Practical Data Science with Python (COSC 2670/2738)
Assignment 2: Data Modelling and Analysis

Predicting Drug-Induced Autoimmunity Using Molecular Descriptors

Statement of Originality:

I declare that this report is my own work and has not been submitted for assessment elsewhere. All sources used have been properly acknowledged.

Chia-Cheng Chang

RMIT University

Date: May 25, 2025

Table of Contents

• An abstract/executive summary

This study applies machine learning techniques to predict drug-induced autoimmunity (DIA) using molecular descriptors calculated with RDKit. In result of this study, the feature fr_anilines (Number of anilines) is the strongest predictor of DIA risk with the highest positive corelation score. However, FractionCSP3 (The fraction of C atoms that are SP3 hybridized.) has the strongest negative correlation with the of DIA risk indicating that it may provide pretective effects. Last but not least,K- Nearest Neighbors classifier have better prediction than Decision Trees in cross-validation.

• Introduction

Drug-induced autoimmunity (DIA) occurs when a medication triggers the immune system to mistakenly attack the body's own tissues, leading to autoimmune-like symptoms. As a result, predicting the risk associated with a drug's molecular composition has become a critical step in pharmaceutical research and development.The task focuses on building the predictive model and identifying potential autoimmune risks associated with drug candidates. The dataset provided facilitates the development of interpretable models for drug toxicity prediction, contributing to advancements in computational toxicology and drug safety assessment.

• **Methodology**

1: data preparation

1.1 Check for Missing Values

1.2 Check for duplicated values

1.3 Check for Data integrity(All molecular descriptors within descriptor file)

1.4 Drop the columns that give no signal

2: Data Exploration

2.1 Look into RDKit_ChemDes and select 10 important features to explore

2.2 Calculate the descriptive statistics

2.3 Create histogram showing frequency distribution for each selected feature

2.4 Observe the graph

2.5 Select 10 important features pairs to explore

2.6 Calculate the coorelation of each pair

2.7 Create scatter plot to visualize relation of each pair using different color to show the label.

3: Data Modelling

3.1 Use coorelation function to find top 60 strong correlation with DIA label, store the features into a list.

3.2 Train KNeighbors and DecisionTree Classifier using columns from the features list

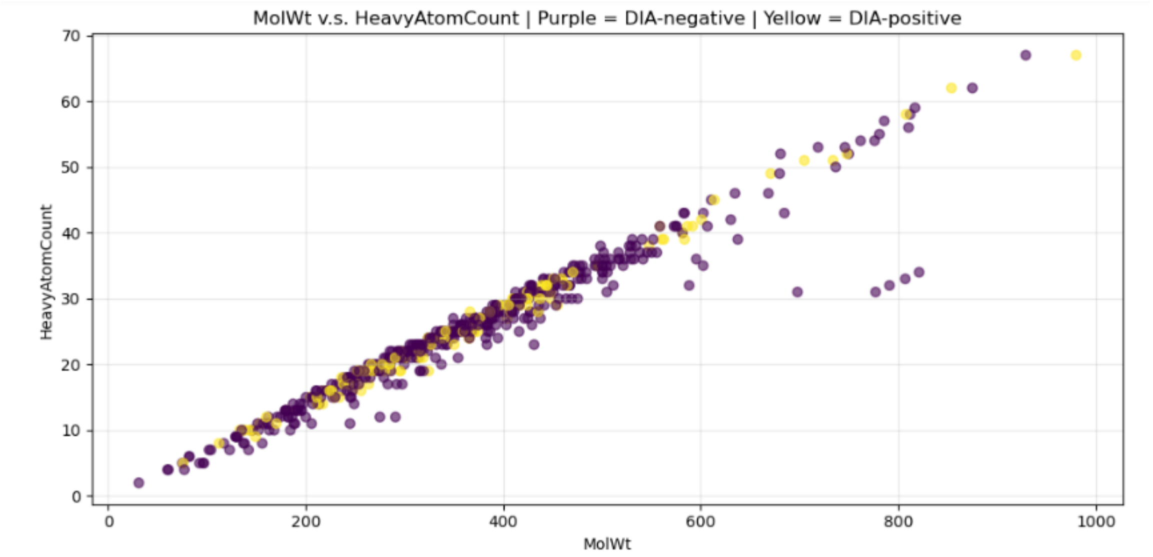
3.3 Get the mean values to compare two models

• Results

1. MolWt', 'HeavyAtomCount has strong correlation. Molecular weight is highly related to h eavy atom count.

2. The top 3 DIA Risk factors are "fr_aniline(corr: 0.216)", "fr_priamide (corr: 0.171)" and "SlogP_VSA10(corr: 0.152)"

3. K- Nearest Neighbors classifier have better prediction than Decision Trees in cross-validation.



- Discussion

Decision Tree outperformed KNN **o**n average, likely due to: 1. Decision Trees can learn complex split patterns, while KNN relies on how close the data are. 2. Trees pick only the most useful features and ignore the rest, reducing noise. KNN uses all features — even unhelpful ones. 3. Decision Trees reduce overfitting and find clear split rules**.**

- Conclusion

Decision Tree classifier, combined with feature selection, demonstrated robust performance on predicting drug-induced autoimmunity using RDKit descriptors.. These results offer a valuable step toward safer drug screening using machine learning.

- References

W3Schools. (n.d.). *Matplotlib Subplot*. Retrieved May 25, 2025, from https://www.w3schools.com/python/matplotlib_subplot.asp

GeeksforGeeks. (n.d.). *Enumerate in Python*. Retrieved May 25, 2025, from https://www.geeksforgeeks.org/enumerate-in-python/

GeeksforGeeks. (n.d.). *matplotlib.pyplot.tight_layout() in Python*. Retrieved May 25, 2025, from https://www.geeksforgeeks.org/matplotlib-pyplot-tight_layout-in-python/

pandas. (n.d.). *pandas.DataFrame.nunique*. Retrieved May 25, 2025, from https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.nunique.html