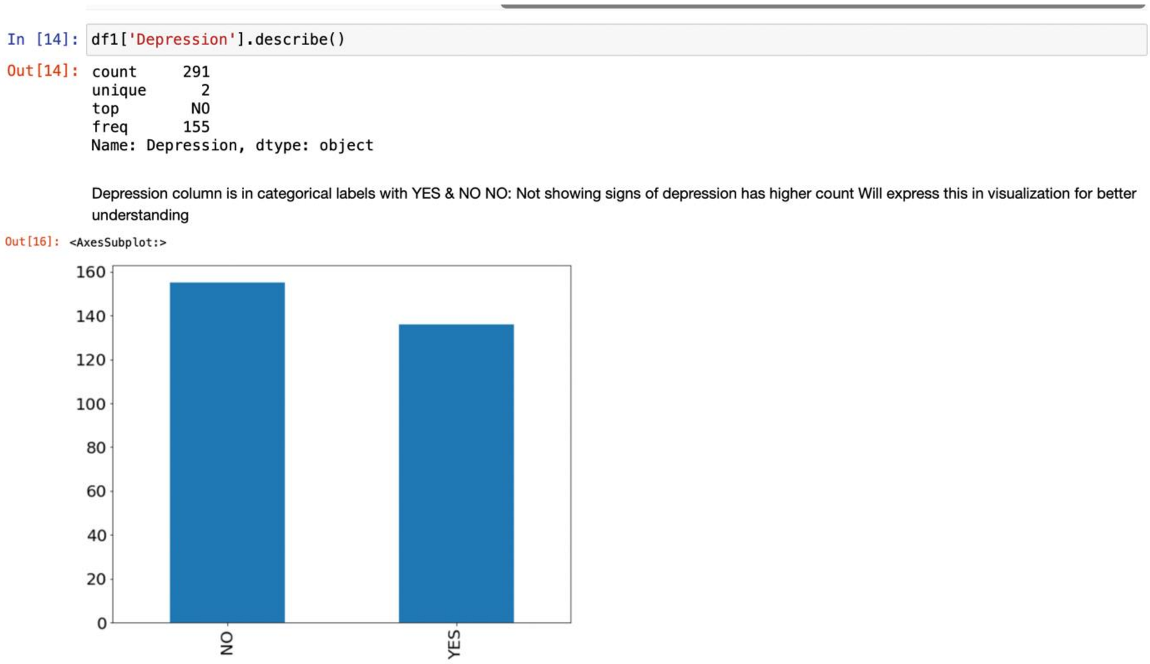
1. Report the class distribution. Is this a balanced or unbalanced data sheet?

With sample size of 291:
NO: 155 people
YES: 146 people

Class NO with higher number of people. As the difference between the classes is only 9 people, this is a pretty balanced data sheet.



```
In [14]: df1['Depression'].describe()

Out[14]: count     291
         unique      2
         top        NO
         freq      155
         Name: Depression, dtype: object
```

Depression column is in categorical labels with YES & NO NO: Not showing signs of depression has higher count Will express this in visualization for better understanding

```
Out[16]: <AxesSubplot:>
```

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

Confusion matrix is selected to evaluate the performance of classification model.
Classification Accuracy: Determine accuracy of classification problem
Misclassification Rate: Determine how often the model gives wrong predictions
Precision: Number of correct outputs provided by model out of all positive cases
Recall: Number of correct model prediction out of total positive classes
F-Measure: Evaluate the recall & precision at the same time

c. Given the size of the data set, which validation option (e.g., percentage split, k-fold cross validation) 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.

In this study, k-fold cross validation is used. As it is suitable to be applied to train and test selected models for a number of k-times on different subsets of training data and build up an estimate of the performance of a machine learning model on unseen data. For KNN, the best performing model parameters tuning is performed to find the best k value which is recorded in the Result & Discussion section.

Result & Discussion

Dummy Classifier:

	Freq-PHO-Binary		Norm-PHO-Binary	
	Validation (%)	Test (%)	Validation (%)	Test (%)
Dummy Classifier	48.7	54.2	46.6	45.8

test_size selected = 0.2, k-fold cross validation = 10

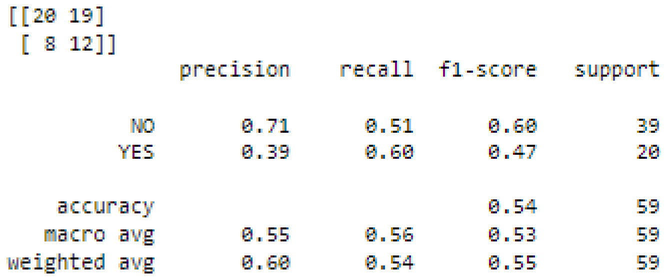
Both datasets in the study are labelled data. Hence, supervised machine learning models consists of K Nearest Neighbor (KNN), Decision Tree (DT), Bayesian Classifier (Bernoulli, Multinomial, Gaussian) and Support Vector Machine (Linear, Poly, RBF, Sigmoid) are evaluated, parameters tuning are performed to search the optimum parameter that give the highest accuracy and the performance comparisons are summarized in this section. As the dataset is multiclass, classification, KNN is expected to be the best performing model for the datasets. The performance of evaluated machine learning model is recorded in this section.

	Freq-PHO-Binary		Norm-PHO-Binary	
	Validation (%)	Test (%)	Validation (%)	Test (%)
k = 1	53.9	44.1	60.3	54.2
k = 25	58.1*	54.2*	N/A	N/A
k = 13	N/A	N/A	62.1*	61.0*
Manhattan p =1, k = 25	56.4		N/A	N/A
Manhattan p =1, k = 13	N/A	N/A	60.3	57.6

*Indicate best classifier, k

Freq-PHO-Binary:

Observation: KNN is the best performing model for this dataset. Test and validation accuracy significantly improved with best k = 25, a promising of 58.2% and 54.2% respectively. Since the difference of train & test accuracy is not big, it can be concluded that the classifier does not suffer from overfitting and generalizes well with new data. In this case, model is better at predicting at class NO (0.60) vs class YES (0.47). From medical practice it is important to predict true positive case (depression class YES) accurately so that patients receive appropriate treatment early. In this case, it can be observed that precision for class YES is low; with high number of false positive.

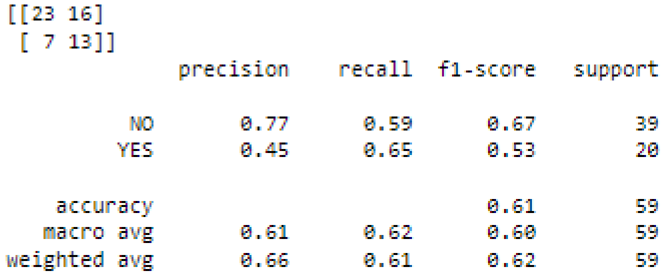


```
[[20 19]
 [ 8 12]]
              precision    recall  f1-score   support

          NO       0.71      0.51      0.60        39
         YES       0.39      0.60      0.47        20

    accuracy                           0.54        59
   macro avg       0.55      0.56      0.53        59
weighted avg       0.60      0.54      0.55        59
```

Norm-PHO-Binary

Observation: KNN is the best performing model for this dataset. Test & validation accuracy improved with best k = 13, with 62.1% and 61.0% respectively.

Since the difference of train & test accuracy is only 1.1%, it can be concluded that the classifier does not suffer from overfitting and generalizes well with new data. In comparison to Freq-PHO-Binary, this dataset gives a better performance in terms of accuracy %. Based on confusion matrix, this model is better at class NO (0.67) vs class YES (0.53). Similarly, the precision for class YES is low (0.45); with high number of false positive

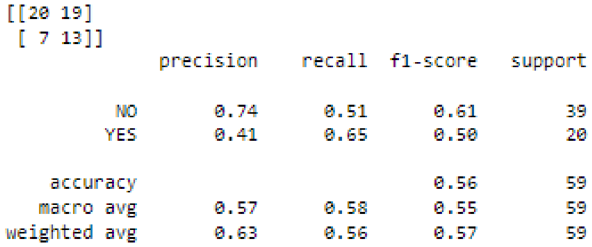


```
[[23 16]
 [ 7 13]]
              precision    recall  f1-score   support

          NO       0.77      0.59      0.67        39
         YES       0.45      0.65      0.53        20

    accuracy                           0.61        59
   macro avg       0.61      0.62      0.60        59
weighted avg       0.66      0.61      0.62        59
```

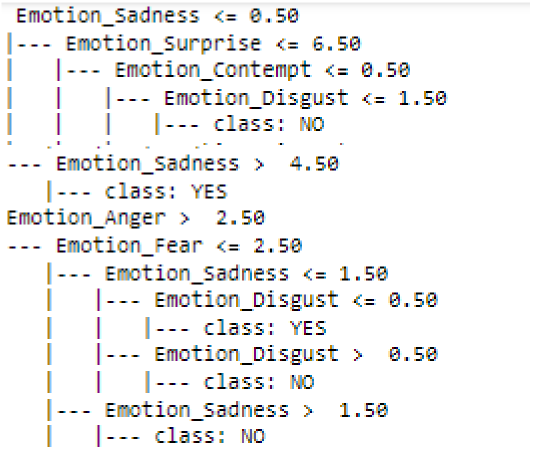
1. Decision Tree

	Freq-PHO-Binary		Norm-PHO-Binary	
	Validation (%)	Test (%)	Validation (%)	Test (%)
Default Parameters	53.9	67.7	59.6	52.5
Entropy, max_depth = 7	56.1	55.9	N/A	N/A
Entropy, max_depth = 5	N/A	N/A	52.1	47.5

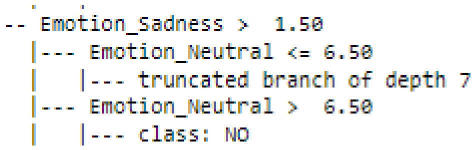
Freq-PHO-Binary
Observation: Decision Tree in this dataset with test and validation accuracy at 56.1% and 55.9% respectively. Since the difference of validation accuracy & test accuracy is not big, it can be concluded that the classifier does not suffer from overfitting and generalizes well with new data. From confusion matrix, the model is better at predicting class NO (0.61%) than YES (0.51%) based on F1 value. Interesting observation from Decision Tree, Emotion_Sadness is dependent on combinations of other emotions in predicting depression classification (refer to images below)



```
[[20 19]
 [ 7 13]]
              precision    recall  f1-score   support

          NO       0.74      0.51      0.61        39
         YES       0.41      0.65      0.50        20

    accuracy                           0.56        59
   macro avg       0.57      0.58      0.55        59
weighted avg       0.63      0.56      0.57        59
```



```
 Emotion_Sadness <= 0.50
|--- Emotion_Surprise <= 6.50
|   |--- Emotion_Contempt <= 0.50
|   |   |--- Emotion_Disgust <= 1.50
|   |   |   |--- class: NO
--- Emotion_Sadness >  4.50
    |--- class: YES
Emotion_Anger >  2.50
--- Emotion_Fear <= 2.50
    |--- Emotion_Sadness <= 1.50
    |   |--- Emotion_Disgust <= 0.50
    |   |   |--- class: YES
    |   |--- Emotion_Disgust >  0.50
    |   |   |--- class: NO
    |--- Emotion_Sadness >  1.50
    |   |--- class: NO
```



```
-- Emotion_Sadness >  1.50
   |--- Emotion_Neutral <= 6.50
   |   |--- truncated branch of depth 7
   |--- Emotion_Neutral >  6.50
   |   |--- class: NO
```

2. Bayesian Classifier
Using k fold cross validation = 10

	Freq-PHO-Binary		Norm-PHO-Binary	
	Validation (%)	Test (%)	Validation (%)	Test (%)
Bernoulli	58.3	71.2	59.0	66.1
Multinomial	61.1	66.1	58.6*	67.8*
Gaussian	52.6*	71.1*	56.0	62.7

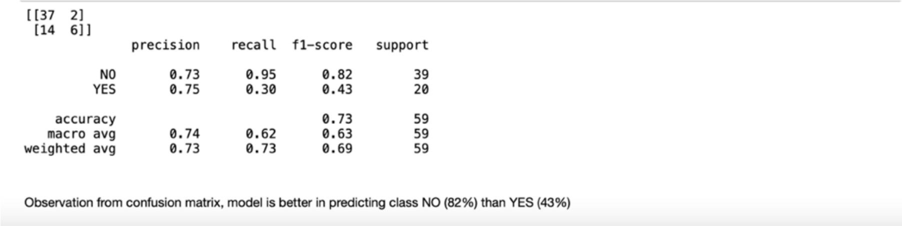
*Highest accuracy %

From the result obtained above, the test accuracy % is significantly higher than validation accuracy %. Hence it can be concluded that the datasets suffer from underfitting and does not generalize will on new data.

3. Support Vector Machine (SVM)

	Freq-PHO-Binary		Norm-PHO-Binary	
	Validation (%)	Test (%)	Validation (%)	Test (%)
Linear (Default)	59.5	67.8	60.0	62.7
Linear, best: C = 10	59.5	67.8	-	-
Linear, best: C = 100	-	-	**60.8***	**64.4***
Poly (Default)	**56.9**	**72.9***	**61.2***	**59.3***
Poly, best C = 1	**56.9**	**72.9***	-	-
Poly, best C = 10	-	-	61.2	57.6
RBF (Default)	53.9	64.4	63.4	59.3
RBF, best C = 10	59.3	54.8	-	-
RBF, best C = 1	-	-	63.4	59.3
Sigmoid (Default)	51.7	67.8	54.3	59.3
Sigmoid, best C = 100	54.8	66.1		
Sigmoid, best C = 1			54.3	59.3

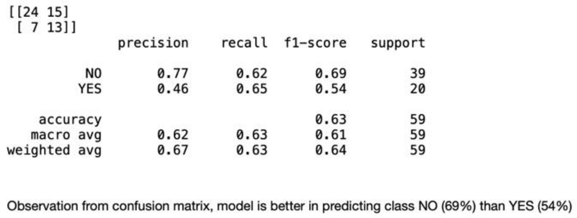
Freq-PHO-Binary:
Observation: The test accuracy % is significantly higher than validation accuracy %. Based on confusion matrix, model is significantly better in predicting class NO (84%) than YES (43%). Despite of highest test accuracy is recorded for this model, dataset suffers from underfitting and does not generalize will on new data. Hence, it can be concluded that this is not suitable to be chosen as the best performing model.



```
[[37  2]
 [14  6]]
              precision    recall  f1-score   support

          NO       0.73      0.95      0.82        39
         YES       0.75      0.30      0.43        20

    accuracy                           0.73        59
   macro avg       0.74      0.62      0.63        59
weighted avg       0.73      0.73      0.69        59
```

Observation from confusion matrix, model is better in predicting class NO (82%) than YES (43%)

Norm-PHO-Binary:
Observation: With SVM linear default parameters, test accuracy is 62.7% slightly improved to 64.4% using best C = 100. Similarly, model is better in predicting class NO (69%) than YES (54%).



```
[[24 15]
 [ 7 13]]
              precision    recall  f1-score   support

          NO       0.77      0.62      0.69        39
         YES       0.46      0.65      0.54        20

    accuracy                           0.63        59
   macro avg       0.62      0.63      0.61        59
weighted avg       0.67      0.63      0.64        59
```

Observation from confusion matrix, model is better in predicting class NO (69%) than YES (54%)

Conclusion & Future Work

From the study, it has been proven that KNN is the best performing model for Freq-PHO-Binary (58.1%) & Norm-PHO-Binary (62.1%) respectively, both are higher than dummy classifier baseline. However, both models did not reach at least 80% of accuracy rate. In future work, it is advisable to include more data (participants) in the study to form a stronger correlation of emotion attributes with depression along with feature selection to find out the best relationship subset of attribute with target (example: Emotion_Sadness, Emotion_Fear, Emotion_Disgust is often linked to depression).