WEST BENGAL COUNCIL OF HIGHER SECONDARY EDUCATION
SYLLABUS FOR CLASSES XI AND XII
SUBJECT : DATA SCIENCE (DTSC)

Course Objectives

The objective of course is:

- To Impart knowledge about basic computer fundamentals and programming languages for data science.
- To Impart knowledge about mathematical and statistical methods for data analysis.
- To Empower students with data visualization techniques and tools.
- To impart knowledge about the basics of data management and Business Theory.
- To impart knowledge about various machine learning techniques used for data analysis.
- To enable students to develop data-based machine learning models for solving real-word applications.
- To enable students to gain practical experience in programming languages and statistical and machine learning tools for data sciences.

Course Outcomes

Upon completion of this course, the student should be able to:

- Explain the importance of and be able to formulate a data analysis problem
- Explain various data types, and data formats , and identify and appropriately acknowledge sources of various types of data
- Gain an ability to apply mathematical and statistical methods in data science applications
- Apply basic data cleaning techniques to prepare data for analysis
- Demonstrate proficiency in using appropriate tools and technology to collect, process, transform, summarize, and visualize data.
- Apply various machine learning algorithms in data-based decision-making applications , and draw accurate and useful conclusions through data analysis
- Demonstrate some skills in data retrieval using Structured Query language (SQL)
- Explain the basics of Business Theory
- Demonstrate skill in basic exploratory data analysis using unsupervised learning
- Demonstrate proficiency in implementing supervised machine learning algorithms for predictive data analysis using the latest programming languages and software tools.
- Differentiate between ethical and unethical uses of data science.

CLASS - XI

SEMESTER – I

SUBJECT: DATA SCIENCE (DTSC)

FULL MARKS: 35 **CONTACT HOURS: 60 Hours**

COURSE CODE: THEORY

UNIT NO.	SUB	TOPICS	CONTACT HOURS	MARKS
Unit -1 Computer Fundamentals (15)	1a.	History of computer, Basic Computer hardware, input and output devices, Basic computer architecture, input output devices, memory and CPU, networking of machines (overview of LAN, MAN, WAN, Internet, Wifi etc), types of computer (workstation, desktop, Smartphone, embedded system, etc.), Overview of Software (system software and application software with examples (mention names only)), Definition of Operating System and functions (mention names of some popular operating systems like Windows, Linux, Android, etc).	**8**	**5**
	1b.	Bit, Byte and Word, Number System (Base, Binary, Decimal, Octal, Hexadecimal), Conversion of number systems, Boolean logic (Boolean Gates), Boolean operators (OR, AND and NOT), ASCII code, Concept of Algorithm and Flowchart.	**6**	**5**
	1c.	Basics of Computer Programming (three levels: high level language, assembly language, machine language, definition and block diagrams), Overview of Compiler and Interpreter (definition and mention name of major compiled (e.g., C, C++) and interpreted languages (e.g., Python), Overview of procedural and object oriented programming (key features and just the basic differences, mention names of some popular procedural (e.g., BASIC, FORTRAN, C) and object oriented programming languages (e.g., C++, Java, Python).	**10**	**5**
Unit -2 Introduction to Python Programming (15)	2a.	Basics of Python programming (with a simple 'hello world' program, process of writing a program, running it, and print statement), Concept of class and object, Data-types (integer, float, string), notion of a variable, Operators (assignment, logical, arithmetic etc.), accepting input from console, conditional statements (If else and Nested If else), Collections (List, Tuple, Sets and Dictionary), Loops (For Loop, While Loop & Nested Loops), iterator, string and fundamental string operations (compare, concatenation, sub-string etc.), Function, recursion.	**12**	**5**

UNIT NO.	SUB	TOPICS	CONTACT HOURS	MARKS
	2b	Overview of linear and nonlinear data structure (definition, schematic view and difference), array (1D, 2D and its relation with matrix, basic operations: access elements using index, insert, delete, search), stack (concept of LIFO, basic operations: Push, Pop, peek, size), queue (concept of FIFO, basic operations: Enqueue, Dequeue, peek, size), use of List methods in Python for basic operations on array, stack and queue, overview of NumPy library and basic array operations (arrange(), shape(), ndim(), dtype() etc.), binary tree (definition and schematic view only) .	**12**	**6**
	2c	Linear search and binary search algorithm, sorting algorithm (bubble sort only)	**4**	**4**
Unit -3 History of AI and Introduction to Linear Algebra (5)	3a	History of AI: AlanTuring and cracking enigma, mark 1 machines, 1956- the birth of the term AI, AI winter of 70's, expert systems of 1980s, skipped journey of present day AI. Distinction between terms AI, Pattern recognition and Machine Learning. (Note: it should be taught as a story more than flow of information World war 2, Enigma and Alan Turing, the birth of modern computers)	**2**	**2**
	3b	Basic matrix operations like matrix addition, subtraction, multiplication, transpose of matrix, identity matrix. A brief introduction to vectors, unit vector, normal vector, Euclidean space.	**6**	**3**

NB : Additional 10 hours for Remedial and/or Tutorial classes

CLASS - XI

SEMESTER – II

SUBJECT: DATA SCIENCE (DTSC)

FULL MARKS: 35 **CONTACT HOURS: 60 HOURS**

COURSE CODE: THEORY

UNIT NO.	SUB	TOPICS	CONTACT HOURS	MARKS
Unit -4 History of data science and statistics (15)	4a	Brief history of data science, data science as conjunction of computer science statistics and domain knowledge. Definition of data science, data science life cycle - capture, maintain, process, analyze, communicate	**6**	**3**
	4b.	Probability distribution, frequency, mean, median and mode, variance and standard deviation, Gaussian distribution, Random sampling by uniform distribution and students-t distribution hypothesis testing, Distance function, Euclidean norm, distance between two points in 2D and 3D and extension of idea to *n* dimensions	**10**	**5**
	4c.	Basic ideas of different Data Science Toolkit: Excel, Weka, R	**12**	**7**
Unit - 5 Data Visualization(10)	5	Types of data: textual data (reviews, comments blogs), signal data (time series, audio, sensor data) visual data (image and video, remote sensing data, feeds etc.) Introduction to data dimension and modality, their representations in computer science. Data cleaning - Representation of data in textual form, tokens, sentences, word histograms, reading from web pages using crawlers - Representation format of audio data, uncompressed wav format and compressed mp3 format (just the description of the pipeline, no maths) - Representation of visual data in RGB pixels, storing in raw format and compressed format (just the description of the pipeline, no maths) - Representation of other forms of data like time series values from different sensors, remote sensing image data etc. - Introduction to the concept of multimodality *i.e*. different modes of data from the same information source (example audio and video generated when filming) - Data dimension (resolution for image, frequency bins and sampling rate for audio, word histograms for text) - Concept of data cleaning, removal of abnormal, incomplete and corrupted or garbage data as a preprocessing stage.	**16**	**10**

UNIT NO.	SUB	TOPICS	CONTACT HOURS	MARKS
Unit -6 Database Management (5)	6	Brief introduction to relational database, tables for keeping data, brief introduction to SQL - Introduction to the concept of database - Relational database, table, schema as columns and tuple as rows - Some basic SQL statements such as CREATE, SELECT, INSERT, UPDATE, DELETE (Simple query examples)	**8**	**5**
Unit -7 Basics of Business Theory (5)	7	Business theory basics: Different business models B2B, B2C. Aggregator type business, manufacturing type business, consultancy and turnkey service based businesses, social media type and general digital platform type business, content hosting businesses. Definitions of profit, loss, revenue, break-even, valuation etc. **[NO LAB COMPONENT]** - The basic business types, product based and service based - Business classification by clients, the B2B and B2C models - Types of business who use DS extensively: software product and service, aggregator (cab, food delivery, groceries, online market), manufacturing and banking - Consultancy type business and service profiling - Social media business and targeted advertising based business model - Basic business terminologies, refer (https://getsling.com/blog/business-terms/)	**8**	**5**

NB : Additional 10 hours for Remedial and/or Tutorial classes

CLASS: XI

SUBJECT: DATA SCIENCE (DTSC)

COURSE CODE: PRACTICAL

FULL MARKS: 30 **CONTACT HOURS: 60 HOURS**

Sub Topic

1. **Computer Fundamentals** [No marks]	● Visit to Computer Lab and familiarization with computers and peripherals and different networking devices (e.g., modem, switch, router). ● Opening of the CPU box/cabinet and identification of different parts (e.g., Motherboard, CPU/Processor, RAM, Hard Disk, power supply).	no marks (6 hours)
2. **Introduction to Python Programming [10 Marks]**		
2a.	● Introduction to installation and running of python codes with hello world and simple accessing user inputs from console examples. ● Menu driven arithmetic calculator ● Simple logical and mathematical programs (e.g., printing patterns, Conversion of binary to decimal and vice versa, computing GCD of two numbers, Finding prime numbers, Generating Fibonacci sequence, Computing factorial –iterative and recursive etc.) ● Finding max, min, avg, sum, length of a list ● Use of basic string methods like upper(), lower(), count(), find(), join(), replace(), split() etc.	3 Marks (4 hours)
2b.	● Use of Python List methods for Stack and Queue implementation, for examples, append() and pop() ● Use of NumPy array methods: arrange(), shape(), ndim(), size(), add(), subtract(), multiply(), divide(), mat() etc. ● Use of NumPy matrix multiplication methods: dot(), matmul(), multiply() etc. ● Linear search and binary search in an array ● Bubble sort in an array	5 Marks (4 hours)
2c.	Creating data frame from .csv file , excel sheet , python dictionary, python list, tuple operation on data frame.	2 Marks (4 hours)

3. **Foundation for AI and Data Science [5 Marks]**	● Generation of random numbers in python following a certain distribution and filling up random arrays ● Introduction to matplotlib to plot arrays as histograms ● Computation of mean, median and mode ● Computing CDF from PDF and plotting using matplotlib ● Plotting Gaussian distribution with a given mean and standard deviation ● Plotting mixture of Gaussian distributions	5 Marks (10 hours)
4. Data Visualization [10 marks]	Using Scipy, opencv and NLTK libraries run codes for the following ● Visualization of audio data as spectrogram ● Visualization of image data by zooming into pixels ● Visualization of word histograms	10 Marks (12 Hours)
5. Database Management [5 marks]	● Use of MySQL database for Creating tables ● Running retrieval, insertion, deletion and updation queries	5 Marks (8 hours)

NB : Additional 10 hours for Remedial and/or Tutorial classes

CLASS - XII

SEMESTER – III

SUBJECT: DATA SCIENCE (DTSC)

FULL MARKS: 35 **CONTACT HOURS: 60 Hours**

COURSE CODE: THEORY

UNIT NO.	SUB	TOPICS	CONTACT HOURS	MARKS
Unit 1: Foundation of statistics for machine learning(5)	1.	Distance between distributions - Euclidean norm, Pearson's correlation coefficient, basic concepts of (not in detail) chi-square distance, Bayes' theorem and Bayesian probability ● Real n-dimensional space (R^n) and vector algebra, dot product of two vectors, vector projections. ● Product moment correlation coefficient (Pearson's coefficient) its use in determining relation between two sets of data ● Chi-square and use in finding distance between two distributions ● Conditional probability and Bayes' theorem , conditional independence	**10**	**5**
Unit 2: Introduction to machine learning (15)	2a.	• **What is Machine Learning?** • Difference between traditional programming and machine learning • Relation of machine learning with AI • Applications of machine learning. • Why should machines have to learn? Why not design machines to perform as desired in the first place? ● Types of Machine Learning Supervised, Unsupervised, Semi-supervised and Reinforcement learning), ● Concept of training, testing and validation, Concepts of training examples, Linear Regression with one variable , hypothesis representation, hypothesis space, Learning Requires Bias, Concept of Loss function ● Training methods for linear regression model: Iterative trial-and-error process that machine learning algorithms may use to train a model, Disadvantages of iterative training method, gradient descent algorithm. ● Effect of learning rate on reducing loss. Importance of feature scaling(mini-max normalization)	**18**	**10**

UNIT NO.	SUB	TOPICS	CONTACT HOURS	MARKS
	2b.	• **<u>What is feature or attribute?</u>** • Definition and meaning of feature in various kinds of data (e.g., structured data, unstructured data(text data, image data)) • Types of features(continuous, categorical) • Representation of training examples with multiple features • Linear regression with multiple attributes (multiple features) • Feature cross and polynomial regression	**10**	**5**
Unit 3: Supervised learning (15)	3a.	● Difference between regression and classification. Examples of some real world classification problems ● Linear classification and threshold classifier, Concept of input space and linear separator, Drawback of threshold classifier, use of logistic function in defining hypothesis function for logistic regression model. ● Probabilistic interpretation of output of the logistic regression model, use of logistic regression model in binary classification task. Multi-class classification using One *vs* all strategy.	**12**	**7**
	3b.	Probabilistic classifier: Bayesian Learning, conditional independence Naive-Bayes classifier	**4**	**3**
	3c.	Measuring Classifier performance: ● Confusion matrix, true positive, true negative, false positive, false negative, error, accuracy, precision, recall, F-measure, sensitivity and specificity ● K-fold cross validation	**6**	**5**

NB : Additional 10 hours for Remedial and/or Tutorial classes

CLASS - XII

SEMESTER – IV

SUBJECT: DATA SCIENCE (DTSC)

FULL MARKS: 35 **CONTACT HOURS: 60 HOURS**

COURSE CODE: THEORY

UNIT NO.	SUB	TOPICS	CONTACT HOURS	MARKS
Unit 4: Decision tree learning and Unsupervised learning (10)	4a.	↬ Concept of entropy for measuring purity (impurity) of a collection of training examples. and information gain as a measure of the effectiveness of an attribute in classifying the training data (just basics and equation) . ↬ Inducing decision tree from the training data using ID3 algorithm, an illustrative example showing how the ID3 algorithm works. ↬ Concept of overfitting, reduced error pruning ↬ Discretizing continuous-valued attributes using information gain-based method (binary split only) ↬ Differences between supervised and unsupervised learning	**12**	**5**
	4b.	● What is unsupervised learning? ● Difference between supervised and unsupervised learning. ● What is clustering? ● Why is clustering an unsupervised learning technique? ● Some examples of real world application of clustering, ● Difference between clustering and classification ● K-means clustering algorithm. Simple use cases	**10**	**5**
Unit 5: Data visualization technique (10)	5.	↬ What is the need for data visualization? ↬ Visualization techniques: visualization of a small number of attributes (Stem and leaf plots, 1D Histogram and 2D Histogram, Box Plots, Pie chart, Scatter Plots) ↬ Visualizing Spatio -temporal Data (Contour plots, Surface plots) ↬ Visualizing higher dimensional data (Plot of data matrix) ↬ Heatmap visualization ↬ Introduction to data visual platform- Tableau and Google Chart	**12**	**10**
Unit 6: Artificial neural network (10)	6.	↬ Biological motivation for Artificial Neural Networks(ANN) ↬ A simple mathematical model of a neuron (McCulloch and Pitts(1943)) ↬ Concept of activation function: threshold function and Sigmoid function, ↬ Perceptron as a linear classifier, perceptron training rule ↬ Representations of AND and OR functions of two inputs using threshold perceptron. Equation of a linear separator in the input space, Representational power of perceptrons		

UNIT NO.	SUB	TOPICS	CONTACT HOURS	MARKS
		⊗ Training *unthresholded perceptron* using *Delta rule*, Need for hidden layers , XOR example, ⊗ Why do we need non-linearity? Network structures: feed forward networks and recurrent networks (basic concept only) ⊗ Training multiplayer feed-forward neural networks using *Backpropagation algorithm* (Concepts only and no derivation). ⊗ Generalization, overfitting, and stopping criterion, overcoming the overfitting problem using a set of validation data ⊗ An Illustrative example of an ANN architecture for handwritten digit recognition (Only input representation, output representation and a block diagram of the network) ⊗ Need for automatic feature learning, difference between the conventional feed-forward neural networks and CNN, role of convolution layer in CNN, An example of 2D convolution, function of pooling layer ⊗ A block diagram illustrating CNN applied to handwritten digit recognition task	**20**	**10**
Unit 7: Case studies in data science (5)	7.	Some case studies: ⊗ Weather forecasting using some statistical and machine learning tools (consider the ML algorithms covered in the theoretical subjects) ⊗ Sentiment Analysis using some machine learning tools (consider the ML algorithms covered in the theoretical subjects) ⊗ A simple collaborative filtering-based recommendation System	**6**	**5**

NB : Additional 10 hours for Remedial and/or Tutorial classes

CLASS: XII

SUBJECT: DATA SCIENCE (DTSC)

COURSE CODE: PRACTICAL

FULL MARKS: 30 **CONTACT HOURS: 60 HOURS**

Sub Topic

SUB	TOPICS	CONTACT HOURS	MARKS
1. Foundation of Statistics for Machine Learning [2 marks]	Consider a table of data about *n* persons with two attributes– age and income and find Pearson correlation coefficient using a python program. Do not use any built-in library function for directly calculating Pearson correlation coefficient.	4 hrs	2
2. Introduction to Machine Learning [5 Marks]			
2a.	● Introduction to python libraries like scipy ● Revisit matrix operations using scipy (basic matrix operations of addition, subtraction, multiplication, transpose)	4 hrs	2
2b.	• Generation of random (x, y) pairs where y = f(x) + d (d varies from -r to +r , a random value), *f* being a linear function • Linear regression or line fitting of the data • Optimizing the function using gradient descent	6 hrs	3
3. Supervised Learning [7 Marks]	● Loading csv file-based datasets using file-read operation in python ● Introduction to pandas library and loading csv and json files ● Building Logistic regression model for binary classification of Diabetes Data set downloadable from the UCI machine learning repository ● Building a decision tree classifier and testing on the Diabetes Data ● Introduction to the IRIS dataset, building a logistic regression model for multi-class classification and testing the model on the IRIS dataset downloadable from UCI Machine Learning Repository ● Building K-nearest neighbor classifier and testing on the IRIS dataset (Use Scikit-learn open source data analysis library for implementing the models)	10 hrs	7

SUB	TOPICS	CONTACT HOURS	MARKS
4. Unsupervised Learning [3 Marks]	Using Scikit-learn library to use k_means algorithm for clustering IRIS data and its visualization	8 Hrs	3
5. Data Visualization techniques [5 Marks]	Introduction to plotly library in python and plotting different types of plot using the library refer this(https://plotly.com/python/plotly-express/) • Stem and leaf plots • 1D Histogram of four attributes of the IRIS dataset • 2D Histogram(considering the IRIS dataset, plot 2D histogram of petal length and width) • Box Plots (Considering the IRIS dataset, show the Box plots of attributes for IRIS attributes and species) • Plot the Pie chart, showing the distribution of IRIS flowers (use IRIS dataset) • Scatter Plots for each pair of attributes of the IRIS dataset • Heatmap	12 Hrs	5
6. Artificial Neural Network [5 Marks]	• Using MLP from Scikit-learn library, develop a handwritten digit recognition model using MLP and MNIST dataset • Using CNN from Keras library, develop a handwritten digit recognition model using CNN and MNIST dataset	10 Hrs	5
7. Case studies in Data Science [3 Marks]	Case Study: sentiment analysis of movie reviews. Use machine learning tools from Scikit-learn library and the IMDB dataset	6 Hrs	3