Problem Statement: Predicting Molecular Mutagenicity Using kNN for SPR Modeling

Mutagenicity, the ability of a substance to induce genetic mutations, is a critical property to evaluate for environmental, health, and safety considerations, particularly in the development of novel chemicals like drugs or solvents. This competition challenges participants to develop a k-Nearest Neighbors (kNN) classification model to predict whether a molecule is mutagenic based on its molecular descriptors.

Dataset

The dataset consists of molecular descriptors and mutagenicity data derived from experimental results on Salmonella typhimurium (Ames test). It includes:

- **Features:** Precomputed molecular descriptors such as Total Polar Surface Area (TPSA), molecular weight (MolWt), and BalabanJ index.
- **Target:** A binary label indicating whether a molecule is mutagenic (1) or non-mutagenic (0).

Objective

Participants are tasked with:

1. Building a kNN-based Quantitative Structure-Property Relationship (QSPR) model.
2. Optimizing the hyperparameter *k* using techniques like cross-validation.
3. Evaluating the model using the F1-score to balance precision and recall, given the importance of minimizing false positives and false negatives in mutagenicity prediction.

Deliverables

- A trained kNN model capable of predicting molecular mutagenicity.
- A presentation detailing the methodology, feature selection, hyperparameter optimization, and model evaluation.

Success Metrics

Submissions will be evaluated based on:

- **Primary Metric:** F1-score on the test dataset.
- **Secondary Metrics:** Accuracy, precision, recall, and clarity of methodology.
- **Bonus:** Novel and effective approaches to feature engineering or hyperparameter tuning.

Deadline: 2nd FEB

Are you ready to identify mutagenic molecules and make chemical innovations safer?