Healthcare Chatbot with Artificial Intelligence for Medical Diagnosis

Overview of the Model

Loading

Loading the csv datasets

Data Preprocessing

To find input features and labels

Data Splitting

Split the data into train and test

Decision Tree Classification

Define and train Decision tree Classifier

Binary Search in Tree

This algorithm is used to search for the disease based on user details.

Print disease to the user

Map the classifier output to the appropriate label.

Using Decision Tree
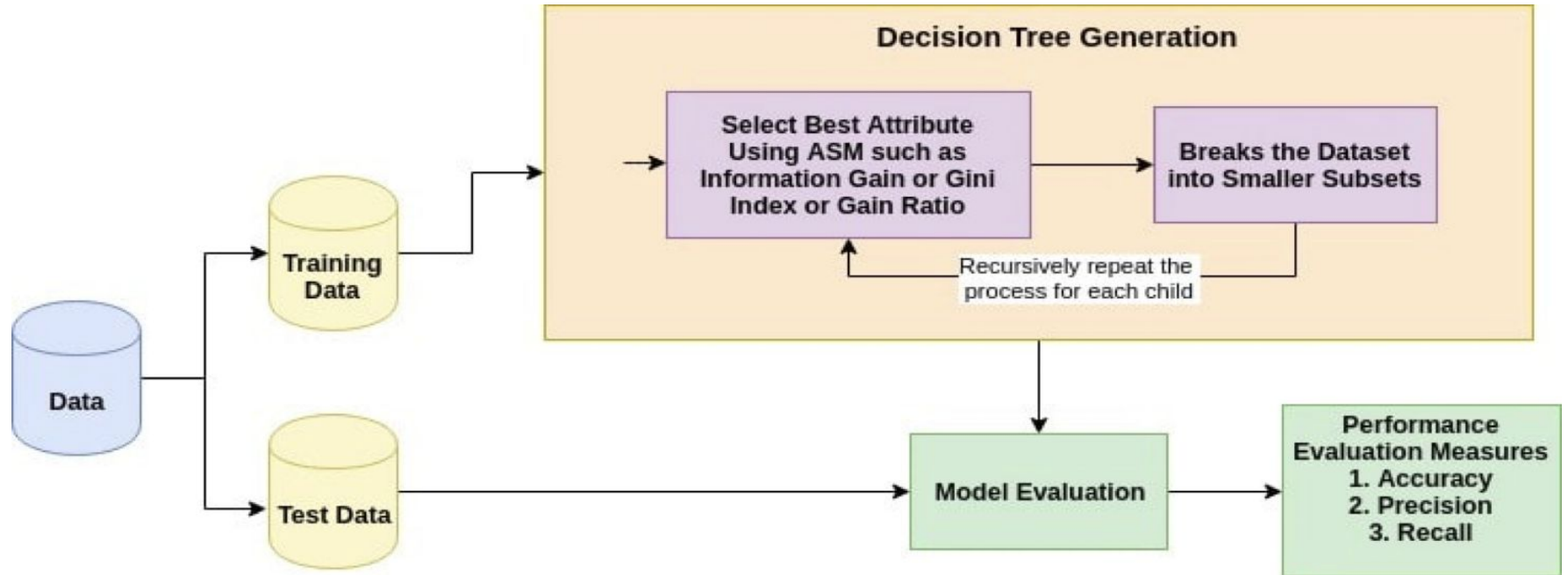
Factors to Consider

Information gain:The decrease in entropy is known as information gain. Based on specified attribute values, information gain computes the difference between entropy before split and average entropy after split of the dataset. (In a system, entropy is the randomness)

$$\text{Info(D)} = - \sum_{i=1}^{m} pi \log_2 pi$$

Gini Index :The Gini Index takes each attribute into account as a binary split. A weighted total of each partition's impurity can be calculated.

$$\text{Gini(D)} = 1 - \sum_{i=1}^{m} Pi^2$$

Dataset

- Training dataset : 113 features , 4920 samples
- Testing dataset: 113 features, 41 samples
- As a label, the word "prognosis" is utilized.
- Label Encoder is a program that converts labels into numerical values.

Before Label Encoder	After Label Encoder
Migraine	1
Heart Attack	2
Typhoid	3
Acne	4

Cross Validation, Testing and Training

clf1 = DecisionTreeClassifier()

clf2 = clf1.fit(x_train,y_train)

print(clf2.score(xtest,ytest))

Cross-validation:
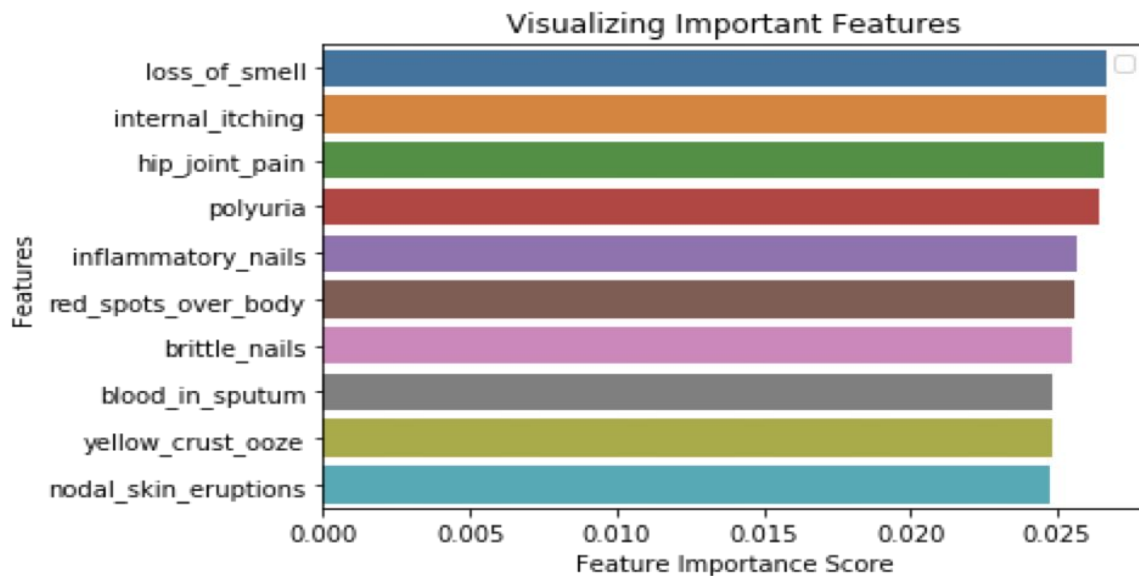
scr = cross_val_score(clf2, x_test, y_test, cv=3)

Print(scr.mean())

Results

Accuracy of training: 0.97

Accuracy of testing:0.94
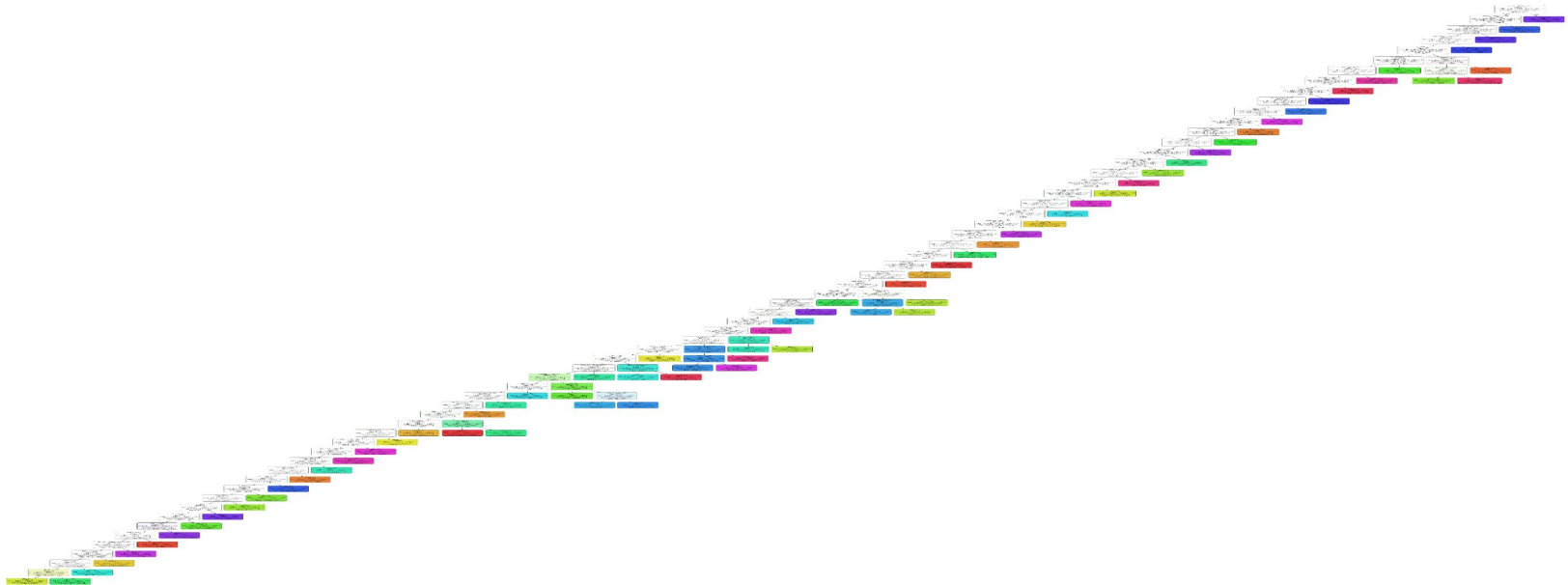
Important Features

Visualizing Important Features

Decision tree

Important Function(code)

Binary search in tree(): A Decision_tree_bot() subfunction that does binary search based on user responses to the bot's questions.
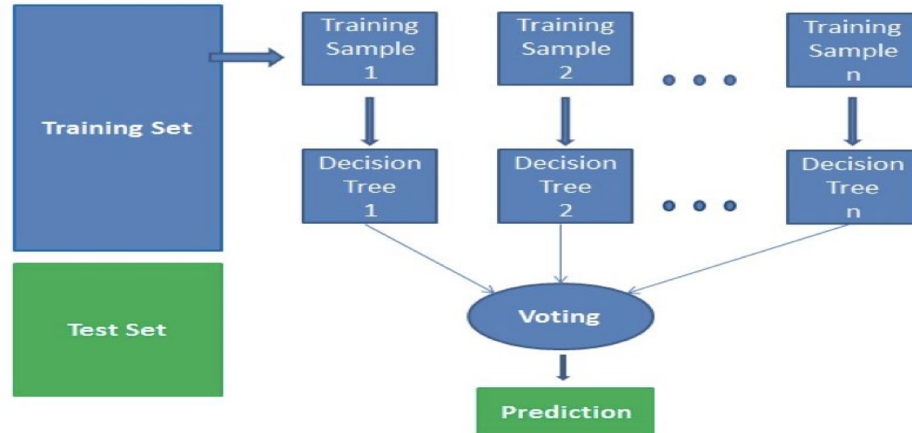
Print disease to user(): This function converts the decision tree output into one of the user's diseases.

Decision tree bot(): The Decision Tree Primary Logic

Main() : When the program is run, the main() function is executed.

Future Enhancements

As Decision Trees overfit noisy data, ensemble techniques such as Random Forest can be applied. Ensemble approaches, on the other hand, necessitate extremely large datasets.

thank you