CS677 – Term Project –

Behavioral Risk Factor Surveillance System

Gaurav Tungare

*Aug 21, 2021*

1.Objective

An opportunity to build and test a classifier models using a real-world data using python. The data to be used is a part of the 2018 BRFSS (Behavioral Risk Factor Surveillance System ) Survey Data prepared by CDC.

2. Background of data

The project-2018-BRFSS-arthritis.csv 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 **- 7995**
* Total **“Test”** Record Count **- 3938**

3. Preprocessing

No preprocessing was needed as the data came with preprocessed csv from CDC

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

For each attribute selection method below classification algorithm were run

* KNN ( K =1 to K = 27)
* Naïve Bayes
* Logistic
* Random Forest ( with different value of N= 1 to N=10 and estimation parameter d=1 to d=5)
* Linear kernel SVM
* Gaussian SVM
* Polynomial kernel SVM of degree=3

5. Test results of all the models

* KNN ( K =1 to K = 27)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| K | TP | FP | TN | FN | Accuracy | TPR | TNR |
| 1 | 2322 | 487 | 799 | 330 | 0.792534 | 0.876 | 0.621 |
| 3 | 2416 | 479 | 845 | 198 | 0.828085 | 0.924 | 0.638 |
| 5 | 2466 | 450 | 862 | 160 | 0.845099 | 0.939 | 0.657 |
| 7 | 2455 | 523 | 852 | 108 | 0.839766 | 0.958 | 0.62 |
| 9 | 2465 | 553 | 831 | 89 | 0.836973 | 0.965 | 0.6 |
| 11 | 2542 | 470 | 821 | 105 | 0.853987 | 0.96 | 0.636 |
| 13 | 2518 | 511 | 816 | 93 | 0.846623 | 0.964 | 0.615 |
| 15 | 2514 | 485 | 853 | 86 | 0.855003 | 0.967 | 0.638 |
| 17 | 2491 | 538 | 836 | 73 | 0.844845 | 0.972 | 0.608 |
| 19 | 2509 | 536 | 834 | 59 | 0.848908 | 0.977 | 0.609 |
| 21 | 2514 | 496 | 848 | 80 | 0.853733 | 0.969 | 0.631 |
| 23 | 2525 | 508 | 848 | 57 | 0.856526 | 0.978 | 0.625 |
| 25 | 2547 | 498 | 834 | 59 | 0.858558 | 0.977 | 0.626 |
| 27 | 2583 | 484 | 816 | 55 | 0.863128 | 0.979 | 0.628 |

* Naïve Bayes

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| TP | FP | TN | FN | Accuracy | TPR | TNR |
| 1462 | 322 | 995 | 1159 | 0.623921 | 0.558 | 0.756 |

* Logistic

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| TP | FP | TN | FN | Accuracy | TPR | TNR |
| 2612 | 1326 | 0 | 0 | 0.663281 | 1 | 0 |

* Random Forest ( with different value of N= 1 to N=10 and estimation parameter d=1 to d=5)

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| N | d | TP | FP | TN | FN | Accuracy | error\_rate | TPR | TNR |
| 1 | 1 | 2372 | 1044 | 294 | 228 | 0.676993 | 0.382246 | 0.912 | 0.22 |
| 1 | 2 | 2425 | 883 | 455 | 175 | 0.731336 | 0.391035 | 0.933 | 0.34 |
| 1 | 3 | 2297 | 965 | 373 | 303 | 0.678009 | 0.394778 | 0.883 | 0.279 |
| 1 | 4 | 2600 | 540 | 798 | 0 | 0.862875 | 0.404706 | 1 | 0.596 |
| 1 | 5 | 2195 | 545 | 793 | 405 | 0.758761 | 0.437257 | 0.844 | 0.593 |
| 2 | 1 | 2577 | 1282 | 56 | 23 | 0.668614 | 0.346195 | 0.991 | 0.042 |
| 2 | 2 | 2250 | 797 | 541 | 350 | 0.708735 | 0.412274 | 0.865 | 0.404 |
| 2 | 3 | 2549 | 452 | 886 | 51 | 0.87227 | 0.416018 | 0.98 | 0.662 |
| 2 | 4 | 2355 | 876 | 462 | 245 | 0.715338 | 0.397301 | 0.906 | 0.345 |
| 2 | 5 | 2528 | 57 | 1281 | 72 | 0.967242 | 0.449871 | 0.972 | 0.957 |
| 3 | 1 | 2539 | 1096 | 242 | 61 | 0.706196 | 0.364424 | 0.977 | 0.181 |
| 3 | 2 | 2600 | 168 | 1170 | 0 | 0.957339 | 0.434979 | 1 | 0.874 |
| 3 | 3 | 2545 | 143 | 1195 | 55 | 0.949721 | 0.441489 | 0.979 | 0.893 |
| 3 | 4 | 2547 | 466 | 872 | 53 | 0.868207 | 0.415041 | 0.98 | 0.652 |
| 3 | 5 | 2453 | 353 | 985 | 147 | 0.873032 | 0.431886 | 0.943 | 0.736 |
| 4 | 1 | 2600 | 175 | 1163 | 0 | 0.955561 | 0.434409 | 1 | 0.869 |
| 4 | 2 | 2600 | 236 | 1102 | 0 | 0.940071 | 0.429445 | 1 | 0.824 |
| 4 | 3 | 2594 | 243 | 1095 | 6 | 0.93677 | 0.429364 | 0.998 | 0.818 |
| 4 | 4 | 2512 | 368 | 970 | 88 | 0.884205 | 0.425864 | 0.966 | 0.725 |
| 4 | 5 | 2514 | 271 | 1067 | 86 | 0.909345 | 0.433595 | 0.967 | 0.797 |
| 5 | 1 | 2557 | 1219 | 119 | 43 | 0.679533 | 0.35295 | 0.983 | 0.089 |
| 5 | 2 | 2275 | 782 | 556 | 325 | 0.718893 | 0.411461 | 0.875 | 0.416 |
| 5 | 3 | 2214 | 643 | 695 | 386 | 0.7387 | 0.427736 | 0.852 | 0.519 |
| 5 | 4 | 2597 | 43 | 1295 | 3 | 0.988319 | 0.445395 | 0.999 | 0.968 |
| 5 | 5 | 2590 | 41 | 1297 | 10 | 0.987049 | 0.446128 | 0.996 | 0.969 |
| 6 | 1 | 2591 | 1308 | 30 | 9 | 0.665566 | 0.34294 | 0.997 | 0.022 |
| 6 | 2 | 2322 | 917 | 421 | 278 | 0.696546 | 0.39665 | 0.893 | 0.315 |
| 6 | 3 | 2600 | 413 | 925 | 0 | 0.895124 | 0.415041 | 1 | 0.691 |
| 6 | 4 | 2560 | 350 | 988 | 40 | 0.900965 | 0.423423 | 0.985 | 0.738 |
| 6 | 5 | 2520 | 277 | 1061 | 80 | 0.909345 | 0.432619 | 0.969 | 0.793 |
| 7 | 1 | 2600 | 612 | 726 | 0 | 0.844591 | 0.398847 | 1 | 0.543 |
| 7 | 2 | 2425 | 799 | 539 | 175 | 0.752666 | 0.39787 | 0.933 | 0.403 |
| 7 | 3 | 2595 | 263 | 1075 | 5 | 0.931945 | 0.427655 | 0.998 | 0.803 |
| 7 | 4 | 2600 | 3 | 1335 | 0 | 0.999238 | 0.448406 | 1 | 0.998 |
| 7 | 5 | 2587 | 206 | 1132 | 13 | 0.944388 | 0.432944 | 0.995 | 0.846 |
| 8 | 1 | 2577 | 1267 | 71 | 23 | 0.672423 | 0.347416 | 0.991 | 0.053 |
| 8 | 2 | 2441 | 912 | 426 | 159 | 0.728035 | 0.387373 | 0.939 | 0.318 |
| 8 | 3 | 2504 | 549 | 789 | 96 | 0.836211 | 0.411786 | 0.963 | 0.59 |
| 8 | 4 | 2564 | 335 | 1003 | 36 | 0.90579 | 0.424318 | 0.986 | 0.75 |
| 8 | 5 | 2596 | 34 | 1304 | 4 | 0.99035 | 0.446209 | 0.998 | 0.975 |
| 9 | 1 | 2600 | 117 | 1221 | 0 | 0.970289 | 0.439129 | 1 | 0.913 |
| 9 | 2 | 2497 | 811 | 527 | 103 | 0.767902 | 0.391035 | 0.96 | 0.394 |
| 9 | 3 | 2599 | 123 | 1215 | 1 | 0.968512 | 0.438722 | 1 | 0.908 |
| 9 | 4 | 2538 | 71 | 1267 | 62 | 0.966227 | 0.447918 | 0.976 | 0.947 |
| 9 | 5 | 2550 | 122 | 1216 | 50 | 0.956323 | 0.442791 | 0.981 | 0.909 |
| 10 | 1 | 2535 | 1123 | 215 | 65 | 0.698324 | 0.362552 | 0.975 | 0.161 |
| 10 | 2 | 2600 | 203 | 1135 | 0 | 0.948451 | 0.432131 | 1 | 0.848 |
| 10 | 3 | 2575 | 118 | 1220 | 25 | 0.963687 | 0.441082 | 0.99 | 0.912 |
| 10 | 4 | 2592 | 213 | 1125 | 8 | 0.94388 | 0.431968 | 0.997 | 0.841 |
| 10 | 5 | 2593 | 47 | 1291 | 7 | 0.986287 | 0.445395 | 0.997 | 0.965 |

* Linear kernel SVM

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| TP | FP | TN | FN | Accuracy | TPR | TNR |
| 2583 | 1355 | 0 | 0 | 0.655917 | 1 | 0 |

* Gaussian SVM

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| TP | FP | TN | FN | Accuracy | TPR | TNR |
| 2516 | 1422 | 0 | 0 | 0.638903 | 1 | 0 |

* Polynomial kernel SVM of degree=3

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| TP | FP | TN | FN | Accuracy | TPR | TNR |
| 2687 | 1251 | 0 | 0 | 0.682326 | 1 | 0 |

**The best result is given by Random Forest with N= 7 and estimation parameter d=4**

6. Best Model - Justification

The above model has **Accuracy** of 0.999238 which indicate that the model is able to correctly predict the value of some form of arthritis, rheumatoid arthritis, gout, lupus, or fibromyalgia, accuracy of >0.90 indicate a good performing model

x

The above model has **TP rate** of 1 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

Round 2/ Iteration 2

7. Attribute selection for round 2

* I ran another iteration for Attribute selection method Information Gain- Information gain was done in python to the select the attribute with best info gain
* This is new attribute selection is going to be used for same set of classifiers and results are compared

|  |  |
| --- | --- |
| Col | infogain |
| x.llcpwt | 10.14927 |
| x.llcpwt2 | 9.527179 |
| x.ststr | 9.455642 |
| x.bmi5 | 9.106152 |
| x.wt2rake | 7.139886 |
| x.strwt | 6.938341 |
| weight2 | 6.389888 |
| wtkg3 | 6.292746 |
| x.age80 | 5.806313 |
| x.state | 5.505707 |
| Iday | 4.922568 |
| height3 | 4.229899 |
| htm4 | 4.043475 |
| htin4 | 4.014331 |
| x.ageg5yr | 3.72204 |
| x.drnkwek | 3.688193 |
| Imonth | 3.578902 |
| Fmonth | 3.576333 |
| alcday5 | 3.28688 |
| drocdy3. | 3.18827 |
| income2 | 3.101019 |

8. 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: **ststr** - five digit number that combines the values for \_STATE

This column has high information gain 9.455642, this indicates a chronic health condition , this will drive the prediction significantly so which mean the state which you stay does play a role in your health condition

1. Variable Name: **BMI5** – Body Mass Index (BMI)

This column has high information gain 9.106152, 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: \_**WT2RAKE** weight used in raking

This column has high information gain 7.139886, 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 WT2RAKE

1. Variable Name: \_ **AGE80YR** Reported age in two age groups calculated variable

This column has high information gain 7.139886 , this is important column as it will drive other columns to predict the health attribute correctly

1. Variable Name: \_ **HEIGHT5** Reported Height in Feet and Inches

This column has high information gain 7.139886, which indicates relation between person height and arthritis

9. Test results for round 2 – with highest information gain attribute selection

* KNN ( K =1 to K = 27)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| K | TP | FP | TN | FN | Accuracy | TPR | TNR |
| 1 | 1806 | 727 | 633 | 772 | 0.61935 | 0.701 | 0.465 |
| 3 | 1972 | 763 | 556 | 647 | 0.64195 | 0.753 | 0.422 |
| 5 | 2057 | 827 | 525 | 529 | 0.655663 | 0.795 | 0.388 |
| 7 | 2115 | 818 | 538 | 467 | 0.673692 | 0.819 | 0.397 |
| 9 | 2115 | 821 | 518 | 484 | 0.668614 | 0.814 | 0.387 |
| 11 | 2143 | 822 | 502 | 471 | 0.671661 | 0.82 | 0.379 |
| 13 | 2176 | 807 | 489 | 466 | 0.676739 | 0.824 | 0.377 |
| 15 | 2191 | 862 | 430 | 455 | 0.665566 | 0.828 | 0.333 |
| 17 | 2237 | 849 | 476 | 376 | 0.688928 | 0.856 | 0.359 |
| 19 | 2233 | 843 | 493 | 369 | 0.69223 | 0.858 | 0.369 |
| 21 | 2266 | 846 | 455 | 371 | 0.69096 | 0.859 | 0.35 |
| 23 | 2247 | 864 | 454 | 373 | 0.685881 | 0.858 | 0.344 |
| 25 | 2232 | 888 | 468 | 350 | 0.685627 | 0.864 | 0.345 |
| 27 | 2282 | 901 | 416 | 339 | 0.685119 | 0.871 | 0.316 |

* Naïve Bayes

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| TP | FP | TN | FN | Accuracy | TPR | TNR |
| 1199 | 272 | 1073 | 1394 | 0.576943 | 0.462 | 0.798 |

* Logistic

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| TP | FP | TN | FN | Accuracy | TPR | TNR |
| 2203 | 760 | 559 | 416 | 0.701371 | 0.841 | 0.424 |

* Decision Tree

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| TP | FP | TN | FN | Accuracy | TPR | TNR |
| 1880 | 689 | 610 | 759 | 0.632301 | 0.712 | 0.47 |

**\*\*\* The attribute selection with best information gain do not necessarily translate into better model accuracy**

10. Lessons learned from this project.

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

Let me summarize on each of the stuff

Classification

Learnt various aspect of classification on various classification algorithms

Naïve Bayes, Logistic, Decision Tree, Randomforest, KNN and SVN.

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 which were analyzed ,various accuracy, sensitivity, specificity, precision, 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.

From Attribute Selection

Attribute selection methodology, Information Gain 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 but the result of model accuracy were not necessarily giving the best result with attribute

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