*Study Datasets:* BELL, CHILLER, and FREESTYLE fonts from given website.

BELL contains 956 observations/ instances and 412 features.

CHILLER contains 962 observations/ instances and 412 features.

FREESTYLE contains 956 observations/ instances and 412 features.

*Preliminary treatment:*

Unneeded excess features, non-numerical values, and non- defined classes discarded.

BELL\_CLEAN contains 239 observations/ instances and 404 features.

CHILLER\_CLEAN contains 238 observations / instances and 404 features.

FREESTYLE\_CLEAN contains 239 observations/ instances and 404 features.

Then defining the three classes (CL1, CL2. And CL3) to each dataset and unionizing them into a full dataset.

DATA then contains 716 observations / instances and 404 features.

*Part 0*

Computing the mean(where m1 = mean(x1)…mean(400)=m400) and standard deviation (s1= sd(x1)..sd(x400) =sd400) of the full data set. Here we standardize and scale the data. We want to standardize the features to the center at 0 and a standard deviation of 1 because different variables can be measured at different scales leading to bias.

We can see here that the standard deviation values are not one, so we will then normalize the data by the function yj = (xj – mj)/ sj. After standardizing the data, a columns standard deviation will be extracted. Column 3’s standard deviation is one, confirming that the data set has been rescaled and standardized, which we then name SDATA. To view the table containing all means and standard deviation prior to the standardization refer to the appendix section.

We then compute the correlation matrix of the 400 features. Finding the correlation matrix will show us the correlation coefficients between the 400 features. We want to display the ten highest absolute correlation coefficient values (shown in table below). A correlation matrix only calculates the correlation between numeric features, so we took out all non-numeric data, leaving us with a 400 x 400 array.

|  |  |  |
| --- | --- | --- |
| **Pixel Positions** | | **Correlation Coefficient** |
| **Xi, Xj, Pair** | |
| r19c16 | r19c15 | 0.91627 |
| r0c4 | r0c3 | 0.91518 |
| r15c16 | r14c16 | 0.91107 |
| r19c15 | r19c14 | 0.90619 |
| r15c17 | r14c17 | 0.90612 |
| r14c18 | r13c18 | 0.90470 |
| r11c18 | r10c18 | 0.90164 |
| r0c5 | r0c4 | 0.89906 |
| r12c1 | r11c1 | 0.89357 |
| r0c3 | r0c2 | 0.88880 |

*Part 1*

1. *- Creating a TRAINSET and TESTSET*

Taking the SDATA set, we will classify CL to SETROW, where CL1 is equal to SETROW1, CL2 is equal to SETROW2, and CL3 is equal to SETROW3. We then separate each class into individual subsets of SETROW1, SETROW2, and SETROW3 to create the proper train and test set distribution. Each subset will then be split arbitrarily where TRAIN will be about 80% of the subset and TEST will be about 20% of the subset. After each subset has been split into train and test sets, all train sets will be unionized into a full TRAINSET (trainsetcl1, trainsetcl2, and trainset cl3). The same will be done to the TESTSET (testsetcl1, testsetcl2, and testsetcl3).

1.1 - *Using the K- Nearest Neighbor algorithm (KNN) on the classification of CL1, CL2 , CL3.*

To apply the KNN algorithm, a matrix with the predictors in the TRAINSET is created labeled TRAIN\_no. Similarly, a matrix containing the predictors in the TESTSET is created labeled

TEST\_no. Then two vectors are created that contain the class labels for the training observations and testing observations which are labeled TRAIN\_label and TEST\_label respectively. We then apply the knn() function on the TRAINSET and TESTSET to predict the CL classifications of each font based on grey level image intensity for pixel at each position. We set a random seed before applying knn() because if several observations are tied as nearest neighbor then R will break the tie. Here we used k = 12.

Percent correct classifications

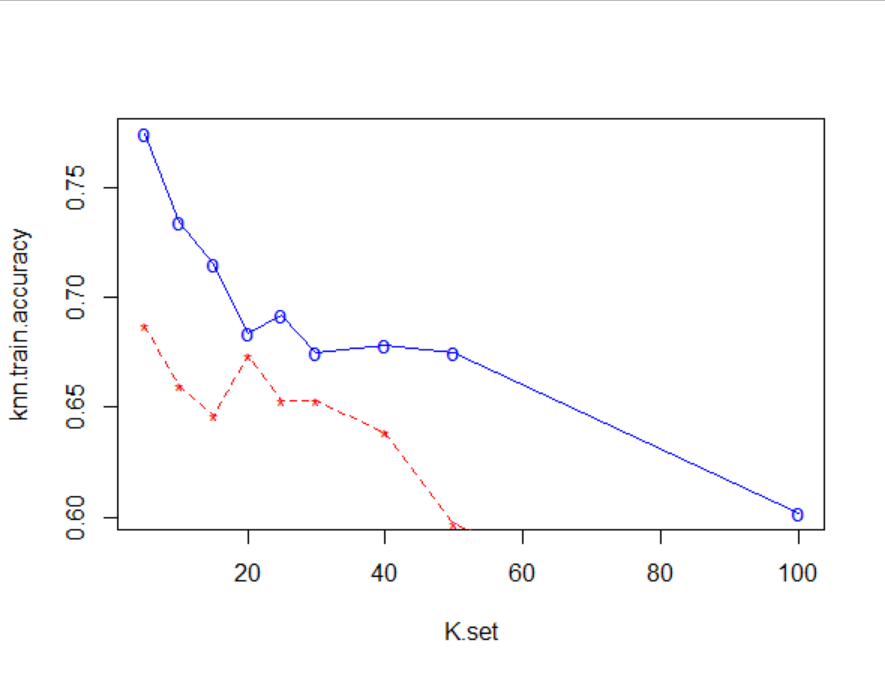
*Trainperf12* = 0.733

*Testperf12* = 0.687

We can see that the testperf accuracy is lower compared to trainperf. This makes sense because in the case of trainperf, the knn algorithm is trained and tested on the same data, while in the case of testperf, the algorithm has not been exposed to the data it is testing. Additionally, our test set is too small to conclude that it would perform similarly on new data. As we change the k value, the train set error could decline, however the test error may not. As we increase the K value, the variance will decrease but there will be a higher bias in classifier. If we decrease the K-value we might overfit the boundaries. So further exploration is needed to understand if increasing or decreasing the k value will decrease the error rate.

1.2 - *Replicating KNN algorithm for K = 5,10,15,20,30,40,50 and 100 and visualizing it.*

By using the same prior KNN function, to replicate this for values 5,10,15,20,30,40, 50 and

100 a loop will be created to identify the percent accuracy. This will be done to both the train and the test set. Then plotted to visualize the best range of k value. The plot of the k values will indicate if the values will overfit between each set.

Percent accuracy of each K value *Testperf5=* 68.75%

*Testperf10=* 66.66%

*Testperf15=* 66.66%

*Testper20=* 68.05%

*Testperf30=* 61.80%

*Testperf40=* 59.02%

*Testperf50=* 58.33%

*Testperf100=* 58.33%

Based on these values we can conclude

that where k is [5 < k < 20] is at least 68%

accuracy.

*Figure 1:plot of percent accuracy with varying k with both test and train sets.*

A plot showing a negative linear correlation of percent accuracy which is obtained by using k = 5,10, 15, 20,30,40,50

100. The blue line is the trainset and the red line is the testset values. Each dot indicates the correlated k value and accuracy.

According to the curves of train and test sets, we can see that we have slightly overfit model, where the test error is higher than the train error by (3%). The discrepancy between the 2 curves show the magnitude of overfit. Allowing us to select knn =5 as the best fit of accuracy.

1.3 - *Identifying the “best” k value within the prior range.*

From the prior loop, we will run it again using a sequence function from the range [5:100] in sets of

5. Thus, testing it on 5,10,15,20,25…100. Which we can conclude that the “ideal” k value will be 5 with

the percent accuracy of 68.75%

1.4 - *KNN algorithm using K = 5 and visualizing the respected correlation matrices and confidence intervals*

Based on the prior loop, we determine that k = 5. Then, we will run KNN algorithm on both the

TESTSET and TRAINSET. After receiving the accuracy values, we will then do a 3x3 confusion

matrix on both TESTSET and TRAINSEST to better visualize the classification for each class.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Train |  | PREDICTION | | |
|  | **Test K = 5** | **CL1** | **CL2** | **CL3** |
| TRUE | **CL1** | 71.63% | 20.00% | 8.37% |
| **CL2** | 14.63% | 80.49% | 4.88% |
| **CL3** | 6.74% | 7.77% | 85.49% |

A screenshot of a cell phone

Description automatically generated

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Test |  | PREDICTION | | |
|  | **Test K = 5** | **CL1** | **CL2** | **CL3** |
| TRUE | **CL1** | 64.29% | 28.57% | 7.14% |
| **CL2** | 21.21% | 69.70% | 9.09% |
| **CL3** | 9.09% | 16.36% | 74.55% |

A screenshot of a cell phone

Description automatically generated

*Figure 2: Confusion matrices and percentage accuracy of TESTSET and TRAINSET*

Top left and bottom left are the confusion matrix for train set and test set, which the rows are the predictions and the columns are the actual values. Right side of the figure will be the corresponding accuracy percentage for each classifier.

*Train set*

Based on the figure, we can see that 154 instances of “CL1” are classified correctly as “CL1”. Then, 132

Instances of “CL2” are also classified correctly as “CL2”. “CL3” was classified correctly 165 instances.

61 instances of “CL1” were classified incorrectly as “CL2” or “CL3”. 32 instances of “CL2” were classified incorrectly as either “CL1” or “CL3”. Lastly, 28 instances of “CL3” were classified incorrectly as either “CL1” or “CL2”. Based on the percentages we can see that “CL3” was classified with the highest accuracy of 85.5%, comparatively to “CL2” at 80.5% and “CL1” at 71.6%.

*Testset*

Based on the figure, we can see that 36 instances of “CL1” are classified correctly as “CL1”. Then, 23

Instances of “CL2” are also classified correctly as “CL2”. “CL3” was classified correctly 41 instances.

20 instances of “CL1” were classified incorrectly as “CL2” or “CL3”. 10 instances of “CL2” were classified incorrectly as either “CL1” or “CL3”. Lastly, 14 instances of “CL3” were classified incorrectly as either “CL1” or “CL2”. Based on the percentages we can see that “CL3” was classified with the highest accuracy of 74.5% comparatively to “CL2” at 69.7% and “CL1” at 64.3%

Overall, we can see that “CL3” had the “best” prediction with the k value at 5.

1.5 - *Determining the confidence interval for both TESTSET and TRAINSET.*

After determining the confusion matrix, we seek to find the confidence intervals diagonals to determine they overlap or are nearly disjoint with each other.

*Confidence interval*

Trainset: 95% CI : (0.7531, 0.8199)

Testset: 95% CI : (0.6150, 0.7638)

We can see that the confidence intervals overlap with each other, which is evidence that quantities are

close to each other. However, we cannot conclude that trainset quantile is greater than testset quantile.

1.6 - Creating the individual packages from the standardize full data set (SDATA).

We going to separate the data into 4 packs (PACK1, PACK2, PACK3, and PACK 4). Each pack will have 100 features corresponded to the 100 pixel intensities. We will divide the pack by rLcM as follows:

PACK 1: L = 0, 1, 2, 3, 4, 5, 6, 7, 8, 9

M = 0, 1, 2, 3, 4, 5, 6, 7, 8, 9

PACK 2: L = 0, 1, 2, 3, 4, 5, 6, 7, 8, 9

M = 10, 11, 12, 13, 14, 15, 16, 17, 18, 19

PACK 3: L = 10, 11, 12, 13, 14, 15, 16, 17, 18, 19

M = 10, 11, 12, 13, 14, 15, 16, 17, 18, 19

PACK 4: L = 10, 11, 12, 13, 14, 15, 16, 17, 18, 19

M = 0, 1, 2, 3, 4, 5, 6, 7, 8, 9

For each pack, we create train (80%) and test (20%) subsets. Then we will apply KNN classification using Kbest = 5 to each pack and assign the accuracy percentage as weighted values w1, w2, w3, and w4.

Accuracy of PACK 1

*W1 = testperfbestk = 58%*

1.7 - Finding the accuracy for each PACK using KNN algorithm, where K= 5.

Using the same script from 1.6, replicating for the following PACKS.

Accuracy of PACKS

*W2 = testperfbestk = 75%*

*W3 = testperfbestk = 63%*

*W4 = testperfbestk = 66%*

According to the KNN classification to each pack, we can see that pack 2 has the highest accuracy rate of 75%. On the other hand, pack 1 resulted to be the lowest accuracy rate of 58%. This is 5 and 8 percentual points behind packs 3 and 4 respectively.

1.8 - Finding the KNN with the weighted features on global performance and TESTSET

We use the accuracy rates we found in 1.7 to weight each pack before standardizing and combining them into one data set. Pack 2 will have more weight than the other packs because it has the highest accuracy rate. The non-weighted set assumes that all features are equally important when they are not. Hence, we weigh the features to account for the importance of different features.

Weighted Trainset Confusion Matrix Weighted Trainset Confusion Matrix

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Train |  | PREDICTION | | | |
|  | **Test K = 5** | | **CL1** | **CL2** | **CL3** |
| TRUE | **CL1** | | 73.89% | 17.73% | 8.37% |
| **CL2** | | 17.05% | 77.84% | 5.11% |
| **CL3** | | 5.70% | 8.81% | 85.49% |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Test |  | PREDICTION | | |
|  | **Test K = 5** | **CL1** | **CL2** | **CL3** |
| TRUE | **CL1** | 75.47% | 15.09% | 9.43% |
| **CL2** | 8.51% | 78.72% | 12.77% |
| **CL3** | 9.09% | 6.82% | 84.09% |

The global train set accuracy = 79.07%

The global test set accuracy = 79.42%

As we see, the global test accuracy is higher than our trainset, which is a good sign that there is no overfitting. There is an increase in accuracy in CL1 and CL2 when we move from trainset to test set, while there is a slight loss of accuracy in CL3 when we move from train set to test set.

Ordinary Distance KNN Test Set Weighted Distance KNN Test Set

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Weighted |  | PREDICTION | | |
| TEST | **Test K = 5** | **CL1** | **CL2** | **CL3** |
| TRUE | **CL1** | 75.47% | 15.09% | 9.43% |
| **CL2** | 8.51% | 78.72% | 12.77% |
| **CL3** | 9.09% | 6.82% | 84.09% |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Ordinary |  | PREDICTION | | |
| TEST | **Test K = 5** | **CL1** | **CL2** | **CL3** |
| TRUE | **CL1** | 64.29% | 28.57% | 7.14% |
| **CL2** | 21.21% | 69.70% | 9.09% |
| **CL3** | 9.09% | 16.36% | 74.55% |

Ordinary KNN: Global test set accuracy = 69.51%

Weighted KNN: Global test set accuracy = 79.42%

The performance of the knn algorithm improved drastically after implementing the weighted distance knn. Our global performance increased by approximately 10%. Similarly, all three classes see an increase in accuracy by around 10%

.

*Appendix:*

Mean table:

A picture containing table, calendar

Description automatically generated

A picture containing table, calendar

Description automatically generated

Standard Deviation table:

A picture containing table

Description automatically generated

A picture containing table

Description automatically generated