Assignment 1

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

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

It is a balanced dataset. Figure 1 shows 136 individuals (46.7%) who have depression and 155 individuals (53.3%) who do not have depression. Due to only 19 individuals (6.6%) difference, it is still considered a balanced dataset.

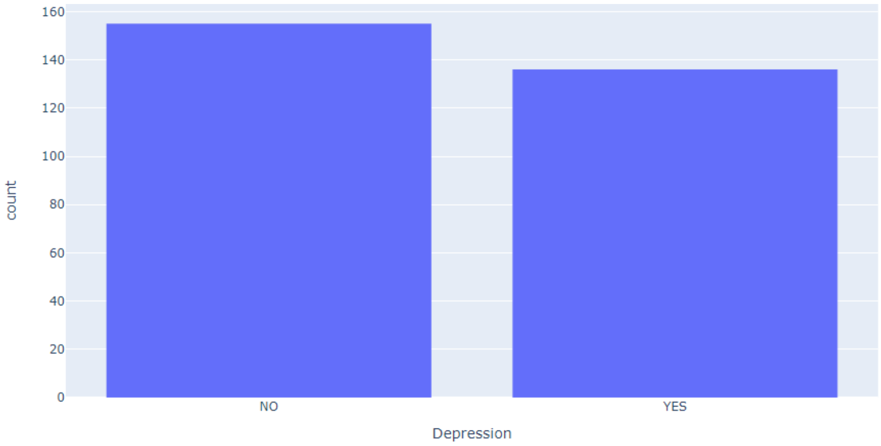

Figure 1: Class Distribution

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

Weighted recall is used as the evaluation matrix. The goal is to identify the depressed person from a group. So, it is acceptable if a non-depressed person is tagged as depressed, but a depressed person should not be labeled non-depress. Recall is used when output-sensitive predictions are needed. Due to the distribution dataset not exactly equal, weighted is used.

c) Specify the validation option you are selecting for your machine learning experiments. Briefly explain the reason for using the validation option.

5-fold cross-validation technique is used.

The data used is small and the small data size will directly affect the reliability of the result. Hence, K-fold validation is used to avoid the overfitting and bias that may cause due to small data set.

In theory, a higher k value leads to less bias and less overfitting as the models have more available data. However, the runtime and the computational cost will also increase when the value of k increases. 5 is chosen as the K value to balance the computation cost and the overfitting issue.

2) Decide at least 3 classification algorithms you are going to run on the Freq-PHOBinary and Norm-PHO-Binary respectively. Report the machine learning experiments you ran.

K-Nearest Neighbors, Decision Tree, and Support Vector Machine are the classification algorithms chosen to explore. A total of 682 hyperparameter sets run on the frequency [341 sets] and normalized [341 sets] representation. Table 1 shows the details of the hyperparameter that are tested in each algorithm.

	Dummy Classifier	K-Nearest Neighbors
Hyperparameter	strategy: [uniform]	n_neighbors: [1, 2, 3, 4, 5, 6, 7, 8, 9, 10] weights: [uniform, distance] algorithm: [auto, ball_tree, kd_tree, brute] p: [1-3]
	Decision Tree	Support Vector Machine

Hyperparameter	criterion: [gini, entropy]
	splitter: [best, random]
	max_depth: [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]
		C = [0.1, 1, 10, 100]
		kernel = [linear, poly, rbf, sigmoid]
		degree = [1, 2, 3, 4, 5]
		gamma = [auto, scale]
		
		*degree only used in poly
		*gamma used in all type of kernel except linear

Table 1: Details of hyperparameter that are tested in each algorithm

Identify the best performing model using Freq-PHO-Binary and also the best performing model using Norm-PHO-Binary (both the best performing models should yield higher performance than your random baseline)

Table 2 shows the best performing model in the frequency and normalized representation. Both the best performing models have a higher performance than the random baseline.

Feature Represent	Machine Learning Algorithm	Parameters	Validation Recall (Weighted Avg)	Random Baseline
Freq	SVM	kernel=poly; C=0.1; degree=1; gamma=auto	0.636	0.488
Norm	DecisionTree	criterion=gini; splitter=best; max_depth=7	0.615	0.502

Table 2: Result

Which feature representation produces a better model?

Between the two best performing model in the frequency and normalized representation, the performance of SVM in frequency representation performs slightly better. Due to there is no significant difference between the model performance in frequency representation and the normalized representation, a threshold of 60% of validation recall is set as the baseline. There have 13 hyperparameter sets in frequency representation able to hit this target whereas only 7 hyperparameter sets in normalized representation can this target. Overall, the frequency presentation produces a better model.

Explain how you determine the best performing model based on the performance metric you have selected.

The model which has the highest recall is chosen as the best performing model. Highest recall indicates the model is able to identify most of the true positives compared to other models. No more than one model shares the highest recall. Hence, no further justification is needed.

Can you explain why one feature representation is better than the other?

From the result, frequency representation performs better than normalization representation.

Normalization should not be used in this dataset because the features are all comparable. The difference between the features matter. Reducing the difference between them bring a negative effect.

Normalization is needed when the features have a different range. However, in this emotional dataset, no feature is in a very different range. Hence, normalization is not necessary.

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 overall best performing model not able to achieve a very promising result.

- Data Collection.
  The current dataset only collects the students' emotions for two weeks and students can record their emotions as many times as they like per day. This data collection system will have two problems:
  - The dataset from one student to another will be imbalanced.
  - Noises are included if record all emotions changes.

  Having more data help the model learns the pattern well and predict the target value correctly. However, if the student records every emotional change and some are just tiny little things that the student will forget very soon, those data will be considered as noises. Hence, it is better if every student only records 1 to 2 emotions per day. If emotion changes frequently that day, the student only needs to choose the emotion that impacts him/her the most.

- Feature Selection.
  There are eight types of emotion recorded in this dataset. Not all these emotions are relevant and some may bring a negative impact on the model. Domain knowledge is needed to filter those emotions that are irrelevant or did not contribute more to the target value. This process not only increases the accuracy of the model and also reduces the training time.