Data Mining

Assignment-2

Introduction

Decision tree:

Decision tree methodology is the development of prediction algorithms for a target variable, or the establishment of classification systems based on numerous variables are two prominent applications of the decision tree technique.

Naïve Bayes Algorithm:

Naive Bayes employs a comparable technique to forecast the likelihood of various classes based on various attributes. This approach is mostly employed in text categorization and when dealing with issues involving several classes.

The simple form of the calculation for Bayes Theorem is as follows:

$$P(Class/Data) = P(Data/Class) * P(Class)/P(Data)$$

Naïve Bayes is one of the fast and easy data mining algorithms to predict a class of datasets. It can be used for binary as well as multiclass Classification. It performs well in multi-class predictions s compared to

the other algorithms. It is the most popular choice for text classification problems.

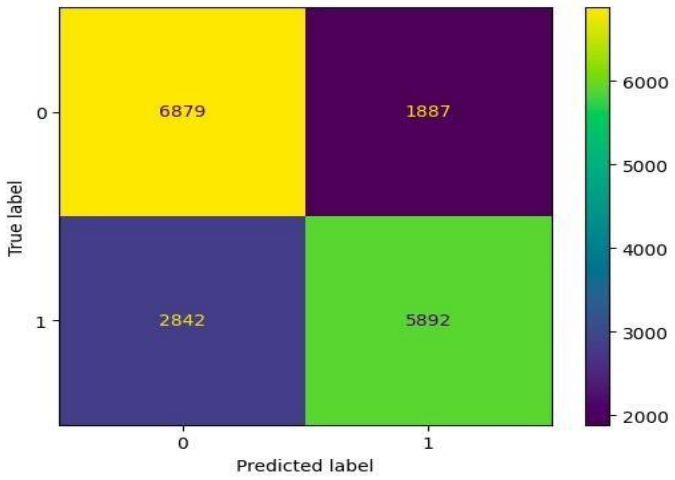
A generative model is naive Bayes. (Gaussian) According to Naive Bayes, each class has a Gaussian distribution. Naive Bayes (Gaussian) implies feature independence; hence the covariance matrices are diagonal matrices.

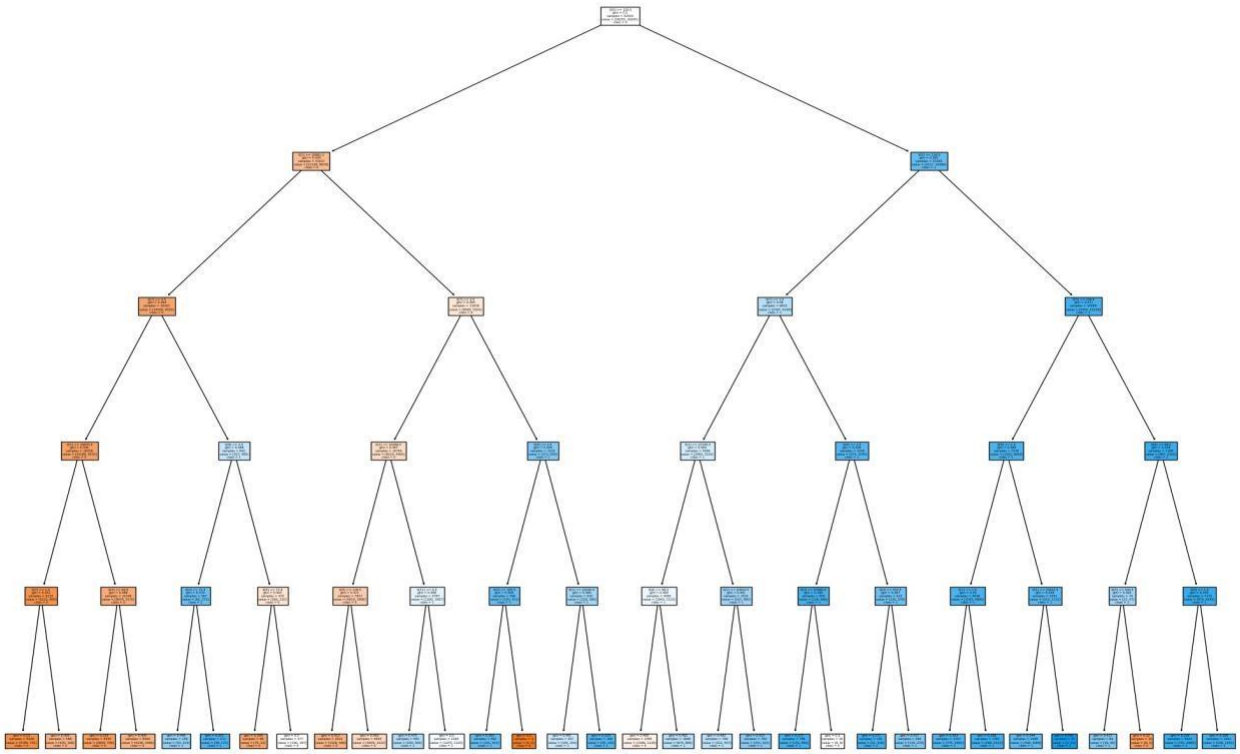
Describing dataset:

The given data was taken during the moment of medical examination. The data is focused on the whether a person's external (Age, height, weight, gender), internal (cholesterol, systolic blood pressure, diastolic blood pressure, glucose) and other (smoking, alcohol intake, physical activity) characteristics determine the presence of cardiovascular disease.

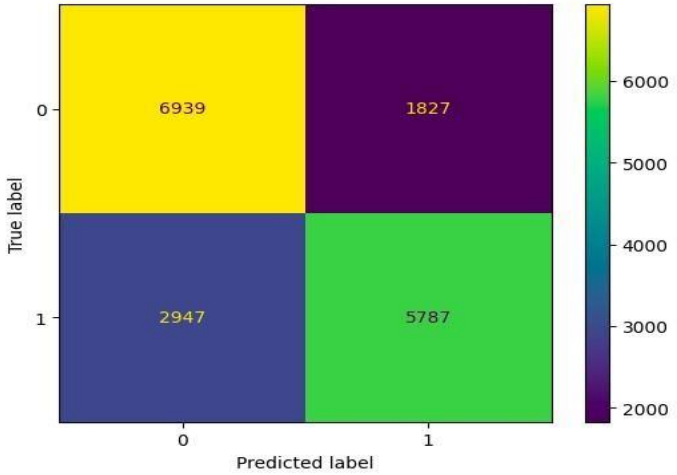
Visualization of Decision Tree:

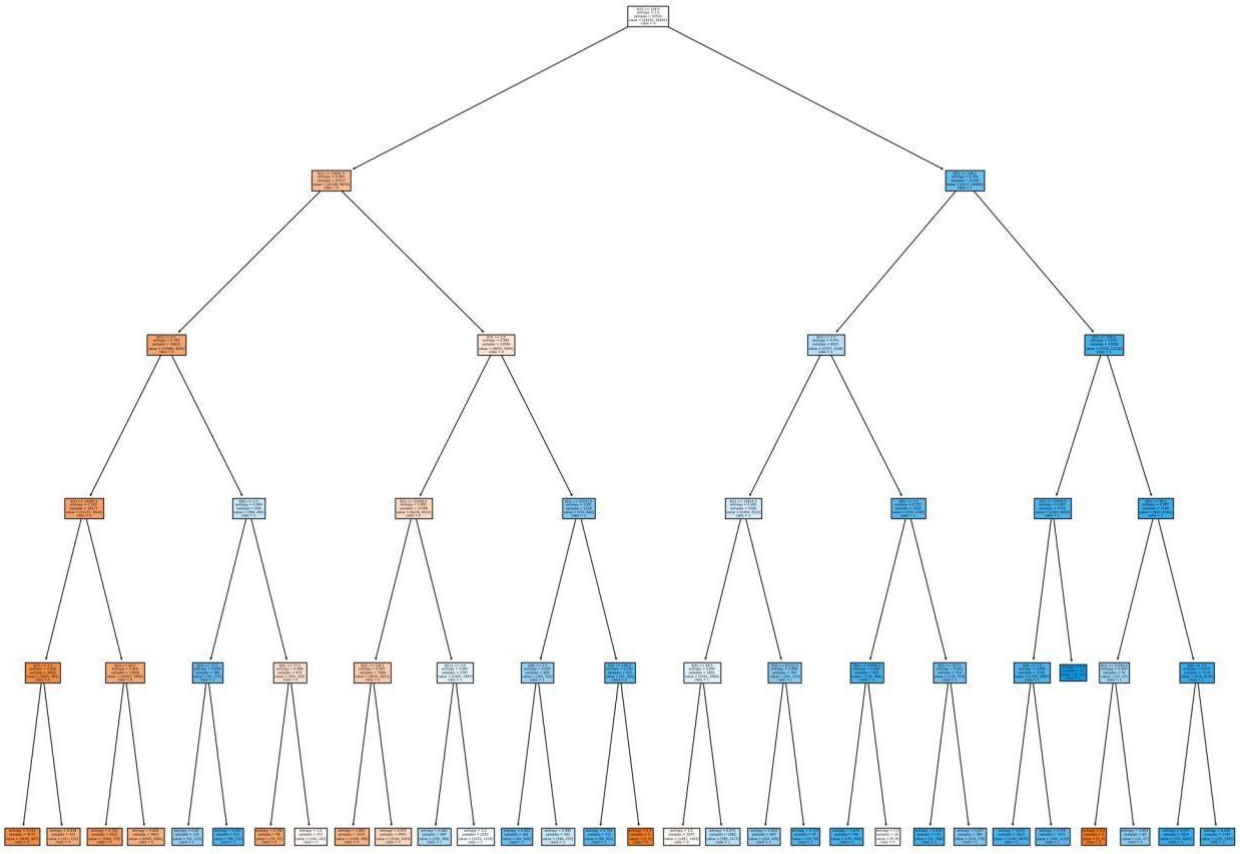
For Gini


6879
1887
2842
5892
0
1
True label
0
1
Predicted label
6000
5000
4000
3000
2000



For Entropy





Comparing Gini and Entropy:

In Gini, we have the most effecting variables when compared with target variable is age, systolic blood pressure, id respectively.

In Entropy, we have the most effecting variables when compared with target variable is age, id, systolic blood pressure respectively.

Here we can observe that the most effective variables differ from one another in gini and entropy, it can also be seen in the figures provided below.

Comparing the result of DT and NB:

With Accuracy:

Gini

```
Accuracy: 0.7297714285714286
```

Entropy

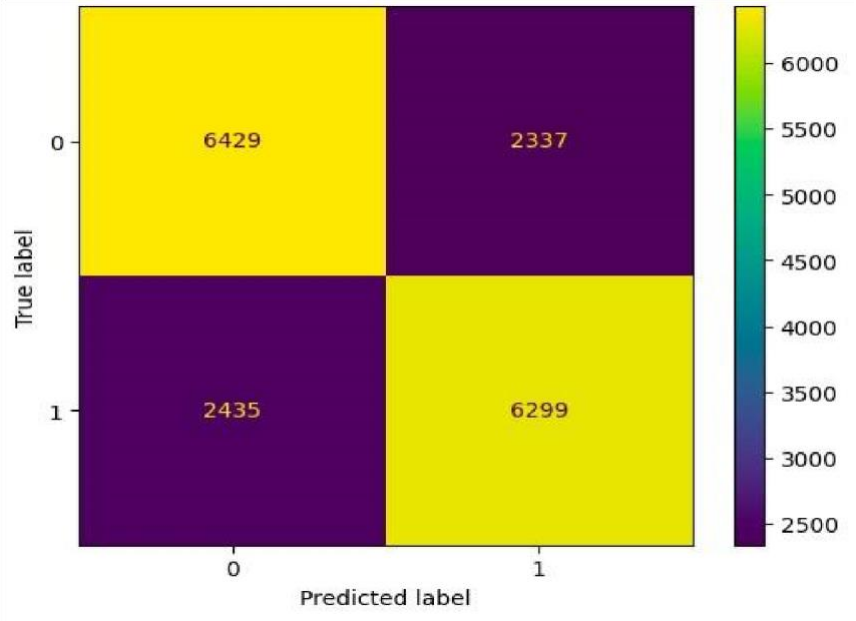
```
Accuracy: 0.7272
```

Naïve bayes

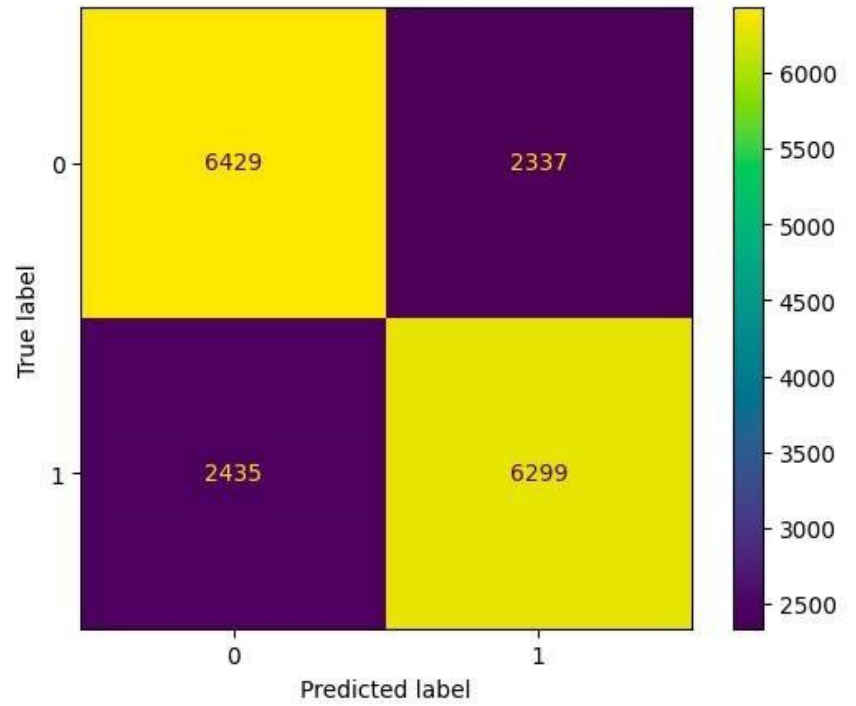
```
Accuracy: 0.5630857142857143
```

With Confusion matrix:

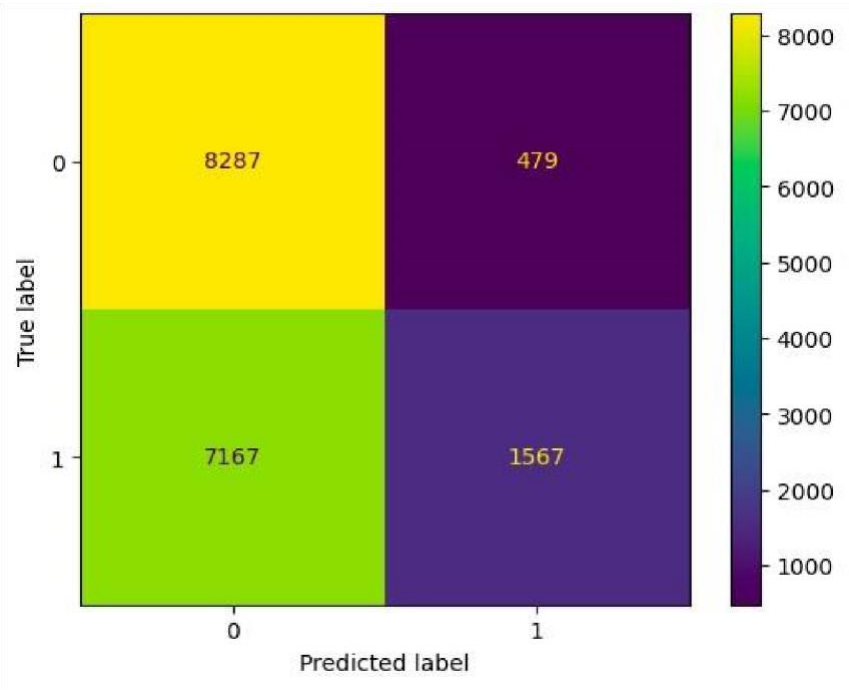
Gini



Entropy



Naïve bayes



Visualization of Decision Tree:

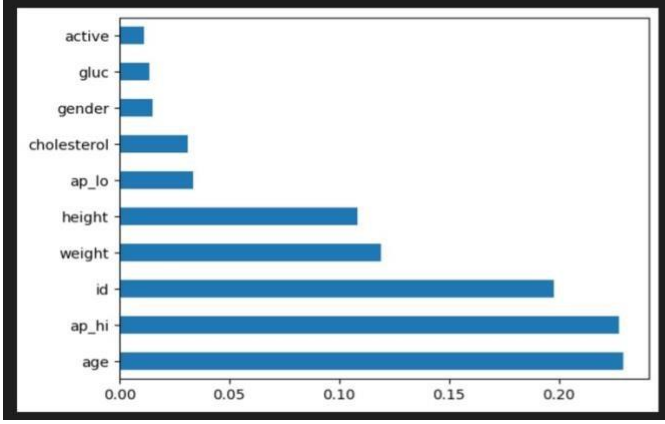
For Gini

```python
#1) Using gini to measure the quality of a split [2 points]

X = data.drop(['cardio'],axis=1)
y = data['cardio']
model = DecisionTreeClassifier(criterion='gini')
model.fit(X , y)
feature_imp = pd.Series(model.feature_importances_ , index = X.columns)
feature_imp.nlargest(10).plot(kind = 'barh')
```

✓ 0.6s Python

```
<AxesSubplot: >
```



For Entropy

```python
#1) Using entropy to measure the quality of a split [2 points]

X = data.drop(['cardio'],axis=1)
y = data['cardio']
model = DecisionTreeClassifier(criterion='entropy')
model.fit(X , y)
feature_imp = pd.Series(model.feature_importances_ , index = X.columns)
feature_imp.nlargest(10).plot(kind = 'barh')
```

✓ 0.8s Python

```
<AxesSubplot: >
```