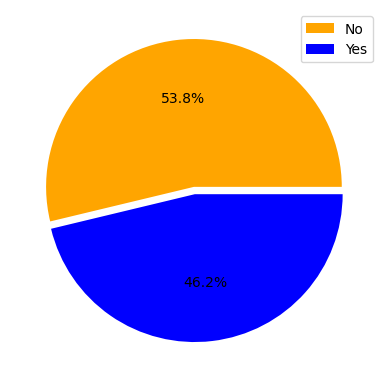
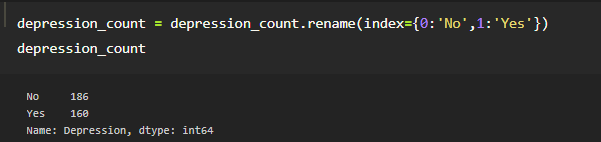
1) Study the data set carefully and answer the questions below:

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 adopted to evaluate the 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.