PREDICTING MOLECULAR MUTAGENICITY USING KNN FOR SPR MODELING

-BY TANYA SINGH
231086

INTRODUCTION

Mutagenicity refers to a substance's ability to cause genetic mutations. It is essentiaL in drug discovery, environmental safety, and chemical risk assessment.

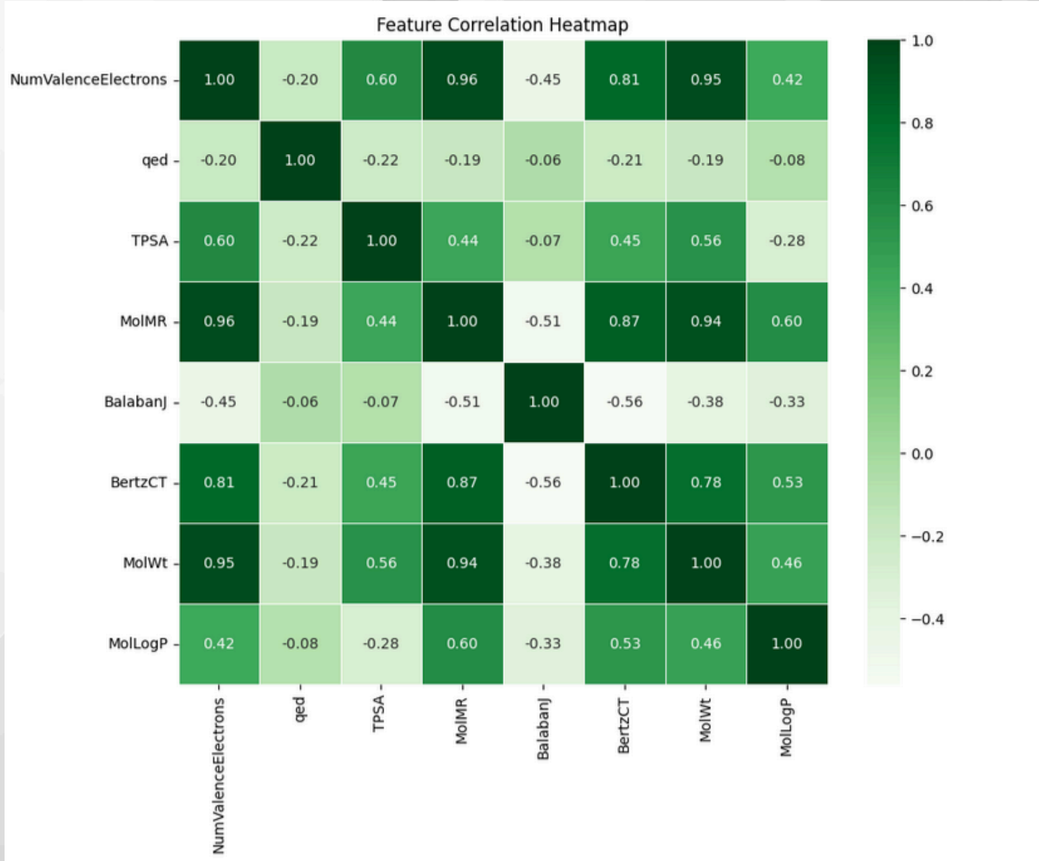
This study aims to predict molecular mutagenicity using a k-Nearest Neighbors (kNN) model based on molecular descriptors.

DATASET OVERVIEW

- **Features**: Molecular descriptors influencing mutagenicity:
    a. Total Polar Surface Area (TPSA) – Affects permeability and absorption.
    b. Molecular Weight (MolWt) – Higher values can increase mutagenic potential.
    c. Balaban J Index – Related to molecular complexity.
    d. Number of Valence Electrons – Affects reactivity with DNA.
- Target: Binary classification(Experimental value) (1 = Mutagenic, 0 = Non-mutagenic).

DATA PREPROCESSING & FEATURE CORRELATION

- Steps Taken:
  - Dropped irrelevant columns (Id, CAS, SMILES, etc.).
  - Standardized features using StandardScaler.
  - Handled missing values by imputing the mean.
- Feature Correlation Analysis:
  - Heatmap to visualize relationships between descriptors.



METHODOLOGY - MODEL SELECTION

- Why kNN?
  - Simple and interpretable.
  - Works well for molecular similarity-based classification.
- Data Splitting:
  - 80% training, 20% test (stratified sampling for class balance).

HYPERPARAMETER OPTIMIZATION & FEATURE SELECTION

- Goal: Find the best k value to balance bias-variance tradeoff.
- Method:
  - Used GridSearchCV with 5-fold cross-validation.
  - Tuned k in range 1 to 25.
  - Scoring Metric: F1-score

- Feature Selection Techniques Used:
  - Correlation Analysis: Removed highly correlated features.
    - SelectKBest (ANOVA F-test): Selectedwq top descriptors based on significance.

MODEL TRAINING & EVALUATION

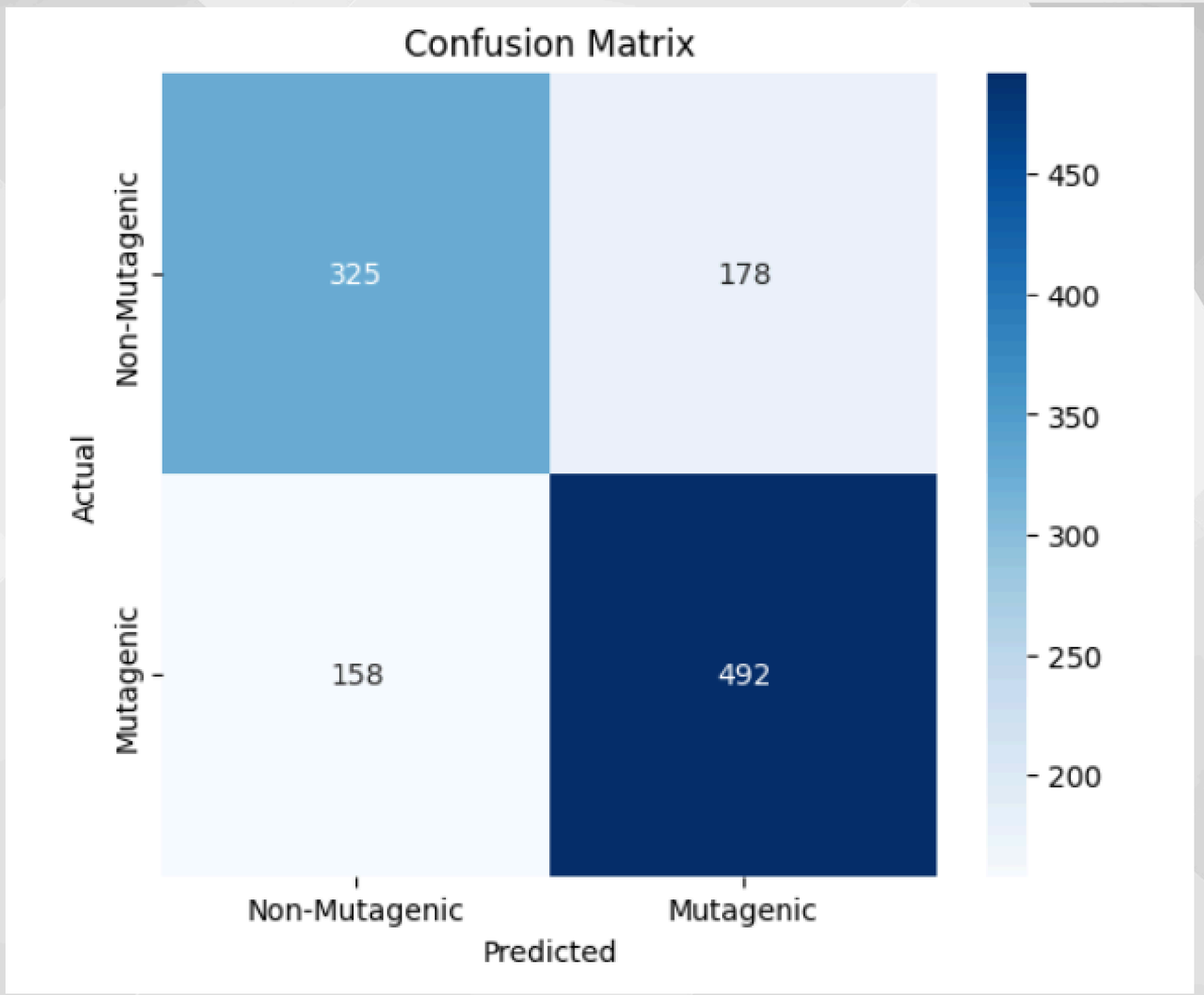
- Evaluation Metrics Used:
  - Accuracy: Measures overall correctness.
  - Precision: How many predicted positives were actually positive?
  - Recall: How many actual positives were correctly predicted?
  - F1-score: Harmonic mean of precision & recall.
- Results:

  Accuracy: 0.7086

  Precision: 0.7343

  Recall: 0.7569

CONFUSION MATRIX ANALYSIS



Performance Improvements & Next Steps

- Identified Limitations:
  - kNN is sensitive to noisy data.
  - Distance metric choice impacts accuracy.
- Future Improvements:
  - Try different distance metrics (Manhattan, Minkowski).
  - Implement weighted kNN (weights='distance').
  - Compare performance with ensemble models (Random Forest, XGBoost).

CONCLUSION

- Successfully built a kNN-based QSPR model for mutagenicity prediction.
- Used cross-validation to optimize k.
- F1-score optimized model balances false positives & false negatives.
- Future work will include advanced feature selection and distance-weighted kNN.

THANK YOU