CS 6375

ASSIGNMENT 2

Names of students in your group:

Pragalbha Vasudevan- pxv142730

Sri Harsha Lenka – sxl162530

Number of free late days used: 1  
Note: You are allowed a **total** of 4 free late days for the **entire semester**. You can use at most 2 for each assignment. After that, there will be a penalty of 10% for each late day.

Please list clearly all the sources/references that you have used in this assignment.

**Part-1 (Data Set)**

**Summary:-**

The chosen data set is about dermatology- (skin problem). The present data set mainly deals about the various skin diseases and their symptoms. The following is the information regarding the dataset taken from <http://archive.ics.uci.edu/ml/datasets.html>.

* Number of instances: 366
* Number of attributes: 33
* Number of classes in the predicted Variable: 6
* Number of null values: 8

**Dealing With Null Values:** We are using mean imputation to substitute the missing values as the number of missing values are small.

**Pre-Processing:**  We use scaling to make the data more efficient for classification, as data doesn’t have much inconstancies or null values further preprocessing is not required.

**Part-II(Understand the dataset)**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Data Set Characteristics:** | Multivariate | **Number of Instances:** | 366 | **Area:** | Life |
| **Attribute Characteristics:** | Categorical, Integer | **Number of Attributes:** | 33 | **Date Donated** | 1998-01-01 |
| **Associated Tasks:** | Classification | **Missing Values?** | Yes | **Number of Web Hits:** | 86661 |

**Source of the dataset:**

[**http://archive.ics.uci.edu/ml/datasets/Dermatology**](http://archive.ics.uci.edu/ml/datasets/Dermatology)

**Dataset Information:**

This database contains 34 attributes, 33 of which are linear valued and one of them is nominal.

The differential diagnosis of erythemato-squamous diseases is a real problem in dermatology. They all share the clinical features of erythema and scaling, with very little differences. The diseases in this group are psoriasis, seboreic dermatitis, lichen planus, pityriasis rosea, cronic dermatitis, and pityriasis rubra pilaris. Usually a biopsy is necessary for the diagnosis but unfortunately these diseases share many histopathological features as well. Another difficulty for the differential diagnosis is that a disease may show the features of another disease at the beginning stage and may have the characteristic features at the following stages. Patients were first evaluated clinically with 12 features. Afterwards, skin samples were taken for the evaluation of 22 histopathological features. The values of the histopathological features are determined by an analysis of the samples under a microscope. In the dataset constructed for this domain, the family history feature has the value 1 if any of these diseases has been observed in the family, and 0 otherwise. The age feature simply represents the age of the patient. Every other feature (clinical and histopathological) was given a degree in the range of 0 to 3. Here, 0 indicates that the feature was not present, 3 indicates the largest amount possible, and 1, 2 indicate the relative intermediate values. The names and id numbers of the patients were recently removed from the database.

**Attribute Information**

Clinical Attributes: (take values 0, 1, 2, 3, unless otherwise indicated)

1: erythema

2: scaling

3: definite borders

4: itching

5: koebner phenomenon

6: polygonal papules

7: follicular papules

8: oral mucosal involvement

9: knee and elbow involvement

10: scalp involvement

11: family history, (0 or 1)

34: Age (linear)

Histopathological Attributes: (take values 0, 1, 2, 3)

12: melanin incontinence

13: eosinophils in the infiltrate

14: PNL infiltrate

15: fibrosis of the papillary dermis

16: exocytosis

17: acanthosis

18: hyperkeratosis

19: parakeratosis

20: clubbing of the rete ridges

21: elongation of the rete ridges

22: thinning of the suprapapillary epidermis

23: spongiform pustule

24: munro microabcess

25: focal hypergranulosis

26: disappearance of the granular layer

27: vacuolisation and damage of basal layer

28: spongiosis

29: saw-tooth appearance of retes

30: follicular horn plug

31: perifollicular parakeratosis

32: inflammatory monoluclear inflitrate

33: band-like infiltrate

**Experiments**

For supervised concept learning (classification) tasks, the classification accuracy of the classifier is one measure of performance. The most commonly used metric for classification accuracy is the percentage of correctly classified test instances over all test instances. To measure the classification accuracy, 10-fold cross-validation technique is used in the experiments. That is, the whole dataset is partitioned into 10 subsets. The 9 of the subsets is used as the training set, and the tenth is used as the test set. This process is repeated 10 times once for each subset being the test set. Classification is the average of these 10 runs. This technique ensures that the training and test sets are disjoint. The VFI5 algorithm achieved 96:2% accuracy on the Dermatology dataset, which means that, out of 37, only about 1 test instance is misclassified by the VFI5 algorithm.

**Relavant Papers:**

G. Demiroz, H. A. Govenir, and N. Ilter, "Learning Differential Diagnosis of Eryhemato-Squamous Diseases using Voting Feature Intervals", Aritificial Intelligence in Medicine

**Part-III(Pre-Processing)**

**Strategy:-** For preprocessing after the substation of missing values we scale the data

**Code:-**

**Plotting:**

#install.packages("mice")

library(mice)

#install.packages("VIM")

library(mice)

aggr\_plot <- aggr(DataRaw, col=c('navyblue','red'), numbers=TRUE, sortVars=TRUE, labels=names(DataRaw), cex.axis=.9, gap=1, ylab=c("Histogram of missing data","Pattern"))

DataRaw$Age[which(is.na(DataRaw$Age))]<-mean(na.omit(DataRaw[["Age"]]))

**Correlation Plot:**

#install.packages("corrplot")

library(corrplot)

M <- cor(DataRaw[1:10])

corrplot(M, method="number")

**Data Distribution:**

hist(DataRaw$class, breaks=12, col="red")

**Separation of training set and target:**

target<-DataRaw$class

DataRaw<-subset(DataRaw, select=-class)

**Preprocessing scaling data:**

#install.packages("caret")

library(lattice)

library(ggplot2)

library(caret)

preprocessParams <- preProcess(DataRaw, method=c("scale"))

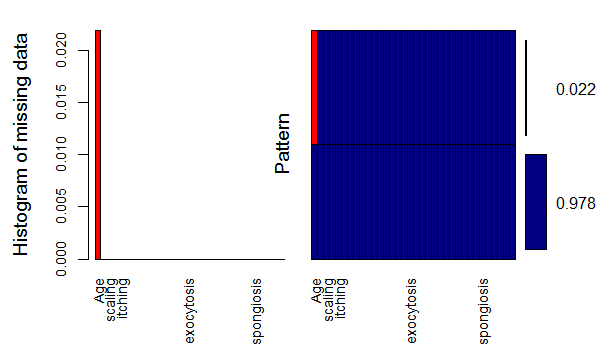
print(preprocessParams)

transformed<-predict(preprocessParams,DataRaw)

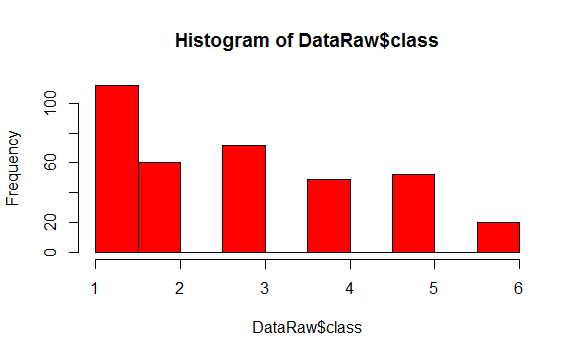
summary(transformed)

**Plots:-**

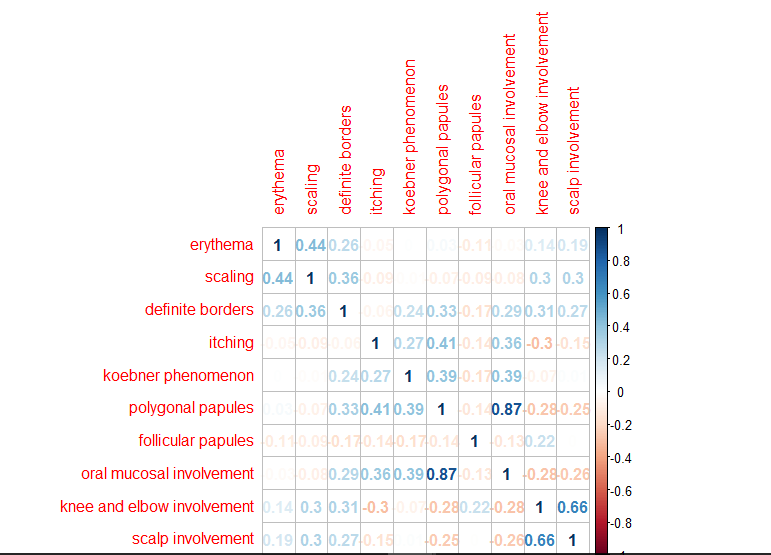
**Missing Data:**



**Data Distribution Plot Over Class:**



**Correlation Plot:-**

****

**Part-IV**

**Experiment Logs:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Experiment** | **Classifier** | **Train/Test Ratio** | **minsplit** | **CP** | **Outcome** |
| **1** | Decision Tree | 80/20 | 20 | 0.01 | 90.41% |
| **2** | Decision Tree | 80/20 | 5 | 0.5 | 36.98% |
| **3** | Decision Tree | 80/20 | 40 | 0.005 | 82.2% |
| **4** | Decision Tree | 80/20 | 10 | 1 | 26% |
| **5** | Decision Tree | 80/20 | 100 | 0.04 | 89.04% |
| **6** | Decision Tree | 80/20 | 2 | 0.3 | 30.13% |
| **7** | Decision Tree | 80/20 | 55 | 1.5 | 32.87% |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Experiment** | **Classifier** | **Train/Test Ratio** | **Threshold** | **Repetition** | **Outcome** |
| **1** | Perceptron | 80/20 | 0.01 | 1 | 69% |
| **2** | Perceptron | 80/20 | 0.5 | 5 | 64.3% |
| **3** | Perceptron | 80/20 | 0.001 | 10 | 56.2% |
| **4** | Perceptron | 80/20 | 0.8 | 2 | 50.3% |
| **5** | Perceptron | 80/20 | 0.35 | 4 | 58.90% |
| **6** | Perceptron | 80/20 | 0.2 | 8 | 64.0 % |
| **7** | Perceptron | 80/20 | 0.09 | 12 | 58.9% |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Experiment** | **Classifier** | **Train/Test Ratio** | **Hidden** | **Threshold** | **Repetition** | **Outcome** |
| **1** | Ann | 80/20 | 3 | 0.01 | 1 | 90.4% |
| **2** | Ann | 80/20 | 4 | 0.5 | 5 | 91% |
| **3** | Ann | 80/20 | 8 | 0.001 | 10 | 76.7% |
| **4** | Ann | 80/20 | 10 | 0.8 | 2 | 90.34% |
| **5** | Ann | 80/20 | 15 | 0.35 | 4 | 83.56% |
| **6** | Ann | 80/20 | 7 | 0.2 | 8 | 91.7% |
| **7** | Ann | 80/20 | 6 | 0.09 | 12 | 83.56% |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Experiment** | **Classifier** | **Train/Test Ratio** | **COST** | **GAMMA** | **Kernel** | **Outcome** |
| **1** | SVM | 80/20 | 1 | 2 | radial | 23.29% |
| **2** | SVM | 80/20 | 4 | 6 | polynomial | 64.38% |
| **3** | SVM | 80/20 | 10 | 8 | sigmoid | 0% |
| **4** | SVM | 80/20 | 10 | - | linear | 52.54% |
| **5** | SVM | 80/20 | 15 | 0.6 | polynomial | 63.01% |
| **6** | SVM | 80/20 | 7 | - | linear | 61.64% |
| **7** | SVM | 80/20 | 6 | 0.09 | polynomial | 54.79% |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Experiment** | **Classifier** | **Train/Test Ratio** | **Laplace** | **Threshold** | **Outcome** |
| **1** | Naive Bayes | 80/20 | 2 | 0.001 | 76.7% |
| **2** | Naive Bayes | 80/20 | 4 | 0.5 | 84.93% |
| **3** | Naïve Bayes | 80/20 | 0.2 | 0.1 | 78.08% |
| **4** | Naïve Bayes | 80/20 | 10 | 0.09 | 84.98% |
| **5** | Naïve Bayes | 80/20 | 6 | 0.3 | 78.21% |
| **6** | Naïve Bayes | 80/20 | 0.08 | 0.87 | 76.7% |
| **7** | Naïve Bayes | 80/20 | 0.9 | 0.7 | 80.82% |

**Part-V**

**Result Analysis: -**

In my opinion Naïve Bayes algorithm is the best as it’s accuracy rates vary very less with respective to parameter changes. This would allow us to obtain a very accurate data with good enough parameters.

And from the above set of observations its obvious that svm has very low accuracy rate even with change of parameters to make it better. With sigmoid function as parameter for kernel it outputs almost 0% accuracy.

**Comparison Authors Analysis Vs Our Result Set:**

Here the author talks about VFI5 classification algorithm from the result analysis of the paper we can observe that it has high accuracy rate and shorter prediction compared to the algorithms we use. But one major drawback is it ignores data rows containing null values which might lead to the less accurate data set.

There are few ways to cope up with the author's accuracy like using very low threshold value, a good number of hidden nodes increase accuracy. Even though we might catch up in accuracy but the time taken to give out results will grow exponentially with increase in data compared to the authors.

**REFERENCES:**

G. Demiroz, H. A. Govenir, and N. Ilter, "Learning Differential Diagnosis of Eryhemato-Squamous Diseases using Voting Feature Intervals", Aritificial Intelligence in Medicine

<http://www-bcf.usc.edu/~gareth/ISL/ISLR%20First%20Printing.pdf>

<http://personal.disco.unimib.it/Vanneschi/McGrawHill_-_Machine_Learning_-Tom_Mitchell.pdf>