*A Project Report*

*on*

***Drug Classification***

carried out as part of the course CSE CS3203

submitted by

*Anshika Sahu 209301654*

*Armaan Sidhu 209301656*

*Parv Valiramani 209301657*

*Khushi Singh 209301633*

From Sixth Semester Section I

*of*

**BACHELOR OF TECHNOLOGY**

In

**Computer Science & Engineering**

Logo

Description automatically generated with low confidence

**Department of Computer Science & Engineering,**

**School of Computer Science and Engineering,**

**Manipal University Jaipur,**

***March 2023***

**Acknowledgement**

This project would not have completed without the help, support, comments, advice, cooperation, and coordination of various people. However, it is impossible to thank everyone individually; We are hereby making a humble effort to thank some of them.

We acknowledge and express our deepest sense of gratitude to our course instructor Dr Satya Prakash Maurya Sirfor his constant support, guidance, and continuous engagement. I highly appreciate his technical comments, suggestions, and criticism during the progress of this project “Drug Classification”. He also helped us to find the correct data set for our model.

We owe my profound gratitude to each of our teammatesfor their precious support and cooperation during the development of this project.

Finally, I extend my heartfelt appreciation to my classmates for their help and encouragement.

**Date of submission:**

**15th April 2023**

**By team:**

**Anshika Sahu 209301654**

**Armaan Sidhu 209301656**

**Parv Valiramani 209301657**

**Khushi Singh 209301633**

1. **Introduction of Problem**

The project on drug classification aims to develop a machine learning model that can accurately classify drugs based on their molecular features. This classification is essential for identifying new drug targets and optimizing drug development processes. The model will use various medical features measured by various pathological tests on a large group of patients to classify drugs into different classes, as per our given data set. The successful implementation of this project will contribute to the development of more effective and efficient drug prescription strategy and development processes, leading to the development of new and improved drugs for the treatment of various diseases and medical conditions.

1. **Problem Statement**

Problem Statement: Our project aims to predict that out of the five classes of drugs present in our dataset namely drug A, drug B, drug C, drug X, drug Y, which one is likely to be prescribed based on some other set of medical attributes of the patient such as Na to K ratio, Blood Pressure level, Cholesterol level, age, and gender.

1. **Data set Description**

Our dataset contains six attributes namely Age, Sex, BP, Cholesterol, Na\_to\_K, Drug.

Total number of records: 200

Target Attribute: “Drug”

Categorical Attributes: Sex, BP, Cholesterol, Drug

Numerical Attribute: Age, Na\_to\_K

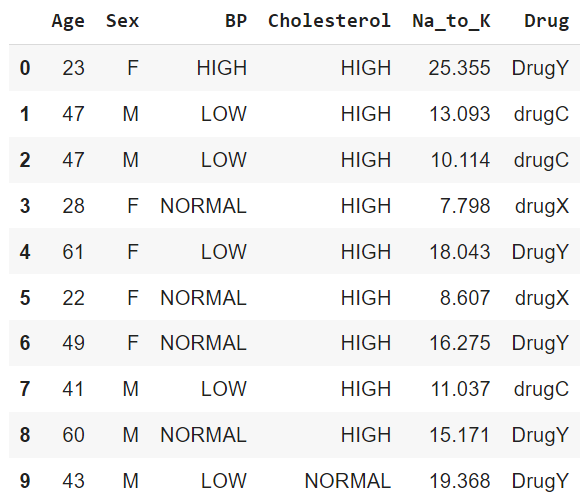
Missing Value: None

The attributes other than “Drug” attribute are the set of medical parameters probably collected from various patients by means of medical tests in pathology and the five classes of drugs were prescribed to them based on the values of these parameters.

Distribution of number of records in the target classes is as follows:

|  |  |
| --- | --- |
| Drug | Number of Records |
| Drug A | 23 |
| Drug B | 16 |
| Drug C | 16 |
| Drug X | 54 |
| Drug Y | 91 |

Sample Values:



Data Pre-processing:

1. We choose x as an array which contains values of attributes age, sex, BP, Cholesterol, Na\_to\_K as

X [0]: `Age`

X [1]: `Sex`

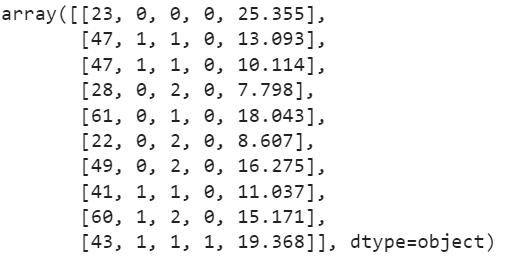
X [2]: `BP`

X [3]: `Cholesterol`

X [4]: `Na\_to\_K`

This array can be divided into test and training set.

1. We wanted all the attributes to be of numerical datatype for easy computation. Thus, on applying label encoder on all the categorical attributes:



1. We also introduced an extra attribute called “Dclass” to the dataset as a numerical version of the Drug column.
2. **Algorithm Used**

*•Decision Tree Classifier* using Gini index as criterion.

•Given a collection of records (our training set) which contains a set of attributes in which one of the attributes is the target class, a decision tree finds a model for that target class attribute as a function of the values of other attributes.

•The goal is to assign previously unseen records to a class as accurately as possible.

•Greedy approach to construct a decision tree is to select nodes with homogenous class distribution or less impurity which can be measured using: “GINI INDEX”

•Gini Index for a given node t:

GINI(t)=1-

•When a node p is split into k partitions (children), the quality of split is computed as:

=GINI(i)

Where, =number of records at child I,

n=number of records at node p

1. **Data Visualization**

Number of records in each target class:

Chart, bar chart

Description automatically generated

• Analysis of categorical attributes

Chart, scatter chart

Description automatically generated

* Here we can analyse that Drug A and Drug B is prescribed only to people with high BP and Na to K ratio less than 15.
* Whereas Drug C is prescribed inly to people with low BP and Na to K ratio less than 15.
* Drug Y is given only to patient with Na to K ratio greater than 15.

Chart, scatter chart

Description automatically generated

-Drug C is prescribed only to people with high cholesterol.

-Again, Drug Y is given to people with Na to K ratio with value greater than 15.

Chart, scatter chart

Description automatically generated

We do not get any classification clue from this attribute except that Drug Y was always prescribed for Na to K values greater than 15.

• Pairwise plot of each attribute with the target attribute is

Chart

Description automatically generated with medium confidence

From the pairwise plot age attribute has scattered distribution of records so no interesting information from this attribute and as discovered earlier here also drug Y records (green dots) fall only in the region of Na to K value greater than 15.

1. **Code Link:**

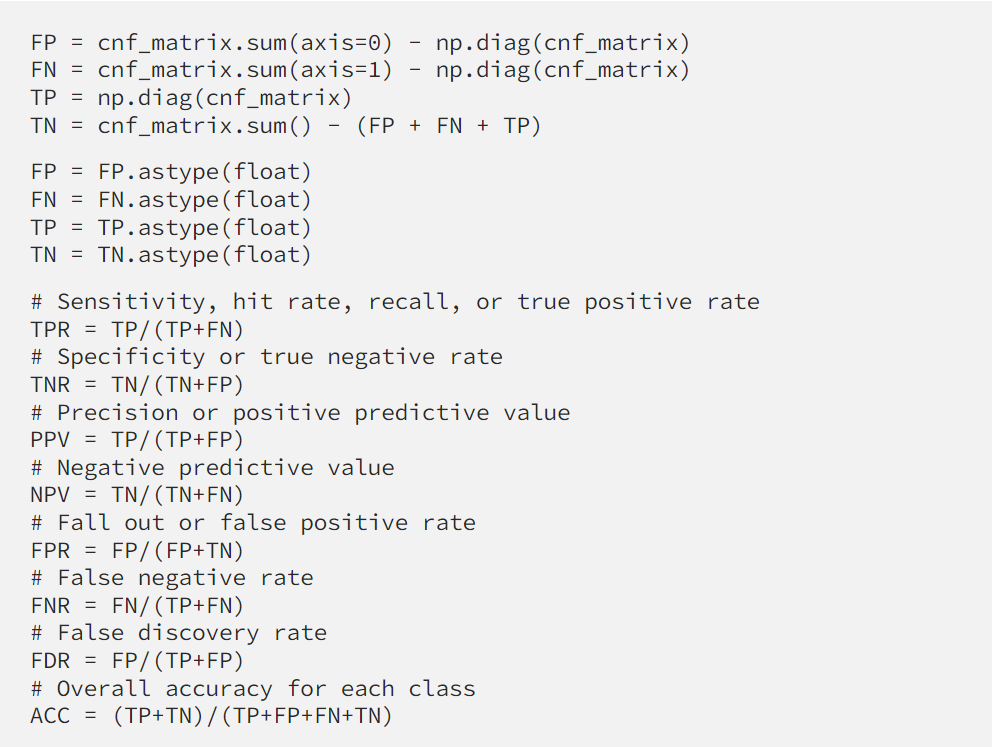
<https://colab.research.google.com/drive/1Sed5HAv1trHDYLfwKF2bfkagm8pbutOH?usp=sharing>

1. **Data Set Link:**

[**https://drive.google.com/file/d/1odrEvxgQ\_bDO-ZIAu0K\_a3\_w0mEsQegt/view?usp=sharing**](https://drive.google.com/file/d/1odrEvxgQ_bDO-ZIAu0K_a3_w0mEsQegt/view?usp=sharing)

1. **Result Analysis**

**•** Formula used for Calculation of Evaluation Measures:



•**Ratio 1: 70:30**

1. Confusion Matrix

Graphical user interface, application

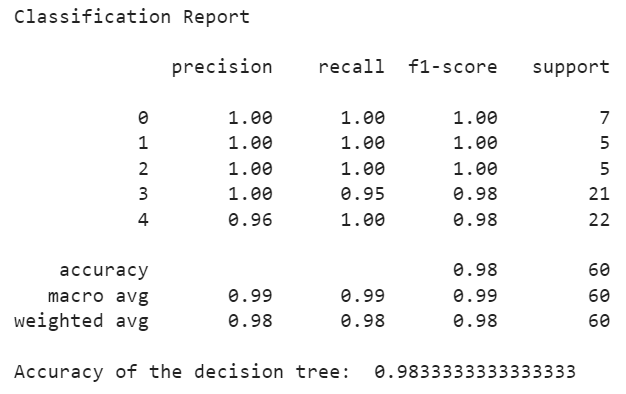
Description automatically generated

1. ROC

Chart

Description automatically generated

1. Evaluation Parameter



* Evaluation Measures
* Table

  Description automatically generated with low confidence

1. Result:

• Text Representation:

Text

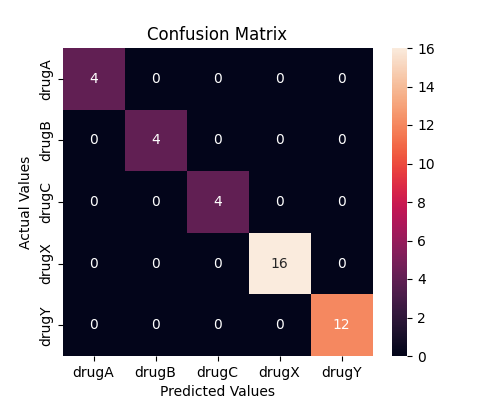
Description automatically generated

• Tree PlotA picture containing qr code

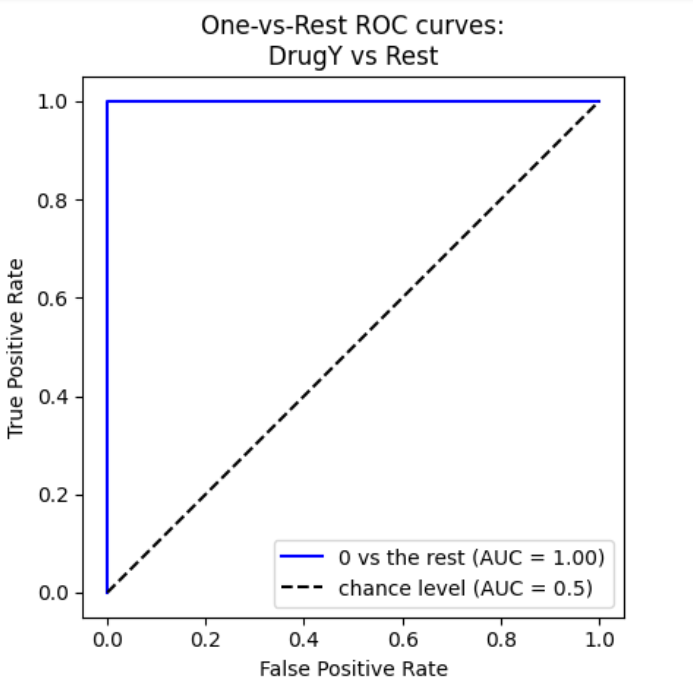
Description automatically generated

**• Ratio 2: 80:20**

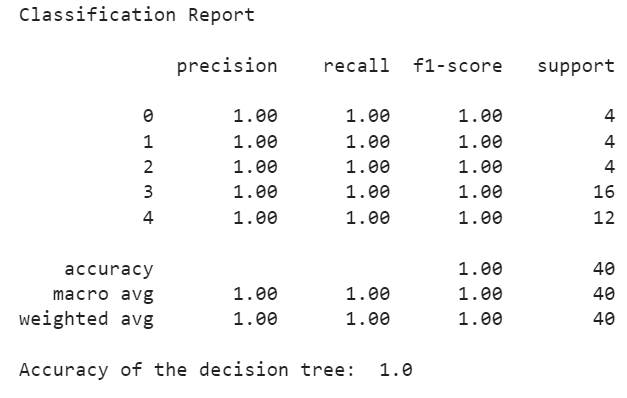
1. Confusion Matrix:



1. ROC



1. Evaluation Parameter



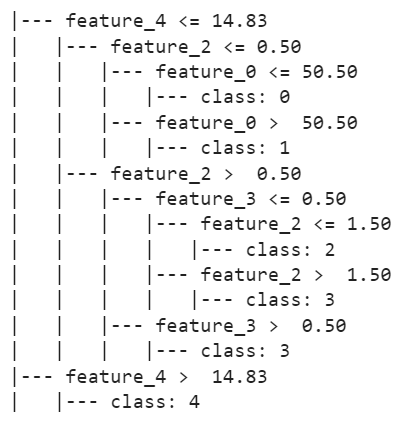
-Evaluation Measures

Text

Description automatically generated

1. Result

• Text Representation



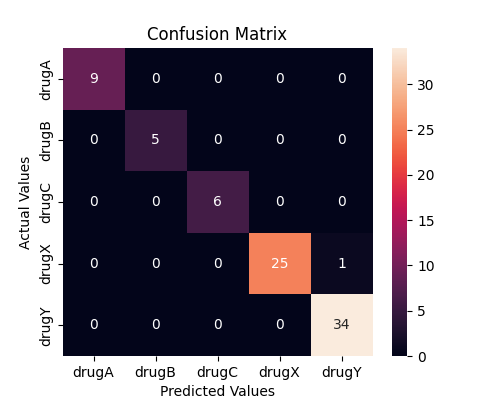
• Tree Plot

A picture containing qr code

Description automatically generated

**• Ratio 3: 60:40**

1. Confusion Matrix



1. ROC

Chart

Description automatically generated

1. Evaluation Parameter

A picture containing text, receipt, screenshot

Description automatically generated

-Evaluation Measures

Table

Description automatically generated

1. Result

• Text Representation

Text

Description automatically generated with low confidence

• Tree Plot

A picture containing qr code

Description automatically generated

1. **Conclusion**

Comparison Table:

|  |  |  |  |
| --- | --- | --- | --- |
| Algorithms  Measures | Train – test  Ratio 1 | Train – test  Ratio 2 | Train – test  Ratio 3 |
| Specificity | 0.9947 | 1.0000 | 0.9956 |
| Sensitivity | 0.9905 | 1.0000 | 0.9923 |
| Accuracy | 0.9834 | 1.0000 | 0.9875 |
| Precision | 0.9913 | 1.0000 | 0.9943 |
| FPR | 0.0053 | 0.0000 | 0.0043 |
| FNR | 0.0095 | 0.0000 | 0.0077 |
| NPV | 0.9950 | 1.0000 | 0.9963 |
| FDR | 0.0086 | 0.0000 | 0.0057 |
| F1-Score | 0.98 | 1.00 | 0.99 |
| MCC | 0.9857 | 1.0000 | 0.9893 |

Based on the results of the project, the decision tree classifier using Gini index as the criterion proved to be an effective algorithm for predicting which drug is more likely to be prescribed based on the set of medical attributes of a patient. The evaluation measures for all three train-test ratios (70:30, 80:20, and 60:40) showed high accuracy, sensitivity, specificity, precision, F1 score, and Matthew’s correlation coefficient (MCC) values, indicating that the model performs well in predicting the target attribute (Drug) based on the other medical attributes. The visualization of the data revealed some interesting patterns, such as the correlation between Drug Y and a high Na to K ratio, or the fact that Drug C is prescribed only to people with high cholesterol. These insights can be useful for medical professionals in understanding the prescribing patterns and making informed decisions. Overall, the project successfully demonstrated the potential of decision tree classifiers in predicting the prescription of drugs based on medical attributes, which can ultimately lead to more personalized and effective treatments for patients.