LOGISTIC REGRESSION & VISUALIZATION OF DATA (by Dilhara Liyanaaratchi)

Summary of the Data Set

```
clear all
clc

tic
T = readtable("Invistico_Airline.csv"); % Importing the data as a table
```

```
Warning: Column headers from the file were modified to make them valid MATLAB identifiers before creating
variable names for the table. The original column headers are saved in the VariableDescriptions property.
Set 'VariableNamingRule' to 'preserve' to use the original column headers as table variable names.
```

```
summary(T);
```

```
Variables:

    satisfaction: 129880×1 cell array of character vectors

        Properties:
            Description:  satisfaction
    Gender: 129880×1 cell array of character vectors

        Properties:
            Description:  Gender
    CustomerType: 129880×1 cell array of character vectors

        Properties:
            Description:  Customer Type
    Age: 129880×1 double

        Properties:
            Description:  Age
        Values:

            Min           7
            Median       40
            Max          85

    TypeOfTravel: 129880×1 cell array of character vectors

        Properties:
            Description:  Type of Travel
    Class: 129880×1 cell array of character vectors

        Properties:
            Description:  Class
    FlightDistance: 129880×1 double

        Properties:
            Description:  Flight Distance
        Values:

            Min           50
            Median      1925
            Max         6951

    SeatComfort: 129880×1 double

        Properties:
```

```
        Description:  Seat comfort
    Values:

        Min          0
        Median       3
        Max          5

Departure_ArrivalTimeConvenient: 129880×1 double

    Properties:
        Description:  Departure/Arrival time convenient
    Values:

        Min          0
        Median       3
        Max          5

FoodAndDrink: 129880×1 double

    Properties:
        Description:  Food and drink
    Values:

        Min          0
        Median       3
        Max          5

GateLocation: 129880×1 double

    Properties:
        Description:  Gate location
    Values:

        Min          0
        Median       3
        Max          5

InflightWifiService: 129880×1 double

    Properties:
        Description:  Inflight wifi service
    Values:

        Min          0
        Median       3
        Max          5

InflightEntertainment: 129880×1 double

    Properties:
        Description:  Inflight entertainment
    Values:

        Min          0
        Median       4
        Max          5

OnlineSupport: 129880×1 double

    Properties:
        Description:  Online support
    Values:

        Min          0
        Median       4
        Max          5

EaseOfOnlineBooking: 129880×1 double

    Properties:
        Description:  Ease of Online booking
    Values:

        Min          0
        Median       4
        Max          5

On_boardService: 129880×1 double

    Properties:
        Description:  On-board service
    Values:

        Min          0
        Median       4
        Max          5

LegRoomService: 129880×1 double

    Properties:
        Description:  Leg room service
    Values:

        Min          0
        Median       4
        Max          5

BaggageHandling: 129880×1 double

    Properties:
        Description:  Baggage handling
    Values:

        Min          1
        Median       4
        Max          5

CheckinService: 129880×1 double

    Properties:
        Description:  Checkin service
    Values:

        Min          0
        Median       3
        Max          5

Cleanliness: 129880×1 double

    Properties:
        Description:  Cleanliness
    Values:

        Min          0
        Median       4
        Max          5

OnlineBoarding: 129880×1 double
```

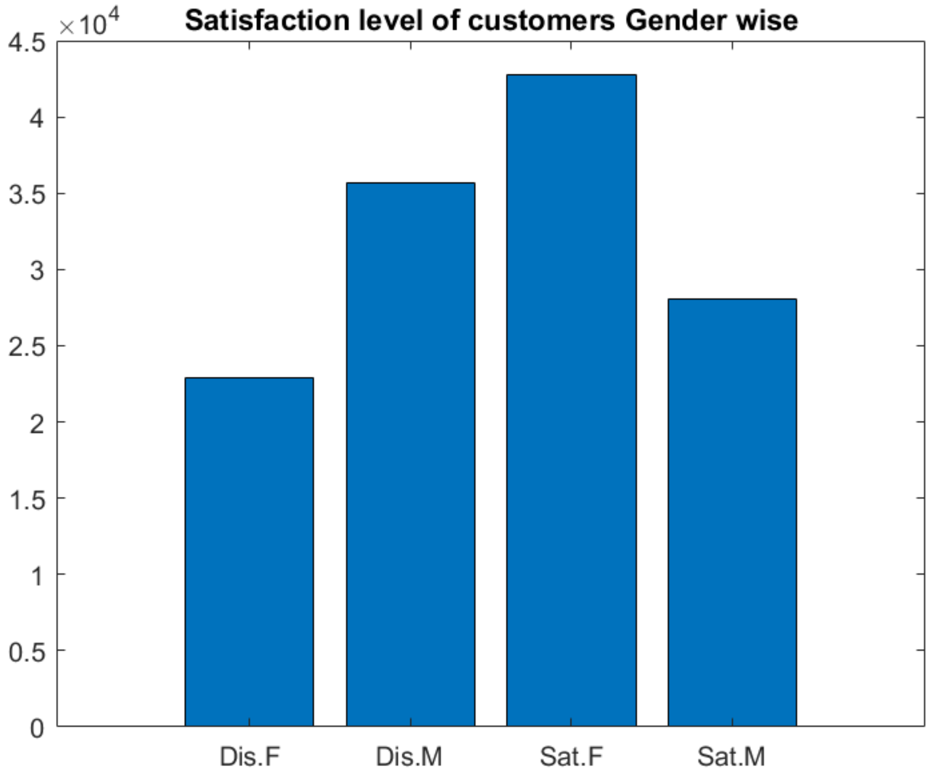
```
        Properties:
            Description:  Online boarding
        Values:

            Min          0
            Median       4
            Max          5

    DepartureDelayInMinutes: 129880×1 double

        Properties:
            Description:  Departure Delay in Minutes
        Values:

            Min            0
            Median         0
            Max         1592

    ArrivalDelayInMinutes: 129880×1 double

        Properties:
            Description:  Arrival Delay in Minutes
        Values:

            Min             0
            Median          0
            Max             1584
            NumMissing      393
```

Missing values

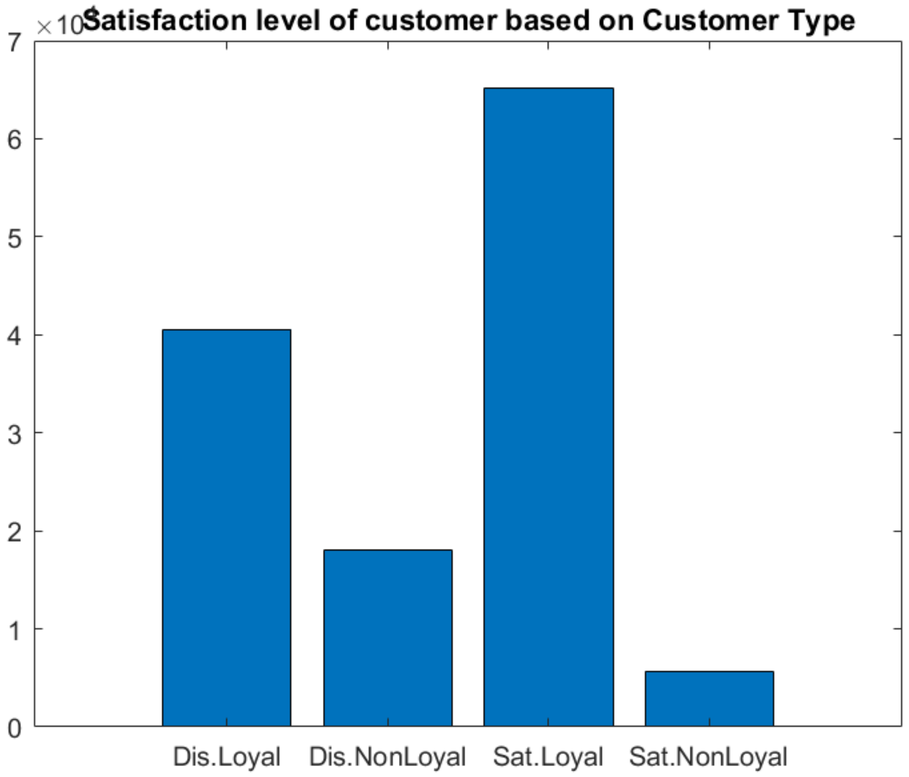
```
% As per the summary it is identified that there are 393 missing values in
% the attribute Arrival delay in minutes. And as compared it to the number
% of observations since it is a small amount we can remove the missing
% observations.

CD = rmmissing(T,'MinNumMissing',1); % removed the rows with missing data
```

Attribute summary via visualization

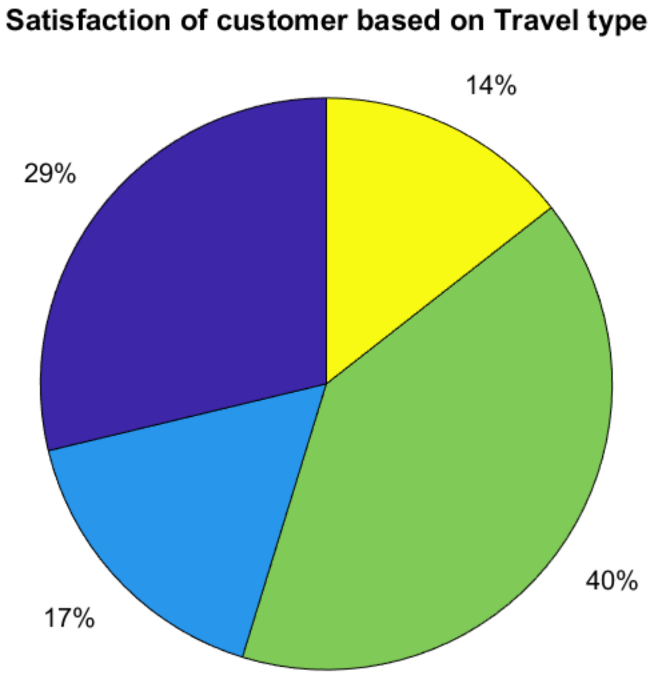
```
Gen = groupcounts(CD,{'satisfaction','Gender'});
CT = groupcounts(CD,{'satisfaction','CustomerType'});
ToT = groupcounts(CD,{'satisfaction','TypeOfTravel'});
C = groupcounts(CD,{'satisfaction','Class'});

GenL = {'Dis.F','Dis.M','Sat.F','Sat.M'};
bar(Gen.GroupCount);
xticklabels(GenL);
title('Satisfaction level of customers Gender wise');
```



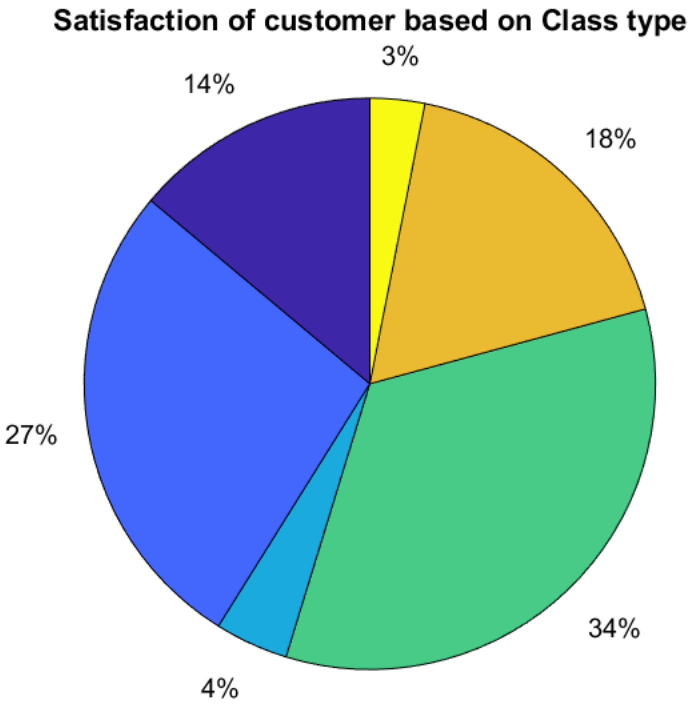
```
CTL = {'Dis.Loyal','Dis.NonLoyal','Sat.Loyal','Sat.NonLoyal'};
bar(CT.GroupCount);
xticklabels(CTL);
title('Satisfaction level of customer based on Customer Type')
```



```
ToTL = {'Dis.Bus','Dis.Per','Sat.Bus','Sat.Per'};
pie(ToT.Percent);
xticklabels(ToTL);
title('Satisfaction of customer based on Travel type');
```



```
pie(C.Percent);
title('Satisfaction of customer based on Class type');
```



Convert Categories to nominal variables

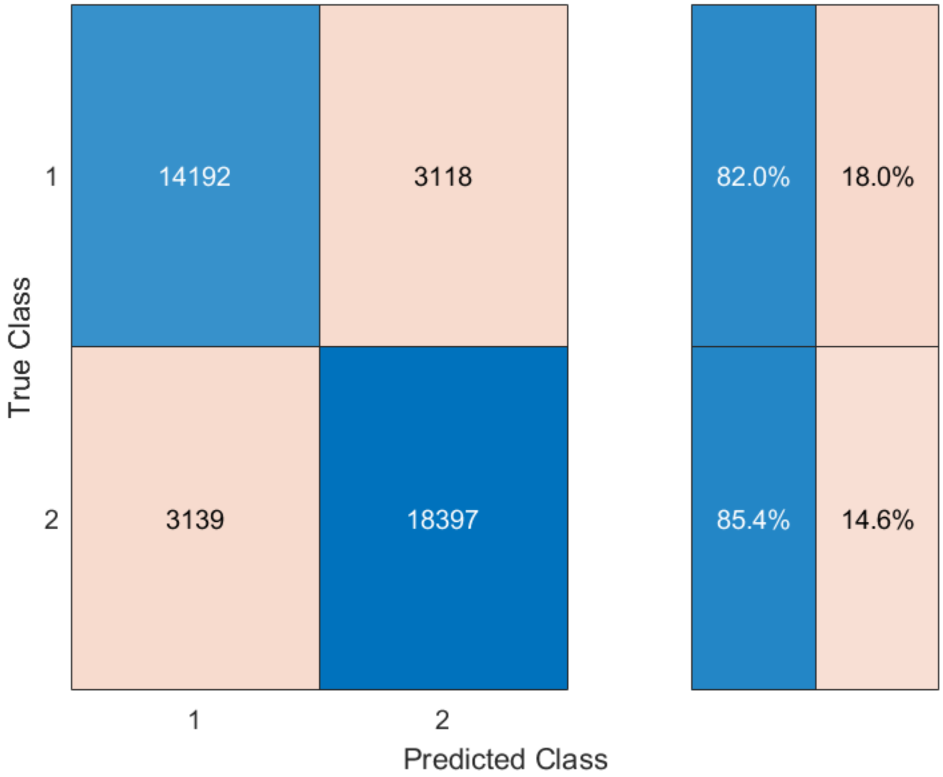
```
% Converting Gender

[grpG,genderVals] = findgroups(CD.Gender);
CD.Gender = grpG;

% Converting Satisfaction
[grpS,satisVals] = findgroups(CD.satisfaction);
CD.satisfaction = grpS;

% Converting Customer Type
[grpCT,CTVals] = findgroups(CD.CustomerType);
CD.CustomerType = grpCT;

% Converting Type of Travel
[grpToT,ToTVals] = findgroups(CD.TypeOfTravel);
CD.TypeOfTravel = grpToT;

% Converting Class
[grpC,ClassVals] = findgroups(CD.Class);
CD.Class = grpC;
```

Logistic Regression Analysis

```
% Splitting the data

Y = CD.satisfaction;
X = [CD.Gender, CD.CustomerType, CD.Age, CD.TypeOfTravel, CD.Class, CD.FlightDistance, CD.Seat


Y = double(Y)-1;

rng(1)
cv = cvpartition(length(X),'holdout',0.3);

% Training set
Xtrain = X(training(cv),:);
Ytrain = Y(training(cv),:);

% Testing set
Xtest = X(test(cv),:);
Ytest = Y(test(cv),:);

mdl_lr = fitglm(Xtrain,Ytrain,'Distribution','binomial','Link','logit');

ptest = predict(mdl_lr,Xtest);

Y_ptest = round(ptest) + 1
```

```
Y_ptest = 38846×1
     1
     1
     2
     1
     1
     1
     1
     1
     1
     1
     ⋮
```

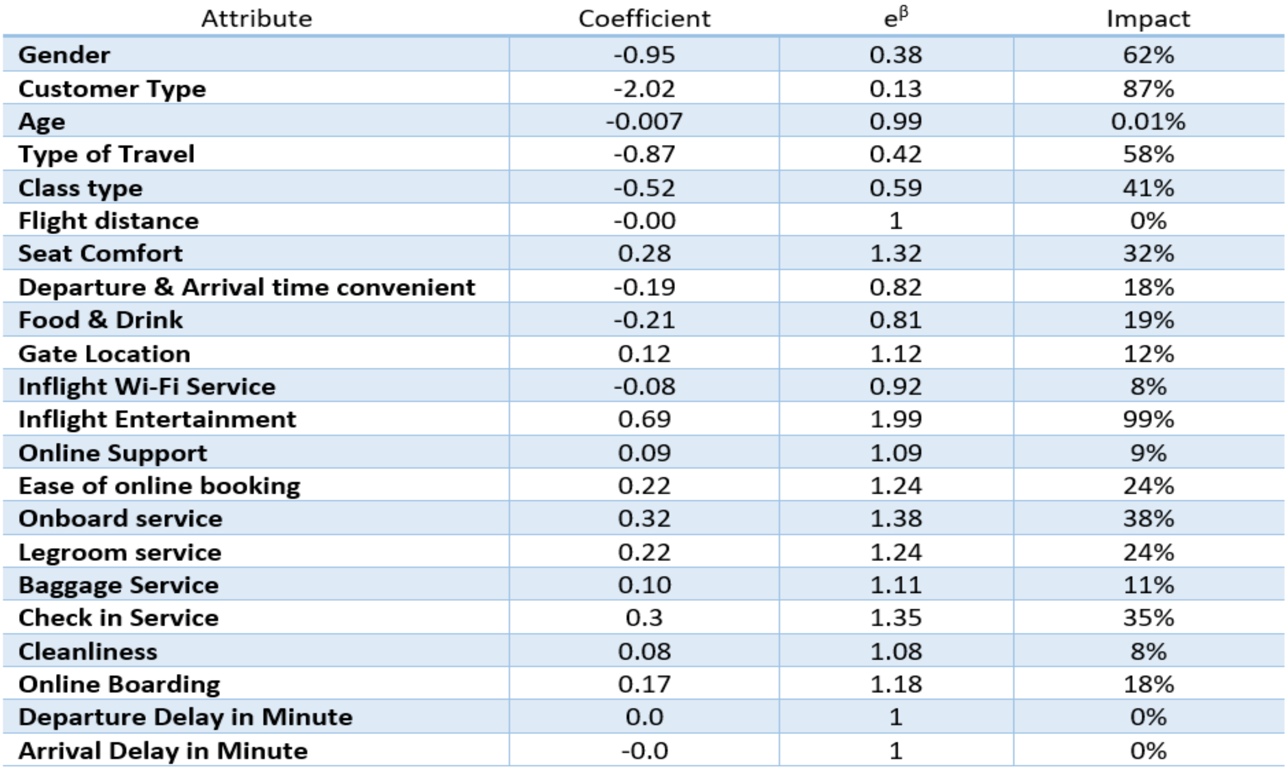
```
cmglm = confusionchart(double(Ytest)+1,Y_ptest,'RowSummary','row-normalized');
```



```
[xxtest,yytest,Tresholds,auctest] = perfcurve(Ytest,ptest,1);

toc
```

```
Elapsed time is 5.908907 seconds.
```

Calculating the Influence on each attributes



Attribute	Coefficient	e^{β}	Impact
Gender	-0.95	0.38	62%
Customer Type	-2.02	0.13	87%
Age	-0.007	0.99	0.01%
Type of Travel	-0.87	0.42	58%
Class type	-0.52	0.59	41%
Flight distance	-0.00	1	0%
Seat Comfort	0.28	1.32	32%
Departure & Arrival time convenient	-0.19	0.82	18%
Food & Drink	-0.21	0.81	19%
Gate Location	0.12	1.12	12%
Inflight Wi-Fi Service	-0.08	0.92	8%
Inflight Entertainment	0.69	1.99	99%
Online Support	0.09	1.09	9%
Ease of online booking	0.22	1.24	24%
Onboard service	0.32	1.38	38%
Legroom service	0.22	1.24	24%
Baggage Service	0.10	1.11	11%
Check in Service	0.3	1.35	35%
Cleanliness	0.08	1.08	8%
Online Boarding	0.17	1.18	18%
Departure Delay in Minute	0.0	1	0%
Arrival Delay in Minute	-0.0	1	0%

DECISION TREE (by Shageerthana Sathiyamoorthy)

```
%Importing the data
T1 = readtable ("Invistico_Airline.csv");
```

```
Warning: Column headers from the file were modified to make them valid MATLAB identifiers before creating
variable names for the table. The original column headers are saved in the VariableDescriptions property.
Set 'VariableNamingRule' to 'preserve' to use the original column headers as table variable names.
```

```
rng(1);
summary(T1)
```

```
Variables:

    satisfaction: 129880×1 cell array of character vectors

        Properties:
            Description:  satisfaction
    Gender: 129880×1 cell array of character vectors

        Properties:
            Description:  Gender
    CustomerType: 129880×1 cell array of character vectors

        Properties:
            Description:  Customer Type
    Age: 129880×1 double

        Properties:
            Description:  Age
        Values:
```

```
        Min           7
        Median       40
        Max          85

TypeOfTravel: 129880×1 cell array of character vectors

    Properties:
        Description:  Type of Travel
Class: 129880×1 cell array of character vectors

    Properties:
        Description:  Class
FlightDistance: 129880×1 double

    Properties:
        Description:  Flight Distance
    Values:

        Min           50
        Median      1925
        Max         6951

SeatComfort: 129880×1 double

    Properties:
        Description:  Seat comfort
    Values:

        Min          0
        Median       3
        Max          5

Departure_ArrivalTimeConvenient: 129880×1 double

    Properties:
        Description:  Departure/Arrival time convenient
    Values:

        Min          0
        Median       3
        Max          5

FoodAndDrink: 129880×1 double

    Properties:
        Description:  Food and drink
    Values:

        Min          0
        Median       3
        Max          5

GateLocation: 129880×1 double

    Properties:
        Description:  Gate location
    Values:

        Min          0
        Median       3
        Max          5

InflightWifiService: 129880×1 double
```

```
    Properties:
        Description:  Inflight wifi service
    Values:

        Min          0
        Median       3
        Max          5

InflightEntertainment: 129880×1 double

    Properties:
        Description:  Inflight entertainment
    Values:

        Min          0
        Median       4
        Max          5

OnlineSupport: 129880×1 double

    Properties:
        Description:  Online support
    Values:

        Min          0
        Median       4
        Max          5

EaseOfOnlineBooking: 129880×1 double

    Properties:
        Description:  Ease of Online booking
    Values:

        Min          0
        Median       4
        Max          5

On_boardService: 129880×1 double

    Properties:
        Description:  On-board service
    Values:

        Min          0
        Median       4
        Max          5

LegRoomService: 129880×1 double

    Properties:
        Description:  Leg room service
    Values:

        Min          0
        Median       4
        Max          5

BaggageHandling: 129880×1 double

    Properties:
        Description:  Baggage handling
    Values:
            Min          1
            Median       4
            Max          5

    CheckinService: 129880×1 double

        Properties:
            Description:  Checkin service
        Values:

            Min          0
            Median       3
            Max          5

    Cleanliness: 129880×1 double

        Properties:
            Description:  Cleanliness
        Values:

            Min          0
            Median       4
            Max          5

    OnlineBoarding: 129880×1 double

        Properties:
            Description:  Online boarding
        Values:

            Min          0
            Median       4
            Max          5

    DepartureDelayInMinutes: 129880×1 double

        Properties:
            Description:  Departure Delay in Minutes
        Values:

            Min            0
            Median         0
            Max         1592

    ArrivalDelayInMinutes: 129880×1 double

        Properties:
            Description:  Arrival Delay in Minutes
        Values:

            Min             0
            Median          0
            Max             1584
            NumMissing      393
```

```
T = rmmissing(T1,'MinNumMissing',1); %remove the missing values

a = (T.satisfaction); %Response variable
T.satisfaction = [];
Y = grp2idx(a); % convert categorical to numbers
m = grp2idx(T.Gender);
n = grp2idx(T.CustomerType);
```

```
o = grp2idx(T.TypeOfTravel);
p = grp2idx(T.Class);
X = [m,n,T.Age,o,p,T.FlightDistance,T.SeatComfort,T.Departure_ArrivalTimeConvenient,T.FoodAndDr

%Divide the data into training and testing
cv = cvpartition(length(X),'holdout',0.4);
Xtrain = X(training(cv),:);
Xtest = X(test(cv),:);
Ytrain = Y(training(cv),:);
Ytest = Y(test(cv),:);
%Decision tree modelling
t = fitctree(Xtrain,Ytrain)
```

```
t =
  ClassificationTree
             ResponseName: 'Y'
    CategoricalPredictors: []
               ClassNames: [1 2]
           ScoreTransform: 'none'
          NumObservations: 77693


  Properties, Methods
```

```
Y_t = predict(t,Xtest);
%confusion matrix
cmtree = confusionchart(Ytest,Y_t)
```

```
cmtree =
  ConfusionMatrixChart with properties:

    NormalizedValues: [2×2 double]
         ClassLabels: [2×1 double]

  Show all properties
```

```
%AUC
[Xt1,Yt1,Thresholds,AUCt] = perfcurve(Ytest,Y_t,1);
```

SUPPORT VECTOR MACHINE (by Daniela Maldonado Sada)

```
clear all
clc
T =readtable('Invistico_Airline.csv',"VariableNamingRule","preserve");
% convert cell variables to categorical
names = T.Properties.VariableNames;
[nrows, ncols] = size(T);
category = false(1,ncols);

for i = 1:ncols
if isa(T.(names{i}),'cell')
category(i) = true;
T.(names{i}) = categorical(T.(names{i}));
end
end
```

```
rng('default'); %making sure the results are the same every time
D = dummyvar(T.Class);% encode categorical variables Class
D = array2table(D);
D.Properties;
T = [T ,D];% add new variable to cars
T.Class =[];
% T.satisfaction = double(T.satisfaction);
T.Gender = double(T.Gender);
T.('Customer Type') = double(T.('Customer Type'));
T.('Type of Travel') = double(T.('Type of Travel'));
% remove missing data, there are few missing data in arrival delay
completedata=rmmissing(T,'MinNumMissing',1);

% define model inputs and target
X = table2array(completedata(:,2:end));
Y = dummyvar(completedata.satisfaction);
Y=Y(:,1);
```

Spliting data in traing and test

```
rng 'default'
[L W]=size(X);
XSVM = cvpartition(L,'holdout',0.40);
% c = cvpartition(n,'KFold',k)

%Training set
Xtrain = X(training(XSVM),:);
Ytrain = Y(training(XSVM),:);

%Test set
Xtest = X(test(XSVM),:);
Ytest = Y(test(XSVM),:);

disp('Training Set');
```

```
Training Set
```

```
tabulate(Ytrain)
```

```
  Value    Count    Percent
      0    42468     54.66%
      1    35225     45.34%
```

```
disp('Test Set');
```

```
Test Set
```

```
tabulate(Ytest);
```

```
  Value    Count    Percent
      0    28414     54.86%
      1    23380     45.14%
```

Modeling SVM

```
cvp = cvpartition(Ytrain, 'KFold', 5);
mdlSVM = fitcsvm(Xtrain,Ytrain,'Standardize',1,'KernelFunction','RBF',...
    'KernelScale','auto');
CVSVMModel = crossval(mdlSVM);
loss_vector_svm=kfoldLoss(CVSVMModel)
```

```
loss_vector_svm = 0.0512
```

```
accuracy_vector_svm=1-loss_vector_svm
```

```
accuracy_vector_svm = 0.9488
```

Predicting

```
%confusion chart
[predicted_classes_svm, Posterior_svm] = kfoldPredict(CVSVMModel);
```

Makig the AUC curve to see the performance

```
[Xsvm,Ysvm,~,AUC_Svm] = perfcurve(Ytrain,Posterior_svm(:,2),'1');
figure;
plot(Xsvm,Ysvm)
xlabel('False positive rate'); ylabel('True positive rate');
CorX=corr(X)
```

```
CorX = 24×24
    1.0000   -0.0308    0.0090    0.0092    0.1208   -0.0721    0.0520   -0.0591 ···
   -0.0308    1.0000   -0.2843   -0.3082    0.0190   -0.0430   -0.1861   -0.0489
    0.0090   -0.2843    1.0000   -0.0449   -0.2494    0.0085    0.0389    0.0155
    0.0092   -0.3082   -0.0449    1.0000   -0.1232    0.0173    0.1915   -0.0314
    0.1208    0.0190   -0.2494   -0.1232    1.0000   -0.0425    0.0014   -0.0048
   -0.0721   -0.0430    0.0085    0.0173   -0.0425    1.0000    0.4349    0.7160
    0.0520   -0.1861    0.0389    0.1915    0.0014    0.4349    1.0000    0.5276
   -0.0591   -0.0489    0.0155   -0.0314   -0.0048    0.7160    0.5276    1.0000
   -0.0110   -0.0003   -0.0008   -0.0138   -0.0023    0.4054    0.5443    0.5235
   -0.0316   -0.0736    0.0140   -0.0189    0.0123    0.1292   -0.0016    0.0261
       ⋮
```

Making the confusion chart to see the true and false positive

```
Y_svm = predict (mdlSVM, Xtest)
```

```
Y_svm = 51794×1
     0
     0
     0
     0
     0
     0
     0
     0
     0
     1
     ⋮
```

```
cmsvm = confusionchart(Ytest,Y_svm,'RowSummary','row-normalized')
```

ARTIFICIAL NEURAL NETWORK (by Ali Izadkhah)

```
clear all
clc

% read data table
T =readtable('Invistico_Airline.csv',"VariableNamingRule","preserve");

% get insight from the data
summary(T)

% convert cell variables to categorical
names = T.Properties.VariableNames;
[nrows, ncols] = size(T);

category = false(1,ncols);
for i = 1:ncols
if isa(T.(names{i}),'cell')
category(i) = true;
T.(names{i}) = categorical(T.(names{i}));
end
end

rng('default');

% encode categorical variables Class
D = dummyvar(T.Class);
D = array2table(D);
D.Properties;
% add new variable to cars
T = [T ,D];
T.Class =[];

%T.satisfaction = double(T.satisfaction);
T.Gender = double(T.Gender);
T.('Customer Type') = double(T.('Customer Type'));
T.('Type of Travel') = double(T.('Type of Travel'));

% remove missing data, there are few missing data in arrival delay
completedata=rmmissing(T,'MinNumMissing',1);

% calculate the training time
tic

% define model inputs and target
```

```
inputs = table2array(completedata(:,2:end))';
targets = dummyvar(completedata.satisfaction)';;

% Initialize neural network
hiddenLayerSize = 17;
net = patternnet(hiddenLayerSize);

% divide data to train, vailidation and test sets
net.divideParam.trainRatio = 70/100;
net.divideParam.valRatio = 20/100;
net.divideParam.testRatio = 10/100;

% Train the network
[net,tr] = train(net,inputs,targets);

% Predict response
scoreTest = net(inputs(:,tr.testInd));
[~,yPred] = max(scoreTest);

% Evaluate classification with confusion matrix
cstest = completedata.satisfaction(tr.testInd);
yTrue = double(cstest);
confusionchart(yTrue,yPred','RowSummary','row-normalized');

% Determine validation error
cstest = completedata.satisfaction(tr.testInd);
validErr = 100*nnz(yPred' ~= double(cstest))/length(cstest)

% calculate model accuracy
accuracy_final_model= 100-validErr

% calculate recall and precision
tp = sum((yPred' == 1) & (yTrue == 1));
fp = sum((yPred' == 1) & (yTrue == 2));
fn = sum((yPred' == 2) & (yTrue == 1));

precision = tp / (tp + fp);
recall = tp / (tp + fn);
F1 = (2 * precision * recall) / (precision + recall);
recall
precision

% calculate ROC and AUC
[xTr, yTr, TTr, aucTr] = perfcurve(double(completedata.satisfaction(tr.testInd)), scoreTest(1,
aucTr

% Plot ROC curve
figure
plot(xTr, yTr, 'LineWidth',4);
xlabel('False positive rate'); ylabel('True positive rate');
title('ROC Curve with ANN model')

toc
```