



***Report on***

**“Breast Cancer Prediction”**

*Submitted in partial fulfillment of the requirements for Sem IV*

**IMAGE PROCESSING AND DATA VISUALIZATION USING  
MATLAB**

**Bachelor of Technology  
in  
Computer Science & Engineering**

*Submitted by:*

<b>Deepa Shree C V</b>	<b>PES2UG19CS105</b>
<b>Pallavi Prabhu</b>	<b>PES2UG19CS274</b>
<b>Swati Maste</b>	<b>PES2UG19CS419</b>
<b>Chinmayi Shetty</b>	<b>PES2UG19CS901</b>

*Under the guidance of*

**Prof. Swati Pratap Jagdale**  
Assistant Professor  
PES University, Bengaluru

**January – May 2021**

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

**PES UNIVERSITY**

(Established under Karnataka Act No. 16 of 2013)  
100ft Ring Road, Bengaluru – 560 085, Karnataka, India

## TABLE OF CONTENTS

<b>Chapter No.</b>	<b>Title</b>	<b>Page No.</b>
<b>1.</b>	<b>Abstract</b>	<b>3</b>
<b>2.</b>	<b>Problem statement</b>	<b>3-5</b>
<b>3.</b>	<b>Module Description</b>	<b>5</b>
<b>4.</b>	<b>High Level Design/Architecture</b>	<b>6</b>
<b>5.</b>	<b>Implementation (Algorithm, platform, s/w, tools used etc...)</b>	<b>7-8</b>
<b>6.</b>	<b>Result snapshots</b>	<b>9-14s</b>

## **1. Abstract:**

Women are seriously threatened by breast cancer with high morbidity and mortality. The lack of robust prognosis models results in difficulty for doctors to prepare a treatment plan that may prolong patient survival time. Hence, the requirement of time is to develop the technique which gives minimum error to increase accuracy.

Our project aims to predict the type of breast cancer tumor with machine learning using matlab. Four algorithms SVM, Decision tree, Discriminant analysis and KNN which predict the breast cancer outcome have been used on our dataset. These models predict the type of breast cancer given the test dataset with 70 percent accuracy. We have implemented data visualization using matlab for better understanding and analysis of our dataset.

## **2. Problem Statement**

### **Breast Cancer Prediction Using Machine Learning:**

This project aims to predict whether the cancer tumor of the patient is benign or malignant using classification algorithms in the machine learning toolbox of Matlab.

The dataset is a collection of 3 csv files, which contains data about patients who have recovered, died and are still undergoing treatment. We aim to combine from all these three datasets and predict if the tumor is cancerous.

All these datasets contain 30 variables, out of which 4 are numerical variables and the rest are categorical. They contain 1134 rows, all combined in total. The columns contain information about the patient like gender, education, age, weight, thickness of the tumor, pregnancy, menopausal age and other important parameters related to a cancer patient. The dataset contains around 10% of null values.

We want to find out if the patient has a Benign or malignant tumor

We have three datasets with the same features and target, but for three categories such as

- The patients who are under treatment
- The patients who are recovered
- The patients who died

The features could be described as bellow:

- patient\_id: the id of the patient
- gender: the gender of the patient which “Female” is 0 and “Male” is 1
- education: the education levels of the patient which Illiterate=0, Elementary= 1, Middle School =2 , High School =3 , Diploma = 4, Associate =5 , Bachelor =6 , Master = 7
- treatment\_date = the date(year) which the patient would receive the treatment
- id\_healthcenter:: is the id for the healthcare center
- idtreatmentregion: the region which the patient would receive the treatment
- hereditary\_history: the patient has the hereditary history of disease which 1 means “Yes” and 0 means “No”
- birth\_date: birth date (year) of the patient
- age: the age of the patient
- weight: the weight of the patient
- thickness\_tumor: the thickness of the tumor detected in the patient body
- marital\_status: the marital status of the patient includes : 1 means married and 0 means single
- marital\_length: the age of the martial status includes 0 means under 10 years, 1 means above 10 years
- pregnancy\_experience: the patient has the pregnancy experience which 0 means has not experience and 1 means has experience
- giving\_birth: the patient has experienced giving the birth. Each number means the number of giving birth
- age\_FirstGivingBirth : in which age the patient has the first experience of giving a birth, which before age 30 equals 0 and after age 30 equals 1
- abortion: the patient has experience of abortion which 0 means has not and 1 means has
- blood: the types of bloods A+ =0, A- = 1, AB+ = 2, AB- = 3, B+ = 4, B- = 5, O+ = 6, O- = 7
- taking\_heartMedicine: it says if the patient takes the heart medicine or not which 0 means does not and 1 means does
- takingbloodpressure\_medicine: it says if the patient takes the blood pressure or not which 0 means does not and 1 means does
- takinggallbladder\_disease\_medicine: it says the patient takes the gallbladder medicine or not which 0 means does not and 1 means does
- smoking: it says if the patient smokes or not which 0 means does not and 1 means does
- alcohol: it says if the patient drinks alcohol or not which 0 means does not and 1 means does
- breast\_pain: if the patient has pain in the breast part which 0 means has not and 1 means has
- radiation\_history: if the patient has experience with radiation in the breast area which 0 means has not and 1 means has
- Birth\_control(Contraception): the patient takes actions for birth control which 0 means does not and 1 means does

- **menstrual\_age**: at which age the patient starts natural menstrual which 0 means the patient does not start menstrual, 1 means under age 12, and 2 means above age 12
- **menopausal\_age**: at which age the patient starts natural menopausal which 0 means does not start, 1 means at under age 50, and 2 means above age 50
- **condition**: the condition of the patient which categorized into different categories such as under treatment, recovered, death
- **Benignmalignanttumor**: is the target of our datasets, the type of tumor which Benign is 0 and malignant is 1

**Source of the data:** [kaggle.com](https://www.kaggle.com)

### 3. Module Description

**1.pre-processing.m:** This file includes the code that helps in pre-processing the input data. This is required to clean the data and make it much easier for further analysis.

**2. data\_vis.mlx:** This file is used for data visualisation which helps in better understanding and interpretation of the data set.

**3.ml\_input\_data.m:** This file contains the code that helps to split the data set into training and testing datasets.

**4.discriminant\_analysis.m:** This file has the code that trains the data set using the Discriminant analysis model. The method `fitcdiscr()` helps to do so.

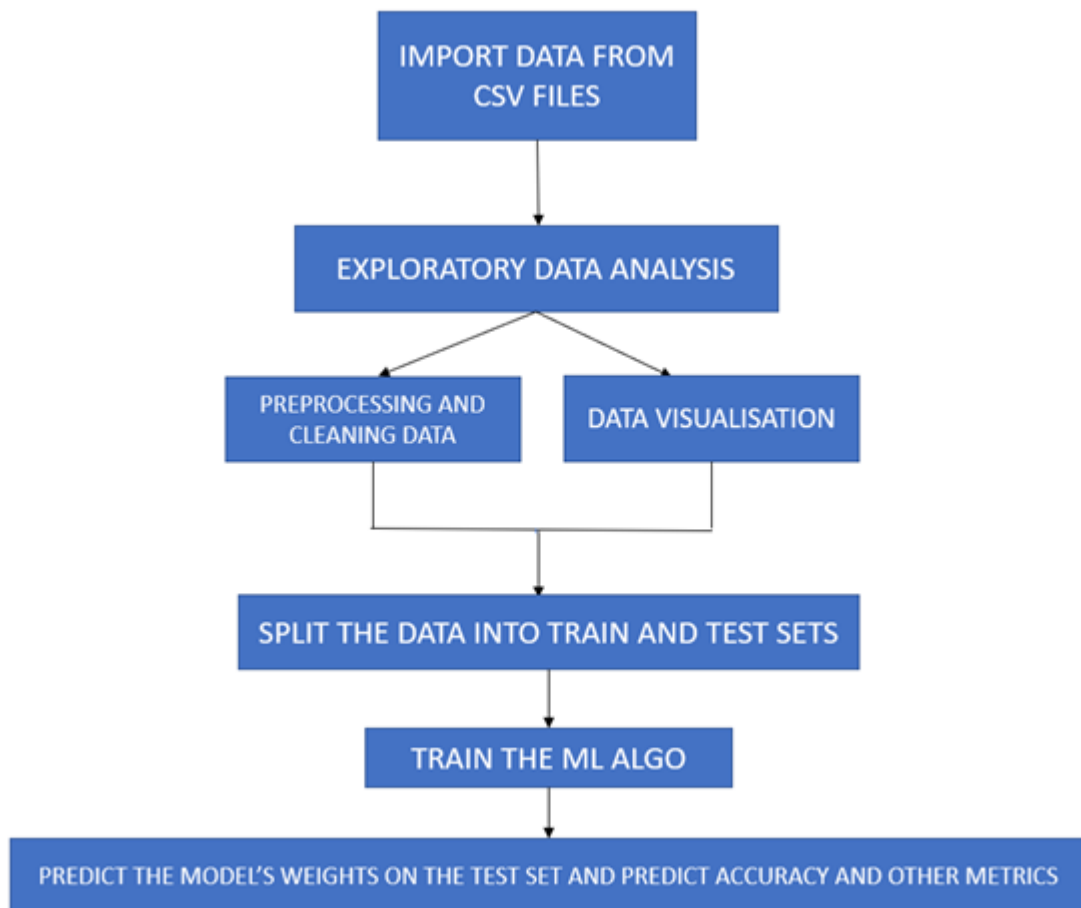
**5.decision\_tree.m:** This file has the code that trains the data set using the Discriminant analysis model. The method `fitctree()` helps to do so.

**6.knn.m:** This file has the code that trains the data set using the K nearest neighbours. The method `fitcknn()` helps to do so.

**7.svm.m:** This file has the code that trains the data set using the Support Vector machine. The method `fitsvm()` helps to do so.

**8.modle\_outputs.m:** This file will produce the output of the matrix of all the models .

#### 4. High Level Architecture



## 5. Implementation

We have made use of the following functions for data pre processing, visualisation and to train the data set to predict whether the tumour is malignant or benign.

### Reading data:

- Standard modules like:
  - readtable: Create table from a csv file
  - vertcat: to merge all the csv files into one single table

### Pre-Processing:

- isnan(): to find the null values
- table2cell(): to convert table to cell to perform relational operations
- corrcoef(): to find the correlation between numerical variables.
- summary(): to get an overview of the dataset
- isoutlier(): to detect outliers in the dataset
- mean(), median(), mode(): to find measures of central tendency
- std(), var(), max(), min(): measures of deviation

### Data Visualisation:

- colormap(), colorbar(), imagesc(), corrcoef(): to plot the correlation matrix on a heatmap
- pie(): to plot 2-D pie charts.
- pie3(): to plot categorical data on 3-D pie charts
- scatter(): to plot a 2-D scatter plot to find relationship between age, menstrual age and type of tumor.
- bar(): to find the frequencies of each categorical variable in the dataset and plot grouped bar graph as well as bar graphs.

## **Machine learning:**

Different classifier algorithms of the machine learning and statistics toolbox to be used to find the tumor type.

- `fitctree()`: fit classifier tree to train the dataset
- `fitcknn(tbl,y)`: it returns k-nearest neighbour classification model based on the predictor variables in the table. The predictor variables in the table 'tbl' and the response array 'y'. In our code tbl is x\_train and y is y\_train.
- `fitcdiscr(x_train,y_train)`: It returns a fitted discriminant analysis model based on the input variables contained in the table x\_train and the response y\_train.
- `fitcsvm()`: to fit the support vector machine
- `predict()`: to run the classifier model on the test set
- `confusionchart()`: to plot the confusion matrix for the predictions

Metrics like accuracy on test and train sets, precision, recall and F1 score have also been calculated based on the results of predict and confusion chart.



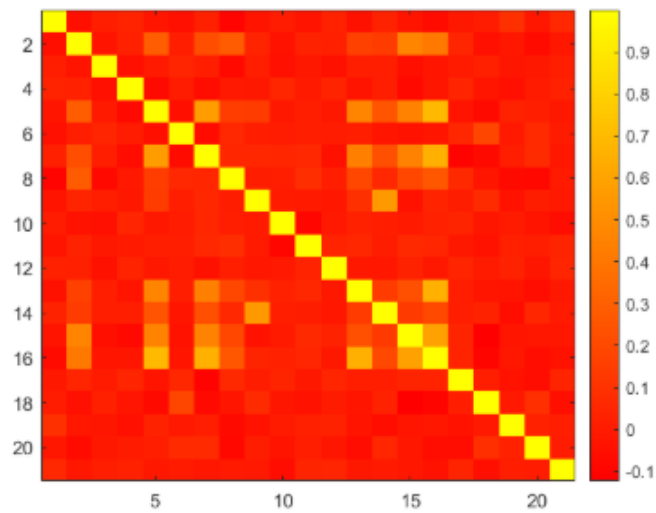
## Result snapshots

Pre-processing:

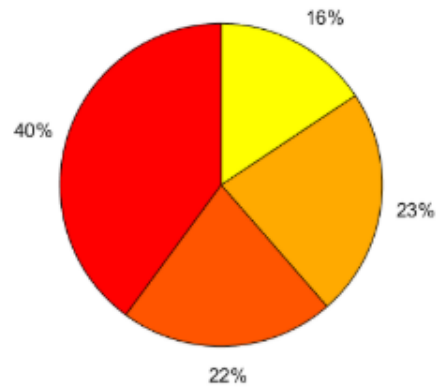
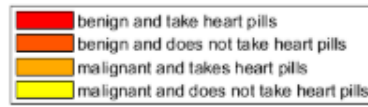
Data visualisation:

### 1. Correlation matrix:

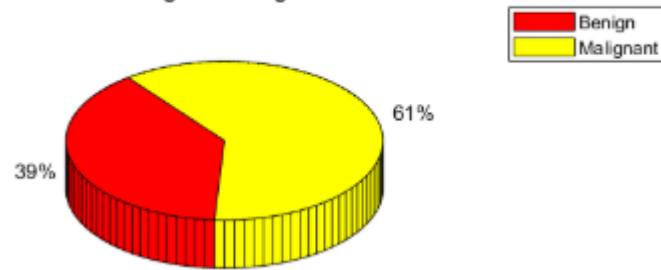
```
1.Benign_malignant_cancer
2.Birth_control(Contraception)
3.breast_pain
4.giving_birth
5.alcohol
6.age_FirstGivingBirth
7.abortion
8.age
9.weight
10.thickness_tumor
11.hereditary_history
12.marital_status
13.menopausal_age
14.menstrual_age
15.pregnancy_experience
16.radiation_history
17.smoking
18.taking_blood_pressure_medicine
19.taking_gallbladder_disease_medicine
20.taking_heartMedicine
```



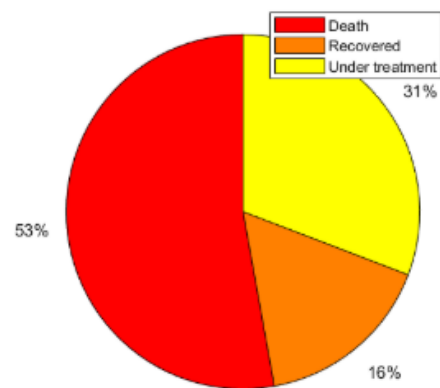
### 2. Pie Charts(2-D & 3-D)



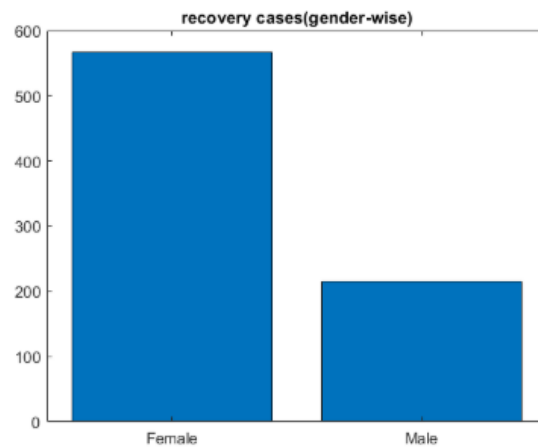
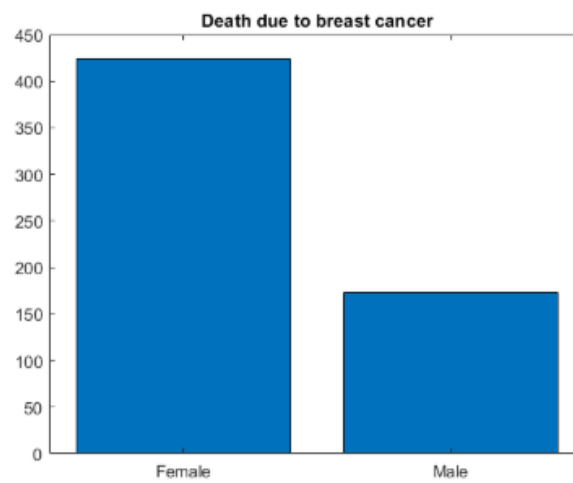
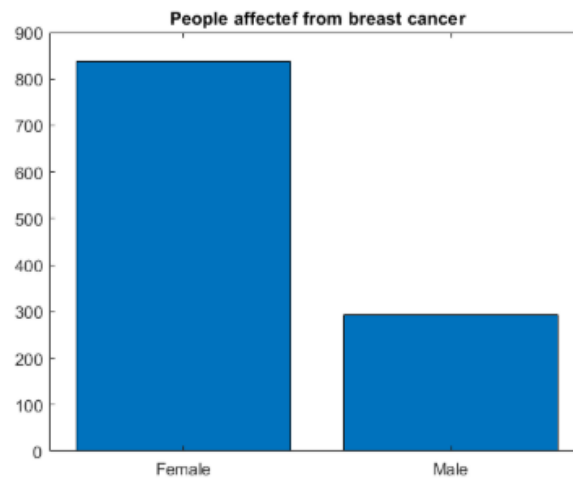
**Benign Vs Malignant**

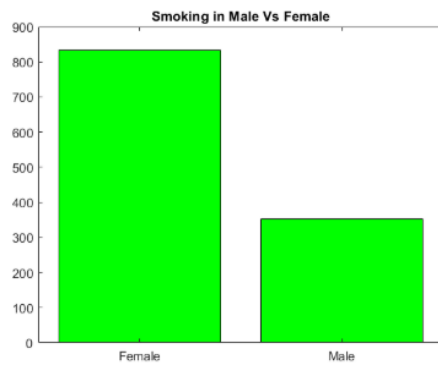


**Condition**

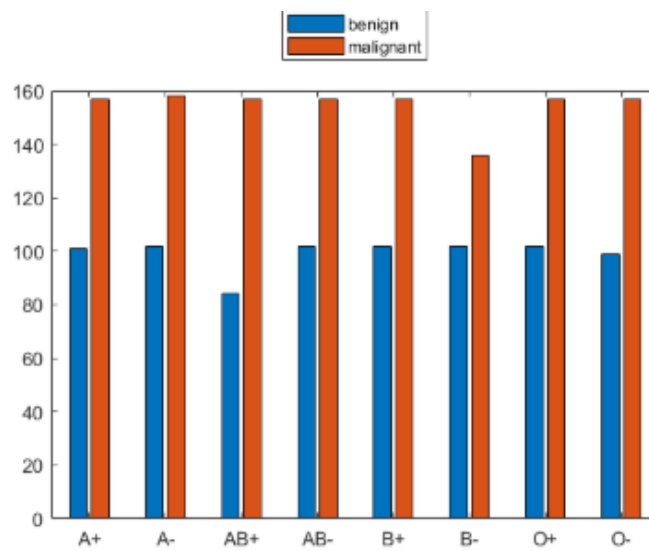


### 3. Bar graph

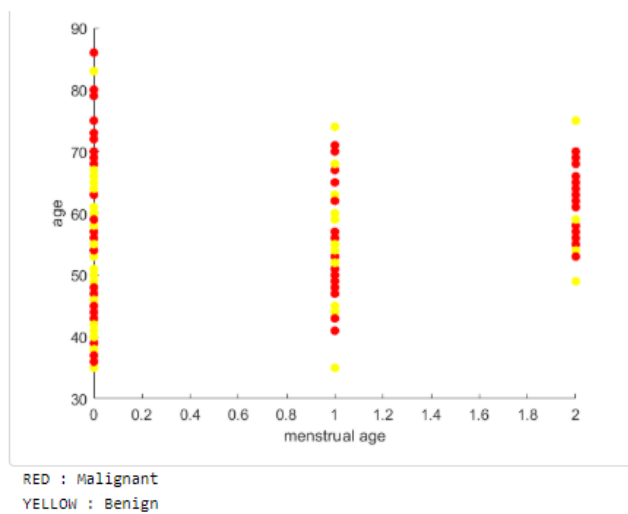




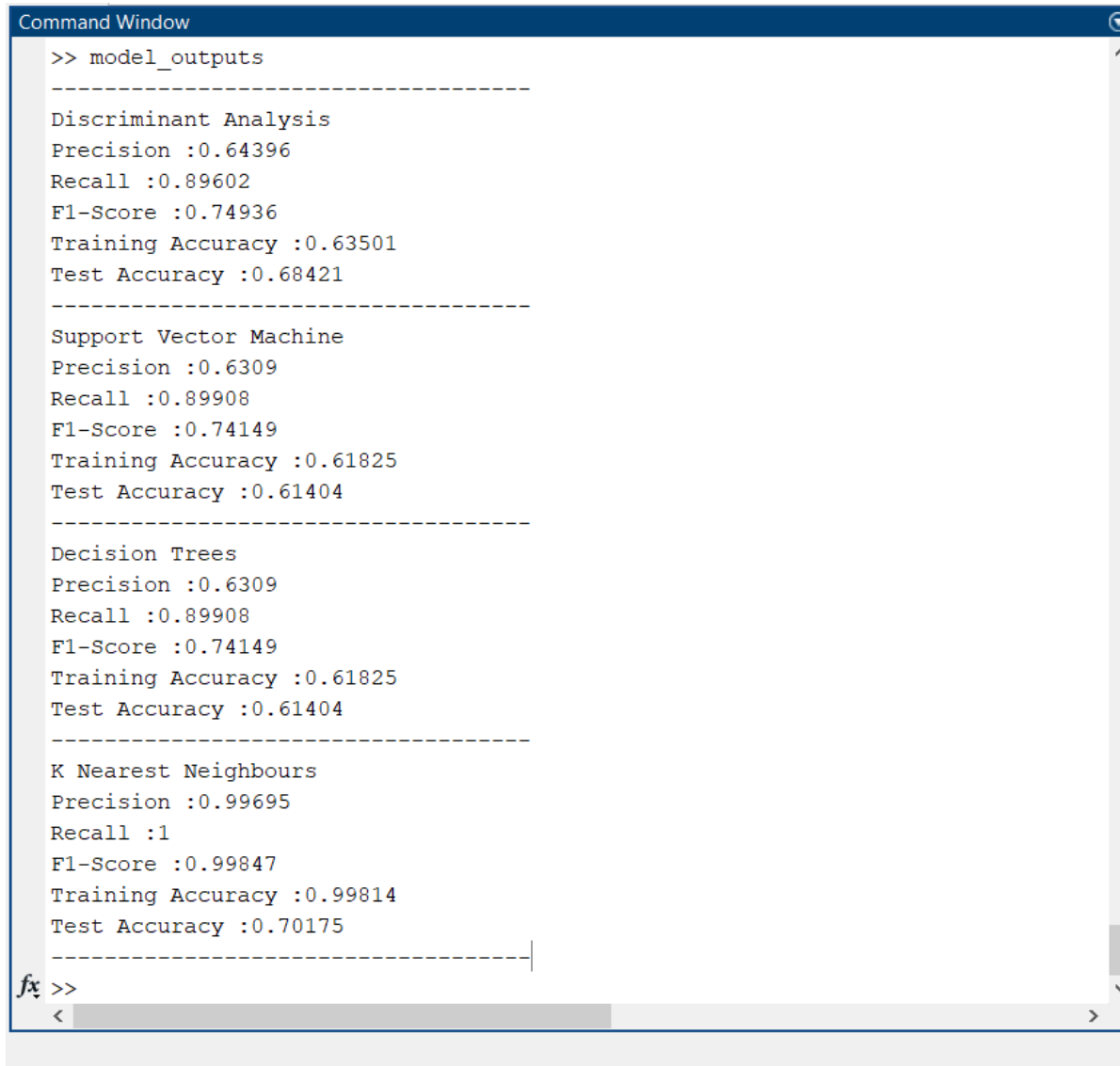
#### 4. Grouped bar graph



#### 5. Scatter plot



## Machine learning:



```
>> model_outputs

-----

Discriminant Analysis
Precision :0.64396
Recall :0.89602
F1-Score :0.74936
Training Accuracy :0.63501
Test Accuracy :0.68421
-----

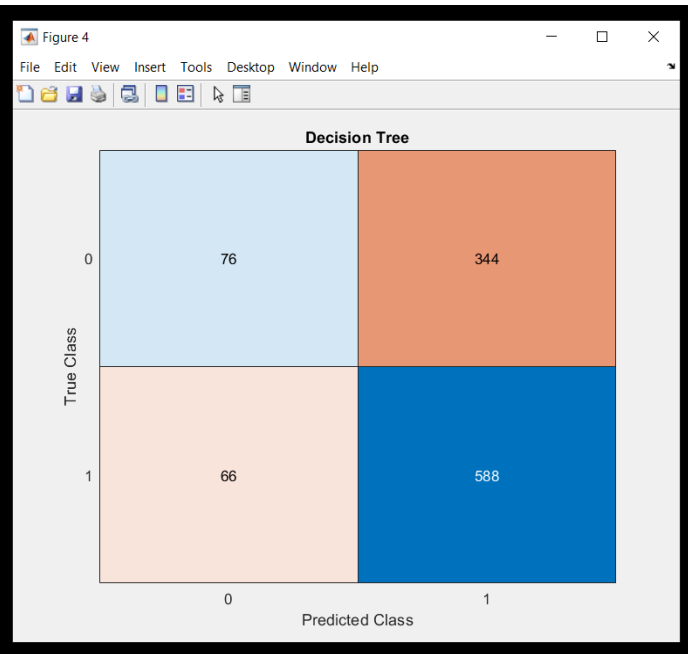
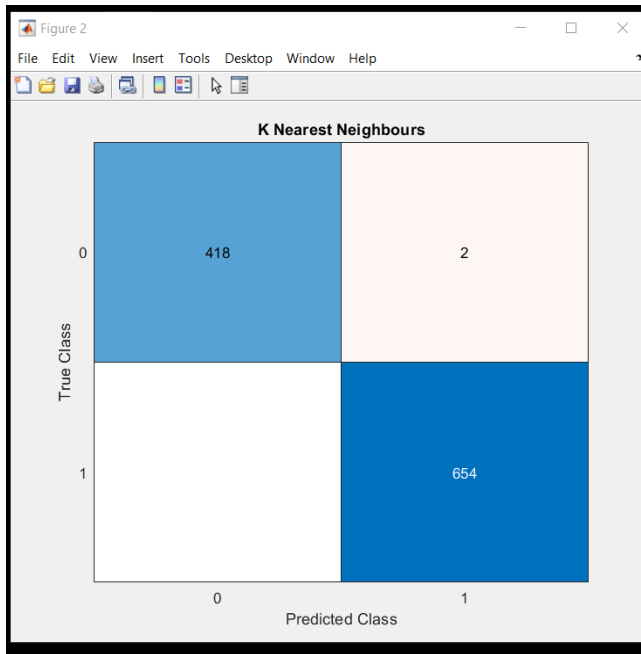
Support Vector Machine
Precision :0.6309
Recall :0.89908
F1-Score :0.74149
Training Accuracy :0.61825
Test Accuracy :0.61404
-----

Decision Trees
Precision :0.6309
Recall :0.89908
F1-Score :0.74149
Training Accuracy :0.61825
Test Accuracy :0.61404
-----

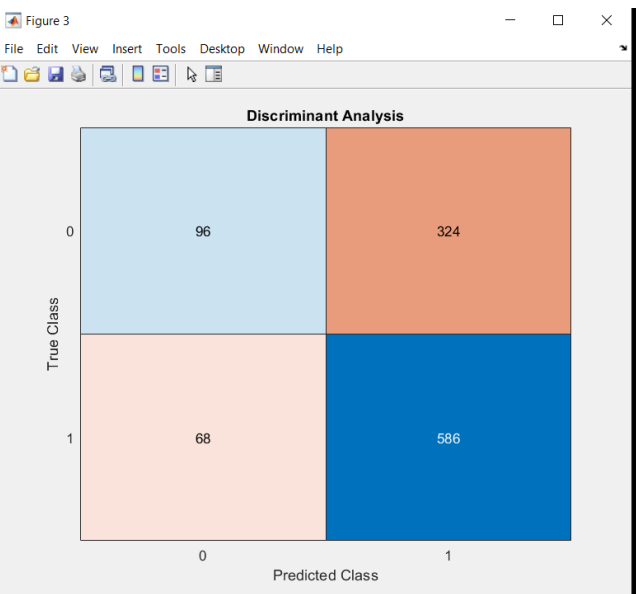
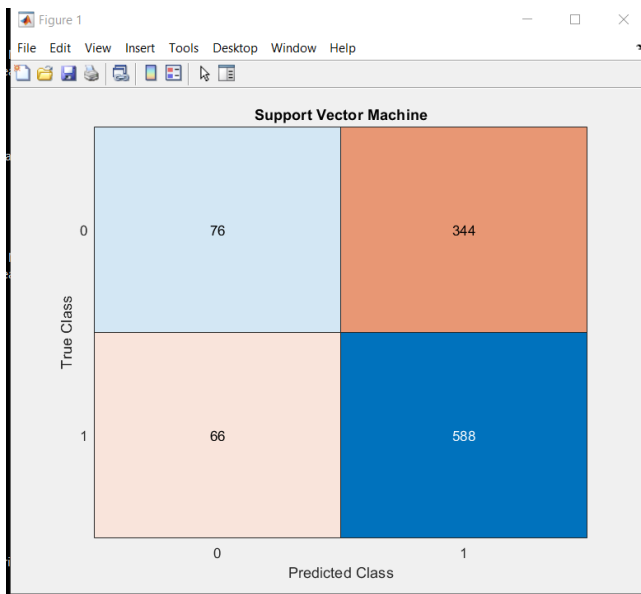
K Nearest Neighbours
Precision :0.99695
Recall :1
F1-Score :0.99847
Training Accuracy :0.99814
Test Accuracy :0.70175
-----

fx >>
```

Metrics of all the four classifier models



Confusion charts for KNN and Decision Tree



Confusion charts for SVM and Discriminant Analysis

## **CONCLUSIONS:**

The K Nearest Neighbours model turns out to be the best out of the four models, with a test accuracy of around 70 percent and recall of 1. There is usually a tradeoff between precision and recall, based on the problem. The metric used to evaluate these models is recall, because we need to take into consideration the number of false negatives that are present here. The model also gives an F1 score of 0.99 which is close to 1.