May 13, 2015

muhammad zeeshan karamat

Machine Learning Assignment

2011-ee-134

**Two class problem**

**Abstract**: Dataset contains cases from study conducted on the survival of patients who had undergone surgery for breast cancer.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Data Set Characteristics:** | Multivariate | **Number of Instances:** | 306 |  |  |
| **Attribute Characteristics:** | Integer | **Number of Attributes:** | 3 |  |  |
| **Associated Tasks:** | Classification | **Missing Values?** | No |  |  |

**Source:**

Donor:   
  
Tjen-Sien Lim (limt **'@'** stat.wisc.edu)

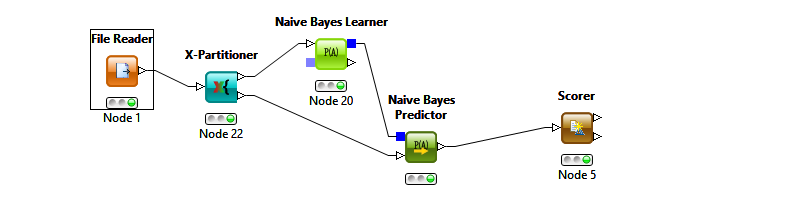
**Data Set Information:**

The dataset contains cases from a study that was conducted between 1958 and 1970 at the University of Chicago's Billings Hospital on the survival of patients who had undergone surgery for breast cancer.

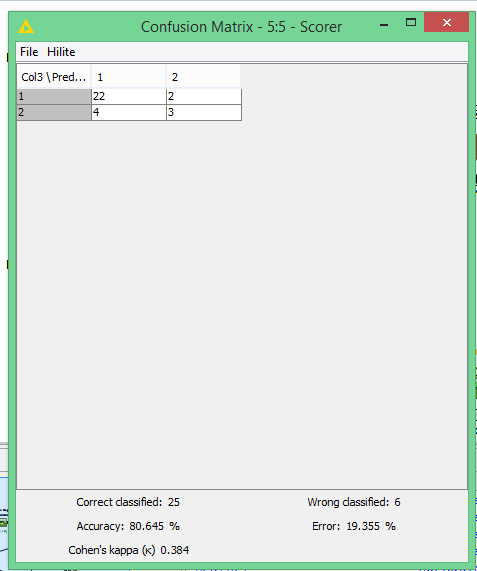
**Attribute Information:**

1. Age of patient at time of operation (numerical)   
2. Patient's year of operation (year - 1900, numerical)   
3. Number of positive axillary nodes detected (numerical)   
4. Survival status (class attribute)   
-- 1 = the patient survived 5 years or longer   
-- 2 = the patient died within 5 year

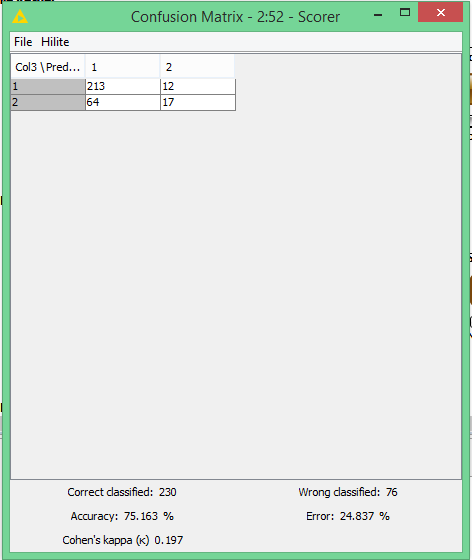
**Using Naïve Bayes Approach:**



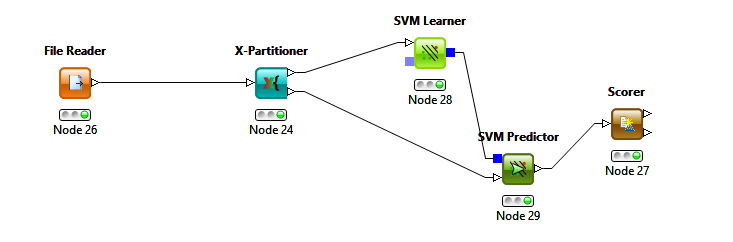
And accuracy of this approach is



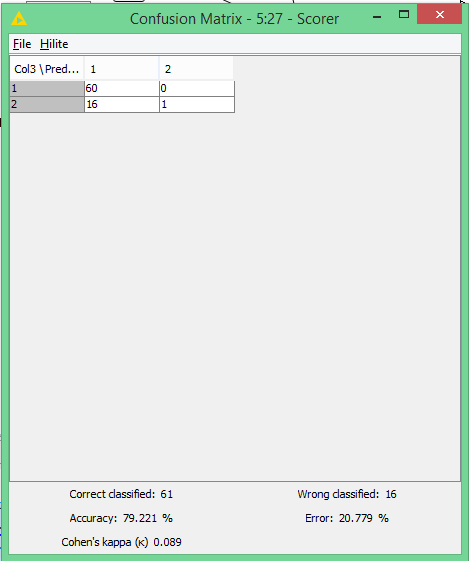
Applying cross-validation on this,



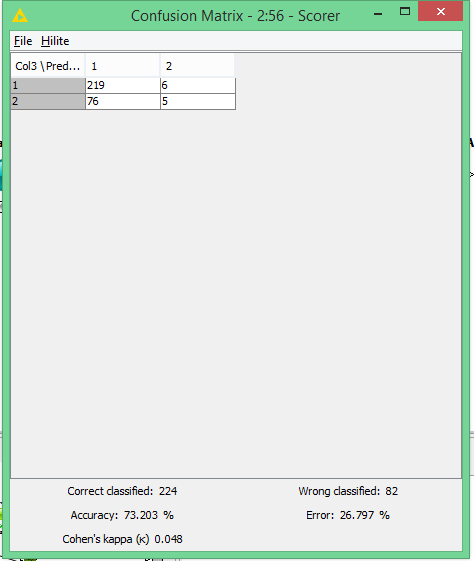
**Using SVM Approach**:



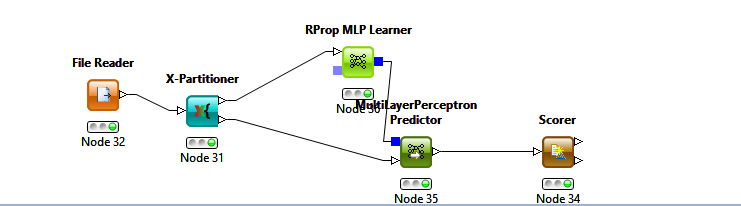
And this approach results in accuracy of



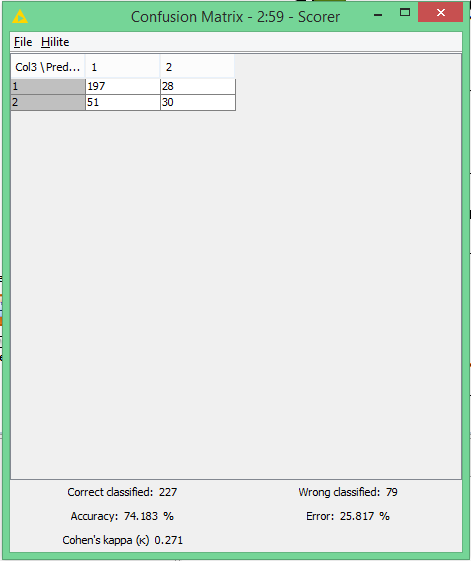
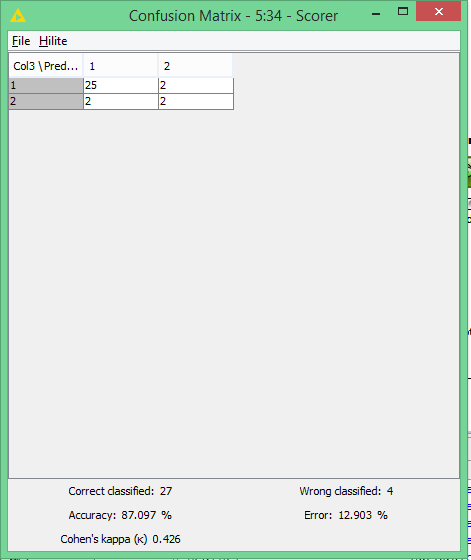
Applying cross-validation on this approach,



**Using Multi-layer perceptron approach:**



Applying cross-validation on this approach Without cross-validation accuracy



**Multi-class Problem:**

**Abstract**: Famous database; from Fisher, 1936

|  |  |  |  |
| --- | --- | --- | --- |
| **Data Set Characteristics:** | Multivariate | **Number of Instances:** | 150 |
| **Attribute Characteristics:** | Real | **Number of Attributes:** | 4 |
| **Associated Tasks:** | Classification | **Missing Values?** | No |

**Source:**

Creator:   
  
R.A. Fisher   
  
Donor:   
  
Michael Marshall (MARSHALL%PLU **'@'** io.arc.nasa.gov)

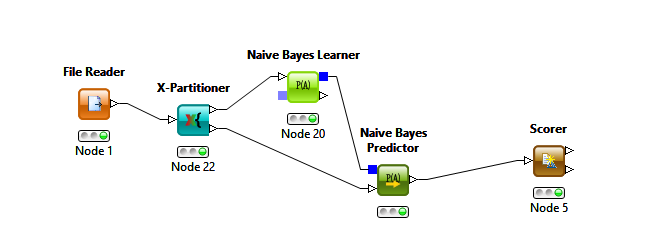
**Data Set Information:**

This is perhaps the best known database to be found in the pattern recognition literature. Fisher's paper is a classic in the field and is referenced frequently to this day. (See Duda & Hart, for example.) The data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant. One class is linearly separable from the other 2; the latter are NOT linearly separable from each other.   
  
Predicted attribute: class of iris plant.   
  
This is an exceedingly simple domain.   
  
This data differs from the data presented in Fishers article (identified by Steve Chadwick, spchadwick **'@'** espeedaz.net ). The 35th sample should be: 4.9,3.1,1.5,0.2,"Iris-setosa" where the error is in the fourth feature. The 38th sample: 4.9,3.6,1.4,0.1,"Iris-setosa" where the errors are in the second and third features.

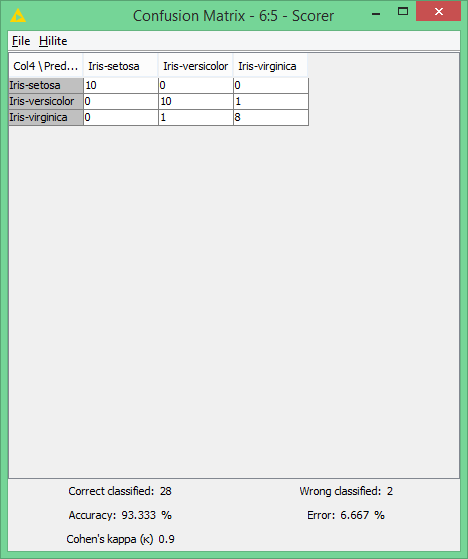
**Attribute Information:**

1. sepal length in cm   
2. sepal width in cm   
3. petal length in cm   
4. petal width in cm   
5. class:   
-- Iris Setosa   
-- Iris Versicolour   
-- Iris Virginica

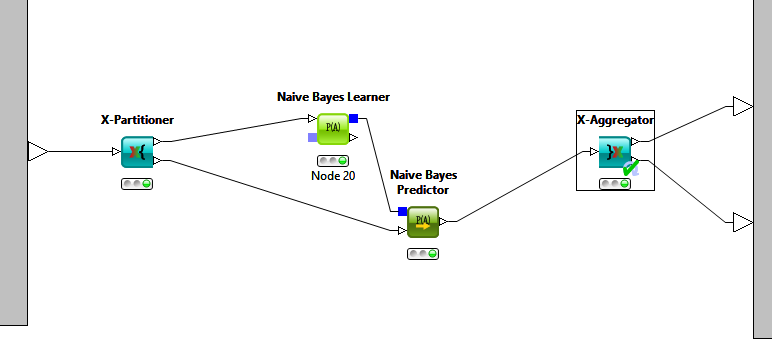
**Using Naïve Bayes Approach:**

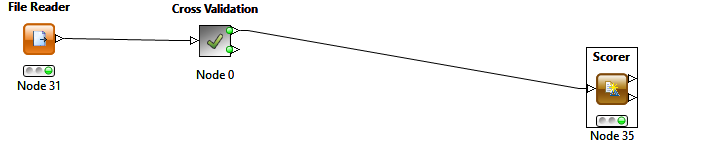


And this approach results in

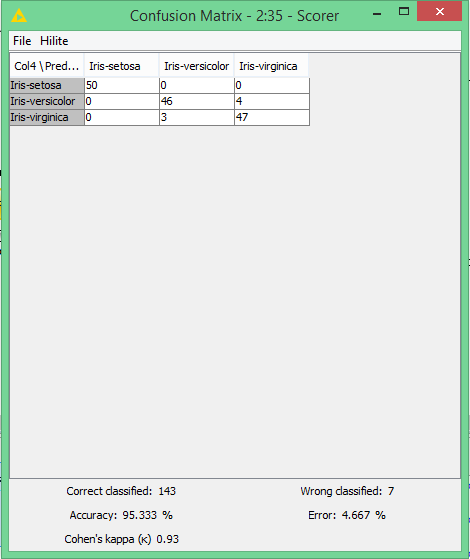


And applying cross-validation on this results in

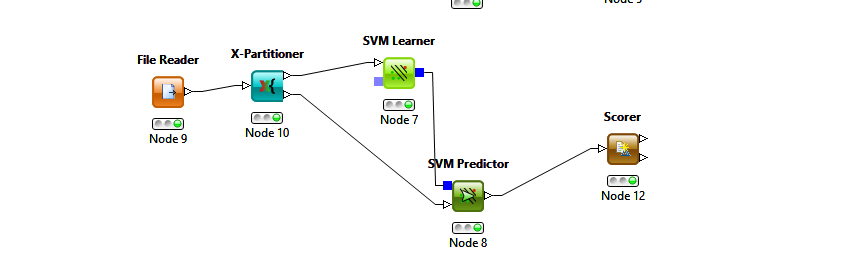




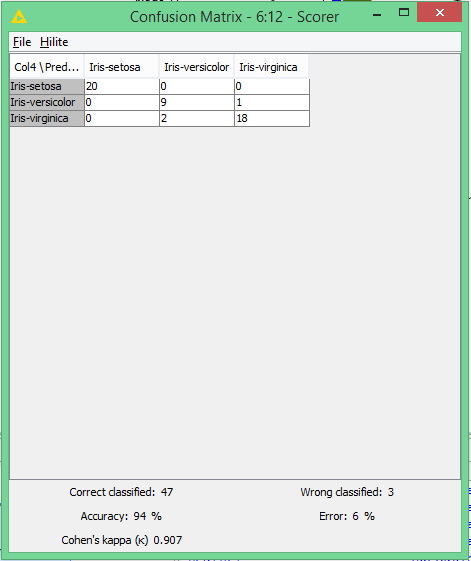
And this results in,



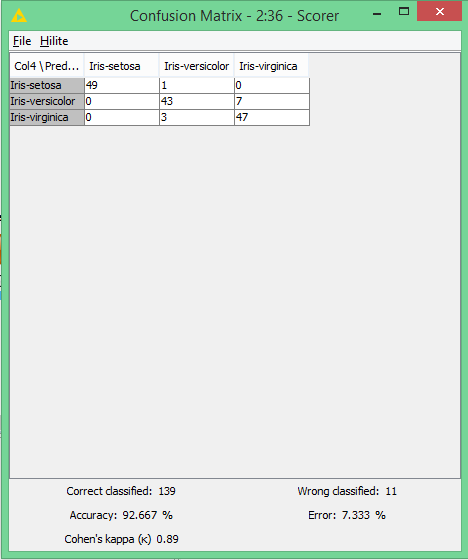
**Using SVM approach:**



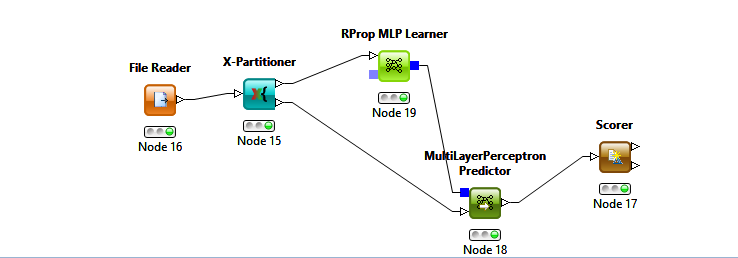
And this approach results in,



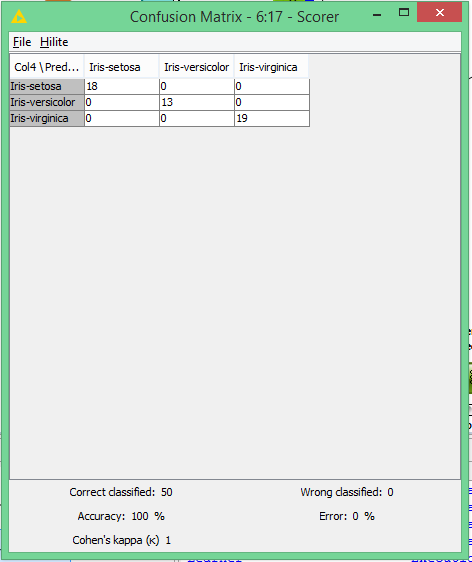
And applying cross validation on this approach results in ,



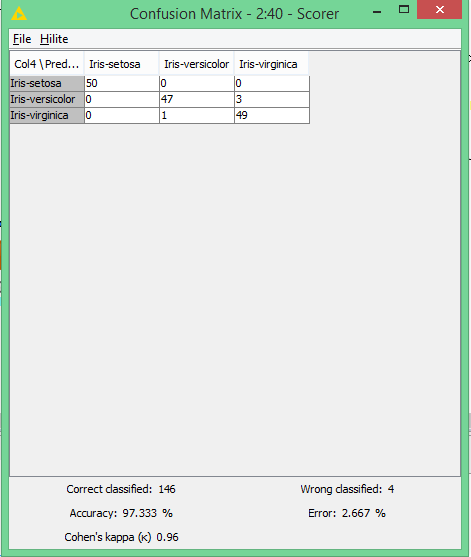
**Using MLP approach:**



And this approach results in



And applying cross-validation on this approach results in,



**Problem having more than 100+ columns:**

**Abstract**: 1593 handwritten digits from around 80 persons were scanned, stretched in a rectangular box 16x16 in a gray scale of 256 values.

|  |  |  |  |
| --- | --- | --- | --- |
| **Data Set Characteristics:** | Multivariate | **Number of Instances:** | 1593 |
| **Attribute Characteristics:** | Integer | **Number of Attributes:** | 256 |
| **Associated Tasks:** | Classification | **Missing Values?** | N/A |

**Source:**

The dataset was created by Tactile Srl, Brescia, Italy ([http://www.tattile.it](http://www.tattile.it/)) and donated in 1994 to Semeion Research Center of Sciences of Communication, Rome, Italy ([http://www.semeion.it](http://www.semeion.it/)), for machine learning research.   
  
For any questions, e-mail Massimo Buscema (m.buscema **'@'** semeion.it) or Stefano Terzi (s.terzi **'@'** semeion.it)

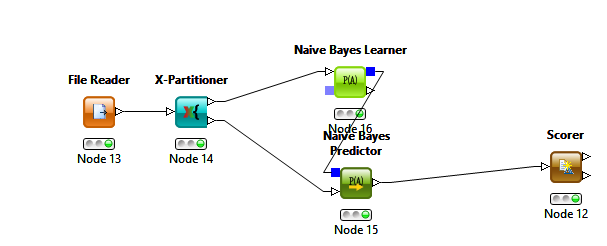
**Data Set Information:**

1593 handwritten digits from around 80 persons were scanned, stretched in a rectangular box 16x16 in a gray scale of 256 values.Then each pixel of each image was scaled into a bolean (1/0) value using a fixed threshold.   
  
Each person wrote on a paper all the digits from 0 to 9, twice. The commitment was to write the digit the first time in the normal way (trying to write each digit accurately) and the second time in a fast way (with no accuracy).   
  
The best validation protocol for this dataset seems to be a 5x2CV, 50% Tune (Train +Test) and completly blind 50% Validation

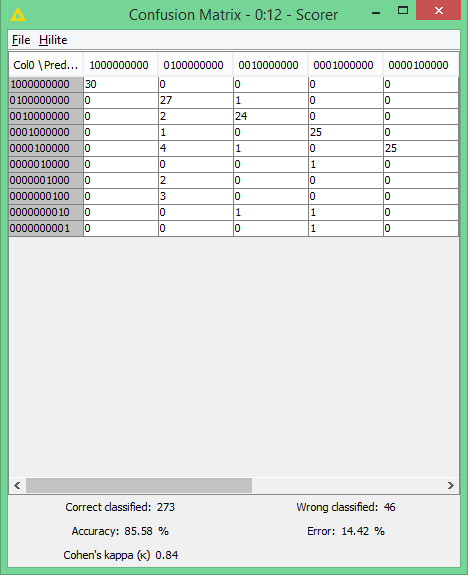
**Attribute Information:**

This dataset consists of 1593 records (rows) and 256 attributes (columns).   
  
Each record represents a handwritten digit, orginally scanned with a resolution of 256 grays scale (28).   
  
Each pixel of the each original scanned image was first stretched, and after scaled between 0 and 1 (setting to 0 every pixel whose value was under tha value 127 of the grey scale (127 included) and setting to 1 each pixel whose orinal value in the grey scale was over 127).   
  
Finally, each binary image was scaled again into a 16x16 square box (the final 256 binary attributes).

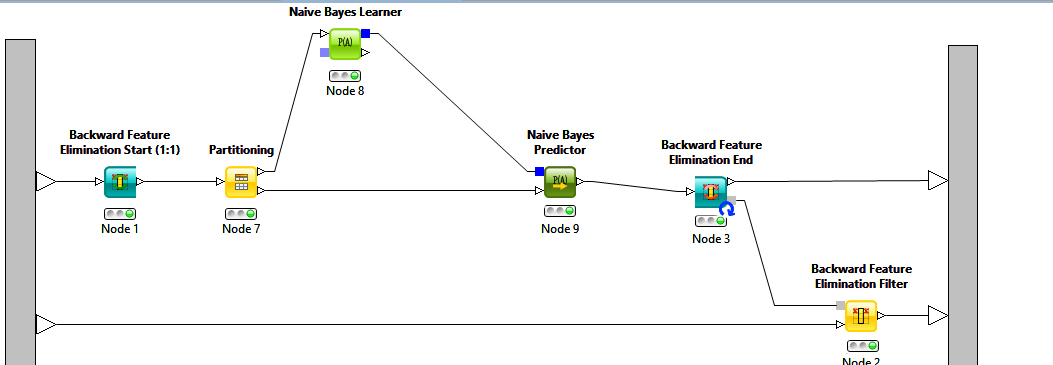
**Using Naïve Bayes Approach:**



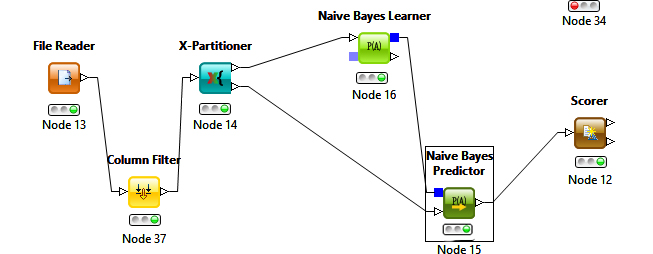
And this approach results in



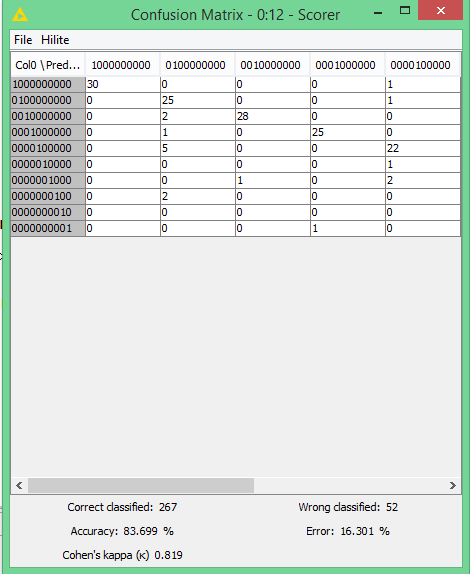
And now applying backward feature elimination method on this dataset for selecting useful attributes,



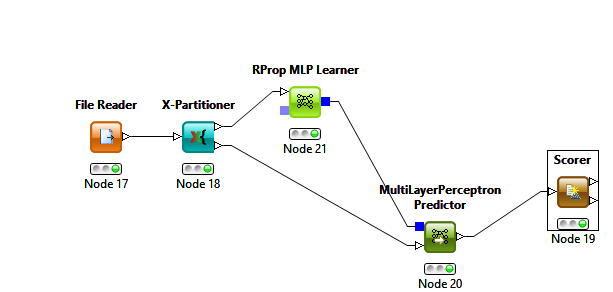
And now applying column filter on input for selecting attributes that give good accuracy



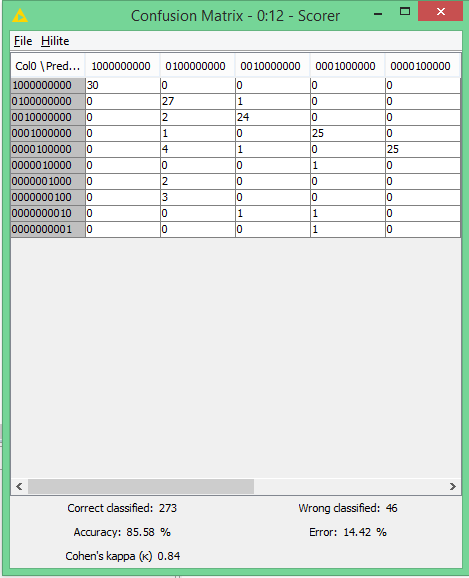
and this results in accuracy of ,



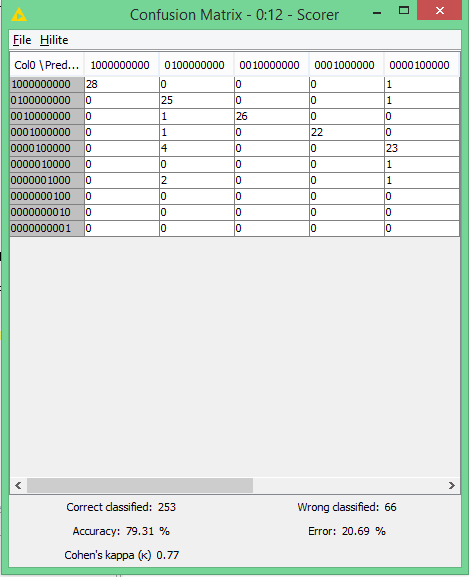
**Using MLP Approach:**



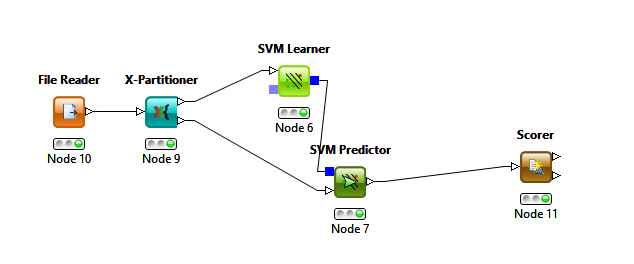
And this results in



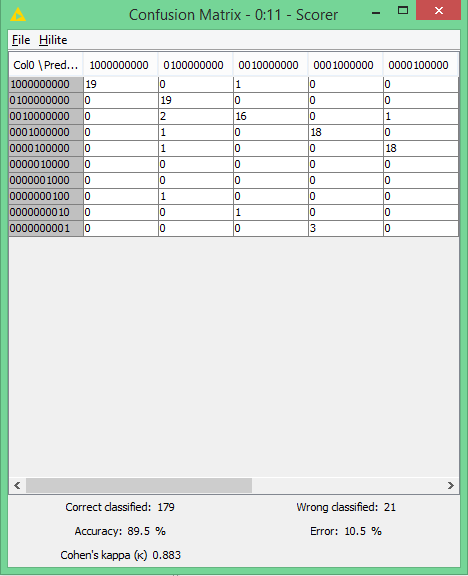
And now applying backward feature elimination method on this dataset results in accuracy of,



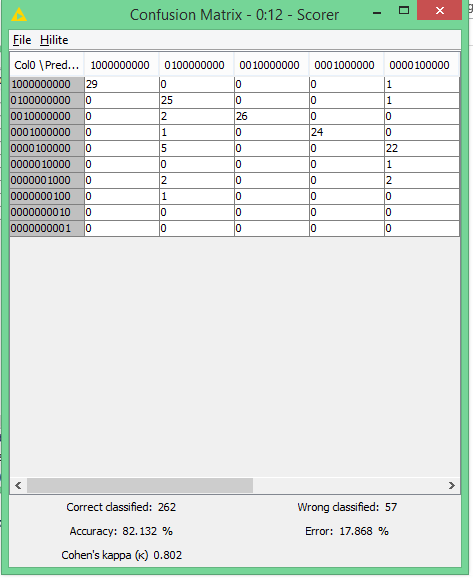
**Using SVM Approach:**



And this approach results in:



And now applying backward elimination on this dataset results in accuracy of,



**Conclusion:**

In this assignment I uses different blocks for predicting my output for various data sets, and then applying various techniques like

* Naïve Bayes,
* Support vector machine,
* Multi-layer perceptron

And uses various other concepts like cross-validation for improving my accuracy and seeing various impact of this on my analysis. Then, I uses backward elimination for selecting useful attributes for problem in which I have a lot of attributes in which some of them are not so useful.