Assignment: Predicting Molecular Mutagenicity Using KNN for SPR Modelling.

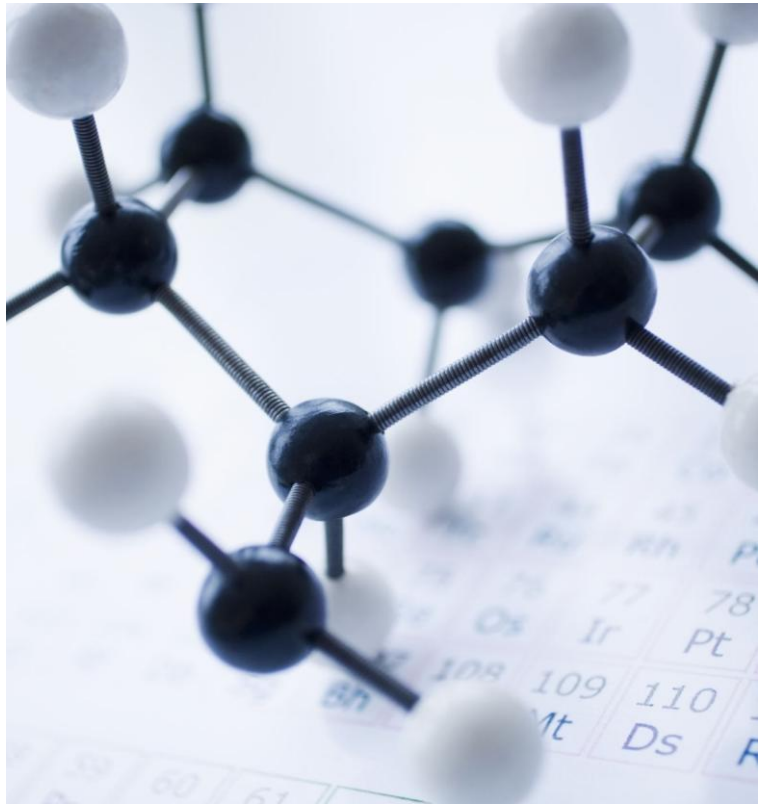
Name: Arnav Harshit

Roll No: 230200

Date: 31 Jan 2025

Problem Statement:

- Mutagenicity is the ability of a substance to induce genetic mutations.
- Critical for environmental health, and safety considerations.
- Objective:
  - Predict Whether a molecule is mutagenic using molecular descriptors.
- Dataset:
  - Molecular descriptors (TPSA, MolWT) and binary labels (mutagenic and non-mutagenic).



Dataset Overview

- Features:

NumValenceElectrons	Qed	TPSA	MolMR	BalabanJ	BertzCT	MolWT	MolLogP

- Target:
  - Binary Label: 1 (mutagenic) or 0 (non-mutagenic).
- **Dataset Size:** 5764
- **Source:** Experimental results on Salmonella typhimurium (Ames test).

Methodology:

- Steps:
  1. **Data Preprocessing:** Standardization of features.
  2. **Model:** k-Nearest Neighbors (kNN) classifier.
  3. **Hyperparameter Tuning:** GridSearchCV to find the best k.
  4. **Evaluation:** F1-score, accuracy, precision, and recall.

Data Preprocessing:

Train-Test Split:
80% training, 20% testing.

Standardization:

- Features scaled using **StandardScalar** to have zero mean and unit variance.

Model Building

- **Algorithm:** k-Nearest Neighbors (kNN).
- **Hyperparameter Tuning:**
  - GridSearchCV used to find the optimal k (number of neighbors).
  - Search range: k = 1 to k = 20.
- **Best K:** 15

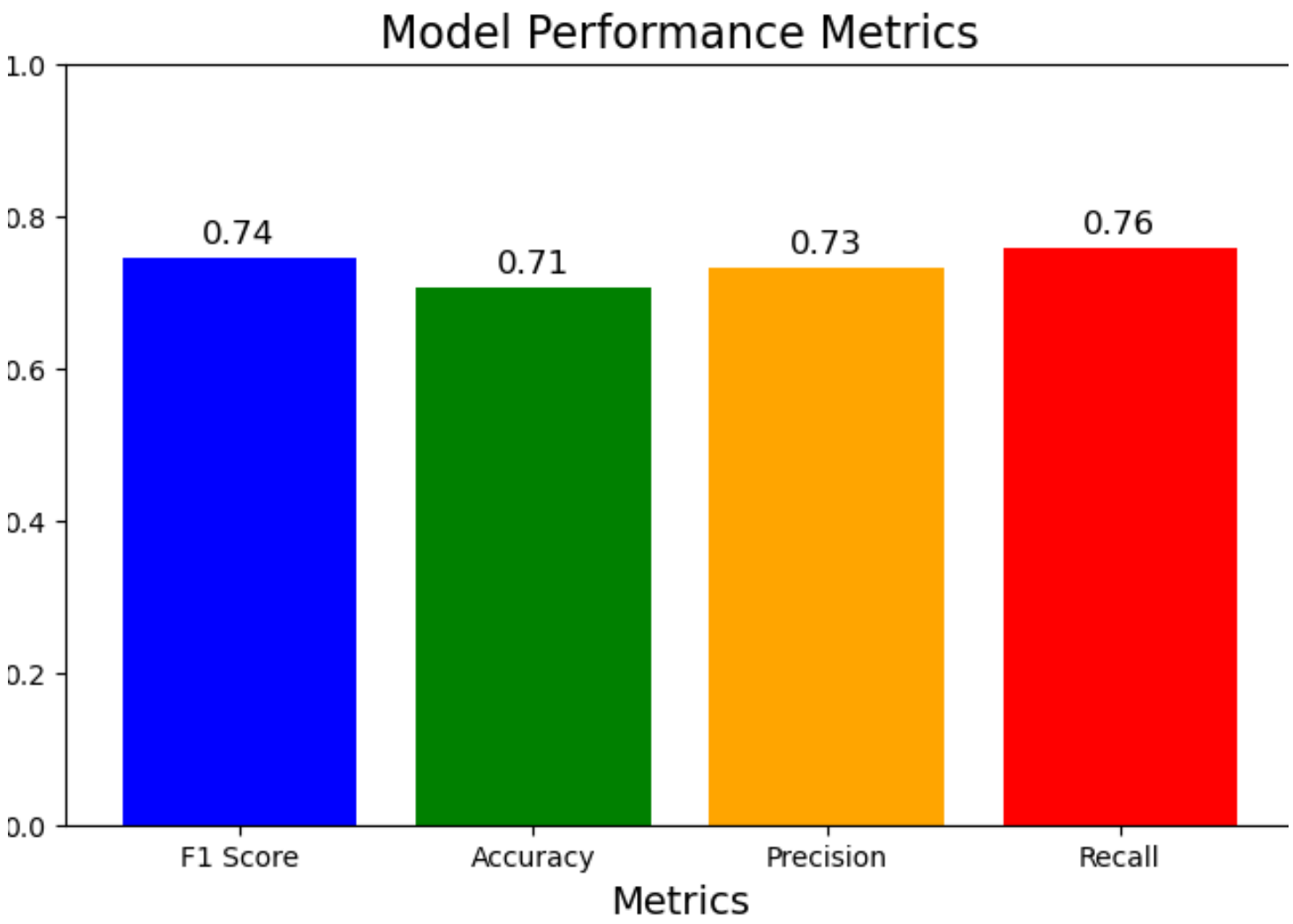
Model Evaluation:

- Metrics:
  - F1 Score: 0.744
  - Accuracy: 0.706
  - Precision: 0.731
  - Recall: 0.757
- Classification Report:

	Precision	recall	F1-Score	Support
0	0.67	0.64	0.66	503
1		0.76	0.74	650
Accuracy			0.71	1153
Macro avg	0.70	0.70	0.70	1153
Weighted avg	0.70	0.71	0.71	1153

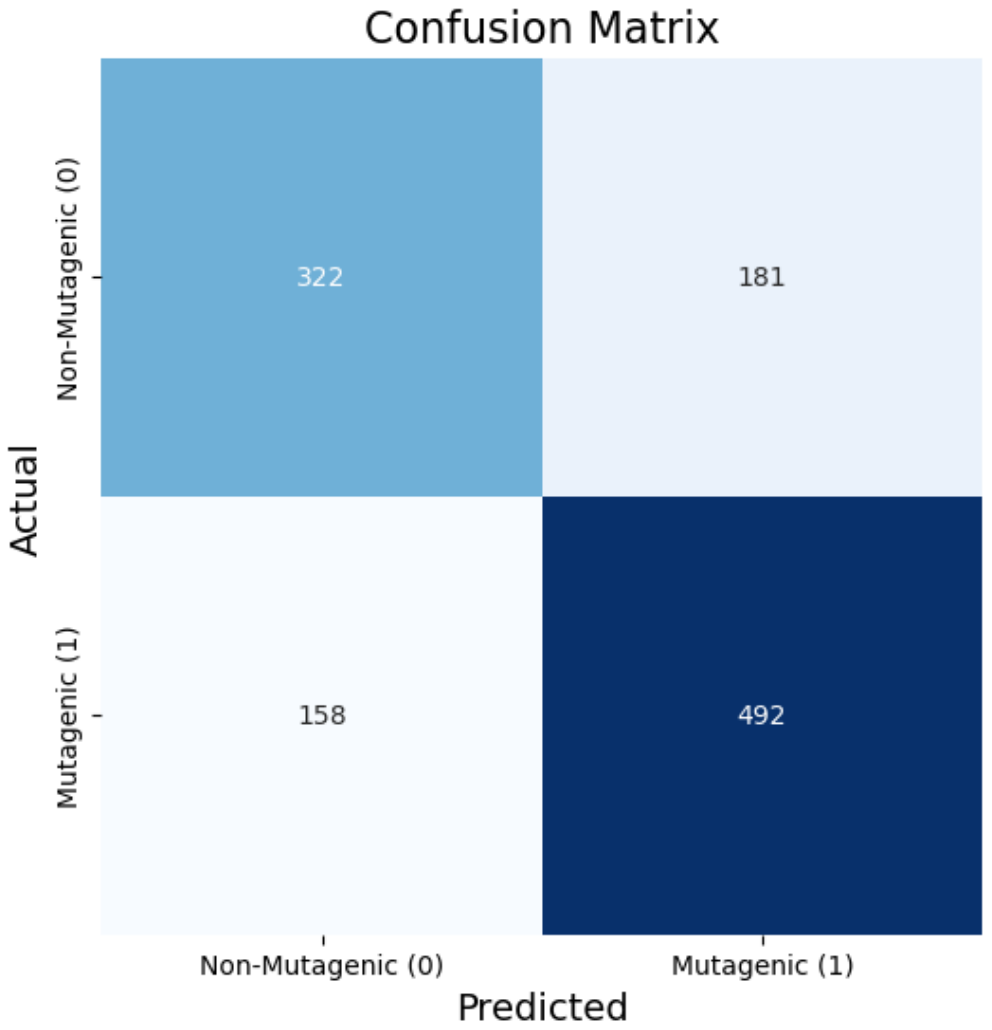
Model Performance Metrics

- The model achieved an F1-Score of 0.74 , accuracy of 0.71, precision of 0.73, and recall of 0.76.



Confusion Matrix

- The confusion matrix shows the number of correct and incorrect predictions for each class (Non-Mutagenic and Mutagenic).



Results

Best Model Performance:

- F1 Score: 0.74
- Accuracy: 0.71

Key insights:

- The model performs well in predicting mutagenicity.

Challenges and Learnings:

Challenges:

- Selecting the optimal k for kNN.

Learnings:

- Importance of hyperparameter tuning.
- kNN is sensitive to the choice of k and distance metric.

Conclusion:

- Successfully built a kNN - based SPR model for mutagenicity prediction.
- Achieved F1 Score: 0.744 on the test dataset.
- Demonstrated the importance of hyperparameter tuning.