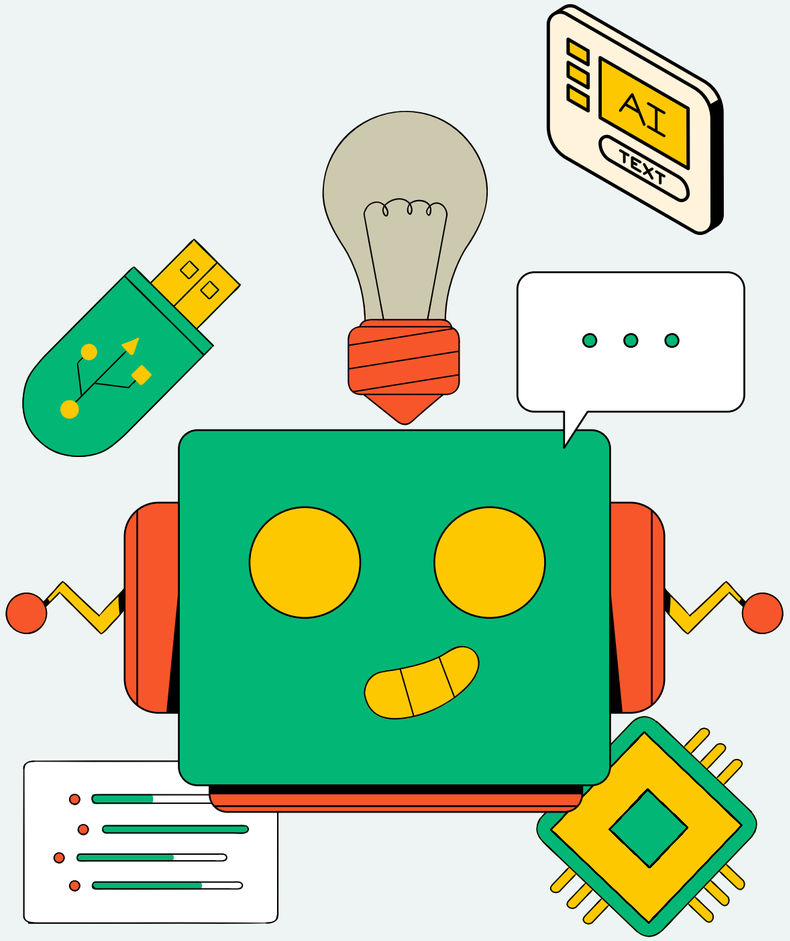
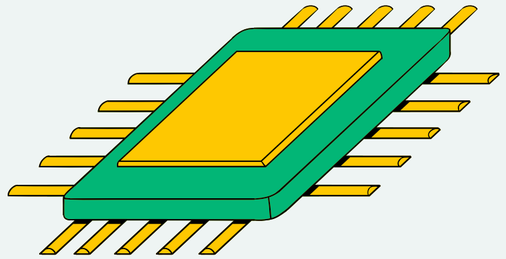


PREDICTING MOLECULAR MUTAGENICITY USING KNN FOR SPR MODELING

PRESENTATION

By – Aryan Kadam



INTRODUCTION

Objective: To build a kNN-based Quantitative Structure-Activity Relationship (QSAR) model to predict the mutagenicity of molecules using molecular descriptors.

Goal: Optimize the model by tuning hyperparameters and evaluating performance using various metrics, including F1-score, accuracy, precision, recall, and confusion matrix.

DATASET OVERVIEW

Dataset: mutagenicity.csv (file used in the analysis)

Source: The dataset contains molecular descriptors and their corresponding mutagenicity (whether a molecule is mutagenic or not).

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) .

Aryan Kadam

METHODOLOGY

k-Nearest Neighbors (kNN) Algorithm:

- A simple, yet powerful classification algorithm based on distance measures.
- For a given test point, the kNN algorithm finds the 'k' nearest neighbors in the training set and predicts the class based on majority voting.
- Parameters: n_neighbors, weights, metric, algorithm.

Modeling Workflow:

a. Data Loading: Load the dataset from CSV.
b. Feature Selection: Identify relevant molecular descriptors for prediction.
c. Preprocessing: Standardize the features to ensure they are on the same scale.
d. Model Training: Train a kNN model using the training dataset.
e. Hyperparameter Tuning: Optimize the hyperparameters using GridSearchCV.
f. Model Evaluation: Evaluate the performance using various metrics

FEATURE SELECTION

- **Selected Features:**
- Features such as NumValenceElectrons, MolLogP, and other molecular properties are selected for the model.
- These features capture important characteristics of molecules that may influence their mutagenicity.

- **Preprocessing:**
- Standardization: The features are standardized using StandardScaler, which scales the data to have zero mean and unit variance.
- This ensures that all features contribute equally to the model, preventing dominance by variables with larger scales.

HYPERPARAMETER OPTIMIZATION

Hyperparameters for kNN:

- n_neighbors: The number of nearest neighbors to consider.
- weights: Method to weight the neighbors ('uniform' or 'distance').
- metric: The distance metric used for computing neighbors ('euclidean', 'manhattan', 'minkowski').
- algorithm: The algorithm used to find neighbors ('ball_tree', 'kd_tree', 'brute').

Optimization Process:

- GridSearchCV: Used for hyperparameter tuning, searching through the specified grid of parameters.
- Cross-validation: 5-fold cross-validation is used to evaluate the model for each combination of hyperparameters.
- Scoring Metric: The F1-score is used as the scoring metric since it balances precision and recall, which are important for mutagenicity prediction.

MODEL EVALUATION

- **Performance Metrics:**
- **F1-Score:** The primary metric used to evaluate the model. It balances precision and recall, minimizing both false positives and false negatives.

- **Accuracy:** The proportion of correct predictions (true positives and true negatives).

- **Precision:** The proportion of correctly predicted positive cases (mutagenic molecules).

- **Recall:** The proportion of actual positive cases correctly predicted.

- **Confusion Matrix:** Used to visualize the performance of the classification model in terms of true positives, true negatives, false positives, and false negatives.

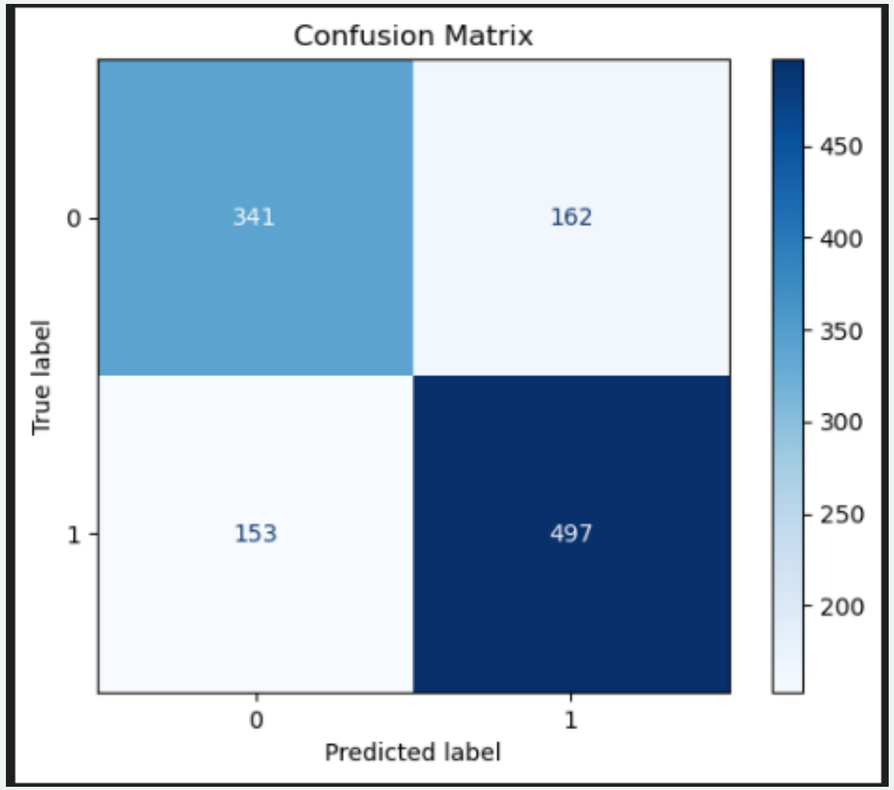
- **Classification Report:** Provides a summary of key classification metrics including F1-score, precision, recall, and support.

RESULTS

- **Best Hyperparameters:** After performing grid search, the best combination of hyperparameters was found.
  - Best n_neighbors: 14
  - Best weights: 'distance'
  - Best metric: 'euclidean'
  - Best algorithm: 'ball_tree'

- **Model Evaluation on Test Set:**
- **F1-Score on Test Set:** 0.7594
- **Accuracy:** 0.7268
- **Precision:** 0.7542
- **Recall:** 0.7646

CONFUSION MATRIX
&
CLASSFIATION REPORT



Classification Report:

	precision	recall	f1-score	support
0	0.69	0.68	0.68	503
1	0.75	0.76	0.76	650
accuracy			0.73	1153
macro avg	0.72	0.72	0.72	1153
weighted avg	0.73	0.73	0.73	1153

Aryan Kadam

CONCLUSION

- **Summary:**

- A kNN-based QSAR model was successfully built to predict the mutagenicity of molecules.

- Hyperparameter tuning via GridSearchCV optimized the kNN model's performance.

- The model showed a balanced F1-score, with competitive accuracy, precision, and recall values.

THANK YOU !