HACETTEPE UNIVERSITY

ELECTRICAL AND ELECTRONICS ENGINEERING

ELE 489-Fundamentals of Machine Learning

Homework 2

Akanay Boran Özcan

2200357021

Question 1:

Class 0 : Not playing outside

Class 1 : Playing outside

Gini index Formula : $1 - \sum_{i=0}^{C-1} P_i(t)^2$

Class 0 : 3 records

Class 1 : 3 records $\quad P(0) = \frac{1}{2}$, $P(1) = \frac{1}{2}$

$\text{Gini (Dataset)} = 1 - \left(\frac{1}{2}^2 + \frac{1}{2}^2\right) = \frac{1}{2} = 0.5$

Weather

Sunny (2 records)

Class (0) : 2 $\quad P(0) = 1$

Class (1) : 0 $\quad P(1) = 0$

$\text{Gini(Sunny)} = 1 - (1^2 + 0^2) = 0$

Overcast (2 records)

class (0) : 0 $\quad P(0) = 0$

class (1) : 2 $\quad P(1) = 1$

$\text{Gini (Overcast)} = 1 - (0^2 + 1^2) = 0$

Rainy (2 records)

Class 0 : 1 $\quad P(0) = \frac{1}{2}$

Class 1 : 1 $\quad P(1) = \frac{1}{2}$

$\text{Gini(rainy)} = 1 - \left(\frac{1}{2}^2 + \frac{1}{2}^2\right) = \frac{1}{2} = 0.5$

Wind

weak (3 records)

Class 0 : 1 $\quad P(0) = \frac{1}{3}$

Class 1 : 2 $\quad P(1) = \frac{2}{3}$

$\text{Gini (weak)} = 1 - \left(\frac{1}{3}^2 + \frac{2}{3}^2\right) = \frac{4}{9}$

Strong (3 records)

Class 0 : 2 $\quad P(0) = \frac{2}{3}$

Class 1 : 1 $\quad P(1) = \frac{1}{3}$

$\text{Gini (Strong)} = 1 - \left(\frac{2}{3}^2 + \frac{1}{3}^2\right) = \frac{4}{9}$

$\text{weighted - Gini (weather)} = \frac{1}{3} \times 0 + \frac{1}{3} \times 0 + \frac{1}{3} \times 0.5 = \frac{1}{6}$

$\text{weighted - Gini (wind)} = \frac{1}{2} \cdot \frac{4}{9} + \frac{1}{2} \cdot \frac{4}{9} = \frac{4}{9}$

Question 2:

1-)

Variance: Measures the spread or dispersion of pixel values in an image. It gives an indication of how much the pixel values deviate from the mean value of the image. High variance means a more complex image with greater differences in pixel values, while low variance indicates a simpler, more uniform image.

Skewness: Measures the asymmetry of the image's histogram of pixel intensities. It tells you whether the distribution of pixel values is skewed towards the lower or higher end of the intensity scale.

Kurtosis: It tells you how extreme the outliers are in the image. High kurtosis means the image has extreme pixel values (outliers), while low kurtosis means the image has fewer extreme pixel values.

Entropy: Measures the randomness or disorder in the pixel intensity distribution of an image. It gives an indication of the amount of information or unpredictability in the image. A higher entropy means the image contains more details or randomness (like a noisy image), while a lower entropy suggests the image is more uniform or predictable.

2-)

I downloaded the dataset and visualized it as pandas data frame.

You can see the data frame in Figure 1.

	Variance	Skewness	Kurtosis	Entropy	Class
0	3.62160	8.66610	-2.8073	-0.44699	0
1	4.54590	8.16740	-2.4586	-1.46210	0
2	3.86600	-2.63830	1.9242	0.10645	0
3	3.45660	9.52280	-4.0112	-3.59440	0
4	0.32924	-4.45520	4.5718	-0.98880	0
...	...	...	...	...	...
1367	0.40614	1.34920	-1.4501	-0.55949	1
1368	-1.38870	-4.87730	6.4774	0.34179	1
1369	-3.75030	-13.45860	17.5932	-2.77710	1
1370	-3.56370	-8.38270	12.3930	-1.28230	1
1371	-2.54190	-0.65804	2.6842	1.19520	1

Figure 1.

After loading the data set into a pandas data frame, I visualized the features in groups of two. You can see the pair plot in Figure 2.

As can be seen in Figure 2 Variance feature is distinguishable than the other features in the plot.

I think it is a good decision to run these features in a decision tree algorithm because most of the features are really close to each other even inside of each other so instead of k-NN algorithm it should be a great choice.

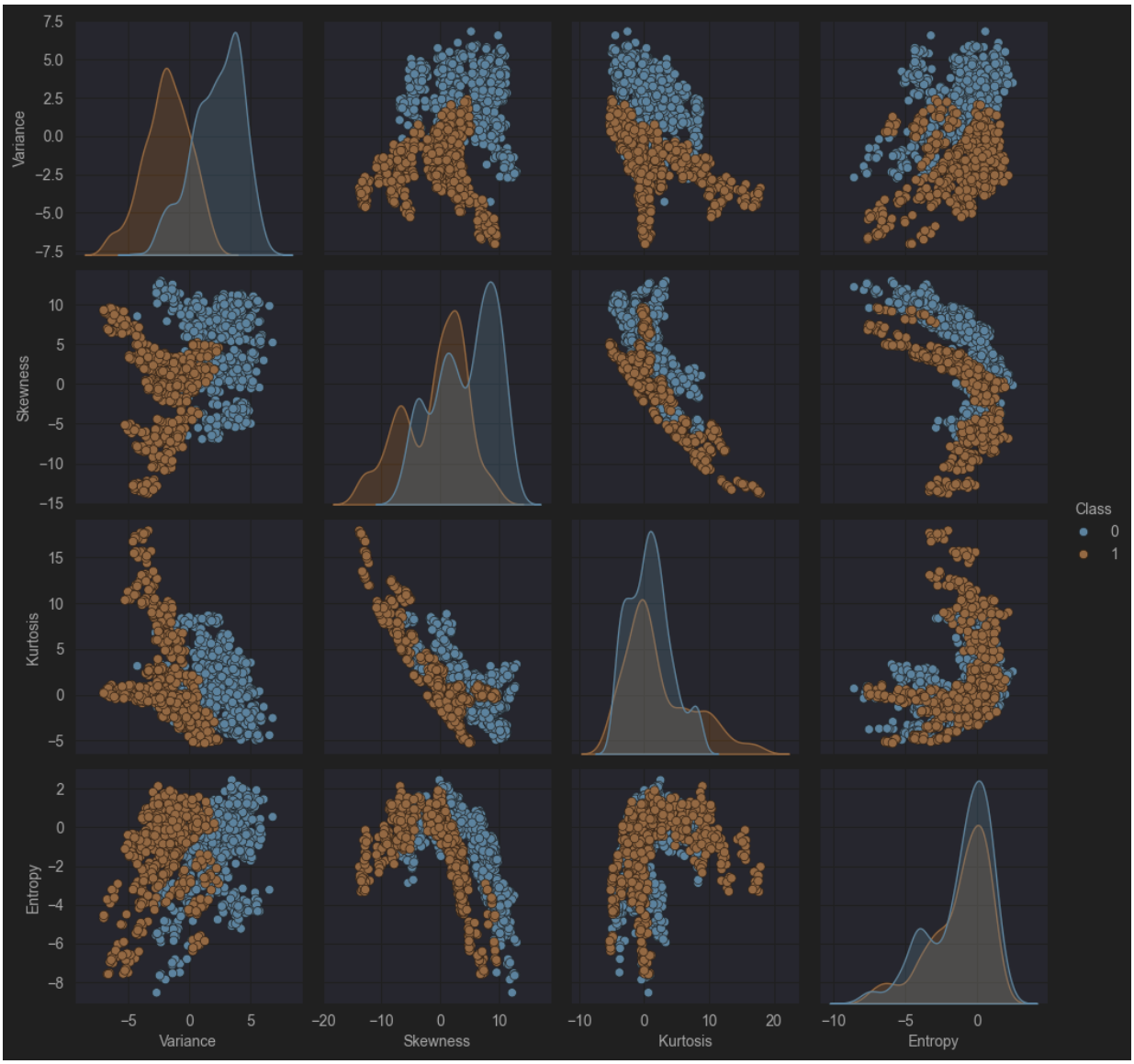

Figure 2.

3-)

Confusion Matrix:

[[155 2]

 [0 118]]

We can say from this that 2 of the real banknotes are classified as fake.

TP = 155, FP = 2, FN = 0, TN = 118.

Classification Report:

	precision	recall	f1-score	support
0	1.00	0.99	0.99	157
1	0.98	1.00	0.99	118
accuracy			0.99	275
macro avg	0.99	0.99	0.99	275
weighted avg	0.99	0.99	0.99	275

Figure 3.

4-) Using the plot_tree() function from sklearn.tree I got the tree plot, in figure 4 the tree is according to Gini index.

Entropy and Gini index is giving quite similar outputs however entropy gives the same outputs with less tree depths in this dataset.

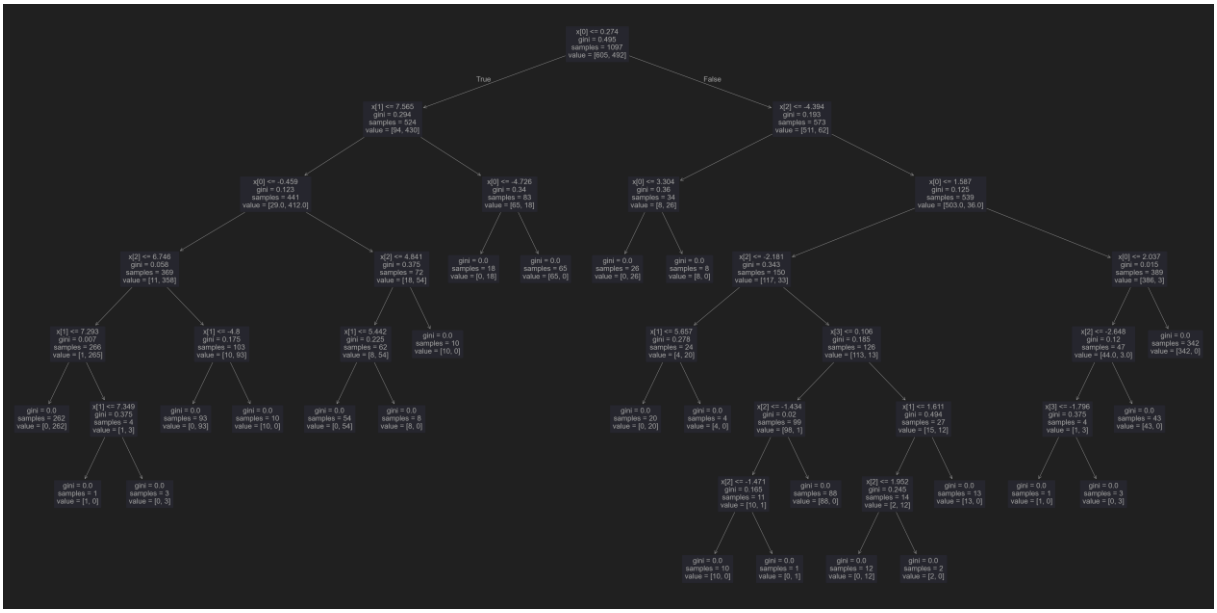

Figure 4.

5-) In figure 5 you can see the feature importance table which seems the variance have the highest importance.

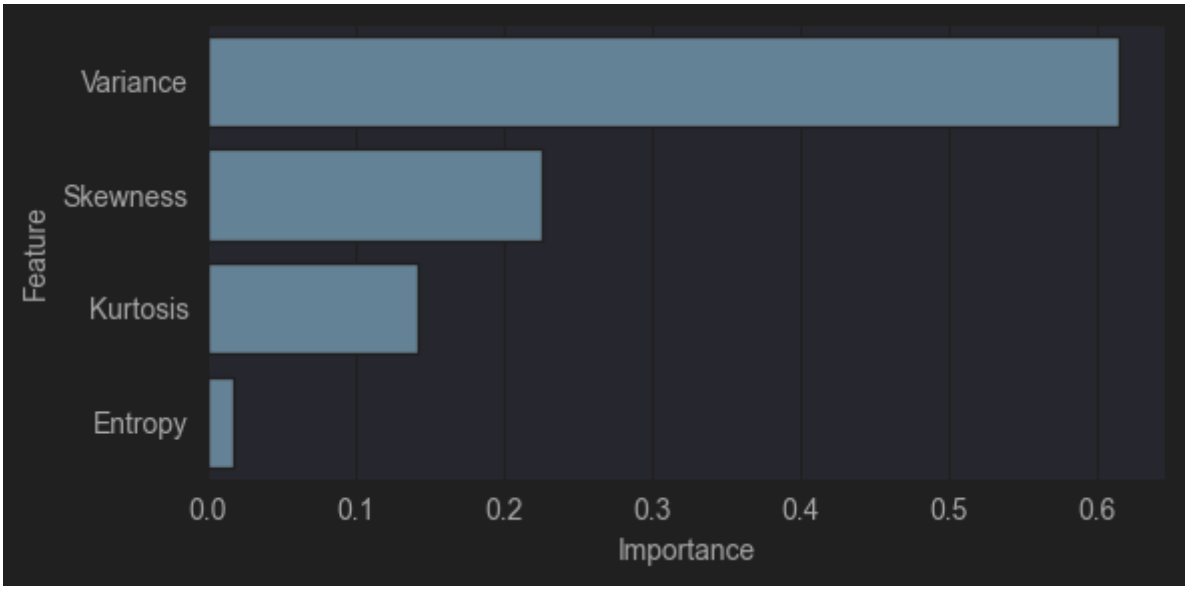

Figure 5.

6-) I learned the Gini index and entropy calculations, underfitting and overfitting, how to make pair plots. I still think the decision tree is a good model for this data because with only 7 or 8 depth we got pretty good outputs at the classification report.

My GitHub profile:

https://github.com/Akanay?tab=repositories