Project Assignment

Aim: To design and implement an appropriate data mining algorithm to provide solution to a real time problem

Roll No.: K014	Name:Heet Gala
Class:B.Tech Cyber Sec	Batch:K1
Date of Tutorial:3/4/23	Date of Submission:3/4/23
Grade:	

1. Brief description about the data set being used.

The dataset being used is a collection of medical data from patients, with the goal of predicting which drug might be most appropriate for a given patient. The dataset contains six variables:

1.]Age: the age of the patient (numerical)

2.]Sex: the sex of the patient (categorical: "M" for male or "F" for female)

3.]BP: the blood pressure of the patient (categorical: "HIGH", "NORMAL", or "LOW")

4.]Cholesterol: the cholesterol level of the patient (categorical: "HIGH" or "NORMAL")

5.]Na_to_K: the ratio of sodium to potassium in the patient's blood (numerical)

6.]Drug: the drug that was prescribed for the patient (categorical: "drugA", "drugB", "drugC", "drugX", or "drugY")

The dataset includes 200 instances of patient data.

The given dataset is best for the CART algorithm because the CART algorithm is a decision tree algorithm used for classification problems. In this dataset, the drug column is the target variable, and the rest of the columns are independent variables used to predict the target variable.

Since the target variable is categorical (drug type), and the independent variables are a mixture of categorical and continuous variables, the CART algorithm can be used to predict the drug type based on the values of the independent variables.

The other three algorithms that could be used for this dataset are k-nearest neighbors (KNN), logistic regression, and random forest. KNN and logistic regression can be used for classification problems, but they might not perform as well as the CART algorithm since the independent variables are a mixture of categorical and continuous variables. Random forest is a decision tree-based algorithm like CART, but it might not perform as well as the CART algorithm since it builds multiple trees and averages the results, which can lead to overfitting. Therefore, the CART algorithm is the best choice for this dataset

2. Description of algorithm being used.

CODE FOR CART ALGO:

```
import os
import pandas as pd
import pydotplus
from IPython.display import Image
from sklearn.preprocessing import LabelEncoder
from sklearn.tree import DecisionTreeClassifier, export_graphviz

#CART stands for Classification And Regression Tree.

# read in the CSV file
data = pd.read_csv("drug200.csv")

# encode non-numeric values using label encoding
le = LabelEncoder()
for col in data.columns:
    if data[col].dtype == 'object':
        data[col] = le.fit_transform(data[col])

# separate the features (X) and target variable
X = data.iloc[:, :-1]
y = data.iloc[:, -1]

# train a CART model
cart_model = DecisionTreeClassifier()
cart_model.fit(X, y)

# define class_names`
```

```
class_names = ["drugA", "drugB", "drugC", "drugX", "drugY"]

# visualize the resulting decision tree
export_graphviz(cart_model, out_file="tree.dot", feature_names=X.columns,
class_names=class_names,rounded=True,filled=True)

# set the PATH variable for Graphviz
os.environ["PATH"] += os.pathsep + '/usr/local/Cellar/graphviz/2.49.0/bin'

# convert the DOT file to a PNG image
graph = pydotplus.graph_from_dot_file("tree.dot")

# save the image to a file
graph.write_png("tree.png")

# display the image
Image(graph.create_png())
```

This algorithm is using the Classification and Regression Tree (CART) algorithm.

The CART algorithm is a decision tree-based classification algorithm used to build a model that can predict a categorical (classification) or continuous (regression) dependent variable based on a set of independent variables. It works by recursively splitting the data into subsets based on the most significant variable at each level of the tree until a stopping criterion is met.

Here, the algorithm is being used to build a classification model to predict which drug is most appropriate for a patient based on their medical information. The input data is read from a CSV file and non-numeric values are encoded using label encoding. The resulting decision tree is visualized using Graphviz, a visualization software, and the resulting image is displayed.

CART is a predictive algorithm used in Machine learning and it explains how the target variable's values can be predicted based on other matters.

In the decision tree, nodes are split into sub-nodes on the basis of a threshold value of an attribute.

The CART algorithm works via the following process:

- The best split point of each input is obtained.

- Based on the best split points of each input in Step 1, the new "best" split point is identified.
- Split the chosen input according to the "best" split point.
- Continue splitting until a stopping rule is satisfied or no further desirable splitting is available.

Gini index

It stores the sum of squared probabilities of each class.

The degree of the Gini index varies from 0 to 1,

- Where 0 depicts that all the elements are allied to a certain class, or only one class exists there.
- The Gini index of value 1 signifies that all the elements are randomly distributed across various classes, and
- A value of 0.5 denotes the elements are uniformly distributed into some classes.

Advantages of CART

- Results are simplistic.
- Classification and regression trees are Nonparametric and Nonlinear.
- Classification and regression trees implicitly perform feature selection.
- Outliers have no meaningful effect on CART.
- It requires minimal supervision and produces easy-to-understand models.

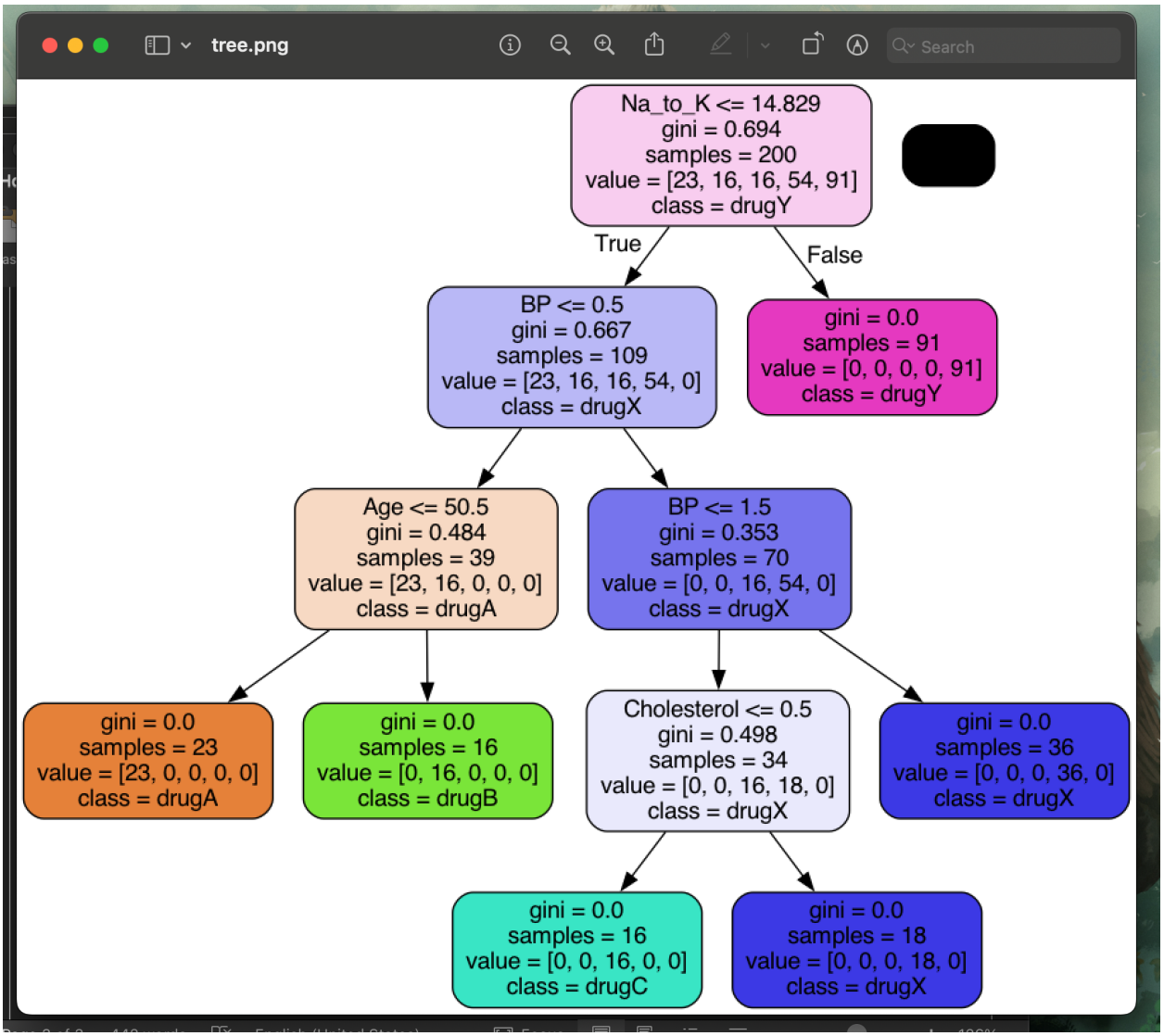
Limitations of CART

- Overfitting.
- High Variance.
- low bias.
- the tree structure may be unstable.

3. Results of implementation.

   The implementation results in a decision tree that can be visualized as shown above. The decision tree has 10 nodes

and represents a classification model for predicting the drug class for a patient based on their Na_to_K ratio, blood pressure (BP), age, and cholesterol levels. The decision tree has a depth of 4. The root node (node 0) splits the data based on the Na_to_K ratio, which has a threshold of 14.829. The gini impurity at this node is 0.694, which means there is a significant level of impurity in the data. The left child node of the root node (node 1) corresponds to the instances with Na_to_K ratio less than or equal to 14.829, and the right child node corresponds to the instances with Na_to_K ratio greater than 14.829. The left child node (node 1) splits the data further based on the blood pressure (BP) attribute, which has a threshold of 0.5. The gini impurity at this node is 0.667. The left child of node 1 (node 2) corresponds to the instances with BP less than or equal to 0.5, and the right child node corresponds to the instances with BP greater than 0.5. Node 2 further splits the data based on age, which has a threshold of 50.5. The gini impurity at this node is 0.484. The right child of node 2 (node 4) corresponds to the instances with age greater than 50.5 and belongs to the drug class B. The left child node of node 2 (node 3) corresponds to the instances with age less than or equal to 50.5 and belongs to drug class A. The right child node of node 1 (node 5) corresponds to the instances with BP greater than 0.5. Node 5 splits the data based on cholesterol level, which has a threshold of 1.5. The gini impurity at this node is 0.353. The left child of node 5 (node 6) corresponds to the instances with cholesterol level less than or equal to 0.5, and the right child node corresponds to the instances with cholesterol level greater than 0.5. Node 6 further splits the data based on the drug class. The left child of node 6 (node 7) corresponds to the instances that belong to drug class C, and the right child node (node 8) corresponds to the instances that belong to drug class X. Node 7 has no child nodes since it corresponds to a pure class. Node 8 also corresponds to a pure class (drug class X) and has no child nodes. The right child of node 5 (node 9) corresponds to the instances with cholesterol level greater than 1.5 and belongs to drug class X. The right child node of the root node (node 10) corresponds to the instances with Na_to_K ratio greater than 14.829 and belongs to drug class Y.


tree.png
Search
Na_to_K <= 14.829
gini = 0.694
samples = 200
value = [23, 16, 16, 54, 91]
class = drugY
True
False
BP <= 0.5
gini = 0.667
samples = 109
value = [23, 16, 16, 54, 0]
class = drugX
gini = 0.0
samples = 91
value = [0, 0, 0, 0, 91]
class = drugY
Age <= 50.5
gini = 0.484
samples = 39
value = [23, 16, 0, 0, 0]
class = drugA
BP <= 1.5
gini = 0.353
samples = 70
value = [0, 0, 16, 54, 0]
class = drugX
gini = 0.0
samples = 23
value = [23, 0, 0, 0, 0]
class = drugA
gini = 0.0
samples = 16
value = [0, 16, 0, 0, 0]
class = drugB
Cholesterol <= 0.5
gini = 0.498
samples = 34
value = [0, 0, 16, 18, 0]
class = drugX
gini = 0.0
samples = 36
value = [0, 0, 0, 36, 0]
class = drugX
gini = 0.0
samples = 16
value = [0, 0, 16, 0, 0]
class = drugC
gini = 0.0
samples = 18
value = [0, 0, 0, 18, 0]
class = drugX