Classification of Key Risk Factors for Alzheimer's From Questionnaire Data

Leah Zhang and Victoria Zhang

Period 4

Table of contents

01

Project Objective

Alzheimer's Disease

"...a general term for memory loss and other cognitive abilities serious enough to interfere with daily life" (Alzheimer's Association).

- not a normal part of aging
- no known cure
- if no medical breakthrough is made, 13.8 million Americans could have Alzheimer's by 2060 (NIH)

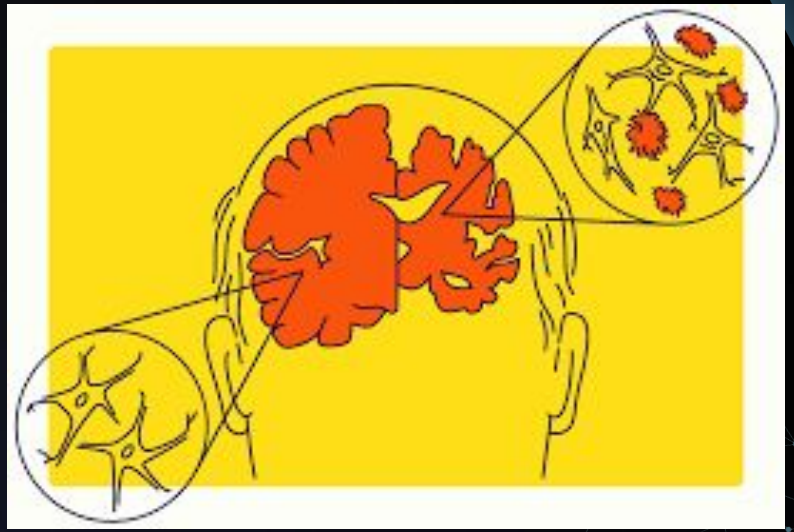


Image source: Alzheimer's Disease Causes, Symptoms, Treatments

Objective: To categorize results from a telephone questionnaire into six risk factors for Alzheimer’s.

02

Background

Datasource

- Collected from questionnaire by the Behavioral Risk Factor Surveillance System
- The ‘Question’ attribute asked questions regarding health
    - Health-related risk behaviors
    - chronic health conditions
    - health-care access
- 30 attributes; 284142 instances

Overall Classification

- Each instance is an individual
- classifying “Class” attribute based on age, location, datasource, etc
- Name is misleading;
- Not classifying Alzheimer's, instead indicator of healthy aging

Attributes

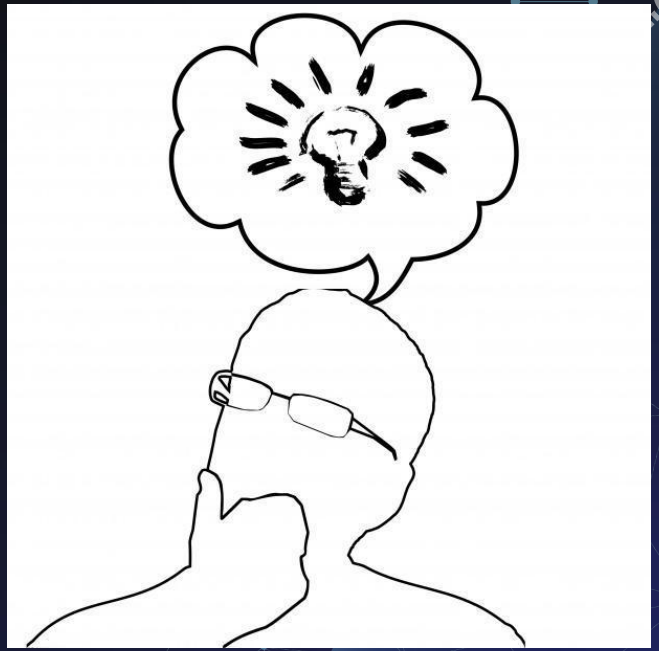
RowId	Dataset row identifier
YearStart	Year Start
YearEnd	Year End
LocationAbbr	Location Abbreviation
LocationDesc	Location Description
Datasource	Data Source
Class	Class description
Topic	Topic description

Question	Question
Data_Value_Unit	The unit, such as "%" for percentage
DataValueTypeID	Identifier for the Data Value Type
Data_Value_Type	The data value type, such as age-adjusted prevalence or crude prevalence
Data_Value	Data Value, such as 14.7
Data_Value_Alt	Equal to data value, but format is numeric
Data_Value_Footnote_Symbol	Footnote Symbol

Class

Health indicators:

- Screenings and Vaccines 62153
- Nutrition/Physical Activity/Obesity 33194
- Caregiving 25493
- Mental Health 22184
- Smoking and Alcohol Use 22183
- Cognitive Decline 22182



03

Preprocessing

Cleaning Up Redundant Columns

```
['Data_Value_Footnote_Symbol', 'Data_Value_Footnote', 'ClassID',
'QuestionID', 'TopicID', 'RowId', 'Data_Value_Alt',
'StratificationCategoryID1', 'StratificationCategoryID2', 'Data_Value_Alt',
'LocationAbbr', 'LocationID', 'StratificationID1', 'StratificationID2']
```

- Attributes that described other columns in the dataset
- Fourteen columns removed

Processing Strings

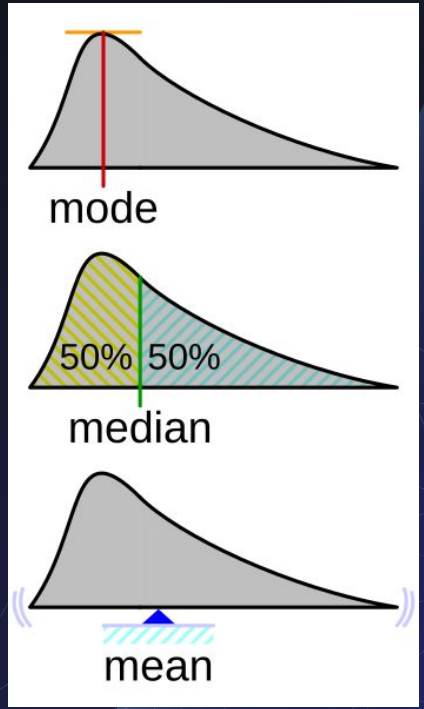
```
df['Stratification2'] = df['Stratification2'].str.replace(',', '',
regex=False)
df['DataValueTypeID'] = df['DataValueTypeID'].str.replace(',', '',
regex=False)
df['Question'] = df['Question'].str.replace(',', '', regex=False)
```

- Commas, quotation marks, blank attribute names
- Pandas add an extra column

Replacing Nans/Null values

- Quantitative data filled by mean
- Qualitative data filled by mode
- Skipping columns without missing values by using

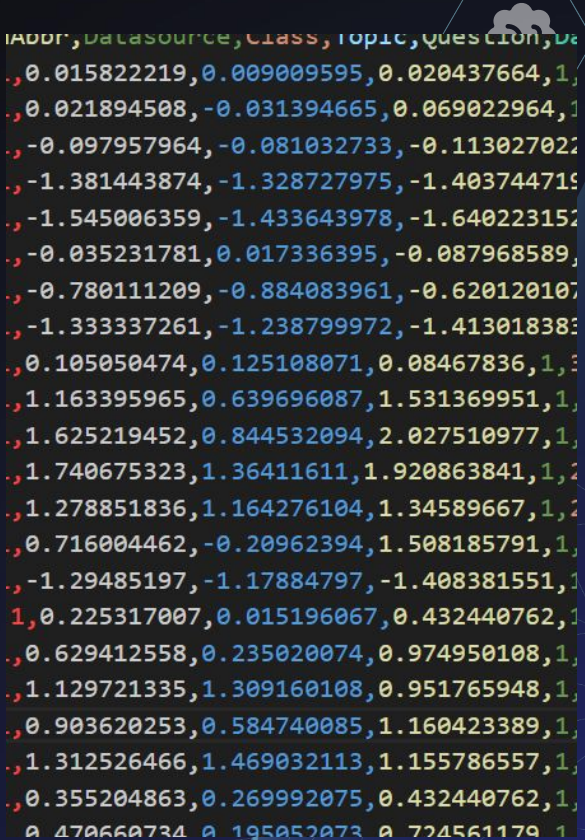
```
df[column].isnull().sum() == 0:
```



Normalizing Features

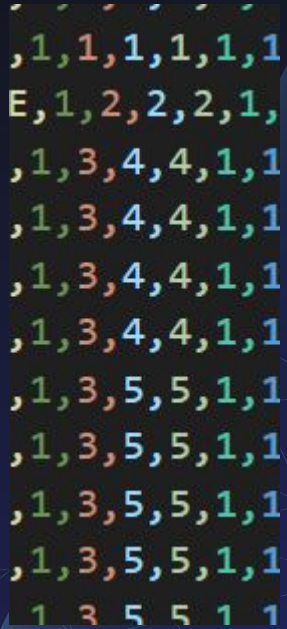
- We normalize attributes that are non-categorical
- `['Data_Value', 'Low_Confidence_Limit', 'High_Confidence_Limit']`
- We use the Z-score normalization from scipy.stats

```
df[val] = zscore(df[val])
```



Transforming Non-numerical Data

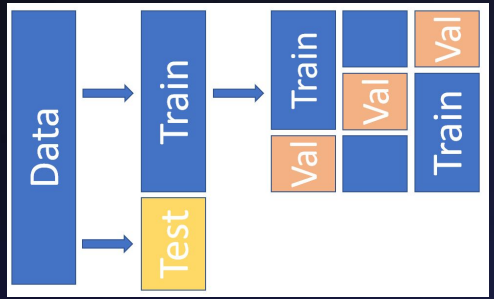
- When data is not transformed to numerical
- Weka is limited
- For loop through each column, check if is qualitative or quantitative
- use `pd.factorize(df[column])[0] + 1`



Splitting the Dataset

- We will do a 70%/20%/10% split for the train and test datasets.
- Use sklearn's train_test_split to split the data with the stratify argument to ensure the ratios of the class are the same.
- The lengths are 198899, 57112, and 28131 respectively. 284,142 total.

```
train_test_split(X_temp, y_temp, stratify=y_temp, test_size=0.33, random_state=42)
```



04

Attribute Selection

Attribute Selection Algorithms Used

Kendall Rank Correlation Coefficient

CorrelationAttributeEval

- Finds correlation based on rank

$$\tau = \frac{\text{Number of concordant pairs} - \text{Number of discordant pairs}}{\frac{n(n-1)}{2}}$$

Concordant Pair: (x1,y1)(x_1, y_1)(x1,y1) and (x2,y2)(x_2, y_2)(x2,y2) is **concordant** if it satisfies one of the following conditions:

x1>x2x_1 > x_2x1>x2 and y1>y2y_1 > y_2y1>y2

x1<x2x_1 < x_2x1<x2 and y1<y2y_1 < y_2y1<y2

Discordant Pair: A pair of observations (x1,y1)(x_1, y_1)(x1,y1) and (x2,y2)(x_2, y_2)(x2,y2) is **discordant** if it satisfies one of the following conditions:

x1>x2x_1 > x_2x1>x2 and y1<y2y_1 < y_2y1<y2

1<x2x_1 < x_2x1<x2 and y1>y2y_1 > y_2y1>y2

n: The total number of observations.

Attribute Selection Algorithms Used

Pearson's Correlation

CorrelationAttributeEval

- ranks attributes based on correlation coefficient

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\, n\Sigma x^2 - (\Sigma x)^2 \,][\, n\Sigma y^2 - (\Sigma y)^2 \,]}}$$

Subset Analysis

CfsSubsetEval

Selects attributes by:

1. grouping attributes as subsets
2. in each subset, evaluate attributes individually
3. choose a subset with high correlation with class but low intercorrelation

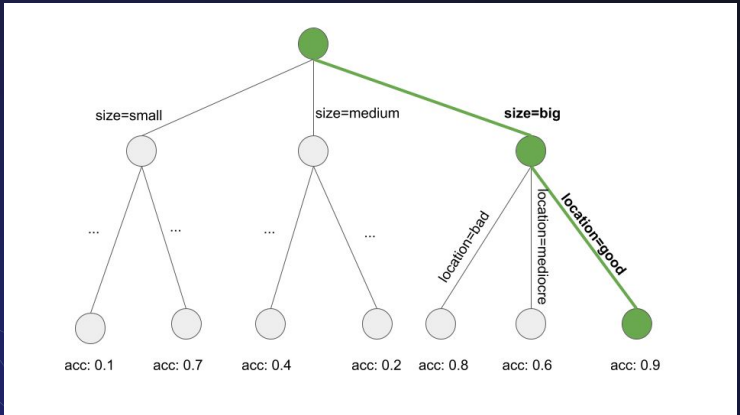
Searches with greedy-stepwise search

Attribute Selection Algorithms Used

Selection by OneR

OneRAttributeEval

- ranks attributes based on classification error from the OneR algorithm



Entropy/Information Gain

InfoGainAttributeEval

- ranks attributes based on information gained with respect to the class

Attributes Selected: Kendall Correlation

```
YearStart SignificanceResult(statistic=-0.008292277413063754, pvalue=9.933366399952563e-09)
YearEnd SignificanceResult(statistic=-0.037674185690728605, pvalue=1.8876229756547483e-149)
LocationDesc SignificanceResult(statistic=0.004162443388828436, pvalue=0.002471294877482304)
Datasource SignificanceResult(statistic=nan, pvalue=nan)
Topic SignificanceResult(statistic=-0.14172617218857977, pvalue=0.0)
Question SignificanceResult(statistic=-0.14172617218857977, pvalue=0.0)
Data_Value_Unit SignificanceResult(statistic=0.03031410117404321, pvalue=1.2063404836511375e-73)
DataValueTypeID SignificanceResult(statistic=0.03031410117404321, pvalue=1.2063404836511375e-73)
Data_Value_Type SignificanceResult(statistic=0.03031410117404321, pvalue=1.2063404836511375e-73)
Data_Value SignificanceResult(statistic=0.21041334613980775, pvalue=0.0)
Low_Confidence_Limit SignificanceResult(statistic=0.23429182517060937, pvalue=0.0)
High_Confidence_Limit SignificanceResult(statistic=0.18575483791844988, pvalue=0.0)
StratificationCategory1 SignificanceResult(statistic=nan, pvalue=nan)
Stratification1 SignificanceResult(statistic=-0.0009736636955514779, pvalue=0.5362844239255413)
StratificationCategory2 SignificanceResult(statistic=-0.03917874396411915, pvalue=9.978459679775013e-122)
Stratification2 SignificanceResult(statistic=-0.02873497235392164, pvalue=4.8475202075739715e-86)
```

0.10 and -0.10

as bounds

Chosen Attributes: Topic, Question, Data_Value, Low_Confidence_Limit, High_Confidence_Limit

Attributes Selected: Pearson's Correlation

```
Ranked attributes:
 0.22043  12 Low_Confidence_Limit
 0.21785   7 Question
 0.21785   6 Topic
 0.21248   9 DataValueTypeID
 0.21248   8 Data_Value_Unit
 0.21248  10 Data_Value_Type
 0.20816  11 Data_Value
 0.18774  13 High_Confidence_Limit
 0.02513  18 Geolocation
 0.02459   3 YearEnd
 0.02426  16 StratificationCategory2
 0.01836   2 YearStart
 0.01821  17 Stratification2
 0.003     4 LocationDesc
 0.00213   1 Num
 0.00113  15 Stratification1
 0        14 StratificationCategory1
 0         5 Datasource

Selected attributes: 12,7,6,9,8,10,11,13,18,3,16,2,17,4,1,15,14,5 : 18
```

cutoff = 0.20

Attributes Selected: Subset Analysis

```
=== Attribute Selection on all input data ===

Search Method:
        Greedy Stepwise (forwards).
        Start set: no attributes
        Merit of best subset found:    0.767

Attribute Subset Evaluator (supervised, Class (nominal): 18 Class):
        CFS Subset Evaluator
        Including locally predictive attributes

Selected attributes: 5,16,17 : 3
                     Topic
                     Stratification2
                     Geolocation
```

Attributes Selected: OneR

```
Ranked attributes:
100         5 Topic
100         6 Question
 61.4298   10 Data_Value
 61.3478   11 Low_Confidence_Limit
 60.9674   12 High_Confidence_Limit
 37.7632   17 Geolocation
 35.7117   16 Stratification2
 34.0509    4 Datasource
 34.0509    3 LocationDesc
 34.0509    2 YearEnd
 34.0509    9 Data_Value_Type
 34.0509    7 Data_Value_Unit
 34.0509    8 DataValueTypeID
 34.0509   15 StratificationCategory2
 34.0509   14 Stratification1
 34.0509   13 StratificationCategory1
 34.0509    1 YearStart

Selected attributes: 5,6,10,11,12,17,16,4,3,2,9,7,8,15,14,13,1 : 17
```

cutoff = 37.0

Attributes Selected: Information Gain

```
Ranked attributes:
 2.5445     5 Topic
 2.5445     6 Question
 1.24906   10 Data_Value
 1.23631   11 Low_Confidence_Limit
 1.22586   12 High_Confidence_Limit
 0.14572   17 Geolocation
 0.13029    8 DataValueTypeID
 0.13029    7 Data_Value_Unit
 0.13029    9 Data_Value_Type
 0.07083    1 YearStart
 0.04949    2 YearEnd
 0.03301   16 Stratification2
 0.02841    3 LocationDesc
 0.00292   15 StratificationCategory2
 0         13 StratificationCategory1
 0         14 Stratification1
 0          4 Datasource

Selected attributes: 5,6,10,11,12,17,8,7,9,1,2,16,3,15,13,14,4 : 17
```

cutoff = 0.13

05

Results + Analysis

Classification Algorithms Used

OneR

Rule-based method selecting ONE rule based on the attribute with least classification error

Decision Table

Rule-based method that selects MULTIPLE rules based on interactions between multiple attributes

J48

Tree-based method based on the usage of the C4.5 algorithm to generate decision trees

Random Forest

Tree-based method that creates a forest of random trees

Att. Selection Algorithm	Classification Algorithm	Accuracy	MAE	RMSE	Precision	Recall
Kendall's Rank (Non-Weka)	OneR	100%	0	0	1.0	1.0
	Decision Table	100%	0.0002	0.0003	1.0	1.0
	J48	100%	0	0	1.0	1.0
	Random Forest	99.9929%	0	0.0037	1.0 (0.99 in one category)	1.0
Pearson's Correlation	OneR	100%	0	0	1.0	1.0
	Decision Table	100%	0	0	1.0	1.0
	J48	100%	0	0	1.0	1.0
	Random Forest	99.9964%	0.0001	0.0031	1.0	1.0

Att. Selection Algorithm	Classification Algorithm	Accuracy	MAE	RMSE	Precision	Recall
Subset Analysis	OneR	100%	0	0	1.0	1.0
	Decision Table	100%	0.0002	0.0003	1.0	1.0
	J48	100%	0	0	1.0	1.0
	Random Forest	100%	0.0002	0.0011	1.0	1.0
OneR Algorithm	OneR	100%	0	0	1.0	1.0
	Decision Table	100%	0.002	0.003	1.0	1.0
	J48	100%	0	0	1.0	1.0
	Random Forest	99.9929%	0.0001	0.004	1.0 (0.99 in one class)	1.0

Att. Selection Algorithm	**Classification Algorithm**	Accuracy	MAE	RMSE	Precision	Recall
Information Gain	OneR	100%	0	0	1.0	1.0
	Decision Table	100%	0.0002	0.0003	1.0	1.0
	J48	100%	0	0	1.0	1.0
	Random Forest	99.9929%	0.0001	0.0039	1.0 (0.99) in a class	1.0

Confusion Matrix Comparisons

```
=== Confusion Matrix ===

   a    b    c    d    e    f    g   <-- classified as
2196    0    0    0    0    0    0 |    a = Cognitive Decline
   0 6154    0    0    0    0    0 |    b = Screenings and Vaccines
   0    0 9579    0    0    0    0 |    c = Overall Health
   0    0    0 2524    0    0    0 |    d = Caregiving
   0    0    0    0 3286    0    0 |    e = Nutrition/Physical Activity/Obesity
   0    0    0    0    0 2196    0 |    f = Mental Health
   0    0    0    0    0    0 2196 |    g = Smoking and Alcohol Use
```

A perfect confusion matrix

Subset Analysis with OneR

```
=== Confusion Matrix ===

   a    b    c    d    e    f    g   <-- classified as
2196    0    0    0    0    0    0 |    a = Cognitive Decline
   0 6152    0    0    0    0    2 |    b = Screenings and Vaccines
   0    0 9579    0    0    0    0 |    c = Overall Health
   0    0    0 2524    0    0    0 |    d = Caregiving
   0    0    0    0 3286    0    0 |    e = Nutrition/Physical Activity/Obesity
   0    0    0    0    0 2196    0 |    f = Mental Health
   0    0    0    0    0    0 2196 |    g = Smoking and Alcohol Use
```

Confusion matrix with some errors

Kendall's Correlation with Random Forest

Kendall's Correlation	Pearson's Correlation	Subset Analysis	OneR Algorithm	Entropy/Info Gain
Topic	Low_Confidence_Limit	Topic	Topic	Topic
Question	Question	Stratification2	Question	Question
Data_Value	Topic	Geolocation	Data_Value	Data_Value
Low_Confidence_Limit	Data_Value_Type		Low_Confidence_Limit	Low_Confidence_Limit
High_ConfidencE_Limit	Data_Value_Unit		High_Confidence_Limit	High_Confidence_Limit
	DataValueTypeID		Geolocation	Geolocation
	Data_Value			DataValueTypeID
				Data_Value_Unit
				Data_Value_Type

Why did most algorithms perform perfectly?

What if we remove ‘Topic’?

```
=== Summary ===

Correctly Classified Instances       28131               100      %
Incorrectly Classified Instances         0                 0      %
Kappa statistic                          1
Mean absolute error                      0
Root mean squared error                  0
Relative absolute error                  0      %
Root relative squared error              0      %
Total Number of Instances            28131

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Cognitive Decline
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Screenings and Vaccines
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Overall Health
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Caregiving
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Nutrition/Physical Activity/Obesity
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Mental Health
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Smoking and Alcohol Use
Weighted Avg.    1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000

=== Confusion Matrix ===

    a    b    c    d    e    f    g   <-- classified as
 2196    0    0    0    0    0    0 |    a = Cognitive Decline
    0 6154    0    0    0    0    0 |    b = Screenings and Vaccines
    0    0 9579    0    0    0    0 |    c = Overall Health
    0    0    0 2524    0    0    0 |    d = Caregiving
    0    0    0    0 3286    0    0 |    e = Nutrition/Physical Activity/Obesity
    0    0    0    0    0 2196    0 |    f = Mental Health
    0    0    0    0    0    0 2196 |    g = Smoking and Alcohol Use
```

From the Kendall Rank-selected attributes with OneR:

What if we ALSO remove ‘Question’?

```
=== Summary ===

Correctly Classified Instances       17132               60.9008 %
Incorrectly Classified Instances     10999               39.0992 %
Kappa statistic                          0.4982
Mean absolute error                      0.1117
Root mean squared error                  0.3342
Relative absolute error                 49.1059 %
Root relative squared error             99.1022 %
Total Number of Instances            28131

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.365    0.000    1.000      0.365    0.535      0.588    0.682     0.414     Cognitive Decline
                 0.752    0.110    0.658      0.752    0.702      0.613    0.821     0.549     Screenings and Vaccines
                 0.688    0.236    0.601      0.688    0.641      0.439    0.726     0.520     Overall Health
                 0.342    0.005    0.867      0.342    0.491      0.521    0.669     0.356     Caregiving
                 0.540    0.067    0.516      0.540    0.528      0.464    0.737     0.333     Nutrition/Physical Activity/Obesity
                 0.617    0.060    0.464      0.617    0.530      0.490    0.779     0.317     Mental Health
                 0.509    0.033    0.567      0.509    0.536      0.500    0.738     0.327     Smoking and Alcohol Use
Weighted Avg.    0.609    0.120    0.645      0.609    0.603      0.508    0.745     0.450

=== Confusion Matrix ===

    a    b    c    d    e    f    g   <-- classified as
  801  296  503   15  237  207  137 |    a = Cognitive Decline
    0 4629 1149    9  306   42   19 |    b = Screenings and Vaccines
    0 1663 6591   38  794  297  196 |    c = Overall Health
    0  297  698  864  247  295  123 |    d = Caregiving
    0  147 1042   22 1774  217   84 |    e = Nutrition/Physical Activity/Obesity
    0    1  465   24   56 1356  294 |    f = Mental Health
    0    4  523   24   21  507 1117 |    g = Smoking and Alcohol Use
```

From the Kendall Rank-selected attributes with OneR:

Att. Selection Algorithm	**Classification Algorithm**	Accuracy	MAE	RMSE	Precision	Recall
Kendall's Rank (Non-Weka)	OneR	60.9002%	0.1117	0.3342	0.645	0.609
	Decision Table	62.2445%	0.1254	0.2505	0.626	0.622
	J48	62.2516%	0.1193	0.2646	0.621	0.623
	Random Forest	61.5655%	0.1153	0.2705	0.615	0.616
Pearson's Correlation	OneR	60.9008%	0.1117	0.3342	0.645	0.609
	Decision Table	63.8904%	0.1203	0.2447	0.649	0.639
	J48	63.318%	0.1172	0.2556	0.632	0.633
	Random Forest	61.5691%	0.1144	0.2683	0.616	0.616

Att. Selection Algorithm	**Classification Algorithm**	Accuracy	MAE	RMSE	Precision	Recall
Subset Analysis	OneR	37.7022%	0.178	0.4219	?	0.377
	Decision Table	39.085%	0.2152	0.3279	?	0.391
	J48	39.085%	0.2175	0.3287	?	0.391
	Random Forest	39.117%	0.2151	0.3282	?	0.391
OneR Algorithm	OneR	66.0268%	0.1117	0.3342	0.645	0.609
	Decision Table	66.0268%	0.1237	0.243	0.661	0.660
	J48	67.5767%	0.1026	0.2588	0.675	0.676
	Random Forest	68.405%	0.1017	0.2391	0.683	0.684

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
0.000	0.000	?	0.000	?	?	0.500	0.078
0.114	0.016	0.667	0.114	0.195	0.214	0.549	0.270
0.947	0.844	0.367	0.947	0.529	0.150	0.552	0.366
0.241	0.035	0.406	0.241	0.302	0.262	0.603	0.166
0.000	0.000	?	0.000	?	?	0.500	0.117
0.101	0.024	0.260	0.101	0.146	0.120	0.538	0.096
0.000	0.000	?	0.000	?	?	0.500	0.078
0.377	0.296	?	0.377	?	?	0.541	0.232

Explanation of Subset Analysis

Att. Selection Algorithm	Classification Algorithm	Accuracy	MAE	RMSE	Precision	Recall
Information Gain	OneR	60.9008%	0.1117	0.3342	0.645	0.609
	Decision Table	66.9724%	0.1218	0.24	0.672	0.670
	J48	68.2663%	0.1003	0.2554	0.682	0.683
	Random Forest	**68.6716%**	**0.1008**	**0.2363**	**0.686**	**0.687**

We choose **Information Gain Attribute Selection with Random Forest classification to be our best model.**

Steps for Reproducibility:

1. In Weka, load the dataset titled Alzh_train.csv. This dataset automatically includes all preprocessing steps done in Python, as outlined above.
2. Under Select Attributes, under AttributeEvaluator, choose InfoGainAttributeEval. Weka should automatically select the Ranker Search method.
3. On the left side of the screen, select the down arrow by "No class" and select "(Nom) Class" as the class variable. Click start to begin attribute selection.
4. Noting down the attributes with a value of more than 0.13, go to the Preprocess tab and remove all attributes below 0.13 by selecting such attributes and removing them. Do not remove the class. The remaining attributes should be identical to those found in the file Alzh_InfoTrain.arff.
5. Open Alzh_test.csv and remove the same attributes removed in Alzh_train.csv. The resulting dataset should be the same found in Alzh_InfoTestarff.
6. Furthermore remove 'Topic' and 'Question' from both of the above training and testing sets if you wish to analyze without those two attributes. The resulting datasets should be equivalent to Alzh_InfoTrain2.arff and Alzh_InfoTest2.arff
7. Return to the Alzh_InfoTrain2.arff dataset. Click on the Classify tab, and under Supplied Test Set, choose Alzh_InfoTest2.arff.
8. Repeat step 3, choosing the correct class to be "Class".
9. Select the Random Forest model under rules.
10. Click Start. This model can be found at Info_RandomForest_BestModel.model

06

Conclusion

Reflection

- Understanding class variable
- Stratified sampling in code
- Dealing with qualitative data
- Being mindful of data

Next Steps

- Improve preprocessing.
    - use min-max and decimal scaling to replicate the exact scale.
- Further analysis of performance, model, and feature selector.
    - possible investigation of intercorrelation between attributes

References

Alzheimer's Association. (2023). 2023 Alzheimer’s Disease Facts and Figures. *Alzheimer’s & Dementia*, *19*(4), 1598–1695. https://doi.org/10.1002/alz.13016

CDC Division of Population Health. (2024). *Alzheimer’s Disease and Healthy Aging Data | Data | Centers for Disease Control and Prevention*. Chronicdata.cdc.gov. https://chronicdata.cdc.gov/Healthy-Aging/Alzheimer-s-Disease-and-Healthy-Aging-Data/hfr9-rurv/about_data

CfsSubsetEval. (n.d.). Weka.sourceforge.io. https://weka.sourceforge.io/doc.dev/weka/attributeSelection/CfsSubsetEval.html

GeeksForGeeks. (2020, May 15). *Python | Kendall Rank Correlation Coefficient*. GeeksforGeeks. https://www.geeksforgeeks.org/python-kendall-rank-correlation-coefficient/

GreedyStepwise. (2022, January 28). Sourceforge.io. https://weka.sourceforge.io/doc.dev/weka/attributeSelection/GreedyStepwise.html

https://www.health.com/condition/alzheimers-overview. (2022, August 15). *What Is Alzheimer’s Disease?*