CDS503-Assignment_1 Report

1)a.Report the class distribution. Is this a balanced or unbalanced data set?

This is an unbalanced data set.

We can see that it has 88 samples in the classification report, and the target of class_label(0:1=55:33) with using 0 for no and 1 for yes.

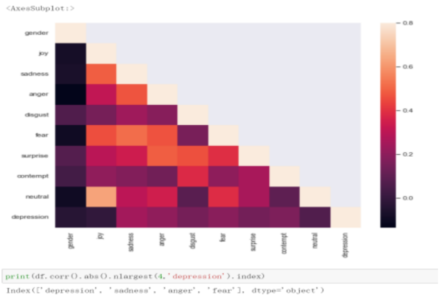
b.Please select and justify a suitable metric to evaluate the performance of your classification model.

Using a confusion matrix and the classification report ,where including weighted average of accuracy, weighted average of F1-score, weighted average of recall and weighted average of precision to judge the classification model.

c.Given the size of the data set,which validation option, do you think is suitable to be used in your machine learning experiments. Specify the validation option you are selecting for your machine learning experiments. Briefly explain the reason for using the validation option.

As the dataset has a total amount of 291 samples and 10 columns. There are no large amount of data as training samples. First, the dataset was pretested using DummyClassifier. After that, I chose to split the dataset with the test set size is 0.3. When using the SVM with Linear Kernel model to predict Freq-PHO-Binary, setting 5-fold cross validation, then the weighted average of accuracy and weighted average of F1-score can be obtained. When using the SVM with RBF Kernel, setting gamma=0.001 and 5-fold cross validation, then the weighted average of accuracy and weighted average of F1-score can be obtained. Through the KNN model, the Decision Tree model and the SVM model, they make their classification reports. For this dataset, the SVM model can be better reflected by the weighted average of F1-score. The ratio between the test set and the training set is 3:7.

2. Which feature representation produces a better model? Explain how you determine the best performing model based on the performance metric you have selected. Can you explain why one feature representation is better than the other?



```
<AxesSubplot:>
```

```
print(df.corr().abs().nlargest(4,'depression').index)
Index(['depression', 'sadness', 'anger', 'fear'], dtype='object')
```

View the relationship among the features and the target by using the correlation .We can conclude that sadness, anger, and fear have a high correlation with depression. And the sadness is the highest correlation with depression.

Using the SVM with Linear Kernel model to predict Freq-PHO-Binary and judge the “sadness” to classify labels. The weighted average of validation accuracy is 0.72, the weighted average of validation precision is 0.73, the weighted average of validation recall is 0.72, and the weighted average F1-score is 0.72. The high weighted average of accuracy and weighted average of recall show that the model can handle the data classification very well.

Using the SVM with RBF Kernel model to predict Norm-PHO-Binary and judge the “sadness” to classify labels. The weighted average of validation accuracy is 0.66, the weighted average of validation precision is 0.69, the weighted average of validation recall is 0.66, and the weighted average F1-score is 0.66. The high weighted average of accuracy and weighted average of recall show that the model can handle the data classification very well.

The emotion dataset is designed to collect a person's mood changes and determine whether the person is prone to depression, so that the negative emotions can better reflect the tendency of depression. The mood of sadness can be used as the main representative. The more sadness a person is, the more likely is the emotion of depression.

4. Is your overall best performing model able to achieve very promising results (reach at least 0.8 of your selected performance metric)? Provide two suggestions on future strategies that can be used to improve the performance of your best performing model.

This dataset can be achieved using SVM model.

(1) Collecting more data and generating more data, especially collecting more data with depression of labels, making the dataset classify labels samples counterbalanced.

(2) Using the resampling way, a minor number of samples are sampled to increase the number of samples.