MACHINE LEARNING: CLASSIFICATION

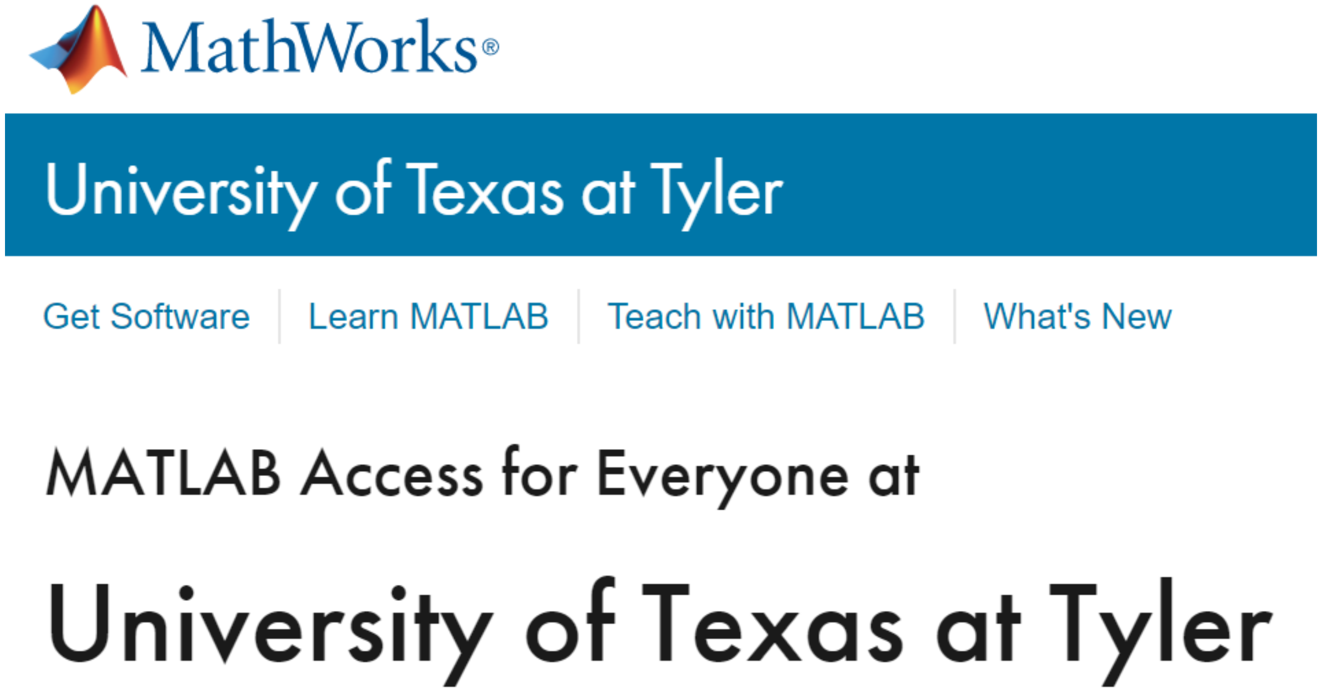
PREMANANDA INDIC, PH.D.

DEPARTMENT OF ELECTRICAL ENGINEERING

The University of Texas at TYLER Center for Health Informatics & Analytics

ORS Research Design & Data Analysis Lab
Office of Research and Scholarship

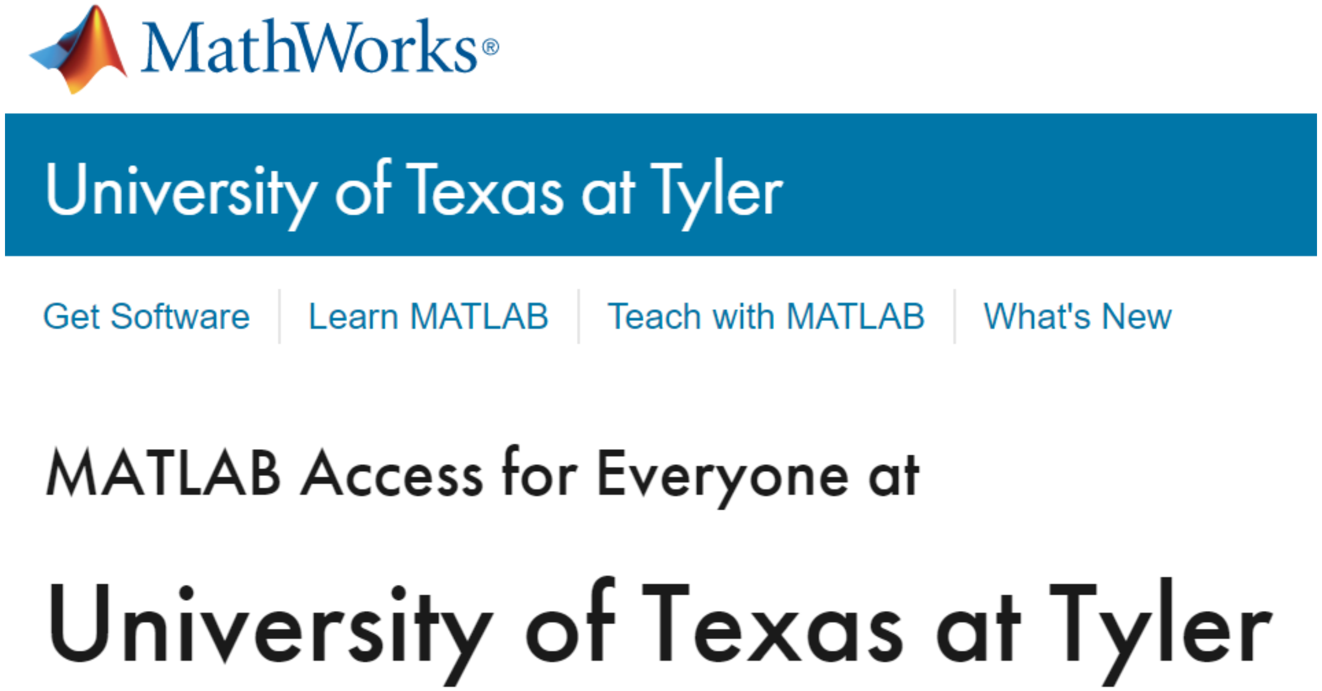
ANALYSIS PLATFORM

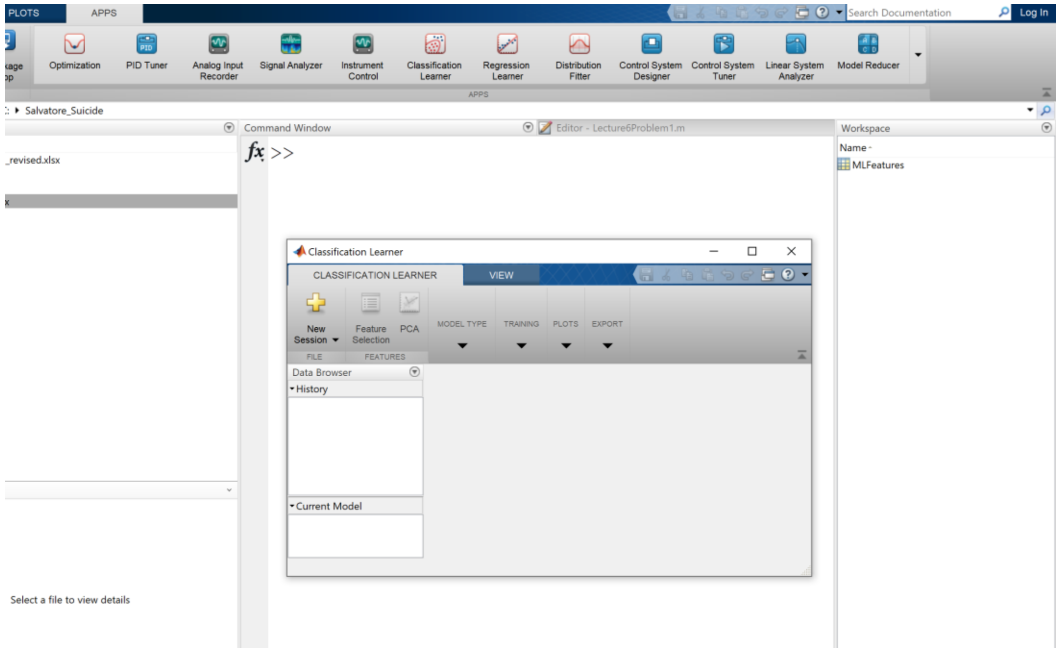


MATLAB R2021b

https://www.mathworks.com/academia/tah-portal/university-of-texas-at-tyler-1108545.html

ANALYSIS PLATFORM





https://www.mathworks.com/academia/tah-portal/university-of-texas-at-tyler-1108545.html

OUTLINE

- INTRODUCTION
- DIFFERENT CLASSIFIERS
- EXAMPLES

OUTLINE

- INTRODUCTION
- DIFFERENT CLASSIFIERS
- EXAMPLES

INTRODUCTION

- What is Machine Learning ?

- Machine Learning is a field of study that gives computers the ability to "learn" without being explicitly programmed
  - Prediction
  - Classification

Samuel AL, IBM J. Research & Development, 1959, vol. 3 (3), 210-229

INTRODUCTION

- What is Machine Learning ?

- Machine Learning is a field of study that gives computers the ability to "learn" without being explicitly programmed
  - Prediction
  - Classification

Samuel AL, IBM J. Research & Development, 1959, vol. 3 (3), 210-229

OUTLINE

- INTRODUCTION
- DIFFERENT CLASSIFIERS
- EXAMPLES

APPROACHES

- SUPERVISED LEARNING
- UNSUPERVISED LEARNING

APPROACHES

- SUPERVISED LEARNING (Classification / Prediction)

Provide training set with features and solutions

APPROACHES

- STANDARD MACHINE LEARNING

- ADVANCED MACHINE LEARNING

  Based on Artificial Neural Networks (Deep Learning)

APPROACHES

- CLASSIFICATION
  - Logistic Regression
  - Support Vector Machine

APPROACHES

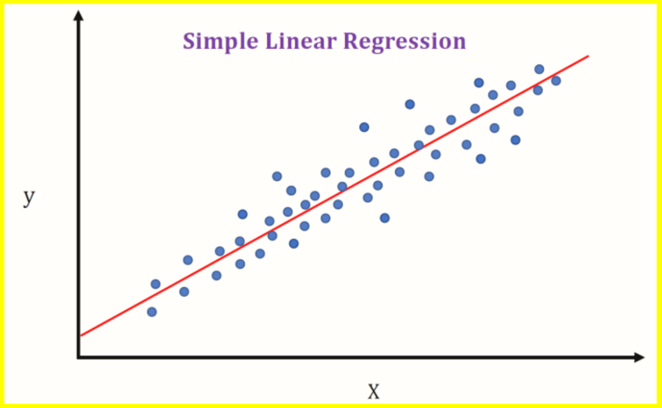
- CLASSIFICATION
  - Logistic Regression
  - Support Vector Machine

APPROACHES

- Linear Regression

$$\hat{y}^i = \theta_0 + \theta_1 x_1^i + \theta_2 x_2^i + \cdots \dots \dots . + \theta_n x_n^i \qquad i = 1,2, \dots \dots . . m$$

$$\hat{Y} = \Theta^T X$$



https://medium.datadriveninvestor.com/machine-learning-101-part-1-24835333d38a

- Gradient Descent by **Louis Augustin Cauchy** in 1847

Cost Function to Minimize

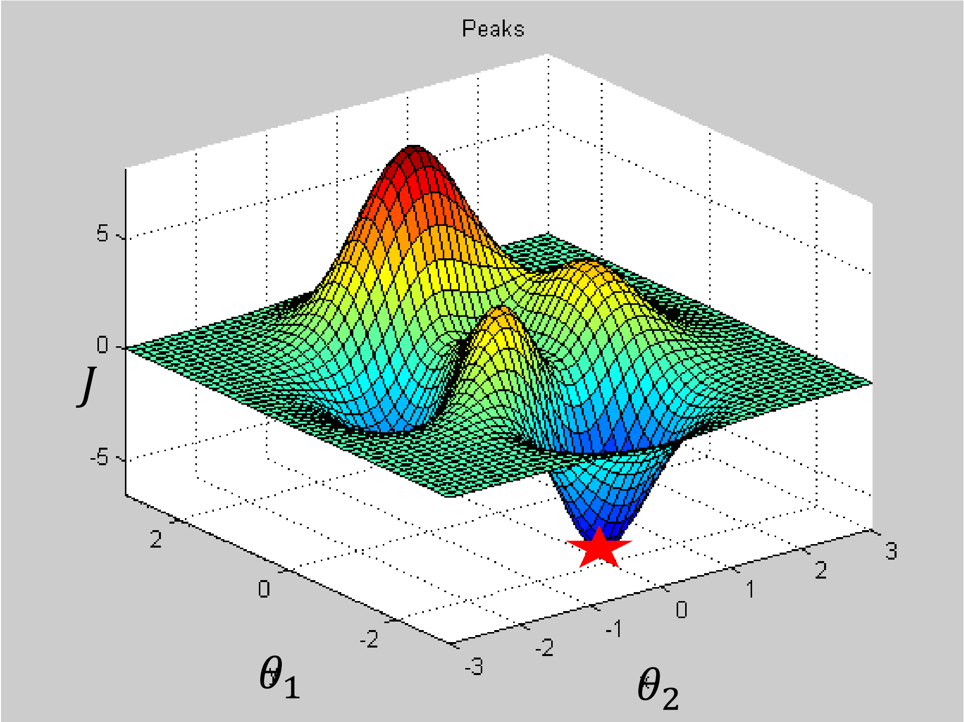
$$J = \left\langle \left(\hat{y}^i - y^i\right)^2 \right\rangle = \left(\hat{Y} - Y\right)^T \left(\hat{Y} - Y\right) = \frac{1}{m} \sum_{i=1}^{m} (\theta^T X^i - y^i)^2$$

APPROACHES

- Linear Regression

$$\Theta^{k+1} = \Theta^{k} - \gamma \nabla_{\Theta} J(\Theta)$$

$$\nabla_{\Theta} J(\Theta) = \frac{2}{m} X^{T}(X\Theta - Y)$$

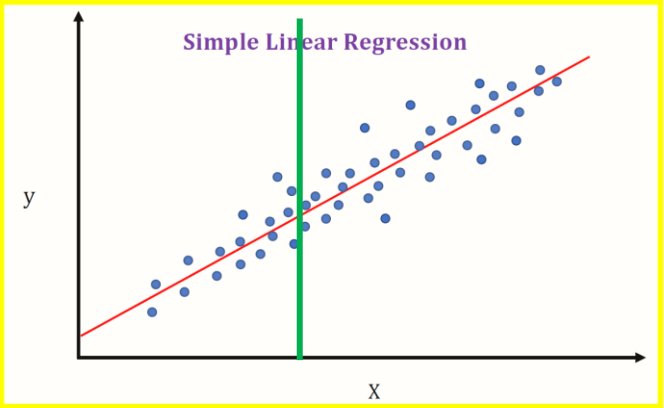


APPROACHES

- Logistic Regression

Two class $y = 1 \quad or\ y = 0$

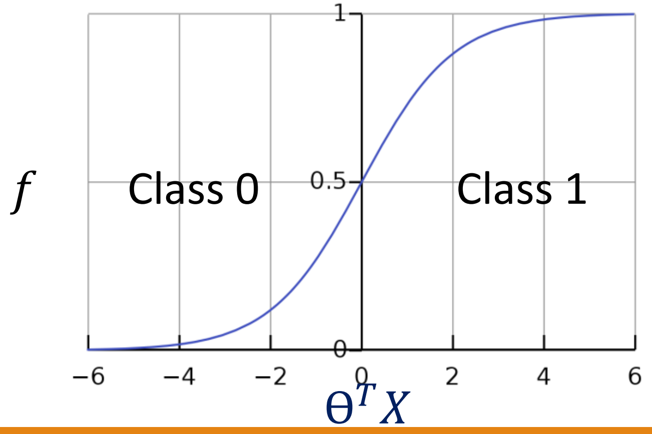
$$\hat{p} = f(\Theta^T X) = \frac{1}{1+e^{-\Theta^T X}}$$



https://medium.datadriveninvestor.com/machine-learning-101-part-1-24835333d38a

$$\hat{y} = 1\ if\ \hat{p} < 0.5;\ \ \hat{y} = 0\ if\ \hat{p} \geq 0.5$$

$$J = \frac{1}{m}\sum_{i=1}^{m}\left[y^i \log\left(\hat{p}^i\right) + \left(1 - y^i\right)\log(1 - \hat{p}^i)\right]$$

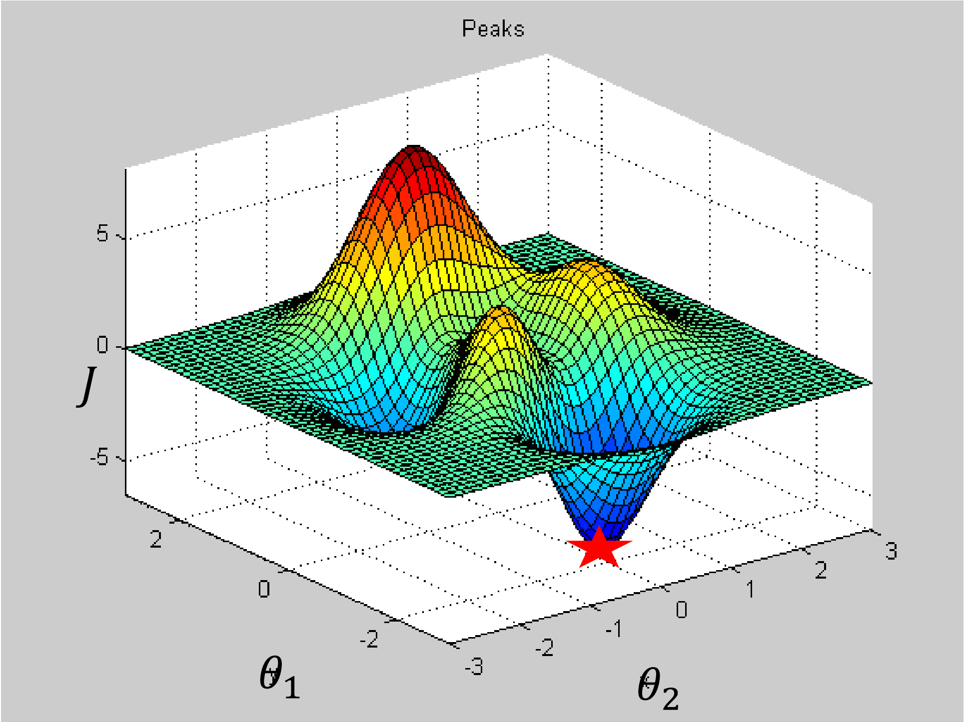


APPROACHES

- Logistic Regression

$$\Theta^{k+1} = \Theta^{k} - \gamma \nabla_{\Theta} J(\Theta)$$

$$\frac{\partial}{\partial \theta_j} J(\Theta) = \frac{1}{m} \sum_{i=1}^{m} \left(f\left(\Theta^T X^i\right) - y^i \right) x_j^i$$



APPROACHES

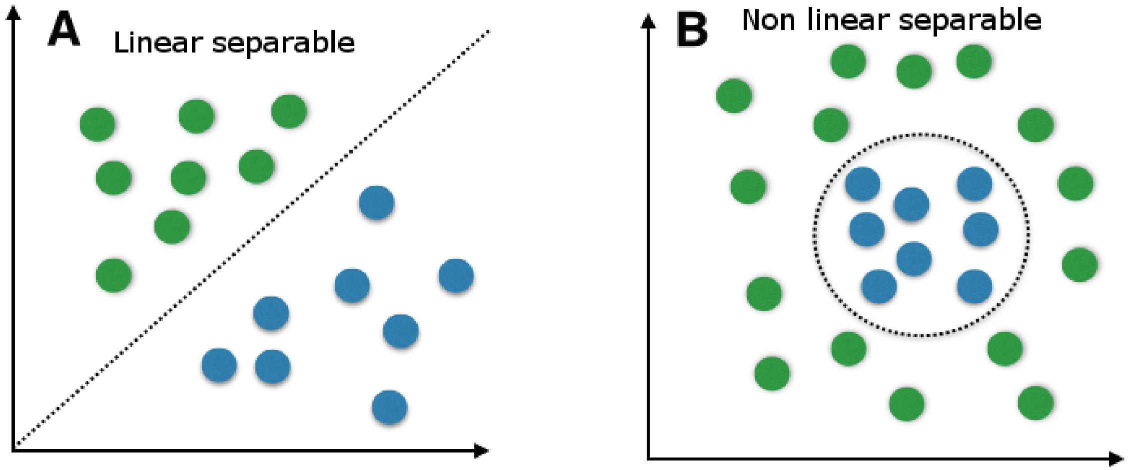
- Support Vector Machine

$$G(x_j, x_k) = \exp\left(-\|x_j - x_k\|^2\right)$$

$G(x_j, x_k) = (1 + x_j{}'x_k)^q$, where q is in the set {2,3,...}.

$$f(X) = w^T X - b$$

x_2



x_1

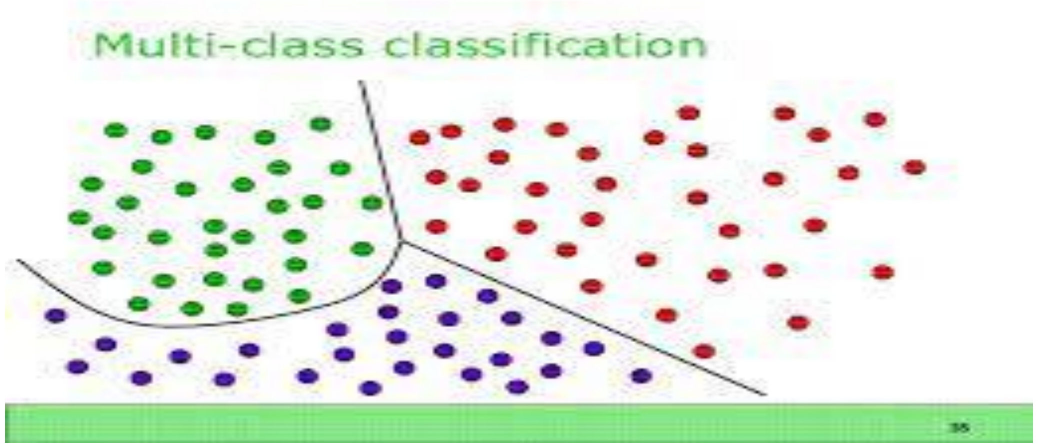
https://medium.com/@LSchultebraucks/introduction-to support-vector-machines-9f8161ae2fcb

APPROACHES

- SUPERVISED LEARNING (Classification / Prediction)
  - Support Vector Machine (SVM)

Used for regression as well as classification

x_2



x_1

https://www.mathworks.com/matlabcentral/fileexchange/62061-multi-class-svm

APPROACHES

- SUPERVISED LEARNING (Classification)
  - Logistic Regression
  - Support Vector Machines
  - k-Nearest Neighbors
  - Decision Trees and Random Forests

SECTION 1: Learner App

- **Home Value Classification:** 9 features to classify high vs low medianHouseValue

longitude: A measure of how far west a house is; a higher value is farther west

latitude: A measure of how far north a house is; a higher value is farther north

housingMedianAge: Median age of a house within a block; a lower number is a newer building

totalRooms: Total number of rooms within a block

totalBedrooms: Total number of bedrooms within a block

population: Total number of people residing within a block

households: Total number of households, a group of people residing within a home unit, for a block

medianIncome: Median income for households within a block of houses (measured in tens of thousands of US Dollars)

medianHouseValue: Median house value for households within a block (measured in US Dollars)

oceanProximity: Location of the house w.r.t ocean/sea

https://www.kaggle.com/camnugent/california-housing-prices

Demo with N=5000
70% Training Data
30% Test Data
Models Trained:
Logistic Regression
SVM

SECTION 1: Learner App

- Prediction of House Price Classification Problem

Confusion Matrix

True Class		Predicted 1	Predicted 0	
	1	True Positive	False Negative	→ Total Positive
	0	False Positive	True Negative	→ Total Negative

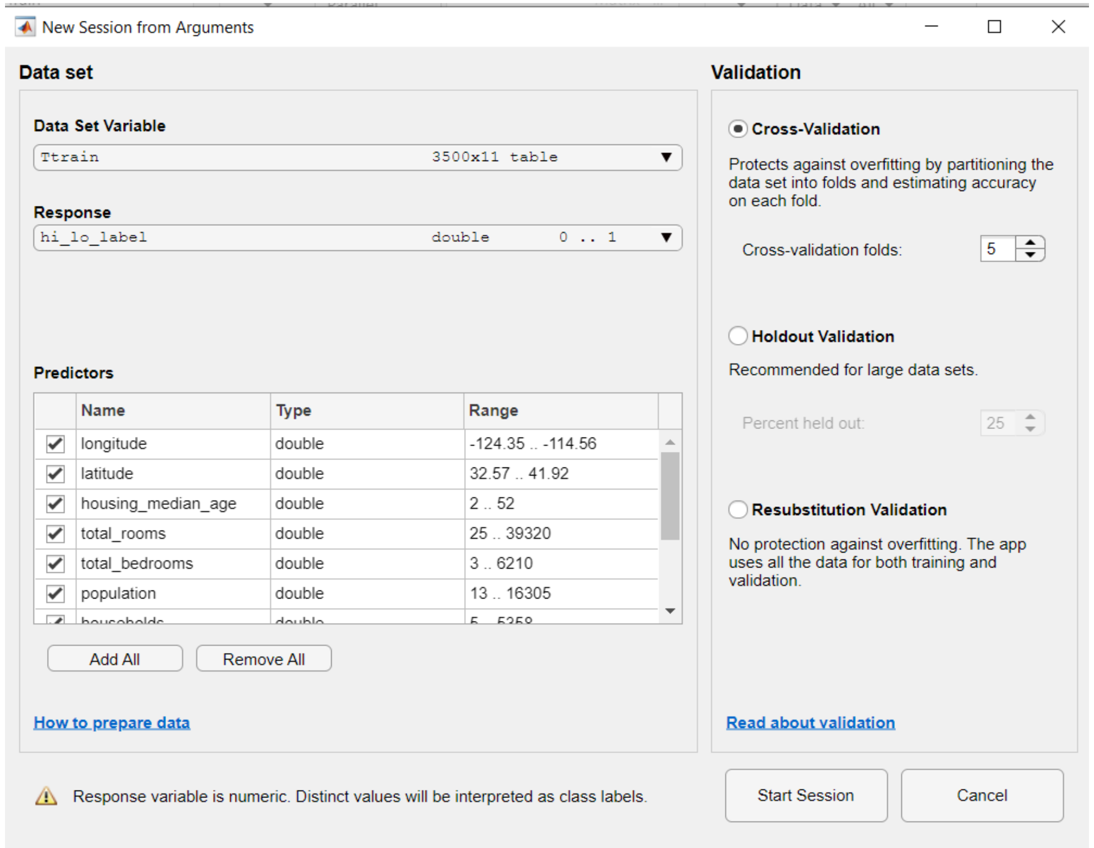
Predicted Class

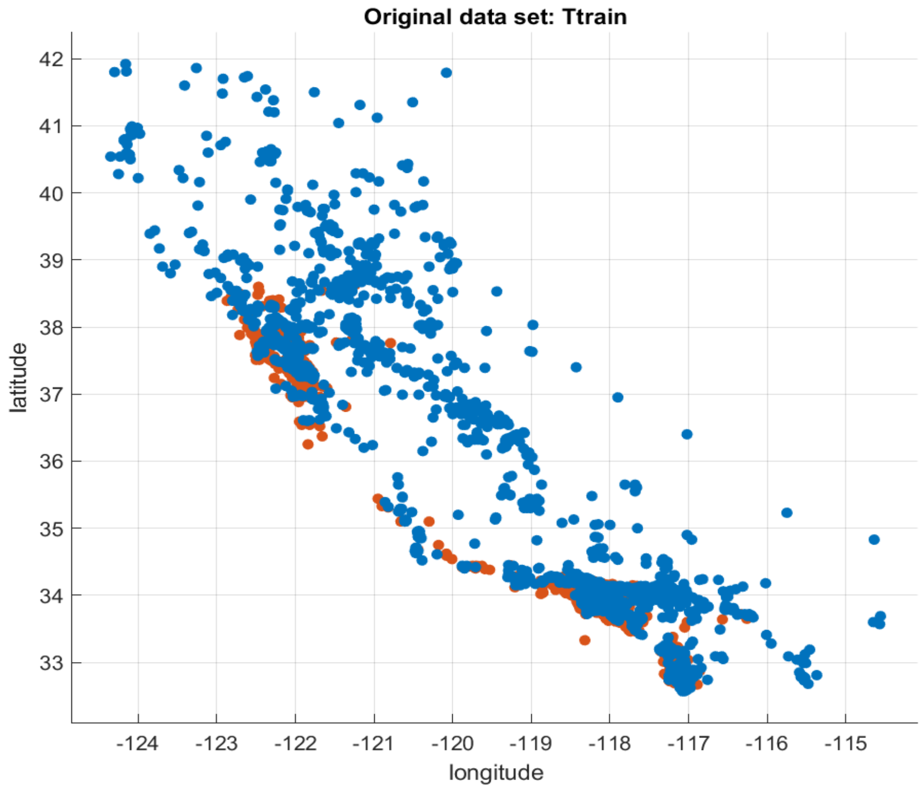
True Positive Rate = True Positive / Total Positive

True Negative Rate = True Negative / Total Negative = 1 – False Positive Rate

SECTION 1: Learner App

➢ DATA IMPORT & CLASSIFICATION LEARNER INITIALIZATION





SECTION 1: Learner App

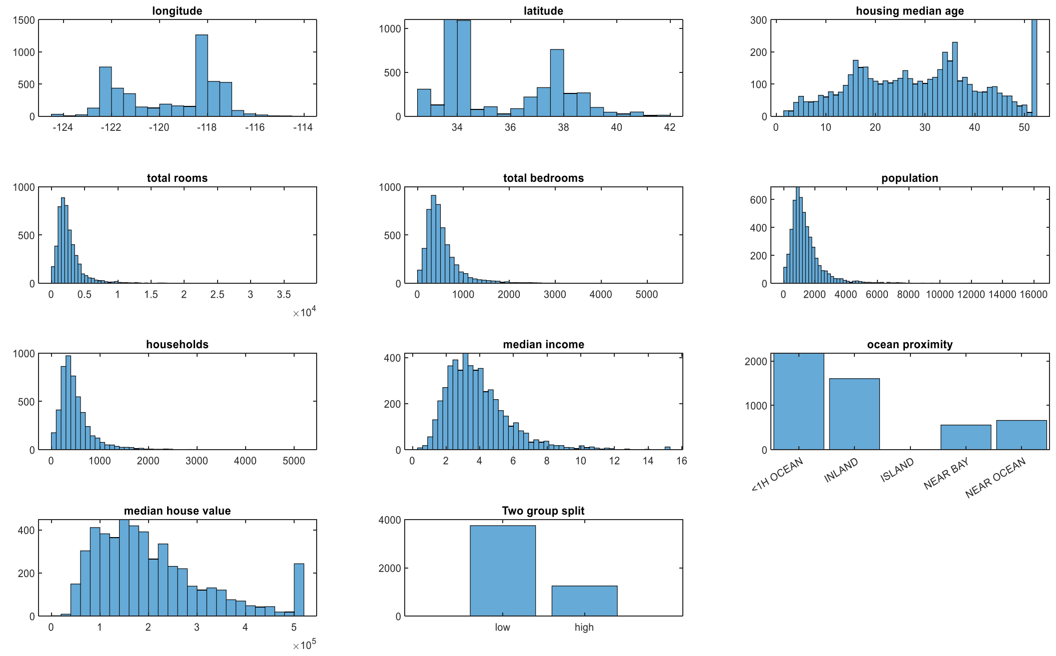
- DATA IMPORT & CLASSIFICATION LEARNER INITIALIZATION

```
classificationLearner(Ttrain,'hi_lo_label');
```

Demo with logistic regression and linear SVM

SECTION 2: Raw Data Analysis

Visualize the data, Summarize variables, data cleaning, pre-processing if needed



207 Missing values, replace with median values

ocean_proximity: 20636×1 categorical

Values:

<1H OCEAN 9135

INLAND 6550

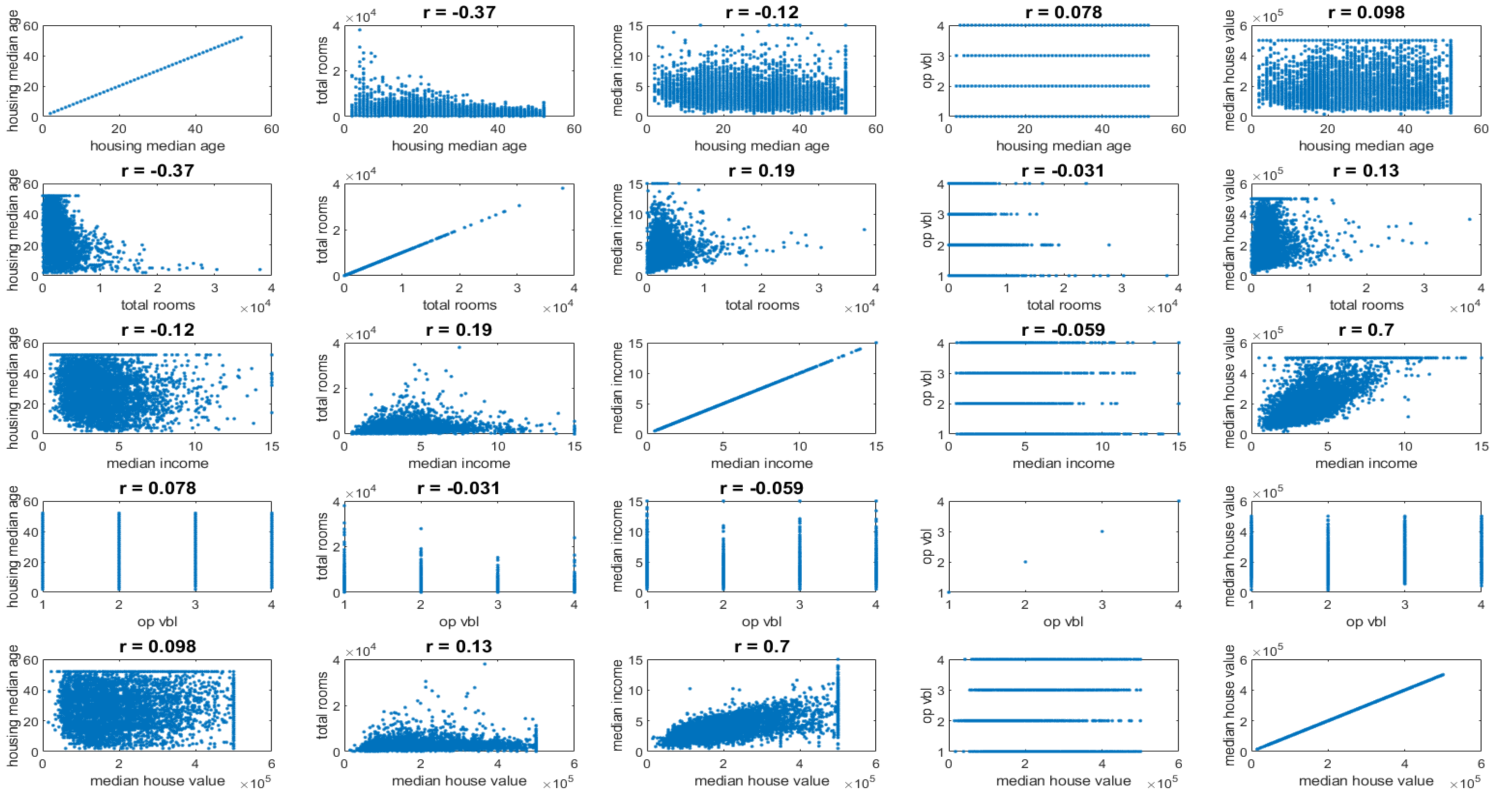
ISLAND 5

NEAR BAY 2289

NEAR OCEAN 2657

SECTION 3: Correlation Analysis

FIND VARIABLE CORRELATIONS TO EACH OTHER AND THE MEDIAN HOUSE VALUE




```
[R,pp] = corr(table2array(T1(:,select_vars)));
```

SECTION 4: Logistic Regression

SPLIT INTO TRAINING AND TEST DATA AND FIT LOGISTIC REGRESSION MODEL

Estimated Coefficients:

	Estimate	SE	tStat	pValue
(Intercept)	-154.19	14.421	-10.692	1.1065e-26
longitude	-1.7683	0.17448	-10.135	3.8752e-24
latitude	-1.8133	0.18885	-9.6018	7.8546e-22
housing_median_age	0.044239	0.0051484	8.5928	8.4901e-18
total_rooms	0.0003444	9.7387e-05	3.5364	0.00040561
total_bedrooms	0.00080298	0.00084259	0.95299	0.3406
population	-0.0023529	0.00020995	-11.207	3.7737e-29
households	0.0039573	0.00094559	4.185	2.8517e-05
median_income	1.0172	0.053904	18.87	2.0101e-79
ocean_proximity_INLAND	-0.053285	0.24937	-0.21368	0.8308
ocean_proximity_ISLAND	0	0	NaN	NaN
ocean_proximity_NEAR BAY	-0.10616	0.19861	-0.53449	0.593
ocean_proximity_NEAR OCEAN	0.11076	0.15948	0.6945	0.48737

3500 observations, 3488 error degrees of freedom
Dispersion: 1
Chi^2-statistic vs. constant model: 1.83e+03, p-value = 0

```
mdl = fitglm([Ttrain(:,1:9)
table(y)],'Distribution','binomial');
```

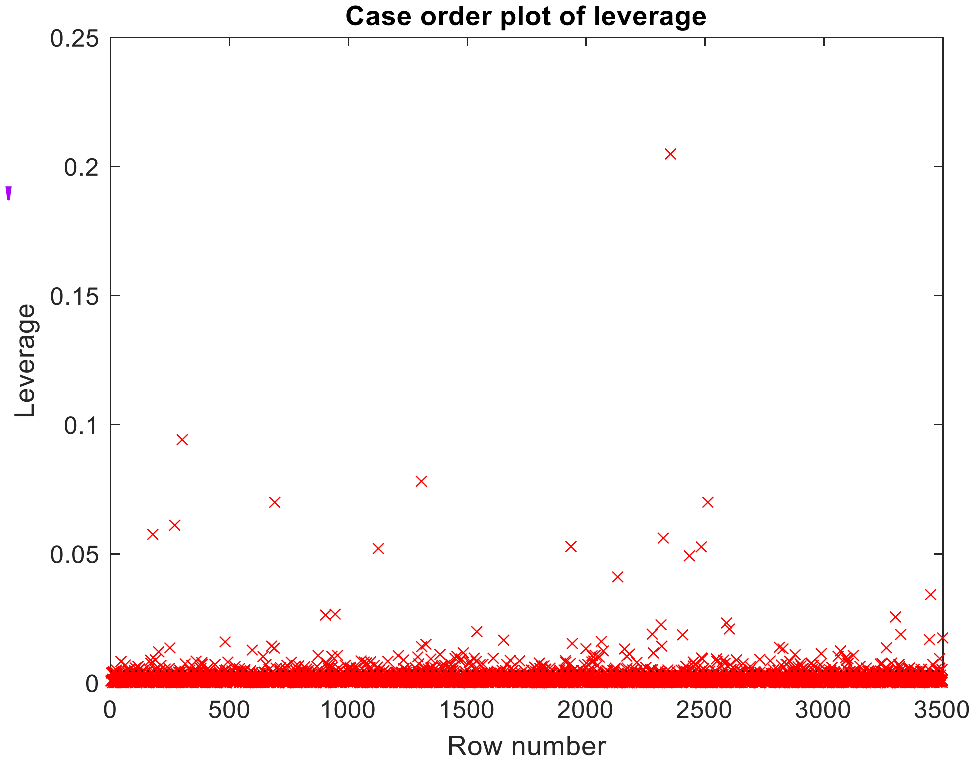
Remove Insignificant features

SECTION 5: Outliers

DIAGNOSTICS OF MODELS- IDENTIFY OUTLIERS

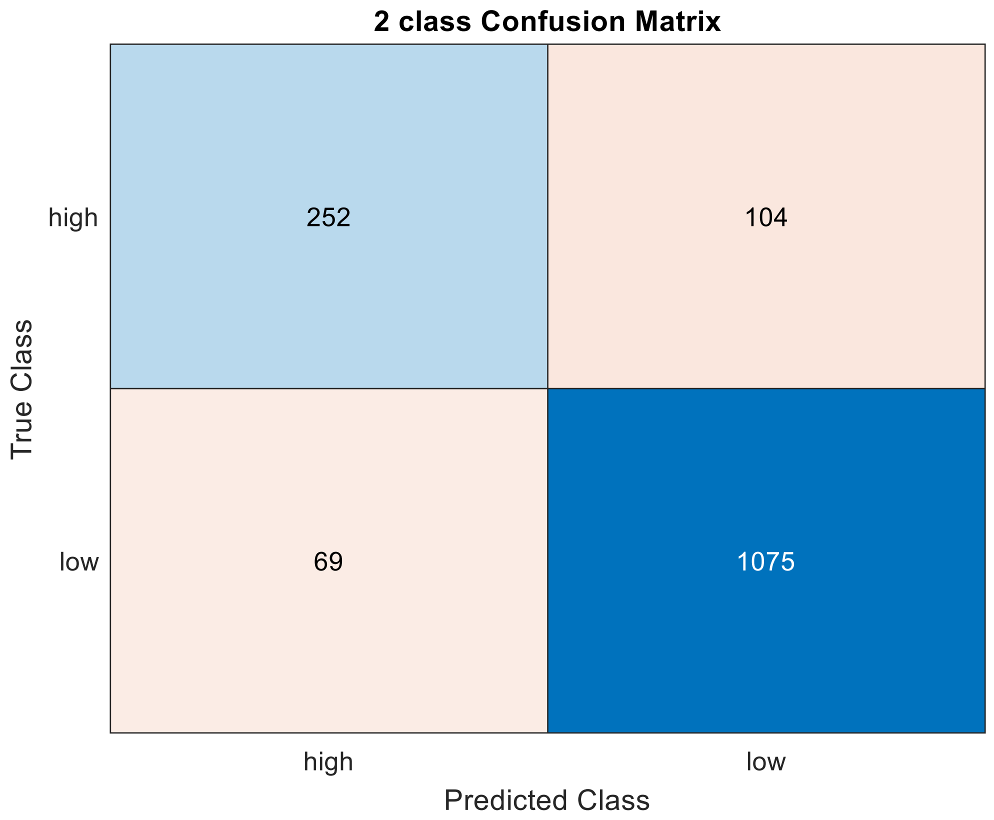
```
mdl1 = fitglm([Ttrain(:,[1:4 6:8])
table(y,'variablenames',{'Hi_lo_label'})],'
Distribution','binomial');
```

```
plotDiagnostics(mdl1,'leverage')
```



SECTION 6: Classification (Clean Data)

TEST MODEL FOR TWO CLASS CLASSIFICATION (Logistic Regression)

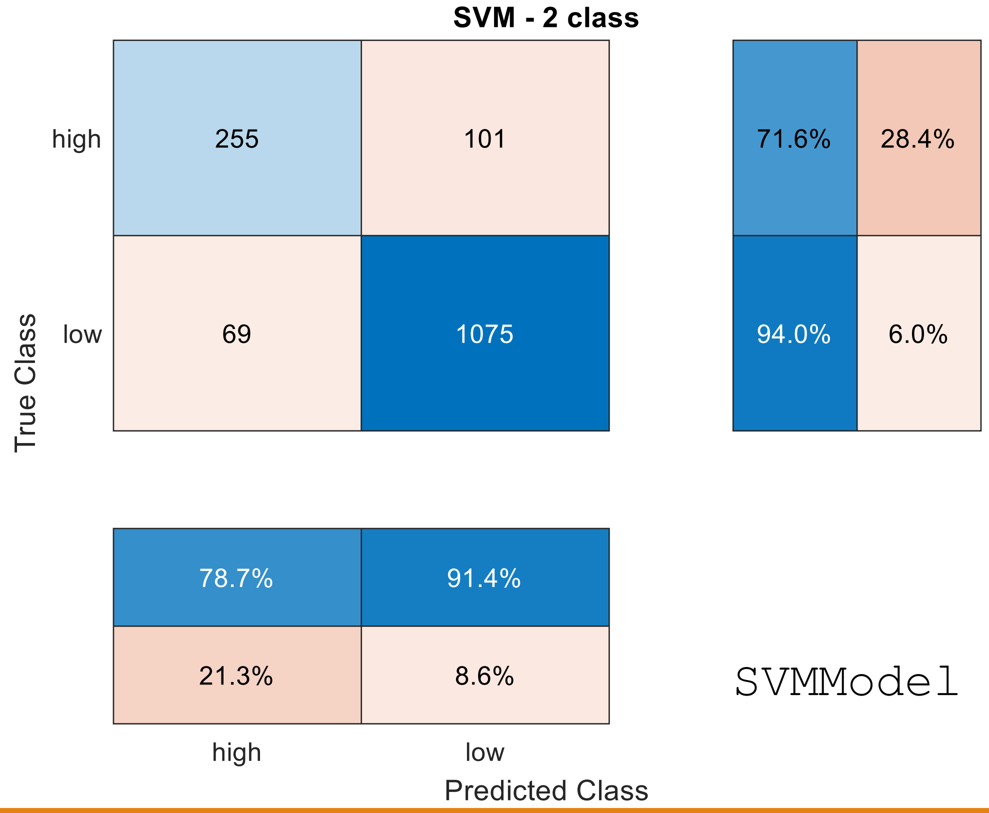


Test Data N = 1500
(30% of 5000)

Missing Values
Insignificant Features
Outliers

SECTION 7: SVM Classification

REGULARIZATION OF VARIABLES DONE AUTOMATICALLY, NO NEED TO CHOOSE FEATURES SEPARATELY AS WAS DONE EARLIER FOR LOGISTIC REGRESSION



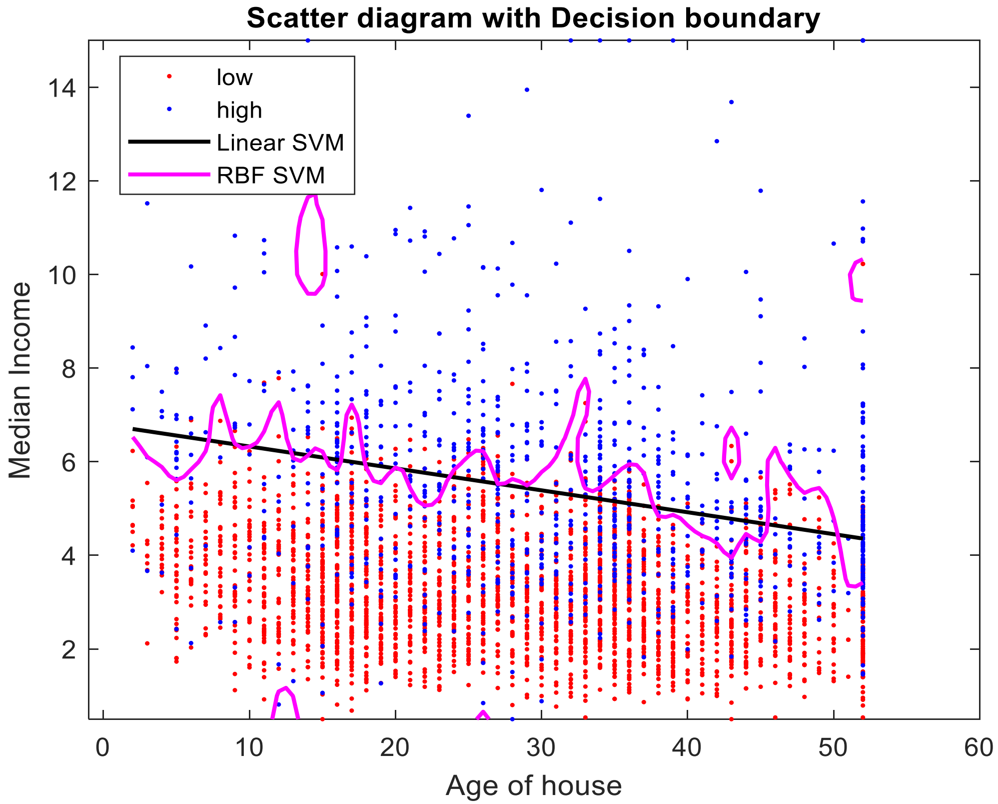
Test Data N = 1500
(30% of 5000)

Linear SVM

```
SVMModel = fitcsvm(Ttrain(:,1:9),y,'standardize',true);
```

SECTION 8: SVM Classification

LINEAR vs RADIAL BASIS FUNCTION (RBF) KERNEL



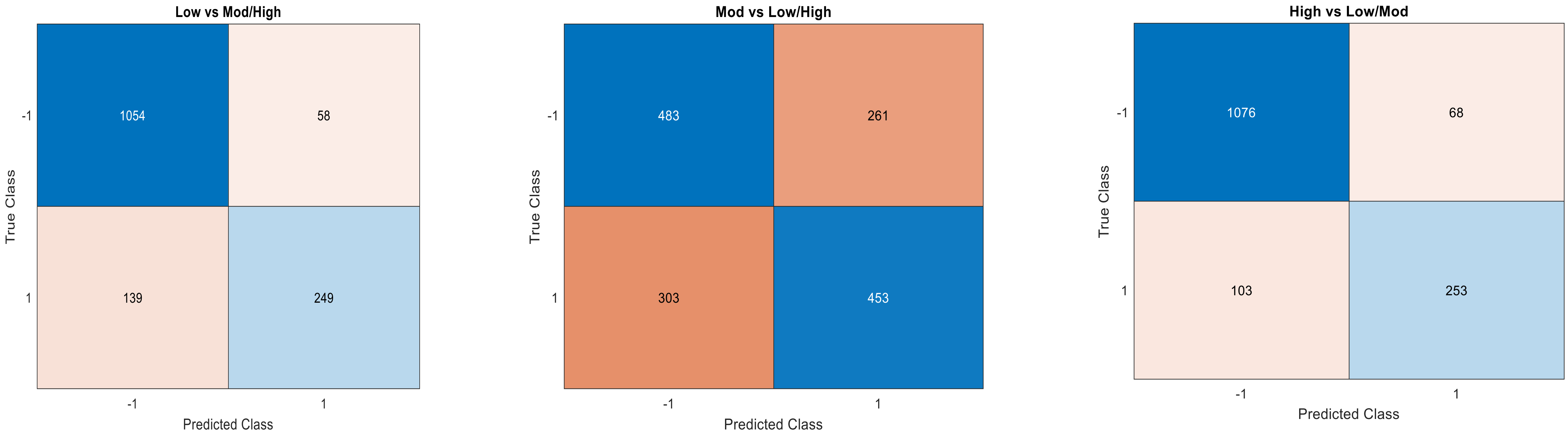

```
fitcsvm([x1 x2],y1);

fitcsvm([x1 x2],y1,'KernelFunction','rbf');
```

x1: Age of House
X2: Median Income

SECTION 9: Multiclassification (SVM)

ONE CLASS vs REST

Also perform one to one class

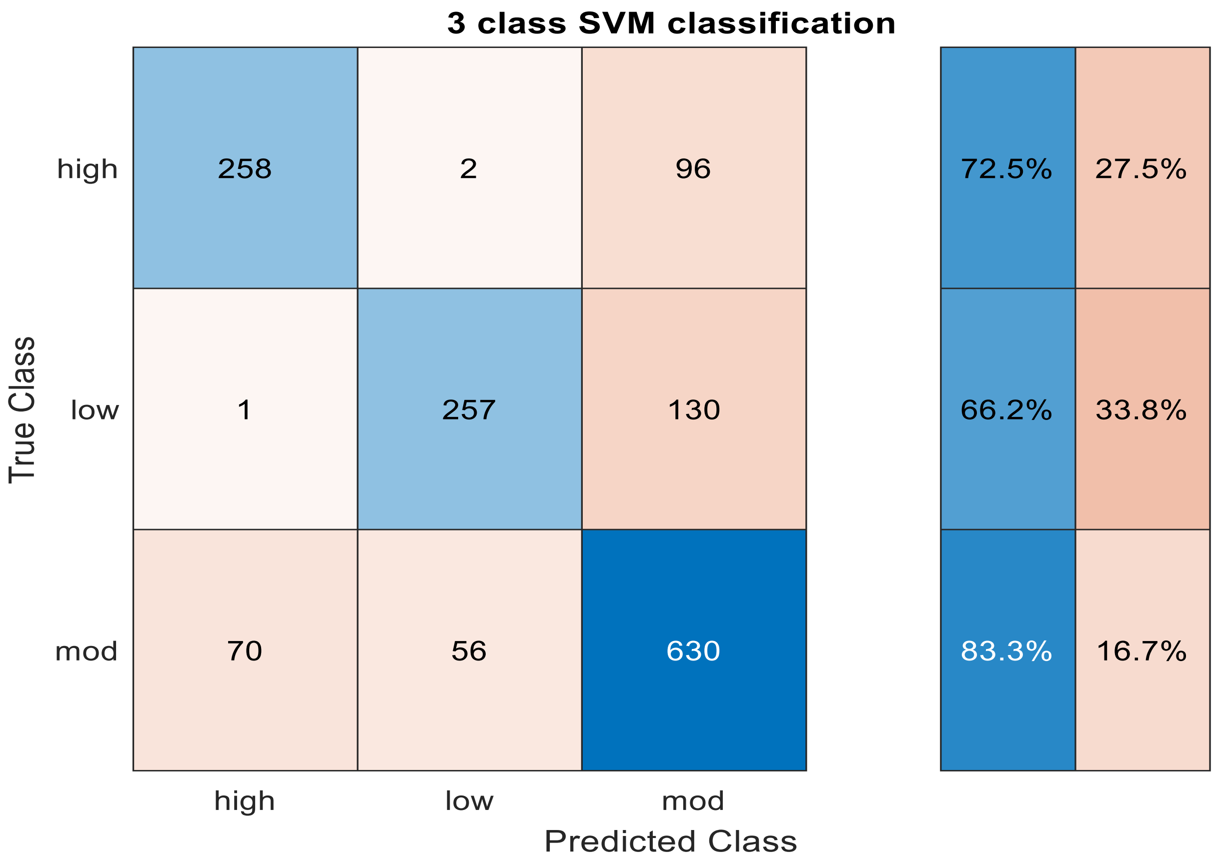



```
Mdl =
fitcecoc(Ttrain(:,1:8),y,'Learners',t,'Coding',coding,'ResponseName',responseName,...
    'PredictorNames',predictorNames,'ClassNames',classNames);
```

SECTION 10: Multiclassification (SVM)

LOW vs MOD vs HIGH CLASS

```
Mdlp =
fitcecoc(Ttrain(:,1:8),y,'Learner
s',t,'FitPosterior',true,...

'ClassNames',{'low','mod','high'}
,...
     'Verbose',2);
```



CONCLUSION

- Classification divides the data into different groups
- Look at the raw data and understand features in relation to class designation
- Several codes are available to perform classification

THANK YOU

ORS Research Design & Data Analysis Lab

Office of Research and Scholarship

SBIR: RAE (Realize, Analyze, Engage) - A digital biomarker based detection and intervention system for stress and carvings during recovery from substance abuse disorders.
PIs: M. Reinhardt, S. Carreiro, P. Indic

STARs Award

The University of Texas System
P. Indic (PI, UT Tyler)

Department of Veterans Affairs

Design of a wearable sensor system and associated algorithm to track suicidal ideation from movement variability and develop a novel objective marker of suicidal ideation and behavior risk in veterans.
Clinical Science Research and Development Grant (approved for funding),
P. Indic (site PI, UT-Tyler)
E.G. Smith (Project PI, VA)
P. Salvatore (Investigator, Harvard University)

Design of a wearable biosensor sensor system with wireless network for the remote detection of life threatening events in neonates.

National Science Foundation Smart & Connected Health Grant
P. Indic (Lead PI, UT-Tyler)
D. Paydarfar (Co PI, UT-Austin)
H. Wang (Co PI, UMass Dartmouth)
Y. Kim (Co PI, UMass Dartmouth)

Pre-Vent

National Institute Of Health Grant
P. Indic (Analytical Core PI, UT-Tyler)
N. Ambal (PI, Univ. of Alabama, Birmingham)

ViSiOn
P. Indic (site PI, UT-Tyler)
P. Ramanand (Co-I, UT Tyler
N. Ambal, (PI, Univ. of Alabama, Birmingham)

QUESTIONS