Foundations of Data Analytics (CS910/ CS430) COURSEWORK 2

Akshara Kannan (2045986) MSc Computer Science

December 8 2020

1 Analysis of Breast Cancer Dataset

1.1 Study of Individual Attributes

a)

Class is of binary attribute type which takes two values, recurrence and no-recurrence.

b)

- Differentiating the types of data present in a dataset is vital to understand and study the patterns it exhibits. The attribute types describe the kind of data held in that attribute. Data types can be qualitative or quantitative in nature.
- There are no quantitative attribute types in the Breast Cancer dataset, neither continuous nor discrete. All of the attributes are qualitative in nature.
- The **menopause** attribute is categoric in nature, because the values it takes are categories of the age groupless that 40('lt40'), greater than 40('gt40') and pre-menopausal('premeno'). Similarly, **deg-malig** and **breast-quad** are also categoric in nature.
- The attributes **node-caps, breast, irradiate** and **class** are binary attributes, because they take two opposing values as labels.
- The attributes **age, tumour-size** and **inv-nodes** are ordered type attributes, because they take up intervals of values in a sequential order. For example, **age** has - 10-19, 20-29 etc

Attribute	Attribute Types
age, tumour-size, inv-nodes	ordered
menopause, breast-quad, deg-malig	categoric
node-caps, breast, irradiate, Class	binary

c)

The shape of the histogram of Age attribute (refer to Figure 1) suggests that it follows the Gaussian or Normal distribution. The parameters of a Gaussian distribution are Mean(μ) and Standard Deviation(σ). When we visualise the **'age'** attribute we see the data displayed in the form of a histogram.

It's clear from the histogram that the age interval 50-59 has the highest Count or frequency, so we can estimate the mean falling in this interval.

Metric	Value
Mean	51.143
Standard Deviation	10.118
Variance	102.3776

Figure 1: Age Attribute

d)

To decide the attributes that exhibit a dependence on class values, we can visualise the graphs of all attributes against the Class attribute (Figure 2). For a particular attribute, if we observe a difference in the proportion of instances which record a recurrence, and instances which record a non-recurrence, then we can conclude that the the class values have dependency on these attributes.

For example, consider the attribute **tumour-size**, for very low and very high tumour sizes (intervals upto 19, and greater than 30), the instances with a recurrence are very less, whereas the mid-interval ranges boast of a high recurrence of cancer. Hence, tumour-size shows a strong dependency.

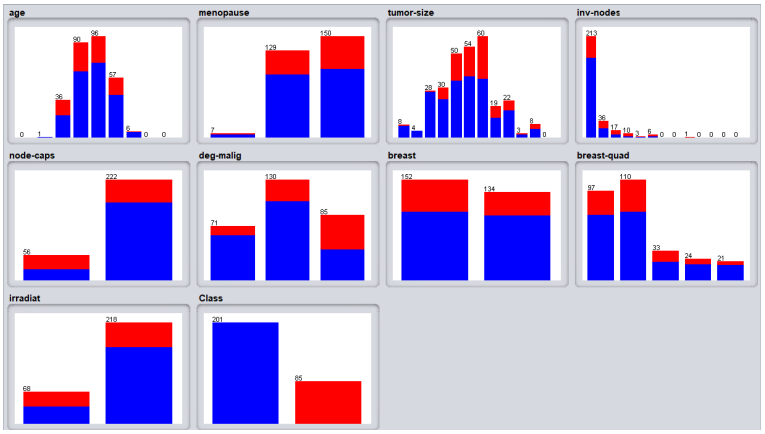

Figure 2: Dependence of Class values on attributes

We can observe a similar strong dependency in the attribute **deg-malig**; it's clear that recurrence of cancer increases for higher degree of malignancy.

In **menopause**, the data shows that for less than 40, the number of occurrences of cancer are very less, leave alone recurrence. The rate of cancer increases for greater than 40 and premenopausal categories, and have an increased occurrence of recurrence between them too.

The attributes **node-caps** and **irradiat** show a weak dependence to the class value.

1.2 Study of Classifiers

a) ZeroR

The ZeroR classifier on the Breast-Cancer dataset results in a model that predicts the class value as non-recurrence events with an accuracy of 65.97% for correctly classified instances. Every instance is classified as a non-recurrence event. The confusion matrix is printed below.

Confusion Matrix :-

a	b	<– Classified as
64	0	a = non-recurrence
33	**0**	b = recurrence

b) Naive Bayes

The Naive Bayes classifier performs considerably better than ZeroR, at the classification task, resulting in an accuracy of 71.134% for correctly classified instances.

Confusion Matrix :-

a	b	<– Classified as
53	11	a = non-recurrence
17	**16**	b = recurrence

c) K Nearest Neighbour I

The values of k were incremented from 1 to 10, and the performance of the classifier was studied. Highest accuracy is observed for k values of 4 to 7.

Confusion Matrix :-

Value of k	Accuracy for correctly classified instances
1	72.16%
2	69.0722%
3	70.10%
4 to 7	**73.19%**
8	72.16%
9	70.10%
10	70.10%

The distance measure used here is Euclidean distance.

c) K Nearest Neighbour II

The values of k were incremented from 1 to 10, and the performance of the classifier was studied.

Here the Manhattan distance metric was set to calculate the distances. We can see the values of accuracies are exactly the same. This is because the attributes are categorical in nature. Both distance metrics work the same way on non-numeric data.

Confusion Matrix :-

Value of k	Accuracy for correctly classified instances
1	72.16%
2	69.0722%
3	70.10%
4 to 7	**73.19%**
8	72.16%
9	70.10%
10	70.10%

e) Decision Trees I

The J48 classifer was performed on the dataset using its default parameters. The accuracy of the model obtained was 68.04% for correctly classified instances.

The algorithm chooses **node-caps** as the attribute at the root of the tree, and classifies an instance as non-recurrent if the node-caps is 'no'. A total of 228.39 instances reached this node, among which 53.4 of them were misclassified. The other child of **node-caps** is the attribute **deg-malig** which is chosen as the decision variable. When deg-malig value was 1 and 3, the instance was classified as recurrent, and as non-recurrent when deg-malig value was 2.

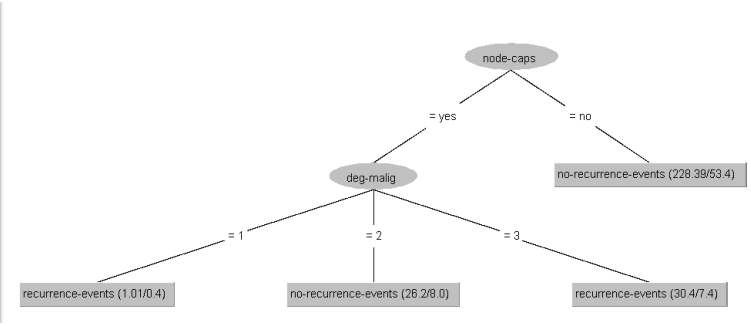

Figure 3: Decision tree

Overfitting is a phenomenon that occurs when a model learns the patterns in the data too specifically, beyond the general rules. It can be identified in a model through metrics such as training and test accuracies as- when the train accuracy is high because the model has learnt the data as well as the noise present in it, whereas the test accuracy is significantly poor, because the model has failed to generalise the patterns in data.

This is not observed with our J48 decision tree. Using a random seed of 42, the dataset was first randomized, using the randomize filter on Weka Preprocess Tab.

- 66% (97 instances)of the dataset was used for training, resulting in an accuracy of 78.3505% for correctly classified instances.
- the rest of the data was used as the testing dataset, and the accuracy obtained was 75% for correctly classified instances.

Therefore, overfitting did not occur.

f) Support Vector Machine

The SVM classifier performs with an accuracy of 70.10%. Magnitude wise, **node-caps** and **deg-malig** have the highest valued weights/coefficients in the model produced by the SVM algorithm, 0.5631 and 0.6353 respectively. This implies how important these two attributes are, to the separation of instances in the SVM classification, ie., the class value exhibits a strong dependency on these attributes. Moreover, these weights are oppositely signed, -0.5631 and +0.6353, which indicates that they classify for the non-recurrence(negative class) and recurrence(positive) class respectively.

By executing the J48 decision tree algorithm, we get the same results; node-caps and deg-malig are the two decision variables chosen for performing the classification.

1.3 Analysing the Classifiers for Breast-Cancer Dataset

For the Breast-cancer dataset, the following classifiers were built and studied:-

1. ZeroR
2. Naive Bayes

3. K Nearest Neighbour
4. Decision Tree
5. Support Vector Machine

The table summarises the accuracies, precision and false positives for all the models.

Classifiers on Breast-Cancer Dataset

Classifier	Accuracy	Weighted Precision	False Positives(Non Recurrence)
ZeroR	65.97%	Undefined	33
Naive Bayes	71.134%	0.701	**17**
KNN (k = 5,6,7,8)	**73.1959**%	**0.809**	26
Decision Tree(J48)	68.0412%	0.657	23
Support Vector Machine(SMO)	70.1031%	0.689	18

Among all of classifiers, we can see that the K-Nearest Neighbour has the highest value of accuracy and Precision.

But, as high as 26 instances are wrongly misclassified as Non-recurrent forms of cancer- which are false positives for the non-recurrent class. In simple words, 26 cases of cancer which were recurrent by nature, were wrongly classified as non-recurrent. In a real-world scenario, possibly, these 26 were supposed to be on intensive medication, but due to such a misclassification, were just assured of non-recurrence of their cancer. The purpose of using a classifier for knowing cancer patients who could suffer from a relapse is defeated here.

The Decision Tree classifier does not have an impressive accuracy nor precision, and has a high value of false positives too. Therefore, it is not the best classifier for this dataset.

Finally, the Naive Bayes Classifier exhibits the most suitable characteristics to be adopted in a real-world scenario. It has the least acceptable value of False positives as well as a high value of Accuracy.

2 Analysis of Car Dataset

2.1 Study of Attributes

The car dataset contains 1728 instances with 7 attributes- buying, maint, doors, persons, lug_boot, safety, and class.

Attributes buying, maint and doors have four labels each and the rest of the attributes-persons, lug_boot and safety, have 3 labels. If we study every attribute individually, on Weka Preprocess Tab, we can see that the instances are split equally among all labels in each attribute.

For example, refer to Figure 4 below, which shows the details about buying attribute. The four labels of this attribute are- vhigh, high, med, low. We can see that there are 432 instances each for all these attribute labels.

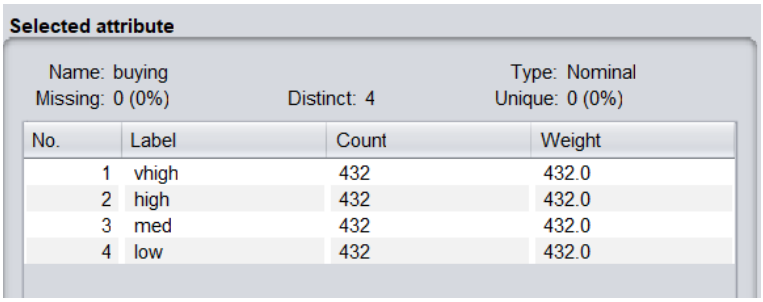

Figure 4: Details about attribute-buying

Another example is considered below, the attribute, lug_boot. It has three labels, and each of them contain an equal split of 576 instances.

It is clear that this dataset has been curated and not recorded through observations. The uniformity in the number of labels among all attributes, and how the instances are equally distributed suggests that the data has been curated through a hierarchical structure like a decision tree. By generating the dataset in this manner by traversing the tree from root to leaf, the dataset owner has ensured that every label of an attribute is represented equally in the dataset.

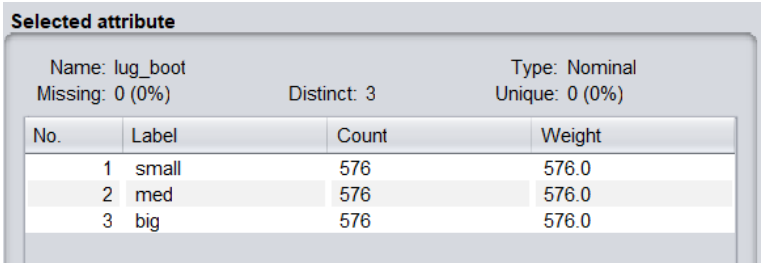

Figure 5: Details about attribute-lug_boot

2.2 Study of Classifiers

Let us go through the performances of each classifier one by one:-

1. Performing the ZeroR classifier on the dataset, the accuracy obtained is 69.5578% for correctly classified instances. Since the model chooses the class which has the maximum instances, it performs poorly for all other classes (acc, good, vgood) other than the majority class (unacc). Therefore it is not a suitable classifier for this dataset.

2. Naive bayes in its default parameter setting (batch size = 100) has a correctly classified accuracy of 87.585% . A higher and lower value of batch size (batch size = 200, 50 respectively) was checked, with no change in the value of the accuracy. The Naive Bayes model is not a suitable model for this classification task as there is a high misclassfication for instances of 'good' class.

3. K Nearest Neighbour model was tested for two different distance measures- Euclidean Distance and Manhattan Distance. For each distance measure, values of k from 1 to 10 were tried. First of all, both distance measures produced the same accuracies for correctly classified instances. This lack of dependency can be attributed to the categorical nature of attributes. With respect to k values, the model performs best with an accuracy of 90.6463% for k = 1,2,3,4, and 5. The accuracy decreases as k increases to 10. The KNN Classifier is also not a suitable model for this classification task as there is a high misclassfication for instances of 'good' class.

4. Decision tree (J48) classifier was trained on the dataset for the parameter settings of unpruned = FALSE, and unpruned = TRUE. The correctly classified accuracy for an unpruned decision tree is 92.6871% and pruned tree (default setting) is 90.9864%. The unpruned setting results a better accuracy. We can see from the confusion matrix, that majority of the instances belonging to each class is classified into their respective classes, hence establishing that J48 is a candidate for the most suitable classifier for this dataset.

5. Finally, different kernels were chosen for performing classification with Support Vector Machines (SMO). The four kernels and the resulting accuracies are listed below:-

   - Poly Kernel - 93.3673%
   - RBF Kernel - 82.1429%
   - Puk - 89.4558%
   - Normalised Poly Kernel - 96.4286%

   The Normalised PolyKernel provides us with the best accuracy observed for this classification task. SVM is the best classifier for this dataset, because it shows in the confusion matrix that it has learnt to differentiate the classes in the most precise manner better than J48 has, resulting in only insignificant misclassifications.

2.3 Analysing 'class' attribute in the Car dataset

2.4 Predicting Evaluation of the car

In the updated dataset, where the Class attribute takes two labels- **unacc** and **acc**, according to the legend below, it is seen that blue represents the **unacc** class, and red represents the **acc** class.

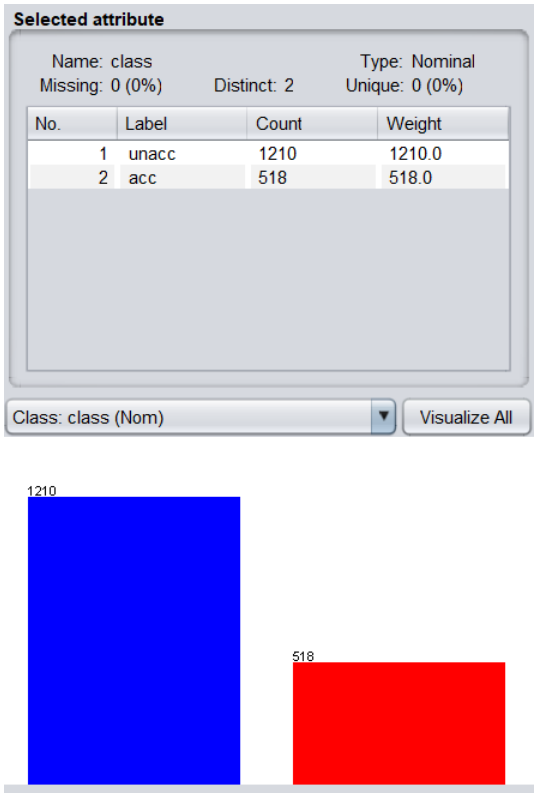

Figure 6: Legend of the new binary class Car dataset

We can see from the histogram visualisation of safety vs Class that we get a pure class of unacc, when safety attribute takes 'low' label. Thus we can say that the low attribute value deterministically results in an unaccepted car. Similarly, the car can be predicted as unacceptable when the person takes '2' label. Refer Figures 7 and 8.

b)

The SVM model trained on the car dataset gives an accuracy of 94.3878% for correctly classified instances, and the model produced is a mathematical one with coefficients for each of the categorical labels of attributes. When the decision tree (J48) model is trained on the dataset, we get an accuracy of 94.2177%, and the model produced is a tree.

We call a model as interpretable when it serves to be easy and intuitive to human eye in understanding and application. In terms of interpretability, the decision tree ranks higher because, it is easy to perform the classification of a new instance by traversing from root to leaf node to obtain a prediction. But in the case of a mathematical model, classification requires meticulous effort and is thus prone to human errors- leading to a misclassification.

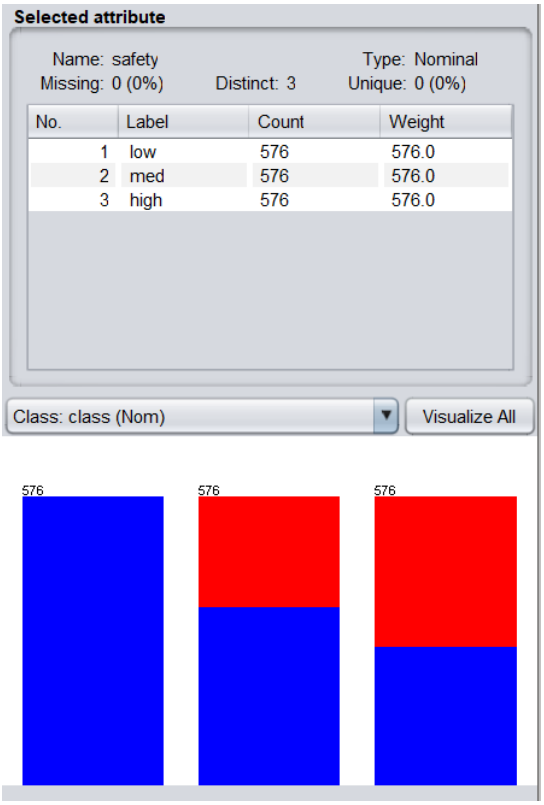

Figure 7: Safety attribute

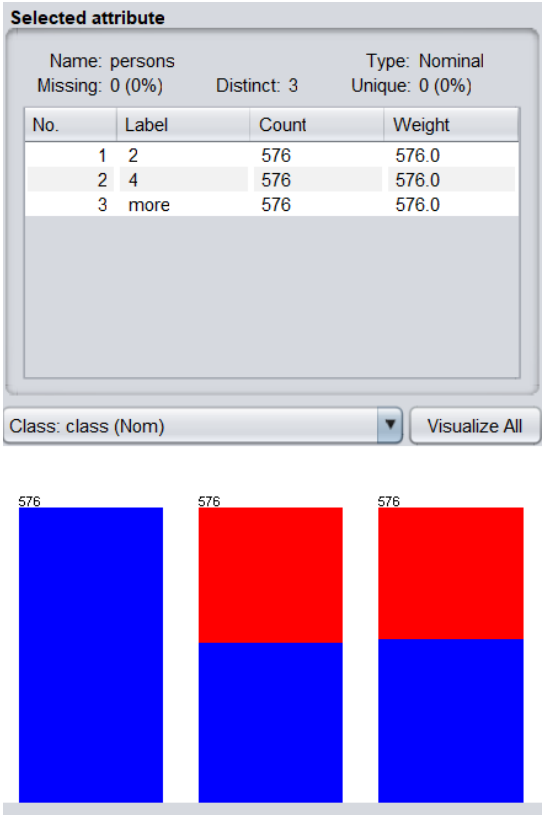

Figure 8: Person attribute