**Student’s Name: Nikhil**

**Roll Number: B20219**

**Mobile No: 8949463760**

**Branch: CSE**

# a.

|  |  |  |
| --- | --- | --- |
|  | **Prediction Outcome** | |
| **True Label** | 93 | 25 |
| 19 | 200 |

Figure 1 KNN Confusion Matrix for K = 1

|  |  |  |
| --- | --- | --- |
|  | **Prediction Outcome** | |
| **True Label** | 92 | 26 |
| 9 | 210 |

Figure 2 KNN Confusion Matrix for K = 3

|  |  |  |
| --- | --- | --- |
|  | **Prediction Outcome** | |
| **True Label** | 92 | 26 |
| 10 | 209 |

Figure 3 KNN Confusion Matrix for K = 5

**b.**

Table 1 KNN Classification Accuracy for K = 1, 3 and 5

|  |  |
| --- | --- |
| **K** | **Classification**  **Accuracy (in %)** |
| 1 | **86.9** |
| 3 | **89.6** |
| 5 | **89.3** |

# Inferences:

1. The highest classification accuracy is obtained with K =3.
2. Increasing the value of K increases the prediction accuracy (however, in the above case accuracy for k=5 is less than k=3 but the difference is very slight and can be ignored. On a long-term basis, we will see more accuracy for higher k greater number of times.)
3. Increasing the value of K increases the prediction accuracy because considering larger number of neighbors will result in a higher chance of greater number of neighbors being closer to the mean of the distribution of each class, hence reducing the possibility of the neighbors being less probable to lie in their respective class, or in simple words, it reduces the possibility of the neighbors being outliers.
4. As the accuracy increases with increase in k, the number of diagonal elements increase.
5. Greater accuracy means greater number of predictions to be right. Since the diagonal elements in a confusion matrix indicate the number of correct predictions, greater accuracy follows from increase in diagonal elements.
6. As the accuracy increases with increase in k, the number of off diagonal elements decrease.
7. Greater accuracy means greater number of predictions to be right. Since the diagonal elements in a confusion matrix indicate the number of correct predictions, greater accuracy follows from increase in diagonal elements, which leads to a decrease in off-diagonal elements.

# a.

|  |  |  |
| --- | --- | --- |
|  | **Prediction Outcome** | |
| **True Label** | 111 | 7 |
| 6 | 213 |

Figure 4 KNN Confusion Matrix for K = 1 post data normalization

|  |  |  |
| --- | --- | --- |
|  | **Prediction Outcome** | |
| **True Label** | 112 | 6 |
| 4 | 215 |

Figure 5 KNN Confusion Matrix for K = 3 post data normalization

|  |  |  |
| --- | --- | --- |
|  | **Prediction Outcome** | |
| **True Label** | 112 | 6 |
| 3 | 216 |

Figure 6 KNN Confusion Matrix for K = 5 post data normalization

**b.**

Table 2 KNN Classification Accuracy for K = 1, 3 and 5 post data normalization

|  |  |
| --- | --- |
| **K** | **Classification**  **Accuracy (in %)** |
| 1 | **96.1** |
| 3 | **97.0** |
| 5 | **97.3** |

# Inferences:

1. Data normalization increases the classification accuracy.
2. Since we are classifying based on Euclidean distance, greater accuracy will be there if distances with respect to all the attributes are considered equally significant. Equal significance can only be achieved if the spread of data in all attributes is same. Same spread can be achieved by scaling the spread in all the attributes to a common spread, and min-max normalization does this job. That’s why we see a greater accuracy.
3. The highest classification accuracy is obtained with K =5.
4. Increasing the value of K increases the prediction accuracy.
5. Increasing the value of K increases the prediction accuracy because considering larger number of neighbors will result in a higher chance of greater number of neighbors being closer to the mean of the distribution of each class, hence reducing the possibility of the neighbors being less probable to lie in their respective class, or in simple words, it reduces the possibility of the neighbors being outliers.
6. As the accuracy increases with increase in k, the number of diagonal elements increase.
7. Greater accuracy means greater number of predictions to be right. Since the diagonal elements in a confusion matrix indicate the number of correct predictions, greater accuracy follows from increase in diagonal elements.
8. As the accuracy increases with increase in k, the number of off diagonal elements decrease.
9. Greater accuracy means greater number of predictions to be right. Since the diagonal elements in a confusion matrix indicate the number of correct predictions, greater accuracy follows from increase in diagonal elements, which leads to a decrease in off-diagonal elements.

|  |  |  |
| --- | --- | --- |
|  | **Prediction Outcome** | |
| **True Label** | 102 | 16 |
| 3 | 216 |

Figure 7 Confusion Matrix obtained from Bayes Classifier

* The classification accuracy obtained from Bayes Classifier is **96.5%.**

Table 3 Mean for class 0 and class 1

|  |  |  |  |
| --- | --- | --- | --- |
| **S. No.** | **Attribute Name** | **Mean** | |
| **Class 0** | **Class 1** |
|  | X\_Maximum | 273.418 | 723.656 |
|  | Y\_Maximum | 1583169.659 | 1431588.69 |
|  | Pixels\_Areas | 7779.663 | 585.967 |
|  | X\_Perimeter | 393.835 | 54.491 |
|  | Y\_Perimeter | 273.183 | 45.658 |
|  | Sum\_of\_Luminosity | 843350.275 | 62191.126 |
|  | Minimum\_of\_Luminosity | 53.326 | 96.236 |
|  | Maximum\_of\_Luminosity | 135.762 | 130.452 |
|  | Length\_of\_Conveyer | 1382.762 | 1480.018 |
|  | Steel\_Plate\_Thickness | 40.073 | 104.214 |
|  | Edges\_Index | 0.123 | 0.385 |
|  | Empty\_Index | 0.459 | 0.427 |
|  | Square\_Index | 0.592 | 0.513 |
|  | Outside\_X\_Index | 0.108 | 0.02 |
|  | Edges\_X\_Index | 0.55 | 0.608 |
|  | Edges\_Y\_Index | 0.523 | 0.831 |
|  | Outside\_Global\_Index | 0.288 | 0.608 |
|  | LogOfAreas | 3.623 | 2.287 |
|  | Log\_X\_Index | 2.057 | 1.227 |
|  | Log\_Y\_Index | 1.848 | 1.318 |
|  | Orientation\_Index | -0.314 | 0.136 |
|  | Luminosity\_Index | -0.115 | -0.116 |
|  | SigmoidOfAreas | 0.925 | 0.543 |

In Fig. 8 and 9 representing covariance matrices for class 0 and class 1 respectively the column numbers and row numbers correspond to attribute with serial number as in Table 3:

Table

Description automatically generated

Figure 8: Covariance matrix for class 0

Table

Description automatically generated

Figure 9: Covariance matrix for class 1

# Inferences:

1. The accuracy of Bayes Classifier is 94.362% which is less than the normalized K-NN model accuracy. This is because bayes classifier assumes the attributes to be independent of each other. But they might not be that independent, and this is indeed the case here as we are getting lower accuracy.
2. From the covariance matrix we see that diagonal entries tend to become smaller as we increase the dimension. The reason for this is the data is not normalized and the attributes towards the end are such that they have lesser absolute value of spread.
3. The off diagonal values represent covariance between the attributes. 2 pairs of attributes with maximum value of covariance are – i) Y\_maximum & Sum\_of\_Luminosity; ii) Sum\_of\_Luminosity & Pixel\_Areas for both the classes. 2 pairs of attributes with minimum covariance are i) Square\_Index & Edges\_Y\_Index; ii) Square\_Index & LogOfAreas for class 0, and i) Luminosity\_Index & LogXIndex; ii) Luminosity\_Index & Outside\_X\_Index for class 1,

Table 4 Comparison between classifiers based upon classification accuracy

|  |  |  |
| --- | --- | --- |
| **S. No.** | **Classifier** | **Accuracy (in %)** |
|  | KNN | 89.614 |
|  | KNN on normalized data | 97.329 |
|  | Bayes | 94.362 |

# Inferences:

1. The K-NN model on normalized data shows the highest accuracy while K-NN model on actual data shows the lowest accuracy.
2. Accuracy of KNN on normalized data > Bayes > KNN model on actual data
3. For KNN, since we are classifying based on Euclidean distance, greater accuracy will be there if distances with respect to all the attributes are considered equally significant. Equal significance can only be achieved if the spread of data in all attributes is same. Same spread can be achieved by scaling the spread in all the attributes to a common spread, and min-max normalization does this job. That’s why we see a greater accuracy for normalized KNN than KNN on actual data.
4. The accuracy of Bayes Classifier is 94.362% which is less than the normalized K-NN model accuracy. This is because bayes classifier assumes the attributes to be independent of each other. But they might not be that independent, and this is indeed the case here as we are getting lower accuracy.
5. Accuracy of KNN model on actual data is less than Bayes Classifier because it finds Euclidian distance without normalizing. Greater accuracy will be there if distances with respect to all the attributes are considered equally significant. But this is not the case in the given dataset. Some attributes lie between 0 and 1 while others range to values in millions. This leads to huge fall in accuracy for KNN model on actual data.