10/11/22
Sehyun Cho
CAP4612

Homework 3

Tasks:

1. Evaluation metrics(Accuracy, Precision, Recall, F1, AUC)
2. Apply those metrics on both SVM and RF classifiers.

- **Describe the algorithms/approaches/tools used:**
  - **a. <u>What it is or What it does</u>**

    Tools used: used SVM and RandomForestClassifier for classifiers declaration, for evaluation metrics needed to import sklearn libraries, and bar plot from matplotlib.pyplot. Also, used confusion matrix library.

  - **b. <u>How it does & Application</u>**

    1. From the raw data table, filtered the 'Ensembl_ID' and 'Class' label, to get pure data.

       Figure 1 – Reading raw data table

	Ensembl_ID	ENSG00000005206.15	ENSG00000083622.8	ENSG00000088970.14	ENSG00000099869.7	ENSG00000100181.20	ENSG00000104691.13
0	TCGA-05-4244-01A	2.979519	0.00000	1.894481	0.000000	0.094936	1.601225
1	TCGA-05-4250-01A	1.761075	0.00000	1.512506	0.000000	0.063790	2.260509
2	TCGA-05-4382-01A	2.527333	0.00000	1.473132	0.080562	0.314608	1.695952
3	TCGA-05-4384-01A	2.300864	0.39099	1.507538	0.029133	2.307563	2.058446
4	TCGA-05-4389-01A	2.388600	0.00000	1.870401	0.000000	0.119019	1.681496

5 rows × 12311 columns

Figure 2 – Filtered dataset

	ENSG00000005206.15	ENSG00000083622.8	ENSG00000088970.14	ENSG00000099869.7	ENSG00000100181.20	ENSG00000104691.13
0	2.979519	0.00000	1.894481	0.000000	0.094936	1.601225
1	1.761075	0.00000	1.512506	0.000000	0.063790	2.260509
2	2.527333	0.00000	1.473132	0.080562	0.314608	1.695952
3	2.300864	0.39099	1.507538	0.029133	2.307563	2.058446
4	2.388600	0.00000	1.870401	0.000000	0.119019	1.681496

5 rows × 12309 columns

    2. Save class table to the different array, since these are string values that contains only 'LUAD' and 'LUSC', make those two classes are positive(LUAD) and negative(LUSC) class, which positive for 1 and negative for 0.

       Figure 3 – Set to an array positive(LUAD) and negative(LUSC) class.

```
array([1, 1, 1, ..., 0, 0, 0])
```

    3. Set X for data table and y for class label, which target class.

4. Split the data by using train_test_split method, set 80 percent for training data, and 20 percent for test data.
5. Apply feature scaling before applying SVM classifier.
6. Create SVM classifier object and fit the data.
7. Create confusion matrix for SVM.

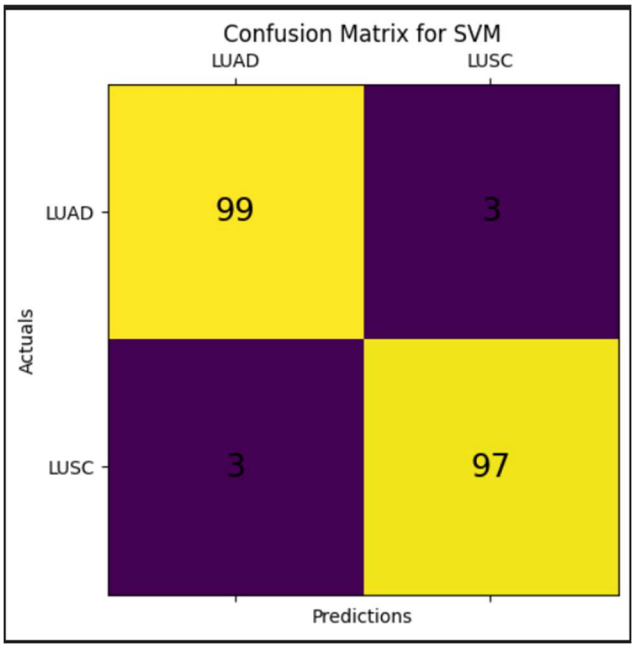

Figure 4 – Confusion matrix of SVM

8. Calculate evaluation metrics: accuracy, precision, recall, F1, AUC from the fitted data by SVM.

Figure 5 – Result of the evaluation metrics from SVM

```
Evaluation metrics in Support Vector Machine
Accuracy:  97.02970297029702
Precision:  97.0
Recall:  97.0
F1 Score:  97.0
AUC evaluation:  97.02941176470588
```

9. Visualize ROC curve with information of SVM.

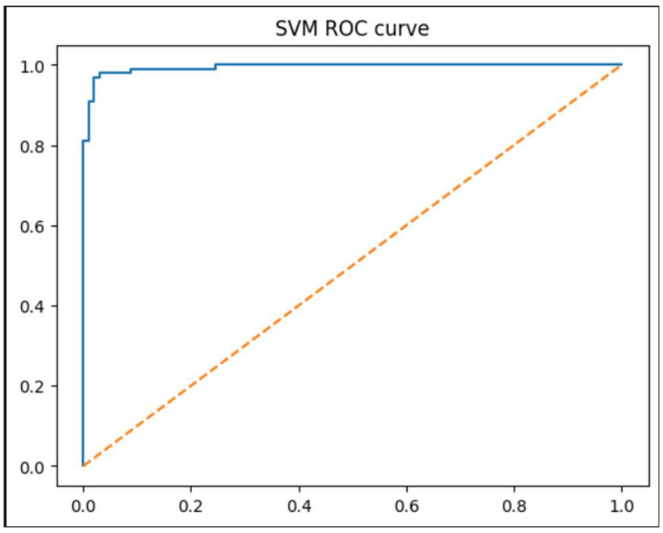

Figure 6 – SVM ROC curve

10. Now create object for Random Forest classifier to test evaluation metrics.
11. Same as SVM, set 80 percent for training data, and 20 percent for test data.
12. Feature scaling before applying Random Forest classifier.
13. Fit the data for Random Forest.
14. Create confusion matrix for Random forest.

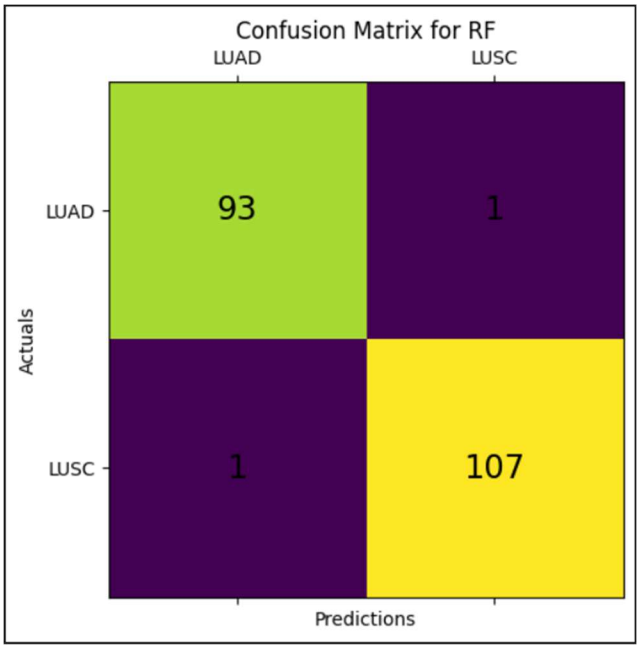

Figure 7 – Confusion matrix of Random forest

15. Calculate the evaluation metrics of the Random Forest.

Figure 8 – Result of the evaluation metrics from Random Forest

```
Evaluation metrics in Random forest
Accuracy:  99.00990099009901
Precision:  99.07407407407408
Recall:  99.07407407407408
F1 Score:  99.07407407407408
AUC evaluation:  99.00512214342002
```

16. Visualize ROC curve with information of Random Forest.

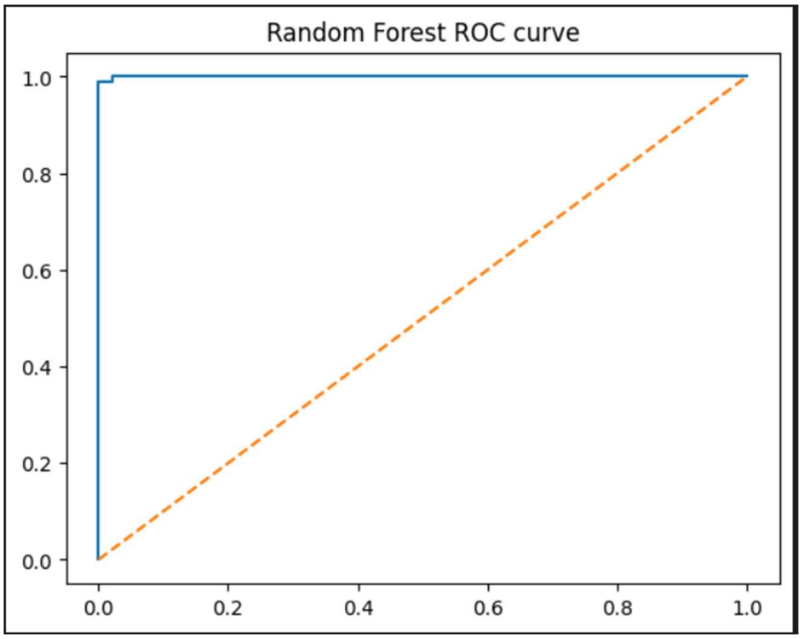

Figure 9 – Random Forest ROC curve

17. Compare the evaluation metrics between SVM and RF classifiers(Result is below).

- **Describe results:**

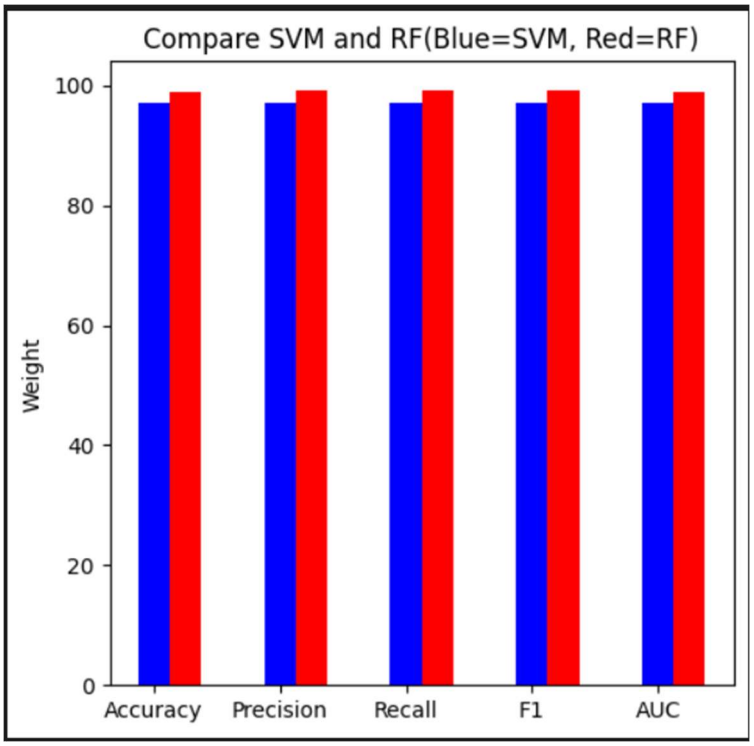

Figure 10 – Evaluation metrics comparison both SVM and RF classifier

1. **Describe the figure and table.**
   From this graph above, I made blue bars are indicated SVM and red bars indicated for RF. The x-axis shows each evaluation metrics, and y-axis shows the percentage weight.
2. **Your observation about the figure and table.**

From my observation, Random forest classifier overall performance for the evaluation metrics are having better than Support Vector Machine.

3. **Conclusion.**

   In conclusion, I could see both classifiers worked well according to the scores of evaluation metrics, but RF had slightly better results than SVM.