**Data Mining & Predictive Analytics**

**Final Examination**

# Spring 2022

*INSTRUCTIONS*: Please answer any **four** questions in the exam.

The completed exam must be returned in no more than 24 hrs. (Note: although the problem statements are long, the questions are straight forward. The exam can be completed in a few hours. So you should not have much problem with it)

Completed exams must be posted in your Assignments’s tool in the ilearn site in a zip file labeled *lastname\_firstname\_finalexam.*

In preparing your answers, you may use and reference published or unpublished sources, the World Wide Web, and/or your own notes. However, you may not ask for or receive help from any person on the exam. Write-ups should be delivered in Word, PDF and/or Excel format.

When you return the exam, you will be required to include a cover sheet (look in the next page) with a statement indicating that the exam is entirely your own work. Your name should appear on this cover sheet.

A comment on academic honesty (may sound obsessive, but it is good advice):

As I have mentioned in previous occasions, the nature of online work sometimes leads to questionable decision making. This is an exam, and therefore it is individual work. **Please avoid the temptation of working in groups**. It is not only morally wrong, it is also myopic, as group work presented as individual work stands out, no matter how much you try to disguise it.

**Before doing something silly, please read Marist College Academic Integrity Policy.** <https://www.marist.edu/academic-resources/advising/academic-integrity-policy>Good luck!!

EL //

**VERY IMPORTANT: Please complete the Team Peer Evaluation Form that I have included in the zip file, and submit it with your final exam. Please note that this is a mandatory requirement.**

**Data Mining & Predictive Analytics**

**Final Examination**

# Spring 2022

I hereby certify that I have completed the attached examination materials, using only my own efforts. I have not asked for or received help from any person in completing this exam.

Swetha Adike 5/8/2022

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_

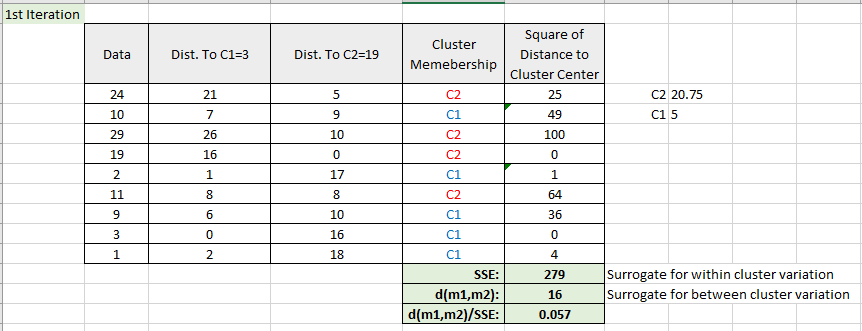
(Name) (date)

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):

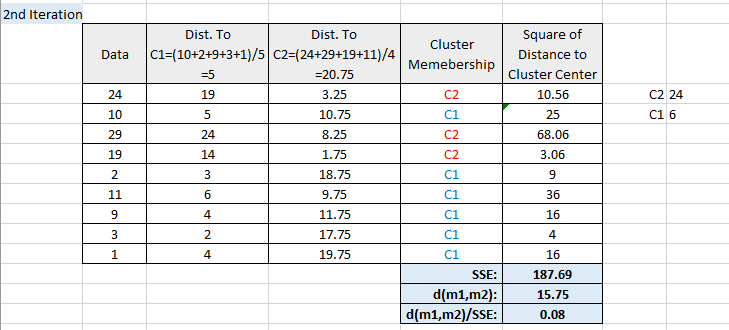
24, 10, 29, 19, 2, 11, 9, 3, 1

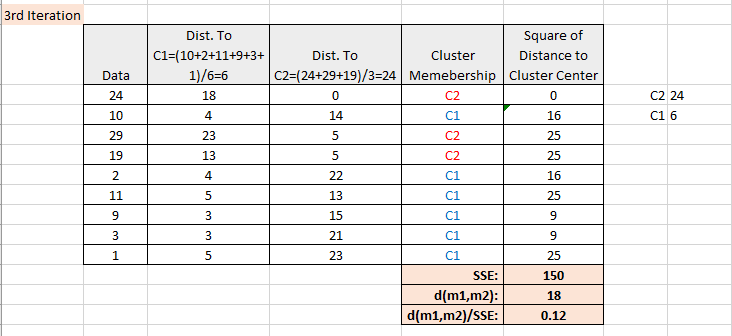
Do it manually, without software

**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-xi) and (19-xi) where xi is corresponding data point. The data is then segregated to cluster C1 and C2 as shown below figure



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





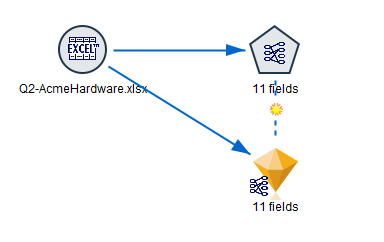
Since the cluster grouping has not changed for 2nd and 3rd iterations, the data points {10 2 11 9 3 1} belong to one cluster and {24, 29 19} belong to another.

1. 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%

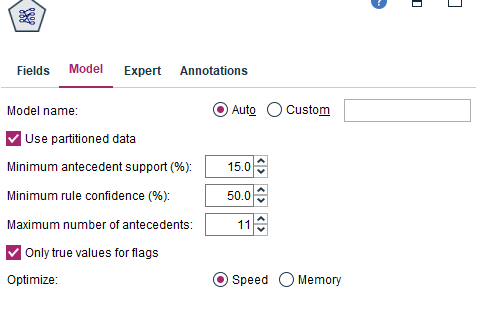
Minimum Rule Confidence: 50%

**Solution 2:** The following stream is created in SPSS modeler using Apriori node in modeling tab.

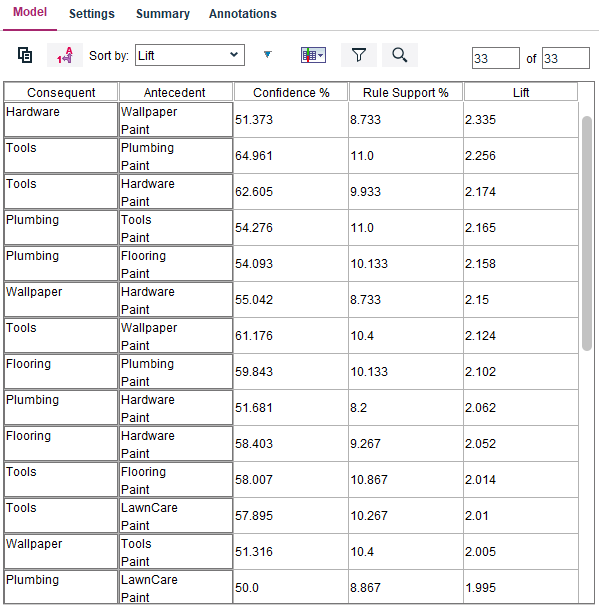


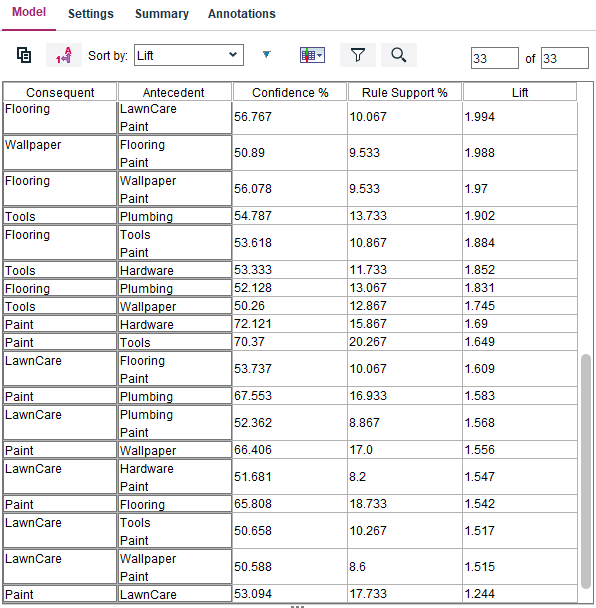
The minimum antecedent (LHS) support is taken as 15% and minimum rule confidence as 50%. Since there are 11 fields in the dataset, maximum number of antecedents is taken as 11.

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.





|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 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

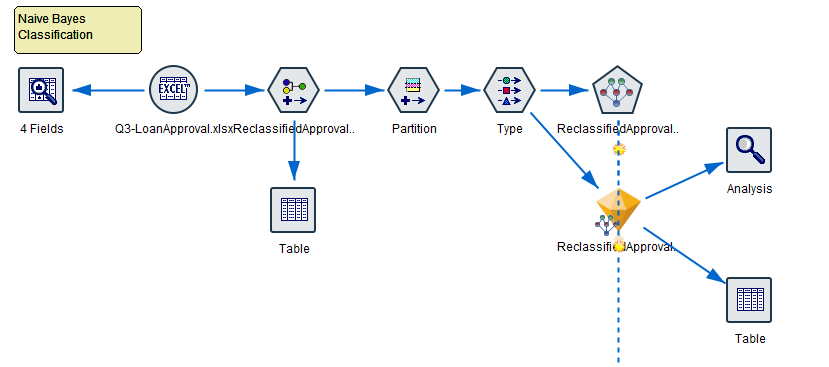
Support (RHS = Hardware) is computed by counting hardware in total 1500 records

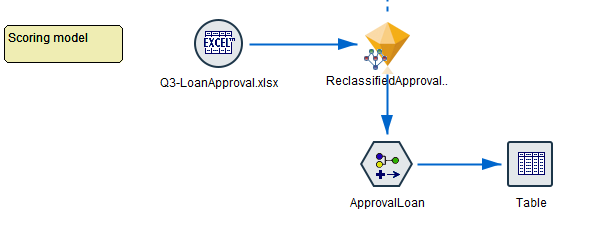
Lift = Confidence / Support (RHS = Hardware)

1. 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.

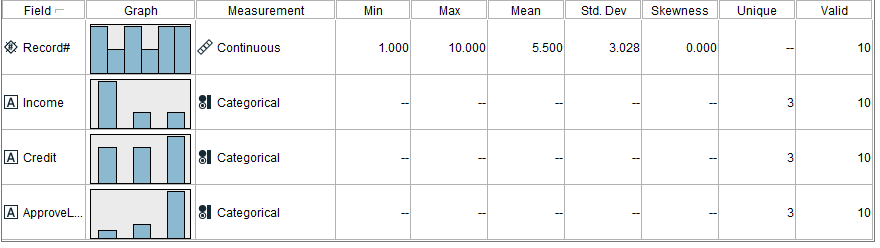
|  |  |  |  |
| --- | --- | --- | --- |
| **Record#** | **Income** | **Credit** | **ApproveLoan** |
| 1 | High | Excellent | yes |
| 2 | High | Good | yes |
| 3 | Medium | Excellent | yes |
| 4 | High | Good | yes |
| 5 | High | Good | yes |
| 6 | Low | Excellent | yes |
| 7 | High | Bad | yes |
| 8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

**Solution 3:** The following is the stream created for Naïve Bayes classification

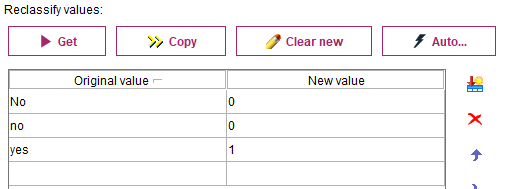




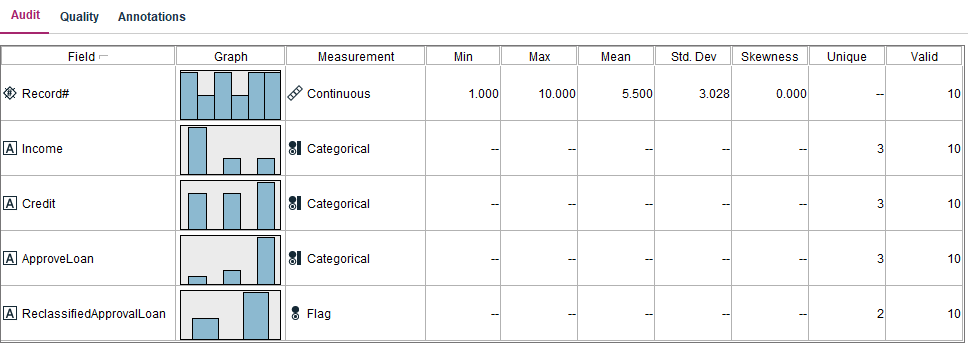
When the data audit node is run for the original given dataset, the following is shown in Audit tab



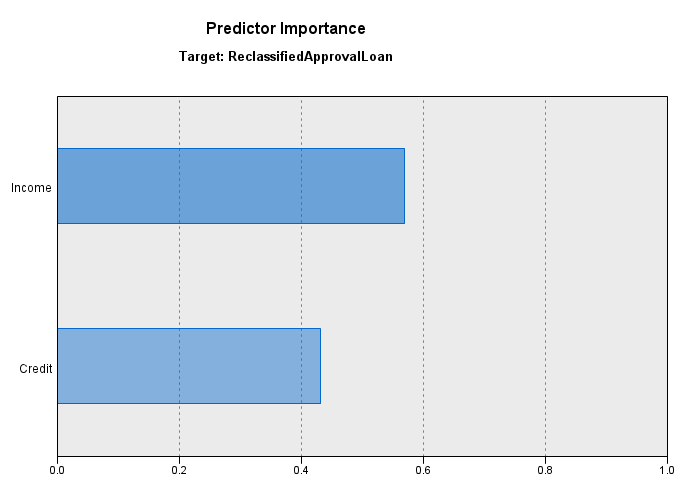
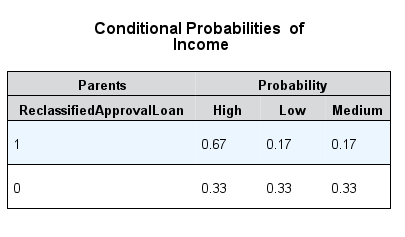
Since there are only 2 flag values for ‘ApproveLoan’ (showing as 3 categories in above figure) this attribute is reclassified to two Boolean values.



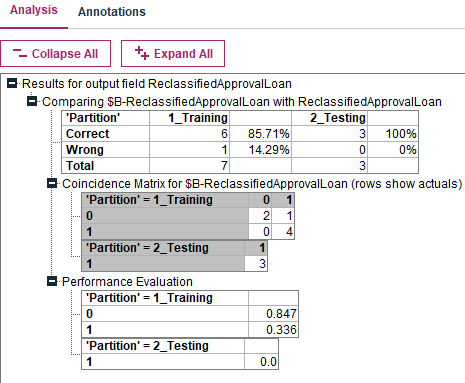
The resulting data audit node now gives relevant details

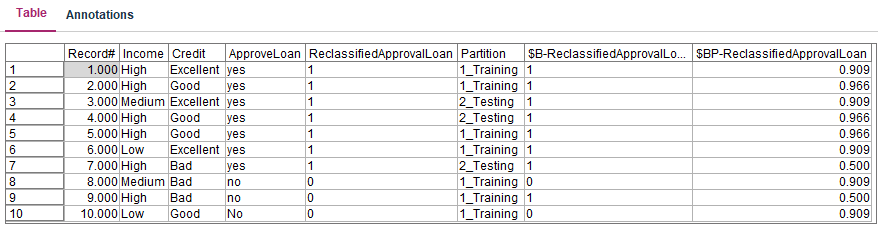


Since data set is very small, 50% of the data is taken as training and 30% for testing. The predictor importance and conditional probabilities of Naïve Bayesian classification on this data gives the following result on taking the reclassified ApproveLoan as target with Income and Credit as inputs

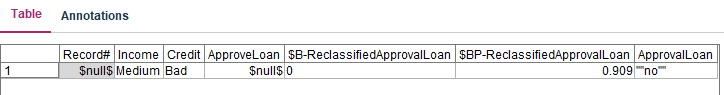
 

The analysis node gives the following accuracy values





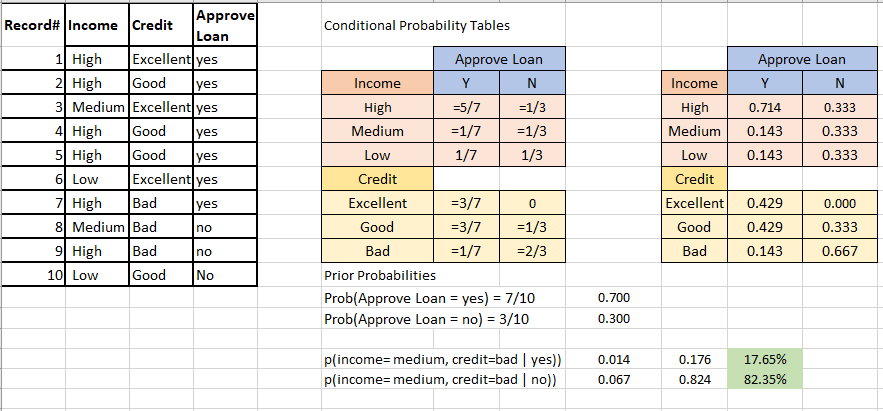
The prediction result is



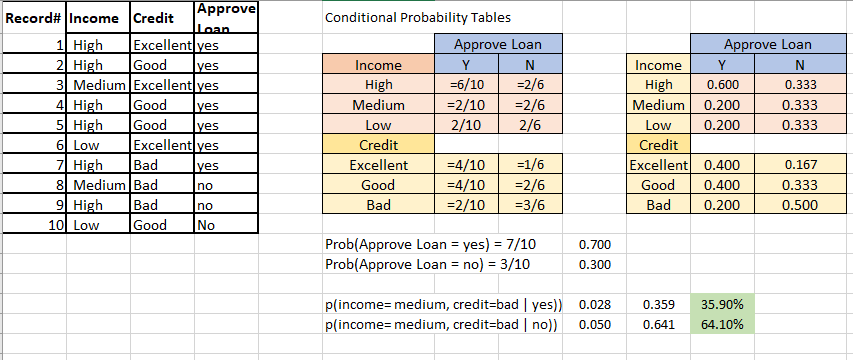
* 1. Build the classifier (manually, without software), by calculating the prior and conditional probability tables

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



Probability tables after zero frequency



* 1. 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

1. The waste.xlsx file attachedcontains 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

1. 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).
2. 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.*

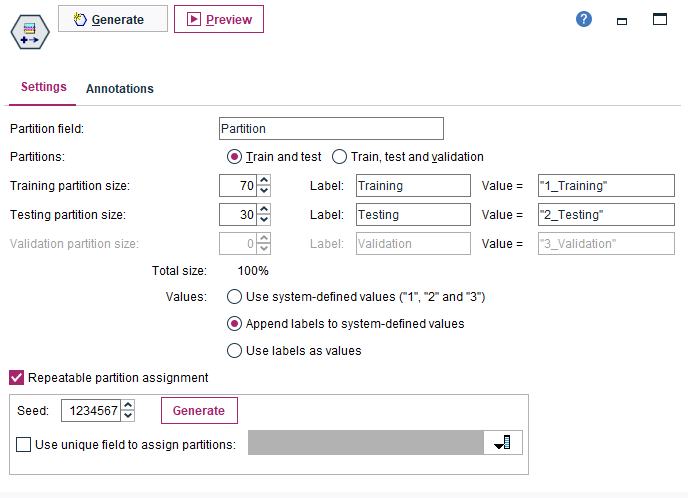
• Number of Instances: 699 o Sample id number o Class: 2(benign) – 4 (malignant) o Clump Thickness 1-10 o Uniformity of Cell Size 1-10 o Uniformity of Cell Shape 1-10 o Marginal Adhesion 1-10 o Single Epithelial Cell Size 1-10

o Bare Nuclei 1-10 o Bland Chromatin 1-10 o Normal Nucleoli 1-10 o Mitosis 1-10

Using SPSS Modeler for this question:

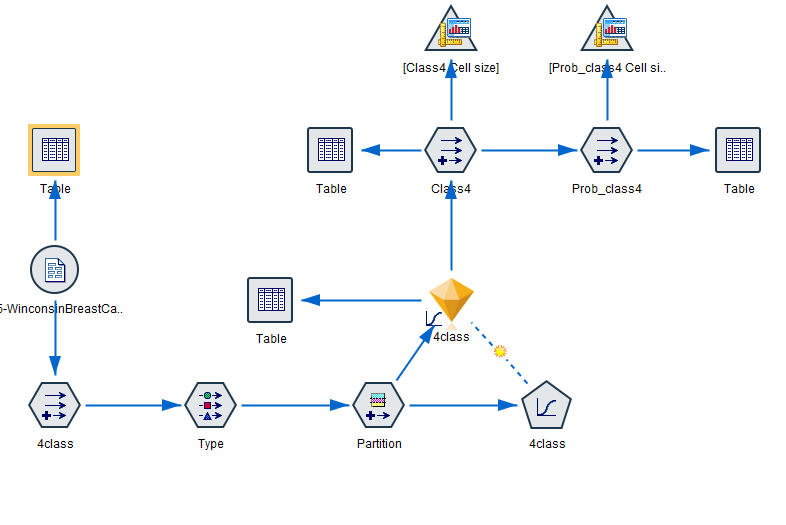
1. Partition the data 70/30.

The data is partitioned for 70 training and 30% testing

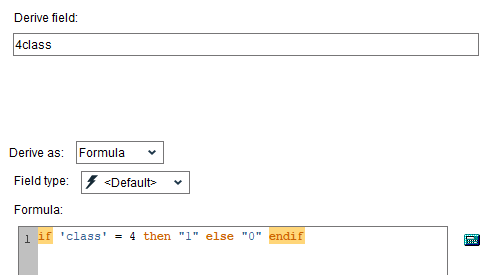


1. Build a binary logistic regression model using forward stepwise method to predict malignant tumors. Describe the model.

The model attached as stream file as ‘Solution5’ as below figure

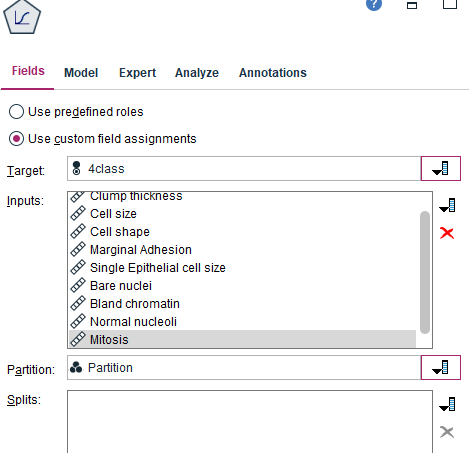


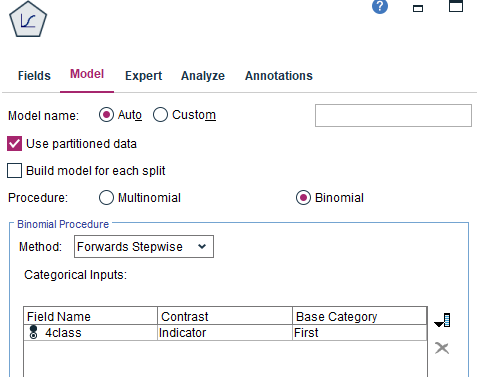
The malignant class is extracted from the data by derive node.



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

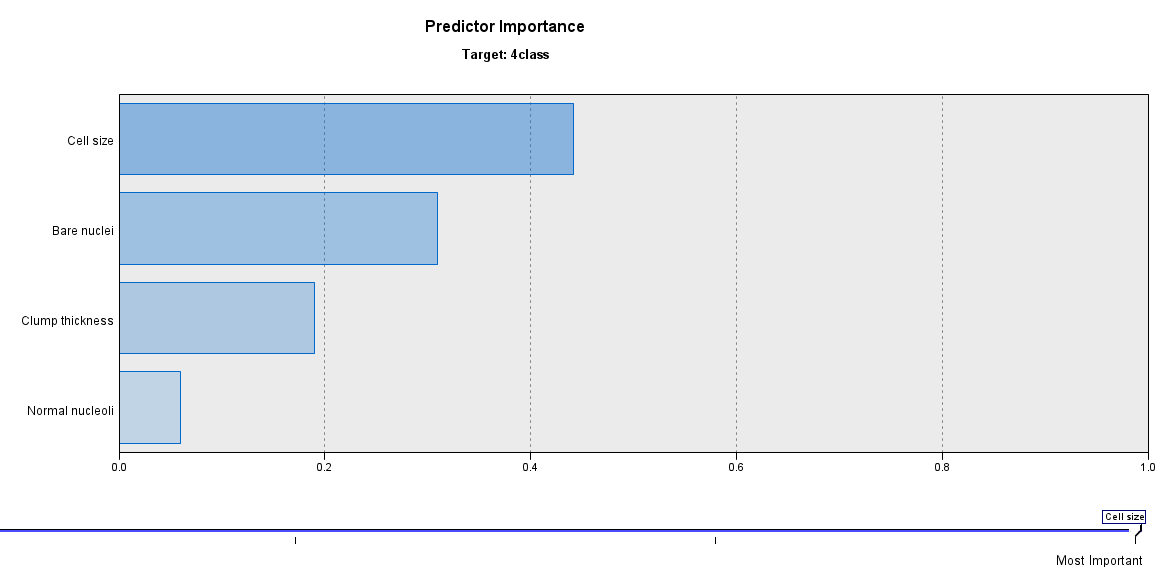
Using logistic regression model, class field is set to target with remaining all attributes under model inputs using partition data.

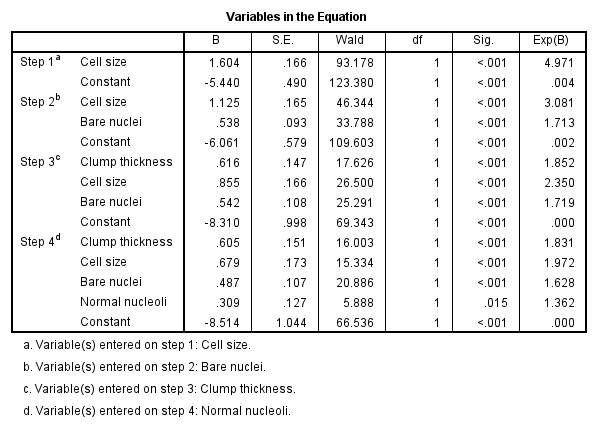




As shown in above, forward stepwise method is used in order to predict the malignant tumors.

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.



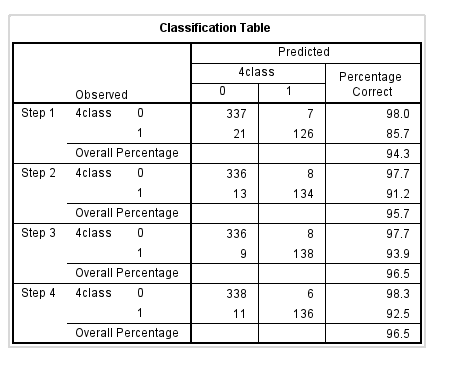


From the above, Logit of class malgnant = -5.440 + 1.604\*cell size -6.061 + 1.125\* cell size + bare nuclei \* 0.538 -8.310 + 0.616 \* clump thickness + 0.855 \* cell size + 0.542 \* bare nuclei -8.514 + 0.605\*clump thickness + 0.679 \*cekk size+ 0.487\*Bare nuclei+ 0.309\* normal nucleoli

This logit is used to derive the probabilities by the formula,  . As the logit increases, the probability tends to 1.

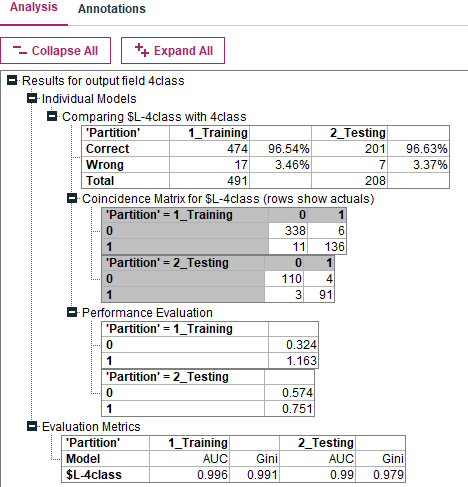
The positive coefficients in the logit indicate higher value on that predictor and associated with higher chance of malignant tumor and negative coefficients indicate that a higher value on the corresponding predictor is associated with lower probability. This can also be verified by the predictor importance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

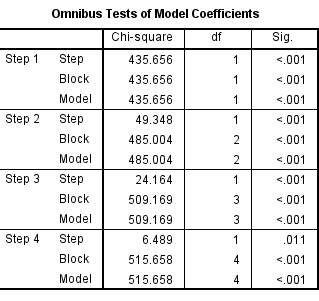


Above table gives the accuracy classification on training data and confusion matrix.

The predictive performance is as below tells that accuracy and AUC are high



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



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

1. 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.

