HACETTEPE UNIVERSITY

DEPARTMENT OF ELECTRICAL & ELECTRONICS ENGINEERING

ELE489 Fundamentals Of Machine Learning

Hasan KARAŞALLI / 2210357107

Homework 2

Report

GitHub Link: https://github.com/KarasalliHasan/ELE489_HW2.git

Q1)

To solve the question,the given steps can be followed:

1. **Calculate the Gini index for the entire dataset.**

There are two possible output variables for “Playing outside?”: **Yes** and **No**

The data has 3 instances of **Yes** and 3 instance of **No.**The Gini index for the entire dataset can be calculated as: $Gini_{overall}(S) = 1 - \left(\left(\frac{3}{6}\right)^2 + \left(\frac{3}{6}\right)^2\right) = 0.5$

2. **Compute the weighted Gini index for splitting on weather.**

It has three possible values of **Sunny**(2 examples), **Overcast**(2 examples) and **Rainy**(2 examples).

For Weather = Sunny, there are 2 examples of **No:** $Gini_{sunny}(S) = 1 - \left(\left(\frac{2}{2}\right)^2\right) = 0$

For Weather = Overcast there are 2 examples of **Yes:** $Gini_{overcast}(S) = 1 - \left(\left(\frac{2}{2}\right)^2\right) = 0$

For Weather = Rainy, there are 1 **No** and 1 **Yes** example**:** $Gini_{rainy}(S) = 1 - \left(\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2\right) = 0.5$

Weighted Gini index of weather: $Gini_{weather}(S) = \left(0\left(\frac{2}{6}\right) + 0\left(\frac{2}{6}\right) + 0.5\left(\frac{2}{6}\right)\right) = \mathbf{0.167}$

3. **Compute the weighted Gini index for splitting on wind.**

It has two possible values of **Weak**(3 examples) and **Strong**(3 examples).

For Wind = Weak, there are 1 **No** and 2 **Yes** examples**:** $Gini_{weak}(S) = 1 - \left(\left(\frac{2}{3}\right)^2 + \left(\frac{1}{3}\right)^2\right) = 0.444$

For Wind = Strong there 2 **No** and 1 **Yes** examples**:** $Gini_{strong}(S) = 1 - \left(\left(\frac{1}{3}\right)^2 + \left(\frac{2}{3}\right)^2\right) = 0.444$

Weighted Gini index of wind: $Gini_{wind}(S) = \left(0.444\left(\frac{3}{6}\right) + 0.444\left(\frac{3}{6}\right)\right) = \mathbf{0.444}$

4. **Choose the feature with the lowest weighted Gini index as the root node.**

The weighted Gini index of Weather is 0.167 while the weighted Gini index of wind is 0.444.

Since the weighted Gini index of weather is smaller,it is chosen as the root node.

Q2

1.

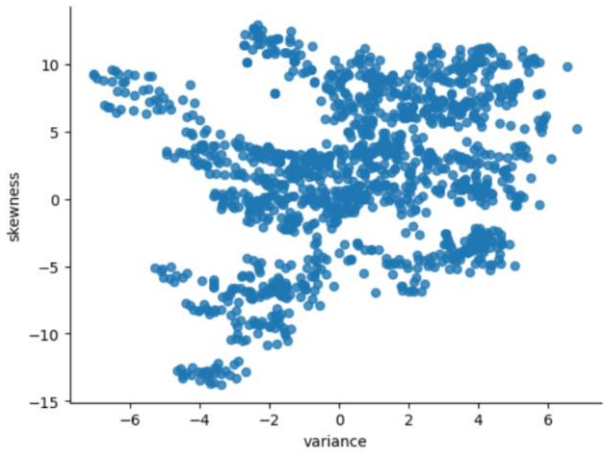
- ***Variance:*** Measures the spread of pixel intensity values around the mean.It can be found by calculating the distance between the pixel intensity and mean of the image.It gives the information about the contrast of the image(higher variance means more contrast).

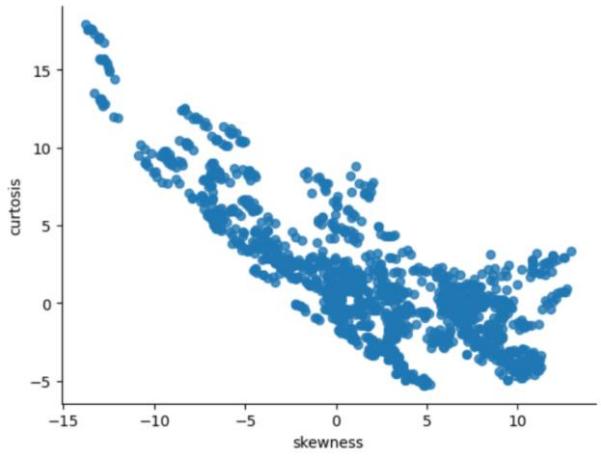
- ***Skewness:*** Measures the asymmetry of the pixel intensity distribution.It is equal to a zero value if the image has symmetric distribution.It can be computed by using the mean and standard deviation informations of the image.

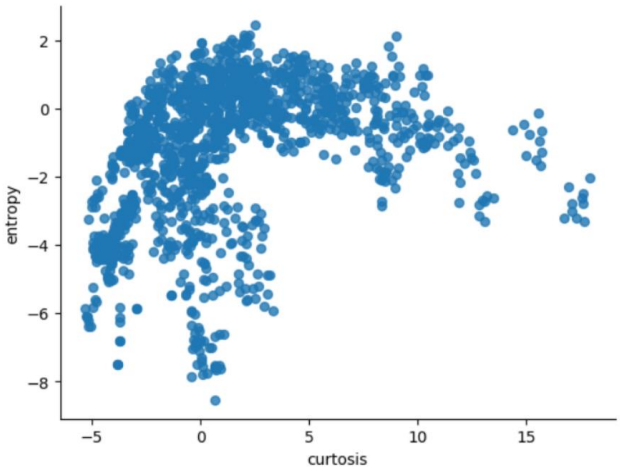
- ***Kurtosis:*** Measures the "tailedness" of the intensity distribution.It reflects the sharpness of the peak and the presence of extreme pixel values in an image.It can be computed with mean and standard deviation values of the image.

- ***Entropy:*** Measures the randomness of the image.It can be computed with the probability of the intensity level.It indicates whether the image has uniform distribution(mostly black or white) or has richer detail and more information.

2. Visualization of the features in group of two







Q:*Do you think the decision tree is a good algorithm to run for these features?*

Comments:

Yes,the decision tree is a good algorithm to run these features because the relationships between features are not linear, and decision trees can handle the non-linear decision boundaries successfully. Also the decision trees dont need feature scaling like normalization,it can handle the process without it and makes the analysis easier.

The decision trees are already used in classification and specifically in decision analysis.

3.

The effects of the different values of ***max_depth,min_samples_split*** *and* ***criterion*** *can be seen from the table given below:*

	Criterion	max_depth	min_samples_split	Accuracy
24	entropy	6.0	2	0.989091
15	entropy	NaN	2	0.989091
27	entropy	9.0	2	0.989091
21	entropy	5.0	2	0.985455
22	entropy	5.0	5	0.985455
23	entropy	5.0	8	0.981818
25	entropy	6.0	5	0.981818
16	entropy	NaN	5	0.981818
28	entropy	9.0	5	0.981818
17	entropy	NaN	8	0.978182
29	entropy	9.0	8	0.978182
26	entropy	6.0	8	0.978182
0	gini	NaN	2	0.967273
1	gini	NaN	5	0.967273
13	gini	9.0	5	0.967273
12	gini	9.0	2	0.967273
11	gini	6.0	8	0.967273
2	gini	NaN	8	0.967273
14	gini	9.0	8	0.967273
10	gini	6.0	5	0.967273
9	gini	6.0	2	0.963636
20	entropy	3.0	8	0.960000
18	entropy	3.0	2	0.960000
19	entropy	3.0	5	0.960000
8	gini	5.0	8	0.952727
6	gini	5.0	2	0.952727
7	gini	5.0	5	0.952727
4	gini	3.0	5	0.949091
5	gini	3.0	8	0.949091
3	gini	3.0	2	0.949091

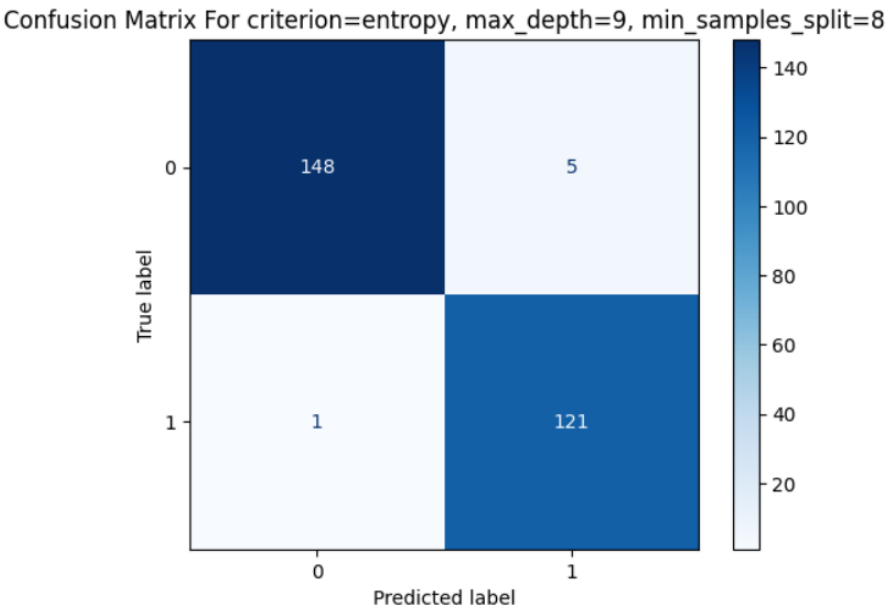
The following inferences can be made from the table above:

- *Entropy criterion's accuracy is always higher than the Gini criterion's when the other parameters are kept same.It means that Entropy criterion gives higher performance.*
- *By looking the top 4 rows,for same criterion and min_sample_split values,the values of max_depth,None-6-9,reaches to same accuracy result but the max_depth=5 has slightly less accuracy then the others.*
  It means that the decision tree model can be modeled with max_depth=6 value even there is no limitation.(It can be verified from the result of plot_tree() function below)
- *By looking 20-18-19 rows,the depth has huge impact on accuracy of the decision tree.For the max_depth=3 value,the models with entropy criterion falls behind the models with gini criterion.*
- *min_sampler_split can decrease the accuracy if it has higher values.The rows 21 and 22 has same accuracy with different min_sampler_split values(2-5) but the accuracy change for the higher min_sampler_split value(row23) when the other parameters are equal to each other.*

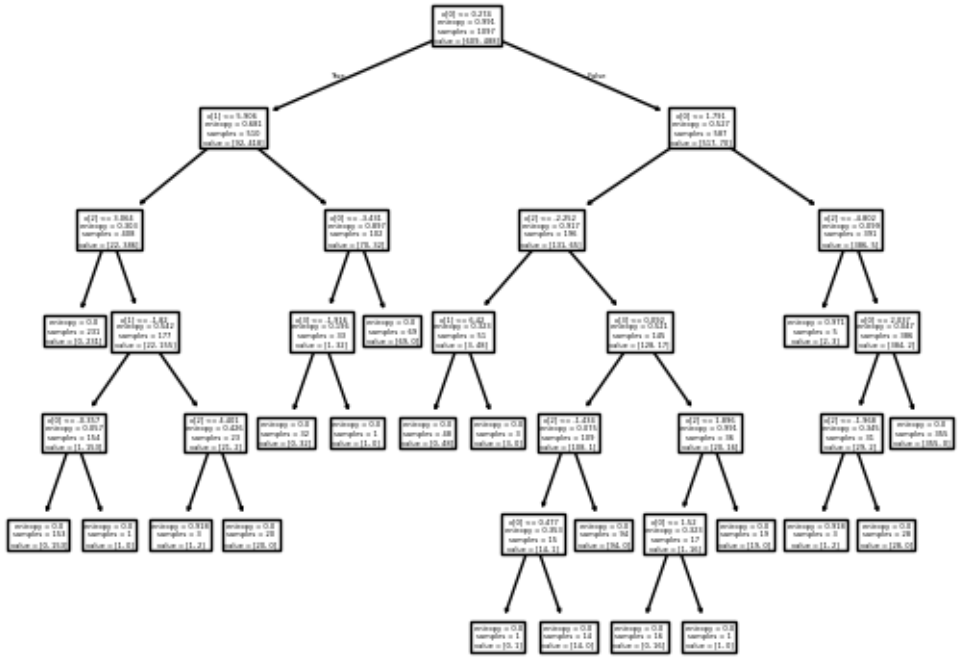
Classification Report:

```
Classification Report For criterion=entropy, max_depth=9, min_samples_split=8:
              precision    recall  f1-score   support

           0       0.99      0.97      0.98       153
           1       0.96      0.99      0.98       122

    accuracy                           0.98       275
   macro avg       0.98      0.98      0.98       275
weighted avg       0.98      0.98      0.98       275
```

Confusion Matrix:

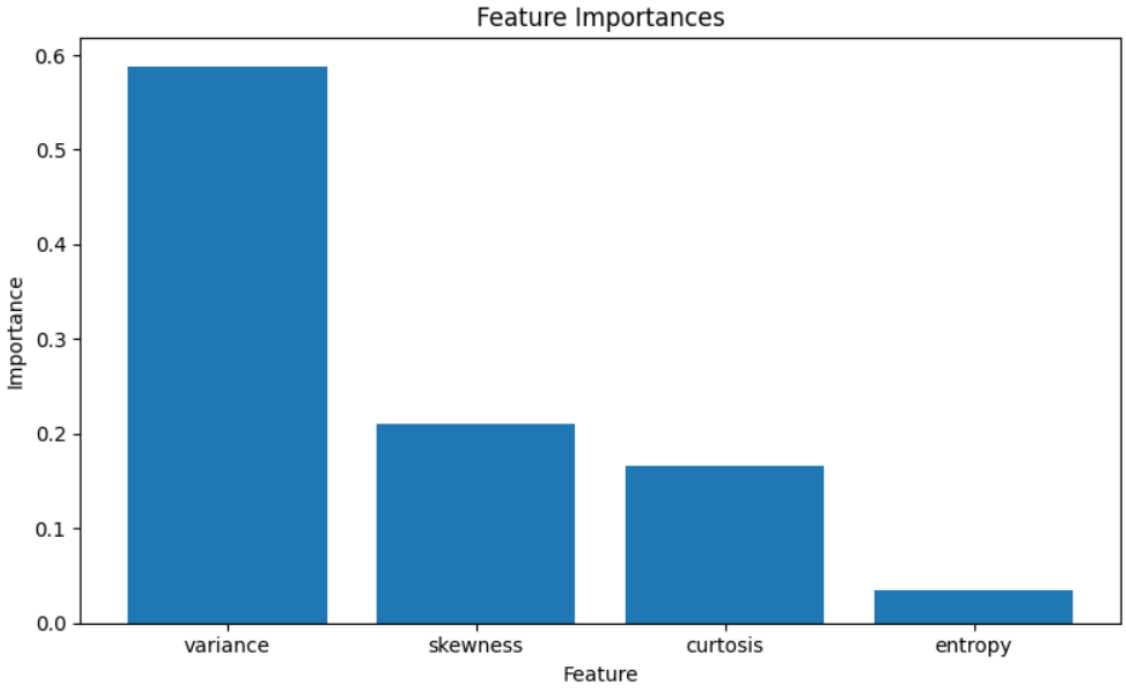


4.



As the tree depth increases, the model becomes harder to interpret due to the increasing number of the branches and it gets more complicated toward to leaf nodes.

5.



The variance is most important feature while the entropy is least important in the training of the decision tree model according to Feature Importance plot.

6.

I learned the terms used in the image processing such as variance, skewness, kurtosis, and entropy. I also gained a deeper knowledge of the decision tree algorithm and how its parameters affects the model's performance.

Additionally, by analyzing the feature_importances_ plot, I learned that not all features contribute equally to a model and some features more important than the others.

I think the decision tree is a good model for this dataset because it performed well for evaluation metrics like the confusion matrix, classification report, and accuracy score.

The accuracy ranged between 94% and 98% which means that model can distinguish the real and fake banknotes succesfully.