1) Study the data set carefully and answer the questions below:

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

A classification problem with skewed data set is considered imbalanced data set. For example, a dataset shows whether a patient is diagnosed with cancer with more instances of Yes than No, is considered imbalanced data set.

In the above example, the patient diagnosed with cancer is called majority class while those make up smaller proportion is called minority class. The degree of imbalanced is highlighted below:

Degree of imbalance	Proportion of Minority Class
Mild	20-40% of the data set
Moderate	1-20% of the data set
Extreme	<1% of the data set

Figure 1 Degree of imbalanced
Source: Google Developers

In this assignment, the target feature is 'Depression", with "No" represents no depression and "Yes" represents depression. It later transforms into 0 and 1 respectively using label encoder. To check the data distribution, the method "value count" is used on the Depression feature, getting the result of 155 No and 136 Yes. Since the minority class is more than 40%, the data set is considered balanced.

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

The suitable metric to evaluate the performance of classification model would be F1 score.

There are four types of performance metrics: accuracy, precision, recall and F1 score. Accuracy is measure of how many instances are correctly predicted over total instances; Precision is a measure how good is the model at whatever it predicted; Recall is a measure how good is the model at picking the correct items; F1 is harmonic mean of precision and recall.

In this assignment our task is to predict the depression patient. In the case of false positive (the patient is diagnosed with Depression when he is not), he will undergo unnecessary treatment. In the case of false negative (the patient is diagnosed as No Depression when he is), he will not receive any treatment. We don't want any of the situation happen, therefore F1 is a better choice since it takes the harmonic mean between recall and precision.

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

The size of data set is 291 entries. Given the data set is tiny, it will be more appropriate to use k-fold cross validation.

Percentage split splits the data into training, validation, and testing set. Given our data entries only 291, the splitting of data will result in only small range of data been utilized into each stage, decreasing the performance of the model.

K-fold cross validation suggests that data is divided into K fold, one serves as training, remaining serve as testing and continue till every fold has been testing fold. The advantage of this method is less data is required but needs more computation power. Considering we have only 291 entries, it will be ideal in this case

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?

The classification algorithms that are going to run in this assignment are K Nearest Neighbour, Naïve Bayes and Support Vector Machine. These three algorithms are run on the Freq-PHO-binary and Norm-PHO-Binary respectively.

The first step is data cleaning and data preparation. After load data into the data frame, we check is there any missing value in the data set using df.info() function. The result is shown as below:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 291 entries, 0 to 290
Data columns (total 10 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   Gender            291 non-null    object
 1   Emotion_Joy       291 non-null    float64
 2   Emotion_Sadness   291 non-null    float64
 3   Emotion_Anger     291 non-null    float64
 4   Emotion_Disgust   291 non-null    float64
 5   Emotion_Fear      291 non-null    float64
 6   Emotion_Surprise  291 non-null    float64
 7   Emotion_Contempt  291 non-null    float64
 8   Emotion_Neutral   291 non-null    float64
 9   Depression        291 non-null    object
dtypes: float64(8), object(2)
memory usage: 22.9+ KB
```

Using this function, we understand that the data has no missing values, and all data are numerical discrete except gender and depression which is categorical norminal data.

Next, we use value count function to examine the distribution of dataset. We know that the dataset is quite balanced with more than 40% are minority group.

Next, since the machine learning requires all features represented in numerical form, one hot encoding is applied to column Gender and label encoding is applied to column Depression, where “No” is labelled as “0” while “Yes” are labelled as “1”. While label encoding is to assign numerical value to categorical data, one hot encoding is to convert categorical data into a new column and assign binary value to it. Machine algorithm often treat order of number as order of significant which means number “1” will have more weight than number “0”. To avoid this hierarchy issue, one hot encoding is applied to categorical nominal data – Gender.

Since the assignment mentions the test data will be released later, all data will be used as training set. The K-fold cross validation is used to compute the accuracy, precision, recall and F1 score of the model. The result is shown as below:

Feature Representat	Machine Learning Al	Parameters	Validation Option	Validation Accuracy	Validation Precision	Validation Recall (We	Validation F1 (Weigh	Avg)
Freq	DummyClassifier	strategy=uniform	K-fold cross validation (k =10)	0.5259	0.4371	0.4687	0.4502	
	KNN	K = 29	K-fold cross validation (k =14)	0.5634	0.5371	0.419	0.4626	
	Naïve Bayes	Bernoulli	K-fold cross validation (k = 2)	0.5567	0.5276	0.5735	0.5342	
	SVM	Linear, c = 100	K-fold cross validation (k =8)	0.6323	0.7302	0.4265	0.5051	
		Poly, c =1	K-fold cross validation (k =8)	0.5976	0.84	0.1912	0.2877	
		Rbf, c =1	K-fold cross validation (k =8)	0.6193	0.7441	0.3607	0.4437	
		Sigmoid, c = 1	K-fold cross validation (k =7)	0.5743	0.669	0.2842	0.367	
Norm	DummyClassifier	strategy=uniform	K-fold cross validation (k =10)	0.5013	0.4739	0.4429	0.5462	
	KNN	K = 29	K-fold cross validation (k =14)	0.6003	0.5688	0.6198	0.5852	
	Naïve Bayes	Bernoulli	K-fold cross validation (k = 2)	0.5616	0.5331	0.5735	0.5362	
	SVM	Linear, c = 1	K-fold cross validation (k =10)	0.5845	0.5691	0.4951	0.518	
		Poly, c =1	K-fold cross validation (k =30)	0.6226	0.5758	0.6567	0.5971	
		RBF, c = 1	K-fold cross validation (k =7)	0.6151	0.6032	0.5588	0.5705	
		Sigmoid, c = 1	K-fold cross validation (k =3)	0.5458	0.4966	0.3199	0.3821	

Figure 2 Performance of model

The best performing model is determined by looking at F1 score. The best performing algorithm has been highlighted in yellow which is Naïve Bayes in Frequency Feature Represented Model and SVM Polynomial in Normalized Feature Representation model. By comparing F1 score, we know that the Normalized Feature Representation model with mean F1 score of 43% (exclude Dummy Classifier) is better than Frequency feature Represented Model with mean F1 score of 53% (exclude Dummy Classifier). The best performing algorithm in Normalized Feature Representation model also performs better than the one in Frequency Represented Model.

The normalized feature produces better result. The normalized feature allows the dataset to use a common scale. When one of the features is far greater than another, the larger feature will affect result

more than the smaller feature. Thus, normalized feature makes both features contributes equally to the model without bias to any features. Therefore, it is performing better in this case.

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.

No, the best performing model is not able to reach very promising results.

The best prediction model produces not more 60% of accuracy. It indicates the model suffers from underfitting. The result shows that model is suffering from high bias low variance.T

The high bias low variance suggests the model is inaccurate but produce consistence result. The model is unable to learn the relationship between variables. In this case, adding more features will help the model to improve. This is due to increasing feature means the model has more data to learn from. Examples of features that can be added into this data set are demographic data such as age, gender, occupations, and clinical data such as The Beck Depression Inventory (BDI).

The next suggestion is to add polynomial features. Polynomial features are features which are created by raising existing features to an exponent. It can help the model fits data better. For example, a new column can be created by increasing the values in column “Emotion_Joy” into exponential of power 2. The exponential offers greater flexibility to model, lowering its bias.