**NAME**

**COLLEGE NUMBER**

**Introduction**

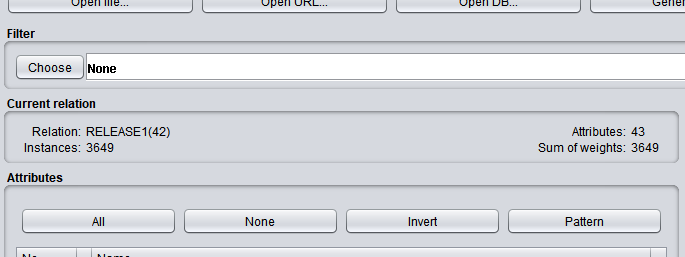
This project is based on an ideal software classification case scenario for determining software metrics and fault data.

Objective

* To build and validate production models based on the datasets provided for metrics 28 and matrices 42
* To establish the model associated with the prediction algorithm for the two datasets
* To determine which classification model is best suited to forecast outcomes based on the two datasets

**Dataset description**

The dataset obtained for this project were obtained from a list of two sets that are called as the Data28 matrices and Data42 matrices. The first set contains the product and the execution whereas the second dataset contains the product, process and execution. All the dataset share in the format of .arff. A sample set R1.arff from the 28metrcies revealed a count of 3649 records with 29 column variables. Also, a sample set from the 42 matrices revealed the total records of 3649 with 43 column values and variables.



With the provided datasets are independent and dependent variables that will remain constant during the whole classification journey on the 5 models while the dependent variables will keep varying and outputs observed and recorded.

**Sample naïve Bayes on Release 1 on 42 dataset**

=== Run information ===

Scheme:       weka.classifiers.bayes.NaiveBayes

Relation:     RELEASE1(42)

Instances:    3649

Attributes:   43

              PRS\_PRS

              VO\_PRS

              INT

              VO

              EXT

              FTUPD

              UPD

              DIFFFT

              SGRTHTOT

              SMODTOT

              UNQUPDID

              VLOUPDSM

              LOUPDSM

              UPD\_CAR

              USAGE

              CALUNQ

              CAL2

              CNDSPNSM

              CNDSPNMX

              CTRNSTMX

              FILINCUQ

              KNT

              LOC

              CNDNOT

              IFTH

              LOP

              NDSENT

              NDSEXT

              NDSPND

              NDSINT

              LGPATH

              STMCTL

              STMDEC

              STMEXE

              VARGLBUS

              VARSPNSM

              VARSPNMX

              VARUSDUQ

              VARUSD2

              RESCPU

              BUSCPU

              TANCPU

              CLASSID

Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

Naive Bayes Classifier

                    Class

Attribute               1          2

                   (0.94)     (0.06)

=====================================

PRS\_PRS

  mean               1.018      2.569

  std. dev.         1.7403     3.4327

  weight sum          3420        229

  precision           1.85       1.85

VO\_PRS

  mean              0.0413     0.4076

  std. dev.         0.2836     1.1166

  weight sum          3420        229

  precision         1.3333     1.3333

INT

  mean              1.1621     2.0556

  std. dev.          1.682     2.6489

  weight sum          3420        229

  precision         1.5333     1.5333

VO

  mean               0.048      0.179

  std. dev.         0.2842     0.5018

  weight sum          3420        229

  precision              1          1

EXT

  mean              0.0989     0.4687

  std. dev.         0.3937     1.0024

  weight sum          3420        229

  precision         1.1667     1.1667

FTUPD

  mean              0.9605     3.3537

  std. dev.         1.9434     5.8064

  weight sum          3420        229

  precision          2.087      2.087

UPD

  mean              2.9424     6.2617

  std. dev.         2.2494     6.9671

  weight sum          3420        229

  precision         1.9643     1.9643

DIFFFT

  mean              1.0679     2.1056

  std. dev.         1.1258     2.0119

  weight sum          3420        229

  precision         1.0909     1.0909

SGRTHTOT

  mean             28.2322   158.1127

  std. dev.        300.505   815.1294

  weight sum          3420        229

  precision        45.5444    45.5444

SMODTOT

  mean             61.1511   251.4507

  std. dev.       288.3385  1006.2906

  weight sum          3420        229

  precision        32.7544    32.7544

UNQUPDID

  mean              2.0096     4.0546

  std. dev.         1.3232     3.5266

  weight sum          3420        229

  precision            1.5        1.5

VLOUPDSM

  mean              0.4643     0.8515

  std. dev.         0.8069     1.1912

  weight sum          3420        229

  precision              1          1

LOUPDSM

  mean              0.8779     1.6719

  std. dev.         1.2935     1.8636

  weight sum          3420        229

  precision         1.1429     1.1429

UPD\_CAR

  mean             303.711   590.7369

  std. dev.       431.2337   936.9773

  weight sum          3420        229

  precision         7.5222     7.5222

USAGE

  mean              0.6249     0.8585

  std. dev.         0.3969     0.2155

  weight sum          3420        229

  precision         0.0031     0.0031

CALUNQ

  mean            101.2222   224.5119

  std. dev.       151.2644    259.424

  weight sum          3420        229

  precision         3.2333     3.2333

CAL2

  mean             89.0599   209.4679

  std. dev.       170.3378   288.9603

  weight sum          3420        229

  precision         7.8649     7.8649

CNDSPNSM

  mean            514.9816   1348.613

  std. dev.      1039.7947  1949.7277

  weight sum          3420        229

  precision         6.7108     6.7108

CNDSPNMX

  mean              33.036    68.4298

  std. dev.        47.3541    71.9907

  weight sum          3420        229

  precision         1.4682     1.4682

CTRNSTMX

  mean              6.6329    10.1318

  std. dev.         4.6916       6.72

  weight sum          3420        229

  precision         1.2381     1.2381

FILINCUQ

  mean             32.0124    54.5739

  std. dev.        18.6495    23.8707

  weight sum          3420        229

  precision         1.2143     1.2143

KNT

  mean             157.321   358.3578

  std. dev.       677.2248  1042.9585

  weight sum          3420        229

  precision        37.8873    37.8873

LOC

  mean           1395.7369  3372.9632

  std. dev.      2132.7126  3742.7717

  weight sum          3420        229

  precision        11.5509    11.5509

CNDNOT

  mean             303.626   776.0188

  std. dev.       468.0225   887.4639

  weight sum          3420        229

  precision         5.7598     5.7598

IFTH

  mean            107.7508   283.9952

  std. dev.       169.5471   324.4615

  weight sum          3420        229

  precision         3.5054     3.5054

LOP

  mean             10.7953    26.4859

  std. dev.        19.4696    42.0812

  weight sum          3420        229

  precision         3.1639     3.1639

NDSENT

  mean             22.8764    41.5037

  std. dev.        35.2513    48.6532

  weight sum          3420        229

  precision         2.6284     2.6284

NDSEXT

  mean             61.7444   126.2428

  std. dev.        96.0326   135.8804

  weight sum          3420        229

  precision         3.7016     3.7016

NDSPND

  mean              8.4515    22.5305

  std. dev.        21.1083     34.627

  weight sum          3420        229

  precision         5.5478     5.5478

NDSINT

  mean            247.4686   653.6664

  std. dev.       386.2877   758.2925

  weight sum          3420        229

  precision         5.3629     5.3629

LGPATH

  mean             10.7533     20.814

  std. dev.         9.0538    17.1466

  weight sum          3420        229

  precision         0.0619     0.0619

STMCTL

  mean            175.3005   436.3251

  std. dev.       270.6152   494.6142

  weight sum          3420        229

  precision         4.2213     4.2213

STMDEC

  mean             61.4149    141.707

  std. dev.       100.4657   158.4884

  weight sum          3420        229

  precision         4.0267     4.0267

STMEXE

  mean            927.8033  2196.2597

  std. dev.      1395.8922   2487.944

  weight sum          3420        229

  precision         9.3987     9.3987

VARGLBUS

  mean             762.261  1841.6997

  std. dev.       1157.172  2024.2192

  weight sum          3420        229

  precision         9.1794     9.1794

VARSPNSM

  mean          20020.6437 79659.9155

  std. dev.     49796.9073127669.3008

  weight sum          3420        229

  precision        285.743    285.743

VARSPNMX

  mean            298.9922   725.7127

  std. dev.       390.9225   753.4889

  weight sum          3420        229

  precision          5.429      5.429

VARUSDUQ

  mean            569.0885  1296.3427

  std. dev.       845.5481  1423.0471

  weight sum          3420        229

  precision         7.0589     7.0589

VARUSD2

  mean            845.4925  2274.0485

  std. dev.      1385.5814  2647.1897

  weight sum          3420        229

  precision        10.9746    10.9746

RESCPU

  mean               1.852     7.0877

  std. dev.        20.0379    22.8289

  weight sum          3420        229

  precision         2.7793     2.7793

BUSCPU

  mean              2.0459     7.2625

  std. dev.        21.8159     23.069

  weight sum          3420        229

  precision         2.8724     2.8724

TANCPU

  mean               1.433     3.1874

  std. dev.        16.8744    11.6524

  weight sum          3420        229

  precision         2.4827     2.4827

Time taken to build model: 0.03 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances        3171               86.9005 %

Incorrectly Classified Instances       478               13.0995 %

Kappa statistic                          0.2042

Mean absolute error                      0.1308

Root mean squared error                  0.3571

Relative absolute error                110.9436 %

Root relative squared error            147.2415 %

Total Number of Instances             3649

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class

                 0.901    0.616    0.956      0.901    0.928      0.216    0.784     0.979     1

                 0.384    0.099    0.207      0.384    0.269      0.216    0.783     0.185     2

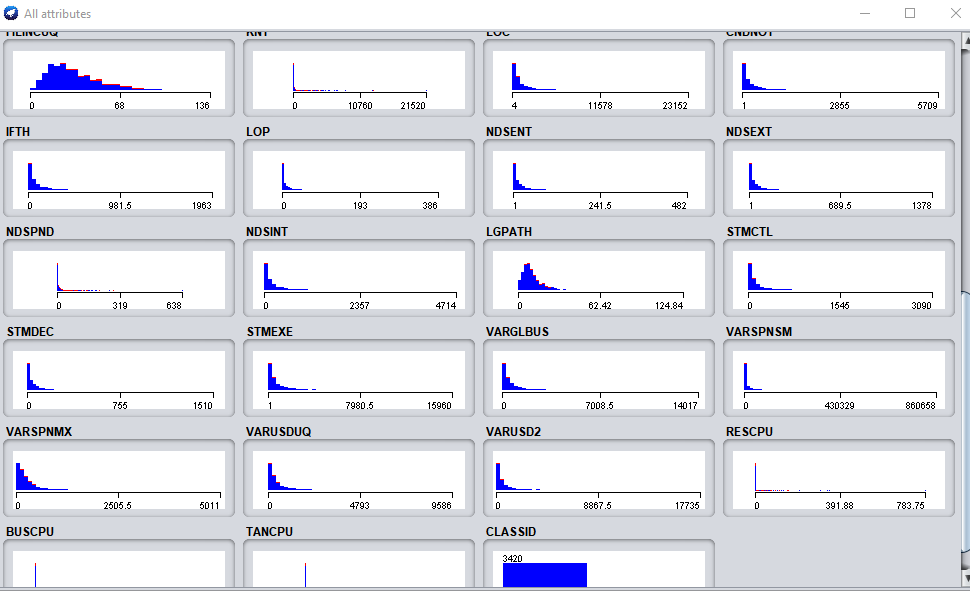
Weighted Avg.    0.869    0.583    0.909      0.869    0.887      0.216    0.784     0.930

=== Confusion Matrix ===

    a    b   <-- classified as

 3083  337 |    a = 1

  141   88 |    b = 2



**28 Dataset**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Datasets | Naïve Bayes | MLP | KNN | Random Forest | LR |
| Release 1 | 477  Correctly classified instances | 3298 | 3298 | 93% | 0.9 |
| Release 2 | 3527 | 3714 | 3156 | 98% | 0.9 |
| Release 3 | 3212 | 3311 | 3578 | 88% | 0.6 |
| Release 4 | 3569 | 3462 | 3211 | 87% | 0.8 |

**42 Dataset**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Datasets | Naïve Bayes | MLP | KNN | Random Forest | LR |
| Release 1 | 3372 | 3311 | 3311 | 93% | 0.8 |
| Release 2 | 3526 | 3117 | 3116 | 90% | 0.7 |
| Release 3 | 3211 | 3321 | 3512 | 88% | 0.8 |
| Release 4 | 3462 | 3462 | 3361 | 94% | 0.77 |

Wheareas across the datasets measured and classified, dataset 42 had more record count than 28 dataset and upon application of the 5 model algorithms could have been seen to be more accurate. The theory is true because the larger the population the, the easier it is to get a closer matching prediction. Logistic regression model can be use used to get the moving direction of the dataset and as two whether a change in one variable would ultimately affected the other. This is perfect for linear variables that have some movement in direction. MLP on the other hand is highly effective in determining natural language patterns as it tends to detect matching patterns and group them together in that match.

Random forest effectively gave the percentage chance of classifying instance to their respective branches, with branches that are almost close to each other getting a random forest match almost equal to 100%. Finally, Naïve Bayes consequently, grouped the datasets that were supplied to it and tried to give out the corresponding closest match of these instances by classifying them and matching them. How this can be interpreted is that suppose the Naïve Theorem gave a correct classification of say 3372, its means that the subsets under this section were closed predicted to match the original target class, and this percentage seems higher.

In conclusion we can say that the five models of classification and predictive analytics all proved worthwhile in this experiment, with each giving a particular result consistent with the research questions in study, however, special emphasis should be made on the fact that each model is suited for its own purpose and would work best on the correct dataset chosen.