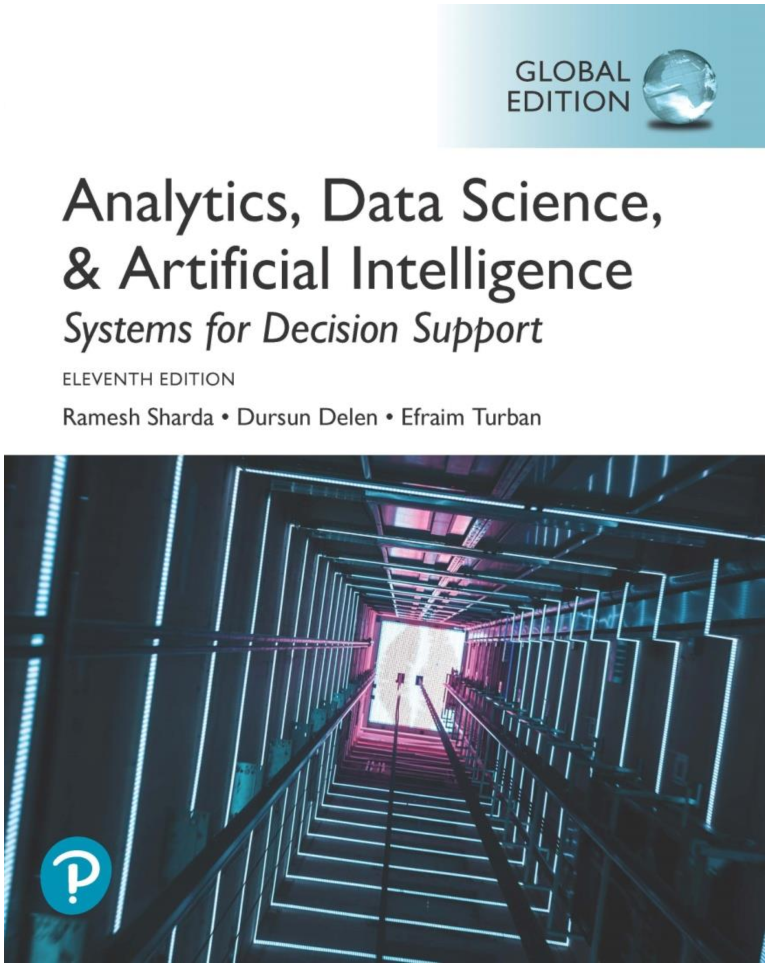
Analytics, Data Science and AI: Systems for Decision Support

Eleventh Edition, Global Edition



Chapter 5

Machine-Learning Techniques for Predictive Analytics

Learning Objectives

5.1 Understand the basic concepts and definitions of artificial neural networks (ANN)

5.2 Understand the concept and structure of support vector machines (SVM)

5.3 Understand the concept and formulation of *k*-nearest neighbor (*k*NN) algorithm

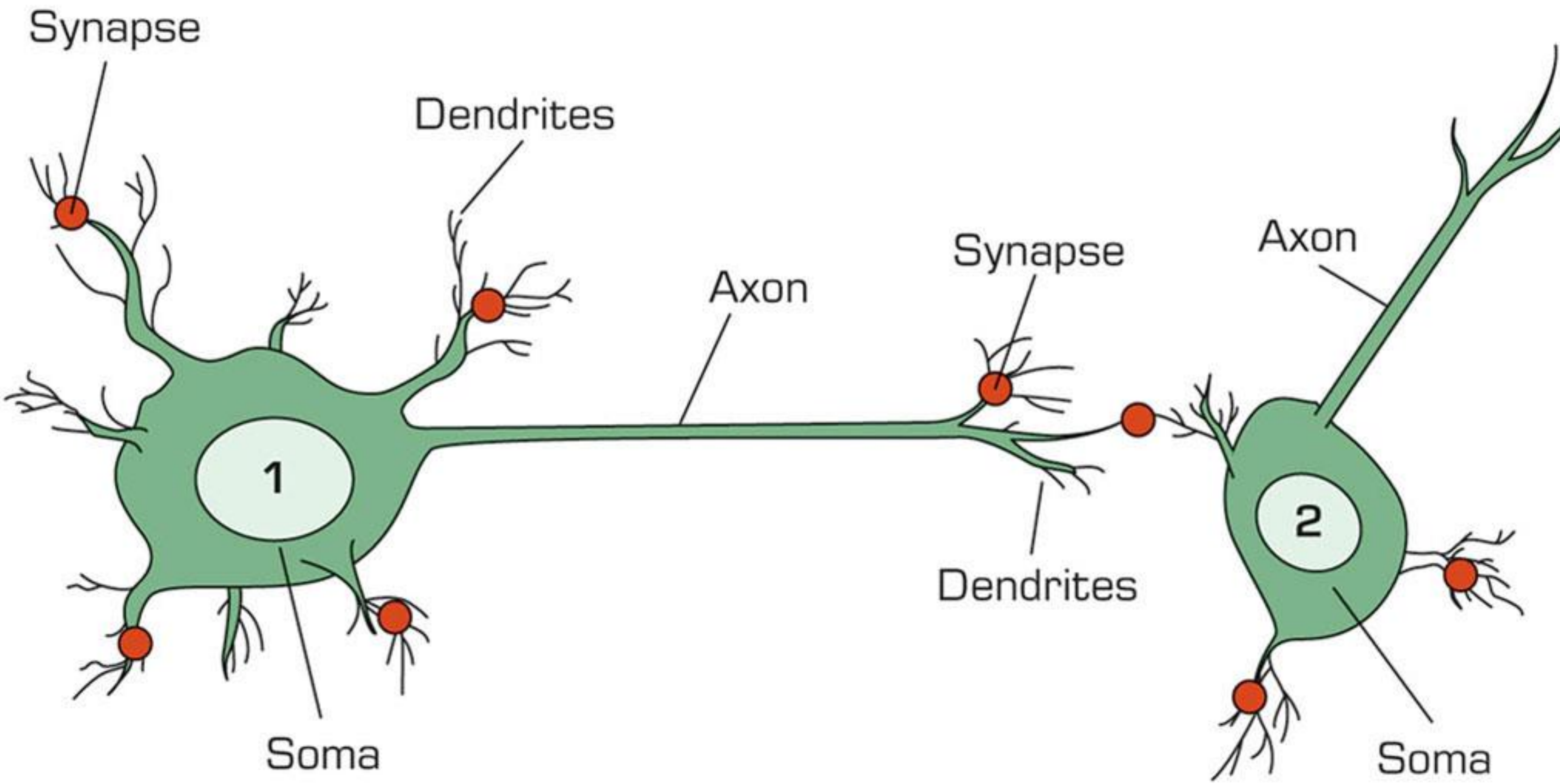
5.4 Understand the basic principles of Bayesian learning and Naïve Bayes algorithm

5.5 Understand different types of ensemble models and their pros and cons in predictive analytics

Neural Network

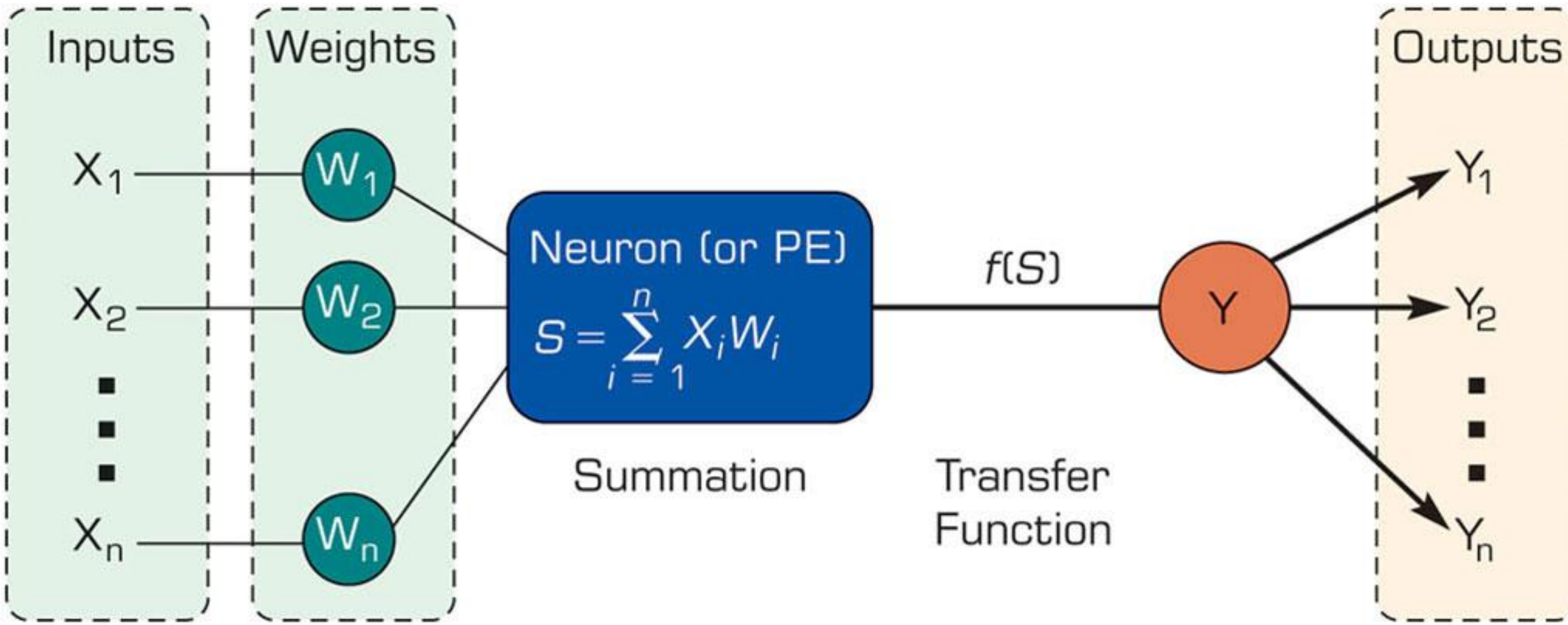
- Neural networks (NN): a human brain metaphor for information processing
- Neural computing
- Artificial neural network (ANN)
- Many uses for ANN for
  - pattern recognition, forecasting, prediction, and classification
- Many application areas
  - finance, marketing, manufacturing, operations, information systems, and so on

Biological Neural Networks



- Two interconnected brain cells (neurons)

Processing Information in ANN



- A single neuron (processing element – PE) with inputs and outputs

Biology Analogy

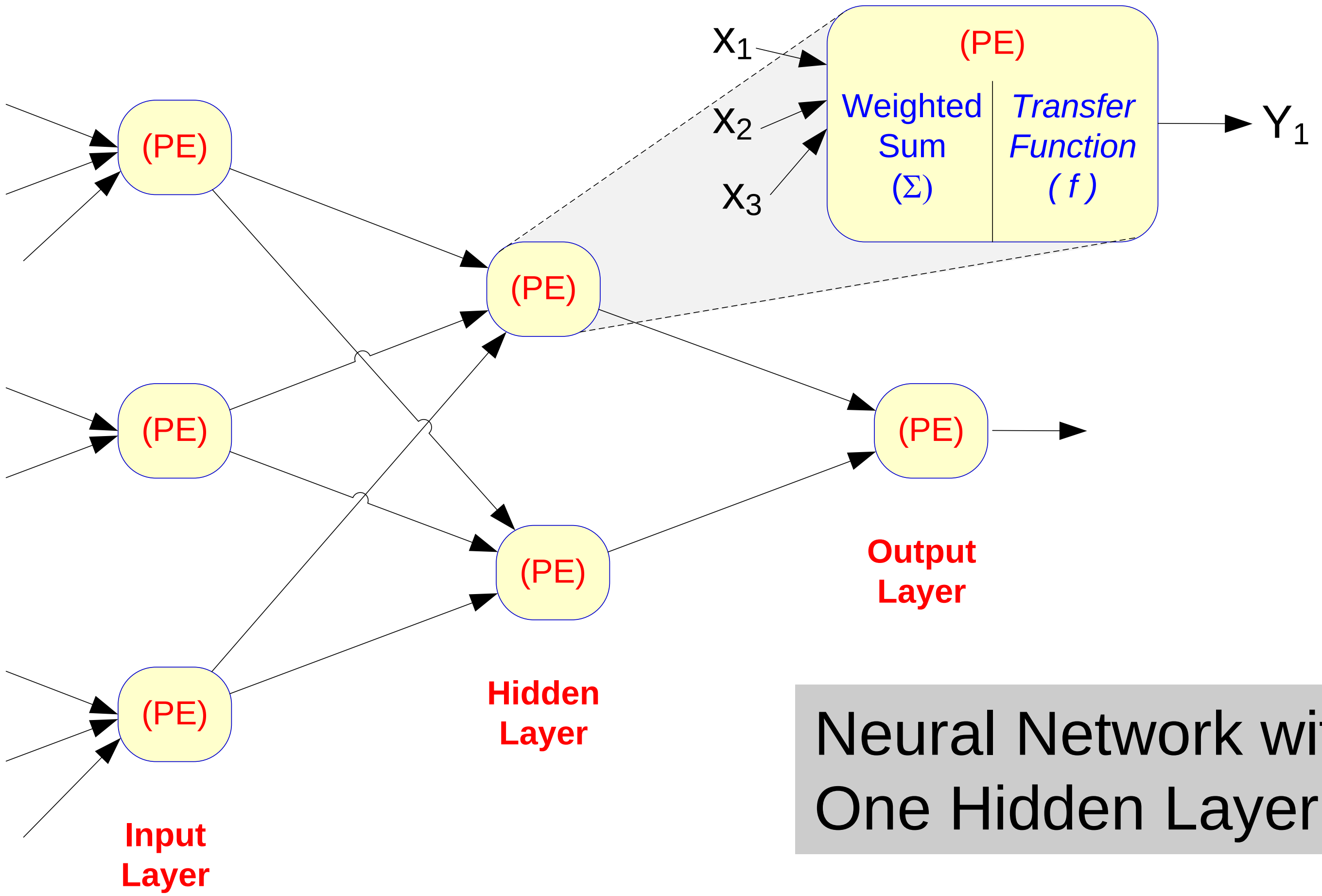
Biological	Artificial
Soma	Node
Dendrites	Input
Axon	Output
Synapse	Weight
Slow	Fast
Many neurons (10^9)	Few neurons (a dozen to hundreds of thousands)

Pearson

Elements of ANN

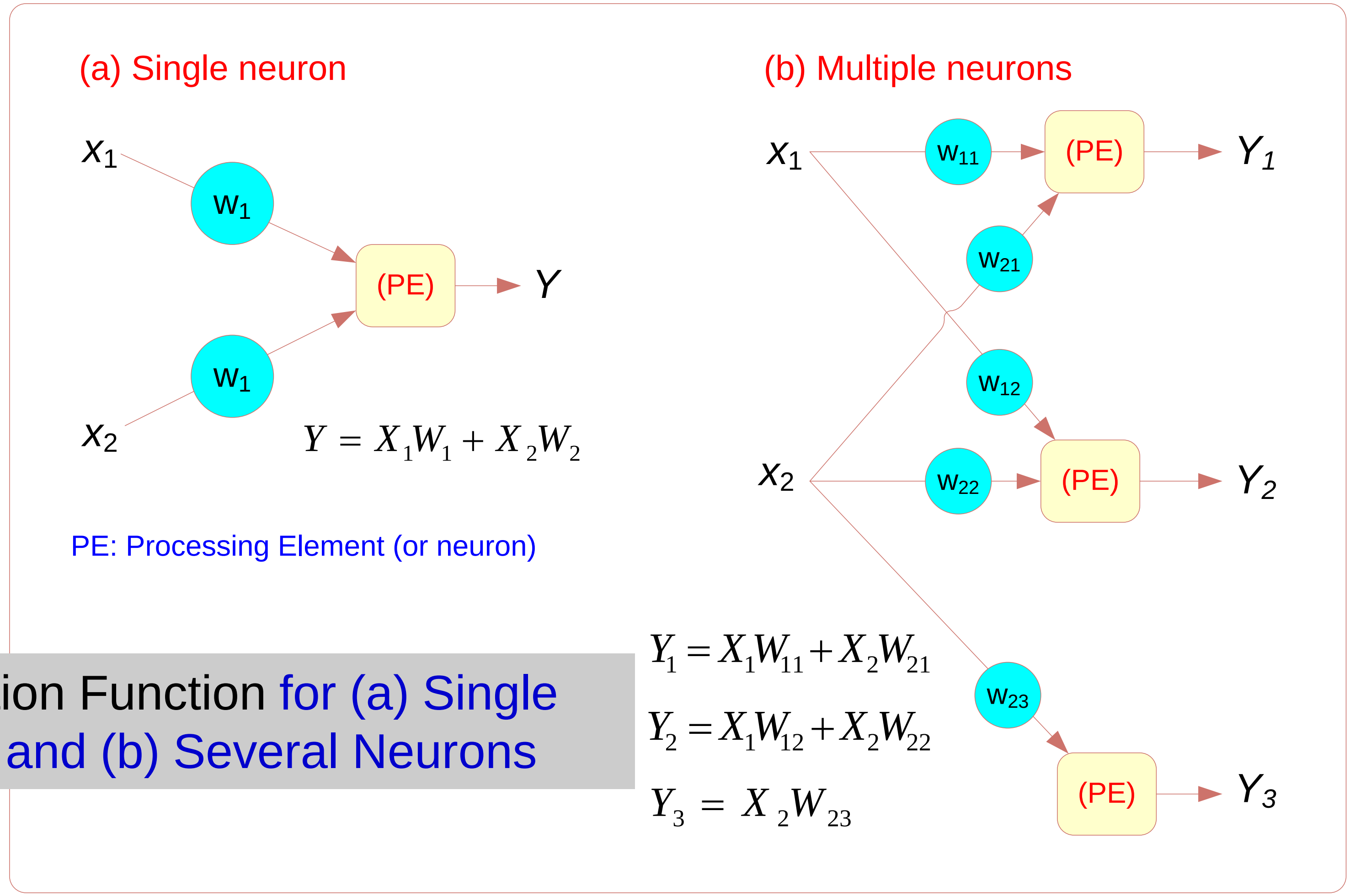
- Processing element (PE)
- Network architecture
  - Hidden layers
  - Parallel processing
- Network information processing
  - Inputs
  - Outputs
  - Connection weights
  - Summation function

Elements of ANN



Neural Network with One Hidden Layer

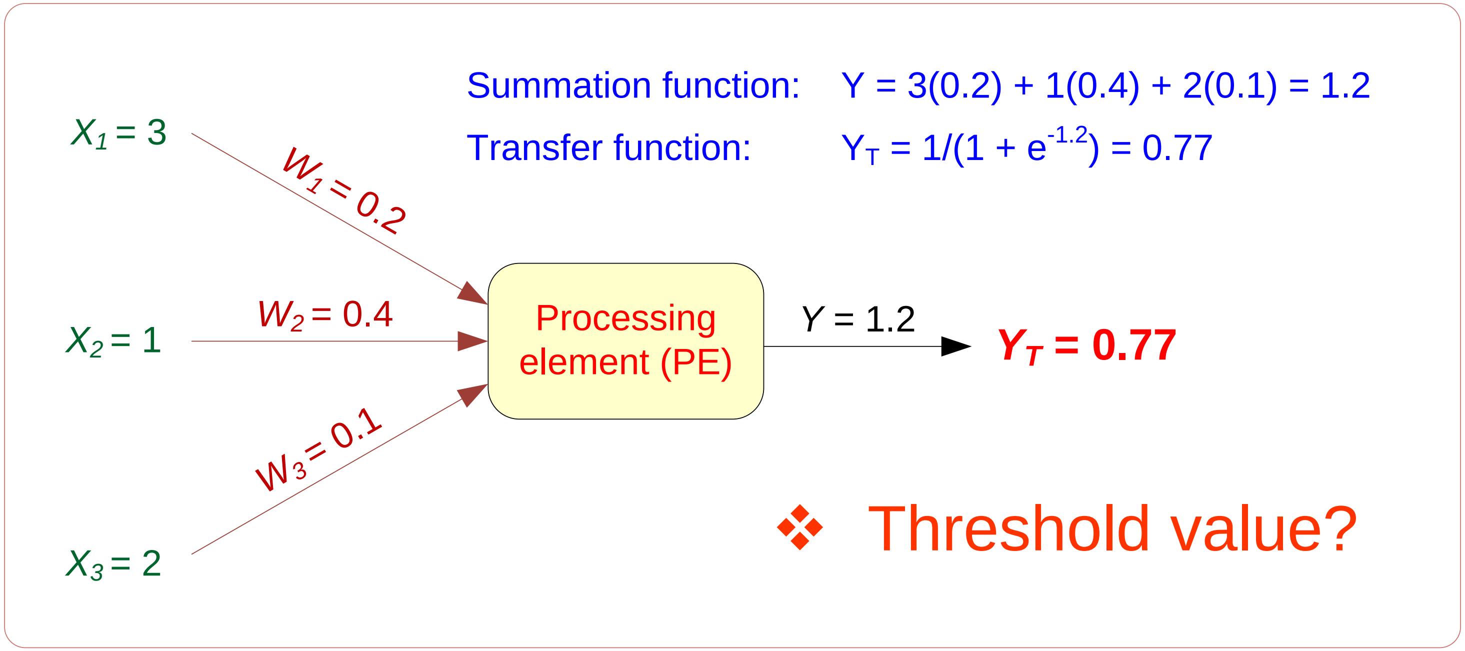
Elements of ANN



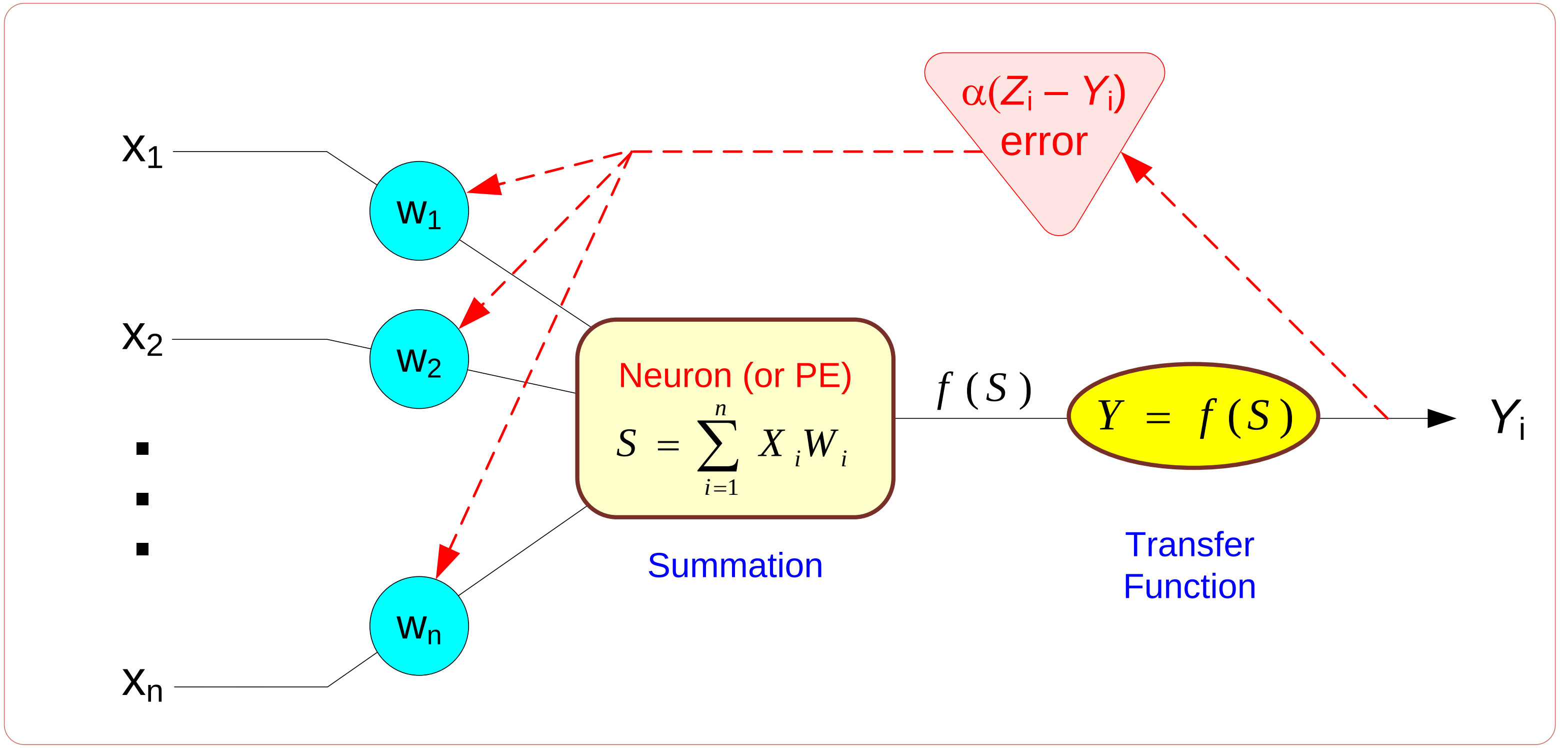
Summation Function for (a) Single Neuron, and (b) Several Neurons

Elements of ANN

- Transformation (Transfer) Function
  - Linear function
  - Sigmoid (logical activation) function [0 1]



Backpropagation Learning



Backpropagation Learning

- The learning algorithm procedure
  1. Initialize weights with random values and set other network parameters: w_1, w_2, …
  2. Read in the inputs and the desired outputs: x_1, x_2, …, z
  3. Compute the actual output (by working forward through the layers):), $y= w_1 * x_1 + w_2 * x_2$, 1 if y>0.5
  4. Compute the error (difference between the actual and desired output): $d = z-y$
  5. Change the weights by working backward through the hidden layers: $w_i(\text{updated}) = w_i(\text{initial}) + \text{alpha} * d * x_i$
  6. Repeat steps 2-5 until weights stabilize

How a Network Learns

- Example: single neuron that learns the inclusive OR operation

	Inputs		
Case	X_1	X_2	Desired Results
1	0	0	0
2	0	1	1 (positive)
3	1	0	1 (positive)
4	1	1	1 (positive)

Example

				Initial Weights				Updated Weights	
Step	X_1	X_2	Z	W_1	W_2	Y	*d*	W_1	W_2
1	0	0	0	0.1	0.3	0	0	0.1	0.3
	0	1	1	0.1	0.3	0	1	0.1	0.5
	1	0	1	0.1	0.5	0	1	0.3	0.5
	1	1	1	0.3	0.5	1	0	0.3	0.5
2	0	0	0	0.3	0.5	0	0	0.3	0.5
	0	1	1	0.3	0.5	0	1	0.3	0.7
	1	0	1	0.3	0.7	0	1	0.5	0.7
	1	1	1	0.5	0.7	1	0	0.5	0.7
3	0	0	0	0.5	0.7	0	0	0.5	0.7
	0	1	1	0.5	0.7	1	0	0.5	0.7
	1	0	1	0.5	0.7	0	1	0.7	0.7
	1	1	1	0.7	0.7	1	0	0.7	0.7
4	0	0	0	0.7	0.7	0	0	0.7	0.7
	0	1	1	0.7	0.7	1	0	0.7	0.7
	1	0	1	0.7	0.7	1	0	0.7	0.7
	1	1	1	0.7	0.7	1	0	0.7	0.7

Alpha (*a*)=0.2; $Y=w_1*x_1+w_2*x_2$, 1 if y>0.5; *d*=Z-Y;

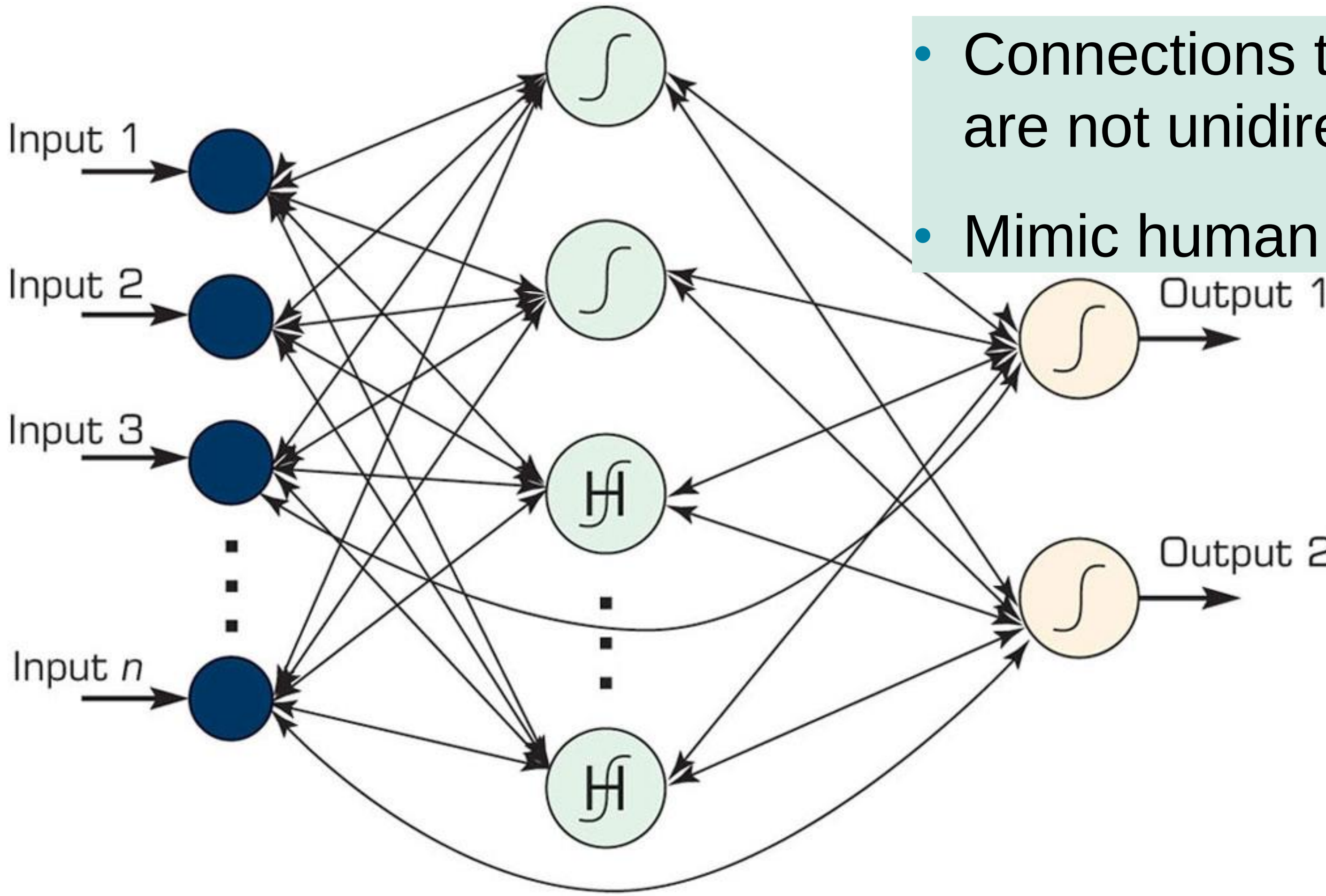
$w_i(\text{update}) = w_i(\text{initial}) + \text{alpha} * d * x_i$

Neural Network Architectures

- Architecture of a neural network is driven by the task it is intended to address
  - Classification, regression, clustering, general optimization, association
- Feedforward, multi-layered perceptron with backpropagation learning algorithm
- Other ANN Architectures – Recurrent, self-organizing feature maps, hopfield networks, …

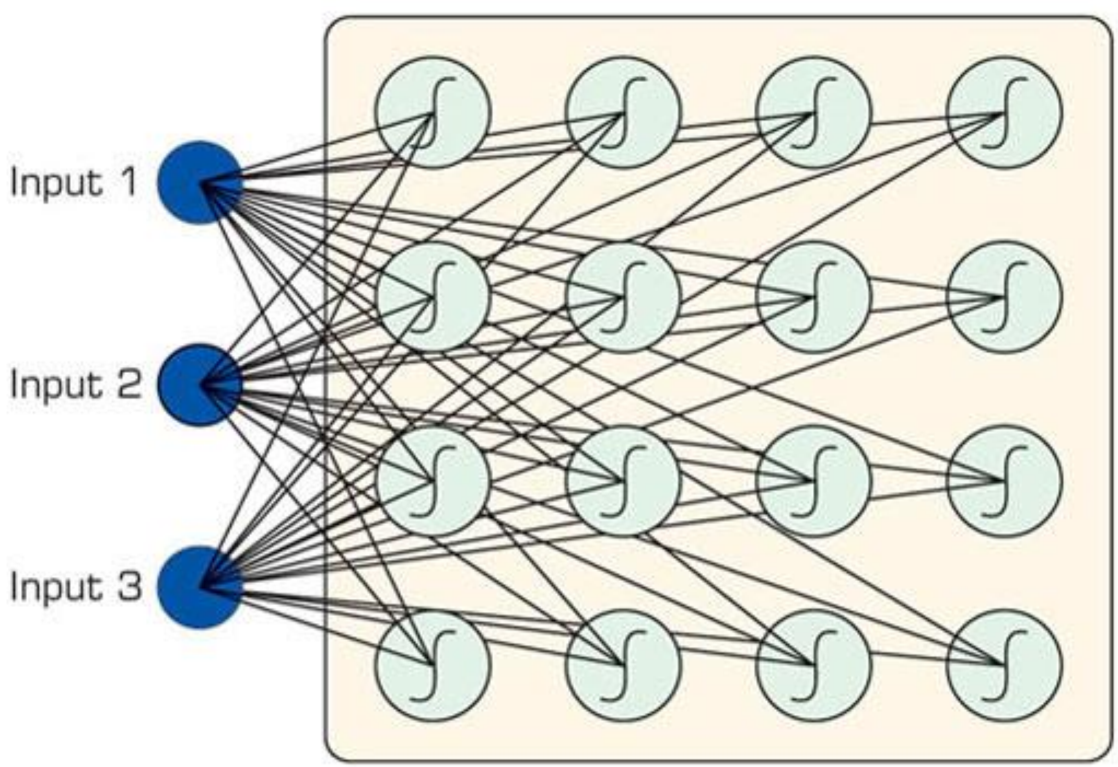
Neural Network Architectures Recurrent Neural Networks

- Connections to nodes are not unidirectional
- Mimic human brain



*H: indicates a "hidden" neuron without a target output

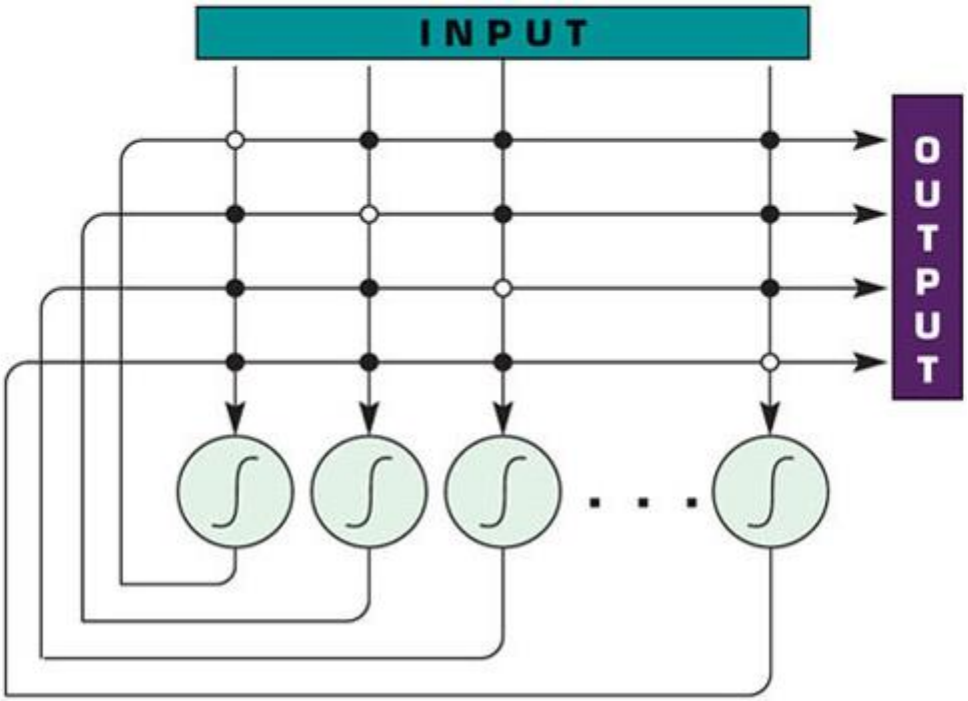
Other Popular ANN Paradigms Self Organizing Maps (SOM)



(a) Kohonen Network (SOM)

- First introduced by the Finnish Professor Teuvo Kohonen
- Applies to clustering type problems (unsupervised learning)

Other Popular ANN Paradigms Hopfield Networks



(b) Hopfield Network

- First introduced by John Hopfield
- Single layer with highly interconnected neurons
- Applies to solving complex computational problems (e.g., optimization problems)

Support Vector Machines (SVM)
(1 of 4)

- SVM are among the most popular machine-learning techniques.
- SVM belong to the family of generalized linear models… (capable of representing non-linear relationships in a linear fashion)
- SVM achieve a classification or regression decision based on the value of the linear combination of input features.
- Because of their architectural similarities, SVM are also closely associated with ANN.

Support Vector Machines (SVM) (2 of 4)

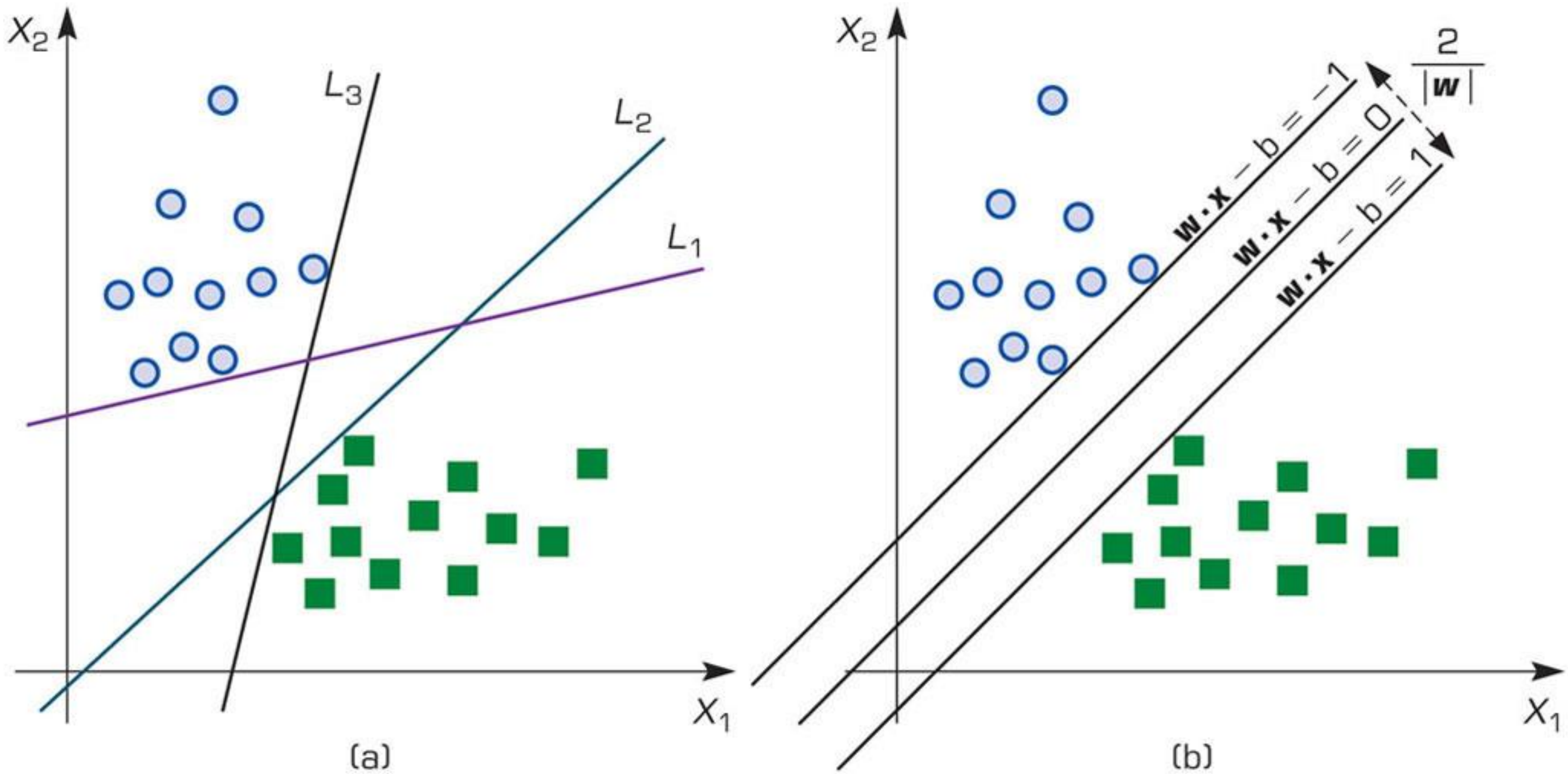
- Goal of SVM: to generate mathematical functions that map input variables to desired outputs for classification or regression type prediction problems.
  - First, SVM uses nonlinear kernel functions to transform non-linear relationships among the variables into linearly separable feature spaces.
  - Then, the maximum-margin hyperplanes are constructed to optimally separate different classes from each other based on the training dataset.
- SVM has solid mathematical foundation!

Support Vector Machines (SVM) (3 of 4)

- A hyperplane is a geometric concept used to describe the separation surface between different classes of things.
  - In SVM, two parallel hyperplanes are constructed on each side of the separation space with the aim of maximizing the distance between them.
- A kernel function in SVM uses the kernel trick (a method for using a linear classifier algorithm to solve a nonlinear problem)
  - The most commonly used kernel function is the radial basis function (RBF).

Pearson

Support Vector Machines (SVM)
(4 of 4)

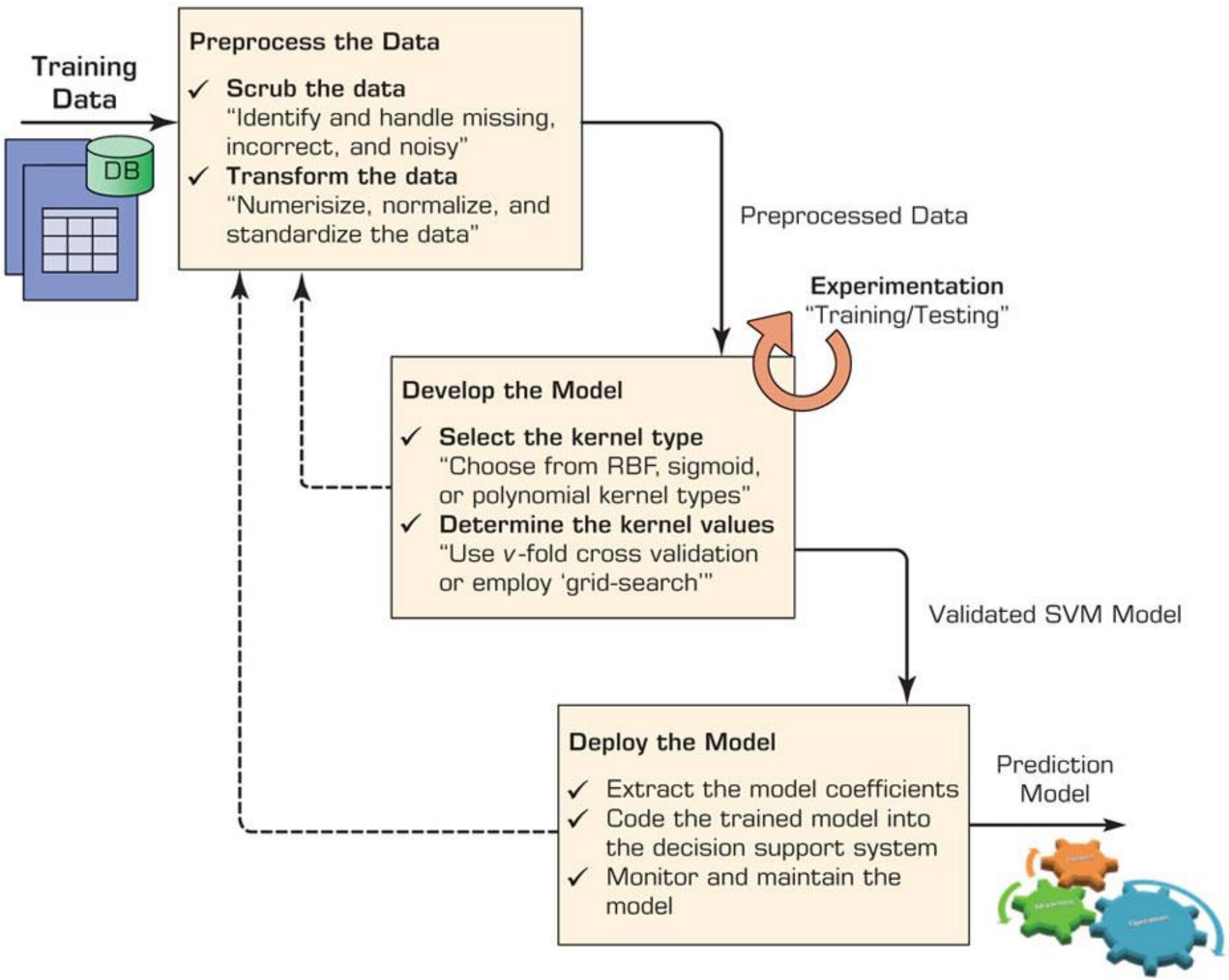


- Many linear classifiers (hyperplanes) may separate the data

How Does a SVM Works?

- Following a machine-learning process, a SVM learns from the historic cases.
- The Process of Building SVM
  1. Preprocess the data
     - Scrub and transform the data.
  2. Develop the model.
     - Select the kernel type (RBF is often a natural choice).
     - Determine the kernel parameters for the selected kernel type.
     - If the results are satisfactory, finalize the model, otherwise change the kernel type and/or kernel parameters to achieve the desired accuracy level.
  3. Extract and deploy the model.

The Process of Building a SVM

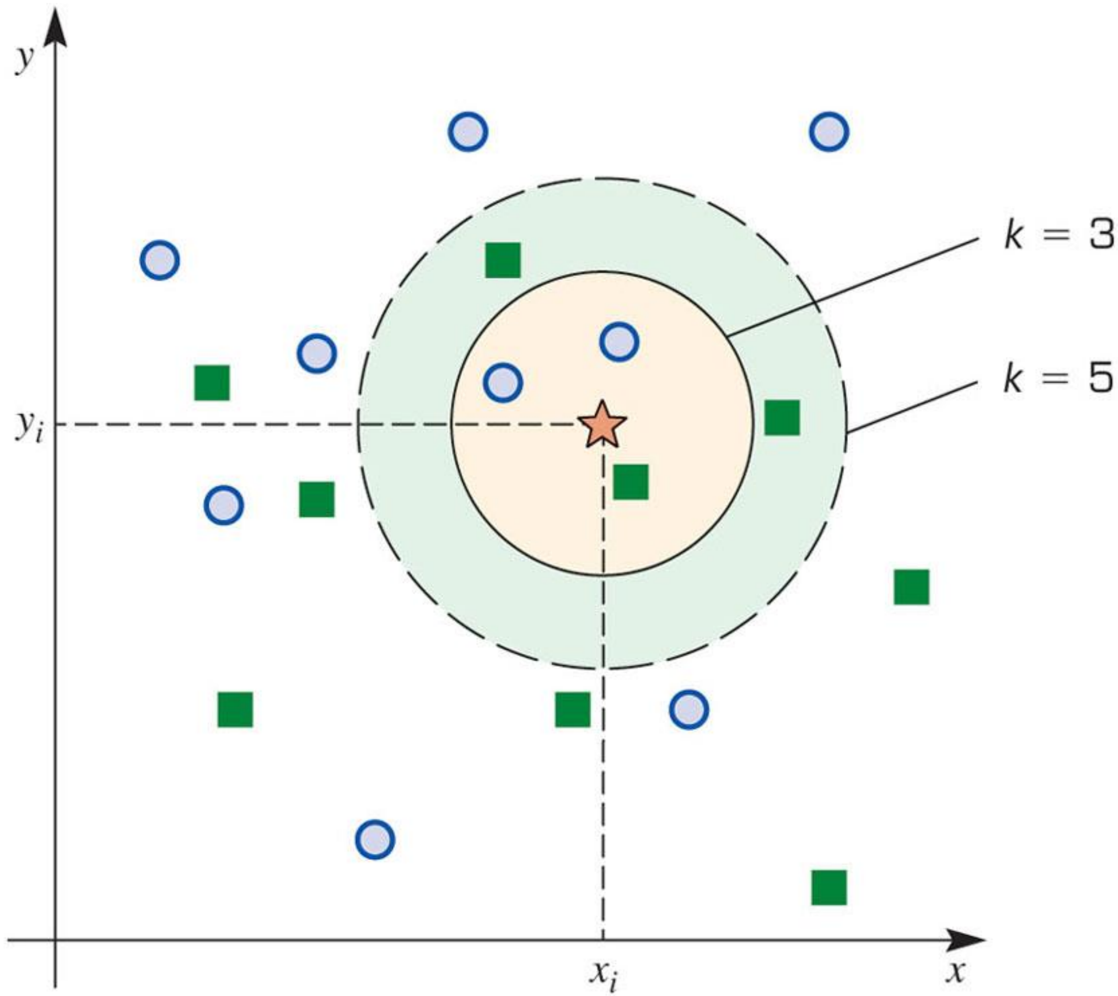


Pearson

k-Nearest Neighbor Method (k-NN) (1 of 2)

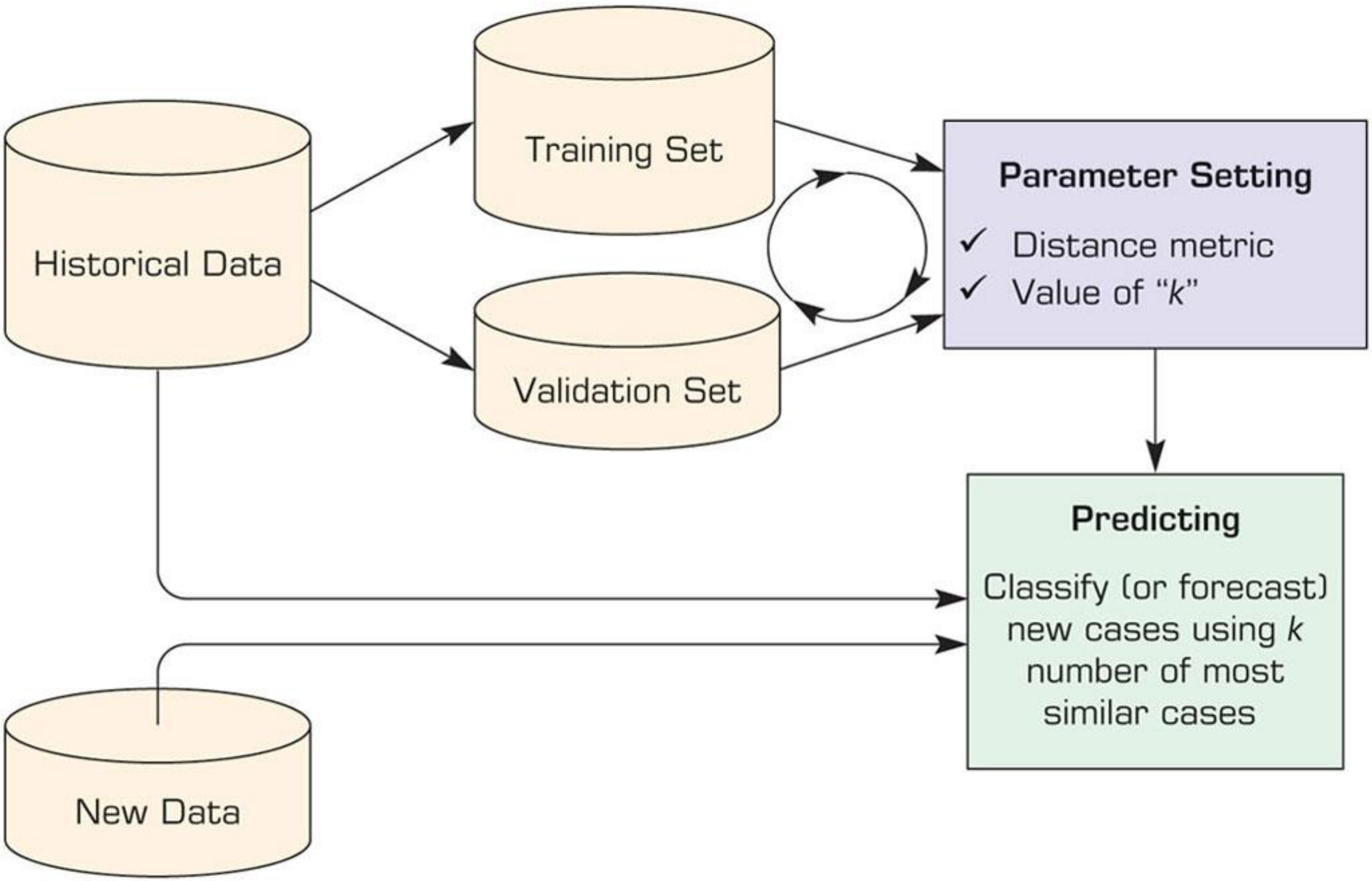
- ANNs and SVMs → time-demanding, computationally intensive iterative derivations
- *k*-NN a simplistic and logical prediction method, that produces <u>very competitive</u> results
- *k*-NN is a prediction method for classification as well as regression types (similar to ANN & SVM)
- *k*-NN is a type of instance-based learning (or lazy learning) – most of the work takes place at the time of prediction (not at modeling)
- *k* : the number of neighbors used in the model

k-Nearest Neighbor Method (k-NN) (2 of 2)



- The answer to "which class a data point belongs to?" depends on the value of *k*

The Process of *k*-NN Method



k-NN Model Parameter (1 of 2)

1. Similarity Measure: The Distance Metric

Minkowski distance

$$d(i,j) = \sqrt[q]{(|x_{i1} - x_{j1}|^q + |x_{i2} - |x_{j2}|^q + \ldots + |x_{ip} - x_{jp}|^q)}$$

If $q = 1$, then d is called Manhatten distance

$$d(i,j) = \sqrt{|x_{i1} - x_{j1}|| + \; |x_{i2} - x_{j2}| + \ldots + |x_{ip} - x_{jp}|}$$

If $q = 2$, then d is called Euclidean distance

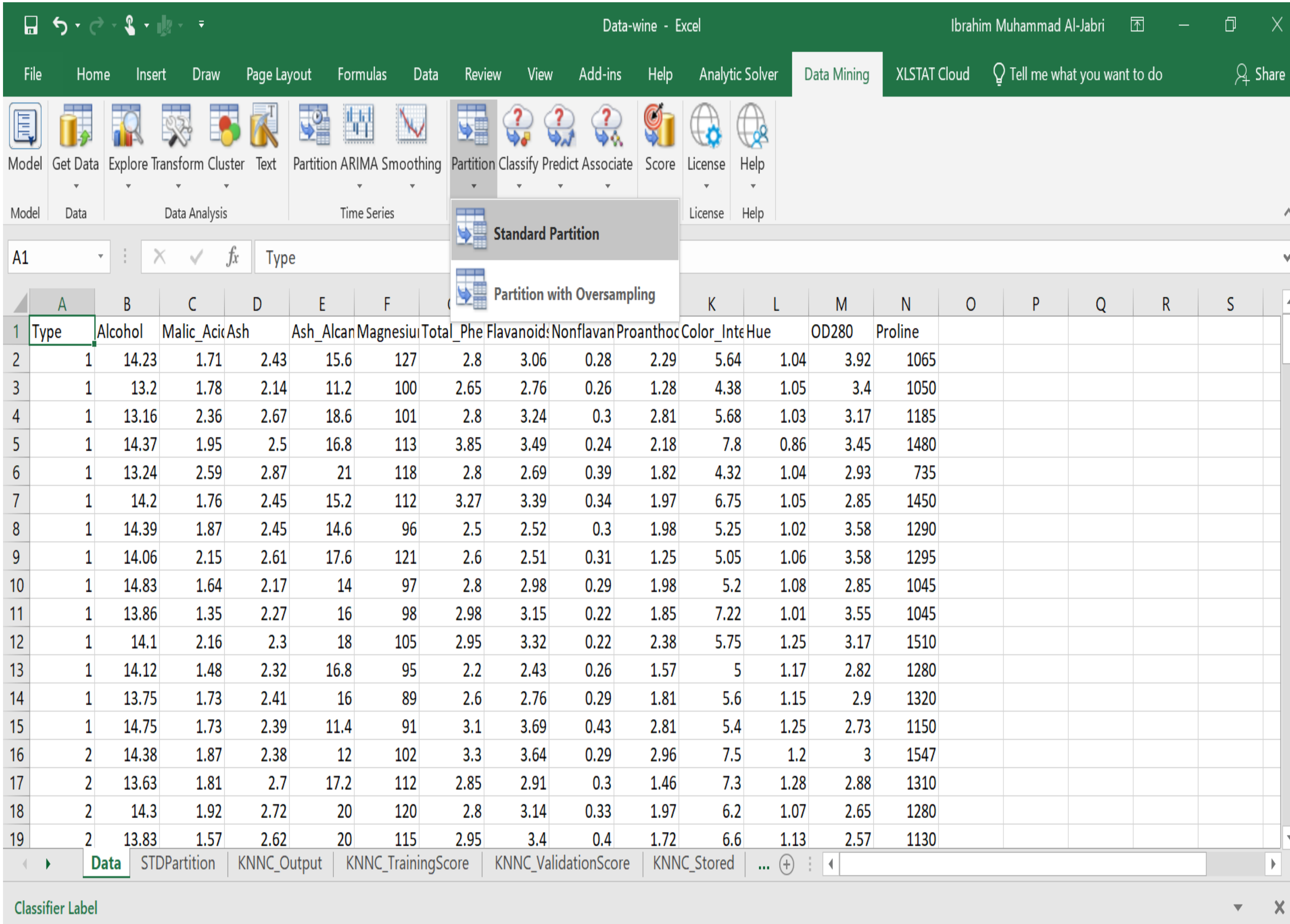
$$d(i,j) = \sqrt{(|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \ldots + |x_{ip} - x_{jp}|^2)}$$

– Numeric versus nominal values?

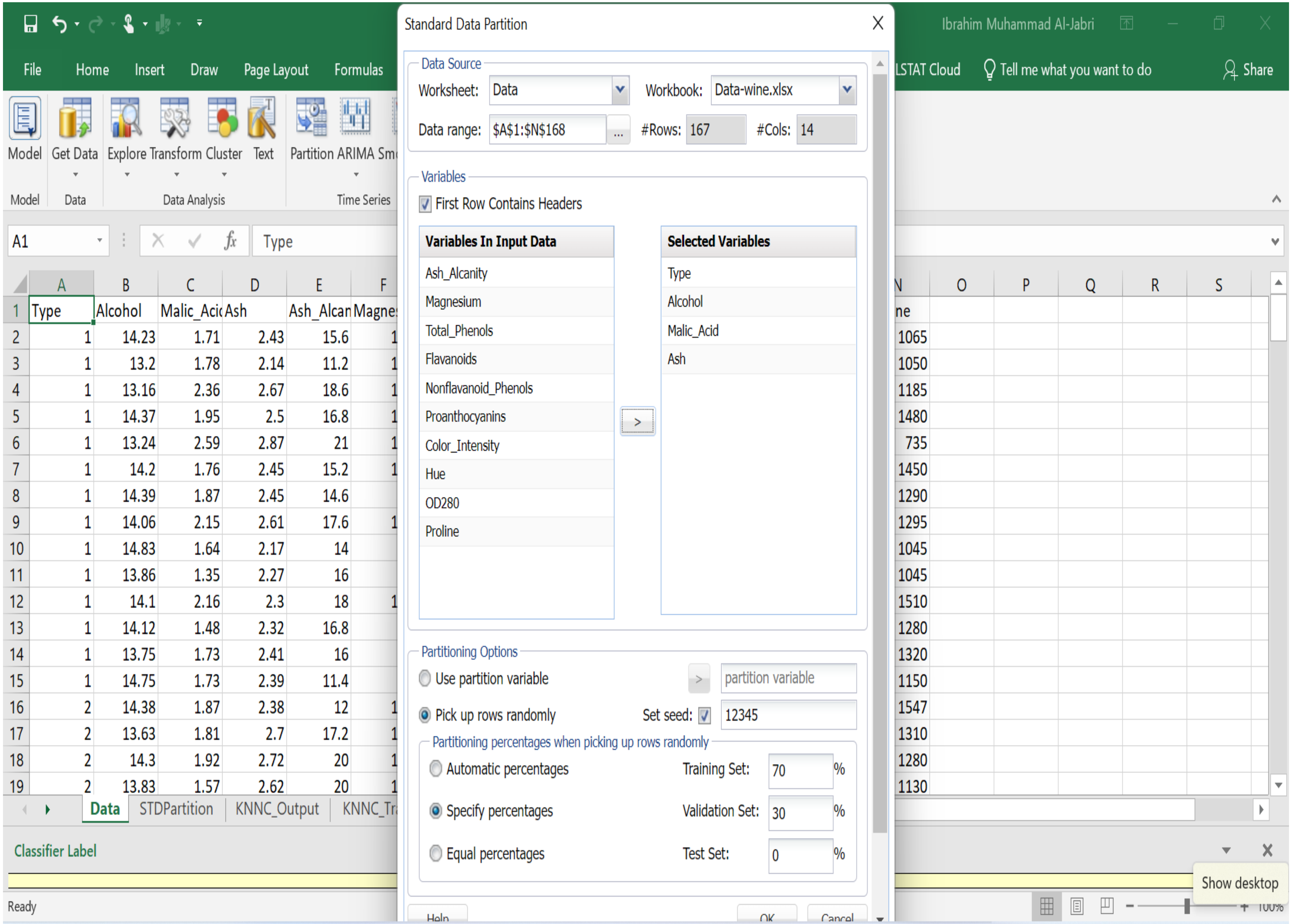
k-NN Model Parameter (2 of 2)

2. Number of Neighbors (the value of *k*)
   - The best value depends on the data
   - Larger values reduces the effect of noise but also make boundaries between classes less distinct
   - An “optimal” value can be found heuristically

- Cross Validation is often used to determine the best value for k and the distance measure



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Type	Alcohol	Malic_Acic	Ash	Ash_Alcan	Magnesiu	Total_Phe	Flavanoid	Nonflavan	Proanthoc	Color_Inte	Hue	OD280	Proline
2	1	14.23	1.71	2.43	15.6	127	2.8	3.06	0.28	2.29	5.64	1.04	3.92	1065
3	1	13.2	1.78	2.14	11.2	100	2.65	2.76	0.26	1.28	4.38	1.05	3.4	1050
4	1	13.16	2.36	2.67	18.6	101	2.8	3.24	0.3	2.81	5.68	1.03	3.17	1185
5	1	14.37	1.95	2.5	16.8	113	3.85	3.49	0.24	2.18	7.8	0.86	3.45	1480
6	1	13.24	2.59	2.87	21	118	2.8	2.69	0.39	1.82	4.32	1.04	2.93	735
7	1	14.2	1.76	2.45	15.2	112	3.27	3.39	0.34	1.97	6.75	1.05	2.85	1450
8	1	14.39	1.87	2.45	14.6	96	2.5	2.52	0.3	1.98	5.25	1.02	3.58	1290
9	1	14.06	2.15	2.61	17.6	121	2.6	2.51	0.31	1.25	5.05	1.06	3.58	1295
10	1	14.83	1.64	2.17	14	97	2.8	2.98	0.29	1.98	5.2	1.08	2.85	1045
11	1	13.86	1.35	2.27	16	98	2.98	3.15	0.22	1.85	7.22	1.01	3.55	1045
12	1	14.1	2.16	2.3	18	105	2.95	3.32	0.22	2.38	5.75	1.25	3.17	1510
13	1	14.12	1.48	2.32	16.8	95	2.2	2.43	0.26	1.57	5	1.17	2.82	1280
14	1	13.75	1.73	2.41	16	89	2.6	2.76	0.29	1.81	5.6	1.15	2.9	1320
15	1	14.75	1.73	2.39	11.4	91	3.1	3.69	0.43	2.81	5.4	1.25	2.73	1150
16	2	14.38	1.87	2.38	12	102	3.3	3.64	0.29	2.96	7.5	1.2	3	1547
17	2	13.63	1.81	2.7	17.2	112	2.85	2.91	0.3	1.46	7.3	1.28	2.88	1310
18	2	14.3	1.92	2.72	20	120	2.8	3.14	0.33	1.97	6.2	1.07	2.65	1280
19	2	13.83	1.57	2.62	20	115	2.95	3.4	0.4	1.72	6.6	1.13	2.57	1130


Standard Data Partition
Data Source
Worksheet: Data
Workbook: Data-wine.xlsx
Data range: $A$1:$N$168
#Rows: 167
#Cols: 14
Variables
First Row Contains Headers
Variables In Input Data
Ash_Alcanity
Magnesium
Total_Phenols
Flavanoids
Nonflavanoid_Phenols
Proanthocyanins
Color_Intensity
Hue
OD280
Proline
Selected Variables
Type
Alcohol
Malic_Acid
Ash
Partitioning Options
Use partition variable
partition variable
Pick up rows randomly
Set seed: 12345
Partitioning percentages when picking up rows randomly
Automatic percentages
Specify percentages
Equal percentages
Training Set: 70 %
Validation Set: 30 %
Test Set: 0 %
File Home Insert Draw Page Layout Formulas
Model Get Data Explore Transform Cluster Text Partition ARIMA
Data Analysis
Time Series
Tell me what you want to do
Share
Ibrahim Muhammad Al-Jabri
Data STDPartition KNNC_Output
Classifier Label
Ready
Show desktop

Model | Get Data | Explore | Transform | Cluster | Text | Partition | ARIMA | Smoothing | Partition | Classify | Predict | Associate | Score | License | Help

Model | Data | Data Analysis | Time Series | Data Mining | Tools | License | Help

C31 | Training

Ratio - Training	0.7
Ratio - Validation	0.3

Partition Summary

Partition	# Records
Training	117
Validation	50

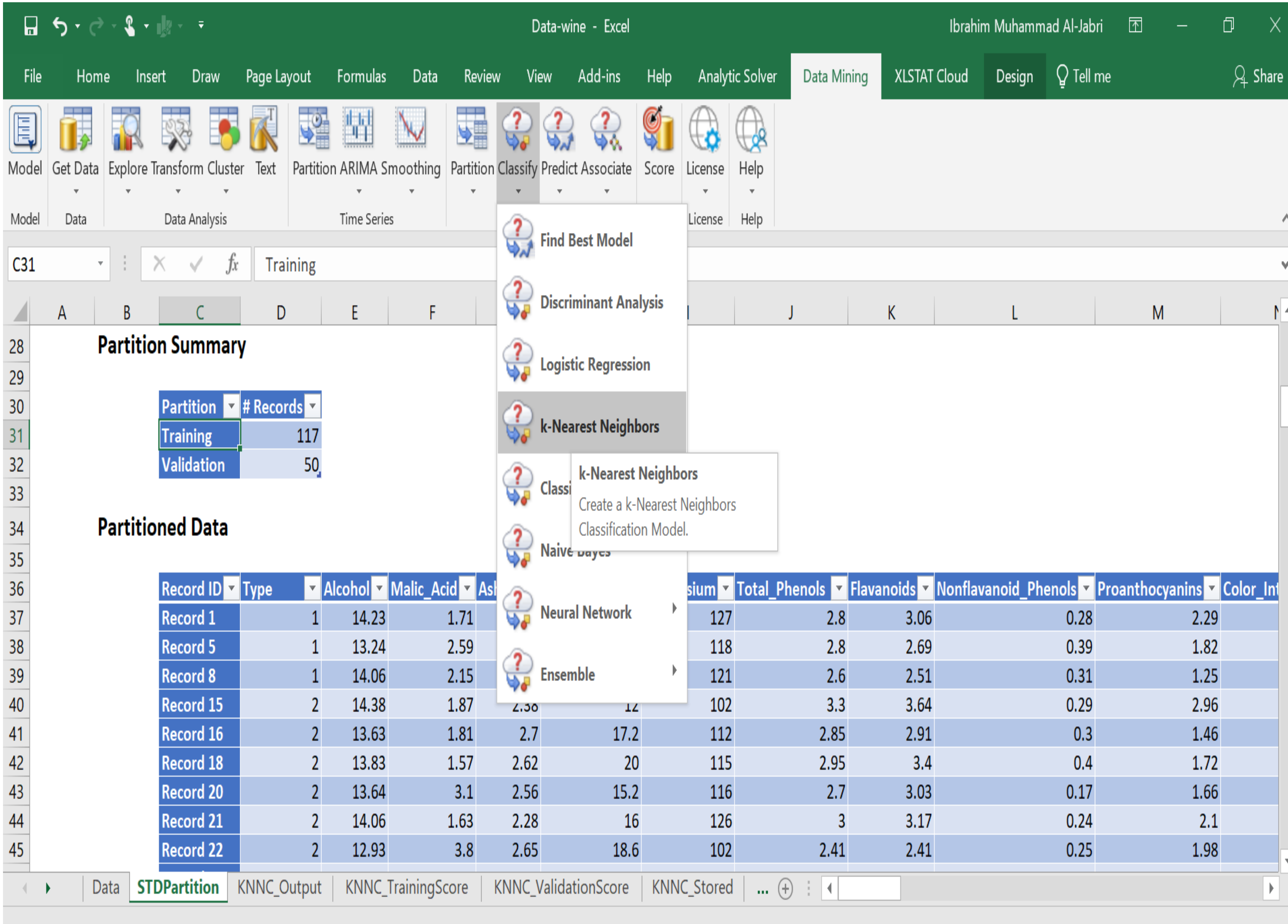
Partitioned Data

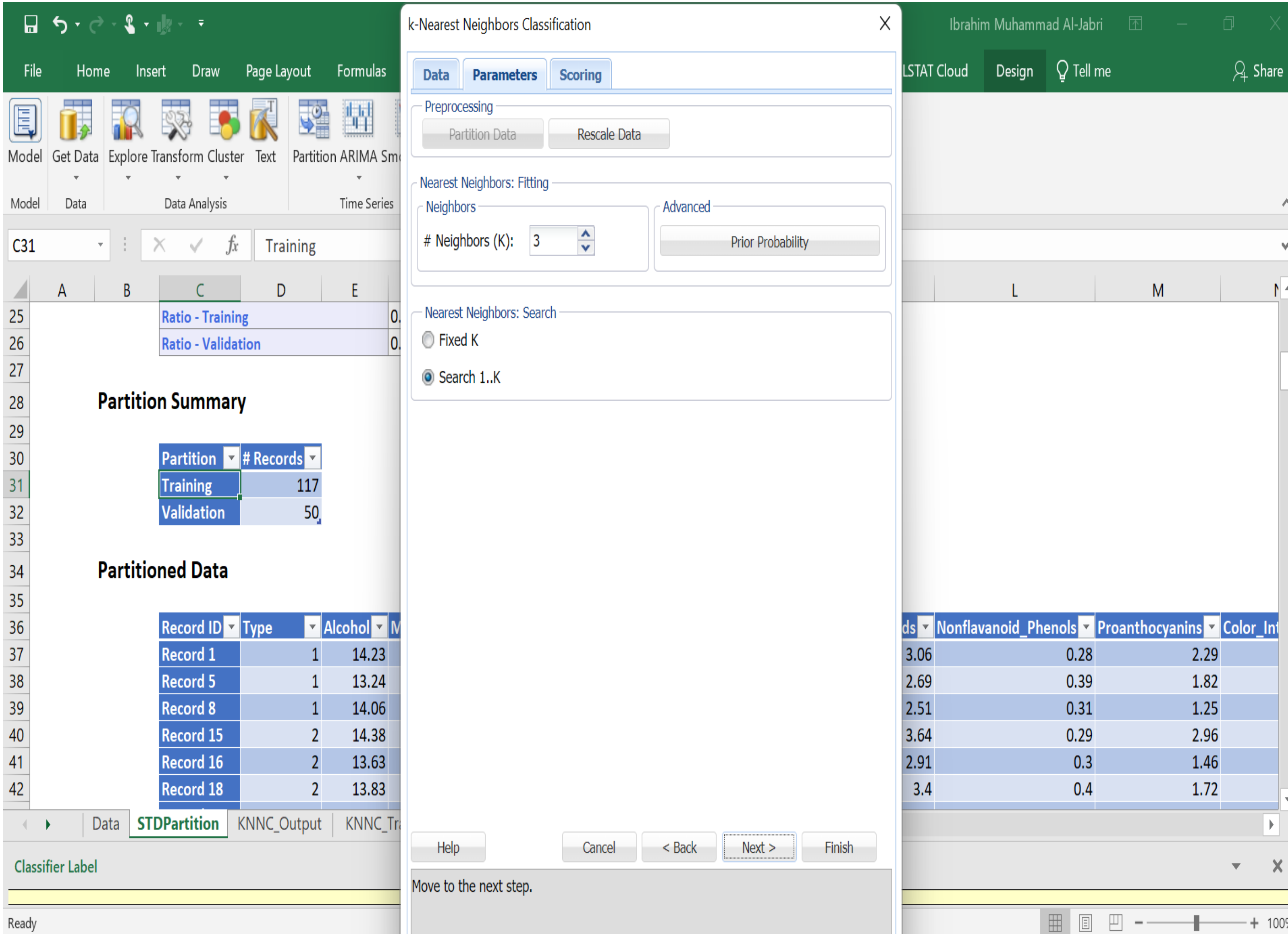
Record ID	Type	Alcohol	Malic_Acid	Ash	Ash_Alcanity	Magnesium	Total_Phenols	Flavanoids	Nonflavanoid_Phenols	Proanthocyanins	Color_Int
Record 1	1	14.23	1.71	2.43	15.6	127	2.8	3.06	0.28	2.29	
Record 5	1	13.24	2.59	2.87	21	118	2.8	2.69	0.39	1.82	
Record 8	1	14.06	2.15	2.61	17.6	121	2.6	2.51	0.31	1.25	
Record 15	2	14.38	1.87	2.38	12	102	3.3	3.64	0.29	2.96	
Record 16	2	13.63	1.81	2.7	17.2	112	2.85	2.91	0.3	1.46	
Record 18	2	13.83	1.57	2.62	20	115	2.95	3.4	0.4	1.72	

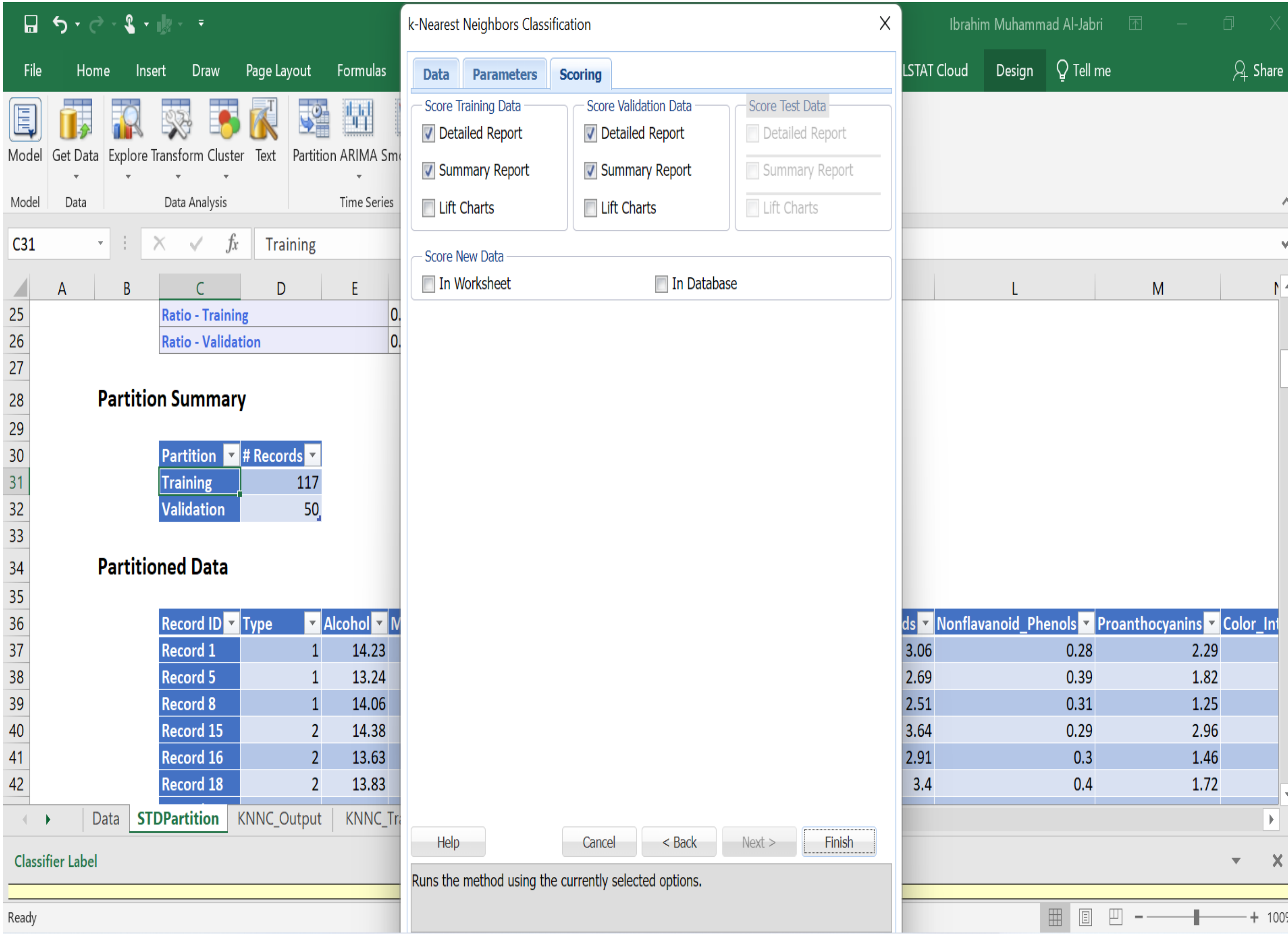
Data | STDPartition | KNNC_Output | KNNC_TrainingScore | KNNC_ValidationScore | KNNC_Stored | ...

Classifier Label

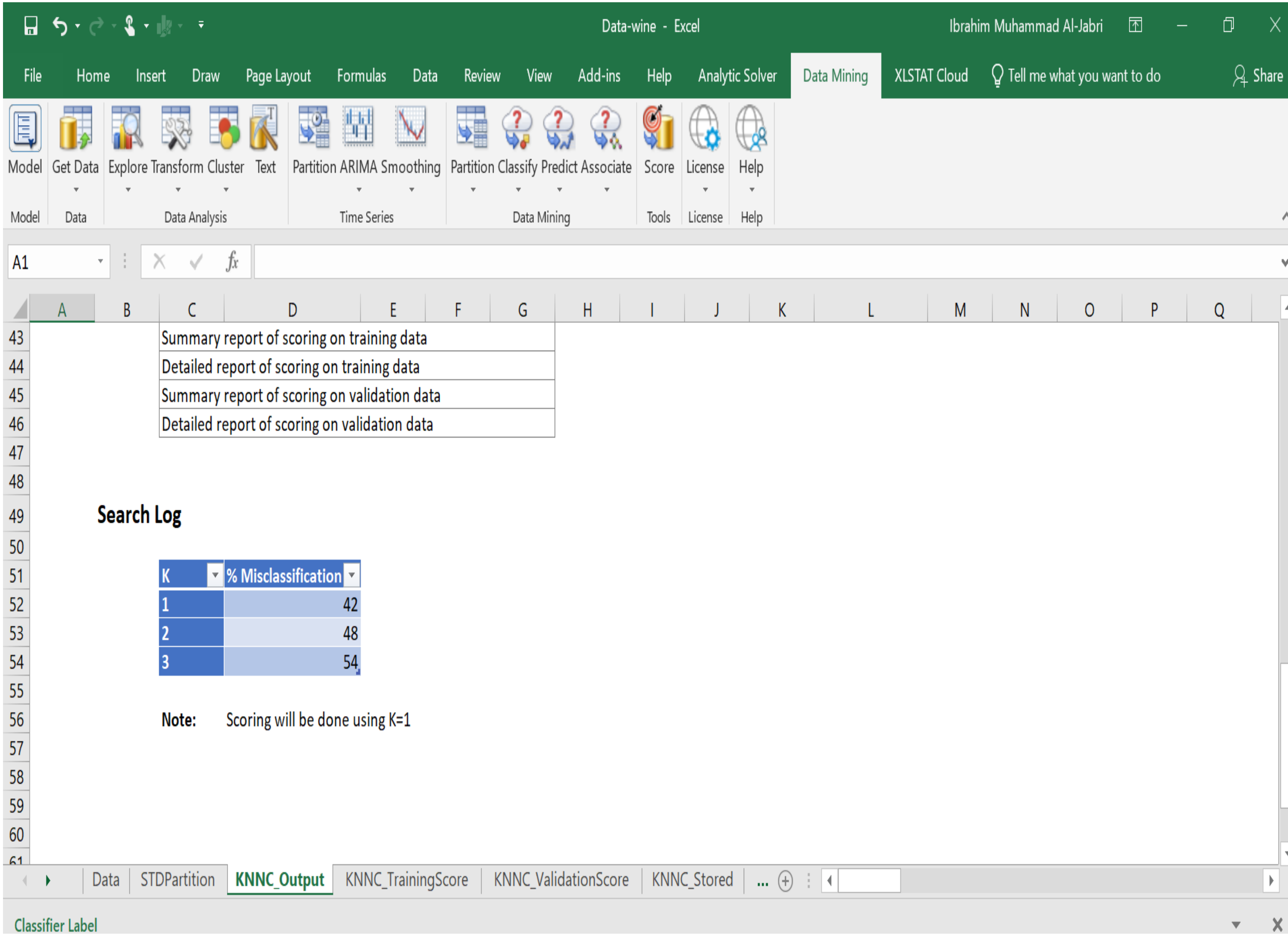
Pearson


Data-wine - Excel
Ibrahim Muhammad Al-Jabri
File Home Insert Draw Page Layout Formulas Data Review View Add-ins Help Analytic Solver Data Mining XLSTAT Cloud Design Tell me Share
Model Get Data Explore Transform Cluster Text Partition ARIMA Smoothing Partition Classify Predict Associate Score License Help
Model Data Data Analysis Time Series License Help
Find Best Model
Discriminant Analysis
Logistic Regression
k-Nearest Neighbors
Classi
Naive Bayes
Neural Network
Ensemble
k-Nearest Neighbors
Create a k-Nearest Neighbors Classification Model.
C31
Training
Partition Summary
Partition # Records
Training 117
Validation 50
Partitioned Data
Record ID Type Alcohol Malic_Acid Ash sium Total_Phenols Flavanoids Nonflavanoid_Phenols Proanthocyanins Color_Int
Record 1 1 14.23 1.71 127 2.8 3.06 0.28 2.29
Record 5 1 13.24 2.59 118 2.8 2.69 0.39 1.82
Record 8 1 14.06 2.15 121 2.6 2.51 0.31 1.25
Record 15 2 14.38 1.87 102 3.3 3.64 0.29 2.96
Record 16 2 13.63 1.81 2.7 17.2 112 2.85 2.91 0.3 1.46
Record 18 2 13.83 1.57 2.62 20 115 2.95 3.4 0.4 1.72
Record 20 2 13.64 3.1 2.56 15.2 116 2.7 3.03 0.17 1.66
Record 21 2 14.06 1.63 2.28 16 126 3 3.17 0.24 2.1
Record 22 2 12.93 3.8 2.65 18.6 102 2.41 2.41 0.25 1.98
Data STDPartition KNNC_Output KNNC_TrainingScore KNNC_ValidationScore KNNC_Stored


k-Nearest Neighbors Classification
Data
Parameters
Scoring
Preprocessing
Partition Data
Rescale Data
Nearest Neighbors: Fitting
Neighbors
Neighbors (K): 3
Advanced
Prior Probability
Nearest Neighbors: Search
Fixed K
Search 1..K
Help
Cancel
< Back
Next >
Finish
Move to the next step.
File
Home
Insert
Draw
Page Layout
Formulas
Model
Get Data
Explore
Transform
Cluster
Text
Partition
ARIMA
Data
Data Analysis
Time Series
Design
Tell me
Share
Ibrahim Muhammad Al-Jabri
C31
Training
Ratio - Training
Ratio - Validation
Partition Summary
Partition
Records
Training
117
Validation
50
Partitioned Data
Record ID
Type
Alcohol
Nonflavanoid_Phenols
Proanthocyanins
Record 1
Record 5
Record 8
Record 15
Record 16
Record 18
Data
STDPartition
KNNC_Output
Classifier Label
Ready
100%


k-Nearest Neighbors Classification
Data
Parameters
Scoring
Score Training Data
Detailed Report
Summary Report
Lift Charts
Score Validation Data
Detailed Report
Summary Report
Lift Charts
Score Test Data
Detailed Report
Summary Report
Lift Charts
Score New Data
In Worksheet
In Database
Help
Cancel
< Back
Next >
Finish
Runs the method using the currently selected options.
File
Home
Insert
Draw
Page Layout
Formulas
Ibrahim Muhammad Al-Jabri
LSTAT Cloud
Design
Tell me
Share
Model
Get Data
Explore
Transform
Cluster
Text
Partition
ARIMA
Data Analysis
Time Series
C31
Training
Ratio - Training
Ratio - Validation
Partition Summary
Partition
Records
Training 117
Validation 50
Partitioned Data
Record ID
Type
Alcohol
Nonflavanoid_Phenols
Proanthocyanins
Color_Int
Record 1 1 14.23 3.06 0.28 2.29
Record 5 1 13.24 2.69 0.39 1.82
Record 8 1 14.06 2.51 0.31 1.25
Record 15 2 14.38 3.64 0.29 2.96
Record 16 2 13.63 2.91 0.3 1.46
Record 18 2 13.83 3.4 0.4 1.72
Data
STDPartition
KNNC_Output
Classifier Label
Ready
100%

Pearson


Data-wine - Excel
Ibrahim Muhammad Al-Jabri
File
Home
Insert
Draw
Page Layout
Formulas
Data
Review
View
Add-ins
Help
Analytic Solver
Data Mining
XLSTAT Cloud
Tell me what you want to do
Share
Model
Get Data
Explore
Transform
Cluster
Text
Partition
ARIMA
Smoothing
Partition
Classify
Predict
Associate
Score
License
Help
Model
Data
Data Analysis
Time Series
Data Mining
Tools
License
Help
A1
Summary report of scoring on training data
Detailed report of scoring on training data
Summary report of scoring on validation data
Detailed report of scoring on validation data
Search Log
K
% Misclassification
1
42
2
48
3
54
Note:
Scoring will be done using K=1
Data
STDPartition
KNNC_Output
KNNC_TrainingScore
KNNC_ValidationScore
KNNC_Stored
Classifier Label

Pearson

Model | Get Data | Explore | Transform | Cluster | Text | Partition | ARIMA | Smoothing | Partition | Classify | Predict | Associate | Score | License | Help

Model | Data | Data Analysis | Time Series | Data Mining | Tools | License | Help

A1

Validation: Classification Summary

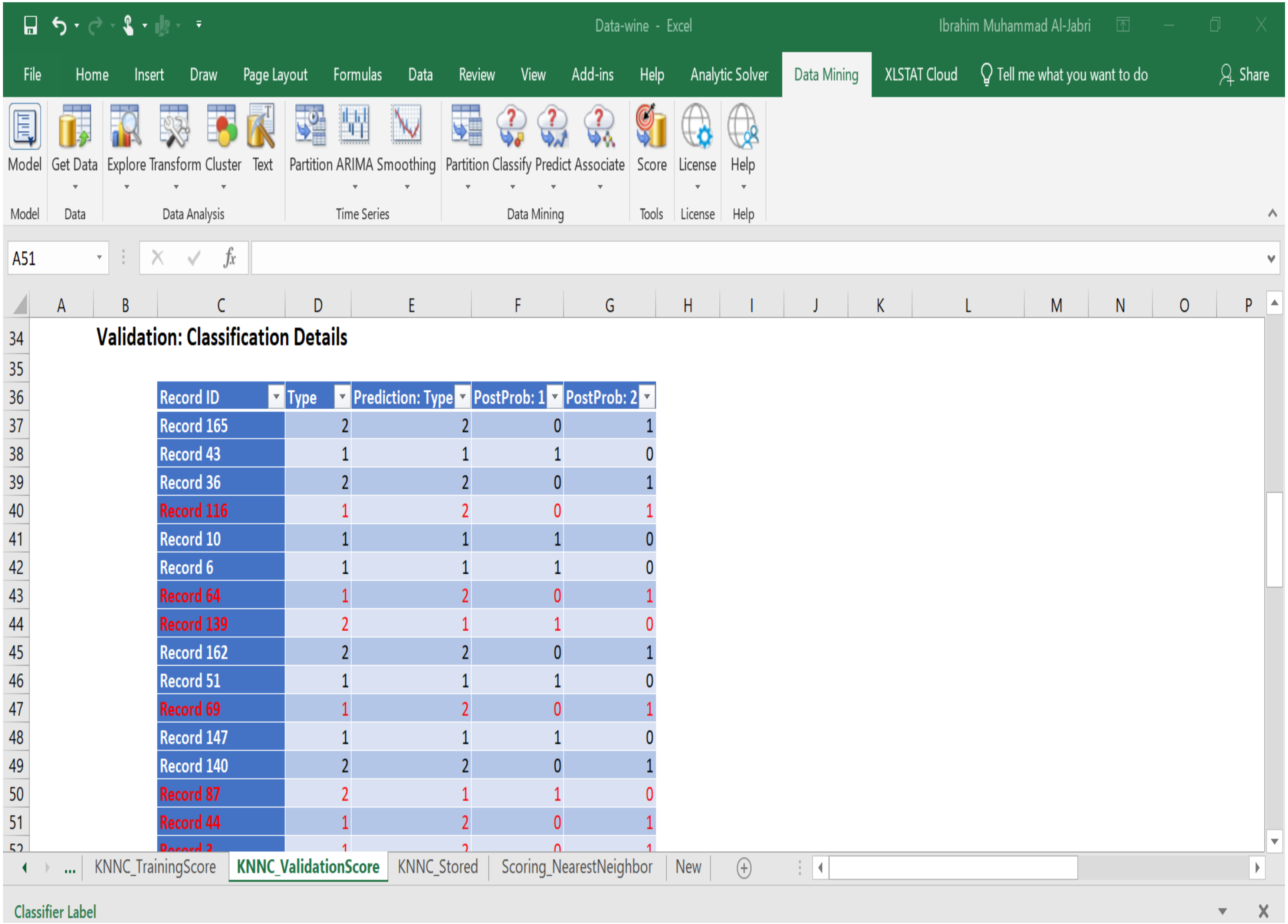
Confusion Matrix		
Actual\Predicted	1	2
1	13	13
2	8	16

Error Report			
Class	# Cases	# Errors	% Error
1	26	13	50
2	24	8	33.33333333
Overall	50	21	42

Metrics	
Metric	Value
Accuracy (#correct)	29
Accuracy (%correct)	58
Specificity	0.666667

... KNNC_TrainingScore | KNNC_ValidationScore | KNNC_Stored | Scoring_NearestNeighbor | New

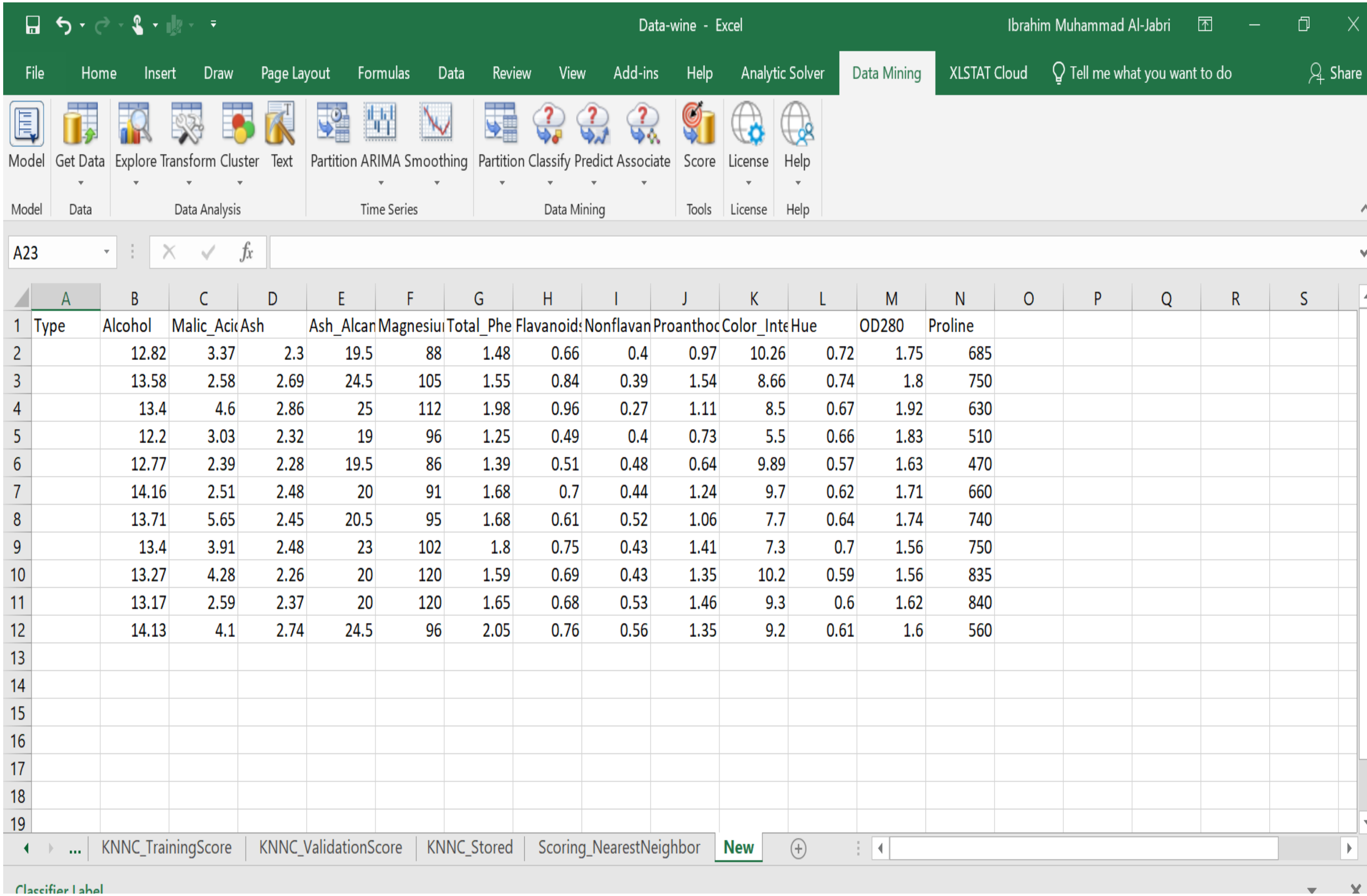
Classifier Label



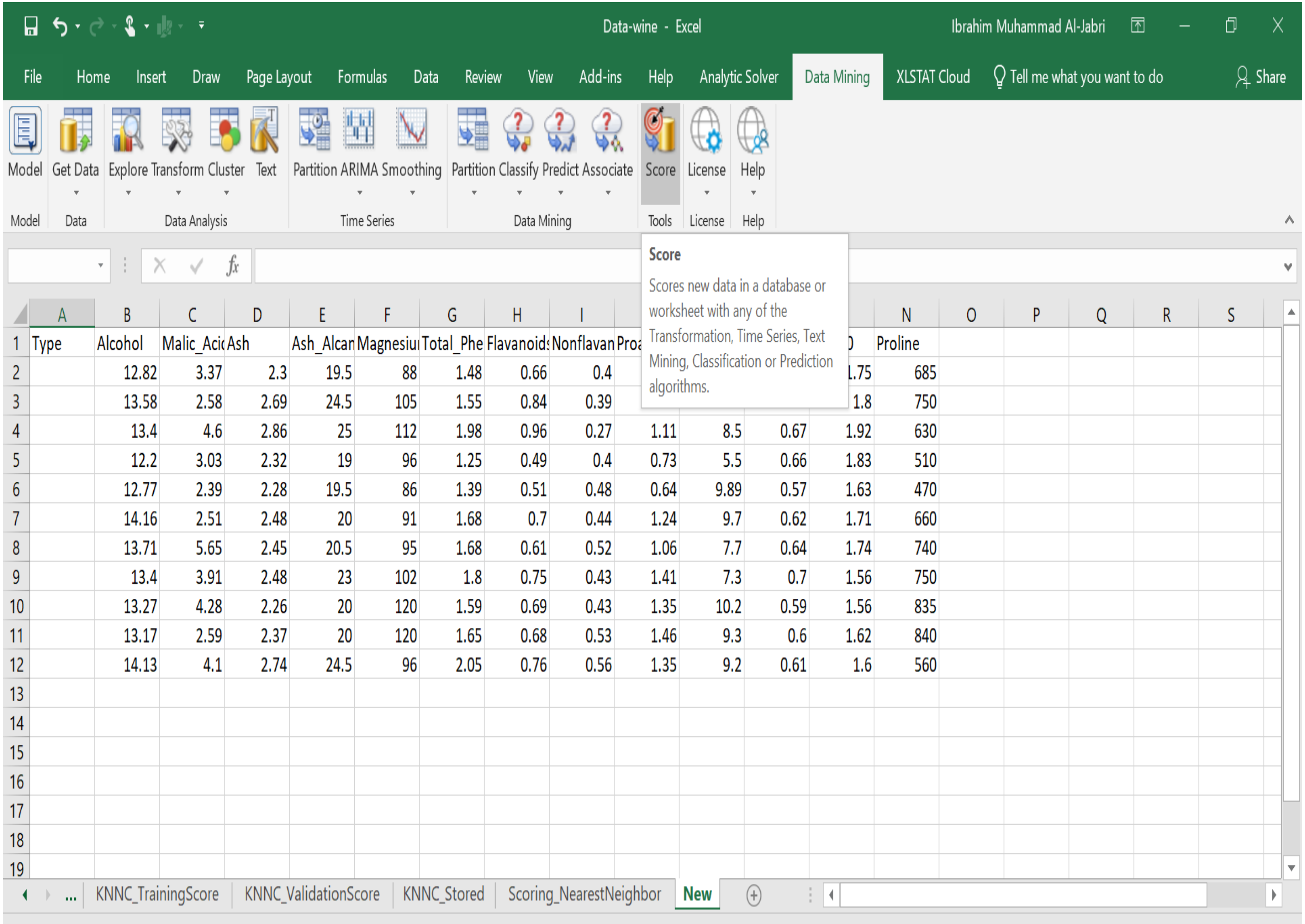
Validation: Classification Details

Record ID	Type	Prediction: Type	PostProb: 1	PostProb: 2
Record 165	2	2	0	1
Record 43	1	1	1	0
Record 36	2	2	0	1
Record 116	1	2	0	1
Record 10	1	1	1	0
Record 6	1	1	1	0
Record 64	1	2	0	1
Record 139	2	1	1	0
Record 162	2	2	0	1
Record 51	1	1	1	0
Record 69	1	2	0	1
Record 147	1	1	1	0
Record 140	2	2	0	1
Record 87	2	1	1	0
Record 44	1	2	0	1
Record 3	1	2	0	1

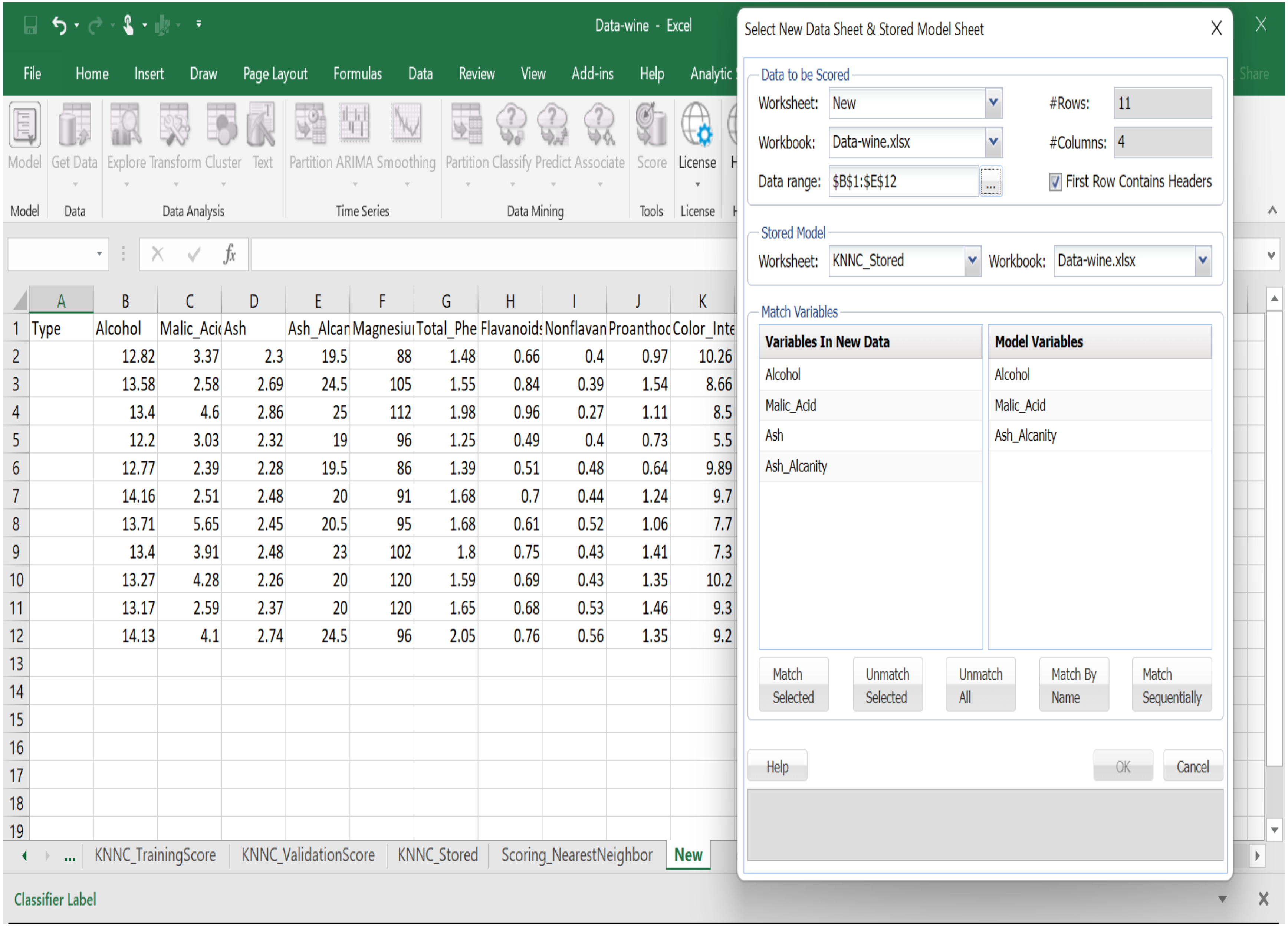
Pearson

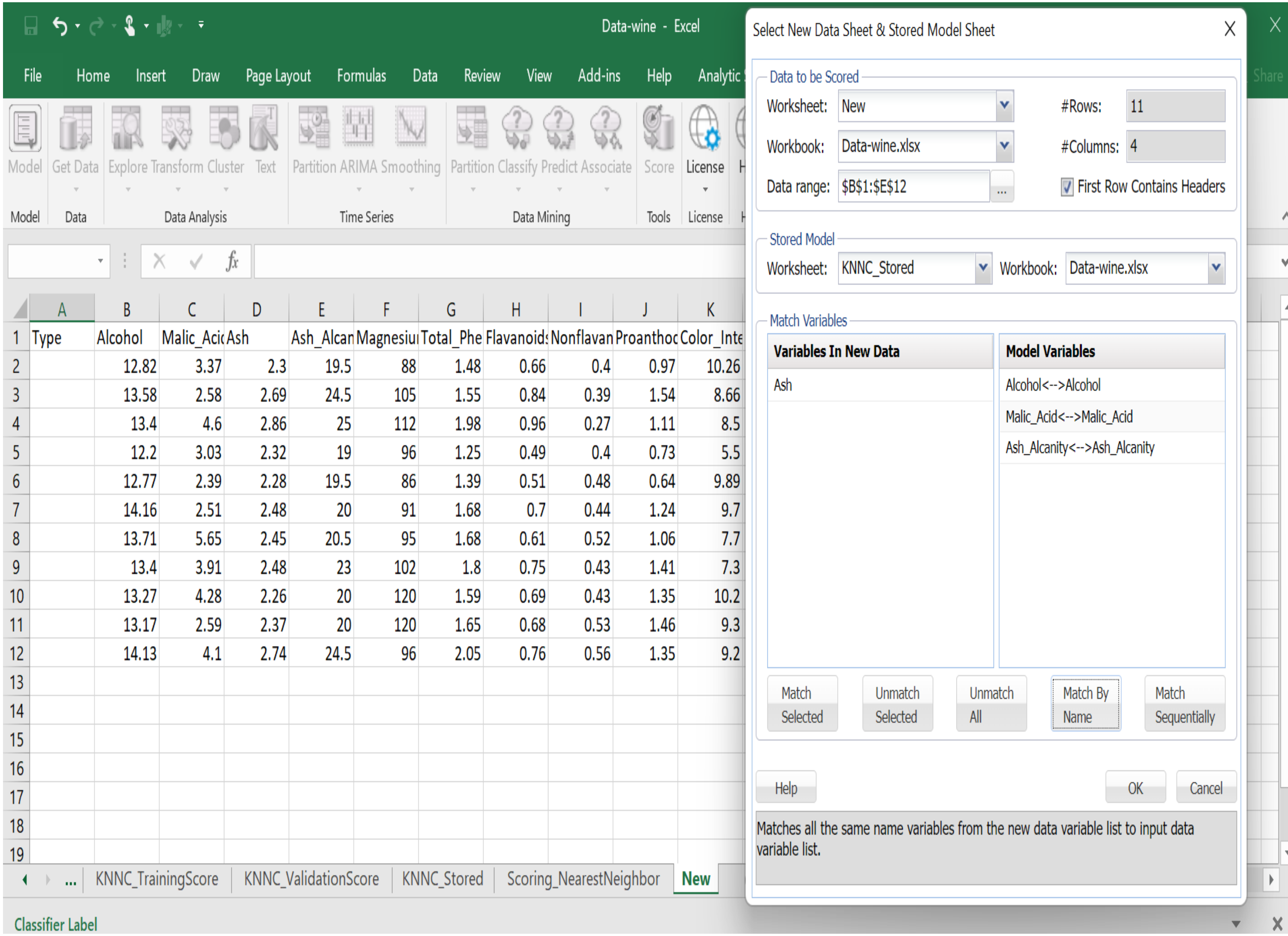


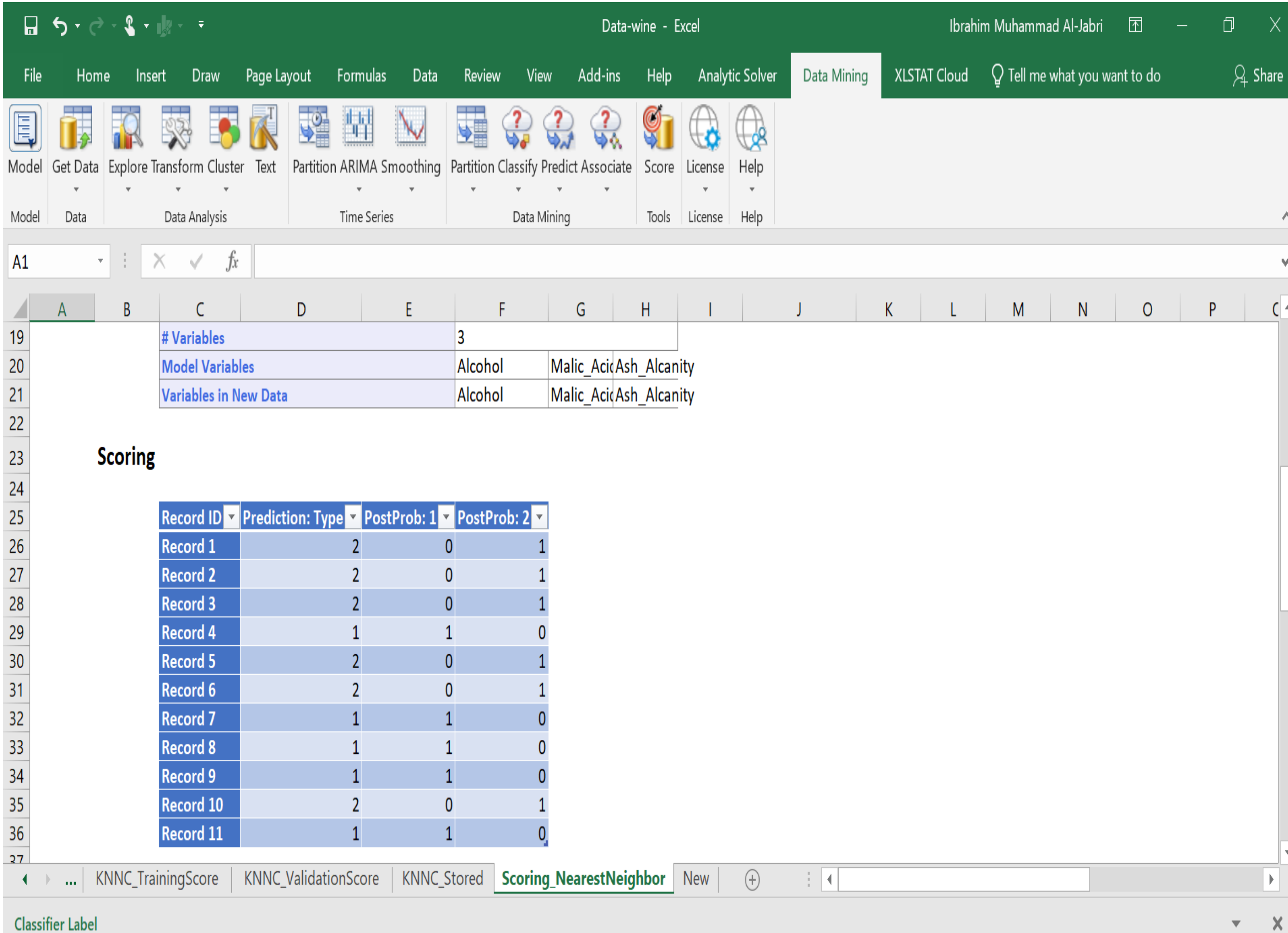
	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Type	Alcohol	Malic_Acic	Ash	Ash_Alcan	Magnesiu	Total_Phe	Flavanoid:	Nonflavan	Proanthoc	Color_Inte	Hue	OD280	Proline
2		12.82	3.37	2.3	19.5	88	1.48	0.66	0.4	0.97	10.26	0.72	1.75	685
3		13.58	2.58	2.69	24.5	105	1.55	0.84	0.39	1.54	8.66	0.74	1.8	750
4		13.4	4.6	2.86	25	112	1.98	0.96	0.27	1.11	8.5	0.67	1.92	630
5		12.2	3.03	2.32	19	96	1.25	0.49	0.4	0.73	5.5	0.66	1.83	510
6		12.77	2.39	2.28	19.5	86	1.39	0.51	0.48	0.64	9.89	0.57	1.63	470
7		14.16	2.51	2.48	20	91	1.68	0.7	0.44	1.24	9.7	0.62	1.71	660
8		13.71	5.65	2.45	20.5	95	1.68	0.61	0.52	1.06	7.7	0.64	1.74	740
9		13.4	3.91	2.48	23	102	1.8	0.75	0.43	1.41	7.3	0.7	1.56	750
10		13.27	4.28	2.26	20	120	1.59	0.69	0.43	1.35	10.2	0.59	1.56	835
11		13.17	2.59	2.37	20	120	1.65	0.68	0.53	1.46	9.3	0.6	1.62	840
12		14.13	4.1	2.74	24.5	96	2.05	0.76	0.56	1.35	9.2	0.61	1.6	560



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Type	Alcohol	Malic_Acic	Ash	Ash_Alcan	Magnesiu	Total_Phe	Flavanoid	Nonflavan	Proa			)	Proline
2		12.82	3.37	2.3	19.5	88	1.48	0.66	0.4				1.75	685
3		13.58	2.58	2.69	24.5	105	1.55	0.84	0.39				1.8	750
4		13.4	4.6	2.86	25	112	1.98	0.96	0.27	1.11	8.5	0.67	1.92	630
5		12.2	3.03	2.32	19	96	1.25	0.49	0.4	0.73	5.5	0.66	1.83	510
6		12.77	2.39	2.28	19.5	86	1.39	0.51	0.48	0.64	9.89	0.57	1.63	470
7		14.16	2.51	2.48	20	91	1.68	0.7	0.44	1.24	9.7	0.62	1.71	660
8		13.71	5.65	2.45	20.5	95	1.68	0.61	0.52	1.06	7.7	0.64	1.74	740
9		13.4	3.91	2.48	23	102	1.8	0.75	0.43	1.41	7.3	0.7	1.56	750
10		13.27	4.28	2.26	20	120	1.59	0.69	0.43	1.35	10.2	0.59	1.56	835
11		13.17	2.59	2.37	20	120	1.65	0.68	0.53	1.46	9.3	0.6	1.62	840
12		14.13	4.1	2.74	24.5	96	2.05	0.76	0.56	1.35	9.2	0.61	1.6	560


Data-wine - Excel
File
Home
Insert
Draw
Page Layout
Formulas
Data
Review
View
Add-ins
Help
Analytic
Model
Get Data
Explore
Transform
Cluster
Text
Partition
ARIMA
Smoothing
Partition
Classify
Predict
Associate
Score
License
Model
Data
Data Analysis
Time Series
Data Mining
Tools
License
Type Alcohol Malic_Acid Ash Ash_Alcan Magnesiu Total_Phe Flavanoid Nonflavan Proanthoc Color_Inte
12.82 3.37 2.3 19.5 88 1.48 0.66 0.4 0.97 10.26
13.58 2.58 2.69 24.5 105 1.55 0.84 0.39 1.54 8.66
13.4 4.6 2.86 25 112 1.98 0.96 0.27 1.11 8.5
12.2 3.03 2.32 19 96 1.25 0.49 0.4 0.73 5.5
12.77 2.39 2.28 19.5 86 1.39 0.51 0.48 0.64 9.89
14.16 2.51 2.48 20 91 1.68 0.7 0.44 1.24 9.7
13.71 5.65 2.45 20.5 95 1.68 0.61 0.52 1.06 7.7
13.4 3.91 2.48 23 102 1.8 0.75 0.43 1.41 7.3
13.27 4.28 2.26 20 120 1.59 0.69 0.43 1.35 10.2
13.17 2.59 2.37 20 120 1.65 0.68 0.53 1.46 9.3
14.13 4.1 2.74 24.5 96 2.05 0.76 0.56 1.35 9.2
KNNC_TrainingScore
KNNC_ValidationScore
KNNC_Stored
Scoring_NearestNeighbor
New
Classifier Label
Select New Data Sheet & Stored Model Sheet
Data to be Scored
Worksheet: New
Workbook: Data-wine.xlsx
Data range: $B$1:$E$12
#Rows: 11
#Columns: 4
First Row Contains Headers
Stored Model
Worksheet: KNNC_Stored
Workbook: Data-wine.xlsx
Match Variables
Variables In New Data
Alcohol
Malic_Acid
Ash
Ash_Alcanity
Model Variables
Alcohol
Malic_Acid
Ash_Alcanity
Match Selected
Unmatch Selected
Unmatch All
Match By Name
Match Sequentially
Help
OK
Cancel


Data-wine - Excel
File Home Insert Draw Page Layout Formulas Data Review View Add-ins Help Analytic
Model Get Data Explore Transform Cluster Text Partition ARIMA Smoothing Partition Classify Predict Associate Score License
Model Data Data Analysis Time Series Data Mining Tools License
Type Alcohol Malic_Acic Ash Ash_Alcan Magnesiu Total_Phe Flavanoid Nonflavan Proanthoc Color_Inte
KNNC_TrainingScore KNNC_ValidationScore KNNC_Stored Scoring_NearestNeighbor New
Classifier Label
Select New Data Sheet & Stored Model Sheet
Data to be Scored
Worksheet: New
Workbook: Data-wine.xlsx
Data range: $B$1:$E$12
#Rows: 11
#Columns: 4
First Row Contains Headers
Stored Model
Worksheet: KNNC_Stored
Workbook: Data-wine.xlsx
Match Variables
Variables In New Data
Ash
Model Variables
Alcohol<-->Alcohol
Malic_Acid<-->Malic_Acid
Ash_Alcanity<-->Ash_Alcanity
Match Selected
Unmatch Selected
Unmatch All
Match By Name
Match Sequentially
Help
OK
Cancel
Matches all the same name variables from the new data variable list to input data variable list.



# Variables	3		
Model Variables	Alcohol	Malic_Aci	Ash_Alcanity
Variables in New Data	Alcohol	Malic_Aci	Ash_Alcanity

Scoring

Record ID	Prediction: Type	PostProb: 1	PostProb: 2
Record 1	2	0	1
Record 2	2	0	1
Record 3	2	0	1
Record 4	1	1	0
Record 5	2	0	1
Record 6	2	0	1
Record 7	1	1	0
Record 8	1	1	0
Record 9	1	1	0
Record 10	2	0	1
Record 11	1	1	0

Naïve Bayes Method for Classification

(1 of 2)

- Naïve Bayes is a simple probability-based classification method
  - Naïve - assumption of independence among the input variables
- Can use both numeric and nominal input variables
  - Numeric variables need to be discretized
- Can be used for both regression and classification
- Naïve based models can be developed very efficiently and effectively
  - Using maximum likelihood method

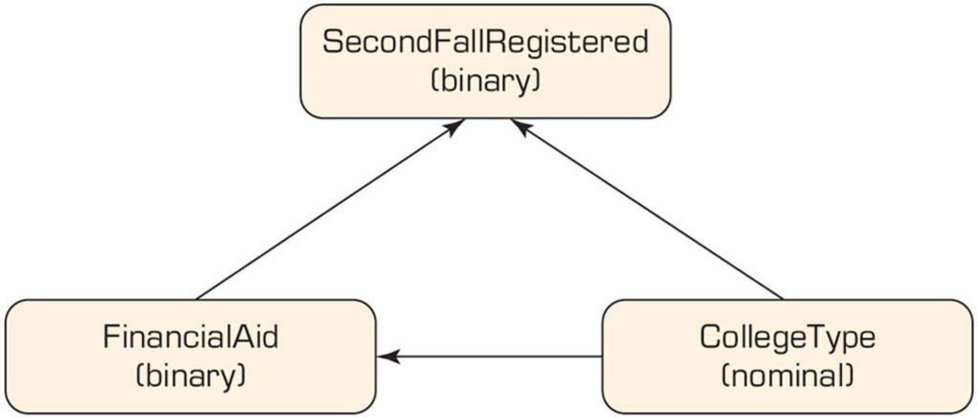
Naïve Bayes Method for Classification (2 of 2)

- Process of Developing a Naïve Bayes Classifier
- Training Phase
  1. Obtain and pre-process the data
  2. Discretize the numeric variables
  3. Calculate the prior probabilities of all class labels
  4. Calculate the likelihood for all predictor variables/values
- Testing Phase
  - Using the outputs of Steps 3 and 4 above, classify the new samples
    - See the numerical example in the book…

Pearson

Bayesian Networks (1 of 5)

- A tool for representing dependency structure in a graphical, explicit, and intuitive way
  - A directed acyclic graph whose nodes correspond to the variables and arcs that signify conditional dependencies between variables and their possible values
  - Direction of the arc matter
  - A partial causality link in student retention



Bayesian Networks (2 of 5)

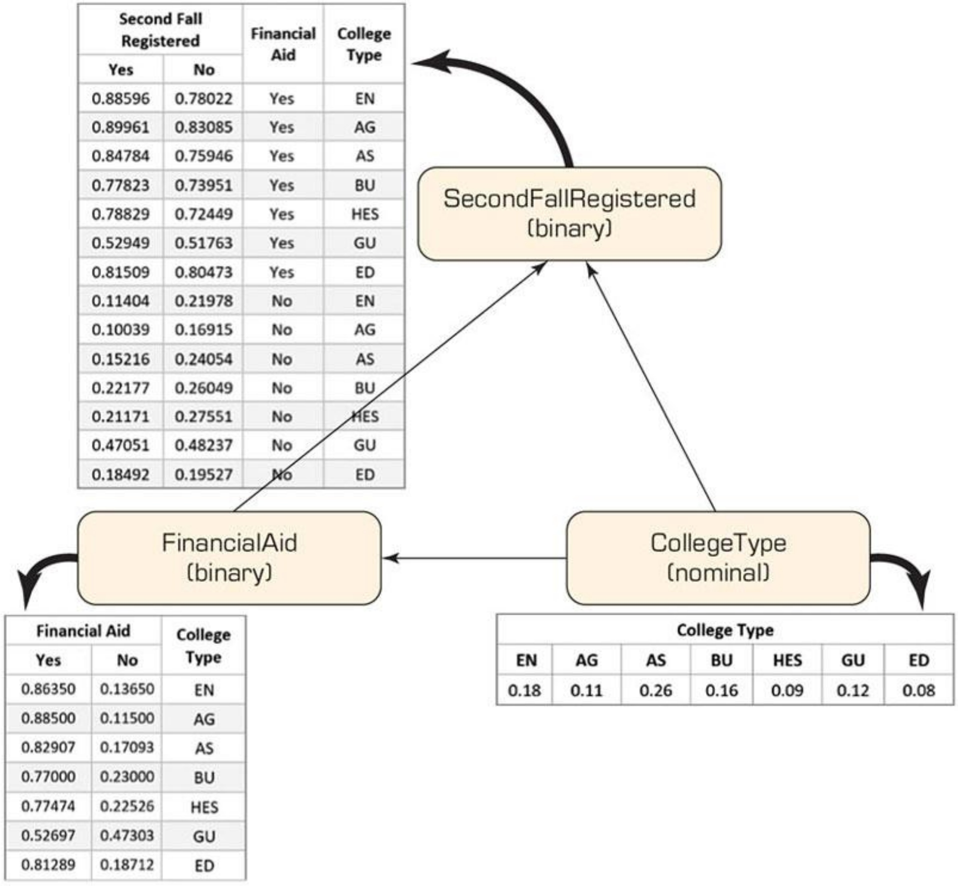
How can BN be constructed?

1. Manually
   - By an engineer with the help of a domain expert
   - Time demanding, expensive (for large networks)
   - Experts may not even be available
2. Automatically
   - Analytically …
   - By learning/inducing the structure of the network from the historical data
     - Availability high-quality historical data is imperative

Bayesian Networks (3 of 5)

How can BN be constructed?

- Analytically



Second Fall Registered		Financial Aid	College Type
Yes	No		
0.88596	0.78022	Yes	EN
0.89961	0.83085	Yes	AG
0.84784	0.75946	Yes	AS
0.77823	0.73951	Yes	BU
0.78829	0.72449	Yes	HES
0.52949	0.51763	Yes	GU
0.81509	0.80473	Yes	ED
0.11404	0.21978	No	EN
0.10039	0.16915	No	AG
0.15216	0.24054	No	AS
0.22177	0.26049	No	BU
0.21171	0.27551	No	HES
0.47051	0.48237	No	GU
0.18492	0.19527	No	ED

Financial Aid		College Type
Yes	No	
0.86350	0.13650	EN
0.88500	0.11500	AG
0.82907	0.17093	AS
0.77000	0.23000	BU
0.77474	0.22526	HES
0.52697	0.47303	GU
0.81289	0.18712	ED

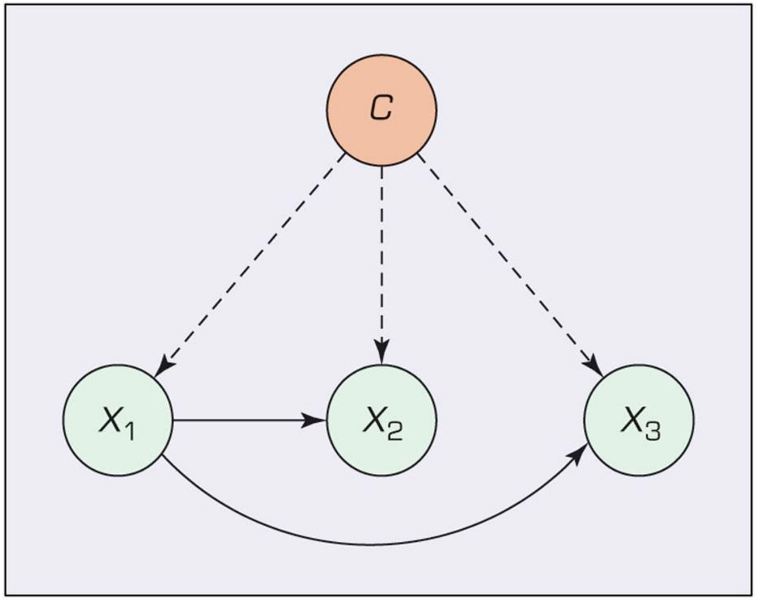
College Type						
EN	AG	AS	BU	HES	GU	ED
0.18	0.11	0.26	0.16	0.09	0.12	0.08

Bayesian Networks (4 of 5)

How can BN be constructed?

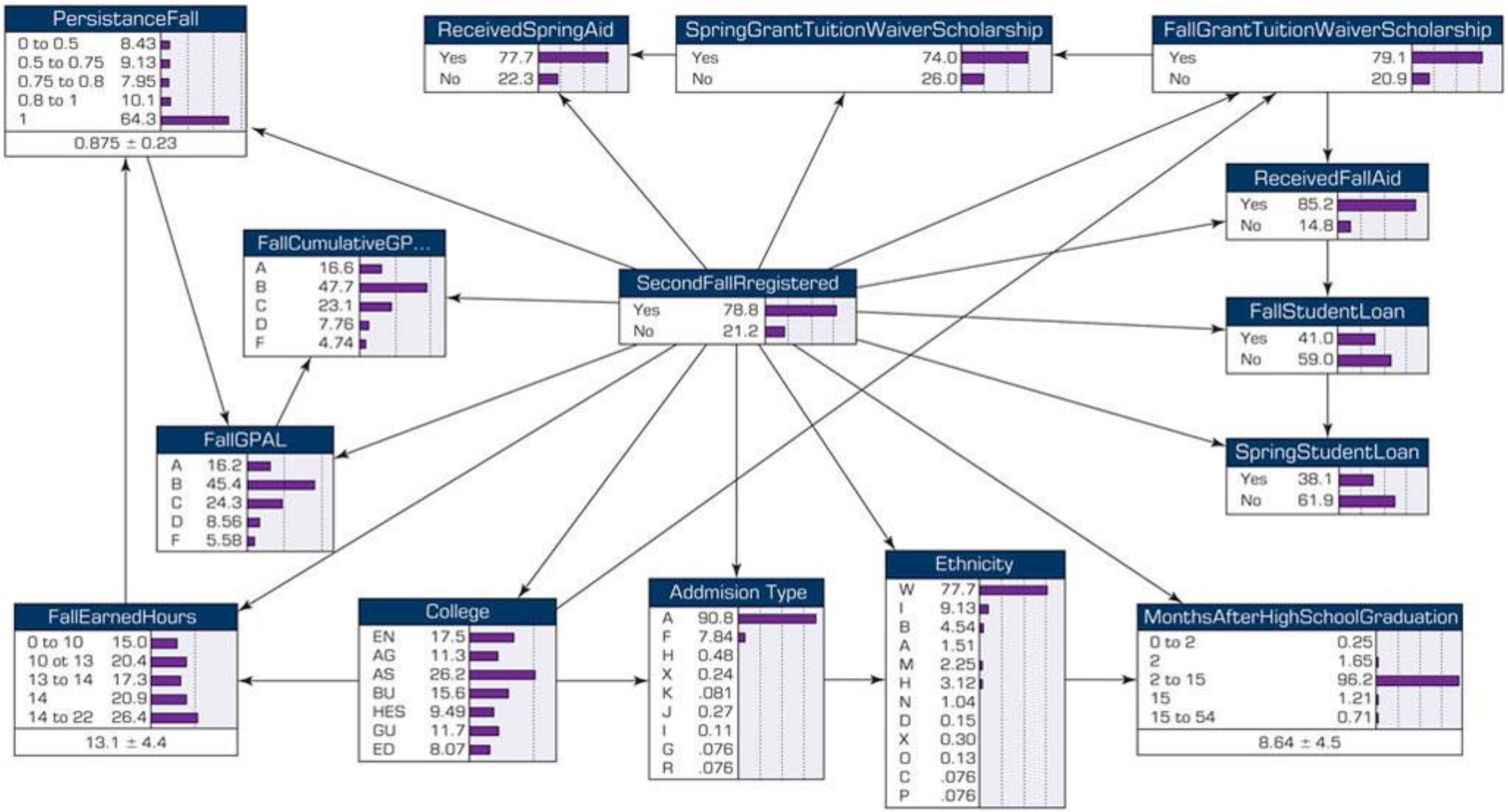
Tree Augmented Naïve Bayes Network Structure

1. Compute information function
2. Build the undirected graph
3. Build a spanning tree
4. Convert the undirected graph into a directed one
5. Construct a TAN model



Tree Augmented Naïve (TAN) Bayes Network Structure

Bayesian Networks (5 of 5)



- EXAMPLE: Bayesian Belief Network for Predicting Freshmen Student Attrition

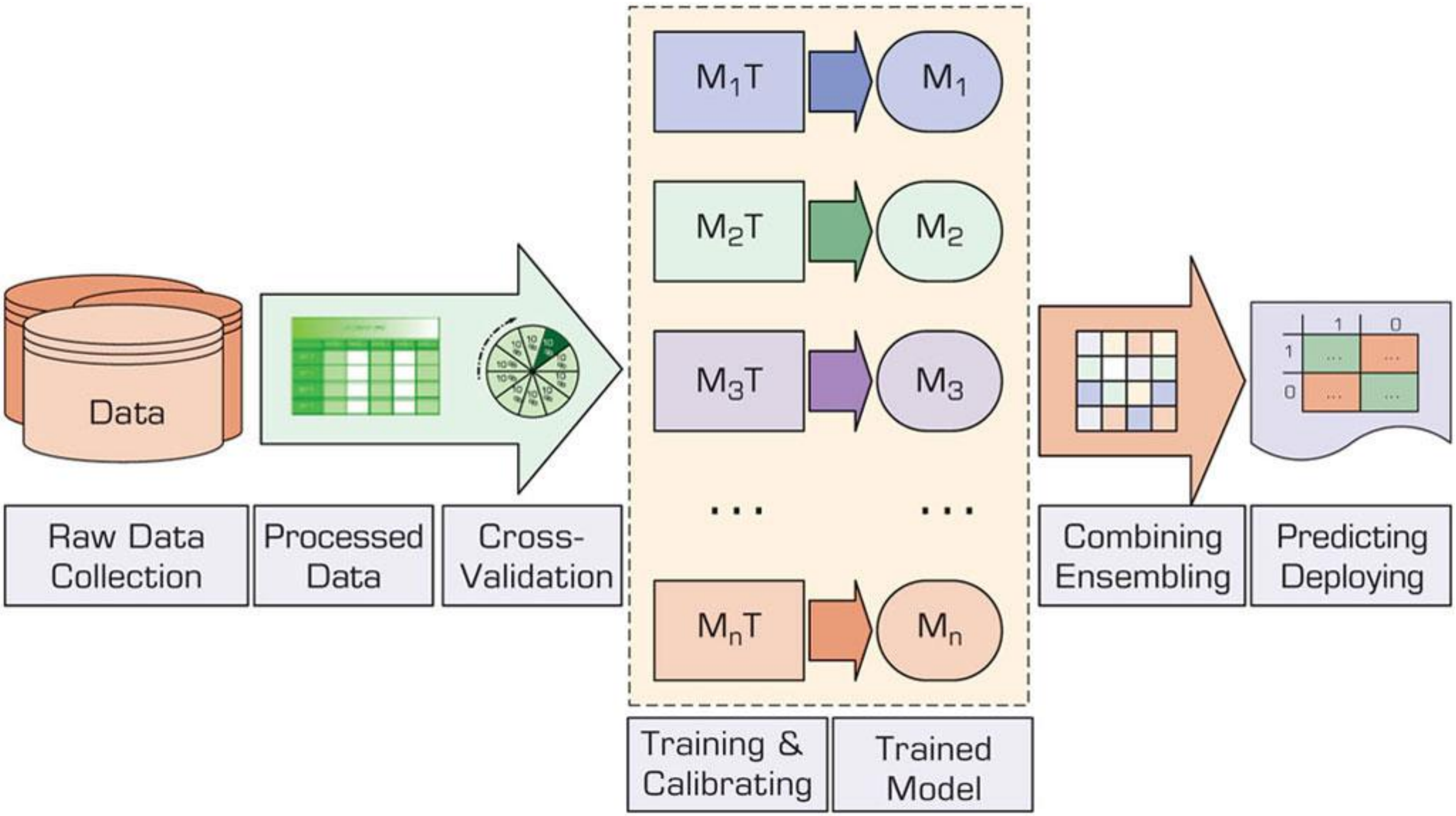
Pearson

Ensemble Modeling (1 of 3)

- Ensemble – combination of models (or model outcomes) for better results
- Why do we need to use ensembles:
  - Better accuracy
  - More stable/robust/consistent/reliable outcomes
- Reality: ensembles wins competitions!
  - Many recent competitions at Kaggle.com
- The Wisdom of Crowds

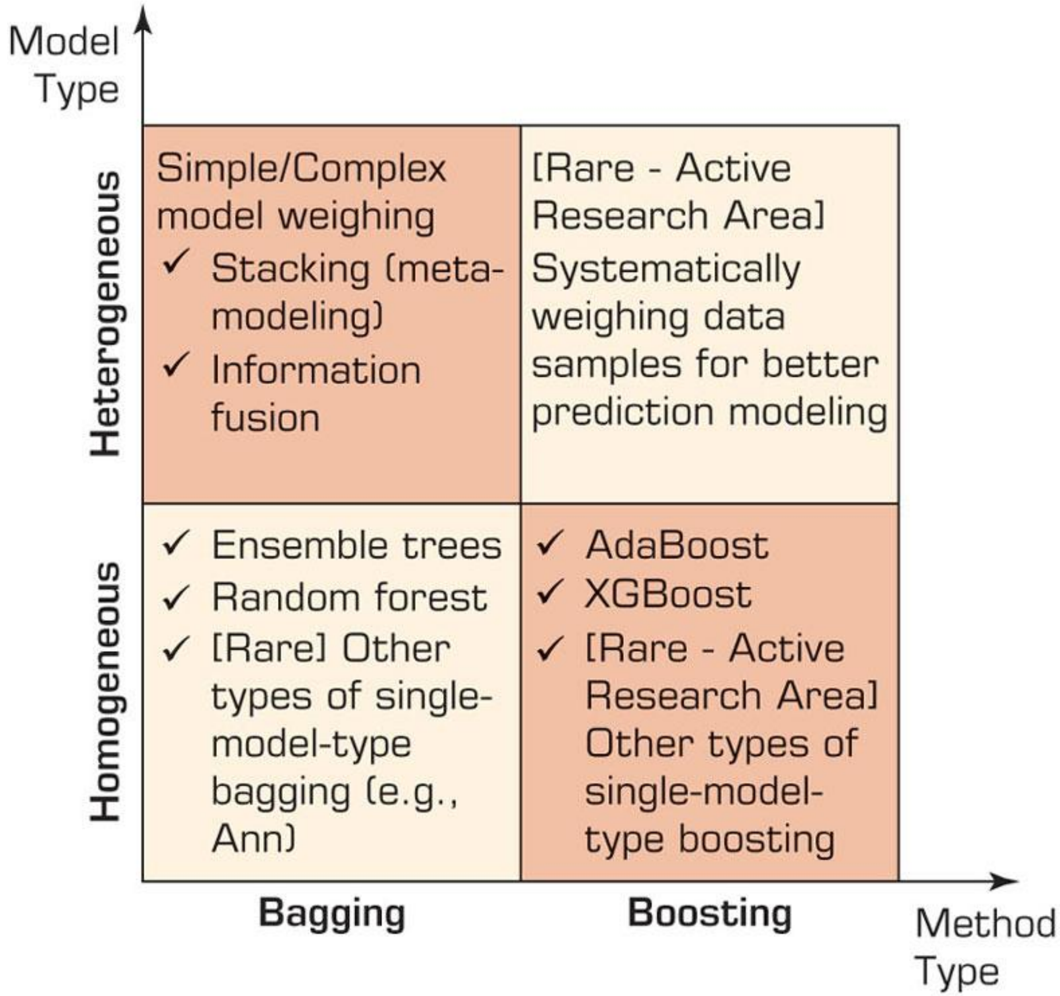
Ensemble Modeling (2 of 3)

Figure 5.19 Graphical Depiction of Model Ensembles for Prediction Modeling.



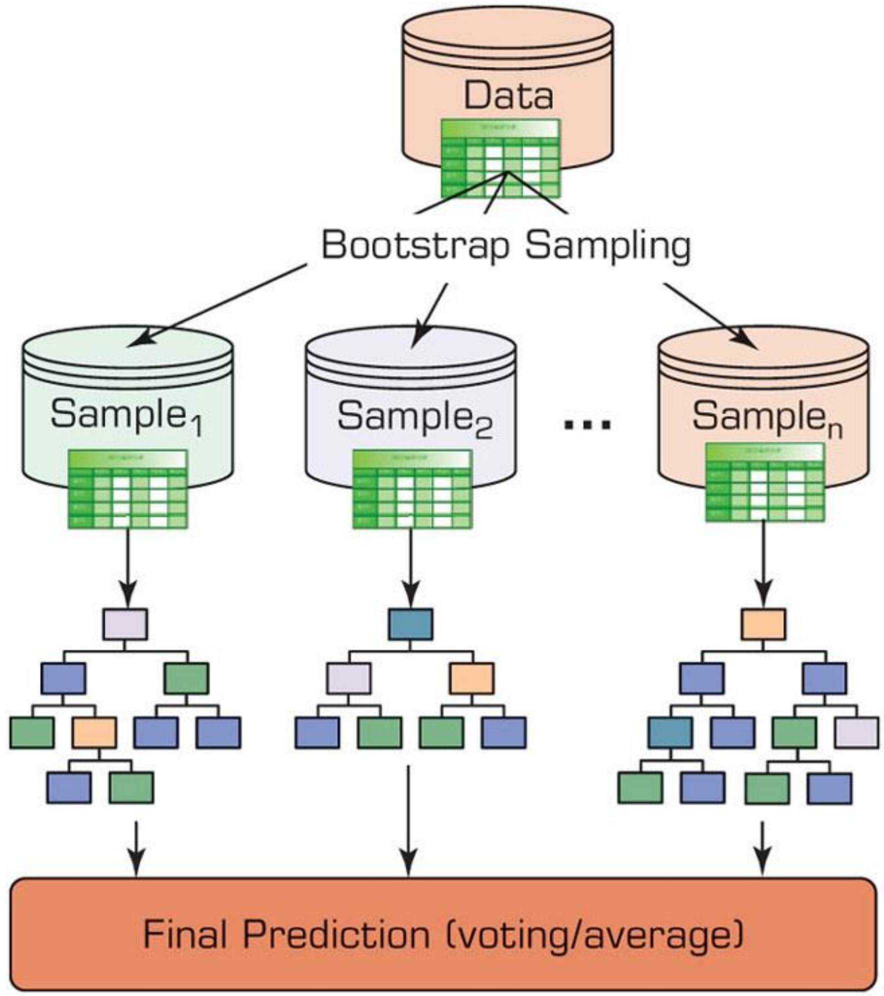
Types of Ensemble Modeling (1 of 4)

Figure 5.20 Simple Taxonomy for Model Ensembles.



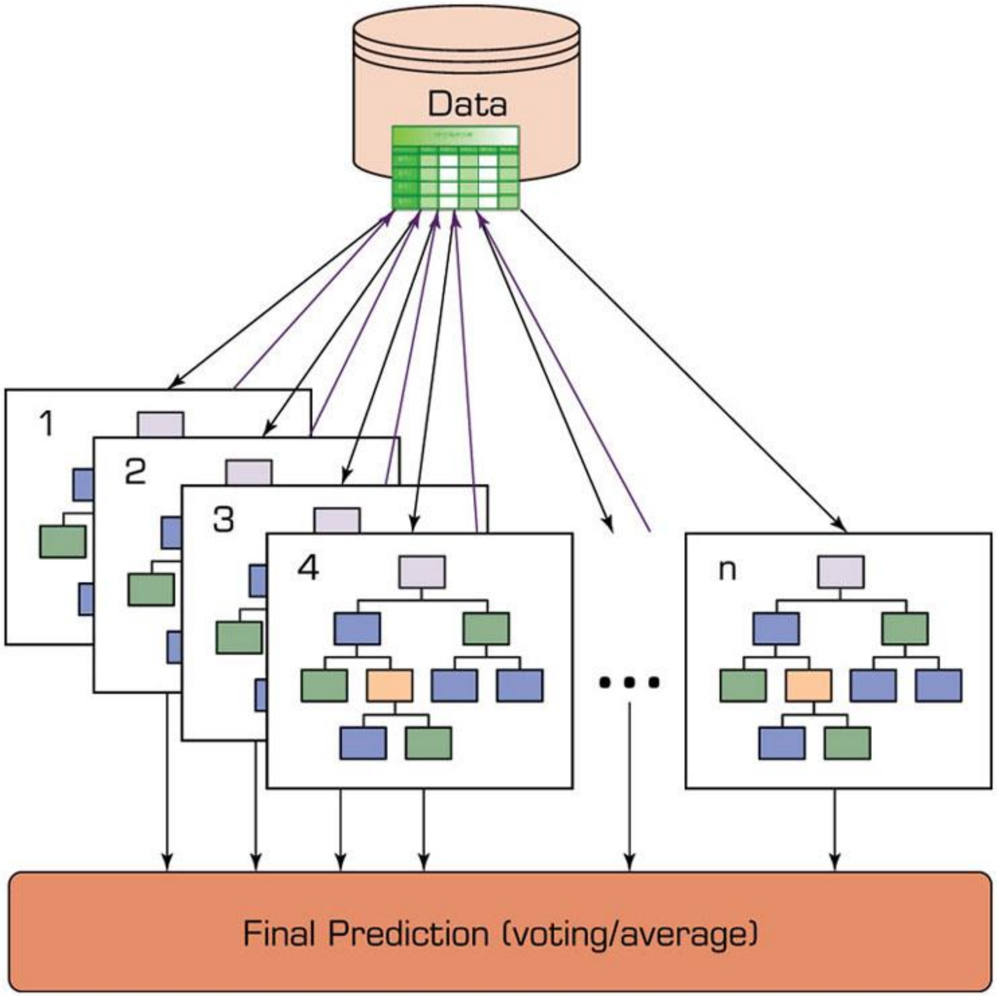
Types of Ensemble Modeling (2 of 4)

Figure 5.20 Bagging-Type Decision Tree Ensembles.



Types of Ensemble Modeling (3 of 4)

Figure 5.20 Boosting-Type Decision Tree Ensembles.



Ensemble Modeling (3 of 3)

- Variants of Bagging & Boosting (Decision Trees)
  - Decision Trees Ensembles
  - Random Forest
  - Stochastic Gradient Boosting

Homogeneous model types (decision trees)

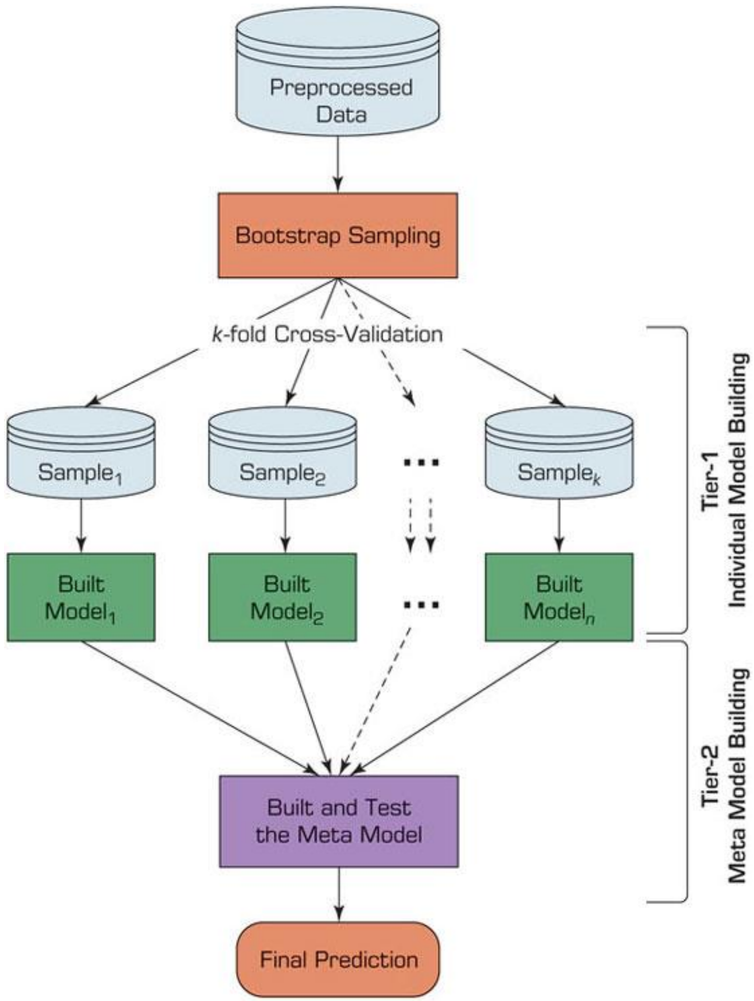
- Stacking
  - Stack generation or super learners
- Information Fusion
  - Any number of any models
  - Simple/weighted combining

Homogeneous model types (decision trees)

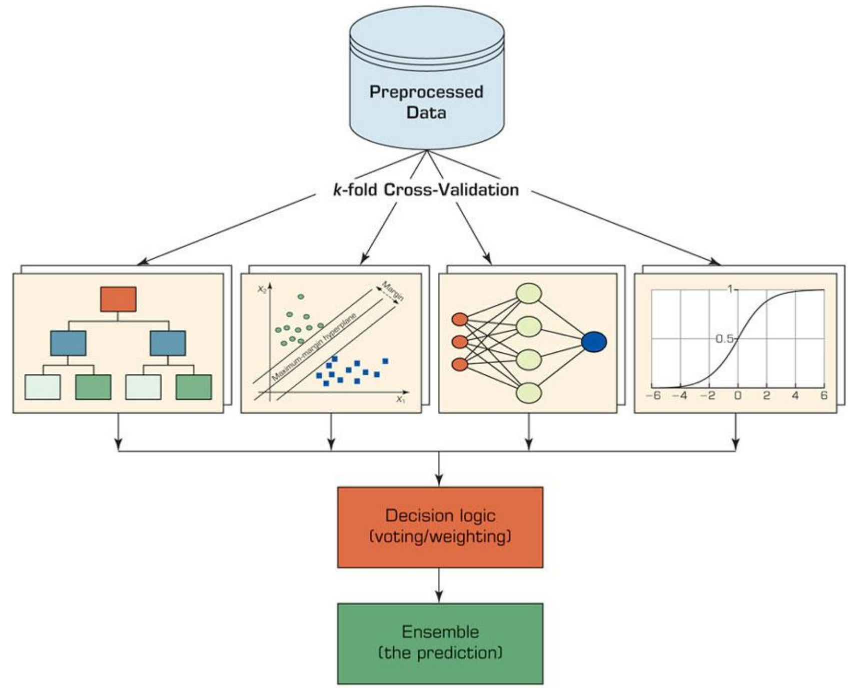
Pearson

Types of Ensemble Modeling (4 of 4)

- STACKING



- INFORMATION FUSION



Ensembles – Pros and Cons

Table 5.9 Brief List of Pros and Cons of Model Ensembles Compared to Individual Models.

PROS (Advantages)	Description
• Accuracy	Model ensembles usually result in more accurate models than individual models.
• Robustness	Model ensembles tend to be more robust against outliers and noise in the data set than individual models.
• Reliability (stable)	Because of the variance reduction, model ensembles tend to produce more stable, reliable, and believable results than individual models.
• Coverage	Model ensembles tend to have a better coverage of the hidden complex patterns in the data set than individual models.
CONS (Shortcomings)	**Description**
• Complexity	Model ensembles are much more complex than individual models.
• Computationally expensive	Compared to individual models, ensembles require more time and computational power to build.
• Lack of transparency (explainability)	Because of their complexity, it is more difficult to understand the inner structure of model ensembles (how they do what they do) than individual models.
• Harder to deploy	Model ensembles are much more difficult to deploy in an analytics-based Managerial decision-support system than single models.

Q & A