Predictive Drug Classification:

Leveraging Patient Data for Personalized Treatment

Mrityunjay Sharma

210107054

Submission Date: April 25, 2024

Final Project submission

Course Name : Applications of AI and ML in chemical engineering

Course Code: CL653

Contents

1 Executive Summary

The project focuses on drug classification using machine learning models on medical test data gathered from various people. The proposed solution utilizes classification algorithms like ***logistic regression, KNN, SVM, Naïve Bayes, decision trees, and random forests***, along with techniques like ***hyperparameter tuning*** to optimize model performance. The methodology involves data preprocessing steps like cleaning, normalization, and class balancing, followed by applying the classification models and tuning their hyperparameters. The expected outcome is ***improved drug classification accuracy*** after hyperparameter tuning, identifying the most effective models and configurations for this task. The goal is to provide an accurate and reliable solution for drug classification based on medical test data to aid in better diagnosis and treatment.

2 Introduction

Background:

The project aims to develop a drug classification methodology that classifies drugs based on their usage to facilitate the drug repositioning process. Drug repositioning involves finding new therapeutic uses for existing drugs, but identifying novel side effects of drugs for repurposing is challenging due to the vast number of drugs and diseases. The proposed classification methodology provides an efficient way to explore new applications for existing drugs by matching them based on usage. Additionally, detecting potential adverse drug reactions is crucial as unexpected adverse effects can impact patient health. While the context is pharmacology, the drug development lifecycle intersects with chemical engineering principles like formulation, manufacturing, and modeling, so streamlining repositioning through robust classification could potentially aid chemical engineering aspects of repurposing drugs.

Problem Statement:

The project aims to develop a methodology to classify existing drugs based on their intended therapeutic usage or applications. This is to facilitate the process of drug repositioning, which involves finding new therapeutic uses for existing drugs beyond their original indications. A major challenge is efficiently identifying the potential novel side effects (both beneficial and adverse) of existing drugs across the vast number of diseases, which is difficult due to the immense number of drug-disease combinations to evaluate. By classifying drugs based on usage, the proposed methodology seeks to provide an efficient framework to match existing drugs with new potential disease applications for repositioning based on their classified profiles. These are the references Explainable Machine Learning for Drug Classification, Research on Drug Classification Using Machine Learning Model, Drug Classification using Machine Learning and Interpretability.

Objectives:

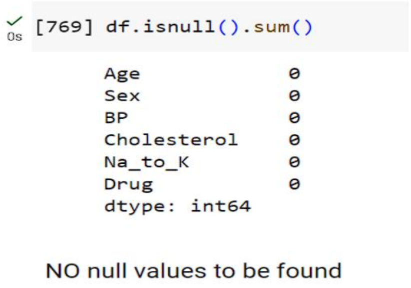
The main objective of the project is to develop a machine learning-based model that can accurately classify drugs based on medical test results obtained from patients, leveraging machine learning techniques to analyze the test data and map drugs to their appropriate therapeutic classes or intended usages. The goal is to create a robust model that can effectively handle varying test result values and patterns to ensure precise drug classification, ultimately aiding in drug repositioning, patient care, drug development, and analysis of potential side effects.

3 Methodology

Data Source: The dataset is taken from Kaggle.com which is a Drug Classification dataset. The target feature is Drug type. The feature sets are: ***Age, Sex, Blood Pressure Levels (BP), Cholesterol Levels, Na to K Ratio***.

Data Preprocessing:

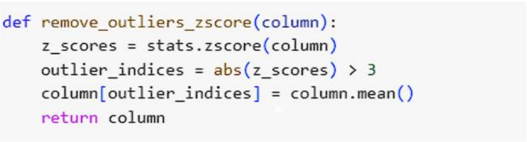
1. **Data Cleaning:** Handling of missing data points, which may involve imputation techniques such as mean substitution or deletion of records with missing values.



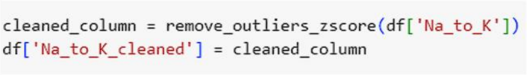
```
[769] df.isnull().sum()

Age            0
Sex            0
BP             0
Cholesterol    0
Na_to_K        0
Drug           0
dtype: int64
```

NO null values to be found

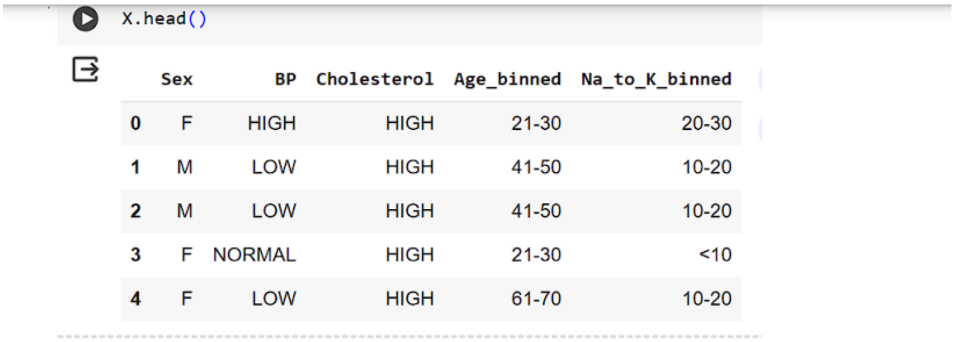


```
def remove_outliers_zscore(column):
    z_scores = stats.zscore(column)
    outlier_indices = abs(z_scores) > 3
    column[outlier_indices] = column.mean()
    return column
```



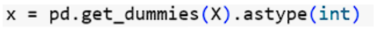
```
cleaned_column = remove_outliers_zscore(df['Na_to_K'])
df['Na_to_K_cleaned'] = cleaned_column
```

2. **Text Preprocess: One-hot encoding or Label encoding** for converting categorical features to numerical features.

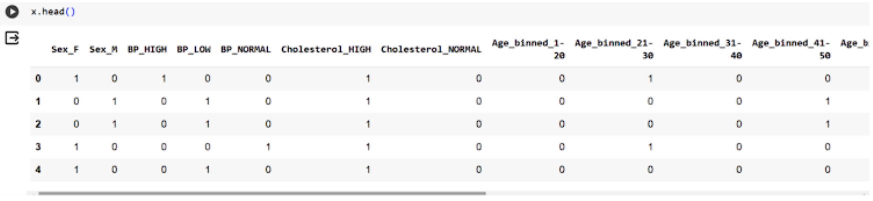


```
X.head()
```

	Sex	BP	Cholesterol	Age_binned	Na_to_K_binned
0	F	HIGH	HIGH	21-30	20-30
1	M	LOW	HIGH	41-50	10-20
2	M	LOW	HIGH	41-50	10-20
3	F	NORMAL	HIGH	21-30	<10
4	F	LOW	HIGH	61-70	10-20

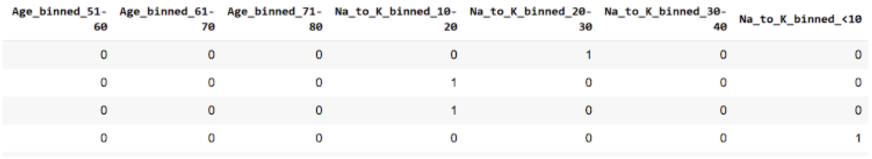


```
x = pd.get_dummies(X).astype(int)
```



```
x.head()
```

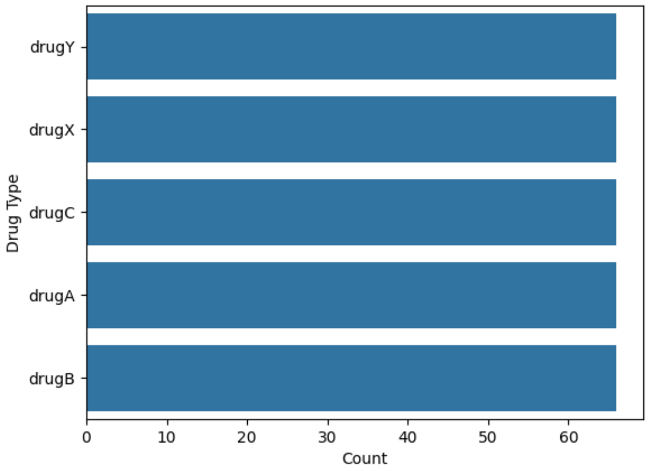
	Sex_F	Sex_M	BP_HIGH	BP_LOW	BP_NORMAL	Cholesterol_HIGH	Cholesterol_NORMAL	Age_binned_1-20	Age_binned_21-30	Age_binned_31-40	Age_binned_41-50	Age_b
0	1	0	1	0	0	1	0	0	1	0	0	
1	0	1	0	1	0	1	0	0	0	0	1	
2	0	1	0	1	0	1	0	0	0	0	1	
3	1	0	0	0	1	1	0	0	1	0	0	
4	1	0	0	1	0	1	0	0	0	0	0	



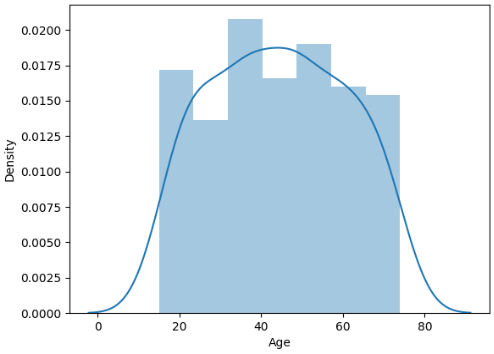
Age_binned_51-60	Age_binned_61-70	Age_binned_71-80	Na_to_K_binned_10-20	Na_to_K_binned_20-30	Na_to_K_binned_30-40	Na_to_K_binned_<10
0	0	0	0	1	0	0
0	0	0	1	0	0	0
0	0	0	1	0	0	0
0	0	0	0	0	0	1

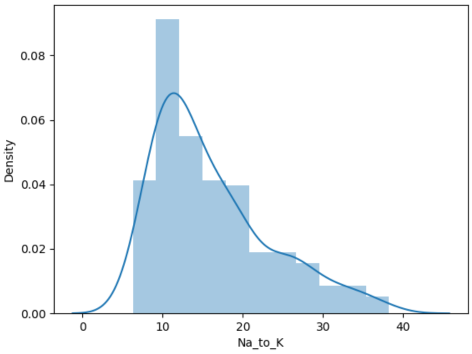
3. **Handling Imbalanced Data:** Apply techniques like oversampling or undersampling to balance class distributions if significant imbalances are observed.

Applied **SMOTE** for class balancing

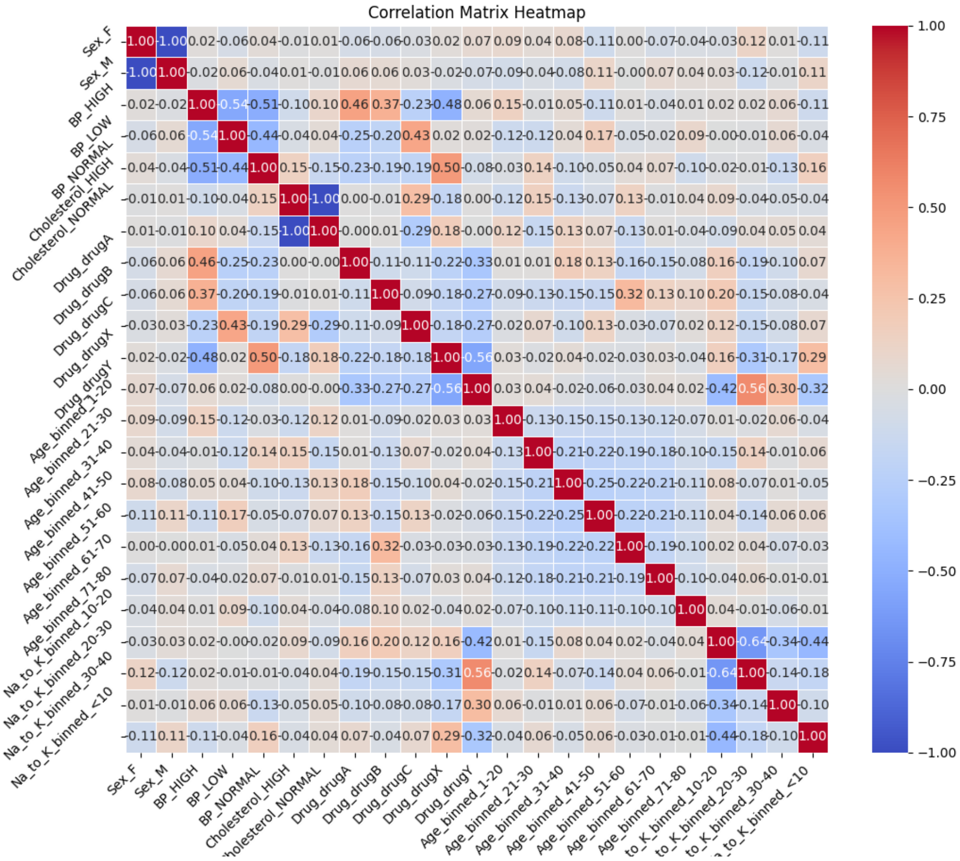


4. **Exploratory Data Analysis (EDA):** Conduct exploratory data analysis to gain insights into the relationships between features and the target variable. Visualize distributions, correlations, and patterns in the data using techniques like histograms, scatter plots, or correlation matrices. Identify any potential trends or outliers that may impact model training and interpretation.





The distribution of **'Age'** column is **symetric**, since the skewness value between -0.5 and 0.5. The distribution of **'Na_to_K'** column is **moderately skewed**, since the skewness value is ***between 0.5 and 1***. It can also be seen from the histogram for 'Na_to_K' column



Model Architecture:

The project employs various machine learning models, including logistic regression, support vector machines (SVM), k-nearest neighbors (KNN), decision trees, and random forests, for the task of drug classification based on medical test results.

Among these models, the random forest algorithm emerged as the best-performing model for this problem. The random forest architecture is well-suited to solve the drug classification problem because it can effectively model the complex relationships between medical test results and drug classes, while providing robustness to noise, handling high-dimensional data, and offering interpretability through feature importance rankings. Additionally, the ensemble nature of random forests helps to improve the overall prediction accuracy and generalization performance of the model.

Tools and Technologies:

Programming Languages: ***Python***

Machine Learning Libraries/Frameworks: ***scikit-learn*** (machine learning library for Python)

Data Manipulation and Analysis:

- ***Pandas*** (data manipulation and analysis library for Python)
- ***NumPy*** (numerical computing library for Python)
- ***Matplotlib/Seaborn*** (data visualization libraries for Python)

Integrated Development Environment (IDE): ***Google Colab***

Additional Python Libraries (as needed):

- ***scikit-optimize*** (for hyperparameter tuning)
- ***imbalanced-learn*** (for handling imbalanced datasets)

4 Implementation Plan

Development Phases:

Phase 1: Project Planning and Data Acquisition

Phase 2: Data Preprocessing and Feature Engineering

Phase 3: Model Development and Training

Phase 4: Model Selection and Interpretation

Model Training:

- Experiment with various machine learning algorithms (e.g., logistic regression, random forests, SVMs)
- Split the data into training and testing sets
- Train and evaluate the models using appropriate evaluation metrics
- Perform hyperparameter tuning to optimize model performance
- Implement cross-validation techniques (e.g., k-fold cross-validation)

Model Evaluation: *Accuracy, Confusion Matrix*

5 Testing and Deployment

Testing Strategy:

1. **Hold-Out Test Set:** Reserve a portion of the dataset (e.g., 20-30%) as a hold-out test set that is not used during model training or validation. This test set should be representative of the real-world data distribution and remain untouched until the final model evaluation.

2. **Stratified Sampling:** When creating the hold-out test set, employ stratified sampling techniques to ensure that the class distributions in the test set are representative of the overall dataset.

3. **Cross-Validation:** During the model development and tuning phase, use cross-validation techniques like k-fold or stratified k-fold cross-validation to evaluate the model's performance on different subsets of the training data.

4. **Real-World Data Testing:** If possible, obtain a separate real-world dataset that was not used during model training or validation. This dataset should ideally come from a different source or represent a different patient population. Evaluating the model's performance on this external dataset can provide valuable insights into its generalization capabilities.

Deployment Strategy:

1. **Scalability:** Harness cloud platforms (e.g., AWS, GCP, Azure) for scalable computational resources. Employ containerization (e.g., Docker) and orchestration (e.g., Kubernetes) for efficient scaling and deployment.

2. **Performance Optimization:** Apply model optimization techniques (e.g., quantization, pruning) for streamlined inference. Utilize hardware accelerators (e.g., GPUs, TPUs) to expedite model inference.

3. **Maintenance and Updates:** Implement CI/CD pipelines for automated model updates and deployments. Ensure version control for model artifacts and codebase changes.

Ethical Considerations:

1. **Data Privacy and Security:**

- Adhere strictly to data privacy regulations (e.g., GDPR, HIPAA) for patient medical data handling.

- Implement robust encryption and access controls to safeguard sensitive information.

2. **Transparency and Interpretability:**

- Utilize model interpretation methods (e.g., SHAP, LIME) to understand decision-making processes.

3. **Human Oversight and Control:**

- Ensure models aren't sole determinants of critical medical decisions.

- Engage domain experts in decision-making, using the model as support.

- Implement mechanisms for human intervention when needed.

4. **Continuous Monitoring and Improvement:**

- Monitor model performance regularly, gathering feedback to address ethical concerns.

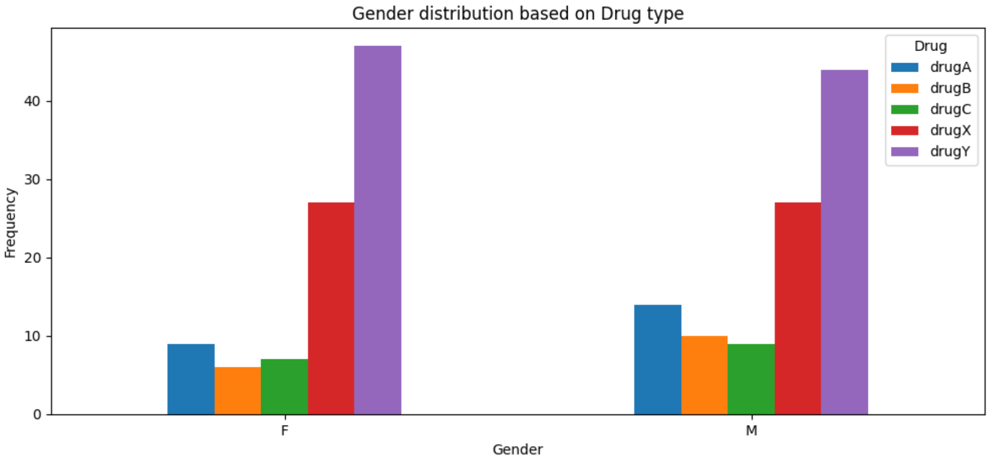
- Continuously refine models and deployment strategies to uphold ethical standards and minimize risks.

6 Results and Discussion

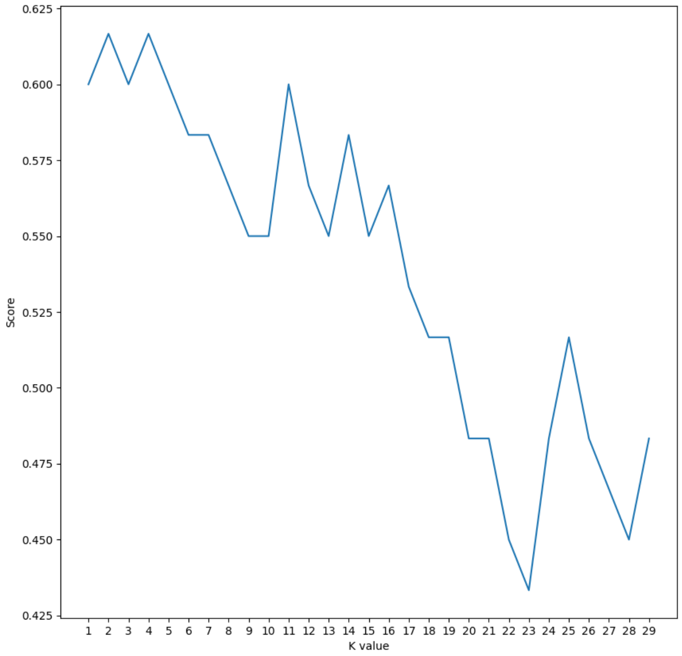
Findings:

1. **Model Performance:** Random Forest outperformed other models with an accuracy of 80%.

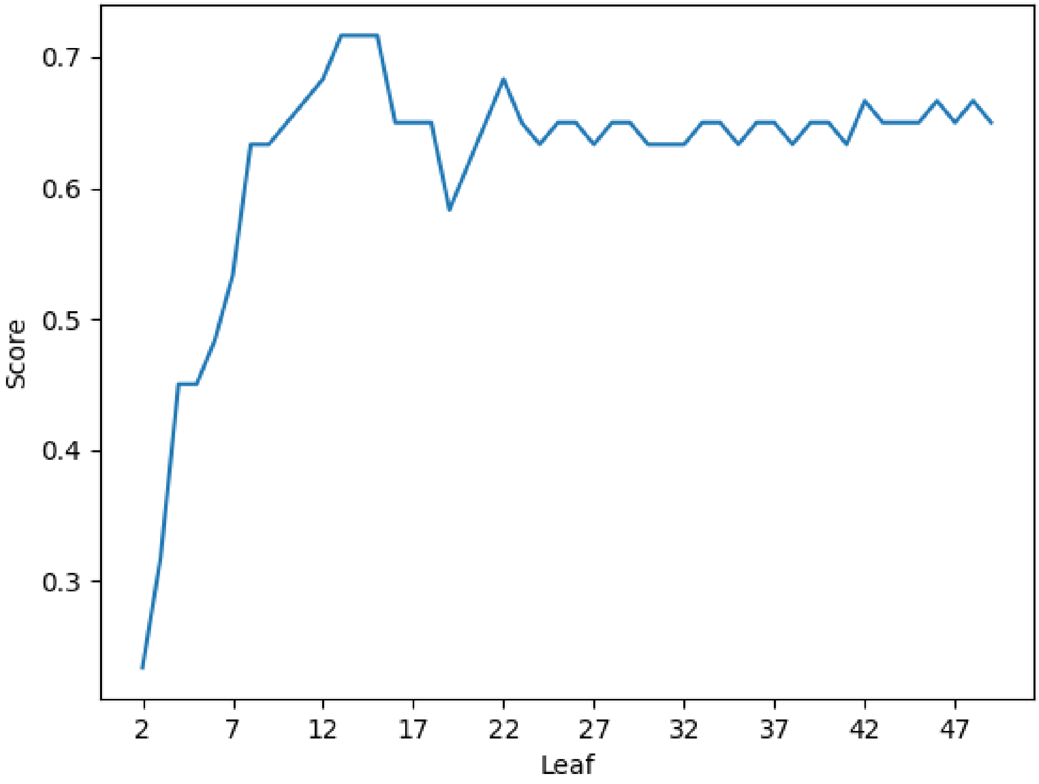
2. **Gender Distribution:** Drug usage distribution was nearly identical between males and females, indicating gender might not significantly influence drug usage in the dataset.



3. **KNN Performance:** KNN's performance decreased with an increase in the number of neighbors, implying reduced generalization possibly due to overfitting or increased noise.

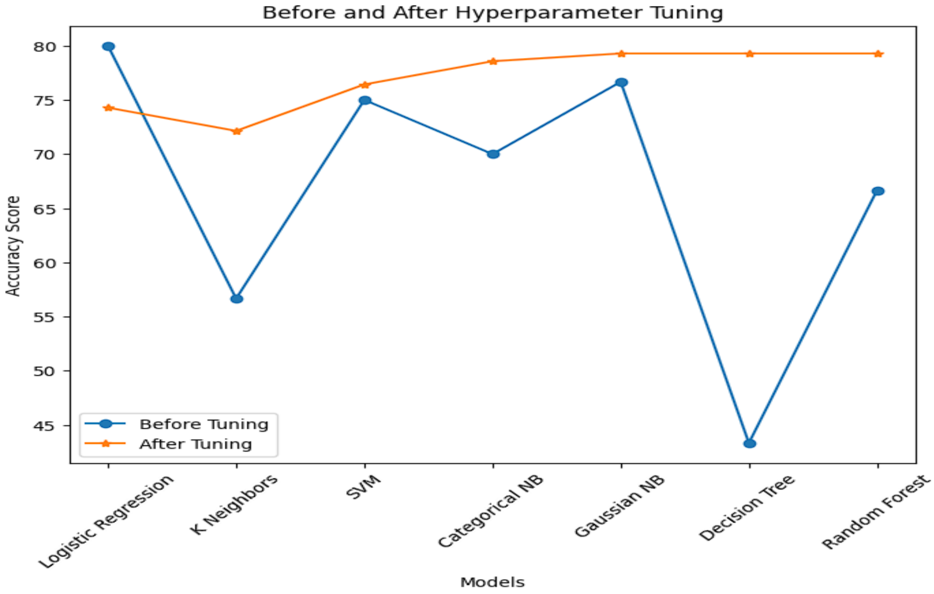


4. **Decision Tree Performance:** The Decision Tree's score increased with more leaf nodes, indicating improved performance. However, this improvement plateaued after around 50 leaf nodes, suggesting diminishing returns with increased complexity.



Comparative Analysis:

Scores after Hyperparameter tuning and before tuning.



Challenges and Limitations:

1. **Insufficient Data:** Limited data points hindered the model's ability to generalize effectively.

2. **Limited Features:** The dataset lacked essential feature columns, potentially limiting the model's predictive power.

3. **Ideal Dataset:** The overly ideal dataset may not reflect real-world complexities, affecting the model's performance in practical settings.

4. **Solution Specificity:** The solution's applicability may be restricted to the specific dataset used, limiting its adaptability to other datasets or contexts.

7 Conclusion and Future Work

The project focuses on developing a drug classification model tailored to chemical engineering, catering to the need for personalized medicine. Leveraging patient data such as age, sex, blood pressure, cholesterol levels, and Na to Potassium ratio, the model accurately classifies drugs, optimizing therapeutic outcomes. Open-source tools like NumPy, Pandas, Matplotlib, Seaborn, and Scikit-learn are utilized for robust model development and deployment.

Emphasizing seamless integration with existing systems, user-friendly interfaces, and regular maintenance ensures continued effectiveness. Scalability and performance optimization strategies are implemented to handle larger datasets and enhance model efficiency. Real-world applications encompass pharmaceutical manufacturing, drug development, and personalized medicine, promising benefits like optimized formulations, expedited drug discovery, and cost reduction.

The project's innovation lies in bridging chemical engineering with personalized medicine, offering tailored drug solutions that advance healthcare delivery and pharmaceutical research.

Future Directions:

1. **Enhanced Personalization:** Further research could focus on refining the model to incorporate more diverse patient data for even more personalized drug recommendations.

2. **Incorporation of Genetic Data:** Integrating genetic information into the model could enhance its predictive power and enable even more precise drug recommendations.

3. **Clinical Trials Optimization:** Explore using the model to optimize clinical trial design and patient selection, potentially reducing trial duration and costs.

4. **Adoption in Healthcare Systems:** Work on integrating the model into healthcare systems to assist clinicians in prescribing the most effective drugs for individual patients, improving patient outcomes.

5. **Exploration of Novel Drug Classes:** Investigate the application of the model to classify and predict the effectiveness of novel drug classes, aiding in drug discovery and development processes.

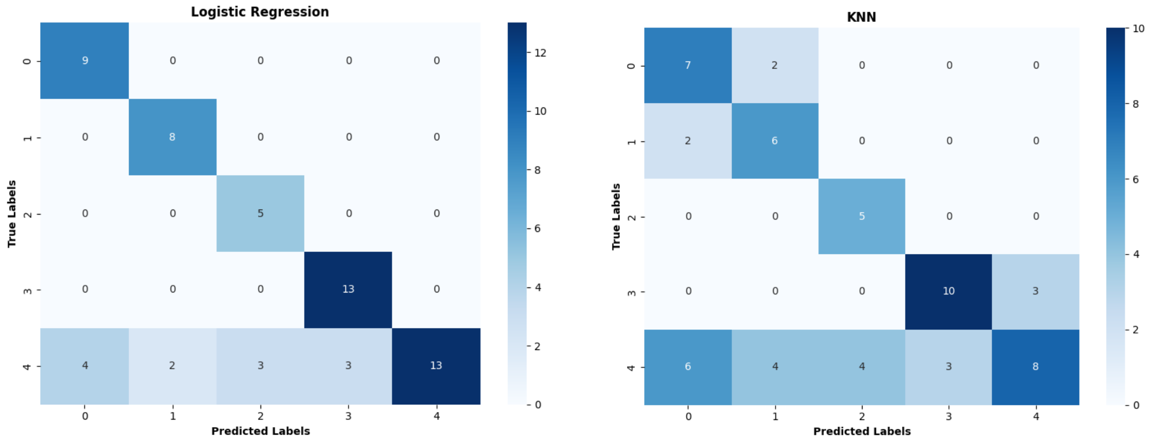
8 References

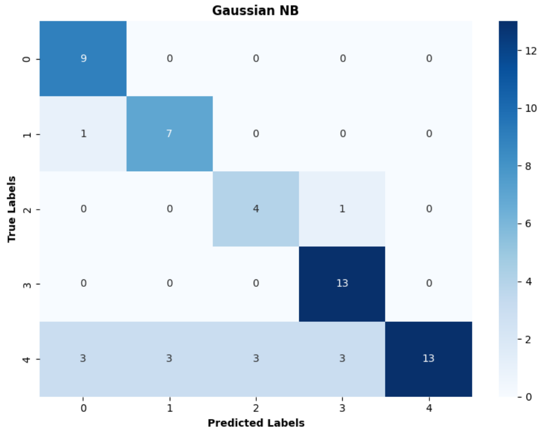
- Dhrumil Vinil Gala, Vaibhav Bharat Gandhi, Vedant Amit Gandhi and Vinaya Sawant, "Drug Classification Using Machine Learning and Interpretability", IEEE Smart Technologies, Communication and Robotics (STCR), 2021. DOI: 10.1109/STCR51658.2021.9588972

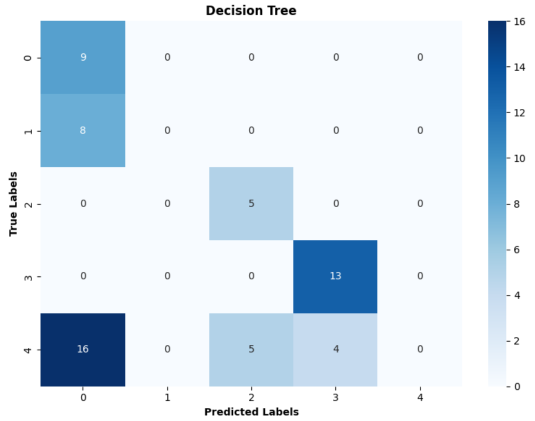
- Changlin Chen , "Research on Drug Classification Using Machine Learning Model", ResearchGate, January 2024, DOI: 10.54097/nfpj0845, License: CC BY-NC 4.0.
- Krishna Mridha, Suborno Deb Bappon, Shahriar Mahmud Sabuj, Tasnim Sarker, " Explainable Machine Learning for Drug Classification", ResearchGate, Feb 2024, DOI: 10.1007/978-981-99-8661-3_48

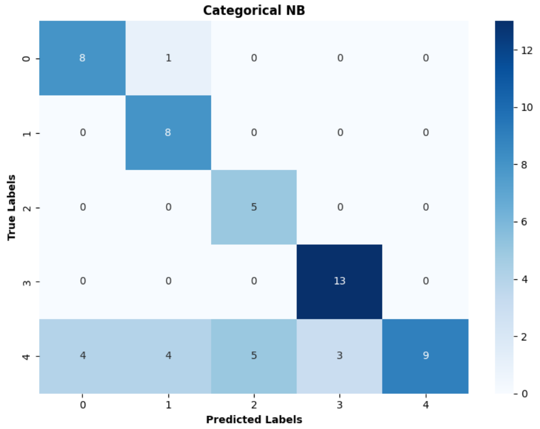
9 Appendices

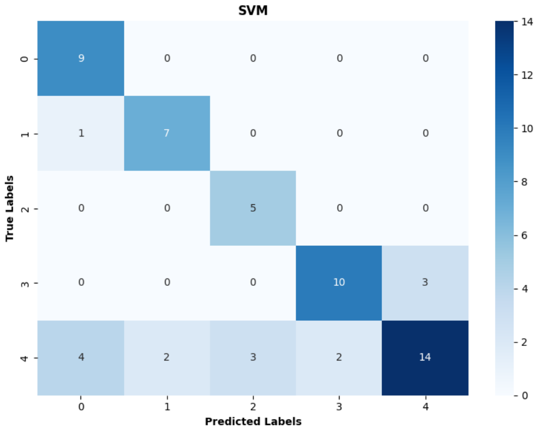
Confusion Matrices of different models

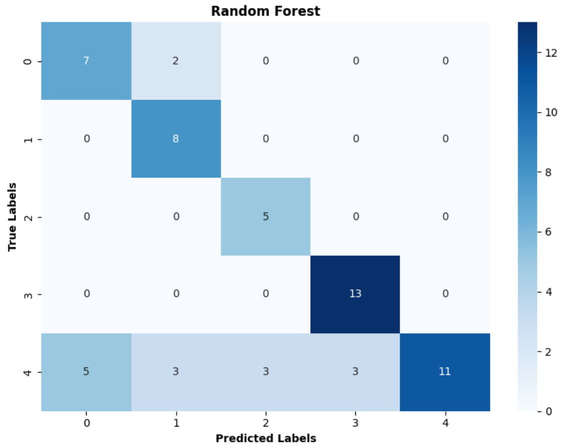




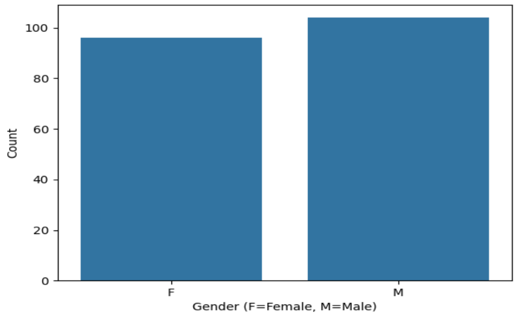


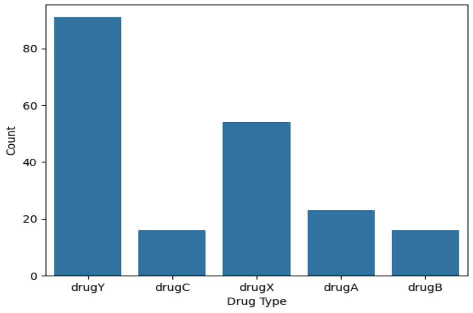


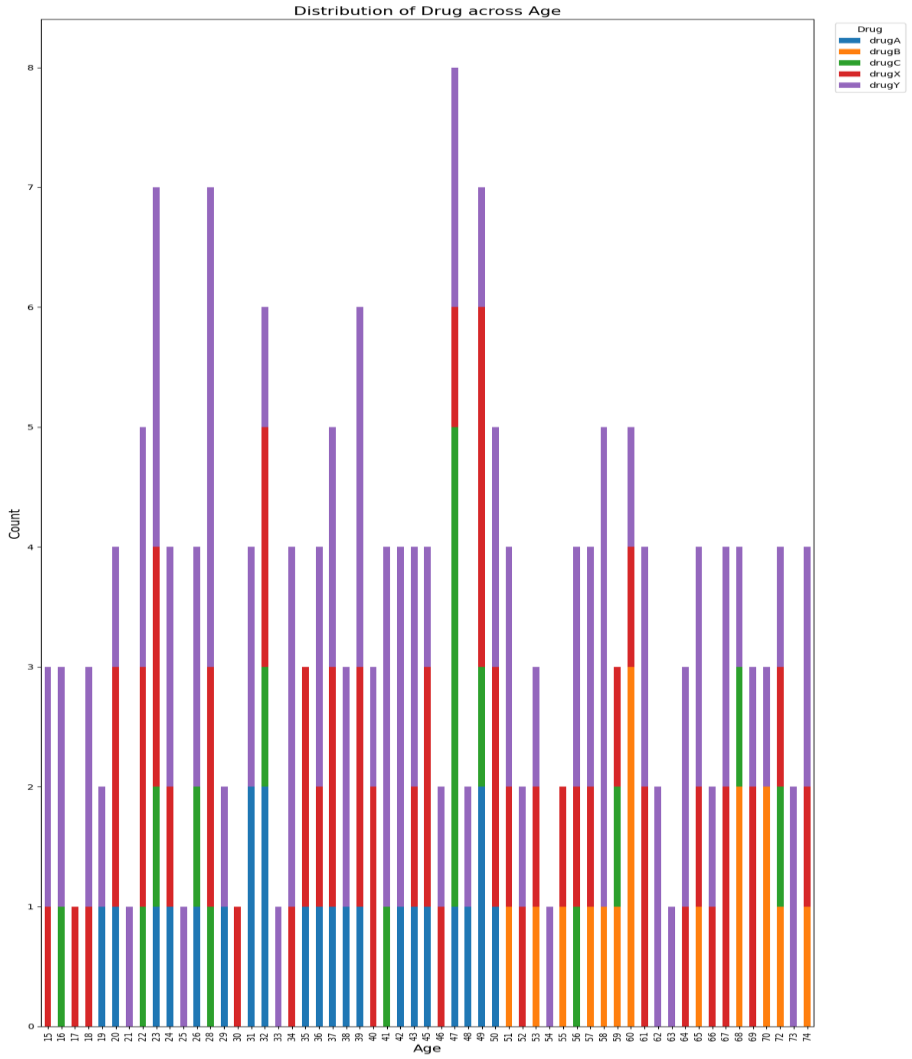


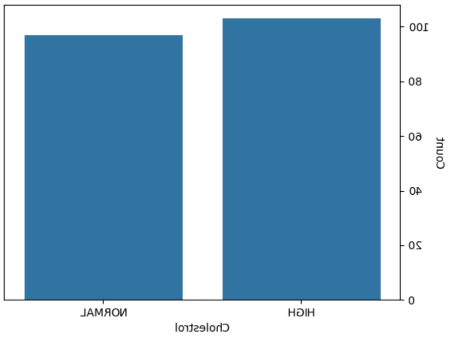


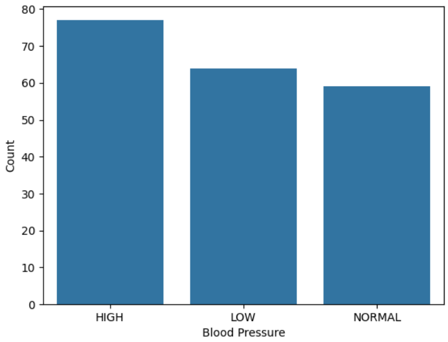
Different Distributions

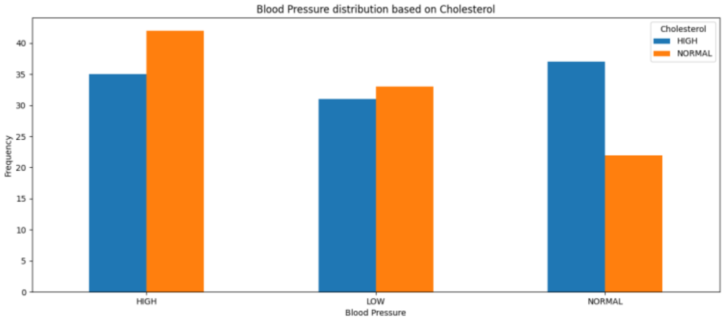





Distribution of Drug across Age
Count
0
1
2
3
4
5
6
7
8
Age
15 16 17 18 19 20 21 22 23 24 25 26 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 72 73 74
Drug
drugA
drugB
drugC
drugX
drugY


Count
0
20
40
60
80
100
NORMAL
HIGH
Cholesterol


Count
0
10
20
30
40
50
60
70
80
HIGH
LOW
NORMAL
Blood Pressure


Blood Pressure distribution based on Cholesterol
Frequency
0
5
10
15
20
25
30
35
40
HIGH
LOW
NORMAL
Blood Pressure
Cholesterol
HIGH
NORMAL

10 Auxiliaries

Data Source: Dataset

Python file: file link