Title of Paper Scala for DS

Sr. No.	Heading	Particulars
1	**Description of the course :** **Including but Not limited to :**	This course provides hands-on experience with Scala and its ecosystem for data analysis and machine learning. Students will learn statistical methods, machine learning algorithms, and data processing techniques using Breeze and Apache Spark. The course also covers time-series analysis, feature engineering, and building scalable data pipelines. Through practical exercises, students will gain proficiency in implementing regression models and clustering while handling real-world datasets effectively.
2	**Vertical :**	Minor
3	**Type :**	Practical
4	**Credit**:	2 credits (1 credit = 15 Hours for Theory or 30 Hours of Practical work in a semester)
5	**Hours Allotted :**	30 Hours
6	**Marks Allotted**:	50 Marks
7	**Course Objectives**: CO1: To set up and configure Scala, SBT, and Apache Spark for programming, data analysis, and large-scale data processing. CO2: To perform statistical calculations, including correlation, frequency distribution, and moving averages using Scala and Breeze, and visualize data insights with Breeze-viz. CO3: To implement machine learning models such as linear regression, logistic regression, and k-means clustering, along with feature engineering for predictive modeling. CO4: To utilize the Breeze library for numerical computations, matrix operations, and time-series data analysis to extract meaningful insights.	
8	**Course Outcomes**: OC1: Students will set up a functional Scala development environment with SBT and execute basic programs for data analysis. OC2: Students will utilize Breeze for numerical operations, matrix manipulations, and statistical computations such as correlation and frequency distribution. OC3: Students will create data visualizations using Breeze-viz and implement machine learning models, including regression and clustering, using Breeze. OC4: Students will work with Apache Spark for large-scale data processing, machine learning pipelines, and time-series analysis to extract meaningful insights.	

9	**Modules**:-
	Module 1:
	0. Set up Scala and SBT on your system. 1. Write a simple Scala program that prints a welcome message for data scientists. 2. Calculate mean, median, and mode of a list of numbers. Implement basic statistical calculations using Scala collections. 3. Generate a random dataset of 10 numbers and calculate its variance and standard deviation. 4. Create a dense vector using Breeze and calculate its sum, mean, and dot product with another vector. 5. Generate a random matrix using Breeze and compute its transpose and determinant. 6. Slice a Breeze matrix to extract a sub-matrix and calculate its row and column sums. 7. Write a program to perform element-wise addition, subtraction, multiplication, and division of two Breeze matrices. 8. Read a CSV file and calculate basic statistics for each numeric column. Use the scala-csv library or similar tools. 9. Handle missing values in a dataset. Replace missing values with the column mean. 10. Filter rows in a dataset where a specific column value exceeds a threshold. 11. Write a program to tokenize and count the frequency of words in a text file. 12. Implement one-hot encoding for a categorical column in a dataset. 13. Create a scatter plot of random data using Breeze-viz. Label the axes and customize the color of points. 14. Generate a histogram of a dataset using Breeze-viz. Experiment with different bin sizes. 15. Plot a line graph for a dataset showing a trend over time. 16. Combine two plots (e.g., scatter and line plot) in a single visualization using Breeze-viz.
	Module 2:
	1. Find the correlation between two lists of numbers. Implement the formula for Pearson correlation coefficient. 2. Calculate the moving average of a time series data using Scala collections. 3. Write a program to compute frequency distribution and cumulative frequency of a dataset. 4. Sort a dataset by a specific column and extract the top 5 rows. 5. Implement linear regression using Breeze. Fit a model to a small dataset and predict a value. 6. Perform logistic regression using Breeze. Classify a dataset with binary labels. 7. Compute the Euclidean distance between two Breeze vectors. Use it for nearest neighbor classification. 8. Cluster a dataset into two groups using k-means clustering in Breeze. 9. Set up Apache Spark locally and count the frequency of words in a text file. 10. Filter rows in a CSV file using Spark DataFrames where a numeric column exceeds a threshold. 11. Perform a group-by operation in Spark DataFrames to compute the average of each group. 12. Join two CSV files in Spark DataFrames based on a common column and write the output to a file.

	13. Create a simple Spark MLlib pipeline to classify data. Use logistic regression or decision trees. 14. Perform basic time series analysis in Scala. Generate synthetic time series data (e.g., daily sales over a month). 15. Create polynomial features from a dataset. Given a list of numbers (e.g., [1, 2, 3]), generate polynomial features up to degree 3 (e.g., [1, 1^2, 1^3, 2, 2^2, 2^3, 3, 3^2, 3^3]).	
10	**Text Books:** 1. Scala for Data Science, by Pascal Bugnion, Packt Publishing, 1st edition (28 January 2016) 2. Mastering Scala by Dennis Alexander, Packt Publishing, 1st edition (2023) 3. Scala 3 Mastery by John Hunt, Apress, 1st edition (2023) 4. Mastering Scala 3 by John Hunt, Apress, 1st edition (2023)	
11	**Reference Books:** 1. Programming Scala by Dean Wampler and Alex Payne, O'Reilly Media, 3rd edition (2021) 2. Scala Cookbook by Alvin Alexander, O'Reilly Media, 2nd edition (2021) 3. Functional Programming in Scala by Paul Chiusano and Rúnar Bjarnason, Manning Publications, 2nd edition (2023)	
12	**Internal Continuous Assessment: 40%**	**External, Semester End Examination 60% Individual Passing in Internal and External Examination**
13	**Continuous Evaluation through:** Students are expected to attend each practical and submit the written practical of the previous session. Performing Practical and writeup submission will be continuous internal evaluation.	