Introduction to Machine learning for tabular data

Dr. Ishaan Gupta*

Department of Biochemical Engineering and Biotechnology, Indian Institute of Technology Delhi, New Delhi, India

Email: Ishaan.Gupta@dbeb.iitd.ac.in

Machine learning (ML) methods have become indispensable for discovering patterns and predictive relationships in biomedical and translational datasets. Unlike images or sequences, most clinical and experimental data—such as gene expression matrices, proteomic profiles, patient metadata, and electronic health records—are inherently tabular in structure. This talk will introduce key ML frameworks designed for tabular data, including classical algorithms such as SVM, random forests, gradient boosting (XGBoost, LightGBM, CatBoost), and emerging deep learning models tailored to structured inputs. Emphasis will be placed on how these approaches handle heterogeneous variables, missing values, and class imbalance commonly encountered in biological datasets. Through translational applications, we will explore how models built on tabular data can predict patient outcomes, drug responses, or biomarker signatures. The session will also highlight practical strategies for feature selection, data preprocessing, model evaluation, and interpretability—critical elements for ensuring biological and clinical relevance. By the end of the session, participants will understand how to structure, analyze, and interpret tabular data using ML tools, enabling them to translate heterogeneous biomedical information into actionable knowledge for precision medicine and disease research.