Introduction to Big Data in biology

Semester No	Code	Credit Hours
7-8	BI -429	3 – 0

Course Description

This course focuses on understanding the statistical structure of large-scale biological datasets using ML algorithms. We cover the basics of ML and study their scalable versions for implementation on a distributed computing framework. We pursue distributed ML algorithms for matrix factorization, convex optimization, dimensional reduction, clustering, classification, graph analytics and deep learning. This course is project driven (3 to 4 small projects) with source material from genomic sciences, structural biology, drug discovery, systems modeling and biological imaging. Students are expected to design, implement and test their ML solutions in Apache Spark.

Text And Material

1. Relevant research papers currently in the field

Course Learning Outcomes:

After completing this course, a student will be able to:

1. Explain theories and methods relevant for handling and analysing massive datasets in life science.
2. Use modern systems for handling and analysis of massive datasets in life science;
3. Analyse properties of data-intensive life science applications and, suggest suitable strategies and architectures that meet application needs.

Assessment System

Quizzes	10-15%
Assignments	5-10%
Midterms	30-40%
ESE	40-50%

Week wise Lecture Plan:

Week No	Description	Quizzes	Assignment

1	Introduction		
2 - 3	Methodology in life science applications	01	01
4 -5	Methods for large-scale data management		
6 - 7	Introduction to Mapreduce, ApacheSpark	02	
8	Processing of life science data using Mapreduce		
9	**MIDTERMS**		
10	Batch system on computational clusters	03	
11	Reproducible data analysis using workflow systems		02
12 -13	Analysis using software containers and micro-service frameworks such as Kubernetes		
14	Use of large-scale storage systems	04	
15	Use of virtualised environments.		
16 -17	Examples of applications in life science include genomics, proteomics, metabolomic & pharmaceutical development		03
18	**END SEMESTER EXAMINATION**		