Skill Enhancement Course for FYIMP/FYUGP

Course Title**: Data Analysis for Chemists**

Prerequisites: None

Format: Lectures (3 h/week), Laboratory (2 h/week)

Course Description: The course provides students with an introduction to the essential tools and techniques used in the analysis and interpretation of scientific data. Emphasizing both theoretical and practical aspects of chemical analysis, the course covers core statistical concepts such as probability distributions, hypothesis testing, regression analysis, and experimental design. The course also explores the use of data repositories and introduces chemometric methods for extracting meaningful patterns from complex datasets, particularly in the context of chemistry and related sciences. Through lectures, practical exercises, and projects, students will develop the skills needed to clean, visualize, and analyze scientific data effectively. By the end of the course, students will be equipped with a foundational toolkit for applying chemical data analysis in research and professional scientific contexts.

Learning Outcomes:

- Visualize data using charts (histograms, boxplots, scatter, bar, line, pie) to analyze distributions and communicate scientific findings.
- Compute and interpret descriptive and inferential statistics, including mean, variance, t-tests, p-values, and correlations.
- Understand core cheminformatics concepts.
- Analyze environmental or health datasets and conduct case studies using multivariate methods (PCA, PLS, clustering) on scientific data.
- Apply advanced preprocessing (e.g., smoothing, baseline correction) to spectroscopic data and use specialized tools for XRD, NMR, and EPR analysis.

Course developed by:

1. Prof. Rupam J Sarma, Department of Chemistry, Gauhati University
2. Prof. Sonit K. Gogoi, Department of Chemistry, Gauhati University
3. Dr. Himangshu P. Goswami, Department of Chemistry, Gauhati University

Course Title: **Data Analysis for Chemists (2L-0T-1P)**

Unit 1.1: Analytical Data and Statistics in Chemical Analysis (10 hours)

Statistics of data sampling and dependence on size; Accuracy and Precision; Systematic & Random Errors; Expression of results to significant figures; Error propagation (using partial derivatives & covariance)

Unit 2.2: Chemometrics and Multivariate Data Analysis (10 hours)

Analysis of complex chemical data using multivariate techniques; Topics include deconvolution, principal component analysis (PCA), partial least squares (PLS) regression, clustering; Preprocessing methods like baseline correction, smoothing (e.g., Savitzky–Golay or equivalent), and normalization. Applications include spectroscopy (IR, NMR) and chromatography datasets.

Unit 2.3: Working with Scientific Repositories and File Formats (10 hours)

Explores open data repositories (e.g., PubChem, PDB, NCBI etc.) and standard scientific file formats (EXCEL, CSV, SDF, JCAMP-DX or equivalent). Students learn to fetch, parse, and organize data using APIs and libraries like RDKit or Biopython etc.

Suggested Laboratory Experiments

- **Descriptive Statistics of Experimental Data**
  Aim: To compute and interpret mean, median, mode, variance, and standard deviation for a sample scientific dataset (e.g., plant growth, chemical yields).
- **Data Visualization of a Chemical/Biological Dataset**
  Aim: To create histograms, boxplots, and scatter plots to visualize the distribution and relationships data (e.g. distribution of molecular speeds, molecular weight of polymers, and heart rate vs. age).
- **Hypothesis Testing in Environmental Data**
  Aim: Employ the t-test to compare pollutant concentrations between two locations using an air quality dataset.
- **Data Cleaning and Preprocessing**
  Aim: To handle missing values, detect outliers, and normalize a tabular dataset (e.g., soil or water quality data).
- **Analysis of Spectral Data**
  Aim: To perform deconvolution and/or Principal Component Analysis (PCA) on infrared or NIR spectral data and interpret variance explained by components.
- **Regression Model on Physical Data**
  Aim: To build and evaluate a simple linear regression model predicting temperature from pressure in a physical chemistry dataset (e.g. reaction kinetics).

NB: The course Instructor(s) may introduce or change the suggested experiments as per requirements.

References:

- **Navidi, W. (2015).** *Statistics for Engineers and Scientists* (4^{th} Ed.). McGraw-Hill.
  Applied statistics with examples drawn from real scientific and engineering problems.
- **Leach, A. R., & Gillet, V. J. (2007).** *An Introduction to Chemoinformatics* (2nd ed.). Springer.
- **Kirkup, L. (2012)**, *Data Analysis for Physical Scientists*, Cambridge University Press.
- **O'Boyle, N. M., et al. (2011).** *Open Babel: An open chemical toolbox*. Journal of Cheminformatics, 3(1).