C	M
3	0	0	3	100

COURSE OBJECTIVES:

This course will enable the student to:

- Understand the basics in R programming in terms of constructs, control statements, string functions
- Understand the use of R for Big Data analytics
- Learn to apply R programming for Text processing

- Apply the R programming from a statistical perspective

UNIT I – INTRODUCTION 9

Introducing to R – Installation of Libraries; Constants and Variables; Numbers; R Data Structures – Help functions in R – Vectors: Numeric Vectors - Scalars – Declarations – recycling – Vectorized operation: Using all and any, NA and NULL values, Filtering, Vectorized if-then else, Vector Equality, Vector Element names – Arithmetic and Boolean operations – conditional and loop statement in R – Functions and Recursions in R – Packages in R.

UNIT II – MATRICES, ARRAYS AND LISTS 9

Creating matrices – Matrix operations – Applying Functions to Matrix Rows and Columns: Adding and deleting rows and columns – Higher Dimensional arrays - Vector/Matrix Distinction – Avoiding Dimension Reduction - Characters and Strings - String vector - String operations and functions – List – Creating lists – General list operations – Accessing list components and values – applying functions to lists – recursive lists – Different R operations using a List, matrix, Array;

UNIT III – DATA FRAMES 9

Overview on Data Frames – Create it in scratch - Matrix-like operations in frames – Merging Data Frames – Applying functions to Data frames – Factors and Tables – factors and levels – Common functions used with factors – Working with tables - Math and Simulations in R - Reading a datafile directly into a dataframe - EDA using R - Reading different file formats.

UNIT IV – OOPS AND STATISTICAL MODELS 9

S3 Classes – S4 Classes – Managing your objects – Input/Output – accessing keyboard and monitor – reading and writing files – accessing the internet – String Manipulation – Statistical analysis: Basic Statistics – Linear Model – Generalized Linear models – Non-linear models - R functions for statistical analysis - Graphics: Creating Graphs – Customizing Graphs – Saving graphs to files – Creating three-dimensional plots – interfacing: Interfacing R to other languages – Parallel R – Time Series and Auto-correlation – Clustering.

UNIT V – VISUALIZATION AND LEARNING TECHNIQUES 9

Introduction to GGPlot2 – Library - Factors – Aesthetics – Plotting with Layers – Overriding Aesthetics – Mapping vs Setting – Histograms – Density Charts – Statistical Transformation – Facets – Coordinates – Themes. Learning Techniques - Supervised Learning: Linear Regression; Logistic Regression; Decision Trees; Random Forests; K-Nearest Neighbours (k-NN); Supprt Vector Machine. Unsupervised Learning: K-Means Clustering; Hierarchical clustering.

TOTAL = 45 Hours

COURSE OUTCOMES:

This course will enable the student to:

- Describe the features of R Programming.
- Use the various data structures in R.
- Apply data frames, control statements and functions for the simulation.
- Develop Oops based classes and apply graphic techniques.
- Identify the statistical methods applied in R.

REFERENCES

1. Norman Matloff , “The Art of R Programming: A Tour of Statistical Software Design”, No Starch Press, 2011
2. Jared P. Lander, “R for Everyone: Advanced Analytics and Graphics”, Addison-Wesley Data & Analytics Series, 2013.
3. Mark Gardener, “ Beginning R – The Statistical Programming Language”, Wiley, 2013
4. Robert Knell, “Introductory R: A Beginner's Guide to Data Visualisation, Statistical Analysis and Programming in R”, Amazon Digital South Asia Services Inc, 2013.

MCAADDAXX - R PROGRAMMING LABORATORY

L	T	P	C	M
0	0	2	1	100

COURSE OBJECTIVES:

This course will enable the student to:

- Understand the fundamentals of R Programming.
- Use mathematical and statistical manipulations in R using the functions that perform the specialized task in R.
- Design and manage the various data structures, frames in R.
- Illustrate the advanced statistical methods in R.

Experiments:

1. Creating and displaying Data.
2. Use of R as a calculator, functions and matrix operations, missing data and logical operators.
3. Conditional executions and loops, data management with sequences.
4. Creating and manipulating a List, Array and Strings - Data management with display paste, split, find and replacement, manipulations with alphabets, evaluation of strings, data frames.
5. Creating a Data Frame and Matrix-like Operations on a Data Frame, Merging two Data Frames
6. Applying functions to Data Frames, import of external data in various file formats, statistical functions, compilation of data.

7. Using Functions with Factors
8. Accessing the Internet
9. Visualization Effects
10. Plotting with Layers
11. Overriding Aesthetics
12. Histograms and Density Charts
13. Simple Linear Regression – Fitting, Evaluation and Visualization
14. Multiple Linear Regression, Lasso and Ridge Regression
15. Use the following scenarios:
    a. Use the Diabetes data set from UCI and Pima Indians Diabetes data set for per forming the following:
        i. Univariate Analysis: Frequency, Mean, Median, Mode, Variance, Standard Deviation, Skewness and Kurtosis.
        ii. Bivariate Analysis: Linear and logistic regression modeling.
        iii. Multiple Regression Analysis
        iv. Also Compare the results of the above analysis for the two data sets.
    b. Data Modelling
        i. Apply Bayesian and SVM techniques on Iris and Diabetes data set.
        ii. Apply and explore various plotting functions on UCI data sets.

Semester III

MCAADDAXX - BIG DATA AND HADOOP

L	T	P	C	M
3	0	0	3	100

COURSE OBJECTIVES:

This course will enable the student to:

- Understand the competitive advantages of big data analytics
- Understand the big data frameworks
- Learn data analysis methods
- Learn stream computing
- Gain knowledge on Hadoop related tools such as HBase, Cassandra, Pig, and Hive for big data analytics

UNIT I – INTRODUCTION TO BIG DATA 9

Big Data – Definition, Characteristic Features – Big Data Applications - Big Data vs Traditional Data - Risks of Big Data - Structure of Big Data - Challenges of Conventional Systems - Web Data – Evolution of Analytic Scalability - Evolution of Analytic Processes, Tools and methods - Analysis vs Reporting - Modern Data Analytic Tools.

UNIT II – HADOOP FRAMEWORK 9

Distributed File Systems - Large-Scale File System Organization – Hadoop Eco-System - HDFS Architecture - HDFS concepts - MapReduce Execution, Algorithms using MapReduce, Matrix-

Vector Multiplication – Data Serialization - Hadoop YARN – Use Cases.

UNIT III – DATA ANALYSIS **9**

Statistical Methods: Regression modeling, Multivariate Analysis - Classification: SVM & Kernel Methods - Rule Mining - Cluster Analysis, Types of Data in Cluster Analysis, Partitioning Methods, Hierarchical Methods, Density Based Methods, Grid Based Methods, Model Based Clustering Methods, Clustering High Dimensional Data - Predictive Analytics – Data analysis using R.

UNIT IV – MINING DATA STREAMS **9**

Streams: Concepts – Stream Data Model and Architecture - Sampling data in a stream - Mining Data Streams and Mining Time-series data - Real Time Analytics Platform (RTAP) Applications - Case Studies - Real Time Sentiment Analysis, Stock Market Predictions.

UNIT V – BIG DATA FRAMEWORKS **9**

Introduction to NoSQL – Aggregate Data Models – HBase: Data Model and Implementations – Hbase Clients – Examples – .Cassandra: Data Model – Examples – Cassandra Clients – Hadoop Integration. Pig – Grunt – Pig Data Model – Pig Latin – developing and testing Pig Latin scripts. Hive – Data Types and File Formats – HiveQL Data Definition – HiveQL Data Manipulation – HiveQL Queries - Sqoop & Flume- Data Ingestion; Oozie;

TOTAL = 45 Hours

COURSE OUTCOMES:

This course will enable the student to:

- Understand how to leverage the insights from big data analytics
- Understand Hadoop frameworks to store and manage the big data.
- Analyze data by utilizing various statistical and data mining approaches
- Perform analytics on real-time streaming data
- Understand the various NoSQL alternative database models

REFERENCES

1. Bill Franks, ―Taming the Big Data Tidal Wave: Finding Opportunities in Huge Data Streams with Advanced Analytics‖, Wiley and SAS Business Series, 2012.
2. David Loshin, "Big Data Analytics: From Strategic Planning to Enterprise Integration with Tools, Techniques, NoSQL, and Graph", 2013.
3. Michael Berthold, David J. Hand, ―Intelligent Data Analysis‖, Springer, Second Edition, 2007.
4. Michael Minelli, Michelle Chambers, and Ambiga Dhiraj, "Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends for Today's Businesses", Wiley, 2013.

5. P. J. Sadalage and M. Fowler, "NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot Persistence", Addison-Wesley Professional, 2012.

6. Richard Cotton, "Learning R – A Step-by-step Function Guide to Data Analysis, , O‘Reilly Media, 2013.

MCAADDAXX - MACHINE LEARNING

L	T	P	C	M
3	0	0	3	100

COURSE OBJECTIVES:

This course will enable the student to:

- Understand the concepts of Machine Learning.
- Appraise the supervised learning techniques and their applications.
- Illustrate the concepts and algorithms of unsupervised learning.
- Describe the perceptrons algorithms.
- Explain the concepts and algorithms of advanced learning.

UNIT I – INTRODUCTION 9

Machine learning -Examples of Machine Learning applications-Learning Associations-Classification Regression-Unsupervised Learning-Reinforcement Learning-Supervised learning: Learning a class from Examples-Regression-Model Selection and Generalization- Case Study: Familiarity with R tool and Python programming language and libraries

UNIT II – LEARNING AND DECISION-TREE LEARNING 9

Concept Learning - Concept learning Task – Concept Learning as search –Finding a maximally specific hypothesis – Version Spaces and Candidate elimination Algorithm –Inductive Bias Decision Tree Learning - Decision Tree representation –Problems for Decision Tree Learning – Hypothesis Search space – Inductive Bias in Decision Tree Learning – Issues in Decision Tree Learning- Case Study: Implementation of decision tree algorithm for problems in Retail Domain

UNIT III – MULTILAYER PERCEPTRONS AND DEEP LEARNING 9

The Perceptron-Training a Perceptron-Learning Boolean Functions-Multilayer Perceptrons-MLP as Universal Approximator- Back propagation Algorithm-Training Procedures Convolution Networks –The Convolution Operation-Pooling-Convolution and Pooling as an infinitely strong prior –Variants of the Basic Convolution Function –Structured Outputs –Data Types –Efficient Convolution Algorithms –Random and Unsupervised features- Case Study: Implementation of Back propagation algorithm for problems in financial domain.

UNIT IV – CLUSTERING 9

Similarity-Based Clustering-Unsupervised learning problems-Hierarchical Agglomerative

Clustering (HAC)-Single-link, complete-link, group-average similarity- K-Means and Mixtures of Gaussians-Flat clustering k-Means algorithms-Mixture of Gaussian model-EM-algorithm for mixture of Gaussian model -Case Study: Implementation of clustering algorithm for problems in financial/insurance/health care domain

UNIT V – REINFORCEMENT LEARNING 9

Association Rule Learning – Apriori – Eclat - Reinforcement Learning - Learning task – Q learning – The Q function – Algorithm for Q learning –convergence – experimentation strategies – updating sequence –Non deterministic rewards and actions –Temporal difference learning – Generalizing from examples –relationship to dynamic programming – Upper Confidence Bound – Thompson Sampling. Case Study: Implementation of Q learning algorithm/reinforcement learning for problems in automotive domain/games

TOTAL = 45 Hours

COURSE OUTCOMES:

This course will enable the student to:

- Acquire Knowledge in various learning techniques like decision tree, Analytical, Inductive and Reinforced learning
- Design a learning model appropriate to the application.
- Use a tool to implement typical Clustering algorithms for different types of applications
- Identify and apply the appropriate machine learning techniques for classification, Pattern recognition, optimization and decision problems.
- Development of techniques in information science applications by applying Computational intelligence and appropriate machine learning techniques

REFERENCES

1. Ethem Alpaydin, "Introduction to Machine Learning", The MIT Press, September 2014, ISBN 978- 0-262-02818-9.(Units 1,3(Multilayer Perceptrons) & 4)
2. Mitchell, Tom, "Machine Learning", New York, McGraw-Hill, First Edition, 2003.(Units 2,5)
3. Ian GoodFellow,Yoshua Bengio,Aaron Courville ,"Deep Learning",MIT Press Book (Unit 3 - Convolutional Networks)
4. Christopher Bishop, “Pattern Recognition and Machine Learning” Springer, 2007.
5. Kevin P. Murphy, “Machine Learning: A Probabilistic Perspective”, MIT Press, 2012.
6. Ethem Alpaydin, “Introduction to Machine Learning”, MIT Press, Third Edition, 2014.
7. Stephen Marshland, "Machine Learning: An Algorithmic Perspective", Chapman & Hall/CRC 2009.
8. Mehryar Mohri, Afshin Rostamizadeh, Ameet Talwalkar, "Foundations of Machine Learning", MIT Press (MA) 2012.
9. Tom Mitchell, "Machine Learning", McGraw-Hill, 1997.

MCADA2211 - BIG DATA AND HADOOP LABORATORY

L	T	P	C	M
0	0	4	2	100

COURSE OBJECTIVES:

This course will enable the student to:

- Set up single and multi-node Hadoop Clusters.
- Solve Big Data problems using Map Reduce Technique.
- Design algorithms that use Map Reduce Technique to apply on Unstructured and structured data.
- Learn NoSQL query.
- Learn Scalable machine learning using Mahout.

Experiments:

1. Implement the following Data structures in Java a)Linked Lists b) Stacks c) Queues d) Set e) Map
2. Perform setting up and Installing Hadoop in its three operating modes: Standalone, Pseudo distributed, fully distributed and use web based tools to monitor your Hadoop setup.
3. Implement the following file management tasks in Hadoop:
   • Adding files and directories • Retrieving files • Deleting files
4. MapReduce Application for Word Counting, mines weather data and Implement Matrix Multiplication.
5. Write Pig Latin scripts to sort, group, join, project, and filter your data
6. Run Hive then use Hive to create, alter, and drop databases, tables, views, functions, and indexes
7. Unstructured data into NoSQL data and do all operations such as NoSQL query with API.
8. Page Rank Computation
9. Mahout machine learning library to facilitate the knowledge build up in big data analysis.
10. Application of Recommendation Systems using Hadoop/mahout libraries
11. Sentiment Analysis of Social Media Data
12. Log Analyzer Application development
13. Financial Fraud Detection
14. Targeted Market Campaign Planning
15. Customer 360 Degree View Generation

COURSE OUTCOMES:

This course will enable the student to:

- Set up single and multi-node Hadoop Clusters.
- Apply Map Reduce technique for various algorithms.
- Design new algorithms that use Map Reduce to apply on Unstructured and structured data.
- Develop Scalable machine learning algorithms for various Big data applications using Mahout.
- Represent NoSQL data

MCAADDAXX -- MACHINE LEARNING WITH PYTHON LABORATORY

L	T	P	C	M
0	0	4	2	100

COURSE OBJECTIVES:

This course will enable the student to:

- Understand the machine learning concepts in Python.
- Describe the data structures and series in Pandas.
- Apply dataframe manipulation operations.
- Explain visualization techniques in Python.
- Understand the statistical views in data analytics.

Experiments:

- **Numpy:** Functions to create array; Numpy operations - dtypes, size, shape, reshape, itemsize; Indexing array; Slicing array;
- **Pandas;** Python pandas; Data structures; Creating a series; Manipulating series; Usage if .loc and .iloc; Creating a dataframe
- **Dataframe Manipulation**: Manipulating dataframes; Indexing a dataframe; Read data from various sources; Concatenate the dataframes; Merge using inner join; Merge using outer join; Merge using right join; Merge using left join; Reshape using melt() function; Check for duplicates;
- **Visualization;** Plots using Matplotlib; Line plot; Scatter plot; Bar plot; Pie plot; Histogram; Box plot; Plots using Seaborn; Strip plot; Pair plot; Distribution plot; Count plot; Heatmap
- **EDA:** Summary Statistics; Missing Value Treatment; Dataframe analysis using groupby(); Advanced Data Explorations
- Program involving Regular Expressions
- The probability that it is Friday and that a student is absent is 3 %. Since there are 5 school days in a week, the probability that it is Friday is 20 %. What is the probability that a student is absent given that today is Friday? Apply Baye's rule in python to get the result.
- Extract the data from database using python
- Implement k-nearest neighbours classification using python
- Given the following data, which specify classifications for nine combinations of VAR1 and VAR2 predict a classification for a case where VAR1=0.906 and VAR2=0.606, using the result of k-means clustering with 3 means (i.e., 3 centroids)

VAR1	VAR2	CLASS
1.713	1.586	0
0.180	1.786	1
0.353	1.240	1
0.940	1.566	0
1.486	0.759	1
1.266	1.106	0
1.540	0.419	1
0.459	1.799	1
0.773	0.186	1

- The following training examples map descriptions of individuals in to high, medium and low credit-worthiness.

Income Range	**Sports Category**	**Work Role**	**Marital Status**	**Age group**	**Own Residential**	**Risk**
Medium	Skiing	Design	Single	Twenties	No	High Risk

High	Golf	Trading	Married	Forties	Yes	Low Risk
Low	Speedway	Transport	Married	Thirties	Yes	Medium Risk
Medium	Football	Banking	Single	Thirties	Yes	Low Risk
High	Flying	Media	Married	Fifties	Yes	High Risk
Low	Football	Security	Single	Twenties	No	Medium Risk
Medium	Golf	Media	Single	Thirties	Yes	Medium Risk
Medium	Golf	Transport	Married	Forties	Yes	Low Risk
High	Skiing	Banking	Single	Thirties	Yes	High Risk
Low	Golf	Unemployed	Married	Forties	Yes	High Risk

Find the unconditional probability of `golf' and the conditional probability of `single' given `medRisk' in the dataset?

- Implement linear regression using python.
- Implement Naïve Bayes theorem to classify the English text
- Implement an algorithm to demonstrate the significance of genetic algorithm
- Combine and merge the data sets.

COURSE OUTCOMES:

This course will enable the student to:

- Define arrays and matrices for an efficient calculation using Numpy packages.
- Apply pandas for data manipulation and analysis.
- Design a powerful data structures in Pandas.
- Create an attractive graphs for given data set in python.
- Analyze the data set using EDA technique in Python.

Semester IV

MCAADDAXX - DATA VISUALIZATION USING TABLEAU & POWER BI

L	T	P	C	M
2	0	2	3	100

COURSE OBJECTIVES:

This course will enable the student to:

- Understand how to accurately represent voluminous complex data set in web and from other data sources.
- Understand the methodologies used to visualize large data sets.
- Understand the various processes involved in data visualization.
- Get used to with using interactive data visualization.
- Understand the different security aspects involved in data visualization.

UNIT I – INTRODUCTION 9

Context of data visualization – Definition, Methodology, Visualization design objectives. Key Factors – Purpose, visualization function and tone, visualization design options – Data representation, Data Presentation, Seven stages of data visualization, widgets, data visualization tools.

UNIT II – TOOLS AND CASE STUDIES 9

Tableau - Intro to Tableau Interface - Connecting to Data - Visual Analytics – Mapping – Calculations - Dashboard and Stories - Power BI - PowerBI - Visualisation with BI - Data Analysis Expressions.

UNIT III - VISUALIZING DATA METHODS

Mapping - Time series - Connections and correlations – Indicator-Area chart-Pivot table- Scatter charts, Scatter maps - Tree maps, Space filling and non-space filling methods-Hierarchies and Recursion - Networks and Graphs-Displaying Arbitrary Graphs-node link graph-Matrix representation for graphs- Info graphics.

UNIT IV – DATA PROCESSING TECHNIQUES 9

Acquiring data, - Where to Find Data, Tools for Acquiring Data from the Internet, Locating Files for Use with Processing, Loading Text Data, Dealing with Files and Folders, Listing Files in a Folder ,Asynchronous Image Downloads, Advanced Web Techniques, Using a Database, Dealing with a Large Number of Files. Parsing data - Levels of Effort, Tools for Gathering Clues, Text Is Best, Text Markup Languages, Regular Expressions (regexps), Grammars and BNF Notation, Compressed Data, Vectors and Geometry, Binary Data Formats, Advanced Detective Work.

UNIT V – INTERACTIVE DATA VISUALIZATION AND SECURITY MECHANISM 9

Drawing with data – Scales – Axes – Updates, Transition and Motion – Interactivity - Layouts –

Geo-mapping – Exporting, Framework – T3, .js, tablo - Port scan visualization - Vulnerability assessment and exploitation - Firewall log visualization - Intrusion detection log visualization - Attacking and defending visualization systems – Creating security visualization system.

List of Experiments:

- Histograms, Density Charts, Simple and Multiple Bar Charts, Pie Charts
- Factors and Aesthetics, Plotting with Layers and Overriding Aesthetics
- Facets and Themes
- Plotting and Controlling Graphs, Adding Text to Graphs and Getting and Setting Values in Graphs
- Load, prepare, model Data and design reports in Power BI/Tableau
- Create a sales report dashboard with three pages (Overview, Profit and My Performance) in Power BI/Tableau.

TOTAL = 45 Hours

COURSE OUTCOMES:

This course will enable the student to:

- Define complex and voluminous data.
- Design and use various methodologies present in data visualization.
- Apply the various process and tools used for data visualization.
- Use interactive data visualization to make inferences.
- Analyze the process involved and security issues present in data visualization.

REFERENCES

- Scott Murray, “Interactive data visualizationfor the web”, O‟Reilly Media, Inc., 2013.
- Ben Fry, “Visualizing Data”, O‟Reilly Media, Inc., 2007.
- Greg Conti, “Security Data Visualization: Graphical Techniques for Network Analysis”, NoStarch Press Inc, 2007.

MCAADDAXX - BIG DATA ANALYTICS THROUGH SPARK

L	T	P	C	M
2	0	2	3	100

COURSE OBJECTIVES:

This course will enable the student to:

- Understand the Spark Ecosystem.
- Apply spark transformation in Python.
- Write SQL operations for Spark based data set.
- Overview the streaming process in real time analytics.
- Explain Machine Learning concepts with Spark.

UNIT I – INTRODUCTION TO SPARK 9

Apache Spark Ecosystem - Setting up the Spark Python Environment – Execution of a PySpark Program – Resilient Distributed Datasets – Spark Architecture – Spark Project Workflow.

UNIT II – SPARK PROGRAMMING WITH PYTHON 9

Loading and Storing Data – Transformations – Actions – Key-Value Resilient Distributed Datasets – Local Variables – Broadcast Variables – Accumulators – Partitioning – Persistence.

UNIT III – SPARK SQL 9

Overview of Spark SQL – Spark Session – Data Frames – Schema of a Data Frame – Operations supported by Data Frames – Filter, Join, GroupBy, Agg operations – Nesting the Operations – Temporary Tables – Viewing and Querying Temporary Tables.

UNIT IV – SPARK STREAMING 9

Use Cases for Realtime Analytics – Transferring, Summarizing, Analyzing Real time data – Data Sources supported by Spark Streaming – Flat files, TCP/IP – Flume – Kafka – Kinesis – Streaming Context – DStreams – Dstream RDDs – Dstream Processing.

UNIT V – MACHINE LEARNING WITH SPARK 9

GraphX - Spark ML – MLib – SparkR - Notebooks with Spark – Jupyter – Zeppelin - Linear Regression – Decision Tree Classification – Principal Component Analysis – Random Forest Classification – Text Pre-processing with TF-IDF – Naïve Bayes Classification – K-Means Clustering – Recommendation Engines.

List of Experiments:

- Program involving Resilient Distributed Datasets
- Program involving Transformations and Actions
- Program involving Key-Value Resilient Distributed Datasets
- Program involving Local Variables, Broadcast Variables and Accumulators
- Program involving Filter, Join, GroupBy, Agg operations
- Viewing and Querying Temporary Tables
- Transferring, Summarizing and Analysing Twitter data
- Program involving Flume, Kafka and Kinesis
- Program involving DStreams and Dstream RDDs
- Linear Regression
- Decision Tree Classification
- Principal Component Analysis
- Random Forest Classification
- Text Pre-processing with TF-IDF
- Naïve Bayes Classification
- K-Means Clustering

TOTAL = 45 Hours

COURSE OUTCOMES:

This course will enable the student to:

- Explain the spark architecture and workflow.
- Prepare data sets and write spark programming in python.
- Apply basic operation required for data analytics in spark.
- Analyze the streaming process in real time scenario.
- Apply machine Learning concepts in Spark.

REFERENCES

1. Tomasz Drabos, “Learning PySpark”, PACKT, 2017.
2. Padma Priya Chitturi, “Apache Spark for Data Science”, PACKT, 2017.
3. Holden Karau, “ Learning Spark”. PACKT, 2016.
4. Sandy Riza, “Advanced Analytics with Spark”, O’ Reilly, 2016.
5. Romeo Kienzler, “Mastering Apache Spark”, PACKT, 2017.