Intro

In this task I implemented the K-Nearest Neighbors algorithm to classify Iris flowers. The main focus was on making 3 functions: calculating Euclidean distance (the core function), then finding the nearest neighbors and making predictions.

The Iris dataset has 150 rows, 4 features, and 3 types of flowers (Iris Setosa, Iris Versicolour, Iris Virginica).

Implementation and results

For writing the algorithm I just used the instructions provided in the assignment description (you can find more details in my ipynb file that is also attached to the report).
The main goal was to test KNN on a new data point [7.0, 3.1, 1.3, 0.7], and the model predicted ‘new_dp’ as ‘Iris-setosa’ but when I visualized the data things got tricky. 'new_dp' is indeed closer to the Iris-Setosa data points (especially if we consider petal width; closer distances to Iris-setosa data points).

At the same time the test point looked quite far from other points in some of the feature spaces. This lead me to an assumption that we are touching limits of the method or the dataset itself.

Observations

Technically, the KNN implementation looks correct and was created by using the recommendation provided. When I looked at the data visually and plotted them, I started doubting the results. The new data point is repeatedly classified as Iris-setosa for smaller k values (which should be reasonable for this dataset), but looking at the plot it seems data point [7.0, 3.1, 1.3, 0.7] could belong to a completely different class that is not described in the dataset.

I think the current dataset might not fully describe this new point accurately, and the results are not very sensitive to metric choices (k, number of neighbours).

Summary

As I understood this exercise intended us to notice that even everything seems fine technically it is important to understand the limitations of both the method and the dataset. I can easily visualize the data and see the patterns here with 4 features and look at the results, but with more features and a bigger dataset it can get much harder to find issues like these and control the final results (in real-world with more complex and for example not ideally preprocessed data could lead to much more time-consuming solutions without the understanding the dataset and it becomes hard to trust the results).

In short, while the method works technically and provides the result, visualizing the data raised several questions. This shows the importance of being awared about the dataset and the potential limitations of chosen algorithms.