Data Analytics and Mining

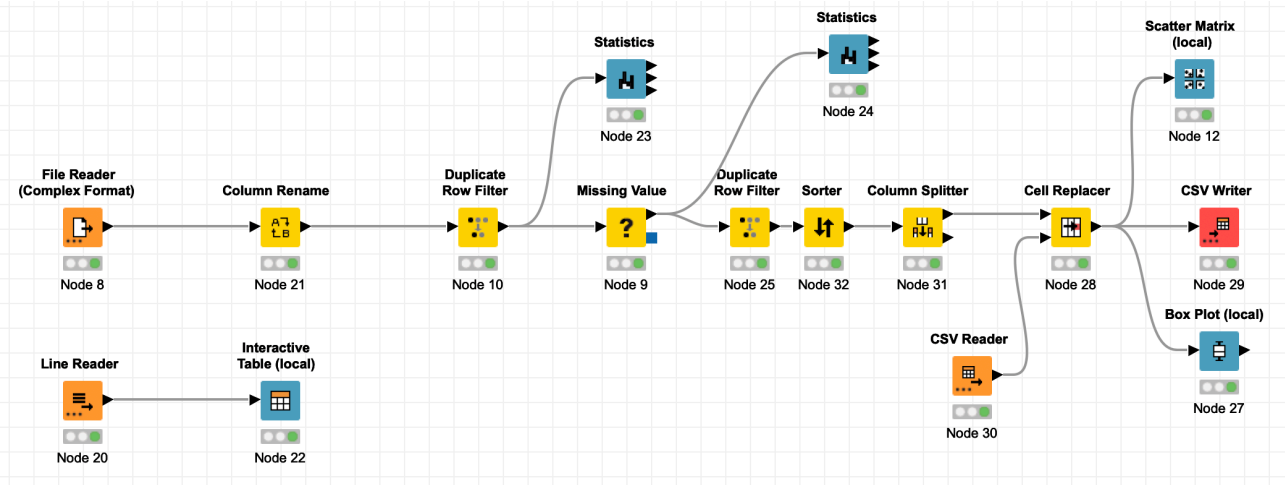
Breast Cancer Diagnosis using SVM and Naïve Bayes Models with Cross validation

Task 1

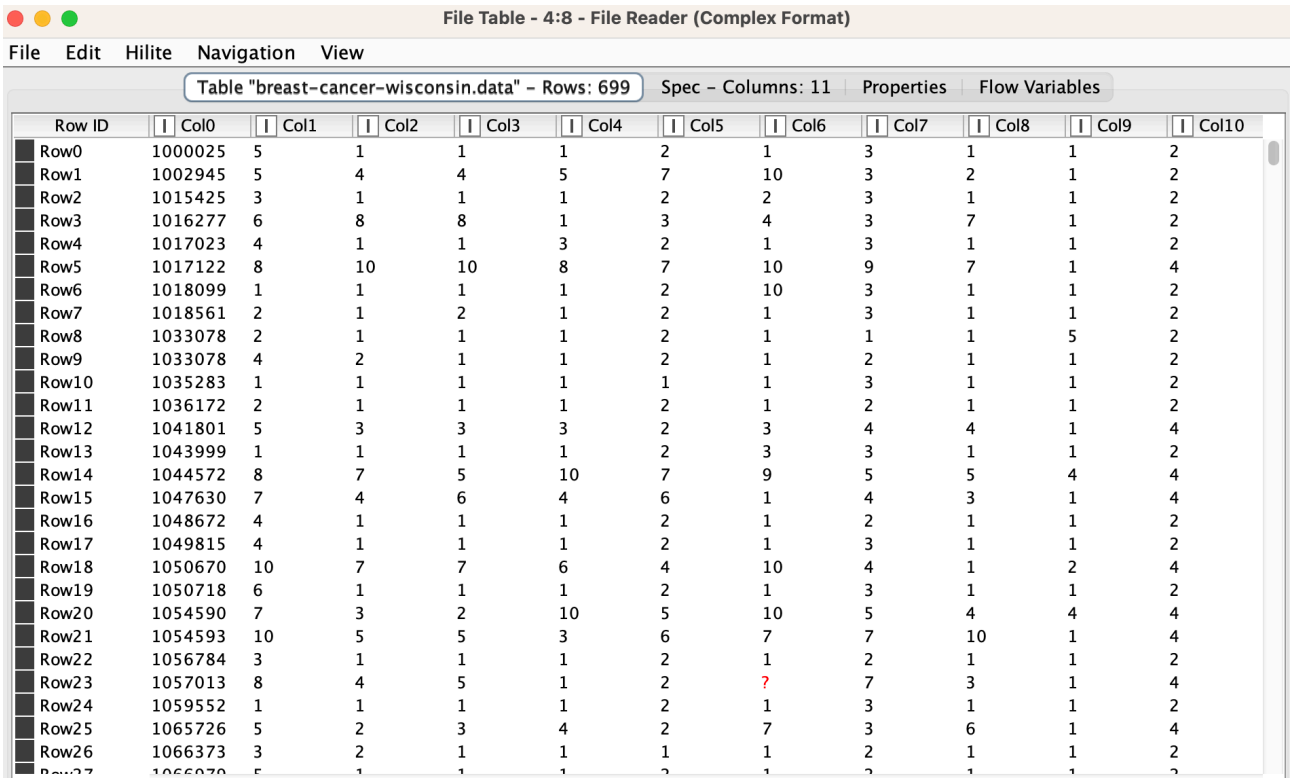
Knime Workflow is presented in **Figure 1.1**. Discussion on the dataset is discussed further.

Figure 1.1: Workflow for Data pre-processing and data-cleaning

1. Referring to **Figure 1.1** the dataset provided in "breast-cancer-wisconsin.data" was in .csv format was read - it was read using the using a file reader node **(node 8 in Figure 1.1)**. **Figure**



1.2 shows how data is represented when is fetched from the .csv file. We can observe that the column names are not mentioned.



Row ID	Col0	Col1	Col2	Col3	Col4	Col5	Col6	Col7	Col8	Col9	Col10
Row0	1000025	5	1	1	1	2	1	3	1	1	2
Row1	1002945	5	4	4	5	7	10	3	2	1	2
Row2	1015425	3	1	1	1	2	2	3	1	1	2
Row3	1016277	6	8	8	1	3	4	3	7	1	2
Row4	1017023	4	1	1	3	2	1	3	1	1	2
Row5	1017122	8	10	10	8	7	10	9	7	1	4
Row6	1018099	1	1	1	1	2	10	3	1	1	2
Row7	1018561	2	1	2	1	2	1	3	1	1	2
Row8	1033078	2	1	1	1	2	1	1	1	5	2
Row9	1033078	4	2	1	1	2	1	2	1	1	2
Row10	1035283	1	1	1	1	1	1	3	1	1	2
Row11	1036172	2	1	1	1	2	1	2	1	1	2
Row12	1041801	5	3	3	3	2	3	4	4	1	4
Row13	1043999	1	1	1	1	2	3	3	1	1	2
Row14	1044572	8	7	5	10	7	9	5	5	4	4
Row15	1047630	7	4	6	4	6	1	4	3	1	4
Row16	1048672	4	1	1	1	2	1	2	1	1	2
Row17	1049815	4	1	1	1	2	1	3	1	1	2
Row18	1050670	10	7	7	6	4	10	4	1	2	4
Row19	1050718	6	1	1	1	2	1	3	1	1	2
Row20	1054590	7	3	2	10	5	10	5	4	4	4
Row21	1054593	10	5	5	3	6	7	7	10	1	4
Row22	1056784	3	1	1	1	2	1	2	1	1	2
Row23	1057013	8	4	5	1	2	?	7	3	1	4
Row24	1059552	1	1	1	1	2	1	3	1	1	2
Row25	1065726	5	2	3	4	2	7	3	6	1	4
Row26	1066373	3	2	1	1	1	1	2	1	1	2

Figure 1.2: View of the file table from File Reader node

2. To determine the column names we read the "breast-cancer-wisconsin.names" files using a line reader node **(node 20 in Figure 1.1)**. A screenshot of the output is shown in **Figure 1.3** and **Figure 1.4**. From the figures below we can see the configuration and column names. In **Figure 1.4**, the column names can be observed

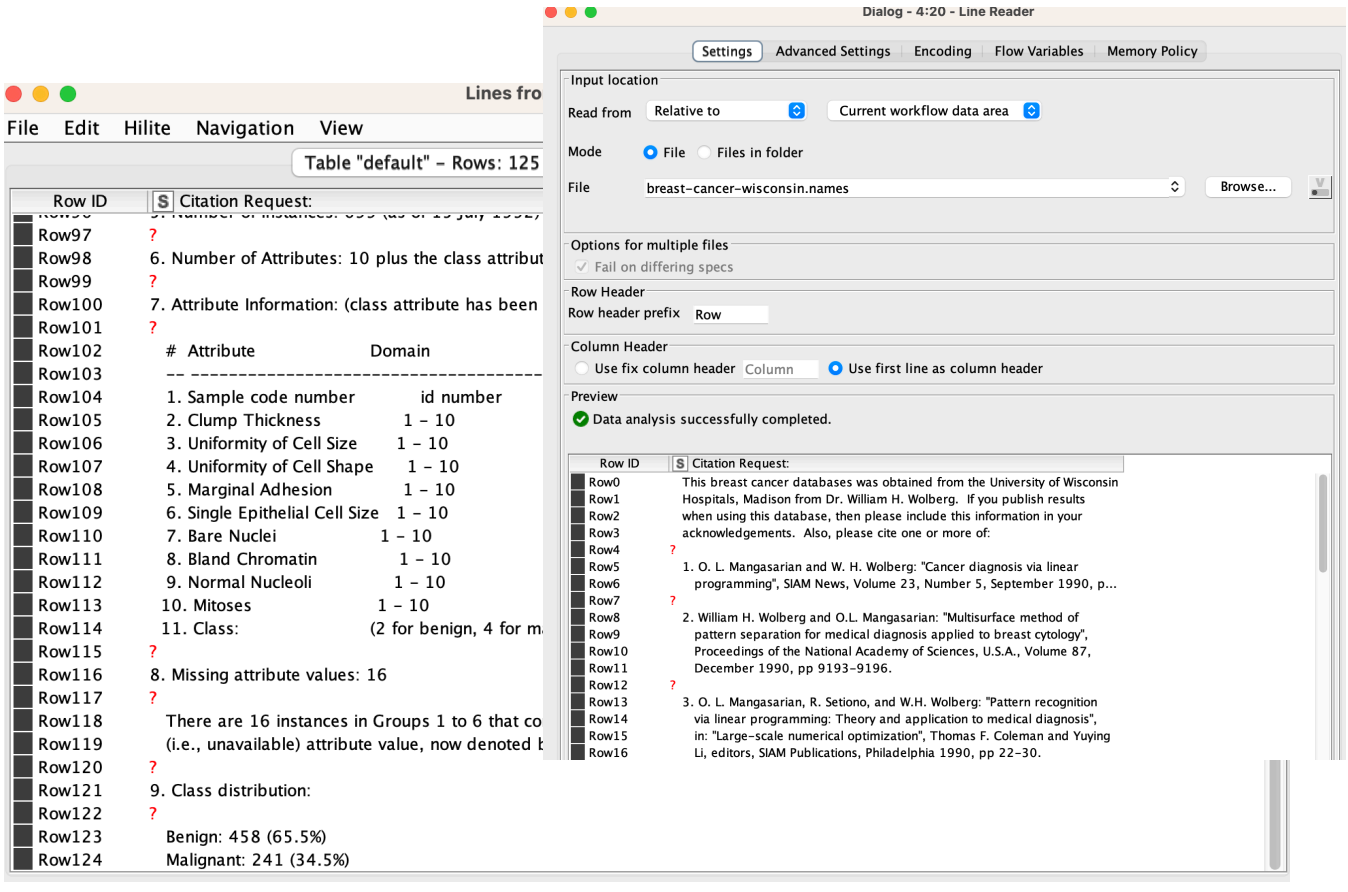


Figure 1.3: Data fetched from "breast-cancer-wisconsin.names" file

Figure 1.4: Data fetched from "breast-cancer-wisconsin.names" file

3. The column names are added to the data table using Column Rename node **(Node 21 in Figure 1.1)**. Configuration and the resulting table is shown in **Figure 1.5** and once the column is renamed the output is shown in **Figure 1.6**

Figure 1.5: Configuration for Column rename node (Node 21)

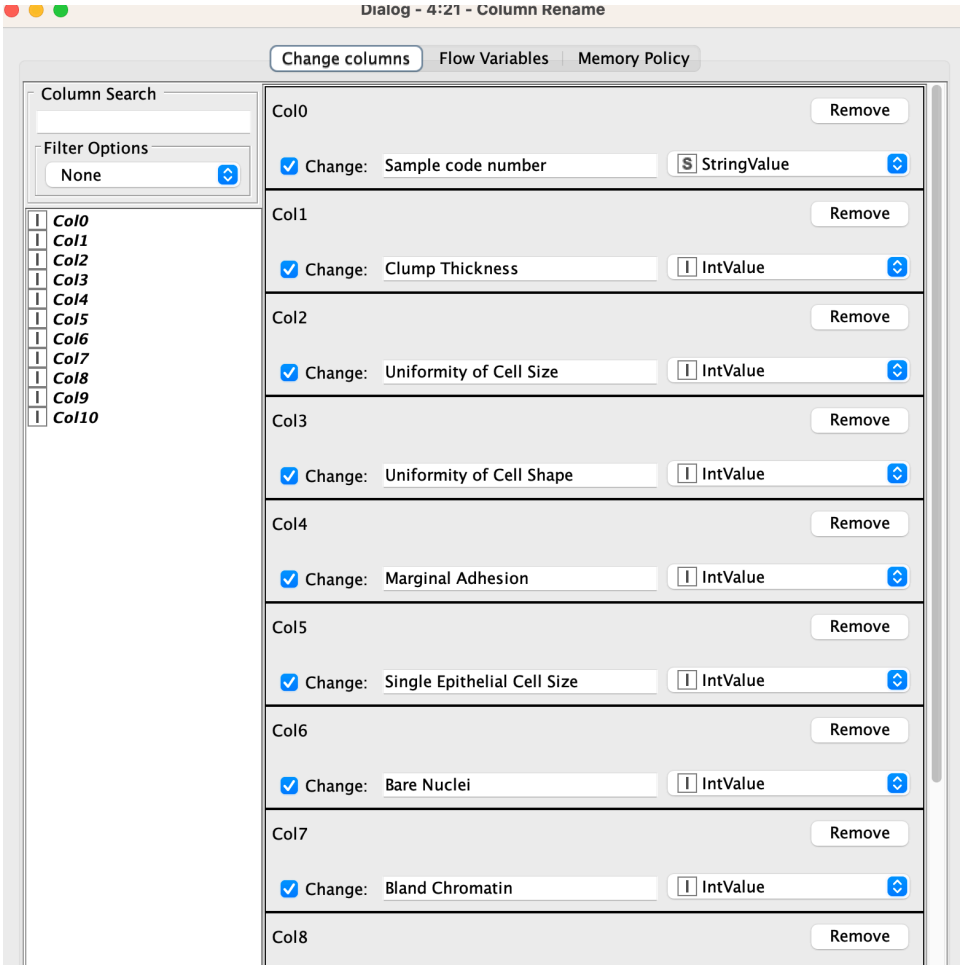
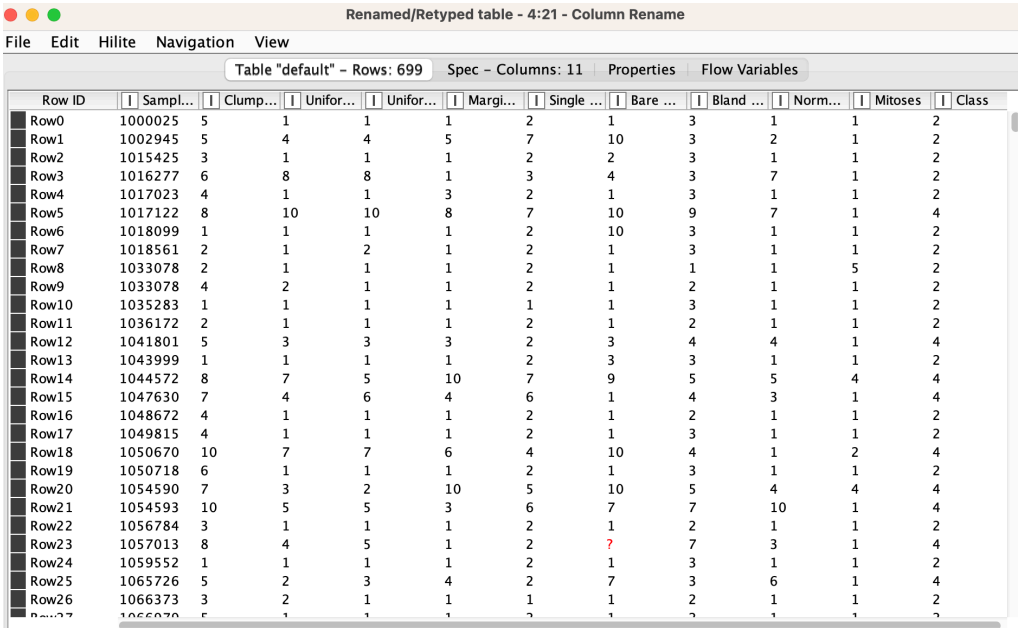


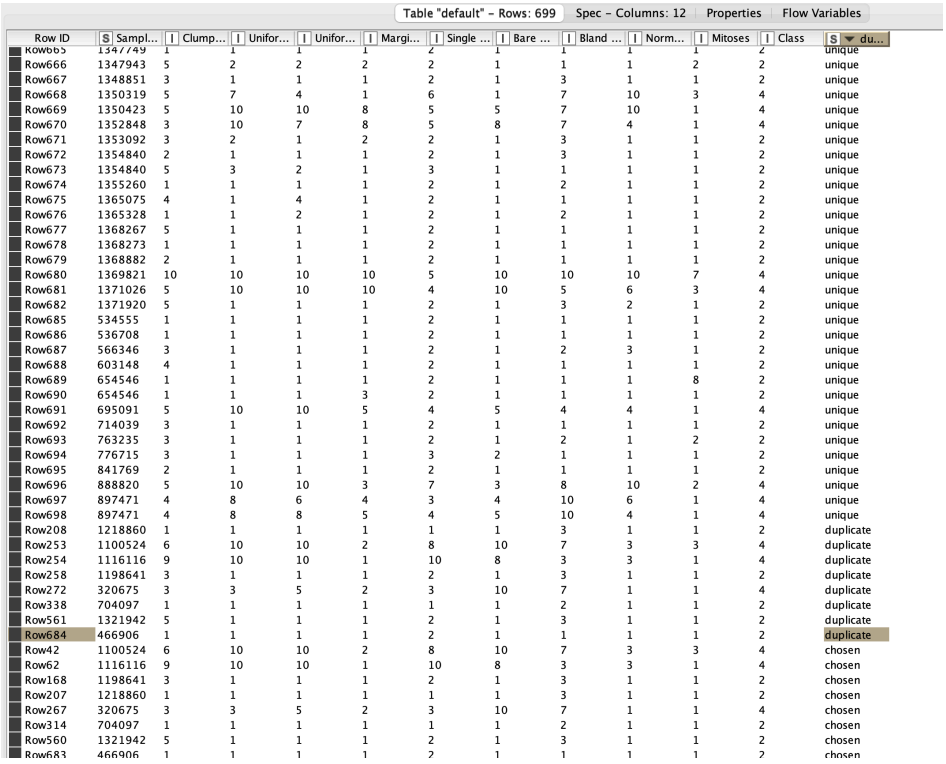
Figure 1.6: Output after columns are renamed



Row ID	Sampl...	Clump...	Unifor...	Unifor...	Margi...	Single ...	Bare ...	Bland ...	Norm...	Mitoses	Class
Row0	1000025	5	1	1	1	2	1	3	1	1	2
Row1	1002945	5	4	4	5	7	10	3	2	1	2
Row2	1015425	3	1	1	1	2	2	3	1	1	2
Row3	1016277	6	8	8	1	3	4	3	7	1	2
Row4	1017023	4	1	1	3	2	1	3	1	1	2
Row5	1017122	8	10	10	8	7	10	9	7	1	4
Row6	1018099	1	1	1	1	2	10	3	1	1	2
Row7	1018561	2	1	2	1	2	1	3	1	1	2
Row8	1033078	2	1	1	1	2	1	1	1	5	2
Row9	1033078	4	2	1	1	2	1	2	1	1	2
Row10	1035283	1	1	1	1	1	1	3	1	1	2
Row11	1036172	2	1	1	1	2	1	2	1	1	2
Row12	1041801	5	3	3	3	2	3	4	4	1	4
Row13	1043999	1	1	1	1	2	3	3	1	1	2
Row14	1044572	8	7	5	10	7	9	5	5	4	4
Row15	1047630	7	4	6	4	6	1	4	3	1	4
Row16	1048672	4	1	1	1	2	1	2	1	1	2
Row17	1049815	4	1	1	1	2	1	3	1	1	2
Row18	1050670	10	7	7	6	4	10	4	1	2	4
Row19	1050718	6	1	1	1	2	1	3	1	1	2
Row20	1054590	7	3	2	10	5	10	5	4	4	4
Row21	1054593	10	5	5	3	6	7	7	10	1	4
Row22	1056784	3	1	1	1	2	1	2	1	1	2
Row23	1057013	8	4	5	1	2	?	7	3	1	4
Row24	1059552	1	1	1	1	2	1	3	1	1	2
Row25	1065726	5	2	3	4	2	7	3	6	1	4
Row26	1066373	3	2	1	1	1	1	2	1	1	2

4. Next, the duplicate row filter **(node 10 in Figure 1.1)** is used to remove duplicate entries. In configuration, duplicate rows are retained and a separate column is added to show the unique, chosen and duplicate entries. Once this node is executed, unique chosen and duplicate rows can be viewed - refer **Figure 1.7**

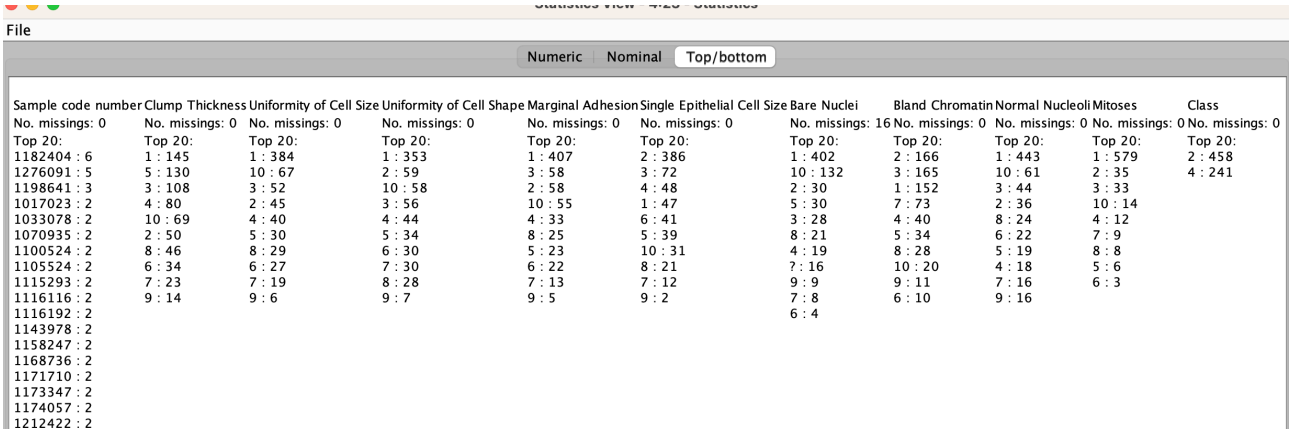
Figure 1.7: Showing duplicate row entries



Row ID	S Sampl...	Clump...	Unifor...	Unifor...	Margi...	Single ...	Bare ...	Bland ...	Norm...	Mitoses	Class	S du...
Row665	1347749	1	1	1	1	2	1	1	1	1	2	unique
Row666	1347943	5	2	2	2	2	1	1	1	2	2	unique
Row667	1348851	3	1	1	1	2	1	3	1	1	2	unique
Row668	1350319	5	7	4	1	6	1	7	10	3	4	unique
Row669	1350423	5	10	10	8	5	5	7	10	1	4	unique
Row670	1352848	3	10	7	8	5	8	7	4	1	4	unique
Row671	1353092	3	2	1	2	2	1	3	1	1	2	unique
Row672	1354840	2	1	1	1	2	1	3	1	1	2	unique
Row673	1354840	5	3	2	1	3	1	1	1	1	2	unique
Row674	1355260	1	1	1	1	2	1	2	1	1	2	unique
Row675	1365075	4	1	4	1	2	1	1	1	1	2	unique
Row676	1365328	1	1	2	1	2	1	2	1	1	2	unique
Row677	1368267	5	1	1	1	2	1	1	1	1	2	unique
Row678	1368273	1	1	1	1	2	1	1	1	1	2	unique
Row679	1368882	2	1	1	1	2	1	1	1	1	2	unique
Row680	1369821	10	10	10	10	5	10	10	10	7	4	unique
Row681	1371026	5	10	10	10	4	10	5	6	3	4	unique
Row682	1371920	5	1	1	1	2	1	3	2	1	2	unique
Row685	534555	1	1	1	1	2	1	1	1	1	2	unique
Row686	536708	1	1	1	1	2	1	1	1	1	2	unique
Row687	566346	3	1	1	1	2	1	2	3	1	2	unique
Row688	603148	4	1	1	1	2	1	1	1	1	2	unique
Row689	654546	1	1	1	1	2	1	1	1	8	2	unique
Row690	654546	1	1	1	3	2	1	1	1	1	2	unique
Row691	695091	5	10	10	5	4	5	4	4	1	4	unique
Row692	714039	3	1	1	1	2	1	1	1	1	2	unique
Row693	763235	3	1	1	1	2	1	2	1	2	2	unique
Row694	776715	3	1	1	1	3	2	1	1	1	2	unique
Row695	841769	2	1	1	1	2	1	1	1	1	2	unique
Row696	888820	5	10	10	3	7	3	8	10	2	4	unique
Row697	897471	4	8	6	4	3	4	10	6	1	4	unique
Row698	897471	4	8	8	5	4	5	10	4	1	4	unique
Row208	1218860	1	1	1	1	1	1	3	1	1	2	duplicate
Row253	1100524	6	10	10	2	8	10	7	3	3	4	duplicate
Row254	1116116	9	10	10	1	10	8	3	3	1	4	duplicate
Row258	1198641	3	1	1	1	2	1	3	1	1	2	duplicate
Row272	320675	3	3	5	2	3	10	7	1	1	4	duplicate
Row338	704097	1	1	1	1	1	1	2	1	1	2	duplicate
Row561	1321942	5	1	1	1	2	1	3	1	1	2	duplicate
Row684	466906	1	1	1	1	2	1	1	1	1	2	duplicate
Row42	1100524	6	10	10	2	8	10	7	3	3	4	chosen
Row62	1116116	9	10	10	1	10	8	3	3	1	4	chosen
Row168	1198641	3	1	1	1	2	1	3	1	1	2	chosen
Row207	1218860	1	1	1	1	1	1	3	1	1	2	chosen
Row267	320675	3	3	5	2	3	10	7	1	1	4	chosen
Row314	704097	1	1	1	1	1	1	2	1	1	2	chosen
Row560	1321942	5	1	1	1	2	1	3	1	1	2	chosen
Row683	466906	1	1	1	1	2	1	1	1	1	2	chosen

The 'Duplicate Row Filter' output is used as an input for the 'Statistics' node **(node 23 in Figure 1.1)** to show the number of missing values. It can observed in **Figure 1.8** that there are 16 missing

entries under the ‘Bare Nuclei’ column. This number is consistent with our “breast-cancer-wisconsin.names” file.



File

Numeric | Nominal | Top/bottom

Sample code number	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
No. missings: 0	No. missings: 0	No. missings: 0	No. missings: 0	No. missings: 0	No. missings: 0	No. missings: 16	No. missings: 0	No. missings: 0	No. missings: 0	No. missings: 0
Top 20:	Top 20:	Top 20:	Top 20:	Top 20:	Top 20:	Top 20:	Top 20:	Top 20:	Top 20:	Top 20:
1182404 : 6	1 : 145	1 : 384	1 : 353	1 : 407	2 : 386	1 : 402	2 : 166	1 : 443	1 : 579	2 : 458
1276091 : 5	5 : 130	10 : 67	2 : 59	3 : 58	3 : 72	10 : 132	3 : 165	10 : 61	2 : 35	4 : 241
1198641 : 3	3 : 108	3 : 52	10 : 58	2 : 58	4 : 48	2 : 30	1 : 152	3 : 44	3 : 33	
1017023 : 2	4 : 80	2 : 45	3 : 56	10 : 55	1 : 47	5 : 30	7 : 73	2 : 36	10 : 14	
1033078 : 2	10 : 69	4 : 40	4 : 44	4 : 33	6 : 41	3 : 28	4 : 40	8 : 24	4 : 12	
1070935 : 2	2 : 50	5 : 30	5 : 34	8 : 25	5 : 39	8 : 21	5 : 34	6 : 22	7 : 9	
1100524 : 2	8 : 46	8 : 29	6 : 30	5 : 23	10 : 31	4 : 19	8 : 28	5 : 19	8 : 8	
1105524 : 2	6 : 34	6 : 27	7 : 30	6 : 22	8 : 21	? : 16	10 : 20	4 : 18	5 : 6	
1115293 : 2	7 : 23	7 : 19	8 : 28	7 : 13	7 : 12	9 : 9	9 : 11	7 : 16	6 : 3	
1116116 : 2	9 : 14	9 : 6	9 : 7	9 : 5	9 : 2	7 : 8	6 : 10	9 : 16		
1116192 : 2						6 : 4				
1143978 : 2										
1158247 : 2										
1168736 : 2										
1171710 : 2										
1173347 : 2										
1174057 : 2										
1212422 : 2										

Figure 1.8: Statistics output table showing number of missing values

The missing value can be observed below highlighted in the green circle in **Figure 1.9.**

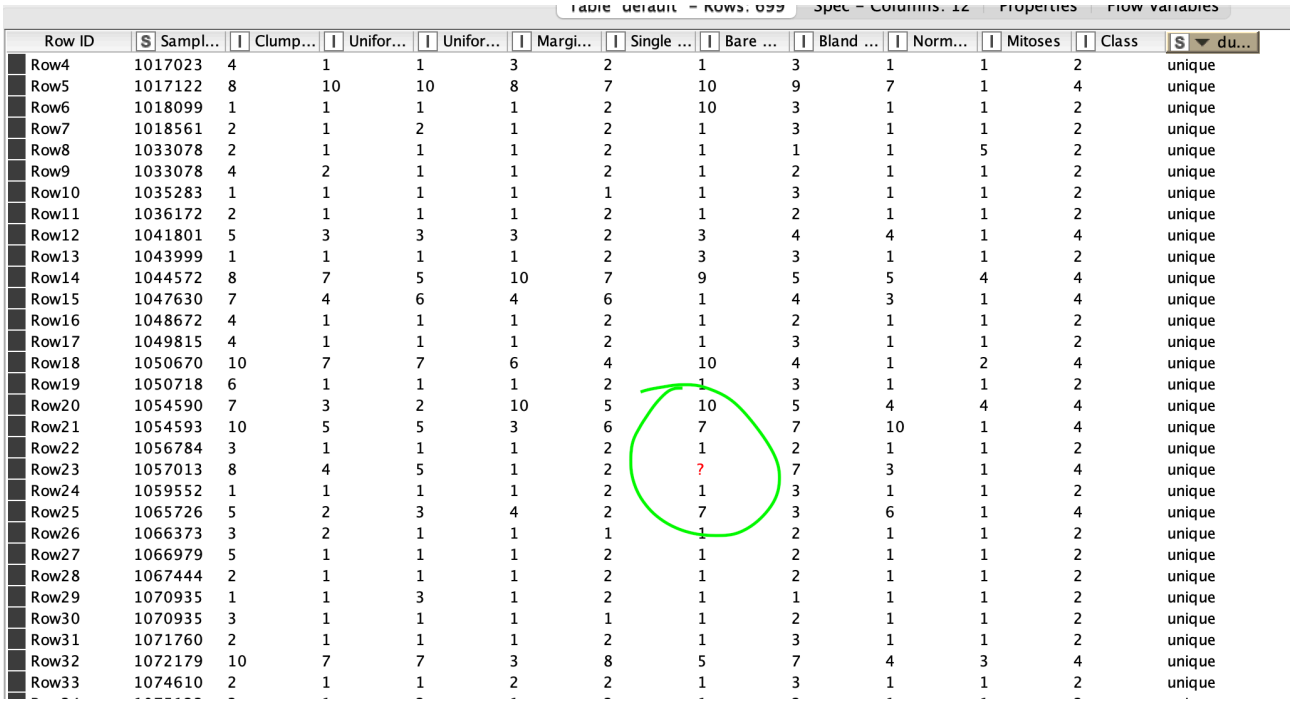


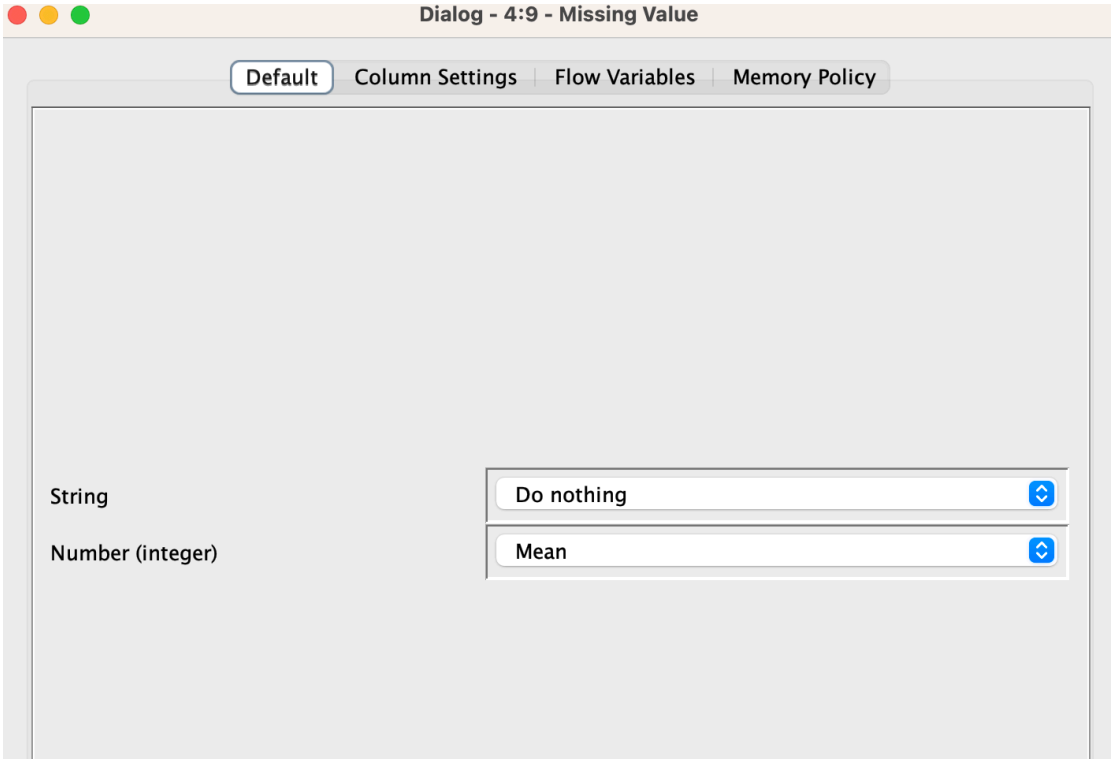
Table "default" – Rows: 699 | Spec – Columns: 12 | Properties | Flow Variables

Row ID	Sampl...	Clump...	Unifor...	Unifor...	Margi...	Single ...	Bare ...	Bland ...	Norm...	Mitoses	Class	du...
Row4	1017023	4	1	1	3	2	1	3	1	1	2	unique
Row5	1017122	8	10	10	8	7	10	9	7	1	4	unique
Row6	1018099	1	1	1	1	2	10	3	1	1	2	unique
Row7	1018561	2	1	2	1	2	1	3	1	1	2	unique
Row8	1033078	2	1	1	1	2	1	1	1	5	2	unique
Row9	1033078	4	2	1	1	2	1	2	1	1	2	unique
Row10	1035283	1	1	1	1	1	1	3	1	1	2	unique
Row11	1036172	2	1	1	1	2	1	2	1	1	2	unique
Row12	1041801	5	3	3	3	2	3	4	4	1	4	unique
Row13	1043999	1	1	1	1	2	3	3	1	1	2	unique
Row14	1044572	8	7	5	10	7	9	5	5	4	4	unique
Row15	1047630	7	4	6	4	6	1	4	3	1	4	unique
Row16	1048672	4	1	1	1	2	1	2	1	1	2	unique
Row17	1049815	4	1	1	1	2	1	3	1	1	2	unique
Row18	1050670	10	7	7	6	4	10	4	1	2	4	unique
Row19	1050718	6	1	1	1	2	1	3	1	1	2	unique
Row20	1054590	7	3	2	10	5	10	5	4	4	4	unique
Row21	1054593	10	5	5	3	6	7	7	10	1	4	unique
Row22	1056784	3	1	1	1	2	1	2	1	1	2	unique
Row23	1057013	8	4	5	1	2	?	7	3	1	4	unique
Row24	1059552	1	1	1	1	2	1	3	1	1	2	unique
Row25	1065726	5	2	3	4	2	7	3	6	1	4	unique
Row26	1066373	3	2	1	1	1	1	2	1	1	2	unique
Row27	1066979	5	1	1	1	2	1	2	1	1	2	unique
Row28	1067444	2	1	1	1	2	1	2	1	1	2	unique
Row29	1070935	1	1	3	1	2	1	1	1	1	2	unique
Row30	1070935	3	1	1	1	1	1	2	1	1	2	unique
Row31	1071760	2	1	1	1	2	1	3	1	1	2	unique
Row32	1072179	10	7	7	3	8	5	7	4	3	4	unique
Row33	1074610	2	1	1	2	2	1	3	1	1	2	unique

Figure 1.9: Missing values identified

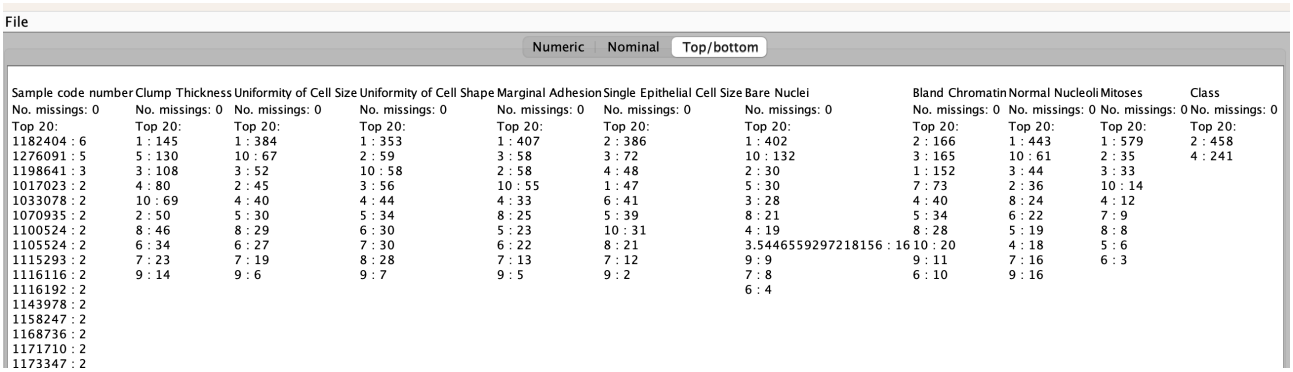
5. The missing values are populated using the mean function in Missing value node **(node 9 in Figure 1.1**). Since all data in the column is an integer hence ‘mean’ is used. Configuration and output is shown in **Figure 1.10**

Figure 1.10:



Configuration for missing value node

On the missing values node is execute, it can be observed that there are no more missing values in the ‘bare nuclei’ column - **Figure 1.11**



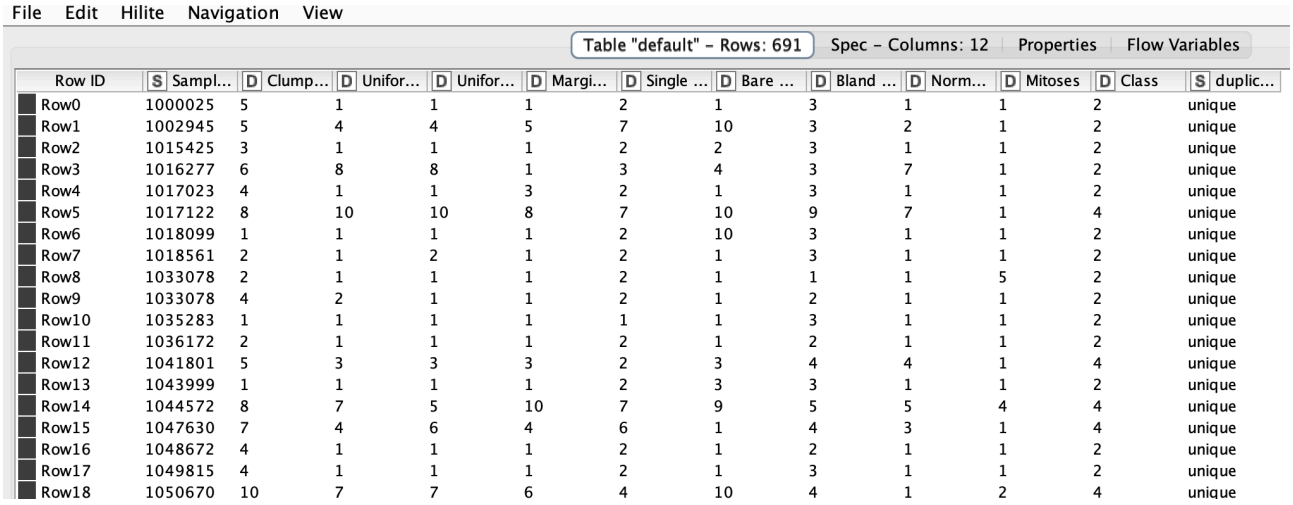
File

Numeric | Nominal | Top/bottom

Sample code number	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
No. missings: 0	No. missings: 0	No. missings: 0	No. missings: 0	No. missings: 0	No. missings: 0	No. missings: 0	No. missings: 0	No. missings: 0	No. missings: 0	No. missings: 0
Top 20:	Top 20:	Top 20:	Top 20:	Top 20:	Top 20:	Top 20:	Top 20:	Top 20:	Top 20:	Top 20:
1182404 : 6	1 : 145	1 : 384	1 : 353	1 : 407	2 : 386	1 : 402	2 : 166	1 : 443	1 : 579	2 : 458
1276091 : 5	5 : 130	10 : 67	2 : 59	3 : 58	3 : 72	10 : 132	3 : 165	10 : 61	2 : 35	4 : 241
1198641 : 3	3 : 108	3 : 52	10 : 58	2 : 58	4 : 48	2 : 30	1 : 152	3 : 44	3 : 33	
1017023 : 2	4 : 80	2 : 45	3 : 56	10 : 55	1 : 47	5 : 30	7 : 73	2 : 36	10 : 14	
1033078 : 2	10 : 69	4 : 40	4 : 44	4 : 33	6 : 41	3 : 28	4 : 40	8 : 24	4 : 12	
1070935 : 2	2 : 50	5 : 30	5 : 34	8 : 25	5 : 39	8 : 21	5 : 34	6 : 22	7 : 9	
1100524 : 2	8 : 46	8 : 29	6 : 30	5 : 23	10 : 31	4 : 19	8 : 28	5 : 19	8 : 8	
1105524 : 2	6 : 34	6 : 27	7 : 30	6 : 22	8 : 21	3.5446559297218156 : 16	10 : 20	4 : 18	5 : 6	
1115293 : 2	7 : 23	7 : 19	8 : 28	7 : 13	7 : 12	9 : 9	9 : 11	7 : 16	6 : 3	
1116116 : 2	9 : 14	9 : 6	9 : 7	9 : 5	9 : 2	7 : 8	6 : 10	9 : 16		
1116192 : 2						6 : 4				
1143978 : 2										
1158247 : 2										
1168736 : 2										
1171710 : 2										
1173347 : 2										

Figure 1.11 - Missing values are removed

6. As there are duplicate values in the data, hence 'Duplicate Row filter' node **(Node 25 in Figure 1.1)** is used to remove the duplicate values. A duplicate row filter node was also used earlier only to identify the presence of duplicate entries, hence this column 'duplicate-type-classifer' was not used as a reference in the configuration settings. All columns were referenced to identify and remove the duplicate values. Once the node is executed, it can be observed in **Figure 1.12** that 8 rows have been removed from the dataset. The total count is now reduced to to 691 rows.



File Edit Hilite Navigation View

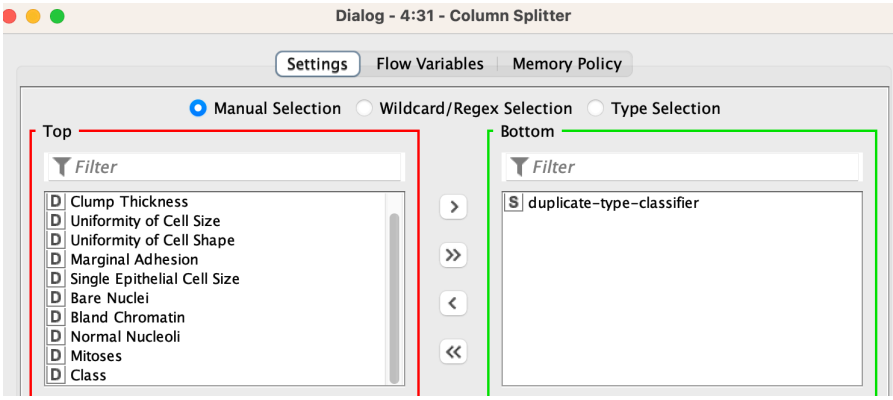
Table "default" - Rows: 691 | Spec - Columns: 12 | Properties | Flow Variables

Row ID	S Sampl...	D Clump...	D Unifor...	D Unifor...	D Margi...	D Single ...	D Bare ...	D Bland ...	D Norm...	D Mitoses	D Class	S duplic...
Row0	1000025	5	1	1	1	2	1	3	1	1	2	unique
Row1	1002945	5	4	4	5	7	10	3	2	1	2	unique
Row2	1015425	3	1	1	1	2	2	3	1	1	2	unique
Row3	1016277	6	8	8	1	3	4	3	7	1	2	unique
Row4	1017023	4	1	1	3	2	1	3	1	1	2	unique
Row5	1017122	8	10	10	8	7	10	9	7	1	4	unique
Row6	1018099	1	1	1	1	2	10	3	1	1	2	unique
Row7	1018561	2	1	2	1	2	1	3	1	1	2	unique
Row8	1033078	2	1	1	1	2	1	1	1	5	2	unique
Row9	1033078	4	2	1	1	2	1	2	1	1	2	unique
Row10	1035283	1	1	1	1	1	1	3	1	1	2	unique
Row11	1036172	2	1	1	1	2	1	2	1	1	2	unique
Row12	1041801	5	3	3	3	2	3	4	4	1	4	unique
Row13	1043999	1	1	1	1	2	3	3	1	1	2	unique
Row14	1044572	8	7	5	10	7	9	5	5	4	4	unique
Row15	1047630	7	4	6	4	6	1	4	3	1	4	unique
Row16	1048672	4	1	1	1	2	1	2	1	1	2	unique
Row17	1049815	4	1	1	1	2	1	3	1	1	2	unique
Row18	1050670	10	7	7	6	4	10	4	1	2	4	unique

Figure 1.12 - Revised data set with 691 rows

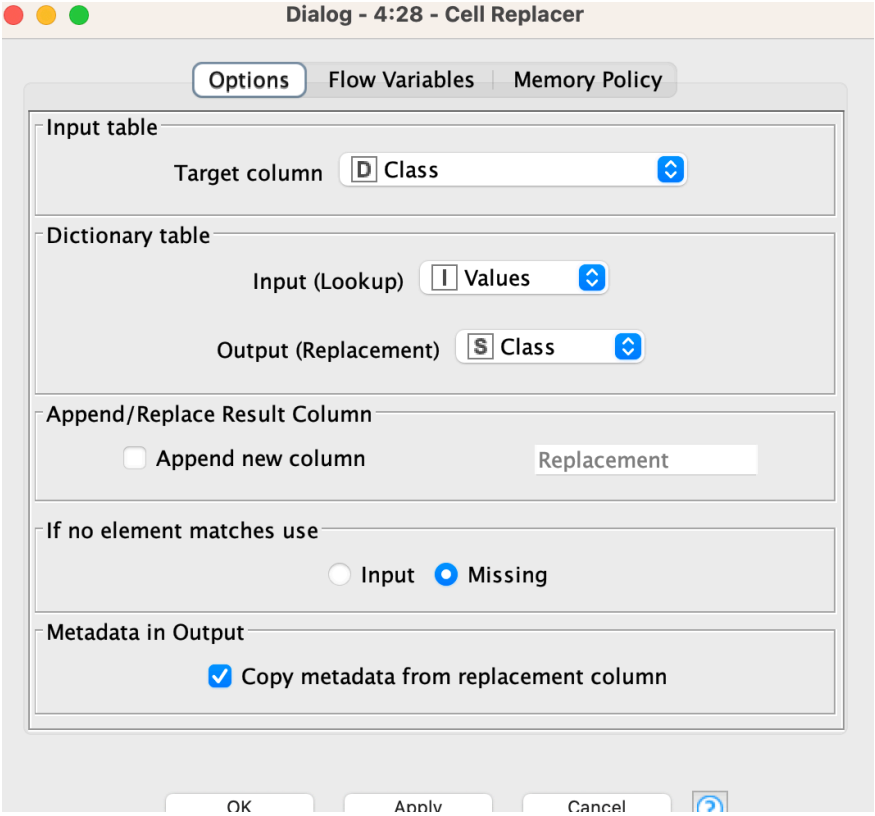
7. Sorter **(Node 32 in Figure 1.1**) and Column Splitter nodes **(Node 31 in Figure 1.1**) are used to sort the rows and then to remove the 'duplicate-type-classifier' column. The output table is exported to a cell splitter node **(Node 28 in Figure 1.1).** Configuration for cell replacer node is shown **Figure 1.13**

Figure 1.13: Column Splitter node configuration



8. Cell Replacer node **(Node 28 in Figure 1.1)** is used to replace values in the data - **Figure 1.16**. Cell replacer takes two inputs. First is our dataset and second is the value replacement table. From the final output it can be observed that class value 2 and 4 is replaced with benign and malignant values respectively

Figure 1.16: Configuration from Cell replacer node (Node 28 in Figure 1.1

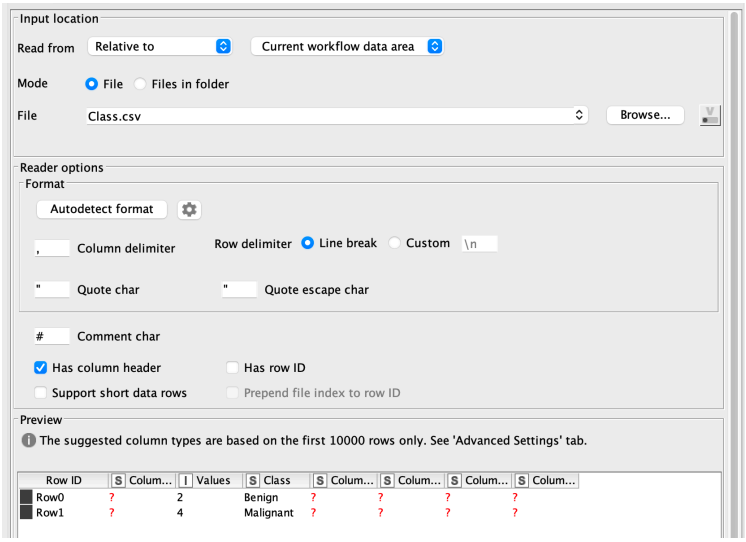


9. The ‘class’ attribute is numeric in nature and it had values 2 and 4. This is changed to ‘Benign’ and Malignant as shown in the configuration and output table in **Figure 1.15**. A new CSV file was created ‘Filename: Class’ which was read by the CSV reader node **(Node 30 in Figure 1.1)**. Here the data was filled for class 2 and 4 as benign and malignant respectively. This is shown below along with the figuration of this node in **Figure 1.15**

Figure 1.15: Configuration for Node 30 and File name: Class

Class

	Values	Class				
	2	Benign				
	4	Malignant				



10. CSV file using a CSV Writer node **(Node 29 in Figure 1.1)**. The configuration of both nodes is shown in **Figure 1.13** and **Figure 1.14**

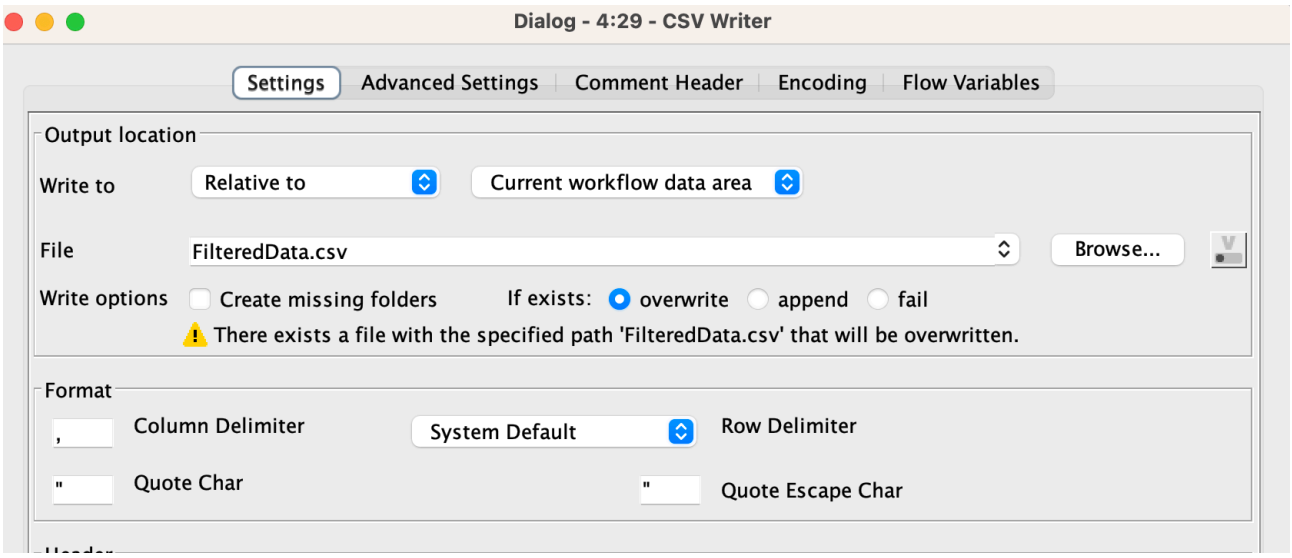


Figure 1.14: CSV Writer node configuration

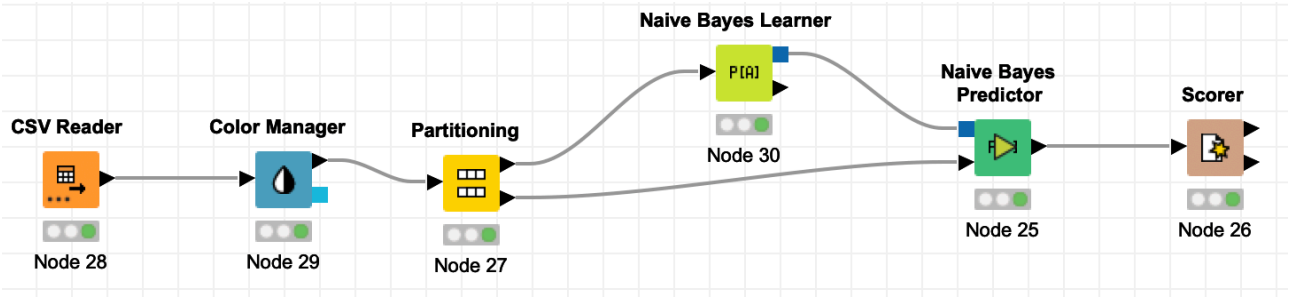
Task 2

The two classification models used are Naive Bayes and SVM.

Naive Bayes

The work flow is shown in **Figure 2.1**

Figure 2.1: Knime workflow flow for Naive Bayes model



Advantages of using Naive Bayes algorithm:

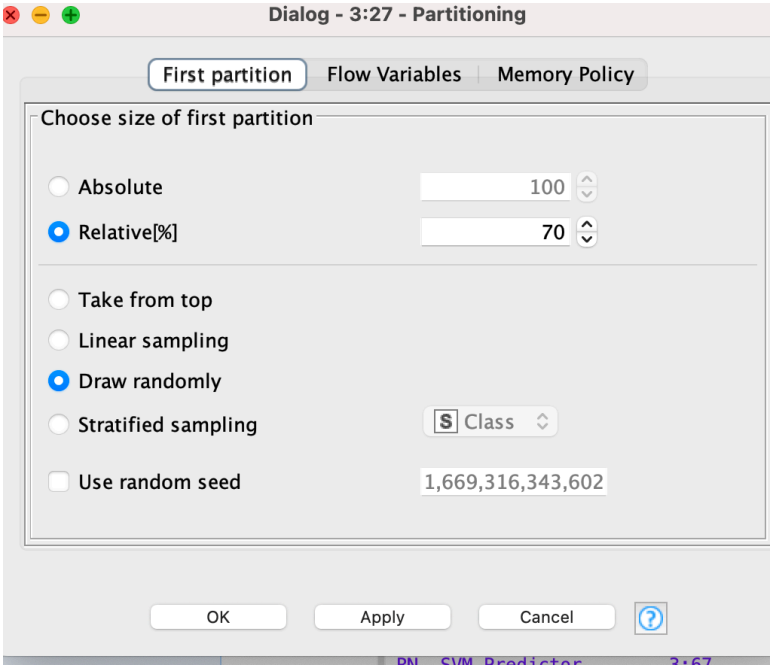
- It is used for high risk for certain diseases and conditions, such as cancer
- It requires less training data and easy to implement as it handles both numerical and categorical data
- It is fast and not sensitive to irrelevant factors
- It supports missing values

Disadvantages of using Naive Bayes algorithm:

- It assumes that all attributes are independent of each other. We are not sure if this case is true for our dataset. More medical expertise is required to understand what attributes contribute to a 'malignant' outcome.

Referring to **Figure 2.1** the CSV file produced in Task 1 i.e. "FilteredData.csv" is read using CSV Reader **(node 28 in Figure 2.1**). Color Manager node **(Node 29 in Figure 2.1)** is used to segregate to the two binary attributes i.e. 'benign' and 'malignant'. Then Naive Bayes Learner node **(node 30 in Figure 2.1)** is applied with 70% training data and data drawn randomly - refer **Figure 2.2**. This data is split using the Partitioning node **(node 27 in Figure 2.1)**.

Figure 2.2 with configuration of Portioning node that takes 70% training data



Once the model is learned, the training data and test data are inputted to the Naive Bayes Predictor node **(node 25 in Figure 2.1)**. Output from Naive Bayes Learner and Partitioning node is provided to Naive Bayes Predictor node. The output from the model is finally captured by the scorer node **(node 26 in Figure 2.1)**.

Another execution of Naive Bayes model **(Figure 2.3)** is applied using X-partitioning node with 'leave one out' in configuration setting. This helps in improving the accuracy of the model. Since the algorithm traverses each node hence and X-Aggregator node is used to capture the data.

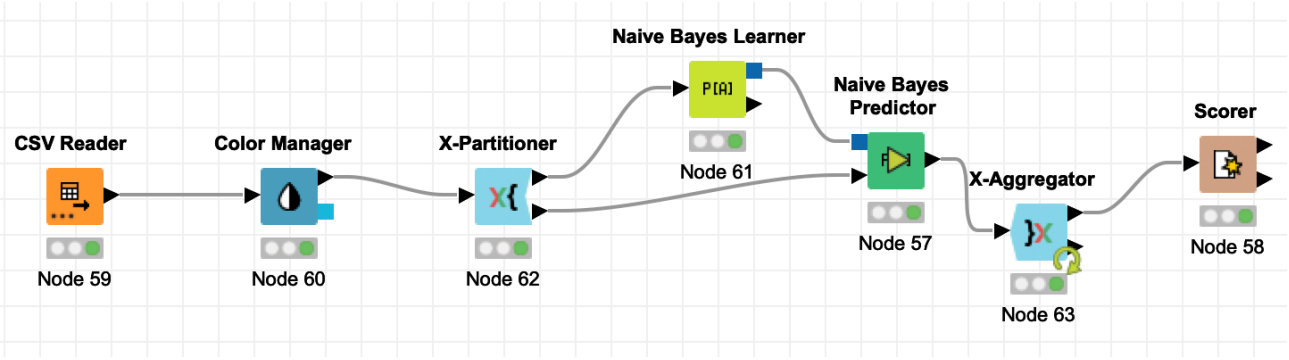


Figure 2.3: Naive Bayes using X-partitioner and X-Aggregator nodes

Super Vector Machines - SVM

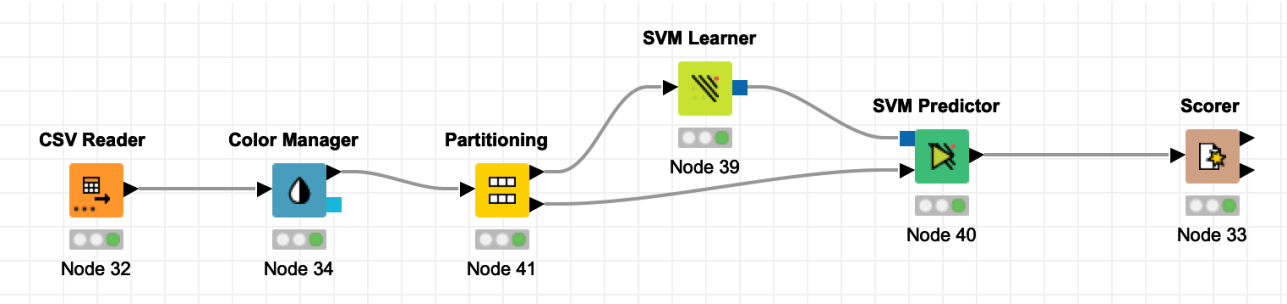


Figure 2.4 using SVM model

Advantages of using SVM model:

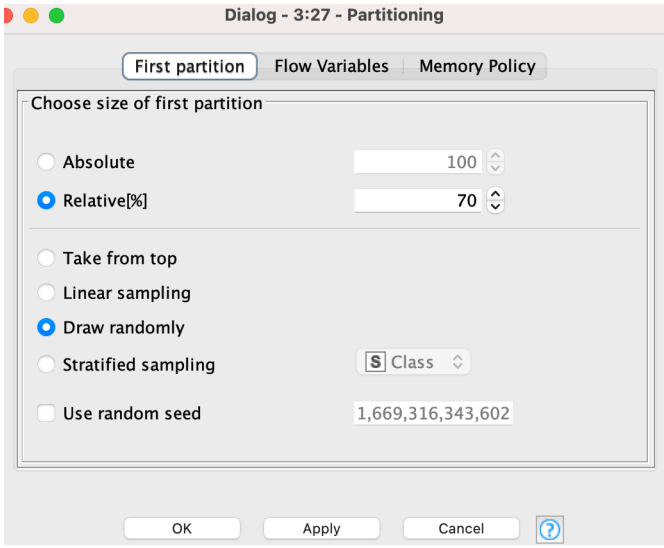
- This effective for data that has a large number of features however, the downside is that takes a lot more time to learn especially for large datasets. The number of attributes are large. The numbers of features were not very large here however, the the accuracy of the model was good

Disadvantages of using SVM model:

- It takes more time to train
- It does not support missing values - hence these have be addressed in data pre-processing

Referring to **Figure 2.4** the CSV file produced in Task 1 i.e. "FilteredData.csv" is read using CSV Reader **(node 32 in Figure 2.2**). Color Manager node **(Node 34 in Figure 2.2)** is used to segregate to the two binary attributes i.e. 'benign' and 'malignant'. Then SVM Learner node **(node 39 in Figure 2.2)** is applied with 70% training data and data drawn randomly - refer **Figure 2.5**.

Figure 2.5 with configuration for Partitioning node



This data is split using the Partitioning node **(node 41 in Figure 2.4)**. Once the model is learned, the training data and test data are inputted to the SVM Predictor node **(node 40 in Figure 2.4)**. Since 'Class' is the attribute that we use for decision hence it is selected in the Class column in the Node 39. decision attribute is the Output from SVM Learner and Partitioning node is provided to

SVM Predictor node **(Node 40 in Figure 2.4)**. The output from the model is finally captured by the scorer node **(node 33 in Figure 2.4)**.

Task 3

The algorithms were run a few times with different partitioning and traversing techniques. Form the confusion matrix, two parameters are critical:

- Accuracy: how accurate the model is on testing data
- Precision or Error rate: What proportion of actual malignant and benign records was predicted correctly. This is very important as the cost for misdiagnosing a malignant record is high i.e. the model predicting that the patient does not have malignant cancer when they actually have.

Naive Bayes

Trial 1 with holdout method - 70% training data

- Accuracy is 97.6%

Row ID	TrueP...	FalseP...	TrueN...	False...	Recall	Precisi...	Sensiti...	Specifi...	F-me...	Accur...	
Benign	136	2	67	3	0.978	0.986	0.978	0.971	0.982	?	?
Malignant	67	3	136	2	0.971	0.957	0.971	0.978	0.964	?	?
Overall	?	?	?	?	?	?	?	?	?	0.976	(

- Precision for malignant attribute is 95.7%

Trial 2 with holdout method - 90% training data

- Accuracy is 94.3%

Table "default" - Rows: 3 | Spec - Columns: 11 | Properties | Flow Variables

Row ID	TrueP...	FalseP...	TrueN...	False...	Recall	Precisi...	Sensiti...	Specifi...	F-me...	Accur...
Benign	45	2	21	2	0.957	0.957	0.957	0.913	0.957	?
Malignant	21	2	45	2	0.913	0.913	0.913	0.957	0.913	?
Overall	?	?	?	?	?	?	?	?	?	0.943

- Precision for malignant attribute is 91.3%

Trial 3 with holdout method - 50% training data

- Accuracy is 96.8%

Row ID	TrueP...	FalseP...	TrueN...	False...	Recall	Precisi...	Sensiti...	Specifi...	F-me...	Accur...
Benign	211	4	124	7	0.968	0.981	0.968	0.969	0.975	?
Malignant	124	7	211	4	0.969	0.947	0.969	0.968	0.958	?
Overall	?	?	?	?	?	?	?	?	?	0.968

- Precision for malignant attribute is 94.7%

Trial 4 using leave one out method Figure 2.6

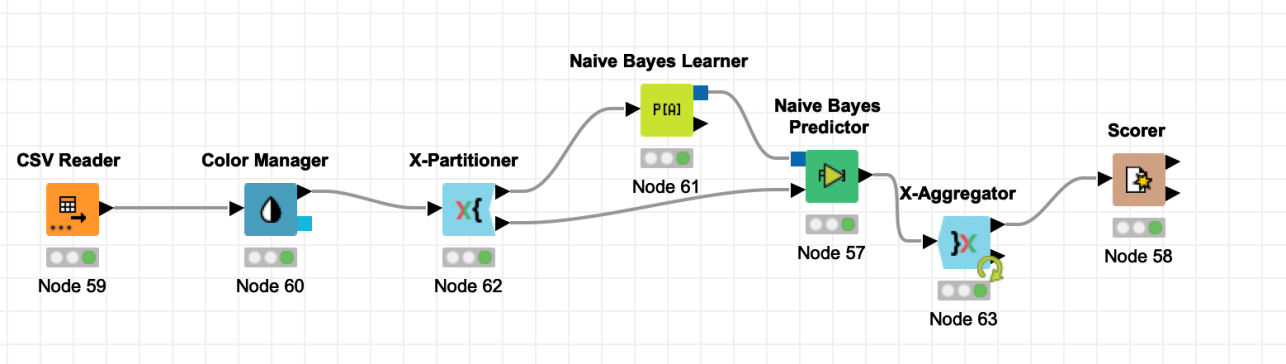


Figure 2.6: Naive Bayes approach with Leave one out method

- This approach took a long time to execute
- Accuracy is 96.8%
- Precision for malignant attribute is 93.9%

SVM Model

Trial 1 with hold-out method - 70% training data

- Accuracy is 94.7%

Row ID	TrueP...	FalseP...	TrueN...	False...	Recall	Precisi...	Sensiti...	Specifi...	F-me...	Accur...
Benign	127	6	70	5	0.962	0.955	0.962	0.921	0.958	?
Malignant	70	5	127	6	0.921	0.933	0.921	0.962	0.927	?
Overall	?	?	?	?	?	?	?	?	?	0.947

- Precision for malignant attribute is 93.3%

Trial 2 with hold-out method - 90% training data

- Accuracy is 97.1%

Row ID	TrueP...	FalseP...	TrueN...	False...	Recall	Precisi...	Sensiti...	Specifi...	F-me...	Accur...
Benign	44	1	24	1	0.978	0.978	0.978	0.96	0.978	?
Malignant	24	1	44	1	0.96	0.96	0.96	0.978	0.96	?
Overall	?	?	?	?	?	?	?	?	?	0.971

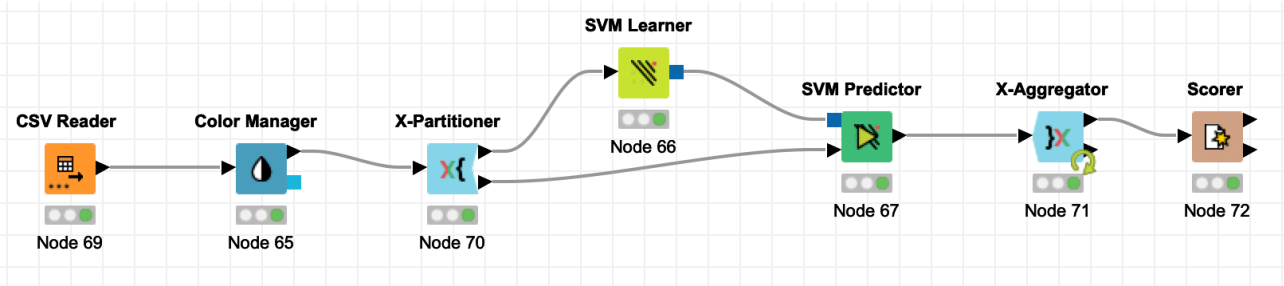
- Precision for malignant attribute is 96%

Trial 3 with hold-out method - 50% training data

- Accuracy is 97.7%

Row ID	TrueP...	FalseP...	TrueN...	False...	Recall	Precisi...	Sensiti...	Specifi...	F-me...	Accur...
Benign	213	4	125	4	0.982	0.982	0.982	0.969	0.982	?
Malignant	125	4	213	4	0.969	0.969	0.969	0.982	0.969	?
Overall	?	?	?	?	?	?	?	?	?	0.977

- Precision for malignant attribute is 96.9%



Trial 4 with Leave one out method

- This took too long to execute
- Accuracy is 96.5%
- Precision for malignant attribute is 94.2%

Row ID	TrueP...	FalseP...	TrueN...	False...	Recall	Precisi...	Sensiti...	Specifi...	F-me...	Accur...
Benign	439	10	228	14	0.969	0.978	0.969	0.958	0.973	?
Malignant	228	14	439	10	0.958	0.942	0.958	0.969	0.95	?
Overall	?	?	?	?	?	?	?	?	?	0.965

Conclusion

Considering our dataset which deals with cancers it is very important that all malignant cases should be identified. That is, even false negatives in benign categories are acceptable. Hence the Precision score for malignant patients is the most important metric. Hence the most effective algorithm is SVM with training data to be at 50%