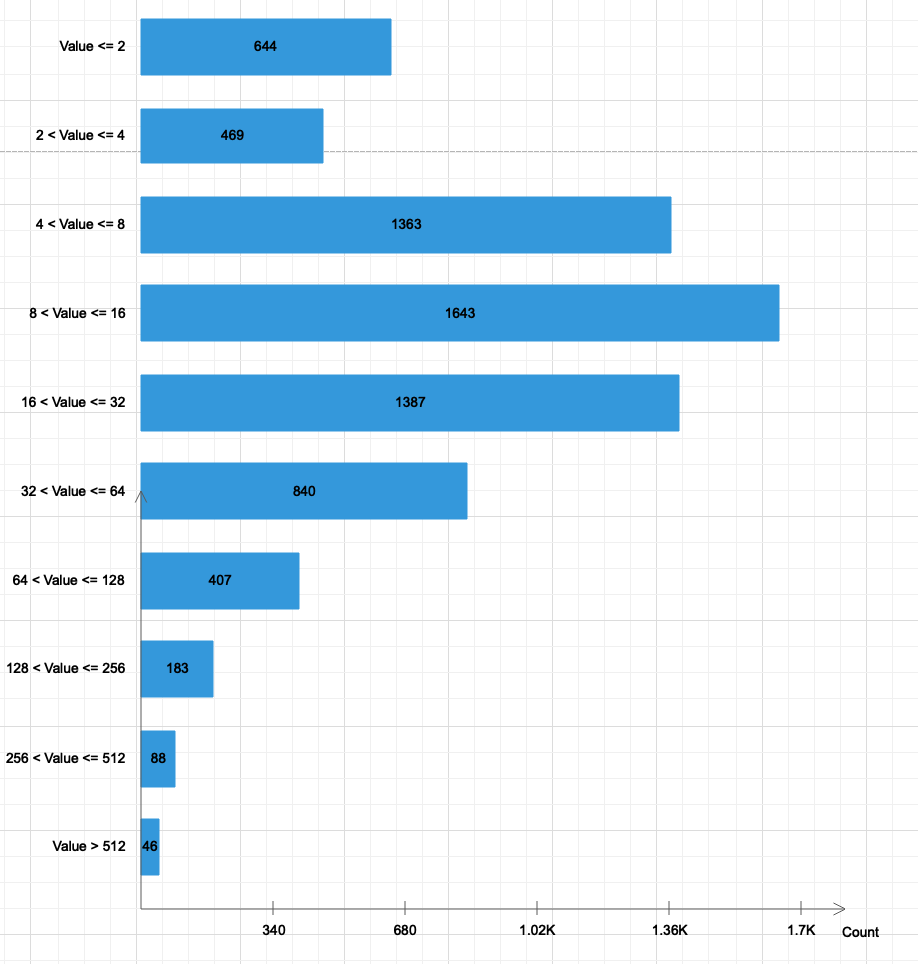
Sam Young, Joseph Munsterteiger, Noah Hoberg

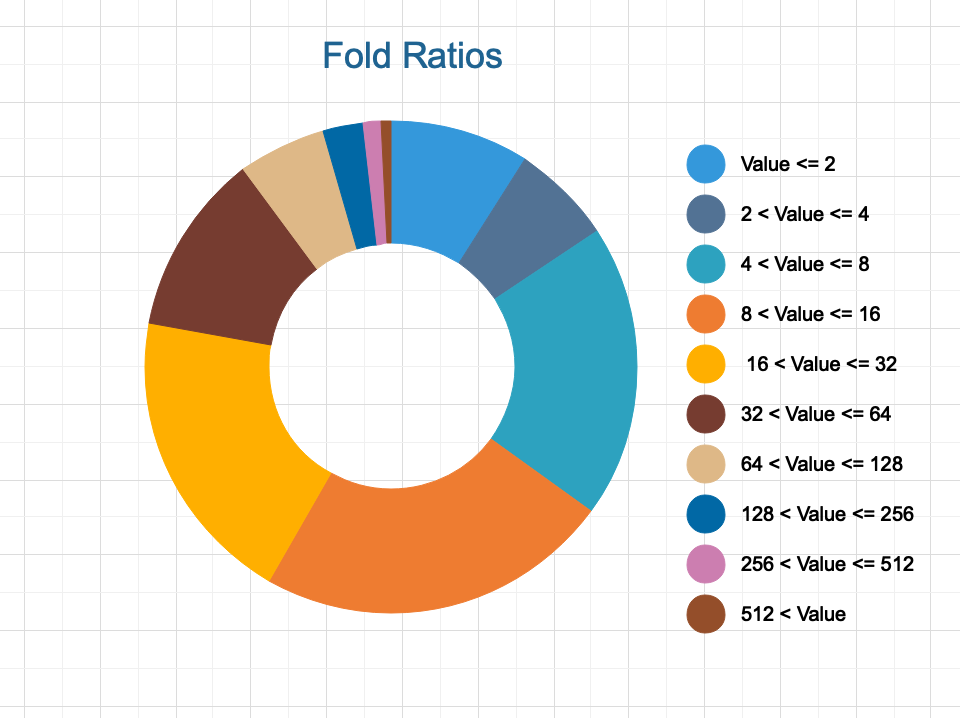
Step 1: Examining Gene Variation

1. Calculate Gene Fold Ratio Values.
2. The largest fold ratio is 800 found in 17 genes.
3. The smallest fold ratio is 1 found in 477 genes.

|  |  |
| --- | --- |
| Range | Count |
| value <= 2 | 644 |
| 2 < value <= 4 | 469 |
| 4 < value <= 8 | 1363 |
| 8 < value <= 16 | 1643 |
| 16 < value <= 32 | 1387 |
| 32 < value <= 64 | 840 |
| 64 < value <= 128 | 407 |
| 128 < value <= 256 | 183 |
| 256 < value <= 512 | 88 |
| 512 < value | 46 |

e. Used EDraw to create these graphs:





Step 2: Finding the most significant genes

1. Compute averages, standard deviations, T-value and S2N
2. Top 50 genes based on S2N:
   1. M55150\_at 1.4676
   2. X95735\_at 1.4445
   3. U50136\_rna1\_at 1.4217
   4. U22376\_cds2\_s\_at 1.3339
   5. M81933\_at 1.2040
   6. M16038\_at 1.2032
   7. M84526\_at 1.2014
   8. M23197\_at 1.1960
   9. U82759\_at 1.1926
   10. Y12670\_at 1.1847
   11. D49950\_at 1.1437
   12. M27891\_at 1.1334
   13. X59417\_at 1.1246
   14. X52142\_at 1.1226
   15. M28170\_at 1.1168
   16. X17042\_at 1.1060
   17. U05259\_rna1\_at 1.1040
   18. Y00787\_s\_at 1.0820
   19. M96326\_rna1\_at 1.0772
   20. U12471\_cds1\_at 1.0697
   21. U46751\_at 1.0641
   22. M80254\_at 1.0444
   23. M92287\_at 1.04305
   24. L13278\_at 1.0420
   25. U09087\_s\_at 1.0363
   26. L08246\_at 1.0348
   27. X74262\_at 1.0278
   28. M62762\_at 1.0234
   29. M31211\_s\_at 1.0229
   30. M28130\_rna1\_s\_at 1.0013
   31. D26156\_s\_at 0.9891
   32. M63138\_at 0.9836
   33. M31523\_at 0.9712
   34. M57710\_at 0.9677
   35. X15949\_at 0.9601
   36. M69043\_at 0.9589
   37. S50223\_at 0.9567
   38. U32944\_at 0.9549
   39. M81695\_s\_at 0.9531
   40. L47738\_at 0.9513
   41. M83652\_s\_at 0.9475
   42. X85116\_rna1\_s\_at 0.9464
   43. M11147\_at 0.9458
   44. Z15115\_at 0.9452
   45. M21551\_rna1\_at 0.9420
   46. M19045\_f\_at 0.9381
   47. X04085\_rna1\_at 0.9305
   48. L49229\_f\_at 0.9207
   49. X14008\_rna1\_f\_at 0.9150
   50. M91432\_at 0.9135
3. Top 50 genes based on T- Value:
   1. M55150\_at 8.0919
   2. U22376\_cds2\_s\_at 7.9043
   3. X59417\_at 6.8031
   4. U50136\_rna1\_at 6.4359
   5. M31211\_s\_at 6.2942
   6. M28170\_at 6.2540
   7. U82759\_at 6.2430
   8. M92287\_at 6.2174
   9. U09087\_s\_at 6.1824
   10. U05259\_rna1\_at 6.1751
   11. M81933\_at 6.1650
   12. U12471\_cds1\_at 6.1462
   13. D26156\_s\_at 6.0971
   14. L13278\_at 6.0213
   15. X74262\_at 6.0163
   16. S50223\_at 5.8586
   17. X52142\_at 5.8331
   18. X95735\_at 5.7276
   19. L47738\_at 5.7150
   20. M31523\_at 5.6775
   21. Z15115\_at 5.6155
   22. M11147\_at 5.5886
   23. D49950\_at 5.5880
   24. X63469\_at 5.5833
   25. D38073\_at 5.5630
   26. U32944\_at 5.4430
   27. U26266\_s\_at 5.4252
   28. J05243\_at 5.4063
   29. L49229\_f\_at 5.3982
   30. X15949\_at 5.3896
   31. X17042\_at 5.3886
   32. M21551\_rna1\_at 5.3773
   33. M31303\_rna1\_at 5.3706
   34. Y08612\_at 5.3666
   35. U20998\_at 5.3612
   36. AF012024\_s\_at 5.3333
   37. X56411\_rna1\_at 5.3274
   38. Y12670\_at 5.3250
   39. U29175\_at 5.3227
   40. M91432\_at 5.3009
   41. HG1612-HT1612\_at 5.2769
   42. M13792\_at 5.2223
   43. D63874\_at 5.2069
   44. U72342\_at 5.2067
   45. X97267\_rna1\_s\_at 5.2030
   46. X76648\_at 5.1952
   47. U35451\_at 5.1444
   48. Z69881\_at 5.1438
   49. D63880\_at 5.1251
   50. M62762\_at 5.1222
4. 31 Genes are common in both top 50 lists:
   1. D26156\_s\_at
   2. D49950\_at
   3. L13278\_at
   4. L47738\_at
   5. L49229\_f\_at
   6. M11147\_at
   7. M21551\_rna1\_at
   8. M28170\_at
   9. M31211\_s\_at
   10. M31523\_at
   11. M55150\_at
   12. M62762\_at
   13. M81933\_at
   14. M91432\_at
   15. M92287\_at
   16. S50223\_at
   17. U05259\_rna1\_at
   18. U9087\_s\_at
   19. U12471\_cds1\_at
   20. U22376\_cds2\_s\_at
   21. U32944\_at
   22. U50136\_rna1\_at
   23. U8279\_at
   24. X15949\_at
   25. X17042\_at
   26. X52142\_at
   27. X59417\_at
   28. X74262\_at
   29. X95735\_at
   30. Y12670\_at
   31. Z15115\_at
5. 69 different genes are in either top 50 list or in both:
   1. AF012024\_s\_at
   2. D26156\_s\_at
   3. D38073\_at
   4. D49950\_at
   5. D63874\_at
   6. D63880\_at
   7. HG1612-HT1612\_at
   8. J05243\_at
   9. L08246\_at
   10. L13278\_at
   11. L47738\_at
   12. L49229\_f\_at
   13. M11147\_at
   14. M13792\_at
   15. M16038\_at
   16. M19045\_f\_at
   17. M21551\_rna1\_at
   18. M23197\_at
   19. M27891\_at
   20. M28130\_rna1\_s\_at
   21. M28170\_at
   22. M31211\_s\_at
   23. M31303\_rna1\_at
   24. M31523\_at
   25. M55150\_at
   26. M57710\_at
   27. M62762\_at
   28. M63138\_at
   29. M69043\_at
   30. M80254\_at
   31. M81695\_s\_at
   32. M81933\_at
   33. M83652\_s\_at
   34. M84526\_at
   35. M91432\_at
   36. M92287\_at
   37. M96326\_rna1\_at
   38. S50223\_at
   39. U05259\_rna1\_at
   40. U09087\_s\_at
   41. U12471\_cds1\_at
   42. U20998\_at
   43. U22376\_cds2\_s\_at
   44. U26266\_s\_at
   45. U29175\_at
   46. U32944\_at
   47. U35451\_at
   48. U46751\_at
   49. U50136\_rna1\_at
   50. U72342\_at
   51. U82759\_at
   52. X04085\_rna1\_at
   53. X14008\_rna1\_f\_at
   54. X15949\_at
   55. X17042\_at
   56. X52142\_at
   57. X56411\_rna1\_at
   58. X59417\_at
   59. X63469\_at
   60. X74262\_at
   61. X76648\_at
   62. X85116\_rna1\_s\_at
   63. X95735\_at
   64. X97267\_rna1\_s\_at
   65. Y00787\_s\_at
   66. Y08612\_at
   67. Y12670\_at
   68. Z15115\_at
   69. Z69881\_at

**Step 3: Building the classifier**

1. See python program
2. See python program
3. See python program
4. **Confusion matrixes for every distance, k value and class label combination. Actual values are across the top and predicted values are vertical.**

**ALL matrix for k = 3 and Euclidean distance**

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | TP:16 | FP:6 |
| Negative | FN:5 | TN:7 |

**ALL matrix for k = 5 and Euclidean distance**

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | TP:16 | FP:6 |
| Negative | FN:5 | TN:7 |

**ALL matrix for k = 7 and Euclidean distance**

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | TP:17 | FP:6 |
| Negative | FN:4 | TN:7 |

**ALL matrix for k = 9 and Euclidean distance**

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | TP:17 | FP:6 |
| Negative | FN:4 | TN:7 |

**ALL matrix for k =11 and Euclidean distance**

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | TP:17 | FP:7 |
| Negative | FN:4 | TN 6 |

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

**ALL matrix for k = 3 and Manhattan distance**

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | TP:16 | FP:6 |
| Negative | FN:5 | TN:7 |

**ALL matrix for k = 5 and Manhattan distance**

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | TP:15 | FP:6 |
| Negative | FN:6 | TN:7 |

**ALL matrix for k = 7 and Manhattan distance**

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | TP:15 | FP:6 |
| Negative | FN:6 | TN:7 |

**ALL matrix for k = 9 and Manhattan distance**

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | TP:17 | FP:6 |
| Negative | FN:4 | TN:7 |

**ALL matrix for k = 11 and Manhattan distance**

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | TP:17 | FP:8 |
| Negative | FN:4 | TN:5 |

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

**ALL matrix for k = 3 and Chebyshev distance**

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | TP:14 | FP:6 |
| Negative | FN:7 | TN:7 |

**ALL matrix for k = 5 and Chebyshev distance**

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | TP:16 | FP:6 |
| Negative | FN:5 | TN:7 |

**ALL matrix for k = 7 and Chebyshev distance**

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | TP:16 | FP:6 |
| Negative | FN:5 | TN:7 |

**ALL matrix for k = 9 and Chebyshev distance**

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | TP:16 | FP:6 |
| Negative | FN:5 | TN:7 |

**ALL matrix for k = 11 and Chebyshev distance**

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | TP:16 | FP:7 |
| Negative | FN:5 | TN:6 |

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

**ALL matrix for k = 3 and Cosine similarity**

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | TP:9 | FP:8 |
| Negative | FN:12 | TN:5 |

**ALL matrix for k = 5 and Cosine similarity**

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | TP:8 | FP:8 |
| Negative | FN:13 | TN:5 |

**ALL matrix for k = 7 and Cosine similarity**

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | TP:9 | FP:8 |
| Negative | FN:12 | TN:5 |

**ALL matrix for k = 9 and Cosine similarity**

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | TP:9 | FP:8 |
| Negative | FN:12 | TN:5 |

**ALL matrix for k = 11 and Cosine similarity**

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | TP:8 | FP:8 |
| Negative | FN:13 | TN:5 |

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

**CLASS LABEL CHANGE**

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

**AML matrix for k = 3 and Euclidean distance**

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | TP: 7 | FP:5 |
| Negative | FN:6 | TN:16 |

**AML matrix for k = 5 and Euclidean distance**

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | TP:7 | FP:5 |
| Negative | FN:6 | TN:16 |

**AML matrix for k = 7 and Euclidean distance**

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | TP:7 | FP:4 |
| Negative | FN:6 | TN:17 |

**AML matrix for k = 9 and Euclidean distance**

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | TP:7 | FP:4 |
| Negative | FN:6 | TN:17 |

**AML matrix for k =11 and Euclidean distance**

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | TP:6 | FP:4 |
| Negative | FN:7 | TN:17 |

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

**AML matrix for k = 3 and Manhattan distance**

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | TP: 7 | FP: 5 |
| Negative | FN: 6 | TN: 16 |

**AML matrix for k = 5 and Manhattan distance**

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | TP: 7 | FP: 6 |
| Negative | FN: 6 | TN: 15 |

**AML matrix for k = 7 and Manhattan distance**

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | TP: 7 | FP: 6 |
| Negative | FN: 6 | TN: 15 |

**AML matrix for k = 9 and Manhattan distance**

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | TP: 7 | FP: 4 |
| Negative | FN: 6 | TN: 17 |

**AML matrix for k = 11 and Manhattan distance**

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | TP: 5 | FP: 4 |
| Negative | FN: 8 | TN: 17 |

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

**AML matrix for k = 3 and Chebyshev distance**

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | TP:7 | FP:7 |
| Negative | FN:6 | TN:14 |

**AML matrix for k = 5 and Chebyshev distance**

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | TP:7 | FP:5 |
| Negative | FN:6 | TN:16 |

**AML matrix for k = 7 and Chebyshev distance**

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | TP:7 | FP:5 |
| Negative | FN:6 | TN:17 |

**AML matrix for k = 9 and Chebyshev distance**

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | TP:7 | FP:5 |
| Negative | FN:6 | TN:16 |

**AML matrix for k = 11 and Chebyshev distance**

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | TP:7 | FP:5 |
| Negative | FN:6 | TN:16 |

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

**AML matrix for k = 3 and Cosine similarity**

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | TP: 5 | FP: 12 |
| Negative | FN: 8 | TN: 9 |

**AML matrix for k = 5 and Cosine similarity**

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | TP: 5 | FP: 13 |
| Negative | FN: 8 | TN: 8 |

**AML matrix for k = 7 and Cosine similarity**

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | TP: 5 | FP: 12 |
| Negative | FN: 8 | TN: 9 |

**AML matrix for k = 9 and Cosine similarity**

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | TP:5 | FP:12 |
| Negative | FN:8 | TN:9 |

**AML matrix for k = 11 and Cosine similarity**

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | TP:5 | FP:13 |
| Negative | FN:8 | TN:8 |

1. **Tables for precision, recall and F-1 measures for both class labels.**

**Precision measures for class label ALL**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Precision** | ***Euclidean*** | ***Chebyshev*** | ***City block*** | ***Cosine Similarity*** |
| *k=3* | *.727272727273* | *.7* | *.727272727273* | *.592411764705* |
| *k=5* | *.727272727273* | *.727272727273* | *.714285714286* | *.5* |
| *k=7* | *.739130434783* | *.727272727273* | *.714285714286* | *.592411764706* |
| *k=9* | *.73913043873* | *.727272727273* | *.73913043873* | *.592411764706* |
| *K=11* | *.70833333333* | *.695652173913* | *.68* | *.5* |

**Recall measures for class label ALL**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Recall** | ***Euclidean*** | ***Chebyshev*** | ***City block*** | ***Cosine Similarity*** |
| *k=3* | *.6956521* | *.66666666666* | *.695652173* | *.64285714* |
| *k=5* | *.695652173913* | *.6956521* | *.6818181818* | *.615384615* |
| *k=7* | *.708333333333* | *.6956521* | *.6818181818* | *.64285714* |
| *k=9* | *.708333* | *.695652* | *.708333* | *.64285* |
| *K=11* | *.73913* | *.7272727272* | *.7727272727* | *.6153846* |

**F-1 measures for class label ALL**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **ALL** | ***Euclidean*** | ***Chebyshev*** | ***City block*** | ***Cosine Similarity*** |
| *k=3* | *.711111111111* | *.682926829268* | *.711111111111* | *.58460516129* |
| *k=5* | *.711111111111* | *.711111111111* | *.697674118605* | *.551724137931* |
| *k=7* | *.723404255319* | *.711111111111* | *.697674118605* | *.58064516129* |
| *k=9* | .723404255319 | *.711111111111* | .723404255319 | *.58460516129* |
| *K=11* | .723404255319 | *.711111111111* | .723404255319 | *.551724137931* |

**Precision measures for class label AML**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **AML** | ***Euclidean*** | ***Chebyshev*** | ***City block*** | ***Cosine Similarity*** |
| *k=3* | *.583333333333* | *.5* | *.583333333333* | *.294117647059* |
| *k=5* | *.583333333333* | *.583333333333* | *.538461538462* | *.277777777778* |
| *k=7* | *.636363636364* | *.583333333333* | *.538461538462* | *.294117647059* |
| *k=9* | *.636363636364* | *.583333333333* | *.636363636364* | *. 294117647059* |
| *K=11* | *.6* | *.545454545455* | *.555555555556* | *. 277777777778* |

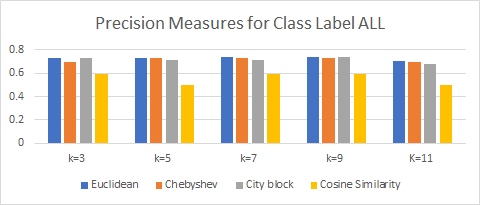
**Recall measures for class label AML**

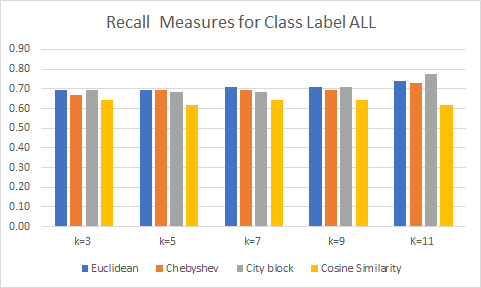
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **AML** | ***Euclidean*** | ***Chebyshev*** | ***City block*** | ***Cosine Similarity*** |
| *k=3* | *.30434782* | *.33333333333* | *.304347826087* | *.357142857143* |
| *k=5* | *.304347826* | *.30434782* | *.31818181818* | *.384615384615* |
| *k=7* | *.291666666* | *.30434782* | *.318181818* | *.357142857143* |
| *k=9* | *.291666666* | *.30434782* | *.291666666* | *.357142857143* |
| *K=11* | *.260869565217* | *.2727272727* | *.22727272* | *.384615384615* |

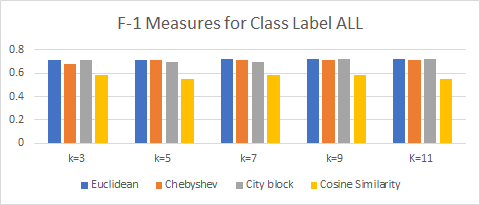
**F-1 measures for class label AML**

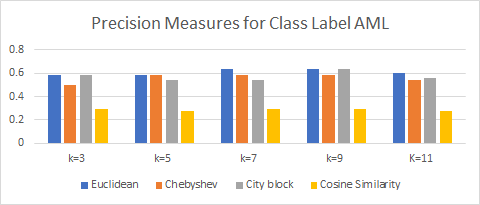
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **AML** | ***Euclidean*** | ***Chebyshev*** | ***City block*** | ***Cosine Similarity*** |
| *k=3* | *.4* | *.4* | *.4* | *.322580645161* |
| *k=5* | *.4* | *.4* | *.4* | *.322580645161* |
| *k=7* | *.4* | *.4* | *.4* | *.322580645161* |
| *k=9* | *.4* | *.4* | *.4* | *.322580645161* |
| *K=11* | *.363636363636* | *.363636363636* | *.322580645161* | *.322580645161* |

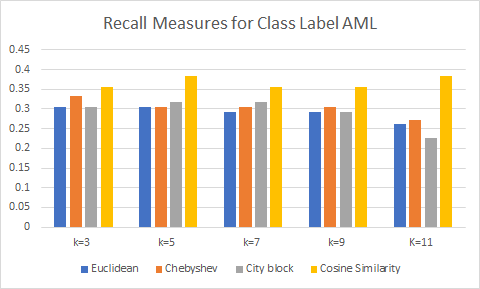
1. Graphs using data from part e

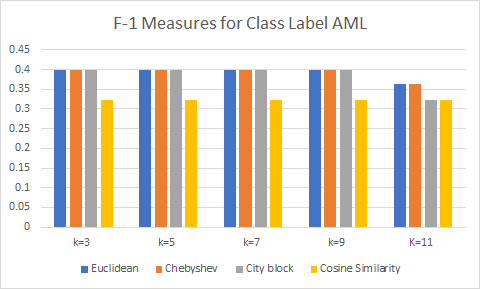












1. Repeat A-F but for all data instead of the preprocessed data. A-C is using the same program, so those parts can be seen in the python code.

**ALL matrix for k = 3 and Euclidean distance**

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | TP:15 | FP:6 |
| Negative | FN:6 | TN:7 |

**ALL matrix for k = 5 and Euclidean distance**

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | TP:15 | FP:7 |
| Negative | FN:6 | TN:6 |

**ALL matrix for k = 7 and Euclidean distance**

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | TP:16 | FP:7 |
| Negative | FN:5 | TN:6 |

**ALL matrix for k = 9 and Euclidean distance**

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | TP:17 | FP:10 |
| Negative | FN:4 | TN:3 |

**ALL matrix for k =11 and Euclidean distance**

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | TP:17 | FP:10 |
| Negative | FN:4 | TN 3 |

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

**ALL matrix for k = 3 and Manhattan distance**

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | TP:16 | FP:7 |
| Negative | FN:5 | TN:6 |

**ALL matrix for k = 5 and Manhattan distance**

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | TP:15 | FP:10 |
| Negative | FN:6 | TN:3 |

**ALL matrix for k = 7 and Manhattan distance**

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | TP:17 | FP:10 |
| Negative | FN:4 | TN:3 |

**ALL matrix for k = 9 and Manhattan distance**

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | TP:20 | FP:10 |
| Negative | FN:1 | TN:3 |

**ALL matrix for k = 11 and Manhattan distance**

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | TP:20 | FP:10 |
| Negative | FN:1 | TN:3 |

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

**ALL matrix for k = 3 and Chebyshev distance**

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | TP:16 | FP:7 |
| Negative | FN:5 | TN:6 |

**ALL matrix for k = 5 and Chebyshev distance**

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | TP:16 | FP:8 |
| Negative | FN:5 | TN:5 |

**ALL matrix for k = 7 and Chebyshev distance**

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | TP:14 | FP:7 |
| Negative | FN:7 | TN:6 |

**ALL matrix for k = 9 and Chebyshev distance**

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | TP:14 | FP:8 |
| Negative | FN:7 | TN:5 |

**ALL matrix for k = 11 and Chebyshev distance**

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | TP:15 | FP:10 |
| Negative | FN:6 | TN:3 |

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

**ALL matrix for k = 3 and Cosine similarity**

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | TP:18 | FP:10 |
| Negative | FN:3 | TN:3 |

**ALL matrix for k = 5 and Cosine similarity**

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | TP:18 | FP:11 |
| Negative | FN:3 | TN:2 |

**ALL matrix for k = 7 and Cosine similarity**

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | TP:16 | FP:11 |
| Negative | FN:5 | TN:2 |

**ALL matrix for k = 9 and Cosine similarity**

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | TP:15 | FP:12 |
| Negative | FN:6 | TN:1 |

**ALL matrix for k = 11 and Cosine similarity**

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | TP:19 | FP:11 |
| Negative | FN:2 | TN:2 |

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

**CLASS LABEL CHANGE**

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

**AML matrix for k = 3 and Euclidean distance**

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | TP: 7 | FP:6 |
| Negative | FN:6 | TN:15 |

**AML matrix for k = 5 and Euclidean distance**

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | TP:6 | FP:6 |
| Negative | FN:7 | TN:15 |

**AML matrix for k = 7 and Euclidean distance**

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | TP:6 | FP:5 |
| Negative | FN:7 | TN:16 |

**AML matrix for k = 9 and Euclidean distance**

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | TP:3 | FP:4 |
| Negative | FN:10 | TN:17 |

**AML matrix for k =11 and Euclidean distance**

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | TP:3 | FP:4 |
| Negative | FN:10 | TN:17 |

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

**AML matrix for k = 3 and Manhattan distance**

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | TP: 6 | FP: 5 |
| Negative | FN: 7 | TN: 16 |

**AML matrix for k = 5 and Manhattan distance**

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | TP: 3 | FP: 6 |
| Negative | FN: 10 | TN: 15 |

**AML matrix for k = 7 and Manhattan distance**

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | TP: 3 | FP: 4 |
| Negative | FN: 10 | TN: 17 |

**AML matrix for k = 9 and Manhattan distance**

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | TP: 3 | FP: 1 |
| Negative | FN: 10 | TN: 20 |

**AML matrix for k = 11 and Manhattan distance**

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | TP:3 | FP:1 |
| Negative | FN:10 | TN:20 |

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

**AML matrix for k = 3 and Chebyshev distance**

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | TP:6 | FP:5 |
| Negative | FN:7 | TN:16 |

**AML matrix for k = 5 and Chebyshev distance**

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | TP:5 | FP:5 |
| Negative | FN:8 | TN:18 |

**AML matrix for k = 7 and Chebyshev distance**

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | TP:6 | FP:7 |
| Negative | FN:7 | TN:14 |

**AML matrix for k = 9 and Chebyshev distance**

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | TP:5 | FP:7 |
| Negative | FN:8 | TN:14 |

**AML matrix for k = 11 and Chebyshev distance**

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | TP:3 | FP:6 |
| Negative | FN:10 | TN:15 |

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

**AML matrix for k = 3 and Cosine similarity**

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | TP: 3 | FP: 3 |
| Negative | FN: 10 | TN: 18 |

**AML matrix for k = 5 and Cosine similarity**

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | TP: 2 | FP: 3 |
| Negative | FN: 11 | TN: 18 |

**AML matrix for k = 7 and Cosine similarity**

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | TP: 2 | FP: 5 |
| Negative | FN: 11 | TN: 16 |

**AML matrix for k = 9 and Cosine similarity**

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | TP:1 | FP:6 |
| Negative | FN:12 | TN:15 |

**AML matrix for k = 11 and Cosine similarity**

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | TP:2 | FP:2 |
| Negative | FN:11 | TN:19 |

**Precision measures for class label ALL**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Precision** | ***Euclidean*** | ***Chebyshev*** | ***City block*** | ***Cosine Similarity*** |
| *k=3* | *.714285714286* | *.695652173913* | *.695652173913* | *.642857142857* |
| *k=5* | *.681818181818* | *.6* | *.66666666667* | *.620689655172* |
| *k=7* | *.695652173913* | *.666666666667* | *.62962962963* | *.592592592593* |
| *k=9* | *.62962962963* | *.636363636364* | *.66666666667* | *.555555555556* |
| *K=11* | *.62962962963* | *.6* | *.66666666667* | *.633333333333* |

**Recall measures for class label ALL**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Recall** | ***Euclidean*** | ***Chebyshev*** | ***City block*** | ***Cosine Similarity*** |
| *k=3* | *.68181818* | *.72727272* | *.72727272* | *.85714285* |
| *k=5* | *.714285714* | *.7619* | *.833333333* | *.9* |
| *k=7* | *.7272727272* | *.7* | *.85* | *.8888888* |
| *k=9* | *.85* | *.736842105* | *.869565217* | *.9375* |
| *K=11* | *.85* | *.8333333* | *.86956217391* | *.904761904* |

**F-1 measures for class label ALL**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **ALL** | ***Euclidean*** | ***Chebyshev*** | ***City block*** | ***Cosine Similarity*** |
| *k=3* | *.697674418605* | *.711111111111* | *.711111111111* | *.734693877551* |
| *k=5* | *.697674418605* | *.711111111111* | *.697674418605* | *.734693877551* |
| *k=7* | *.711111111111* | *.68292689268* | *.723404255139* | *.711111111111* |
| *k=9* | *.723404255139* | *.68292689268* | .754716981132 | *.697674418605* |
| *K=11* | *.723404255139* | *.697674418605* | .754716981132 | *.745098039216* |

**Precision measures for class label AML**

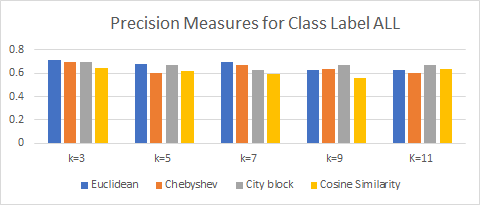
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **AML** | ***Euclidean*** | ***Chebyshev*** | ***City block*** | ***Cosine Similarity*** |
| *k=3* | *.538461538462* | *.545454545455* | *.545454545455* | *.5* |
| *k=5* | *.5* | *.333333333333* | *.5* | *.4* |
| *k=7* | *.545454545455* | *.461538461538* | *.428571428571* | *.285714285714* |
| *k=9* | *.428571428571* | *.416666666667* | *.75* | *.142857142857* |
| *K=11* | *. 428571428571* | *.333333333333* | *.75* | *.5* |

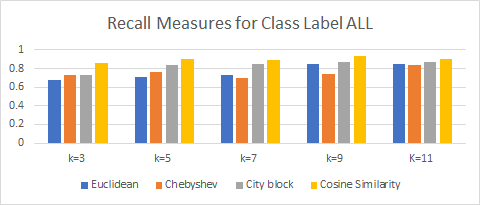
**Recall measures for class label AML**

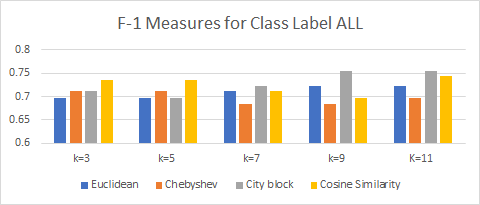
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **AML** | ***Euclidean*** | ***Chebyshev*** | ***City block*** | ***Cosine Similarity*** |
| *k=3* | *.318181818* | *.27272727* | *.2727272727272* | *.14285714* |
| *k=5* | *.285714* | *.23809523* | *.1666666* | *.1* |
| *k=7* | *.272727272727* | *.3* | *.15* | *.1111111* |
| *k=9* | *.15* | *.26315789* | *.130434782609* | *.0625* |
| *K=11* | *.15* | *.1666666667* | *.130434782* | *.0952380952* |

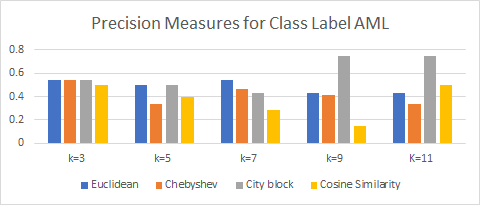
**F-1 measures for class label AML**

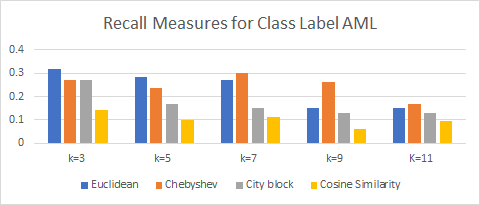
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **AML** | ***Euclidean*** | ***Chebyshev*** | ***City block*** | ***Cosine Similarity*** |
| *k=3* | *.4* | *.363636363636* | *.363636363636* | *.2222222222222* |
| *k=5* | *.363636363636* | *.322580645161* | *.222222222222* | *.16* |
| *k=7* | *.363636363636* | *.363636363636* | *.222222222222* | *.16* |
| *k=9* | *.222222222222* | *.322580645161* | *.222222222222* | *.0869565217391* |
| *K=11* | *.22222222222* | *.222222222222* | *.222222222222* | *.16* |

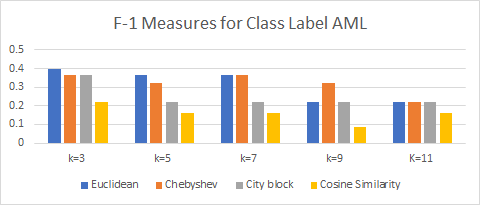












1. **Answer the following questions:**
   1. *For each dataset above, what distance/similarity measure and k value combination are best for predicting ALL class?*

For original data, the best distance and k value combination for predicting ALL is city block distance and either k = 9 or k = 11. Both of these share the highest F-1 value

For preprocessed data, the best combinations for predicting ALL is Euclidean at k = 7, Euclidean and city block at k = 9, and Euclidean and city block at k = 11. All of these share the highest F-1 value.

* 1. *For each dataset above, what distance/similarity measure and k value combination are best for predicting AML class?*

For original data, the best distance and k value combination for predicting AML is Euclidean at k=3.

For preprocessed data, the best distance and k value combination is any combination of k equaling 3, 5, 7 or 9 along with using Euclidean, Chebyshev or city block. Every combination of these shares the highest F-1 value.

* 1. *Overall, which dataset is better for our purposes: original or preprocessed?*

Original Preprocessed

Highest ALL .754 Highest ALL .723

Highest AML .4 Highest AML .4

Although the original dataset has the higher F-1 measure, the difference is negligible, and with large datasets the difference would not be worth the computational expenses.

Who Did What and Percentages

Due to the change in group we will be breaking apart what we did for part 3 and 4 as parts 1 and 2 had both been completed prior to our group formation.

**Work Day 1\*:**

Joseph and Samuel met to figure out how to read and process the data that got formed into ARFF files. Joe and Sam thought of how to compute the distances between 2 files but got stuck and decided to schedule a meeting with Imad.

**Work Day 2\*:**

All 3 group members met after the meeting with Imad. During the second work day, we were able to load the arff files in, but we still didn’t get the distances completely right because we didn’t fully understand how the file was being read in 3 huge rows, and we were having troubles with imports.

**Work Day 3\*:**

Joseph and Samuel met to fix the previous distance formulas mistakes. The code became much less computationally expensive.

**Work Day 4\*:**

All 3 group members met to use the distances to find the k nearest neighbors, to compute the weighted distances, and to make the predictions afterwards.

**Work Day 5\*:**

All 3 group members met to construct tables for the confusion matrixes, precision, recall, and F-1 measures, along with graphs.

**Work Day 6\*:**

All 3 group members met to construct even more tables and more graphs. Also, we made a readme, refactored our code to make it more legible, and finished our writeups as well as this document.

Noah Hoberg :30% Joseph Munsterteiger: 35% Samuel Young: 35%