



**UNIVERSITY OF TURKISH AERONAUTICAL
ASSOCIATION
ENGINEERING FACULTY
DEPARTMENT OF COMPUTER ENGINEERING**

**CENG 463, Data Analysis Project Fall
Ecoli Bacteria Data Set**

Project members:

Buğra Küçük 170444023

Safa Özcan 170444004

Anıl Erbaşı 180444072

Project Advisor:

ASSIST. PROF. DR. SHADI AL SHEHABI

Dataset Description

Ecoli Bacteria Data Set cover the 9 type of bacteria and it is title known as Protein Localization Sites it. Creator of this Dataset is “Kenta Nakai” from Institue of Molecular and Cellular Biology Osaka, University.

Number Of Instances and Attirubutes

Number of Instances: 336 for the E.coli dataset

Number of Attributes for Ecoli dataset: 8 (7 predictive, 1 name)

Attribute Information

Sequence Name: Accession number for the SWISS-PROT database (We did not use this column in our Project because it is not a necessary anywhere in project so i deleted it).

mcg: McGeoch's method for signal sequence recognition.

gvh: von Heijne's method for signal sequence recognition.

lip: von Heijne's Signal Peptidase II consensus sequence score. Binary attribute.

chg: Presence of charge on N-terminus of predicted lipoproteins. Binary attribute.

aac: score of discriminant analysis of the amino acid content of outer membrane and periplasmic proteins.

alm1: score of the ALOM membrane spanning region prediction program.

alm2: score of ALOM program after excluding putative cleavable signal regions from the sequence.

Class: Name of the bacterias.

We imported attributes as “Double” variable type except the Class attribute it is “Categorical”.

There are no missing Attribute of this Dataset.

Class Distrubition

cp (cytoplasm)	143
im (inner membrane without signal sequence)	77
pp (perisplasm)	52
imU (inner membrane, uncleavable signal sequence)	35
om (outer membrane)	20
omL (outer membrane lipoprotein)	5
imL (inner membrane lipoprotein)	2
imS (inner membrane, cleavable signal sequence)	2

Classes Numeric Values

You can see this chart on the code too but i still wanted to add it here. As you can see there are numeric values added to each Class to use the dataset more efficent on the code.

There are two datasets for this project because there were no other options to importdata with categorical values so we changed class names to numeric values.

```
%*****Classların numeric karşılığı*****  
%   cp   (cytoplasm)                                1  
%   im   (inner membrane without signal sequence)    2  
%   pp   (perisplasm)                                3  
%   imU  (inner membrane, uncleavable signal sequence) 4  
%   om   (outer membrane)                            5  
%   omL  (outer membrane lipoprotein)                6  
%   imL  (inner membrane lipoprotein)                7  
%   imS  (inner membrane, cleavable signal sequence)  8  
%*****Classların numeric karşılığı*****
```

MATLAB Code

In this part there will be explanation for the each part of the code. Although there are already explanations inside the code we wanted to explain on here again.

Also we added code and the dataset Zip file.

```
clc;          (This part clears the code on command window deletes old data's and closes
clear;          Old figures running on matlab)
close all;
%*****Classların numeric karşılığı*****
%   cp   (cytoplasm)                                1
%   im   (inner membrane without signal sequence)    2
%   pp   (perisplasm)                                3
%   imU  (inner membrane, uncleavable signal sequence) 4
%   om   (outer membrane)                            5
%   omL  (outer membrane lipoprotein)                6
%   imL  (inner membrane lipoprotein)                7
%   imS  (inner membrane, cleavable signal sequence)  8
%*****Classların numeric karşılığı*****

%Reads the chosen dataset
load('ecoli.data')          Loads the ecoli.data set
fprintf('**What you want to do?**');
fprintf(2, '\nFor the operations(1)\nFind centered data
matrix(2)\nFor PCA algorithm(3)\nSample data to find specific
category(4)\n');
todo=input('Choose a thing to do: ');    This part hold todo attribute
                                          to let user choose the desired
                                          Attribute column

switch todo
    case 1

%Chose the attribute from dataset
fprintf('**Attributes**');
fprintf(2, '\nmcg(1)\ngvh(2)\nlip(3)\nchg(4)\naac(5)\nalm1(6)\nalm2(
7)\nClasses(8)\n');
Attribute=input('Choose an attribute: ');    This part hold Attribute
                                              Let user chose what
                                              operation they want to use

fprintf('**Operations**');
fprintf(2, '\nmean(1)\nmedian(2)\nsum(3)\nmax(4)\nrange(5)\nskewness
(6)\nkurtosis(7)\nboxplot(8)\noutliers(9)\nNothing(0)\n');
Operation=input('Choose an operation: ');
    switch Operation
        case 1
            Mean=mean(ecoli(:,Attribute))    This is the first operation
                                              and it fins the desired operations
        case 2
            Median=median(ecoli(:,Attribute))
        case 3
            Sum=sum(ecoli(:,Attribute))
```

```

case 4
Max=max(ecoli(:,Attribute))
Min=min(ecoli(:,Attribute))
case 5
Range=range(ecoli(:,Attribute))
case 6
Skewness=skewness(ecoli(:,Attribute))
if skewness(ecoli(:,Attribute))>0
    disp('curve to right')
elseif skewness(ecoli(:,Attribute))<0
    disp('curve to left')
else disp('curve is symetric')
end
case 7
Kurtosis=kurtosis(ecoli(:,Attribute))
if kurtosis(ecoli(:,Attribute))
    disp('curve to platykurtic')
elseif kurtosis(ecoli(:,Attribute))>3
    disp('curve to leptokurtic')
else disp('curve is mesokurtic')
end
case 8
boxplot(ecoli(:,Attribute))
case 9
K=ecoli(:,Attribute);
LL=quantile(K,0.25) -1.5*iqr(K);
UL=quantile(K,0.75) +1.5*iqr(K);
count=0;
for i=1:length(K)
    if(K(i)>UL || K(i)<LL)
        count=count+1;
    end
end
outliers=count

otherwise
    disp('You did not choose anything');
end

```

case 2 Second Operation to find Centered Data Matrix

Blue is for the old data set and **Red** is the Centered Data matrix

```

%Finds the centered data matrix
load('ecoli.data')
ecoli(:,8)=[]; %we remove the class because we are trying to
find centered data matrix
meanecoli=mean(ecoli); %we get mean of the data set
ecolinorm=ecoli-meanecoli; %then we subtract from the data
newmean=mean(ecolinorm); %again we need to take mean of the
new dataset called ecolinorm
scatter(ecoli,meanecoli,'blue');
hold on

```

```

title('Red is scatter of centered data matrix', 'Blue is old
data matrix');
scatter(ecolinorm,newmean, 'red');
hold off

```

case 3

Operation 3 find the pca of the dataset

```

X=ecoli(:, [1,2,3,4,5,6,7]); %inputlarımız
y=ecoli(:, [8]); %outputlarımız(classlar)
m=2;
[U, Z] = pca(X, 'NumComponents',m); %pca uygulanıyor
figure;
hold on
scatter(Z(:,1), Z(:,2), 25, y, 'filled'); %grafik
hold off
colormap(jet); %farklı renklerle daha kolay görmek için

```

case 4

Operation 4 takes some input from you then gives Your the predicted category

This part is generated automatically from matlab to import data i used so much effort to import it normally with “load” operation but it did not work since class attribute had categorical value but i found this solution to upload it as table to the matlab from “.data” extension.

```

% Set up the Import Options and import the data
opts = delimitedTextImportOptions("NumVariables", 8);
% Specify range and delimiter
opts.DataLines = [1, Inf];
opts.Delimiter = ",";
% Specify column names and types
opts.VariableNames = ["mcg", "gvh", "lip", "chg", "aac",
"alm1", "alm2", "Class"];
opts.VariableTypes = ["double", "double", "double", "double",
"double", "double", "double", "categorical"];
% Specify file level properties
opts.ExtraColumnsRule = "ignore";
opts.EmptyLineRule = "read";
% Specify variable properties
opts = setvaropts(opts, "Class", "EmptyFieldRule", "auto");
% Import the data
ecoliKNN = readtable("ecoliKNN.data", opts);
% Clear temporary variables
clear opts

```

Here on this part we used KNN search algorithm to find the predicted class

```

%Predict KNN
modelformed =
fitcknn(ecoliKNN, 'Class~mcg+gvh+lip+chg+aac+alm1+alm2');
modelformed.NumNeighbors=3;
mcg=input('Enter mcg: ');
gvh=input('Enter gvh: ');
lip=input('Enter lip: ');

```

```

chg=input('Enter chg: ');
aac=input('Enter aac: ');
alm1=input('Enter alm1: ');
alm2=input('Enter alm2: ');
predict(modelformed,[mcg,gvh,lip,chg,aac,alm1,alm2])
end

```

Execution Of Code

We will show the running code part by part.

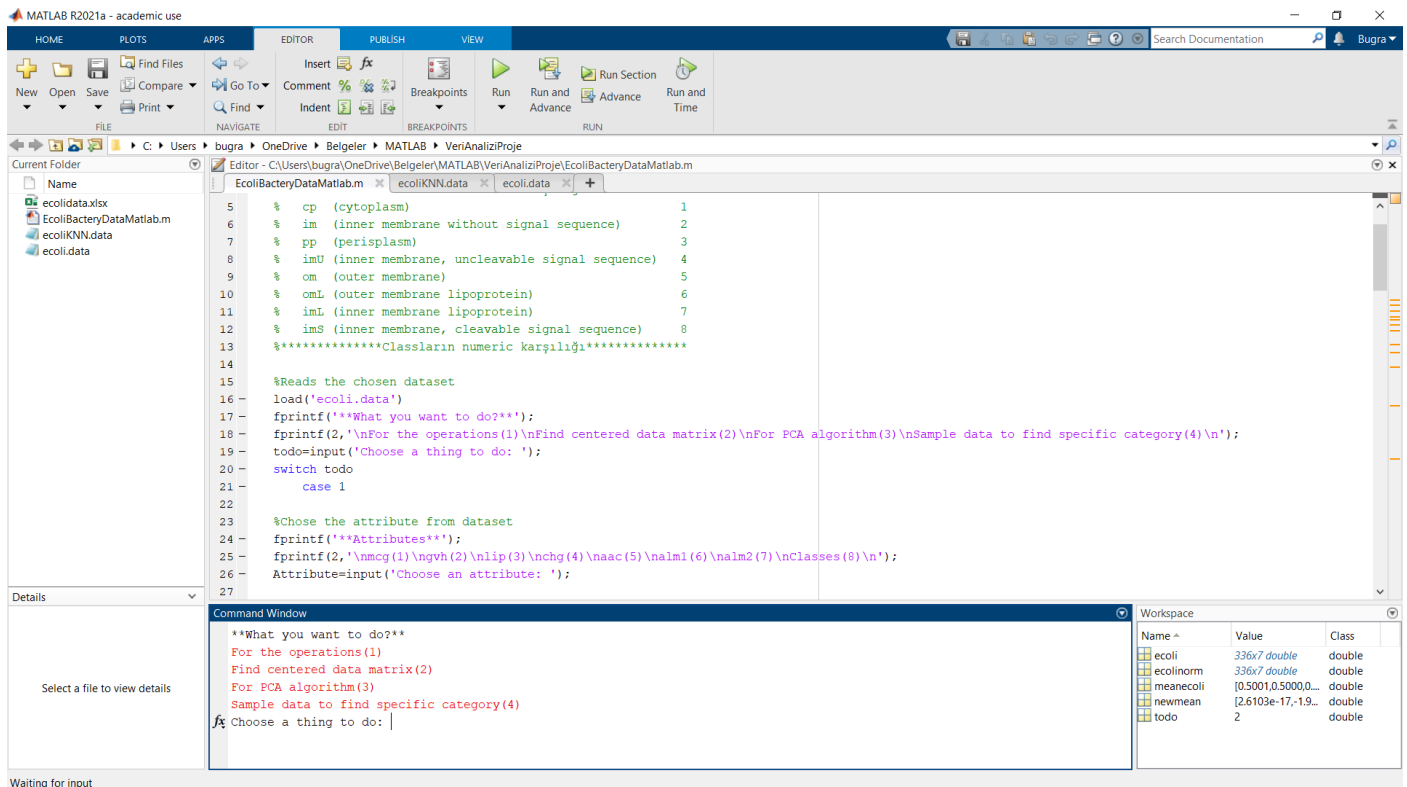


Figure 1: It shows the first thing after pressing “Run” button.

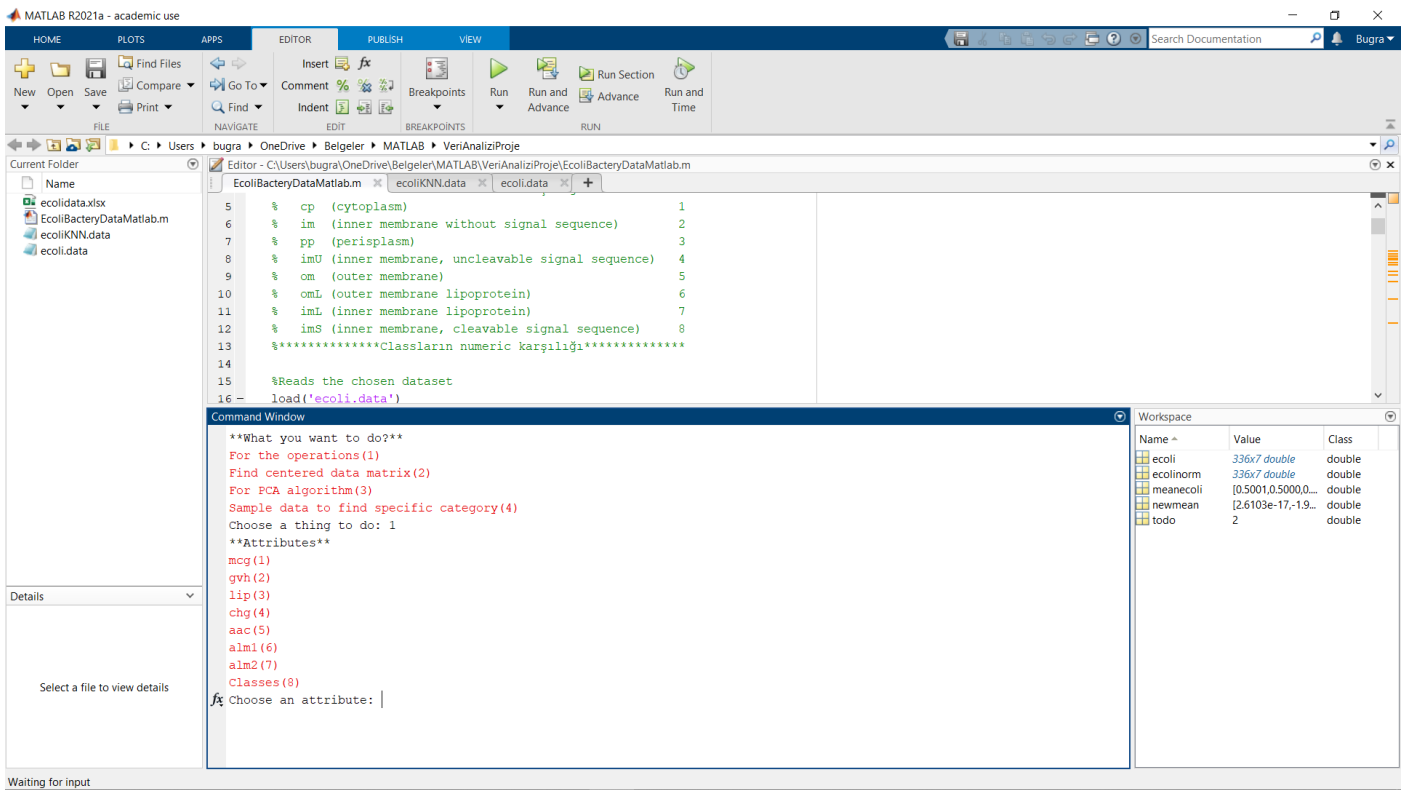


Figure 2: After selecting first operation code wants us to choose the attribute that we want to work on

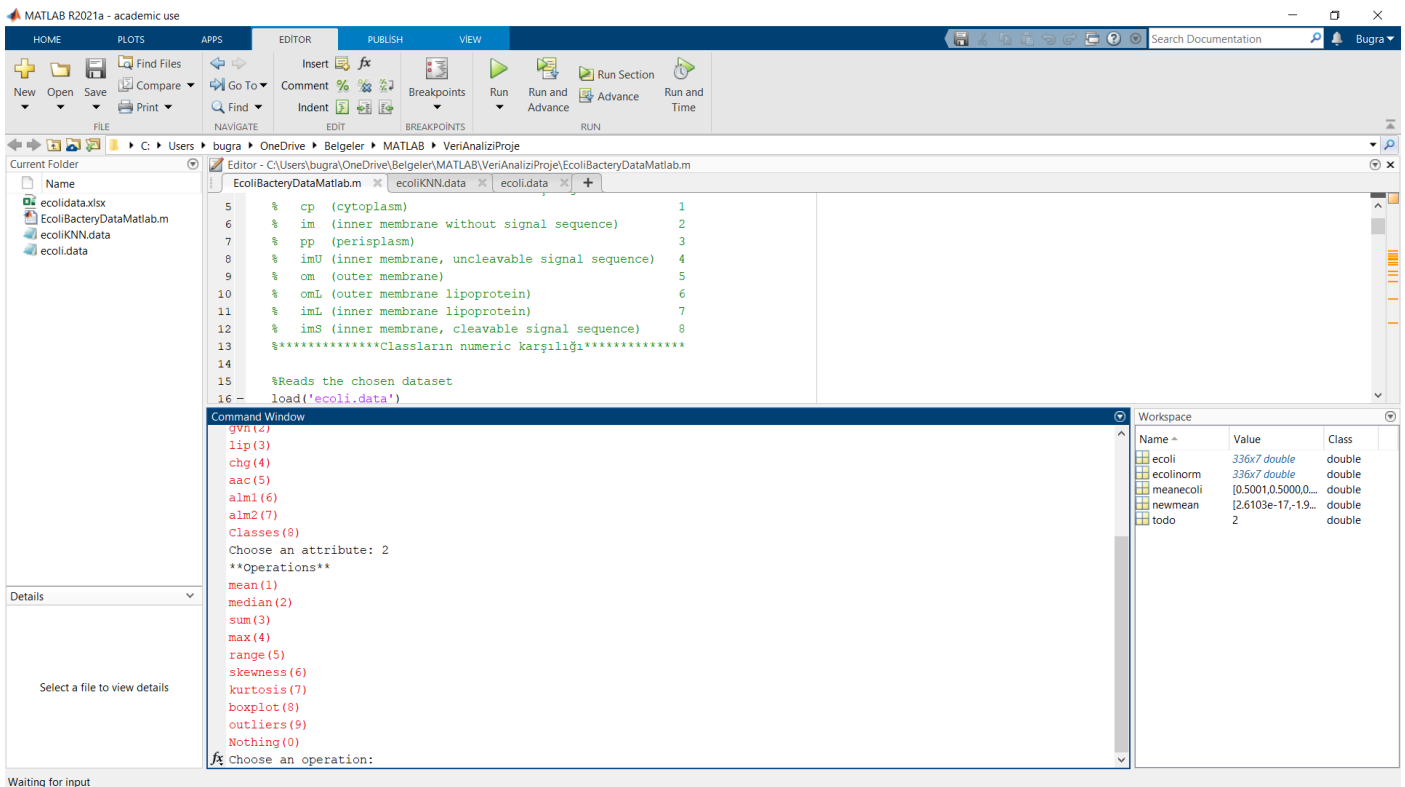


Figure 3: After we selected gvh it shows us the operation part to let us choose the operation that we want to work on with.

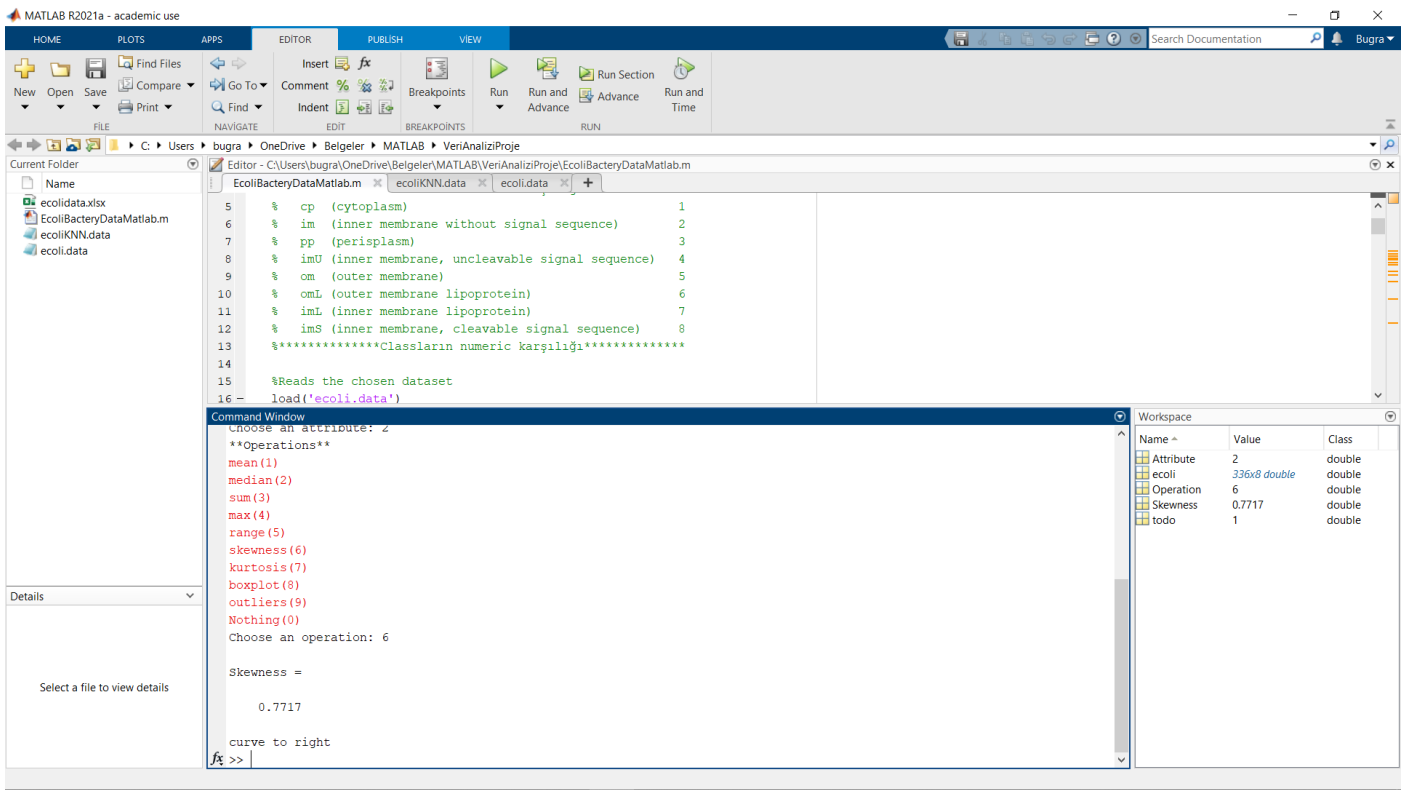


Figure 4: After selecting skewness(6) it shows us the skewness of the gvh and shows the interpret of the skewness.

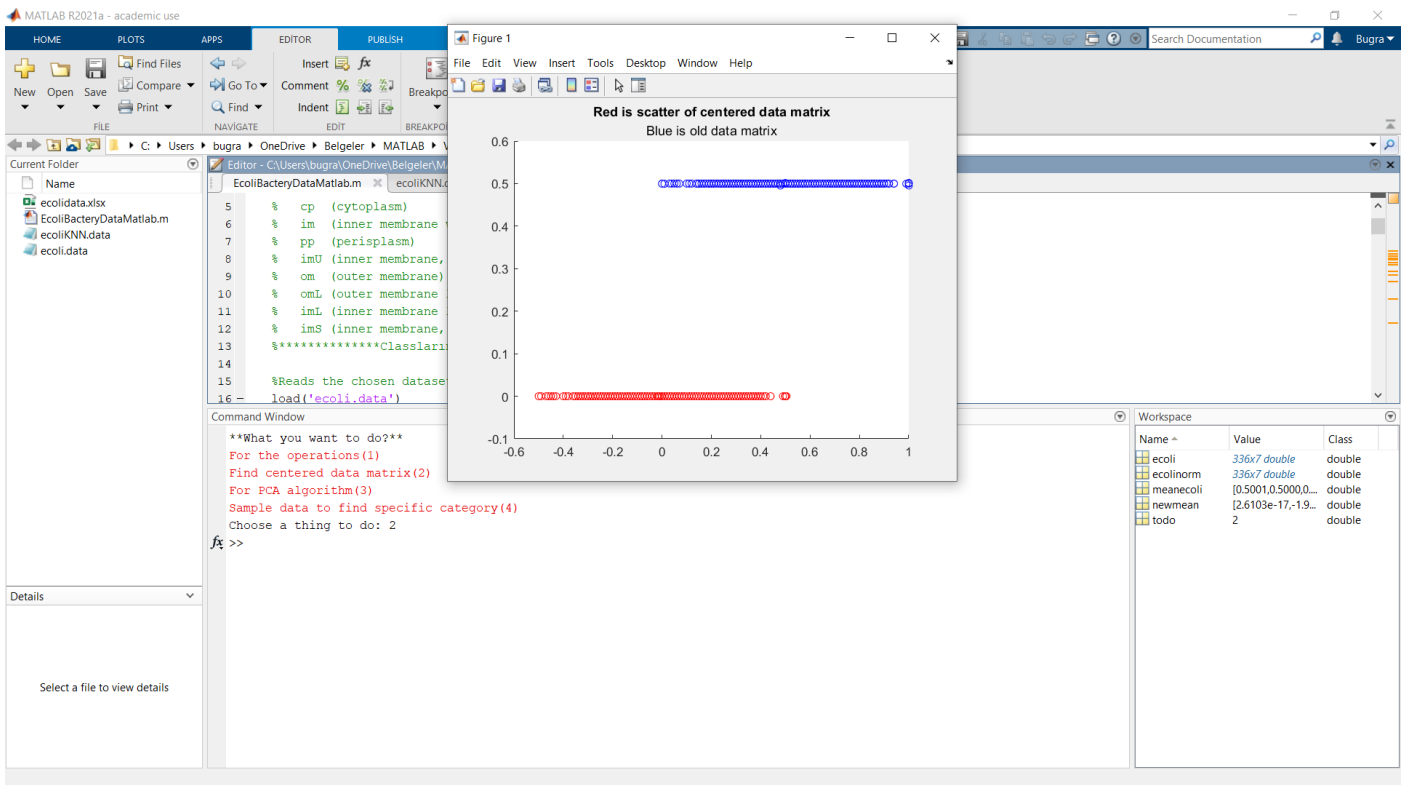


Figure 5: Our second operation shows the centered data matrix. Red is the new centered data matrix and blue is the old matrix also it looks like a little bit broken but if you get closer and check the values of the dots you can see they are different but the reason behind of this linear dots is because of the dataset.

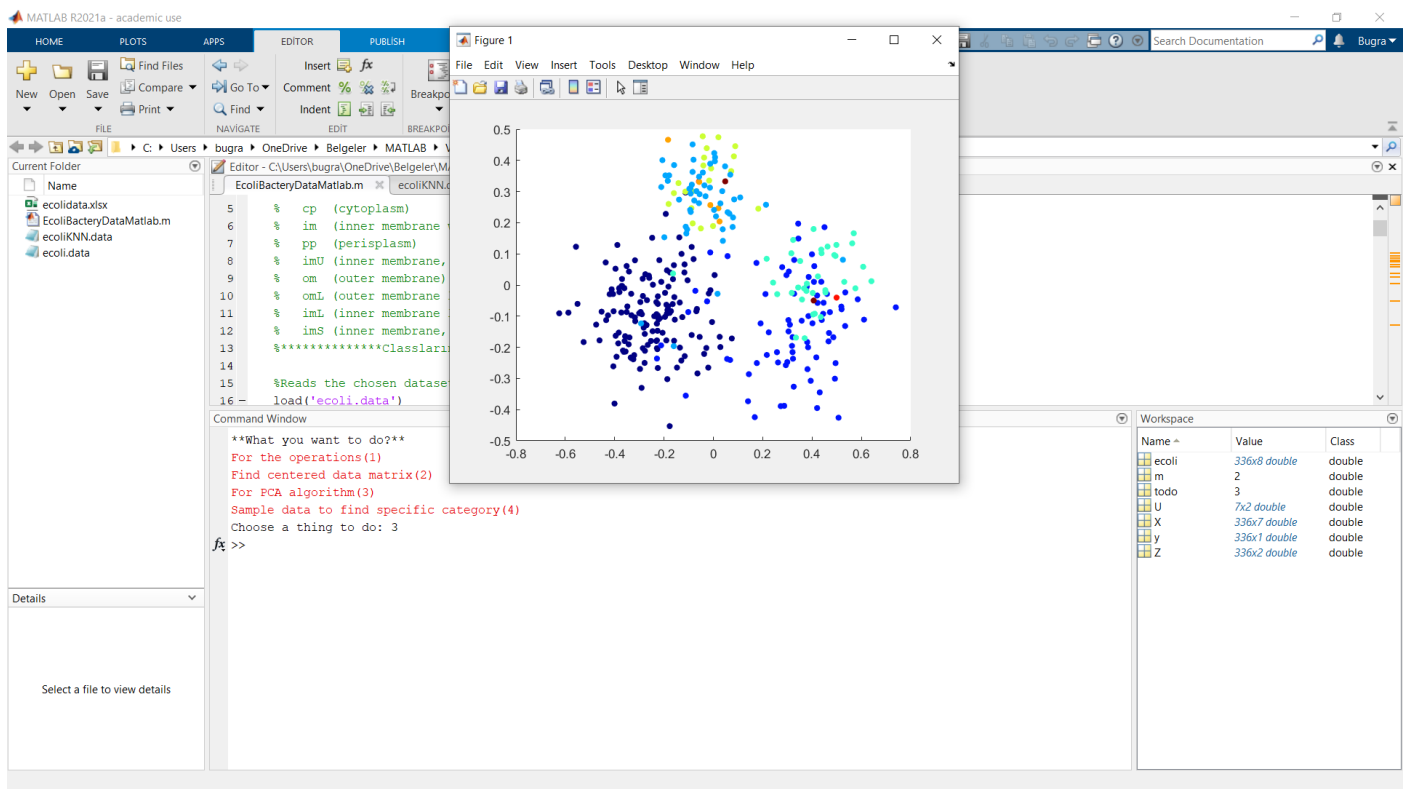


Figure 6: This our Third operation for PCA algorithm if you get close you can see there are colors for the every Class.

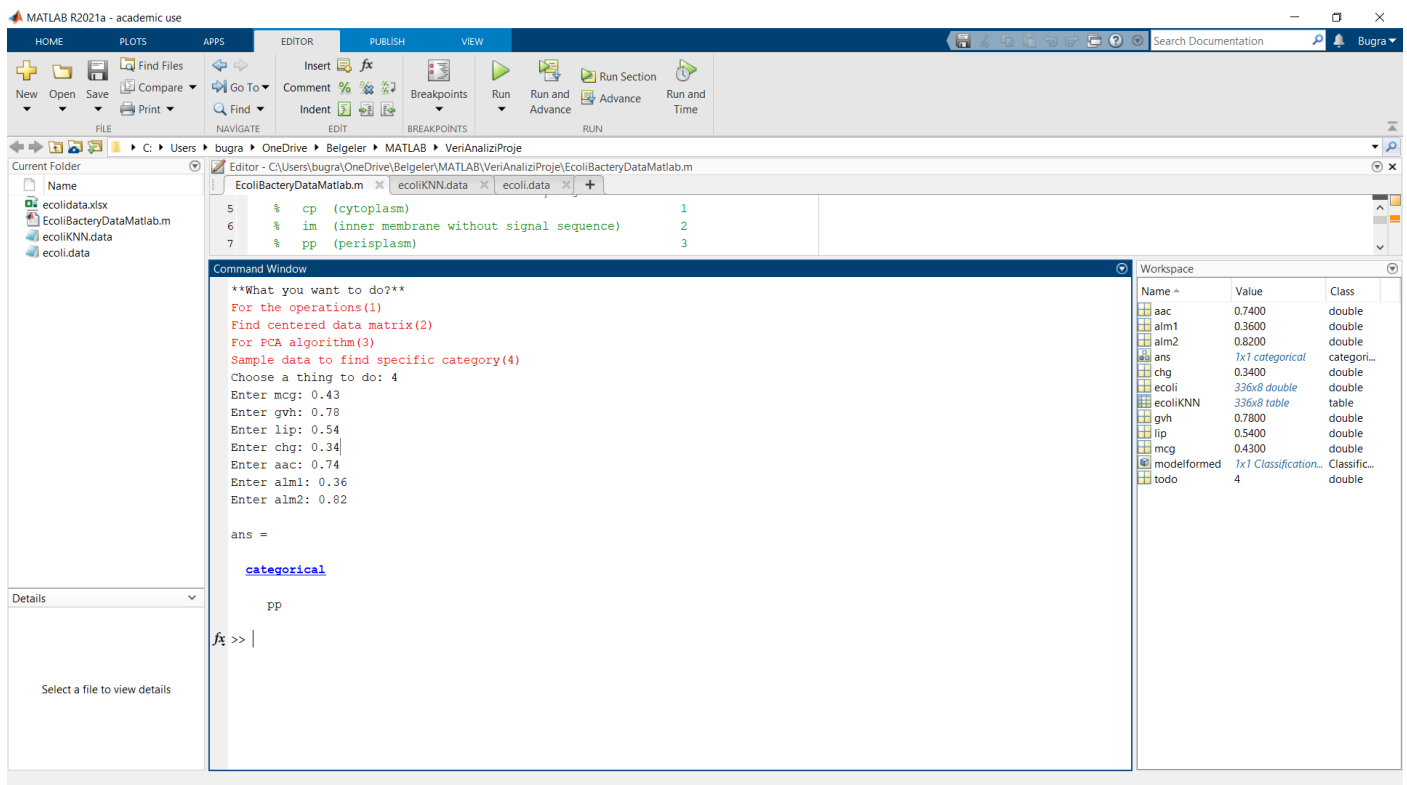
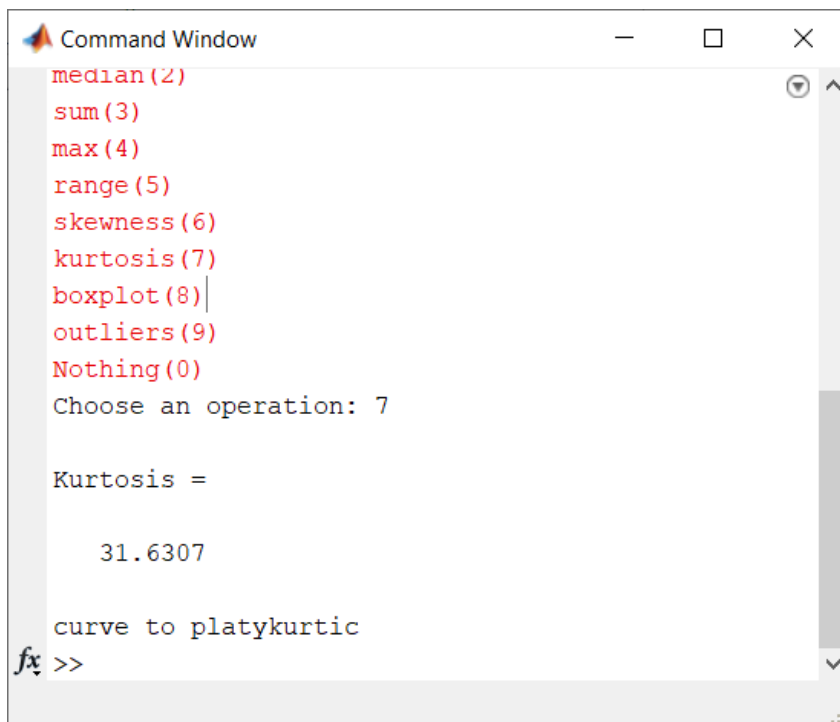


Figure 7: This is our last operation and when you insert numbers for each attribute you get predicted category of the inserted attributes. Also it uses the KNN search algorithm to find the predicted Category.

Interpretation of the skewness and kurtosis values according to the selected attribute.

As you can see on Figure 3 we find the interpretation of skewness and kurtosis values inside **Operation 1**. Anyway we will put some examples of the different combinations.

Kurtosis Of Lip attribute



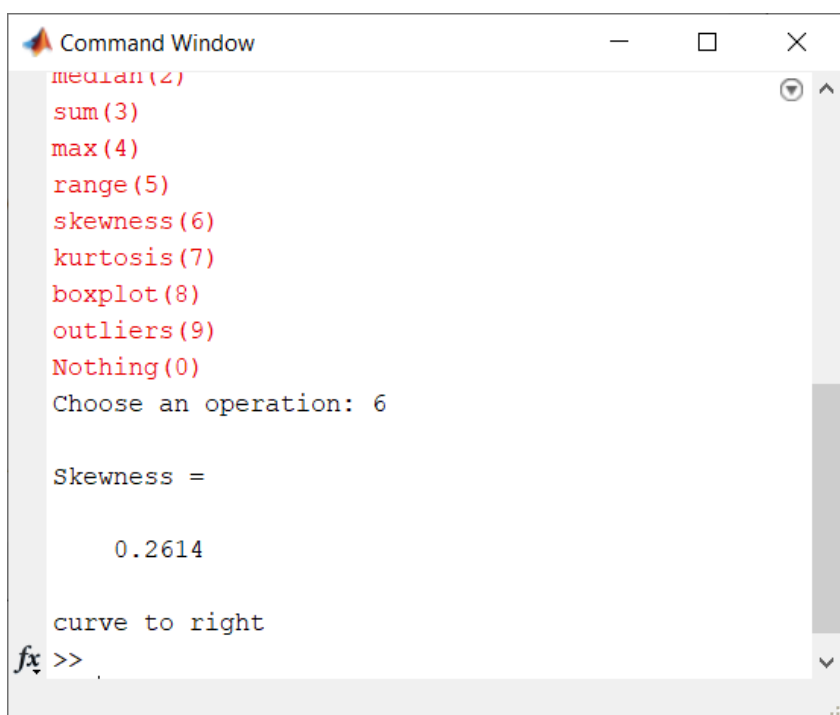
```
Command Window
median(2)
sum(3)
max(4)
range(5)
skewness(6)
kurtosis(7)
boxplot(8)
outliers(9)
Nothing(0)
Choose an operation: 7

Kurtosis =

    31.6307

curve to platykurtic
fx >>
```

Skewness of the alm1 attribute



```
Command Window
median(2)
sum(3)
max(4)
range(5)
skewness(6)
kurtosis(7)
boxplot(8)
outliers(9)
Nothing(0)
Choose an operation: 6

Skewness =

    0.2614

curve to right
fx >>
```

Skewness of aac attribute



```
Command Window
median(2)
sum(3)
max(4)
range(5)
skewness(6)
kurtosis(7)
boxplot(8)
outliers(9)
Nothing(0)
Choose an operation: 7

Kurtosis =

    4.2980

curve to platykurtic
fx >>
```

Kurtosis of the mcg attribute



```
Command Window
median(2)
sum(3)
max(4)
range(5)
skewness(6)
kurtosis(7)
boxplot(8)
outliers(9)
Nothing(0)
Choose an operation: 7

Kurtosis =

    2.1357

curve to platykurtic
fx >>
```