**PROJECT REPORT**

**On DRUG prediction using**

**Decision Tree algorithm**



**Department of Computer Science & Engineering**

**Dr B R Ambedkar National Institute of Technology,**

**Jalandhar**

**Submitted To** **Submitted by**

Dr Kuldeep Kumar Sir Suneeta Singh

21203028

**Table of Content**

1. **Introduction…………………………………………………...3**
   1. **Brief about Machine Learning………………………....3**
      1. **Types of Machine Learning…………………………3**
   2. **Decision Tree…………………………………………....3**
   3. **Application of Decision Tree………….…………………...3**
2. **About Project………………………………………………….4**
   1. **Model works on given Metrics………………………….4**
      1. **GINI Index Metrics…………………………………..4**
      2. **Entropy and information Gain………………………4**
3. **Flowchart……………………………………………………….5**
4. **Steps to build Decision Tree model……………………………6**
   1. **Downloading and Loading the Dataset…………………6**
   2. **Preprocessing of Dataset………………………………....6**
   3. **Setting up the Decision Tree……………………………..7**
   4. **Practice……………………………………………………8**
   5. **Modeling of Tree……………………………………….....8**
   6. **Prediction/Evaluation………………………………….…9**
   7. **Vizualization of Tree…………………………………......10**
   8. **Comparision of Two Results…………………………......11**
5. **CONCLUSION…………………………………………………12**

**2**

**1.INTRODUCTION**

**1.1 MACHINE LEARNING**

Machine learning is a part of Artificial intelligence which used for software applications to become more accurate by predicting outomes without using explicite programming.Machine Learning algorithms take historical data as input to predict output values.

Many of leading companies as facebook,google,uber makes machine learning as a central part of their operaions.Machine learning is often classified into different parts:

**1.1.1 Types**

**1.Supervised Learning:** Algorithm used labeled types of data for training.

**2.Unsupervised learning:** In this type of algorithm ,involves training on unlabeled data.

**3.Semi-Supervised Learning:**This algorithms used mix type of data.Some Used labeled data and some used unlabeled data for training.

Now here we are discussed about Decision tree algorithms.

**1.2 DECISION TREE**

Decision trees are a type of Supervised machine learning that can be used for both classification and regression problems,but mostly used for solving classification problems.Decision tree is powerful analytical model that has the ability to comprehend data with minimal preprocessing time.In decision tree data is splitting continuously according to given parameter.

A Decision tree consists of nodes in which internal nodes represent feature of dataset,branches represent decision rules and the leaf nodes represent the outcome.

Decision trees classify based on sorting them from root to leaf node.Each node in trees specify a test of some attributes and each branch from that nodes corresponds to some values for that attributes.

**1.3 APPLICATION OF DECISION TREE**

**1.**Decision tree modeling can be used for making predictions.

**2**.As many organization had created their database to enhance their services.Deision tree is possible way to extract information from database.

3

**3**.It is an approach to manage costumer relationship that is investigate how individual users access online services.

**2. ABOUT PROJECT**

In this project we are implementing Drug medication cases using Decision tree model. We have collected data about set of patient ,where everyone suffered from same illness.In their course of treatment duration time,each patient had responded to one of the given 5 medications,Drug A,Drug B,Drug C,Drug X,Drug Y.

Now we have to buid a model to find out which drug would be more appropriate for future patient with same sickness.There are different features of this dataset are Sex,Blood Pressure,Age and the Cholesterol of the patient and the target of this is Drug that each patient responded too.We can use the training part of dataset to build model and use it to predict class of new unknown patient.

Here we assume feature of Dataset is X parameter and Y as a Target parameter.

**2.1 METRICS USED**

There are few metrics that we can use to build a decision tree model,but here we would use only two most famous metrics:GINI Index and Entropy and Information gain metrics.

**2.1.1 GINI INDEX**

Gini Index score used to evaluate that how splitting is good by mixed the classes that splits in two groups.It could have a score between 0 and 1 values,where 0 is for all observations belongs to one class and 1 is for random distribution of the observations within the classes.

**2.1.2 ENTROPY AND INFORMATION GAIN**

For entropy we can say that ,measurement of impurity within a dataset.

And Information gain is differenc of entropy before and after the splitting of a feature.

Both the method measured the impurity used to spilitting but in different way. Gini Index calculates the binary split impurity , where as Information Gain measures the entropy before and after the splitting. .

43. **FLOWCHART**

**Dataset**

**Preprocessing**

**Setting up Decision Tree**

1.Devided into X and Y for Prediction

2.Train andTest Split

3.Shape of Training and Testing

**Visualization**

By plotting tree

**Modeling by criterion Entropy**

And Apply Decision tree classifier

**Comparision**

1.Predicted score with testing values

2.Best Score

3.Best Parameter

**Visualization**

By plotting tree

**Prediction/Evaluation**

1.Prediction Tree

2.Confusion Matrix

3.Classification-Report

4.Accuracy Score

**Prediction/Evaluation**

1.Prediction Tree

2.Confusion Matrix

3.Classification-Report

3.Accuracy Score

**Modeling by criterion GINI Index**

And Apply Decision tree classifier

5

**4. Steps to Build Decision Tree model**

**4.1 Downloading and Loading the Dataset**

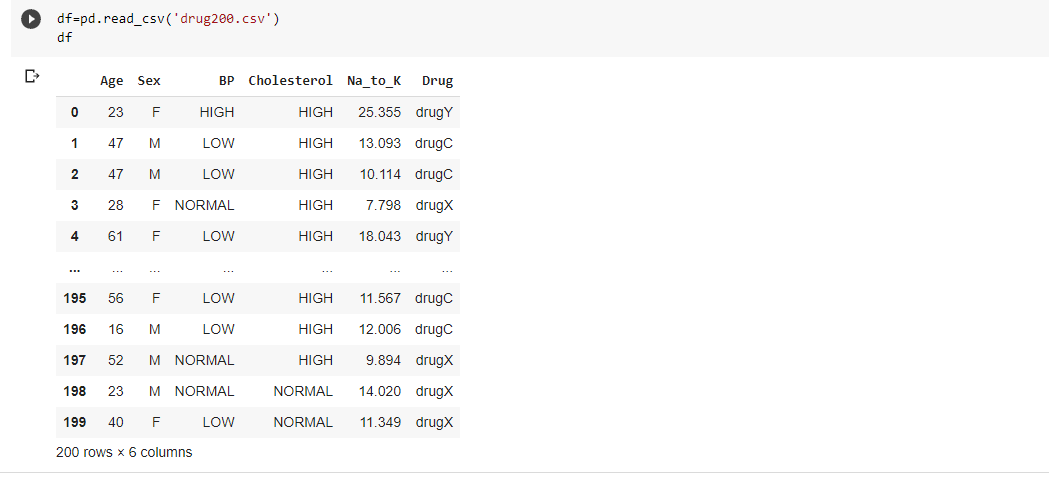
**4.1.1** Downloaded the dataset from <https://drive.google.com/file/d/1Us7u4Sy13MvRMXTMk2RgfE68OGHdGYUX/view?usp=sharing> .

**4.1.2** .Create a new file in Google Collab Plateform.

**4.1.3**  Imported all the libraries numpy,pandas,DecisionTreeClassifier from sklearn.

**4.1.4** Loaed the dataset named as “drug200.csv” into Collab file.

**4.1.5** Open and read the data file using pandas dataframe.



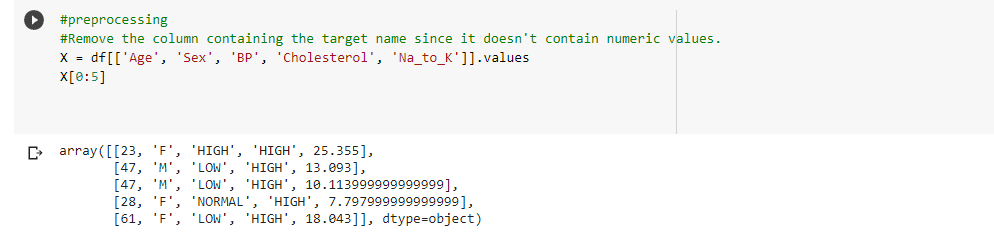
**4.2 Preprocessing of Data**

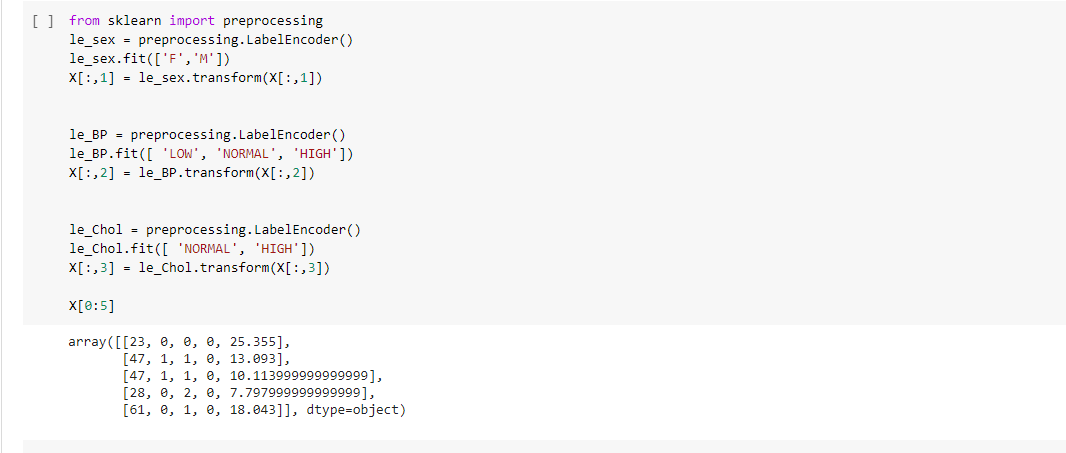
Using as ‘df’ the drug200.csv data read by pandas, and variables are:

X is the Feature Matrix that is input data (data of df) and Y is the response vector (target or output data)

Remove the column of target name because it does not contain numerical values.

6



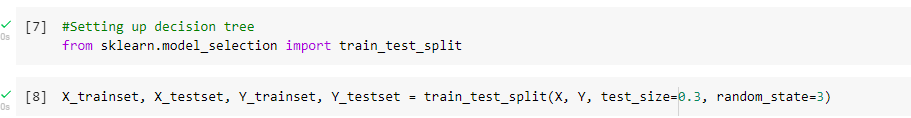


**4.2.1** Fill the target variable.

**4.3 Setting up the Decision Tree**

**4.3.1** Devided the data into X and Y where X is set of attribute used for prediction and Y attribute used for whose value needs to be predicted.

Imported train -test split from library sklearn.model\_selection and this return 4 parameters X\_trainset, X\_testset, Y\_trainset, Y\_testset.

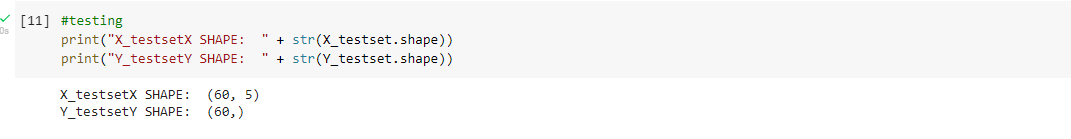
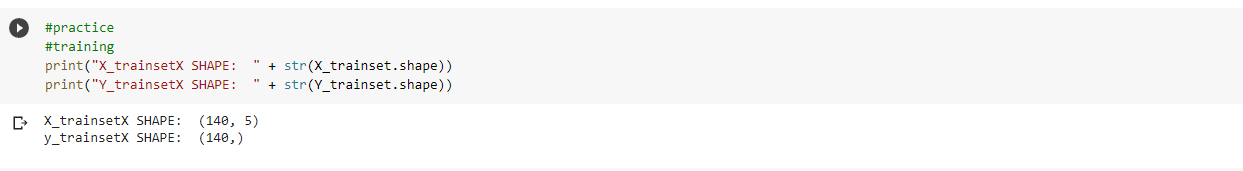


Here test\_size represents the ratio of testing dataset and random\_state ensures that obtatin the same splits.And 30% dataset to testing and 70% dataset to training.

7

**4.4 Practice**

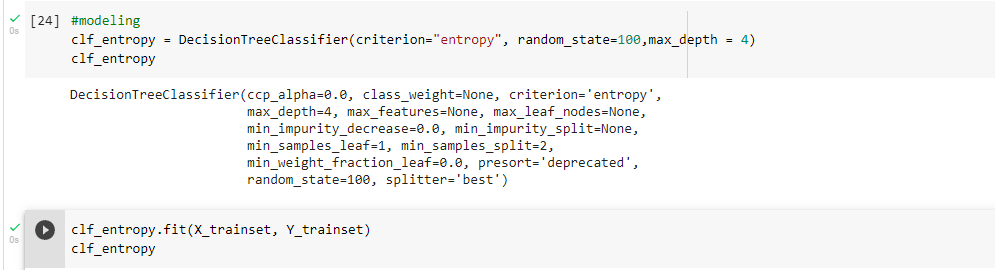
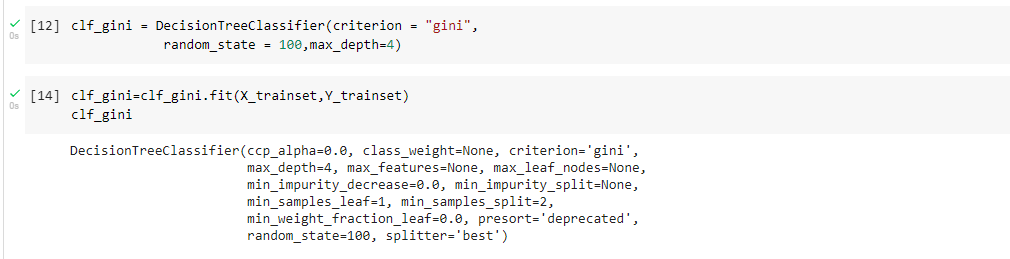
**4.4.1** Print the shape of X\_trainset ,Y\_trainset,X\_testset and Y\_testset and confirm that dimensions match.



**4.5 Modeling**

**4.5.1** Apply Decision Tree Classification into training data and train the model .

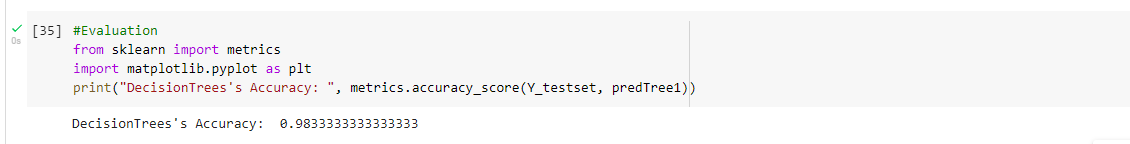
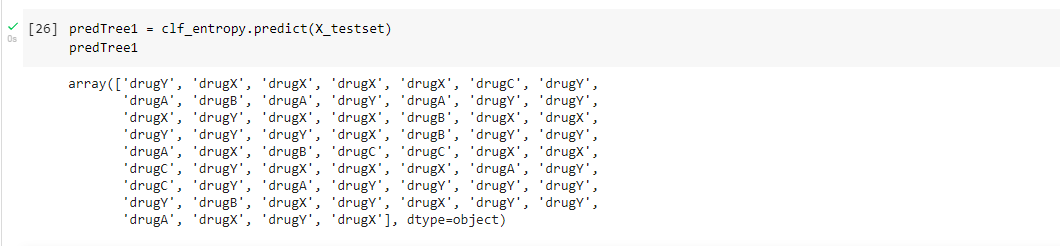
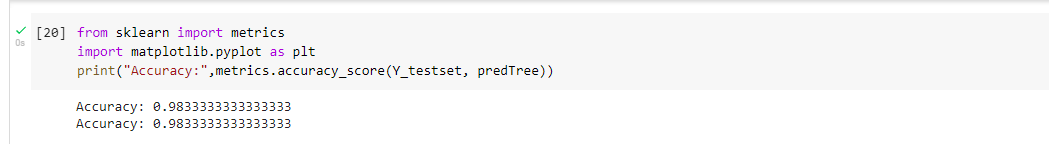
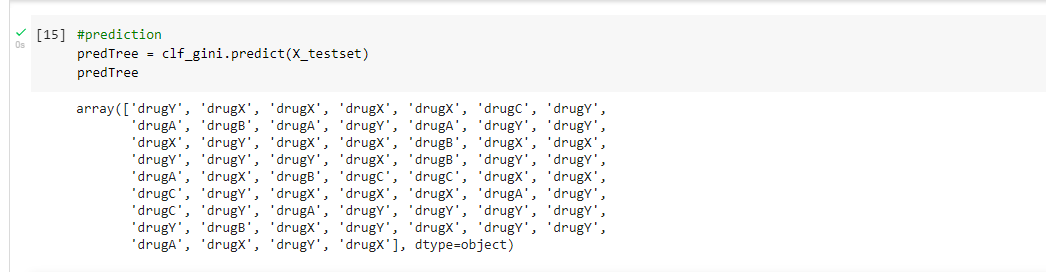
**4.5.2** Did this once using Criterian Gini Index and once with Criterian Entropy.



**8**

**4.6 Prediction /Evaluation**

**4.6.1** Now import the metrics from sklearn.plot confusion matrix and classification report and get the predicted accuracy score of our model by both metrics Gini index and Entropy.

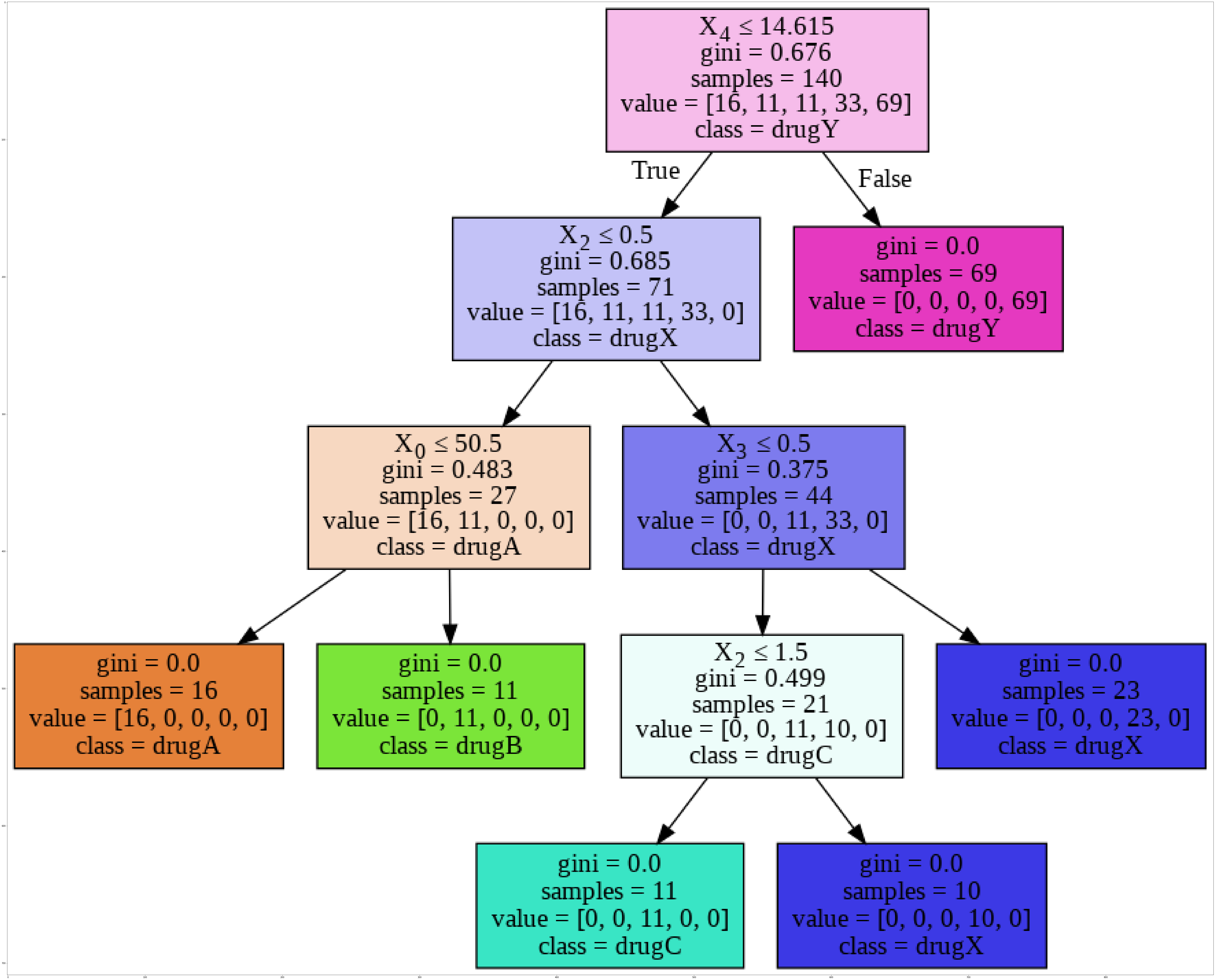


**4.7 Visualization of Model**

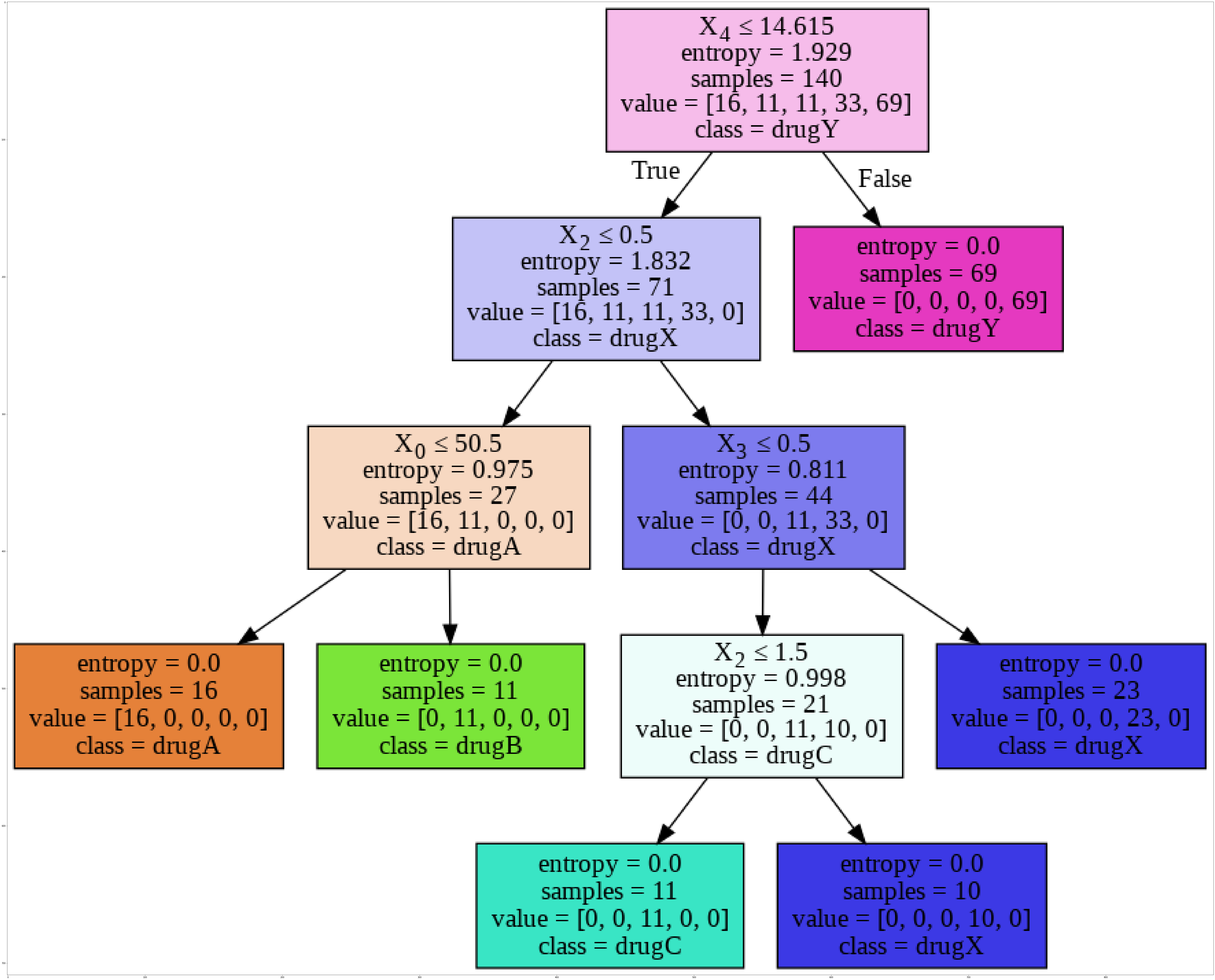
**4.7.1** Imported all the libraries StringIO,pydotplus,matplotlib for plotting the tree.

**4.7.2** Load tree figure as ‘Drugtree.png ‘ and plot both the trees ,once by Gini criterian and other by entropy.

9



10

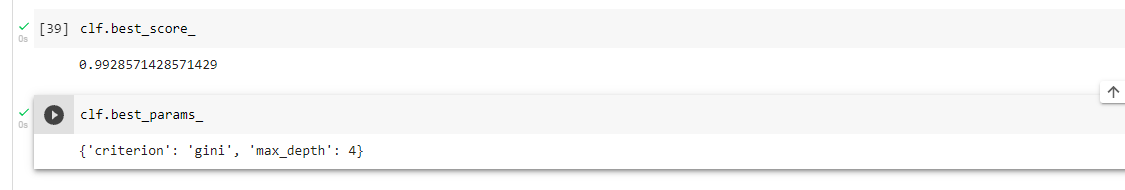


**4.8 Comparision**

**4.8.1**  At last imported GridSearchCV for comparision and compare both the predicted accuracy score with testing values.



**4.8.2**  And it was 99.2 percentile by criterian Gini and maximum depth 4.



So,finalized the Model.

**5. CONCLUSION**

Above Decision Tree model shows how algorithm works and will be useful for future reference.Tree gives visual representation of all possible outcomes **.**

**12**

**Thankyou !**

**13**