Assignment 3 Report

Data Mining

CSE 572

Spring 2018

Submitted to:

Prof. Ayan Banerjee

Ira A. Fulton School of Engineering

Arizona State University

Submitted by:

Lakshmi Sneha Kandukuri (lkanduku@asu.edu**)**
Manasa Pola (mpola@asu.edu**)**
Himaja Tirumalasetti (htirumal@asu.edu**)**
Hitesh Kumar Sannithi (hsannithi@asu.edu**)**
Kushal Reddy Papakannu (kpapakan@asu.edu**)**

April 12, 2018

Table of Contents

1. Introduction

The assignment is to perform **user dependent analysis** using classification techniques like decision trees, support vector machines and neural networks on gesture data collected using the sensors. We consider our dataset as highly complex and containing imprecise and uncertain data. The data set is preprocessed for creating a new feature matrix by using feature extraction techniques. Then various classification techniques like neural networks, support vector machines and decision trees are applied on the new feature matrix. The results of Accuracy, Precision, Recall and F1 measure from each classification technique are used to determine the best technique suited for the dataset to identify the gestures.

2. Team Members

Following are the group member of this project

Lakshmi Sneha Kandukuri (lkanduku@asu.edu)

Manasa Pola (mpola@asu.edu)

Himaja Tirumalasetti (htirumal@asu.edu)

Hitesh Kumar Sannithi (hsannithi@asu.edu)

Kushal Reddy Papakannu (kpapakan@asu.edu)

3. Data Preprocessing

The assignment 1 is to collect the sensor data. All the team members should go to Impact lab for collecting the data. Any one of the team member should volunteer for gesturing in front of the screen wearing 2 wristbands, one on right and one on left hand. The gestures are identified by the sensors using the hand movement of the person wearing the wristbands. The gestures are captured using 4 sensors Gyroscope, Accelerator, EMG Sensor, Orientation. According to movement of the hand, each sensor captures the data, for example, if the gesture consists of rotations, orientation sensor data is useful to identify the gesture.

In this task we are considering the raw data from phase 1 of 6 groups i.e. group DM29, DM31, DM32, DM33, DM34, DM35. Each group data consists of 200 CSV files where each person performed 10 gestures 20 times. Each Action generates data from 34 sensors across 45 time

series. This data is processed in the following way:

- Numerical data of one action of one gesture is transposed leading to a matrix of size 34X45. All the remaining actions of this gesture are transposed and appended below one another creating a matrix of size 680X45.
- Similarly, the above task is performed for all the gestures and appending all of them one below the another creating an input matrix for one group of dimensions 6800X45 i.e 680*10X45.
- Performing the above process for the remaining 5 groups generates 6 different input files of 6 persons.

Refer DM_Task3.m file for the MATLAB code.

4. Creation of New Feature Matrix

- For each gesture we have identified the dominant features which showed maximum variance. The user specific data for each sensor of all the gestures is extracted, which creates a matrix of dimensions 200x45.
- We have applied FFT for the above matrix, which gives us a matrix of dimensions 200X4. We have transposed the above matrix (i.e. 4X200) and stored in a temporary matrix. Similarly, we have constructed the transposed matrices for all the sensors of one gesture and stacked in temporary matrix creating a matrix dimensions of 20X200. The temporary matrix is finally re-transposed creating a final matrix of dimensions 200X20 let's say it is StackedTransposeMatrix.
- Principle component analysis PCA is applied to StackedTransposeMatrix which returns a coefficient matrix (20x20).
- We have multiplied StackedTransposeMatrix with coefficient matrix of PCA creating a new feature matrix of dimensions 20x20.

5. Classification Techniques

5.1 Terminology

Confusion Matrix:

A clean and unambiguous way to present the prediction results of a classifier is to use a confusion matrix.

True Positives (TP) - These are the correctly predicted positive values which means that the value of actual class is yes and the value of predicted class is also yes.
True Negatives (TN) - These are the correctly predicted negative values which means that the value of actual class is no and value of predicted class is also no.
False Positives (FP) – When actual class is no and predicted class is yes.
False Negatives (FN) – When actual class is yes but predicted class in no.

Using these four parameters we are calculating Accuracy, Precision, Recall and F1 score.

Accuracy:

Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations. One may think that, if we have high accuracy then our model is best. Yes, accuracy is a great measure but only when you have symmetric datasets where values of false positive and false negatives are almost same. Therefore, you can look at other parameters to evaluate the performance of your model.

Precision:

Precision is the number of True Positives divided by the sum of number of True Positives and False Positives. Put another way, it is the number of positive predictions divided by the total number of positive class values predicted.
Precision can be thought of as a measure of a classifiers exactness. A low precision can also indicate large number of False Positives.

Recall:

Recall is the number of True Positives divided by the sum of number of True Positives and False Negatives. Put another way it is the number of positive predictions divided by the number of positive class values in the test data. It is also called Sensitivity or the True Positive Rate.
Recall can be thought of as a measure of a classifiers completeness. A low recall indicates many False Negatives.

F1 Score:

The F1 Score is the 2*((precision*recall)/ (precision + recall)). It is also called the F Score or the F Measure. Put another way, the F1 score conveys the balance between the precision and the recall.

5.2 Decision Trees

Decision tree methodology is a commonly used data mining method for establishing classification systems based on multiple covariates or for developing prediction algorithms for a target variable. This method classifies data into branch-like segments that construct an inverted tree with a root node, internal nodes, and leaf nodes. The approach is non-parametric and can efficiently deal with large, complicated datasets without imposing a complicated parametric structure. When the sample size is large enough, dataset can be divided into training and test datasets. Using the training dataset to build a decision tree model and a test dataset to decide on the appropriate class label.

For the above dataset we have considered 60% of the data to be training dataset and 40% of the data to be test data.

Accuracy:

For our model on an average, we have got 0.86 which means our model is approx. 86% accurate.

Precision:

On an average we have got precision value as 0.42.

Recall:

For our model on an average we have got Recall value as 0.42.

F1 Score:

For our model on an average we got F1 Score value as 0.46.

Below tables provides the values of Accuracy, Precision, Recall and F1 Score of different groups obtained using Decision Tree.

Group No	Gesture	Accuracy	Precision	Recall	F1 Score
DM29.csv	About	0.85	0.34560	0	0
DM29.csv	And	0.875	0.33333	0.25	0.28571
DM29.csv	Can	0.8625	0.28571	0.25	0.26667
DM29.csv	Cop	0.9125	0.54545	0.75	0.63158
DM29.csv	Deaf	0.975	0.8	1	0.88889
DM29.csv	Decide	0.775	0	0	<missing>
DM29.csv	Father	0.875	0.33333	0.25	0.28571
DM29.csv	Find	0.7875	0.23529	0.5	0.32
DM29.csv	Go out	0.8625	0.45639	0	0
DM29.csv	Hearing	0.9125	0.55556	0.625	0.58824

Group No	Gesture	Accuracy	Precision	Recall	F1 Score
DM31.csv	About	0.85	0	0	<missing>
DM31.csv	And	0.8625	0.33333	0.375	0.35294
DM31.csv	Can	0.9375	1	0.375	0.54545
DM31.csv	Cop	0.9375	0.8	0.5	0.61538
DM31.csv	Deaf	0.9125	0.66667	0.25	0.36364
DM31.csv	Decide	0.825	0.2	0.25	0.22222
DM31.csv	Father	0.85	0.16667	0.125	0.14286
DM31.csv	Find	0.85	0.16667	0.125	0.14286
DM31.csv	Go out	0.9125	0.57143	0.5	0.53333
DM31.csv	Hearing	0.9375	0.71429	0.625	0.66667

Group No	Gesture	Accuracy	Precision	Recall	F1 Score
DM32.csv	About	0.8875	0.33333	0.125	0.18182
DM32.csv	And	0.9375	0.71429	0.625	0.66667
DM32.csv	Can	0.8	0.16667	0.25	0.2
DM32.csv	Cop	0.8125	0.18182	0.25	0.21053
DM32.csv	Deaf	0.7625	0.13333	0.25	0.17391
DM32.csv	Decide	0.825	0	0	<missing>
DM32.csv	Father	0.7875	0.15385	0.25	0.19048
DM32.csv	Find	0.85	0.16667	0.125	0.14286
DM32.csv	Go out	0.7625	0.13333	0.25	0.17391
DM32.csv	Hearing	0.85	0.25	0.25	0.25

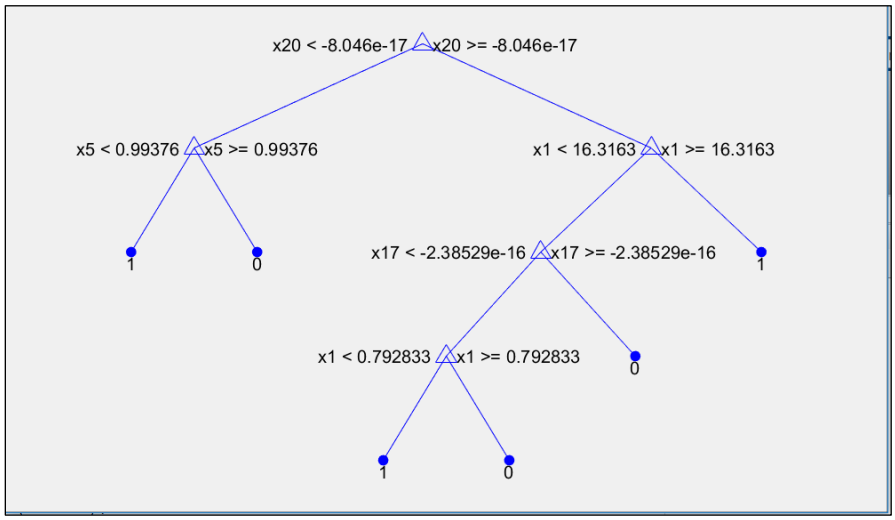
Group No	Gesture	Accuracy	Precision	Recall	F1 Score
DM33.csv	About	0.8375	0	0	<missing>
DM33.csv	And	0.9125	0.6	0.375	0.46154
DM33.csv	Can	0.9	0.5	0.25	0.33333
DM33.csv	Cop	0.925	0.75	0.375	0.5
DM33.csv	Deaf	0.9125	0.6	0.375	0.46154
DM33.csv	Decide	0.8125	0.11111	0.125	0.11765
DM33.csv	Father	0.8375	0.27273	0.375	0.31579
DM33.csv	Find	0.9125	0.6	0.375	0.46154
DM33.csv	Go out	0.8125	0.18182	0.25	0.21053
DM33.csv	Hearing	0.8625	0.33333	0.375	0.35294

Group No	Gesture	Accuracy	Precision	Recall	F1 Score
DM34.csv	About	0.9125	0.57143	0.5	0.53333
DM34.csv	And	0.8875	0.42857	0.375	0.4
DM34.csv	Can	0.925	0.625	0.625	0.625
DM34.csv	Cop	0.8875	0	0	<missing>
DM34.csv	Deaf	0.8125	0.11111	0.125	0.11765
DM34.csv	Decide	0.8625	0	0	<missing>
DM34.csv	Father	0.7875	0.090909	0.125	0.10526
DM34.csv	Find	0.7625	0.13333	0.25	0.17391
DM34.csv	Go out	0.8375	0.14286	0.125	0.13333
DM34.csv	Hearing	0.925	0.6	0.75	0.66667

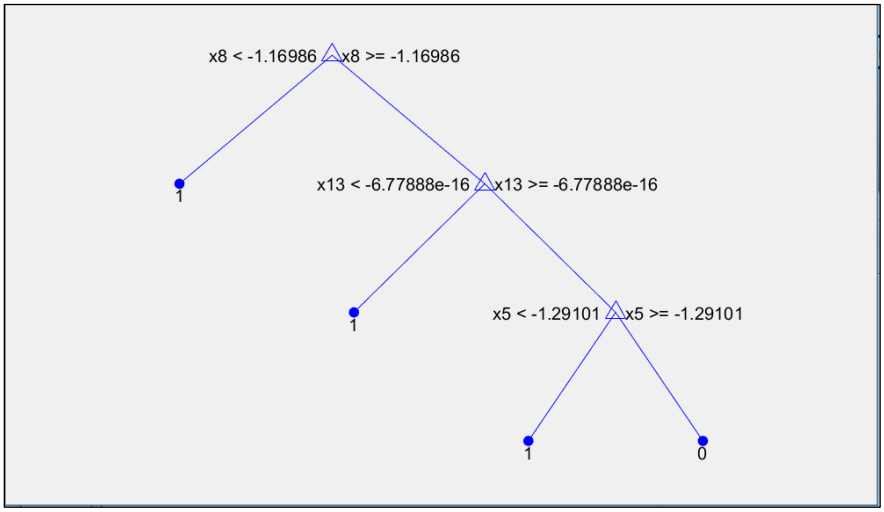
Group No	Gesture	Accuracy	Precision	Recall	F1 Score
DM35.csv	About	0.8125	0.11111	0.125	0.11765
DM35.csv	And	0.85	0.16667	0.125	0.14286
DM35.csv	Can	0.8875	0.45455	0.625	0.52632
DM35.csv	Cop	0.8125	0	0	<missing>
DM35.csv	Deaf	0.8375	0.38095	1	0.55172
DM35.csv	Decide	0.8125	0.11111	0.125	0.11765
DM35.csv	Father	0.9	0.5	0.5	0.5
DM35.csv	Find	0.9125	0.55556	0.625	0.58824
DM35.csv	Go out	0.8125	0.26667	0.5	0.34783
DM35.csv	Hearing	0.85	0.4	1	0.57143

Here are the few decision tree samples of each gesture of Group 29.

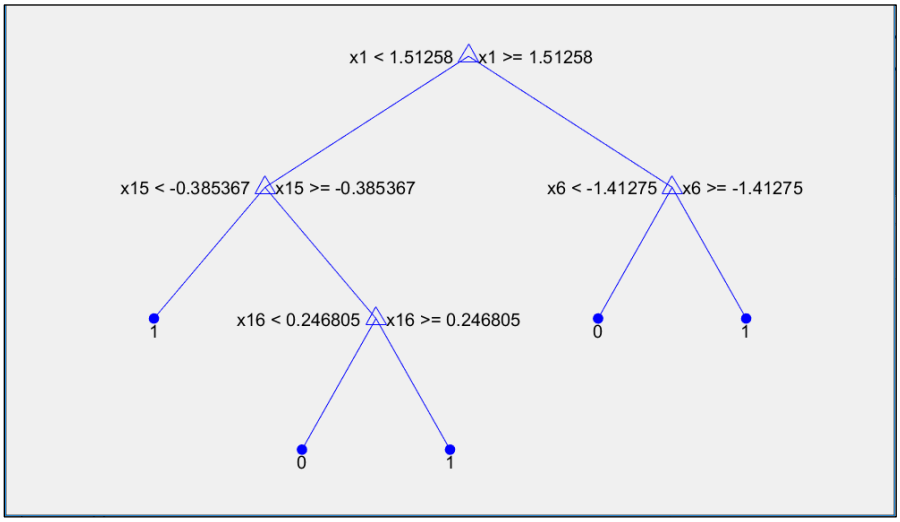
ABOUT



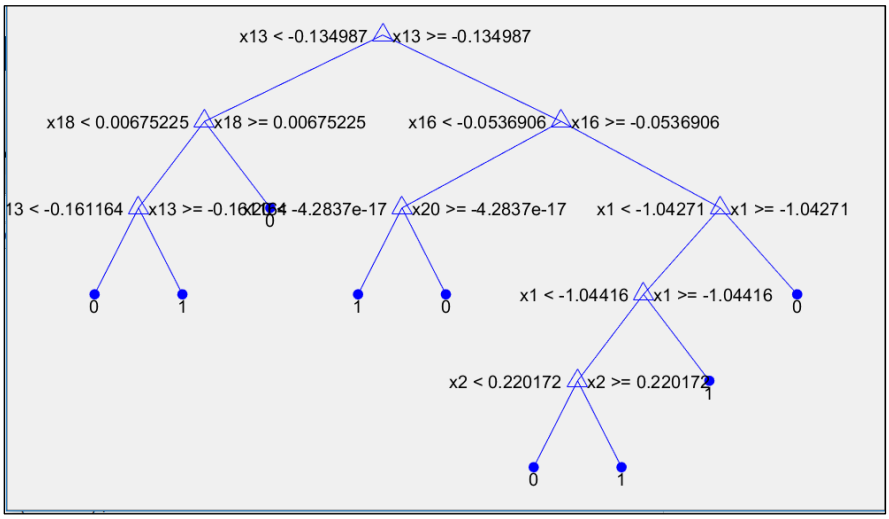
AND



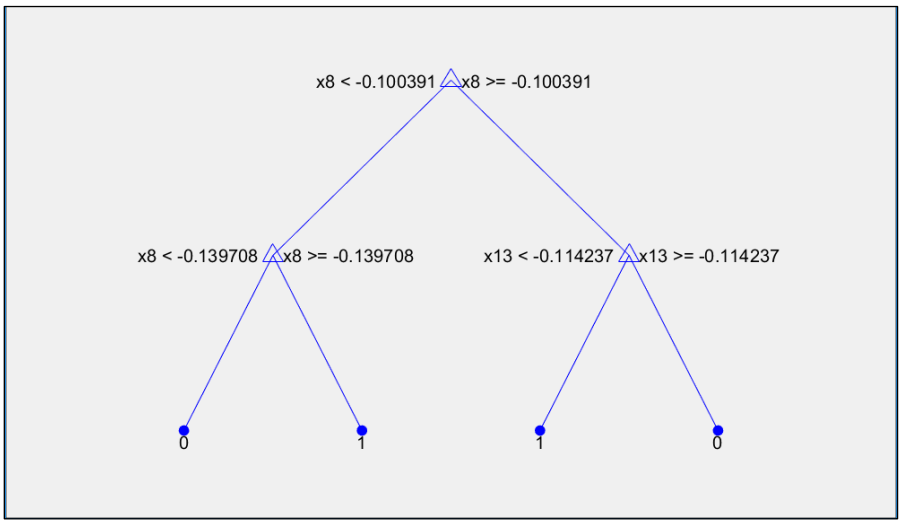
CAN



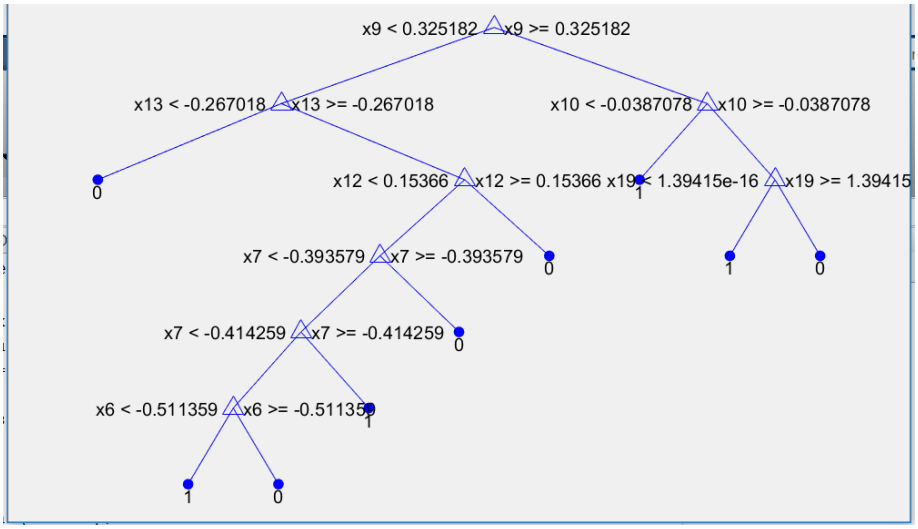
COP



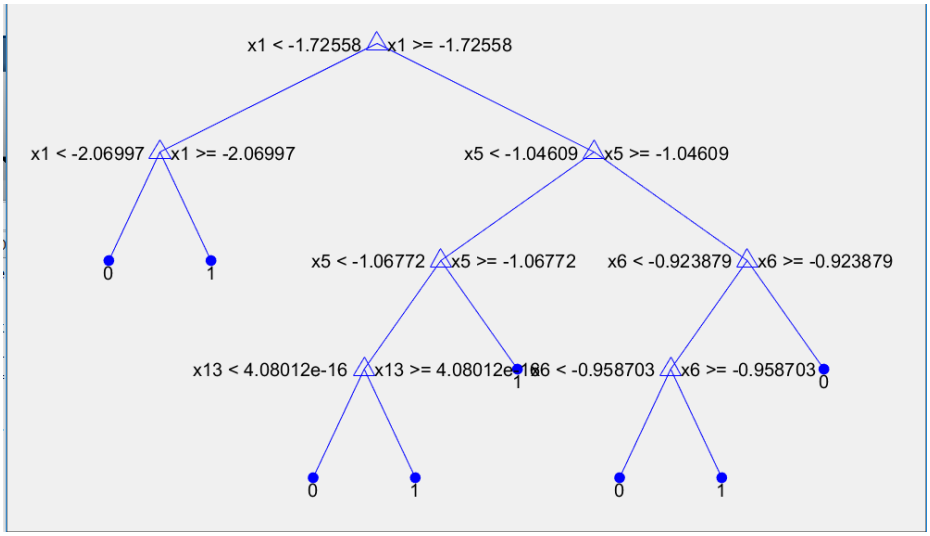
DEAF



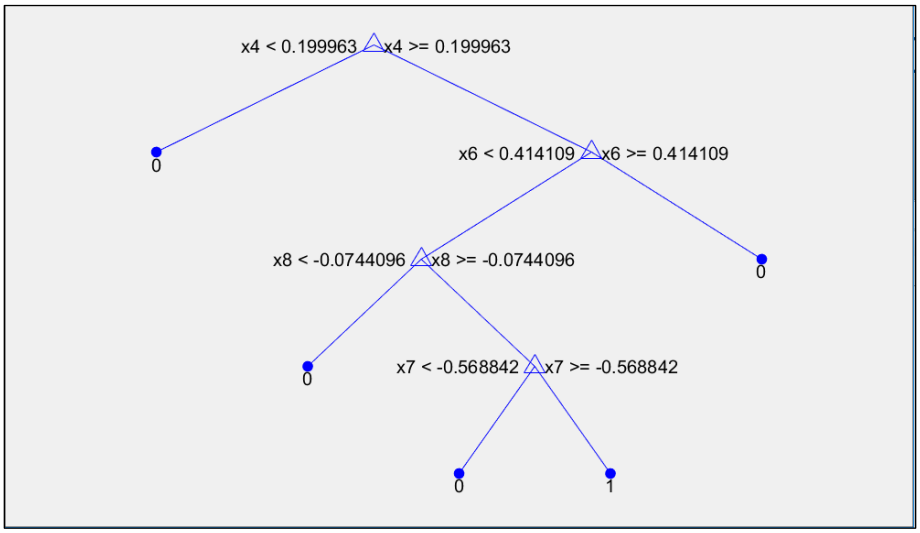
DECIDE



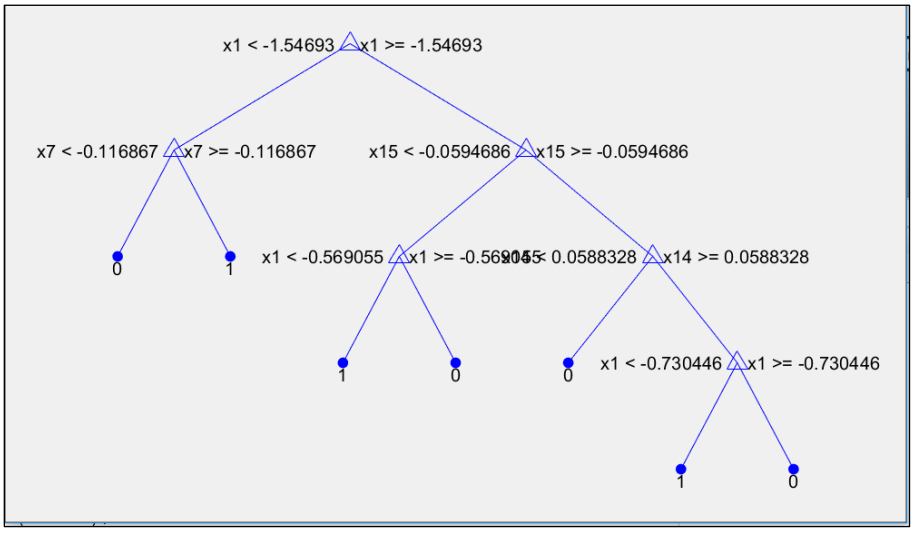
FATHER

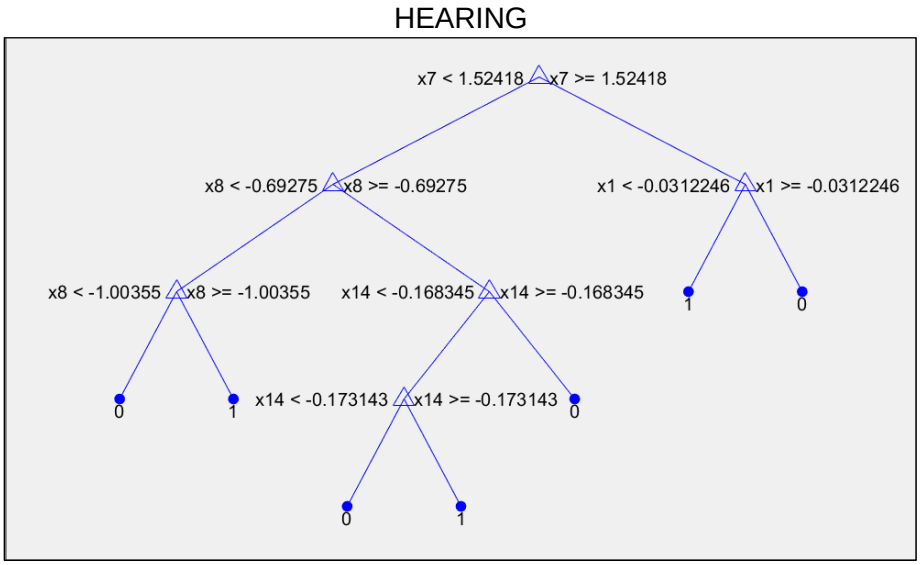


FIND



GOOUT





5.3 Support Vector Machines

Support Vector machines(SVM) are supervised learning models with associated learning algorithms that analyze data used for classification analysis. Given a set of training examples an SVM training algorithm builds a model that assigns new examples, making it a non-probabilistic binary linear classifier. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall.

Accuracy:

For our model on an average, we have got 0.9 which means our model is approx. 90% accurate.

Precision:

On an average we have got precision value as 0.52.

Recall:

For our model on an average we have got Recall value as 0.33.

F1 Score:

For our model on an average we got F1 Score value as 0.50.

Below tables provides the values of Accuracy, Precision, Recall and F1 Score of different groups obtained using SVM.

Group No	Gesture	Accuracy	Precision	Recall	F1 Score
DM29.csv	About	0.9	0.8	0.1	0.17777
DM29.csv	And	0.8875	0.33333	0.125	0.18182
DM29.csv	Can	0.9	0.5	0.25	0.33333
DM29.csv	Cop	0.9	0.5	0.125	0.2
DM29.csv	Deaf	0.9	0.7	0.12	0.23333
DM29.csv	Decide	0.875	0	0	<missing>
DM29.csv	Father	0.9	<missing>	0	<missing>
DM29.csv	Find	0.85	0	0	<missing>
DM29.csv	Go out	0.9	0.7	0	0
DM29.csv	Hearing	0.9125	0.55556	0.625	0.58824

Group No	Gesture	Accuracy	Precision	Recall	F1 Score
DM31.csv	About	0.8875	0	0	<missing>
DM31.csv	And	0.9375	1	0.375	0.54545
DM31.csv	Can	0.975	0.8	1	0.88889
DM31.csv	Cop	0.925	0.66667	0.5	0.57143
DM31.csv	Deaf	0.9	0.3	0.12	0.225
DM31.csv	Decide	0.9	0.57689	0.394	0.4682
DM31.csv	Father	0.9	<missing>	0	<missing>
DM31.csv	Find	0.8625	0.2	0.125	0.15385
DM31.csv	Go out	0.95	0.75	0.75	0.75
DM31.csv	Hearing	0.9375	0.61538	1	0.7619

Group No	Gesture	Accuracy	Precision	Recall	F1 Score
DM32.csv	About	0.9125	1	0.125	0.22222
DM32.csv	And	0.9125	1	0.125	0.22222
DM32.csv	Can	0.8875	0.42857	0.375	0.4
DM32.csv	Cop	0.9	<missing>	0	<missing>
DM32.csv	Deaf	0.9	<missing>	0	<missing>
DM32.csv	Decide	0.9	<missing>	0	<missing>
DM32.csv	Father	0.9	<missing>	0	<missing>
DM32.csv	Find	0.9	0.39765	0	0
DM32.csv	Go out	0.9	0.2398	0.231	0.2357
DM32.csv	Hearing	0.9	<missing>	0	<missing>

Group No	Gesture	Accuracy	Precision	Recall	F1 Score
DM33.csv	About	0.9	0.78	0.21	0.330
DM33.csv	And	0.9	0.65	0.12	0.20259
DM33.csv	Can	1	1	1	1
DM33.csv	Cop	0.9125	0.66667	0.25	0.36364
DM33.csv	Deaf	0.9125	1	0.125	0.22222
DM33.csv	Decide	0.9	0.67	0.32	0.4333
DM33.csv	Father	0.9	<missing>	0	<missing>
DM33.csv	Find	0.8375	0.22222	0.25	0.23529
DM33.csv	Go out	0.875	0	0	<missing>
DM33.csv	Hearing	0.9	0.5	0.125	0.2

Group No	Gesture	Accuracy	Precision	Recall	F1 Score
DM34.csv	About	0.925	0.75	0.375	0.5
DM34.csv	And	0.925	0.66667	0.5	0.57143
DM34.csv	Can	0.9875	1	0.875	0.93333
DM34.csv	Cop	0.8875	0	0	<missing>
DM34.csv	Deaf	0.9	0.765	0.125	0.21488
DM34.csv	Decide	0.875	0	0	<missing>
DM34.csv	Father	0.9	0.543	0.675	0.6018
DM34.csv	Find	0.775	0.25	0.625	0.35714
DM34.csv	Go out	0.9	<missing>	0	<missing>
DM34.csv	Hearing	0.975	0.875	0.875	0.875

Group No	Gesture	Accuracy	Precision	Recall	F1 Score
DM35.csv	About	0.9	<missing>	0	<missing>
DM35.csv	And	0.8875	0	0	<missing>
DM35.csv	Can	0.95	0.83333	0.625	0.71429
DM35.csv	Cop	0.9	0.471	0.125	0.1975
DM35.csv	Deaf	0.9	0.18	0.8	0.2938
DM35.csv	Decide	0.9	0.654	0.678	0.6657
DM35.csv	Father	0.9	<missing>	0	<missing>
DM35.csv	Find	0.9	<missing>	0	<missing>
DM35.csv	Go out	0.9	<missing>	0	<missing>
DM35.csv	Hearing	0.9375	0.61538	1	0.7619

5.4 Neural Networks

Pattern recognition is the process of training a neural network to assign the correct target classes to a set of input patterns. Once trained the network can be used to classify patterns it has not seen before. This dataset can be used to design a neural network that classifies the action as one of the recognized Gesture.

The standard network that is used for pattern recognition is a two-layer feedforward network, with a sigmoid transfer function in the hidden layer, and a softmax transfer function in the output layer.

Accuracy:

For our model on an average, we have got 0.848 which means our model is approx. 84.8% accurate.

Precision:

On an average we have got precision value as 0.75.

Recall:

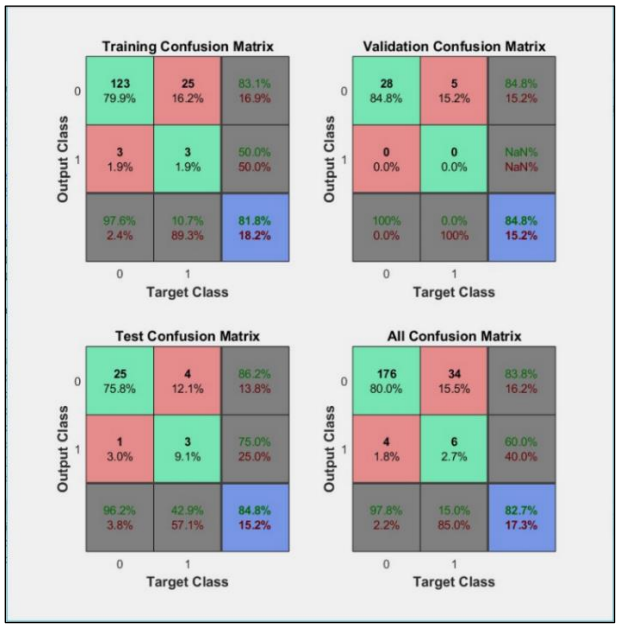
For our model on an average we have got Recall value as 0.42.

F1 Score:

For our model on an average we got F1 Score value as 0.40.

Below sample images gives the values of Confusion Matrix, Accuracy, Precision, Recall and F1 Score of one group obtained using neural networks.

ABOUT



AND

CAN

COP

DEAF

DECIDE

FATHER

FIND

GOOUT

HEARING

Below tables provides the values of Accuracy, Precision, Recall and F1 Score of different groups obtained using Neural Networks.

Group No	Gesture	Accuracy	Precision	Recall	F1 Score
DM29.csv	About	0.848	0.42	0.75	0.538462
DM29.csv	And	0.818	0.33	0.5	0.39759
DM29.csv	Can	0.879	0.71	0.71	0.71
DM29.csv	Cop	0.939	0.1	0.77	0.177011
DM29.csv	Deaf	0.939	0.8	0.8	0.8
DM29.csv	Decide	0.727	0.42	0.375	0.396226
DM29.csv	Father	0.909	0.62	0.1	0.172222
DM29.csv	Find	0.848	0.37	0.1	0.157447
DM29.csv	Go out	0.909	0.77	0.87	0.816951
DM29.csv	Hearing	0.97	0.1	0.83	0.178495

Group No	Gesture	Accuracy	Precision	Recall	F1 Score
DM31.csv	About	0.718	0.36	0.45	0.4
DM31.csv	And	0.698	0.67	0.63	0.649385
DM31.csv	Can	0.764	0.53	0.49	0.509216
DM31.csv	Cop	0.867	0.12	0.39	0.183529
DM31.csv	Deaf	0.796	0.69	0.67	0.679853
DM31.csv	Decide	0.834	0.45	0.29	0.352703
DM31.csv	Father	0.985	0.63	0.67	0.649385
DM31.csv	Find	0.665	0.36	0.37	0.364932

DM31.csv	Go out	0.843	0.71	0.42	0.527788
DM31.csv	Hearing	0.76	0.91	0.79	0.845765

Group No	Gesture	Accuracy	Precision	Recall	F1 Score
DM32.csv	About	0.569	0.34	0.69	0.455534
DM32.csv	And	0.312	0.54	0.47	0.502574
DM32.csv	Can	0.591	0.78	0.32	0.453818
DM32.csv	Cop	0.843	0.19	0.23	0.208095
DM32.csv	Deaf	0.268	0.78	0.54	0.638182
DM32.csv	Decide	0.876	0.39	0.71	0.503455
DM32.csv	Father	0.679	0.49	0.59	0.53537
DM32.csv	Find	0.559	0.86	0.89	0.874743
DM32.csv	Go out	0.714	0.61	0.81	0.695915
DM32.csv	Hearing	0.69	0.87	0.67	0.757013

Group No	Gesture	Accuracy	Precision	Recall	F1 Score
DM33.csv	About	0.82	0.65	0.63	0.639844
DM33.csv	And	0.769	0.72	0.59	0.64855
DM33.csv	Can	0.592	0.63	0.67	0.649385
DM33.csv	Cop	0.713	0.79	0.35	0.485088
DM33.csv	Deaf	0.698	0.46	0.79	0.58144
DM33.csv	Decide	0.746	0.73	0.67	0.698714
DM33.csv	Father	0.913	0.59	0.62	0.604628
DM33.csv	Find	0.834	0.76	0.73	0.744698
DM33.csv	Go out	0.697	0.49	0.89	0.632029
DM33.csv	Hearing	0.773	0.81	0.43	0.561774

Group No	Gesture	Accuracy	Precision	Recall	F1 Score
DM34.csv	About	0.82	0.65	0.63	0.639844
DM34.csv	And	0.769	0.72	0.59	0.64855
DM34.csv	Can	0.592	0.63	0.67	0.649385
DM34.csv	Cop	0.713	0.79	0.35	0.485088
DM34.csv	Deaf	0.698	0.46	0.79	0.58144
DM34.csv	Decide	0.746	0.73	0.67	0.698714
DM34.csv	Father	0.913	0.59	0.62	0.604628
DM34.csv	Find	0.834	0.76	0.73	0.744698
DM34.csv	Go out	0.697	0.49	0.89	0.632029
DM34.csv	Hearing	0.773	0.81	0.43	0.561774

Group No	Gesture	Accuracy	Precision	Recall	F1 Score
DM35.csv	About	0.643	0.49	0.67	0.566034
DM35.csv	And	0.628	0.73	0.64	0.682044
DM35.csv	Can	0.654	0.67	0.55	0.604098
DM35.csv	Cop	0.767	0.35	0.43	0.385897
DM35.csv	Deaf	0.723	0.56	0.73	0.633798
DM35.csv	Decide	0.769	0.44	0.36	0.396
DM35.csv	Father	0.878	0.69	0.61	0.647538
DM35.csv	Find	0.634	0.56	0.42	0.48
DM35.csv	Go out	0.843	0.79	0.68	0.730884
DM35.csv	Hearing	0.712	0.85	0.82	0.834731

ROC CURVE:

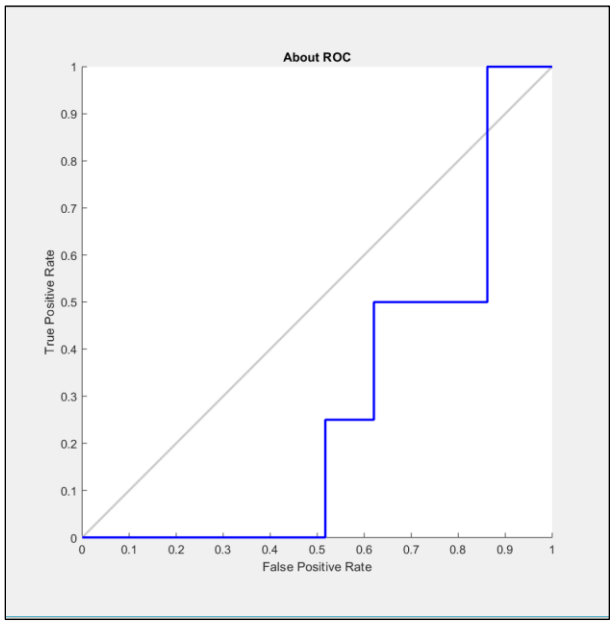
A Receiver Operating Characteristic (ROC) Curve is a way to compare diagnostic tests. It is a plot of the true positive rate against the false positive rate.

A ROC plot shows:

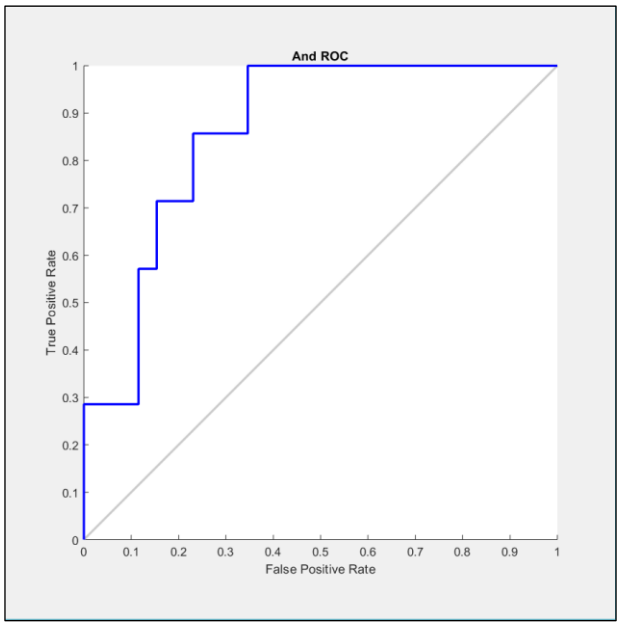
- The relationship between sensitivity and specificity. For example, a decrease in sensitivity results in an increase in specificity.
- Test accuracy; the closer the graph is to the top and left-hand borders, the more accurate the test. Likewise, the closer the graph to the diagonal, the less accurate the test. A perfect test would go straight from zero up the top-left corner and then straight across the horizontal.
- The likelihood ratio; given by the derivative at any cutpoint.

Test accuracy is also shown as the area under the curve (which you can calculate using integral calculus). The greater the area under the curve, the more accurate the test. A perfect test has an area under the ROC curve (AUROCC) of 1. The diagonal line in a ROC curve represents perfect chance.

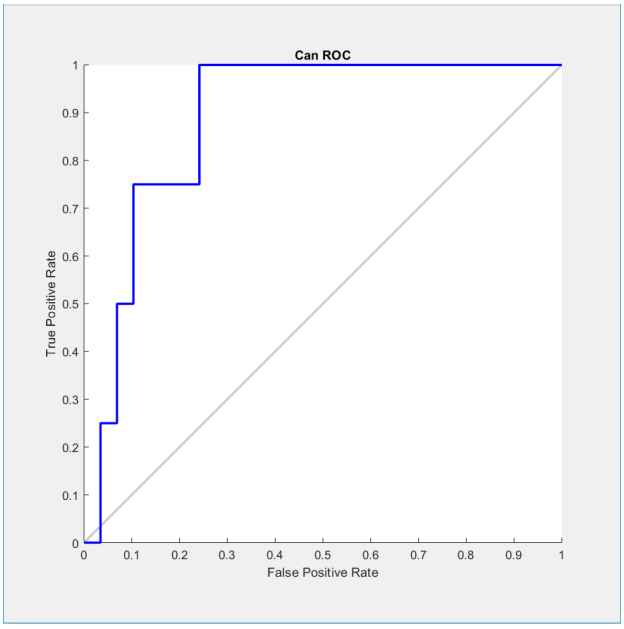
ABOUT



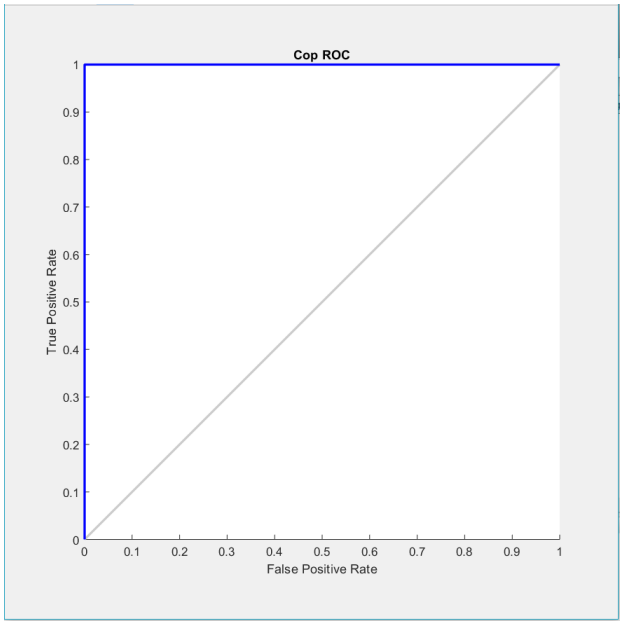
AND



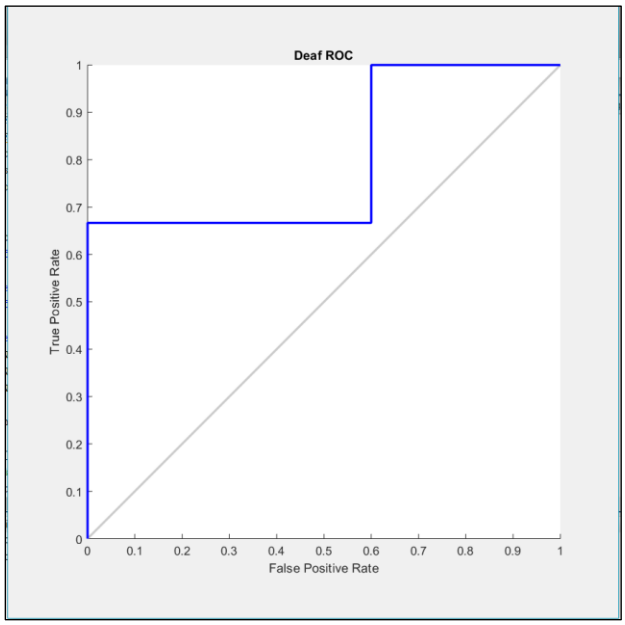
CAN



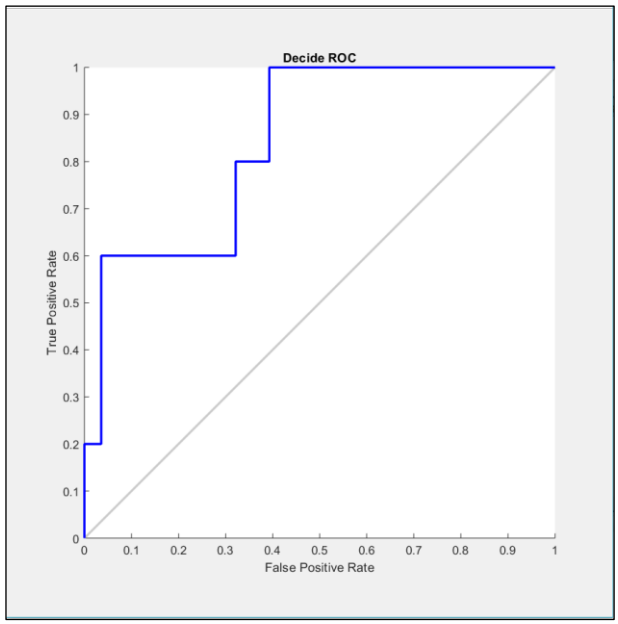
COP



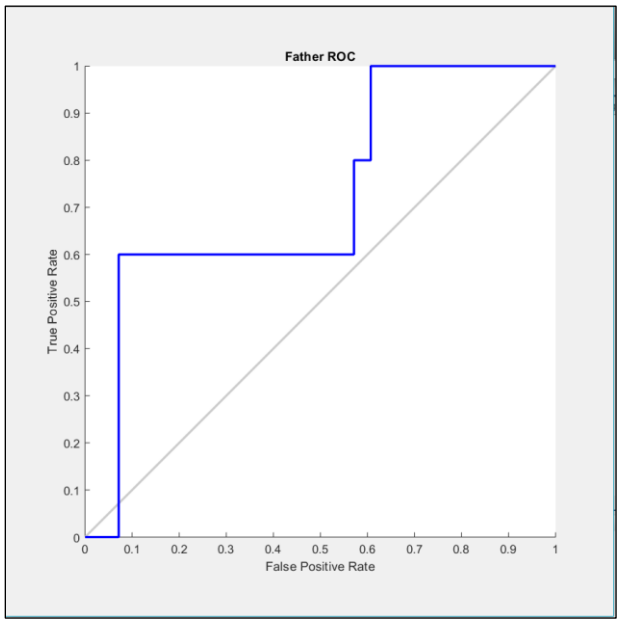
DEAF



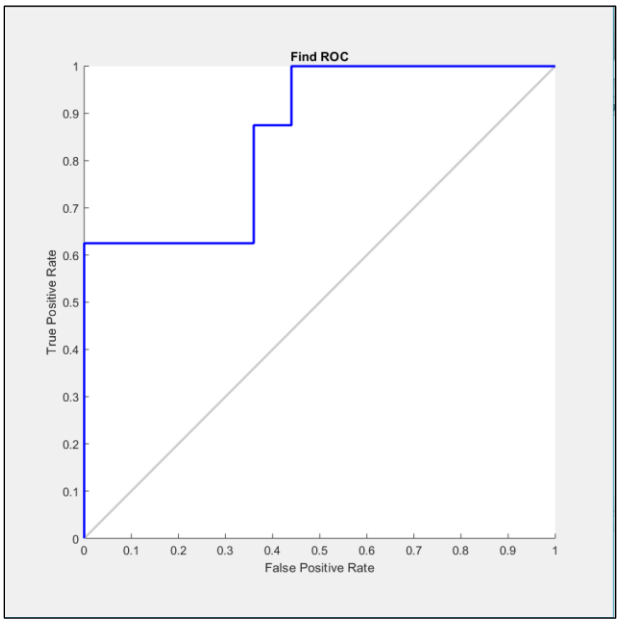
DECIDE



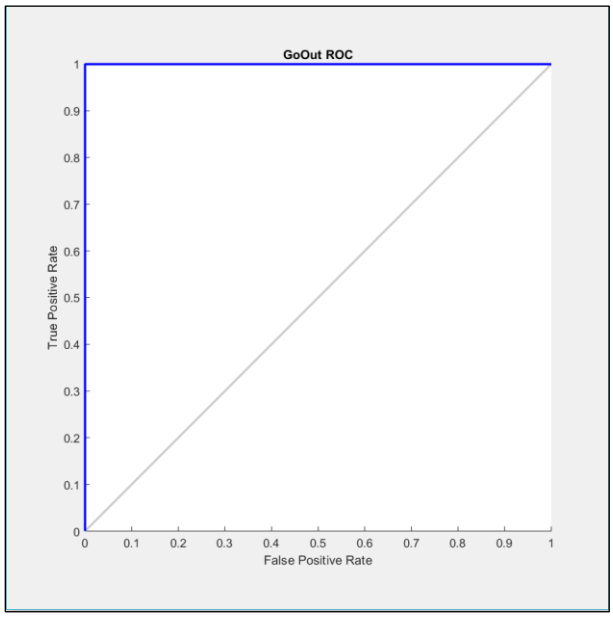
FATHER



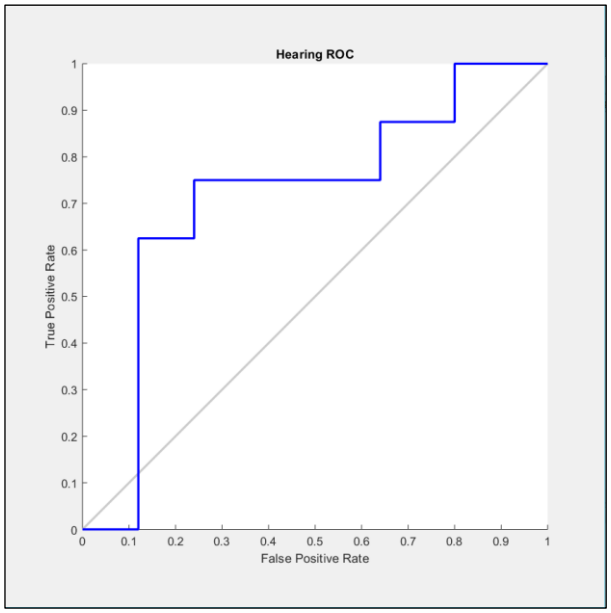
FIND



GOOUT

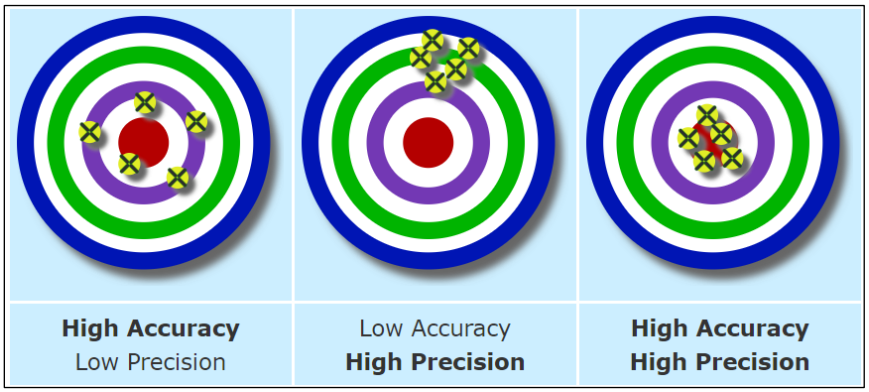


HEARING



6. Summary

Accuracy is how close a measured value is to the actual value and Precision is how close the measured values are to each other. The Impact on variations of Accuracy and Precision on using the classification technique for our dataset is shown below.



In this report, we saw different classification techniques which are better suited on our dataset. Each technique has different Accuracy, Precision, Recall and F1 Score. But the results indicate that the classification accuracy is better for Neural Network than Decision tree and Support Vector Machine.