**Assignment 3**

Machine Learning, SS2021

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

Task 1)

Dataset 1: {'k': 3}

Test Score: 0.6153846153846154

Chart, line chart

Description automatically generatedChart, scatter chart

Description automatically generated

Dataset 2: {'k': 3}

Test Score: 0.4444444444444444

Chart, scatter chart

Description automatically generated

Chart, line chart

Description automatically generated

Dataset 3: {'k': 3}

Chart, scatter chart

Description automatically generatedChart, line chart

Description automatically generatedTest Score: 0.5094339622641509

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: 0.44354203935599285

Test Score for k=1: 0.5555555555555556

Chart, scatter chart

Description automatically generated

Mean score of cross validation: 0.4666070363744782

Test Score for k=30: 0.5555555555555556

Chart, scatter chart

Description automatically generated

Mean score of cross validation: 0.44354203935599285

Test Score for k=100: 0.5555555555555556

The best parameter found by GridSearchCV was {'k': 3} with an accuracy of 0.46815742397137744

Chart, line chart

Description automatically generated