Q1.

Decision Tree

Summary Table

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Classifier | Accuracy | Precision | Recall | F1 | Training Time |
| Decision Tree  Entropy, depth 5 | 0.52 | 0.5 | 0.52 | 0.48 | 0.01106 |
| Decision Tree  Entropy, depth 10 | 0.53 | 0.52 | 0.53 | 0.52 | 0.05246 |
| Decision Tree  Entropy, depth 15 | 0.56 | 0.55 | 0.56 | 0.55 | 0.06496 |
| Decision Tree  Entropy, depth 20 | 0.59 | 0.58 | 0.59 | 0.58 | 0.06730 |
| Decision Tree  Gini, depth 5 | 0.53 | 0.52 | 0.53 | 0.49 | 0.02451 |
| Decision Tree  Gini, depth 10 | 0.55 | 0.54 | 0.55 | 0.54 | 0.04212 |
| Decision Tree  Gini, depth 15 | 0.58 | 0.58 | 0.58 | 0.58 | 0.05089 |
| Decision Tree  Gini, depth 20 | 0.59 | 0.59 | 0.59 | 0.59 | 0.05395 |

We can see for both entropy and Gini computation, when depth increase, both training time and accuracy will increase, as more depth will implies more computation, and more details can be separated. For the difference between Gini and entropy, the training time at first entropy takes shorter time comparing to Gini, but later Gini computes faster than entropy. Entropy should takes longer time as the computation as entropy requires to compute split info while Gini will have one equation less than entropy. But we can see rather Gini is better than Entropy from the result that for all accuracy, precision, recall and F1, the value is a bit higher. In others words, the true-positive case has a more proportion to other cases in prediction result.

Program Output

Entropy

Calendar

Description automatically generated with medium confidenceA picture containing graphical user interface

Description automatically generatedA picture containing calendar

Description automatically generatedA picture containing graphical user interface

Description automatically generated

Gini

A picture containing calendar

Description automatically generatedA picture containing calendar

Description automatically generatedCalendar

Description automatically generated with medium confidenceA picture containing graphical user interface

Description automatically generated

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Classifier | Accuracy | Precision | Recall | F1 | Training Time |
| KNN  # Neighbors 3 | 0.48 | 0.48 | 0.48 | 0.48 | 0.01068 |
| KNN  # Neighbors 5 | 0.49 | 0.47 | 0.49 | 0.47 | 0.01031 |
| KNN  # Neighbors 7 | 0.49 | 0.47 | 0.49 | 0.47 | 0.00941 |
| Random Forest  # Estimator=10 | 0.66 | 0.66 | 0.66 | 0.64 | 0.08291 |
| Random Forest  # Estimator=50 | 0.66 | 0.66 | 0.66 | 0.65 | 0.30515 |
| Random Forest  # Estimator=100 | 0.65 | 0.67 | 0.65 | 0.63 | 0.56236 |

Generally, training time in Random Forest is much higher than KNN. The major reasons as Random Forests will generate lots of classifier and helps to better generalize the problem. And therefore, the accuracy is much higher than KNN, in which KNN is a simple and straight forward classifier.

For KNN, we can see that although accuracy will increase when number of neighbors increase from 3 to 5, while stay the same at 7, the true positive among with false positive, is smaller than false negative, which precision is smaller than recall.

For random forest, in this case we choose sklearn default value for number of estimators, 10 before 0.22 version, and 100 after 0.22 version, and a middle value 50. We can actually see 10 to 50, F1 score increases, in other words, true positive cases have larger proportion, which means the real accuracy increases. As we cannot see an actual growth in accuracy in 10 to 50, then for random forest 100, it is believed that it has been over-generalize. Therefore, no reference value for the cases. In short, increase in number of estimators will increase in accuracy, and training time.

KNN

A picture containing calendar

Description automatically generatedCalendar

Description automatically generated with low confidenceA picture containing graphical user interface

Description automatically generated

Random Forest

Calendar

Description automatically generated with low confidenceA picture containing graphical user interface

Description automatically generatedA picture containing calendar

Description automatically generated

Q2.  
一張含有 文字 的圖片

自動產生的描述

Note: result may various during runtime because it purely depends on how the training data randomly sampled, which follow the instruction in lecture slide. There is commented code to get static training set for training such that the result will be much constant (but it will violate the lecture instruction. However it can get a static minimal number of classifier for classification)

C\*(x) = sign[1.738\*C1 + 2.666\*C2+ 2.565\*C3 + 2.6179\*C4+ 3.744\*C5]

Where

C1: x>1.5, then y = -1, else y = 1

C2: x>0.5, then y = 1, else y = 1

C3: x>8.5, then y = 1, else y = -1

C4: x>6.5, then y = -1, else y = 1

C5: x>4.5, then y = 1, else y = 1