winter project Presentation

VIBHANSHU
231143

Introduction

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.

Objective:

- Develop a k-Nearest Neighbors (kNN) model to predict molecular mutagenicity.
- Optimize hyperparameters for better accuracy.
- Evaluate model performance using cross-validation.

importance:

- Mutagenicity impacts drug design, environmental safety, and regulatory compliance.
- Data derived from Ames test on Salmonella typhimurium.

Mutagenicity and Its Relation to Molecular Descriptors

Mutagenicity refers to the ability of a chemical compound to cause genetic mutations by altering DNA. This property is crucial for assessing drug safety, environmental toxins, and regulatory approvals.

Key Molecular Descriptors and Their Influence on Mutagenicity

The dataset includes several descriptors that influence mutagenicity, including:

1. **Total Polar Surface Area (TPSA)**
   - Represents the sum of polar atomic surface areas in a molecule.
   - High TPSA can impact cell membrane permeability and drug absorption.

- Relation to mutagenicity:
  - Compounds with very high TPSA might be less likely to penetrate cell membranes and reach DNA, potentially reducing mutagenicity.
  - However, some highly polar molecules might interact strongly with DNA, increasing mutagenic potential.

2. Molecular Weight (MolWt)

- Measures the total mass of the molecule.
- Larger molecules often have more functional groups, increasing biological activity.
- Relation to mutagenicity:
  - Higher molecular weight may correlate with higher mutagenic potential, especially in aromatic compounds that intercalate DNA.

3. Balaban J Index

- A topological descriptor capturing molecular connectivity and branching.
- Related to molecular shape and complexity.
- Relation to mutagenicity:
  - Higher values might indicate more complex structures that interact with DNA, potentially increasing mutagenicity.

4. Number of Valence Electrons

- Determines reactivity of a molecule.
- Relation to mutagenicity:
  - Reactive molecules may form electrophilic species that bind to DNA, causing mutations.

Methodology:

- **1. Data Collection & Preprocessing:**
  - Dataset contains molecular descriptors (TPSA, MolWt, etc.) and binary labels (mutagenic: 1, non-mutagenic: 0).
  - Dropped irrelevant columns (Id, CAS, SMILES, etc.).
  - Scaled features using StandardScaler to normalize numerical values.
- **2. Model Selection:**
  - Chose k-Nearest Neighbors (kNN) for classification.
  - Justification: Simple, interpretable, effective for chemical data.
- **3. Data Splitting:**
  - 80% training, 20% test split.
  - Used stratification to maintain class distribution.

Hyperparameter Optimization and feature selection

- Goal: Find the best k value for kNN to balance bias-variance tradeoff.
- Method:
  - Used GridSearchCV with 5-fold cross-validation.
  - Tuned k in range 1 to 25.
  - Scoring metric: F1-score (balances precision & recall).
- Result:
  - Optimal - k = 13 was selected.
- Techniques Used:
  - Correlation Analysis: Removed highly correlated features.
  - SelectKBest (ANOVA F-test): Selected top features based on statistical signifinance.

Model Evaluation

- **Metrics Used:**
  - Accuracy: Measures overall correctness.
  - Precision: How many predicted positives were actually positive?
  - Recall: How many actual positives were correctly predicted?
  - F1-score: Harmonic mean of precision & recall.
- **Results:**
  - Accuracy: 70.77%
  - Precision: 73.32%
  - Recall: 75.69%
  - F1-score: 74.49

Performance Improvements & Next Steps

- **Limitations & Improvements:**
  a. kNN is sensitive to noisy data.
  b. Try different distance metrics (manhattan, minkowski).
  c. Use weighted kNN (weights='distance').
  d. Modify scoring factor for grid search in acuracy or precision basis.

Performance Improvements & observations

- **When we modify scoring factor for grid search in acuracy or precision basis:**

1. **We get the value of best k for both accuracy and precision scoring i.e.**
    a. Optimal k based on Accuracy: 0.689
    b. Optimal k based on Precision: 0.7819
2. **When we use closest integer value of k for each case we get folllowing results:**
    a. **When acuracy is more important**
        i. Accuracy: 71.0321
        ii. Precision: 73.3038
        iii. Recall: 0.7646
        iv. F1-score: 0.7485

a. **When precision is more important**
    i. Accuracy: 71.5525
    ii. Precision: 77.3810
    iii. Recall: 0.7000
    iv. F1-score: 0.7351

2. When we change weights from uniform to distance we get:

- **When we change weights and try various value of k mannually to get best result we get k = 16 and performance is:**
    - Accuracy: 72.2463
    - Precision: 74.4807
    - Recall: 0.7723
    - F1-score: 0.7583

Conclusion

- **Theoritical Conclusion:**
  - Successfully built a kNN-based QSPR model.
  - Used cross-validation to optimize k.
  - Evaluated performance with multiple metrics.
  - Identified potential improvements for future work.

- **Observational conclusion:**
  - When we change scoring parameter we get variation in results but in them f1 score don’t wary that much and for precision one both accuracy and precision is better than for accuracy one.
  - When we change the weight parameter in KNN from uniform to distance and manually adjust k for some values near 13 (from previous part) we get improvement in all the predictions
- **Hence distance work more accurate for this case**

THANK YOU