DRUG CLASSIFICATION USING MACHINE LEARNING TECHNIQUES

Siddhesh Desai
Student ID: 13008482
MSc Data Science and Computational Intelligence, Coventry
United Kingdom
Email: desais10@uni.coventry.ac.uk

Surabhi Jarabandahalli Rammurthy
Student ID: 12801086
MSc Data Science and Computational Intelligence, Coventry
United Kingdom
Email: jarabandas@uni.coventry.ac.uk

***Abstract*—The evolution of drug discovery and design technologies has been worldwide in the last decade. Many different drug-drug interaction models were developed with hypothetical training and testing. The reactions of Adverse drugs might increase the amount of mortality. Hence It is necessary to identify the effect of drugs on unique targets. In this paper appropriate machine learning algorithms: Decision tree, K Nearest Neighbor, and Logistic Regression are applied to the dataset of patients who suffered from the same illness, where during the course of treatment each patient responded to a different set of drugs. The aim of this paper is to develop a model which classifies the drugs that are appropriate for a future patient with the same illness based on feature extraction with good computational efficiency. Applied machine learning techniques will be compared over the performance by expected classification accuracy on the testing dataset. The results of the experiment are based on the dataset obtained from different features of the patients who suffered from the same illness.**

Keywords—Features extraction, Decision tree, Logistic Regression, K Nearest neighbor, Drug Prescription, Classification, Python, Machine Learning.

I. INTRODUCTION

The forecast of the interaction of drugs on the target is a crucial part of drug classification. Each type of drug might have a distinct effect on individuals with the same disease depending on certain attributes of the person. The accurate prediction will speed up the process and provide insights into the mechanism of drug functioning. In spite of great development in the field of technology predicting the effects of drugs on the target is very expensive and long-delayed. Hence accurate prediction models are required in the field of pharmacology. Furthermore. unanticipated side effects may negatively affect a patient's health or even their life, identifying innovative potential unfavourable pharmaceutical responses is also a crucial task. The intent is also to foresee the unwanted side effects of the drug which might be caused to the individual and control them in advance. In order to address the above-mentioned problems this paper will address the classification based on the usage and feature selection.

The use of machine learning is firmly recommended when it comes to the prediction of drug-target interaction over any other methods. Machine learning directs understanding of the experiments and provides knowledge about the properties of drugs as well as the drug targets. Knowing the functioning of drugs and structures will help increase the entire processing of the system, especially in terms of decision trees. We analyse the overall features of the patient to lend decisions on the drugs.

Generally, the importance of machine learning is huge in the field of bioinformatics. It is used on a large scale in Biology, Biomedicine and Health care. Supervised learning is extensively utilized to train the data and predict future outcomes. Regression and classification are among supervised learning. Furthermore, Novel machine-learning strategies are employed in many studies to boost accuracy. In this study, we have employed Decision tree, Logistic regression and K Nearest Neighbour algorithms to build a model of maximum accuracy. The dataset chosen will fully indicate the effectiveness of techniques while stating the potential problems that might occur while using them. The further section will provide insights into the dataset.

II. LITERATURE REVIEW

A few papers have been published employing the machine learning algorithm for the classification of drugs which are demonstrated below.

In this paper (D. V. Gala, V. B. Gandhi, V. A. Gandhi and V. Sawant, 2021) [1] the author has depicted the use of machine learning algorithms to find interpretability. This model uses Random Forest, Logistic Regression and Decision tree algorithms for the prediction. Whereas LIME and SHAP methods are used to analyze the output.

LIME is used to depict the output of any Regressor or classifier in a very practical manner. It alters the feature values of a single data sample, and analyses the impact of these modifications on the result. Whereas SHAP is based on Shapley values which were initially put up as a cooperative game theory solution concept by Lloyd Shapley in 1951. This approach performs well with any machine learning algorithm. These models will depict the features which are responsible for the cause of divergence in results. The author has concluded that a Decision tree and Random Forest are the best methods for the classification of drugs as this shows an accuracy of 0.97. The decision tree and logistic regression ML methods described in this paper will be referred to for the current study. This model has proved a good accuracy and performance.

This study (M. R. Barkat, S. M. Moussa and N. L. Badr, 2021) [2] is based on 'Drug Target interaction (DTI) prediction using machine learning' where huge machine learning and artificial intelligence models are adopted in the drug evolution leading to mutation at all stages which in turn causes the drug target interaction prediction long delayed. Hence a new approach with ML algorithms and feature extraction method is introduced to increase the prediction and accuracy rate. The author has used only 7 features with 21 target features and based on the similarity between the drugs a negative dataset is produced. This will apparently reduce the noise among the data which was a drawback in Random Forest. The model is trained to foresee the target interactions. The proposed method has an average accuracy of 97% which has outperformed most methods produced previously. The study of target interaction is necessary to understand the drugs to be used for a patient with certain medical attributes which in turn controls the negative effect of drugs on the individual. The DTIs will be referred for the current study to increase the overall result of the produced model.

This paper (A. Kumar, C. Janaki, M. V. Hosur and S. N. Pal, 2021) [3] depicts the 'Machine learning techniques to identify potential drug targets for Anti-epileptic drugs' for every model studied there are certain tasks which outperformed whereas the others are not at the best. The idea is to assemble and merge many models to achieve a high rate of accuracy.

The treatment for Anti-epileptic drugs (AED) is available and almost 30% of the patients are resistant. This study applies ML algorithms to predict the potential targets for the expansion of new AED's. Feature selection is done through Physio-Chemical, Structural and post-transitional modifications and drug targets are labelled using the Support Vector Machine, Decision tree and Random Forest algorithms. Three sets of datasets are used to train the model where the unbalanced dataset gave an accuracy of 75% and the balanced dataset achieved an accuracy of 90% as stated by the author. This model is useful for identifying the target for any disease. Hence this approach can be utilized in the current study to find the potential targets for the drug classification.

III. DATA SET

The dataset was taken from Kaggle and the link to the same will be mentioned under Appendix.

A. Description

The dataset used as a part of this paper consists of data of multiple people across various age groups suffering from a similar kind of disease. It contains their medical information like

- *Age*
- *Sex*
- *BP* – This value indicates the Blood pressure level of a particular patient. It is categorized into 3 levels HIGH, NORMAL and LOW. NORMAL blood pressure is an ideal value here and both LOW and HIGH blood pressure could be detrimental to a patient's health.
- *Cholesterol* – This value indicates the Cholesterol level of a particular patient. It is categorized into 2 levels HIGH and NORMAL. A healthy person should ideally show NORMAL cholesterol level.
- *Na_to_K* – This value indicates the ratio of the amount of Sodium to potassium found in the blood. To maintain normal body function, it is important to have a balanced level of potassium and sodium in the blood. Consuming too much sodium and not enough potassium can lead to high blood pressure. Therefore, it is important to consider the balance between these electrolytes when building a machine learning or deep learning model that is related to blood pressure or overall health. A diet that is rich in potassium and low in sodium can help to maintain a healthy balance of these electrolytes and support overall health.
- *Drug* – This indicates the drug which the patient needs to take who is suffering from the disease. There are 5 drugs in the dataset as Drug A, Drug B, Drug C, Drug X and Drug Y.

B. Problem Statement

For a particular disease, there are five Drugs available as medication. These drugs are manufactured by different pharmaceutical companies and are recommended to be consumed by patients considering their medical conditions. The drug that needs to be taken by a particular patient depends on a variety of factors like Age, Gender, Blood Pressure and Cholesterol levels, and Sodium to Potassium ratio.

Based on the past data available in the dataset for multiple patients and their drug prescriptions we need to predict the appropriate drug for a patient considering their medical information.

C. Planned Solution

As a part of this paper, we will be implementing supervised machine learning techniques of classification and predicting the drug to be taken by a particular patient based on the medical information of a patient. The goal is to implement a model and achieve maximum accuracy by passing appropriate features to train and test the model. The drug type would be the target variable and other parameters like Age, Sex, BP, Cholesterol, and Na_to_K would act as features.

We will be making use of the following pre-defined models of classification.

- Logistic Regression
- Decision Tree
- K Nearest Neighbor

We will understand more about these models in the next section.

D. Data Basic Information

Let's see the overview of the dataset.

Field/Column Name	Data Type	Description	Sample Values
Age	Integer	Age of a patient	23,47
Sex	String	Gender of a patient	M,F
BP	String	Blood Pressure level of a patient	LOW,NORMAL,HIGH
Cholesterol	String	Cholesterol level of a patient	HIGH,NORMAL
Na_to_K	Float	Sodium to Potassium ratio level	20.942
Drug	String	The medicinal drug to be taken by a patient	drugA,drugB,drugC,drugX,drugY

Figure 1: Basic information of the dataset

IV. Classification Techniques

A. Decision Tree

A Decision tree (DT) is a supervised learning technique and a classifier in the form of a tree consisting of 2 types of nodes: Decision nodes and Leaf Nodes. While classifying the decision tree will traverse from the root node to the leaf node where the final instance of the classification is found. This is also referred to as the top-down, Recursive partitioning approach. At each decision node, the data will be divided depending on certain decision rules which serves to minimize the Gini impurity in the decision tree. The algorithm that is used to build the tree is called the CART (Classification and Regression tree) algorithm where simply a question or condition is mentioned by the decision tree, depending on the answer (Yes/No) the tree will slitted further into more nodes [4].

The decision trees can be used to handle categorical as well as numerical data to analyse the datasets. The process of cleaning the data is very less in DT's. We can perform the validation using statistical tests which prove that decision trees are the most reliable and give great accuracy.

B. K-Nearest Neighbour

K-Nearest Neighbour or K-NN is a supervised learning algorithm which is non-parametric. This algorithm utilizes proximity to make classifications or regression, but typically this is used in classification. 'K' is the representation of the count of nearest neighbours that need to be categorized in the training data. The classification for any new instance is then decided upon using the most prevalent value for the target value among those k examples hence making it sensitive among local structural data. Neither K is a constant value nor it has any mathematical formula to determine it completely depends on the experiment and the dataset being used where different values of K will be taken to maximize the accuracy of the classification and the performance of the overall model.

This algorithm is also known as the 'Lazy Learner algorithm' as instant learning from the training set is almost impossible in this case whereas at the time of classifying the data it will store the dataset and action is taken on the dataset. The distance between the data instances is calculated using the 'Euclidian Distance' hence the data instances should be processed correctly in order to classify without any impurities.

C. Logistic Regression

Logistic regression is a supervised machine learning technique also referred to as a logit model which will predict the probability of predictive analysis and classification. Instead of using linear regression, which predicts continuous results, it accomplishes this by forecasting categorical outcomes. It is divided into two categories as Binomial where there are only 2 possible outcomes and Multinomial where there are more than two possible outcomes. This model is also referred to as the Discriminative model which will distinguish the classes. When the model has a significant number of predictor variables, logistic regression is especially vulnerable to overfitting. Regularization is utilized when the model has high dimensionality. Logistic regression is very efficient as it is easy to train and implement and the computational power required is very less compared to any other methods in machine learning. Hence making it an idle choice for the model being developed.

V. Experimental Setup

A. System Setup

- Dataset - The dataset was obtained from Kaggle.
- Coding – All the coding is done using Python as the main language and its different libraries for data processing and visualization.
- Platform - Google colab was used as an environment to execute the code.

B. Exploratory Data Analysis

Exploratory Data Analysis is a crucial step in the data analysis process. It allows you to understand the nature and characteristics of your dataset and uncover valuable insights that can help you confirm whether the data is relevant to your business problem. This step can provide valuable information about the dataset and help you ensure that the data makes sense in the context of your analysis.

This data set contains 6 variables in total of which 4 are categorical features and 2 are continuous features. After exploring the dataset, we discovered that there are no null or missing values for any of the variables and the data is already cleaned for processing. The dataset contains around 200 patients' information and the drugs prescribed to them. The numerical features Age and Na_to_K are distributed uniformly and could be depicted in the form of histograms.

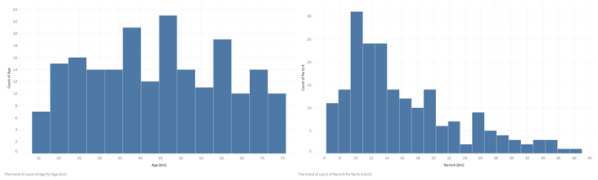

Figure 2: Histogram of 'Age' and 'Na_to_K' numeric feature

We also categorized the 'Age' variable into different Age groups to understand the patients that we are dealing with. Based on their age a patient could either be a 'Young', 'Middle Age' or 'Senior Citizen'.

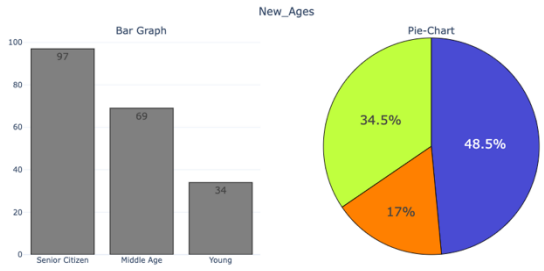

Figure 3: Data Visualization for 'Age'

The findings showed that the majority of the people suffering from the disease were 'Senior Citizens' followed by 'Middle Age' people and the 'Young' patients were very few. The sex ratio also was around 1:1 where the no of male patients was almost similar to the no of female patients.

Let's move to the categorical features, 'BP', 'Cholesterol' and 'Na_to_K' ratio is some important medical information which plays a vital role in drug prescription. Any change in the levels of each of the features would affect the decision of drug to be prescribed so it is very important to explore this data before actually making use of it to train the model.

BP – For visualizing the Blood Pressure levels we replaced the categorical values with numerical constants.

The levels were replaced as below,

- LOW – 0
- NORMAL – 1
- HIGH – 2

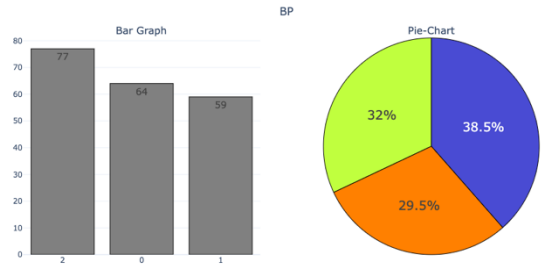

Figure 4: Data Visualization for 'BP'

According to the visualization, the findings specified that the majority of the patients either had High blood pressure or Low blood pressure. Around one-third of the patients had normal blood pressure.

Cholesterol – Cholesterol levels were almost the same for both types but patients with higher levels of cholesterol were on the upper side compared to the patients having low cholesterol levels.

Na_to_K - The expected range for sodium in the blood is 135-145 and the expected range for potassium is 5-5.5 in an adult human. Therefore, the Na_To_K ratio should be between 24.5 and 29. Individuals with a Na_To_K ratio outside of this range are considered to be either good or bad.

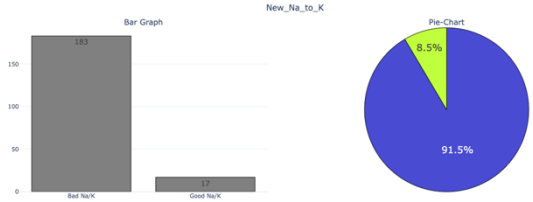

Figure 5: Data Visualization for 'Na_to_K'

The results showed that the majority of patients had a bad sodium-to-potassium ratio. This could be a major factor to be considered while prescribing medicines.

C. Data Pre-Processing

Data preprocessing is the process of cleaning, transforming, and organizing raw data to make it suitable for downstream analysis or modelling. This involves a range of techniques, such as removing missing values, identifying and addressing outliers, and applying appropriate scaling or normalization. Data preprocessing is a crucial step in the data science process, as it ensures that the data used in analysis or modelling is of high quality and well-organized.

As seen earlier this dataset has 2 numeric/continuous features and 4 categorical features so we need to convert the categorical features to continuous features. For converting the categorical features, we will be making use of the Label encoding method. Label Encoding refers to converting the labels into a numeric form so as to convert them into a machine-readable form. Machine learning algorithms can then decide in a better way how those labels must be operated. It is an important pre-processing step for the structured dataset in supervised learning [5]. Since most of the categorical features had distinct defined values so we chose the Label Encoding technique over one hot encoding.

D. Preparing Training and Test Data

After all the features were converted to numeric features, we split the data sets. These sets were then scaled. Scaling is the process of transforming the range of values for the features in a dataset so that they are standardized. This is often necessary for machine learning algorithms to work effectively, as many algorithms require that the input features have the same scale. Scaling the data can also improve the performance of a model by helping the algorithm converge faster. To scale the data, the values for each feature are typically transformed to have a range of values such as between 0 and 1 or between -1 and 1.

E. Modelling

We have applied three supervised learning models of classification and trained and tested them using the datasets which we created in the previous step.

Models trained are as below:

1. Decision Tree
2. K Nearest Neighbor
3. Logistic Regression

VI. Results And Discussion

We trained the models with a variety of combinations of train and test data to check their accuracy and the result was as below

Train Data:|Test Data ratio	Decision Tree	K Nearest Neighbor	Logistic Regression
80|20	100%	70%	85%
70|30	100%	72%	82%
60|40	99%	76%	81%

Figure 6: Accuracy table for all the models

We can see that the Decision tree has consistently shown the maximum accuracy followed by Logistic Regression and the least accuracy was shown by K Nearest Neighbor. Also, we can observe that the less train data is passed to the K Nearest Neighbor algorithm the accuracy keeps on increasing. So, we can say that the K Nearest Neighbor is not the most reliable one.

To dig deeper into evaluating the models we plotted their confusion matrix. A confusion matrix is a tool for evaluating the performance of a classification model. It shows how well the model predicts the true class labels of a set of data. The confusion matrix allows you to see how well the classifier is performing by comparing the predicted class labels with the true class labels.

The results of the confusion matrix when the train data was 60% and the test was 40% were as below.

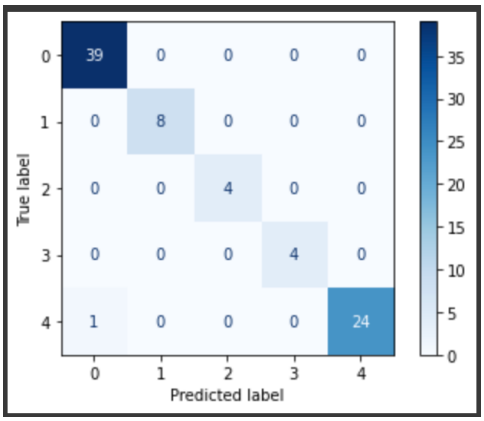

Figure 7: Confusion matrix for Decision Tree model

The number of true positives and true negatives is way too higher and there are no false positives and only one false negative.

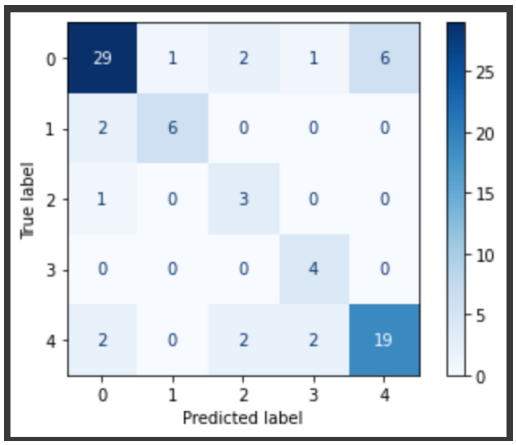

Figure 8: Confusion matrix for K Nearest Neighbor model

The number of true positives and true negatives is comparatively low and a good number of false positives and false negatives could be seen.

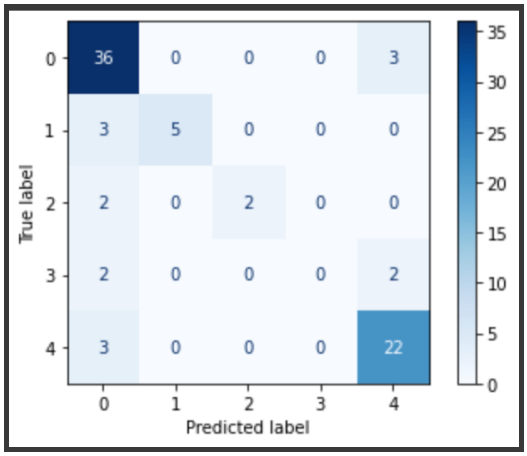

Figure 9: Confusion matrix for Logistic Regression model

For the Logistic Regression model, the true positives and negatives are showing decent numbers and there are not so frequent false positives and false negatives.

VII. Conclusions

After considering all the factors and evaluating the results of the models we can conclude that the Decision Tree is the best fit to solve our problem. With the medical information of a patient, a Decision Tree classifier is successfully able to predict the type of drug that needs to be taken by a patient.

As a part of future work, more medical details of the patients could be added to the dataset and also different drugs could be added for a variety of diseases. An implementation of the Decision Tree model should be able to successfully predict the drug for a patient based on his/her medical condition.

APPENDIX

Link to the Dataset - https://www.kaggle.com/datasets/prathamtripathi/drug-classification?datasetId=830916

Google Drive Link for complete code – https://colab.research.google.com/drive/1vGjTOOuzb54OvmB5S1MFEaxI-Z17kHmd?usp=sharing

This link contains code which is written in Python and can be executed using Google colab. It contains code for all the processes from Data Visualization, Pre-processing the data and Modelling.

Note: Please download the dataset CSV file and upload it to the root folder of the google colab instance. It should be accessible at path '/content/drug200.csv' or else you need to change the path for reading the .csv file from the source.

REFERENCES

[1] Gala, D. V., Gandhi, V. B., Gandhi, V. A., & Sawant, V. (2021, October). Drug Classification using Machine Learning and Interpretability. In 2021 Smart Technologies, Communication and Robotics (STCR) (pp. 1-8). IEEE.

[2] Barkat, M. R., Moussa, S. M., & Badr, N. L. (2021, December). Drug-target Interaction Prediction Using Machine Learning. In 2021 Tenth International Conference on Intelligent Computing and Information Systems (ICICIS) (pp. 480-485). IEEE.

[3] Kumar, A., Janaki, C., Hosur, M. V., & Pal, S. N. (2020, December). Machine Learning techniques to identify potential drug targets for Anti-epileptic drugs. In 2020 IEEE International Conference on Machine Learning and Applied Network Technologies (ICMLANT) (pp. 1-6). IEEE.

[4] Decision tree algorithm in Machine Learning - Javatpoint (no date) www.javatpoint.com. Available at: https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm (Accessed: December 13, 2022).

[5] Anon. (2022) *ML: Label Encoding of Datasets in Python* [online]available from <https://www.geeksforgeeks.org/ml-label-encoding-of-datasets-in-python/> [13 December 2022]

[6] Bilalbora (2022) *Drug Classification- Plotly EDA- ML* [online]available from <https://www.kaggle.com/code/bilalbora/drug-classification-plotly-eda-ml#%F0%9F%A7%AA-INTRODUCTION-THE-DRUG-CLASSIFICATION> [15 December 2022]

[7] Ihsncnkz (2022) *Decision Tree and Random Forest Classifier Models* [online]available from <https://www.kaggle.com/code/ihsncnkz/decision-tree-and-random-forest-classifier-models> [15 December 2022]

[8] Guptankit (2021) *Different Classification Algorithms on Drug Data* [online]available from <https://www.kaggle.com/code/guptankit/different-classification-algorithms-on-drug-data> [15 December 2022]