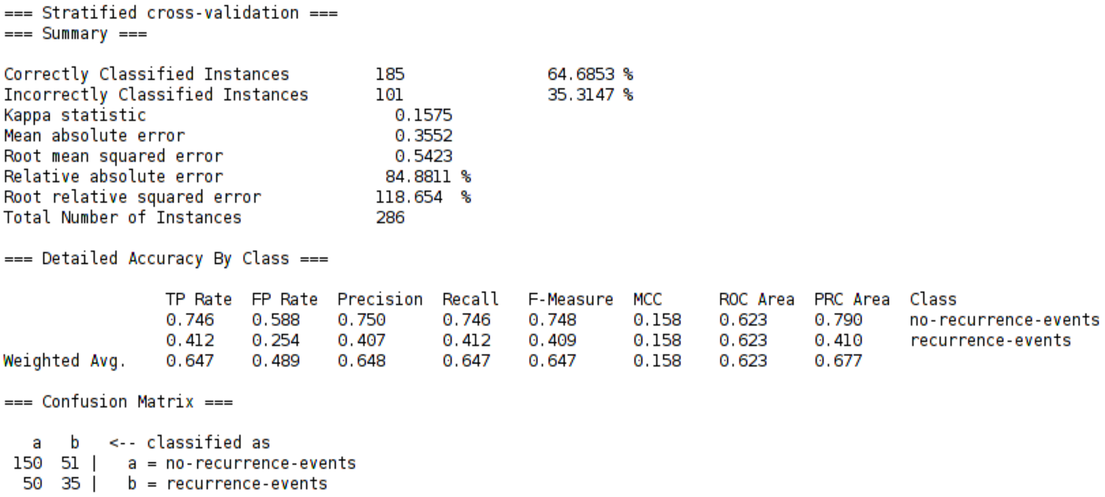
Classification performance evaluation

To evaluate the classification performance I’m running the 5 algorithm of MultilayerPerceptron, Naive Bayes, J48, RandomForest, REPTree each with the 3 data set of breast-cancer.arff, diabetes.arff and iris.arff. For comparative analysis of the classification accuracy i need to evaluate each of every data-set classification rate. To do the thing, need to upload the data in weka.

MultilayerPerceptron classifier with breast-cancer.arff
Process:
Upload data: Preprocess -- Openfile -- Choose your data of format .arff -- open
Classification: Classify -- Classifier choose – weka -- classifier – functions – MultilayerPerception.
Test option: Cross validation choose fold 10 (default value for cross validation)
Run Classification: Run

In the classifier output it will show the summary results of classification.
We will look carefully about the Correctly classified instants. Here for MultilayerPerceptron classification of MultilayerPerceptron of breast-cancer.arff with default setting calculated the **classification accuracy with 64.6853%** of 185 instances.

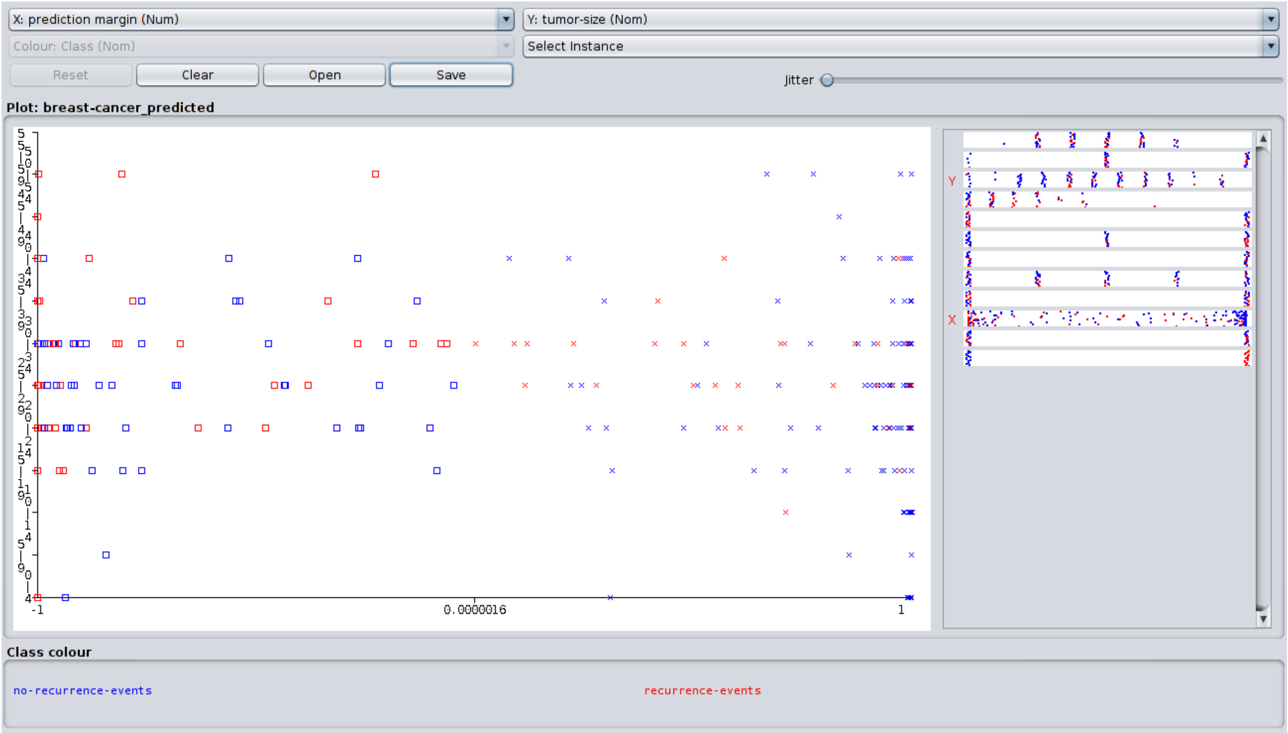


```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         185               64.6853 %
Incorrectly Classified Instances       101               35.3147 %
Kappa statistic                          0.1575
Mean absolute error                      0.3552
Root mean squared error                  0.5423
Relative absolute error                 84.8811 %
Root relative squared error            118.654  %
Total Number of Instances              286

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.746    0.588    0.750      0.746    0.748      0.158    0.623     0.790     no-recurrence-events
                 0.412    0.254    0.407      0.412    0.409      0.158    0.623     0.410     recurrence-events
Weighted Avg.    0.647    0.489    0.648      0.647    0.647      0.158    0.623     0.677

=== Confusion Matrix ===

   a   b   <-- classified as
 150  51 |   a = no-recurrence-events
  50  35 |   b = recurrence-events
```

Figure Summary: In the above figure notice that 185 instants is correctly classified with accuracy of 64.6853 % and miss-classified instant is 101 of inaccuracy of 35.3147 %. The summary comes up with others significance terms along with confusion matrix.

In the classification output it will show details report with each and every attributes with it’s weights value. In the bottom of classification output it will come up with results of Kappa Statistic, Mean absolute error (MSE), Relative absolute error, Root relative square error (RSE), and details accuracy of the class with F-measure which is a very important for model selection. If your F-measure value is near to 1 then you can assume the created model is best. At the very bottom of the classifier output it will show the confusion matrix. Confusion matrix calculate the precision and re-col value of the model accuracy.

To get the classification error:
If you want to see the classification error for more details and want to make a plot with your custom x-label and y-lable. To change the x-label value just click and change the label just right click on the label.
Process: right click of your result list -- visualize classification error.



X: prediction margin (Num)

Y: tumor-size (Nom)

Colour: Class (Nom)

Select Instance

Plot: breast-cancer_predicted

Class colour

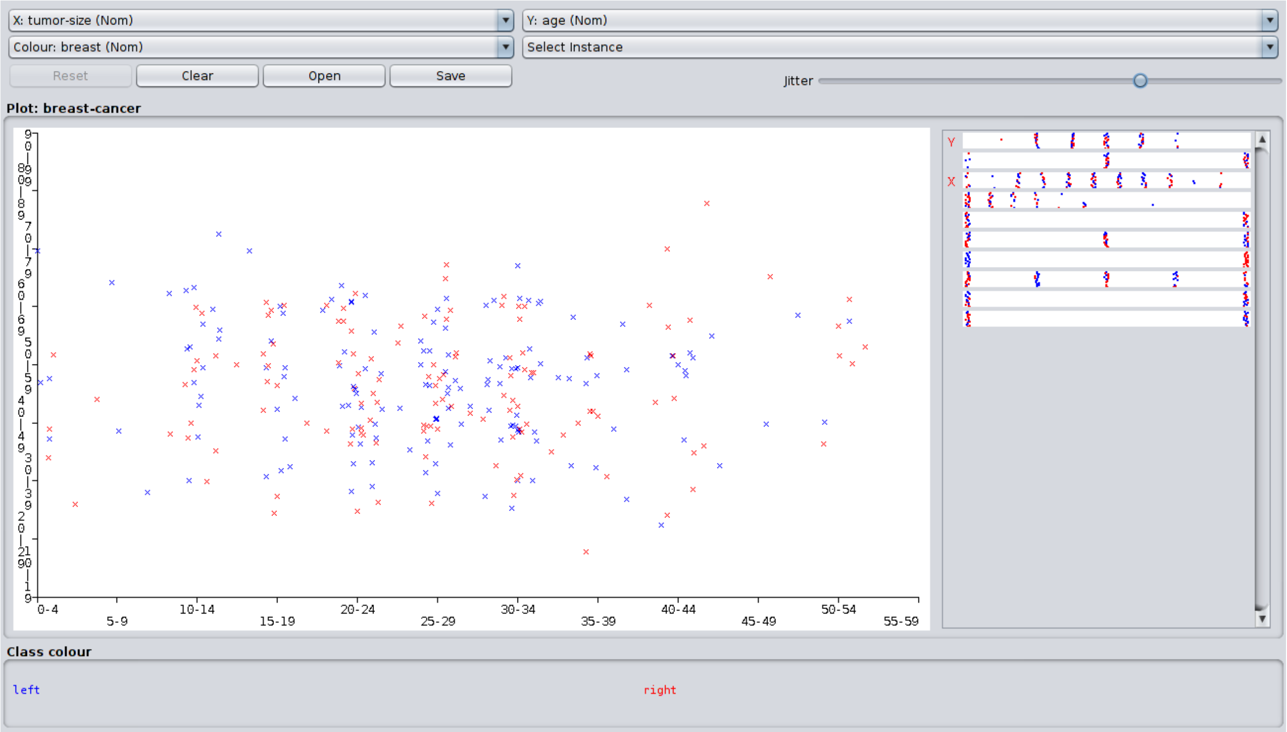
no-recurrence-events

recurrence-events

Figure Summary: The above figure is the classification error with x-label of prediction margin and y-label of tumor-size.

Visualization of data classification:

To visualize the classification go to visualizer it will plot a n*n dimensional visualization plot where n is the number of attributes.



X: tumor-size (Nom)

Y: age (Nom)

Colour: breast (Nom)

Select Instance

Plot: breast-cancer

Class colour

left

right

Figure summary: It’s the picture for classification of tumor-size in x-axis and Age in y-axis. Class color are define with two color of blue and red of shape x, one single type of color represent one class in the range of 0-60 of total 11 classes.

Classification accuracy table for 5 classification algorithm with 3 data set

	Classification Algorithms				
	Multilayer Perceptron	**Naive Bayes**	**J48**	**RandomForest**	***REPTree***
breast-cancer.arff	64.6853	71.6783	***75.5245***	69.5804	70.6294
diabetes.arff	75.3906	***76.3021***	73.8281	75.7813	68.099
iris.arff	***97.3333***	96	96	95.3333	94
Average	237.4092	243.9804	***245.3526***	240.695	232.7284

From the above table we notice that breast-cancer data is giving best classification rate for J48 classification algorithm of 75.5245 % which is relatively good than others algorithms. Naive Bayes classification algorithm gives good results for diabetes data set and Multilayer Perceptron classification algorithm gives relatively good results for iris data set. Average classification rate for the five algorithm J48 classification algorithm gives relatively good rather than others then Naive Bayes algorithm buy REPTree not good for any of the data-set. So by comparing the classification performance we can draw a conclusion that REPTree is not so good for classification.

If we take a close look at the summary table for Multilayer Perceptron for breast-cancer the kappa statistic which adjust the error value of the error matrix in diagonally that value is 0.1575, for any good model this value will be close to zero.
Another error term is Mean absolute error which will calculate error in positive which value is 0.3552, for any good model it seems that this value will be close to zero.
Root mean square, Relative absolute error and Root relative absolute error is also calculating the error value but it has different meaning for each of the error term.

Now in the section of Details accuracy by class section is calculating performance for no-recurrence-events and recurrence-event.

Here is the most important things is that F-measure value which is tell us about the model accuracy and how robust the model. Generally if the F-measure value is close to 1 then it suggest that the model is good enough which will give lower error and higher accuracy for classification . F-measure is calculating form Precision and Recall value.

Precision:
Precision means correctly distinguish the value from the testing data which is the summation of true positive and false negative.
Recall:
Recall means our model miss-classify the test data according to the given data-set. Recall is the summation of false positive and true negative.

F-measure is calculating from precision and recall using the formula

$$\text{F-measure} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

Confusion matrix is known as a error matrix which will tell more about classification algorithm specially for supervised learning algorithm. In the confusion matrix each row is represent the predicted value and each column is represent the actual value. Where the first cell of first row call *true positive*, second cell called as *false positive* and second row of first cell called as *true negative* and second cell called as *false negative*.

Data mining applications

In this ere of internet and rapid developments any resource are generating huge data in each and every day. Data is the most precision and expensive things for Machine Learning and Data Science [3]. Collection of data and store of the data and followed by processing of the data is really challenging and have insightful meaning. So any size of data has great meaning if you do analysis with the data and able to find the hidden pattern of the data to extract the information is really good.

Data mining is a process of data to extract the information form the data-set. For data mining there is number of option is there but to read the data in computer language is easy in Python and R programming if you using tool then Weka is a good for data mining and analysis of the data.

There is so many application of data mining like- Healthcare, Market analysis, Education, Engineering, Fraud detection, Lie Detection, Manufacturing Fault detection, Customer segmentation, Banking, Research, Criminal Investigation, Bio Informatics, etc.

Here I'll describe the major application of data mining and it's application in Healthcare and Customer segmentation.

Data mining application in Healthcare domain:

Healthcare is a great domain for apply data mining [1] [2]. It's a process of prediction or forecasting form the data. Competitive analysis of the data and find the hidden pattern from the data and apply the prediction statistics whats happened with the patients and give an suggestion for further instruction or some information to recover the problem or come out form the situation. If there is too many data so by analysis the data by Machine Learning or Data Science and Neural Network we can explore the complex structure of the data and explore each of the structure to know more about and deploy the predictive analysis structure for successful application.

Source of Healthcare data:

1. RHIhub Rural Health Information hub (https://www.ruralhealthinfo.org/topics/statistics-and-data/data-sources-and-tools)
2. https://healthdata.gov/
3. Agency for Healthcare Research and quality (https://www.ahrq.gov/data/index.html)
4. U.S National Library of Medicin (https://www.nlm.nih.gov/nichsr/stats_tutorial/section3/index.html)

Data collection:

Basically data collection means to keep the record of your data. In early days in healthcare data is collected by pen and paper after that this data is store in a computerized format. But now a days data collection is easy it's most of the part is done by a electronics system with a short span of time it will collect data from body and started doing analysis to find the common pattern and match with the existing one and give report in a glance.

Some application:

Treatment effectiveness:
To predict the treatment effectiveness using data mining very effective. Taking the symptom form patients and doing the comparative analysis on the data and finding the similarities and categorize it. If the category of predictive analysis is match with the critical analysis then suggest the effectiveness of the disease. We can categorize the effectiveness is several group and commanded the undergo operation as advice.

Fraud and abuse detection:
To find the disease and calming the re-commanded medicine, operation or report is use full by using this process. In this way detecting the fraud undergo operation or medicine irrelevant report commanded by doctors easily can be trapped. Where the systematically procedure will give best suggestion for medical recommendation. Using the fraud and abuse detection procedure Texas system recover $2.2 million of there revenue.

Data mining application in Customer Segmentation:

Customer *segmentation* will help for business analysis and market prediction by predicting the characteristic or behaviors of the customers [4]. To target a group of customer instead of individuals and take an action which have a great effectiveness on business analysis. Customer segmentation is the analysis of the data from the back behaviors of the customer and find the opinion of a customer and group all the similar type of customer for further analysis and take an action on them [5] [6]. Customer segmentation allow to monitor over time so the changes become understandable.

Source:
1. data.world (https://data.world/datasets/customer)
2. Keggle Data-sets (https://www.kaggle.com/datasets
3. Website, Bank, Online marketing.

Data Collection:
Early days data is collected from pen and paper resource only. Now in digitize world data is automatically generating form customer behaviors on online may be traveling, internet browsing, banking, shopping, etc.

Insightful ideas form customer data:

Behavioral Analysis:
To segment the customer by their behavioral is importance. It’s possible by data analysis and find the behavior of a customer by his or her financial, opinion, internet browsing.

Shopping or marketing analysis:
By analysis of a customer shopping or card used for shopping it is easy to predict for any customer willingness for shopping or marketing or online purchase and further recommendation can be done with system after analysis of the data.

Product recommendation:
product recommendation for a customer is possible only by analysis the customer purchase that the customer is done before. So by analysis the product the system can give suggestion which type of product is recommendable for any particular customer. When we talk about procedure recommendation then you can take an examples of Netflix movie recommendation, amazon product recommendation, any online shopping website product recommendation.

Appendix:

[1] Data Mining in Healthcare - A Review by Neesha Jothi, Nur’Aini, Abdul Rashid, Wahidah Husain

[2] Study and analysis of data mining for healthcare by Zoubida Alaoui Mdaghri, Morocco, Rabat, Mourad El Yadari, Abdelillah Benyoussef, Abdellah El Kenz.

[3] Medical Analytics for Healthcare Intelligence (https://www.journals.elsevier.com/artificial-intelligence-in-medicine/call-for-papers/medical-analytics-for-healthcare-intelligence)

[4] The Customer Segmentation solution plan (https://www.ibm.com/support/knowledgecenter/en/SSEPGG_9.7.0/com.ibm.datatools.datamining.doc/miningplan_custseg.html)

[5] Research on Customer Segmentation Based on Extension Classification by Chunyan Yang Xiaomei Li, Weihua Li (https://link.springer.com/chapter/10.1007/978-3-642-29426-6_13)

[6] Mining data to discover customer segments by S Kelly (https://link.springer.com/article/10.1057/palgrave.im.4340185)