**MACHINE LEARNING PROJECT**

**Milestone #1**

**SPAMBASE**

**A. Dataset Building and Normalization**

Provide the following details:

1. Source of Dataset: UCI Machine Learning Repository
2. Description of Dataset (include the number of original features and the sample size):

This dataset contains a collection of spam emails from a set of individuals who have filed spam to the owner’s postmaster. Within the dataset, there exists a collection of non-spam emails as well. There are 4601 samples and 58 features.

1. Name, Description and type of each feature. Include here if you had to normalize the feature, the reason for its normalization and the range of values (e.g. 0 to 1).

To lessen confusion, features 1-58 have been labeled alphabetically.

48 continuous real [0,100] attributes of type **word\_freq\_WORD**

= percentage of words in the e-mail that match WORD, i.e. 100 \* (number of times the WORD appears in the e-mail) / total number of words in e-mail. A "word" in this case is any string of alphanumeric characters bounded by non-alphanumeric characters or end-of-string.

6 continuous real [0,100] attributes of type **char\_freq\_CHAR**

= percentage of characters in the e-mail that match CHAR, i.e. 100 \* (number of CHAR occurences) / total characters in e-mail

1 continuous real [1,...] attribute of type **capital\_run\_length\_average**

= average length of uninterrupted sequences of capital letters

1 continuous integer [1,...] attribute of type **capital\_run\_length\_longest**

= length of longest uninterrupted sequence of capital letters

1 continuous integer [1,...] attribute of type **capital\_run\_length\_total**

= sum of length of uninterrupted sequences of capital letters

= total number of capital letters in the e-mail

1 nominal {0,1} class attribute of type **spam**

= denotes whether the e-mail was considered spam (1) or not (0), i.e. unsolicited commercial e-mail.

1. Processed/Normalized dataset If the dataset cannot be uploaded because of its size, provide a link to it (via Dropbox or Google drive)

[**https://www.dropbox.com/sh/xyv2ce9j04iuxdd/AACu0Xr\_GNOyT-YPJPbFZRPva?dl=0**](https://www.dropbox.com/sh/xyv2ce9j04iuxdd/AACu0Xr_GNOyT-YPJPbFZRPva?dl=0)

1. Number of features that were removed and the reason for their removal

0 features were removed. All features are relevant.

**B. Performance of kNN, Decision Trees, Bayesian Network on Different Feature Sets**

1. Performance measures to include are Accuracy, Precision, Recall, F-Measure, AUC and Kappa statistic
2. Create one table for each measure indicated above. Use the table below as your guide. Note that you need to provide the performance when the complete feature set is used for model building as well as when different feature selection techniques are applied.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classifier** | **Complete Feature Set** | **Feature Selection using Forward Search** | **Feature Selection using Backward Search**  **(BACKWARD ELIMINATION)** | **Top 9 features (with higher coefficients) based on 3 Components of Principal Components Analysis (PCA)** |
| **kNN (k=1)** | 90.31 | 89.57 | 91.09 | 72.09 |
| **kNN (k=3)** | 90.35 | 91.92 | 91.33 | 71.85 |
| **kNN (k=5)** | 89.94 | 92.44 | 90.87 | 72.03 |
| **kNN (k=7)** | 89.63 | 92.24 | 91.37 | 71.72 |
| **kNN (k=9)** | 89.24 | 91.91 | 90.85 | 72.03 |
| **Decision Trees** | 88.35 | 90.59 | 90.55 | 82.03 |
| **Bayesian Network (e.g. Naïve Bayes)** | 79.79 | 81.79 | 82.81 | 62.51 |

**Table 1 – Accuracy (%)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classifier** | **Complete Feature Set** | **Feature Selection using Forward Search** | **Feature Selection using Backward Search**  **(BACKWARD ELIMINATION)** | **Top 9 features (with higher coefficients) based on 3 Components of Principal Components Analysis (PCA)** |
| **kNN (k=1)** | 91.92 | 93.01 | 92.66 | 86.03 |
| **kNN (k=3)** | 91.15 | 92.91 | 91.91 | 85.14 |
| **kNN (k=5)** | 90.81 | 93.42 | 91.64 | 84.88 |
| **kNN (k=7)** | 90.49 | 93.48 | 91.98 | 84.37 |
| **kNN (k=9)** | 89.63 | 92.73 | 91.31 | 84.44 |
| **Decision Trees** | 85.81 | 89.13 | 90.14 | 84.21 |
| **Bayesian Network (e.g. Naïve Bayes)** | 96.74 | 79.65 | 95.94 | 97.36 |

**Table 2 – Precision (%)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classifier** | **Complete Feature Set** | **Feature Selection using Forward Search** | **Feature Selection using Backward Search**  **(BACKWARD ELIMINATION)** | **Top 9 features (with higher coefficients) based on 3 Components of Principal Components Analysis (PCA)** |
| **kNN (k=1)** | 92.14 | 89.53 | 92.65 | 64.42 |
| **kNN (k=3)** | 93.15 | 93.83 | 93.97 | 64.89 |
| **kNN (k=5)** | 92.83 | 94.19 | 93.47 | 65.53 |
| **kNN (k=7)** | 92.65 | 93.76 | 93.97 | 65.46 |
| **kNN (k=9)** | 93.04 | 94.05 | 93.87 | 66.03 |
| **Decision Trees** | 96.88 | 96.23 | 94.84 | 86.62 |
| **Bayesian Network (e.g. Naïve Bayes)** | 68.97 | 94.01 | 74.82 | 39.20 |

**Table 3 – Recall (%)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classifier** | **Complete Feature Set** | **Feature Selection using Forward Search** | **Feature Selection using Backward Search**  **(BACKWARD ELIMINATION)** | **Top 9 features (with higher coefficients) based on 3 Components of Principal Components Analysis (PCA)** |
| **kNN (k=1)** | 92.02 | 91.22 | 92.64 | 73.66 |
| **kNN (k=3)** | 92.13 | 93.36 | 92.92 | 73.63 |
| **kNN (k=5)** | 91.79 | 93.79 | 92.54 | 73.93 |
| **kNN (k=7)** | 91.54 | 93.61 | 92.95 | 73.69 |
| **kNN (k=9)** | 91.29 | 93.38 | 92.56 | 74.09 |
| **Decision Trees** | 90.99 | 92.54 | 92.40 | 85.37 |
| **Bayesian Network (e.g. Naïve Bayes)** | 80.51 | 86.22 | 84.05 | 55.85 |

**Table 4 – F-Measure (%)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classifier** | **Complete Feature Set** | **Feature Selection using Forward Search** | **Feature Selection using Backward Search**  **(BACKWARD ELIMINATION)** | **Top 9 features (with higher coefficients) based on 3 Components of Principal Components Analysis (PCA)** |
| **kNN (k=1)** | 0.5 | 0.5 | 0.5 | 0.5 |
| **kNN (k=3)** | 0.919 | 0.925 | 0.919 | 0.756 |
| **kNN (k=5)** | 0.936 | 0.946 | 0.935 | 0.766 |
| **kNN (k=7)** | 0.953 | 0.969 | 0.955 | 0.771 |
| **kNN (k=9)** | 0.954 | 0.958 | 0.958 | 0.769 |
| **Decision Trees** | 0.864 | 0.891 | 0.907 | 0.853 |
| **Bayesian Network (e.g. Naïve Bayes)** | 0.934 | 0.872 | 0.936 | 0.846 |

**Table 5 – AUC**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classifier** | **Complete Feature Set** | **Feature Selection using Forward Search** | **Feature Selection using Backward Search**  **(BACKWARD ELIMINATION)** | **Top 9 features (with higher coefficients) based on 3 Components of Principal Components Analysis (PCA)** |
| **kNN (k=1)** | 0.797 | 0.784 | 0.813 | 0.453 |
| **kNN (k=3)** | 0.797 | 0.830 | 0.817 | 0.446 |
| **kNN (k=5)** | 0.788 | 0.841 | 0.808 | 0.448 |
| **kNN (k=7)** | 0.782 | 0.837 | 0.818 | 0.442 |
| **kNN (k=9)** | 0.772 | 0.830 | 0.807 | 0.447 |
| **Decision Trees** | 0.747 | 0.798 | 0.799 | 0.621 |
| **Bayesian Network (e.g. Naïve Bayes)** | 0.607 | 0.599 | 0.660 | 0.324 |

**Table 6 – Kappa Statistic**

**C. Discussion of Results**

Discuss the results, determine which classifier is better and explain the reasons for choosing such classifier.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | CFS | FORWARD | BACKWARD | TOP 9 |
| ACCURACY (sum) | 617.61 | 630.46 | 628.87 | 504.26 |
| PRECISION (sum) | 636.55 | 634.33 | 645.58 | 606.43 |
| RECALL (sum) | 629.66 | 655.6 | 637.59 | 452.15 |
| F-MEASURE (sum) | 630.27 | 644.12 | 640.06 | 510.22 |
| AUC (sum) | 6.06 | 6.061 | 6.11 | 5.261 |
| KAPPA (sum) | 5.29 | 5.519 | 5.522 | 3.181 |
| **TOTAL** | 2525.44 | 2576.09 | 2563.732 | 2081.502 |

**Table 7 – Sum of Each Measure**

The sum of each measure of Complete Feature Set, Feature Selection using Forward Search, Feature Selection using Backward Search, and Principal Components Analysis (PCA) have been recorded onto Table 7. Based on the results above, the Feature Selection using Forward Search is more dominant compared to the others. Though Forward Search and Backward Search have close values, Forward Search has a slightly bigger Total compared to Backward Search. Since Forward Search is more prevalent in Table 7, Forward Search has been chosen to be the classifier.

**D. Problems Encountered**

By using the previous dataset (Water Quality Complaints), Part B of the Milestone 1 results to very low percentages in its measures. The group have found this unacceptable. In order to solve this problem, the group have attempted to modify the dataset by cleaning it. Unfortunately, this method did not work. Since there was limited time, the group has decided to look for another dataset (Spambase).

The group has encountered lengthy process run time, especially with k-NN. There was no solution to this so the group just had to wait for the process to finish.