Shri Ramdeobaba College of Engineering and Management, Nagpur-13.

Department of Computer Science & Engineering (Data Science)

Minor in Data Science (For students of branches other than CSE, AIML, Cyber Security, Data Science)

SNo.	Semester	Course Code	Course Name	Hrs/Week	Credits	CA	ESE	Total	ESE Duration
1.	III	CDTM301	Programming for Data Science	3	3	40	60	100	3
2.	IV	CDTM401	Databases and SQL for Data Science	3	3	40	60	100	3
3.	V	CDTM501	Data Handling and Visualization	4	4	40	60	100	3
4.	VI	CDTM601	Statistical Machine Learning	4	4	40	60	100	3
5.	VII	CDPM701	Project	8	4	50	50	100	--
			TOTAL	22	18			500	

Honors (Only for CSE and Cyber Security students)

SNo.	Semester	Course Code	Course Name	Hrs/Week	Credits	CA	ESE	Total	ESE Duration
1.	III	CSTH302	Data Science Programming Languages	3	3	40	60	100	3
2	IV	CSTH402	Statistics for Data Analysis	3	3	40	60	100	3
3.	V	CSTH502	Data Engineering	4	4	40	60	100	3
4.	VI	CSTH602-1	Business and Web Analytics	4	4	40	60	100	3
		CSTH602-2	Machine Learning	4	4	40	60	100	3
5.	VII	CSPH702	Project	8	4	50	50	100	--
			TOTAL	**22**	**18**			**500**	

Course Code: CDTM301 **Course: Programming for Data Science**

L: 3 Hrs, T: 0 Hr, P: 0 Hr, Per Week **Total Credits: 03**

Course Objectives:

- To apply fundamental programming concepts, computational thinking and data analysis techniques to solve real-world data science problems.
- To analyse data and perform simple data visualizations using Processing
- To understand and apply introductory programming concepts such as sequencing, iteration and selection.
- To work with variable and external data sets, write statistical functions

Syllabus:

UNIT I: R Basics

What is R?, R Studio Framework, How to use R Directories, In and Out of R-Studio

UNIT II: Data Structures and Data Input

Vector and Matrices, Data structures for Heterogeneous data, Data import with example

UNIT III: Programming with Data

Logical Statements, Loops, Repeats and Functions, Data Wrangling, Interacting with Data in R

UNIT IV: Plotting Data

Vectors and Scatterplots, Histograms, Barcharts, Boxplots, and Grouping Data

UNIT V: Basic Statistical functions

Subsamples and Random variable distributions, Testing, Regression

Unit VI: Version Control

Create a GIT Repo, Review a Repo History

Syllabus for Semester IV(Minor), B. TECH. Computer Science and Engineering (Data Science)

Course Code:	**CDTM401**	**Course:**	**Databases and SQL for Data Science**
L: 3 Hrs, T: 0 Hr, P: 0 Hr, Per Week		**Total Credits:**	**03**

Course Outcomes:

On successful completion of the course, students will be able to:

1. Apply critical programming language concepts such as data types, iteration, control structures, functions, and Boolean operators by writing R programs and through examples
2. Perform statistical analysis of data.
3. Analyze a data set in R and present findings using the appropriate R packages.

Text Books:

1. R Programming for Data Science, Roger D. Peng, Lean Publishing.
2. R for Data Science, Hadley Wickham & Garrett Grolemund, O'Reilly Publishing

Reference Books:

1. Data Visualization and Exploration with R, by Eric Pimpler, Geospatial Training Services.

Course Objectives :

The purpose of this course is to introduce relational database concepts and help students apply foundational knowledge of the SQL language. It is also intended to get them started with performing SQL access in a data science environment.

Syllabus :

Unit I :

Getting Started with SQL : Introduction to databases, basic SQL

Unit II :

Compare and contrast the roles of a database administrator and a data scientist, and explain the differences between one-to-one, one-to-many, and many-to-many relationships with databases, ER models.

Unit III :

Introduction to Relational Databases and Tables : Fundamental concepts behind databases, tables, and the relationships between them. Creation of database instance,

SQL statements to create and manipulate tables.

Unit IV :

Intermediate SQL : use string patterns and ranges to search data , sort and group data in

result sets, Filtering, Sorting, and Calculating Data with SQL, nested queries.

Unit V :

Subqueries and Joins in SQL, Modifying and Analysing Data with SQL

Unit VI :

Accessing Databases using Python

Course Outcomes:

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

1. Create tables using SQL
2. Construct simple, nested, multiple table, and advanced queries for data retrieval.

3. Access databases using Python

Syllabus for Semester V(Minor), B. TECH. Computer Science and Engineering (Data Science)

Course Code:	**CDTM501**	**Course:**	**Data Handling and Visualization**
L: 4Hrs, T: 0 Hr, P: 0 Hr, Per Week		**Total Credits:**	**04**

Text Books :

1. Python and Sql Programming by Tony Coding, Amazon Asia-Pacific Holdings Private Limited
2. SQL Practice Problems by Sylvia Moestl Vasilik, Amazon Asia-Pacific Holdings Private Limited

Reference Books :

1. SQL for Data Analytics: Perform fast and efficient data analysis with the power of SQL by Upom Malik , Matt Goldwasser , Benjamin Johnston , Packt Publishing

Course Objectives:

In this course, students will learn about

- Handling Data with Python Libraries
- Extracting and Analyzing data from different resources
- Numerical Python
- Data Analysis with Pandas
- Data Visualization using matplotlib
- Advanced Visualization with Seaborn

Syllabus:

UNIT I: Overview of Python

Defining variable, structure of code, Defining blocks and Indentation, Play with Strings, Handling, Input and output

UNIT II: Learning Numpy

Intro to Numpy, creating arrays, using arrays and scalars, Indexing arrays, Array Transposition, Universal Array Function, Array Processing, Array Input and Output.

UNIT III: Exploring with Pandas

Data frames, Index objects, Reindex, Drop Entry, Selecting entries, Data Alignment, Rank and Sort, Summary statistics, Missing Data, Index hierarchy.

UNIT IV: Handling Data with Python

Json with Python, Learn how to import and export data from python, Html with python, learn to import html files with panda, Microsoft excel file with python

UNIT V: Working with Data

Merge, Merge on Index, Concatenate, Combining Data Frames, Reshaping, Pivoting, Duplicates in Data Frames, Replace, Rename Index, Binning, Outliers, Permutation, Group By on Data Frames, Group By on Dictionary and Series, Aggregation, Splitting Applying and Combining, Cross Tabulation

UNIT VI: Data Visualization

Installing Seaborn, Histogram, Kernel density estimate plots, Box and violin plots, Regression plots, Heatmap and clustered matrices.

Course Outcomes:

On successful completion of the course, students will be able to:

1. Use python libraries to analyze complex data.
2. Demonstrate different approaches to data visualisation.
3. Analyze the data and can give detailed insights.

Text Books:

1. Data Visualization with Python: Create an impact with meaningful data insights using interactive and engaging visuals by Tim Gromann, Packt Publishing
2. Practical Python Data Visualization, by Ashwin Pajankar, Apress

Reference Books:

1. Mastering Python Data Visualization by Kirthi Raman, Packt Publishing.

Syllabus for Semester VI(Minor), B. Tech. Computer Science and Engineering(Da Science)

Course Code: CDTM71 **Course: Statistical Machine Learning**

L: 4 Hrs, T: 0 Hr, P: 0 Hr, Per Week **Total Credits: 04**

Course Objectives:

1. To introduce the basic concepts and techniques of machine learning.
2. To understand major machine learning algorithms.
3. To identify machine learning techniques suitable for a given problem.

Course Syllabus:

UNIT I:

Introduction, Types of Machine Learning, Supervised Learning, Regression and Classification, Decision Tree Learning, Over-fitting, Cross-Validation, and Experimental Evaluation of Learning Algorithms.

UNIT II:

Instance-Based Learning: k-Nearest neighbor algorithm, Weighted k-Nearest neighbor algorithm, Case-based learning.

Regression: Linear Regression, Logistic Regression.

UNIT III:

Probabilistic Machine Learning: Maximum Likelihood Estimation, MAP, Naive Bayes classifier, Bayes optimal classifiers, Minimum description length principle.

UNIT IV:

Association Rule Mining: Support and Confidence, Apriori Algorithm, FP Tree Algorithm. Association to correlation analysis.

UNIT V:

Clustering and Unsupervised Learning: K-means, Hierarchical Clustering using single linkage, complete linkage and average linkage methods.

UNIT VI:

Support Vector Machine, and Ensemble learning: boosting, bagging, Random Forest.

Course Outcomes:

On successful completion of the course, students will be able to:

1. Apply supervised machine learning model to solve a specific problem.
2. Use probabilistic machine learning model for a given application.

3. Apply unsupervised machine learning model to solve a specific problem.
4. Solve problems using support vector machine, and Ensemble learning.

Text Books:

1. Tom Mitchell; Machine Learning- an Artificial Intelligence Approach, Volume-II; Morgan Kaufmann, 1986.
2. Christopher Bishop, Pattern Recognition and machine learning; Springer Verlag, 2006.

Reference Books:

1. Soumen Chakrabarti; Mining the Web: Discovering Knowledge from Hypertext Data, Morgan-Kaufmann, 2003.
2. A. K. Jain and R. C. Dubes; Algorithms for Clustering Data; Prentice Hall PTR, 1988.
3. Ethem Alpaydin, Introduction to Machine Learning, PHI.

Syllabus for Semester VII (Minor), B. Tech. Computer Science &Engineering(Data Science)

Course Code: CDPM701 **Course: Project**

L: 0 Hrs, T: 0 Hr, P: 8 Hr, Per Week **Total Credits: 04**

Course Outcomes

These learning outcomes are intended to assist students in acquiring a comprehensive understanding of Data Science concepts and their practical applications. On successful completion of the project, students will be able to:

1. Identify, understand, formulate, and solve engineering problems

2. Apply knowledge of Math, Science, and Engineering

3. Participate in a hands-on project involving different Data Science techniques, and the applications of these skills in different domains and settings.

4. Identify and employ appropriate system development tools and techniques

5. Test and deploy the system to solve the societal, industry, and real life problems

6. Function in multi-disciplinary teams

7. Engage in Life-long learning.

8. Employ techniques, skills, and modern engineering tools for presentation, report / paper drafting, and product manual development.

The project will focus on the creation of Data Science based products, utilizing the expertise acquired in previous semesters. This topic primarily centres on the product development cycle and its sequential execution. The acquisition of skills necessary for producing high-quality research papers, project reports, product manuals, patent documents, and presentations is facilitated by engaging in a range of educational activities like as attending technical sessions and seminars, accessing online resources, participating in courses offered by platforms like NPTEL, EDX, and Coursera, and completing associated assessments.

Shri Ramdeobaba College of Engineering and Management, Nagpur-13.

Department of Computer Science & Engineering (Data Science)

Honors (Only for CSE and Cyber Security students)

SNo.	Semester	Course Code	Course Name	Hrs/Week	Credits	CA	ESE	Total	ESE Duration
1.	III	CSTH302	Data Science Programming Languages	3	3	40	60	100	3
2	IV	CSTH402	Statistics for Data Analysis	3	3	40	60	100	3
3.	V	CSTH502	Data	4	4	40	60	100	3

			Engineering						
4.	VI	CSTH602-1	Business and Web Analytics	4	4	40	60	100	3
		CSTH602-2	Machine Learning	4	4	40	60	100	3
5.	VII	CSPH702	Project	8	4	50	50	100	--
			TOTAL	**22**	**18**			**500**	

Syllabus for Semester III

Course Code: CSTH302 **Course: Data Science Programming Languages**

L: 3Hrs, T: 0 Hr, P: 0 Hr, Per Week **Total Credits: 03**

Course Objectives:

1. Learn powerful R tools for solving data problems with greater clarity and ease.
2. Examine your data, generate hypotheses and test them.
3. Learn to transform your datasets into a form convenient for analysis.

Syllabus:

UNIT I: Exploring the R Framework

History and Overview of R, What is R? , What is S?, The S Philosophy, Back to R, Basic Features of R,Free Software, Design of the R, System, Limitations of R, R Resources, Entering Input, Evaluation, R Objects, Numbers, Attributes, Creating Vectors, Mixing Objects, Explicit Coercion, Matrices, Lists, Factors, Missing Values, Data Frames, Names.

UNIT II: Getting Data In and Out of R

Getting Data In and Out of R, Reading and Writing Data, Reading Data Files with read.table(), Reading in Larger Datasets with read.table, Calculating Memory Requirements for R Objects, Using dput and dump( ), Binary formats, File connections, Reading lines of a text file, Reading from a URL connection, Subsetting a Vector, Subsetting a Matrix, Subsetting Lists, Subsetting Nested Elements of a List, Extracting Multiple Elements of a List, Partial Matching, Removing NA Value

UNIT III: Vectorized Operations, Dates and Times and Managing the Data Frames

Vectorized Matrix Operations, Dates in R, Times in R, Operations on Dates and Times, Data Frames, the dplyr Package, dplyr Grammar and commands. Control Structures : if-else, for Loops, Nested for Loops, while loop, repeat loop, next, break.

UNIT IV: Functions and Scoping Rules of R

Functions, Argument Matching, Lazy Evaluation, A Diversion on Binding Values to Symbol , Scoping rules, Lexical scoping, Plotting the likelyhood, coding standards for R.

UNIT V: Loop Functions, Debugging

Looping on the Command Line. lapply(), sapply() , split(), Splitting a Data Frame, tapply, apply(), Col/Row Sums and Means, Other Ways to Apply , mapply(), Vectorizing a Function,

UNIT VI: Profiling R Code and Simulation

Using system.time(), Timing longer expressions, The R Profiler, Using summaryRpof(), Generating Random Numbers, Selecting the random number seed, Simulating a linear model and Random Sampling

Technology: R Programming, R Studio

Course Outcomes:

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

1. Demonstrate critical notions such as data types, operators, data structures.
2. Apply programming language concepts such as control structures and functions
3. Perform data manipulations using R
4. Analyze a data set in R and present findings using the appropriate R packages

Text Books:

1. R Programming for Data Science, Roger D. Peng, Lean Publishing.
2. R for Data Science,Hadley Wickham & Garrett Grolemund, O'Reilly Publishing

Reference Books and Web Resources:

1. Data Visualization and Exploration with R, by Eric Pimpler, Geospatial Training Services.
2. https://github.com/data-datum/learning_R

Syllabus for Semester IV

Course Code:	**CSTH402**	**Course:**	**Statistics for Data Analysis**
L: 3 Hrs, T: 0 Hr, P: 0 Hr, Per Week		**Total Credits: 3**	

Course Objectives:

This course aims to build basic foundations of statistics which are useful for data analysis.

Syllabus:

Unit 1:

Introduction, Variables and Types of data, Data collection and sampling techniques, uses and misuses of statistics, Organizing data, histograms, frequency polygons, ogives, other types of graphs.

Unit 2:

Measures of central tendency, measures of variation, measures of position, five-number summary, boxplots

Unit 3:

Probability distributions, normal distributions, binomial distribution, other types of distributions, central limit theorem.

Unit 4:

Confidence interval and sample size, Hypothesis testing.

Unit 5:

Testing the Difference Between Two Means, Two Proportions, and Two Variances, Correlation and Covariance.

Unit 6:

Simple linear regression, multiple linear regression, logistic regression.

Course Outcomes:

On successful completion of the course, the students will be able to:

1. Apply sampling techniques to generate appropriate samples.
2. Apply descriptive statistical techniques for data analysis.
3. Interpret data to perform hypothesis testing.
4. Apply regression models.

Text Books:

1. Elementary Statistics A Step by Step Approach by Allan G. Bluman , McGraw Hill Publications, Seventh Edition.
2. Practical Statistics for Data Scientists by Peter Bruce and Andrew Bruce, O`Reilly Publications.

Syllabus for Semester V

Course Code: CSTH502 **Course: Data Engineering**

L: 4 Hrs, P: 0 Hr, T: 0 Hr Per Week **Total Credits: 04**

Course Objectives

- To understand data engineering concepts
- To understand cloud computing capabilities and implementations
- To understand big data concepts

Syllabus:

Unit-1

Introduction: Data Engineering, Data Engineering Ecosystem, Data Engineer Lifecycle, Data Engineer Vs Data Science.

Unit-2

Big Data: Streaming Process, Linux , Cloud, spark Data frame API & Spark SQL.

Unit -3

Security and Privacy: SSL Public and Private Gift Certificate, Certificate Authority, GDPR Regulation.

Unit-4

Data Warehousing : Data Warehousing on AWS, Data Lakes, Data Pipeline, Apache Air flow, Monitoring Data Pipeline, Deploying Data Pipeline, Extract Transform and Loads.

Relational Databases: Use SQL & PostgreSQL, when to use NoSQL databases.

Unit-5

Data Processing: Analytics Frameworks, Data Visualization, Hadoop, Apache Kafka (Framework), Docker, API.

Unit-6

Case Study: Data Camp, Data Science @ Twitter, Capstone Project .

Course Outcomes

On successful completion of the course, students will be able to:

1. Evaluate Data Engineering as a discipline of study and differentiate it from Data Science.
2. Summarize cloud computing capabilities and compare cloud computing with on-site implementations.
3. Utilize Linux and the command line to perform computing tasks and explain how Linux is used
4. Describe Hadoop's and Spark's role in big data and explain batch versus in memory processing of big data.
5. Summarize pros and cons of relational databases and SQL and implement a NoSql database .

Text Books

1. The Data Engineering Cookbook Mastering The Plumbing Of Data Science Andreas Kretz May 18, 2019.
2. Data Engineering with Python: Work with massive datasets to design data models and automate data pipelines using Python Paperback – 23 Oct. 2020

Reference Books

1. Spark: The Definitive Guide: Big Data Processing Made Simple by Bill Chambers
2. Big Data, Black Book: Covers Hadoop 2, MapReduce, Hive, YARN, Pig, R, and Data Visualization
3. Data pipleline pocket reference : Moving and processing data for analytics by James Densmore

Syllabus for Semester VI

Course Code: CSTH602-1 **Course: Business and Web Analytics**

L: 4 Hrs, T: 0 Hr, P: 0 Hr, Per Week **Total Credits: 04**

Course Objectives

1. To familiarize students with the concepts of business and web analytics.
2. To enable students to deploy business and web intelligence to improve the outcomes of marketing or business plan.
3. Students will gain an understanding of the strategic and operational aspects of business and web analytics tools and technologies.

SYLLABUS

UNIT – I: Predictive Analytics : Trendlines and Regression analysis, Forecasting techniques

UNIT – II: Simulation and Risk Analysis : Monte Carlo simulation, Random Sampling from probability distributions, dynamic system simulation, spreadsheet modeling and analysis.

UNIT – III: Prescriptive Analytics : Linear Optimization, Interger and nonlinear optimization, optimization analytics.

UNIT – IV: Web Analytic fundamentals: Capturing data: Web logs or JavaScript's tags, Separate data serving and data capture, Type and size of data, Innovation, Integration, Selecting optimal web analytic tool, Understanding click stream data quality, Identifying unique page definition, Using cookies, Link coding issues.

UNIT – V: Web Metrics: Common metrics: Hits, Page views, Visits, Unique visitors, Unique page views, Bounce, Bounce rate, Page/visit, Average time on site, New visits; Optimization (e-commerce, non e-commerce sites): Improving bounce rates, Optimizing adwords campaigns; Real time report, Audience report, Traffic source report, Custom campaigns, Content report, Google analytics, Introduction to KPI, characteristics, Need for KPI, Perspective of KPI, Uses of KPI.

Relevant Technologies: Internet & TCP/IP, Client / Server Computing, HTTP (Hypertext Transfer Protocol), Server Log Files & Cookies, Web Bugs.

UNIT – VI: Web Analytics 2.0:Web analytics 1.0, Limitations of web analytics 1.0, Introduction to analytic 2.0, Competitive intelligence analysis : CI data sources, Toolbar data, Panel data ,ISP data, Search engine data, Hybrid data, Website traffic analysis: Comparing long term traffic trends, Analyzing competitive site overlap and opportunities.

Google Analytics: Brief introduction and working, Adwords, Benchmarking, Categories of traffic: Organic traffic, Paid traffic; Google website optimizer, Implementation technology, Limitations, Performance concerns, Privacy issues.

Course Outcomes:

On completion of the course the student will be able to

1. Apply predictive analytic techniques on real time data.
2. Perform simulation and risk analysis.
3. Explain and discuss web metrics.
4. Examine how different industries across the globe are using web analytics analytics.

Text Books:

1. James R. Evans, Business Analytics Methods, Models and Decisions, Pearson.
2. Clifton B., Advanced Web Metrics with Google Analytics, Wiley Publishing, Inc.2nd ed.
3. Kaushik A., Web Analytics 2.0, The Art of Online Accountability and Science of Customer Centricity, Wiley Publishing, Inc. 1st ed.

4. Sterne J., Web Metrics: Proven methods for measuring web site success, John Wiley and Sons

References:

1. Mathew Ganis, Avinash Koihrkar-Social Media Analytics-IBM Press
2. Jim Sterne, Social Media Metrics, Wiley