INTEGRATED COURSE FOR UNIVERSITY WITH SAMATRIX

COURSES PROPOSED

3 Year BCA with Data Analytics & visualization

Semester	Name of Course	Total Credits
Sem 1	Overview of Data Science and Foundation of Data Analysis	3
Sem 2	Data Analysis Using Python	3
Sem 3	Probabilistic Modelling and Reasoning	3
Sem 3	R Programming for Data Science and Analysis	3
Sem 4	Foundation of Machine Learning and Pattern Recognition	3
Sem 4	Data Visualization using Tableau or other advance-Tools & Techniques	3
Sem 5	Big Data Analytics with Hadoop	3
Sem 5	Scala for Data Science	3
Sem 6	Project	6

INTEGRATED B TECH COURSE IN UNIVERSITY WITH SAMATRIX

SEMESTER: 1	Credit: 2-0-1
Software Req: MS Office 2013/2016 Version	Hours: 4 per Week

OVERVIEW OF DATA SCIENCE AND FOUNDATION OF DATA ANALYTICS

Objectives: The objective of this course is to teach students the vital data science concepts and tasks that occupy the data scientist

UNIT – I

Introduction to Data Science: Defining Data Science and Big Data, Benefits and Uses of Data Science and Big Data, Facets of Data, Structured Data, Unstructured Data, Natural Language, Machine-generated Data, Graph based or Network Data, Audio, Image, Video, Streaming data, Data Science Process, Big data ecosystem and data science, Distributed file systems, Distributed programming framework, data integration framework, machine learning framework, No SQL Databases, scheduling tools, benchmarking tools, system deployments

UNIT – II

Data Science Processes: Six steps of data science processes, define research goals, data retrieval, cleansing data, correct errors as early as possible, integrating – combine data from different sources, transforming data, exploratory data analysis, Data modelling, model and variable selection, model execution, model diagnostic and model comparison, presentation and automation.

UNIT – III

Introduction to Machine Learning: What is Machine Learning, Learning from Data, History of Machine Learning, Big Data for Machine Learning, Leveraging Machine Learning, Descriptive vs Predictive Analytics, Machine Learning and Statistics, Artificial Intelligence and Machine Learning, Types of Machine Learning – Supervised, Unsupervised, Semi-supervised, Reinforcement Learning, Types of Machine Learning Algorithms, Classification vs Regression Problem, Bayesian, Clustering, Decision Tree, Dimensionality Reduction, Neural Network and Deep Learning, Training machine learning systems

UNIT – IV

Introduction to AI: What is AI, Turing test, cognitive modelling approach, law of thoughts, the relational agent approach, the underlying assumptions about intelligence, techniques required to solve AI problems, level of details required to model human intelligence, successfully building an intelligent problem, history of AI

Introduction to Data Handling: Introduction to Data and their uses, Overview of Data analysis, Ribbon|Workbook|Worksheets|Format Cells|Find & Select|Templates|Data Validation|Keyboard Shortcuts|Print|Share|Protect, Working with statistical formulas & functions, Data Validation & data models, Data Analysis using statistical methods

Samatrix.io

DATA ANALYSIS USING PYTHON

SEMESTER: 2	Credit: 2-0-1
Software: Python, NumPy, Pandas, Matplotlib, Seaborn, SciPy	No of Hours : 4 Per Week

Objectives: The objective of this course is to teach students the concepts of Python Programming Language with Libraries

UNIT – I

Python programming Basic: Python interpreter, IPython Basics, Tab completion, Introspection, %run command, magic commands, matplotlib integration, python programming, language semantics, scalar types. Control flow

Data Structure, functions, files: tuple, list, built-in sequence function, dict, set, functions, namescape, scope, local function, returning multiple values, functions are objects, lambda functions, error and exception handling, file and operation systems

UNIT – II

NumPy: Array and vectorized computation: Multidimensional array object. Creating ndarrays, arithmetic with numpy array, basic indexing and slicing, Boolean indexing, transposing array and swapping axes, universal functions, array-oriented programming with arrays, conditional logic as arrays operations, file input and output with array

Pandas: Pandas data structure, series, DataFrame, Index Object, Reindexing, dropping entities from an axis, indexing, selection and filtering, integer indexes, arithmetic and data alignment, function application and mapping, soring and ranking, correlation and covariance, unique values, values controls and membership, reading and writing data in text format

UNIT -III

Visualization with Matplotlib: Figures and subplots, colors, markers, line style, ticks, labels, legends, annotation and drawing on sublots, matplotlib configuration

Plotting with pandas and seaborn: line plots, bar plots, histogram, density plots, scatter and point plots, facet grids and categorical data

Samatrix.io

PROBABILISTIC MODELLING AND REASONING WITH PYTHON

SEMESTER: 3	Credit: 2-0-1
Software: Python, NumPy, Pandas, Matplotlib, Seaborn, SciPy	No of Hours: 4 per week

Objectives: The objective of this course is to teach students the concepts of Statistics, probability, probability distribution, and other statistical methods to solve various engineering problems

UNIT – I

Introduction to Statistics: Introduction to Statistics. Role of statistics in scientific methods, current applications of statistics.

Scientific data gathering: Sampling techniques, scientific studies, observational studies, data management.

Data description: Displaying data on a single variable (graphical methods, measure of central tendency, measure of spread), displaying relationship between two or more variables, measure of association between two or more variables.

UNIT – II

Probability Theory: Sample space and events, probability, axioms of probability, independent events, conditional probability, Bayes' theorem.

Random Variables: Discrete and continuous random variables. Probability distribution of discrete random variables, binomial distribution, poisson distribution. Probability distribution of continuous random variables, The uniform distribution, normal (gaussian) distribution

UNIT -III

Point Estimations: Methods of finding estimators, method of moments, maximum likelihood estimators, bayes estimators. Methods of evaluating estimators, mean squared error, best unbiased estimator, sufficiency and unbiasedness

Interval Estimations: Confidence interval of means and proportions, Distribution free confidence interval of percentiles

UNIT - IV

Test of Statistical Hypothesis and p-values: Tests about one mean, tests of equality of two means, test about proportions, p-values, likelihood ratio test, Bayesian tests

Univariate Statistics using Python: Mean, Mode. Median, Variance, Standard Deviation, Normal Distribution, t-distribution, interval estimation, Hypothesis Testing, Pearson correlation test, ANOVA F-test

Samatrix.io

R PROGRAMMING FOR DATA SCIENCE AND DATA ANALYSIS

SEMESTER: 3	Credit: 2-0-1
Software Req: R Studio	Hours: 4 per Week

Objectives: The objective of this course is to teach students R Programming Language, basic functions in R programming language and critical techniques

UNIT – I

Getting Started with R and R Workspace: Introducing R, R as a programming Language, the need of R, Installing R, RStudio, RStudio's user interface, console, editor, environment pane, history pane, file pane, plots pane, package pane, help and viewer pane

R Workspace, R's working directory, R Project in R Studio, absolute and relative path, Inspecting an Environment, Inspect existing Symbols, View the structure of object, Removing symbols, Modifying Global Options, Modifying warning level, Library of Packages, Getting to know a package, Installing a Package from CRAN, Updating Package from CRAN, Installing package from online repository, Package Function, Masking and name conflicts

UNIT – II

Basic Objects and Basic Expressions: Vectors, Numeric Vectors, Logical Vectors, Character Vectors, subset vectors, Named Vectors, extracting element, converting vector, Arithmetic operators, create Matrix, Naming row and columns, subsetting matrix, matrix operators, creating and subsetting an Array, Creating a List, extracting element from list, subsetting a list, setting value, creating a value of data frame, subsetting a data frame, setting values, factors, useful functions of a data frame, loading and writing data on disk, creating a function, calling a function, dynamic typing, generalizing a function. Assignment Operators, Conditional Expression, using if as expression and statement, using if with vectors, vectorized if: ifelse, using switch, using for loop, nested for loop, while loop

UNIT – III

Working with Basic Objects and Strings: Working with object function, getting data dimensions, reshaping data structures, iterating over one dimension, logical operators, logical functions, dealing with missing values, logical coercion, math function, number rounding functions, trigonometric functions, hyperbolic functions, extreme functions, finding roots, derivatives and integration, Statistical function, sampling from a vector, Working with random distributions, computing summary statistics, covariance and correlation matrix, printing string, concatenating string, transforming text, Formatting text, formatting date and time, formatting date and time to string, finding string pattern, using group to extract data, reading data

UNIT – IV

Working with Data – Visualize and Analyze Data: Reading and Writing Data, importing data using built-in-function, READR package, export a data frame to file, reading and writing Excel worksheets, reading and writing native data files, loading built-in data sets, create scatter plot, bar chart, pie chart, histogram and density plots, box plot, fitting linear model and regression tree

Samatrix.io

FOUNDATION OF MACHINE LEARNING AND PATTERN RECOGNITION

SEMESTER: 4	Credit: 2-0-1
Software: Python, NumPy, Pandas, Matplotlib, Seaborn, SciPy, Scikit-Learn	No of Hours: 4 per week

Objectives: The objective of this course is to teach students the basic concepts of machine learning, supervised learning, unsupervised learning, and reinforcement learning

UNIT – I

Introduction: Learning systems, real world applications of machine learning, why machine learning, variable types and terminology, function approximation

Types of machine learning: Supervised learning, unsupervised learning, reinforcement learning

Important concepts of machine learning: Parametric vs non-parametric models, the trade-off between prediction accuracy and model interpretability, the curse of dimensionality, measuring the quality of fit, bias-variance trade off, overfitting, model selection, no free lunch theorem

UNIT – II

Linear Regression: Linear regression, estimating the coefficients, accessing the accuracy of coefficient estimates, accessing the accuracy of the model

Classification: Logistic regression, estimating regression coefficients, making predictions, multiple logistic regressions, linear discriminant analysis, bayes' theorem of classification,

UNIT – III

Resampling Methods, Model Selection and Regularization: Cross-validation, leave-one-out cross-validation, k-fold cross-validation, the bootstrap, subset selection, shrinkage methods, ridge and lasso regression, dimension reduction methods, principal components regression

Tree Based Methods: Advantages and disadvantages of trees, regression Trees, classification trees, bagging, random forest, boosting

UNIT – IV

Support Vector Machine: Maximum margin classifier, classification using a separating hyperplane, the maximal margin classifier, support vector classifier, support vector machines, classification with non-linear decision boundaries, support vector machine

Unsupervised Learning: Principle component analysis, what are principal components, clustering methods, k-means clustering, hierarchical clustering,

Samatrix.io

DATA VISUALIZATION USING TABLEAU

SEMESTER: 4	Credit: 2-0-1
Software: Microsoft Office 2013 or 2016, Tableau Desktop, Power BI	No of Hours : 4 Per Week

UNIT - I

INTRODUCTION TO DATA HANDLING Overview of Data analysis, Introduction to Data visualization, Working with statistical formulas - Logical and financial functions , Data Validation & data models, Power Map for visualize data , Power BI-Business Intelligence , Data Analysis using statistical methods, Dashboard designing.

UNIT - II

INTRODUCTION TO DATA MANIPULATION USING FUNCTION: Heat Map, Tree Map, Smart Chart, Azure Machine learning , Column Chart, Line Chart , Pie,Bar, Area, Scatter Chart, Data Series, Axes , Chart Sheet , Trendline , Error Bars, Sparklines, Combination Chart, Gauge, Thermometer Chart , Gantt Chart , Pareto Chart etc , Frequency Distribution, Pivot Chart, Slicers , Tables: Structured References, Table Styles , What-If Analysis: Data Tables, Goal Seek, Quadratic Equation , Transportation Problem, Maximum Flow Problem, Sensitivity Analysis, Histogram, Descriptive, Statistics, Anova, F-Test, t-Test, Moving, Average, Exponential Smoothing | Correlation model | Regression model, Practical Lab

UNIT – III

TABLEAU SOFTWARE: GETTING STARTED WITH TABLEAU SOFTWARE: What is Tableau? What does the Tableau product suite comprise of? How Does Tableau Work? Tableau Architecture, What is My Tableau Repository? Connecting to Data & Introduction to data source concepts, Understanding the Tableau workspace, Dimensions and Measures, Data Types & Default Properties, Building basic views, Saving and Sharing your work-overview, Practical Lab

UNIT - IV

TABLEAU BUILDING VIEWS (REPORTS): Date Aggregations and Date parts, Cross tab & Tabular charts, Totals & Subtotals, Bar Charts & Stacked Bars, Trend lines, Forecasting, Filters, Context filters, Line Graphs with Date & Without Date, Tree maps, Scatter Plots

Samatrix.io

BIG DATA ANALYTICS WITH HADOOP

SEMESTER: 5	Credit: 2-0-1
Software: Apache Hadoop, Apache Pig, Apache Hive, Apache Spark, Apache Avro, Ubuntu/Centos, Java	No of Hours: 4 per Week

Objectives: The objective of this course is to teach students the conceptual framework of Big Data, Virtualization, MapReduce, HDFS, Pig, Hive, Spark, ZooKeeper, HBase

UNIT – I

Big Data: Fundamentals of Big Data, defining big data, building successful big data management architecture, big data journey

Big Data Types: Structured and unstructured data types, real time and non-real time requirements

Distributed Computing: History of distributed computing, basics of distributed computing

UNIT – II

Big Data Technology Foundation: Big Data stack, redundant physical infrastructure, security infrastructure, operational databases, organising data services and tools, analytical data warehouse, big data analytics

Virtualization: Basics of virtualization, hypervisor, abstraction and virtualization, implementing virtualization with big data

Cloud and Big Data: Defining cloud, cloud deployment and delivery models, cloud as an imperative for big data, use the cloud for big data

UNIT – III

Operational Databases: Relational database, nonrelational database, key-value pair databases, document databases, columnar databases, graph databases, spatial databases

MapReduce Fundamentals: Origin of MapReduce, map function, reduce function, putting map and reduce together, optimizing map reduce

Hadoop: Discovering Hadoop, Hadoop distributed file system, Hadoop MapReduce, Hadoop file system, dataflow, Hadoop I/O, data integrity, compression, serialization, file-based data structure

UNIT – IV

Avro: Avro data types and schemas, in-memory serialization and deserialization, avro datafiles, schema resolution

Pig: Comparison with databases, pig latin, user defined functions, data processing operators

Hive: Running hive, comparison with traditional databases, HiveQL, tables, querying data, user-defined functions

Spark: Resilient distributed datasets, shared variables, anatomy of a spark job run, executors and cluster managers,

HBase: HBasics, concepts, clients, HBase vs RDBMS, Praxis

ZooKeeper: ZooKeeper services, building application with ZooKeeper

Samatrix.io

SCALA FOR DATA SCIENCE

SEMESTER: 5	Credit: 2-0-1
Software Req: Scala	Hours: 4 per Week

Objectives: The objective of this course is to teach students Scala Programming Language, basic functions in Scala programming language and critical techniques

UNIT – I

Scala Language: Getting to know Scala programming language, Scala and Java, Statically typed language, Apache Spark and Scala, Scala Performance Benefits, Installing Scala, Using Scala REPL/Shell, getting help from Scala shell, Hello World, Paste mode, retrieving history, auto-complete feature, exiting from Scala REPL

UNIT – II

Variables, Data Types, Conditional Statements: Immutability of variables, define mutable and immutable variables, mutability and type safety, Specifying types for variables, Scala Identifier rules, naming conventions, Scala data types, Boolean types, string type, multiline strings, string operations, string concatenation, string interpolation, length of string, splitting string, extracting part of string, index of character of strings, the ANY type, type casting, Boolean expressions, conditional statement in Scala, nested IF/ELSE statement, pattern matching,

UNIT – III

Code Blocks, Functions, Collections: Code Blocks in Scala, Why use functions in Scala, understanding functions in Scala, define and invoke a function, functions with multiple parameters, positional parameters, functions with no argument, single-line function, passing function as argument, anonymous function, Collections in Scala, Understanding List, list size, convert list to string, iterating over list, map function and collection, foreach, reduce operation, list equality, create set, indexing map, manipulating maps, understanding tuples, indexing tuples, mutable collections, nested collections

UNIT – IV

Loops, Packages, Classes and Exceptional Handling: For loop, While loop, Breaking Loop iteration, classes and objects in Scala, Create classes and objects, singleton objects, case classes, equality checks, classes and packages, avoid name space collusion, importing package, fundamental of exception handling, type inferences and exception handling, try, catch, finally, Scala built tool (SBT), Compile Scala applications,