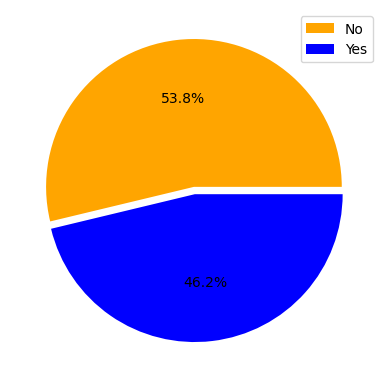
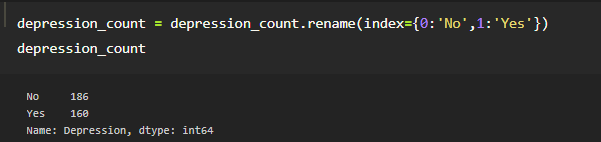
1) Study the data set carefully and answer the questions below:

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



According to the provided pie chart, the dataset appears to be balanced, although there is a slight difference in the occurrence of target classes. Specifically, there are a total of 186 instances classified as 'Yes' and 160 instances classified as 'No'. This indicates a difference of 26 instances, which corresponds to approximately 7.6% of the total instances and highlights the distinction between the two classes. It's worth noting that the dataset is relatively small, and due to its size, each entry can significantly impact the balance and overall performance of the learning model.

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

Accuracy scoring metric is chosen to evaluate the model’s performance in this experiment, due to its ability and an intuitive metric that works well with a balanced dataset, which presents in this case.

As for f-1 metric, it’s not suitable in this dataset as it’s a better suit for unbalanced dataset, where in this experiment may cause consistence of low accuracy during the performance evaluation.

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.

K-fold cross validation is used in this experiment due to the limited instances available in the datasets by maximizing the use of available data for both training and evaluation. Its drawback lies in the computation inefficiency due to training and evaluating the model based on different folds of data multiple times. Though, it can provide a more robust estimate of model performance as it utilizes multiple train-test splits and reduces the impact of data variability in this case.

2.From the records in A1\_Assignment\_Sheet, it clearly suggests that Norm-PHO-Binary produces a better model that has an average accuracy score of 0.65 using the 4 Classifier methods. Meanwhile, Norm-PHO-Binary dataset is able to produce model performing at an average accuracy score of 0.59. Notably while using Norm-PHO-Binary, SVM (Support Vector Machine) model is able to produce an accuracy score of 0.69, while Freq-PHO-Binary produces 2 best results with SVM and KNN (K-Nearest Neighbors) at an average score of 0.64.

To select the best model for each Feature Representation, hyperparameters tuning for each model is one of the crucial steps to increase the model’s baseline performance and effectiveness on the respective datasets. If the hyperparameters are poorly chosen, the model may overfit the training data, meaning it becomes too specialized and fails to perform well on new, unseen examples. There are instances where Validation and Testing scores are far apart, therefore, a lower validation score is chosen to avoid underfitting problem.

Reason why Norm-PHO-Binary dataset is able to produce a better performing model is due to its well-refined data in the features, in this case normalizing the emotions of each day. When a dataset is normalized, it overall improves the interpretability and comparability of its data across different models. Besides, Norm-PHO-Binary is well-refined dataset may help in capturing the underlying patterns and concepts and further reduces the likelihood of overfitting to outliers in the training data. When working with frequencies, the actual values are often specific to the context, which may limit the model generalization, making it harder to compares. Normalized dataset, on the other hand, provides a standardized representation that can be easily understood, compared, and be used in another model.

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.

The best performing SVM model is only able to achieve accuracy of 0.69, despite the dataset is balanced. The sample size in a dataset plays a significant role in the performance of a classifier model.

In this experiment, one of the ways we can further improve the learning model’s performance is by increasing the diversity and the training data. The model should be trained on a more diverse and extensive data from a publicly available dataset that is similar or processed similarly to this. That way, the classifier model will have a broader understanding of the data beforehand.

Another way is to further fine-tune the classifier model by training it on any smaller dataset of a similarly related task. This allows the learning model to adapt properly to the related task and able to perform better when given different dataset of similar task in future.