# Classification

|  |  |
| --- | --- |
| Name | Roll No |
| FARHAN SHOUKAT | 15L-4292 |

## Classification Algorithms used

* kNN
* Decision Trees

## Preprocessing

* Feature Ignorance
* First 42 and last 30 features are all 0. That’s why I ignored them all.
* It reduced predicting time in kNN by almost 10-15%.
* Normalization
* Scaling

## Result

|  |  |  |  |
| --- | --- | --- | --- |
| **Performance of KNN (made using kaggleTestSubset)** | | | |
| **Parameters** | **Time to Train** | **Time to Test (seconds)** | **Accuracy** |
| K=3, Minkowski, p=3, weights=’distance’, normalized | 1 | 60 | 98.834951 |
|  |  |  |  |
| K=3, Minkowski, p=3, weights=’distance’, scaled | 1 | 60 | 98.873786 |
|  |  |  |  |
| K=3, Eucl dist, weights=’distance’, normalized | 1 | 7 | 98.718 |
| K=5, Eucl dist, weights=’distance’, normalized | 1 | 7 | 98.563 |
| K=7, Eucl dist, weights=’distance’, normalized | 1 | 7 | 98.447 |
| K=8, Eucl dist, weights=’distance’, normalized | 1 | 7 | 98.485 |
|  |  |  |  |
| K=3, Eucl dist, weights=’distance’, scaled | 1 | 7 | 98.680 |
| K=5, Eucl dist, weights=’distance’, scaled | 1 | 7 | 98.563 |
| K=7, Eucl dist, weights=’distance’, scaled | 1 | 7 | 98.447 |
| K=8, Eucl dist, weights=’distance’, scaled | 1 | 7 | 98.485 |
|  |  |  |  |
| K=3, Eucl dist, weights=’distance’ | 1 | 7 | 98.330 |
| K=5, Eucl dist, weights=’distance’ | 1 | 7 | 98.291 |
| K=7, Eucl dist, weights=’distance’ | 1 | 7 | 98.214 |
| K=8, Eucl dist, weights=’distance’ | 1 | 7 | 98.097 |
|  |  |  |  |
| K=3, Eucl dist | 1 | 7 | 98.330 |
| K=5, Eucl dist | 1 | 7 | 98.252 |
| K=7, Eucl dist | 1 | 7 | 98.214 |
| K=8, Eucl dist | 1 | 7 | 98.864 |
|  |  |  |  |
| K=3, Manhattan dist | 1 | 7 | 97.786 |
| K=5, Manhattan dist | 1 | 7 | 97.786 |
| K=7, Manhattan dist | 1 | 7 | 97.398 |
| K=8, Manhattan dist | 1 | 7 | 97.204 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Performance of Decision Tree (made using kaggleTestSubset)** | | | |
| **Parameters** | **Time to Train** | **Time to Test (seconds)** | **Accuracy** |
| Entropy, Max\_depth=9, nomalized | 4 | 0 | 93.0097 |
| Entropy, Max\_depth=10, nomalized | 4 | 0 | 94.1359 |
| Entropy, Max\_depth=11, nomalized | 4 | 0 | 94.0194 |
| Entropy, Max\_depth=12, nomalized | 4 | 0 | 94.9417 |
|  |  |  |  |
| Entropy, Max\_depth=9, scaled | 4 | 0 | 93.0485 |
| Entropy, Max\_depth=10, scaled | 4 | 0 | 93.2427 |
| Entropy, Max\_depth=11, scaled | 4 | 0 | 93.0097 |
| Entropy, Max\_depth=12, scaled | 4 | 0 | 93.0834 |
|  |  |  |  |
| Entropy, Max\_depth=9 | 4 | 0 | 93.0874 |
| Entropy, Max\_depth=10 | 4 | 0 | 92.8544 |
| Entropy, Max\_depth=11 | 4 | 0 | 93.0485 |
| Entropy, Max\_depth=12 | 4 | 0 | 92.9709 |
|  |  |  |  |
| Gini, Max\_depth=9 | 4 | 0 | 92.5825 |
| Gini, Max\_depth=10 | 4 | 0 | 92.3495 |
| Gini, Max\_depth=11 | 4 | 0 | 92.7767 |
| Gini, Max\_depth=12 | 4 | 0 | 92.3495 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Best results obtained from whole test data** | | | |
| **Parameters** | **Time to Train** | **Time to Test (seconds)** | **Accuracy** |
| K=3, Minkowski, p=3, weights=’distance’, scaled | 1 | 121 | 98.640 |
|  |  |  |  |
| Entrpy, Max\_depth=10, normalized | 4 | 0 | 93.320 |

## Discussions and Conclusion

* In case of kNN, scaled data with Minkowski, p=3 and K=3 worked best when accuracy is considered but it takes alot of time. If some points of accuracy does not matter then it is preferred to use Euclidian distance because it gives very good accuracy in very short time. Additionally if K is increased or decreased it decresed accuracy and. Same happened for p.
* In case of Decision Tree, normalized data with Entropy, max\_height=10 and 11 worked best. Increasing or decreasing from those values decreased accurracy.
* I have tried many techniques and tried to further improve the results but I did not manage to dot that. I think increasing training data can improve results.
* I also tried following techniques but they gave very poor results. That’s why I did not computed their results
* RadiusNeighborsClassifier
* MinMaxScaler
* QuantileTransformer
* Binarization