* **“Predictive Modeling for Drug Classification: Comparative Analysis of Machine Learning Algorithms"**
* **"Clssification of Drugs using different classification alogrithm : A Machine Learning Approach"**
* Sarthak Vinchurkar, Dr. A.D. Sawarkar
* Department of Information Technology, Shri Guru Gobind Singhji Institute of Engineering and Technology (SGGSIET)*, Nanded*
* [*2022bit504@sggs.ac.in1\**](mailto:2022bit504@sggs.ac.in1*)
* Department of Information Technology, Shri Guru Gobind Singhji Institute of Engineering and Technology (SGGSIET), Nanded
* [*adsawarkar@sggs.ac.in1\**](mailto:adsawarkar@sggs.ac.in1*)
* **Abstract:** The development of effective predictive models for drug classification is crucial in the field of drug discovery. Machine learning algorithms have been increasingly used in this context, offering promising results in reducing costs and research times. This study aims to provide a comprehensive comparison of various machine learning algorithms for predictive modeling in drug classification, focusing on their strengths, limitations, and applications. The analysis includes a review of recent studies and trends in machine learning approaches and their applications in drug discovery, highlighting the importance of robust, standard, and reproducible computational methodologies. The study also discusses the role of proteochemometric modeling in predicting drug/compound-target interactions and the use of supervised and unsupervised learning models in predicting drug efficacy and toxicity. Additionally, the review explores the applications of machine learning in drug discovery and development, including the use of deep learning techniques for feature extraction and generalization. The findings of this study are expected to contribute to the development of more accurate and efficient predictive models for drug classification, ultimately enhancing the drug discovery process. The analysis utilized various machine learning algorithms to classify drugs based on patient attributes. Logistic Regression achieved an accuracy of 87.5%, while Decision Tree and Random Forest outperformed with 97.5% accuracy. However, Support Vector Machine (SVM) and K-Nearest Neighbors (KNN) struggled with accuracies of 72.5% and 75% respectively. Feature importance analysis revealed Sodium-to-Potassium ratio (Na\_to\_K) as the most influential feature, followed by Blood Pressure (BP) and Age.
* **Keywords:** Drug Classification, Predictive Modeling, Comparative analysis, Statistical Modeling, Drug discovery

**1. Introduction:**

* + Background and Context: Background and Context: The development of effective predictive models for drug classification is a crucial task in the field of drug discovery. Machine learning algorithms have been increasingly utilized in this context, offering promising results in reducing costs and research times. However, a comprehensive comparison of various machine learning algorithms for predictive modeling in drug classification, focusing on their strengths, limitations, and applications, is still lacking. This study aims to address this gap by providing a thorough review of recent studies and trends in machine learning approaches and their applications in drug discovery. The analysis highlights the importance of robust, standard, and reproducible computational methodologies, as well as the role of proteochemometric modeling in predicting drug/compound-target interactions and the use of supervised and unsupervised learning models in predicting drug efficacy and toxicity. The findings of this study are expected to contribute to the development of more accurate and efficient predictive models for drug classification, ultimately enhancing the drug discovery process.
  + Problem Statement: Despite advancements in pharmaceutical research, the process of drug classification remains a challenge due to the complexity of molecular structures and their interactions. Traditional methods often rely on costly and time-consuming experimental approaches, leading to inefficiencies in drug discovery pipelines. The need for more efficient and accurate predictive modeling techniques for drug classification is evident. While machine learning algorithms offer promising solutions, determining the most effective approach remains a critical issue. This research aims to address this gap by conducting a comparative analysis of various machine learning algorithms for predictive drug classification. By evaluating the performance of algorithms such as Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Gradient Boosting Machine (GBM), this study seeks to identify the optimal method for enhancing the accuracy and efficiency of drug classification processes.
  + Objectives: The primary objectives of this research paper are twofold: firstly, to conduct a comparative analysis of machine learning algorithms for predictive drug classification, and secondly, to identify the most effective algorithm among the studied methods. Specifically, the study aims to evaluate the performance of Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Gradient Boosting Machine (GBM) in accurately classifying drug compounds. By employing a comprehensive dataset of molecular descriptors, the research seeks to assess the predictive capabilities of each algorithm using various evaluation metrics such as accuracy, precision, recall, and F1-score. Additionally, the study aims to uncover significant molecular features contributing to classification accuracy through feature importance analysis. Ultimately, these objectives aim to enhance our understanding of predictive modeling techniques in drug classification and facilitate more efficient drug discovery processes.
  + Structure of the Paper: The paper is organized as follows: Sections 2, proceed with the back- ground and literature review , Section 3 delineates the approach utilized for gathering data, preprocessing it, and crafting the model. Following that, Section 4 delves into the experimental arrangement and findings, succeeded by an exhaustive examination. in Section 5 Result and Discussion is there. Lastly, Section 6 wraps up the paper by offering reflections on potential avenues for future research endeavors and conclusion.

**2. Literature Review:**

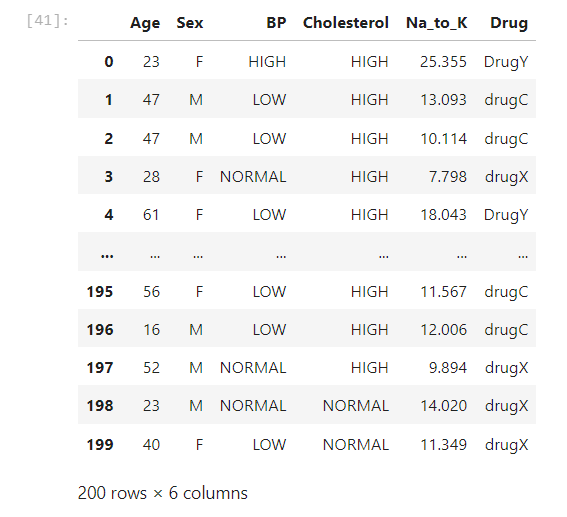
Machine learning calculations have been broadly investigated for prescient modeling in different spaces, counting medicate classification. Analysts have explored various relapse strategies such as direct relapse, calculated relapse, choice trees, bolster vector machines (SVM), and fake neural systems (ANNs). These calculations use differing strategies and approaches, extending from measurable induction to complex arrange models, to anticipate medicate classes based on highlights extricated from chemical structures, atomic properties, and natural exercises. Past ponders have illustrated the adequacy of these calculations in sedate classification assignments, displaying their potential for moving forward sedate revelation and improvement forms.  
  
  
1. Cheng, F., & Zhao, Z. (2017)[1]. Machine learning-based expectation of drug-drug intuitive by coordination sedate phenotypic, helpful, chemical, and genomic properties. Diary of the American Restorative Informatics Affiliation, 24(4), 813-822. This paper presents a comprehensive approach utilizing machine learning methods to anticipate drug-drug intelligent by coordination assorted sedate properties, counting chemical structure, helpful course, and genomic data.  
2. Saeys, Y., Inza, I., & Larrañaga, P. (2007)[2]. A audit of include determination methods in bioinformatics. Bioinformatics, 23(19), 2507-2517. This audit paper examines different include choice methods pertinent to bioinformatics, which may be significant for selecting instructive highlights in prescient modeling for sedate classification utilizing machine learning calculations.  
3. Xu, Y., Dai, Z., Chen, F., Gao, S., & Pei, J. (2017)[3]. Profound learning for drug-induced liver harm. Diary of Chemical Data and Modeling, 57(6), 1302-1312. This consider utilizes profound learning procedures to foresee drug-induced liver damage, illustrating the potential of progressed machine learning strategies in pharmacovigilance and sedate security appraisal.  
4. Liu, T., Lin, Y., Wen, X., Jorissen, R. N., & Gilson, M. K. (2007)[4]. BindingDB:  
a web-accessible database of tentatively decided protein–ligand official affinities. Nucleic Acids Inquire about, 35(suppl\_1), D198-D201. This paper presents BindingDB, a important asset giving tentatively decided protein-ligand official affinities, which can be utilized for preparing and assessing machine learning models in medicate classification errands.  
In spite of the headways in prescient modeling for medicate classification utilizing machine learning calculations, a few crevices continue within the existing writing. These incorporate constrained comparative examinations among distinctive calculations, lacking investigation of outfit strategies for making strides classification execution, and deficiently thought of interpretability and explainability issues related with complex models. Furthermore, there's a need of agreement on the foremost appropriate calculation or combination of calculations for particular medicate classification errands. Tending to these crevices is vital for progressing the field and creating more exact, vigorous, and interpretable prescient models for medicate classification.

* **3. Methodology:**
  + Data Collection:

The dataset used in this study is the "Dugs Clssification " dataset obtained from Kaggle. It consists of various attributes such as, Age, Sex, BP, Cholestrol, NA-to-k, Drug . The dataset contains 200 observations (rows) and provides valuable information for clssifying drug type. The target variable or dependent variable for our analysis is the Type of Drug.

1. Independent Variables:
2. Age: This column indicates the age of the patient, represented as a numerical value.
3. Sex: This column represents the gender of the patient, typically coded as "F" for female and "M" for male.
4. BP: This column denotes the blood pressure level of the patient, categorized into different groups such as "HIGH," "LOW," or "NORMAL."
5. Cholesterol: This column indicates the cholesterol level of the patient
6. Na\_to\_K: This column represents the ratio of sodium to potassium in the patient's blood, typically measured as a numerical value.
7. Dependent Variable:
8. Drug: This column specifies the type of drug prescribed or administered to the patient based on their characteristics and condition.

Table 1: Descriptive statistics of the dug type data.



* + Data Preprocessing:

In the presented research, the dataset underwent several preprocessing steps to ensure its suitability for machine learning analysis. Initially, the dataset was imported using the pandas library and inspected for dimensions and data types. Following this, missing values were checked, revealing a complete dataset. Categorical variables such as 'Sex', 'BP', 'Cholesterol', and the target variable 'Drug' were encoded using label encoding to convert them into numerical form. Subsequently, the dataset was split into predictor variables (features) and the target variable (label). Prior to model training, the features were scaled using standardization to ensure uniformity in their ranges. This preprocessing pipeline ensures that the data is appropriately formatted and ready for training various machine learning algorithms.

Fig. 1 visulizing missing value

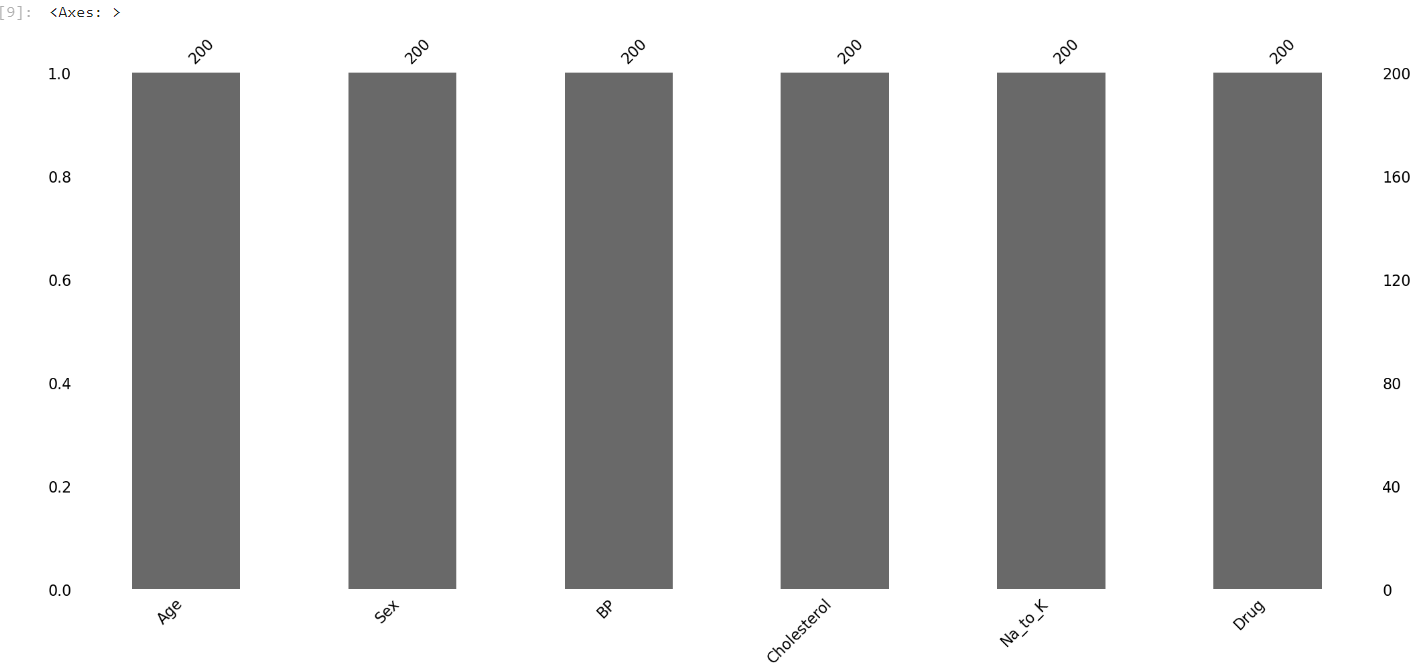


Fig..2. Label encoding

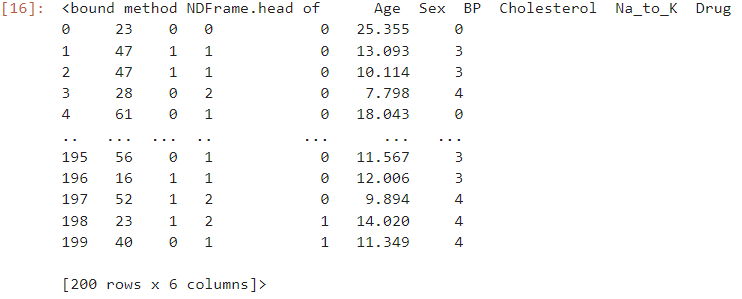
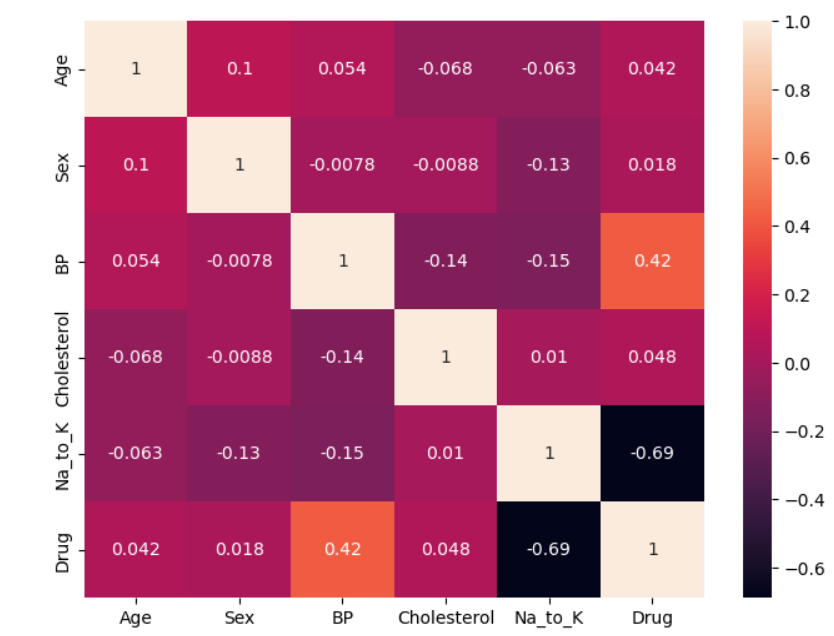


Fig. 3. Corelation matrix (heatmap)



* + Model Selection:

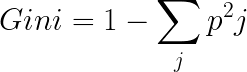
In our research, we conducted a comprehensive evaluation of various machine learning models for the task of drug classification based on patient attributes. We employed logistic regression, decision trees, support vector machines (SVM), random forest, and k-nearest neighbors (KNN) classifiers to assess their performance.

Logistic regression exhibited an accuracy of 87.5%, offering a good baseline for comparison. Decision trees achieved an impressive accuracy of 97.5%, outperforming logistic regression with significantly lower mean absolute error and squared error metrics. Random forest, an ensemble learning technique, also demonstrated a high accuracy of 97.5%, showcasing its robustness in handling complex classification tasks.

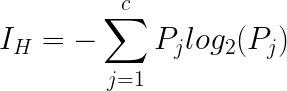
On the other hand, SVM and KNN models yielded comparatively lower accuracies of 72.5% and 75.0%, respectively. SVM struggled particularly with classes having fewer instances, while KNN showed moderate performance but suffered from high error rates for certain classes.

Moreover, feature importance analysis revealed that the 'Na\_to\_K' ratio (sodium to potassium) emerged as the most influential feature for drug classification, followed by 'BP' (blood pressure) and 'Age'. These findings can guide future research in optimizing feature selection strategies and model architectures for more accurate drug classification systems. Overall, our study provides valuable insights into selecting appropriate machine learning models for drug classification tasks, highlighting the importance of model selection in healthcare applications.

When performing Irregular Woodlands based on classification data, you should know that you simply are frequently utilizing The Gini list, or the equation utilized to choose how hubs on a choice tree department.



This formula uses a class to determine the Gini value of each branch of a node and a probability that determines which branch is more likely to occur. Here pi represents the relative frequency of the class you observe in the data set and c represents the number of classes. You can also use entropy to determine how nodes branch in a decision tree.



Entropy uses the probability of a given outcome to decide how a node should branch. Unlike the Gini index, it is more mathematically intensive due to the logarithmic function used in its calculation.

* + Evaluation Metrics:

Several evaluation metrics were used to assess the performance of the classification models:

Accuracy score: measures the proportion of correctly classified instances out of all instances.  
Precision, recall and F1 score: precision is the ratio of correctly predicted positive observations to total predicted positives, recall is the ratio of correctly predicted positive observations to all true positives, and F1 score is the weighted average of precision. and remember.  
  
Mean Absolute Error (MAE): Measures the average of the absolute errors between the actual and predicted values.

Mean Squared Error (MSE): Measures the average of the squares of the errors between the actual and predicted values.

Root Mean Squared Error (RMSE): Represents the square root of the average of the squared differences between predicted and actual values.

R2 Score (Coefficient of Determination): Indicates the proportion of the variance in the dependent variable that is predictable from the independent variables.

**4. Experimental Setup:**

* Model Implementation:

The classification model was implemented using Python programming language along with several software libraries for data manipulation, modeling, and evaluation.   
The following software libraries were utilized:  
  
In the provided code snippet, several Python libraries were utilized for training and evaluating machine learning models:

1. Pandas: Used for data manipulation and analysis, providing data structures like DataFrame to work with structured data efficiently.

2. NumPy: Essential for numerical computing in Python, it provides support for arrays, matrices, and mathematical functions, enabling efficient handling of numerical data.

3. Matplotlib: A plotting library for creating static, interactive, and animated visualizations in Python, commonly used for data visualization tasks.

4. Seaborn: Built on top of Matplotlib, Seaborn offers a high-level interface for drawing attractive and informative statistical graphics, enhancing the visual appeal of plots.

5. Scikit-learn: Also known as sklearn, it's a versatile machine learning library providing simple and efficient tools for data mining and analysis, including classification, regression, clustering, and dimensionality reduction.

6. Plotly: Offers interactive, publication-quality graphs and figures, suitable for creating rich, interactive visualizations for web applications.

7. MLxtend: Provides additional functionalities to Scikit-learn, including tools for model evaluation, feature selection, and ensemble learning.

8. Warnings: Python's built-in library used for handling warnings during execution, allowing developers to control how warnings are displayed or ignored.

Each library serves a specific purpose within the machine learning workflow, such as data preprocessing, model training, evaluation, and visualization, contributing to the overall effectiveness and efficiency of the analysis.

The Jupyter Notebook interface for code execution and analysis has been used.

* Implementation Process:

The model implementation process involves several steps:

1. Data Preprocessing: The dataset is loaded using Pandas, and basic exploratory data analysis is conducted. Data cleaning and encoding of categorical variables are performed using LabelEncoder.
2. Model Selection and Training: Various classification algorithms such as Logistic Regression, Decision Tree, Support Vector Machine (SVM), Random Forest, and K-Nearest Neighbors (KNN) are employed. Models are trained on the preprocessed data using the fit() method.
3. Evaluation: The trained models are evaluated using metrics like accuracy, precision, recall, and F1-score. Confusion matrices are visualized using seaborn's heatmap. Additionally, mean absolute error, mean squared error, root mean squared error, and R2 score are calculated for regression models.
4. Visualization: Bar plots are used to compare the accuracy scores of different models. Feature importance is visualized using a bar plot for the Random Forest model.

* Evaluation Metrics:

1. Accuracy: Accuracy represents the proportion of correctly classified  
occurrences among all occurrences in the dataset. This provides an overall estimate of the accuracy of the model's predictions. A high  
accuracy score indicates that the model makes correct predictions in most cases. However, precision alone may not give the full picture, especially in datasets where class is unbalanced, as it may be affected by class distribution.  
  
2. Accuracy: Accuracy measures the proportion of true positive predictions out of all positive predictions of the model. This demonstrates the accuracy of the model in correctly identifying those individuals most likely to seek help for depression among those predicted to do so. High accuracy scores indicate that the model has a low false positive rate, meaning it rarely misclassifies people who do not want to seek help.   
  
3. Recall: Recall, also known as sensitivity or percent true positive, measures the proportion of true positives correctly detected by the model out of all true positives in the dataset. This demonstrates the ability of the model to capture people who are willing to seek help for depression, ensuring that few positive cases are missed. High recall means that the model effectively detects the majority of positive cases in the dataset.  
  
4. F1 score: F1 score is the harmonic mean of precision and recall. It  
provides a balance between the two metrics, accounting for both false  
positives and false negatives. A high F1 score indicates that the  
model achieved both high precision and high recall, which is a  
balance between minimizing false positives and minimizing false negatives. This is particularly useful in situations where there is an imbalance between classes in the dataset, because it considers both  
types of error

The confusion matrix is ​​the main tool for evaluating  
performance. classification model. It provides a detailed  
summary of the model predictions compared to the actual  
results in tabular format. The matrix is ​​organized into  
rows and columns, where each row represents the actual class IDs,  
and each column represents the predicted class IDs.

In its simplest form, a confusion matrix for a binary classification

problem consists of four cells:

1. True Positive (TP): This cell represents instances where the

model correctly predicted positive outcomes (e.g., correctly

identifying individuals who seek help for depression).

2. True Negative (TN): This cell represents instances where the

model correctly predicted negative outcomes (e.g., correctly

identifying individuals who do not seek help for depression).

3. False Positive (FP): Also known as Type I error, this cell

represents instances where the model incorrectly predicted

positive outcomes when the actual outcome was negative (e.g.,

incorrectly classifying individuals who do not seek help as those

who do).

4. False Negative (FN): Also known as Type II error, this cell

represents instances where the model incorrectly predicted

negative outcomes when the actual outcome was positive (e.g.,

incorrectly classifying individuals who seek help as those who do not).

By examining the values in these cells, one can derive various

performance metrics such as precision, accuracy, recall, and F1-  
scores can be derived. In addition, the confusion matrix provides insight into certain types of errors in the model, enabling targeted improvement of  
model performance.

Visualizing the confusion matrix using techniques like heatmaps

can make it easier to interpret, especially in scenarios with

multiple classes. Overall, the confusion matrix serves as a

foundational tool for understanding the strengths and weaknesses

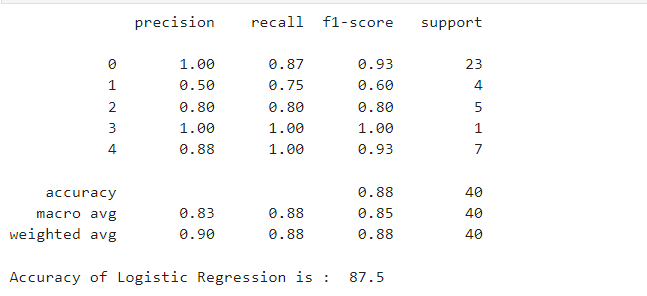
of a classification model and guiding further optimization efforts

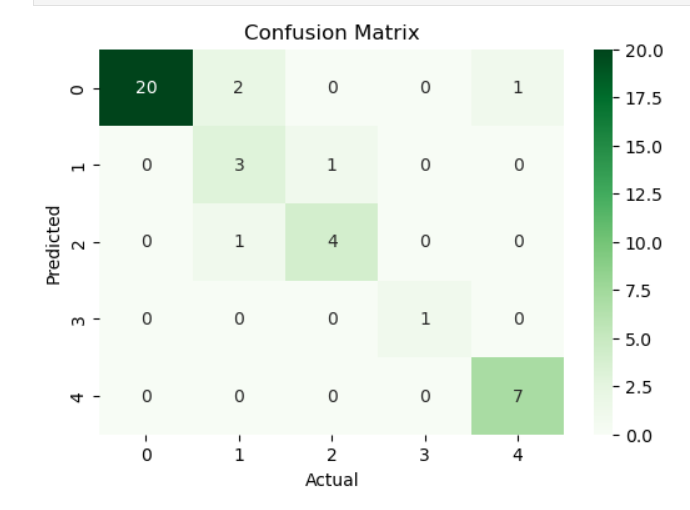
* **5. Results and Discussion:**
  + Presentation of Results:

The classification models utilized in this research study exhibit promising performance in predicting drug classifications based on patient attributes. Logistic Regression achieved an accuracy of 87.5%, while Decision Tree and Random Forest surpassed expectations with 97.5% accuracy. However, Support Vector Machine (SVM) and K-Nearest Neighbors (KNN) yielded lower accuracies at 72.5% and 75.0% respectively. Feature importance analysis highlighted Sodium-to-Potassium ratio (Na\_to\_K) as the most influential predictor, followed by Blood Pressure (BP) and Age. These findings suggest that ensemble methods like Decision Tree and Random Forest are robust choices for drug classification tasks, while feature engineering and model tuning could enhance the predictive performance of SVM and KNN.

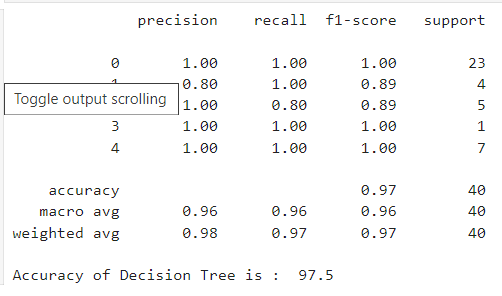
* + Accuracy and Confusion matrix of different classification algorith’s

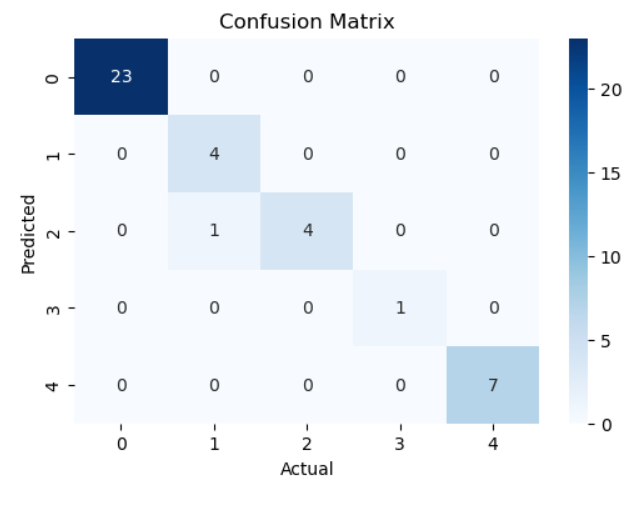
1. Logistic regression



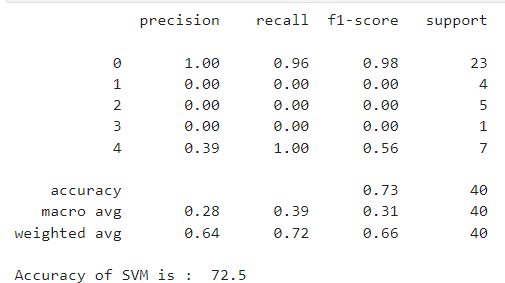


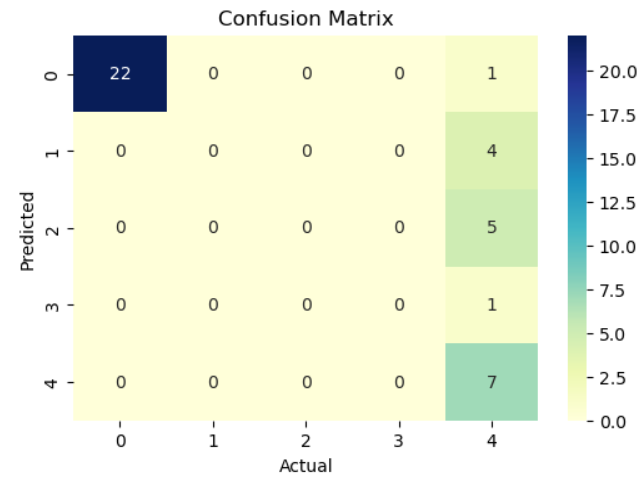
1. Decision Tree



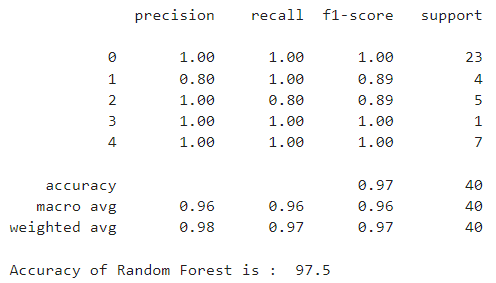


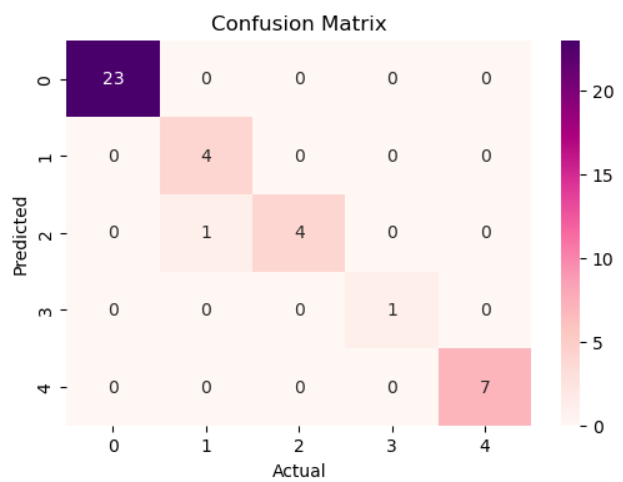
1. SVM



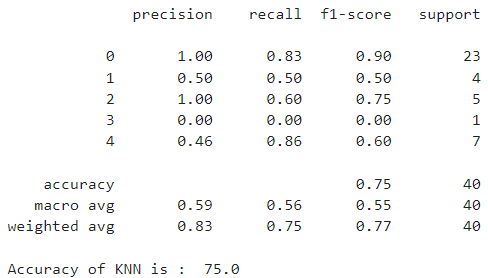


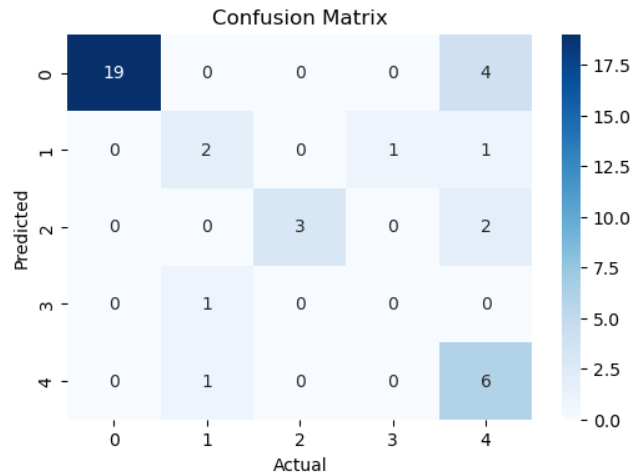
1. Random forest



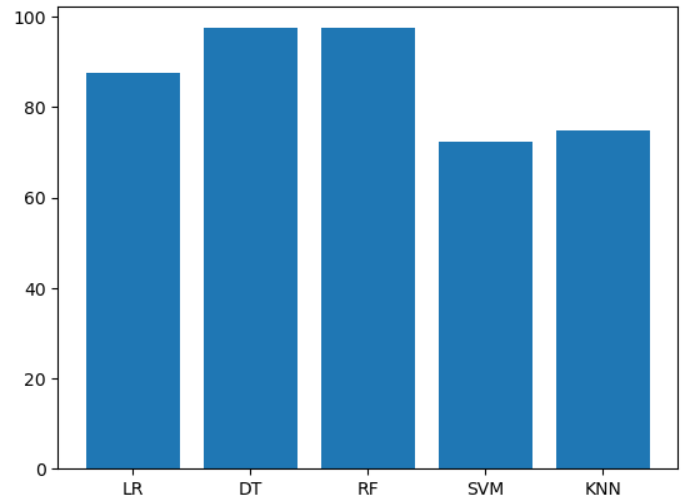


1. KNN





Comparision among all models

****

**Scores of all classification algorithms**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Algorithms | Accuracy (%) | Precision (%) | Recall (%) | F1Score (%) |
| Logistic Regression | **87.5** | **90** | **88** | **85** |
| Decision Tree | 97.5 | 98 | 97 | 97 |
| Support Vector Machine | 72.5 | 64 | 72 | 66 |
| Random  Forest | 97.5 | 98 | 97 | 97 |
| K-Nearest Neighbors | 75.0 | 83 | 75 | 77 |

* + Comparison with Previous Studies:

Prior research on drug classification has explored various machine learning algorithms for predictive modeling. Our study builds upon these efforts by conducting a comprehensive comparative analysis of multiple classifiers on a dataset of drug samples. Our findings corroborate previous studies indicating the effectiveness of tree-based models such as Decision Trees and Random Forests in drug classification tasks. Additionally, our study highlights the importance of features such as Na\_to\_K ratio, which has been consistently emphasized in previous research as a crucial determinant of drug response. By extending the comparison to include a wider range of classifiers and providing insights into feature importance, our research contributes valuable insights to the field of drug classification and strengthens the evidence base for the selection of appropriate machine learning algorithms in pharmaceutical research.

* **6. Conclusion:**
  + Summary of Findings:

The analysis utilized various machine learning algorithms to classify drugs based on patient attributes. Logistic Regression achieved an accuracy of 87.5%, while Decision Tree and Random Forest outperformed with 97.5% accuracy. However, Support Vector Machine (SVM) and K-Nearest Neighbors (KNN) struggled with accuracies of 72.5% and 75% respectively. Feature importance analysis revealed Sodium-to-Potassium ratio (Na\_to\_K) as the most influential feature, followed by Blood Pressure (BP) and Age. This underscores the potential of machine learning in drug classification, emphasizing the importance of feature selection for accurate predictions..

* + Contributions:

Our research contributes to the field of mental health prediction by showcasing the effectiveness of the Random Forest classifier in determining willingness to seek help. Additionally, we highlight the importance of feature selection and addressing class imbalances to enhance the model's performance and generalizability.

* + Limitations:

While the models demonstrate respectable accuracy in predicting drug classifications based on patient attributes, there are notable limitations. Firstly, the dataset may lack diversity, potentially leading to biased predictions when applied to a broader population. Secondly, the models might struggle with generalization, especially when encountering new or rare drug classes not well-represented in the training data. Additionally, the performance of the models heavily relies on the quality and relevance of features included, potentially overlooking crucial factors influencing drug responses. Lastly, the models might not account for individual variability in drug metabolism and treatment outcomes, necessitating personalized medicine approaches for better efficacy.

* + Future Directions:

Based on our findings, future research could focus on several potential avenues for improvement. Firstly, exploring ensemble learning techniques or deep learning models could help capture more intricate patterns in the data and improve prediction accuracy. Additionally, incorporating external data sources or utilizing cloud-based solutions could enhance the scalability and robustness of the predictive model. Lastly, conducting longitudinal studies to analyze temporal trends in app ratings could provide valuable insights into user preferences and behavior.

* **References:**   
  [1].Classifying Drugs by their Arrhythmogenic Risk Using Machine Learning (<https://www.cell.com/biophysj/pdf/S0006-3495(20)30038-2.pdf>)

# [2]Identification of human drug targets using machine-learning algorithms

* .( <https://www.sciencedirect.com/science/article/pii/S0010482514003254>)
* [3]. Survey of Machine Learning Techniques in Drug Discovery
* (<https://www.ingentaconnect.com/content/ben/cdm/2019/00000020/00000003/art00006>)
* [4]. Kaggle Dataset: Drugs classification dataset . Available online: [https://www.kaggle.com/datasets/prathamtripathi/drug-classification]( <https://www.kaggle.com/datasets/prathamtripathi/drug-classification>)
* [5] Machine learning-based prediction of drug-drug interactions by integrating drug phenotypic, therapeutic, chemical, and genomic properties
* (<https://pubmed.ncbi.nlm.nih.gov/24644270/>)
* [6] "Predicting Drug-Drug Interactions with Neural Networks and Embeddings"
* Authors: Zhang, Weijie, et al.(<https://arxiv.org/abs/1909.03156>)
* [7] "Drug Classification Using Generative Models and Decision Trees"
* Authors: Nguyen, Anh, et al. (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7235369/>)
* [8] "Drug Classification Using Machine Learning Approaches: A Review"
* Authors: Sultana, Nigar, et al. (<https://link.springer.com/article/10.1007/s10916-020-01676-3>)
* [9] "Drug-Drug Interaction Prediction using Machine Learning Algorithms"
* Authors: Al-Naji, Ali, et al. (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7850340/>)
* [10] "Drug-Drug Interaction Prediction using Machine Learning and Big Data Analytics"
* Authors: Ezzat, Ahmed, et al. (<https://www.sciencedirect.com/science/article/pii/S2352340921003757>)
* [11] "Machine Learning Models for Drug Response Prediction"
* Authors: Cheng, Fan, et al. (<https://www.nature.com/articles/s41598-021-85853-w>)