**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 3**](#_3ak0k3hdk457)

[**Part 2 - Description of Dataset 3**](#_tzjfbwashhn1)

[**Part 3 - Pre-Processing 4**](#_tbxvv6qolpws)

[Removing Attributes with >70% Missing Values 4](#_4llditr4xoub)

[Filling In Default Values 4](#_l6jg05698rq8)

[Engineering Class Attribute 6](#_syzaw424lu4w)

[Training/Validation/Test Dataset 8](#_2i44rkau51fi)

[**Part 4 - Attribute Selection Algorithms and Model Classifiers Used 8**](#_wrt1ta9rf8go)

[CfsSubsetEval with BestFirst 9](#_jiqzfshs0ey)

[Correlation (Non-Weka Attribute Selection) 9](#_q14occncgp02)

[CorrelationAttributeEval with Ranker 10](#_12dvrc216xmi)

[GainRatioAttributeEval with Ranker 10](#_6xifqgdog0gt)

[SymmetricUncertAttributeEval with Ranker 11](#_39rhf1eiqx7d)

[**Part 5 - Selection Algorithm Results 11**](#_d5l92b28mcxv)

[CfsSubsetEval with BestFirst 11](#_sarpqirw3zrc)

[CorrelationAttributeEval with Ranker 12](#_scbvi6q5hwpn)

[**Part 6 - Model Selection 15**](#_xx6359drwp58)

[**Part 7 - Results 1**](#_p1siz2vm1q3u)**7**

[**Part 8 - Analysis 2**](#_khxu10p4jv8l)**7**

[**Part 9 - Conclusions/Steps for Reproduction**](#_s3q8r0nrzg7q) **30**

[**Part 10 - Teamwork Makes the Dreamwork 3**](#_4ycdwdymrr59)**1**

[**Part 11 - Sources and Citations 3**](#_gb09rg2gnd9a)**1**

## 

## 

## 

## 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) **- 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.

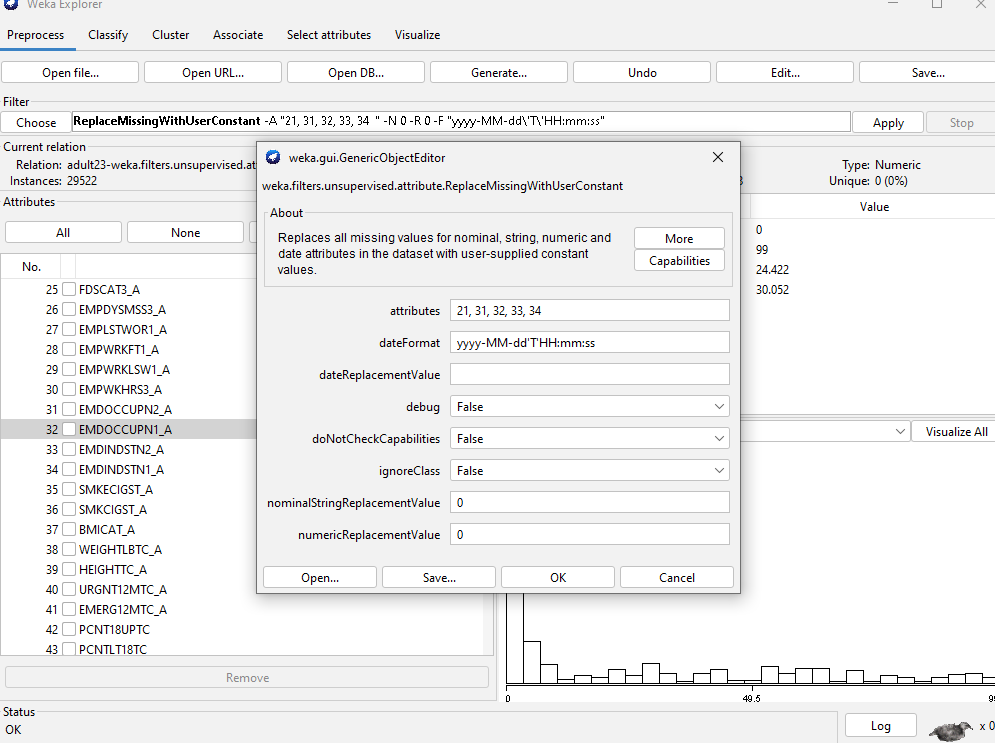
This resulted in the removal of 311 attributes, leaving values with more data useful for predicting the class. While this seems like a lot, we still have 336 attributes to preprocess.

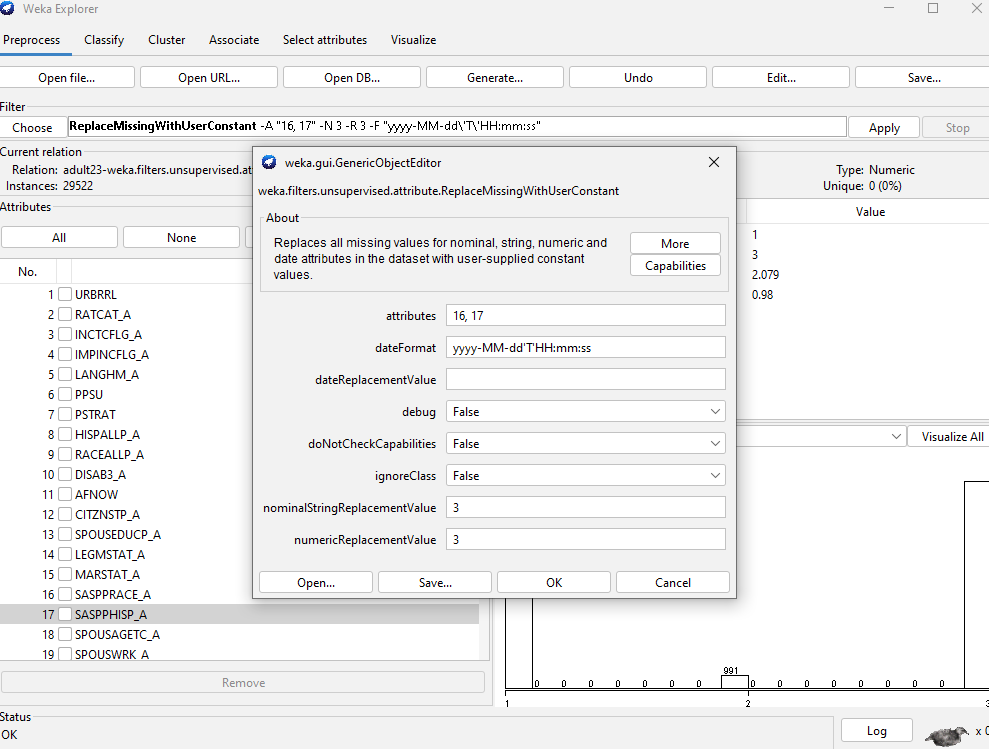
### Filling In Default Values

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, 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 | 9 |
| 13, 18, 30, 50, 62, 103, 186, 187, 191 | 99 |
| 27 | 999 |
| 190 | 9999 |
| 51 | 99999 |

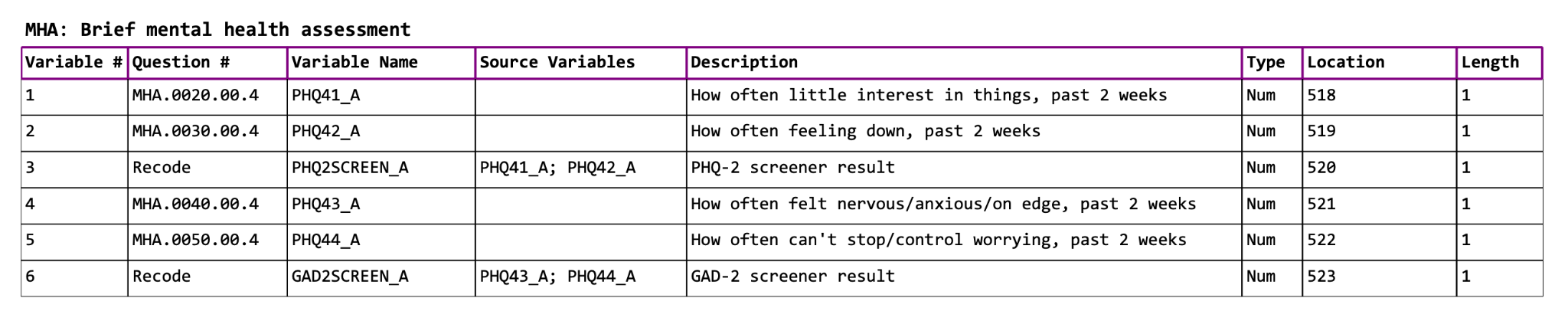
These values were filled in using Weka’s ***ReplaceMissingWithUserConstant*** feature. After completion, there were no more missing values within the dataset.



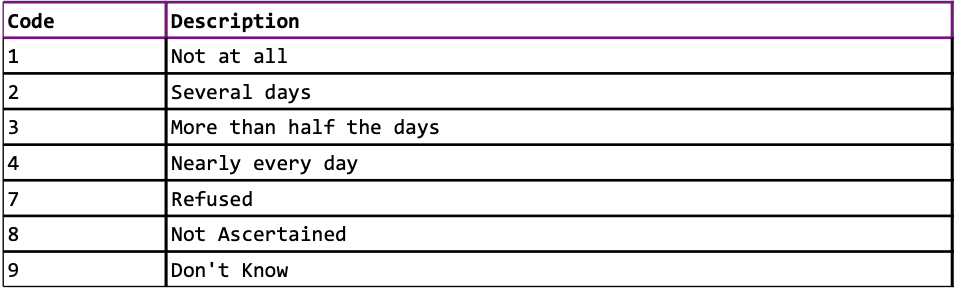


### *Engineering Class Attribute*

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.



The codes for these 4 attributes breakdown as so:

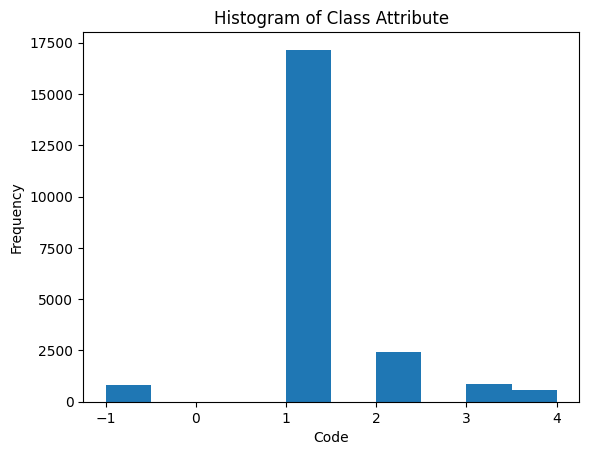


There are multiple approaches we could have taken to combine these 4 attributes. All 4 attributes have 0 missing values, so there is no risk of bias. The values placed in the engineered attribute will be as follows: (1) Average then round values between 2-4 to get the average number of days where the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

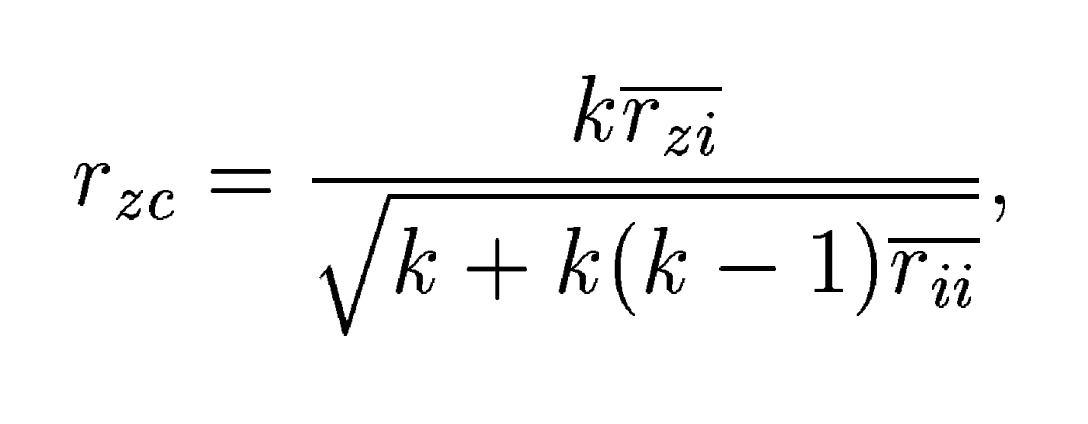
## Part 4 - Attribute Selection Algorithms and Model Classifiers Used

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.



Where *rzc* is the merit of subset C

k = number of features in the subset

*bar(rzi*) is the average correlation between the class and selected features

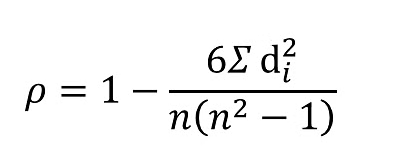
*bar(rii*) is the average correlation among the selected features

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.

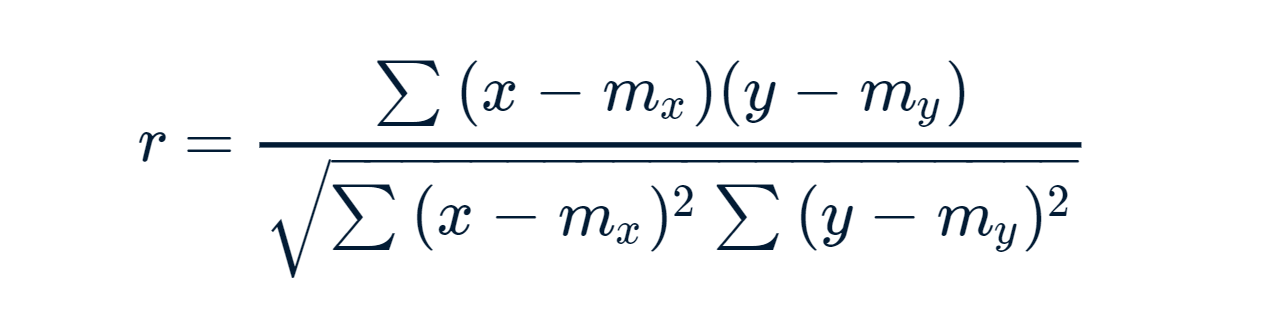


| # Calculate Spearman rank correlations  correlations = df.corr(method='spearman')['engineered\_attribute'].drop('engineered\_attribute')  # Take the absolute value of the correlations  absolute\_correlations = correlations.abs()  # Get the top 10 most correlated attributes  top\_10\_correlations = 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.

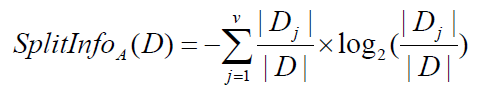


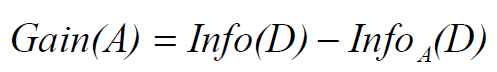
The ranker search method ranks attributes based on their individual merit (correlation scores). They are sorted in descending order of their evaluation score, meaning the most predictive attribute appears at the top of the list. Ranker can also be set to eliminate attributes under a certain threshold,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

### GainRatioAttributeEval with Ranker

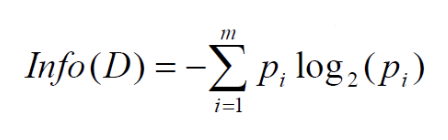
We used Weka for this approach. It utilized the following formulas to calculate the GainRatio for a particular attribute

GainRatio(A) = Gain(A)/SplitInfoA(D)



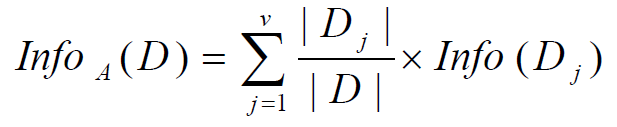


Where the expected information (or entropy) needed to classify a value in D is



where *pi* 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:



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

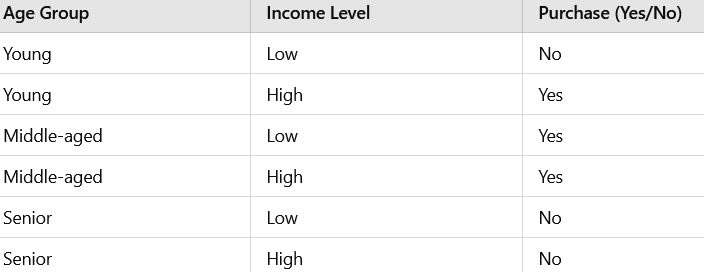
***symmEVAL*** - 318,159,160,261,150,311,156,266,154,153,158

## 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.

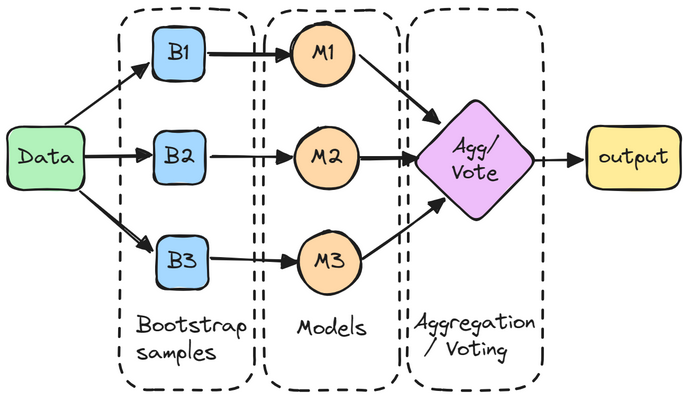


*J48*

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.

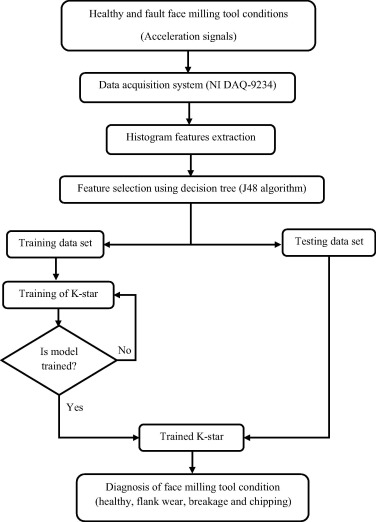
*Bagging*

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.



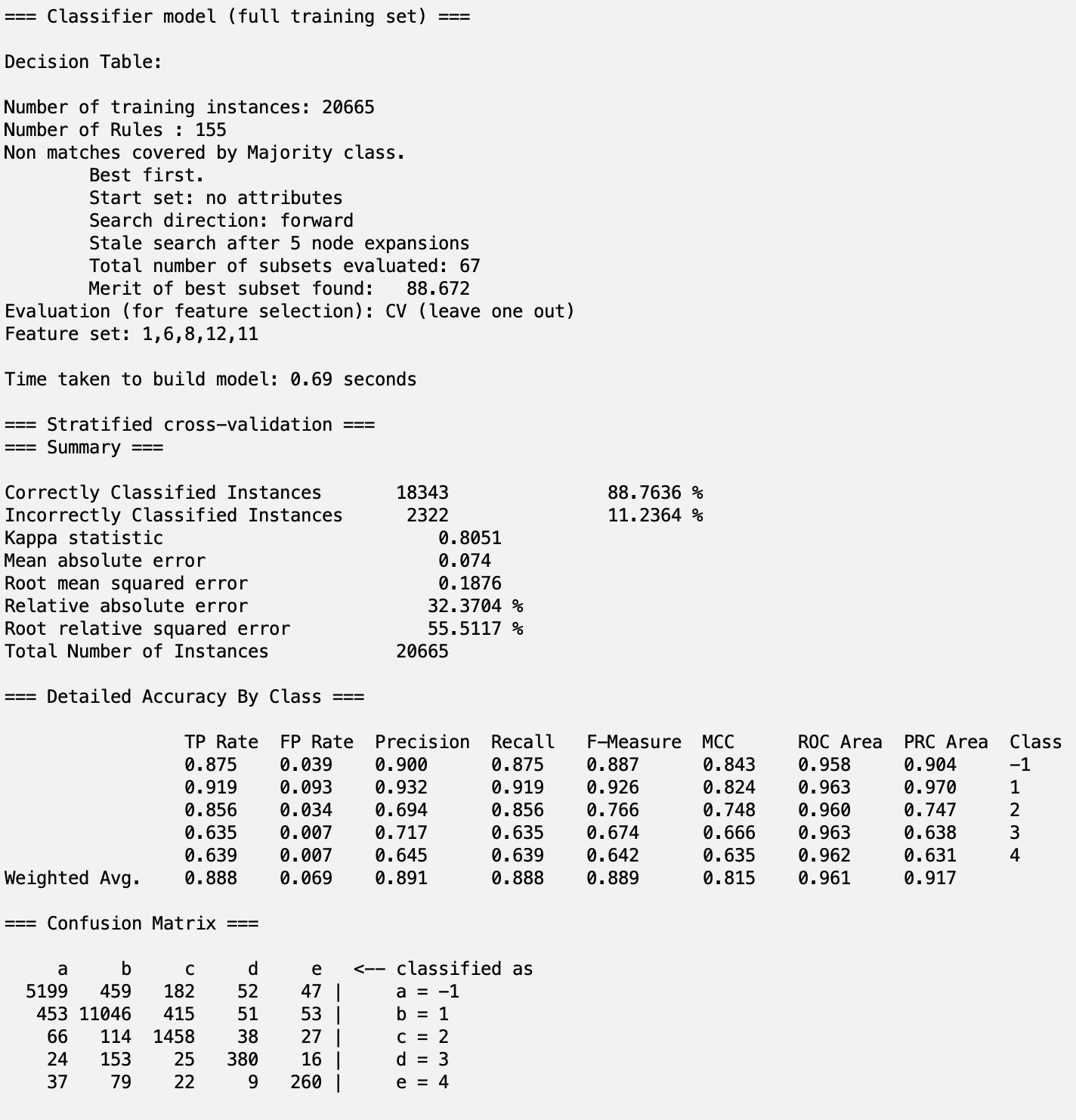
*KStar*

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.

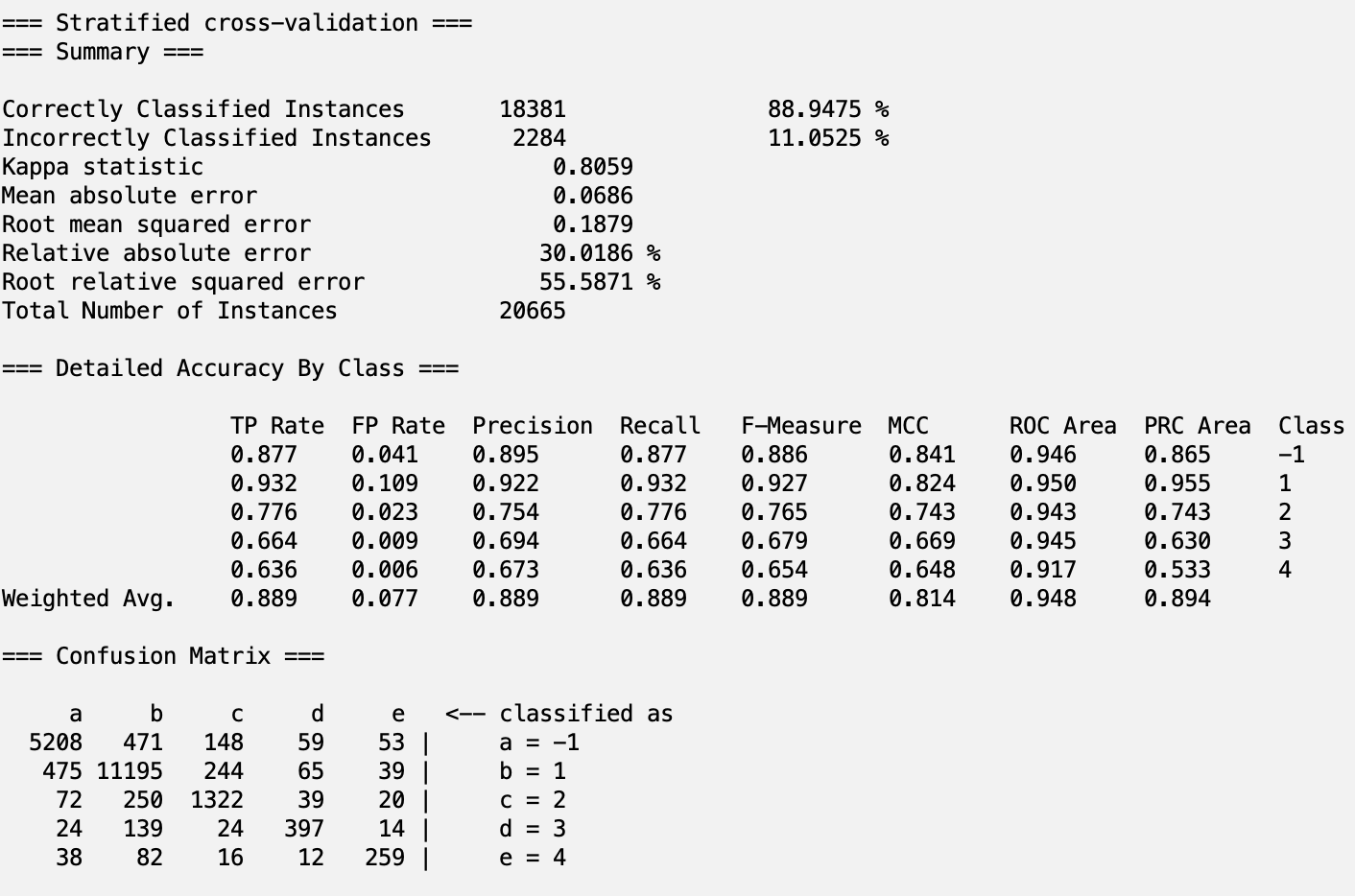


## Part 7 - Results

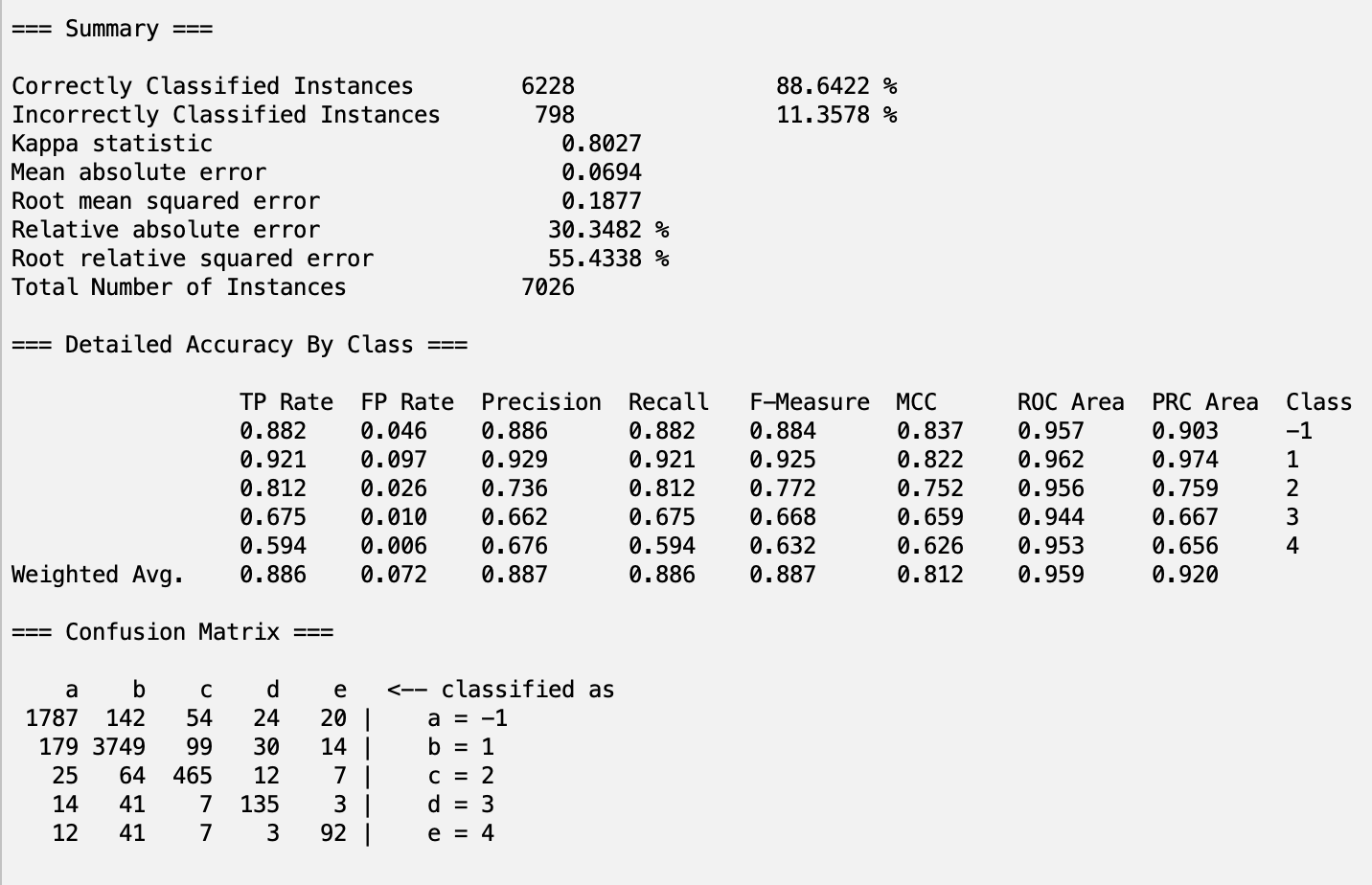
***Decision Table on cfsEVAL***

******

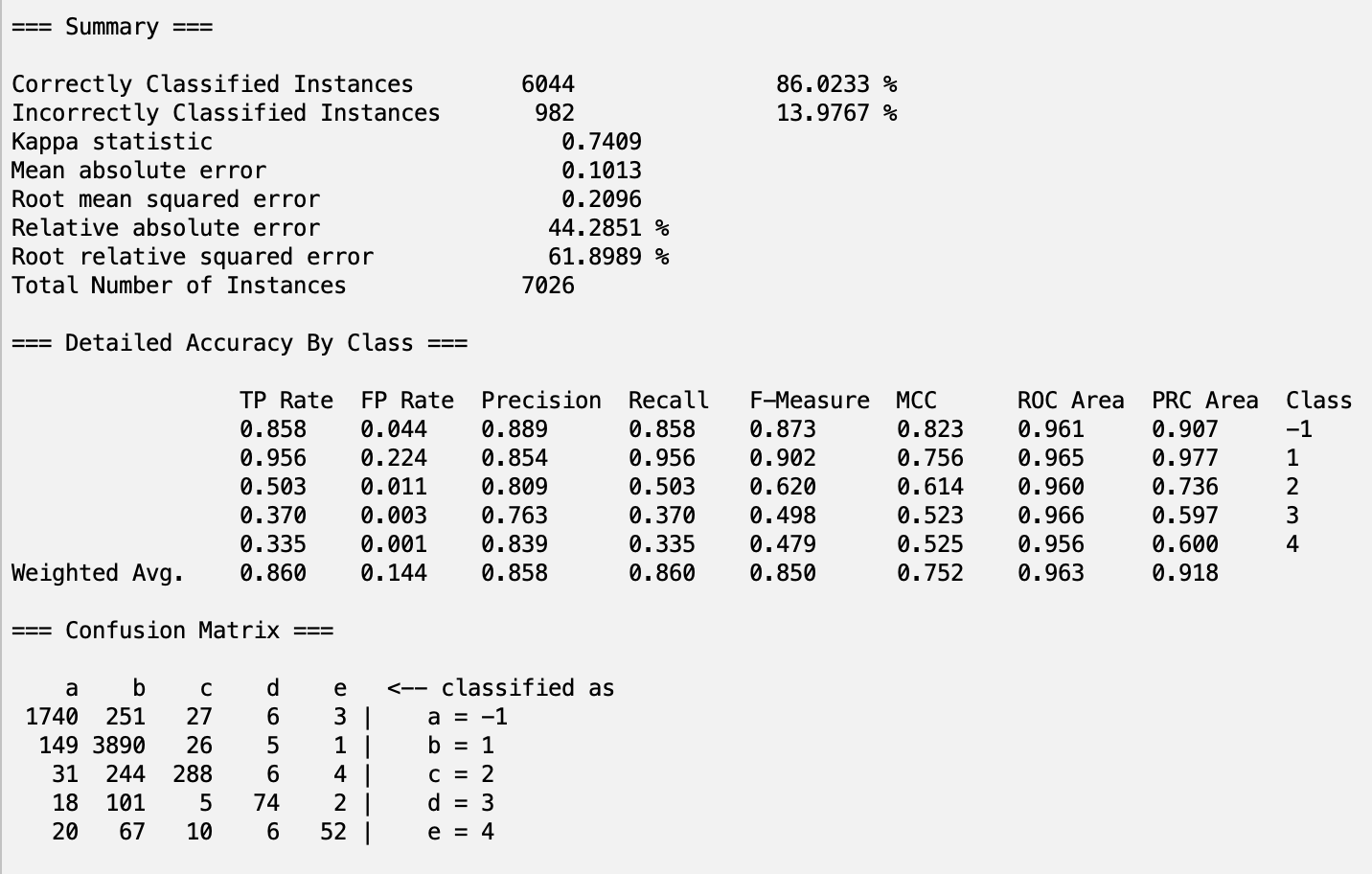
***J48 on cfsEVAL***

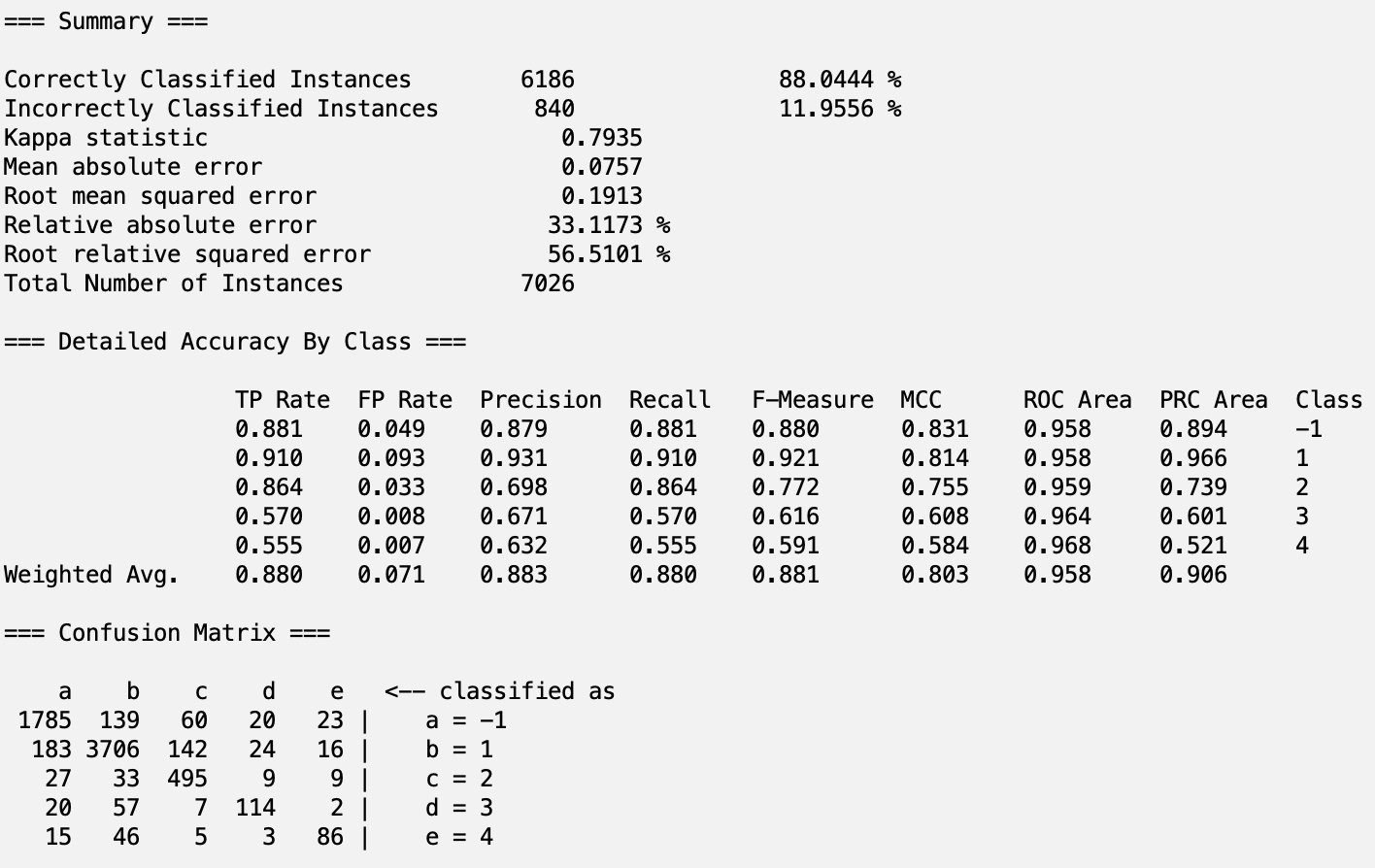
******

***Bagging on cfsEVAL***

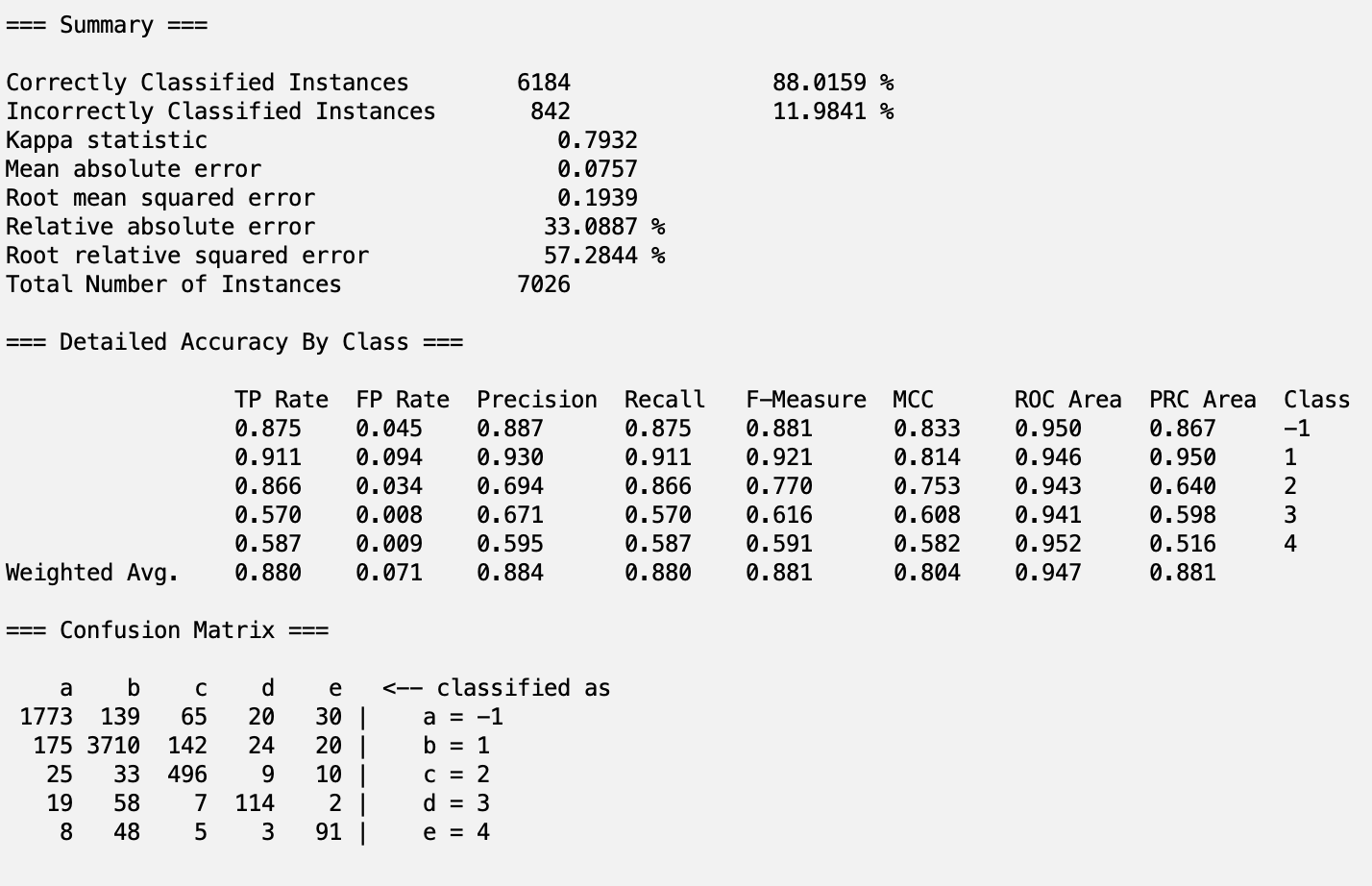
******

***KStar on cfsEVAL***

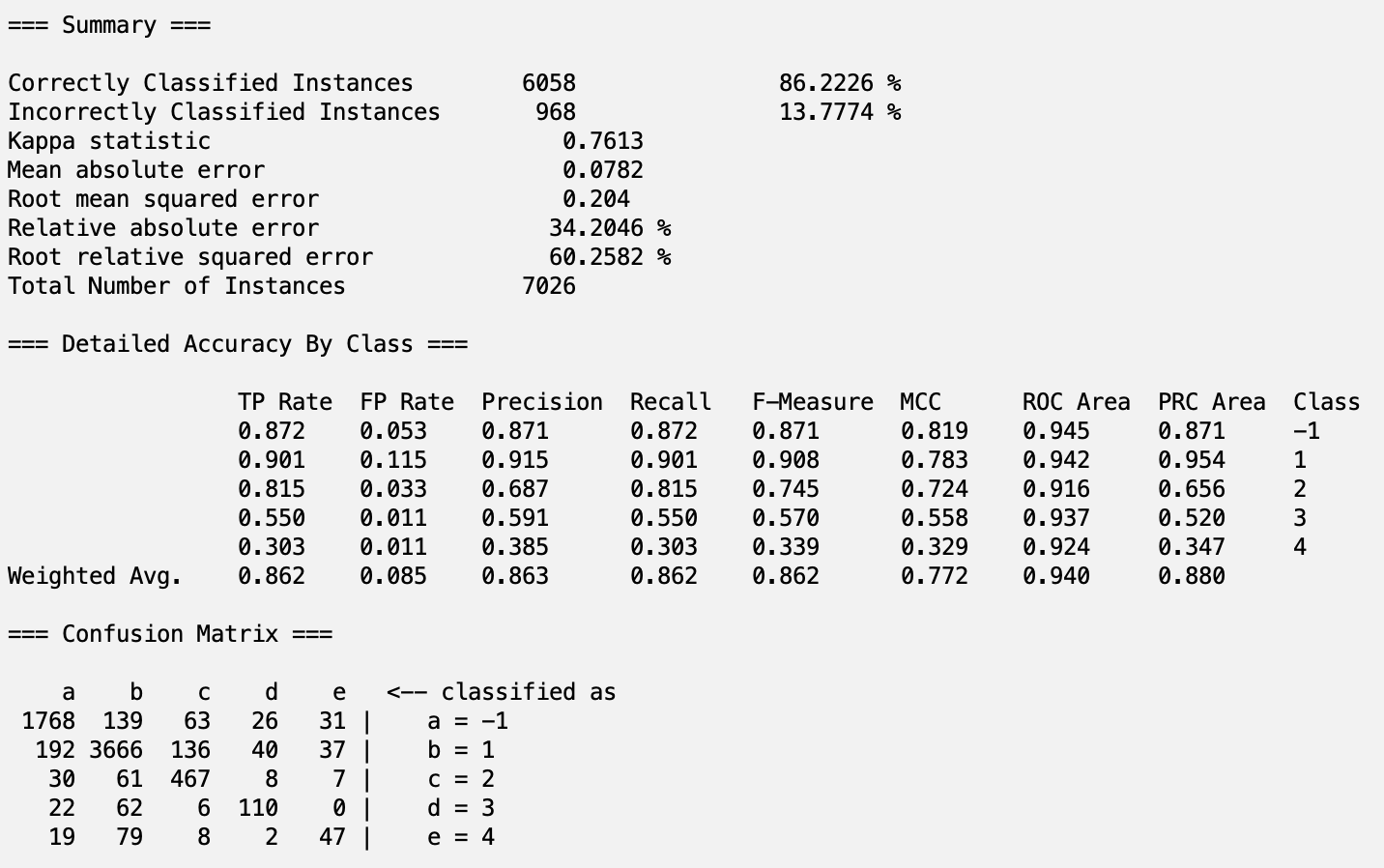
***Decision Table on reliefEVAL***

******

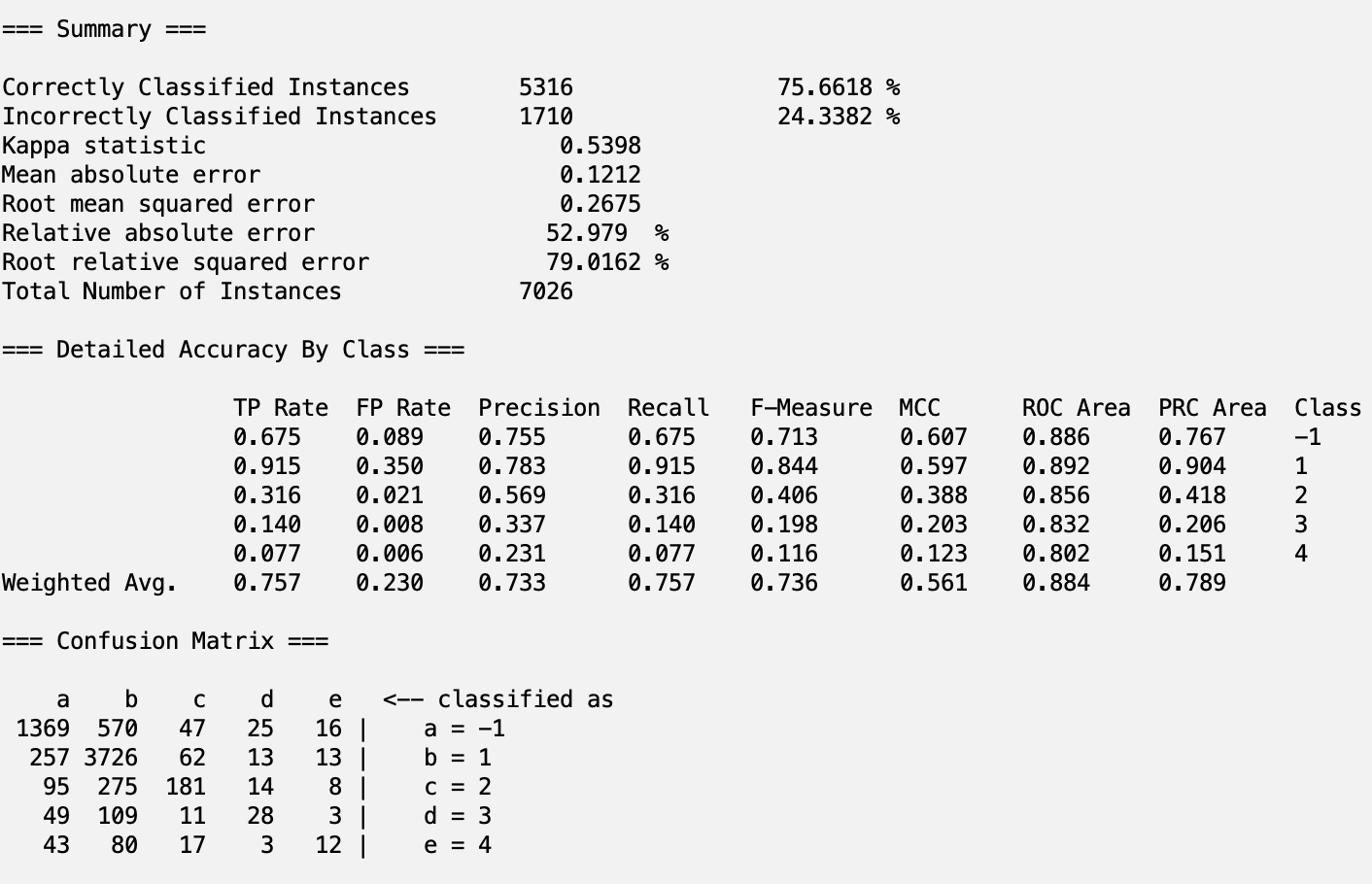
***J48 on reliefEVAL***

******

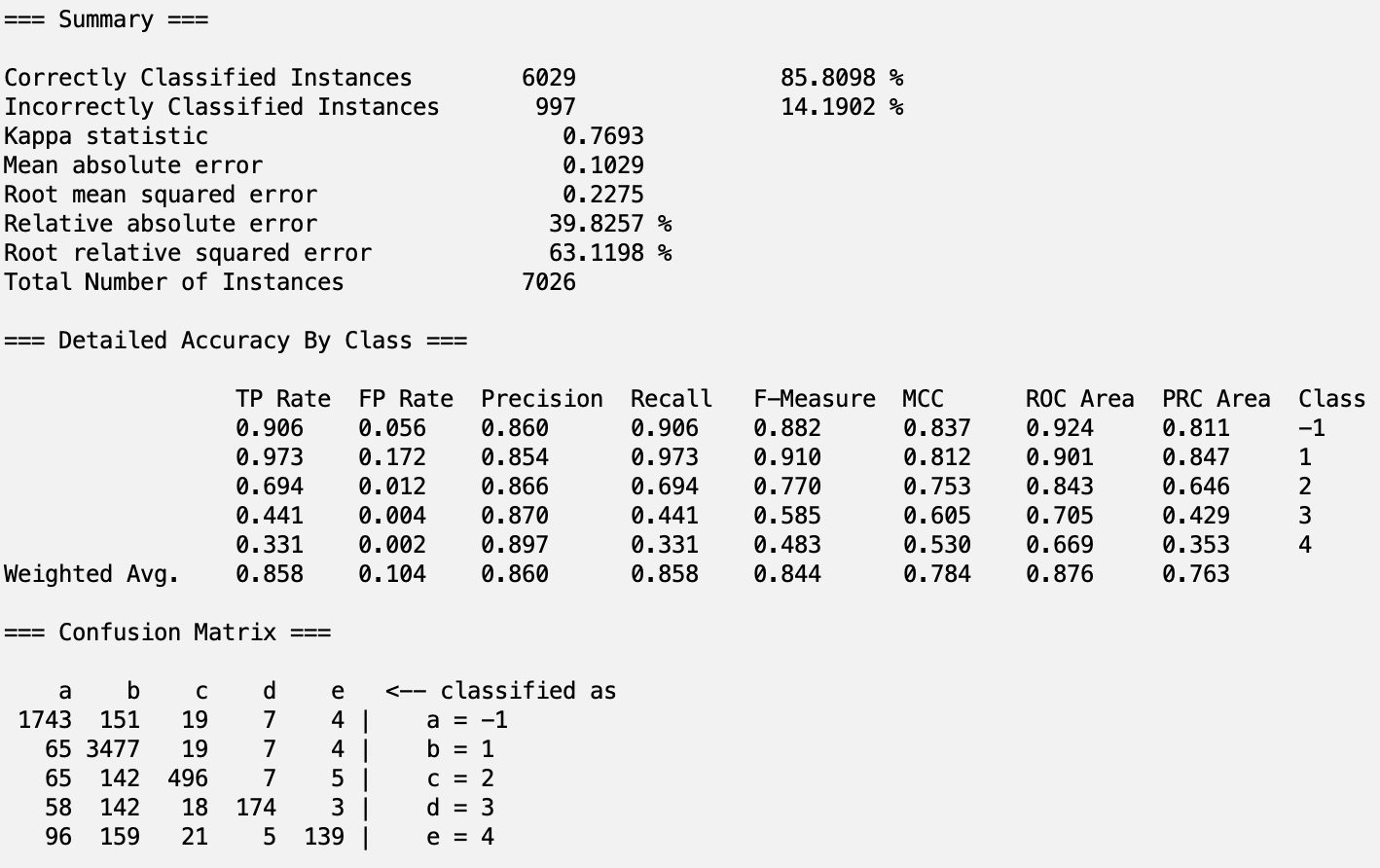
***Bagging on reliefEVAL***

******

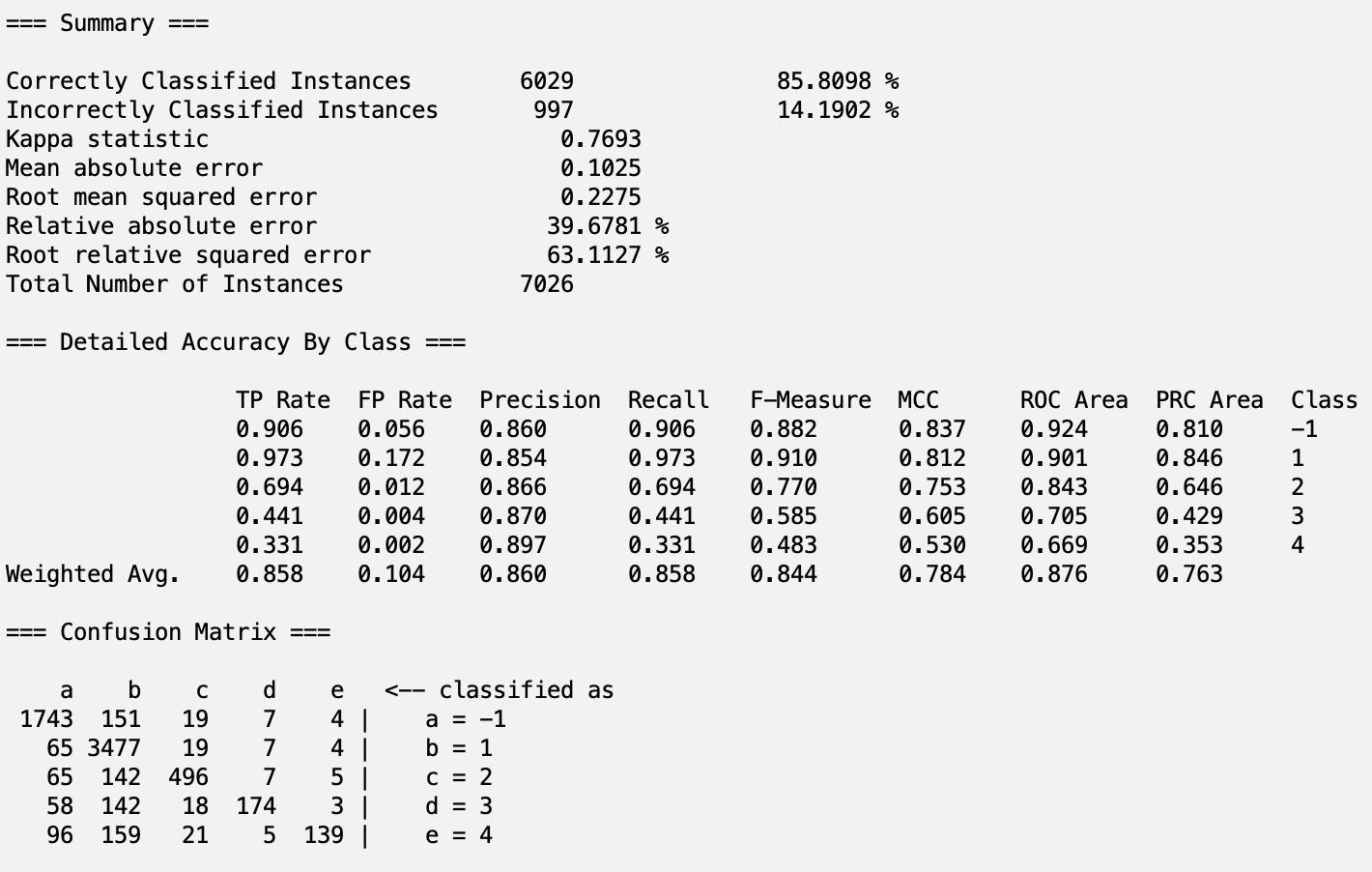
***KStar on reliefEVAL***

******

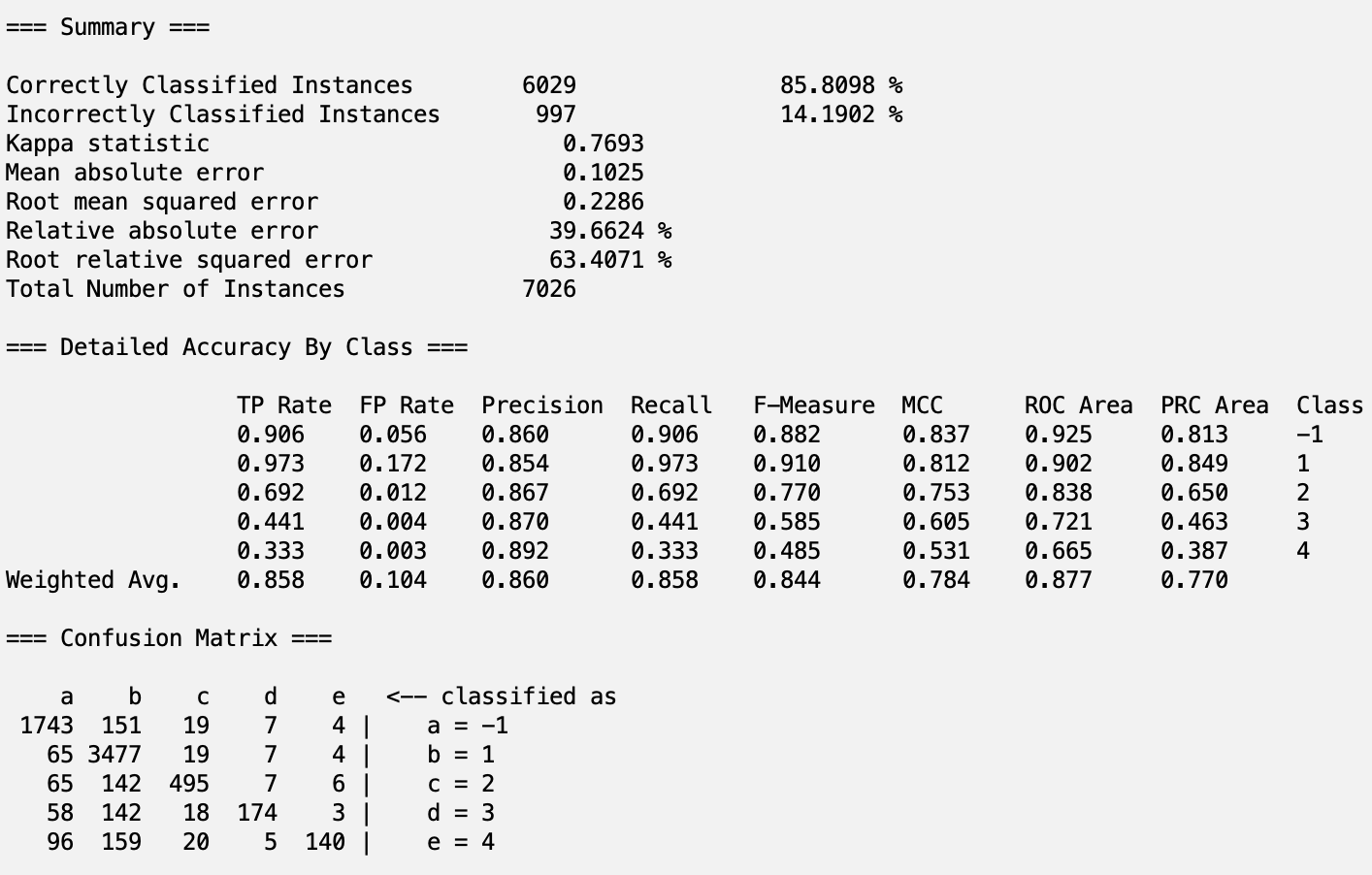
***Decision Table on corrEVAL***

******

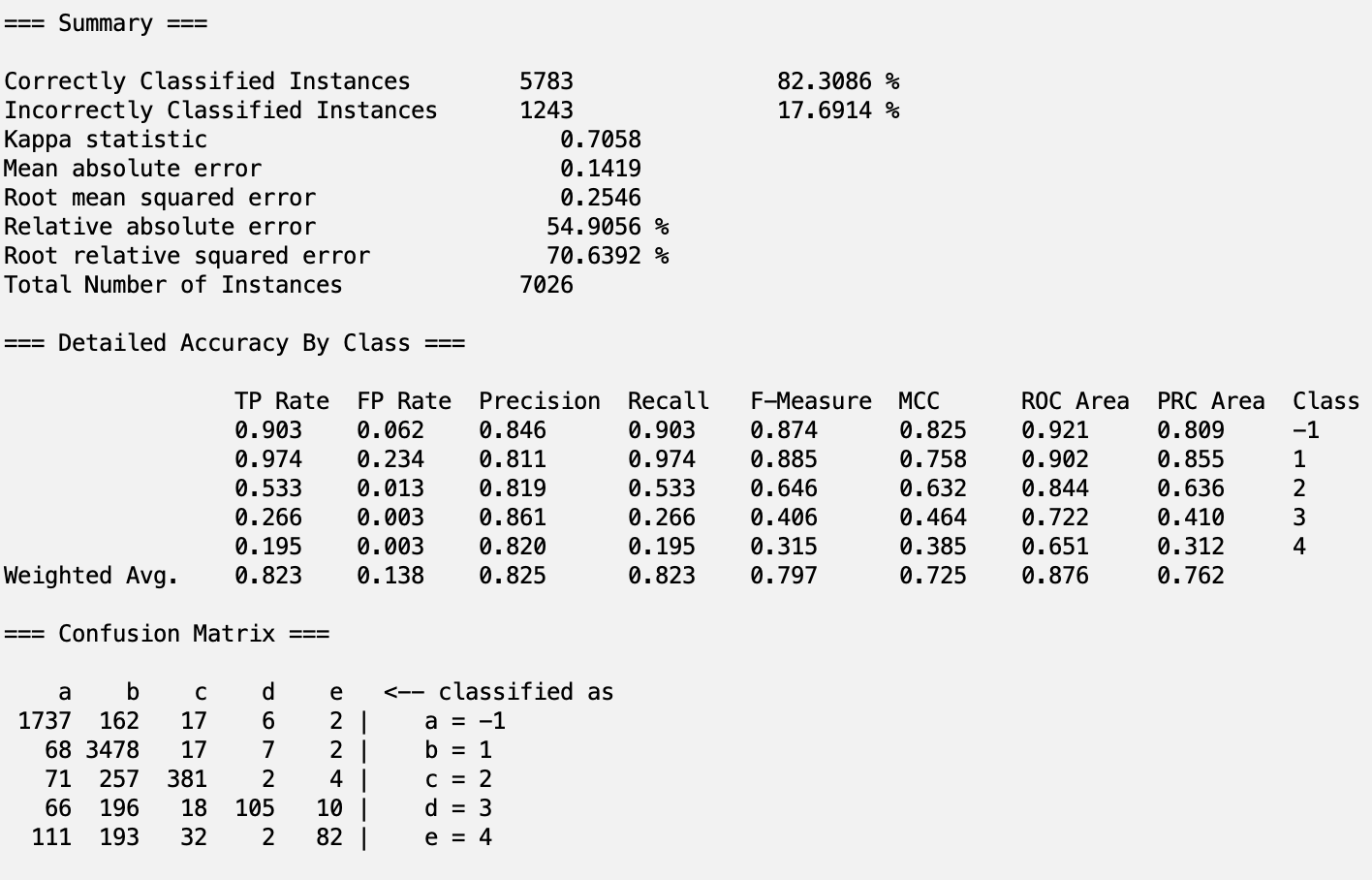
***J48 on corrEVAL***

******

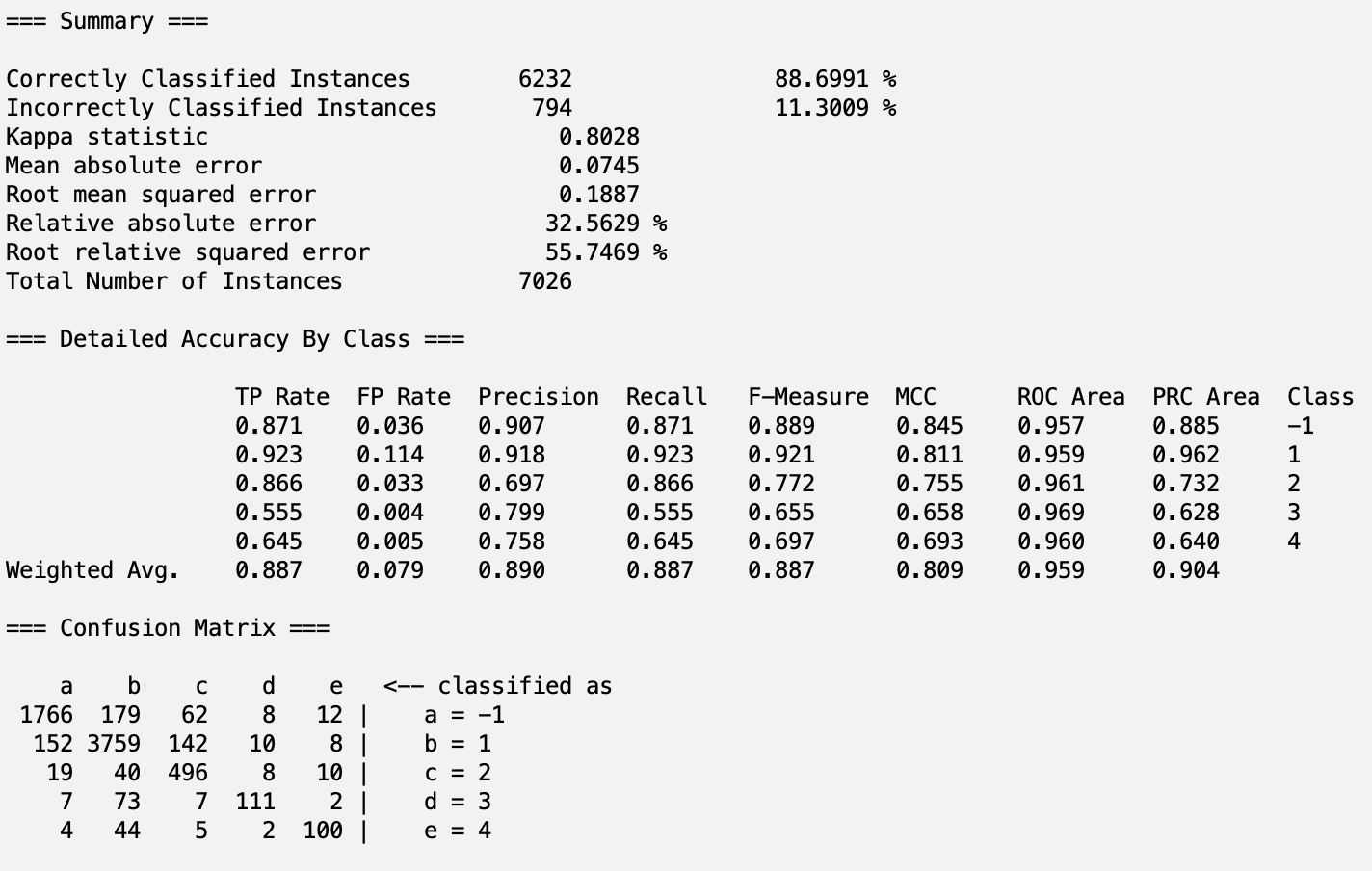
***Bagging on corrEVAL***

******

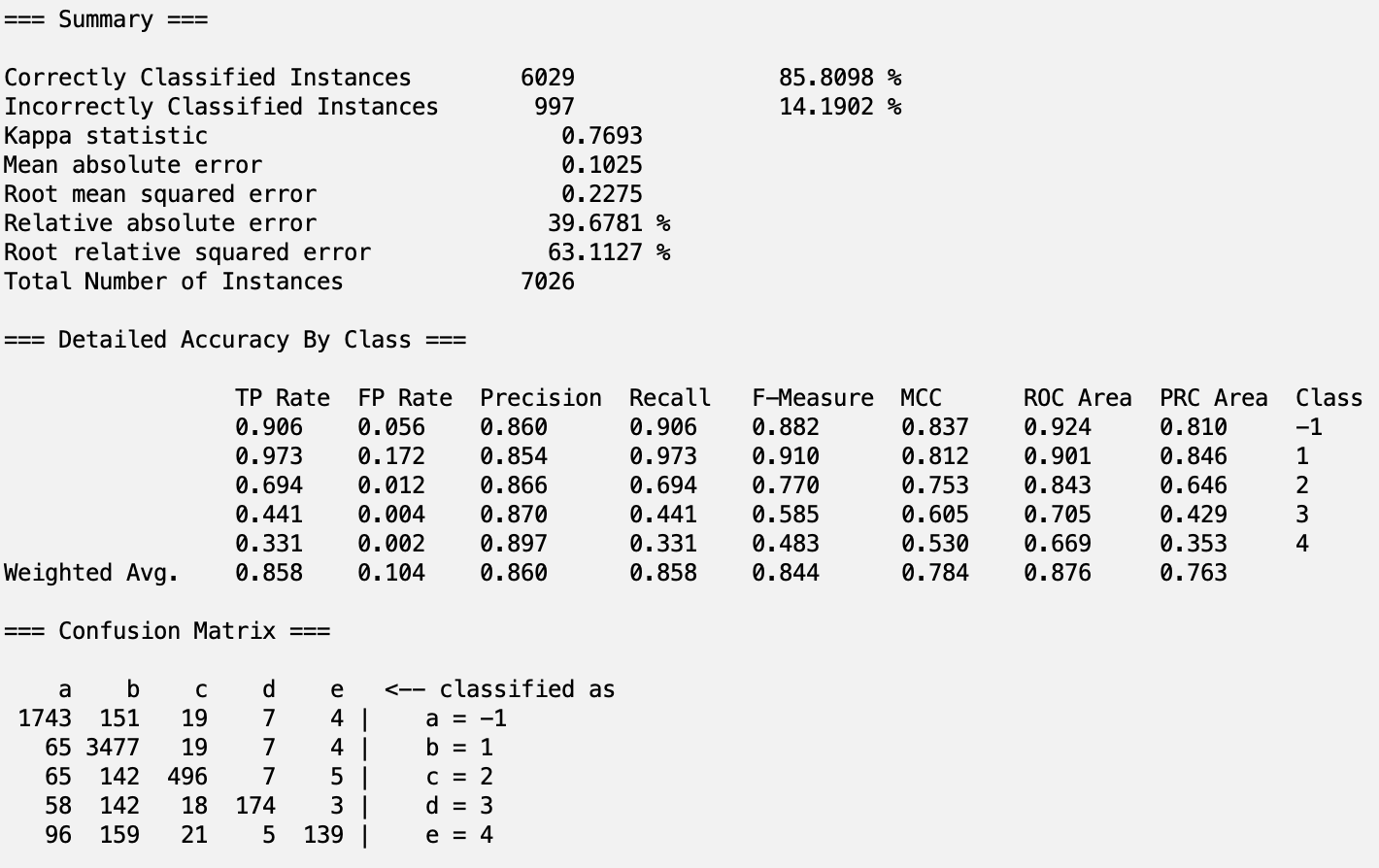
***KStar on corrEVAL***

******

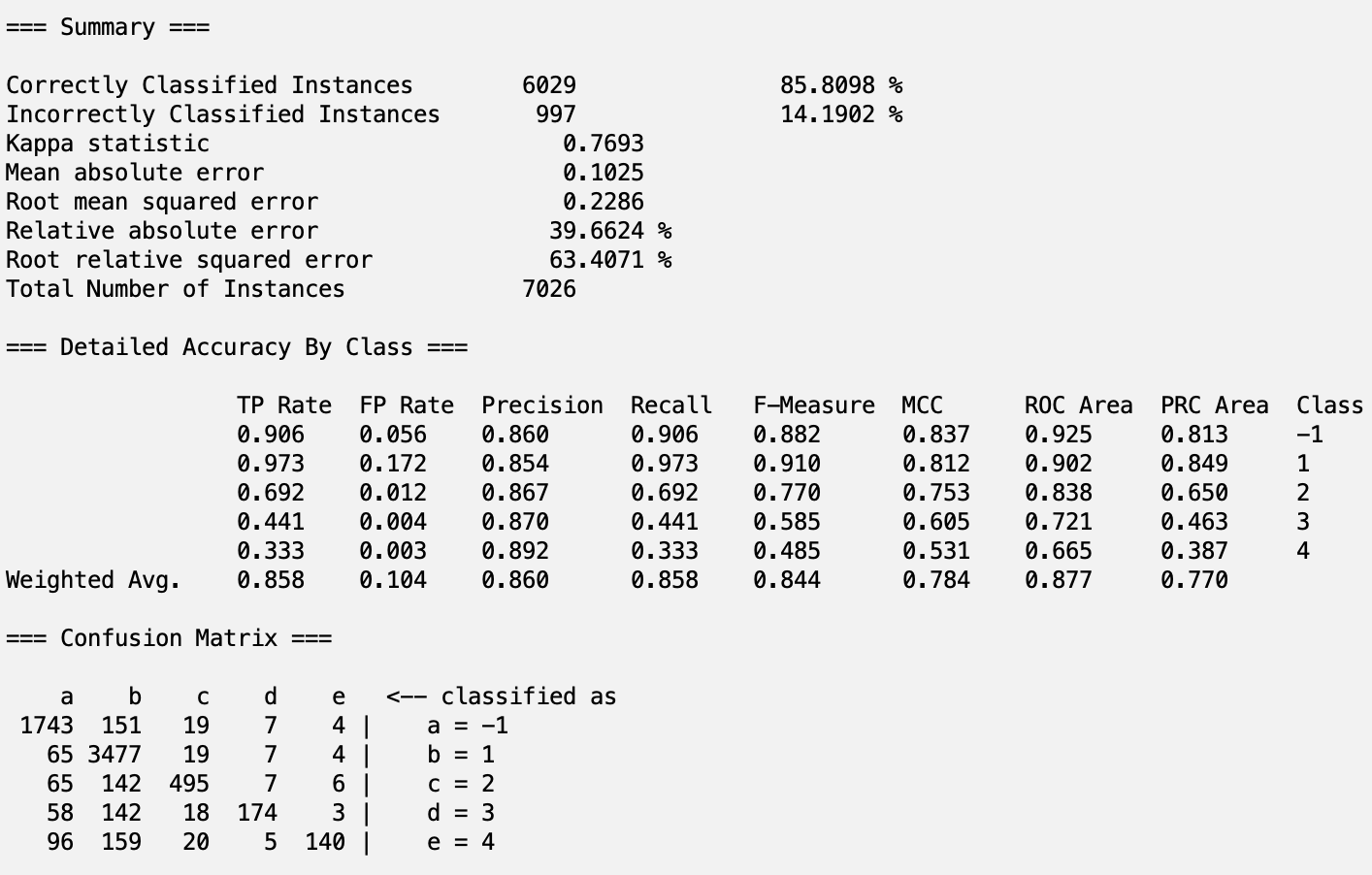
***Decision Table on gainEVAL***

******

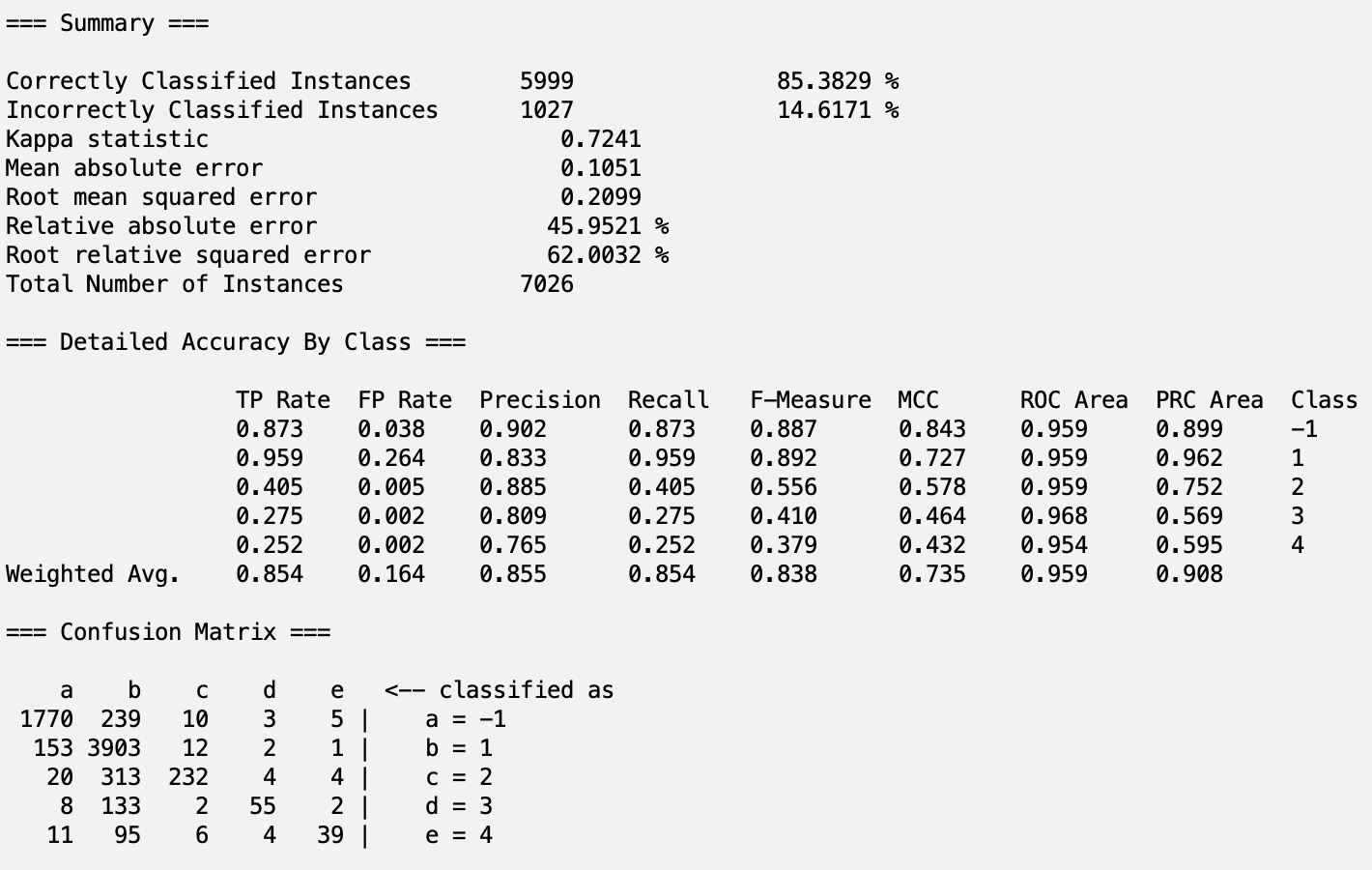
***J48 on gainEVAL***

******

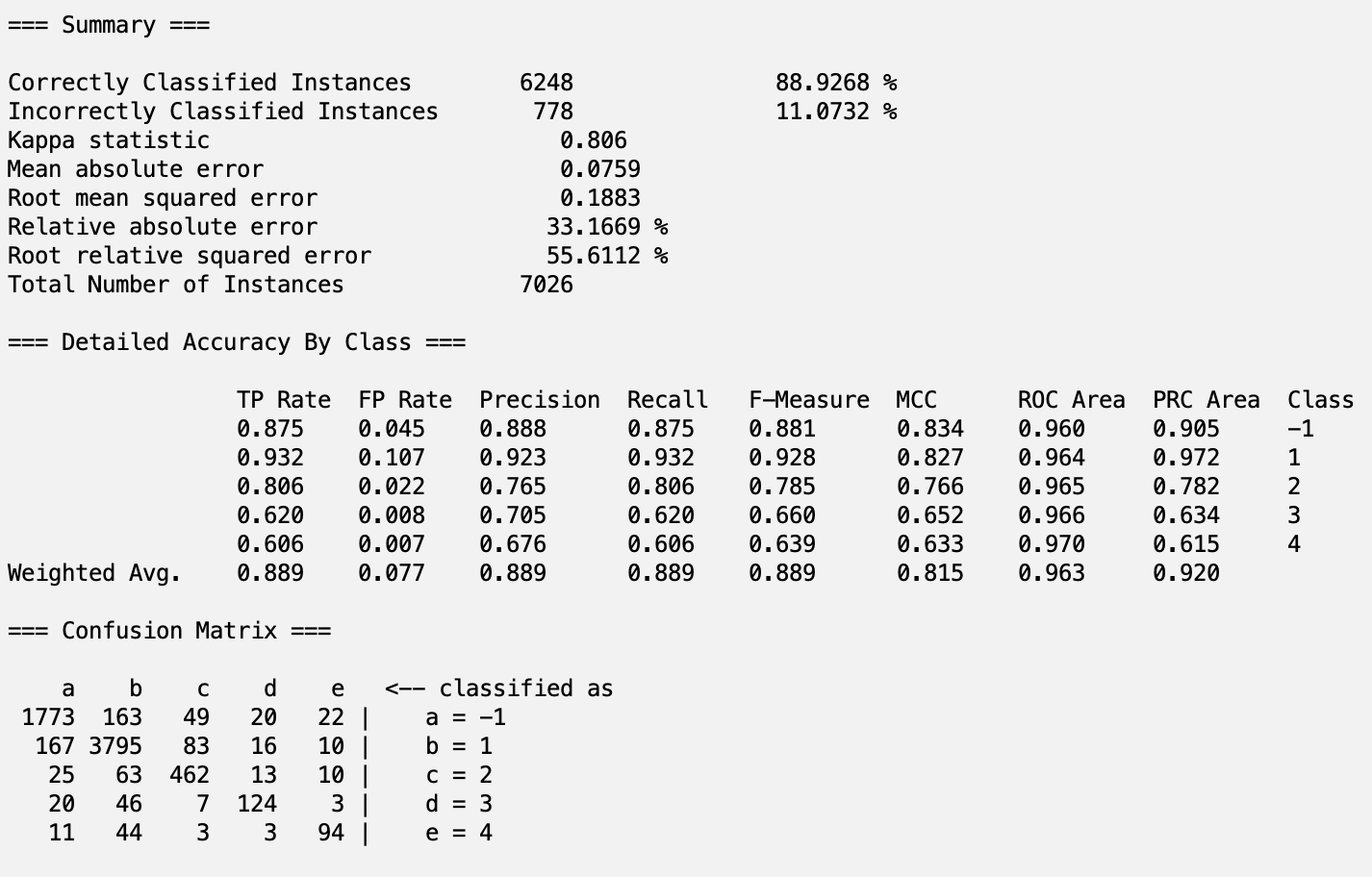
***Bagging on gainEVAL***

******

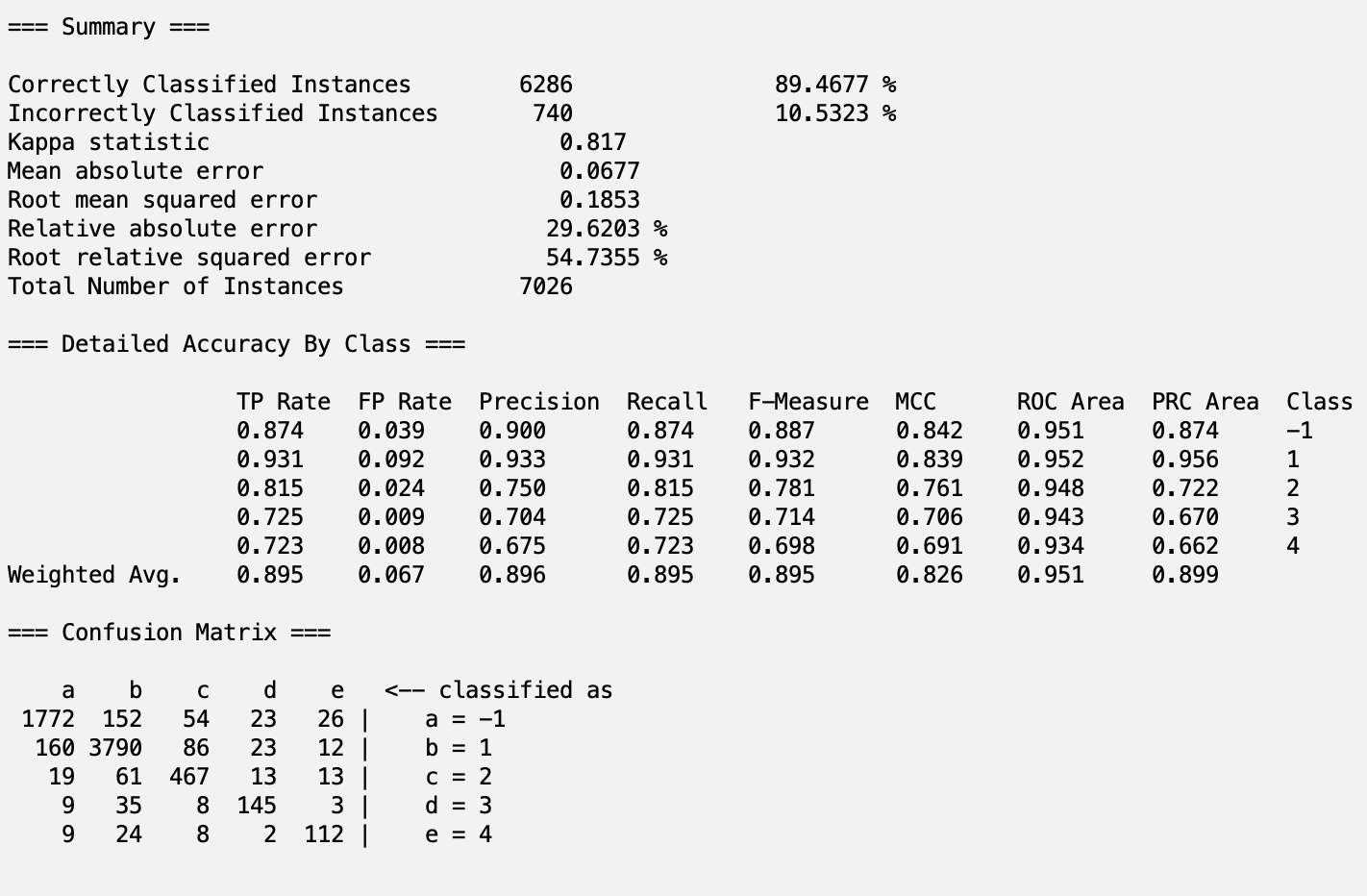
***KStar on gainEVAL***

******

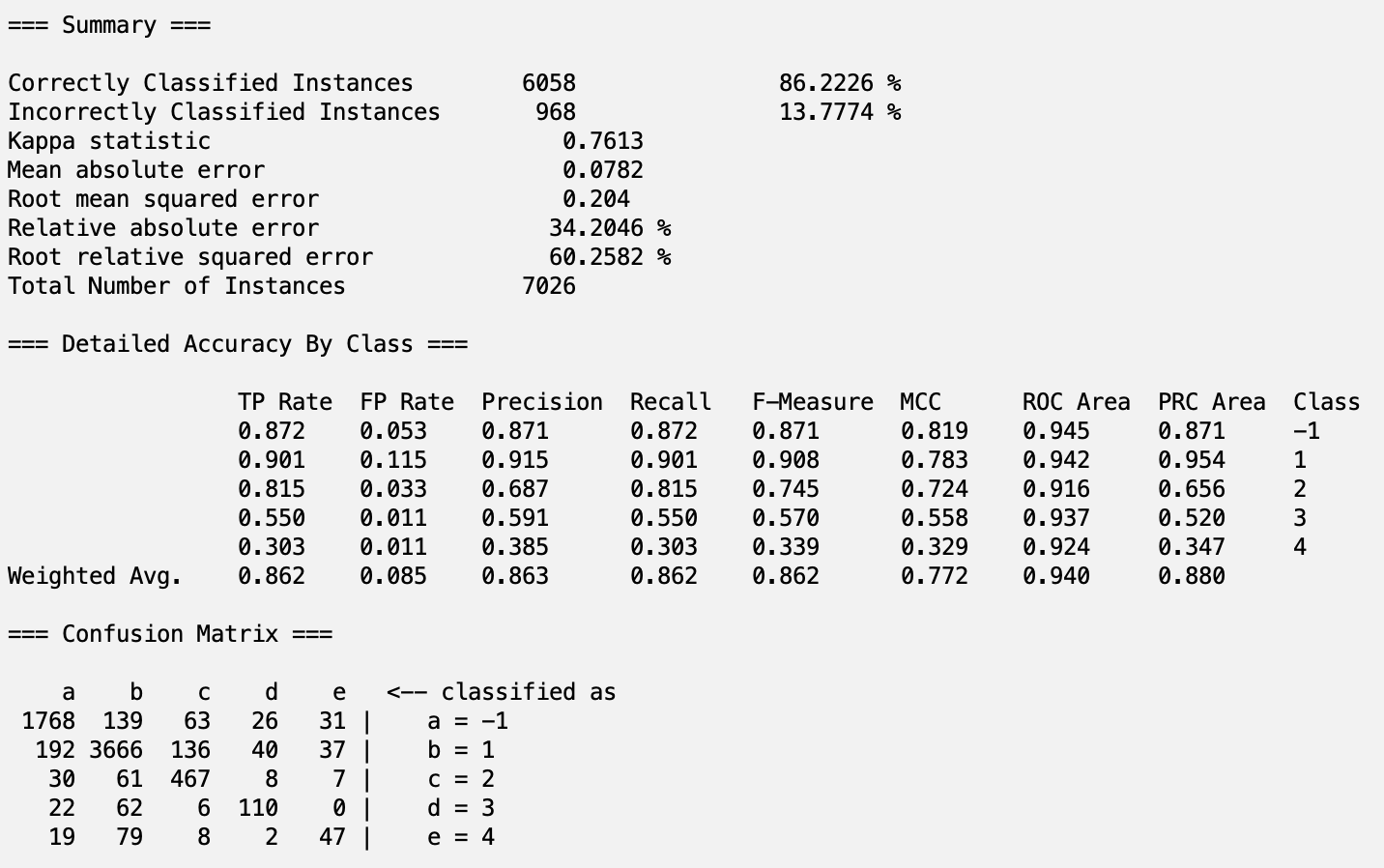
***Decision Table on symmEVAL***

******

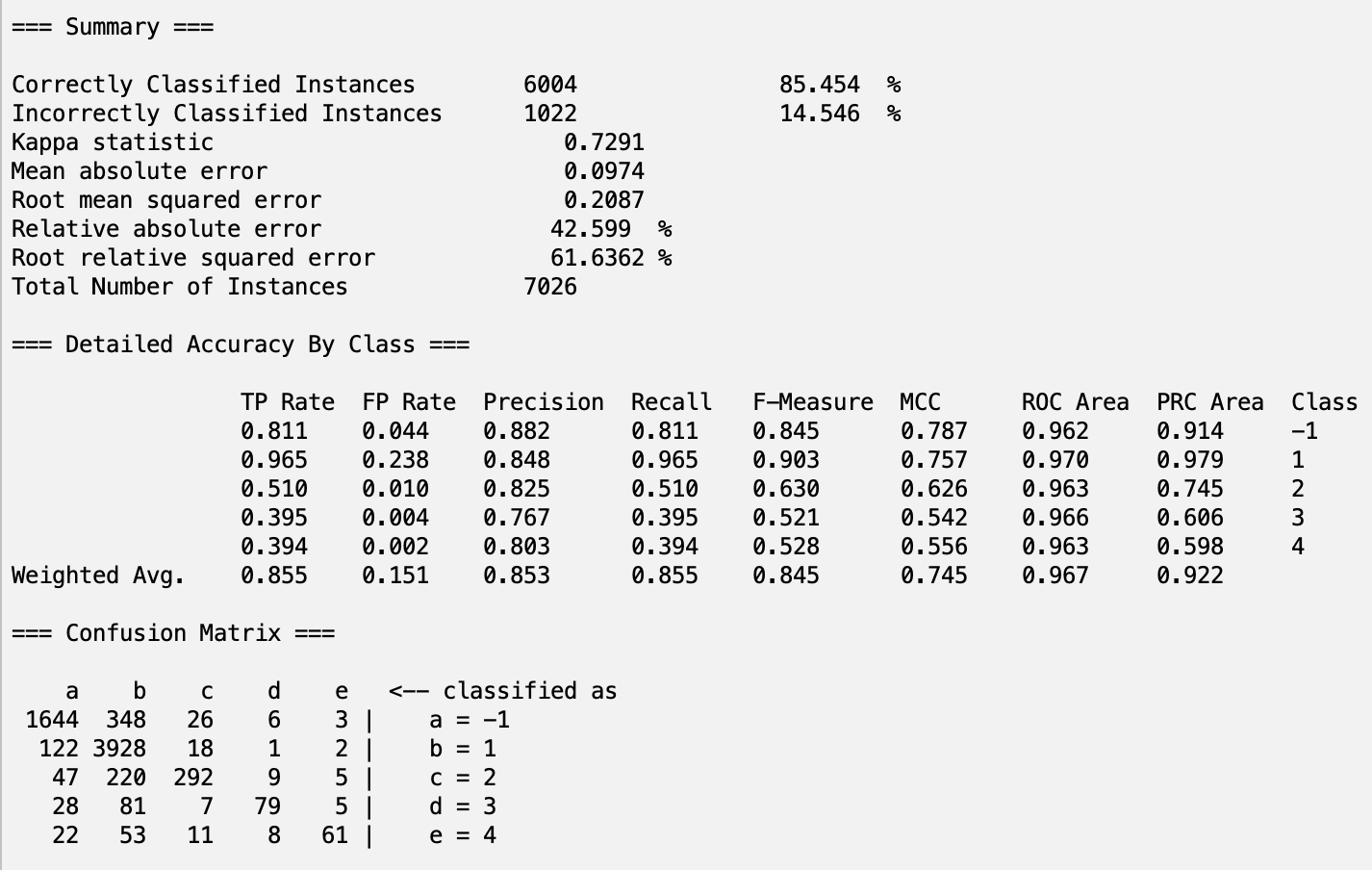
***J48 on symmEVAL***

******

***Bagging on symmEVAL***

******

***KStar on symmEVAL***

******

## 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  MLTFAMFLG\_A  EVRMARRIED\_A  SMKNOW\_A  PAITOOTH3M\_A  VIGIL4\_A  DISCRIM5\_A  MHTHND\_A  MHTHDLY\_A  HYSTEV | HRTESTLAST\_A  EMPDYSMSS3\_A  URBRRL  REGION  PSTRAT  LONGCOVD1\_A  VIGIL4\_A  SMKNOW\_A  INTV\_QRT  VIGIL3\_A | SOCSCLPAR\_A  COGMEMDFF\_A  LSATIS4\_A  PHSTAT\_A  SOCERRNDS\_A  DISCRIM5\_A  VIGIL4\_A  MHTHDLY\_A  MHTHND\_A | MHTHND\_A  MHTHDLY\_A  SOCSCLPAR\_A  TBIHLSBMC\_A  WTFA\_A  TBILCDCMG\_A  HOMEHC12M\_A  RXDG12M\_A  MEDNG12M\_A  MEDDL12M\_A | FAM\_A  WTFA\_A  MHTHND\_A  MHTHDLY\_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.