ASSIGNMENT REPORT

1. a)

Depression	Frequency
NO	155
YES	136

The ratio of NO to YES is 1:0.88, which is not exactly 1:1. Therefore, the class distribution is considered **slightly imbalanced** with more instances of NO (155) than YES (136).

b) **Recall is selected as the metric** to evaluate the performance of the classification models because **reducing false negative is more important** when it comes to determining whether an individual is showing signs of depression. If an individual actually shows signs of depression but is classified as not showing signs of depression, the consequence can be very bad for the individual. We **do not want to miss out** any individual showing signs of depression, for they might need early intervention to prevent the mental health problem from getting worse. Not to mention, the recall metric can **account for the slightly imbalanced class distribution.**

c) The dataset contains 291 instances, which is considered quite small. We **do not have sufficient data for percentage split, and it is even harder to define what would be sufficient**. Splitting such a small dataset into training and validation sets might end up having insufficient number of instances in each set for us to get reliable results. Therefore, **k-fold cross validation is a more suitable validation option** in this assignment in the machine learning experiments because instead of splitting the dataset into two sets, we can use all the data provided for training and validation.

2. Recall, or weighted recall to be more specific, will be the primary performance metric to judge which model is better. Each model is trained and validated using 10-fold cross-validation so that the models are comparable. Recall scores of each model are compared, and the model with the highest validation weighted recall is selected as the best performing model. In this assignment, different classification algorithms, such as KNN Classifier, Decision Tree Classifier, Naïve Bayes Classifier, SVM, and Logistic Regression are performed, and the results are recorded in A1_Experiment_Sheet.xlsx. GridSearch using 10-fold cross-validation is used in tuning the parameters, and the results using different parameters are also recorded in the sheet. Finally, the recall scores are compared to identify the best performing model.

Below are just some of the experiments. More are recorded in the sheet, containing 148 rows of classifiers with different parameters.

Using Freq-PHO-Binary

a. KNeighborsClassifier with n_neighbors = 31 gives 0.6049 recall.
b. DecisionTreeClassifier with criterion=gini, splitter=best, max_depth=None gives 0.5602 recall.
c. SVC with c=0.1, kernel=linear gives 0.6186 recall.

Using Norm-PHO-Binary

a. KNeighborsClassifier with n_neighbors = 31 gives 0.6121 recall.
b. DecisionTreeClassifier with criterion=gini, splitter=best, max_depth=None gives 0.6426 recall.

c. SVC with c=0.1, kernel=linear gives 0.5946 recall.

The best performing model using Freq-PHO-Binary is **Support Vector Machine**(kernel = linear, c=0.1) with recall of 0.6186, while the best performing model using Norm-PHO-Binary is **Decision Tree Classifier**(criterion = gini, splitter = best, and max_depth = None) with recall of 0.6426. Both validation recalls are higher than the same validation recall of 0.5052 produced by the Dummy Classifier using both datasets.

Based on the selected performance metric recall, the feature representation using Norm-PHO-Binary produces a better model in overall. Even though all other performance metrics, such as accuracy, precision, and f1 are also recorded in the excel sheet, following discussion is based on the validation recall score. Generally speaking, it is observed that while applying the same classification algorithm with the same parameters on both datasets, Norm-PHO-Binary gives better performance as compared to Freq-PHO-Binary, especially in KNN and Decision Tree Classifiers. This is because KNN is sensitive to magnitudes, and feature scaling does help to bring every feature to the same weight (so that the data won't have very different values for majority voting to works) to allow for meaningful comparisons. Even though decision trees are not affected by monotonic transformation of the variables, it turns out that models using Norm-PHO-Binary actually gives better performance. In the other algorithms, sometimes Freq-PHO-Binary could give better performance, and it really depends on the parameters used. This actually shed light on that feature scaling won't always guarantee better results, so it's definitely not a one-size-fits-all approach.

One thing to note from this assignment is that every tested model with heavy parameter tuning has not been able to achieve very promising results. Majority of the recalls are between 0.5 and 0.6+. Being unable to achieve very promising results might be a sign of high bias or underfitting, and this is the main issue. Another issue observed is that the recall scores for each experiment varies greatly, and this is a sign of high variance.

The scaled features using Norm-PHO-Binary is better than the other because it allows for faster convergence, all features are weighed in equally (no very different values), and good for ML algorithms that calculate distances between data.

4. The overall best performing model is **Decision Tree Classifier** (criterion = gini, splitter = best, and max_depth = None) using Norm-PHO-Binary with recall of only 0.6426, which is unable to achieve very promising results (at least 0.8 recall score).

Below are some future strategies that can be used to improve the performance of the best performing model.

a. Try getting additional features. Perhaps the current features are not significant for the ML algorithms to learn a good hypothesis. Bringing in additional features which was relevant in predicting whether an individual is showing signs of depression can help to fix the high bias. For example, number of social interactions and intolerance of uncertainty and stress.

b. Try adding polynomial features. Most of the time, real world problems are not simple that it can always be modelled linearly. Adding polynomial features can help fix the high bias by explaining the part which could not be explained by the current data.

c. Perhaps a more intuitive way is to get more labelled training data, but it’s a rather costly measure. When we have more training data, most of the time, we can improve the performance of our model by fixing high variance.

d. Feature selection. A small subset of features containing features significant to target prediction works better than just using all features. During the data exploratory analysis, it is observed that many features contain many outliers. Using a smaller set of features decreases dimensionality, removes insignificant features which were also noisy, and fixes high variance.