CS699 – Term Project

Gaurav Tungare

*Jun 21, 2021*

1.Objective

To give students an opportunity to build and test a

classifier model using a real-world data. The data to be used is a part of the 2018 BRFSS Survey Data prepared by CDC.

2. Background of data

The project-2018-BRFSS-arthritis.arff file has the dataset for the project and it has 11933 tuples and 108 attributes. Each tuple is a person who participated in the survey and each. attribute is an answer to a survey question. The class attribute is havarth3 and its value is

either 1 or 2. The value of 1 means that the person was ever told to have some form of arthritis, rheumatoid arthritis, gout, lupus, or fibromyalgia.

* **Total Record Count – 11933 tuples w/ 108 attributes**

After Train and Test split between dataset

* Total **“Train”** Record Count **- 7997**
* Total **“Test”** Record Count **- 3936**

2. Preprocessing

Preprocessing of data was done using weka on filtered data set for Numeric to Numeric conversion but multiple runs were made using both preprocessed and as it data , the preprocessed data was not giving any gain in the classification so the final run was made without the preprocessed data

Below is the snapshot of the processing done but please note the final model was not applied any preprocessing

Graphical user interface, text, application, email

Description automatically generated

3. Attribute selection method

* **The following attribute selection method is used for the project**

**\*\*\*\* Please click on the attribute selection method to get the details**

* [**GainRationAtributeEval**](#_Attribute_Selection_:)
* [**InfoGainAtributeEval**](#_Attribute_Selection_:_1)
* [**ReliefAtributeEval**](#_Attribute_Selection_:_2)
* [**SymmetricalUncertAttributeEval**](#_Attribute_Selection_:_3)

The attribute selection for each attribute is described in detail for each section. Please click on the link

4. Names of all classifier algorithms used in the **project**

For each attribute selection method below classification algorithm were run

* Naïve Bayes
* Logistic
* OneR
* Decision Tree
* RandomForest
* KNN (K=10)
* KNN (K=12)
* KNN (K=15)

5. Test results of all 24 models

**\*\*\* Please click on the link in column 2 to view detail result**

|  |  |  |  |
| --- | --- | --- | --- |
| **Test Dataset** | | | |
| Sr No | Attrubute Selection | Classification | Correctly Classified Instances |
| 1 | [GainRationAtributeEval](#_1.Using_Naïve_Bayes) | Naïve Bayes | 71.37% |
| 2 | [GainRationAtributeEval](#_2.Using_Logistics) | Logistic | 75.15% |
| 3 | [GainRationAtributeEval](#_3.Using_OneR) | OneR | 73.91% |
| 4 | [GainRationAtributeEval](#_4.Using_Decision_Tree) | Decision Tree | 72.26% |
| 5 | [GainRationAtributeEval](#_5.Using_Random_Forest) | RandomForest | 75.94% |
| 6 | [GainRationAtributeEval](#_6.Using_KNN_() | KNN (K=10) | 72.33% |
| 6A | [GainRationAtributeEval](#_6A.Using_KNN_() | KNN (K=12) | 72.54% |
| 6B | [GainRationAtributeEval](#_6B.Using_KNN_() | KNN (K=15) | 73.12% |
|  |  |  |  |
| 7 | [InfoGainAtributeEval](#_7.Using_Naïve_Bayes) | Naïve Bayes | 71.39% |
| 8 | [InfoGainAtributeEval](#_8.Using_Logistic) | Logistic | 74.82% |
| 9 | [InfoGainAtributeEval](#_9.Using_OneR) | OneR | 73.91% |
| 10 | [InfoGainAtributeEval](#_10.Using_Decision_Tree) | Decision Tree | 72.87% |
| 11 | [InfoGainAtributeEval](#_11.Using_RandomForest) | RandomForest | 75.64% |
| 12 | [InfoGainAtributeEval](#_12.Using_KNN(K=10)) | KNN (K=10) | 71.65% |
| 12A | [InfoGainAtributeEval](#_12A.Using_KNN(K=12)) | KNN (K=12) | 72.10% |
| 12B | [InfoGainAtributeEval](#_12B.Using_KNN(K=15)) | KNN (K=15) | 73.07% |
|  |  |  |  |
| 13 | [ReliefAtributeEval](#_13.Using_Naïve_Bayes) | Naïve Bayes | 71.98% |
| 14 | [ReliefAtributeEval](#_14.Using_Logistic) | Logistic | 74.44% |
| 15 | [ReliefAtributeEval](#_15.Using_OneR) | OneR | 73.91% |
| 16 | [ReliefAtributeEval](#_16.Using_Decision_Tree) | Decision Tree | 73.07% |
| 17 | [ReliefAtributeEval](#_11.Using_RandomForest) | RandomForest | 75.03% |
| 18 | [ReliefAtributeEval](#_18.Using_KNN(K=10)) | KNN (K=10) | 71.75% |
| 18A | [ReliefAtributeEval](#_18A.Using_KNN(K=12)) | KNN (K=12) | 72.31% |
| 18B | [ReliefAtributeEval](#_18B.Using_KNN(K=15)) | KNN (K=15) | 73.45% |
|  |  |  |  |
| 19 | [SymmetricalUncertAttributeEval](#_19.Using_Naïve_Bayes) | Naïve Bayes | 71.09% |
| 20 | [SymmetricalUncertAttributeEval](#_20.Using_Logistic) | Logistic | 74.70% |
| 21 | [SymmetricalUncertAttributeEval](#_21.Using_OneR) | OneR | 73.91% |
| 22 | [SymmetricalUncertAttributeEval](#_22.Using_Decision_Tree) | Decision Tree | 73.22% |
| 23 | [SymmetricalUncertAttributeEval](#_23.Using_RandomForest) | RandomForest | 75.28% |
| 24 | [SymmetricalUncertAttributeEval](#_24.Using_KNN(K=10)) | KNN (K=10) | 71.82% |
| 24A | [SymmetricalUncertAttributeEval](#_24A.Using_KNN(K=12)) | KNN (K=12) | 72.64% |
| 24B | [SymmetricalUncertAttributeEval](#_24B.Using_KNN(K=15)) | KNN (K=15) | 73.07% |

6. Best Model

The best model after 24 iteration is using **GainRatio** Attribute selection with **RandomForest** classification gave the best performance across

**=== Summary ===**

|  |  |  |
| --- | --- | --- |
| Correctly Classified Instances | 2989 | **75.94%** |
| Incorrectly Classified Instances | 947 | 24.06% |
| Kappa statistic | 0.4257 |  |
| Mean absolute error | 0.3435 |  |
| Root mean squared error | 0.4073 |  |
| Relative absolute error | 76.82% |  |
| Root relative squared error | 86.15% |  |
| Total Number of Instances | 3936 |  |

**=== Detailed Accuracy By Class ===**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC | PRC Area | Class |
|  | 0.884 | 0.485 | 0.782 | 0.884 | 0.83 | 0.435 | 0.808 | 0.885 | 2 |
|  | 0.515 | 0.116 | 0.693 | 0.515 | 0.591 | 0.435 | 0.808 | 0.674 | 1 |
| Weighted Avg. | **0.759** | **0.361** | 0.752 | 0.759 | **0.749** | 0.435 | **0.808** | 0.814 |  |

**=== Confusion Matrix ===**

|  |  |  |
| --- | --- | --- |
| A | b | <-- classified as |
| 2305 | 303 | a = 2 |
| 644 | 684 | b = 1 |

Justification

The above model has **TP rate** of 0.759 which indicate that the model is able to correctly predict the value of some form of arthritis, rheumatoid arthritis, gout, lupus, or fibromyalgia, TP rate ~0.80 indicate a good performing model

The **ROC** is 0.80 which indicate that the model is performing very well on the test data set, any ROC value >=0.80 indicate a good performing model

The **FP rate** is 0.36 which indicate the class attribute which were falsely classified as Y as with every model we have FP rate but since the FP rate is low it is indication of a good performing model

The **F-measure** of 0.749 indicate good sensitivity and recall for over all model, the close the F-measure to 1 the prediction across test data set improves, again this measure gives us confidence that the class attribute which the model is predicting is good

7. Five attributes most relevant to the class attribute

The five attribute which are most relevant to class attribute which will help predict the correct value for some form of arthritis, rheumatoid arthritis, gout, lupus, or fibromyalgia

1. Variable Name: **CHCCOPD1** - you have chronic obstructive pulmonary disease, C.O.P.D., emphysema or chronic bronchitis?

This column has high Gain ratio of 0.0650295, this indicates a chronic health condition , this wil drive the prediction significantly

1. Variable Name: **DIFFWALK** - Difficulty Walking or Climbing Stairs

This column has high Gain ratio of 0.1132304 , this indicates overall health of the person predicting the overall well being. If the person has answered yes to the question it would mean underlying health condition

1. Variable Name: \_**RFHLTH** Adults with good or better health

This column has high Gain ratio of 0.0564898 , this indicates overall health of adults with good or better health, which will drive the predicting if the person has some form of issues with health or not

1. Variable Name: \_ **AGE65YR** Reported age in two age groups calculated variable

This column has high Gain ratio of 0.053772, with three categories of age 18 to 64 or Age 65 or older , this is important column as it will drive other columns to predict the health attribute correctly

1. Variable Name: \_ **EMPLOY1** Employment Status

Surprisingly, this column has high Gain ratio of 0.0462859, which indicates relation between if person is employed and presence/absence of disease. You employment status does drive overall health of the person which in turn helps predict the if certain disease if present/absent

8. Lessons learned from this project.

There are various aspect of data mining which I have learnt from this project.

Let me summarize on each of the stuff

From Data cleaning

Before performing any data analysis, the available data has to be cleaned to remove as many errors as possible. The goal is to convert the available data into high quality but in our case data cleaning has not much effect on the classification

From Attribute Selection

Got various attribute selection methodology, Information Gain and Gain Ratio, using various selection works and how different information gain is calculated help in selection best attribute selection over attribute which will give best classification measure and avoids noise as much as possible

Classification

Learnt various aspect of classification on various classification algorithms

Naïve Bayes, Logistic, Decision Tree, OneR, Randomforest & KNN.

On the test dataset the model performance is not varying a lot , the deviation across each model is +/- 3% which indicate that you need to be very careful with attribute selection method

Model Performance

I learnt about general prediction ability of a model all 24 models were analyzed ,various accuracy, sensitivity, specificity, precision, F measure (or F-score). These different measures represent different aspects of the performance of a model which was used to predict some form of

arthritis, rheumatoid arthritis, gout, lupus, or fibromyalgia.

**-----------------------------------------------------------------**

**Section I**

Attribute Selection : GainRationAtributeEval

The following attribute selection method was a

**GainRationAtributeEval**



Any gain ration below **0.004285 where eliminated**, the following columns were eliminated

|  |  |  |
| --- | --- | --- |
| 0.002927 | 28 | lastden4 |
| 0.002787 | 70 | x.educag |
| 0.002735 | 12 | educa |
| 0.002713 | 7 | veteran3 |
| 0.002523 | 56 | x.metstat |
| 0.002183 | 42 | medcost |
| 0.001381 | 23 | usenow3 |
| 0.001381 | 79 | x.rfsmok3 |
| 0.001194 | 18 | hivtst6 |
| 0.00119 | 99 | x.denvst3 |
| 0.001169 | 1 | x.aidtst3 |
| 0.001127 | 73 | x.rfseat2 |
| 0.001117 | 72 | x.rfdrhv6 |
| 0.000773 | 40 | iday |
| 0.000716 | 74 | x.rfseat3 |
| 0.000517 | 17 | seatbelt |
| 0.000218 | 39 | imonth |
| 0.000187 | 83 | x.urbstat |
| 2.79E-05 | 37 | iyear |
| 0 | 38 | Fmonth |
| 0 | 92 | x.ststr |
| 0 | 107 | x.state |

*1.Using Naïve Bayes*

Naïve Bayes

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.26 seconds

=== Summary ===

Correctly Classified Instances 2809 71.3669 %

Incorrectly Classified Instances 1127 28.6331 %

Kappa statistic 0.4091

Mean absolute error 0.291

Root mean squared error 0.5139

Relative absolute error 65.0852 %

Root relative squared error 108.683 %

Total Number of Instances 3936

=== Detailed Accuracy By Class ===

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

0.697 0.254 0.844 0.697 0.763 0.421 0.782 0.861 2

0.746 0.303 0.556 0.746 0.638 0.421 0.782 0.604 1

Weighted Avg. 0.714 0.270 0.747 0.714 0.721 0.421 0.782 0.775

=== Confusion Matrix ===

a b <-- classified as

1818 790 | a = 2

1. 91 | b = 1

Graphical user interface, text, application

Description automatically generated

*2.Using Logistics*

=== Summary ===

Correctly Classified Instances 2958 75.1524 %

Incorrectly Classified Instances 978 24.8476 %

Kappa statistic 0.4128

Mean absolute error 0.3211

Root mean squared error 0.406

Relative absolute error 71.8049 %

Root relative squared error 85.8618 %

Total Number of Instances 3936

=== Detailed Accuracy By Class ===

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

0.868 0.477 0.781 0.868 0.822 0.419 0.809 0.877 2

0.523 0.132 0.669 0.523 0.587 0.419 0.809 0.656 1

Weighted Avg. 0.752 0.361 0.743 0.752 0.743 0.419 0.809 0.803

=== Confusion Matrix ===

a b <-- classified as

2264 344 | a = 2

1. 694 | b = 1

Graphical user interface, text, application

Description automatically generated

*3.Using OneR*

=== Summary ===

Correctly Classified Instances 2909 73.9075 %

Incorrectly Classified Instances 1027 26.0925 %

Kappa statistic 0.3316

Mean absolute error 0.2609

Root mean squared error 0.5108

Relative absolute error 58.3505 %

Root relative squared error 108.0339 %

Total Number of Instances 3936

=== Detailed Accuracy By Class ===

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

0.935 0.645 0.740 0.935 0.826 0.371 0.645 0.735 2

0.355 0.065 0.735 0.355 0.478 0.371 0.645 0.478 1

Weighted Avg. 0.739 0.450 0.738 0.739 0.709 0.371 0.645 0.648

=== Confusion Matrix ===

a b <-- classified as

2438 170 | a = 2

1. 71 | b = 1

Graphical user interface, text, application

Description automatically generated

*4.Using Decision Tree*

=== Summary ===

Correctly Classified Instances 2844 72.2561 %

Incorrectly Classified Instances 1092 27.7439 %

Kappa statistic 0.3456

Mean absolute error 0.335

Root mean squared error 0.4783

Relative absolute error 74.9194 %

Root relative squared error 101.1545 %

Total Number of Instances 3936

=== Detailed Accuracy By Class ===

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

0.844 0.517 0.762 0.844 0.801 0.351 0.665 0.758 2

0.483 0.156 0.613 0.483 0.540 0.351 0.665 0.497 1

Weighted Avg. 0.723 0.395 0.712 0.723 0.713 0.351 0.665 0.670

=== Confusion Matrix ===

a b <-- classified as

2202 406 | a = 2

1. 42 | b = 1

Graphical user interface, text

Description automatically generated

*5.Using Random Forest*

=== Summary ===

Correctly Classified Instances 2989 75.94 %

Incorrectly Classified Instances 947 24.06 %

Kappa statistic 0.4257

Mean absolute error 0.3435

Root mean squared error 0.4073

Relative absolute error 76.8167 %

Root relative squared error 86.1491 %

Total Number of Instances 3936

=== Detailed Accuracy By Class ===

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

0.884 0.485 0.782 0.884 0.830 0.435 0.808 0.885 2

0.515 0.116 0.693 0.515 0.591 0.435 0.808 0.674 1

Weighted Avg. 0.759 0.361 0.752 0.759 0.749 0.435 0.808 0.814

=== Confusion Matrix ===

a b <-- classified as

2305 303 | a = 2

644 684 | b = 1

Graphical user interface, text, application

Description automatically generated

*6.Using KNN ( N=10)*

=== Summary ===

Correctly Classified Instances 2847 72.3323 %

Incorrectly Classified Instances 1089 27.6677 %

Kappa statistic 0.2954

Mean absolute error 0.3356

Root mean squared error 0.4261

Relative absolute error 75.0611 %

Root relative squared error 90.1136 %

Total Number of Instances 3936

=== Detailed Accuracy By Class ===

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

0.918 0.658 0.732 0.918 0.815 0.327 0.764 0.835 2

0.342 0.082 0.679 0.342 0.455 0.327 0.764 0.593 1

Weighted Avg. 0.723 0.464 0.714 0.723 0.693 0.327 0.764 0.753

=== Confusion Matrix ===

a b <-- classified as

2393 215 | a = 2

Graphical user interface, text, application

Description automatically generated

*6A.Using KNN ( N=12)*

=== Summary ===

Correctly Classified Instances 2855 72.5356 %

Incorrectly Classified Instances 1081 27.4644 %

Kappa statistic 0.3012

Mean absolute error 0.3362

Root mean squared error 0.4225

Relative absolute error 75.1784 %

Root relative squared error 89.3541 %

Total Number of Instances 3936

=== Detailed Accuracy By Class ===

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

0.918 0.654 0.734 0.918 0.816 0.332 0.773 0.843 2

0.346 0.082 0.684 0.346 0.460 0.332 0.773 0.604 1

Weighted Avg. 0.725 0.461 0.717 0.725 0.696 0.332 0.773 0.762

=== Confusion Matrix ===

a b <-- classified as

2395 213 | a = 2

868 460 | b = 1

Graphical user interface, text, application

Description automatically generated

*6B.Using KNN ( N=15)*

=== Summary ===

=== Summary ===

Correctly Classified Instances 2878 73.1199 %

Incorrectly Classified Instances 1058 26.8801 %

Kappa statistic 0.3352

Mean absolute error 0.336

Root mean squared error 0.4187

Relative absolute error 75.1385 %

Root relative squared error 88.5575 %

Total Number of Instances 3936

=== Detailed Accuracy By Class ===

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

0.896 0.593 0.748 0.896 0.815 0.354 0.782 0.853 2

0.407 0.104 0.667 0.407 0.505 0.354 0.782 0.621 1

Weighted Avg. 0.731 0.428 0.721 0.731 0.711 0.354 0.782 0.774

=== Confusion Matrix ===

a b <-- classified as

2338 270 | a = 2

788 540 | b = 1

Graphical user interface, text, application

Description automatically generated

**Section II**

Attribute Selection : InfoGainAtributeEval

The following attribute selection method was use **InfoGainAtributeEval**



**Any info below 0.003072 where eliminated**, the following columns were eliminated

|  |  |  |
| --- | --- | --- |
| 0.002227 | 56 | x.metstat |
| 0.001566 | 7 | veteran3 |
| 0.001375 | 18 | hivtst6 |
| 0.001327 | 1 | x.aidtst3 |
| 0.001167 | 99 | x.denvst3 |
| 0.001108 | 79 | x.rfsmok3 |
| 0.001092 | 42 | medcost |
| 0.000779 | 39 | imonth |
| 0.000719 | 72 | x.rfdrhv6 |
| 0.000647 | 73 | x.rfseat2 |
| 0.000581 | 74 | x.rfseat3 |
| 0.000415 | 17 | seatbelt |
| 0.000362 | 23 | usenow3 |
| 0.000115 | 83 | x.urbstat |
| 7.35E-06 | 37 | iyear |
| 0 | 92 | x.ststr |
| 0 | 38 | fmonth |
| 0 | 107 | x.state |

*7.Using Naïve Bayes*

=== Summary ===

=== Summary ===

Correctly Classified Instances 2810 71.3923 %

Incorrectly Classified Instances 1126 28.6077 %

Kappa statistic 0.4081

Mean absolute error 0.2905

Root mean squared error 0.5134

Relative absolute error 64.9655 %

Root relative squared error 108.5876 %

Total Number of Instances 3936

=== Detailed Accuracy By Class ===

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

0.700 0.259 0.841 0.700 0.764 0.419 0.782 0.862 2

0.741 0.300 0.557 0.741 0.636 0.419 0.782 0.604 1

Weighted Avg. 0.714 0.273 0.746 0.714 0.721 0.419 0.782 0.775

=== Confusion Matrix ===

a b <-- classified as

1826 782 | a = 2

344 984 | b = 1

Graphical user interface, text, application

Description automatically generated

*8.Using Logistic*

=== Summary ===

=== Summary ===

Correctly Classified Instances 2945 74.8222 %

Incorrectly Classified Instances 991 25.1778 %

Kappa statistic 0.4065

Mean absolute error 0.3213

Root mean squared error 0.4084

Relative absolute error 71.8524 %

Root relative squared error 86.3826 %

Total Number of Instances 3936

=== Detailed Accuracy By Class ===

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

0.863 0.477 0.780 0.863 0.820 0.412 0.804 0.874 2

0.523 0.137 0.660 0.523 0.583 0.412 0.804 0.651 1

Weighted Avg. 0.748 0.363 0.740 0.748 0.740 0.412 0.804 0.799

=== Confusion Matrix ===

a b <-- classified as

2251 357 | a = 2

634 694 | b = 1

Graphical user interface, text, application

Description automatically generated

*9.Using OneR*

=== Summary ===

=== Summary ===

Correctly Classified Instances 2909 73.9075 %

Incorrectly Classified Instances 1027 26.0925 %

Kappa statistic 0.3316

Mean absolute error 0.2609

Root mean squared error 0.5108

Relative absolute error 58.3505 %

Root relative squared error 108.0339 %

Total Number of Instances 3936

=== Detailed Accuracy By Class ===

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

0.935 0.645 0.740 0.935 0.826 0.371 0.645 0.735 2

0.355 0.065 0.735 0.355 0.478 0.371 0.645 0.478 1

Weighted Avg. 0.739 0.450 0.738 0.739 0.709 0.371 0.645 0.648

=== Confusion Matrix ===

a b <-- classified as

2438 170 | a = 2

857 471 | b = 1

Graphical user interface, text, application

Description automatically generated

*10.Using Decision Tree*

=== Summary ===

Correctly Classified Instances 2868 72.8659 %

Incorrectly Classified Instances 1068 27.1341 %

Kappa statistic 0.3798

Mean absolute error 0.338

Root mean squared error 0.4547

Relative absolute error 75.5937 %

Root relative squared error 96.1736 %

Total Number of Instances 3936

=== Detailed Accuracy By Class ===

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

0.817 0.446 0.783 0.817 0.800 0.381 0.712 0.791 2

0.554 0.183 0.607 0.554 0.580 0.381 0.712 0.536 1

Weighted Avg. 0.729 0.357 0.723 0.729 0.725 0.381 0.712 0.705

=== Confusion Matrix ===

a b <-- classified as

2132 476 | a = 2

592 736 | b = 1

Graphical user interface, text, application

Description automatically generated

*11.Using RandomForest*

=== Summary ===

Correctly Classified Instances 2977 75.6352 %

Incorrectly Classified Instances 959 24.3648 %

Kappa statistic 0.4119

Mean absolute error 0.3496

Root mean squared error 0.408

Relative absolute error 78.1777 %

Root relative squared error 86.2901 %

Total Number of Instances 3936

=== Detailed Accuracy By Class ===

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

0.892 0.511 0.774 0.892 0.829 0.425 0.809 0.885 2

0.489 0.108 0.698 0.489 0.575 0.425 0.809 0.680 1

Weighted Avg. 0.756 0.375 0.749 0.756 0.744 0.425 0.809 0.816

=== Confusion Matrix ===

a b <-- classified as

2327 281 | a = 2

678 650 | b = 1

Graphical user interface, text, application

Description automatically generated

*12.Using KNN(K=10)*

=== Summary ===

Correctly Classified Instances 2820 71.6463 %

Incorrectly Classified Instances 1116 28.3537 %

Kappa statistic 0.2735

Mean absolute error 0.3395

Root mean squared error 0.4288

Relative absolute error 75.9304 %

Root relative squared error 90.6812 %

Total Number of Instances 3936

=== Detailed Accuracy By Class ===

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

0.918 0.679 0.726 0.918 0.811 0.306 0.759 0.832 2

0.321 0.082 0.666 0.321 0.433 0.306 0.759 0.582 1

Weighted Avg. 0.716 0.478 0.706 0.716 0.683 0.306 0.759 0.747

=== Confusion Matrix ===

a b <-- classified as

2394 214 | a = 2

902 426 | b = 1

Graphical user interface, text, application

Description automatically generated

*12A.Using KNN(K=12)*

=== Summary ===

Correctly Classified Instances 2838 72.1037 %

Incorrectly Classified Instances 1098 27.8963 %

Kappa statistic 0.2861

Mean absolute error 0.3387

Root mean squared error 0.4251

Relative absolute error 75.7371 %

Root relative squared error 89.9027 %

Total Number of Instances 3936

=== Detailed Accuracy By Class ===

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

0.920 0.670 0.729 0.920 0.814 0.319 0.768 0.840 2

0.330 0.080 0.678 0.330 0.444 0.319 0.768 0.597 1

Weighted Avg. 0.721 0.471 0.712 0.721 0.689 0.319 0.768 0.758

=== Confusion Matrix ===

a b <-- classified as

2400 208 | a = 2

890 438 | b = 1

Graphical user interface, text, application

Description automatically generated

*12B.Using KNN(K=15)*

=== Summary ===

Correctly Classified Instances 2876 73.0691 %

Incorrectly Classified Instances 1060 26.9309 %

Kappa statistic 0.3282

Mean absolute error 0.339

Root mean squared error 0.4217

Relative absolute error 75.8165 %

Root relative squared error 89.1813 %

Total Number of Instances 3936

=== Detailed Accuracy By Class ===

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

0.904 0.610 0.744 0.904 0.816 0.351 0.776 0.847 2

0.390 0.096 0.674 0.390 0.494 0.351 0.776 0.611 1

Weighted Avg. 0.731 0.436 0.721 0.731 0.708 0.351 0.776 0.768

=== Confusion Matrix ===

a b <-- classified as

2358 250 | a = 2

810 518 | b = 1

Graphical user interface, text, application

Description automatically generated

Section III

Attribute Selection : ReliefAtributeEval

The following attribute selection method was use **ReliefAtributeEval**



**Any info below 0.0000821** where eliminated, the following columns were eliminated

|  |  |  |
| --- | --- | --- |
| -0.0003256 | 85 | x.llcpwt2 |
| -0.000337 | 5 | height3 |
| -0.0003978 | 86 | x.llcpwt |
| -0.0004491 | 93 | x.strwt |
| -0.000462 | 61 | htm4 |
| -0.0005376 | 91 | x.wt2rake |
| -0.0015873 | 57 | htin4 |
| -0.0016971 | 43 | checkup1 |
| -0.0018695 | 19 | hivrisk5 |

*13.Using Naïve Bayes*

=== Summary ===

=== Summary ===

Correctly Classified Instances 2833 71.9766 %

Incorrectly Classified Instances 1103 28.0234 %

Kappa statistic 0.4114

Mean absolute error 0.2824

Root mean squared error 0.5021

Relative absolute error 63.1526 %

Root relative squared error 106.1972 %

Total Number of Instances 3936

=== Detailed Accuracy By Class ===

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

0.721 0.283 0.833 0.721 0.773 0.419 0.788 0.873 2

0.717 0.279 0.567 0.717 0.633 0.419 0.788 0.606 1

Weighted Avg. 0.720 0.282 0.744 0.720 0.726 0.419 0.788 0.783

=== Confusion Matrix ===

a b <-- classified as

1881 727 | a = 2

376 952 | b = 1

Graphical user interface, text, application

Description automatically generated

*14.Using Logistic*

=== Summary ===

=== Summary ===

Correctly Classified Instances 2930 74.4411 %

Incorrectly Classified Instances 1006 25.5589 %

Kappa statistic 0.3972

Mean absolute error 0.3221

Root mean squared error 0.4097

Relative absolute error 72.0367 %

Root relative squared error 86.6492 %

Total Number of Instances 3936

=== Detailed Accuracy By Class ===

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

0.861 0.484 0.777 0.861 0.817 0.403 0.802 0.871 2

0.516 0.139 0.654 0.516 0.577 0.403 0.802 0.651 1

Weighted Avg. 0.744 0.368 0.736 0.744 0.736 0.403 0.802 0.797

=== Confusion Matrix ===

a b <-- classified as

2245 363 | a = 2

643 685 | b = 1

Graphical user interface, text, application

Description automatically generated

*15.Using OneR*

=== Summary ===

Correctly Classified Instances 2909 73.9075 %

Incorrectly Classified Instances 1027 26.0925 %

Kappa statistic 0.3316

Mean absolute error 0.2609

Root mean squared error 0.5108

Relative absolute error 58.3505 %

Root relative squared error 108.0339 %

Total Number of Instances 3936

=== Detailed Accuracy By Class ===

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

0.935 0.645 0.740 0.935 0.826 0.371 0.645 0.735 2

0.355 0.065 0.735 0.355 0.478 0.371 0.645 0.478 1

Weighted Avg. 0.739 0.450 0.738 0.739 0.709 0.371 0.645 0.648

=== Confusion Matrix ===

a b <-- classified as

2438 170 | a = 2

857 471 | b = 1

Graphical user interface, text, application

Description automatically generated

*16.Using Decision Tree*

=== Summary ===

=== Summary ===

Correctly Classified Instances 2876 73.0691 %

Incorrectly Classified Instances 1060 26.9309 %

Kappa statistic 0.3793

Mean absolute error 0.3361

Root mean squared error 0.4581

Relative absolute error 75.1595 %

Root relative squared error 96.8866 %

Total Number of Instances 3936

=== Detailed Accuracy By Class ===

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

0.827 0.459 0.780 0.827 0.803 0.381 0.713 0.795 2

0.541 0.173 0.615 0.541 0.575 0.381 0.713 0.532 1

Weighted Avg. 0.731 0.363 0.724 0.731 0.726 0.381 0.713 0.706

=== Confusion Matrix ===

a b <-- classified as

2158 450 | a = 2

610 718 | b = 1

Graphical user interface, text, application

Description automatically generated

*17.Using RandomForest*

=== Summary ===

Correctly Classified Instances 2953 75.0254 %

Incorrectly Classified Instances 983 24.9746 %

Kappa statistic 0.3948

Mean absolute error 0.3517

Root mean squared error 0.4092

Relative absolute error 78.6547 %

Root relative squared error 86.5504 %

Total Number of Instances 3936

=== Detailed Accuracy By Class ===

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

0.891 0.527 0.769 0.891 0.825 0.408 0.807 0.884 2

0.473 0.109 0.689 0.473 0.561 0.408 0.807 0.677 1

Weighted Avg. 0.750 0.386 0.742 0.750 0.736 0.408 0.807 0.814

=== Confusion Matrix ===

a b <-- classified as

2325 283 | a = 2

700 628 | b = 1

Graphical user interface, text, application

Description automatically generated

*18.Using KNN(K=10)*

=== Summary ===

Correctly Classified Instances 2824 71.748 %

Incorrectly Classified Instances 1112 28.252 %

Kappa statistic 0.2709

Mean absolute error 0.3379

Root mean squared error 0.4301

Relative absolute error 75.5725 %

Root relative squared error 90.9671 %

Total Number of Instances 3936

=== Detailed Accuracy By Class ===

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

0.925 0.691 0.725 0.925 0.813 0.307 0.758 0.831 2

0.309 0.075 0.678 0.309 0.425 0.307 0.758 0.581 1

Weighted Avg. 0.717 0.483 0.709 0.717 0.682 0.307 0.758 0.747

=== Confusion Matrix ===

a b <-- classified as

2413 195 | a = 2

917 411 | b = 1

Graphical user interface, text, application

Description automatically generated

*18A.Using KNN(K=12)*

=== Summary ===

Correctly Classified Instances 2846 72.3069 %

Incorrectly Classified Instances 1090 27.6931 %

Kappa statistic 0.2841

Mean absolute error 0.3392

Root mean squared error 0.4281

Relative absolute error 75.8649 %

Root relative squared error 90.5347 %

Total Number of Instances 3936

=== Detailed Accuracy By Class ===

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

0.931 0.685 0.727 0.931 0.817 0.324 0.763 0.839 2

0.315 0.069 0.699 0.315 0.434 0.324 0.763 0.588 1

Weighted Avg. 0.723 0.477 0.718 0.723 0.688 0.324 0.763 0.754

=== Confusion Matrix ===

a b <-- classified as

2428 180 | a = 2

910 418 | b = 1

Graphical user interface, text, application

Description automatically generated

*18B.Using KNN(K=15)*

=== Summary ===

=== Summary ===

Correctly Classified Instances 2891 73.4502 %

Incorrectly Classified Instances 1045 26.5498 %

Kappa statistic 0.329

Mean absolute error 0.3387

Root mean squared error 0.4237

Relative absolute error 75.7407 %

Root relative squared error 89.6174 %

Total Number of Instances 3936

=== Detailed Accuracy By Class ===

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

0.919 0.628 0.742 0.919 0.821 0.359 0.774 0.846 2

0.372 0.081 0.701 0.372 0.486 0.359 0.774 0.609 1

Weighted Avg. 0.735 0.443 0.728 0.735 0.708 0.359 0.774 0.766

=== Confusion Matrix ===

a b <-- classified as

2397 211 | a = 2

834 494 | b = 1

Graphical user interface, text, application

Description automatically generated

Section IV

Attribute Selection : SymmetricalUncertAttributeEval

The following attribute selection method was use **SymmetricalUncertAttributeEval**



Any info not equal 0 where eliminated, the following columns were eliminated

|  |  |  |
| --- | --- | --- |
| 0 | 92 | x.ststr |
| 0 | 38 | fmonth |
| 0 | 107 | x.state |

*19.Using Naïve Bayes*

=== Summary ===

Correctly Classified Instances 2798 71.0874 %

Incorrectly Classified Instances 1138 28.9126 %

Kappa statistic 0.4026

Mean absolute error 0.2914

Root mean squared error 0.5146

Relative absolute error 65.1673 %

Root relative squared error 108.8399 %

Total Number of Instances 3936

=== Detailed Accuracy By Class ===

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

0.696 0.261 0.840 0.696 0.761 0.414 0.780 0.860 2

0.739 0.304 0.554 0.739 0.633 0.414 0.780 0.602 1

Weighted Avg. 0.711 0.275 0.743 0.711 0.718 0.414 0.780 0.773

=== Confusion Matrix ===

a b <-- classified as

1816 792 | a = 2

346 982 | b = 1

Graphical user interface, text, application

Description automatically generated

*20.Using Logistic*

=== Summary ===

Correctly Classified Instances 2940 74.6951 %

Incorrectly Classified Instances 996 25.3049 %

Kappa statistic 0.4034

Mean absolute error 0.3208

Root mean squared error 0.4089

Relative absolute error 71.7397 %

Root relative squared error 86.475 %

Total Number of Instances 3936

=== Detailed Accuracy By Class ===

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

0.862 0.480 0.779 0.862 0.819 0.409 0.804 0.872 2

0.520 0.138 0.658 0.520 0.581 0.409 0.804 0.651 1

Weighted Avg. 0.747 0.364 0.738 0.747 0.739 0.409 0.804 0.798

=== Confusion Matrix ===

a b <-- classified as

2249 359 | a = 2

637 691 | b = 1

Graphical user interface, text, application

Description automatically generated

*21.Using OneR*

=== Summary ===

Correctly Classified Instances 2909 73.9075 %

Incorrectly Classified Instances 1027 26.0925 %

Kappa statistic 0.3316

Mean absolute error 0.2609

Root mean squared error 0.5108

Relative absolute error 58.3505 %

Root relative squared error 108.0339 %

Total Number of Instances 3936

=== Detailed Accuracy By Class ===

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

0.935 0.645 0.740 0.935 0.826 0.371 0.645 0.735 2

0.355 0.065 0.735 0.355 0.478 0.371 0.645 0.478 1

Weighted Avg. 0.739 0.450 0.738 0.739 0.709 0.371 0.645 0.648

=== Confusion Matrix ===

a b <-- classified as

2438 170 | a = 2

857 471 | b = 1

Graphical user interface, text, application

Description automatically generated

*22.Using Decision Tree*

=== Summary ===

Correctly Classified Instances 2882 73.2215 %

Incorrectly Classified Instances 1054 26.7785 %

Kappa statistic 0.3793

Mean absolute error 0.3377

Root mean squared error 0.4586

Relative absolute error 75.5304 %

Root relative squared error 96.9867 %

Total Number of Instances 3936

=== Detailed Accuracy By Class ===

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

0.834 0.468 0.778 0.834 0.805 0.382 0.710 0.792 2

0.532 0.166 0.620 0.532 0.573 0.382 0.710 0.527 1

Weighted Avg. 0.732 0.366 0.725 0.732 0.727 0.382 0.710 0.702

=== Confusion Matrix ===

a b <-- classified as

2176 432 | a = 2

622 706 | b = 1

Graphical user interface, text, application

Description automatically generated

*23.Using RandomForest*

=== Summary ===

Correctly Classified Instances 2963 75.2795 %

Incorrectly Classified Instances 973 24.7205 %

Kappa statistic 0.4005

Mean absolute error 0.3495

Root mean squared error 0.4072

Relative absolute error 78.1695 %

Root relative squared error 86.12 %

Total Number of Instances 3936

=== Detailed Accuracy By Class ===

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

0.894 0.525 0.770 0.894 0.827 0.415 0.813 0.890 2

0.475 0.106 0.696 0.475 0.565 0.415 0.813 0.681 1

Weighted Avg. 0.753 0.383 0.745 0.753 0.739 0.415 0.813 0.820

=== Confusion Matrix ===

a b <-- classified as

2332 276 | a = 2

697 631 | b = 1

Graphical user interface, text, application

Description automatically generated

*24.Using KNN(K=10)*

=== Summary ===

Correctly Classified Instances 2827 71.8242 %

Incorrectly Classified Instances 1109 28.1758 %

Kappa statistic 0.2727

Mean absolute error 0.3389

Root mean squared error 0.4285

Relative absolute error 75.7827 %

Root relative squared error 90.636 %

Total Number of Instances 3936

=== Detailed Accuracy By Class ===

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

0.926 0.690 0.725 0.926 0.813 0.310 0.760 0.833 2

0.310 0.074 0.681 0.310 0.426 0.310 0.760 0.583 1

Weighted Avg. 0.718 0.482 0.710 0.718 0.683 0.310 0.760 0.749

=== Confusion Matrix ===

a b <-- classified as

2415 193 | a = 2

916 412 | b = 1

Graphical user interface, text, application

Description automatically generated

*24A.Using KNN(K=12)*

=== Summary ===

Correctly Classified Instances 2859 72.6372 %

Incorrectly Classified Instances 1077 27.3628 %

Kappa statistic 0.2934

Mean absolute error 0.3391

Root mean squared error 0.4255

Relative absolute error 75.827 %

Root relative squared error 89.9917 %

Total Number of Instances 3936

=== Detailed Accuracy By Class ===

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

0.933 0.678 0.730 0.933 0.819 0.334 0.769 0.842 2

0.322 0.067 0.708 0.322 0.442 0.334 0.769 0.594 1

Weighted Avg. 0.726 0.472 0.722 0.726 0.692 0.334 0.769 0.758

=== Confusion Matrix ===

a b <-- classified as

2432 176 | a = 2

901 427 | b = 1

Graphical user interface, text, application

Description automatically generated

*24B.Using KNN(K=15)*

=== Summary ===

Correctly Classified Instances 2876 73.0691 %

Incorrectly Classified Instances 1060 26.9309 %

Kappa statistic 0.3243

Mean absolute error 0.34

Root mean squared error 0.4234

Relative absolute error 76.0286 %

Root relative squared error 89.5503 %

Total Number of Instances 3936

=== Detailed Accuracy By Class ===

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

0.910 0.620 0.742 0.910 0.817 0.350 0.775 0.848 2

0.380 0.090 0.681 0.380 0.487 0.350 0.775 0.603 1

Weighted Avg. 0.731 0.442 0.722 0.731 0.706 0.350 0.775 0.765

=== Confusion Matrix ===

a b <-- classified as

2372 236 | a = 2

824 504 | b = 1

Graphical user interface, text, application

Description automatically generated

Appendix

Performance of the model on Train dataset

|  |  |  |
| --- | --- | --- |
| On Train dataset | | |
| Attrubute Selection | Classification | Correctly Classified Instances |
| GainRationAtributeEval | Naïve Bayes | 71.37% |
| GainRationAtributeEval | Logistic | 76.60% |
| GainRationAtributeEval | OneR | 72.83% |
| GainRationAtributeEval | Decision Tree | 71.56% |
| GainRationAtributeEval | RandomForest | 74.39% |
| GainRationAtributeEval | KNN (K=10) | 71.85% |
| GainRationAtributeEval | KNN (K=12) | 71.94% |
| GainRationAtributeEval | KNN (K=15) | 72.76% |
|  |  |  |
| InfoGainAtributeEval | Naïve Bayes | 71.03% |
| InfoGainAtributeEval | Logistic | 74.78% |
| InfoGainAtributeEval | OneR | 72.83% |
| InfoGainAtributeEval | Decision Tree | 72.61% |
| InfoGainAtributeEval | RandomForest | 74.54% |
| InfoGainAtributeEval | KNN (K=10) | 71.28% |
| InfoGainAtributeEval | KNN (K=12) | 71.85% |
| InfoGainAtributeEval | KNN (K=15) | 72.66% |
|  |  |  |
| ReliefAtributeEval | Naïve Bayes | 71.83% |
| ReliefAtributeEval | Logistic | 74.48% |
| ReliefAtributeEval | OneR | 72.83% |
| ReliefAtributeEval | Decision Tree | 72.74% |
| ReliefAtributeEval | RandomForest | 74.09% |
| ReliefAtributeEval | KNN (K=10) | 71.60% |
| ReliefAtributeEval | KNN (K=12) | 71.91% |
| ReliefAtributeEval | KNN (K=15) | 72.71% |
|  |  |  |
| SymmetricalUncertAttributeEval | Naïve Bayes | 70.95% |
| SymmetricalUncertAttributeEval | Logistic | 74.60% |
| SymmetricalUncertAttributeEval | OneR | 72.83% |
| SymmetricalUncertAttributeEval | Decision Tree | 72.29% |
| SymmetricalUncertAttributeEval | RandomForest | 74.17% |
| SymmetricalUncertAttributeEval | KNN (K=10) | 71.30% |
| SymmetricalUncertAttributeEval | KNN (K=12) | 71.64% |
| SymmetricalUncertAttributeEval | KNN (K=15) | 72.60% |