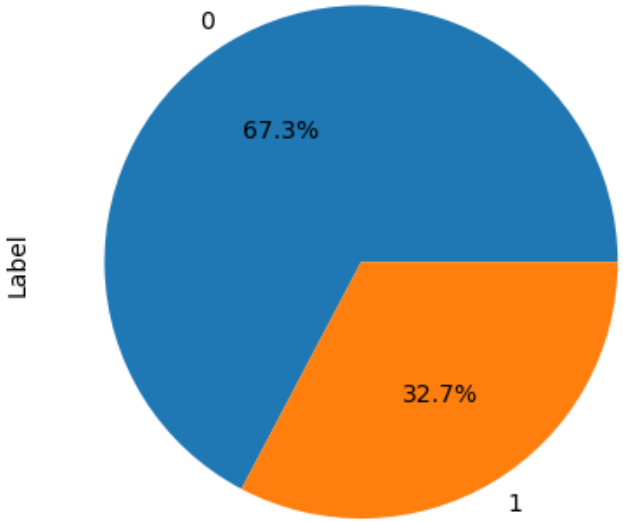
Name: Shayna Nicholas Tuscano

Understanding the Dataset :

The dataset has 5 columns namely Participants, HR, respr, Time(sec) and label corresponding to participant number, heart rate, respiratory rate, timestamps and labels respectively and has 112516 rows.
In this dataset we have Independent variable like Participants, HR, respr, Time(sec) and Label is Dependent variable. The label column has only two distinct values/categories ie. 0 or 1 which can classify if the patient is stressed or is not stressed. It is a binary classification problem. 0=not stressed &1=stressed.
Hence we use classification type of supervised learning to solve this problem.

Data Exploration :

After visually inspecting the label column to check if the dataset is imbalanced or not it was observed that, the occurrence of 0 is 67.185486% and 1 is 32.814514% . The dataset is slightly imbalanced. It won't affect the performance of the model to great extent . Hence no imbalance correction technique needs to be applied on the dataset.



Choosing an ML Package:

For this assignment I will be using Sklearn which is an open source library. This library can handle classification, regression, clustering, dimensionality reduction, model selection, and preprocessing. It is well documented hence easy to use and can efficiently solve machine learning problems. It is build on top of NumPy, SciPy and Matplotlib. It can also be used to check the performance metrics of the ML algorithm**.**

Data Pre-processing :

Following data pre-processing steps are taken:

1. Identify and remove duplicate data - No duplicate data was present .
2. Handling missing values - 44 missing values were found in HR column and they are replaced with mean value.
3. Feature selection - Finding the correlation of features with the label and removing columns like “Participant" and “Time(sec)" as the correlation between label and these column is very less.HR is highly correlated and respr is negatively correlated with label.

Algorithm Selection and Application :

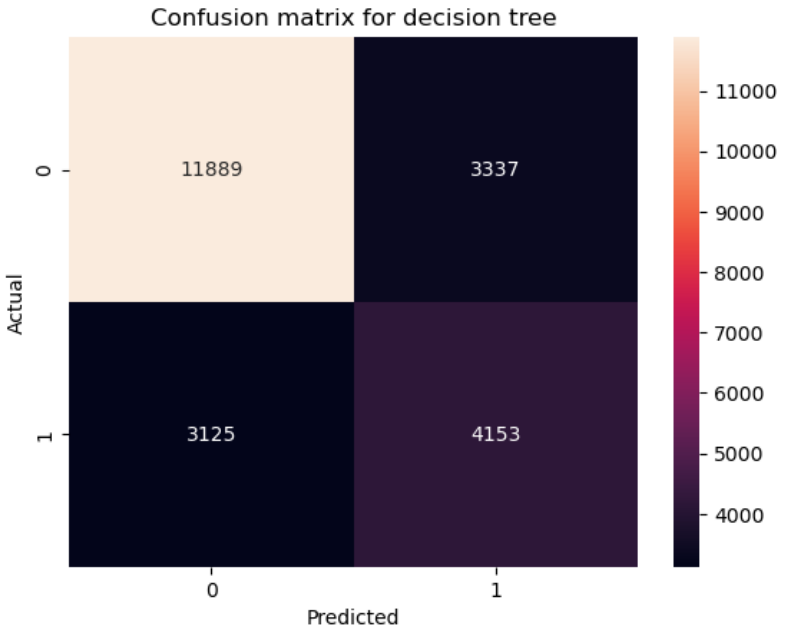
First the dataset is split in Training data (80%) and Test data (20%).For this assignment we will be using Decision Tree & Random Forest Algorithm.

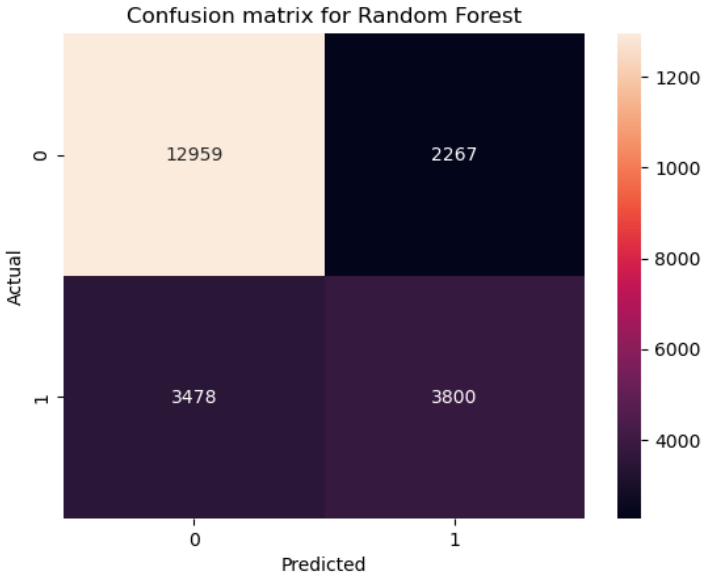
1. **Decision Tree :**Decision Tree is a Supervised learning technique. It can be used for classification and Regression problems.Decision tree consist of nodes (root node,Branch Node and Leaf Node) and branches. Branch nodes are used to make any decision and have multiple branches, whereas Leaf nodes do not contain any further branches. The root node is the initial node where the entire dataset is split based on different conditions. The root node is selected using ASM technique(to find the best feature in a dataset). ASM techniques are Information Gain & Gini Index. If we use information gain then the node which has the max information gain will be considered as root node which is further divided into decision nodes. The ASM techniques aims in creating most homogeneous subset of the data after splitting the data to maximise the information gain.This process is repeated until a stage is reached where you cannot further classify the nodes and nodes are called the as a leaf node. In short the decision tree splits the data recursively using the decision nodes unless its left with pure leaf node.

2. **Random Forest Algorithm:** Random forest can be used for both Classification and Regression problemsIt is based on the concept of ensemble learning in which multiple classifier are combined to improve the performance of the model The random forest algorithm is made up of a multiple decision trees. In random forest the decision forest are build using bagging technique.Bagging (Bootstrap + aggregating).Bootstrap technique involves randomly forming some samples of data from the original dataset with replacement.Random forest uses random and smaller subset of features while training each tree .The final predicted output is based on majority voting instead of relying on one decision tree, the random forest takes the prediction from each decision tree and based on the majority votes of predictions, and it predicts the final output.

Model Evaluation:

After evaluating both the models the accuracy for Random forest was slightly high compare to Decision. Accuracy of Decision Tree was 71.2851 % and Accuracy of Random Forest was 74.4712%.For this assignment our aim is select the model which gives minimum false negative ie. the model should avoid predicting 0 if the actual value is 1, because 0 = Not stressed and if the person is stressed and the model predicts stressed the person might not seek medical help on time and which in future may lead to serious health issues like heart attack ,depression etc.Hence we will use Recall to measure the performance of both the models.Recall takes false negative into account.





The recall for Decision tree is 57.0623% which is better than that of Random Forest which has recall of 52.2121.

Hence for this assignment Decision Tree would be a better choice compare to Random Forest Algorithm.

Comparative Analysis:

The results provided by both the algorithm were not slightly different.

Both the models have strengths and weaknesses.

1. Decision tree can perform classification with less computation where as Random forest may require more computation due to presence of a large number of trees.

2. Decision tree is more prone to overfitting on the other hand random forest prevents overfitting.

Below is a comparative analysis of both algorithms performance on the given dataset.

Parameters	Decision tree	Random Forest
Accuracy	Less than that of Random Forest	More than that of Decision Tree
Training Time	Less than that of Random Forest `0.3211 sec`	More than that of Decision Tree `11.7113 sec`
Prediction Time	Less than that of Random Forest 0.0061sec	More than that of Decision Tree `0.5881 sec`