ELE489 – Fundamentals of Machine Learning

Homework 1 – k-NN Implementation and Analysis

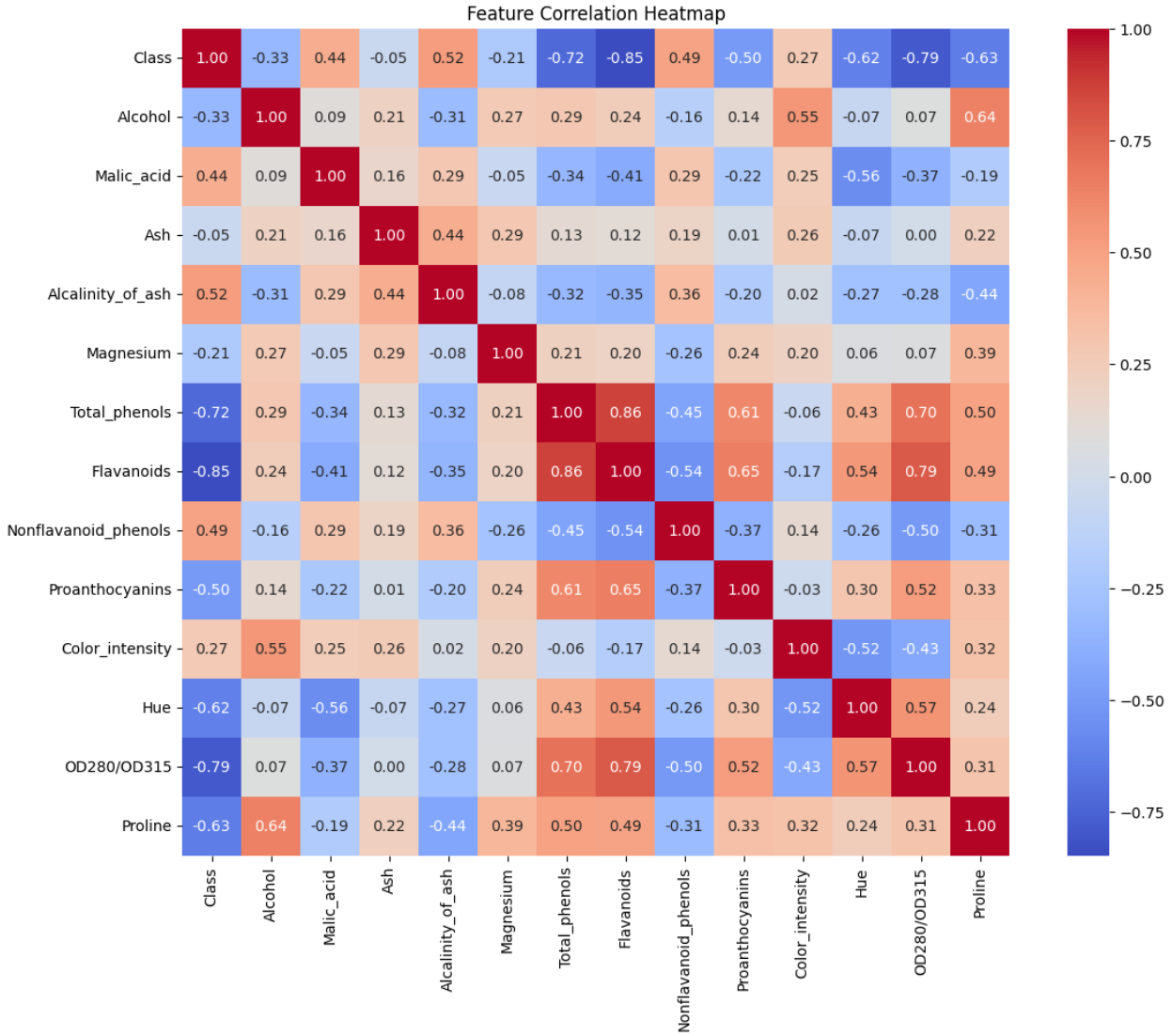
Alperen Özkan

2210357032

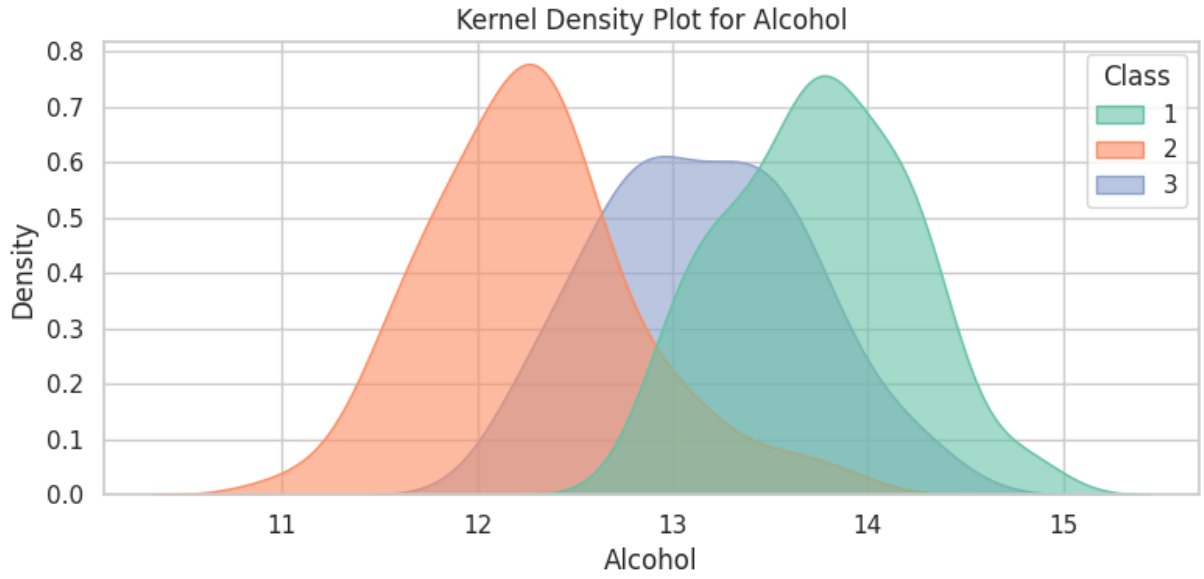
Introduction

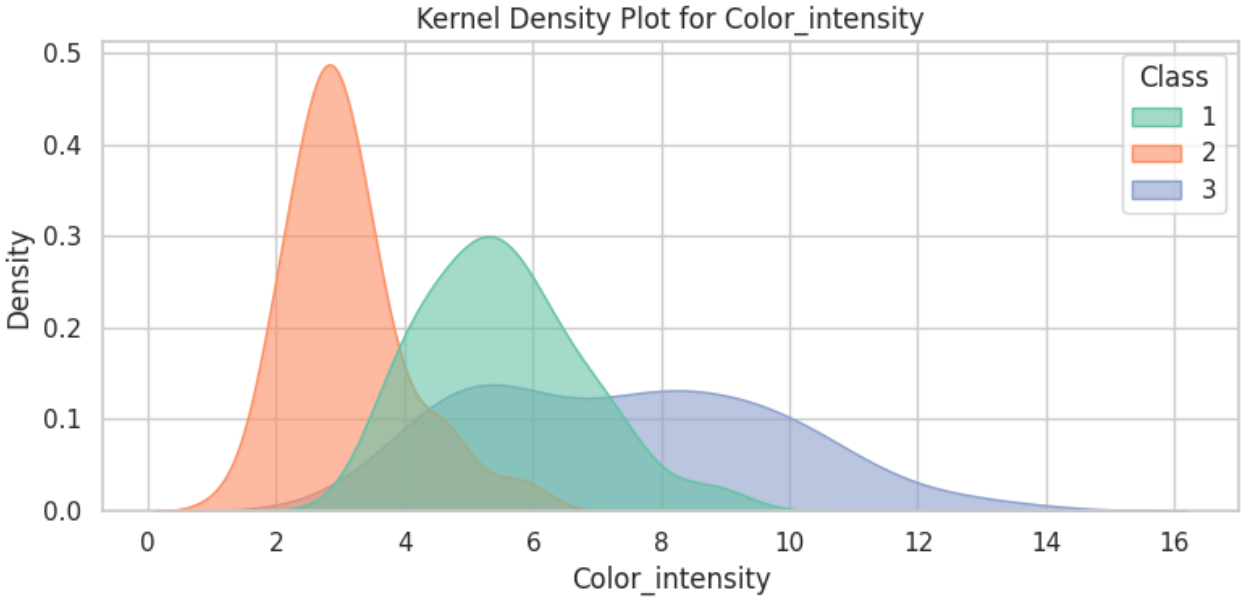
In this assignment, I implemented the k-Nearest Neighbors (k-NN) algorithm from scratch in Python using only basic libraries like NumPy. I didn't use any built-in. I tested my implementation on the Wine dataset and analyzed how different values of **K** and three distance metrics (Euclidean, Manhattan, and Chebyshev) affect the accuracy.

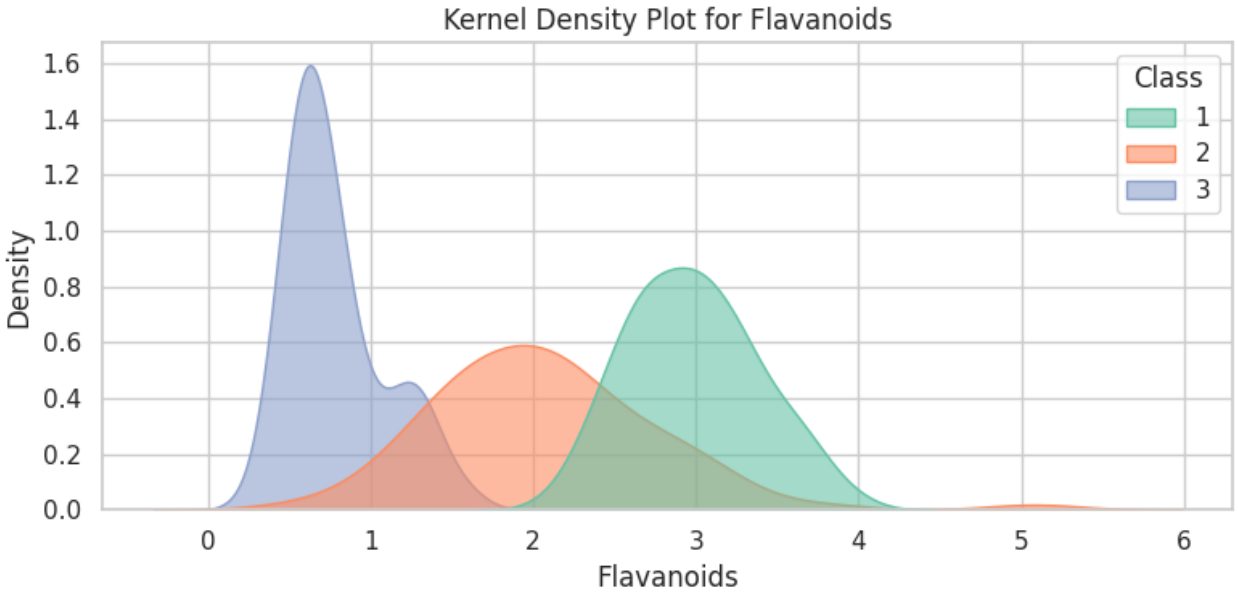
Feature Correlation Heatmap

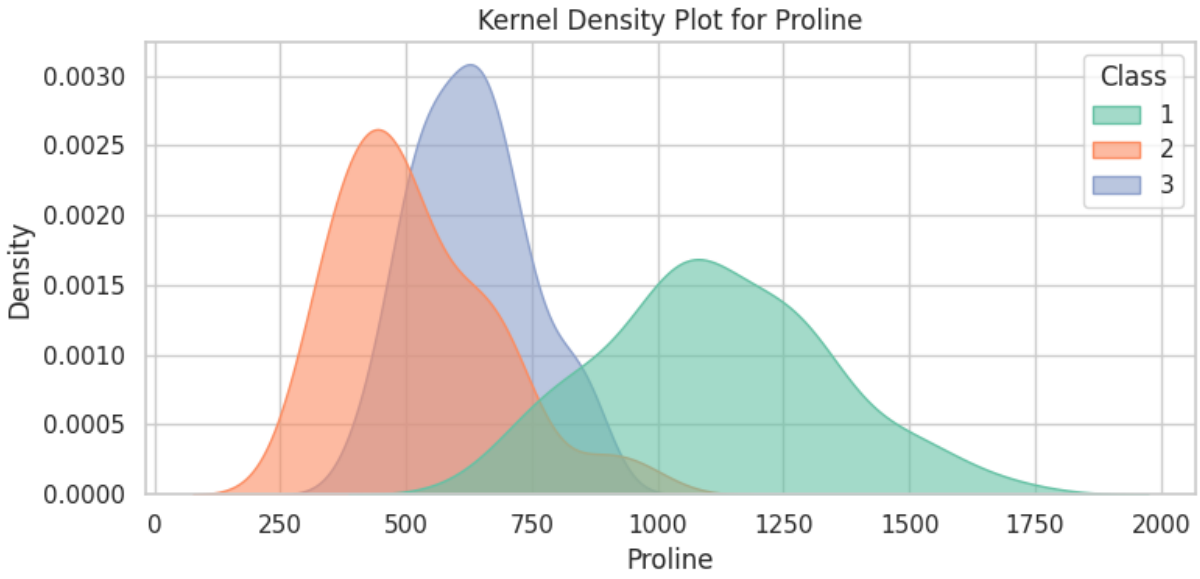


KDE and Pair Plots of Features with Some Classes



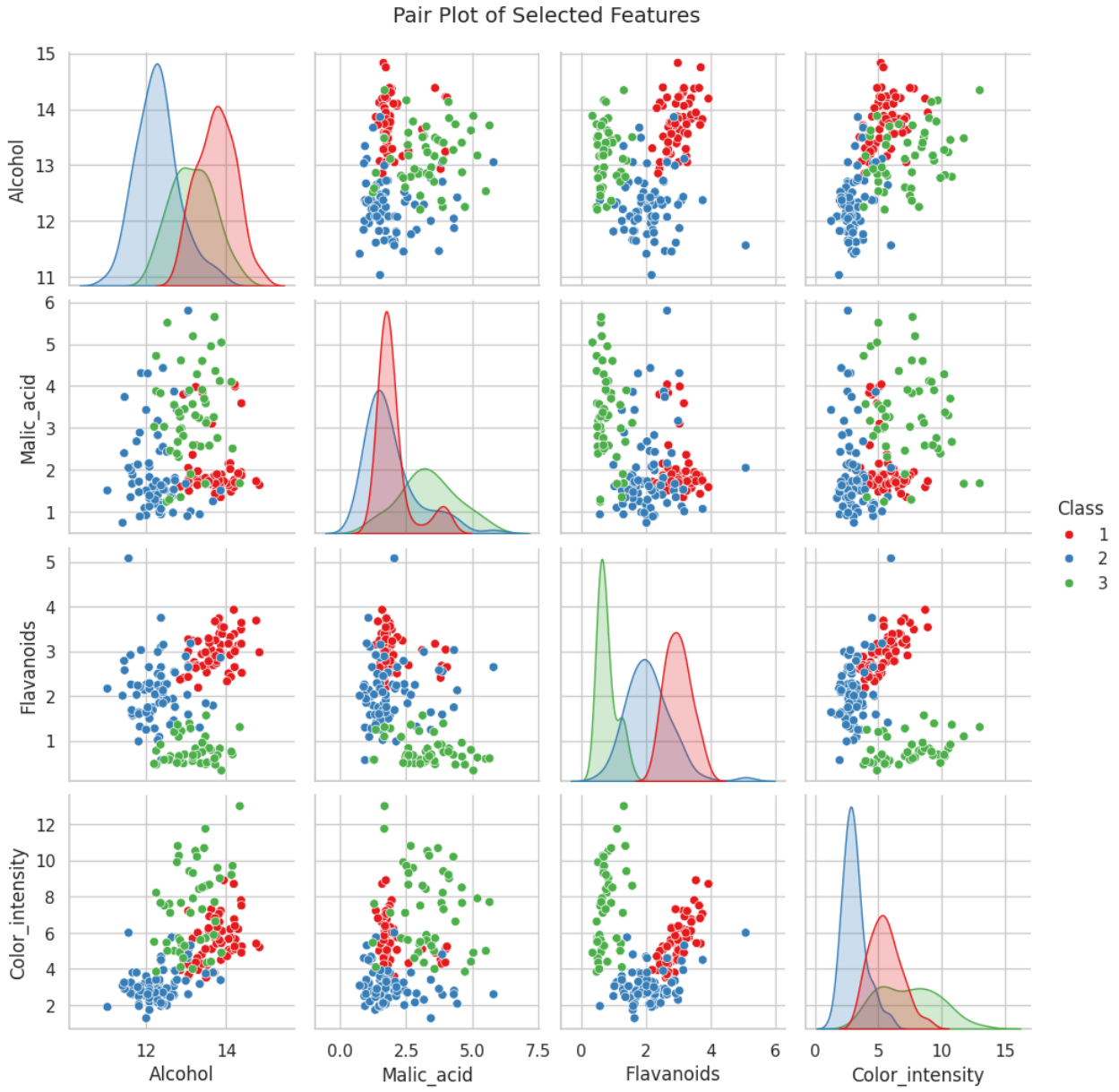




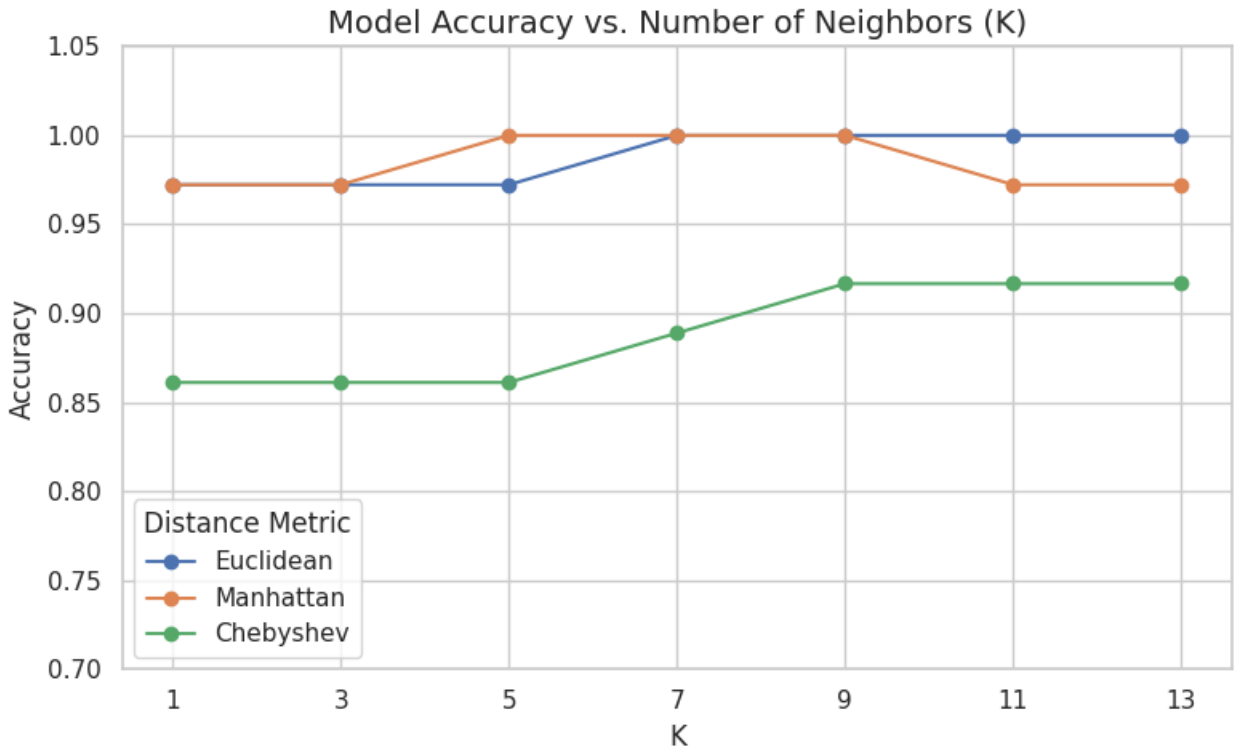


After visualizing the data using KDE and pair plots, I noticed that some features separate the classes better than others. For example, **Flavanoids**, **Alcohol**, and **Color Intensity** show clearer differences between classes, while features like **Malic Acid** have more overlap. This probably means that some features are more important for classification.

Pair Plots of Some Features



Comparison K Values

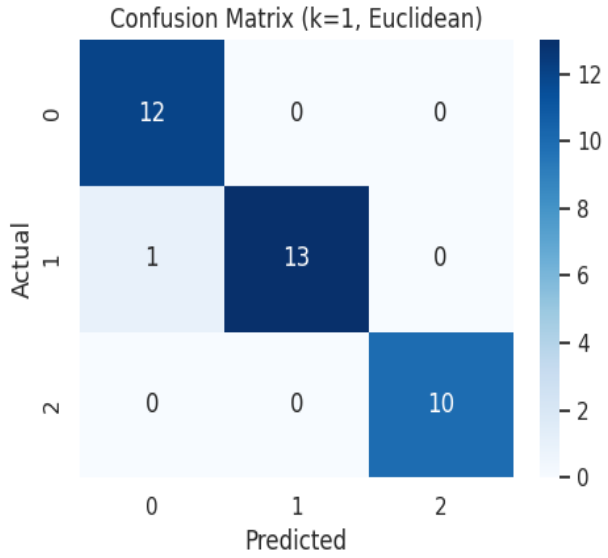


I tested different values of **K** (from 1 to 13). Based on the results:

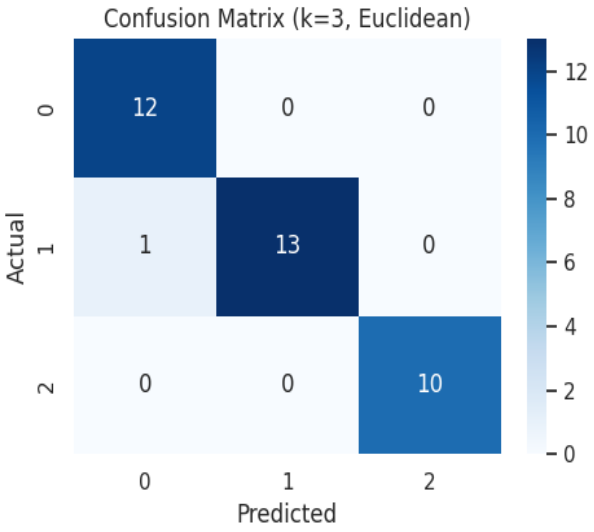
- **Euclidean and Manhattan** distances performed very similarly. Both reached **100% accuracy** between K=5 and K=9.
- **Chebyshev** distance had lower accuracy overall. It started at around **86%** and went up to about **91%** as K increased.
- For **K=1 and K=3**, all metrics performed slightly worse, probably because small K values make the model more sensitive to noise.
- After **K=9**, Manhattan accuracy started to drop a bit, while Euclidean stayed constant at 100%.

Overall, the best performance was seen at **K=7** using **Euclidean or Manhattan** distance. Chebyshev was more unstable, so it might not be the best choice for this dataset.

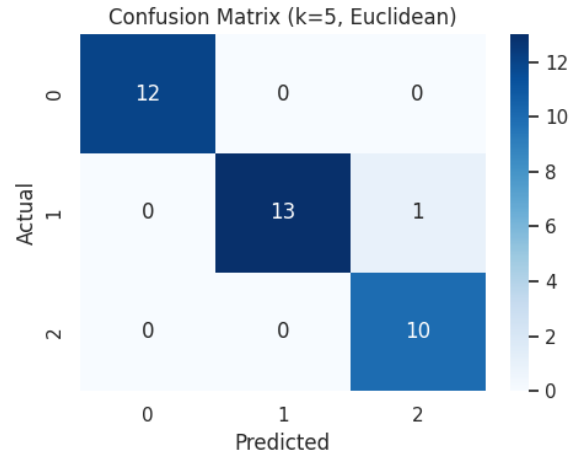
Classification Accuracy for each Value of K(Euclidean)



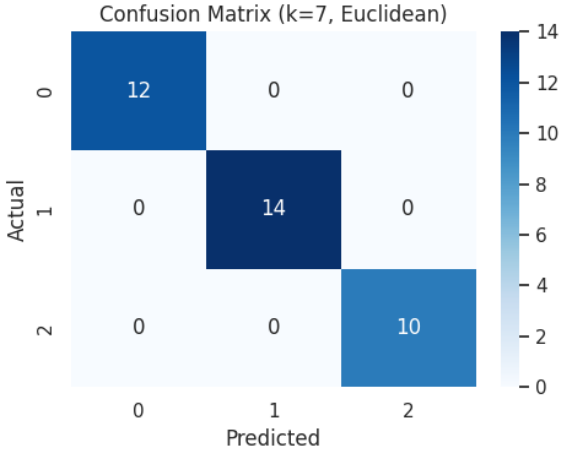
	precision	recall	f1-score	support
1	0.92	1.00	0.96	12
2	1.00	0.93	0.96	14
3	1.00	1.00	1.00	10
accuracy			0.97	36
macro avg	0.97	0.98	0.97	36
weighted avg	0.97	0.97	0.97	36



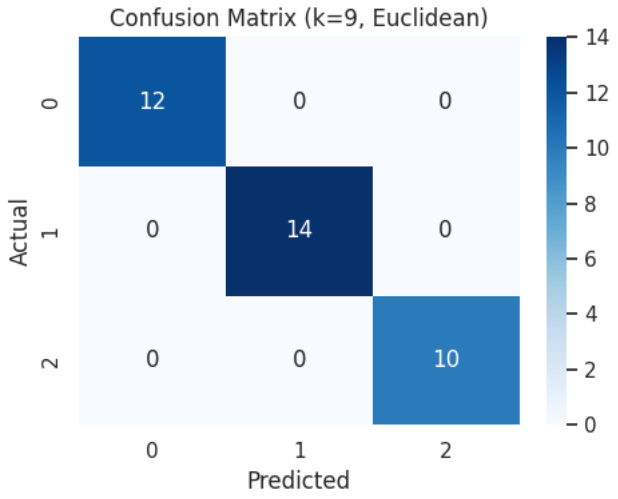
	precision	recall	f1-score	support
1	0.92	1.00	0.96	12
2	1.00	0.93	0.96	14
3	1.00	1.00	1.00	10
accuracy			0.97	36
macro avg	0.97	0.98	0.97	36
weighted avg	0.97	0.97	0.97	36



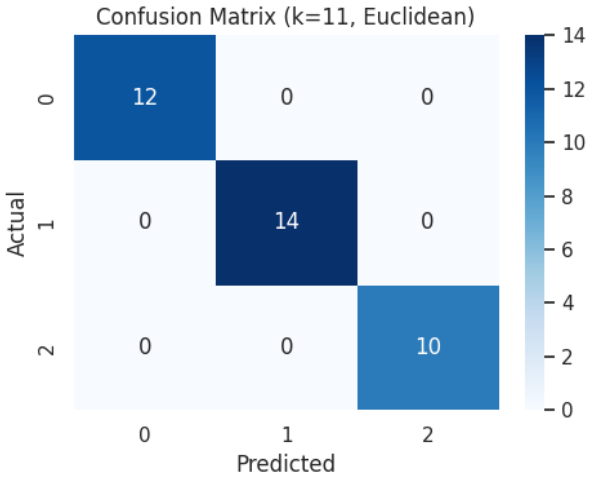
	precision	recall	f1-score	support
1	1.00	1.00	1.00	12
2	1.00	0.93	0.96	14
3	0.91	1.00	0.95	10
accuracy			0.97	36
macro avg	0.97	0.98	0.97	36
weighted avg	0.97	0.97	0.97	36



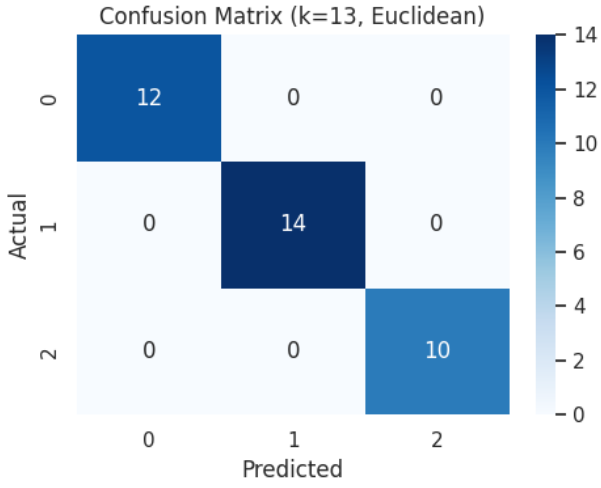
	precision	recall	f1-score	support
1	1.00	1.00	1.00	12
2	1.00	1.00	1.00	14
3	1.00	1.00	1.00	10
accuracy			1.00	36
macro avg	1.00	1.00	1.00	36
weighted avg	1.00	1.00	1.00	36



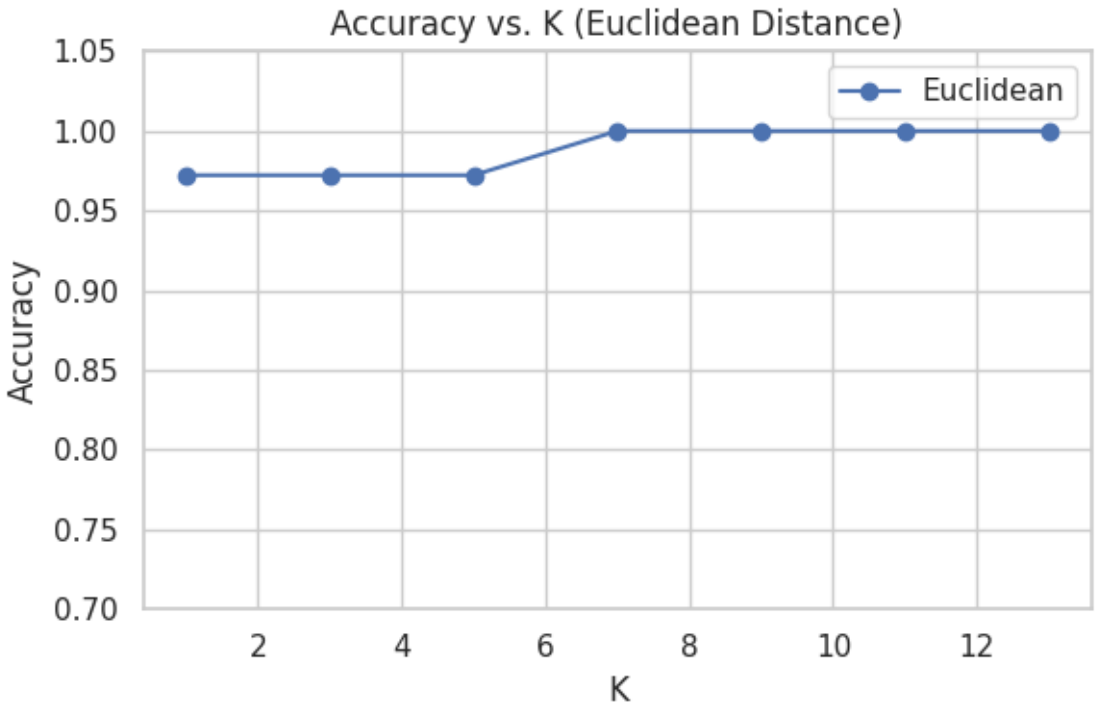
	precision	recall	f1-score	support
1	1.00	1.00	1.00	12
2	1.00	1.00	1.00	14
3	1.00	1.00	1.00	10
accuracy			1.00	36
macro avg	1.00	1.00	1.00	36
weighted avg	1.00	1.00	1.00	36



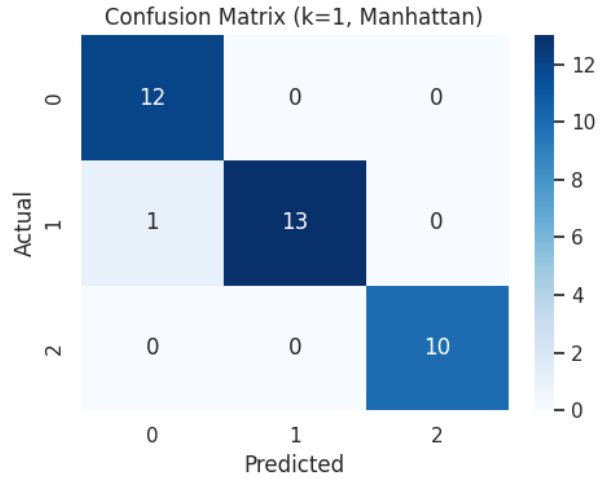
	precision	recall	f1-score	support
1	1.00	1.00	1.00	12
2	1.00	1.00	1.00	14
3	1.00	1.00	1.00	10
accuracy			1.00	36
macro avg	1.00	1.00	1.00	36
weighted avg	1.00	1.00	1.00	36



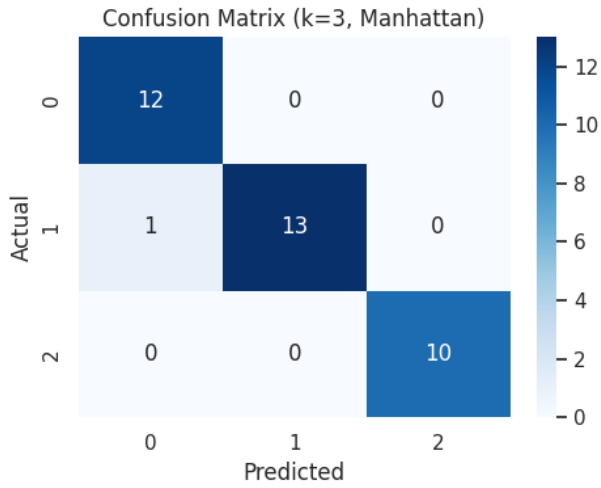
	precision	recall	f1-score	support
1	1.00	1.00	1.00	12
2	1.00	1.00	1.00	14
3	1.00	1.00	1.00	10
accuracy			1.00	36
macro avg	1.00	1.00	1.00	36
weighted avg	1.00	1.00	1.00	36


Accuracy vs. K (Euclidean Distance)
Euclidean
Accuracy
K
1.05
1.00
0.95
0.90
0.85
0.80
0.75
0.70
2
4
6
8
10
12

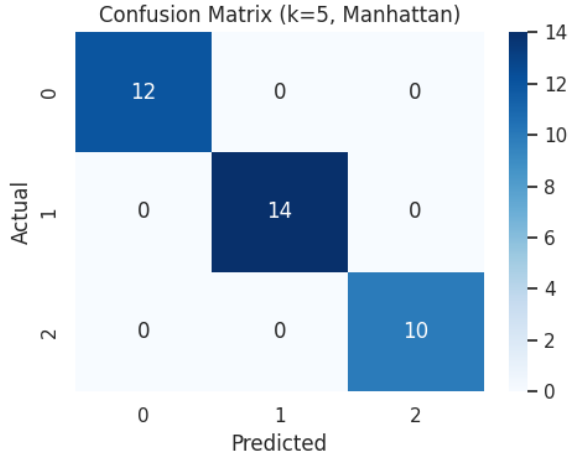
Classification Accuracy for each Value of K(Manhattan)



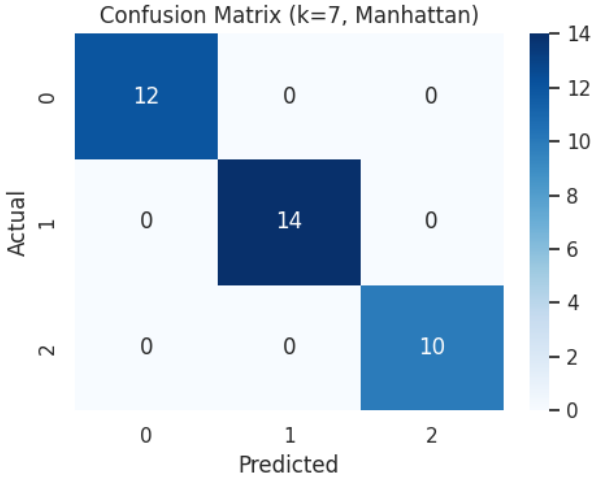
	precision	recall	f1-score	support
1	0.92	1.00	0.96	12
2	1.00	0.93	0.96	14
3	1.00	1.00	1.00	10
accuracy			0.97	36
macro avg	0.97	0.98	0.97	36
weighted avg	0.97	0.97	0.97	36



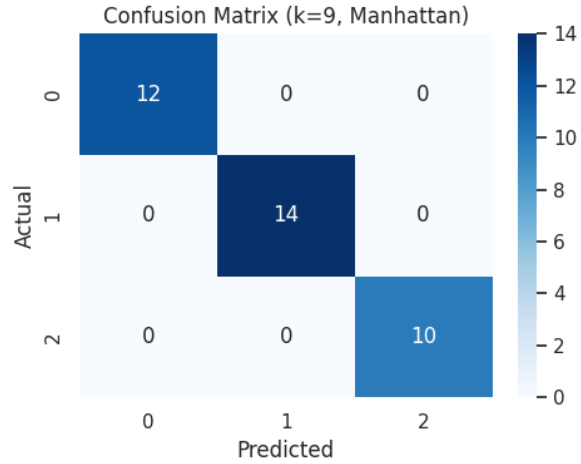
	precision	recall	f1-score	support
1	0.92	1.00	0.96	12
2	1.00	0.93	0.96	14
3	1.00	1.00	1.00	10
accuracy			0.97	36
macro avg	0.97	0.98	0.97	36
weighted avg	0.97	0.97	0.97	36



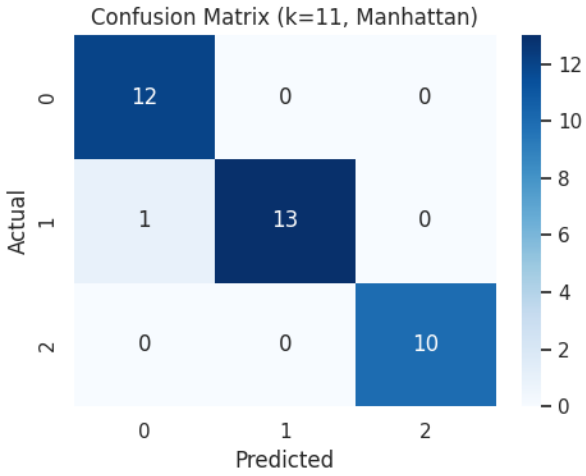
	precision	recall	f1-score	support
1	1.00	1.00	1.00	12
2	1.00	1.00	1.00	14
3	1.00	1.00	1.00	10
accuracy			1.00	36
macro avg	1.00	1.00	1.00	36
weighted avg	1.00	1.00	1.00	36



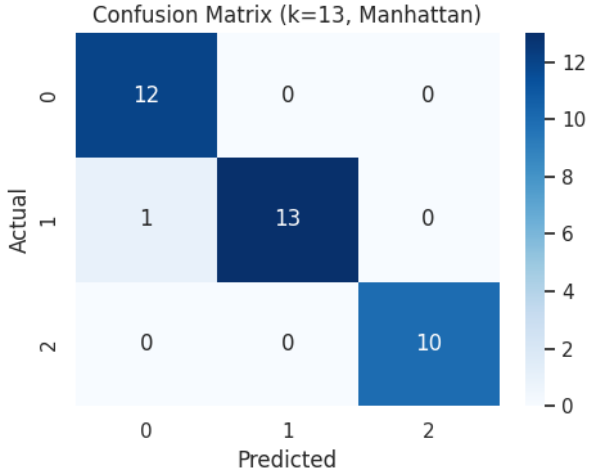
	precision	recall	f1-score	support
1	1.00	1.00	1.00	12
2	1.00	1.00	1.00	14
3	1.00	1.00	1.00	10
accuracy			1.00	36
macro avg	1.00	1.00	1.00	36
weighted avg	1.00	1.00	1.00	36



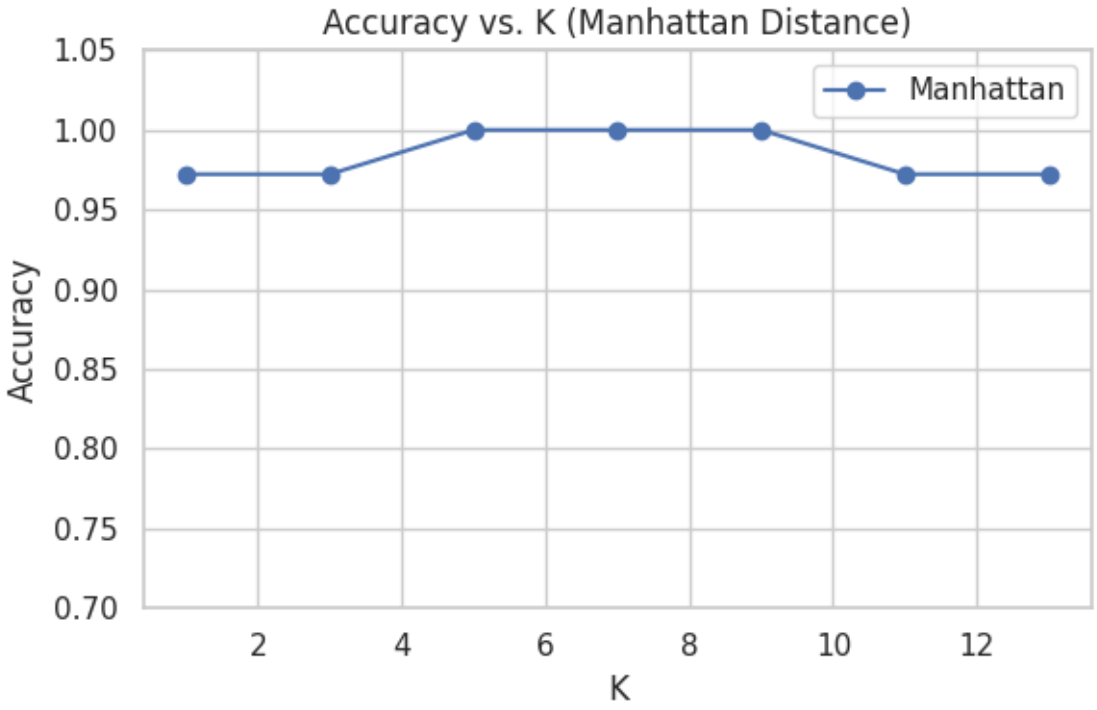
	precision	recall	f1-score	support
1	1.00	1.00	1.00	12
2	1.00	1.00	1.00	14
3	1.00	1.00	1.00	10
accuracy			1.00	36
macro avg	1.00	1.00	1.00	36
weighted avg	1.00	1.00	1.00	36



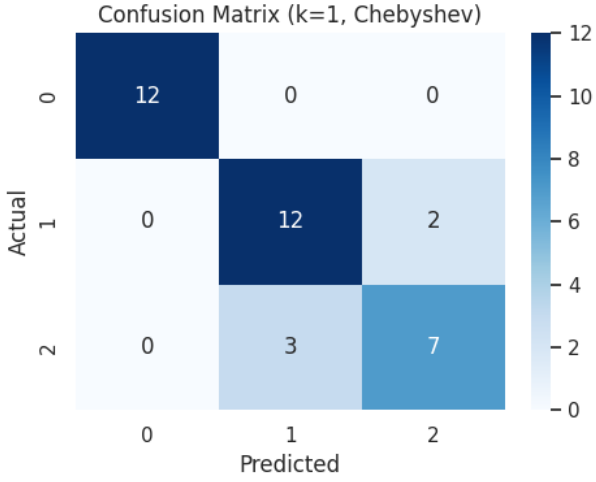
	precision	recall	f1-score	support
1	0.92	1.00	0.96	12
2	1.00	0.93	0.96	14
3	1.00	1.00	1.00	10
accuracy			0.97	36
macro avg	0.97	0.98	0.97	36
weighted avg	0.97	0.97	0.97	36



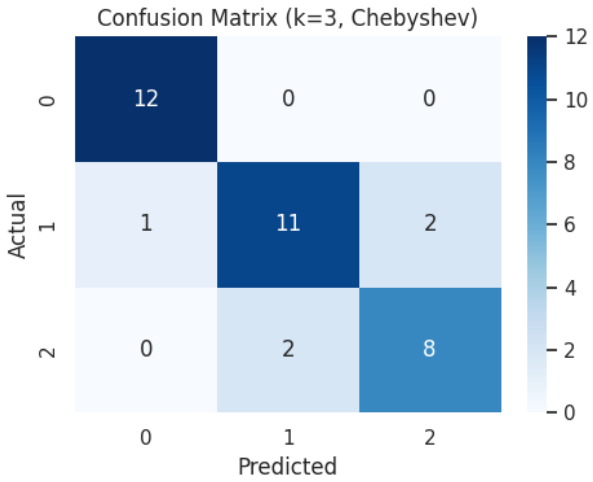
	precision	recall	f1-score	support
1	0.92	1.00	0.96	12
2	1.00	0.93	0.96	14
3	1.00	1.00	1.00	10
accuracy			0.97	36
macro avg	0.97	0.98	0.97	36
weighted avg	0.97	0.97	0.97	36


Accuracy vs. K (Manhattan Distance)
Manhattan
Accuracy
K
1.05
1.00
0.95
0.90
0.85
0.80
0.75
0.70
2
4
6
8
10
12

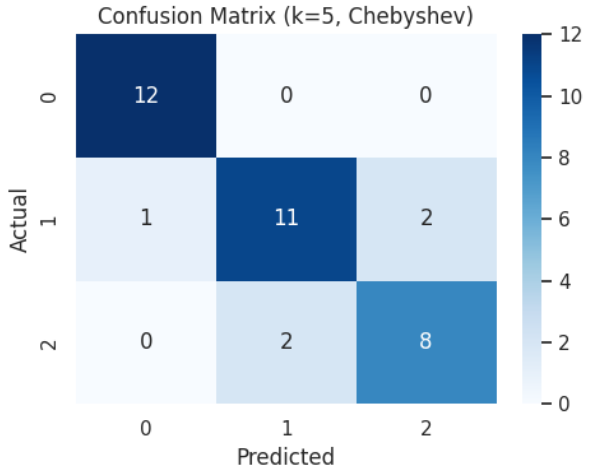
Classification Accuracy for each Value of K(Chebyshev)



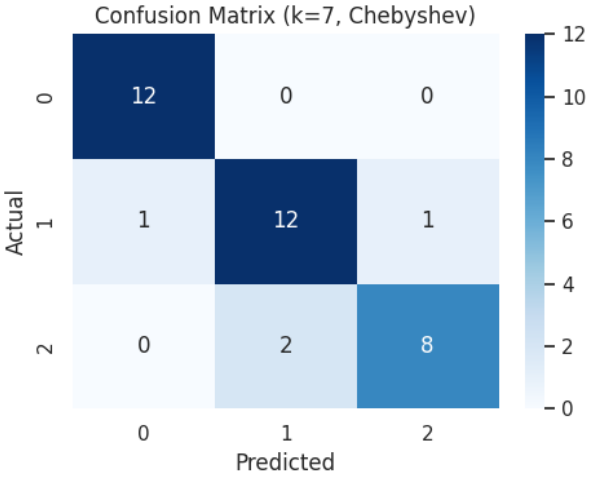
	precision	recall	f1-score	support
1	1.00	1.00	1.00	12
2	0.80	0.86	0.83	14
3	0.78	0.70	0.74	10
accuracy			0.86	36
macro avg	0.86	0.85	0.85	36
weighted avg	0.86	0.86	0.86	36



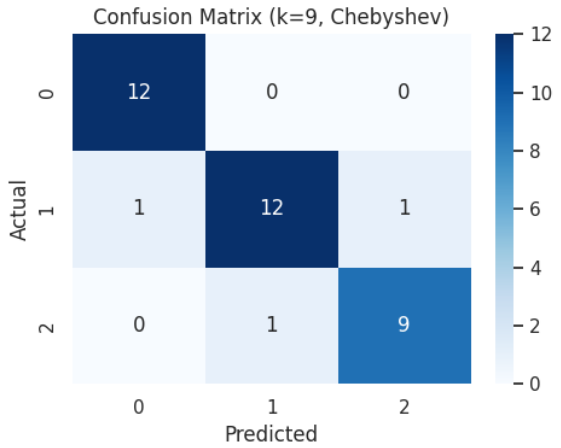
	precision	recall	f1-score	support
1	0.92	1.00	0.96	12
2	0.85	0.79	0.81	14
3	0.80	0.80	0.80	10
accuracy			0.86	36
macro avg	0.86	0.86	0.86	36
weighted avg	0.86	0.86	0.86	36



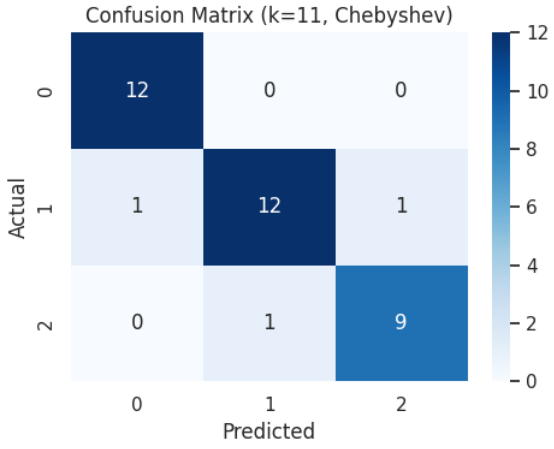
	precision	recall	f1-score	support
1	0.92	1.00	0.96	12
2	0.85	0.79	0.81	14
3	0.80	0.80	0.80	10
accuracy			0.86	36
macro avg	0.86	0.86	0.86	36
weighted avg	0.86	0.86	0.86	36



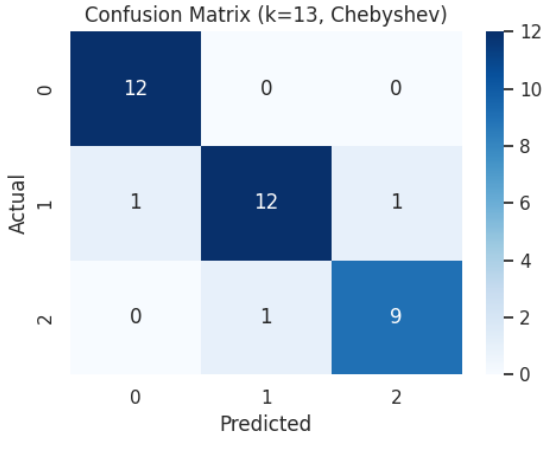
	precision	recall	f1-score	support
1	0.92	1.00	0.96	12
2	0.86	0.86	0.86	14
3	0.89	0.80	0.84	10
accuracy			0.89	36
macro avg	0.89	0.89	0.89	36
weighted avg	0.89	0.89	0.89	36



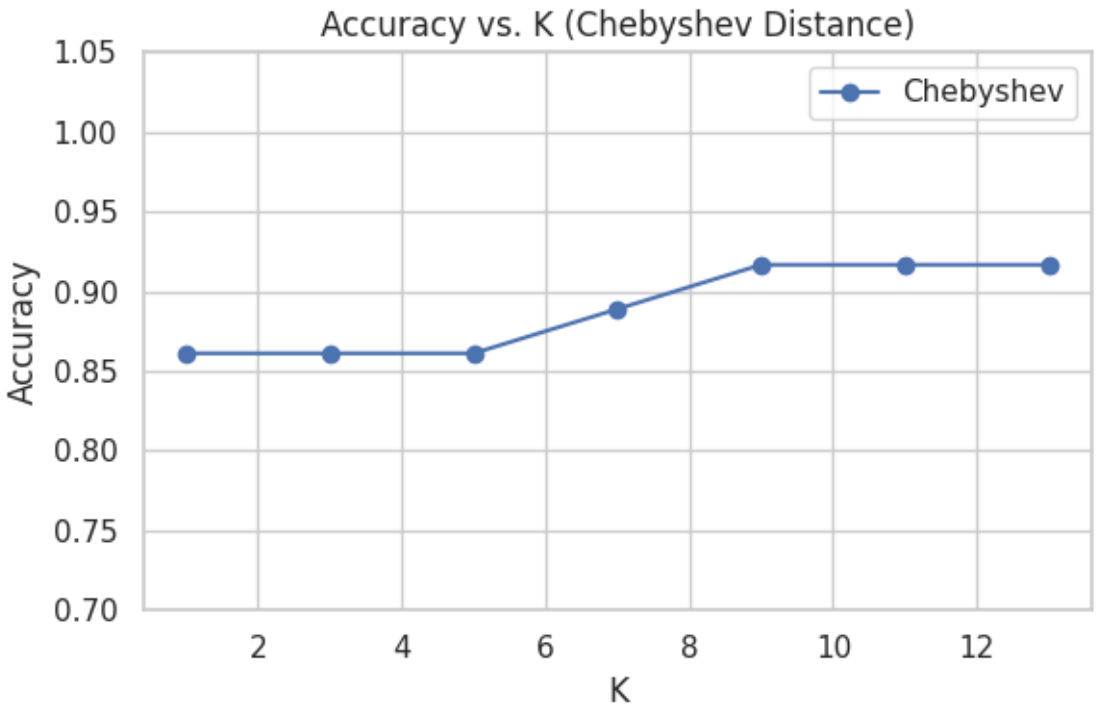
	precision	recall	f1-score	support
1	0.92	1.00	0.96	12
2	0.92	0.86	0.89	14
3	0.90	0.90	0.90	10
accuracy			0.92	36
macro avg	0.92	0.92	0.92	36
weighted avg	0.92	0.92	0.92	36



	precision	recall	f1-score	support
1	0.92	1.00	0.96	12
2	0.92	0.86	0.89	14
3	0.90	0.90	0.90	10
accuracy			0.92	36
macro avg	0.92	0.92	0.92	36
weighted avg	0.92	0.92	0.92	36



	precision	recall	f1-score	support
1	0.92	1.00	0.96	12
2	0.92	0.86	0.89	14
3	0.90	0.90	0.90	10
accuracy			0.92	36
macro avg	0.92	0.92	0.92	36
weighted avg	0.92	0.92	0.92	36


Accuracy vs. K (Chebyshev Distance)
Chebyshev
Accuracy
K
1.05
1.00
0.95
0.90
0.85
0.80
0.75
0.70
2
4
6
8
10
12

Conclusion

To sum up:

- **Euclidean distance** with **K=5 or K=7** gave the best results.
- **Flavanoids**, **Color Intensity**, and **Alcohol** were helpful features for distinguishing classes.
- **Chebyshev** was less reliable and gave lower accuracy overall.
- The model worked best when K wasn't too small or too large—around 5 to 7 seems ideal for this dataset.

This project helped me understand how K and the distance metric affect k-NN performance, and how feature distributions impact classification.

GitHub Repository: https://github.com/aalperennozkann/ELE489-HW1