**Assignment 3**

Machine Learning, SS2021

|  |  |  |
| --- | --- | --- |
| **Team members** | | |
| Last name | First name | Matriculation Number |
| Rodiga | Luca | 1620972 |
| Binford | Bob | 1600002 |

Task 1)

Dataset 1: {'k': 1}

Test Score: 1.0

Chart, line chart

Description automatically generatedChart, scatter chart

Description automatically generated

Dataset 2: {'k': 1}

Test Score: 0.9953703703703703

Scatter chart

Description automatically generated

Chart, line chart

Description automatically generated

Chart, line chart

Description automatically generatedChart, scatter chart

Description automatically generated

Chart, line chart

Description automatically generatedDataset 3: {'k': 3}

Test Score: 0.8867924528301887

1.3)

Choosing a k too small leads to an inefficient and inaccurate clustering performance, where neighbors that matter are not considered. Especially with k = 1, only the nearest neighbor is considered and it’s label is assigned to the test datapoint.

Choosing a k too large leads to inaccurate predictions, because too many data points are considered

Chart, scatter chart

Description automatically generated

Mean score of cross validation: 0.7634585569469291

Test Score for k=1: 0.7222222222222222

Chart, scatter chart

Description automatically generated

Mean score of cross validation: 0.816124031007752

Test Score for k=30: 0.7777777777777778

Chart, scatter chart

Description automatically generated

Mean score of cross validation: 0.7186762075134168

Test Score for k=100: 0.7037037037037037

Chart, line chart

Description automatically generated

The best parameter found by GridSearchCV was {'k': 7} with an accuracy of 0.8269290399522958