Intro to ML for
Chemical Engineers

-

Presentation 2025

# Project presentation

Presented by Shaurya Srivastava

Presentation 2025

-

shaurya srivastava

# Origin of the creative idea

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.

# Project vision and mission

The mission of this final task is to build a knn model to detect the mutation in the compounds using Quantitative Structure-Property Relationship (QSPR) model. where using index like balaban j index and other parameters we classify the compounds

01. First we do some data cleaning and some dataa logarithmic normalization and using scaling we scale the data to use it in model training

02. the we implement the the knn model for classification where we get accuracy around 80% and upon hyperparameter tunning by using grid search cv and cross validation methods we get that k value around 17 is good to get this accuracy

Presentation 2025

-

shaurya srivastava

# Inspiration and creativity

using chat gpt and different ai tools such as perplexity and different articles present on internet such as :

1)https://medium.com/@kad.denuwaraeng/why-is-knn-a-game-changer-for-q-spr-modeling-in-chemical-engineering-7d49943155c8

2)https://pmc.ncbi.nlm.nih.gov/articles/PMC8339974/

# Hyperparameters

## knn

After experimentation, k was set to 9, a balance between accuracy and generalization.

## Distance Weighting

Inverse-distance weighting was used to give closer neighbors more influence on predictions

## Distance Metric

Euclidean distance was chosen for calculating the distance between data points.

## Cross-Validation

5-fold cross-validation was used to assess the model's robustness and generalize its performance.

# Model validation

Accuracy

**79.94 %**

F-1 score

**80.37%**

# conclusion

The KNN model demonstrated strong predictive capability for experimental chemical property classification, achieving an accuracy of 79.94% and an F1 score of 82.37%. Future research could explore alternative classifiers, feature extraction techniques, and hyperparameter tuning to further enhance the model's performance and generalize its predictive capabilities.

# Thank You