PREDICTING MOLECULAR MUTAGENICITY USING KNN FOR SPR MODELING

A KNN Classification Approach

- Mutagenicity is the ability of a substance to cause genetic mutations, an essential property for evaluating chemicals like drugs or solvents.
- The challenge is to predict whether a molecule is mutagenic or non-mutagenic based on molecular descriptors.

- Experimental results on Salmonella typhimurium (Ames test).
- Features: Molecular descriptors such as TPSA, MolWt, and BalabanJ index.

Target: A binary label indicating mutagenicity (1) or non-mutagenicity (0).

OBJECTIVE

- Build a kNN-based Quantitative Structure-Property Relationship (QSPR) model:
- Predict whether a molecule is mutagenic or non-mutagenic based on descriptors.
- Hyperparameter Optimization:
- Use cross-validation to tune the hyperparameter k (number of neighbors).
- Evaluation:
- Model performance will be evaluated using the F1-score, balancin precision and recall.

DISCUSSION

- Dataset Overview
- Features:
- TPSA: Total Polar Surface Area.
- MolWt: Molecular Weight.
- BalabanJ: Balaban J index.
- Other Descriptors: qed, NumValenceElectrons, etc.
- Target: Experimental value (1: Mutagenic, 0: Non-mutagenic).
- Total Rows: 5764 molecular samples with descriptors and corresponding mutagenicity status.

DATA PREPROCESSING

Data Preprocessing

- Step 1: Handling Missing Data:
  - Checked the dataset for missing values (none found).
- Step 2: Feature Selection:
  - Focused on relevant molecular descriptors: TPSA, MolWt BalabanJ, qed, NumValenceElectrons.
- Step 3: Splitting Data:
  - Split data into Training (80%) and Testing (20%) sets.

EXPLORATORY DATA ANALYSIS

- Boxplots for key features:
    - TPSA, MolWt, BalabanJ index—visualized to detect outliers and data distribution.
    - Ensured there were no significant outliers affecting the model's performance.

STANDARDIZATION

Used StandardScaler to scale the features (mean=0, standard deviation=1).

This ensures that all features contribute equally, especially since kNN is distance-based.

HYPERPARAMETER OPTIMIZATION (KNN)

- KNN Overview:
  - A distance-based classifier that assigns labels based on majority vote of k nearest neighbors.
- Hyperparameter: k (number of neighbors):
  - Used GridSearchCV for cross-validation, exploring value k between 1 and 30.
  - Evaluated performance using the F1-score (as it balar precision and recall

Model Evaluation

- Model Evaluation
- F1-Score on Test Data:
- F1-Score = 0.728.
- Classification Report:
- Precision: 0.69 for Mutagenic (class 1).
- Recall: 0.77 for Mutagenic (class 1).
- Accuracy: 68% on test data.
- Support: 650 mutagenic and 503 non-mutagenic samples in the test set

CONCLUSION

- A KNN-based model was successfully built to predic molecular mutagenicity using molecular descriptors
- Hyperparameter tuning (k=17) and feature scaling improved model performance.
- The model achieved a balanced F1-score of 0.728.

THANK YOU