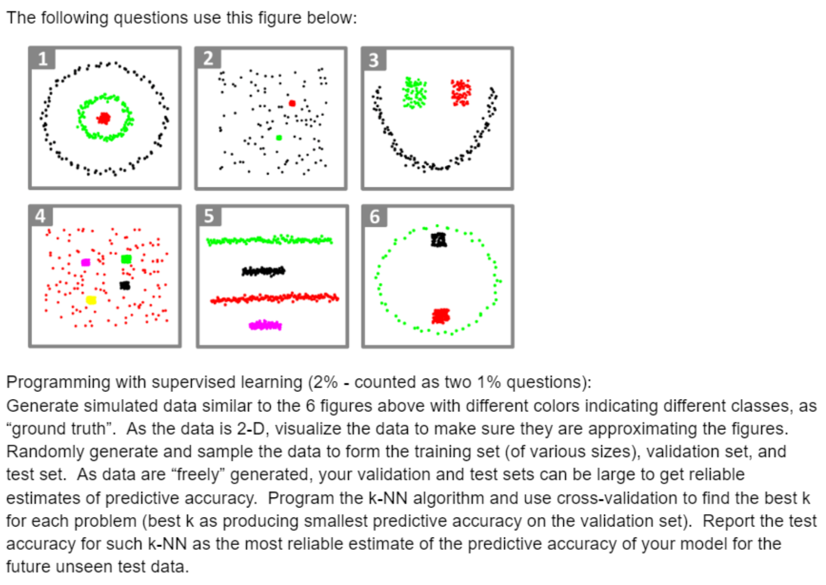
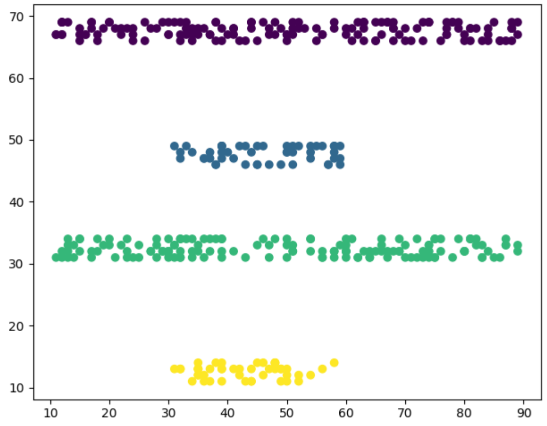
Connor McGoey

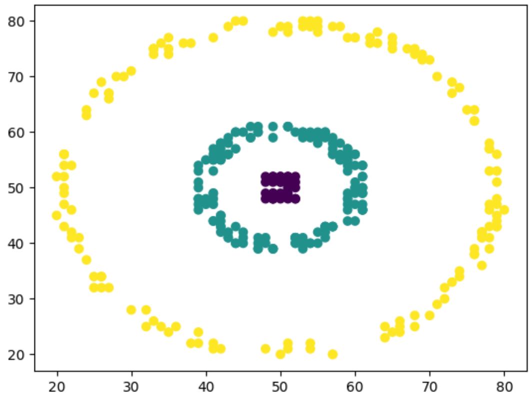


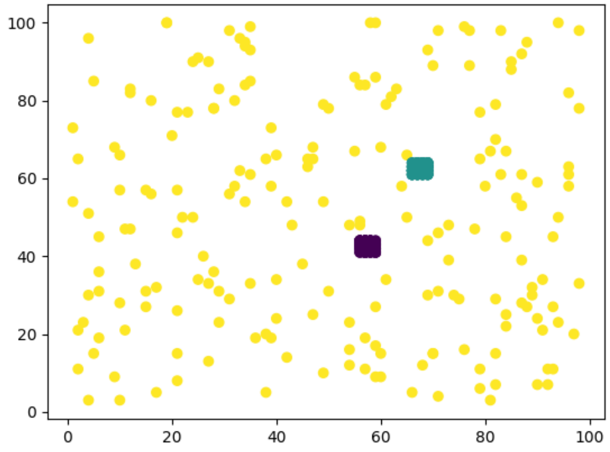
The following questions use this figure below:

Programming with supervised learning (2% - counted as two 1% questions):
Generate simulated data similar to the 6 figures above with different colors indicating different classes, as "ground truth". As the data is 2-D, visualize the data to make sure they are approximating the figures. Randomly generate and sample the data to form the training set (of various sizes), validation set, and test set. As data are "freely" generated, your validation and test sets can be large to get reliable estimates of predictive accuracy. Program the k-NN algorithm and use cross-validation to find the best k for each problem (best k as producing smallest predictive accuracy on the validation set). Report the test accuracy for such k-NN as the most reliable estimate of the predictive accuracy of your model for the future unseen test data.

The data sets I am choosing to work with are: 5, 1, and 2 The following are my data sets that I created in Python mapped with matplotlib.pyplot:

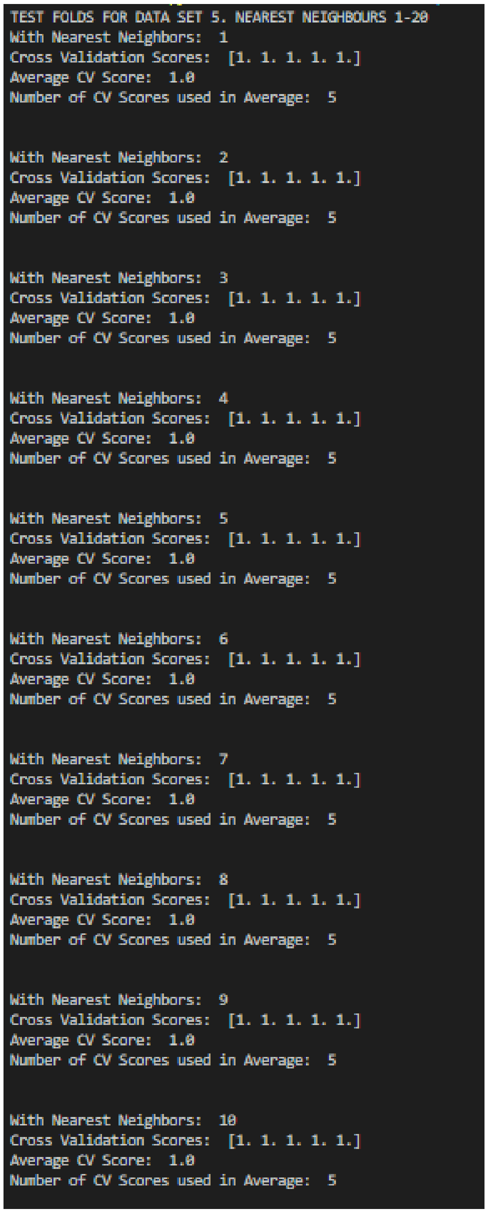




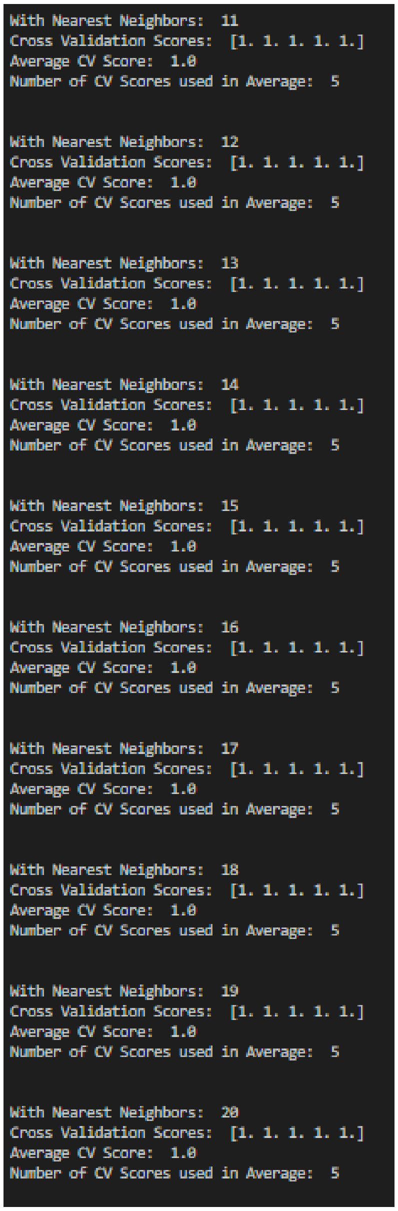


Program Output For KNN

DATA SET 5 (5 FOLDS)

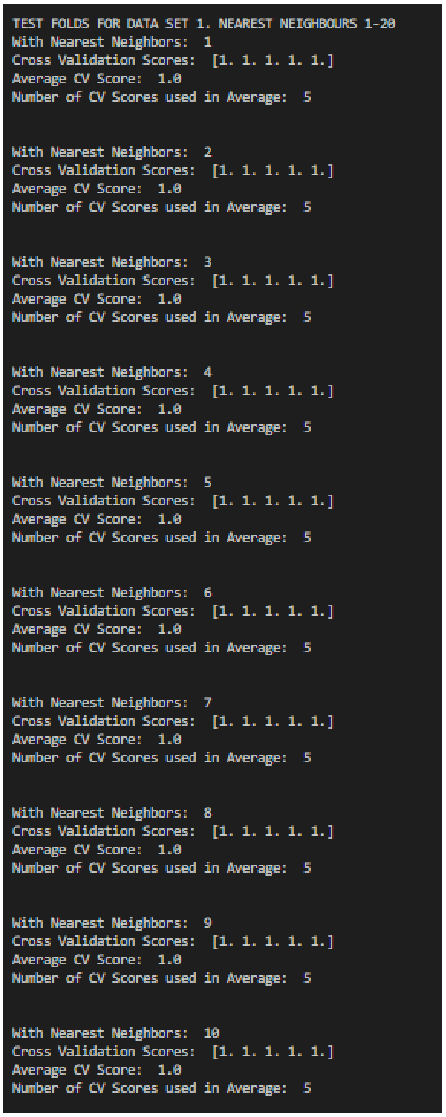


```
TEST FOLDS FOR DATA SET 5. NEAREST NEIGHBOURS 1-20
With Nearest Neighbors:  1
Cross Validation Scores:  [1. 1. 1. 1. 1.]
Average CV Score:  1.0
Number of CV Scores used in Average:  5


With Nearest Neighbors:  2
Cross Validation Scores:  [1. 1. 1. 1. 1.]
Average CV Score:  1.0
Number of CV Scores used in Average:  5


With Nearest Neighbors:  3
Cross Validation Scores:  [1. 1. 1. 1. 1.]
Average CV Score:  1.0
Number of CV Scores used in Average:  5


With Nearest Neighbors:  4
Cross Validation Scores:  [1. 1. 1. 1. 1.]
Average CV Score:  1.0
Number of CV Scores used in Average:  5


With Nearest Neighbors:  5
Cross Validation Scores:  [1. 1. 1. 1. 1.]
Average CV Score:  1.0
Number of CV Scores used in Average:  5


With Nearest Neighbors:  6
Cross Validation Scores:  [1. 1. 1. 1. 1.]
Average CV Score:  1.0
Number of CV Scores used in Average:  5


With Nearest Neighbors:  7
Cross Validation Scores:  [1. 1. 1. 1. 1.]
Average CV Score:  1.0
Number of CV Scores used in Average:  5


With Nearest Neighbors:  8
Cross Validation Scores:  [1. 1. 1. 1. 1.]
Average CV Score:  1.0
Number of CV Scores used in Average:  5


With Nearest Neighbors:  9
Cross Validation Scores:  [1. 1. 1. 1. 1.]
Average CV Score:  1.0
Number of CV Scores used in Average:  5


With Nearest Neighbors:  10
Cross Validation Scores:  [1. 1. 1. 1. 1.]
Average CV Score:  1.0
Number of CV Scores used in Average:  5
```

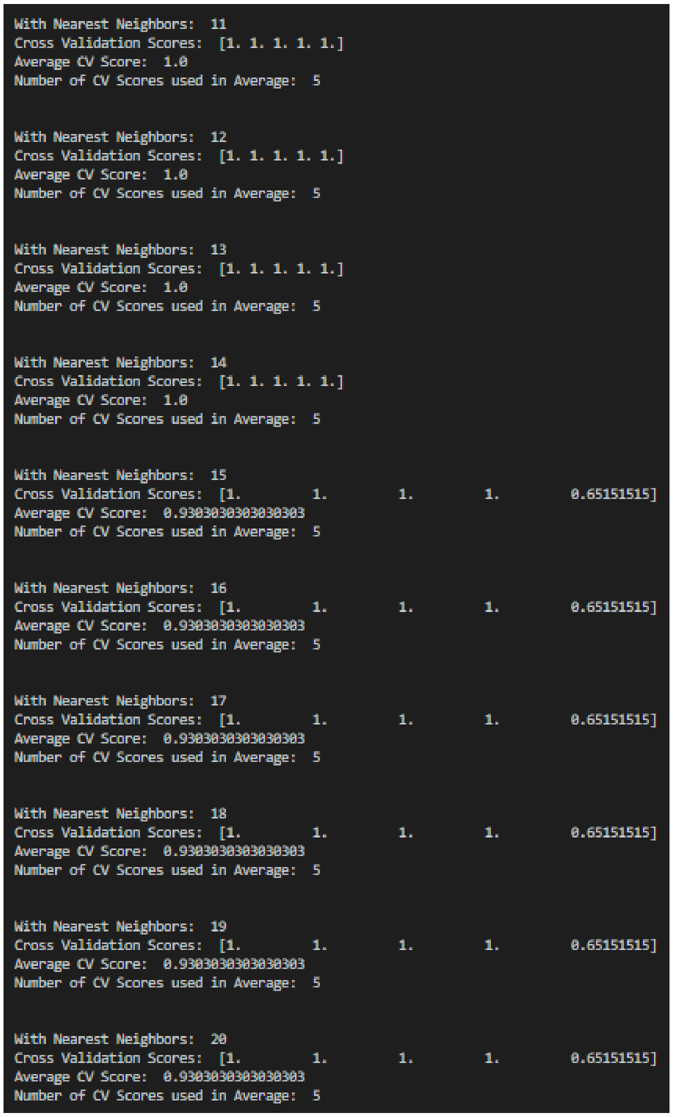


```
With Nearest Neighbors:  11
Cross Validation Scores:  [1. 1. 1. 1. 1.]
Average CV Score:  1.0
Number of CV Scores used in Average:  5


With Nearest Neighbors:  12
Cross Validation Scores:  [1. 1. 1. 1. 1.]
Average CV Score:  1.0
Number of CV Scores used in Average:  5


With Nearest Neighbors:  13
Cross Validation Scores:  [1. 1. 1. 1. 1.]
Average CV Score:  1.0
Number of CV Scores used in Average:  5


With Nearest Neighbors:  14
Cross Validation Scores:  [1. 1. 1. 1. 1.]
Average CV Score:  1.0
Number of CV Scores used in Average:  5


With Nearest Neighbors:  15
Cross Validation Scores:  [1. 1. 1. 1. 1.]
Average CV Score:  1.0
Number of CV Scores used in Average:  5


With Nearest Neighbors:  16
Cross Validation Scores:  [1. 1. 1. 1. 1.]
Average CV Score:  1.0
Number of CV Scores used in Average:  5


With Nearest Neighbors:  17
Cross Validation Scores:  [1. 1. 1. 1. 1.]
Average CV Score:  1.0
Number of CV Scores used in Average:  5


With Nearest Neighbors:  18
Cross Validation Scores:  [1. 1. 1. 1. 1.]
Average CV Score:  1.0
Number of CV Scores used in Average:  5


With Nearest Neighbors:  19
Cross Validation Scores:  [1. 1. 1. 1. 1.]
Average CV Score:  1.0
Number of CV Scores used in Average:  5


With Nearest Neighbors:  20
Cross Validation Scores:  [1. 1. 1. 1. 1.]
Average CV Score:  1.0
Number of CV Scores used in Average:  5
```

DATA SET 1 (5 FOLDS)

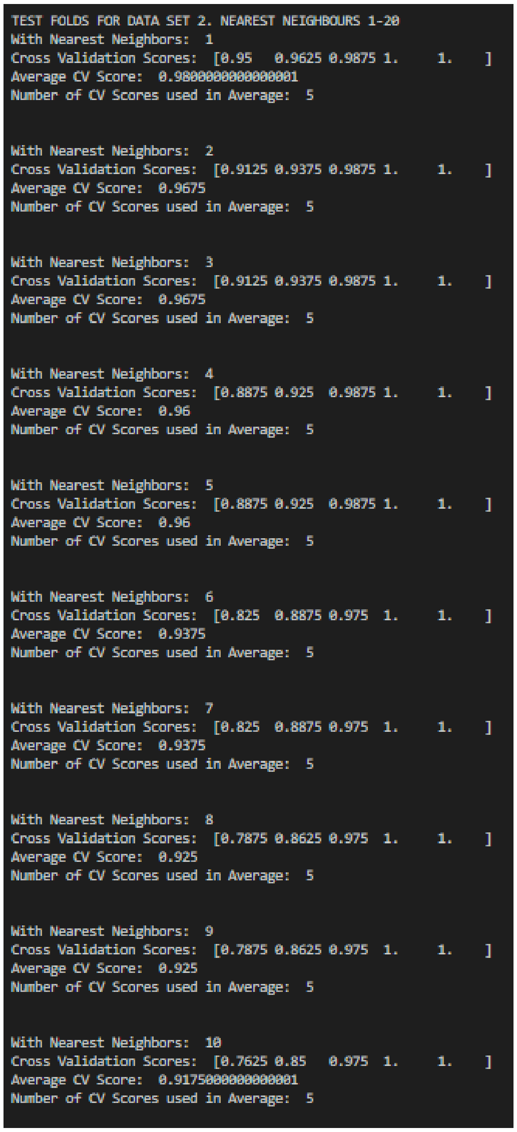


```
TEST FOLDS FOR DATA SET 1. NEAREST NEIGHBOURS 1-20
With Nearest Neighbors:  1
Cross Validation Scores:  [1. 1. 1. 1. 1.]
Average CV Score:  1.0
Number of CV Scores used in Average:  5


With Nearest Neighbors:  2
Cross Validation Scores:  [1. 1. 1. 1. 1.]
Average CV Score:  1.0
Number of CV Scores used in Average:  5


With Nearest Neighbors:  3
Cross Validation Scores:  [1. 1. 1. 1. 1.]
Average CV Score:  1.0
Number of CV Scores used in Average:  5


With Nearest Neighbors:  4
Cross Validation Scores:  [1. 1. 1. 1. 1.]
Average CV Score:  1.0
Number of CV Scores used in Average:  5


With Nearest Neighbors:  5
Cross Validation Scores:  [1. 1. 1. 1. 1.]
Average CV Score:  1.0
Number of CV Scores used in Average:  5


With Nearest Neighbors:  6
Cross Validation Scores:  [1. 1. 1. 1. 1.]
Average CV Score:  1.0
Number of CV Scores used in Average:  5


With Nearest Neighbors:  7
Cross Validation Scores:  [1. 1. 1. 1. 1.]
Average CV Score:  1.0
Number of CV Scores used in Average:  5


With Nearest Neighbors:  8
Cross Validation Scores:  [1. 1. 1. 1. 1.]
Average CV Score:  1.0
Number of CV Scores used in Average:  5


With Nearest Neighbors:  9
Cross Validation Scores:  [1. 1. 1. 1. 1.]
Average CV Score:  1.0
Number of CV Scores used in Average:  5


With Nearest Neighbors:  10
Cross Validation Scores:  [1. 1. 1. 1. 1.]
Average CV Score:  1.0
Number of CV Scores used in Average:  5
```

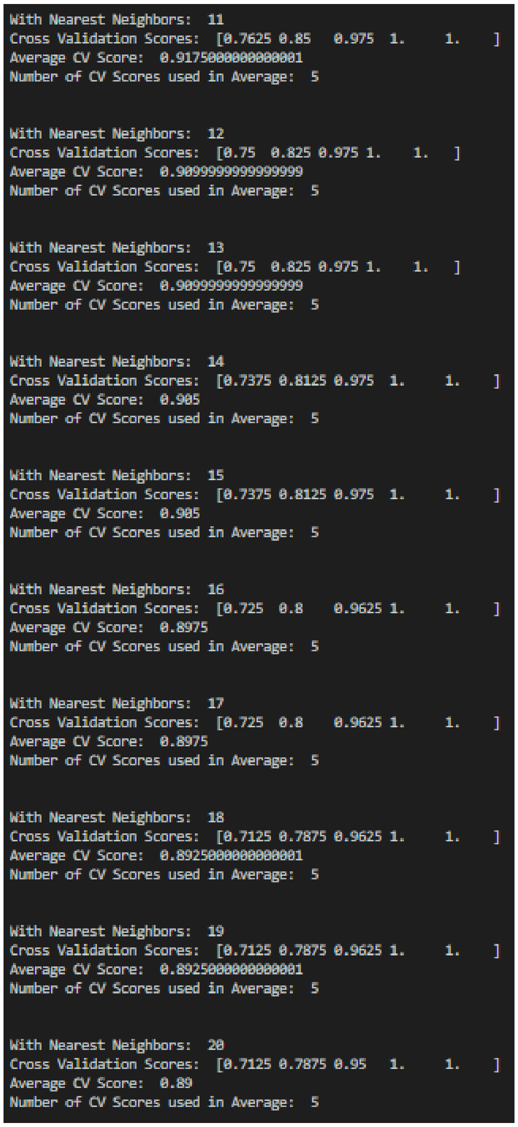


```
With Nearest Neighbors:  11
Cross Validation Scores:  [1. 1. 1. 1. 1.]
Average CV Score:  1.0
Number of CV Scores used in Average:  5


With Nearest Neighbors:  12
Cross Validation Scores:  [1. 1. 1. 1. 1.]
Average CV Score:  1.0
Number of CV Scores used in Average:  5


With Nearest Neighbors:  13
Cross Validation Scores:  [1. 1. 1. 1. 1.]
Average CV Score:  1.0
Number of CV Scores used in Average:  5


With Nearest Neighbors:  14
Cross Validation Scores:  [1. 1. 1. 1. 1.]
Average CV Score:  1.0
Number of CV Scores used in Average:  5


With Nearest Neighbors:  15
Cross Validation Scores:  [1.         1.         1.         1.         0.65151515]
Average CV Score:  0.9303030303030303
Number of CV Scores used in Average:  5


With Nearest Neighbors:  16
Cross Validation Scores:  [1.         1.         1.         1.         0.65151515]
Average CV Score:  0.9303030303030303
Number of CV Scores used in Average:  5


With Nearest Neighbors:  17
Cross Validation Scores:  [1.         1.         1.         1.         0.65151515]
Average CV Score:  0.9303030303030303
Number of CV Scores used in Average:  5


With Nearest Neighbors:  18
Cross Validation Scores:  [1.         1.         1.         1.         0.65151515]
Average CV Score:  0.9303030303030303
Number of CV Scores used in Average:  5


With Nearest Neighbors:  19
Cross Validation Scores:  [1.         1.         1.         1.         0.65151515]
Average CV Score:  0.9303030303030303
Number of CV Scores used in Average:  5


With Nearest Neighbors:  20
Cross Validation Scores:  [1.         1.         1.         1.         0.65151515]
Average CV Score:  0.9303030303030303
Number of CV Scores used in Average:  5
```

DATA SET 2 (5 FOLDS)

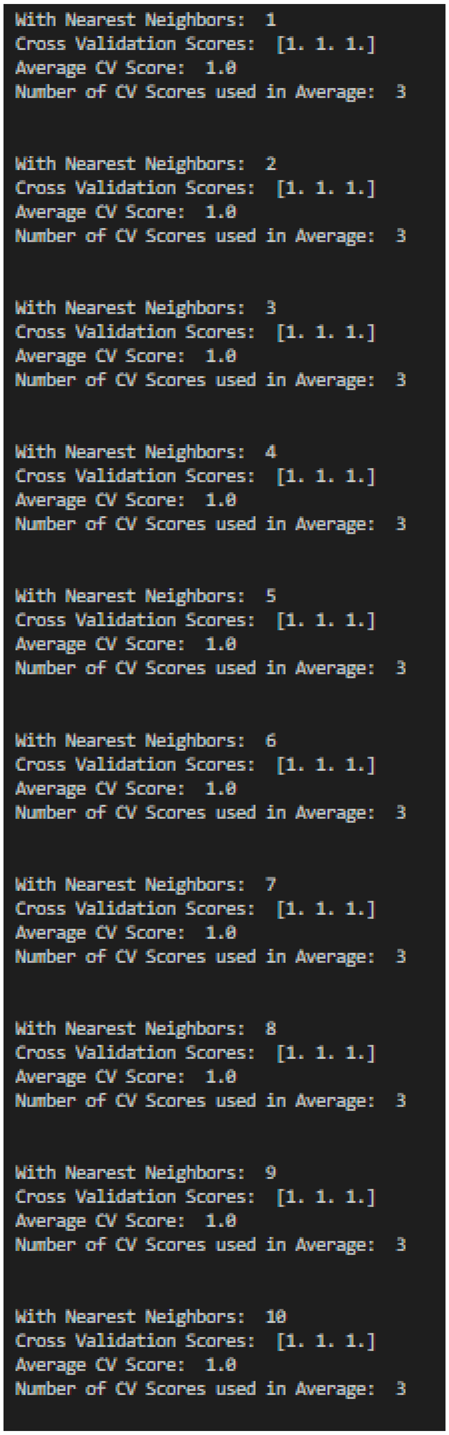


```
TEST FOLDS FOR DATA SET 2. NEAREST NEIGHBOURS 1-20
With Nearest Neighbors:  1
Cross Validation Scores:  [0.95   0.9625 0.9875 1.     1.    ]
Average CV Score:  0.9800000000000001
Number of CV Scores used in Average:  5


With Nearest Neighbors:  2
Cross Validation Scores:  [0.9125 0.9375 0.9875 1.     1.    ]
Average CV Score:  0.9675
Number of CV Scores used in Average:  5


With Nearest Neighbors:  3
Cross Validation Scores:  [0.9125 0.9375 0.9875 1.     1.    ]
Average CV Score:  0.9675
Number of CV Scores used in Average:  5


With Nearest Neighbors:  4
Cross Validation Scores:  [0.8875 0.925  0.9875 1.     1.    ]
Average CV Score:  0.96
Number of CV Scores used in Average:  5


With Nearest Neighbors:  5
Cross Validation Scores:  [0.8875 0.925  0.9875 1.     1.    ]
Average CV Score:  0.96
Number of CV Scores used in Average:  5


With Nearest Neighbors:  6
Cross Validation Scores:  [0.825  0.8875 0.975  1.     1.    ]
Average CV Score:  0.9375
Number of CV Scores used in Average:  5


With Nearest Neighbors:  7
Cross Validation Scores:  [0.825  0.8875 0.975  1.     1.    ]
Average CV Score:  0.9375
Number of CV Scores used in Average:  5


With Nearest Neighbors:  8
Cross Validation Scores:  [0.7875 0.8625 0.975  1.     1.    ]
Average CV Score:  0.925
Number of CV Scores used in Average:  5


With Nearest Neighbors:  9
Cross Validation Scores:  [0.7875 0.8625 0.975  1.     1.    ]
Average CV Score:  0.925
Number of CV Scores used in Average:  5


With Nearest Neighbors:  10
Cross Validation Scores:  [0.7625 0.85   0.975  1.     1.    ]
Average CV Score:  0.9175000000000001
Number of CV Scores used in Average:  5
```

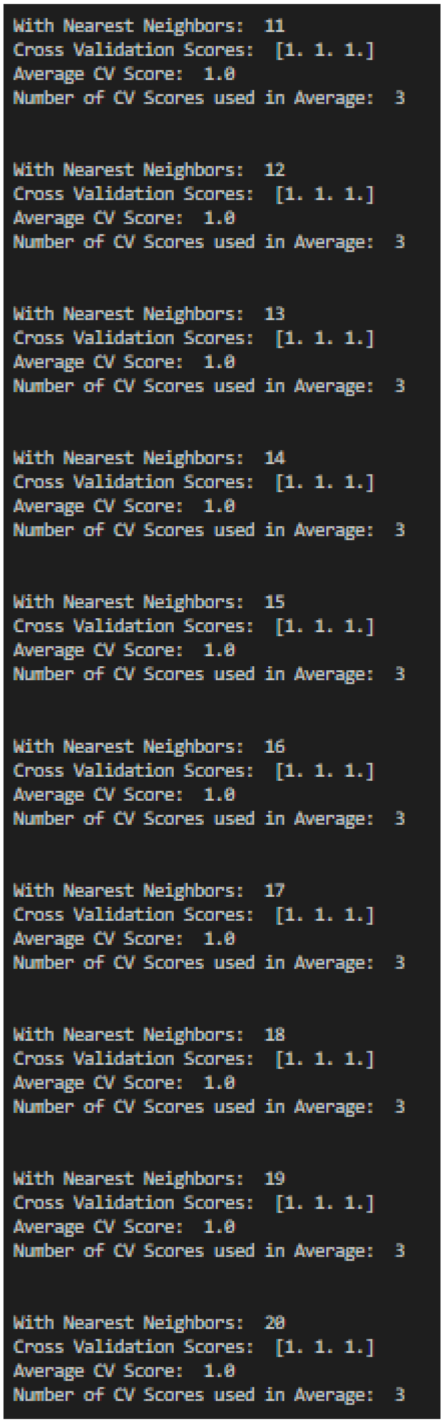


```
With Nearest Neighbors:  11
Cross Validation Scores:  [0.7625 0.85   0.975  1.     1.    ]
Average CV Score:  0.9175000000000001
Number of CV Scores used in Average:  5


With Nearest Neighbors:  12
Cross Validation Scores:  [0.75  0.825 0.975 1.    1.   ]
Average CV Score:  0.9099999999999999
Number of CV Scores used in Average:  5


With Nearest Neighbors:  13
Cross Validation Scores:  [0.75  0.825 0.975 1.    1.   ]
Average CV Score:  0.9099999999999999
Number of CV Scores used in Average:  5


With Nearest Neighbors:  14
Cross Validation Scores:  [0.7375 0.8125 0.975  1.     1.    ]
Average CV Score:  0.905
Number of CV Scores used in Average:  5


With Nearest Neighbors:  15
Cross Validation Scores:  [0.7375 0.8125 0.975  1.     1.    ]
Average CV Score:  0.905
Number of CV Scores used in Average:  5


With Nearest Neighbors:  16
Cross Validation Scores:  [0.725  0.8    0.9625 1.     1.    ]
Average CV Score:  0.8975
Number of CV Scores used in Average:  5


With Nearest Neighbors:  17
Cross Validation Scores:  [0.725  0.8    0.9625 1.     1.    ]
Average CV Score:  0.8975
Number of CV Scores used in Average:  5


With Nearest Neighbors:  18
Cross Validation Scores:  [0.7125 0.7875 0.9625 1.     1.    ]
Average CV Score:  0.8925000000000001
Number of CV Scores used in Average:  5


With Nearest Neighbors:  19
Cross Validation Scores:  [0.7125 0.7875 0.9625 1.     1.    ]
Average CV Score:  0.8925000000000001
Number of CV Scores used in Average:  5


With Nearest Neighbors:  20
Cross Validation Scores:  [0.7125 0.7875 0.95   1.     1.    ]
Average CV Score:  0.89
Number of CV Scores used in Average:  5
```

DATA SET 5 (3 FOLDS)

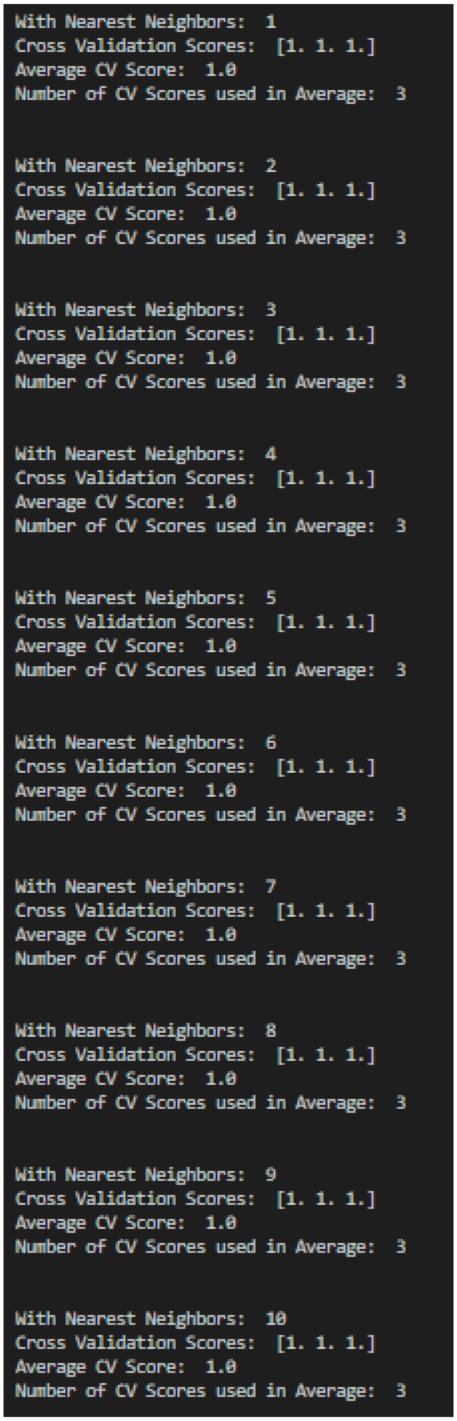


```
With Nearest Neighbors:  1
Cross Validation Scores:  [1. 1. 1.]
Average CV Score:  1.0
Number of CV Scores used in Average:  3


With Nearest Neighbors:  2
Cross Validation Scores:  [1. 1. 1.]
Average CV Score:  1.0
Number of CV Scores used in Average:  3


With Nearest Neighbors:  3
Cross Validation Scores:  [1. 1. 1.]
Average CV Score:  1.0
Number of CV Scores used in Average:  3


With Nearest Neighbors:  4
Cross Validation Scores:  [1. 1. 1.]
Average CV Score:  1.0
Number of CV Scores used in Average:  3


With Nearest Neighbors:  5
Cross Validation Scores:  [1. 1. 1.]
Average CV Score:  1.0
Number of CV Scores used in Average:  3


With Nearest Neighbors:  6
Cross Validation Scores:  [1. 1. 1.]
Average CV Score:  1.0
Number of CV Scores used in Average:  3


With Nearest Neighbors:  7
Cross Validation Scores:  [1. 1. 1.]
Average CV Score:  1.0
Number of CV Scores used in Average:  3


With Nearest Neighbors:  8
Cross Validation Scores:  [1. 1. 1.]
Average CV Score:  1.0
Number of CV Scores used in Average:  3


With Nearest Neighbors:  9
Cross Validation Scores:  [1. 1. 1.]
Average CV Score:  1.0
Number of CV Scores used in Average:  3


With Nearest Neighbors:  10
Cross Validation Scores:  [1. 1. 1.]
Average CV Score:  1.0
Number of CV Scores used in Average:  3
```

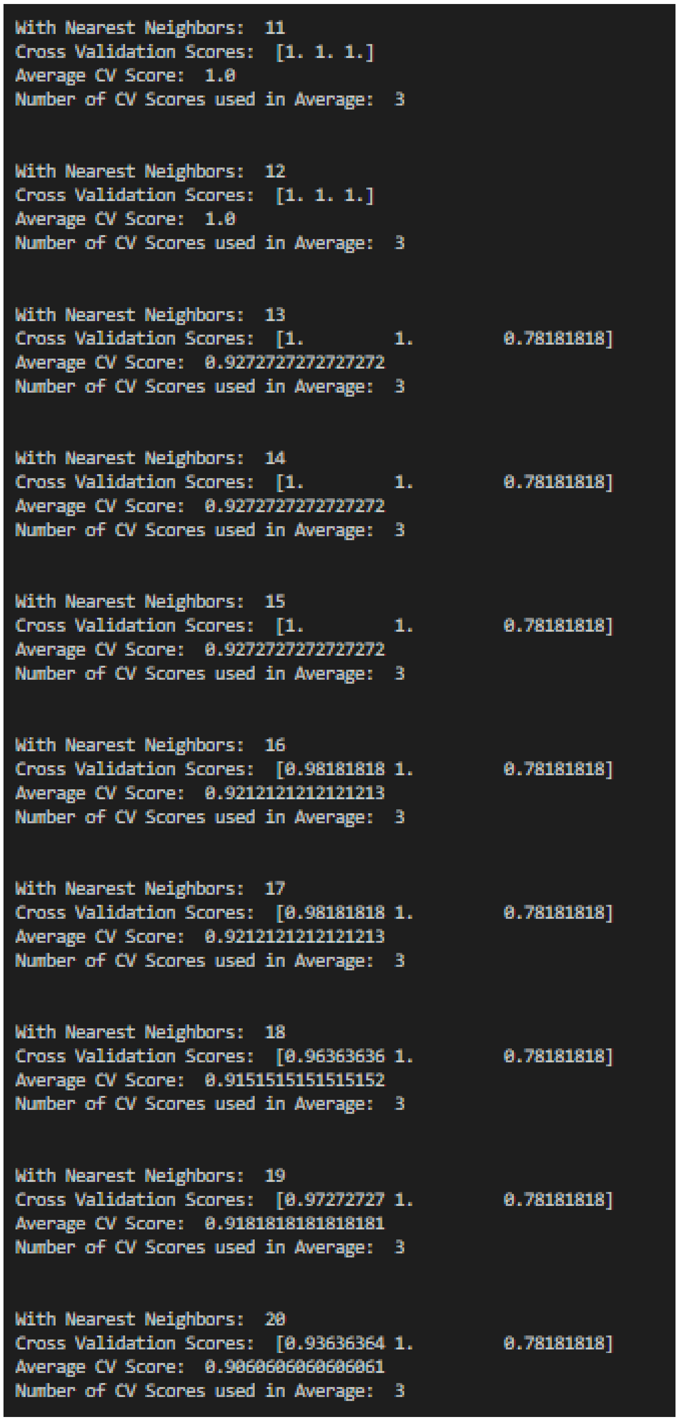


```
With Nearest Neighbors:  11
Cross Validation Scores:  [1. 1. 1.]
Average CV Score:  1.0
Number of CV Scores used in Average:  3


With Nearest Neighbors:  12
Cross Validation Scores:  [1. 1. 1.]
Average CV Score:  1.0
Number of CV Scores used in Average:  3


With Nearest Neighbors:  13
Cross Validation Scores:  [1. 1. 1.]
Average CV Score:  1.0
Number of CV Scores used in Average:  3


With Nearest Neighbors:  14
Cross Validation Scores:  [1. 1. 1.]
Average CV Score:  1.0
Number of CV Scores used in Average:  3


With Nearest Neighbors:  15
Cross Validation Scores:  [1. 1. 1.]
Average CV Score:  1.0
Number of CV Scores used in Average:  3


With Nearest Neighbors:  16
Cross Validation Scores:  [1. 1. 1.]
Average CV Score:  1.0
Number of CV Scores used in Average:  3


With Nearest Neighbors:  17
Cross Validation Scores:  [1. 1. 1.]
Average CV Score:  1.0
Number of CV Scores used in Average:  3


With Nearest Neighbors:  18
Cross Validation Scores:  [1. 1. 1.]
Average CV Score:  1.0
Number of CV Scores used in Average:  3


With Nearest Neighbors:  19
Cross Validation Scores:  [1. 1. 1.]
Average CV Score:  1.0
Number of CV Scores used in Average:  3


With Nearest Neighbors:  20
Cross Validation Scores:  [1. 1. 1.]
Average CV Score:  1.0
Number of CV Scores used in Average:  3
```

DATA SET 1 (3 FOLDS)

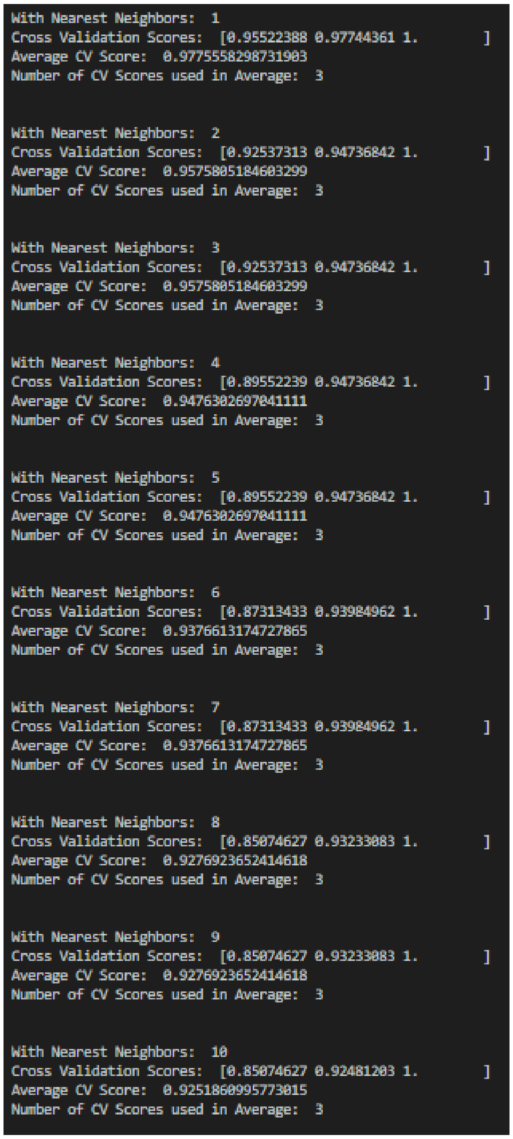


```
With Nearest Neighbors:  1
Cross Validation Scores:  [1. 1. 1.]
Average CV Score:  1.0
Number of CV Scores used in Average:  3


With Nearest Neighbors:  2
Cross Validation Scores:  [1. 1. 1.]
Average CV Score:  1.0
Number of CV Scores used in Average:  3


With Nearest Neighbors:  3
Cross Validation Scores:  [1. 1. 1.]
Average CV Score:  1.0
Number of CV Scores used in Average:  3


With Nearest Neighbors:  4
Cross Validation Scores:  [1. 1. 1.]
Average CV Score:  1.0
Number of CV Scores used in Average:  3


With Nearest Neighbors:  5
Cross Validation Scores:  [1. 1. 1.]
Average CV Score:  1.0
Number of CV Scores used in Average:  3


With Nearest Neighbors:  6
Cross Validation Scores:  [1. 1. 1.]
Average CV Score:  1.0
Number of CV Scores used in Average:  3


With Nearest Neighbors:  7
Cross Validation Scores:  [1. 1. 1.]
Average CV Score:  1.0
Number of CV Scores used in Average:  3


With Nearest Neighbors:  8
Cross Validation Scores:  [1. 1. 1.]
Average CV Score:  1.0
Number of CV Scores used in Average:  3


With Nearest Neighbors:  9
Cross Validation Scores:  [1. 1. 1.]
Average CV Score:  1.0
Number of CV Scores used in Average:  3


With Nearest Neighbors:  10
Cross Validation Scores:  [1. 1. 1.]
Average CV Score:  1.0
Number of CV Scores used in Average:  3
```

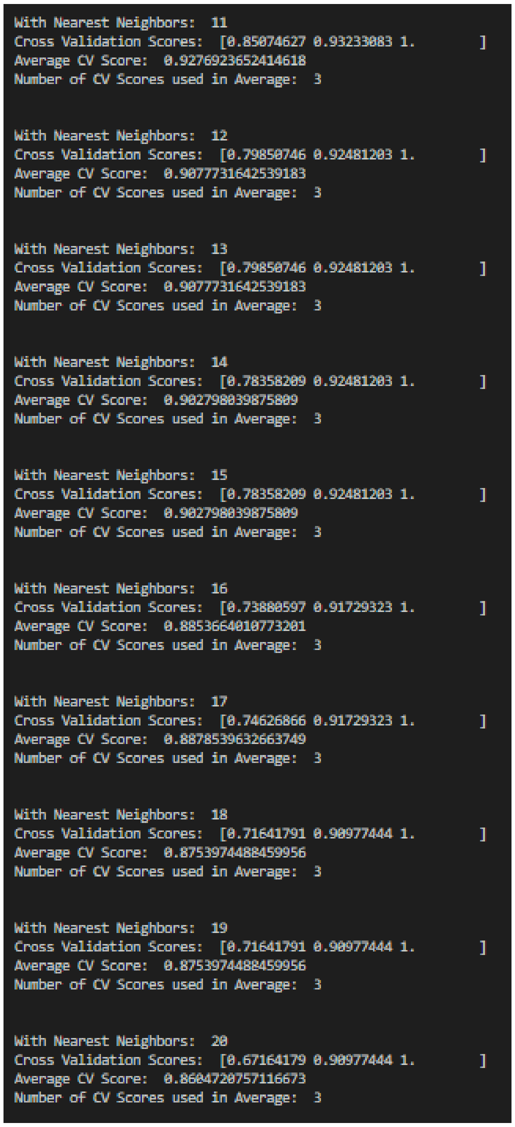


```
With Nearest Neighbors:  11
Cross Validation Scores:  [1. 1. 1.]
Average CV Score:  1.0
Number of CV Scores used in Average:  3


With Nearest Neighbors:  12
Cross Validation Scores:  [1. 1. 1.]
Average CV Score:  1.0
Number of CV Scores used in Average:  3


With Nearest Neighbors:  13
Cross Validation Scores:  [1.         1.         0.78181818]
Average CV Score:  0.9272727272727272
Number of CV Scores used in Average:  3


With Nearest Neighbors:  14
Cross Validation Scores:  [1.         1.         0.78181818]
Average CV Score:  0.9272727272727272
Number of CV Scores used in Average:  3


With Nearest Neighbors:  15
Cross Validation Scores:  [1.         1.         0.78181818]
Average CV Score:  0.9272727272727272
Number of CV Scores used in Average:  3


With Nearest Neighbors:  16
Cross Validation Scores:  [0.98181818 1.         0.78181818]
Average CV Score:  0.9212121212121213
Number of CV Scores used in Average:  3


With Nearest Neighbors:  17
Cross Validation Scores:  [0.98181818 1.         0.78181818]
Average CV Score:  0.9212121212121213
Number of CV Scores used in Average:  3


With Nearest Neighbors:  18
Cross Validation Scores:  [0.96363636 1.         0.78181818]
Average CV Score:  0.9151515151515152
Number of CV Scores used in Average:  3


With Nearest Neighbors:  19
Cross Validation Scores:  [0.97272727 1.         0.78181818]
Average CV Score:  0.9181818181818181
Number of CV Scores used in Average:  3


With Nearest Neighbors:  20
Cross Validation Scores:  [0.93636364 1.         0.78181818]
Average CV Score:  0.9060606060606061
Number of CV Scores used in Average:  3
```

DATA SET 2 (3 FOLDS)



```
With Nearest Neighbors:  1
Cross Validation Scores:  [0.95522388 0.97744361 1.        ]
Average CV Score:  0.9775558298731903
Number of CV Scores used in Average:  3


With Nearest Neighbors:  2
Cross Validation Scores:  [0.92537313 0.94736842 1.        ]
Average CV Score:  0.9575805184603299
Number of CV Scores used in Average:  3


With Nearest Neighbors:  3
Cross Validation Scores:  [0.92537313 0.94736842 1.        ]
Average CV Score:  0.9575805184603299
Number of CV Scores used in Average:  3


With Nearest Neighbors:  4
Cross Validation Scores:  [0.89552239 0.94736842 1.        ]
Average CV Score:  0.9476302697041111
Number of CV Scores used in Average:  3


With Nearest Neighbors:  5
Cross Validation Scores:  [0.89552239 0.94736842 1.        ]
Average CV Score:  0.9476302697041111
Number of CV Scores used in Average:  3


With Nearest Neighbors:  6
Cross Validation Scores:  [0.87313433 0.93984962 1.        ]
Average CV Score:  0.9376613174727865
Number of CV Scores used in Average:  3


With Nearest Neighbors:  7
Cross Validation Scores:  [0.87313433 0.93984962 1.        ]
Average CV Score:  0.9376613174727865
Number of CV Scores used in Average:  3


With Nearest Neighbors:  8
Cross Validation Scores:  [0.85074627 0.93233083 1.        ]
Average CV Score:  0.9276923652414618
Number of CV Scores used in Average:  3


With Nearest Neighbors:  9
Cross Validation Scores:  [0.85074627 0.93233083 1.        ]
Average CV Score:  0.9276923652414618
Number of CV Scores used in Average:  3


With Nearest Neighbors:  10
Cross Validation Scores:  [0.85074627 0.92481203 1.        ]
Average CV Score:  0.9251860995773015
Number of CV Scores used in Average:  3
```



```
With Nearest Neighbors:  11
Cross Validation Scores:  [0.85074627 0.93233083 1.        ]
Average CV Score:  0.9276923652414618
Number of CV Scores used in Average:  3


With Nearest Neighbors:  12
Cross Validation Scores:  [0.79850746 0.92481203 1.        ]
Average CV Score:  0.9077731642539183
Number of CV Scores used in Average:  3


With Nearest Neighbors:  13
Cross Validation Scores:  [0.79850746 0.92481203 1.        ]
Average CV Score:  0.9077731642539183
Number of CV Scores used in Average:  3


With Nearest Neighbors:  14
Cross Validation Scores:  [0.78358209 0.92481203 1.        ]
Average CV Score:  0.902798039875809
Number of CV Scores used in Average:  3


With Nearest Neighbors:  15
Cross Validation Scores:  [0.78358209 0.92481203 1.        ]
Average CV Score:  0.902798039875809
Number of CV Scores used in Average:  3


With Nearest Neighbors:  16
Cross Validation Scores:  [0.73880597 0.91729323 1.        ]
Average CV Score:  0.8853664010773201
Number of CV Scores used in Average:  3


With Nearest Neighbors:  17
Cross Validation Scores:  [0.74626866 0.91729323 1.        ]
Average CV Score:  0.8878539632663749
Number of CV Scores used in Average:  3


With Nearest Neighbors:  18
Cross Validation Scores:  [0.71641791 0.90977444 1.        ]
Average CV Score:  0.8753974488459956
Number of CV Scores used in Average:  3


With Nearest Neighbors:  19
Cross Validation Scores:  [0.71641791 0.90977444 1.        ]
Average CV Score:  0.8753974488459956
Number of CV Scores used in Average:  3


With Nearest Neighbors:  20
Cross Validation Scores:  [0.67164179 0.90977444 1.        ]
Average CV Score:  0.8604720757116673
Number of CV Scores used in Average:  3
```

I ran my tests using SciKit Learn’s machine learning Python library with the help of the following resources to build my test data and implement K-Fold cross-validation:

> https://w3schools.com/python/python_ml_cross_validation.asp
>
> https://www.w3schools.com/python/python_ml_knn.asp

To figure out which K resulted in the best accuracy for each test data, I ran each one with K=1 through K=20.

FINDINGS WITH N=5 FOLDS

DATA SET 5

All different K values returned an average cross-validation score of 1.0

DATA SET 1

All different K values returned an average cross-validation score of 1.0 up to and including K=14. From K=15 to K=20, an average Cross-Validation score of 0.9303030303030303 was reported for all K.

DATA SET 2

Data set 2 proved to be the most sporadic in cross-validation. As K increase, the Cross-Validation scores decreased. The highest average CV score was 0.9800000000000001 when K=1 and the lowest was 0.89 when K=20.

FINDINGS WITH N=3 FOLDS

DATA SET 5

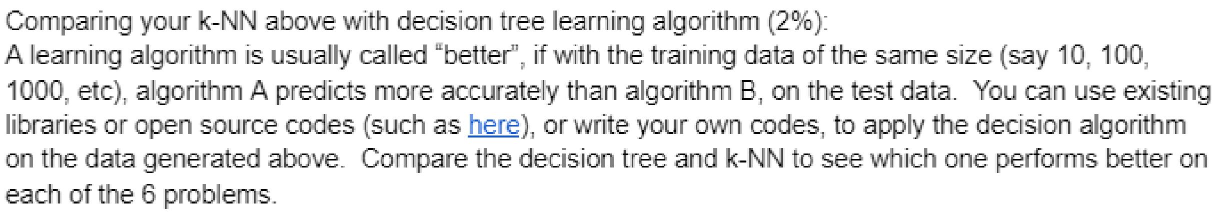
All different K values returned an average cross-validation score of 1.0

DATA SET 1

All K <= 12 returned an average cross-validation score of 1.0. The CV score then remained the same for all 13 <= K <= 17 with a score of 0.9212121212121213. K=18, K=19, and K=20 all dropped in average CV scores with each increment, the lowest being K=20 with an average CV score of 0.906060606060601.

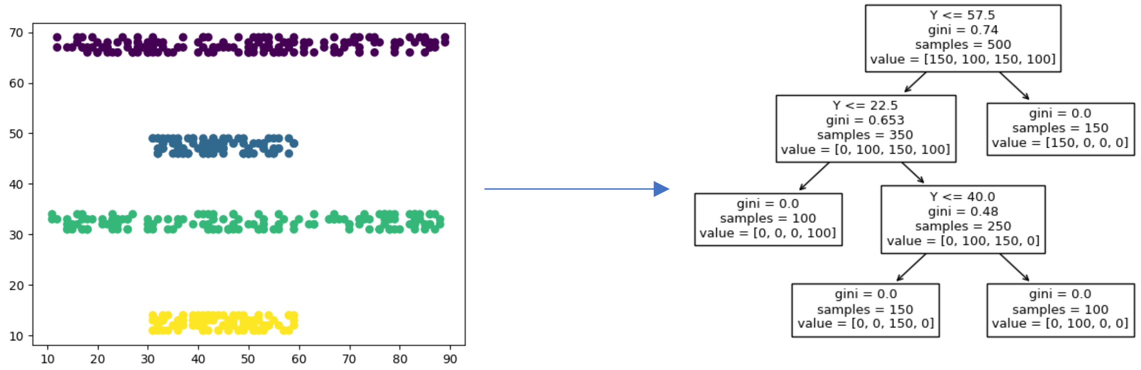
DATA SET 2

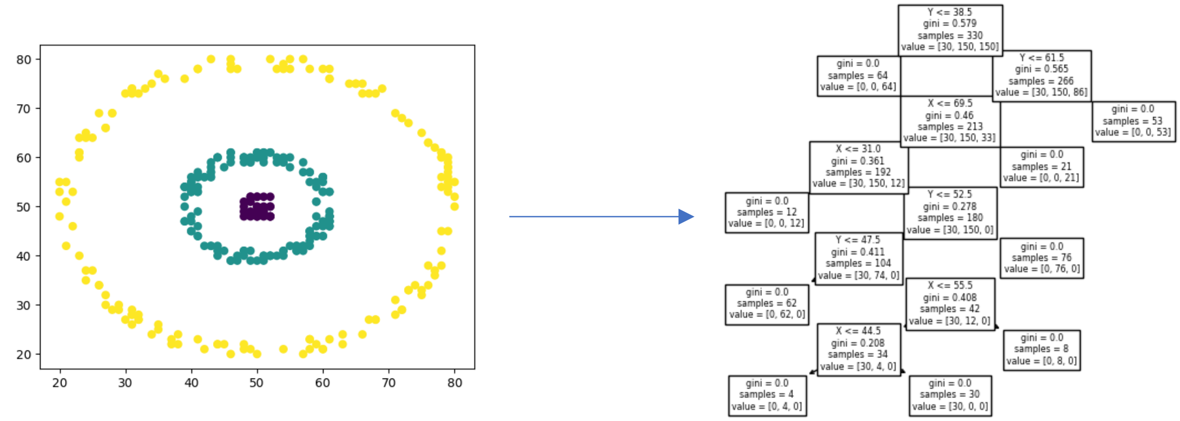
K=1 once again returned the highest cross-validation score with approximately 0.978. As K increased, the average CV score decreased. For example, K=5 returned an average cross-validation score of approximately 0.948, K=10 returned an average cross-validation score of approximately 0.925, K=15 returned an average cross-validation score of approximately 0.90, and K=20 returned an average cross-validation score of approximately 0.86.

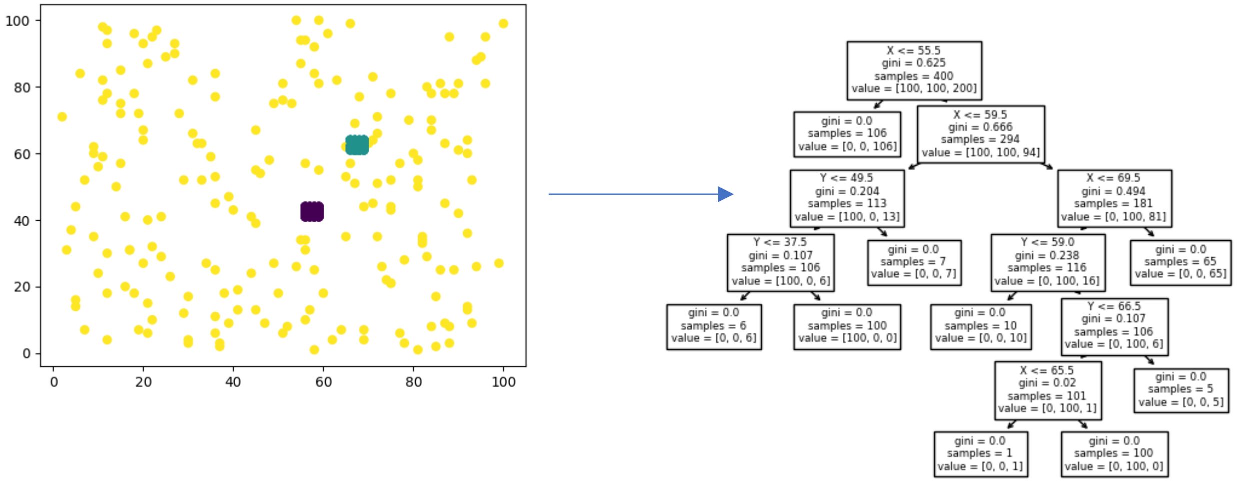


Comparing your k-NN above with decision tree learning algorithm (2%):
A learning algorithm is usually called "better", if with the training data of the same size (say 10, 100, 1000, etc), algorithm A predicts more accurately than algorithm B, on the test data. You can use existing libraries or open source codes (such as here), or write your own codes, to apply the decision algorithm on the data generated above. Compare the decision tree and k-NN to see which one performs better on each of the 6 problems.

I used the same test data as the question above. Again, using SciKit Learn's machine learning library, I called on the decision tree classifier for the three data sets, which produced the following trees:

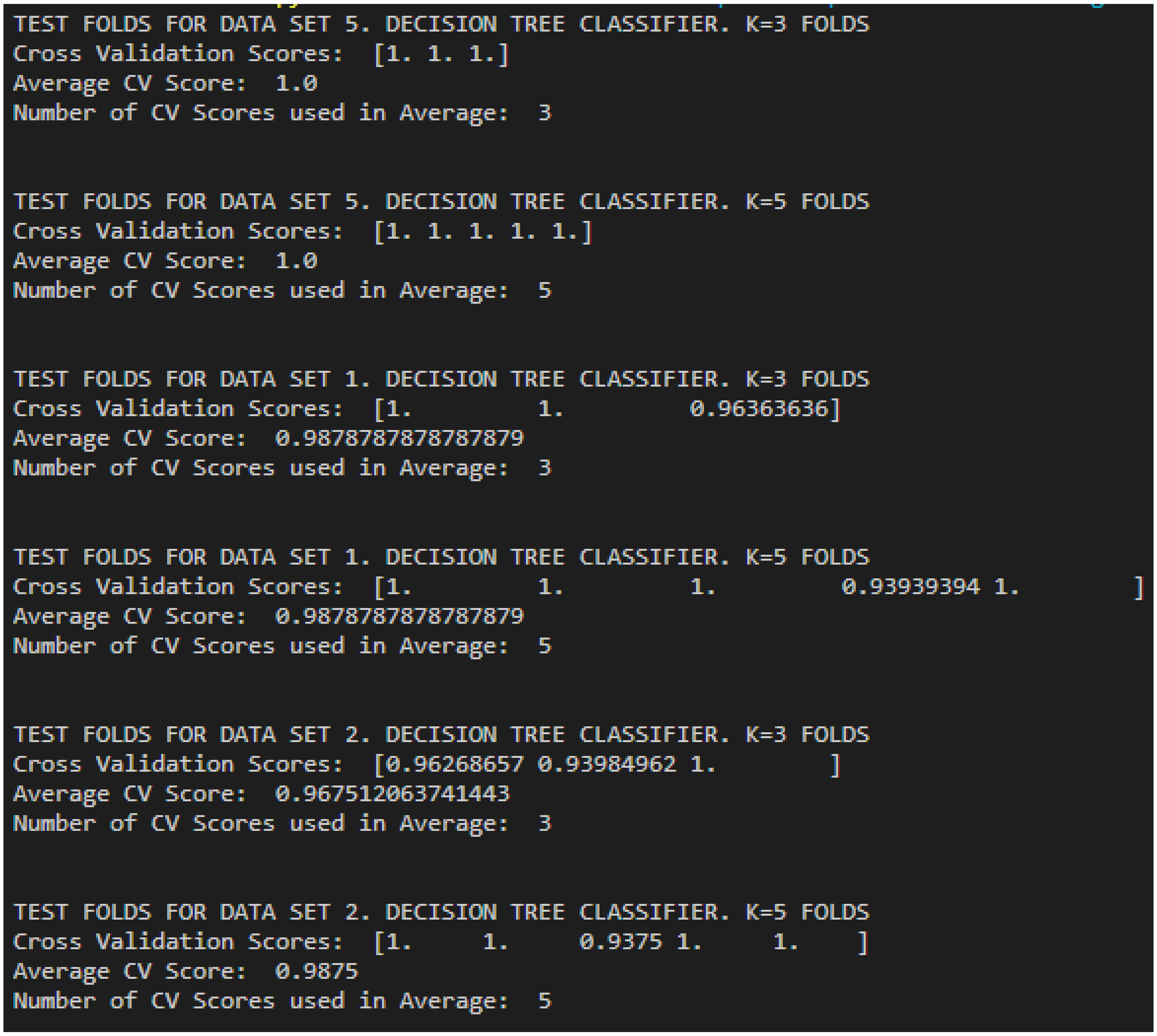






Connor McGoey

PROGRAM OUTPUT FOR DECISION TREE



```
TEST FOLDS FOR DATA SET 5. DECISION TREE CLASSIFIER. K=3 FOLDS
Cross Validation Scores:  [1. 1. 1.]
Average CV Score:  1.0
Number of CV Scores used in Average:  3


TEST FOLDS FOR DATA SET 5. DECISION TREE CLASSIFIER. K=5 FOLDS
Cross Validation Scores:  [1. 1. 1. 1. 1.]
Average CV Score:  1.0
Number of CV Scores used in Average:  5


TEST FOLDS FOR DATA SET 1. DECISION TREE CLASSIFIER. K=3 FOLDS
Cross Validation Scores:  [1.         1.         0.96363636]
Average CV Score:  0.9878787878787879
Number of CV Scores used in Average:  3


TEST FOLDS FOR DATA SET 1. DECISION TREE CLASSIFIER. K=5 FOLDS
Cross Validation Scores:  [1.         1.         1.         0.93939394 1.        ]
Average CV Score:  0.9878787878787879
Number of CV Scores used in Average:  5


TEST FOLDS FOR DATA SET 2. DECISION TREE CLASSIFIER. K=3 FOLDS
Cross Validation Scores:  [0.96268657 0.93984962 1.        ]
Average CV Score:  0.967512063741443
Number of CV Scores used in Average:  3


TEST FOLDS FOR DATA SET 2. DECISION TREE CLASSIFIER. K=5 FOLDS
Cross Validation Scores:  [1.     1.     0.9375 1.     1.    ]
Average CV Score:  0.9875
Number of CV Scores used in Average:  5
```

FINDINGS WITH N=5 FOLDS

DATA SET 5

The cross-validation score with N=5 folds returned an average **CV** (cross-validation) score of 1.0 across all tests done to produce a decision tree, meaning that both KNN and decision tree performed the same on this test data with N=5 folds.

DATA SET 1

The cross-validation score with N=5 folds returned an average **CV** (cross-validation) score of 0.98787878787879. This means that KNN outperformed Decision tree for this test data for all when using 1-14 nearest neighbours. However, if using 15-20 nearest neighbours, Decision tree outperformed KNN.

DATA SET 2

The cross-validation score with N=5 folds returned an average **CV** (cross-validation) score of 0.9875. The KNN classifier returned an average **CV** score of 0.98 at best with K=1 Nearest Neighbours, meaning that the Decision Tree classification performed better than any K for this data set.

FINDINGS WITH N=3 FOLDS

DATA SET 5

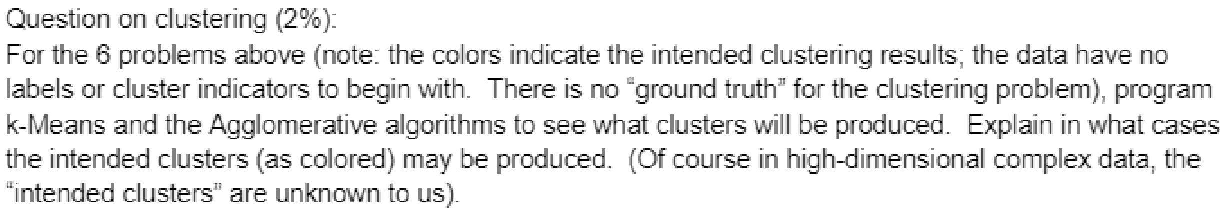
The cross-validation score with N=3 folds returned an average **CV** (cross-validation) score of 1.0 across all tests done to produce a decision tree, meaning that both KNN and decision tree performed better on this test data with N=3 folds.

DATA SET 1

The cross-validation score with N=3 folds returned an average **CV** (cross-validation) score of 0.9878787878787879 for this data set. KNN classifier returned varying results with its highest average CV score being 1.0 for all K <= 12, with all subsequent 13 <= K <= 20 being below the average CV score of the decision tree. This means that Decision Tree classification for this data set is only better than KNN if KNN is using any nearest number of neighbours greater than 12. Otherwise, KNN outperformed decision tree.

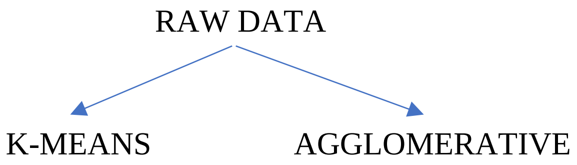
DATA SET 2

The cross-validation score with N=3 folds returned an average **CV** (cross-validation) score of 0.967512063741443 for this data set. KNN classification returned an average **CV** score maximum of 0.9775558298731903 with K=1 Nearest Neighbours, meaning that decision tree classification performed worse on this data set when K=1 nearest neighbours. However, for all other 2 <= K <= 20, decision tree classifier outperformed KNN.



Question on clustering (2%):
For the 6 problems above (note: the colors indicate the intended clustering results; the data have no labels or cluster indicators to begin with. There is no “ground truth” for the clustering problem), program k-Means and the Agglomerative algorithms to see what clusters will be produced. Explain in what cases the intended clusters (as colored) may be produced. (Of course in high-dimensional complex data, the “intended clusters” are unknown to us).

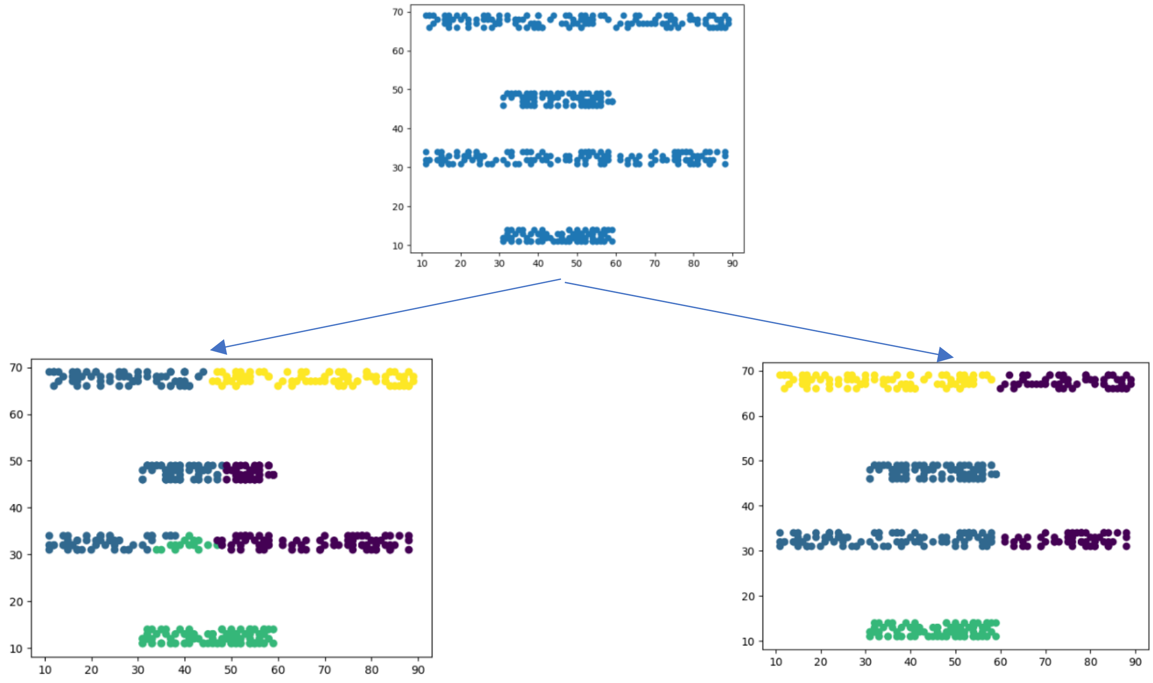
The following are the graphical representations of the clustering algorithms performed on the same sets of data as in the prior 2 questions. They are represented in the order of:



RAW DATA

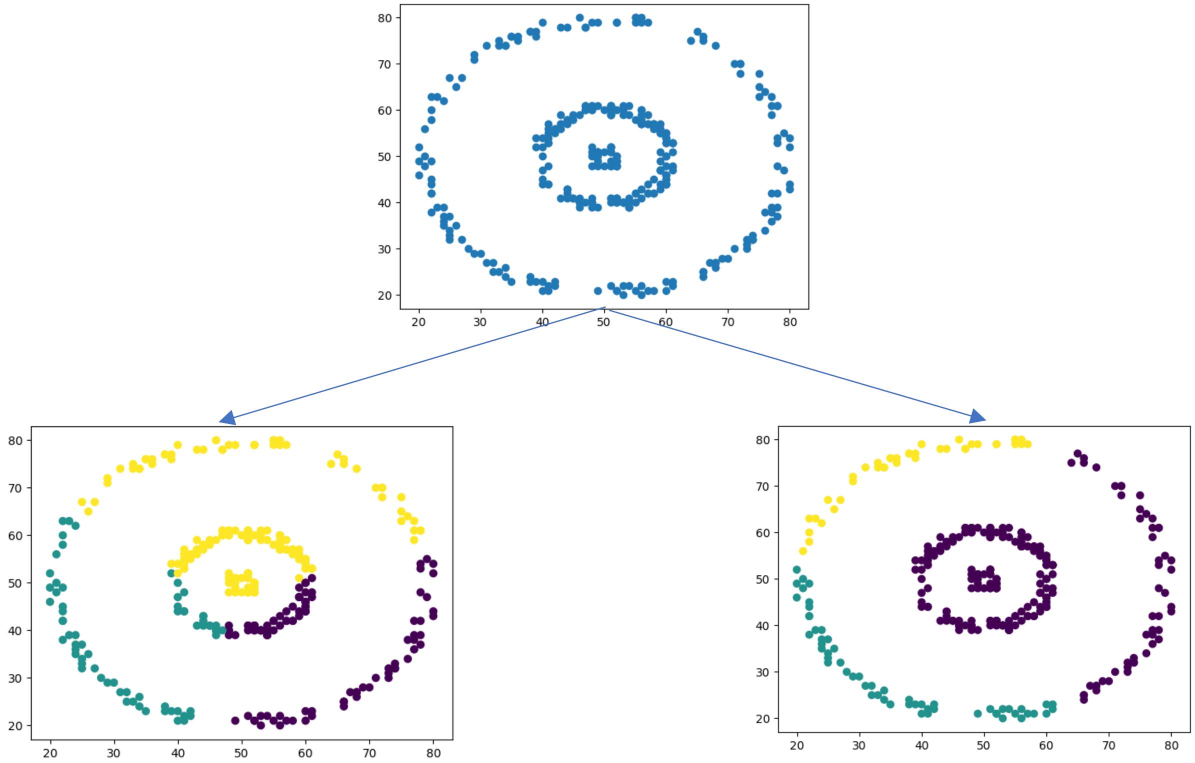
K-MEANS AGGLOMERATIVE

DATA SET 5



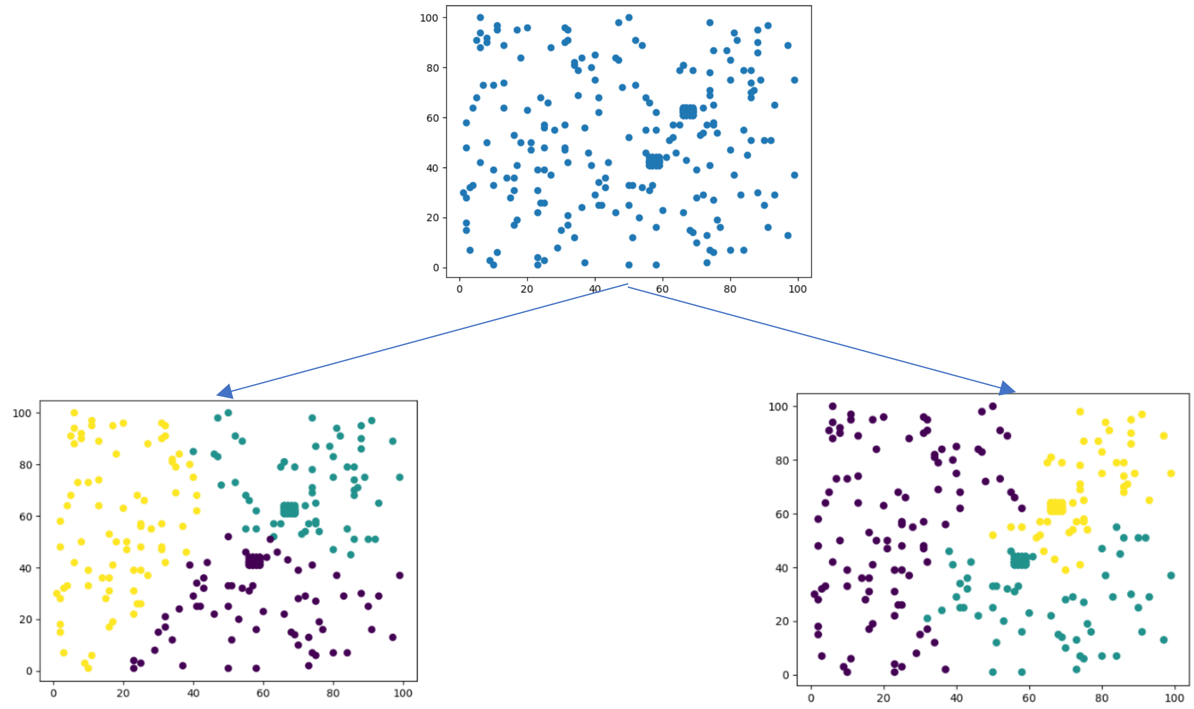
With data set 5, applying k-Means and Agglomerative clustering produced similar results on the top cluster, the cluster where 30 <= y <= 35, and the bottom cluster for most tests done. However, the cluster where 45 <= y <= 50 produced different results in most cases. I believe that for this test data, the intended clusters could be produced for k-Means in the case where the random 4 data points are assigned to each of the four clusters as that would increase the likelihood of expanding to intended clusters, the chances of this are quite low though as there are 500 total data points. For Agglomerative clustering, I believe that the intended clusters could be produced in the cases where there exists 4 pairs of data points that are all closest together than all other data points in the set, where one pair belongs to each of the 4 clusters. Additionally, if those clusters are all in the center of the clusters (x=50), this would prevent the clusters from expanding across the y-axis to an unintended cluster. In this case, I believe the intended clusters could be produced.

DATA SET 1



For data set 1, k-Means clustering produced sporadic results each time it was run, never returning the intended clusters. However, Agglomerative clustering came close to the intended clustering a few times, though usually grouping all center pieces of data together in one cluster. For k-Means, I can only think of one case where there is a possibility of producing the intended clusters. This would be when one of the randomly picked points for an initial cluster is in the center, one is in the inner ring, and the third is chosen to be on the same angle from the center of the circle as the dot chosen for the inner ring. In this case, there is a chance that the outer ring would fill with the third cluster at the same rate as the inner ring, both coming to a close fulling their intended clusters. With Agglomerative clustering, once again there needs to be 3 initial class pairs that are closest together in the 3 individual rings. However, I believe the intended clusters could be produced if the distance between each subsequence point is equal within each cluster (all points in the center are 0.3 Euclidean distance apart aside from the initial pair, all points in the inner ring are 0.7 Euclidean distance apart aside from the initial pair, and all points in the outer ring are 1.5 Euclidean distance apart aside from the initial pair).

DATA SET 2



Data set 2 was similar in both cases of k-Means clustering and Agglomerative cluster for almost all tests I ran. In both clustering classifiers, the two dense clusters both belonged to their intended clusters, but neither were contained to just those dense clusters. There were also no tests ran where the more sporadic cluster spread throughout the graph of random data points all belonged to the same cluster. I do not believe that there is any scenario for either k-Means or Agglomerative clustering where the intended clusters would be produced. The reason I believe this is that there are no limiters for the containment of the two dense clusters, nor are there any conditions in which I believe the sporadic cluster could contain those two dense clusters before they spread (even considering very specific starting cluster locations).