**Mental State Classification using EEG-based Brain**

horizontal line

# 

# 

# Introduction

The purpose of this project is to provide an efficient, parametric, general, and completely automatic real time classification method of electroencephalography (EEG) signals obtained from emotions. The characteristics of the considered high-frequency signals (theta, alpha, beta, gamma) and adapting strategies like the Fourier Transform, Features Extraction (mean, standard deviation, power).

## Problem

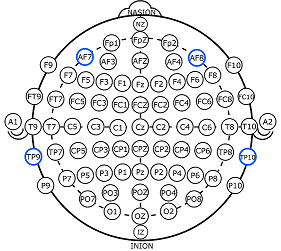
It is difficult to look at the EEG signal and identify the state of Human mind. In this project, EEG Brainwave Dataset is trained to predict the state of mind. the state of mind is predicted in terms of positive, negative and neutral expression.

## The formation of the EEG dataset

An electroencephalogram (EEG) is a test that detects electrical activity in your brain using small metal discs (electrodes) attached to your scalp. Your brain cells communicate via electrical impulses and are active all the time, even when you're asleep.

### Ground Truth

A Muse EEG headband is used which recorded the TP9, AF7, AF8 and TP10 EEG placements via dry electrodes. Non verbal task we implemented to prevent interference of electromyographic signals. It has taken a small window technique as the signals are stationary for a small interval of time.



Our data was collected from two people (1 male, 1 female) for 3 minutes per state - positive, neutral, negative. Six minutes of resting neutral data is also recorded, the stimuli used to evoke the emotions are below

1. Marley and Me - Negative (Twentieth Century Fox) *Death Scene*
2. Up - Negative (Walt Disney Pictures) *Opening Death Scene*
3. My Girl - Negative (Imagine Entertainment) *Funeral Scene*
4. La La Land - Positive (Summit Entertainment) *Opening musical number*
5. Slow Life - Positive (BioQuest Studios) *Nature timelapse*
6. Funny Dogs - Positive (MashupZone) *Funny dog clips*

## Environment

*The code is written in python language using windows 10. Multiple libraries used like pandas, numpy and matlab. StandardScaler is used for data preprocessing. For the classifications there are libraries like pipeline, RandomForestClassifier, KNeighborsClassifier and SVC(support Vector Classifier). PCA is used for dimension reduction*

## Process

*The dataset is first uploaded and then the class distribution. The classes and attributes are split by transferring the labels to label variable and dropped from the brainwave dataset. Correlation between the features are found out using corr method which is later depicted as a heat map.* A [heat map](https://www.data-to-viz.com/graph/heatmap.html) (or heatmap) is a graphical representation of data where the individual values contained in a matrix are represented as colors. For our dataset the +1 correlation is shown in white, -1 is shown in black and rest in shades of yellow and red. Classification takes place using Nearest Neighbour, Random Forest, Logistic Regression and Support Vector Machine. *Principal Component Analysis is used to reduce the dimension of the data for making components out of the data using correlation which is more effective to find the important features and accuracy. Finally a comparison of the models is given.*

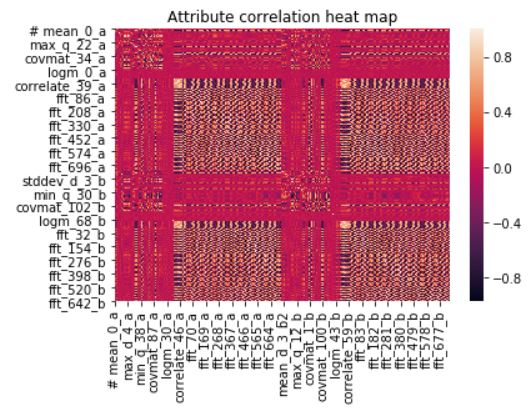
## 

## 

## Method

### Correlation

Correlation refers to the relationship between two variables and how they may or may not change together.The most common method for calculating correlation is [Pearson’s Correlation Coefficient, that assumes a normal distribution of the attributes involved. A correlation of -1 or 1 shows a full negative or positive correlation respectively. Whereas a value of 0 shows no correlation at all. We tried correlation with the brain dataset and got the result which showed in the heatmap.



### Nearest Neighbor Classifier

K-NN or K-Nearest Neighbors is one of the most famous classification algorithms as of now in the industry simply because of its simplicity and accuracy. K-NN is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions). In K-NN, K is the number of nearest neighbors. The number of neighbors is the core deciding factor. K is generally an odd number if the number of classes is 2. When K=1, then the algorithm is known as the nearest neighbor algorithm.

### Random Forest Classifier

RandomForest is a tree & bagging approach-based ensemble classifier. It will automatically reduce the number of features by its probabilistic entropy calculation approach. Random forest (ensemble method) builds multiple decision trees and merges them together to get a more accurate and stable prediction. Random Forest is a flexible, easy to use machine learning algorithm that produces, even without hyper-parameter tuning, a great result most of the time. It is also one of the most used algorithms, because it’s simplicity and the fact that it can be used for both classification and regression tasks.

### Logistic Regression Classifier

Logistic Regression is another technique borrowed by machine learning from the field of statistics. It is the go-to method for binary classification problems (problems with two class values). Logistic regression is like linear regression in that the goal is to find the values for the coefficients that weight each input variable. Unlike linear regression, the prediction for the output is transformed using a nonlinear function called the logistic function. The logistic function looks like a big S and will transform any value into the range 0 to 1.

### Support Vector Machine (SVM)

A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples.

### Principal Component Analysis (PCA)

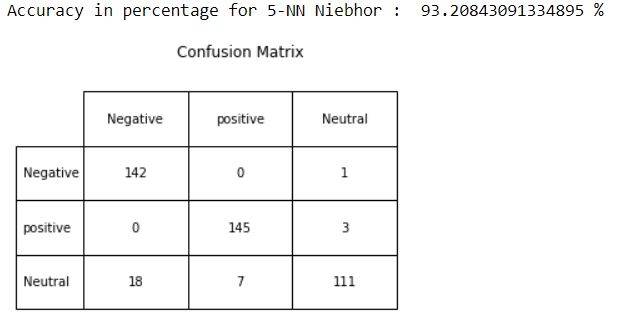
PCA can transform original low level variables to a higher dimensional space and thus reduce the number of required variables. All collinear variables get clubbed together. Here we used PCA to reduce the dimension. Firstly we normalize the data. Then find the components with high variance and use those features and leave the rest. We got 20 PCs for our data. In PCA we can lose information if the left out features have more variance but if it has really less value it can not lose information.

### Cross Validation

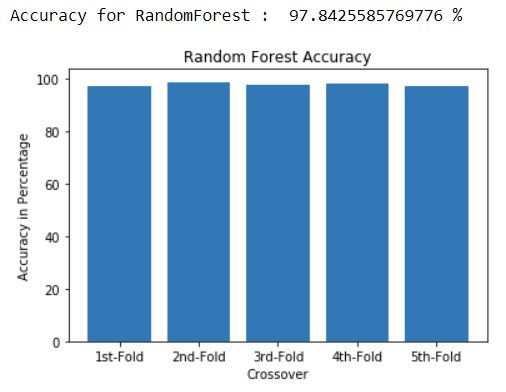
Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample.The procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into. As such, the procedure is often called k-fold cross-validation.

## Results:

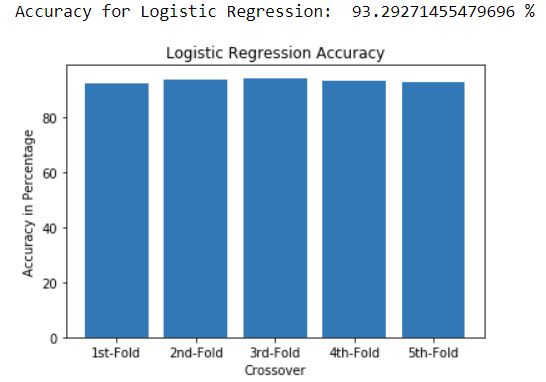
### Nearest Neighbor Classifier



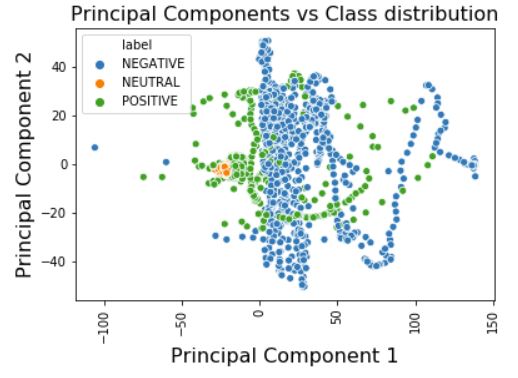
### Random Forest Classifier



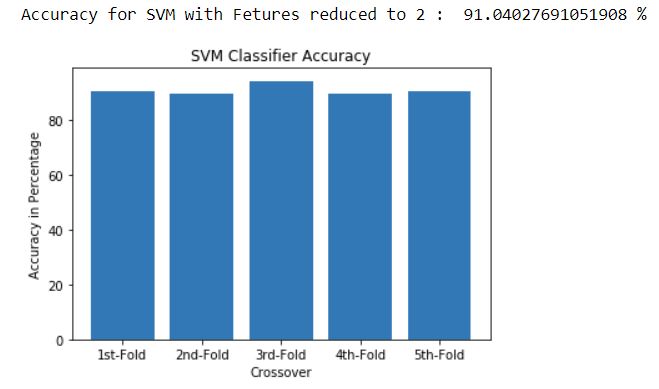
### Logistic Regression Classifier

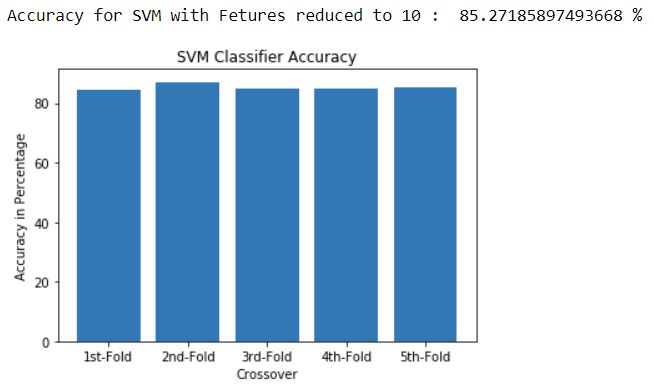


### Principal Component Analysis (PCA)



### Support Vector Machine (SVM)





## Comparison of all the models

