Data mining & Analytics

Project report

Tharun reddy Banda

University Id: 519677

Project Report: Human Emotion Prediction

Introduction:

The purpose of this project is to predict the emotion of a person, whether he/she is in a neutral state, happy, or unhappy, based on the features provided in the dataset. The aim is to use various techniques to determine the most accurate model for the given dataset.

Data:

The data is collected from ECG signals contains 31 features, all independent features are numerical, and a target label which is also encoded numerically.

Step 1:

Import Libraries

- Imported all necessary libraries such as NumPy, Pandas, Matplotlib, Seaborn and Scikit-learn for data analysis and modelling.

Step 2:

Load Data

- Loaded the data into a Pandas DataFrame.

Step 3:

Data Exploration

- visualized the data to gain insights and understanding.
- Noticed that the target variable contains multiple classes, so extracted only the required class data from the total dataset to predict human emotion.

Step 4:

Pre-processing

- Extracted the data into new Dataframe and checked for its properties

- checked the data type of all the features and converted the data type of certain columns from integer to float
- Checked for duplicates in the data. To detect outliers, used the interquartile range. The data points that fall out of the interquartile range (0.25 to 0.75) are considered as outliers.
- Performed the outlier's detection for each class data and replaced the missing value with the median .

Step 5:

Data Visualization

- Visualized the data to gain insights into the features and target label

Step 6:

Co-related Features

- Used a heatmap to check the correlation between features.
- Noticed that the data has highly negatively correlated features with the target, so extracted only the features that are positively linearly correlated with the target variable.

Step 7:

Feature Selection

• For Feature Selection, multiple feature selection techniques were used such as SelectKBest, Select from Model, Recursive Feature Elimination, Sequential Feature Selector, and the embedded method (SelectFromModel) to find the most important features that contribute more weight to the target variable.

• The features that were most commonly predicted by the four feature selection models were selected.

Step 8:

Feature Scaling

• Standard Scaler was used to scale down the features obtained in feature selection.

Step 9:

Feature extraction

• In cases where two or more independent features have the same correlation with the target variable and have equal correlation with each other, the feature selection model fails to select the best feature. In such cases, feature extraction was used to extract the best feature from them.

• PCA was used to reduce the dimensionality (number of features) within the dataset while still retaining as much information as possible. To determine the number of dimensions the PCA (figure 4) should reduce the data, an iteration method was used, which starts by considering one component and increases its components after every iteration. A k-Neighbors Classifier model was also used to find the number of components at which the model accuracy was high. The component at which the model accuracy was maximum was selected.

Step 9:

Label Encoding

• The label data was converted into positive values using Label Encoder from scikit-learn. The label data was originally encoded with negative values.

Step 10:

Data Split

- The data was split into 70% train and 30% test sets for training and testing the model using Train_test_split from scikit-learn

Step 11:

Model Building and Training

- Used varies Machine Learning models to check which model is giving high performance.
- Used classification models to predict the emotion

The following classification models were used:

1. ExtraTreesClassifier
2. Logistic Regression
3. DecisionTreeClassifier
4. SVM Classifier
5. HistGradientBoostingClassifier
6. Gradient Boosting Classifier
7. Stacking classifier

Step 11:

Hyperparameter tunning

Used GridSearchCV for Hyperparameter tunning every model to get the best Hyperparameters which help the model to improve the accuracy score.

Step 12: Model Evaluation and Comparison

Model Evaluation:

• The accuracy scores of all the models were calculated for both the training and testing data

• The model which was found to have the highest accuracy performance in both training and testing data was selected for further evaluation

MEACHINE LEARNING MODELS USED	ACCURACY
Logistic Regression	62%
ExtraTreesClassifier	100%
DecisionTreeClassifier	96%
SVM Classifier	65%
HistGradientBoostingClassifier	100 %
Gradient Boosting Classifier	99.7%
Stacking classifier	56%

Step 13: Confusion Matrix

• The selected model was evaluated using the confusion matrix to check whether precision and recall are balanced

• The f1Score was used to evaluate the precision and recall balance

Step 14: Unsupervised Learning

• Unsupervised learning was used to evaluate the model performance

• KMeans clustering and Hierarchical Clustering were used as unsupervised learning techniques (figure 3) and selected object which is well matched to its own cluster and poorly matched to

neighbouring clusters by using range value 0-1 considered the value which is near to 1.

• Evaluated the accuracy of k-means clustering on the original label data of emotions and the predicted classes

• The accuracy obtained was 62% with k-means and 67% with the Hierarchical Clustering

Step 15: Determining the Optimal Number of Clusters in KMeans

• The elbow method was used to determine the optimal number of clusters in KMeans

• The number of clusters was selected at the point at which the elbow shape was created

Step 16: Silhouette Score

• The Silhouette score (figure 3) was used to evaluate how well the data points were clustered inside the clusters

• The score was calculated as a value between 0 and 1, with 1 being the best score

• The score was used to select the object which was well-matched to its own cluster and poorly matched to neighboring clusters

Step 17: Evaluating the Accuracy of KMeans Clustering

• The accuracy of KMeans clustering was evaluated on the original label data of emotions and the predicted classes

• The accuracy obtained was 62% with KMeans

Step 18: Evaluating the Accuracy of Hierarchical Clustering

• The accuracy of Hierarchical Clustering was also evaluated on the original label data of emotions and the predicted classes

• The accuracy obtained was 67% with Hierarchical Clustering

Results:

The best accuracy among all the models was achieved with the HistGradientBoostingClassifier (100% with training data and 0.5538461538461539 with testing data)

• Unsupervised learning with KMeans clustering and Hierarchical Clustering was also performed, and the accuracy score between the original label data of emotion and predicted classes of KMeans was 62% and Hierarchical Clustering was 67%

Conclusion:

In conclusion, the results of this experiment provide valuable insight into the performance of different machine learning models for binary classification problems. Model HistGradientBoosting Classifier was found to be the most accurate and can be considered for further analysis and implementation in the future. The steps involved in the project, including data pre-processing , feature selection, data split, model building and training, model evaluation and comparison, and unsupervised learning were described in detail.

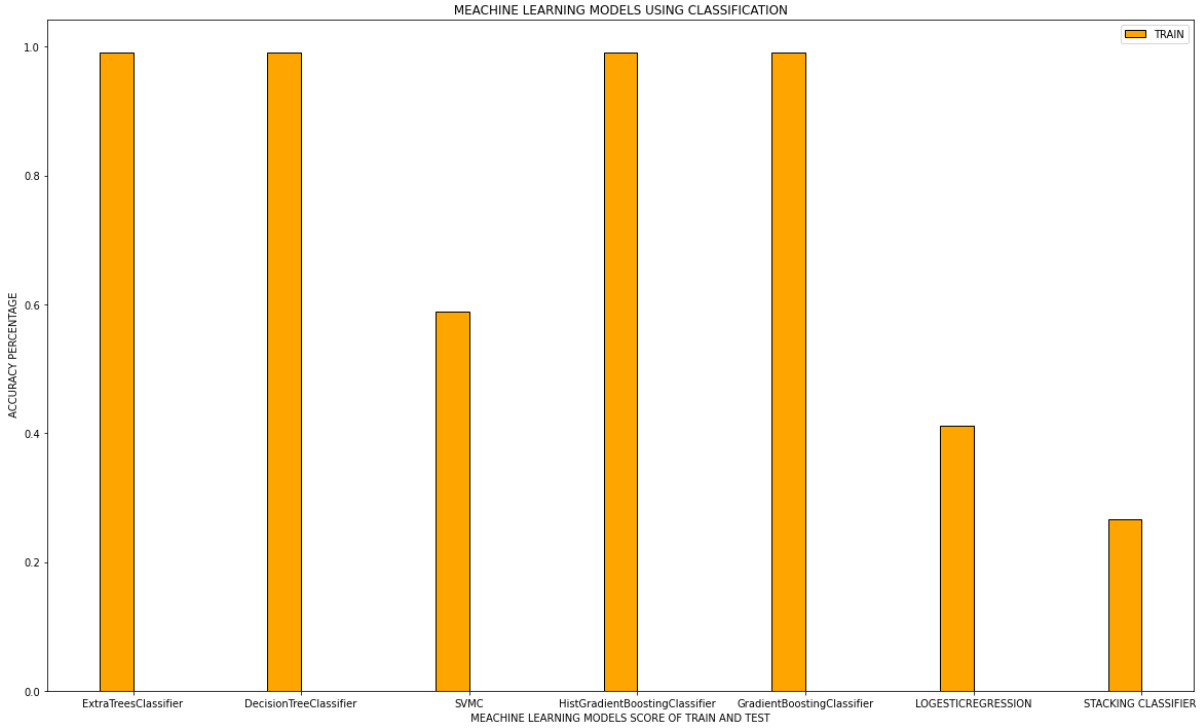
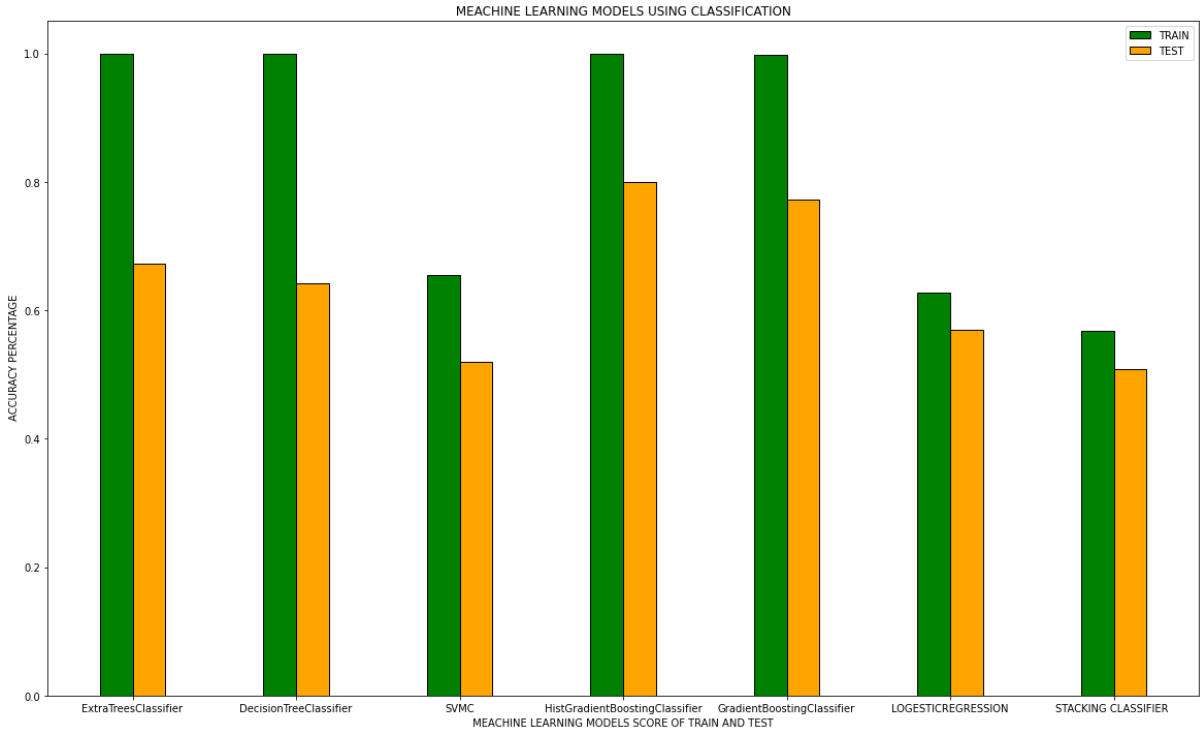

MEACHINE LEARNING MODELS USING CLASSIFICATION
TRAIN
ACCURACY PERCENTAGE
0.0
0.2
0.4
0.6
0.8
1.0
ExtraTreesClassifier
DecisionTreeClassifier
SVMC
HistGradientBoostingClassifier
GradientBoostingClassifier
LOGESTICREGRESSION
STACKING CLASSIFIER
MEACHINE LEARNING MODELS SCORE OF TRAIN AND TEST

Figure-1


MEACHINE LEARNING MODELS USING CLASSIFICATION
TRAIN
TEST
ACCURACY PERCENTAGE
0.0
0.2
0.4
0.6
0.8
1.0
ExtraTreesClassifier
DecisionTreeClassifier
SVMC
HistGradientBoostingClassifier
GradientBoostingClassifier
LOGESTICREGRESSION
STACKING CLASSIFIER
MEACHINE LEARNING MODELS SCORE OF TRAIN AND TEST

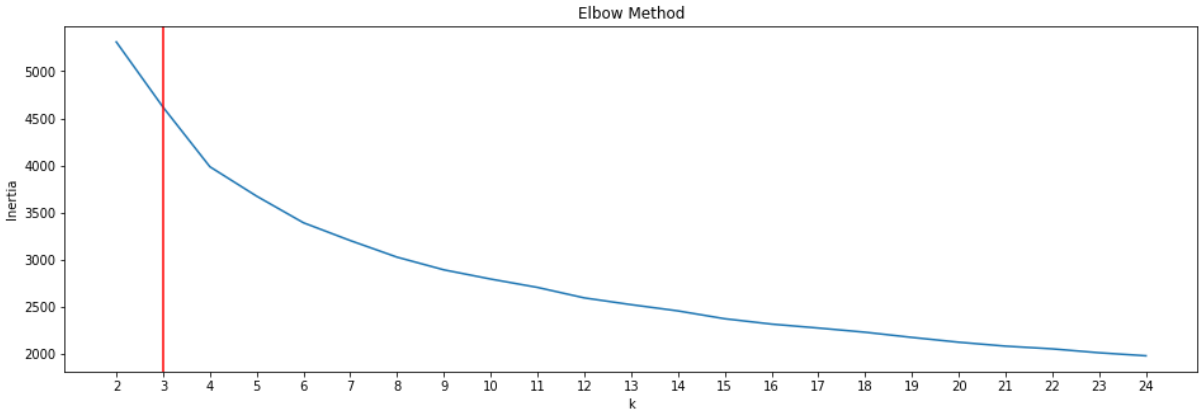


Figure:2

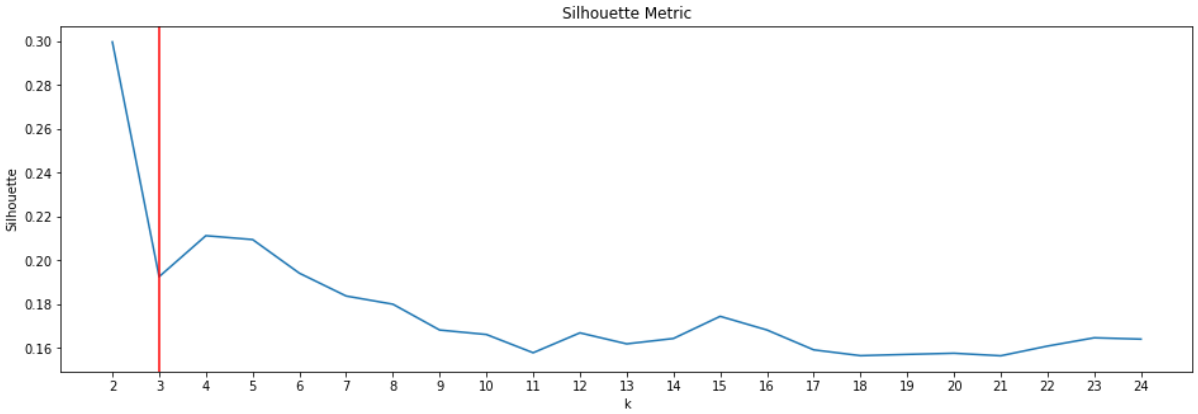


Figure:3

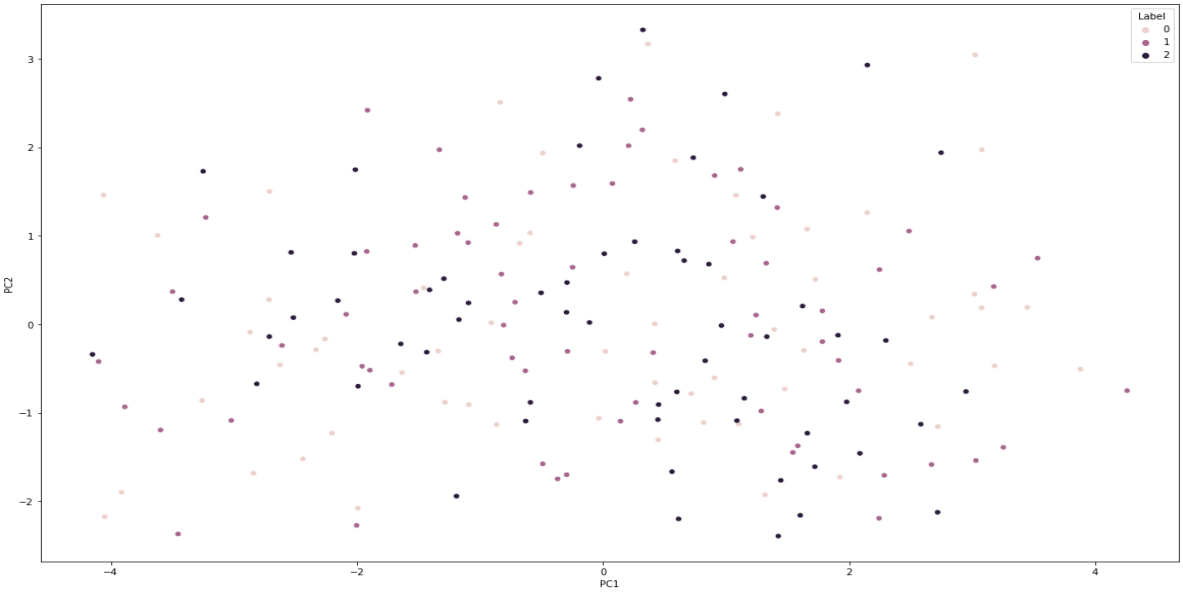


Figure:4