 **Authored By: Muzammil Ahmed Shaik**



**Machine Learning Models**

Ranking of models based on accuracy:

1) Bagging Decision Tree: Accuracy = 92%

score on test: 0.92

score on the train: 0.9466666666666667

Bagging Decision Tree is an optimal model.

2.a) K-Nearest-Neighbors: Accuracy = 84%

score on test: 0.84

score on train: 0.8266666666666667

KNN is an optimal model.

2.b) Random Forest: Accuracy = 84%

score on test: 0.84

score on train: 1.0

Random Forest is an overfitting model.

2.c) Voting Classifier: Accuracy = 84%

score on test: 0.84

score on train: 0.9333333333333333

Voting Classifier is an overfitting model.

3.a) Boosting Decision Tree: Accuracy = 76%

score on test: 0.76

score on train: 1.0

Boosting Decision Tree is an overfitting model.

3.b) Neural Network (Tuning): Accuracy = 76%

Testing accuracy: 0.76

Training accuracy: 0.8133333333333334

The neural network is an optimal model

3.c) XG Boost Tree: Accuracy = 76%

score on test: 0.76

score on train: 1.0

XG boost Tree is an overfitting model

4.a) Logistic Regression: Accuracy = 72%

score on test: 0.72

score on train: 0.8666666666666667

Logistic Regression is an overfitting model.

4.b) Support Vector Machine: Accuracy = 72%

score on test: 0.72

score on train: 0.8

Support Vector Machine model is an optimal model

4.c) Decision Tree: Accuracy = 72%

score on test: 0.72

score on train: 1.0

Decision Tree is an overfitting model

5.a) Multinomial Naive Bayes algorithm: Accuracy = 44%

score on test: 0.44

score on train: 0.68

Multinomial Naive Bayes model is an underfitting model

Out of these models, the best, optimal, and most accurate model is the bagging decision tree which gave 92% accuracy to predict symptoms of anxiety of the patient.

**Confusion Matrix**

a confusion matrix, also known as an error matrix, is a specific table layout that allows visualization of the performance of an algorithm, typically a supervised learning one.

**Chart

Description automatically generated**

Fig 2a. Confusion matrix with necessary data

This confusion matrix is plotted for the Bagging decision tree model which has achieved 92% accuracy in predicting anxiety symptoms when facial and vocal data is given.

The features that have been selected in the model are related to vocal data converted to digits, voice data converted to digits, and facial expression embedded to numbers.

Given these features, using the bagging decision tree as a model we can predict whether a patient suffers from an anxiety disorder. In the real world, a doctor or more precisely a psychiatrist will be able to diagnose anxiety disorder using these 10 features that we have selected for the model. The other features which have not been included in this model and decided to drop them for the prediction were sensible and reasonable.

**Fisher’s Exact test p-value**

Top 10 features selected from ~80k features using select-k-best method are:

a) pcm\_fftMag\_spectralRollOff90.0\_sma\_linregc2\_q1

b) pcm\_fftMag\_spectralRollOff50.0\_sma\_de\_de\_iqr1-3\_q1

c) pcm\_fftMag\_spectralMinPos\_sma\_de\_de\_quartile2\_q1

d) pcm\_fftMag\_spectralCentroid\_sma\_de\_iqr2-3\_q2

e) mfcc\_sma\_de[6]\_percentile95.0\_q6

f) pcm\_fftMag\_spectralRollOff50.0\_sma\_de\_quartile3\_q9

g) pcm\_fftMag\_spectralCentroid\_sma\_de\_iqr2-3\_q9

h) pcm\_fftMag\_spectralCentroid\_sma\_de\_iqr1-3\_q10

i) FaceEmbeddingDim66.min.q1

k) FaceEmbeddingDim101.kurtosis.q7

After the selection of these features, we might think of dropping more features, or are these features significant for the model? These all questions can be answered using Fisher’s exact test p-value. If the p-value is less than 0.05, then that feature is significant.

**Graphical user interface, text, application

Description automatically generated**

Fig 3a. Chi-square test for top 10 features

Seeing the P-values of the top 10 features, we can say that all the features listed below are significant and none of them can be dropped from the dataset.

Finally, we were able to achieve 10 features from ~80k features which is an excessive triumph for building a model. Ideally, a good model has 6-8 features as it is difficult to monitor features more than 10 but 10 is however adequate.

**Methodology**

Eye-catching material in this project was the number of features which were ~80k, to deal with this problem we have used down casting of datatypes each feature to utilize memory efficiently. There were some missing values in the features which were filled in using median and mode imputation. Exploratory data analysis was done on the dataset, feature engineering was implemented efficiently by handling nan and infinite values instead of dropping them. Feature selection was done using random forest feature importance method to get average score of features to drop them. Other methods like correlation and select k best methods were used to wisely select 10 features from 79404. Data visualization was done on every step to know how data is related to the problem and decisions were made accordingly. Classifications models were built with feature transformationions to get precise score to predict anxiety given voice and facial data. Bagging decision tree stands out to be the best one with 92% accuracy. Hypothesis testing was done on selected features to know whether they are significant or not. Finally, correlation matrix was built with complete metrics for a classification model.