Understanding the Effect of Certain Socioeconomic Factors on Mental Health Outcomes

Medha Pappula, Kade Yen

ML1 - Q1 Project

Thomas Jefferson High School for Science and Technology

10/21/24

Part 1 - Statement / Project Goal

Mental health has emerged as a significant global concern in recent years, encompassing various challenges individuals face. The impact of socioeconomic factors on mental well-being has garnered considerable attention, revealing a complex relationship between the two. Individuals with lower socioeconomic status (SES) tend to experience higher rates of mental disorders, encounter barriers in accessing mental health services, and often suffer from increased psychological distress. Conversely, individuals with higher SES generally exhibit lower rates of mental disorders, possess better access to resources and support, and enjoy stronger social networks. Understanding the influence of socioeconomic factors on mental health outcomes is crucial for developing targeted interventions and policies. This overview aims to provide a foundation for further exploration of the connection between mental health and socioeconomic factors, emphasizing the need to comprehend and address these factors to improve mental health outcomes globally.

This leads to our research question: *What specific socioeconomic factors among adults in the United States have a profound impact on mental health severity?*

To answer this question, we will look at the 2023 National Health Interview Survey provided by the CDC. This is a dataset consisting of a broad range of health topics collected through personal household interviews. Specifically, to ensure low variability in results, we will look at adult interviews.

We specifically chose to look at *severity* as opposed to *presence* since it's more important to assess the level of an individual has been affected by mental health as opposed to the presence of it. This can help ensure medical professionals are utilizing their resources to help those who are the most vulnerable.

Part 2 - Description of Dataset

This dataset has 29522 rows of information for 647 attributes. Each row represents an adult interview conducted and their responses to certain questions.
Link to dataset: https://www.cdc.gov/nchs/nhis/2023nhis.htm

Below is an explanation of each attribute:
Attribute Description.pdf - https://drive.google.com/open?id=14JO5mxtoPb2VAMDX68zPNtwcqRXZsrwt

For this study, we are classifying the attributes listed as "brief mental health assessment" which includes the PHQ41_A, PHQ42_A, PHQ44_A, and PHQ44_A attributes. These attributes ask questions related to certain behaviors such as loss of interest and anxiety within the past 2 weeks. It also asks for different medical conditions that an individual may have had in the past to try to correlate that with their mental health.

Classifying for this attribute will be useful because we can directly see the impact of specific socioeconomic factors on the mental health of the individual and use this information to create conclusions on infrastructure/public policy plans to alleviate mental health crises. For example, if the level of education affects mental health outcomes, then we can suggest more plans to keep people in education for longer. When it comes to different health conditions, if a certain cancer or illness leads to a higher correlation of mental health problems, then we would know the relative target area to help stop and find solutions to mitigate this issue and lower overall mental health problems.

Part 3 - Pre-Processing

Removing Attributes with >70% Missing Values

Utilizing the Weka software all attributes with missing percentages greater than 70% were subsequently removed. This is done for a variety of reasons:

1. **Lack of Information:** If 70% of the data is missing, the attribute provides little value or information to the class variable.
2. **Difficulty in imputation:** Attempting to fill in so much missing data would introduce noise or biases rather than useful insight, leading to poor model performance
3. **The Curse of Dimensionality:** Keeping too many incomplete features increases the complexity of the model without any large improvement. Removing these features simplifies the model
4. **Avoiding overfitting:** Filling in a lot of the missing values could result in the model overfitting to the data, where the model learns the values from the filled in values rather than the actual values.

This resulted in the removal of 311 attributes, leaving values with more data useful for predicting the class. While this seems like a lot, we still have 336 attributes to preprocess.

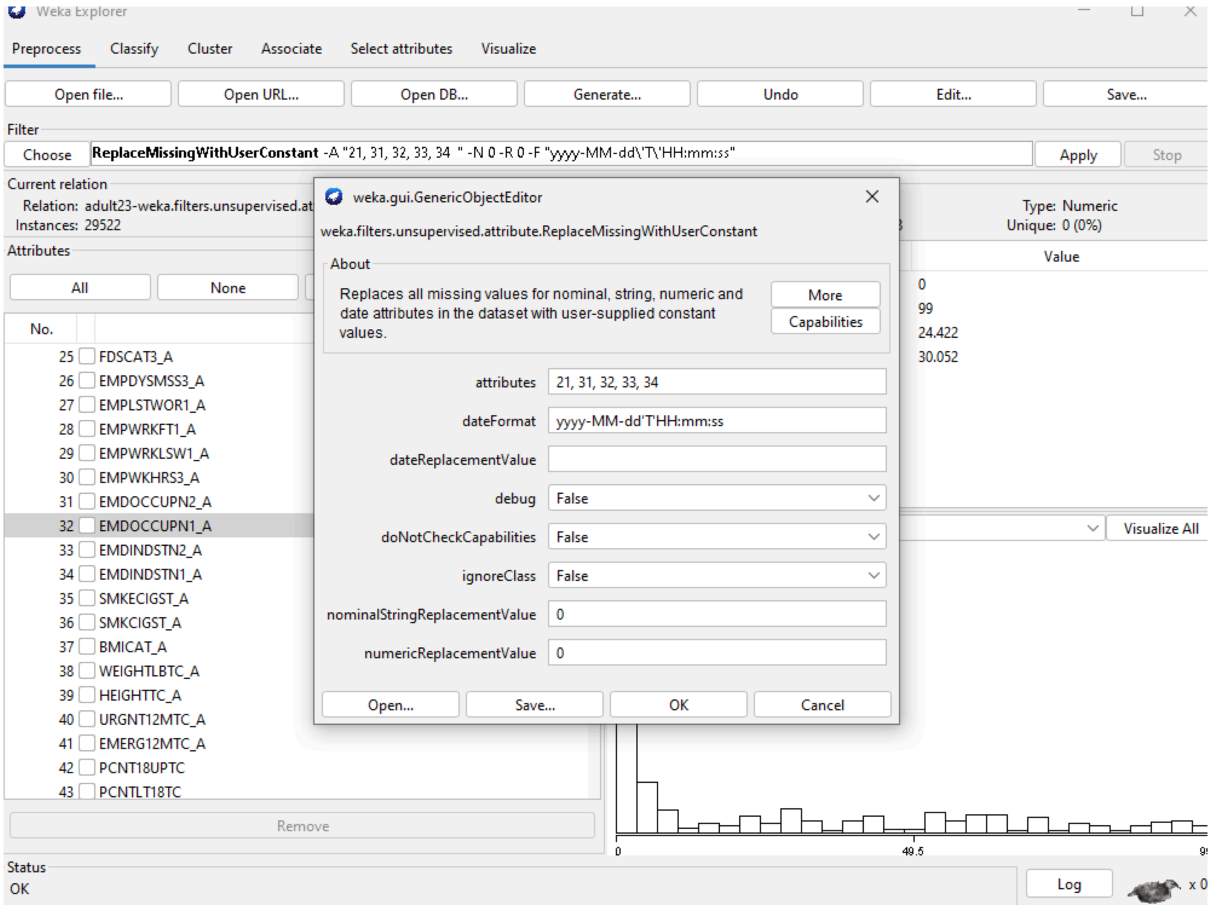
Filling In Default Values

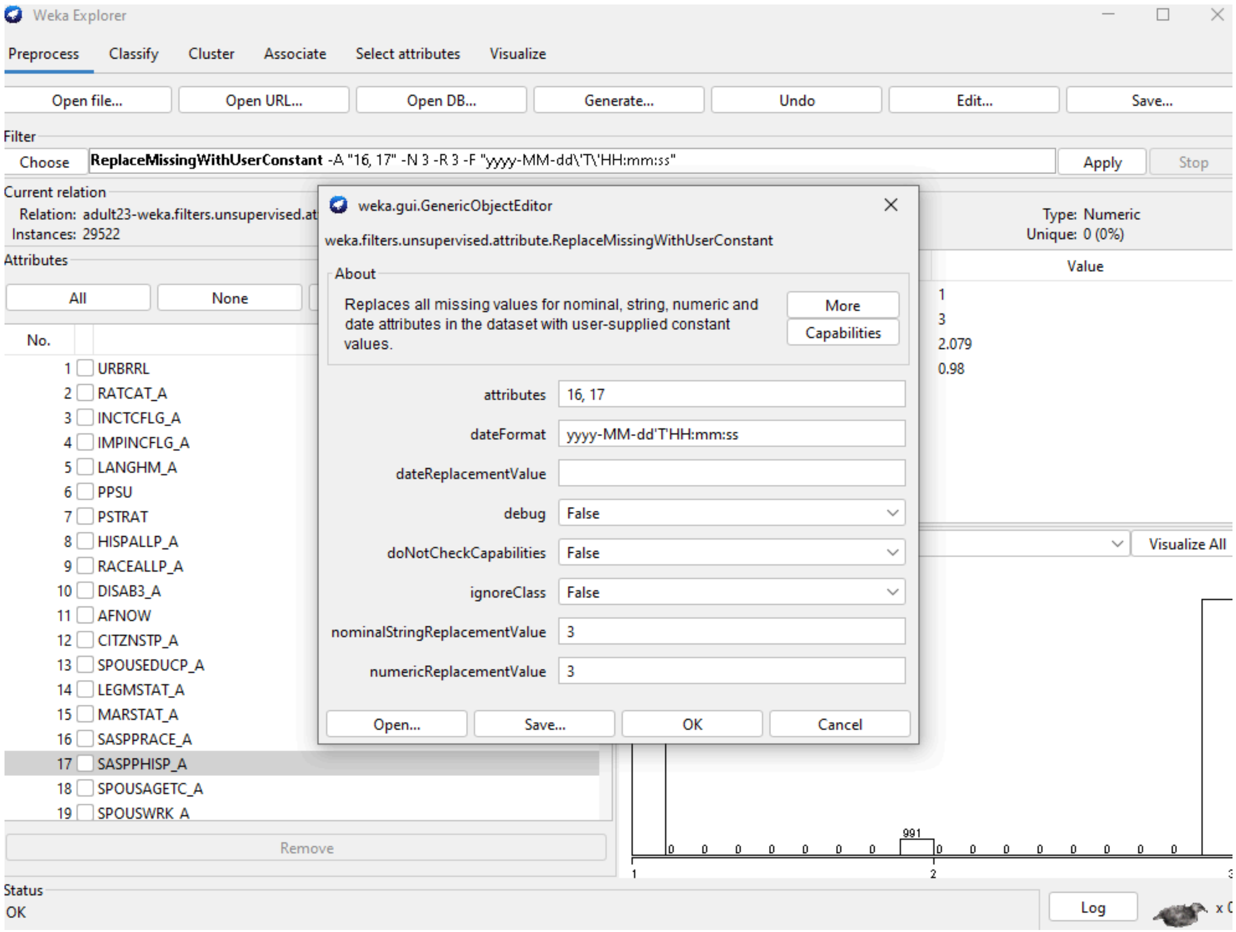
All attributes with missing values have a default value meaning “Don’t Know” according to the codebook. To fill in these values, each attribute with a missing value was compared against the codebook to find this default value, the following attributes have the given default value, which was replaced to remove all missing values from the dataset.

Specified Attributes Index	Default Value
21, 31, 32, 33, 34	0
16, 17	3
47	6
46	7
48, 60	8
5, 11, 19, 20, 27, 28, 87, 89, 90, 91, 92, 93, 96, 98, 99, 100, 101, 102, 104, 109, 110, 116, 117, 122, 125, 127, 136, 142, 143, 144, 145, 146, 147, 148,	9

149, 150, 168, 169, 170, 179, 181, 182, 183, 185, 188, 193, 194, 195, 199, 200, 201, 202, 203, 204, 205, 206, 207, 208, 209, 210, 211, 212, 213, 214, 218, 219, 220, 222, 223, 224, 225, 231, 233, 238, 242, 243, 244, 245, 246, 247, 248, 249, 250, 251, 252, 253, 254, 255, 256, 257, 258, 259, 282, 283, 284, 291, 292, 293, 310, 314, 315, 320, 321, 323, 324, 325	
13, 18, 30, 50, 62, 103, 186, 187, 191	99
27	999
190	9999
51	99999

These values were filled in using Weka's ***ReplaceMissingWithUserConstant*** feature. After completion, there were no more missing values within the dataset.





Engineering Class Attribute

To create the class variable, we combined the following 4 variables about questions used in a mental health assessment. These attributes are similar to each other such as feeling down and having little interest in things for the past 2 weeks.

MHA: Brief mental health assessment

Variable #	Question #	Variable Name	Source Variables	Description	Type	Location	Length
1	MHA.0020.00.4	PHQ41_A		How often little interest in things, past 2 weeks	Num	518	1
2	MHA.0030.00.4	PHQ42_A		How often feeling down, past 2 weeks	Num	519	1
3	Recode	PHQ2SCREEN_A	PHQ41_A; PHQ42_A	PHQ-2 screener result	Num	520	1
4	MHA.0040.00.4	PHQ43_A		How often felt nervous/anxious/on edge, past 2 weeks	Num	521	1
5	MHA.0050.00.4	PHQ44_A		How often can't stop/control worrying, past 2 weeks	Num	522	1
6	Recode	GAD2SCREEN_A	PHQ43_A; PHQ44_A	GAD-2 screener result	Num	523	1

The codes for these 4 attributes breakdown as so:

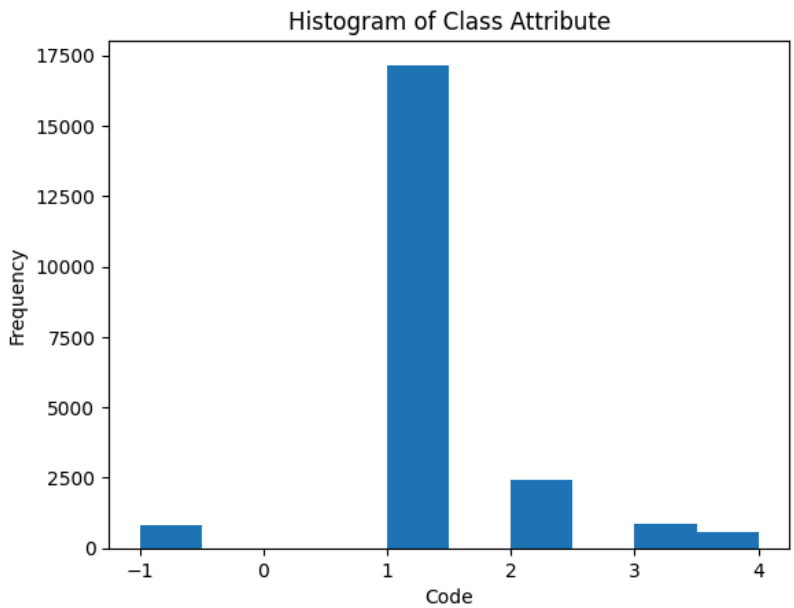
Code	Description
1	Not at all
2	Several days
3	More than half the days
4	Nearly every day
7	Refused
8	Not Ascertained
9	Don't Know

There are multiple approaches we could have taken to combine these 4 attributes. All 4 attributes have 0 missing values, so there is no risk of bias. The values placed in the engineered attribute will be as follows: (1) Average then round values between 2-4 to get the average number of days where the individual faced a mental health issue, (2) 1 if all values are coded as 1 and, -1 if all values are between values 7-9.

This is a complex expression, so we opted to utilize Python via Google Colaboratory to alter the data frame. We utilized Pandas to manipulate the data frame.

```
def engineer_attribute(row):

 """
 Engineers a new attribute based on PHQ41_A, PHQ42_A, PHQ43_A, PHQ44_A.
 row: A pandas Series representing a row in the DataFrame.
 Returns: The engineered attribute value.
 """

 phq_values = [row['PHQ41_A'], row['PHQ42_A'], row['PHQ43_A'], row['PHQ44_A']]
 if all(1 <= val <= 1 and not np.isnan(val) for val in phq_values):
   return 1
 elif all(7 <= val <= 9 and not np.isnan(val) for val in phq_values):
   return -1
 else:
   average = np.nanmean(phq_values)
   if 2 <= average <= 4:
     return round(average)
   else:
     return np.nan  # Or another default value if needed

# Apply the function to create a new column
df['engineered_attribute'] = df.apply(engineer_attribute, axis=1)

# Print the DataFrame with the new column
print(df[['PHQ41_A', 'PHQ42_A', 'PHQ43_A', 'PHQ44_A', 'engineered_attribute']])
```

This results in this new distribution of the dataset.



Training/Validation/Test Dataset

For the next steps of figuring out which features are important and which to choose to build the model, the data is split into a training, testing, and validation dataset. We did this using Python sklearn's train_test_split function.

```
from sklearn.model_selection import train_test_split

# Split the data into training and a temporary set (test + validate)
train_df, temp_df = train_test_split(df, test_size=0.3, random_state=42)

# Split the temporary set into test and validate sets
test_df, validate_df = train_test_split(temp_df, test_size=0.5, random_state=42)

# Now you have three DataFrames: train_df, test_df, and validate_df
print(f"Train size: {len(train_df)}")
print(f"Test size: {len(test_df)}")
print(f"Validate size: {len(validate_df)}")
```

Now that these three have been created, the study can continue.

Part 4 - Attribute Selection Algorithms and Model Classifiers Used

Now that the data has been preprocessed, the most important features can be extracted to build the final algorithm. Below we tested a multitude of algorithms.

CfsSubsetEval with BestFirst

The evaluator selects feature subsets based on their predictive ability and redundancy. The key idea is to find features that have high correlation with the class (predictive power) and have low correlation with each other (redundancy).

This method works on the assumption that a good subset of features contains features highly correlated with the target class but uncorrelated with each other. This is calculated by getting the 'merit' of each subset.

$$r_{zc} = \frac{k\overline{r_{zi}}}{\sqrt{k + k(k-1)\overline{r_{ii}}}}$$

Where r_{zc} is the merit of subset C

k = number of features in the subset

bar(r_{zi}*)* is the average correlation between the class and selected features

bar(r_{ii}*)* is the average correlation among the selected features

Best first is a search strategy used to explore the space of possible feature subsets. It's essentially a greedy hill-climbing algorithm that can backtrack if necessary. The algorithm searches for the best subset by evaluating neighboring subsets (adding removing one feature at a time) and selects the best one based on the CfsSubsetEval merit score. There are three steps:

1. Forward Selection: Empty set of features and add features incrementally
2. Backward Elimination: Start will all features and remove one by one
3. Bidirectional Search: Combine both forward and backward elimination

Correlation (Non-Weka Attribute Selection)

This process was done in Google Colaboratory by first calculating the Spearman rank correlation index of each attribute in relation to the class "engineered attribute". Then by taking the absolute value of the correlation coefficients, we render a list of which attributes are most correlated with the class variable. From this list we take the top 10, around a threshold of 0.6 on both the positive and negative end.

$$\rho = 1 - \frac{6\Sigma\, \mathrm{d}_i^2}{n(n^2-1)}$$

```
# Calculate Spearman rank correlations
correlations = df.corr(method='spearman')['engineered_attribute'].drop('engineered_attribute')
```

```
# Take the absolute value of the correlations
absolute_correlations = correlations.abs()

# Get the top 10 most correlated attributes
top_10_correlations = absolute_correlations.nlargest(10)

print(top_10_correlations)
```

CorrelationAttributeEval with Ranker

CorrelationAttributeEval evaluates attributes based on their correlation with the class label (for classification tasks). It measures the relationship between each feature and target class. This evaluation is done independently for each attribute, meaning that it calculates the correlation of one attribute at a time with the class. In our case, it uses the Pearson correlation.

$$r = \frac{\sum (x - m_x)(y - m_y)}{\sqrt{\sum (x - m_x)^2 \sum (y - m_y)^2}}$$

The ranker search method ranks attributes based on their individual merit (correlation scores). They are sorted in descending order of their evaluation score, meaning the most predictive attribute appears at the top of the list. Ranker can also be set to eliminate attributes under a certain threshold, ensuring the resulting attributes are actually useful for prediction.

GainRatioAttributeEval with Ranker

We used Weka for this approach. It utilized the following formulas to calculate the GainRatio for a particular attribute

$$\text{GainRatio(A)} = \text{Gain(A)/SplitInfo}_{\text{A}}\text{(D)}$$

$$SplitInfo_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2\left(\frac{|D_j|}{|D|}\right)$$

$$Gain(A) = Info(D) - Info_A(D)$$

Where the expected information (or entropy) needed to classify a value in D is

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

where p_i is the probability that a value D is part of class C, m here represents the number of classes.

$Info_A$ uses attribute A to split D into v partitions before using that information to put D in a class:

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j)$$

v here is the number of unique values in attribute 5.

SymmetricUncertAttributeEval with Ranker

SymmetricUncertAttributeEval is an attribute evaluator based on the concept of Symmetrical Uncertainty (SU), which is derived from information gain (IG). This works in combination with a search method like Ranker. Entropy is a measure of the uncertainty or randomness in data. Information gain measures the reduction in entropy when an attribute X is shown. Symmetrical uncertainty is a normalized form of information gain, designed to remove biases towards attributes with many values. Ranker uses these normalized merit scores to rank the top N attributes, which can be selected using a threshold.

Part 5 - Selection Algorithm Results

CfsSubsetEval with BestFirst

```
Search Method:
        Best first.
        Start set: no attributes
        Search direction: forward
        Stale search after 5 node expansions
        Total number of subsets evaluated: 5616
        Merit of best subset found:    0.723

Attribute Subset Evaluator (supervised, Class (numeric): 321 engineered_attribute):
        CFS Subset Evaluator
        Including locally predictive attributes

Selected attributes: 10,63,107,134,140,150,154,159,160,189

                DISAB3_A
                MLTFAMFLG_A
                EVRMARRIED_A
                SMKNOW_A
                PAITOOTH3M_A
                VIGIL4_A
                DISCRIM5_A
                MHTHND_A
                MHTHDLY_A
                HYSTEV2_A
```

Correlation (Non-Weka Attribute Selection)

```
0.634879   121 HRTESTLAST_A
0.610201    26 EMPDYSMSS3_A
0.596229     1 URBRRL
0.593039    70 REGION
0.592348     7 PSTRAT
0.590748   224 LONGCOVD1_A
0.587731   150 VIGIL4_A
0.58613    134 SMKNOW_A
0.582264    71 INTV_QRT
0.579129   151 VIGIL3_A
```

CorrelationAttributeEval with Ranker

All values with correlation above 0.3 (positive) and greater than -0.4 (negative) are included, as those are highly correlated. These are the attributes that are best for predicting for the class.

```
Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (numeric): 321 engineered_attribute):
        Correlation Ranking Filter
Ranked attributes:
 0.633966   261 SOCSCLPAR_A
 0.633344   266 COGMEMDFF_A
 0.568578   311 LSATIS4_A
 0.538037   312 PHSTAT_A
 0.511334   262 SOCERRNDS_A
… [omitted for length]
-0.703882   154 DISCRIM5_A
-0.724287   150 VIGIL4_A
-0.745768   160 MHTHDLY_A
-0.747032   159 MHTHND_A
```

GainRatioAttributeEval with Ranker

```
Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 321 engineered_attribute):
        Gain Ratio feature evaluator
```

```
Ranked attributes:
 0.697292    159 MHTHND_A
 0.6864      160 MHTHDLY_A
 0.632153    261 SOCSCLPAR_A
 0.510287    136 TBIHLSBMC_A
 0.506297    318 WTFA_A
 0.505223    137 TBILCDCMG_A
 0.50145     161 HOMEHC12M_A
 0.593591    203 RXDG12M_A
 0.590783    214 MEDNG12M_A
 0.589613    215 MEDDL12M_A
[... omitted for length]

Selected attributes: 159,160,261,136,318,137,161,203,214,215
```

SymmetricUncertAttributeEval with Ranker

```
Search Method:
                    Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 321 engineered_attribute):
                    Symmetrical Uncertainty Ranking Filter

Ranked attributes:
 0.564391    322 FAM_A
 0.192102    318 WTFA_A
 0.19129     159 MHTHND_A
 0.088267    160 MHTHDLY_A
 0.083587    261 SOCSCLPAR_A
 0.078292    150 VIGIL4_A
 0.09597     311 LSATIS4_A
 0.0568936   156 DISCRIM3_A
 0.064919    266 COGMEMDFF_A
 0.063711    154 DISCRIM5_A
 0.061754    153 VIGIL1_A

[... omitted for length]

Selected attributes: 318,159,160,261,150,311,156,266,154,153,158
```

This leaves the following selected attributes for each attribute selection algorithm, these will from this point in the report be referred to as:
cfsEVAL - 10,63,107,134,140,150,154,159,160,189

reliefEVAL - 121, 26, 1, 70, 7, 224, 150, 134, 71, 151
corrEVAL - 261, 266, 311, 312, 262, 154, 150, 160, 159
gainEVAL - 159,160,261,136,318,137,161,203,214,215
symmEVAL - 318,159,160,261,150,311,156,266,154,153,158

Part 6 - Model Selection

We chose the following 4 models to test the selected attributes:

Decision Table

Decision tables are concise visual representations of which actions to perform based on a given dataset. The structure of a decision table is a condition that is the inputs or features of the model. Each row responds to a combination of feature values, an action/decision that outputs the result of applying the model to the conditions which is the class model, and rules which are the specific values/conditions in which the decision is made which becomes a rule. A decision table is considered balanced if it includes every possible combination of the input variables. This model is good for interpretation and creates rules that might be valuable for analysis, however, it is not suitable for continuous variables and may be impractical for complex models.

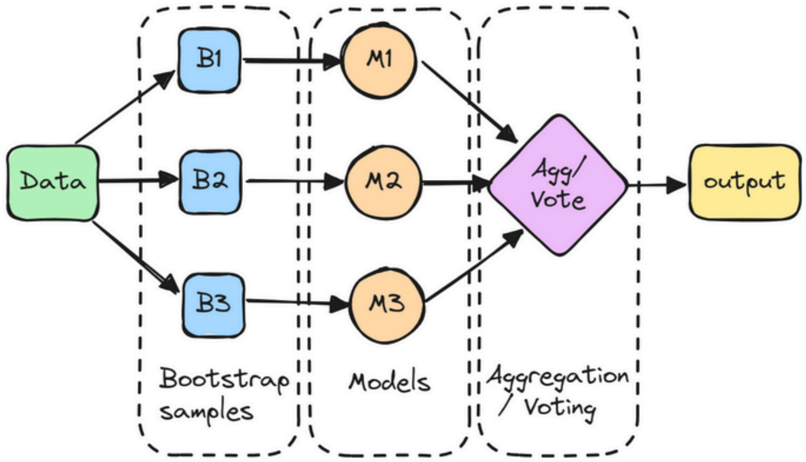
Age Group	Income Level	Purchase (Yes/No)
Young	Low	No
Young	High	Yes
Middle-aged	Low	Yes
Middle-aged	High	Yes
Senior	Low	No
Senior	High	No

J48

This classifier is a subset of existing decision tree algorithms. It is an open-source Java implementation of the C4.5 decision tree algorithm. It is similar to decision tables however it uses a recursive process to build the tree. It also uses information gain to measure how much a feature reduces the uncertainty for the class label. Due to its recursive nature, J48 is great for larger datasets as well as handling both categorical and continuous variables. However, it is prone to overfitting and bias towards certain features due to the use of gain ratios.

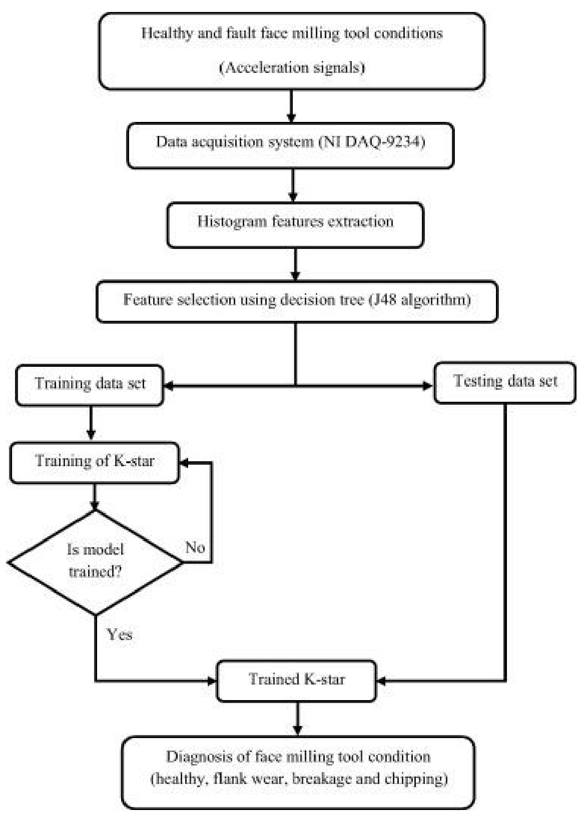
Bagging

Bagging involves training multiple models independently on different subsets of the data. First, the data will be randomly sampled an n amount of times with replacement. Then the model will train it on each of the data samples which would then create predictions. The models' predictions will be combined through simple averaging to make an overall prediction. Bagging can reduce variance and improve stability because it trains on multiple different models with different subsets of data. However, this leads to an increased computation because requires training of multiple models as well as not being useful for low variance models.



KStar

KStar is an instance-based classification model where it stores all the training data and makes predictions only when a new instance is classified. It makes decisions based on distances between new instances and the stored training data. Instances that are closer to the new instance have more influence and more weight on the final prediction. K star is good for complex distributions of data as well as categorical and continuous data. However, it is complex and may be slow for large datasets as well as being memory intensive because it saves all the training instances.



Part 7 - Results

Decision Table on cfsEVAL

```
=== Summary ===

Correctly Classified Instances       18343               88.7636 %
Incorrectly Classified Instances      2322               11.2364 %
Kappa statistic                          0.8051
Mean absolute error                      0.074
Root mean squared error                  0.1876
Relative absolute error                 32.3704 %
Root relative squared error             55.5117 %
Total Number of Instances            20665

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.875    0.039    0.900      0.875    0.887      0.843    0.958     0.904     -1
                 0.919    0.093    0.932      0.919    0.926      0.824    0.963     0.970     1
                 0.856    0.034    0.694      0.856    0.766      0.748    0.960     0.747     2
                 0.635    0.007    0.717      0.635    0.674      0.666    0.963     0.638     3
                 0.639    0.007    0.645      0.639    0.642      0.635    0.962     0.631     4
Weighted Avg.    0.888    0.069    0.891      0.888    0.889      0.815    0.961     0.917

=== Confusion Matrix ===

     a     b     c     d     e   <-- classified as
  5199   459   182    52    47 |     a = -1
   453 11046   415    51    53 |     b = 1
    66   114  1458    38    27 |     c = 2
    24   153    25   380    16 |     d = 3
    37    79    22     9   260 |     e = 4
```

J48 on cfsEVAL

```
=== Summary ===

Correctly Classified Instances       18381               88.9475 %
Incorrectly Classified Instances      2284               11.0525 %
Kappa statistic                          0.8059
Mean absolute error                      0.0686
Root mean squared error                  0.1879
Relative absolute error                 30.0186 %
Root relative squared error             55.5871 %
Total Number of Instances            20665

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.877    0.041    0.895      0.877    0.886      0.841    0.946     0.865     -1
                 0.932    0.109    0.922      0.932    0.927      0.824    0.950     0.955     1
                 0.776    0.023    0.754      0.776    0.765      0.743    0.943     0.743     2
                 0.664    0.009    0.694      0.664    0.679      0.669    0.945     0.630     3
                 0.636    0.006    0.673      0.636    0.654      0.648    0.917     0.533     4
Weighted Avg.    0.889    0.077    0.889      0.889    0.889      0.814    0.948     0.894

=== Confusion Matrix ===

     a     b     c     d     e   <-- classified as
  5208   471   148    59    53 |     a = -1
   475 11195   244    65    39 |     b = 1
    72   250  1322    39    20 |     c = 2
    24   139    24   397    14 |     d = 3
    38    82    16    12   259 |     e = 4
```

Bagging on cfsEVAL

```
=== Summary ===

Correctly Classified Instances        6228               88.6422 %
Incorrectly Classified Instances       798               11.3578 %
Kappa statistic                          0.8027
Mean absolute error                      0.0694
Root mean squared error                  0.1877
Relative absolute error                 30.3482 %
Root relative squared error             55.4338 %
Total Number of Instances             7026

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.882    0.046    0.886      0.882    0.884      0.837    0.957     0.903     -1
                 0.921    0.097    0.929      0.921    0.925      0.822    0.962     0.974     1
                 0.812    0.026    0.736      0.812    0.772      0.752    0.956     0.759     2
                 0.675    0.010    0.662      0.675    0.668      0.659    0.944     0.667     3
                 0.594    0.006    0.676      0.594    0.632      0.626    0.953     0.656     4
Weighted Avg.    0.886    0.072    0.887      0.886    0.887      0.812    0.959     0.920

=== Confusion Matrix ===

    a    b    c    d    e   <-- classified as
 1787  142   54   24   20 |    a = -1
  179 3749   99   30   14 |    b = 1
   25   64  465   12    7 |    c = 2
   14   41    7  135    3 |    d = 3
   12   41    7    3   92 |    e = 4
```

KStar on cfsEVAL

```
=== Summary ===

Correctly Classified Instances        6044               86.0233 %
Incorrectly Classified Instances       982               13.9767 %
Kappa statistic                          0.7409
Mean absolute error                      0.1013
Root mean squared error                  0.2096
Relative absolute error                 44.2851 %
Root relative squared error             61.8989 %
Total Number of Instances             7026

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.858    0.044    0.889      0.858    0.873      0.823    0.961     0.907     -1
                 0.956    0.224    0.854      0.956    0.902      0.756    0.965     0.977     1
                 0.503    0.011    0.809      0.503    0.620      0.614    0.960     0.736     2
                 0.370    0.003    0.763      0.370    0.498      0.523    0.966     0.597     3
                 0.335    0.001    0.839      0.335    0.479      0.525    0.956     0.600     4
Weighted Avg.    0.860    0.144    0.858      0.860    0.850      0.752    0.963     0.918

=== Confusion Matrix ===

    a    b    c    d    e   <-- classified as
 1740  251   27    6    3 |    a = -1
  149 3890   26    5    1 |    b = 1
   31  244  288    6    4 |    c = 2
   18  101    5   74    2 |    d = 3
   20   67   10    6   52 |    e = 4
```

Decision Table on reliefEVAL

```
=== Summary ===

Correctly Classified Instances        6186               88.0444 %
Incorrectly Classified Instances       840               11.9556 %
Kappa statistic                          0.7935
Mean absolute error                      0.0757
Root mean squared error                  0.1913
Relative absolute error                 33.1173 %
Root relative squared error             56.5101 %
Total Number of Instances             7026

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.881    0.049    0.879      0.881    0.880      0.831    0.958     0.894     -1
                 0.910    0.093    0.931      0.910    0.921      0.814    0.958     0.966     1
                 0.864    0.033    0.698      0.864    0.772      0.755    0.959     0.739     2
                 0.570    0.008    0.671      0.570    0.616      0.608    0.964     0.601     3
                 0.555    0.007    0.632      0.555    0.591      0.584    0.968     0.521     4
Weighted Avg.    0.880    0.071    0.883      0.880    0.881      0.803    0.958     0.906

=== Confusion Matrix ===

    a    b    c    d    e   <-- classified as
 1785  139   60   20   23 |    a = -1
  183 3706  142   24   16 |    b = 1
   27   33  495    9    9 |    c = 2
   20   57    7  114    2 |    d = 3
   15   46    5    3   86 |    e = 4
```

J48 on reliefEVAL

```
=== Summary ===

Correctly Classified Instances        6184               88.0159 %
Incorrectly Classified Instances       842               11.9841 %
Kappa statistic                          0.7932
Mean absolute error                      0.0757
Root mean squared error                  0.1939
Relative absolute error                 33.0887 %
Root relative squared error             57.2844 %
Total Number of Instances             7026

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.875    0.045    0.887      0.875    0.881      0.833    0.950     0.867     -1
                 0.911    0.094    0.930      0.911    0.921      0.814    0.946     0.950     1
                 0.866    0.034    0.694      0.866    0.770      0.753    0.943     0.640     2
                 0.570    0.008    0.671      0.570    0.616      0.608    0.941     0.598     3
                 0.587    0.009    0.595      0.587    0.591      0.582    0.952     0.516     4
Weighted Avg.    0.880    0.071    0.884      0.880    0.881      0.804    0.947     0.881

=== Confusion Matrix ===

    a    b    c    d    e   <-- classified as
 1773  139   65   20   30 |    a = -1
  175 3710  142   24   20 |    b = 1
   25   33  496    9   10 |    c = 2
   19   58    7  114    2 |    d = 3
    8   48    5    3   91 |    e = 4
```

Bagging on reliefEVAL

```
=== Summary ===

Correctly Classified Instances        6058               86.2226 %
Incorrectly Classified Instances       968               13.7774 %
Kappa statistic                          0.7613
Mean absolute error                      0.0782
Root mean squared error                  0.204
Relative absolute error                 34.2046 %
Root relative squared error             60.2582 %
Total Number of Instances             7026

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.872    0.053    0.871      0.872    0.871      0.819    0.945     0.871     -1
                 0.901    0.115    0.915      0.901    0.908      0.783    0.942     0.954     1
                 0.815    0.033    0.687      0.815    0.745      0.724    0.916     0.656     2
                 0.550    0.011    0.591      0.550    0.570      0.558    0.937     0.520     3
                 0.303    0.011    0.385      0.303    0.339      0.329    0.924     0.347     4
Weighted Avg.    0.862    0.085    0.863      0.862    0.862      0.772    0.940     0.880

=== Confusion Matrix ===

    a    b    c    d    e   <-- classified as
 1768  139   63   26   31 |    a = -1
  192 3666  136   40   37 |    b = 1
   30   61  467    8    7 |    c = 2
   22   62    6  110    0 |    d = 3
   19   79    8    2   47 |    e = 4
```

KStar on reliefEVAL

```
=== Summary ===

Correctly Classified Instances        5316               75.6618 %
Incorrectly Classified Instances      1710               24.3382 %
Kappa statistic                          0.5398
Mean absolute error                      0.1212
Root mean squared error                  0.2675
Relative absolute error                 52.979  %
Root relative squared error             79.0162 %
Total Number of Instances             7026

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.675    0.089    0.755      0.675    0.713      0.607    0.886     0.767     -1
                 0.915    0.350    0.783      0.915    0.844      0.597    0.892     0.904     1
                 0.316    0.021    0.569      0.316    0.406      0.388    0.856     0.418     2
                 0.140    0.008    0.337      0.140    0.198      0.203    0.832     0.206     3
                 0.077    0.006    0.231      0.077    0.116      0.123    0.802     0.151     4
Weighted Avg.    0.757    0.230    0.733      0.757    0.736      0.561    0.884     0.789

=== Confusion Matrix ===

    a    b    c    d    e   <-- classified as
 1369  570   47   25   16 |    a = -1
  257 3726   62   13   13 |    b = 1
   95  275  181   14    8 |    c = 2
   49  109   11   28    3 |    d = 3
   43   80   17    3   12 |    e = 4
```

Decision Table on corrEVAL

```
=== Summary ===

Correctly Classified Instances        6029               85.8098 %
Incorrectly Classified Instances       997               14.1902 %
Kappa statistic                          0.7693
Mean absolute error                      0.1029
Root mean squared error                  0.2275
Relative absolute error                 39.8257 %
Root relative squared error             63.1198 %
Total Number of Instances             7026

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.906    0.056    0.860      0.906    0.882      0.837    0.924     0.811     -1
                 0.973    0.172    0.854      0.973    0.910      0.812    0.901     0.847     1
                 0.694    0.012    0.866      0.694    0.770      0.753    0.843     0.646     2
                 0.441    0.004    0.870      0.441    0.585      0.605    0.705     0.429     3
                 0.331    0.002    0.897      0.331    0.483      0.530    0.669     0.353     4
Weighted Avg.    0.858    0.104    0.860      0.858    0.844      0.784    0.876     0.763

=== Confusion Matrix ===

    a    b    c    d    e   <-- classified as
 1743  151   19    7    4 |    a = -1
   65 3477   19    7    4 |    b = 1
   65  142  496    7    5 |    c = 2
   58  142   18  174    3 |    d = 3
   96  159   21    5  139 |    e = 4
```

J48 on corrEVAL

```
=== Summary ===

Correctly Classified Instances        6029               85.8098 %
Incorrectly Classified Instances       997               14.1902 %
Kappa statistic                          0.7693
Mean absolute error                      0.1025
Root mean squared error                  0.2275
Relative absolute error                 39.6781 %
Root relative squared error             63.1127 %
Total Number of Instances             7026

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.906    0.056    0.860      0.906    0.882      0.837    0.924     0.810     -1
                 0.973    0.172    0.854      0.973    0.910      0.812    0.901     0.846     1
                 0.694    0.012    0.866      0.694    0.770      0.753    0.843     0.646     2
                 0.441    0.004    0.870      0.441    0.585      0.605    0.705     0.429     3
                 0.331    0.002    0.897      0.331    0.483      0.530    0.669     0.353     4
Weighted Avg.    0.858    0.104    0.860      0.858    0.844      0.784    0.876     0.763

=== Confusion Matrix ===

    a    b    c    d    e   <-- classified as
 1743  151   19    7    4 |    a = -1
   65 3477   19    7    4 |    b = 1
   65  142  496    7    5 |    c = 2
   58  142   18  174    3 |    d = 3
   96  159   21    5  139 |    e = 4
```

Bagging on corrEVAL

```
=== Summary ===

Correctly Classified Instances        6029               85.8098 %
Incorrectly Classified Instances       997               14.1902 %
Kappa statistic                          0.7693
Mean absolute error                      0.1025
Root mean squared error                  0.2286
Relative absolute error                 39.6624 %
Root relative squared error             63.4071 %
Total Number of Instances             7026

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.906    0.056    0.860      0.906    0.882      0.837    0.925     0.813     -1
                 0.973    0.172    0.854      0.973    0.910      0.812    0.902     0.849     1
                 0.692    0.012    0.867      0.692    0.770      0.753    0.838     0.650     2
                 0.441    0.004    0.870      0.441    0.585      0.605    0.721     0.463     3
                 0.333    0.003    0.892      0.333    0.485      0.531    0.665     0.387     4
Weighted Avg.    0.858    0.104    0.860      0.858    0.844      0.784    0.877     0.770

=== Confusion Matrix ===

    a    b    c    d    e   <-- classified as
 1743  151   19    7    4 |    a = -1
   65 3477   19    7    4 |    b = 1
   65  142  495    7    6 |    c = 2
   58  142   18  174    3 |    d = 3
   96  159   20    5  140 |    e = 4
```

KStar on corrEVAL

```
=== Summary ===

Correctly Classified Instances        5783               82.3086 %
Incorrectly Classified Instances      1243               17.6914 %
Kappa statistic                          0.7058
Mean absolute error                      0.1419
Root mean squared error                  0.2546
Relative absolute error                 54.9056 %
Root relative squared error             70.6392 %
Total Number of Instances             7026

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.903    0.062    0.846      0.903    0.874      0.825    0.921     0.809     -1
                 0.974    0.234    0.811      0.974    0.885      0.758    0.902     0.855     1
                 0.533    0.013    0.819      0.533    0.646      0.632    0.844     0.636     2
                 0.266    0.003    0.861      0.266    0.406      0.464    0.722     0.410     3
                 0.195    0.003    0.820      0.195    0.315      0.385    0.651     0.312     4
Weighted Avg.    0.823    0.138    0.825      0.823    0.797      0.725    0.876     0.762

=== Confusion Matrix ===

    a    b    c    d    e   <-- classified as
 1737  162   17    6    2 |    a = -1
   68 3478   17    7    2 |    b = 1
   71  257  381    2    4 |    c = 2
   66  196   18  105   10 |    d = 3
  111  193   32    2   82 |    e = 4
```

Decision Table on gainEVAL

```
=== Summary ===

Correctly Classified Instances        6232               88.6991 %
Incorrectly Classified Instances       794               11.3009 %
Kappa statistic                          0.8028
Mean absolute error                      0.0745
Root mean squared error                  0.1887
Relative absolute error                 32.5629 %
Root relative squared error             55.7469 %
Total Number of Instances             7026

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.871    0.036    0.907      0.871    0.889      0.845    0.957     0.885     -1
                 0.923    0.114    0.918      0.923    0.921      0.811    0.959     0.962     1
                 0.866    0.033    0.697      0.866    0.772      0.755    0.961     0.732     2
                 0.555    0.004    0.799      0.555    0.655      0.658    0.969     0.628     3
                 0.645    0.005    0.758      0.645    0.697      0.693    0.960     0.640     4
Weighted Avg.    0.887    0.079    0.890      0.887    0.887      0.809    0.959     0.904

=== Confusion Matrix ===

    a    b    c    d    e   <-- classified as
 1766  179   62    8   12 |    a = -1
  152 3759  142   10    8 |    b = 1
   19   40  496    8   10 |    c = 2
    7   73    7  111    2 |    d = 3
    4   44    5    2  100 |    e = 4
```

J48 on gainEVAL

```
=== Summary ===

Correctly Classified Instances        6029               85.8098 %
Incorrectly Classified Instances       997               14.1902 %
Kappa statistic                          0.7693
Mean absolute error                      0.1025
Root mean squared error                  0.2275
Relative absolute error                 39.6781 %
Root relative squared error             63.1127 %
Total Number of Instances             7026

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.906    0.056    0.860      0.906    0.882      0.837    0.924     0.810     -1
                 0.973    0.172    0.854      0.973    0.910      0.812    0.901     0.846     1
                 0.694    0.012    0.866      0.694    0.770      0.753    0.843     0.646     2
                 0.441    0.004    0.870      0.441    0.585      0.605    0.705     0.429     3
                 0.331    0.002    0.897      0.331    0.483      0.530    0.669     0.353     4
Weighted Avg.    0.858    0.104    0.860      0.858    0.844      0.784    0.876     0.763

=== Confusion Matrix ===

    a    b    c    d    e   <-- classified as
 1743  151   19    7    4 |    a = -1
   65 3477   19    7    4 |    b = 1
   65  142  496    7    5 |    c = 2
   58  142   18  174    3 |    d = 3
   96  159   21    5  139 |    e = 4
```

Bagging on gainEVAL

```
=== Summary ===

Correctly Classified Instances        6029               85.8098 %
Incorrectly Classified Instances       997               14.1902 %
Kappa statistic                          0.7693
Mean absolute error                      0.1025
Root mean squared error                  0.2286
Relative absolute error                 39.6624 %
Root relative squared error             63.4071 %
Total Number of Instances             7026

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.906    0.056    0.860      0.906    0.882      0.837    0.925     0.813     -1
                 0.973    0.172    0.854      0.973    0.910      0.812    0.902     0.849     1
                 0.692    0.012    0.867      0.692    0.770      0.753    0.838     0.650     2
                 0.441    0.004    0.870      0.441    0.585      0.605    0.721     0.463     3
                 0.333    0.003    0.892      0.333    0.485      0.531    0.665     0.387     4
Weighted Avg.    0.858    0.104    0.860      0.858    0.844      0.784    0.877     0.770

=== Confusion Matrix ===

    a    b    c    d    e   <-- classified as
 1743  151   19    7    4 |    a = -1
   65 3477   19    7    4 |    b = 1
   65  142  495    7    6 |    c = 2
   58  142   18  174    3 |    d = 3
   96  159   20    5  140 |    e = 4
```

KStar on gainEVAL

```
=== Summary ===

Correctly Classified Instances        5999               85.3829 %
Incorrectly Classified Instances      1027               14.6171 %
Kappa statistic                          0.7241
Mean absolute error                      0.1051
Root mean squared error                  0.2099
Relative absolute error                 45.9521 %
Root relative squared error             62.0032 %
Total Number of Instances             7026

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.873    0.038    0.902      0.873    0.887      0.843    0.959     0.899     -1
                 0.959    0.264    0.833      0.959    0.892      0.727    0.959     0.962     1
                 0.405    0.005    0.885      0.405    0.556      0.578    0.959     0.752     2
                 0.275    0.002    0.809      0.275    0.410      0.464    0.968     0.569     3
                 0.252    0.002    0.765      0.252    0.379      0.432    0.954     0.595     4
Weighted Avg.    0.854    0.164    0.855      0.854    0.838      0.735    0.959     0.908

=== Confusion Matrix ===

    a    b    c    d    e   <-- classified as
 1770  239   10    3    5 |    a = -1
  153 3903   12    2    1 |    b = 1
   20  313  232    4    4 |    c = 2
    8  133    2   55    2 |    d = 3
   11   95    6    4   39 |    e = 4
```

Decision Table on symmEVAL

```
=== Summary ===

Correctly Classified Instances        6248               88.9268 %
Incorrectly Classified Instances       778               11.0732 %
Kappa statistic                          0.806
Mean absolute error                      0.0759
Root mean squared error                  0.1883
Relative absolute error                 33.1669 %
Root relative squared error             55.6112 %
Total Number of Instances             7026

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.875    0.045    0.888      0.875    0.881      0.834    0.960     0.905     -1
                 0.932    0.107    0.923      0.932    0.928      0.827    0.964     0.972     1
                 0.806    0.022    0.765      0.806    0.785      0.766    0.965     0.782     2
                 0.620    0.008    0.705      0.620    0.660      0.652    0.966     0.634     3
                 0.606    0.007    0.676      0.606    0.639      0.633    0.970     0.615     4
Weighted Avg.    0.889    0.077    0.889      0.889    0.889      0.815    0.963     0.920

=== Confusion Matrix ===

    a    b    c    d    e   <-- classified as
 1773  163   49   20   22 |    a = -1
  167 3795   83   16   10 |    b = 1
   25   63  462   13   10 |    c = 2
   20   46    7  124    3 |    d = 3
   11   44    3    3   94 |    e = 4
```

J48 on symmEVAL

```
=== Summary ===

Correctly Classified Instances        6286               89.4677 %
Incorrectly Classified Instances       740               10.5323 %
Kappa statistic                          0.817
Mean absolute error                      0.0677
Root mean squared error                  0.1853
Relative absolute error                 29.6203 %
Root relative squared error             54.7355 %
Total Number of Instances             7026

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.874    0.039    0.900      0.874    0.887      0.842    0.951     0.874     -1
                 0.931    0.092    0.933      0.931    0.932      0.839    0.952     0.956     1
                 0.815    0.024    0.750      0.815    0.781      0.761    0.948     0.722     2
                 0.725    0.009    0.704      0.725    0.714      0.706    0.943     0.670     3
                 0.723    0.008    0.675      0.723    0.698      0.691    0.934     0.662     4
Weighted Avg.    0.895    0.067    0.896      0.895    0.895      0.826    0.951     0.899

=== Confusion Matrix ===

    a    b    c    d    e   <-- classified as
 1772  152   54   23   26 |    a = -1
  160 3790   86   23   12 |    b = 1
   19   61  467   13   13 |    c = 2
    9   35    8  145    3 |    d = 3
    9   24    8    2  112 |    e = 4
```

Bagging on symmEVAL

```
=== Summary ===

Correctly Classified Instances        6058               86.2226 %
Incorrectly Classified Instances       968               13.7774 %
Kappa statistic                          0.7613
Mean absolute error                      0.0782
Root mean squared error                  0.204
Relative absolute error                 34.2046 %
Root relative squared error             60.2582 %
Total Number of Instances             7026

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.872    0.053    0.871      0.872    0.871      0.819    0.945     0.871     -1
                 0.901    0.115    0.915      0.901    0.908      0.783    0.942     0.954     1
                 0.815    0.033    0.687      0.815    0.745      0.724    0.916     0.656     2
                 0.550    0.011    0.591      0.550    0.570      0.558    0.937     0.520     3
                 0.303    0.011    0.385      0.303    0.339      0.329    0.924     0.347     4
Weighted Avg.    0.862    0.085    0.863      0.862    0.862      0.772    0.940     0.880

=== Confusion Matrix ===

    a    b    c    d    e   <-- classified as
 1768  139   63   26   31 |    a = -1
  192 3666  136   40   37 |    b = 1
   30   61  467    8    7 |    c = 2
   22   62    6  110    0 |    d = 3
   19   79    8    2   47 |    e = 4
```

KStar on symmEVAL

```
=== Summary ===

Correctly Classified Instances        6004               85.454  %
Incorrectly Classified Instances      1022               14.546  %
Kappa statistic                          0.7291
Mean absolute error                      0.0974
Root mean squared error                  0.2087
Relative absolute error                 42.599  %
Root relative squared error             61.6362 %
Total Number of Instances             7026

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.811    0.044    0.882      0.811    0.845      0.787    0.962     0.914     -1
                 0.965    0.238    0.848      0.965    0.903      0.757    0.970     0.979     1
                 0.510    0.010    0.825      0.510    0.630      0.626    0.963     0.745     2
                 0.395    0.004    0.767      0.395    0.521      0.542    0.966     0.606     3
                 0.394    0.002    0.803      0.394    0.528      0.556    0.963     0.598     4
Weighted Avg.    0.855    0.151    0.853      0.855    0.845      0.745    0.967     0.922

=== Confusion Matrix ===

    a    b    c    d    e   <-- classified as
 1644  348   26    6    3 |    a = -1
  122 3928   18    1    2 |    b = 1
   47  220  292    9    5 |    c = 2
   28   81    7   79    5 |    d = 3
   22   53   11    8   61 |    e = 4
```

Part 8 - Analysis

Summary of all the accuracies

	Model Type				
Attribute Group	Decision Table	J48	Bagging	KStar	Averages
cfsEVAL	88.7636	88.9475	88.6422	86.0233	**88.09415**
reliefEVAL	88.0444	88.0159	86.2226	75.6618	84.486175
corrEVAL	85.8098	86.4098	85.8098	82.3086	85.0845
gainEVAL	88.6991	85.8098	85.8098	85.3829	86.4254
symmEVAL	88.9268	89.4677	86.2226	85.454	**87.517775**
Averages	**88.04874**	**87.73014**	86.5414	82.96612	

This suggests that the Decision Table model is best for this classification task. Additionally, the cfsEVAL attribute group was the best performing subgroup of attributes, suggesting that it captured the output group well. However, the difference between the Decision Table and J48 models' as well as cfsEVAL and symmEVAL groups' averages is minimal, so there needs to be a deeper look into other metrics. One such metric we can look at is the available error scores (mean absolute error, root mean squared error, relative absolute error, and root relative squared error).

Across these scores, **J48 with SymmetricUncertAttributeEval** had the lowest error outputs. This gives confidence that in future datasets, this model will be able to run with similar presion, recall, and accuracy as demonstrated in this test. In other words, the likelihood of false classifications will be small.

The main goal of this study was to identify which socioeconomic factors play a large role in determining the severity of one's mental health status. One way to look at this is by iterating through the branches of the J48 tree generated, however that only considers the values in this specific model. To get a more thorough understanding of selected attributes, we can look at the attribute groups:

cfsEVAL	**reliefEVAL**	**corrEVAL**	**gainEVAL**	**symmEVAL**
DISAB3_A	HRTESTLAST_A	SOCSCLPAR_A	MHTHND_A	FAM_A
MLTFAMFLG_A	EMPDYSMSS3_A	COGMEMDFF_A	MHTHDLY_A	WTFA_A
EVRMARRIED_A	URBRRL	LSATIS4_A	SOCSCLPAR_A	MHTHND_A
SMKNOW_A	REGION	PHSTAT_A	TBIHLSBMC_A	MHTHDLY_A

PAITOOTH3M_A VIGIL4_A DISCRIM5_A MHTHND_A MHTHDLY_A HYSTEV	PSTRAT LONGCOVD1_A VIGIL4_A SMKNOW_A INTV_QRT VIGIL3_A	SOCERRNDS_A DISCRIM5_A VIGIL4_A MHTHDLY_A MHTHND_A	WTFA_A TBILCDCMG_A HOMEHC12M_A RXDG12M_A MEDNG12M_A MEDDL12M_A	SOCSCLPAR_A VIGIL4_A LSATIS4_A DISCRIM3_A COGMEMDFF_A DISCRIM5_A VIGIL1_A

There are some overlaps within each group, such as MHTHND_A and MHTHDLY_A. To get a better understanding, each attribute is defined below.

Attribute	Description
DISAB3_A	The Washington Group Short Set Composite Disability Indicator
MLTFAMFLG_A	Indicator for multifamily households
EVRMARRIED_A	Sample adult has ever been married
SMKNOW_A	Now smoke cigarettes
PAITOOTH3M_A	Toothache or jaw pain
VIGIL4_A	Avoid certain situations and places
DISCRIM5_A	You are threatened or harassed
MHTHND_A	Needed counseling, therapy but did not get it due to cost, past 12 months
MHTHDLY_A	Delayed counseling, therapy due to cost, past 12 months
HRTESTLAST_A	A How long since hearing test
EMPDYSMSS3_A	Days missed work, past 12 months (top-coded)
URBRRL	2013 NCHS Urban-Rural Classification Scheme for Counties
LONGCOVD1_A	Had COVID-19 symptoms for 3 or more months
SMKNOW_A	Now smoke cigarettes
VIGIL3_A	Watch what you say and how you say it

SOCSCLPAR_A	Language socially
COGMEMDFF_A	Difficulty remembering/concentrating
PHSTAT_A	General health status
SOCERRNDS_A	Difficulty doing errands alone
TBIHLSBMC_A	Headache, sensitivities, balance problems or mood change, past 12 months
WTFA_A	Weight - Final Annual
TBILCDCMG_A	A Lost consciousness, dazed or confused, or had gap in memory, past 12 months
HOMEHC12M_A	Received care at home, past 12 months
RXDG12M_A	Needed prescription medication but did not get it due to cost, past 12 months
MEDNG12M_A	Needed medical care but did not get it due to cost, past 12 months
MEDDL12M_A	Delayed medical care due to cost, past 12 months
VIGIL1_A	Prepare for possible insults before leaving home
FAM_A	Number of Emergency Contacts

In general, it can be seen that these attributes fall in 1 of 4 main categories:

- General Health:
  - Have been smoking in past
  - Annual Weight
  - Disabled

- Personality Specifics:
  - Prepares for insults when leaving home
  - Difficulty doing errands alone

- Delay of Medical attention due to cost
  - Needed therapy, but couldn’t get it due to cost
  - Delayed medical care due to cost

- Family Structure
  - Married to someone else
  - Received care at home
  - Lives in a multifamily household

- Lives in urban/suburban/rural area

There are additionally some other attributes that don’t fall in this category, such as having COVID-19 for three or more months and days missed at work for the past 12 months. This data is pulled from the 2023 survey, a time where the effects of COVID-19 still played some role. This could have created a potential bias towards these values being significant since they were of relevance at the time. A future study utilizing data from a more recent study would be better able to tell if the impact of the COVID-19 pandemic still plays a role in the severity of Mental Health. Both of these attributes are NOT present in the SymmetricUncertAttributeEval attribute group, meaning that the final model selected doesn’t include these attributes. This independence means that it can be more generalizable to years without inherent COVID-19 impact, however external testing is needed to validate that claim.

In general, these attributes suggest that health, financial status, family support, and internal thoughts contribute to the severity of mental health. Three of these can be assessed in a non-psychological setting. For example, when a new patient is admitted, a hospital can check what outside family support the individual has, their general health, and how long they waited to come. Using these, hospitals can make recommendations as to sending an individual for a psych eval, ensuring more individuals receive the care they need. Even as mental health becomes a more widely accepted topic, there are many taboos associated with it and this information can help ensure that those who are most vulnerable have no barriers to support.

Part 9 - Conclusions/Steps for Reproduction

As stated above, the J48 model with Symmetric Uncertainty Attribute Evaluation Selection had the best results of the 20 runs for this project. We were successfully able to train and test a predictive classification model that predicted the severity of mental health onset for adult individuals and feel confident about our results. However, there is some potential bias due to the data coming from NIH’s 2023 study, future projects should look into gathering more recent data to properly assess the potential impact of COVID-19 on severity. Future studies could also initially group attributes into subgroups based on relatedness, combining similar attributes to create a stronger model.

Steps to Reproduce Our Model: J48 model with Symmetric Uncertainty Attribute Evaluation Selection:

All csv files can be found in the project folder under “train/test/val files”

OPTIONAL:

1. Open Weka and load the adult23_train+test.csv in the zip file.
2. Under the Proprocess tab, click Filter → Choose → Filters → Unsupervised → Attribute
3. then select NumerictoNominal
4. Click on the white space and ensure that all attributes are selected. Hit Apply.
5. Go to the “Select Attributes” tab and choose the correct class “engineered_attribute”
6. Select SymmetricUncertAttributeEval (Symmetric Uncertainty Attribute Evaluation

Selection) as the Attribute Evaluator, and Ranker as the Search Method
7. Hit Start and wait for the program to finish
8. Take note of top 11 features; keep the index values for these features
9. Go back to the Preprocess tab and click Filter → Choose → Filters → Unsupervised → Remove
10. Click on the white space and paste in the selected attribute indexes, add in 321 as this is the class attribute
11. Set invertSelection to be True
12. Save and Click Apply
13. Click on the Classify tab and click "Percentage Split" under Test Options, write 70%
14. Select the J48 model under trees
15. Click start and wait for it to complete

The final model can be found here:

https://drive.google.com/file/d/10dGNMBCDxjRuOy1RWt79ywOtHUk64zLH/view?usp=sharing

Part 10 - Teamwork Makes the Dreamwork

Medha:
- Finding Data
- Project Statement
- Initial Attribute visualization and understanding
- Engineered class variable in Python
- Running the 20 Models on Attribute Selection Groups
- Information on how Attribute Selection methods worked

Kade:
- Removed Unnecessary Attributes (involved going through 600+ attributes 3 times)
- Filling in Missing values from each attribute
- Generating Attribute Selection Groups
- Information on how Models worked
- Citing sources, proofreading paper

Part 11 - Sources and Citations

- Awan, Abid Ali. "A Guide to Bagging in Machine Learning: Ensemble Method to Reduce Variance and Improve Accuracy." *DataCamp*, DataCamp, 20 Nov. 2023, www.datacamp.com/tutorial/what-bagging-in-machine-learning-a-guide-with-examples.
- "NHIS - 2023 NHIS." *Www.cdc.gov*, 6 Apr. 2023, www.cdc.gov/nchs/nhis/2023nhis.htm. Accessed 10 Apr. 2023.
- *LRI*, www.lri.fr/~pierres/donn%E9es/save/these/articles/lpr-queue/hall99correlationbased.pdf. Accessed 22 Oct. 2024.
- Brownlee, Jason. "How to Perform Feature Selection with Machine Learning Data in

Weka." *MachineLearningMastery.Com*, 12 Dec. 2019, machinelearningmastery.com/perform-feature-selection-machine-learning-data-weka/.

- Sahazada, Sariq. "Correlation-Based Feature Selection in a Data Science Project." *Medium*, Medium, 10 May 2024, medium.com/@sariq16/correlation-based-feature-selection-in-a-data-science-project-3ca08d2af5c6.
- WekaLoverWekaLover10322 silver badges55 bronze badges, and KevinDKevinD 72177 silver badges1414 bronze badges. "How the Selection Happens in 'infogainattributeeval' in Weka Feature Selection (Filter Method)." *Stack Overflow*, 1 Feb. 1961, stackoverflow.com/questions/33982943/how-the-selection-happens-in-infogainattributeeval-in-weka-feature-selection.
- Hudson, Christopher G. "Socioeconomic status and mental illness: Tests of the social causation and selection hypotheses." *American Journal of Orthopsychiatry*, vol. 75, no. 1, 2005, pp. 3–18, https://doi.org/10.1037/0002-9432.75.1.3.
- Lee, G. R., et al. "Gender differences in the depressive effect of widowhood in later life." *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, vol. 56, no. 1, 1 Jan. 2001, https://doi.org/10.1093/geronb/56.1.s56.
- Macintyre, Anna, et al. "What has economics got to do with it? the impact of socioeconomic factors on mental health and the case for collective action." *Palgrave Communications*, vol. 4, no. 1, 30 Jan. 2018, https://doi.org/10.1057/s41599-018-0063-2.
- Nagasu, Miwako, et al. "Association of socioeconomic and lifestyle-related risk factors with mental health conditions: A cross-sectional study." *BMC Public Health*, vol. 19, no. 1, Dec. 2019, https://doi.org/10.1186/s12889-019-8022-4.
- Roy-Byrne, Peter P., et al. "Low socioeconomic status and mental health care use among respondents with anxiety and depression in the NCS-R." *Psychiatric Services*, vol. 60, no. 9, Sept. 2009, pp. 1190–1197, https://doi.org/10.1176/ps.2009.60.9.1190.
- "Weka - Quick Guide." *Tutorialspoint*, www.tutorialspoint.com/weka/weka_quick_guide.htm. Accessed 22 Oct. 2024.
- "Package Weka.Attributeselection." *Weka.attributeSelection*, 28 Jan. 2022, weka.sourceforge.io/doc.dev/weka/attributeSelection/package-summary.html.