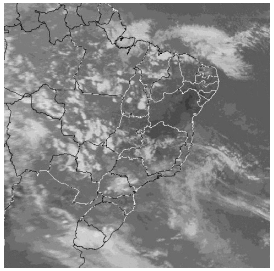


CAP-359 PRINCIPLES AND APPLICATIONS OF DATA MINING

Rafael Santos – rafael.santos@inpe.br
www.lac.inpe.br/~rafael.santos/

Data Mining Concepts and Applications

Classification

Some slides on this introduction adapted from *Introduction to Data Mining*; Pang-Ning Tan, Michael Steinbach and Vipin Kumar (2005).

Classification

- Prediction of a category or discrete label.
- Model or Classifier creation:
  - Input: instances with known classes.
  - Output: model based on the data and algorithm.
- Classification:
  - Input: unlabeled data.
  - Output: labels for the unlabeled data based on the model.
- Post-processing: model evaluation.

Introduction to Data Mining; Pang-Ning Tan, Michael Steinbach and Vipin Kumar (2005).

Classification

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125	No
2	No	Married	100	No
3	No	Single	70	No
4	Yes	Married	120	No
5	No	Divorced	95	Yes
6	No	Married	60	No
7	Yes	Divorced	220	No
8	No	Single	85	Yes
9	No	Married	75	No
10	No	Single	90	Yes

We want to predict who will cheat on taxes based on other attributes.

Sort

Tid	Refund	Marital Status	Taxable Income	Cheat
7	Yes	Divorced	220	No
2	No	Married	100	No
4	Yes	Married	120	No
6	No	Married	60	No
9	No	Married	75	No
1	Yes	Single	125	No
3	No	Single	70	No
5	No	Divorced	95	Yes
8	No	Single	85	Yes
10	No	Single	90	Yes

Classification

Tid	Refund	Marital Status	Taxable Income	Cheat
2	No	Married	100	No
6	No	Married	60	No
9	No	Married	75	No
3	No	Single	70	No
7	Yes	Divorced	220	No
4	Yes	Married	120	No
1	Yes	Single	125	No
5	No	Divorced	95	Yes
8	No	Single	85	Yes
10	No	Single	90	Yes

Sort 2

Sort 1

Sort 1

Sort 2

Tid	Refund	Marital Status	Taxable Income	Cheat
6	No	Married	60	No
3	No	Single	70	No
9	No	Married	75	No
8	No	Single	85	Yes
10	No	Single	90	Yes
5	No	Divorced	95	Yes
2	No	Married	100	No
4	Yes	Married	120	No
1	Yes	Single	125	No
7	Yes	Divorced	220	No

Classification

Tid	Refund	Marital Status	Taxable Income	Cheat
6	No	Married	60	No
3	No	Single	70	No
9	No	Married	75	No
8	No	Single	85	Yes
10	No	Single	90	Yes
5	No	Divorced	95	Yes
2	No	Married	100	No
4	Yes	Married	120	No
1	Yes	Single	125	No
7	Yes	Divorced	220	No

Sort 1

Sort 2

Sort 1

Sort 2

Tid	Refund	Marital Status	Taxable Income	Cheat
6	No	Married	60	No
3	No	Single	70	No
9	No	Married	75	No
8	No	Single	85	Yes
10	No	Single	90	Yes
5	No	Divorced	95	Yes
2	No	Married	100	No
4	Yes	Married	120	No
1	Yes	Single	125	No
7	Yes	Divorced	220	No

Classification

- Bad rule: nobody cheats: 3/10 errors.
- Bad rule: those who don't get refunds, cheat: 4/10 errors.
- Better rule: if 85 ≤ income ≤ 100 then cheat: 1/10 errors.
- Even better rule: if 85 ≤ income ≤ 95 then cheat: 0/10 errors.

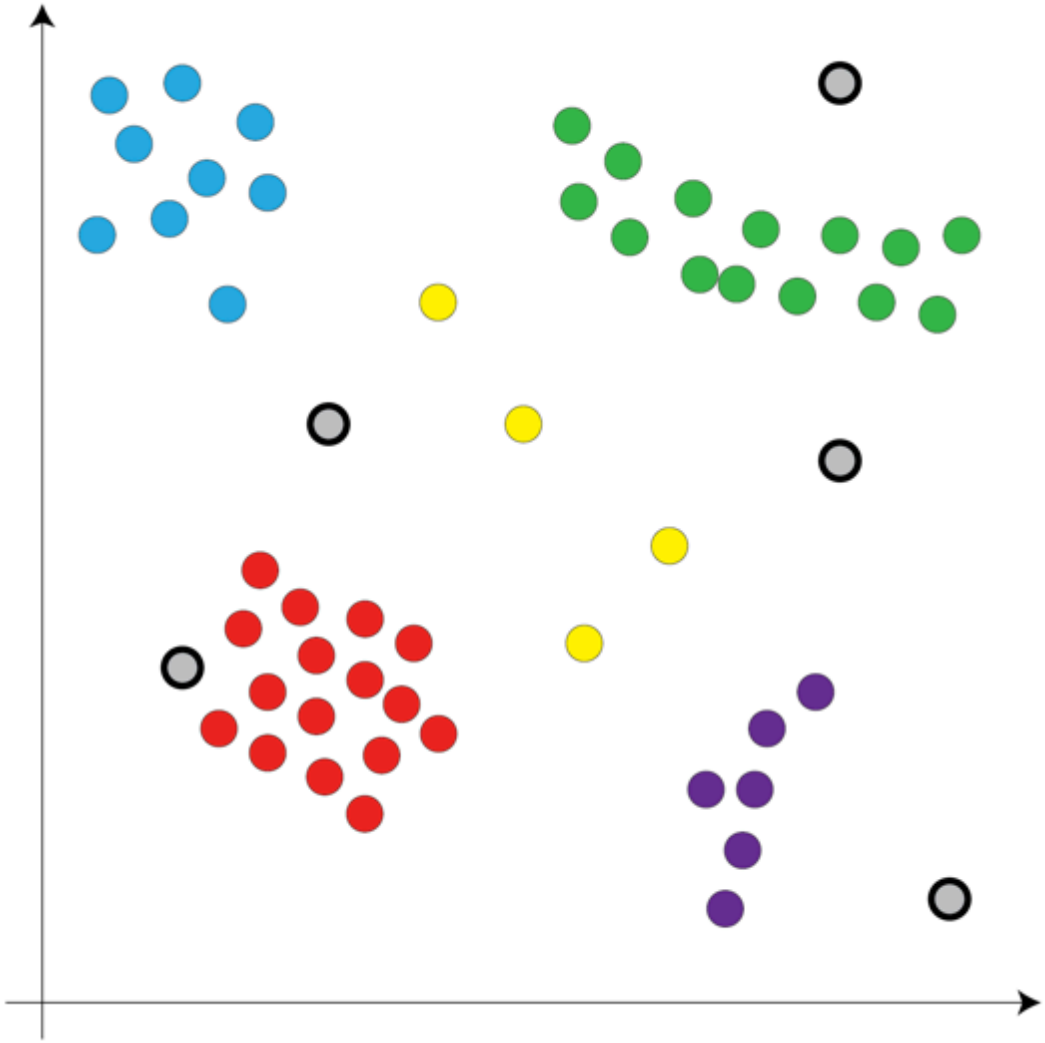
Tid	Refund	Marital Status	Taxable Income	Cheat
6	No	Married	60	No
3	No	Single	70	No
9	No	Married	75	No
8	No	Single	85	Yes
10	No	Single	90	Yes
5	No	Divorced	95	Yes
2	No	Married	100	No
4	Yes	Married	120	No
1	Yes	Single	125	No
7	Yes	Divorced	220	No

- Another perfect rule: if 75 ≤ income ≤ 95 *and* marital status is {single or divorced} then cheat: 0/10 errors.
- Another perfect rule: if 75 ≤ income ≤ 95 *and* marital status is {single or divorced} *and* refund is no then cheat: 0/10 errors.

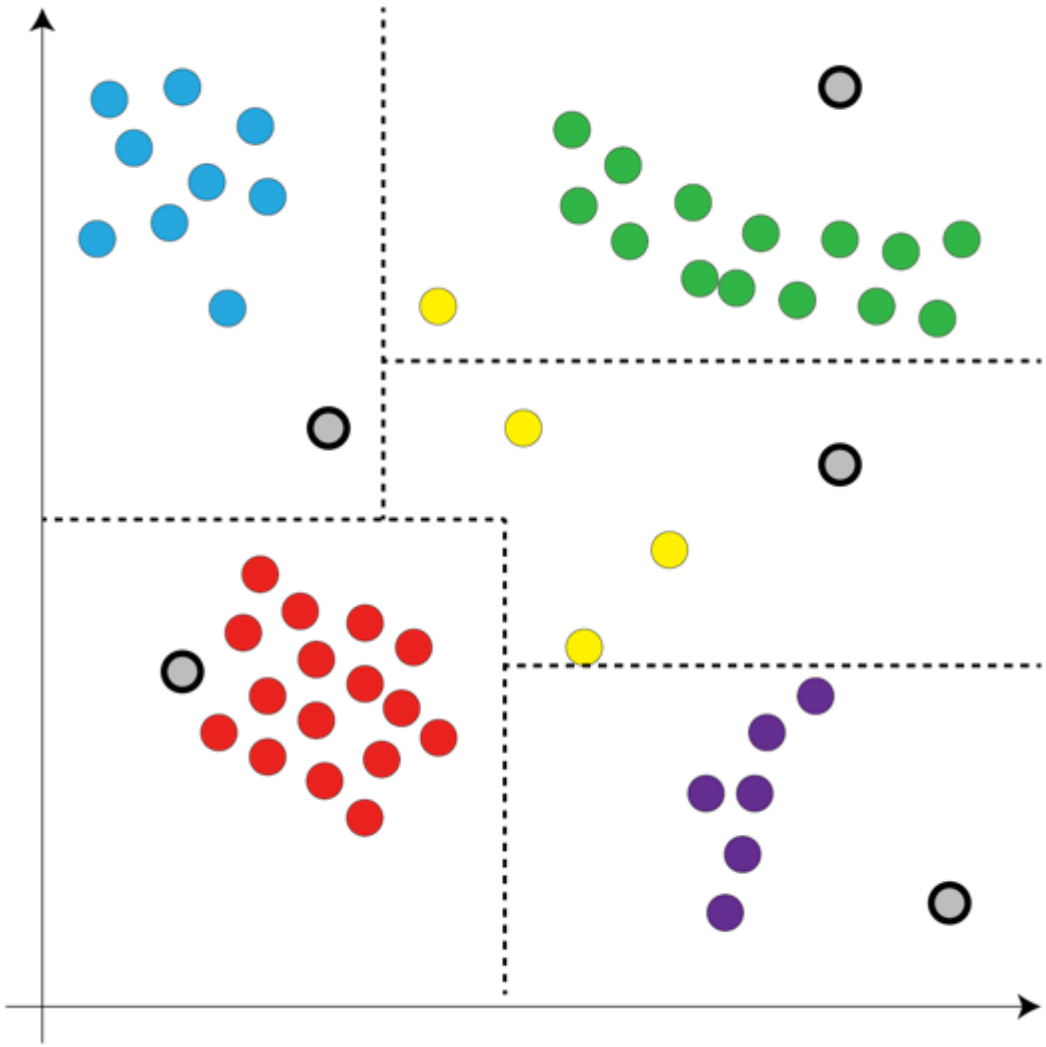
What do we want from a classifier?

- Classify unknown data.
  - Model must be robust enough to deal with previously unknown data – generalization!
- Explain our data, e.g. using statistics and rules.
  - Eventually there is no need to explain all data in intricate details: generalization again!

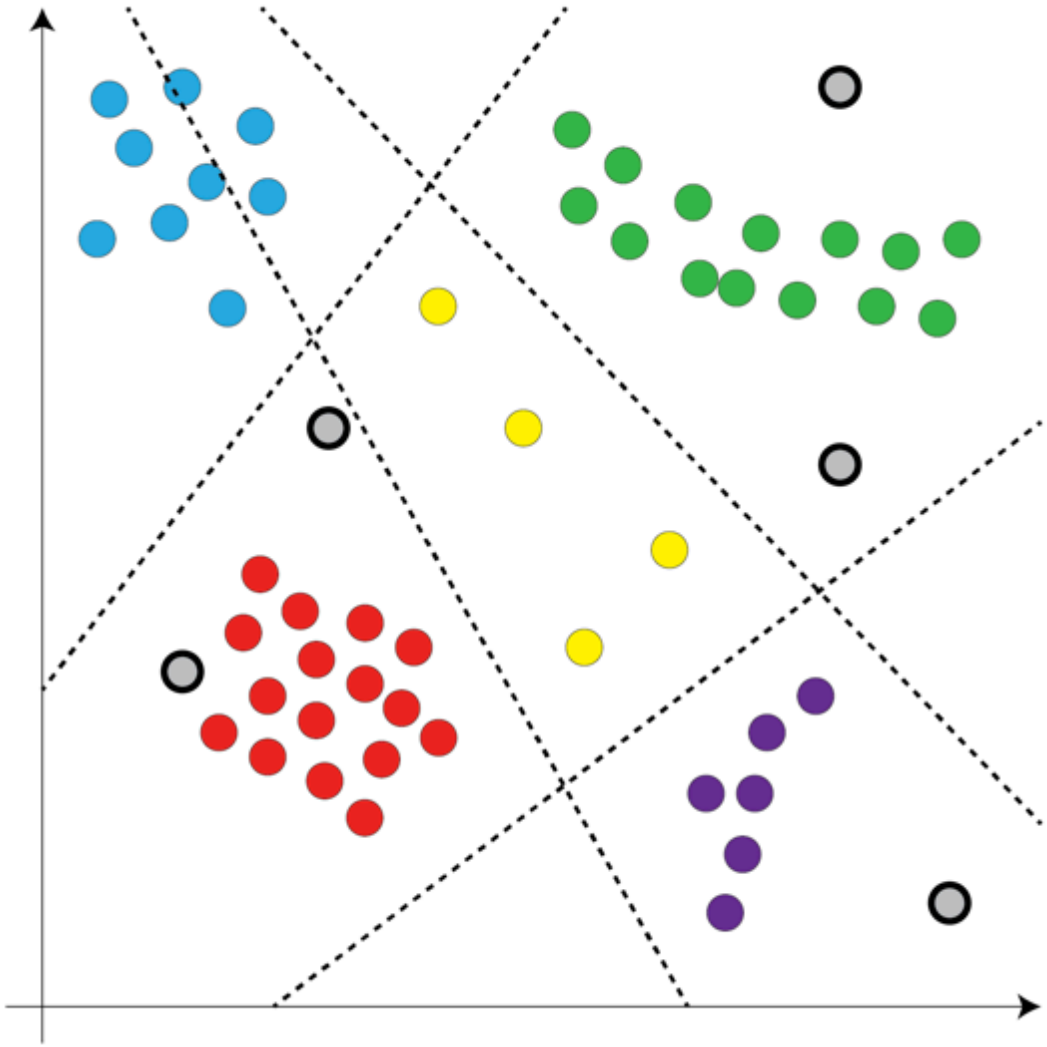
Classification



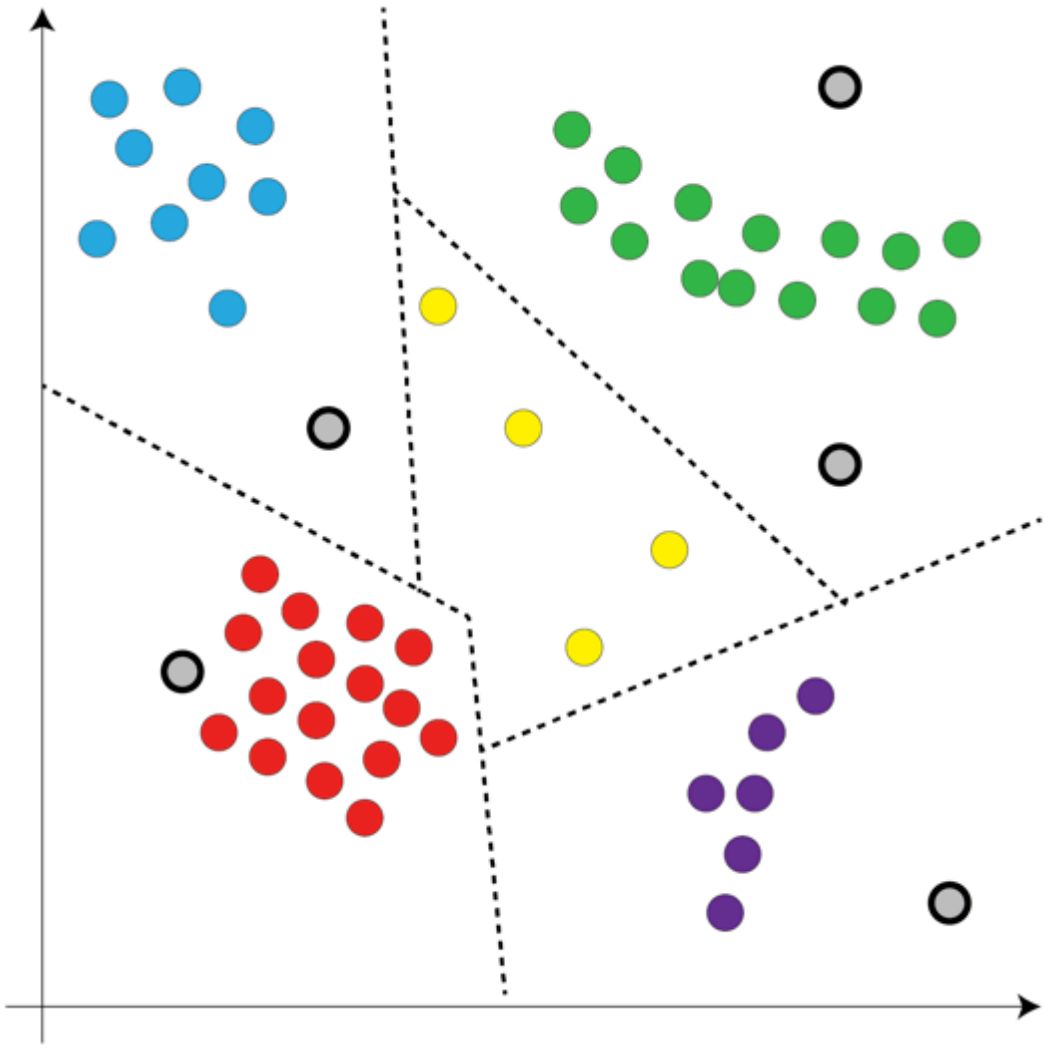
Classification



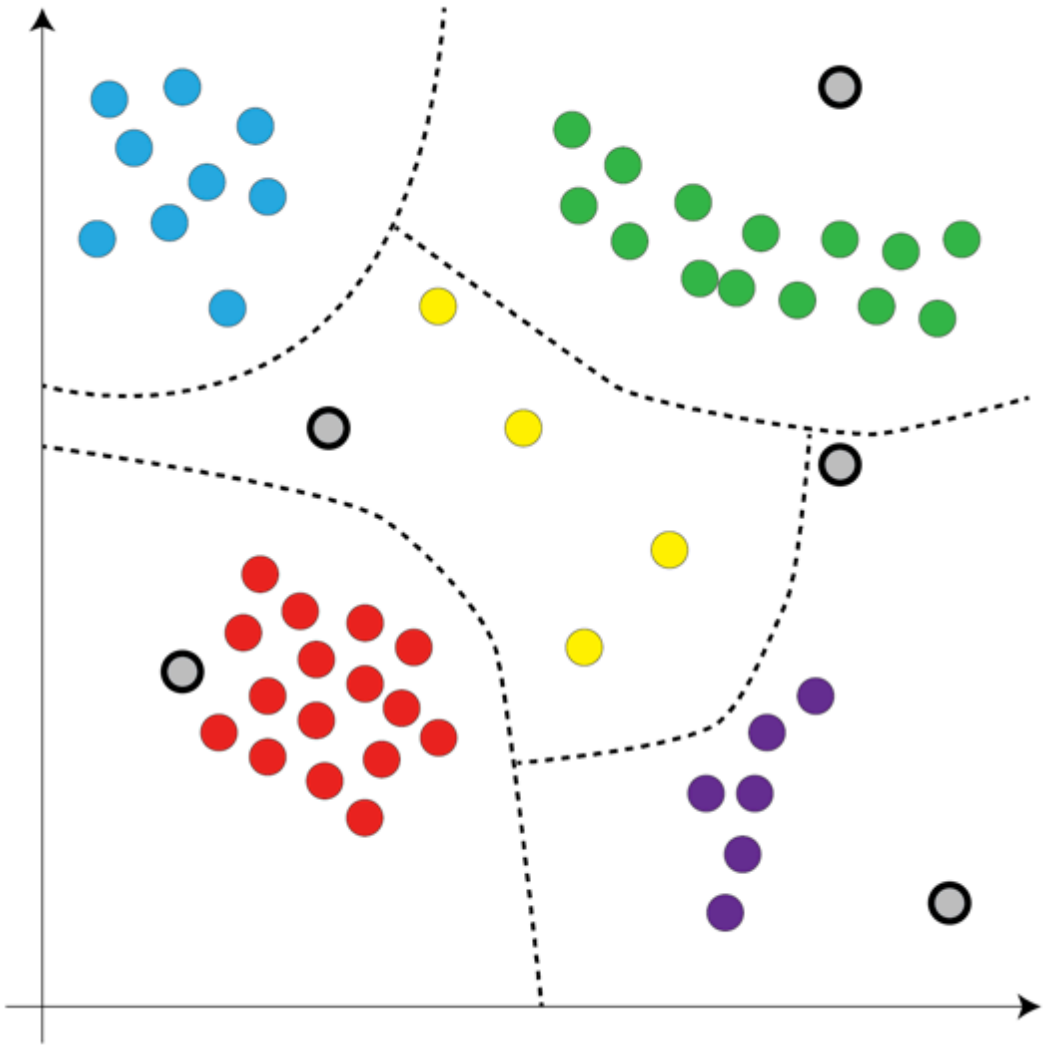
Classification



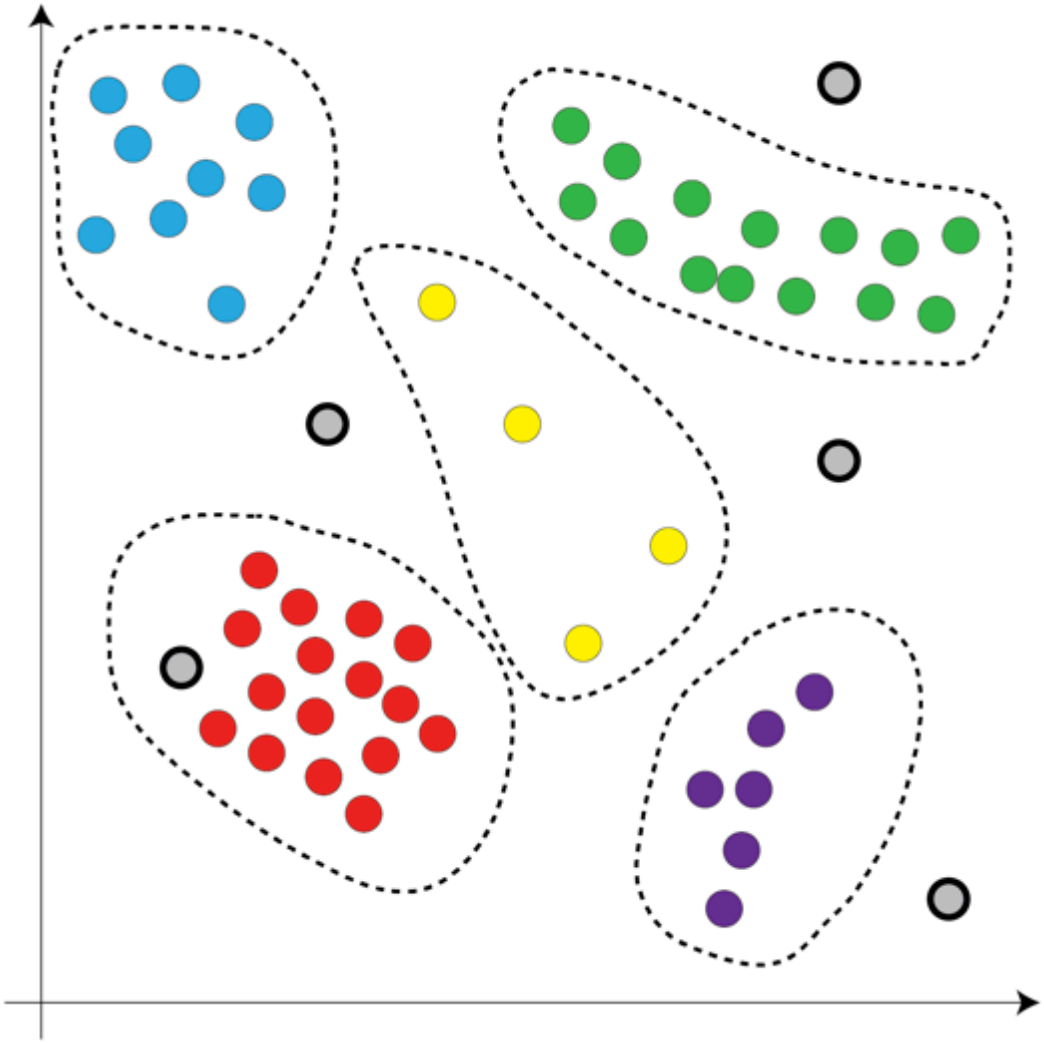
Classification



Classification



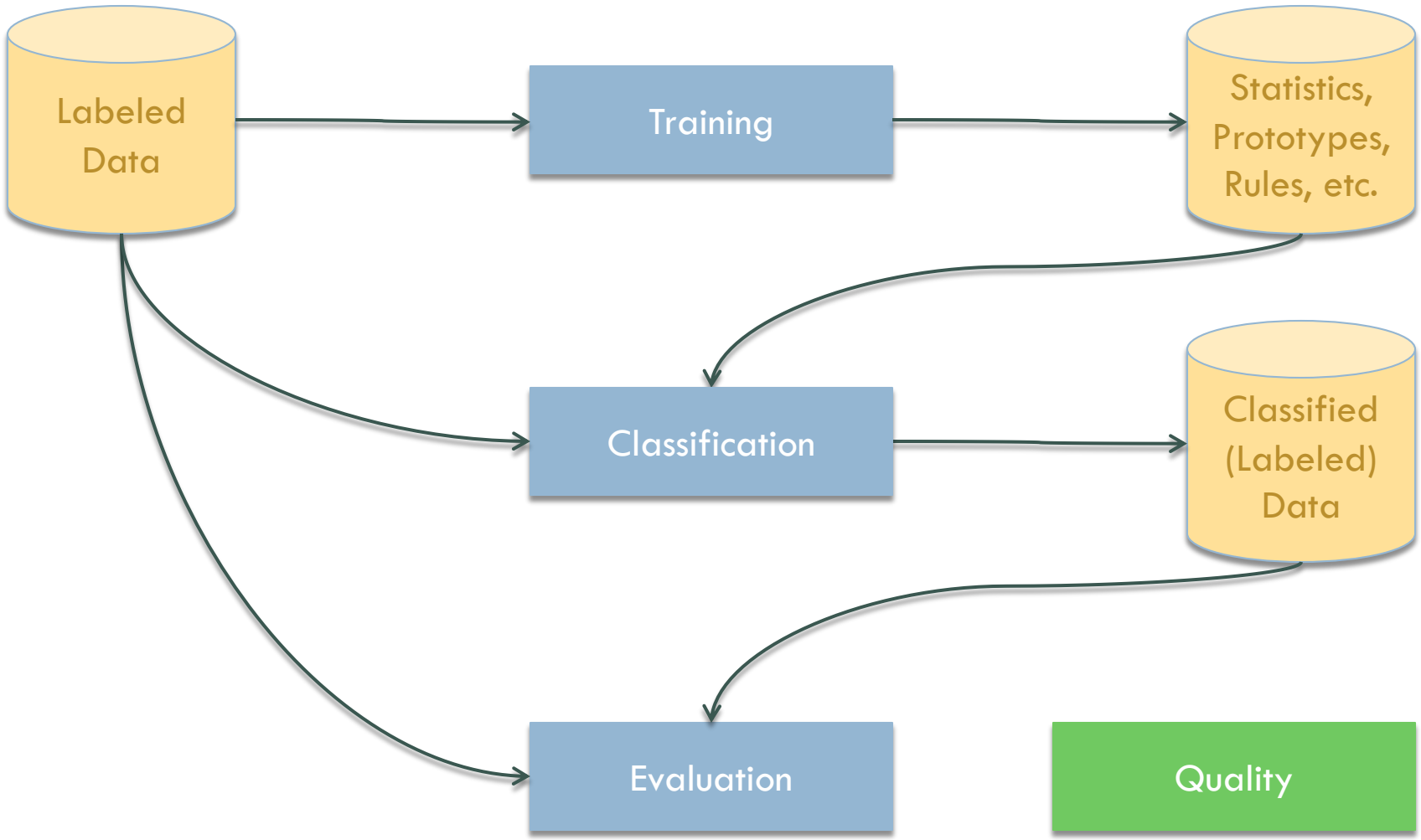
Classification



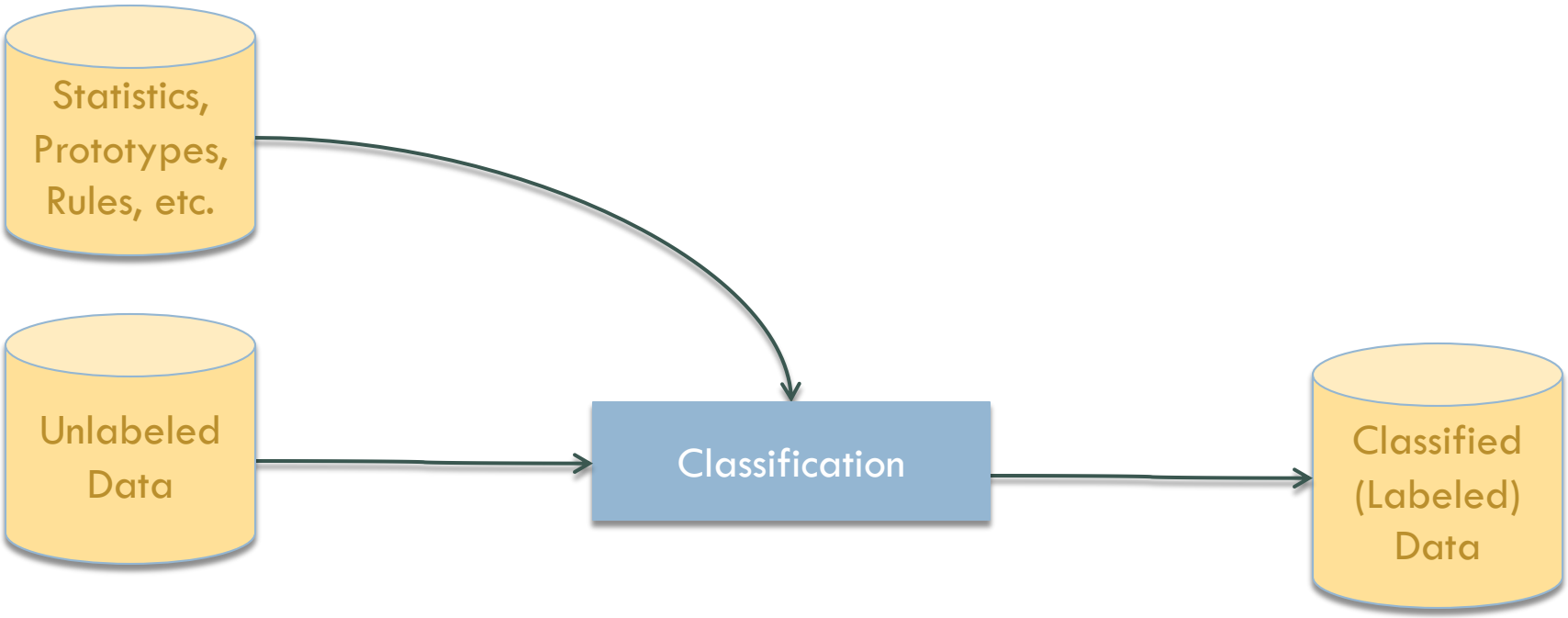
How to create a model?

- Different algorithms creates different models.
- Some models are inherently more precise, some are easier to understand.
- Some models are compact, some are extensive.
- Which is better?

Classification



Classification



Classification: Evaluation

- A simple evaluation technique: confusion matrix.
  - Classify labeled or known data.
  - Usually data used for training or a subset of it

Was classified as

Original Label is

	A	B	C
A	50	0	0
B	20	155	25
C	10	0	190

- Accuracy: Correct Classifications/All Classifications
  - 395/450 = 87.78%

More Metrics from Confusion Matrix

- Recall for a class (sensitivity or true positive rate)
  - Of the classified as ***X*** how many are really ***X*** (i.e. not other classes in X's boundary)?
  - TP/(TP+FN): 0.6250 for A; 1.0000 for B; 0.8837 for C
- Precision for a class (positive predictive value)
  - Of all the ***X*** how many were classified as ***X*** (i.e. not misclassified)?
  - TP/(TP+FP): 1.0000 for A; 0.7750 for B; 0.9500 for C

- True Positives
- True Negatives
- False Positives
- False Negatives

- For ***B***:

Classified as

Original

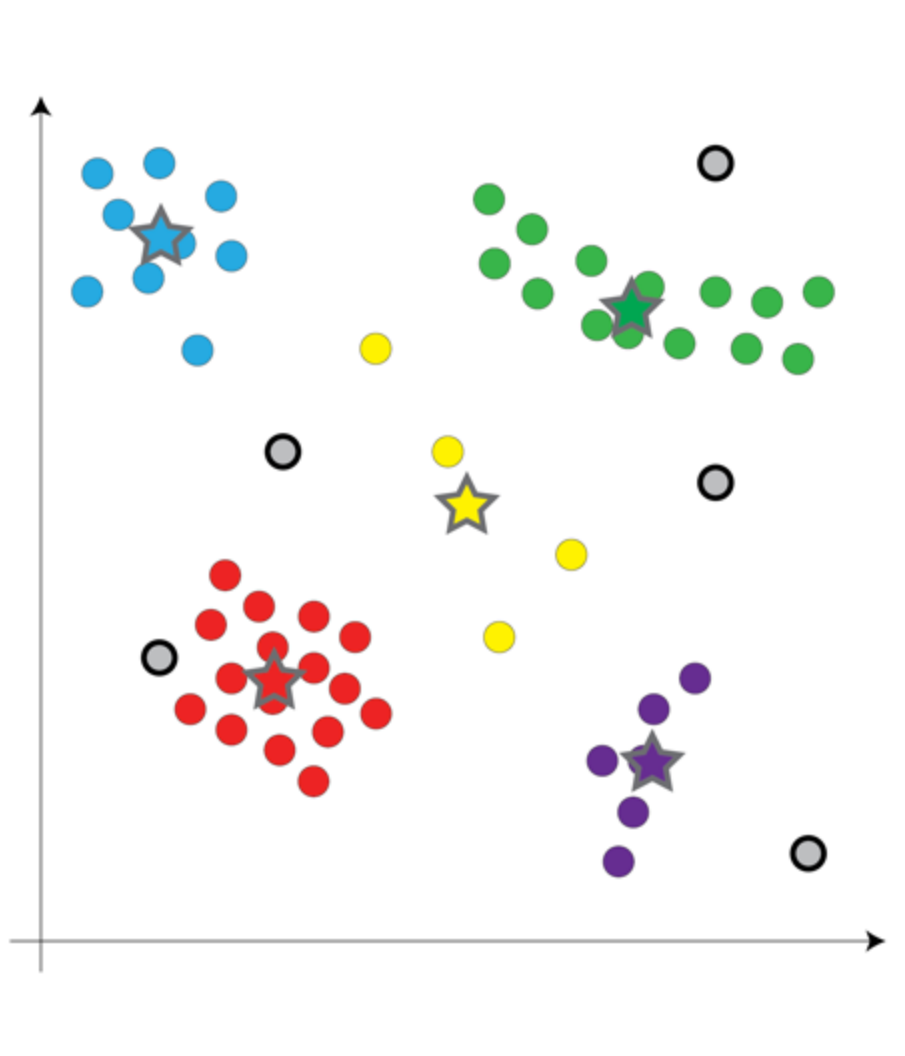
	A	**B**	C
A	50	0	0
B	20	155	25
C	10	0	190

Ideas from the evaluation process

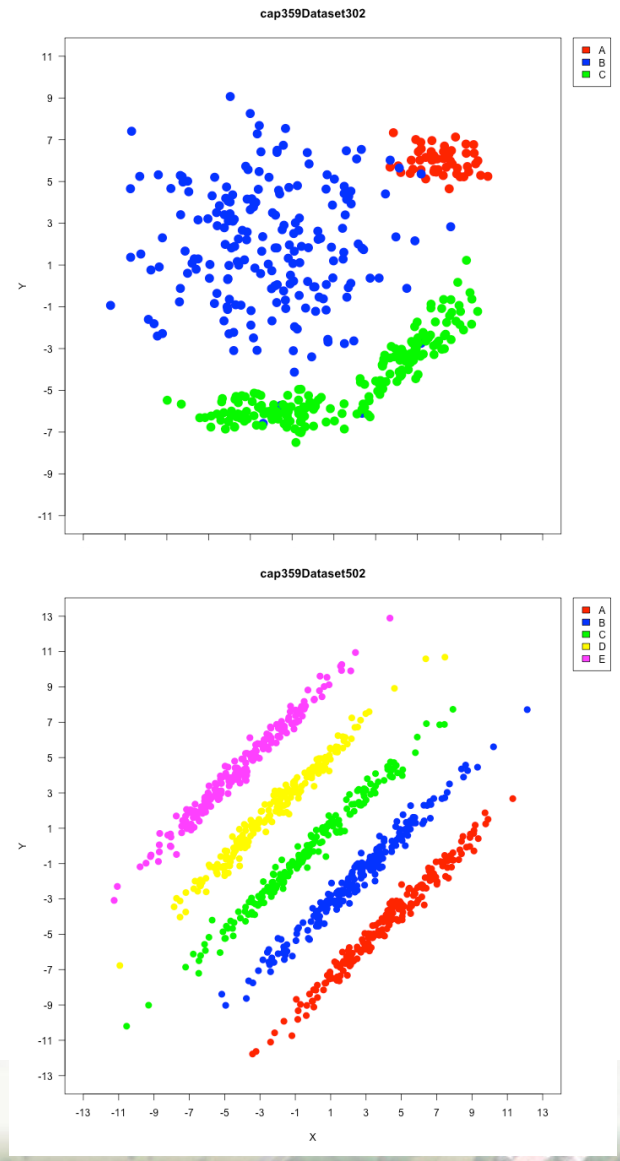
- Labeled data: does it *really* corresponds to samples for a class?
- Are there mixed classes in our labels for class X?
- Are there really N classes (instead of N±n)?

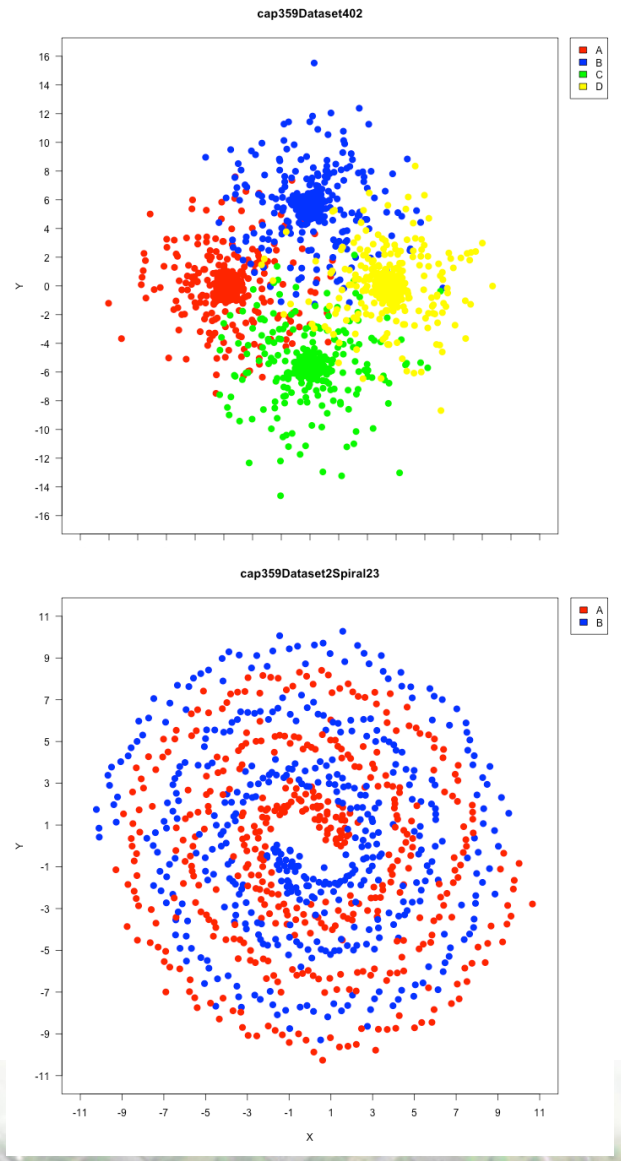
Classification: Minimum Distance

- Model is the average of the data points (geometric center).
- Class is determined from the minimum distance to center.
- Other metrics may be used.



Classification: Minimum Distance



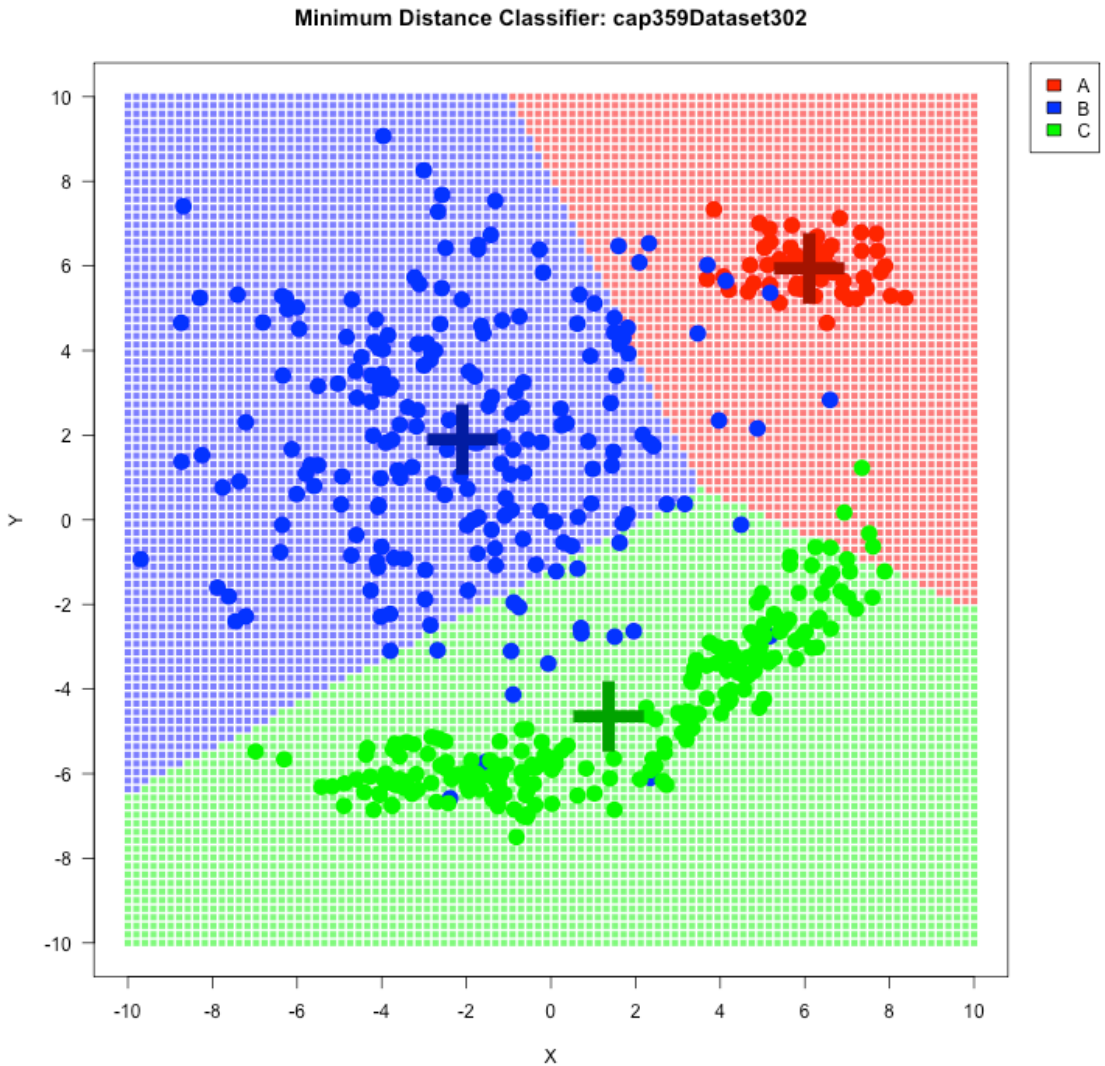


Classification: Minimum Distance

- Decision Boundary with labeled points and classes' prototypes

	Classified as		
	A	B	C
A	50	0	0
B	11	170	19
C	4	0	196

Accuracy: 92.444%

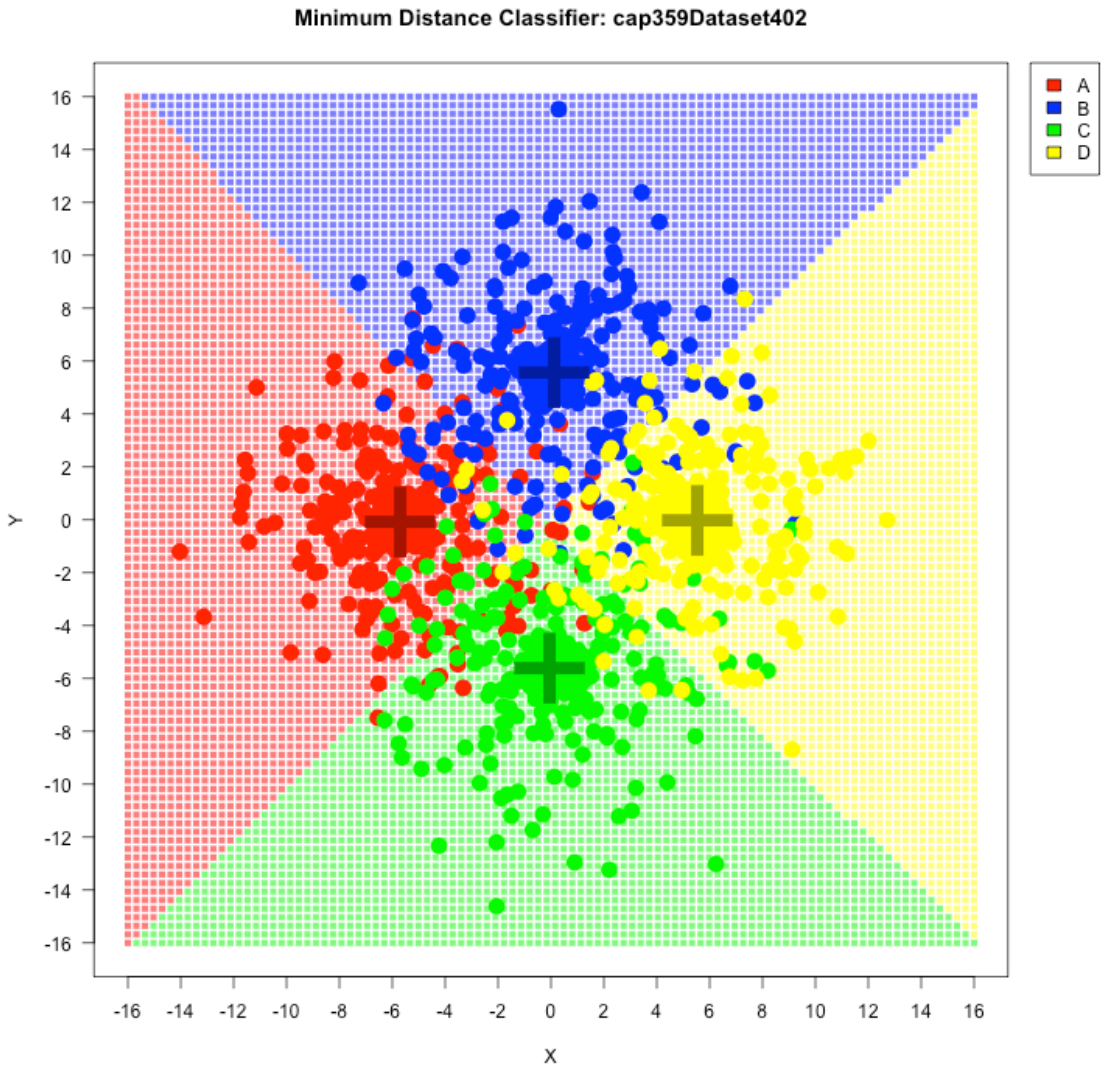


Classification: Minimum Distance

- Decision Boundary with labeled points and classes' prototypes

	Classified as			
	A	B	C	D
A	356	16	26	2
B	19	357	1	23
C	19	0	362	19
D	4	14	15	367

Accuracy: 90.125%

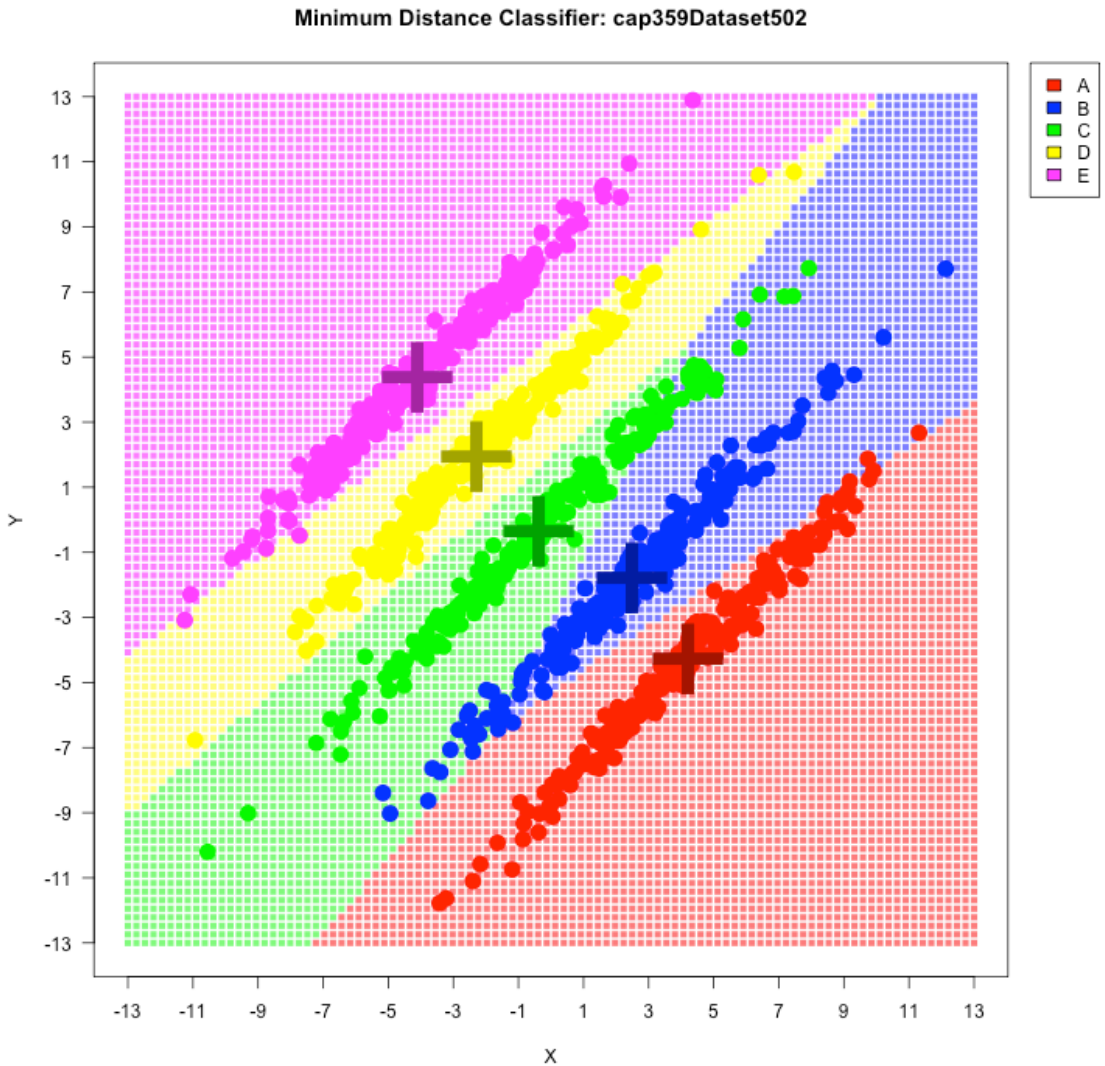


Classification: Minimum Distance

- Decision Boundary with labeled points and classes' prototypes

	Classified as				
	A	B	C	D	E
A	195	5	0	0	0
B	1	171	28	0	0
C	0	29	171	0	0
D	0	0	0	198	2
E	0	0	0	2	198

Accuracy: 93.300%

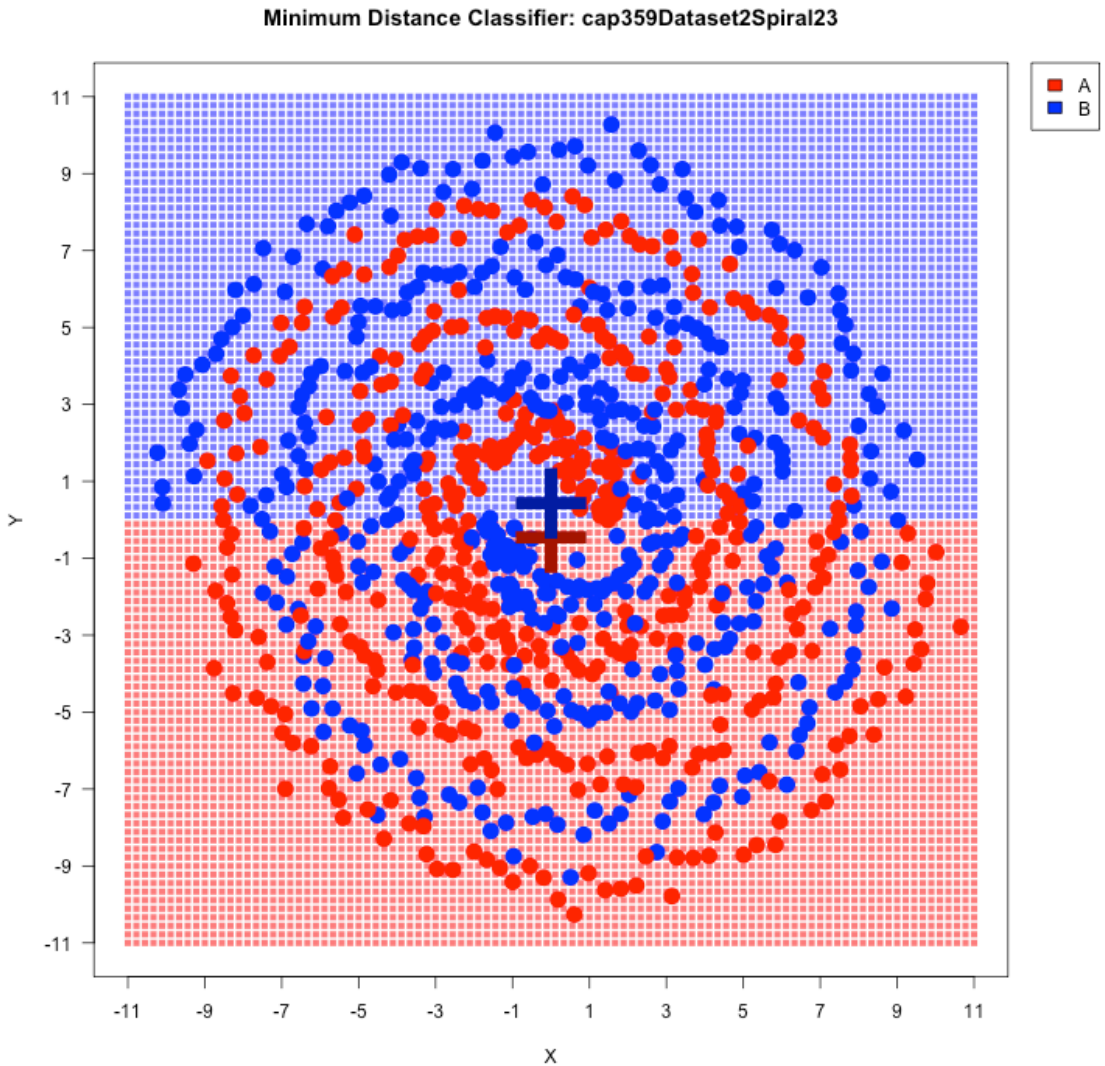


Classification: Minimum Distance

- Decision Boundary with labeled points and classes' prototypes

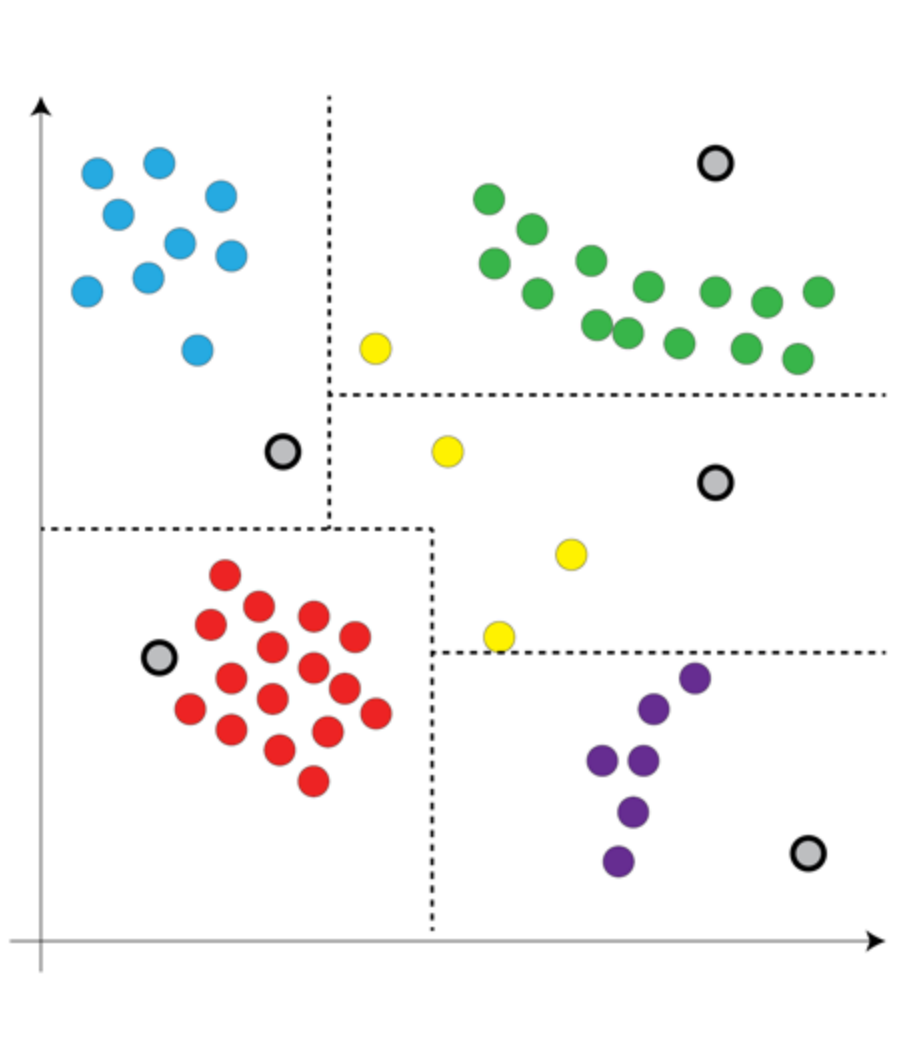
	Classified as	
	A	B
A	227	223
B	224	226

Accuracy: 50.333%

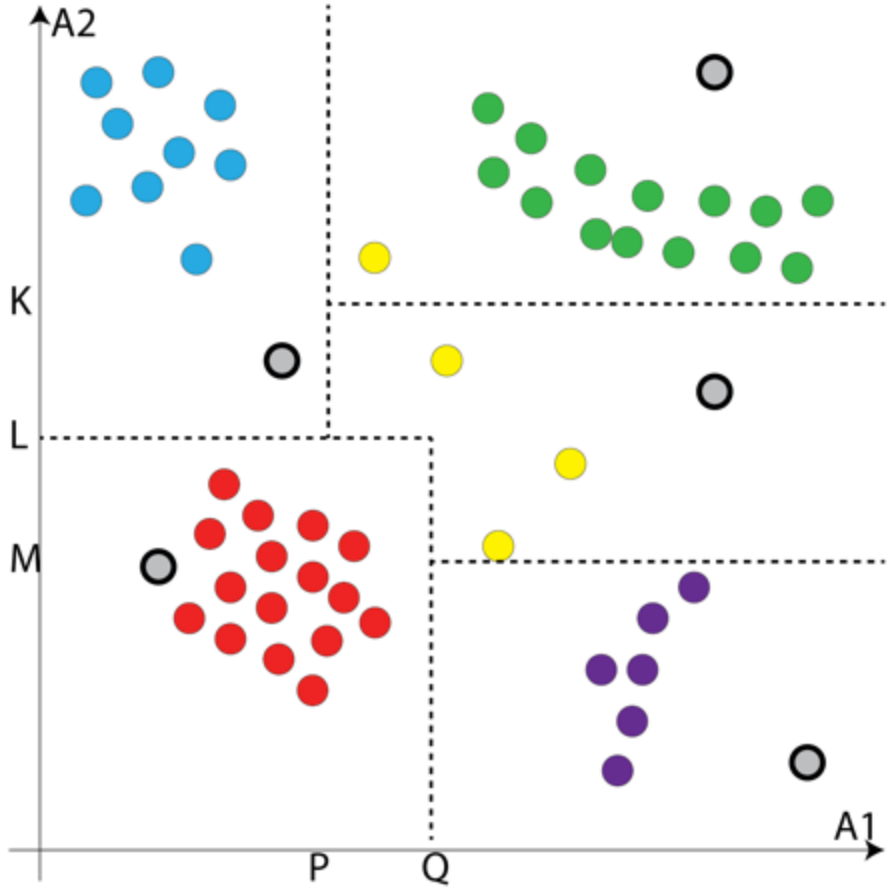


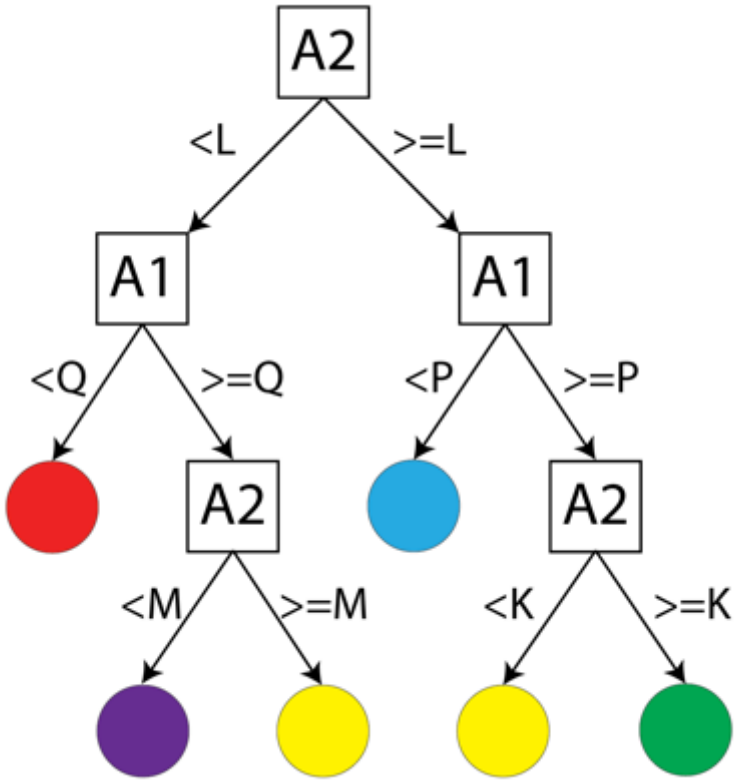
Classification: Decision Tree

- Model is the set of decision rules that best separates the classes.
- Class is determined from evaluation of the rules.

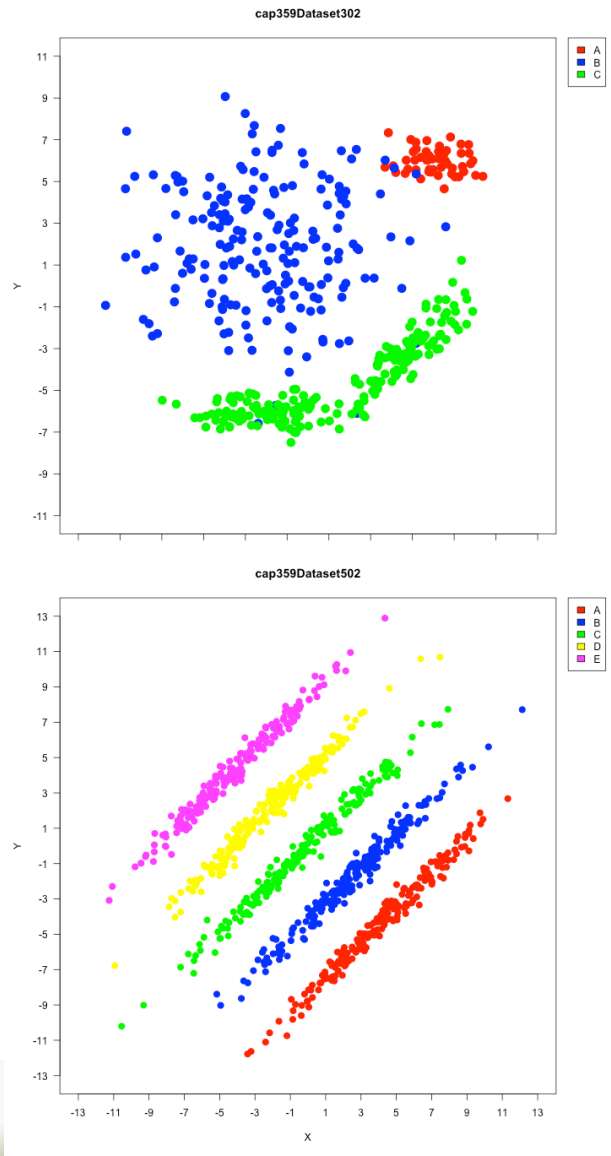


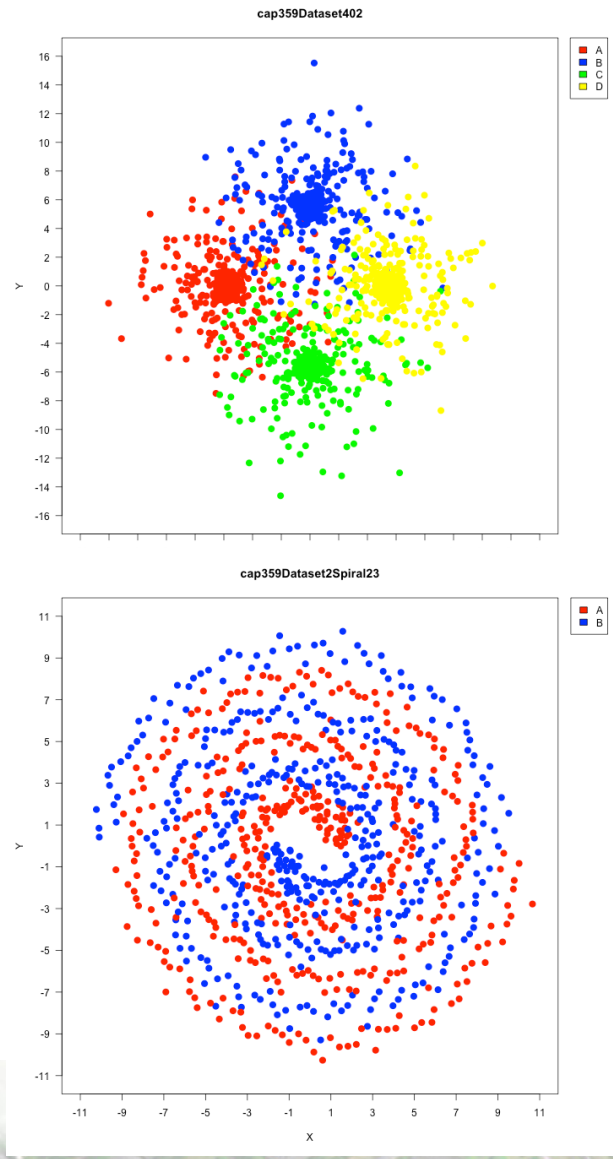
Classification: Decision Tree



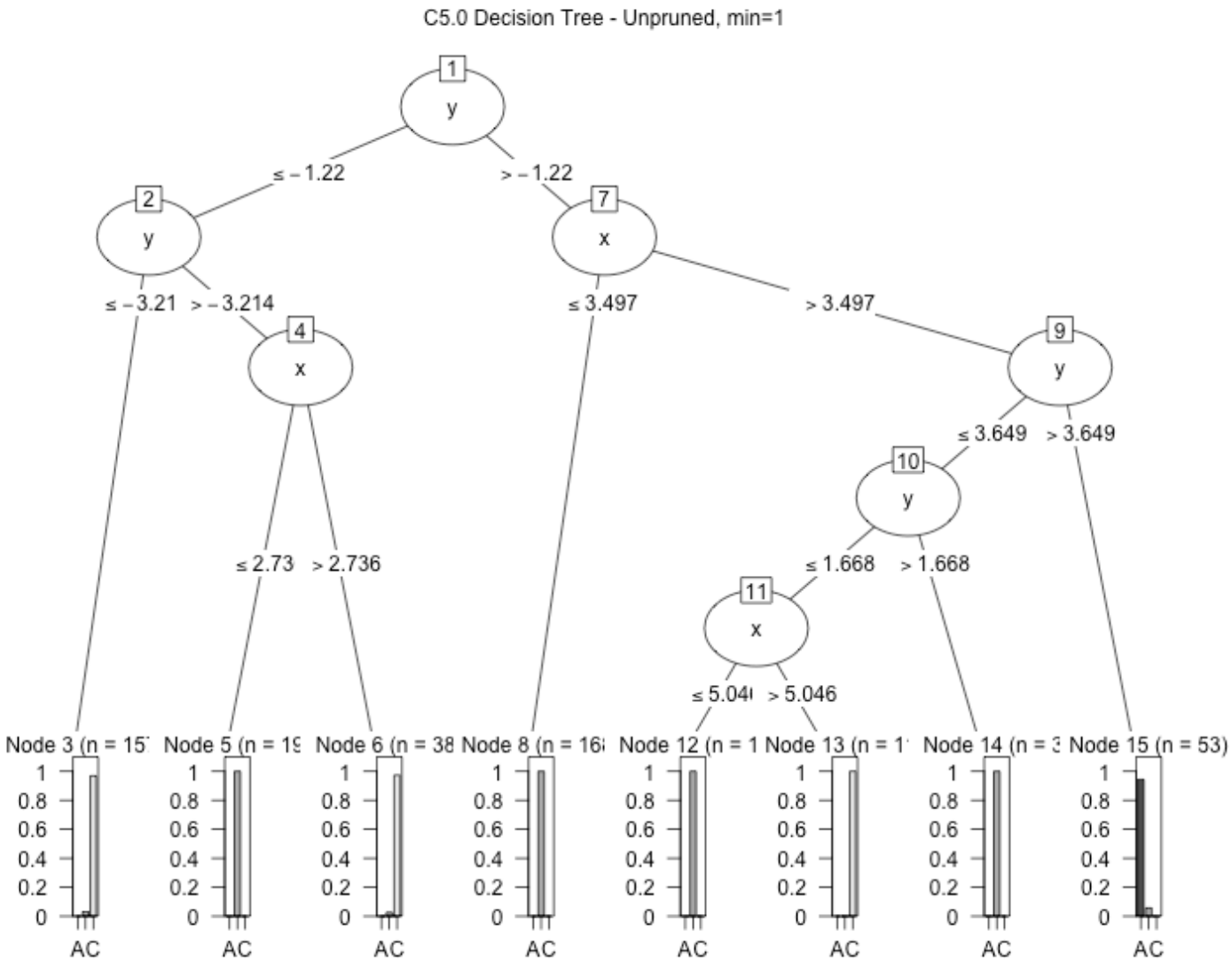


Classification: Decision Tree





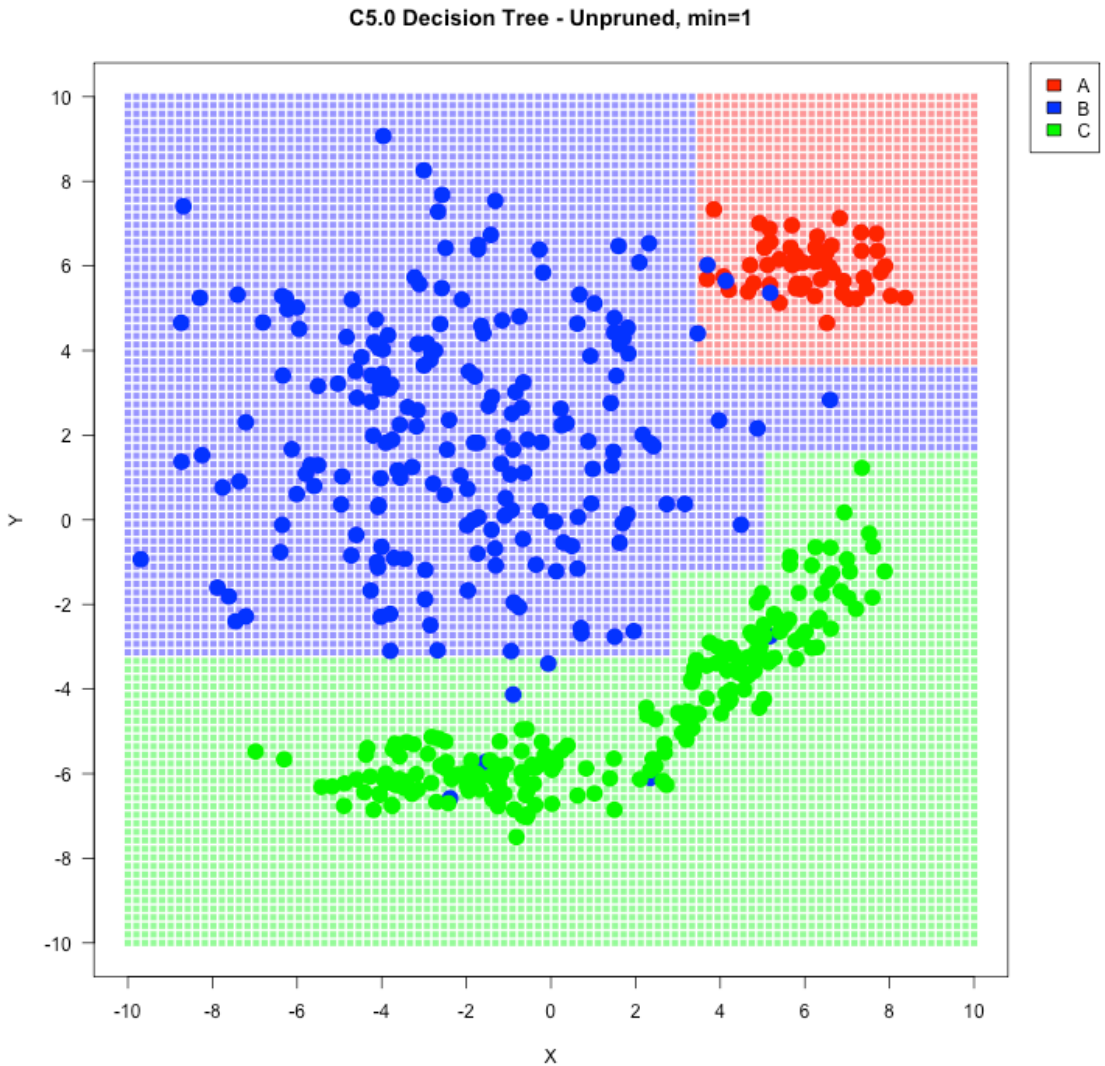
Classification: Decision Tree



Classification: Decision Tree

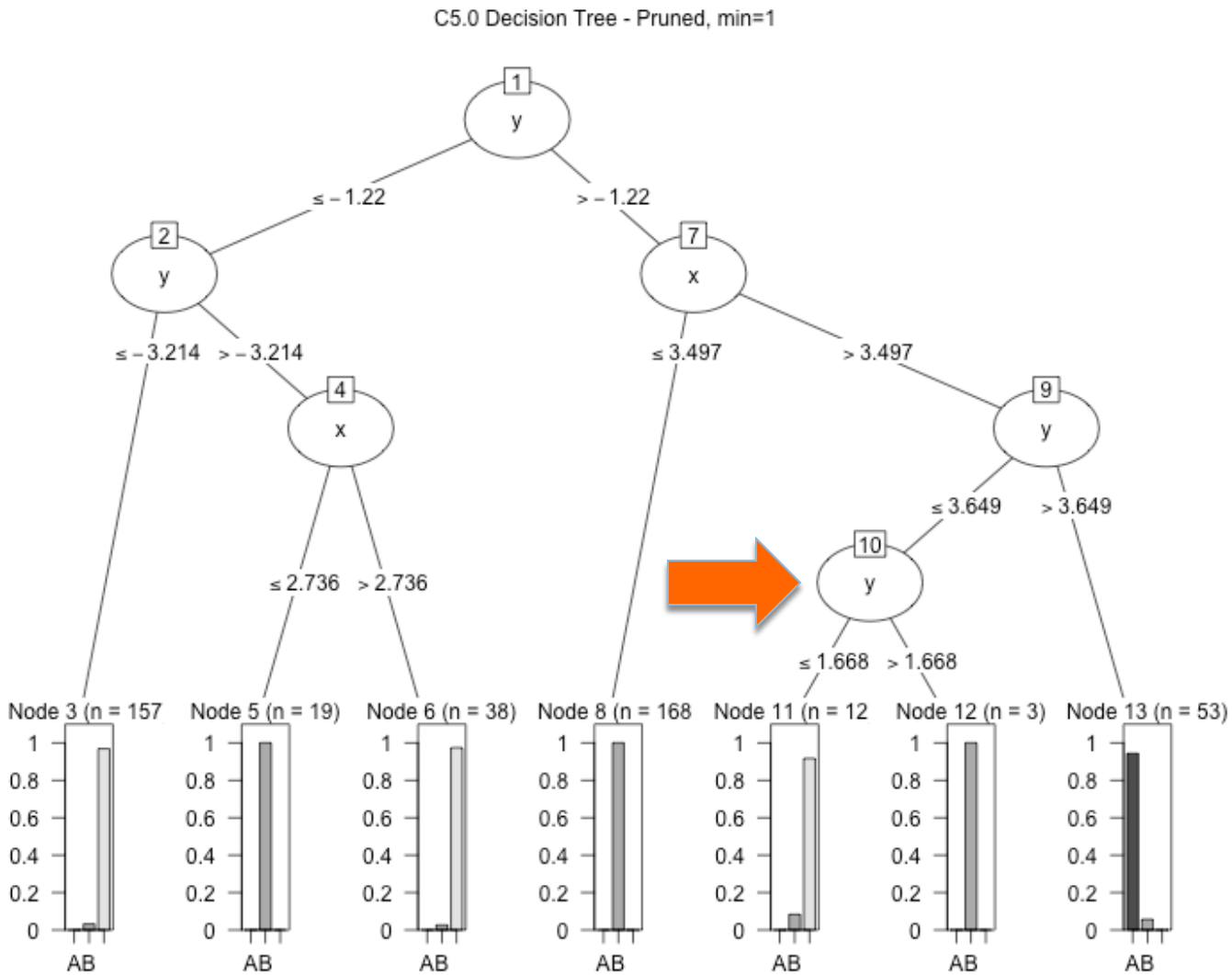
	Classified as		
	A	B	C
A	50	0	0
B	3	191	6
C	0	0	200

Accuracy: 0.98



Classification: Decision Tree

- A pruned tree.

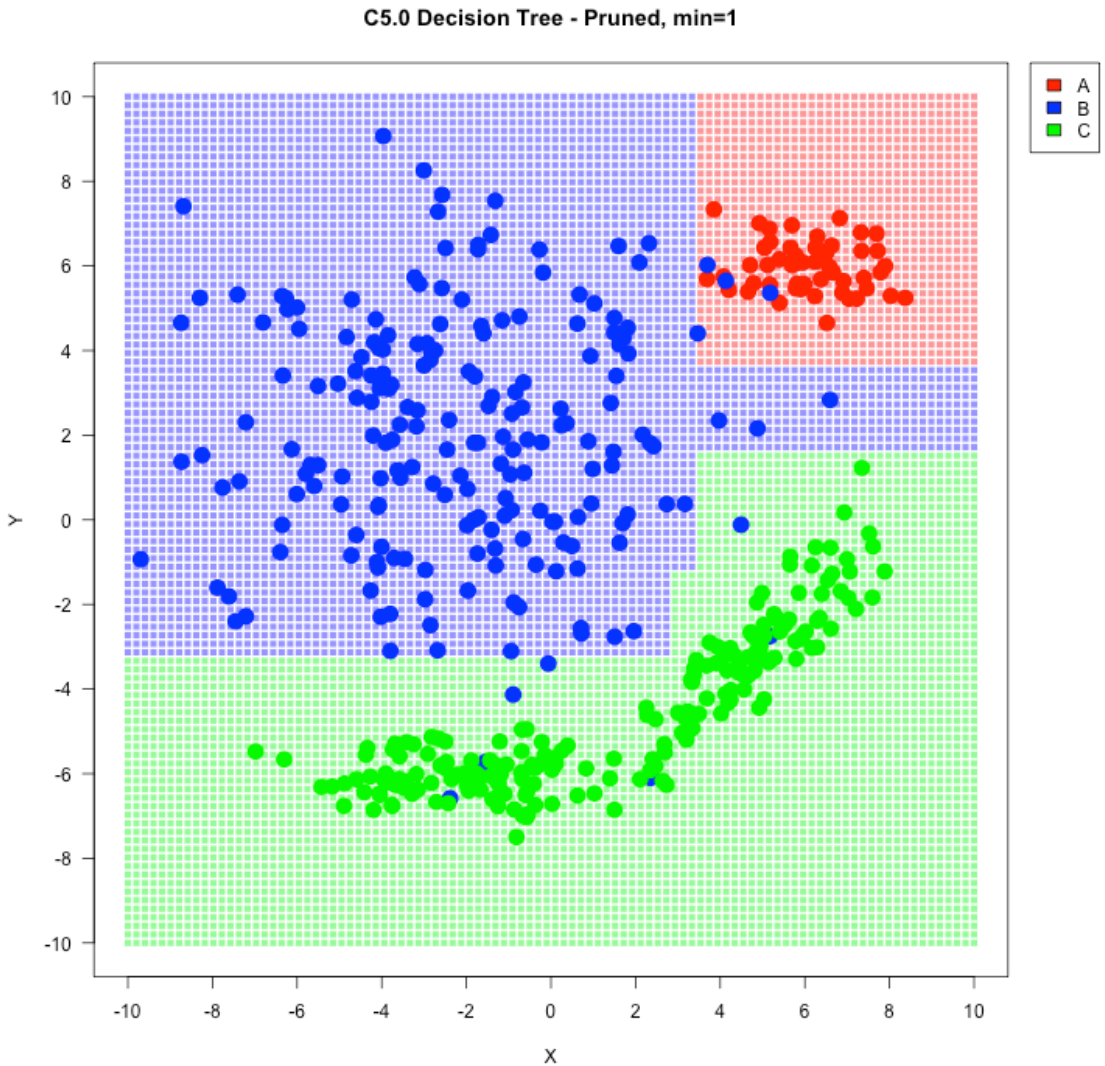


Classification: Decision Tree

- Pruned tree.

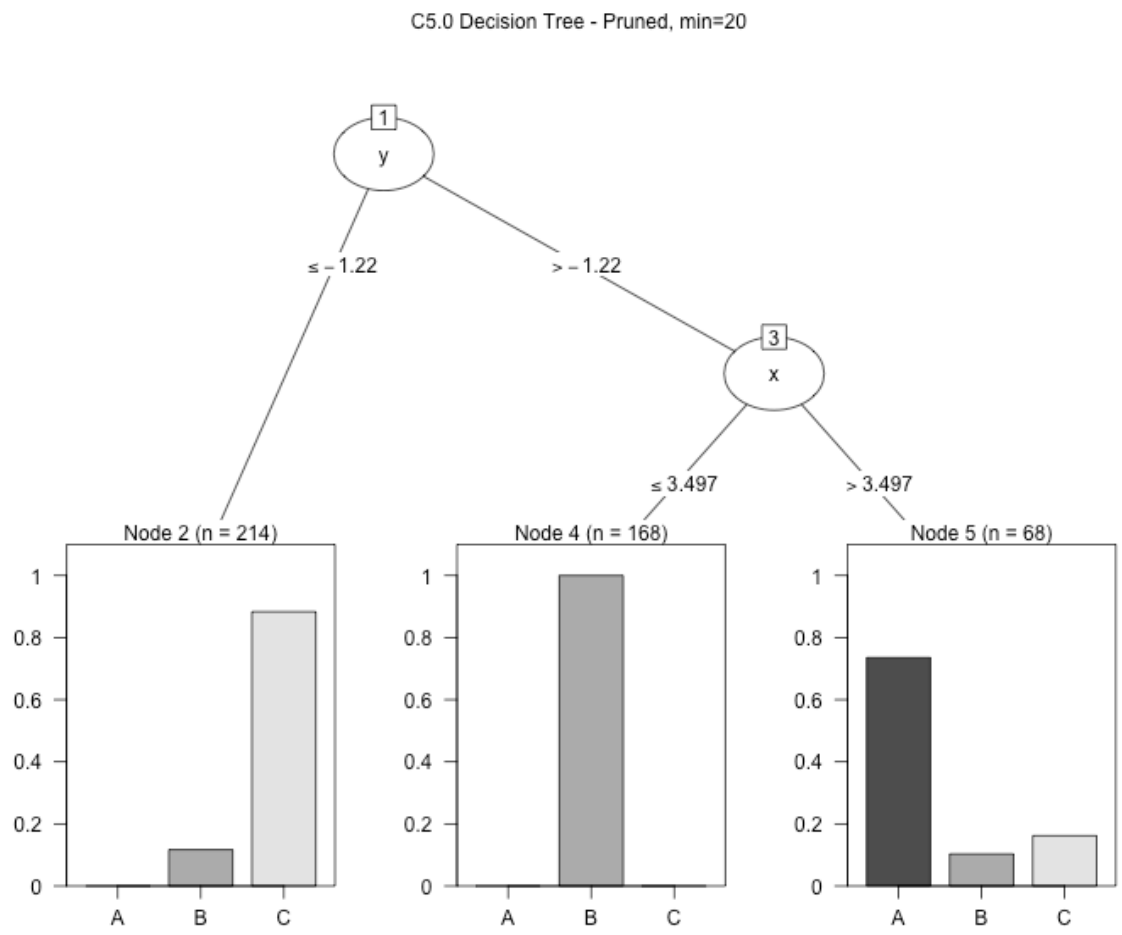
	Classified as		
	A	B	C
A	50	0	0
B	3	190	7
C	0	0	200

Accuracy: 0.9777778



Classification: Decision Tree

- A pruned tree with minimum leaf size.

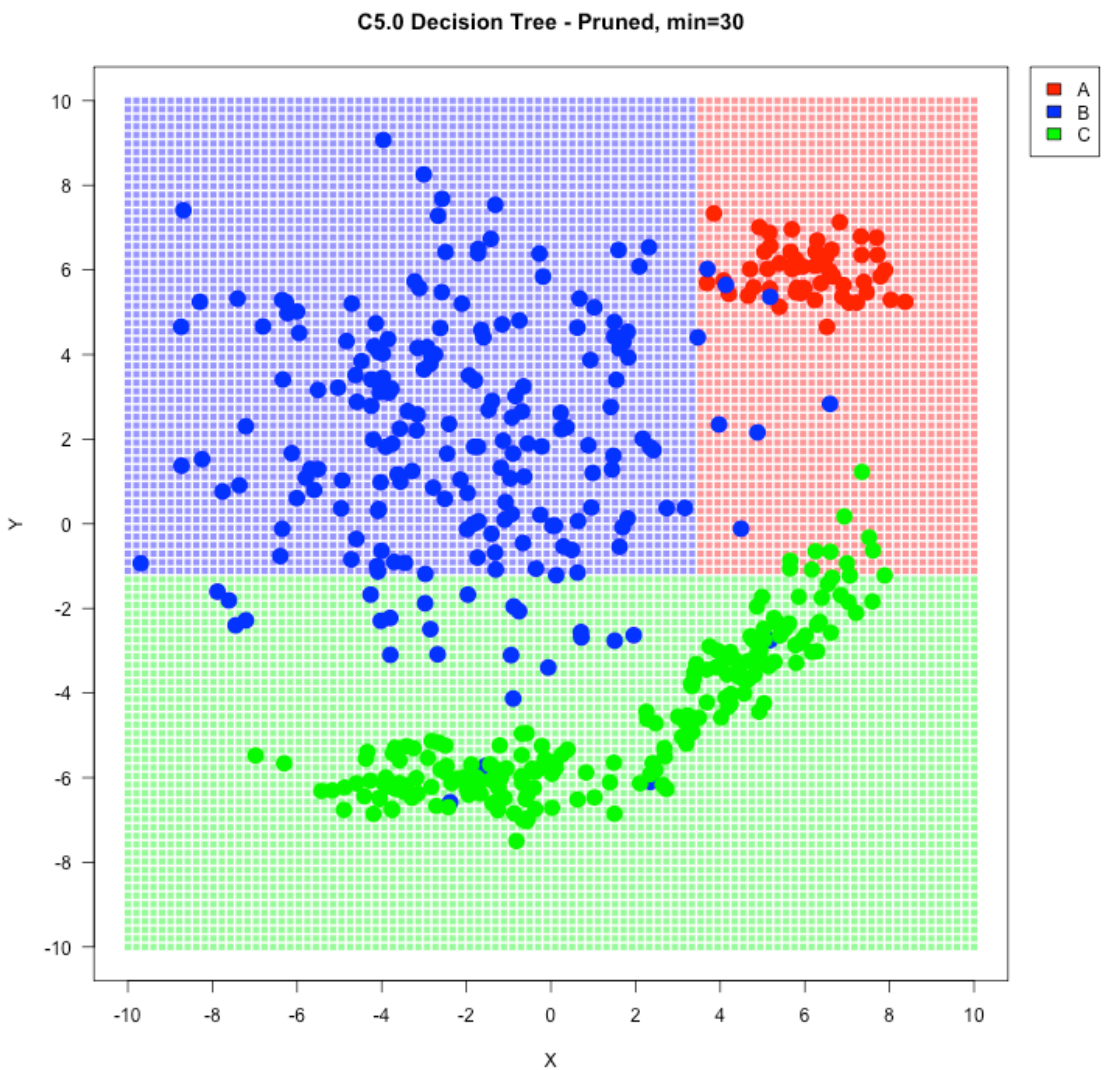


Classification: Decision Tree

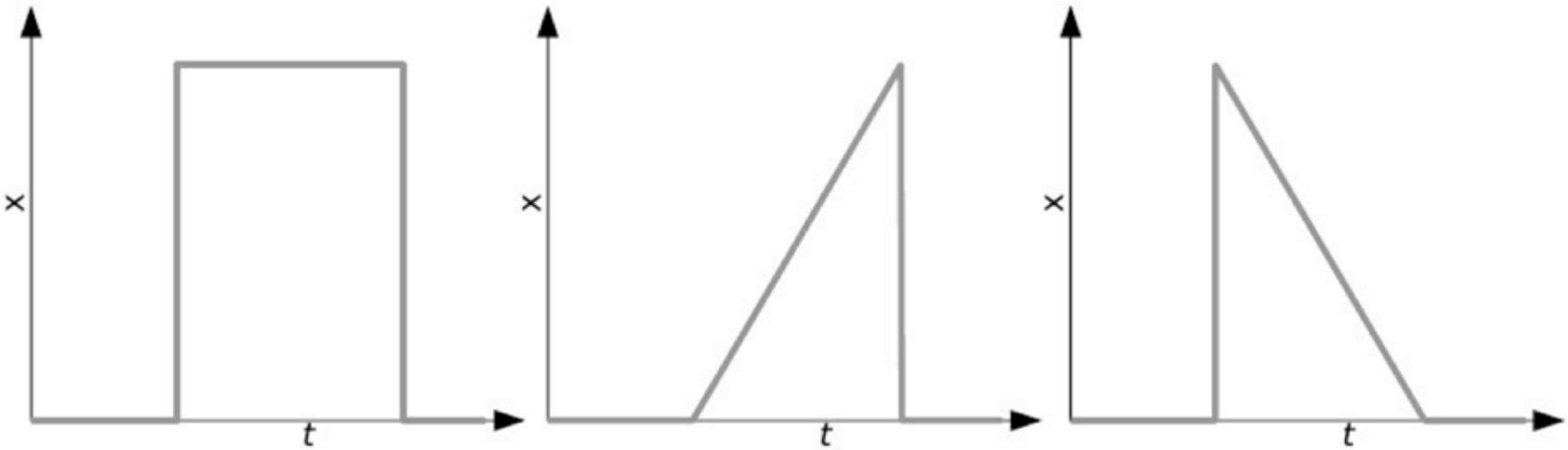
- A pruned tree with minimum leaf size.

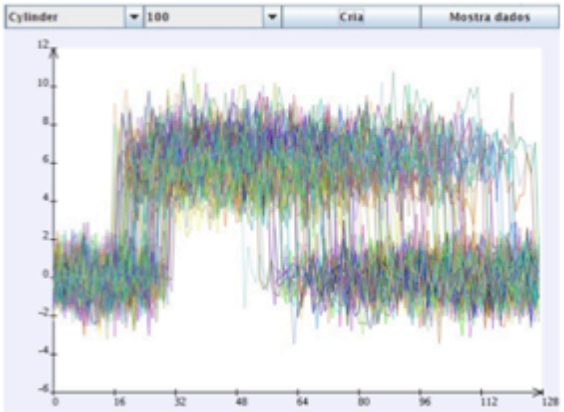
	Classified as		
	A	B	C
A	50	0	0
B	7	168	25
C	10	0	190

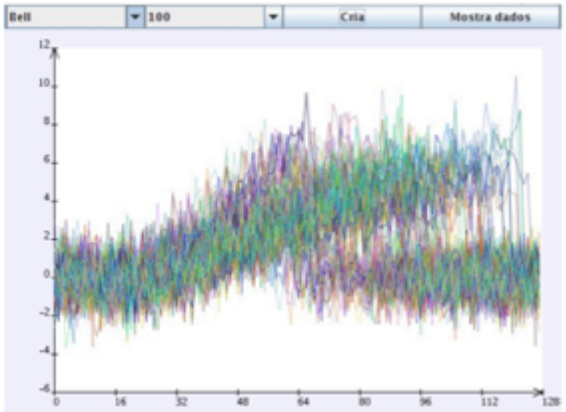
Accuracy: 0.9066667

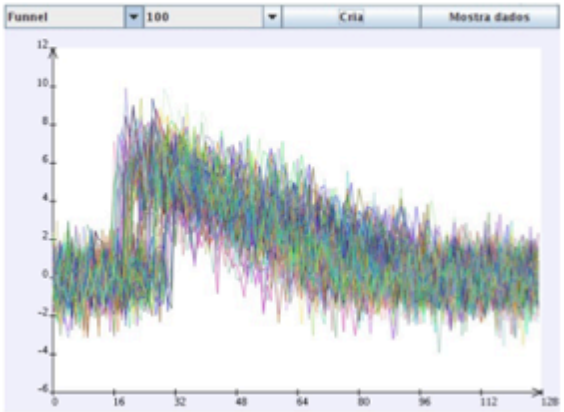


Classification: Decision Tree



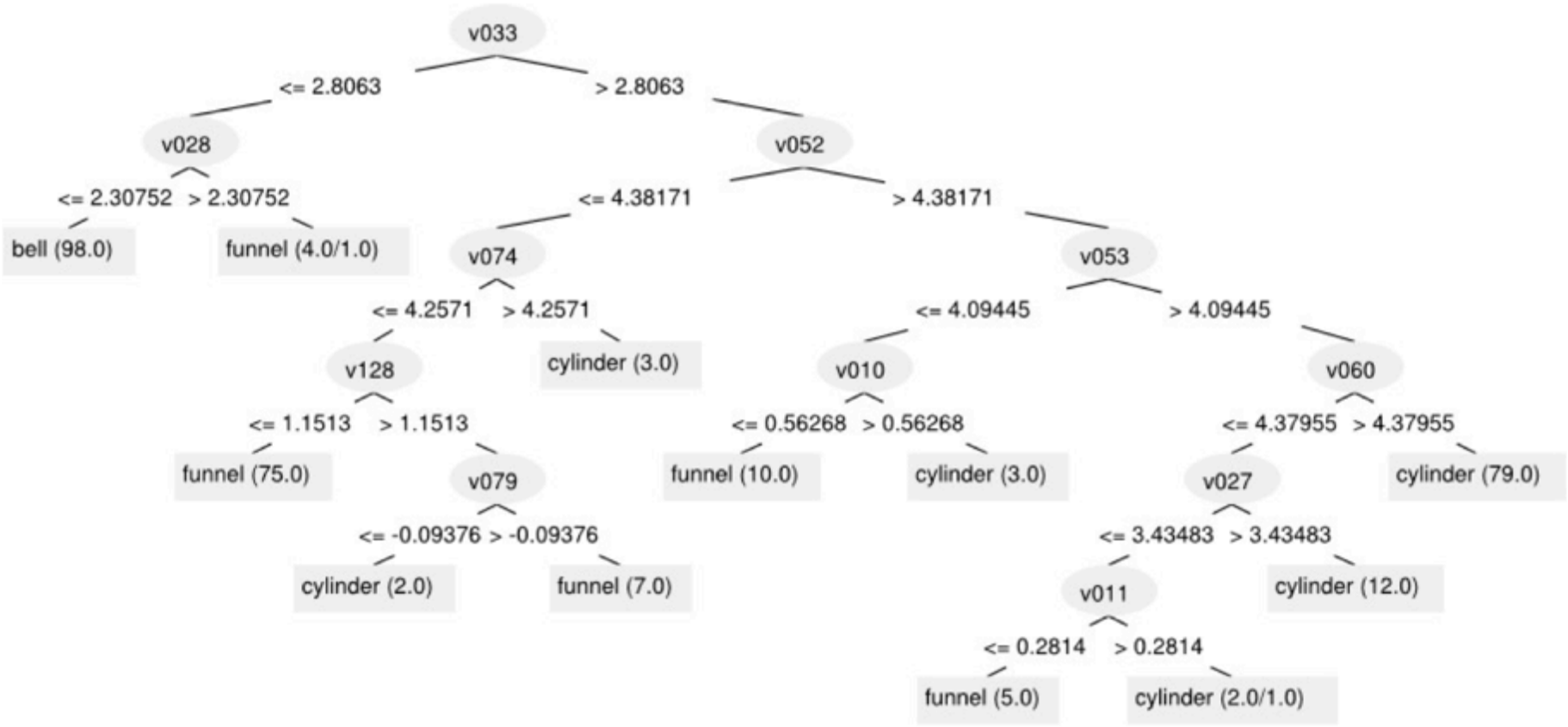






Must implement/document this in R! See next version.

Classification: Decision Tree

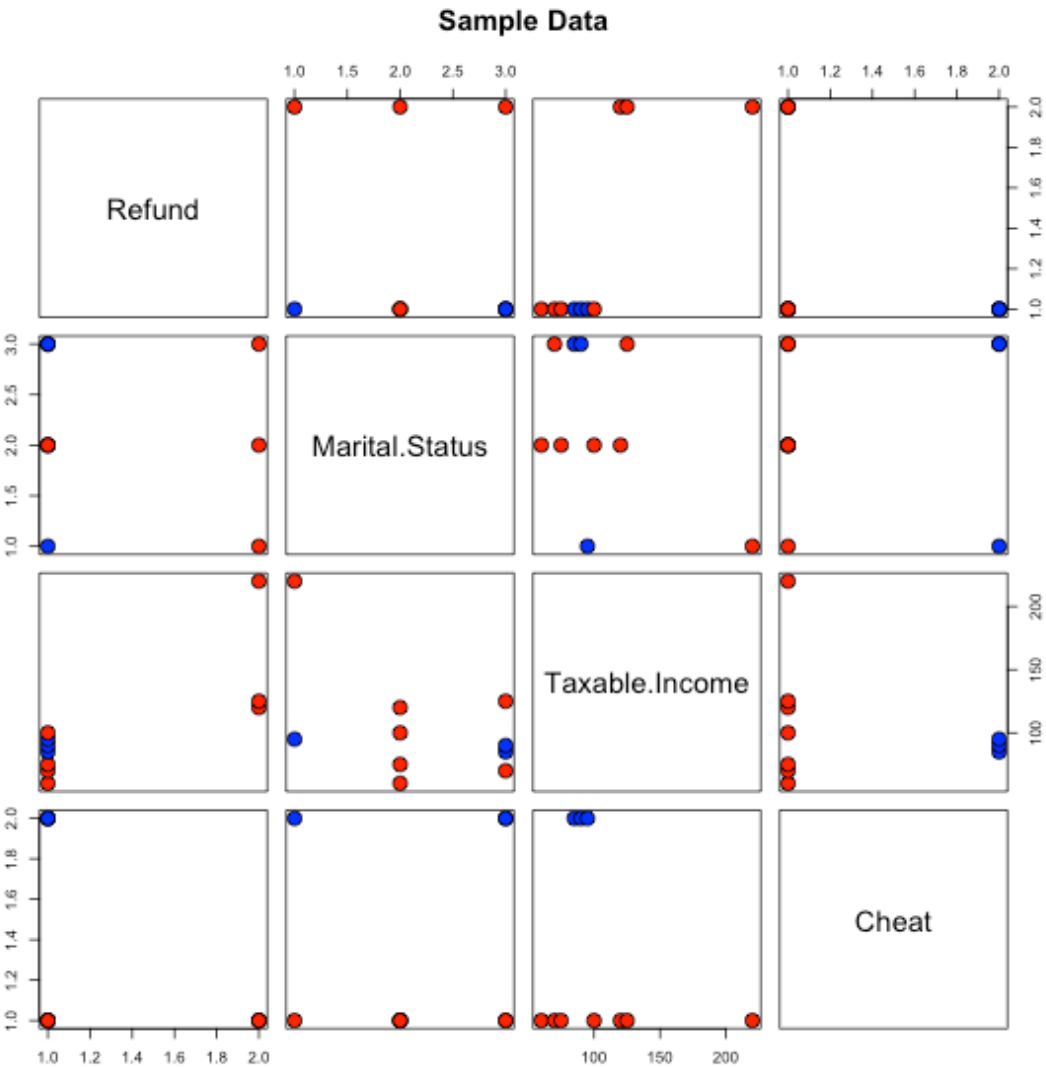


Must implement/document this in R! See next version.

Classification: Decision Tree

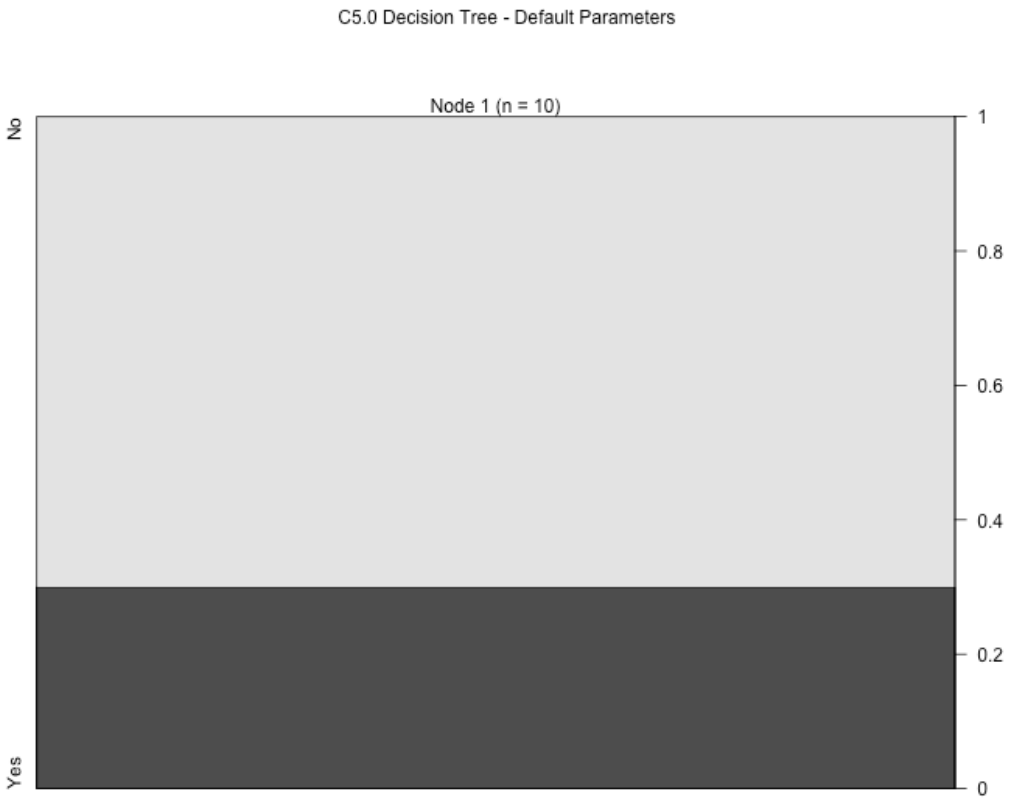
- Our toy problem:

Tid	Refund	Marital Status	Taxable Income	Cheat
6	No	Married	60	No
3	No	Single	70	No
9	No	Married	75	No
8	No	Single	85	Yes
10	No	Single	90	Yes
5	No	Divorced	95	Yes
2	No	Married	100	No
4	Yes	Married	120	No
1	Yes	Single	125	No
7	Yes	Divorced	220	No



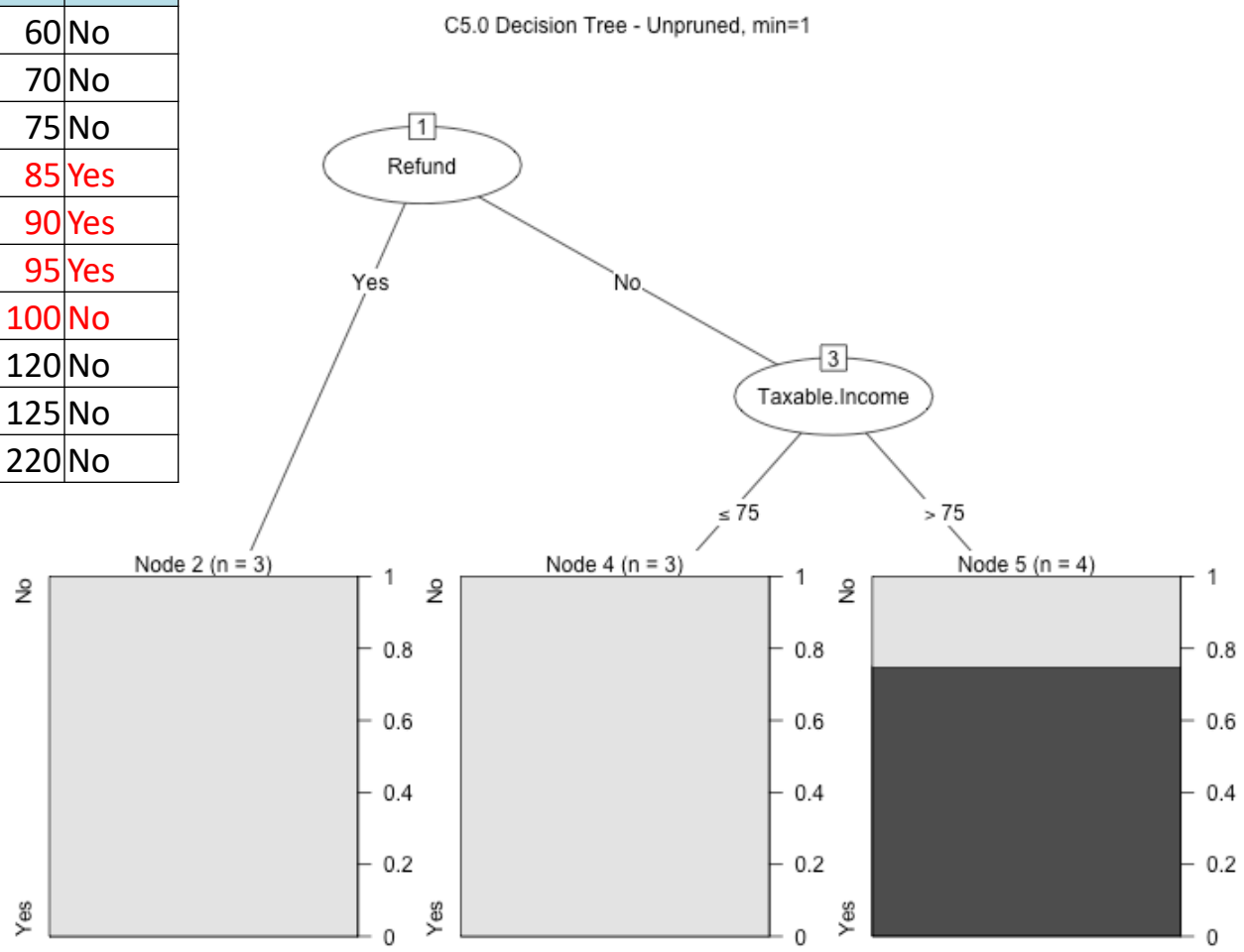
Classification: Decision Tree

Tid	Refund	Marital Status	Taxable Income	Cheat
6	No	Married	60	No
3	No	Single	70	No
9	No	Married	75	No
8	No	Single	85	Yes
10	No	Single	90	Yes
5	No	Divorced	95	Yes
2	No	Married	100	No
4	Yes	Married	120	No
1	Yes	Single	125	No
7	Yes	Divorced	220	No



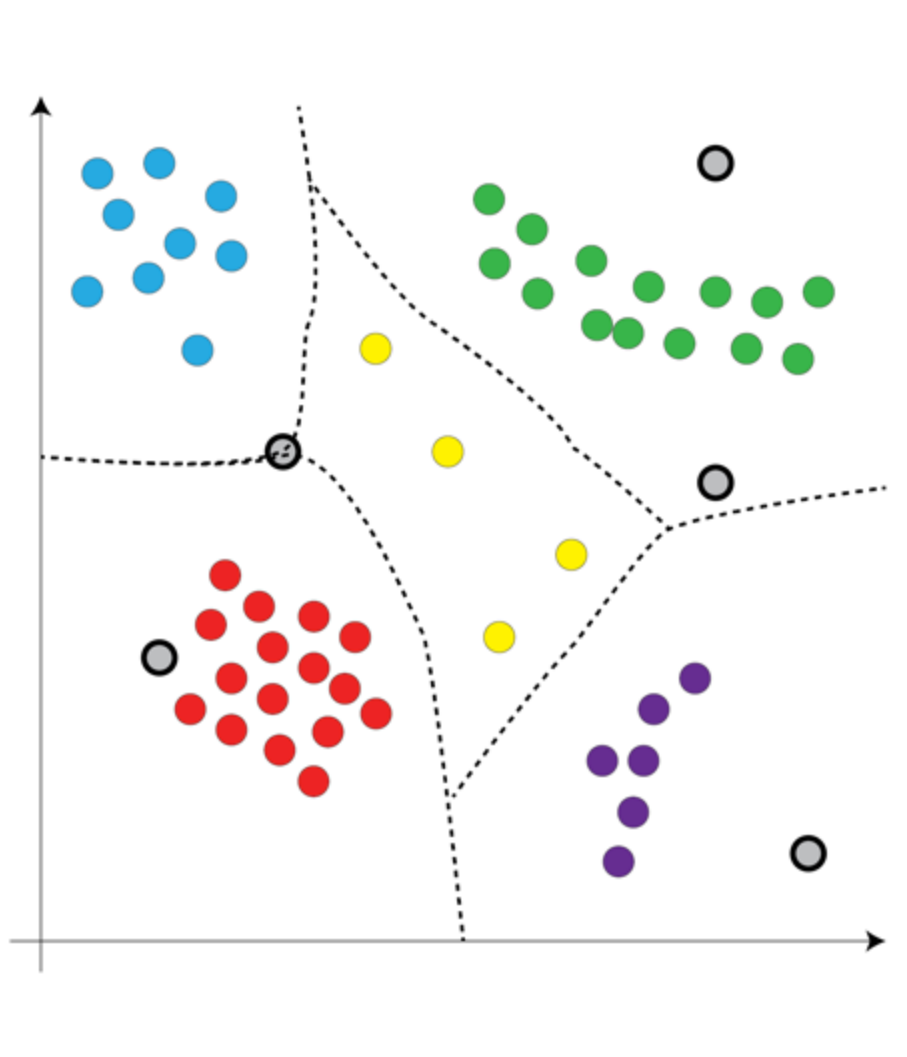
Classification: Decision Tree

Tid	Refund	Marital Status	Taxable Income	Cheat
6	No	Married	60	No
3	No	Single	70	No
9	No	Married	75	No
8	No	Single	85	Yes
10	No	Single	90	Yes
5	No	Divorced	95	Yes
2	No	Married	100	No
4	Yes	Married	120	No
1	Yes	Single	125	No
7	Yes	Divorced	220	No

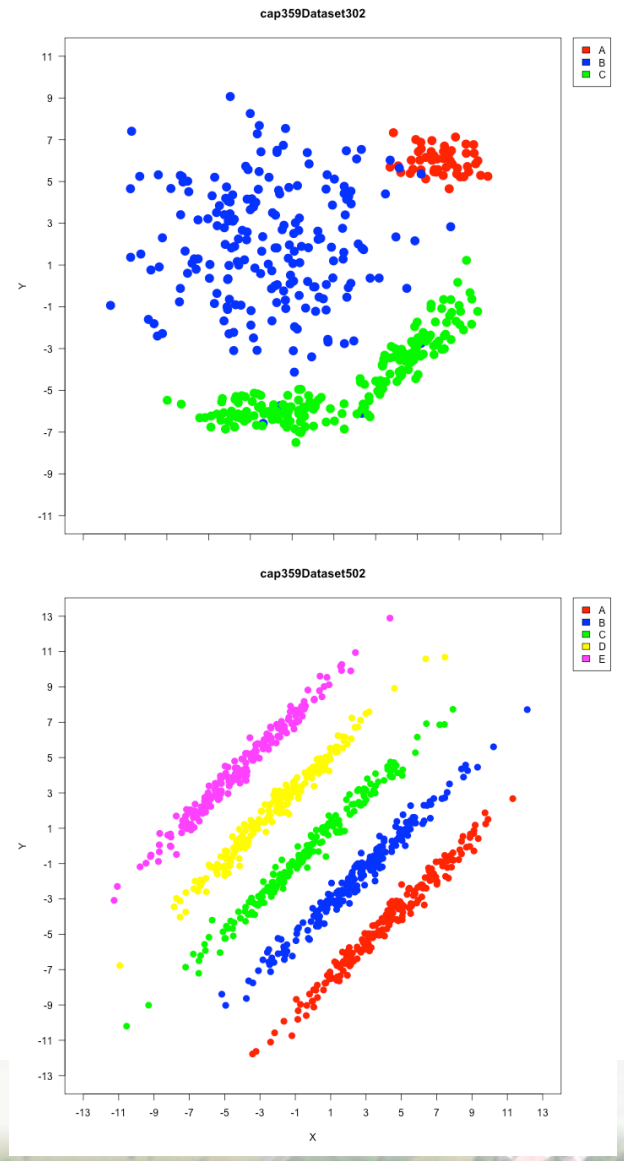


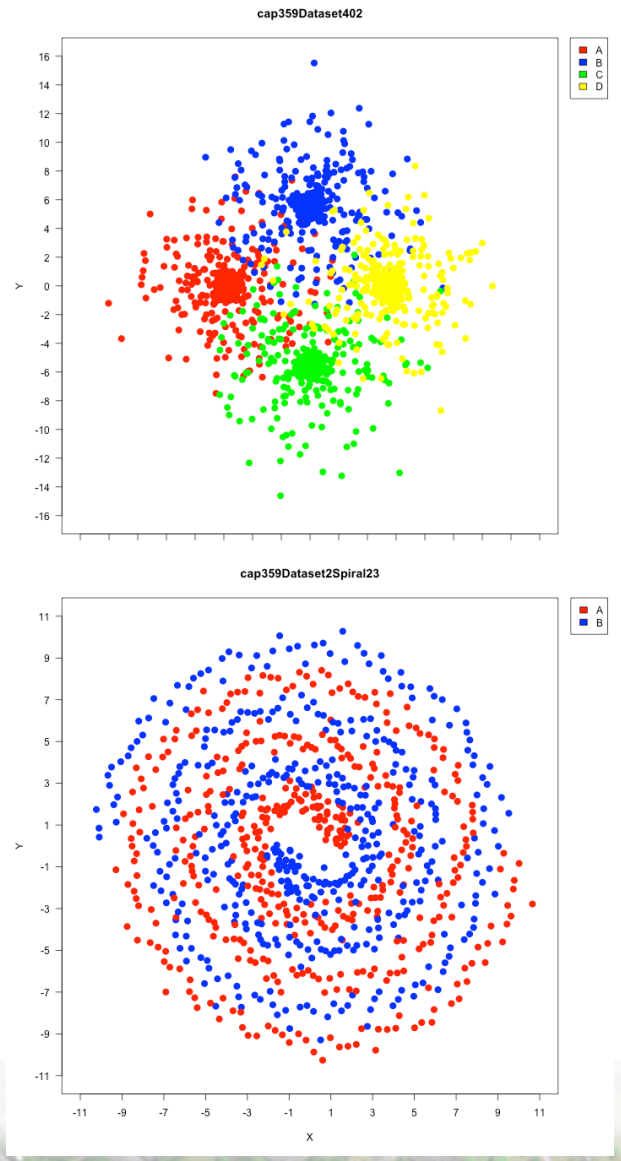
Classification: Nearest Neighbors

- No Model: uses labeled data points themselves.
  - Computationally intensive.
- Class is determined from majority of labeled nearest neighbors.



Classification: Nearest Neighbors

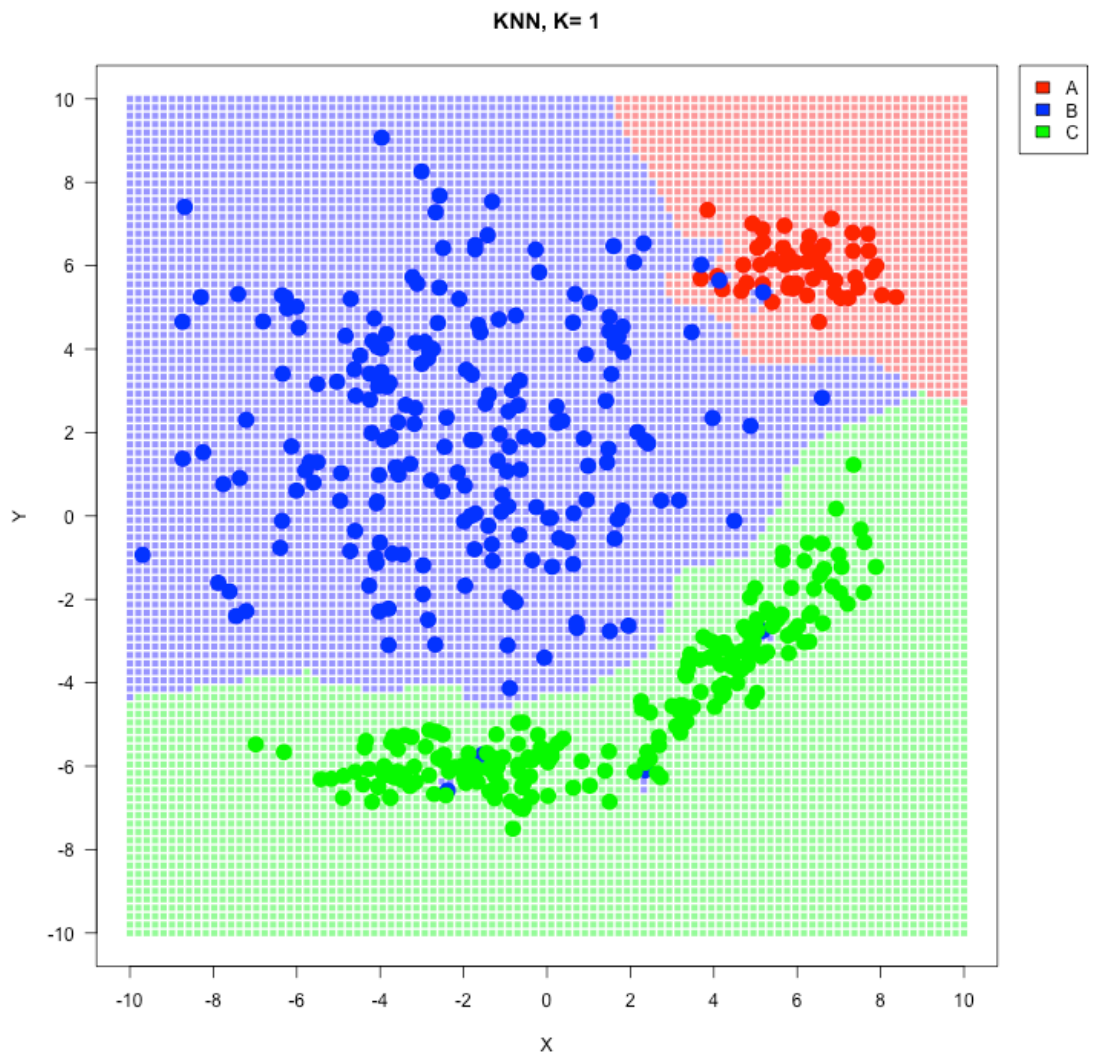




Classification: Nearest Neighbors

- K=1
- A perfect classifier? Accuracy = 1

	Classified as		
	A	B	C
A	50	0	0
B	0	200	0
C	0	0	200

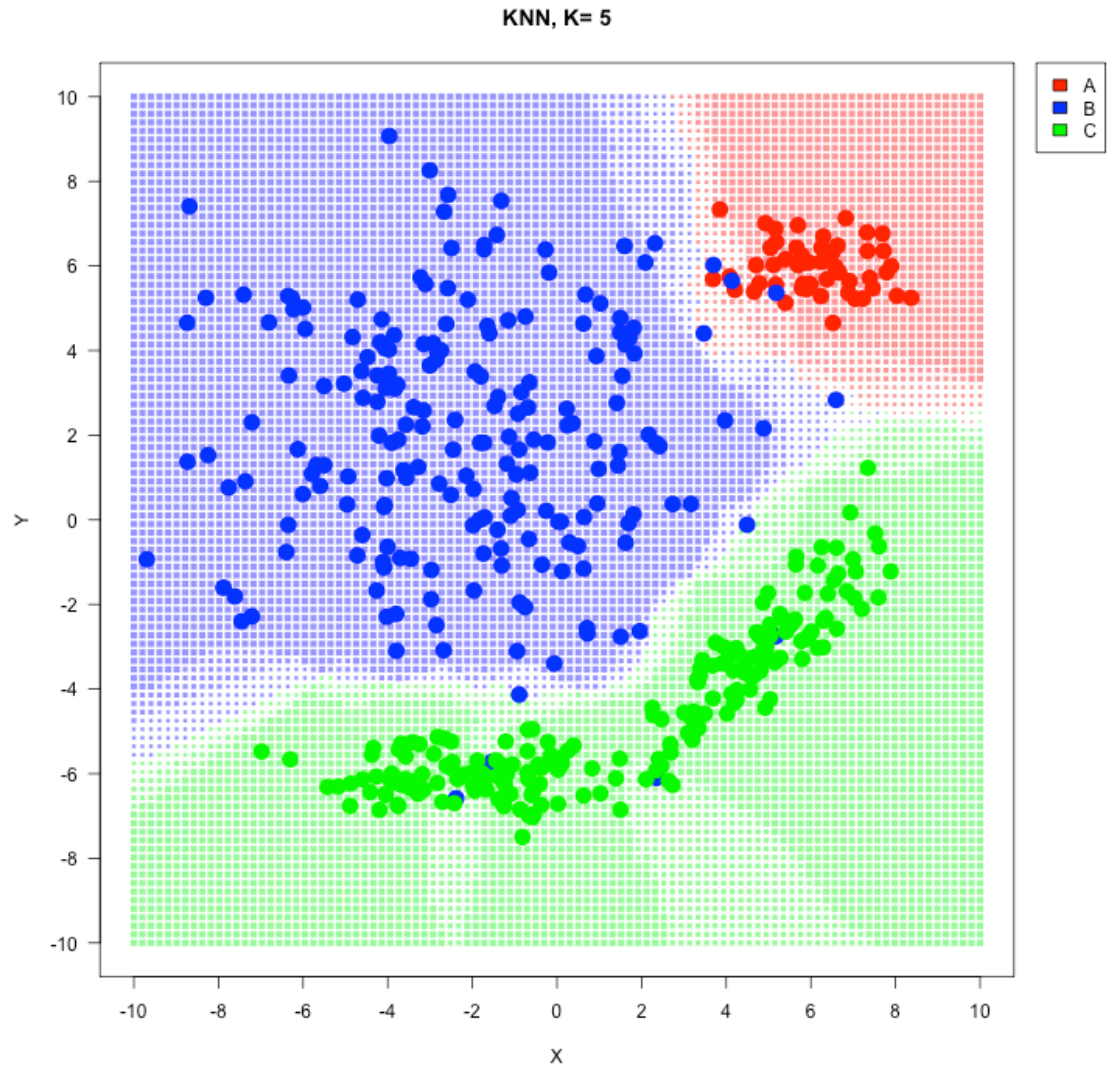


Classification: Nearest Neighbors

- K=5

Accuracy: 0.98

	Classified as		
	A	B	C
A	50	0	0
B	4	191	5
C	0	0	200

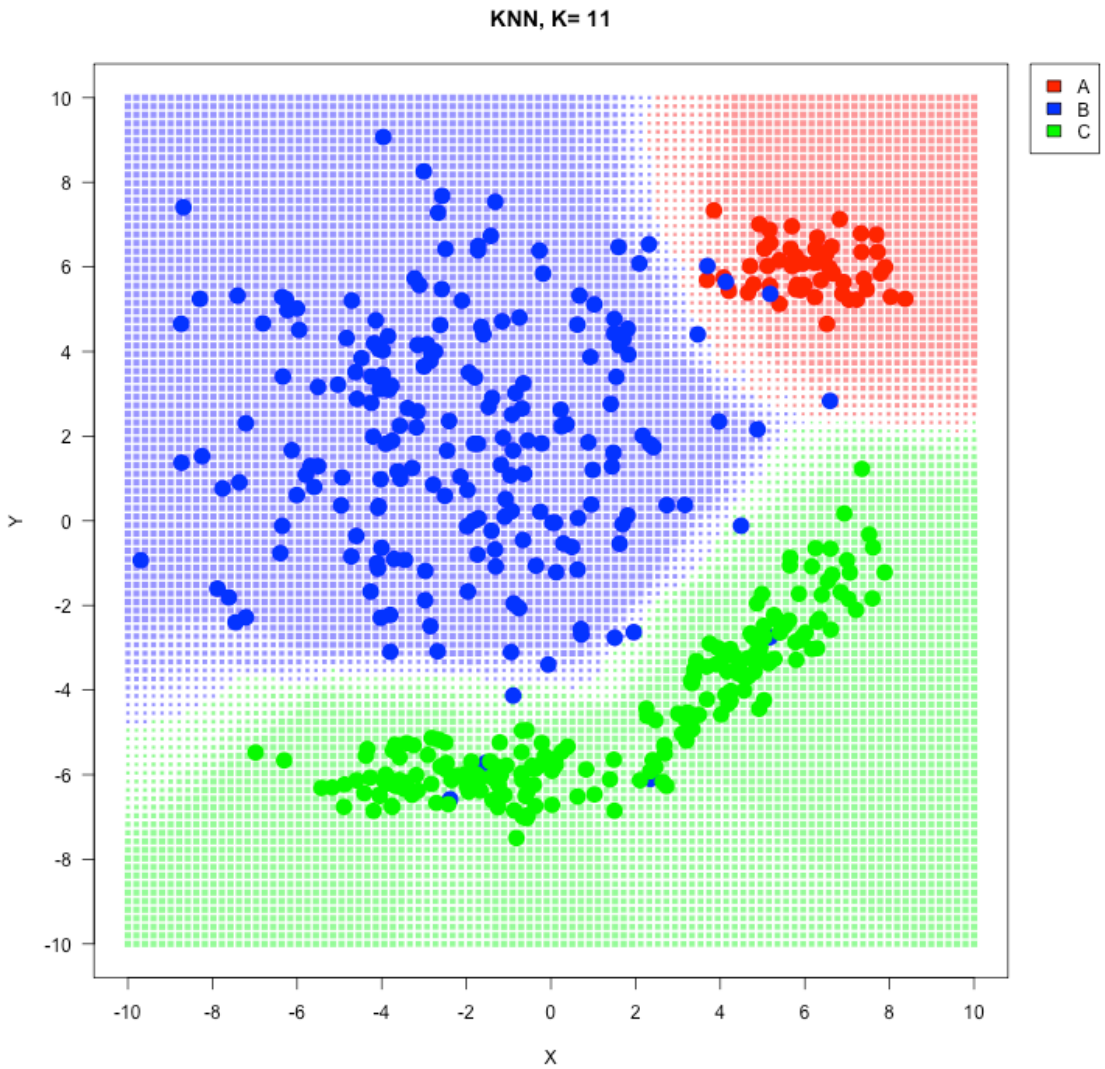


Classification: Nearest Neighbors

- K=11

Accuracy: 0.9755556

	Classified as		
	A	B	C
A	50	0	0
B	4	189	7
C	0	0	200

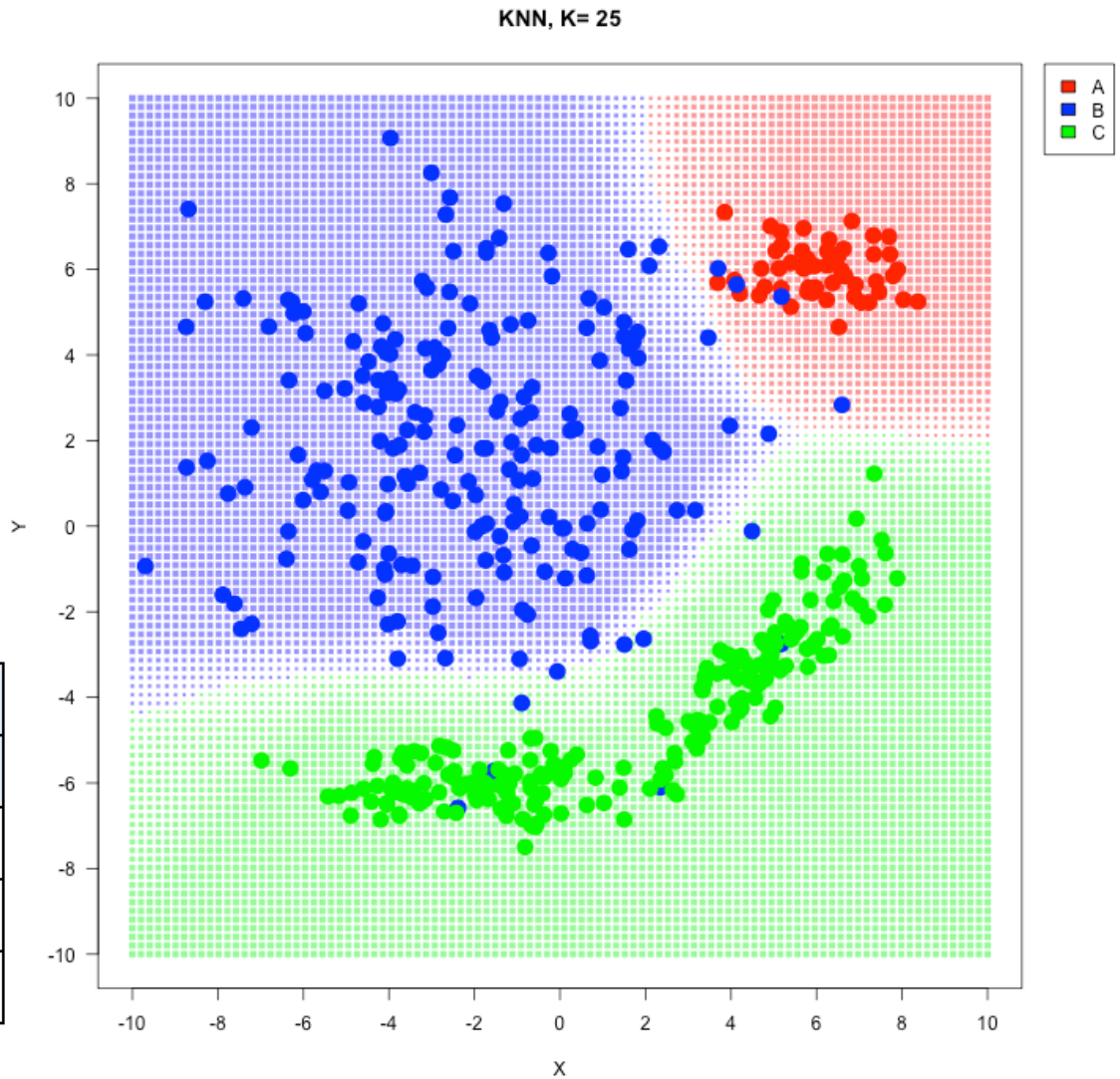


Classification: Nearest Neighbors

- K=25

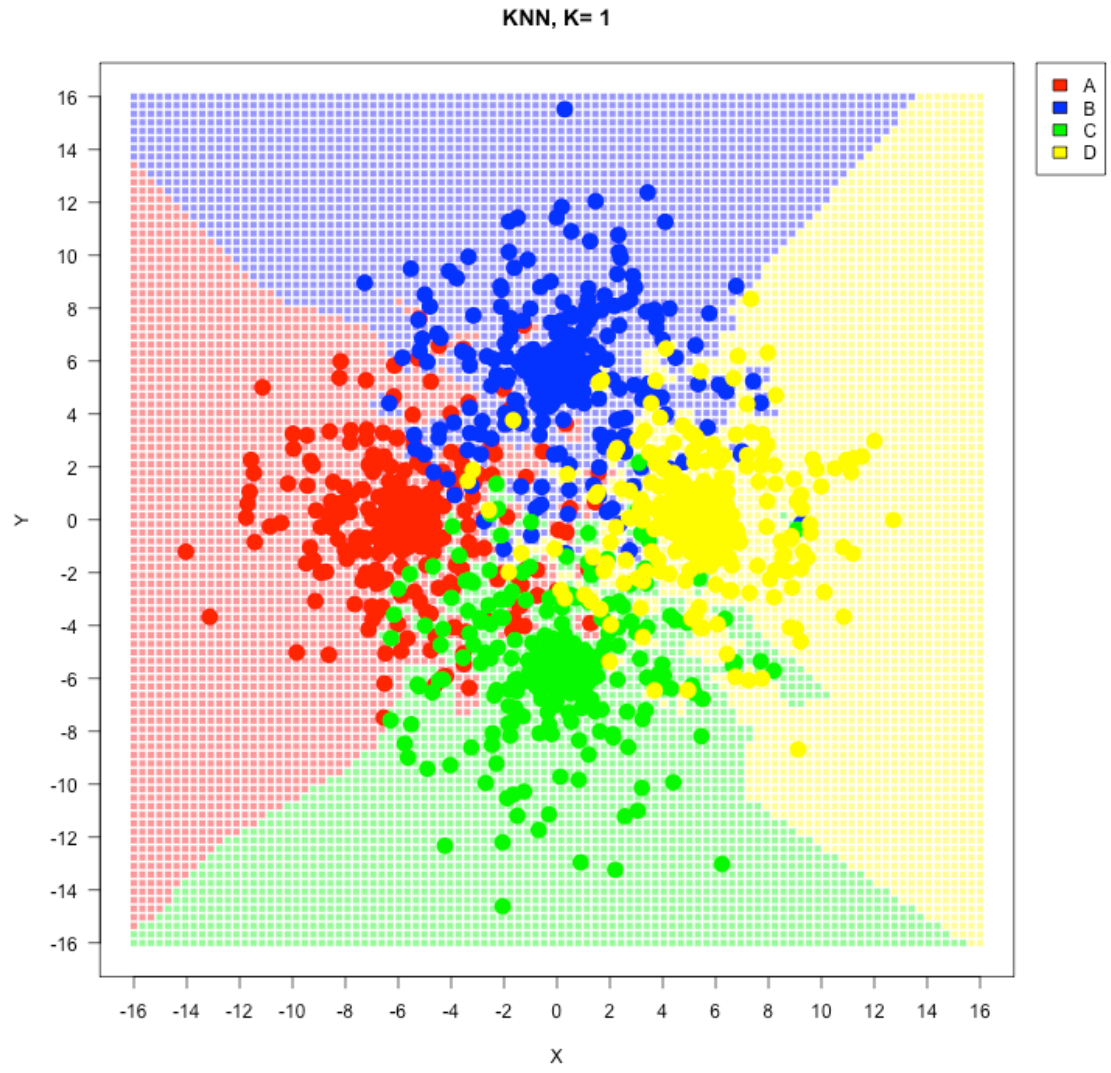
Accuracy: 0.9711111

	Classified as		
	A	B	C
A	50	0	0
B	4	187	9
C	0	0	200



Classification: Nearest Neighbors

- K=1

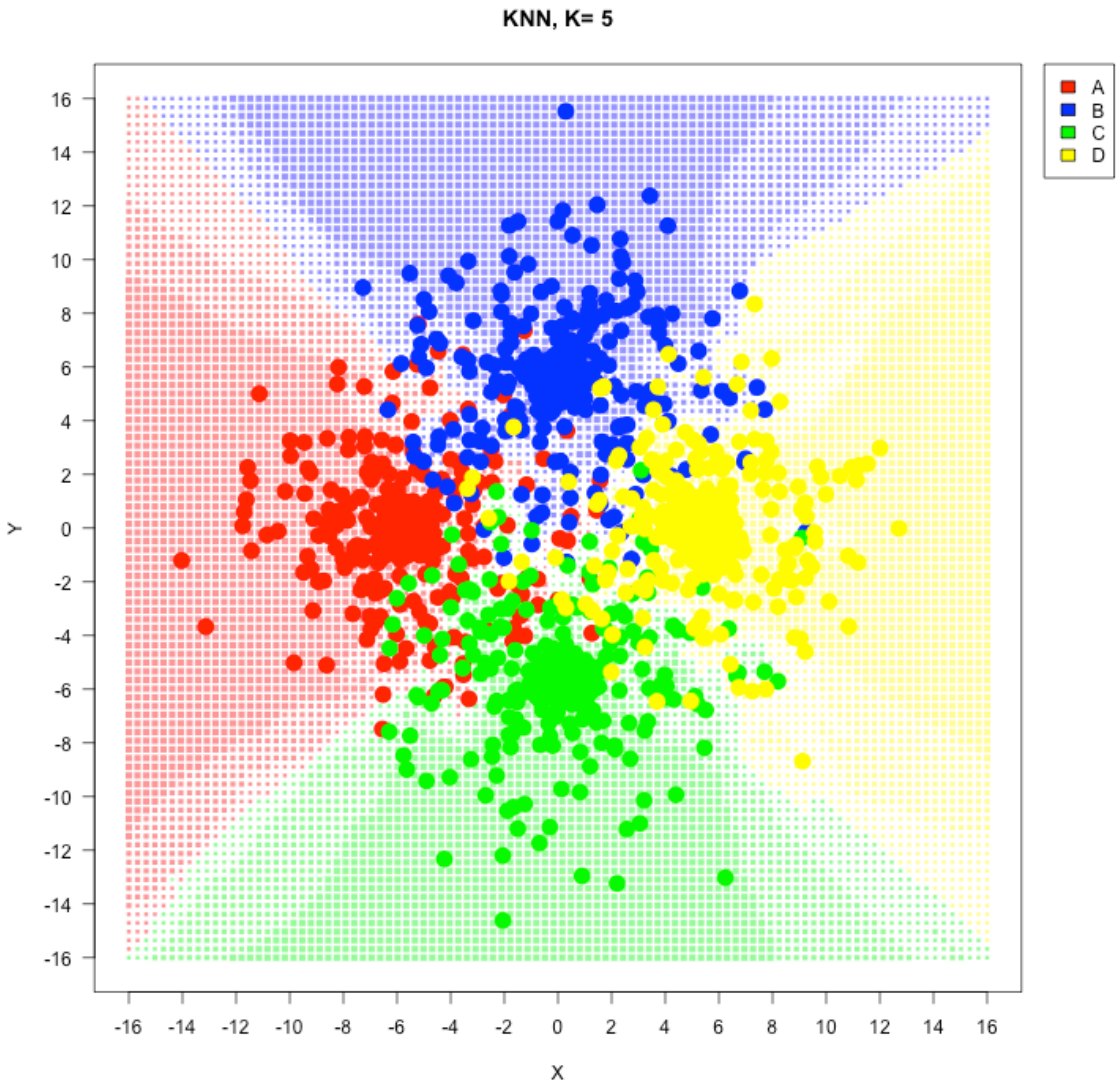


Classification: Nearest Neighbors

- K=5

Accuracy: 0.916875

	Classified as			
	A	B	C	D
A	367	16	16	1
B	13	370	1	16
C	17	1	361	21
D	4	15	12	369

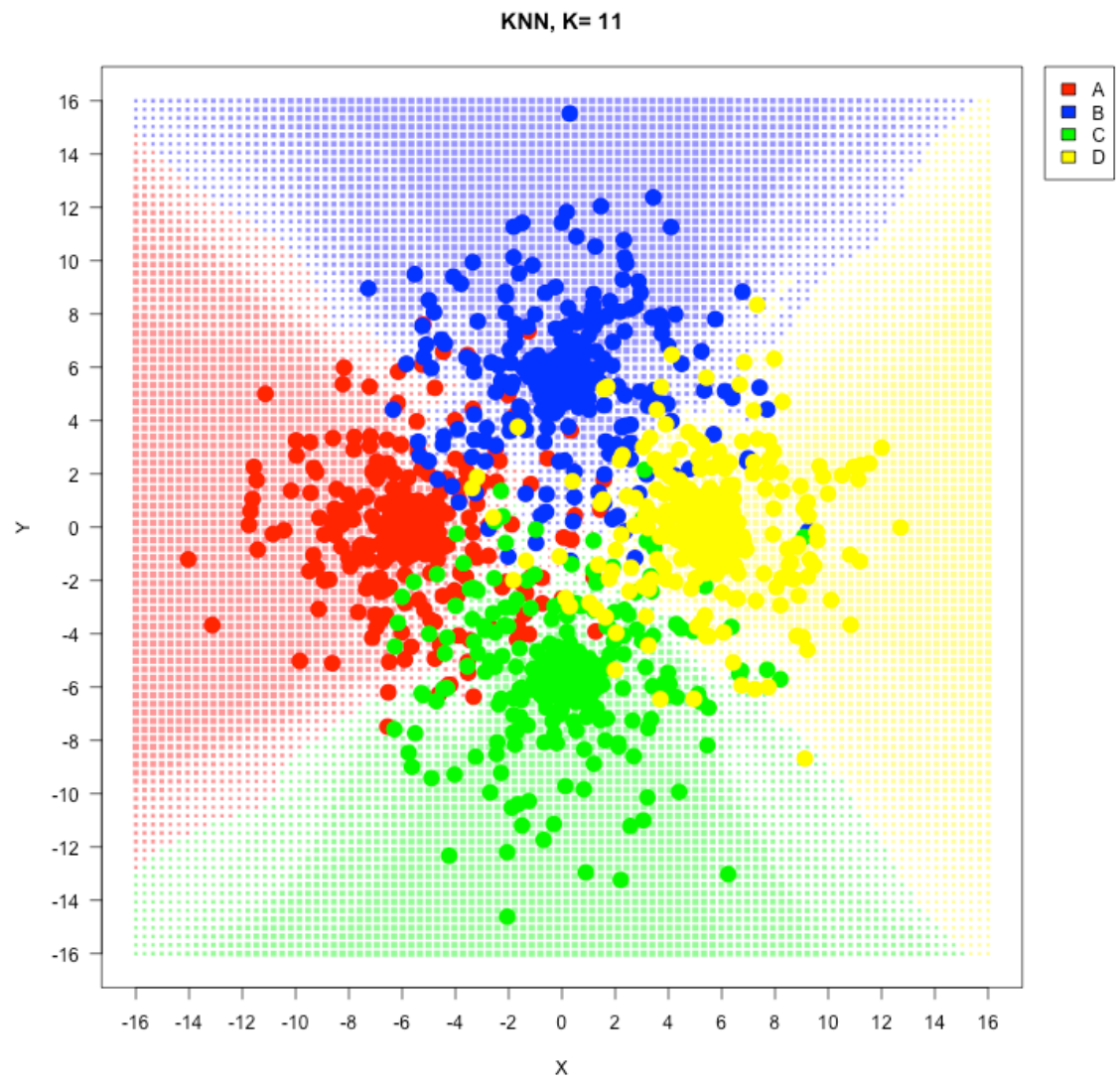


Classification: Nearest Neighbors

- K=11

Accuracy: 0.908125

	Classified as			
	A	B	C	D
A	361	19	20	0
B	12	370	1	17
C	22	1	359	18
D	6	16	15	363

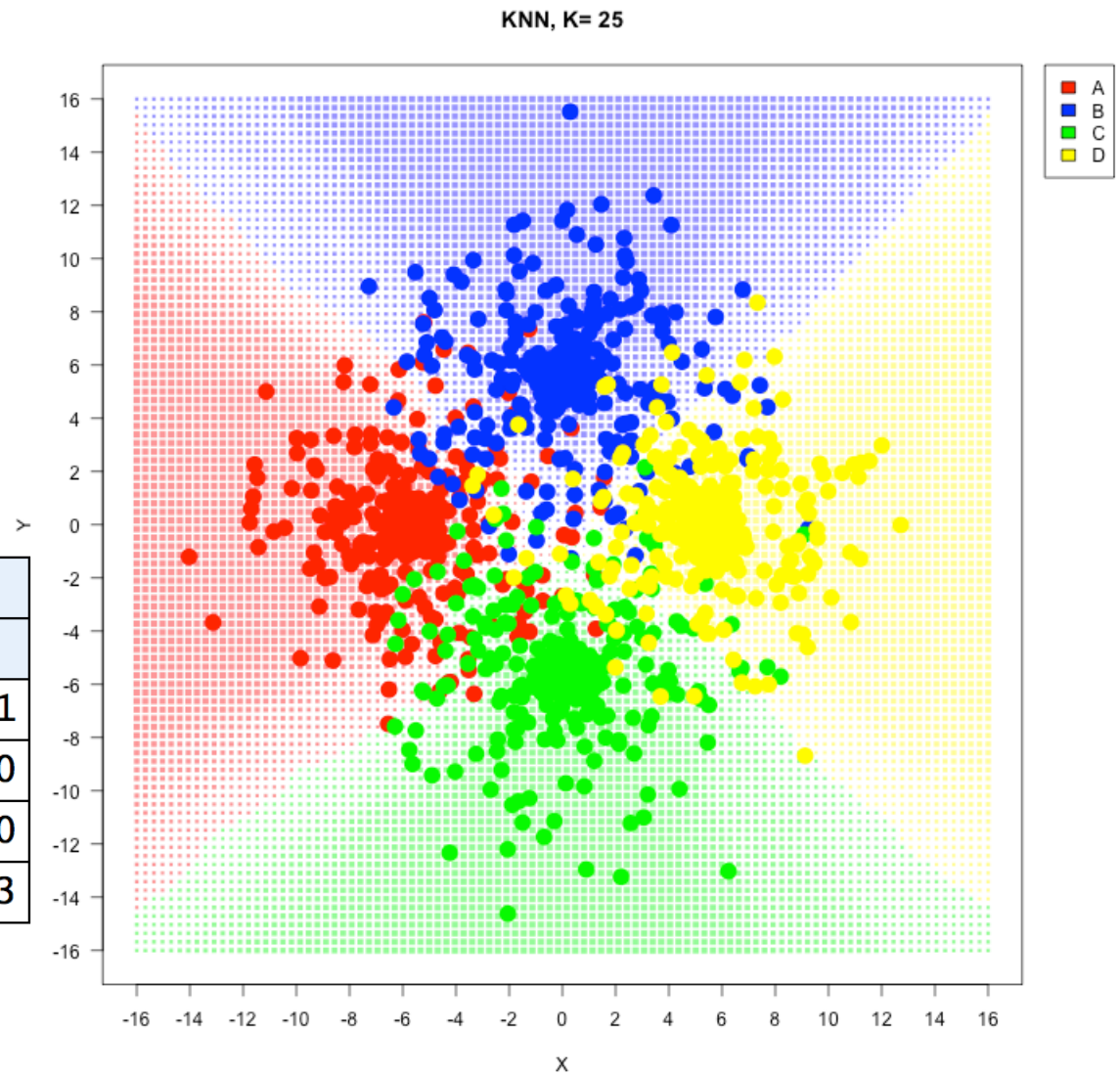


Classification: Nearest Neighbors

- K=25

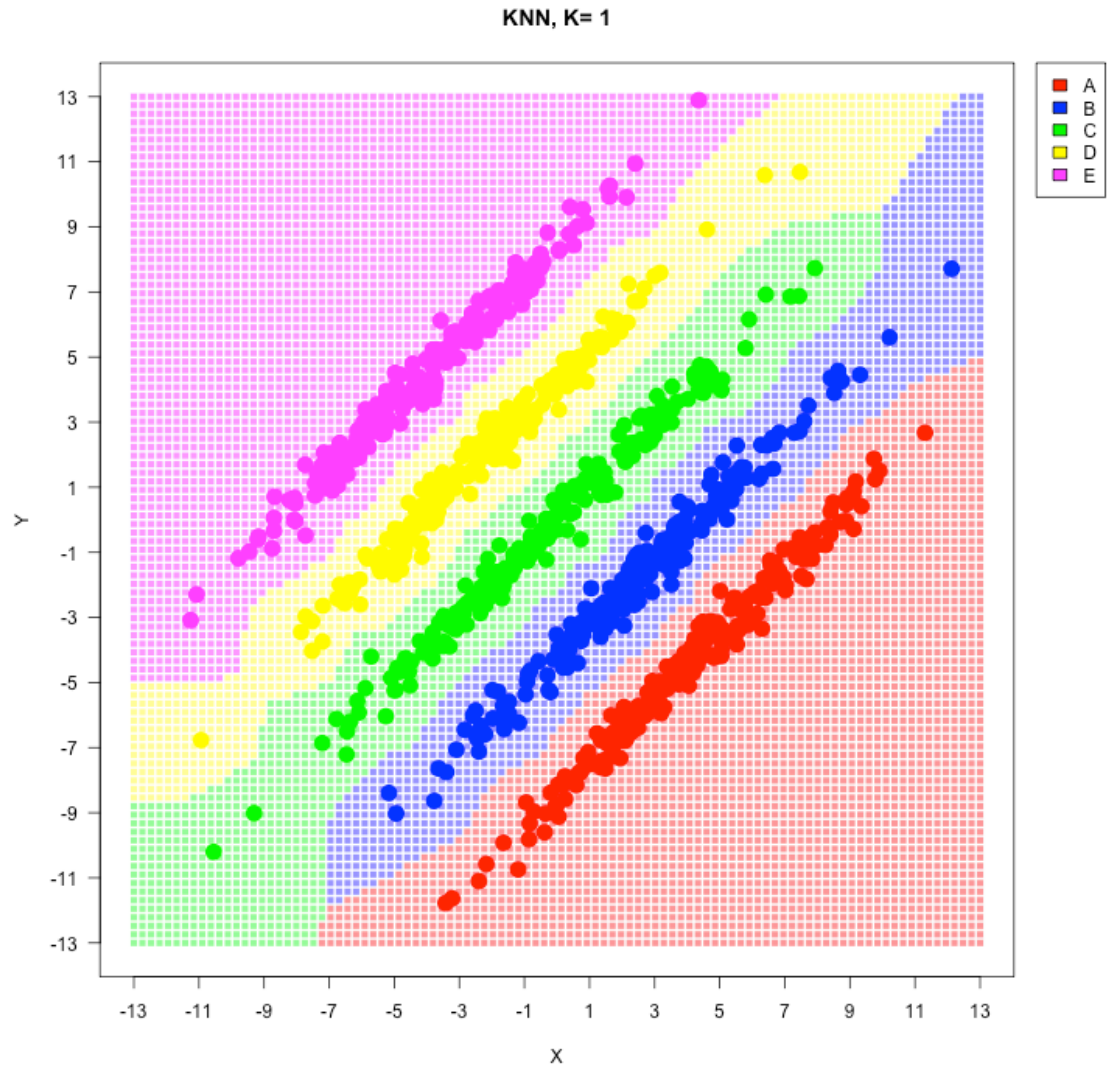
Accuracy: 0.900625

	Classified as			
	A	B	C	D
A	357	20	22	1
B	15	364	1	20
C	22	1	357	20
D	5	17	15	363



Classification: Nearest Neighbors

- K=1

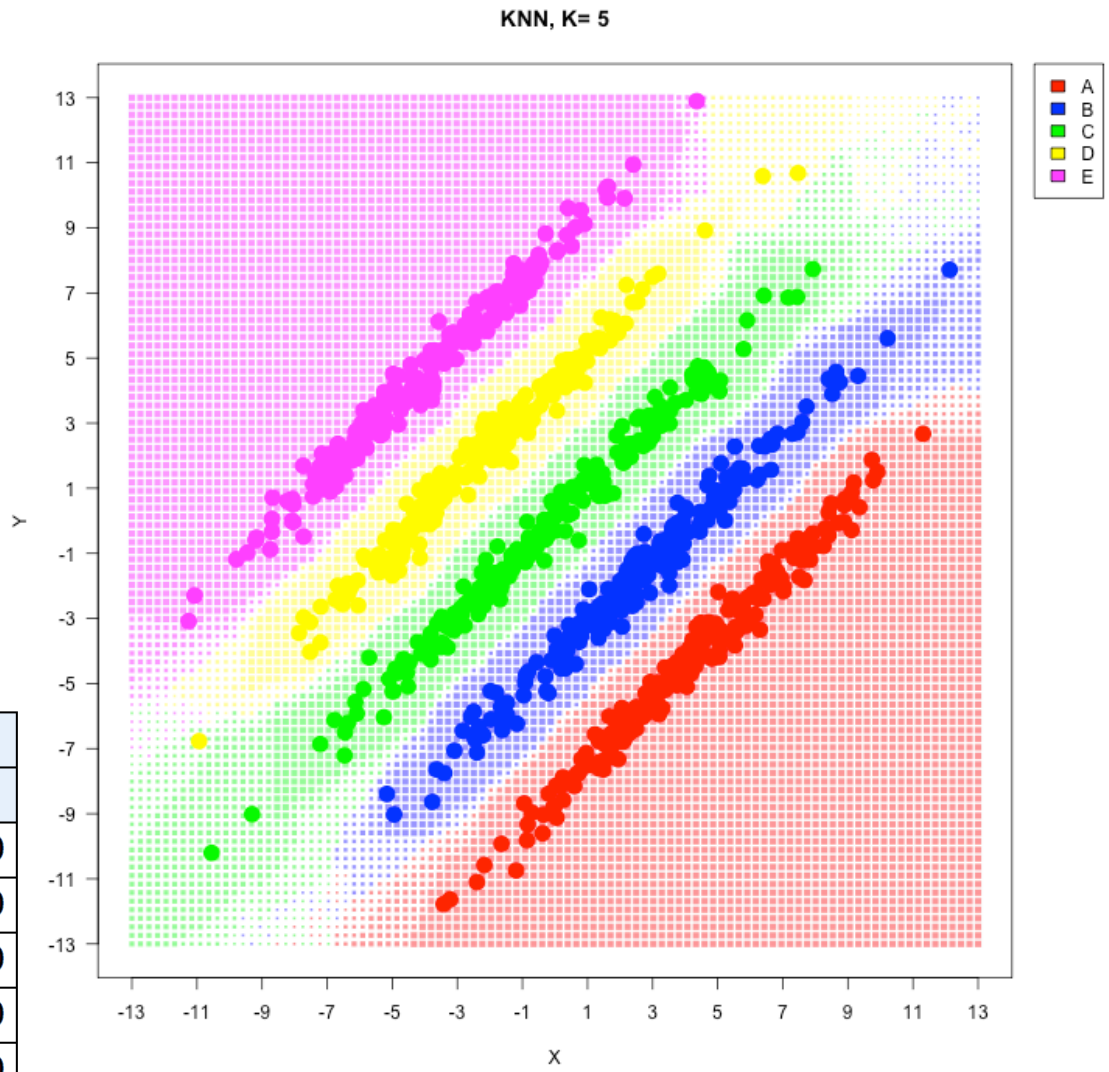


Classification: Nearest Neighbors

- K=5

Accuracy: 0.999

	Classified as				
	A	B	C	D	E
A	200	0	0	0	0
B	0	200	0	0	0
C	0	0	200	0	0
D	0	0	1	199	0
E	0	0	0	0	200

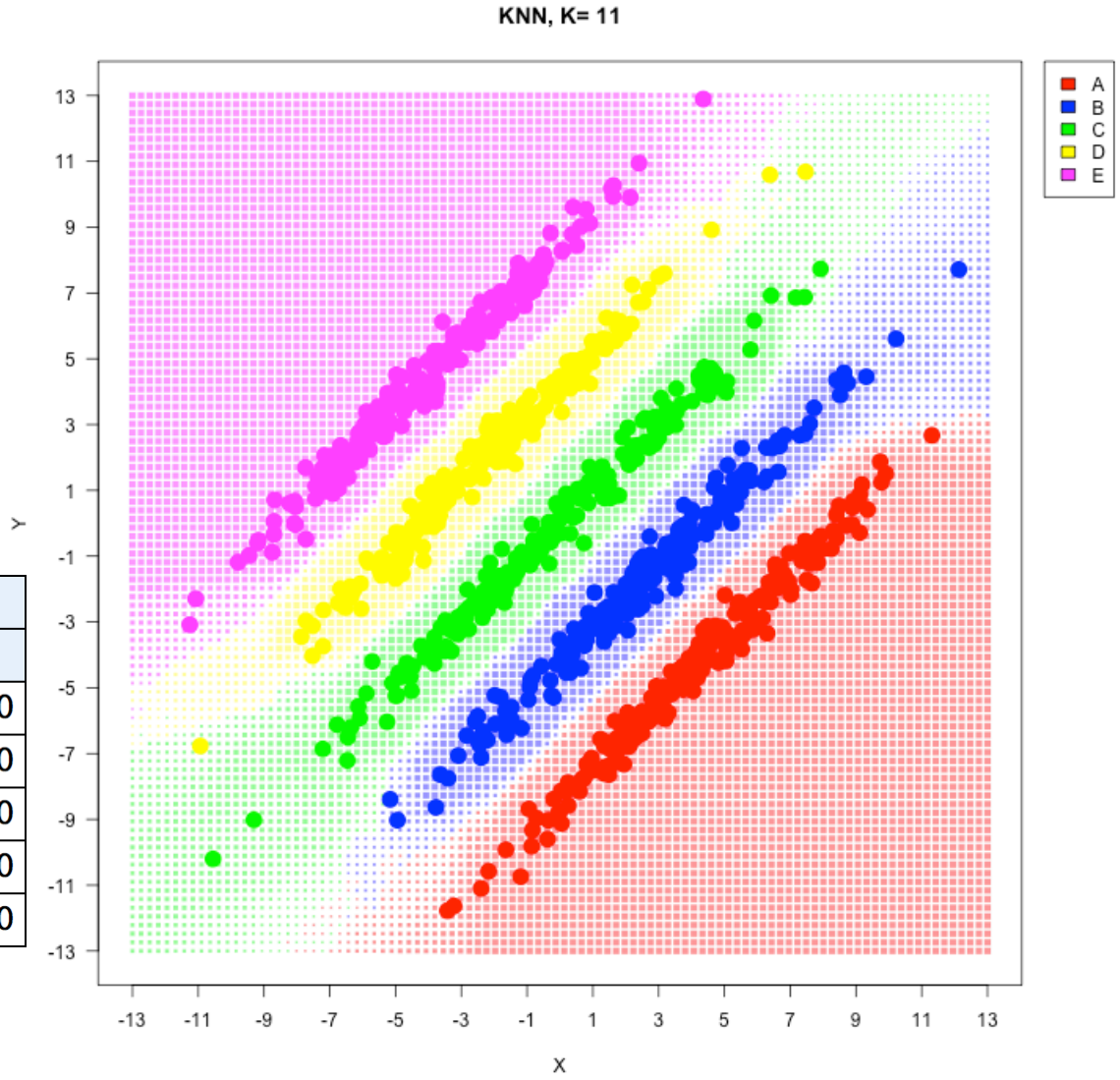


Classification: Nearest Neighbors

- K=11

Accuracy: 0.998

	Classified as				
	A	B	C	D	E
A	200	0	0	0	0
B	0	200	0	0	0
C	0	0	200	0	0
D	0	0	2	198	0
E	0	0	0	0	200

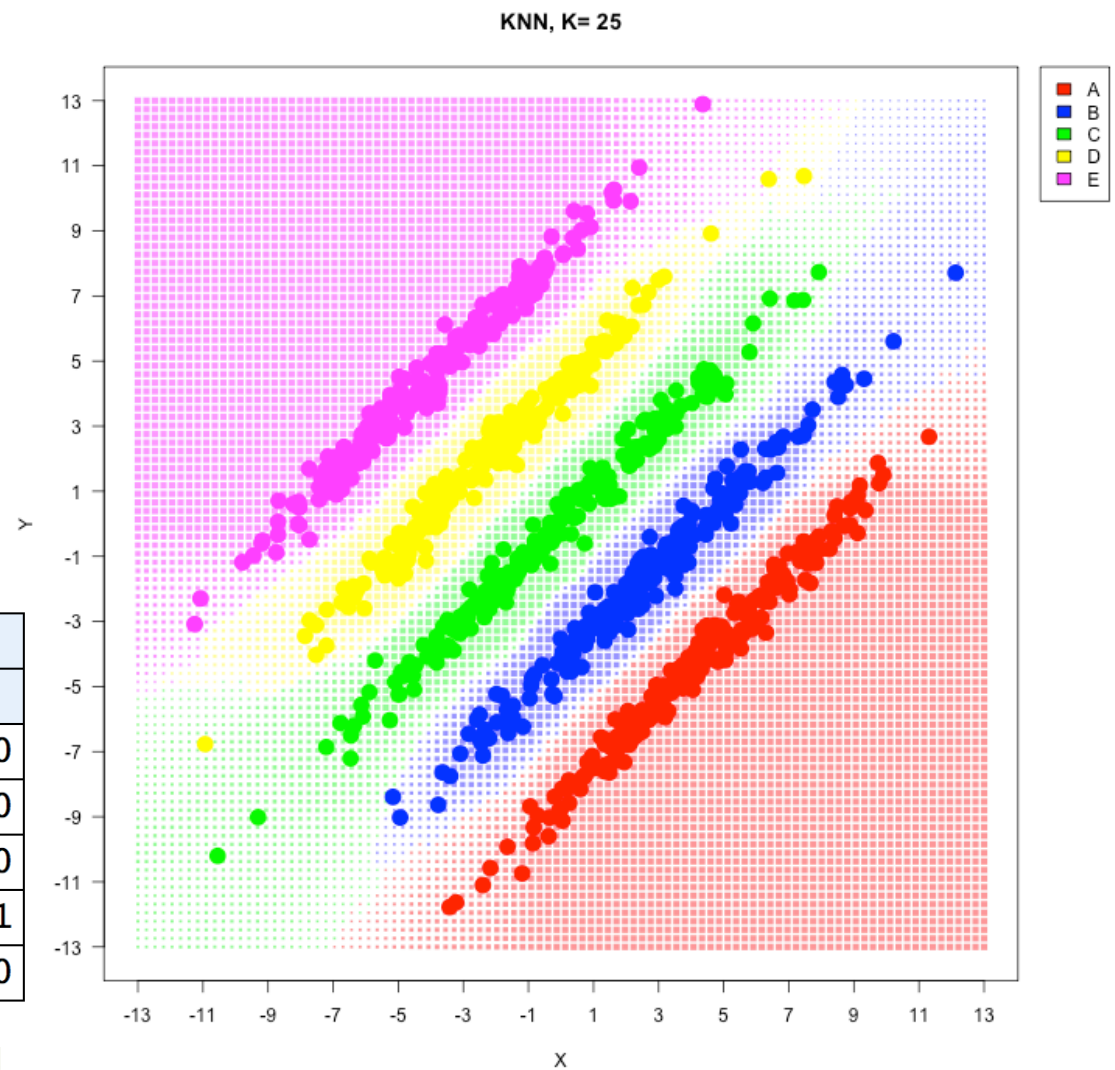


Classification: Nearest Neighbors

- K=25

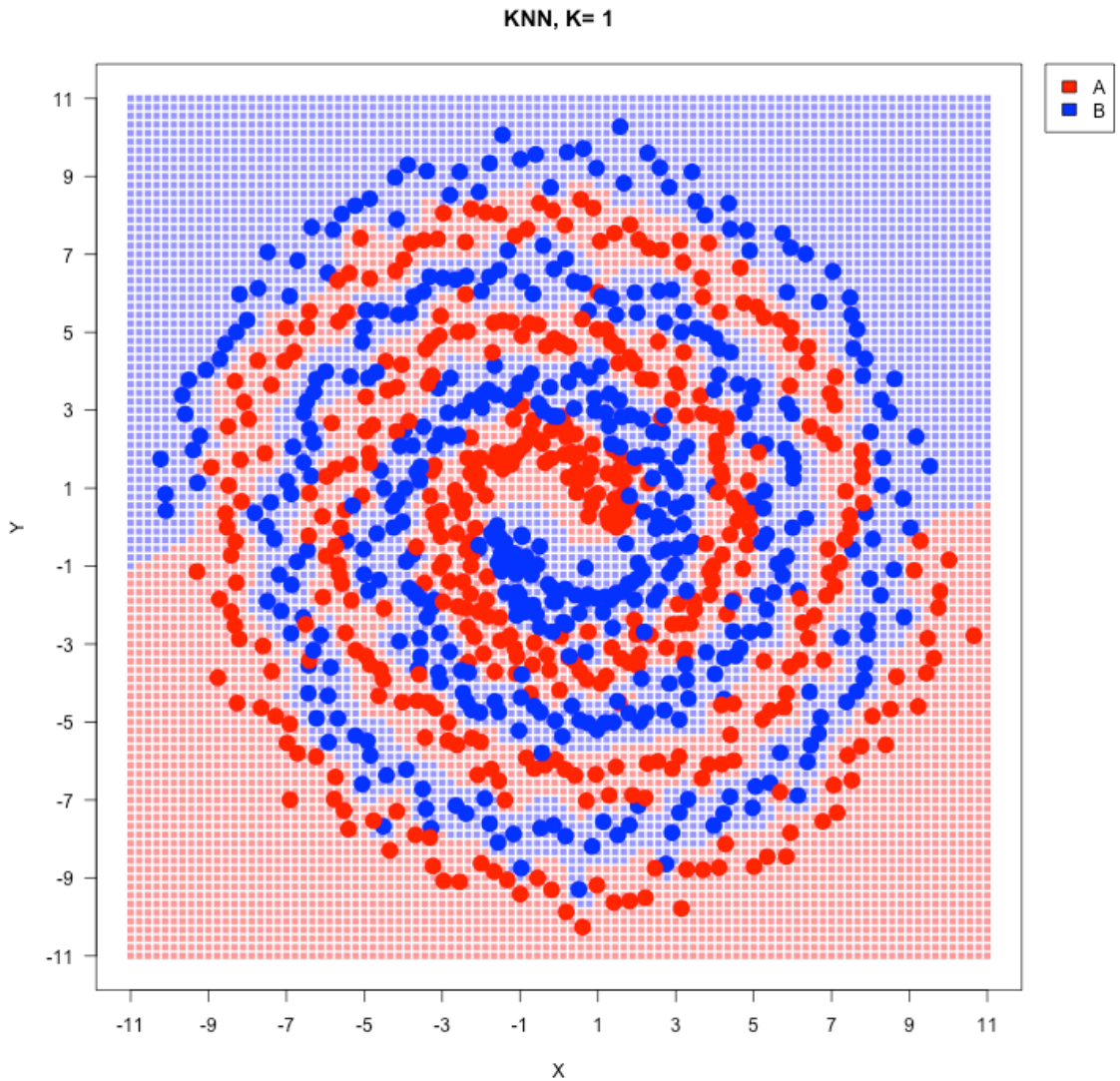
Accuracy: 0.998

	Classified as				
	A	B	C	D	E
A	200	0	0	0	0
B	0	200	0	0	0
C	0	0	200	0	0
D	0	0	1	198	1
E	0	0	0	0	200



Classification: Nearest Neighbors

- K=1

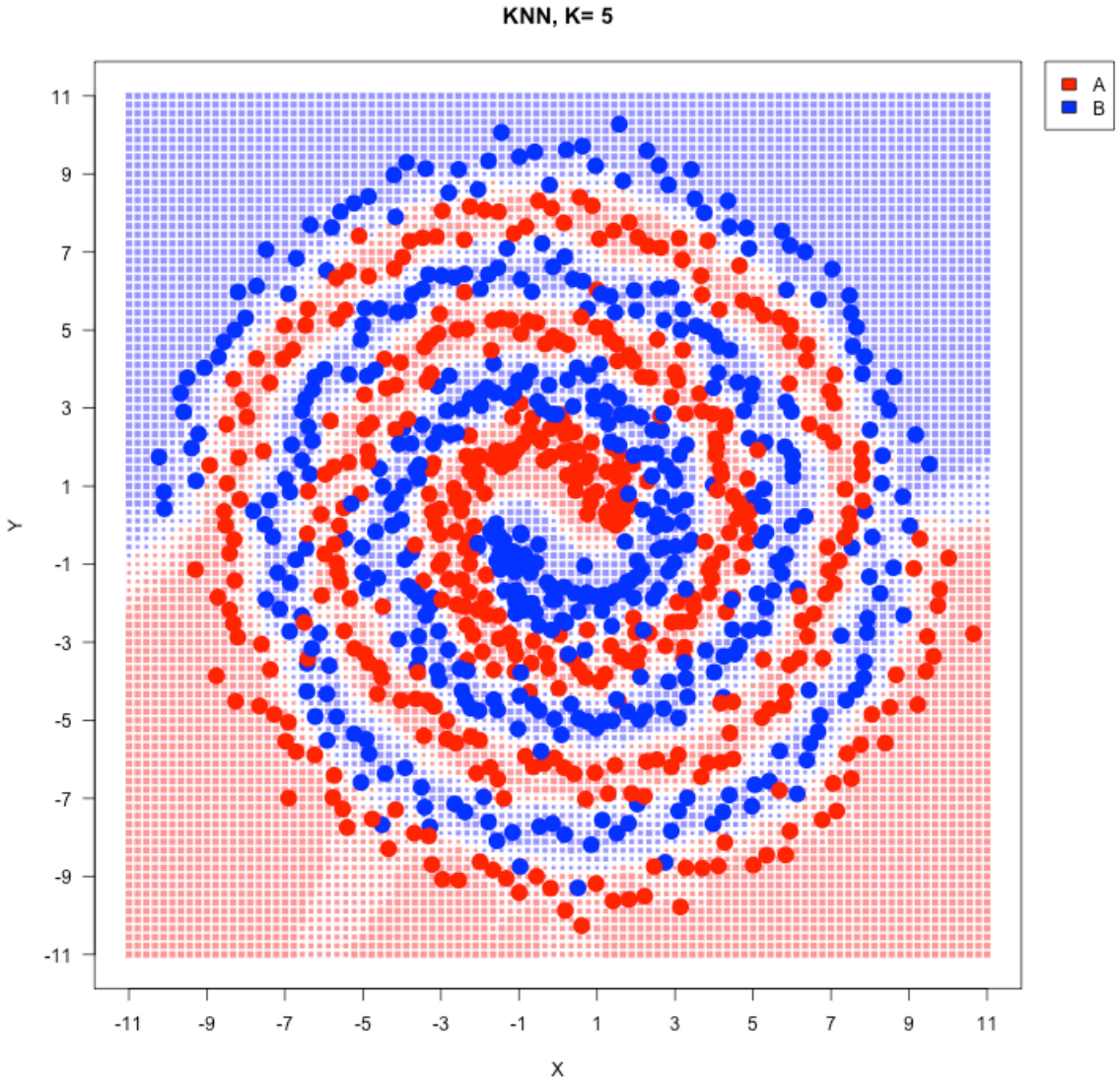


Classification: Nearest Neighbors

- K=5

Accuracy: 0.9277778

	Classified as	
	A	B
A	420	30
B	35	415

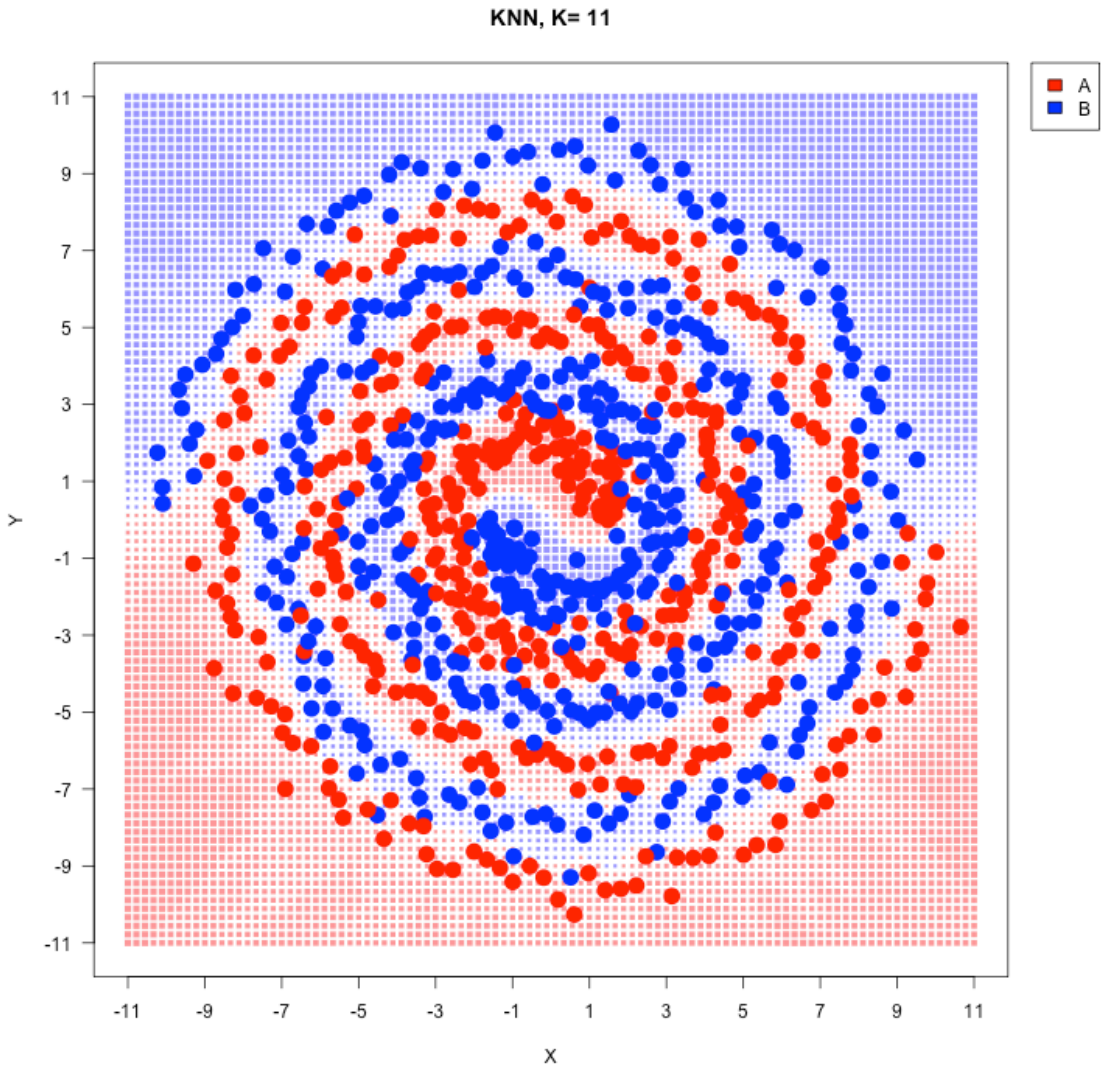


Classification: Nearest Neighbors

- K=11

Accuracy: 0.9055556

	Classified as	
	A	B
A	414	36
B	49	401

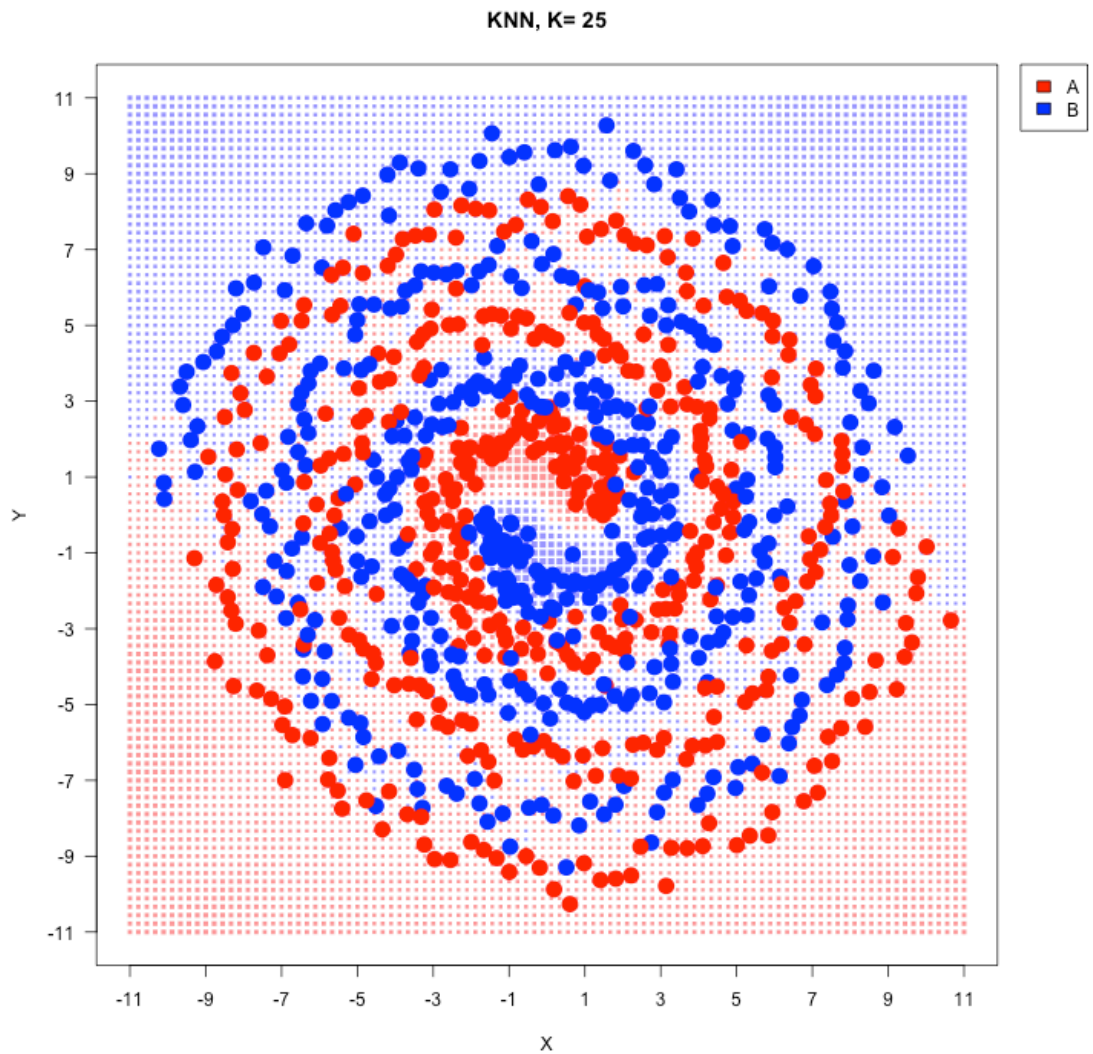


Classification: Nearest Neighbors

- K=25

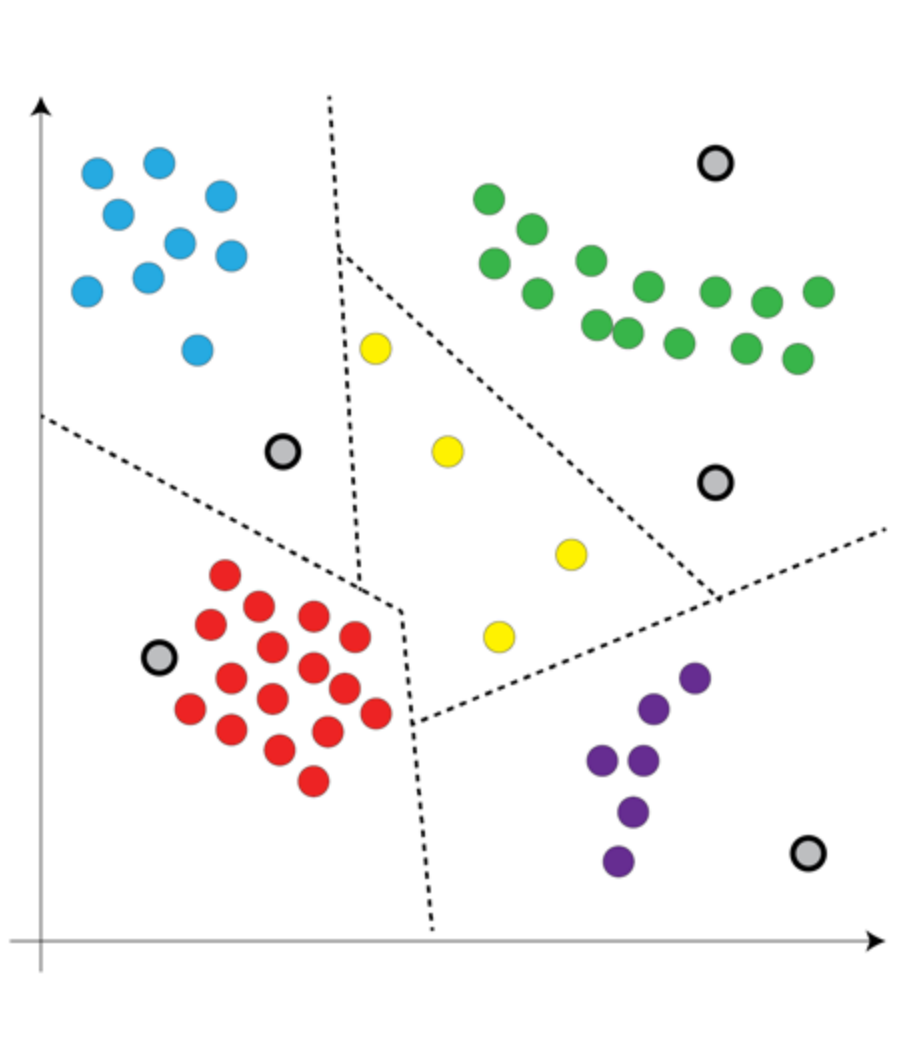
Accuracy: 0.73

	Classified as	
	A	B
A	341	109
B	134	316



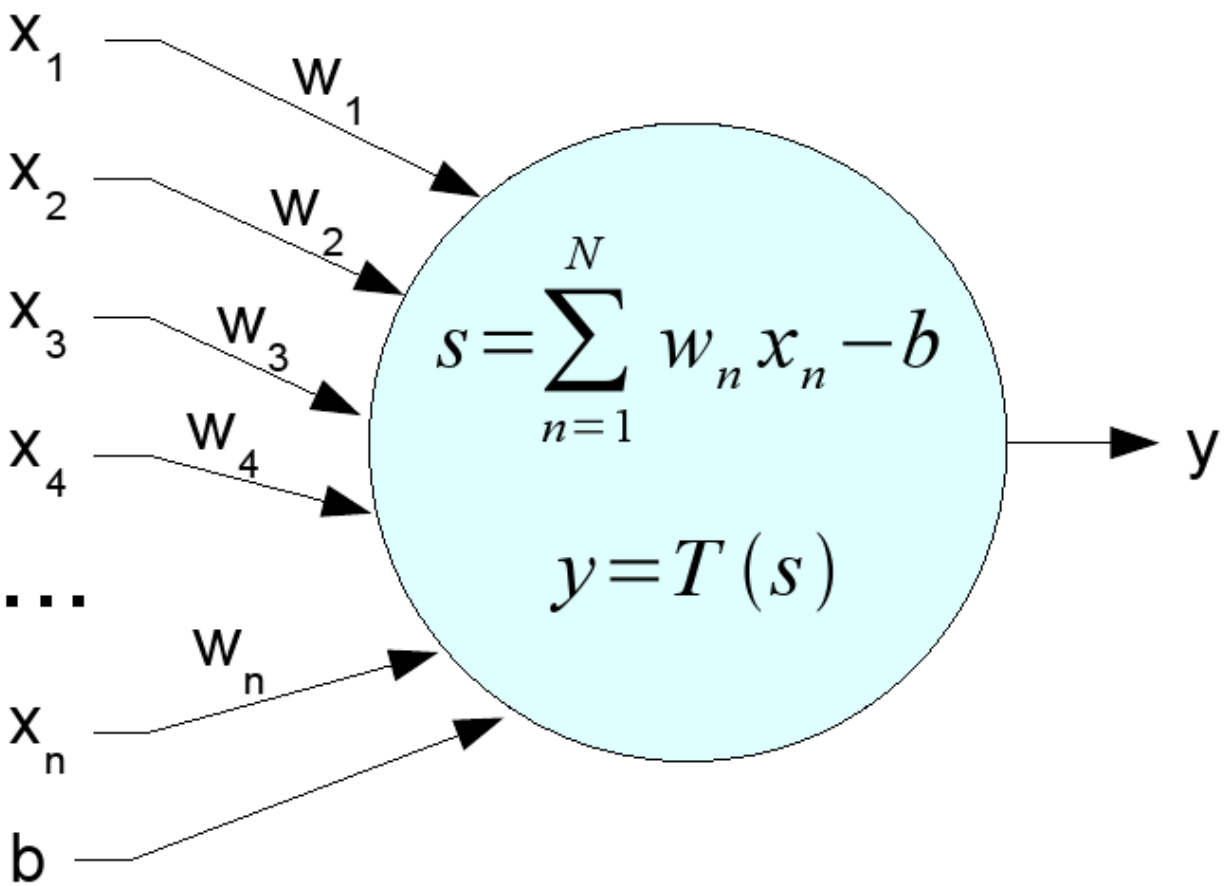
Classification: Neural Networks (MLPs)

- Model: parameters of a neural network, trained to separate classes.
  - Model is hard to understand.
  - Underfitting/Overfitting problem.



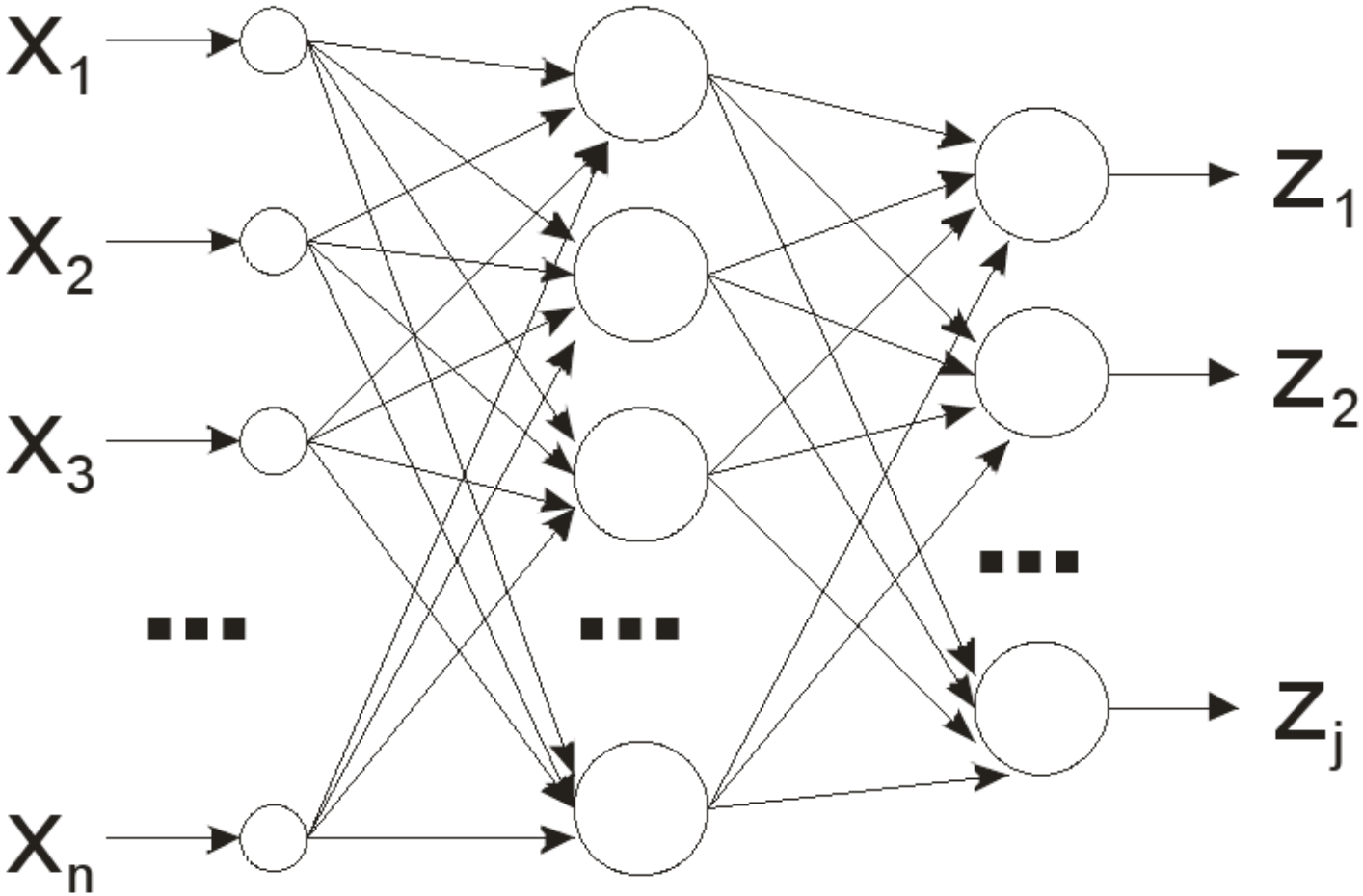
Classification: Neural Networks (MLPs)

- Perceptron

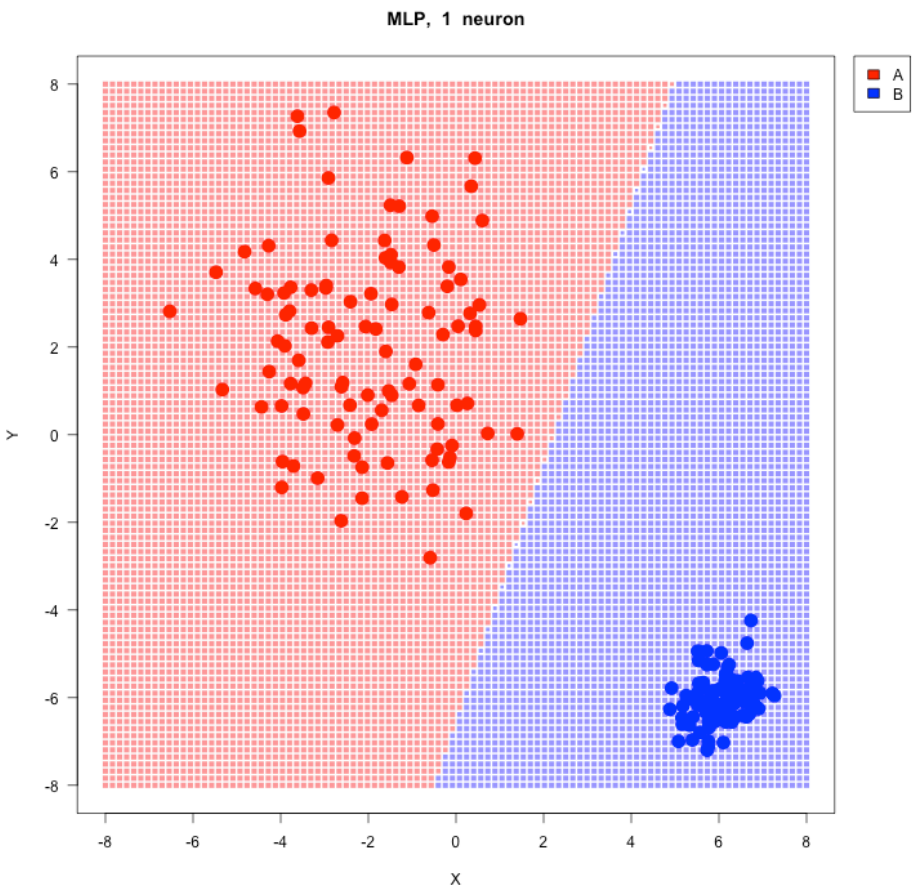


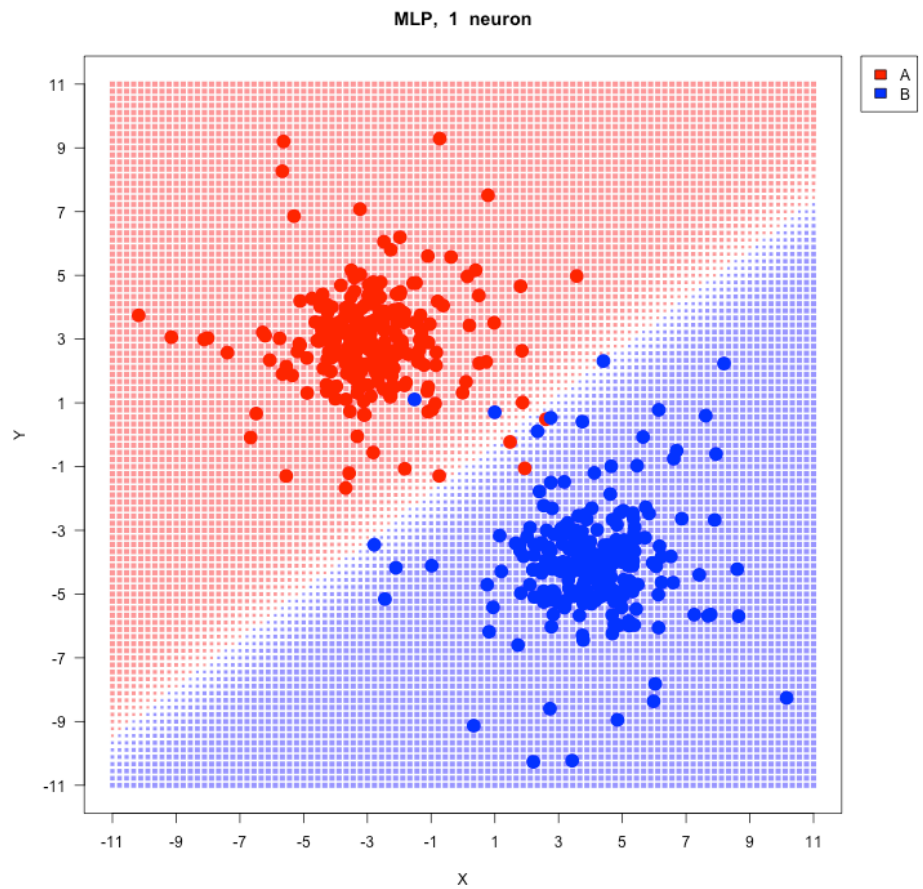
Classification: Neural Networks (MLPs)

- Multilayer Perceptron

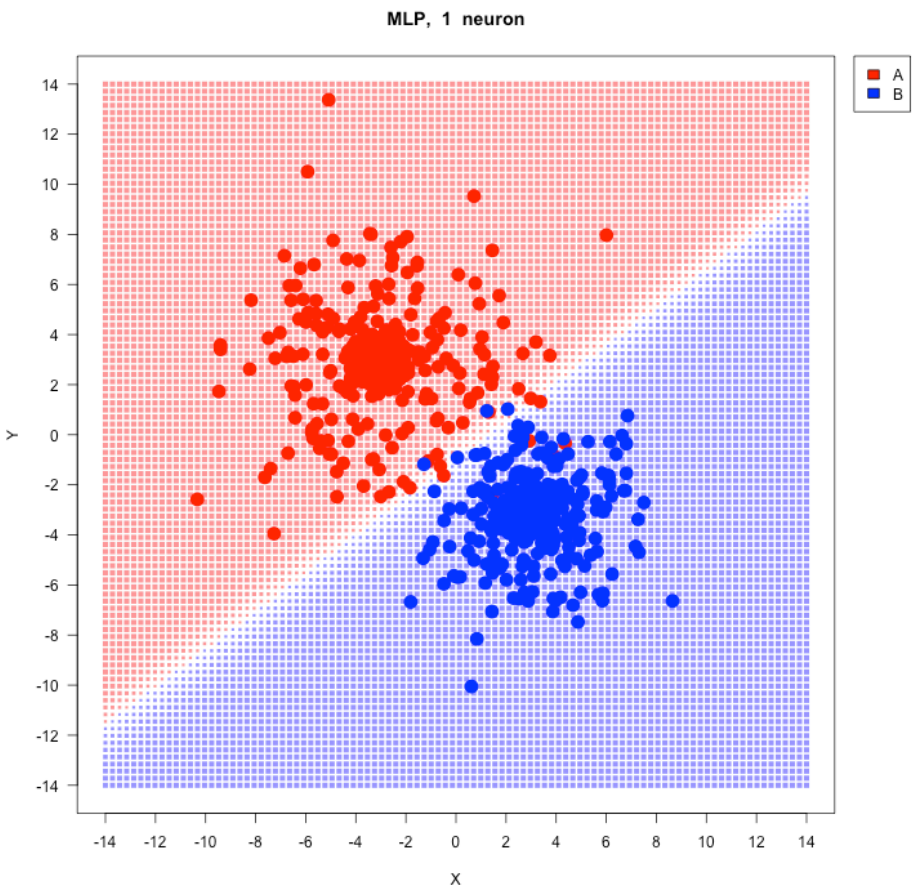


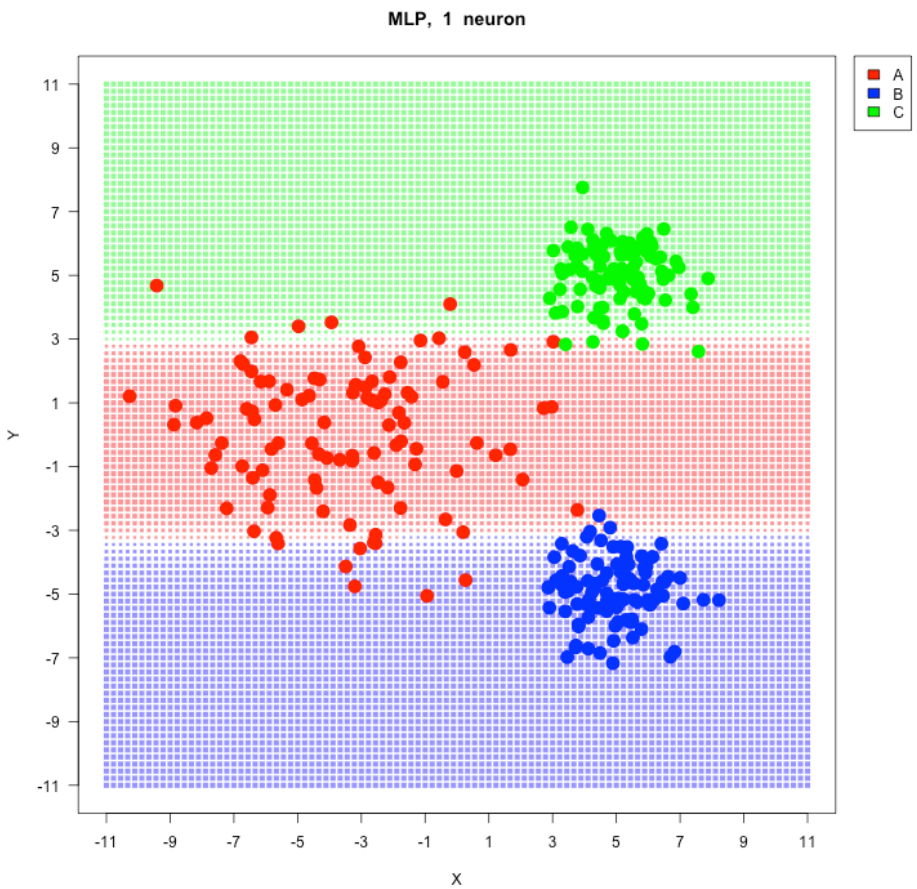
MLPs with a single neuron, single hidden layer





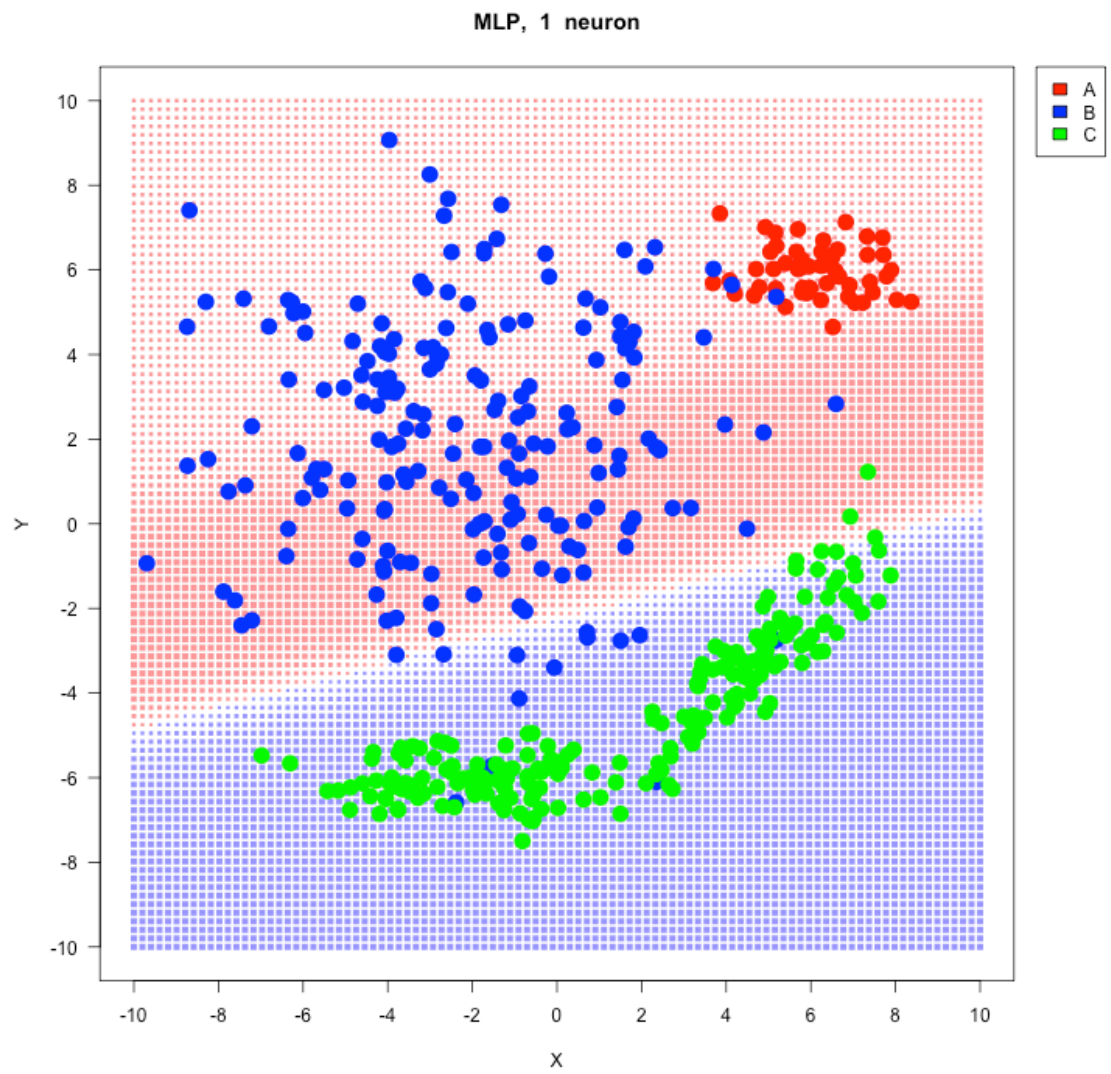
MLPs with a single neuron, single hidden layer



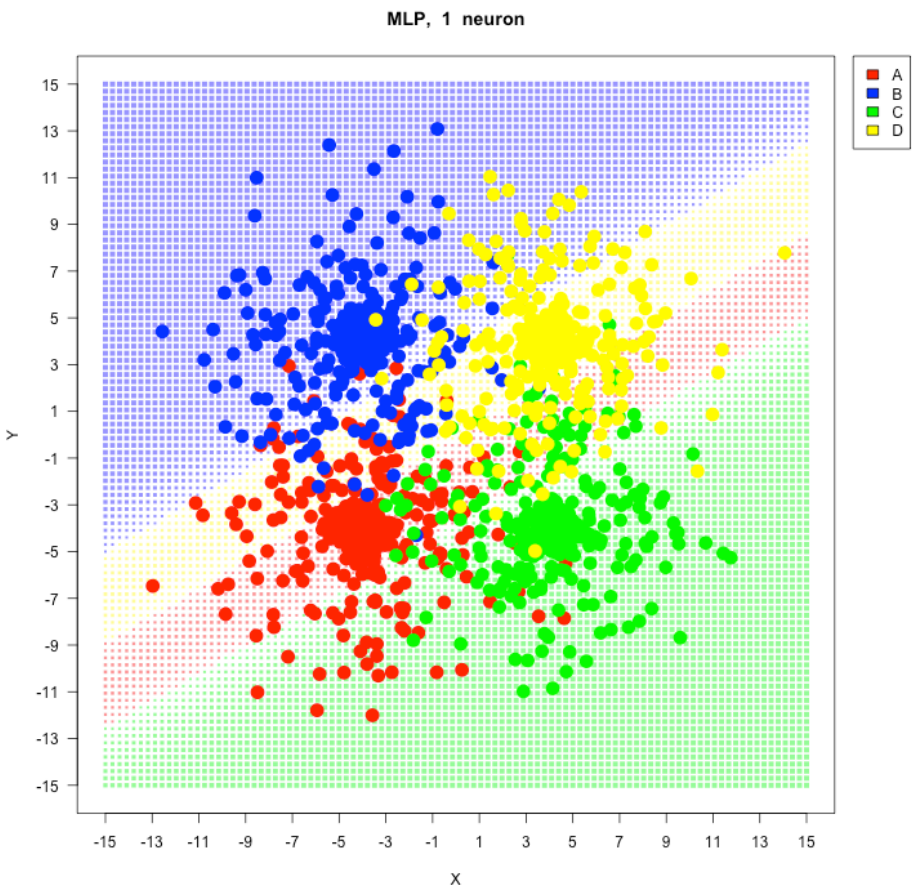


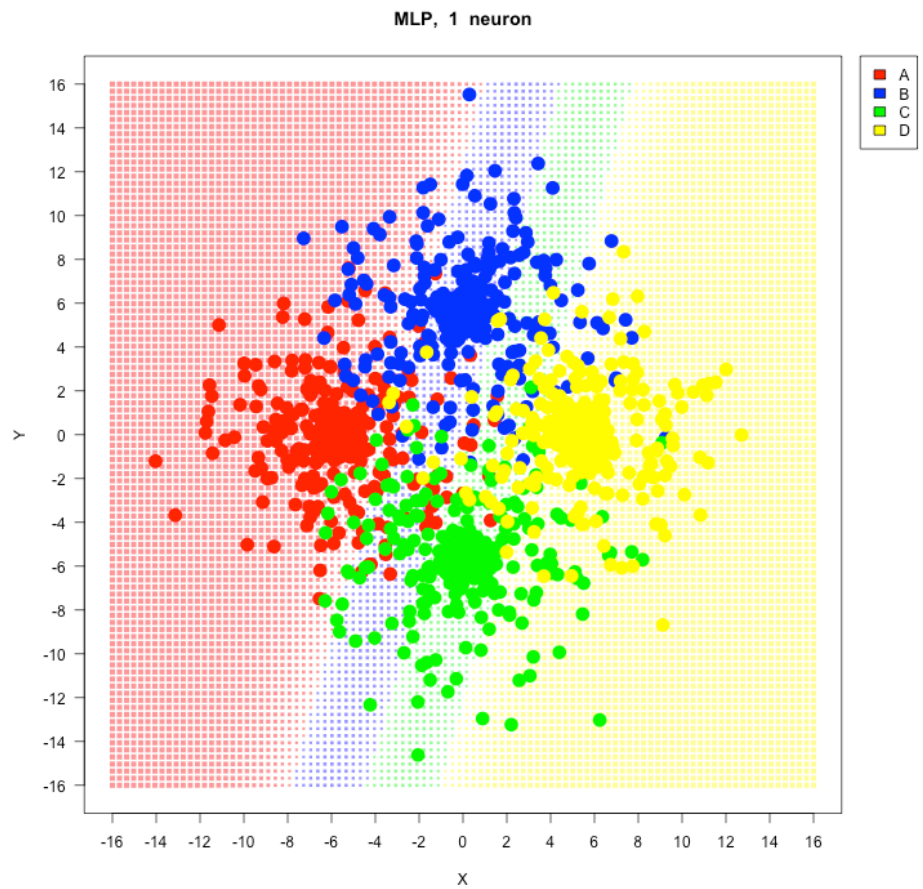
MLPs with a single neuron, single hidden layer

- One neuron cannot separate three classes with non-parallel hyperplanes!

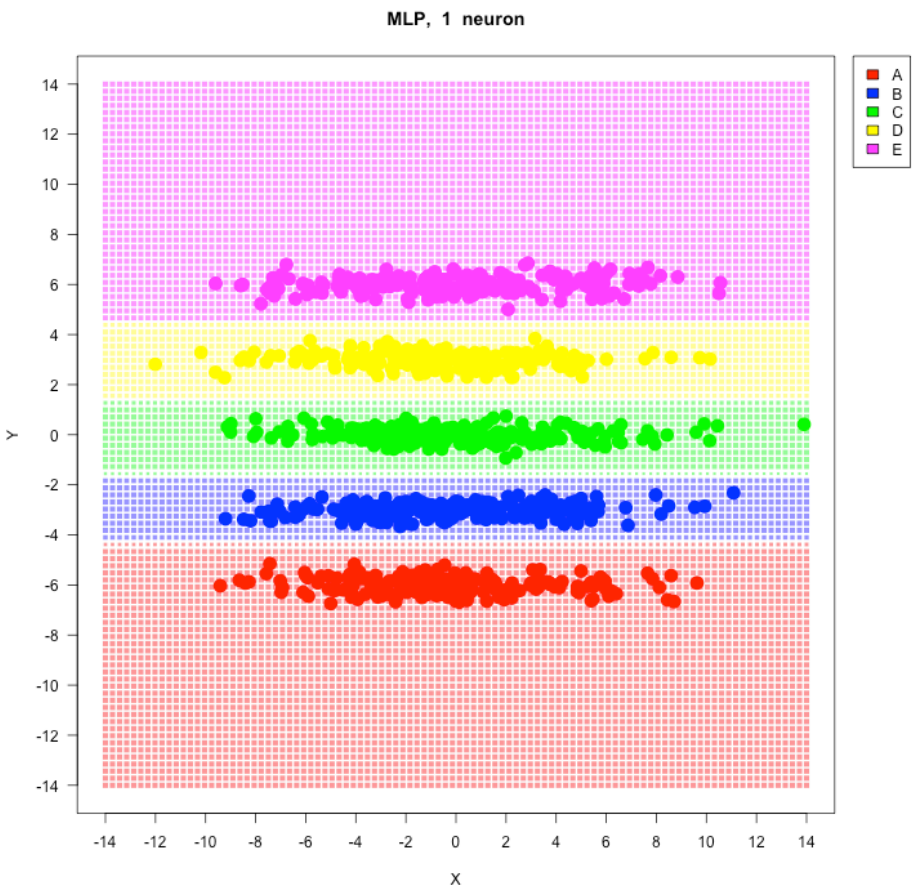


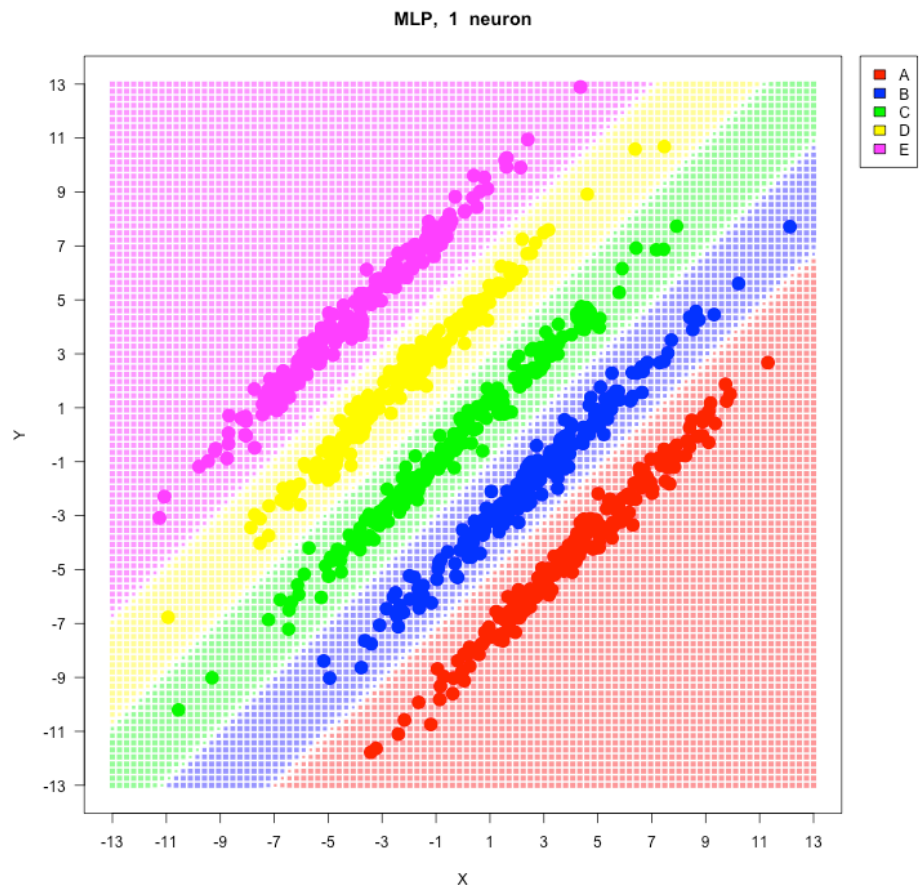
MLPs with a single neuron, single hidden layer



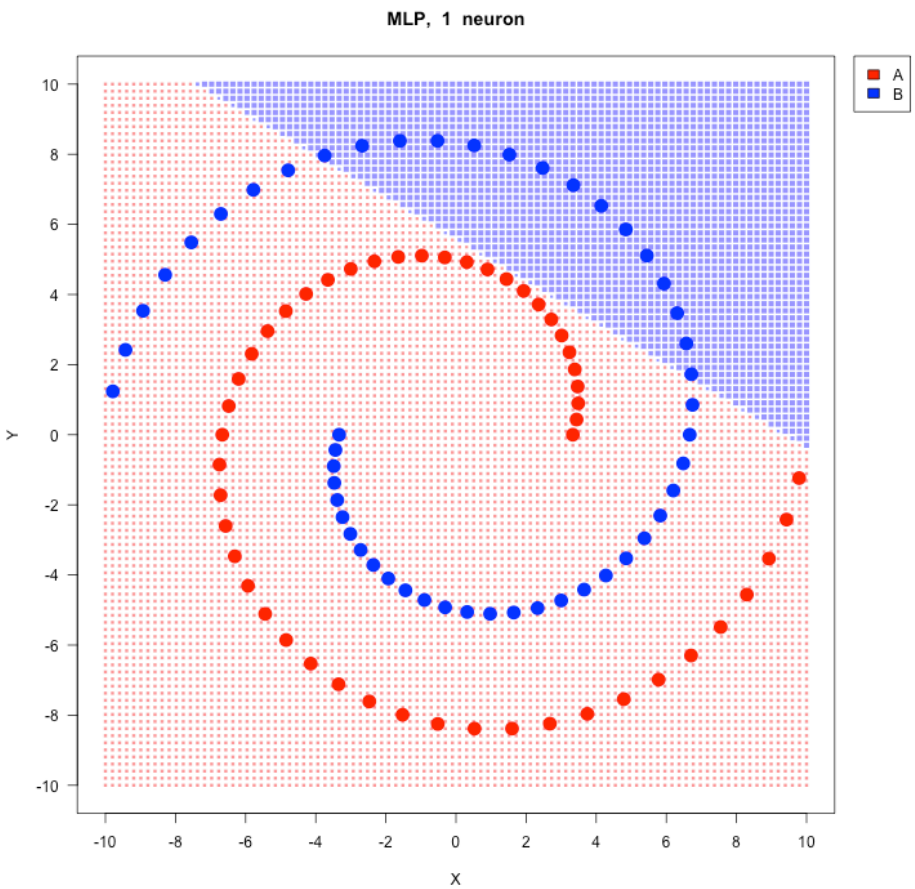


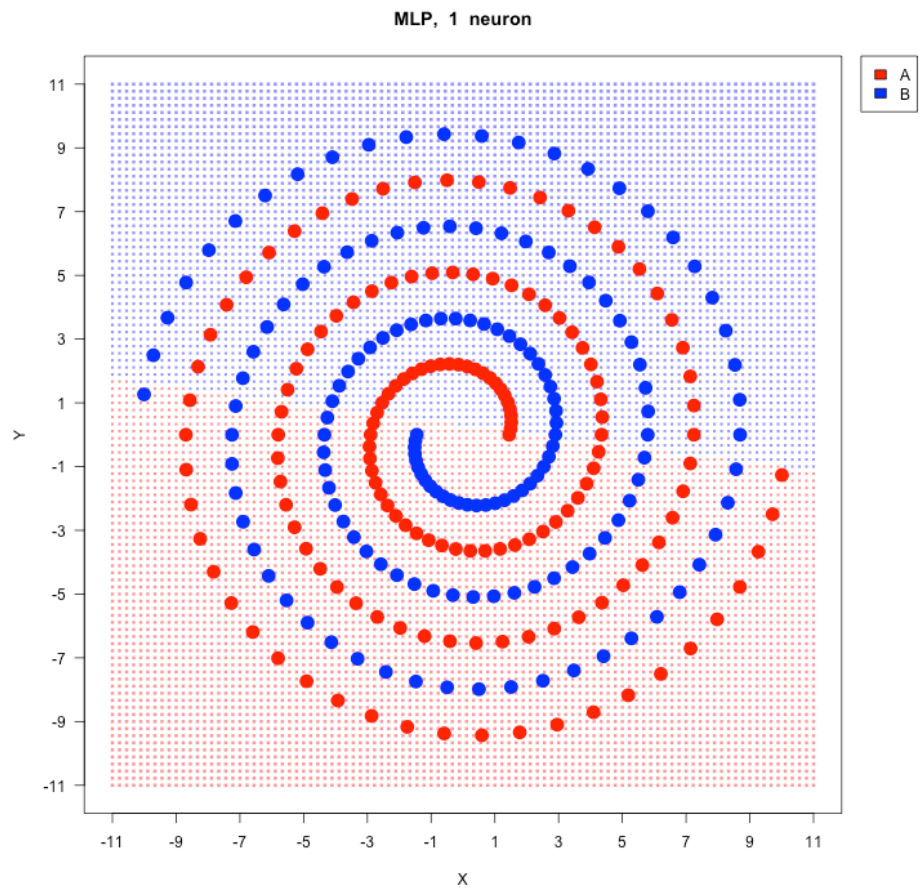
MLPs with a single neuron, single hidden layer





MLPs with a single neuron, single hidden layer

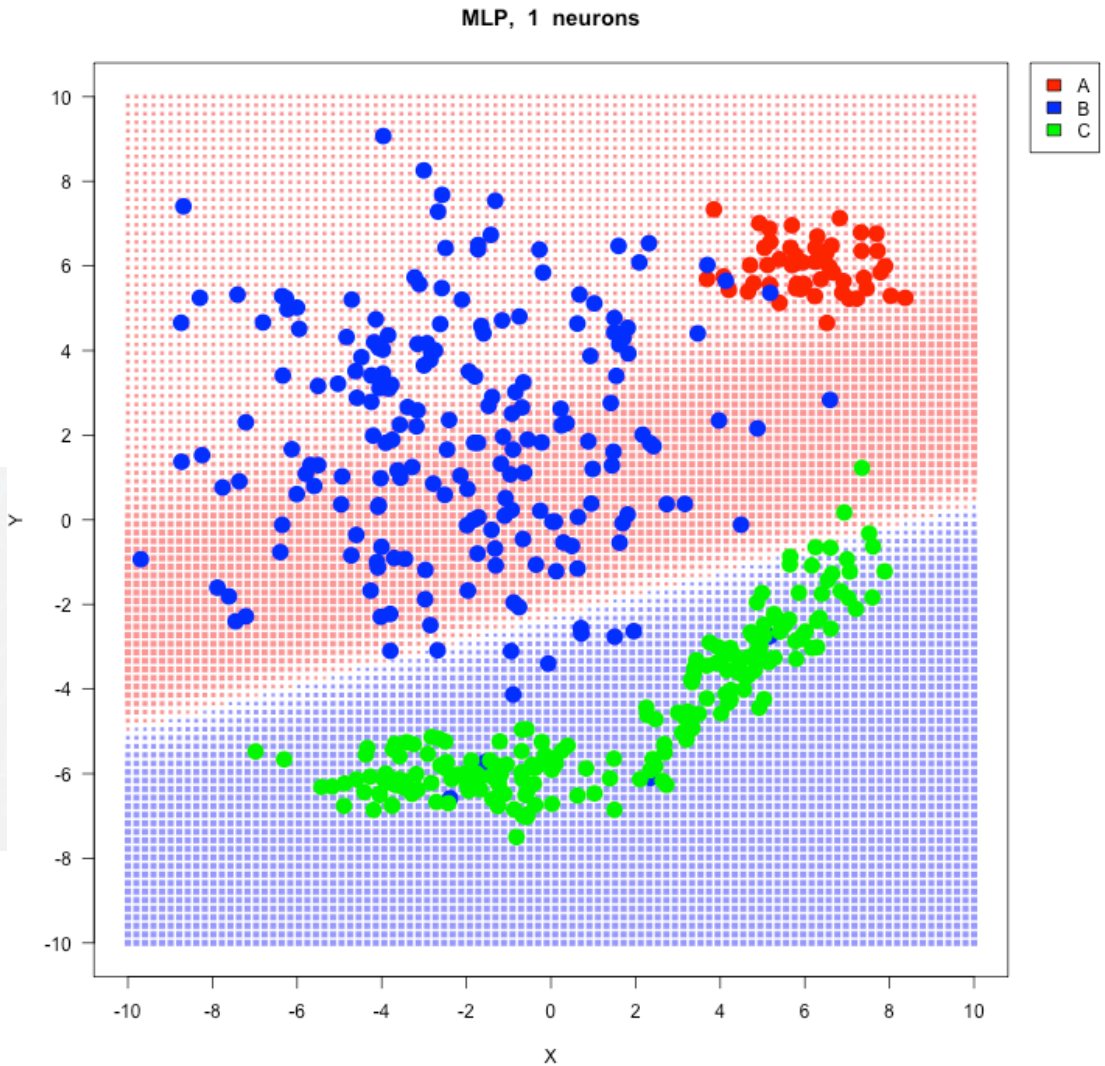




Classification: Neural Networks (MLPs)

- 1 Neuron in the hidden layer~~: unable to calculate the confusion matrix~~

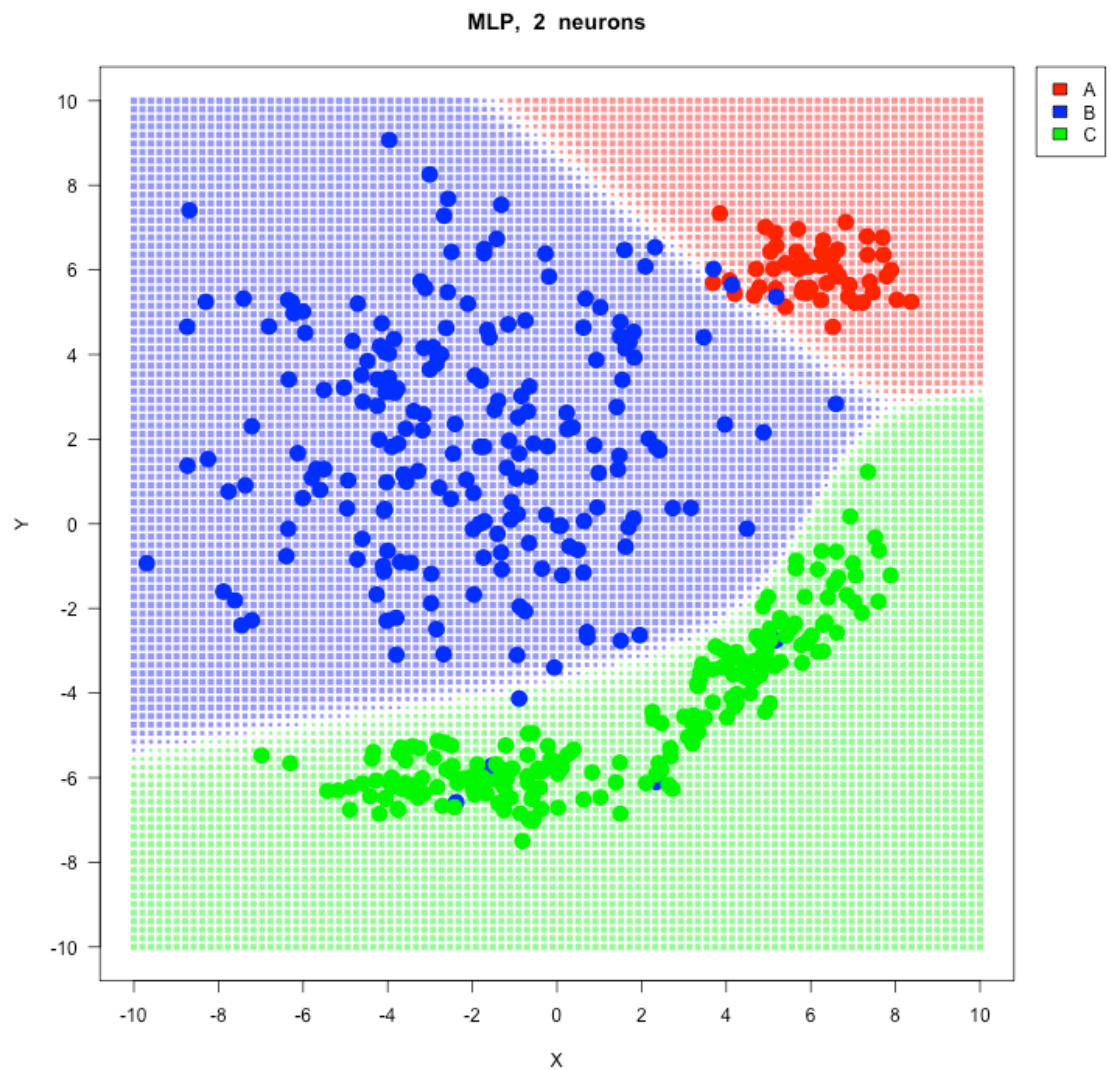
	Classified as		
	A	B	C
A	0	50	0
B	0	192	8
C	0	2	198



Classification: Neural Networks (MLPs)

- 2 Neurons in the hidden layer
- Accuracy: 0.9777778

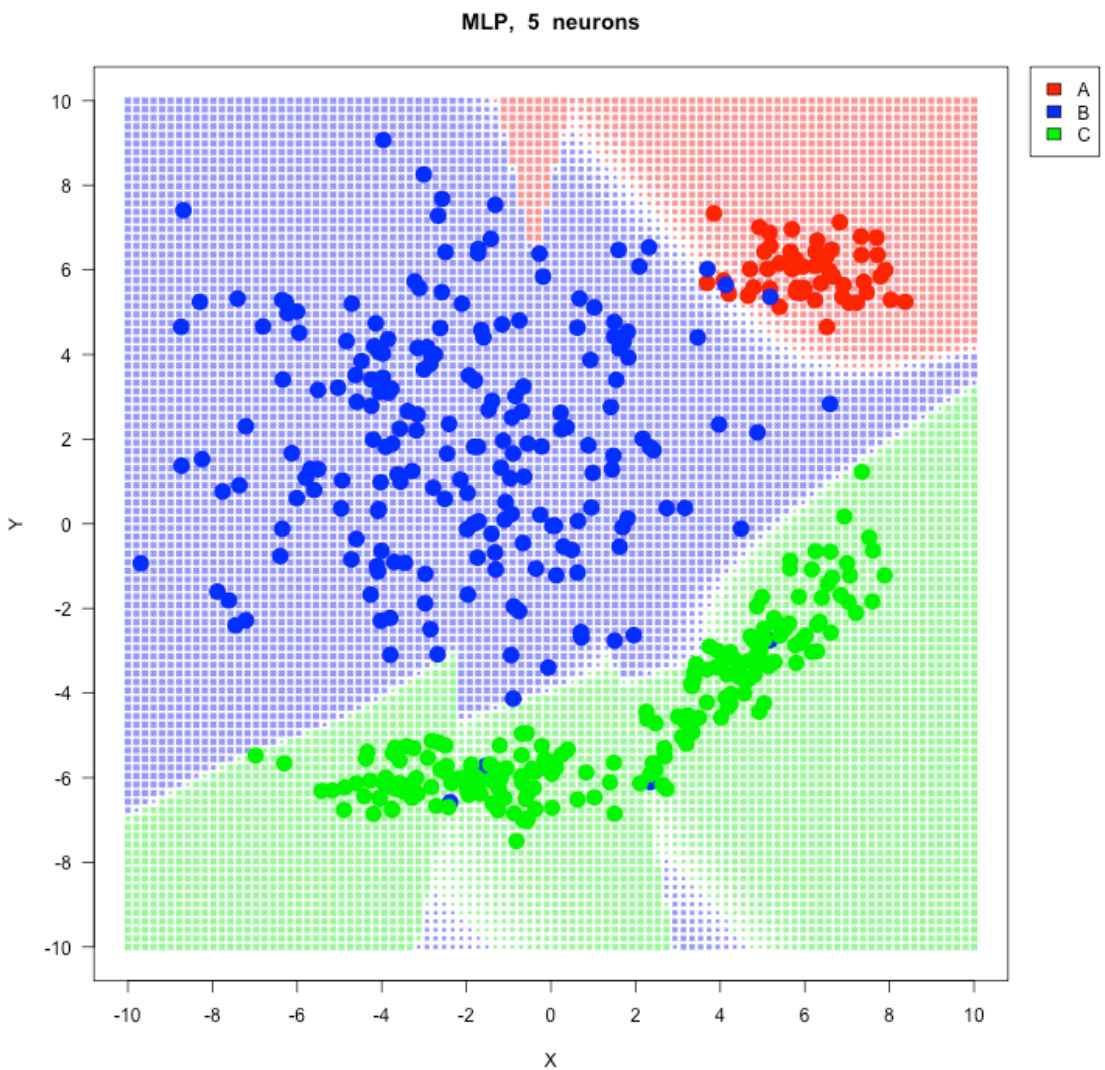
	Classified as		
	A	B	C
A	48	2	0
B	3	192	5
C	0	0	200



Classification: Neural Networks (MLPs)

- 5 Neurons in the hidden layer
- Accuracy: 0.9822222

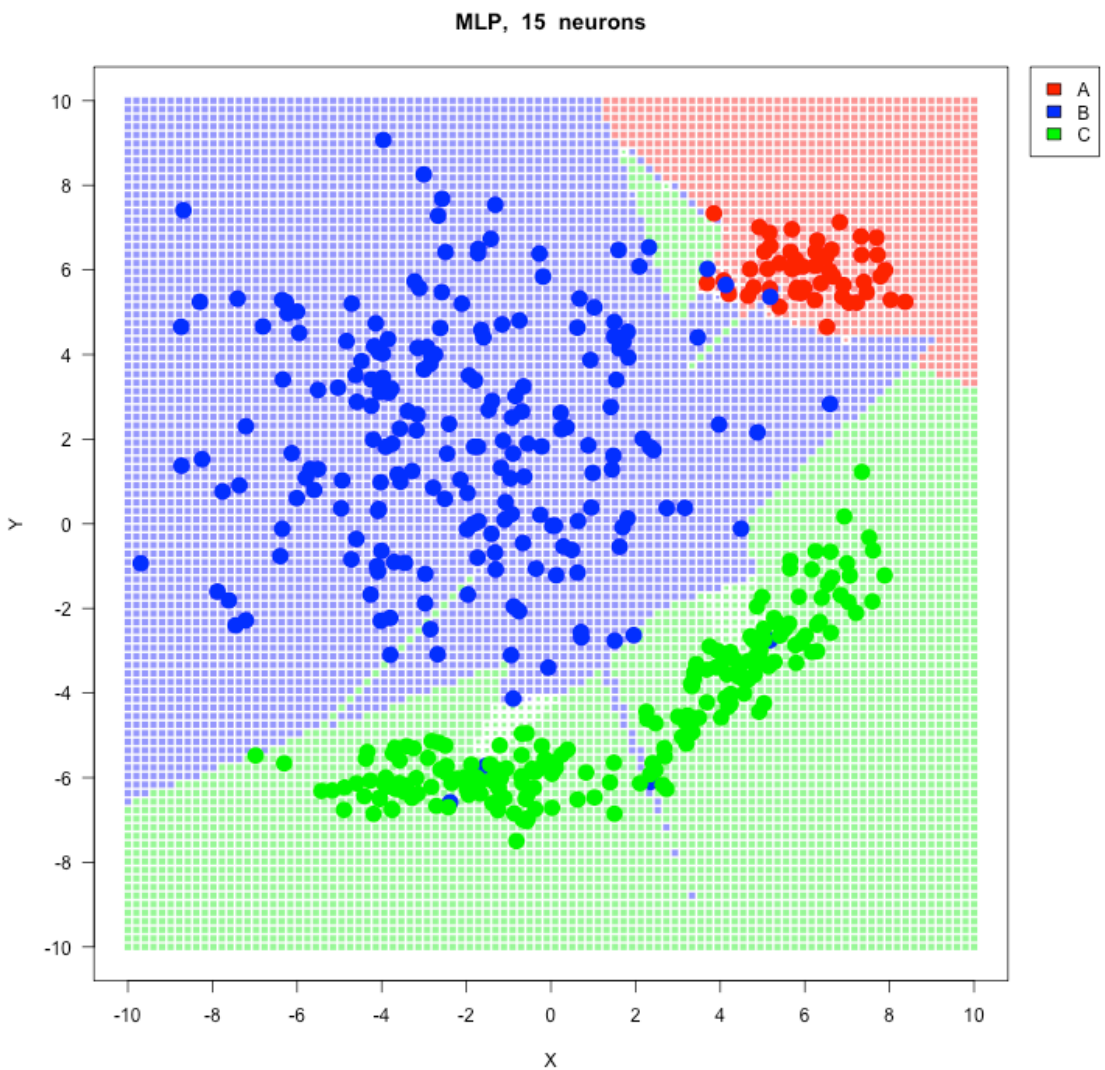
	Classified as		
	A	B	C
A	49	1	0
B	3	193	4
C	0	0	200



Classification: Neural Networks (MLPs)

- 15 Neurons in the hidden layer
- Accuracy: 0.9911111

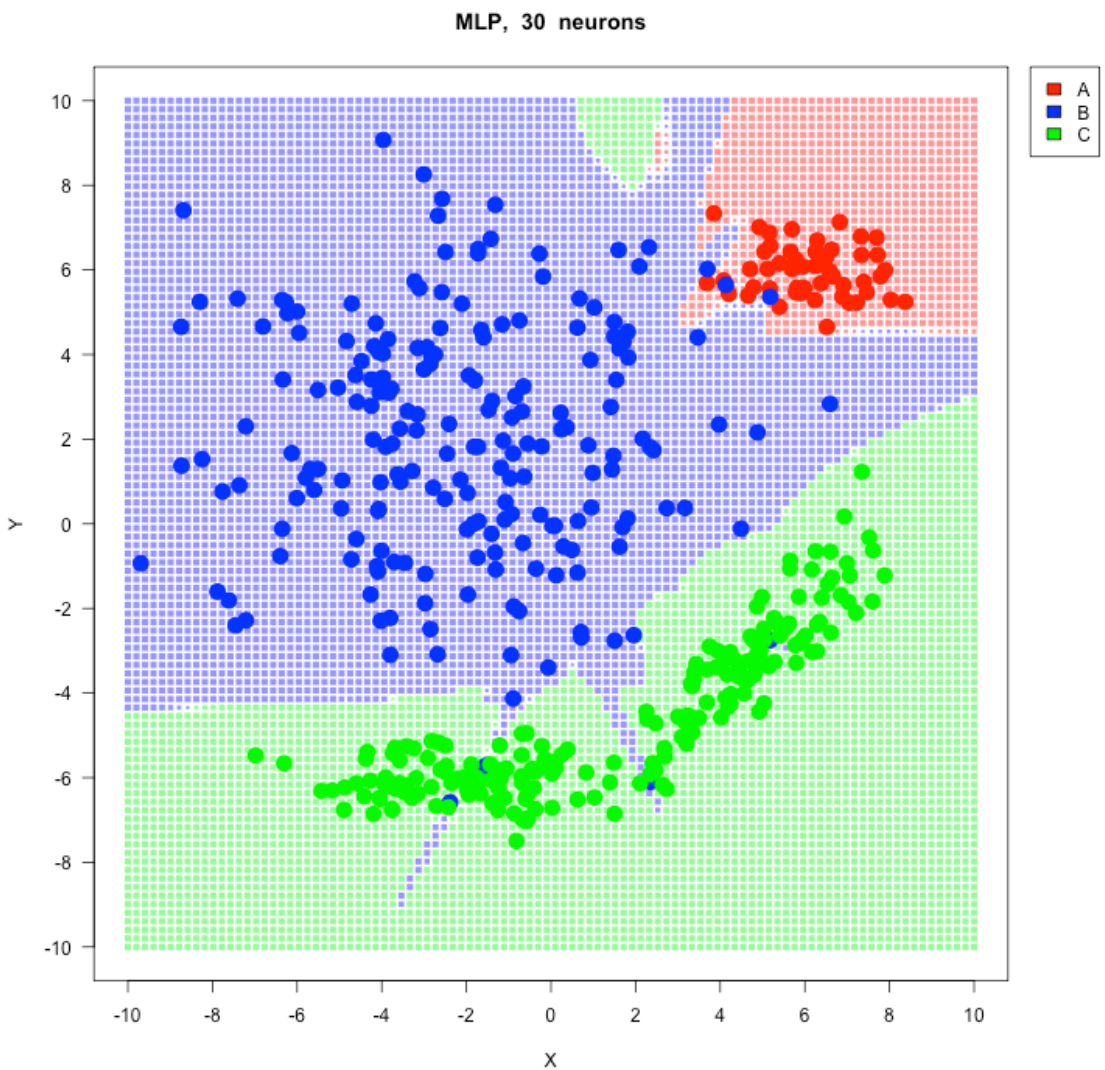
	Classified as		
	A	B	C
A	50	0	0
B	1	196	3
C	0	0	200



Classification: Neural Networks (MLPs)

- 30 Neurons in the hidden layer
- Accuracy: 0.9911111

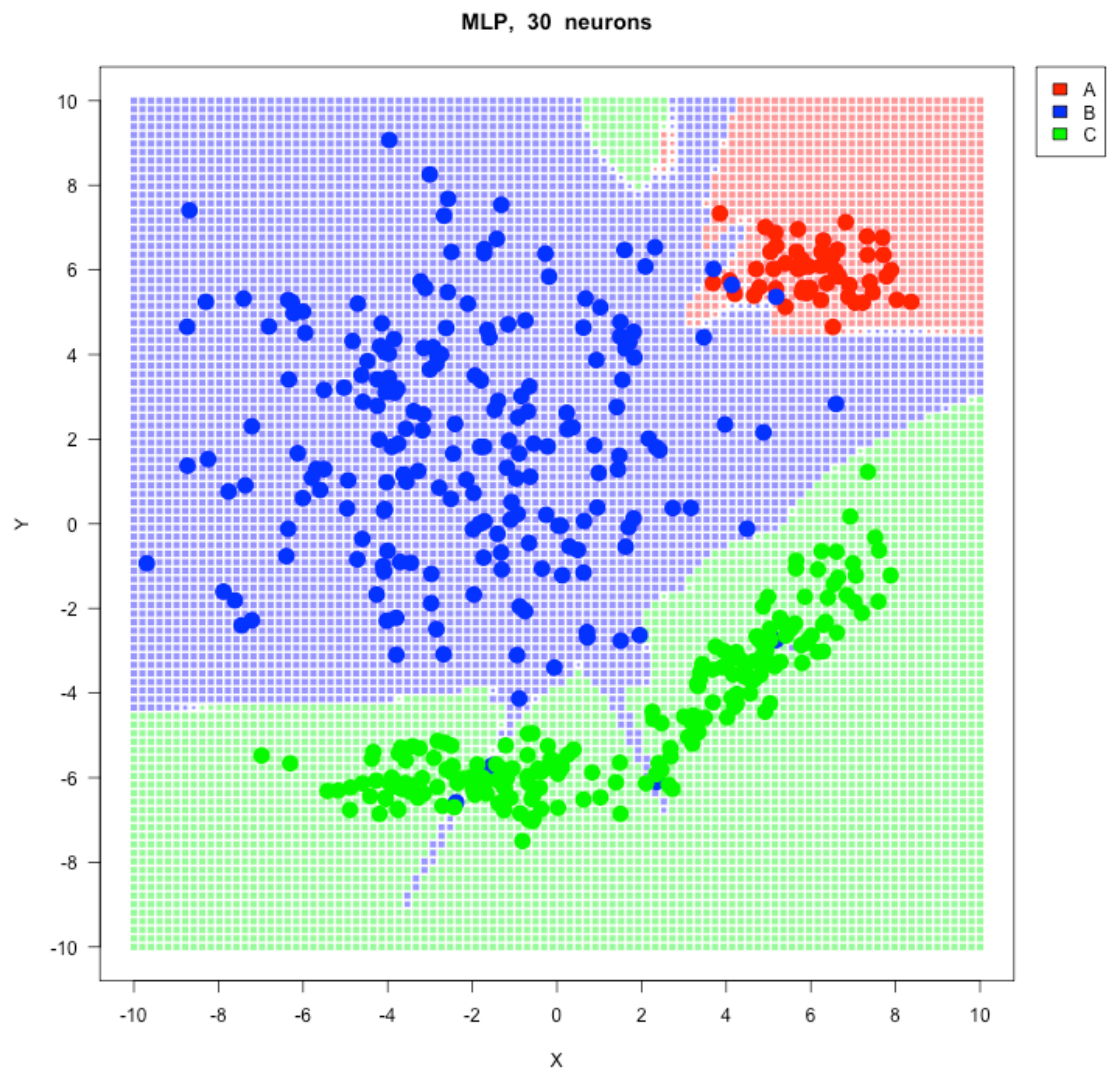
	Classified as		
	A	B	C
A	50	0	0
B	2	197	1
C	0	1	199



Classification: Neural Networks (MLPs)

- 60 Neurons in the hidden layer
- Accuracy: 0.9933333

	Classified as		
	A	B	C
A	50	0	0
B	1	197	2
C	0	0	200



Missing on this version

- Random Forests
  - Ensembles of trees, outputs mode of classes.
- SVM (Support Vector Machines)
  - A binary classification method, can be combined to use with N classes.

What about My Data?

Do I have a classification problem?

- Do you need to predict a class from observational data?
- Do you need to explain the data (rules, trees)?
- Start organizing the data!