Honors Specialization in Computer Science Engineering (AIML & Data Science Track)

(Only for CSE and Cyber Security students)

S. N.	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**	

Syllabus for Semester III, B. TECH. Computer Science and Engineering (Data Science & AIML)

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, B. TECH. Computer Science and Engineering (Data Science & AIML)
Department of Computer Science and Engineering

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, B. TECH. Computer Science and Engineering (Data Science & AIML)
Department of Computer Science and Engineering

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, B. TECH. Computer Science and Engineering (Data Science & AIML)

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

Syllabus for Semester VII, B. TECH. Computer Science and Engineering (Data Science & AIML)

CSPH702 **Course: Project**

T: 0 Hr, P: 8 Hr, Per week **Total Credits: 04**

Outcomes

These learning outcomes are intended to assist students in acquiring a comprehensive understanding of AI-ML and 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 AI- ML and data science related 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 society's and industry's 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 artificial intelligence and machine learning-based products, utilizing the expertise acquired in previous semesters. This topic primarily centers 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 like YouTube lectures, participating in courses offered by platforms like NPTEL, EDX, and Coursera, and completing associated assessments.