Data Mining & Predictive Analytics

Final Examination

1. Assuming K= 2 (2 clusters), apply K-means clustering to the one-field data set below (do it manually, without resorting to SPSS Modeler):

   24, 10, 29, 19, 2, 11, 9, 3, 1

   Do it __manually__, without software

   Solution 1: Let 3 and 19 in the given data be assumed as cluster centroids. The distance of each data point from the cluster centroids 3 and 19 is calculated by $(3-x_i)$ and $(19-x_i)$ where x_i is corresponding data point. The data is then segregated to cluster C1 and C2 as shown below figure

1st Iteration						
	Data	Dist. To C1=3	Dist. To C2=19	Cluster Memebership	Square of Distance to Cluster Center	
	24	21	5	C2	25	C2 20.75
	10	7	9	C1	49	C1 5
	29	26	10	C2	100	
	19	16	0	C2	0	
	2	1	17	C1	1	
	11	8	8	C2	64	
	9	6	10	C1	36	
	3	0	16	C1	0	
	1	2	18	C1	4	
				SSE:	279	Surrogate for within cluster variation
				d(m1,m2):	16	Surrogate for between cluster variation
				d(m1,m2)/SSE:	0.057	

The same calculations are repeated for iteration 2 and 3 as shown in below tables (also attached as solution1.xlsx file

2nd Iteration

Data	Dist. To C1=(10+2+9+3+1)/5 =5	Dist. To C2=(24+29+19+11)/4 =20.75	Cluster Memebership	Square of Distance to Cluster Center	
24	19	3.25	C2	10.56	C2 24
10	5	10.75	C1	25	C1 6
29	24	8.25	C2	68.06	
19	14	1.75	C2	3.06	
2	3	18.75	C1	9	
11	6	9.75	C1	36	
9	4	11.75	C1	16	
3	2	17.75	C1	4	
1	4	19.75	C1	16	
			SSE:	187.69	
			d(m1,m2):	15.75	
			d(m1,m2)/SSE:	0.08	

3rd Iteration

| Data	Dist. To C1=(10+2+11+9+3+1)/6=6	Dist. To C2=(24+29+19)/3=24	Cluster Memebership	Square of Distance to Cluster Center	
24	18	0	C2	0	C2 24
10	4	14	C1	16	C1 6
29	23	5	C2	25	
19	13	5	C2	25	
2	4	22	C1	16	
11	5	13	C1	25	
9	3	15	C1	9	
3	3	21	C1	9	
1	5	23	C1	25	
			SSE:	150	
			d(m1,m2):	18	
			d(m1,m2)/SSE:	0.12	

Since the cluster grouping has not changed for 2nd and 3rd iterations, the data points {10 2 11 9 3 1} belong to one cluster and {24, 29 19} belong to another.

---

2. Use SPSS Modeler and the Apriori algorithm to perform a market basket analysis of a dataset of 1500 transactions containing shopping information from a hardware store (AcmeHardware.xls). The file contains fields that indicate whether or not a customer, during a single visit, purchased a particular product category. Thus each record represents a store visit in which at least one product category was purchased. Find useful rules considering the following thresholds:

Minimum Antecedent (LHS) Support: 15%
Minimum Rule Confidence: 50%

Solution 2: The following stream is created in SPSS modeler using Apriori node in modeling tab.
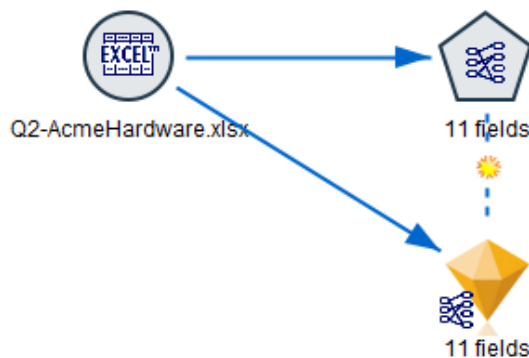
The minimum antecedent (LHS) support is taken as 15% and minimum rule confidence as 50%. Since there are 11 fields in the dataset, maximum number of antecedents is taken as 11. Since we don't what are on left hand side and on right hand side, all the fields are kept as antecedents and consequents.

The following are the transaction rules given all antecedents and consequents. 33 rules are generated out of which 1 rule is less than 1.5 and few are less than 2. Let 2.15 be the threshold greater than 2.15 (or equal) be with the improvement, which means 2.15 times better than random chance, with acceptable confidence and antecedent support can be neglected, as I consider this as the threshold. 13 rules highlighted in the below figure come under this condition.

Sort by: Lift ▼ | 33 of 33

Consequent	Antecedent	Confidence %	Rule Support %	Lift
Hardware	Wallpaper Paint	51.373	8.733	2.335
Tools	Plumbing Paint	64.961	11.0	2.256
Tools	Hardware Paint	62.605	9.933	2.174
Plumbing	Tools Paint	54.276	11.0	2.165
Plumbing	Flooring Paint	54.093	10.133	2.158
Wallpaper	Hardware Paint	55.042	8.733	2.15
Tools	Wallpaper Paint	61.176	10.4	2.124
Flooring	Plumbing Paint	59.843	10.133	2.102
Plumbing	Hardware Paint	51.681	8.2	2.062
Flooring	Hardware Paint	58.403	9.267	2.052
Tools	Flooring Paint	58.007	10.867	2.014
Tools	LawnCare Paint	57.895	10.267	2.01
Wallpaper	Tools Paint	51.316	10.4	2.005
Plumbing	LawnCare Paint	50.0	8.867	1.995

Consequent	Antecedent	Confidence %	Rule Support %	Lift
Flooring	LawnCare Paint	56.767	10.067	1.994
Wallpaper	Flooring Paint	50.89	9.533	1.988
Flooring	Wallpaper Paint	56.078	9.533	1.97
Tools	Plumbing	54.787	13.733	1.902
Flooring	Tools Paint	53.618	10.867	1.884
Tools	Hardware	53.333	11.733	1.852
Flooring	Plumbing	52.128	13.067	1.831
Tools	Wallpaper	50.26	12.867	1.745
Paint	Hardware	72.121	15.867	1.69
Paint	Tools	70.37	20.267	1.649
LawnCare	Flooring Paint	53.737	10.067	1.609
Paint	Plumbing	67.553	16.933	1.583
LawnCare	Plumbing Paint	52.362	8.867	1.568
Paint	Wallpaper	66.406	17.0	1.556
LawnCare	Hardware Paint	51.681	8.2	1.547
Paint	Flooring	65.808	18.733	1.542
LawnCare	Tools Paint	50.658	10.267	1.517
LawnCare	Wallpaper Paint	50.588	8.6	1.515
Paint	LawnCare	53.094	17.733	1.244

Consequent	Antecedent	Confidence %	Rule Support %	Lift
Hardware	Wallpaper and Paint	51.37	8.733	2.33
Tools	Plumbing and Paint	64.96	11.0	2.25
Tools	Hardware and Paint	62.61	9.93	2.17
Plumbing	Tools and Paint	54.28	11.0	2.16
Plumbing	Flooring and Paint	54.09	10.13	2.15
Wallpaper	Hardware and Paint	55.04	8.73	2.15

Since the improvement value for consequent as 'Hardware' and antecedent 'Wallpaper''paint' is high at lift of 2.33 with 51% confidence and rule support of 8.7% and can be ranked as 1, formulated as below

Manually,

LHS = Wallpaper, Paint and RHS = Hardware

Support (RHS = Hardware) is computed by counting hardware in total 1500 records

Lift = Confidence / Support (RHS = Hardware)

---

3. A financial institution has decided to apply data mining techniques to lower the risk associated with the loan approval process. To accomplish this task the company has decided to implement a Naïve Bayes classifier that, learning from historical data, is able to predict which loan applicants should be approved for loans. The table below depicts the data set (10 records) used in the machine learning process.

Record#	Income	Credit	ApproveLoan
1	High	Excellent	yes
2	High	Good	yes
3	Medium	Excellent	yes
4	High	Good	yes
5	High	Good	yes
6	Low	Excellent	yes
7	High	Bad	yes
8	Medium	Bad	no
9	High	Bad	no
10	Low	Good	No

Solution 3: The following is the stream created for Naïve Bayes classification
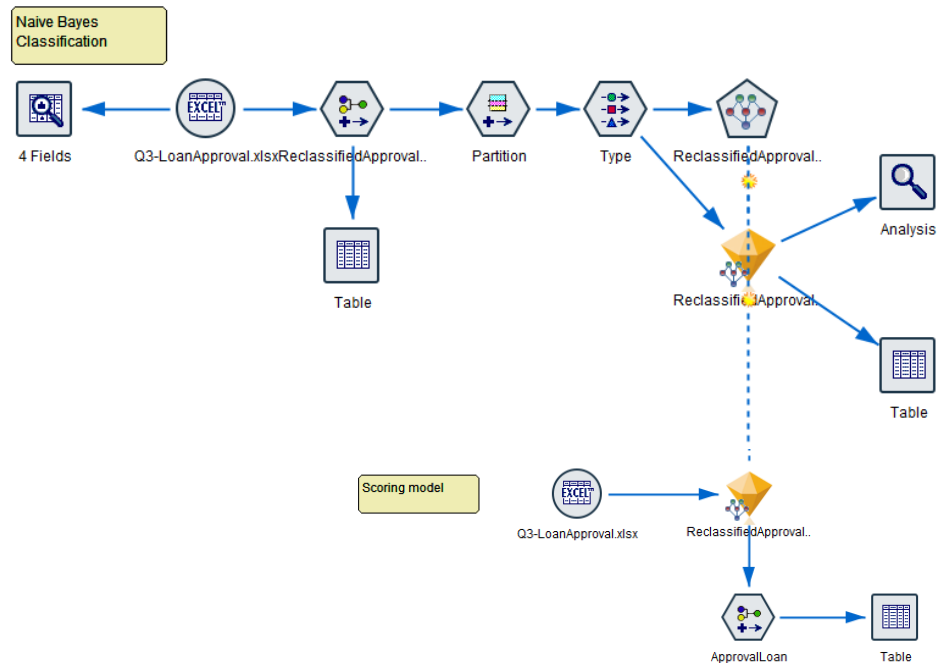
When the data audit node is run for the original given dataset, the following is shown in Audit tab

Field	Graph	Measurement	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
⊗ Record#		⟋ Continuous	1.000	10.000	5.500	3.028	0.000	--	10
A Income		▮ Categorical	--	--	--	--	--	3	10
A Credit		▮ Categorical	--	--	--	--	--	3	10
A ApproveL...		▮ Categorical	--	--	--	--	--	3	10

Since there are only 2 flag values for 'ApproveLoan' (showing as 3 categories in above figure) this attribute is reclassified to two Boolean values.

Reclassify values:

▶ Get	≫ Copy	✐ Clear new	⚡ Auto...

Original value	New value
No	0
no	0
yes	1

The resulting data audit node now gives relevant details

Field	Graph	Measurement	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
⊗ Record#		Continuous	1.000	10.000	5.500	3.028	0.000	--	10
A Income		Categorical	--	--	--	--	--	3	10
A Credit		Categorical	--	--	--	--	--	3	10
A ApproveLoan		Categorical	--	--	--	--	--	3	10
A ReclassifiedApprovalLoan		Flag	--	--	--	--	--	2	10

Since data set is very small, 50% of the data is taken as training and 30% for testing. The predictor importance and conditional probabilities of Naïve Bayesian classification on this data gives the following result on taking the reclassified ApproveLoan as target with Income and Credit as inputs

Predictor Importance
Target: ReclassifiedApprovalLoan

Conditional Probabilities of Income

Parents	Probability		
ReclassifiedApprovalLoan	High	Low	Medium
1	0.67	0.17	0.17
0	0.33	0.33	0.33
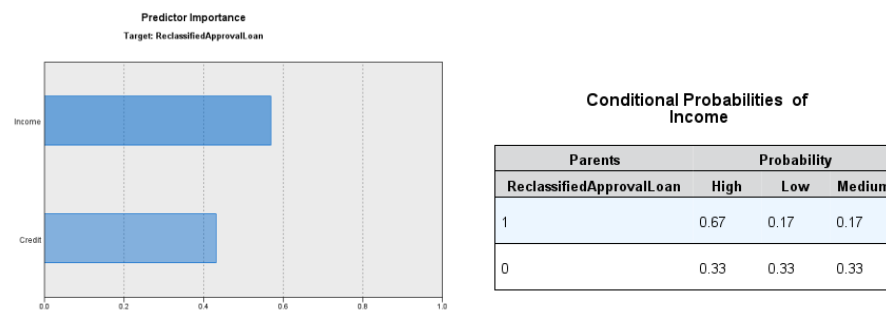
The analysis node gives the following accuracy values

— Collapse All   ++ Expand All

☐ Results for output field ReclassifiedApprovalLoan
  ☐ Comparing $B-ReclassifiedApprovalLoan with ReclassifiedApprovalLoan

'Partition'	1_Training		2_Testing	
Correct	6	85.71%	3	100%
Wrong	1	14.29%	0	0%
Total	7		3	

  ☐ Coincidence Matrix for $B-ReclassifiedApprovalLoan (rows show actuals)

'Partition' = 1_Training	0	1
0	2	1
1	0	4

'Partition' = 2_Testing	1
1	3

  ☐ Performance Evaluation

'Partition' = 1_Training	
0	0.847
1	0.336

'Partition' = 2_Testing	
1	0.0

	Record#	Income	Credit	ApproveLoan	ReclassifiedApprovalLoan	Partition	$B-ReclassifiedApprovalLo...	$BP-ReclassifiedApprovalLoan
1	1.000	High	Excellent	yes	1	1_Training	1	0.909
2	2.000	High	Good	yes	1	1_Training	1	0.966
3	3.000	Medium	Excellent	yes	1	2_Testing	1	0.909
4	4.000	High	Good	yes	1	2_Testing	1	0.966
5	5.000	High	Good	yes	1	1_Training	1	0.966
6	6.000	Low	Excellent	yes	1	1_Training	1	0.909
7	7.000	High	Bad	yes	1	2_Testing	1	0.500
8	8.000	Medium	Bad	no	0	1_Training	0	0.909
9	9.000	High	Bad	no	0	1_Training	1	0.500
10	10.000	Low	Good	No	0	1_Training	0	0.909

The prediction result is

Record#	Income	Credit	ApproveLoan	$B-ReclassifiedApprovalLoan	$BP-ReclassifiedApprovalLoan	ApprovalLoan
1	$null$ Medium	Bad	$null$	0	0.909	""no""

b) Build the classifier (<u>manually,</u> without software), by calculating the prior and conditional probability tables
This is attached in the excel file named 'solution2'. Since the conditional probability of no excellent credit zero, p(Credit=Excellent |no) = 0, the model is to be formulated to zero frequencies.

Probability table prior to zero frequency

Record#	Income	Credit	Approve Loan
1	High	Excellent	yes
2	High	Good	yes
3	Medium	Excellent	yes
4	High	Good	yes
5	High	Good	yes
6	Low	Excellent	yes
7	High	Bad	yes
8	Medium	Bad	no
9	High	Bad	no
10	Low	Good	No

Conditional Probability Tables

	Approve Loan	
Income	Y	N
High	=5/7	=1/3
Medium	=1/7	=1/3
Low	1/7	1/3
Credit		
Excellent	=3/7	0
Good	=3/7	=1/3
Bad	=1/7	=2/3

	Approve Loan	
Income	Y	N
High	0.714	0.333
Medium	0.143	0.333
Low	0.143	0.333
Credit		
Excellent	0.429	0.000
Good	0.429	0.333
Bad	0.143	0.667

Prior Probabilities

Prob(Approve Loan = yes) = 7/10	0.700	
Prob(Approve Loan = no) = 3/10	0.300	
p(income= medium, credit=bad | yes))	0.014	0.176	17.65%
p(income= medium, credit=bad | no))	0.067	0.824	82.35%

Probability tables after zero frequency

Record#	Income	Credit	Approve Loan
1	High	Excellent	yes
2	High	Good	yes
3	Medium	Excellent	yes
4	High	Good	yes
5	High	Good	yes
6	Low	Excellent	yes
7	High	Bad	yes
8	Medium	Bad	no
9	High	Bad	no
10	Low	Good	No

Conditional Probability Tables

	Approve Loan	
Income	Y	N
High	=6/10	=2/6
Medium	=2/10	=2/6
Low	2/10	2/6
Credit		
Excellent	=4/10	=1/6
Good	=4/10	=2/6
Bad	=2/10	=3/6

	Approve Loan	
Income	Y	N
High	0.600	0.333
Medium	0.200	0.333
Low	0.200	0.333
Credit		
Excellent	0.400	0.167
Good	0.400	0.333
Bad	0.200	0.500

Prob(Approve Loan = yes) = 7/10	0.700	
Prob(Approve Loan = no) = 3/10	0.300	
p(income= medium, credit=bad | yes))	0.028	0.359	35.90%
p(income= medium, credit=bad | no))	0.050	0.641	64.10%

c) Predict the outcome of the loan application corresponding to a customer with the following attributes: bad credit, medium income

Let the record: bad credit and medium income be x => x= (income=medium, credit=bad)
Probability of not getting a loan approval,

p(x | no) .p(no) = p(income=medium | no) *p(credit=bad| no) *p(approveloan=no)
= 2/6 * 3/6 *3/10 = 0.333 * 0.500 * 0.300 = 0.050

p(x | yes) .p(yes) = p(income=medium | yes) *p(credit=bad| yes) *p(approveloan=yes)
= 2/10 * 2/10 *7/10 = 0.200 * 0.200 * 0.700 = 0.028
p(no | x) = 0.050/(0.050+0.028) = 64.1%
p(yes |x) = 0.028/(0.050+0.028) = 35.9%
It is predicted that with a probability of 64.1% the load is not approved

---

4. The waste.xlsx file attached contains information from a waste management study in which the amount of solid waste produced within an area was related to type of land usage. Interest is in relating land usage to amount of waste produced for planning purposes. Inputs were found to be highly correlated and the dataset is used to demonstrate principal components regression. The file contains 40 records and the following fields:

INDUST Acreage (US) used for industrial work
METALS Acreage used for fabricated metal
TRUCKS Acreage used for trucking and wholesale trade
RETAIL Acreage used for retail trade
RESTRNTS Acreage used for restaurants and hotels
WASTE Amount of solid waste produced

a) Run a linear regression analysis predicting a target (amount of waste produced) as a function of several related inputs (amount of acreage put to different uses).
b) After examining the regression results, run a principal components analysis and comment on the results. What components would you choose and why?

Use SPSS Modeler for this question

---

5. The breast cancer dataset was obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg:

W.H. Wolberg, W.N. Street, and O.L. Mangasarian. Machine learning techniques to diagnose breast cancer from fine-needle aspirates. Cancer Letters 77 (1994) 163-171.
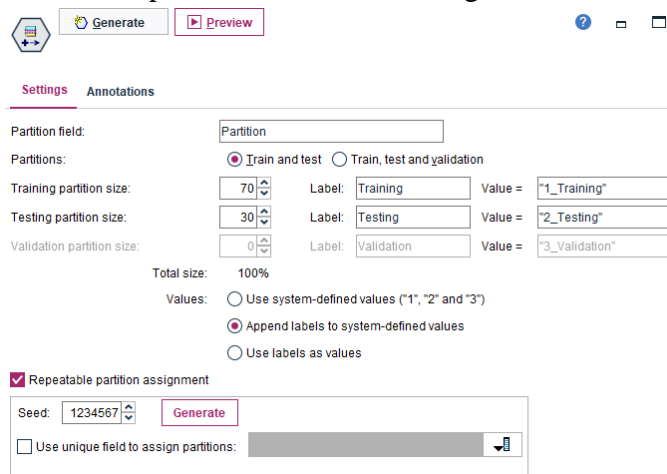
- Number of Instances: 699 ○
    Sample id number ○ Class:
    2(benign) – 4 (malignant) ○
    Clump Thickness 1-10 ○

Uniformity of Cell Size 1-10 ○
Uniformity of Cell Shape 1-10 ○
Marginal Adhesion 1-10 ○
Single Epithelial Cell Size 1-10
○   Bare Nuclei 1-10 ○
Bland Chromatin 1-10 ○
Normal Nucleoli 1-10 ○
Mitosis 1-10

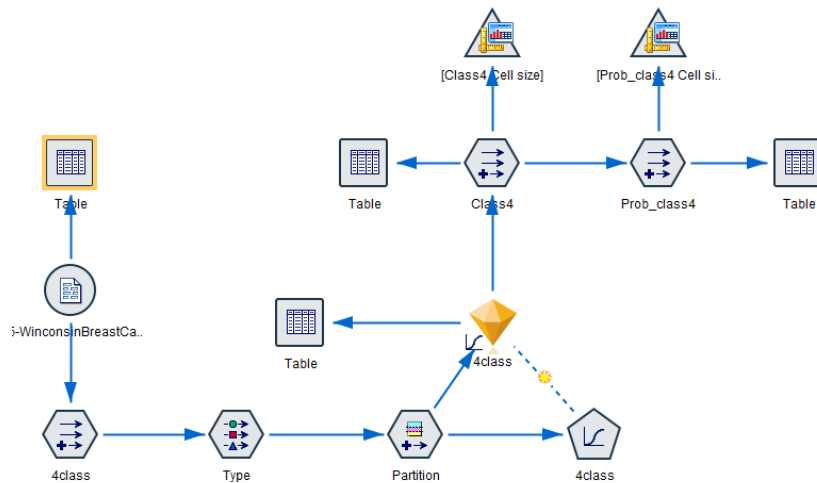Using SPSS Modeler for this question:

a) Partition the data 70/30.
The data is partitioned for 70 training and 30% testing

b) Build a binary logistic regression model using forward stepwise method to predict malignant tumors. Describe the model.
The model attached as stream file as 'Solution5' as below figure

The malignant class is extracted from the data by derive node.

Derive field:

4class

Derive as: Formula

Field type: <Default>

Formula:

```
1 if 'class' = 4 then "1" else "0" endif
```

As mentioned in the solution part a, using a partition node, 70% of the dataset is set for training and 30% on testing

Using logistic regression model, class field is set to target with remaining all attributes under model inputs using partition data.

As shown in above, forward stepwise method is used in order to predict the malignant tumors.

The result predictor importance is as below, which shows that out of all the attributes, Cell size, Bare nuclei, clump thickness and normal nucleoli holds high importance than other fields.

Predictor Importance
Target: 4class

Variables in the Equation
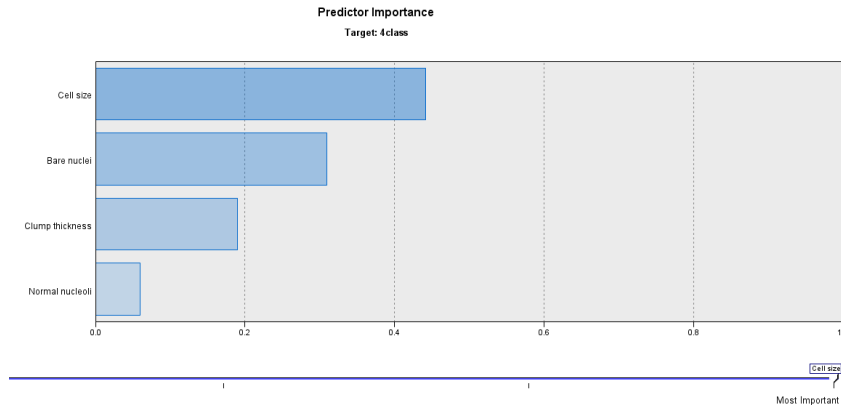
		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Cell size	1.604	.166	93.178	1	<.001	4.971
	Constant	-5.440	.490	123.380	1	<.001	.004
Step 2[b]	Cell size	1.125	.165	46.344	1	<.001	3.081
	Bare nuclei	.538	.093	33.788	1	<.001	1.713
	Constant	-6.061	.579	109.603	1	<.001	.002
Step 3[c]	Clump thickness	.616	.147	17.626	1	<.001	1.852
	Cell size	.855	.166	26.500	1	<.001	2.350
	Bare nuclei	.542	.108	25.291	1	<.001	1.719
	Constant	-8.310	.998	69.343	1	<.001	.000
Step 4[d]	Clump thickness	.605	.151	16.003	1	<.001	1.831
	Cell size	.679	.173	15.334	1	<.001	1.972
	Bare nuclei	.487	.107	20.886	1	<.001	1.628
	Normal nucleoli	.309	.127	5.888	1	.015	1.362
	Constant	-8.514	1.044	66.536	1	<.001	.000

a. Variable(s) entered on step 1: Cell size.

b. Variable(s) entered on step 2: Bare nuclei.

c. Variable(s) entered on step 3: Clump thickness.

d. Variable(s) entered on step 4: Normal nucleoli.

From the above, Logit of class malgnant = -5.440 + 1.604*cell size -6.061 + 1.125* cell size + bare nuclei * 0.538 -8.310 + 0.616 * clump thickness + 0.855 * cell size + 0.542 * bare nuclei -8.514 + 0.605*clump thickness + 0.679 *cekk size+ 0.487*Bare nuclei+ 0.309* normal nucleoli

This logit is used to derive the probabilities by the formula, $\hat{p} = \dfrac{1}{1+e^{-Logit}}$. As the logit increases, the probability tends to 1.

The positive coefficients in the logit indicate higher value on that predictor and associated with higher chance of malignant tumor and negative coefficients indicate that a higher value on the corresponding predictor is associated with lower probability. This can also be verified by the predictor importance figure above.

The corresponding coefficients, for example 1.604 for cell size tell that there is 6% more chance to be suffering with malignant tumor than others with cell size, holding all other parameters constant.

Classification Table

			Predicted		
			4class		Percentage
	Observed		0	1	Correct
Step 1	4class	0	337	7	98.0
		1	21	126	85.7
	Overall Percentage				94.3
Step 2	4class	0	336	8	97.7
		1	13	134	91.2
	Overall Percentage				95.7
Step 3	4class	0	336	8	97.7
		1	9	138	93.9
	Overall Percentage				96.5
Step 4	4class	0	338	6	98.3
		1	11	136	92.5
	Overall Percentage				96.5

Above table gives the accuracy classification on training data and confusion matrix. The predictive performance is as below tells that accuracy and AUC are high

Analysis   Annotations

— Collapse All   + Expand All

□ Results for output field 4class
  □ Individual Models
    □ Comparing $L-4class with 4class

'Partition'	1_Training		2_Testing	
Correct	474	96.54%	201	96.63%
Wrong	17	3.46%	7	3.37%
Total	491		208	

    □ Coincidence Matrix for $L-4class (rows show actuals)

'Partition' = 1_Training	0	1
0	338	6
1	11	136
'Partition' = 2_Testing	0	1
0	110	4
1	3	91

    □ Performance Evaluation

'Partition' = 1_Training	
0	0.324
1	1.163
'Partition' = 2_Testing	
0	0.574
1	0.751

  □ Evaluation Metrics

'Partition'	1_Training		2_Testing	
Model	AUC	Gini	AUC	Gini
$L-4class	0.996	0.991	0.99	0.979

The likelihood ratio test is similar to f test, since the significance of <.001 which is associated with malignant tumors.

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	435.656	1	<.001
	Block	435.656	1	<.001
	Model	435.656	1	<.001
Step 2	Step	49.348	1	<.001
	Block	485.004	2	<.001
	Model	485.004	2	<.001
Step 3	Step	24.164	1	<.001
	Block	509.169	3	<.001
	Model	509.169	3	<.001
Step 4	Step	6.489	1	.011
	Block	515.658	4	<.001
	Model	515.658	4	<.001

Since the above attributes have more predictor importance than other fields, considering the model with these 4 fields. The model is built as follows

c) Compute all relevant predictive performance metrics.

The below are the predictive performance metrics where accuracy is 96.63% which is higher than training and can say that there is no overfitting in the equation.

Recall = 91/(91+3) = 96.8%

Precision = 91/(91+4) = 95.79%

Specificity = 110/(110+4) = 96.50%

1- Specificity = 3.5%

95.79% of the individuals are predicted with malignant tumors are actually with tumor and approximately 3.5% are not with the tumor.

Analysis   Annotations

− Collapse All   ⁺₊ Expand All

⊟ Results for output field 4class
  ⊟ Individual Models
    ⊟ Comparing $L-4class with 4class

'Partition'	1_Training		2_Testing	
Correct	474	96.54%	201	96.63%
Wrong	17	3.46%	7	3.37%
Total	491		208	

⊟ Coincidence Matrix for $L-4class (rows show actuals)

'Partition' = 1_Training	0	1
0	338	6
1	11	136

'Partition' = 2_Testing	0	1
0	110	4
1	3	91

⊟ Performance Evaluation

'Partition' = 1_Training	
0	0.324
1	1.163

'Partition' = 2_Testing	
0	0.574
1	0.751

⊟ Evaluation Metrics

'Partition'	1_Training		2_Testing	
Model	AUC	Gini	AUC	Gini
$L-4class	0.996	0.991	0.99	0.979