Fall 2018 CS286 Data Analysis and Prediction Project Report

Project Title: Predicting Onset of Mild Cognitive Impairment from NACC clinical data.

Group Number: Group 5

Project Team:

Pratik Patil

Shruti Kothari

Mrudula Murali

ABSTRACT

The results of directly comparing the prediction accuracy of different classifiers to predict the onset of Mild Cognitive Impairment from data of National Alzheimer’s Coordinating Center(NACC) are reported. Performance comparisons were carried out using 20 and 30 best subset features obtained from various dimensionality reduction and feature selection techniques. The dataset is obtained from NACC and contains information of various clinical variables for unique subjects. Dimensionality reduction and feature selection has been done on these variables, the results of which are used with different classifiers to make the prediction. A Decision Tree prediction accuracy of 90%(subset of 30 features using fclassif) was the highest accuracy observed. The model presented can help in predicting whether a person will or will not develop MCI thus, making the clinical trials cost effective as taking study participants at high risk for developing MCI, can help in achieving cost-effective prevention trials. Future work includes using deep learning models and doing a further review of the clinical variables which were present in the dataset but the description of which wasn’t available in the NACC data element dictionary [6] referred. We also intend to carry forward the technique of one-hot encoding on categorical columns which wasn’t completed due to time constraints.

Key Words: Mild cognitive impairment, machine learning, dimensionality reduction, NACC dataset, ROC AUC, hypothesis testing.

Table of Contents

List of Tables

List of Figures

INTRODUCTION

Mild cognitive impairment (MCI) is a common disorder, affecting 3–5% of adults over 60 and 15% of adults over 75. It is characterized by a decline in cognitive function that falls between the changes associated with typical aging and those fulfilling the criteria for dementia. Although MCI can convert to Alzheimer's disease (AD), community-based studies suggest that many individuals diagnosed with MCI do not progress to AD at an accelerated rate and may even revert to normal. But once a person develops AD, there is no cure. Thus, it is very important to identify MCI at an early stage, to reduce the risk of progression from MCI to AD.

A publication related to our analysis [1] focuses on enriching clinical trials to retard disease progression during presymptomatic phases of Mild Cognitive Impairment (MCI). It uses the National Alzheimer’s Coordinating Center (NACC) Uniform Data Sets version 2.0 for identifying subjects who are at high risk for developing MCI within 4 years from initial testing. Recruiting such subjects as study participants can aid carrying out cost-effective prevention trials. However, accurately identifying those who are destined to develop MCI is difficult.

Dale et. al, have shown that poor physical health, functional status, and depression are also associated with lower cognitive performance in the general population leading to the onset of deterioration of brain. Therefore, this study involves the use of such indicators as inputs from the subjects. [3]

The aim of our project is to build a model that predicts which subjects are suitable for dementia clinical trials i.e. identifying which subjects are likely to develop MCI within 4 years from initial testing and which are not. This study will help the broader scientific community by greatly reducing the costs involved in clinical trials.

PURPOSE AND QUESTIONS OF INTEREST

The study done by Ming Lin et. al has primarily focused on using different feature selection, however they did not achieve significant out-of-sample accuracy in predicting the onset of MCI [1]. The motivation of our study is use extensive pre-processing of the data by studying the significance of every feature in the available NACC dataset, and then using effective feature selection along with many different predictive models and eventually choosing the best model among them. Since the starting point of our study is [1], after completing the analysis we expect to have a model that is able to predict if an individual will or will not develop Mild Cognitive Impairment (MCI) within 4 years with at least 74% accuracy, which is the highest accuracy the paper [1] achieves using SVM.

DATA COLLECTION

As our analysis focuses on a patient’s medical condition, we used the data collected by NACC (National Alzheimer's Coordinating Center), a center for Alzheimer's disease research. The NACC dataset comprises of demographic data, neuropsychological testing scores, clinical diagnosis etc. Examples of types of data recorded in the dataset are:

- Behavioral - e.g. neuropsychiatric, geriatric, etc.
- Cognitive - e.g. memory, problem solving, judgement, etc.
- Physical - e.g. height, weight, heart rate, blood pressure, etc.

Each of these types of data play an important role in our analysis. For e.g., the logical memory test scores can significantly help in predicting the probability that a subject develops MCI. Lower test scores would mean that the subject is more likely to develop MCI. However, no judgement can be made based on just one particular data. These data in combination with each other can yield more informative results. Thus, this type of data is relevant to perform our analysis.

As we want our model to be able to make predictions as accurately as possible we need sufficiently large amount of data to obtain better accuracy as more data almost always leads to better results. It smooths out the noise and it allows you to get higher confidence in your results.

The data we have used comes from the clinical variables collected in the National Alzheimer’s Coordinating Center (NACC) downloaded April 2015. This dataset contains 31,872 unique subjects. Among them, 7026 subjects have normal cognition at baseline and are being used in our current analysis. Initially this dataset contains 688 clinical variables with an additional 128 derived variables were computed from Uniform Data sets version 2.0 (UDS 2.0).

The data we have has most of the features as categorical nominal. We need a sufficiently high correlation between the predictor variables and the target, however, an extremely high correlation can happen with binary predictor merely because of coincidence. This phenomenon has to be studied careful with this kind of dataset.

DATA CLEANING

As we had a large number of features almost 688 at the beginning, we decided to perform a two-step process for cleaning of our data. The first step being the pre-cleaning filter step wherein we focused on correcting the data using the rules defined in the data dictionary and the second step being the regular data cleaning which imputation, label encoding, text clustering, rounding the continuous values, etc.

- **Pre-cleaning filter**

  The features in our dataset are not self-explanatory as these are based on medical results and understanding what each feature means requires some domain knowledge. Therefore, to get some insight on these attributes we referred the NACC data dictionary [6] which gave a detailed explanation on what each variable means, the type and range of values it takes.

  So, in this step we worked on making our dataset adhere to the data dictionary. Our dataset originally had many discrepancies wherein the dependencies between the features were not in accordance with the dictionary. One such example is the dependency between the features ‘CBTIA’ (Transient ischemic attack) and ‘TIA1YR’, ’TIA2YR’, ’TIA3YR’, ’TIA4YR’ (year in which the attacks occurred). As per the dictionary if a subject doesn’t get CBTIA attack then the value for the CBTIA field should be 0 and all TIA years should be blank. But in our dataset, we had cases where the CBTIA field was 0 but the TIA years had some year values in them which was incorrect.

  After manually checking for all the dataset variables in the dictionary we found and corrected 78 features. We did not correct these 78 features simply by imputing them with any random number rather we handled each of these imputations differently by looking at the allowable codes in the data dictionary.

- **Data cleaning**

  We first dropped all the columns which had no data in them. We had 251 such columns. Then we dropped columns which had a single value for all the records, as a single value in a column cannot possibly contribute in predicting the target class. We dropped certain columns based on our understanding and dictionary reference which we felt were not important in making the prediction.

  We then dropped columns which by coincidence had a high correlation with the target variable. Basically, these were the features that were incorrectly determining the class label and were causing the accuracy to be 1.

  For all the descriptive columns in the dataset our first attempt was to perform NLP techniques to get information from them, we thus did clustering of all the text in columns using K-nearest neighbors and key collision with fingerprint method to reduce it to a single category. Details for this process is mentioned in the appendix section.

Certain ‘YEAR’ columns had fractional values, we rounded off these to the nearest numeric value. We did the same for columns which as per the data dictionary were categorical but still had numerical values in the dataset.

Categorical columns which had blank fields were imputed by 0, where 0 in most cases represented absence of disease.

We also tried one hot encoding all the categorical nominal columns, but this increased the number of columns exponentially as a majority of the columns in our dataset are categorical nominal.

Lastly, we label encoded all the numeric and categorical values before carrying out further analysis.

From the overall data cleaning process, we conclude that this reduction in dataset is optimal for our prediction problem and this was confirmed using the heatmap of the correlation dataset.

VISUALIZATION

- **t-SNE (t-distributed Stochastic Neighbor Embedding):**

Since we had very high dimensional data it was very difficult to visualize it. Therefore, the first thing that we tried was using t-SNE for visualizing the data. t-SNE is a tool to visualize high-dimensional data. It converts similarities between data points to joint probabilities and tries to minimize the Kullback-Leibler divergence between the joint probabilities of the low-dimensional embedding and the high-dimensional data. Below is how our data looked after t-SNE visualization.

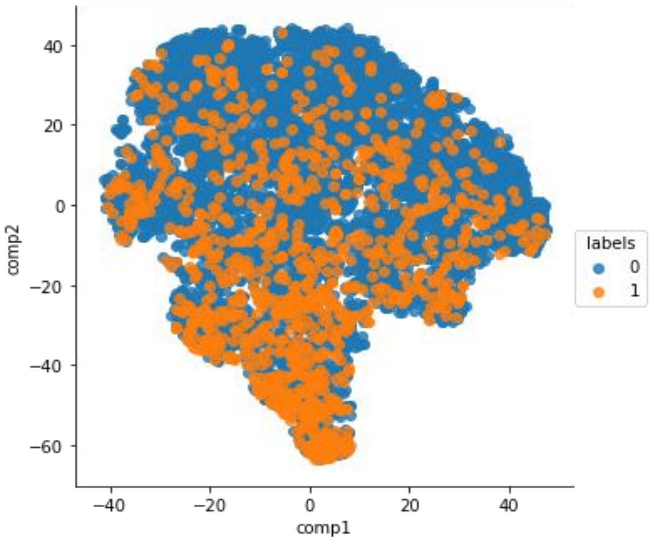


Figure 1 - Visualization of dataset using t-SNE

- **Histogram/Distribution**

  Below we have a histogram of all the continuous variables in the dataset. They seem to be normally distributed.

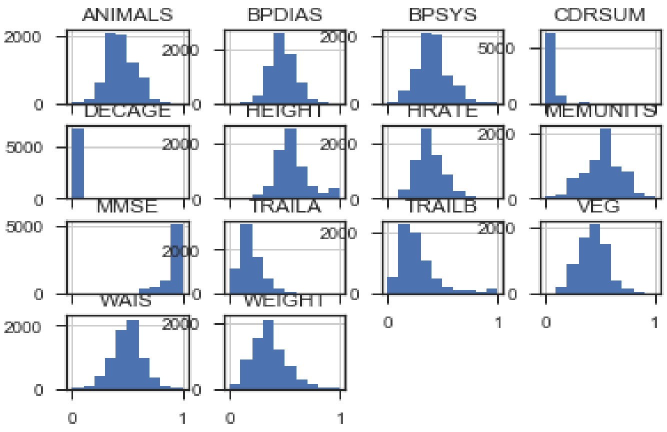


Figure 2 - Histogram of continuous variables

- **Correlation coefficient and Heatmap**

Since we had high dimensional data (700 predictor variables), initially it was very difficult to visualize. Therefore, we tried getting an initial overview using multiple Heatmaps of correlation data. We used Pearson correlation coefficient to calculate the correlation of every column. Many columns had a very high correlation of greater than 0.9, therefore we had a closer look at those columns to find if they have spurious correlation because of the improper data. For instance, a column had 2009 as a value for all the subjects with MCI (label 1) and 2008 for all subjects with no MCI (label 0). Therefore, the correlation for this was exactly 1, which is spurious. We dropped all such columns and eventually retained columns with greater than 0.4 correlation coefficient. Below we have heatmap of first 20 variables before cleaning and second heatmap is constructed after cleaning and final feature selection

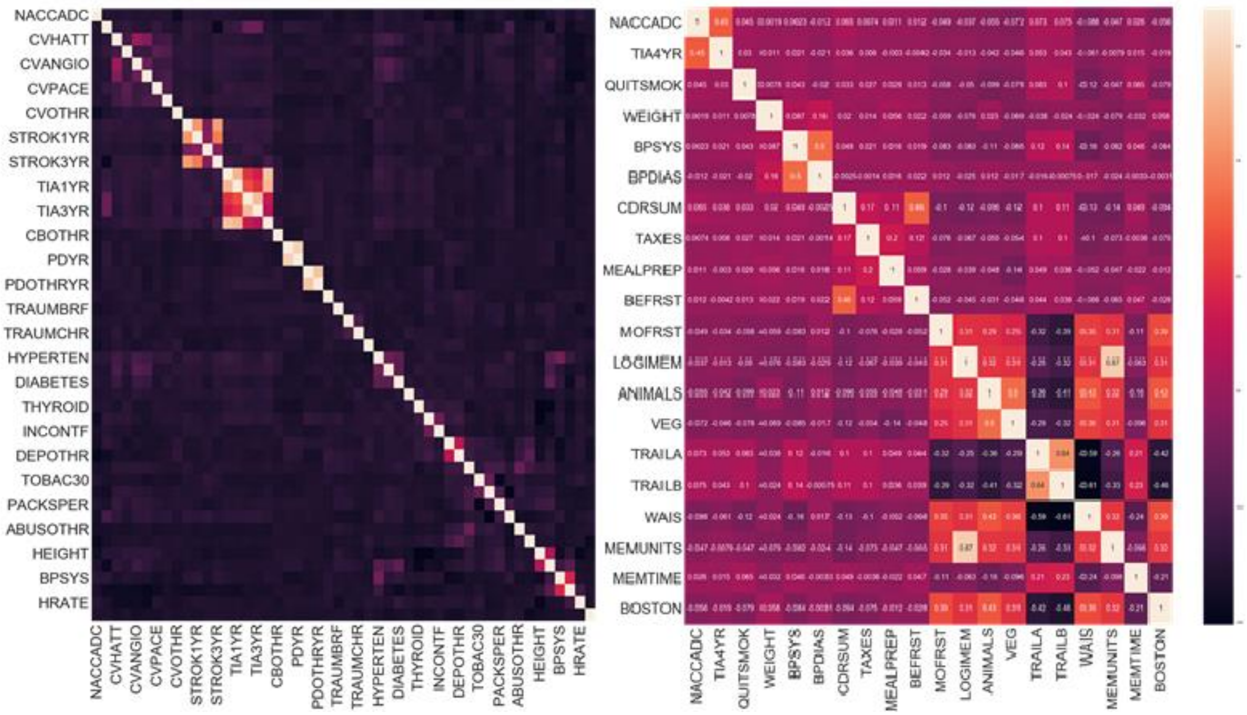


Figure 3 - (L to R) Heatmap of first 20 features before data cleaning, Heatmap of 20 selected features

- **Pair plot**

Most of our predictor variables were categorical and therefore it was difficult to visualize and get descriptive statistics. Therefore, here we focus only on the 14 variable which have continuous values. Below we have a pair plots for those continuous variables. They indicate that these variables have high variance and are correlated with the class variable.

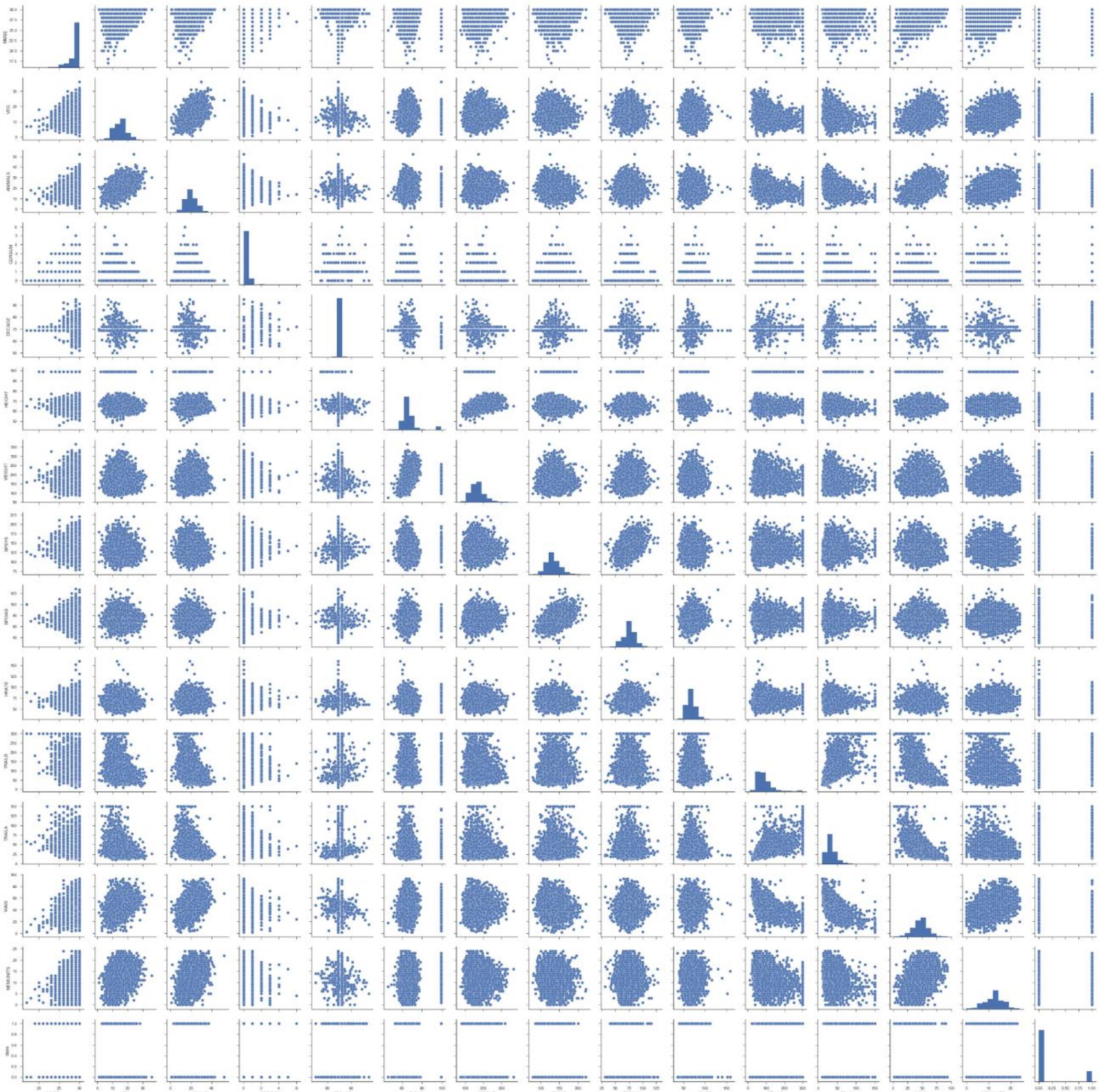


Figure 4 - Pair plot of continuous variables

- **Outliers**

We checked if we have outliers in continuous valued columns using box plots. We observed that huge number of subjects were market as outliers using the formula, (see the box plots below). Therefore, we decided to not remove those outliers as they would reduce the number of data points we have. This decision was taken because the count of all the continuous variables were fractional subset of our whole feature set.

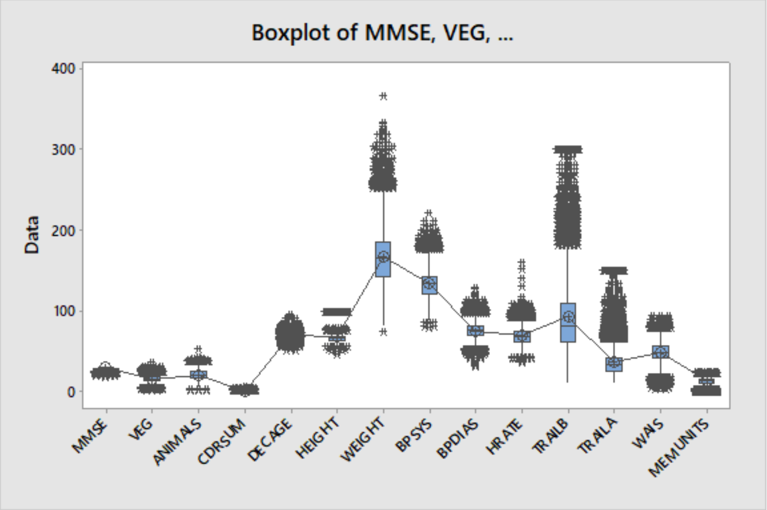


Figure 5- Boxplots showing outliers for continuous variables

PRUNING, REFINEMENT, NORMALIZATION, and IMPUTATION OF DATA

- **Pruning**:

We have used 4 dimensionality reduction techniques to get the best subset of features. Using every technique, we have extracted 20 and 30 best features. The techniques we have used are as below:

1. Principal Component Analysis:
   - Linear dimensionality reduction using Singular Value Decomposition of the data to project it to a lower dimensional space.
2. chi square:
   - This score can be used to select the n_features,features with the highest values for the test chi-squared statistic from X, which must contain only non-negative features such as booleans or frequencies (e.g., term counts in document classification), relative to the classes.
3. f_classif:

   Compute the ANOVA F-value for the provided sample.

4. mutual_info_classif:

   Estimate mutual information for a discrete target variable.Mutual information (MI) between two random variables is a non-negative value, which measures the dependency between the variables. It is equal to zero if and only if two random variables are independent, and higher values mean higher dependency.

- **Imputation**:

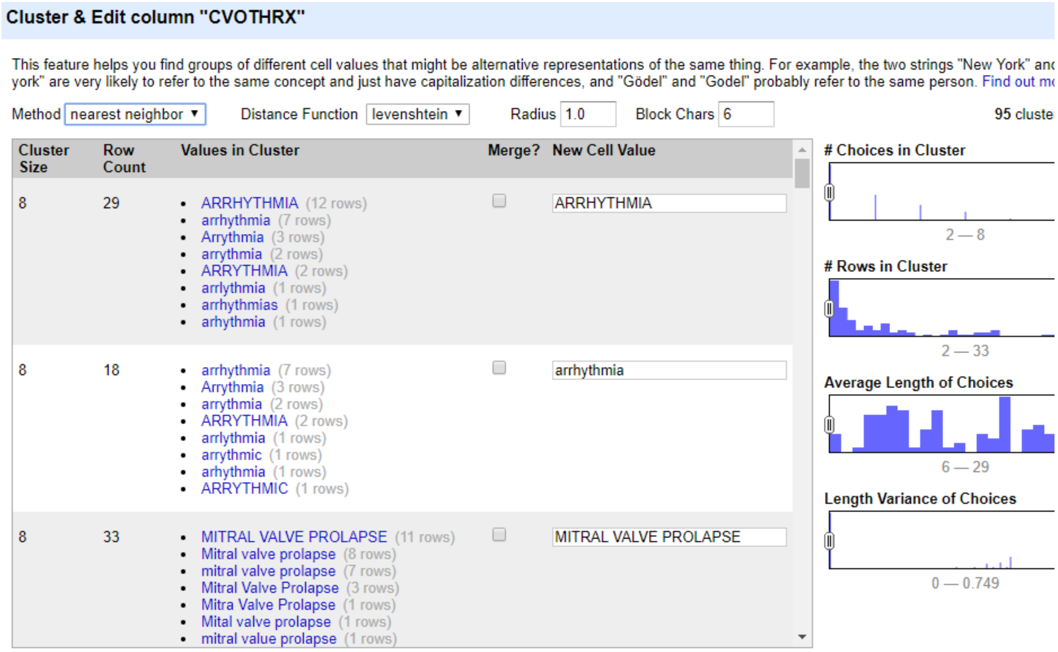
  Majority of our data was categorical data. Only 14 out of 688 columns were continuous. Imputing with mean or any another statistical method would not make sense for imputing categorical variables. Therefore, we imputed missing values with 0.

- **Normalization**:

  We have normalized all the continuous data columns between 0 and 1 using MinMax scaler.

- **Refinement: Text Clustering:**

The dataset had a few columns which were text columns. On close examination we found that most of the values were misspelled even though they mean the same thing. Therefore, we used text clustering using KNN to cluster misspelled words and replaced it with a single correct spelling. Later these corrected texts were label encoded.



Cluster & Edit column "CVOTHRX"

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" anc york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. Find out mc

Method nearest neighbor ▼ Distance Function levenshtein ▼ Radius 1.0 Block Chars 6 95 cluste

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
8	29	• ARRHYTHMIA (12 rows) • arrhythmia (7 rows) • Arrythmia (3 rows) • arrythmia (2 rows) • ARRYTHMIA (2 rows) • arrlythmia (1 rows) • arrhythmias (1 rows) • arhythmia (1 rows)	□	ARRHYTHMIA
8	18	• arrhythmia (7 rows) • Arrythmia (3 rows) • arrythmia (2 rows) • ARRYTHMIA (2 rows) • arrlythmia (1 rows) • arrythmic (1 rows) • arhythmia (1 rows) • ARRYTHMIC (1 rows)	□	arrhythmia
8	33	• MITRAL VALVE PROLAPSE (11 rows) • Mitral valve prolapse (8 rows) • mitral valve prolapse (7 rows) • Mitral Valve Prolapse (3 rows) • Mitra Valve Prolapse (1 rows) • Mital valve prolapse (1 rows) • mitral value prolapse (1 rows)	□	MITRAL VALVE PROLAPSE

Figure 6- Clustering of textual data

DATA ANALYSIS

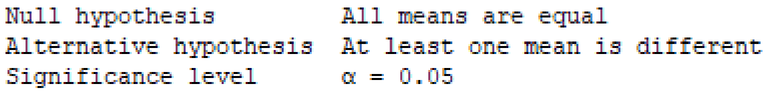
- **Correlation**:

We calculated a correlation matrix where correlation coefficient of every column was calculated for every other column. Since there were a huge number of columns, it was not feasible to manually check if a pair of columns are correlated/ independent. Therefore, we plotted it as a heatmap and evaluated the results. Below is a sample correlation heatmap. We also set a threshold of 0.4 for a correlation coefficient with the predictor class. Therefore any column with a correlation of 0.4 was removed.

- **ANOVA**:

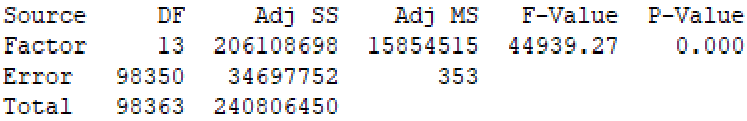
We checked if the continuous variables we have are actually independent using ANOVA test. The test conducted had alpha of 0.05. Below are the results for the ANOVA test. We conclude that the columns are in fact independent, because of the different mean.

1. Hypothesis:

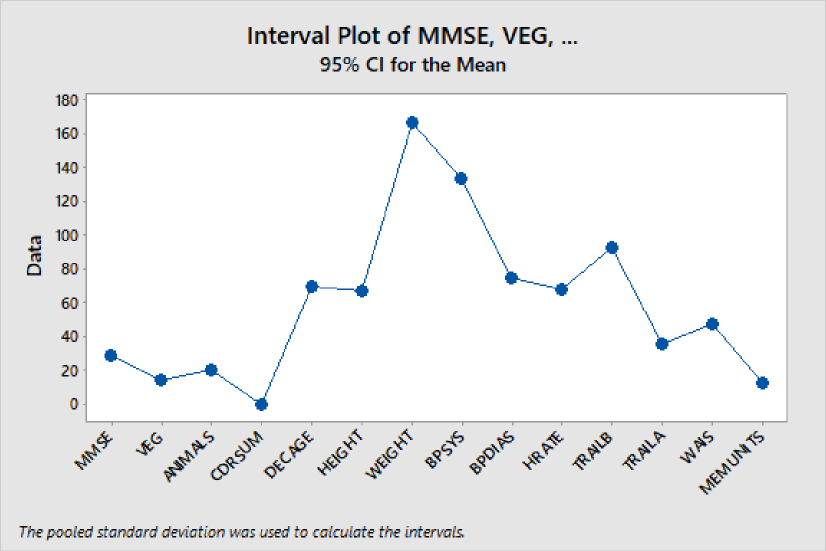


```
Null hypothesis         All means are equal
Alternative hypothesis  At least one mean is different
Significance level      α = 0.05
```

2. Results:



```
Source     DF     Adj SS    Adj MS   F-Value  P-Value
Factor     13  206108698  15854515  44939.27    0.000
Error   98350   34697752       353
Total   98363  240806450
```

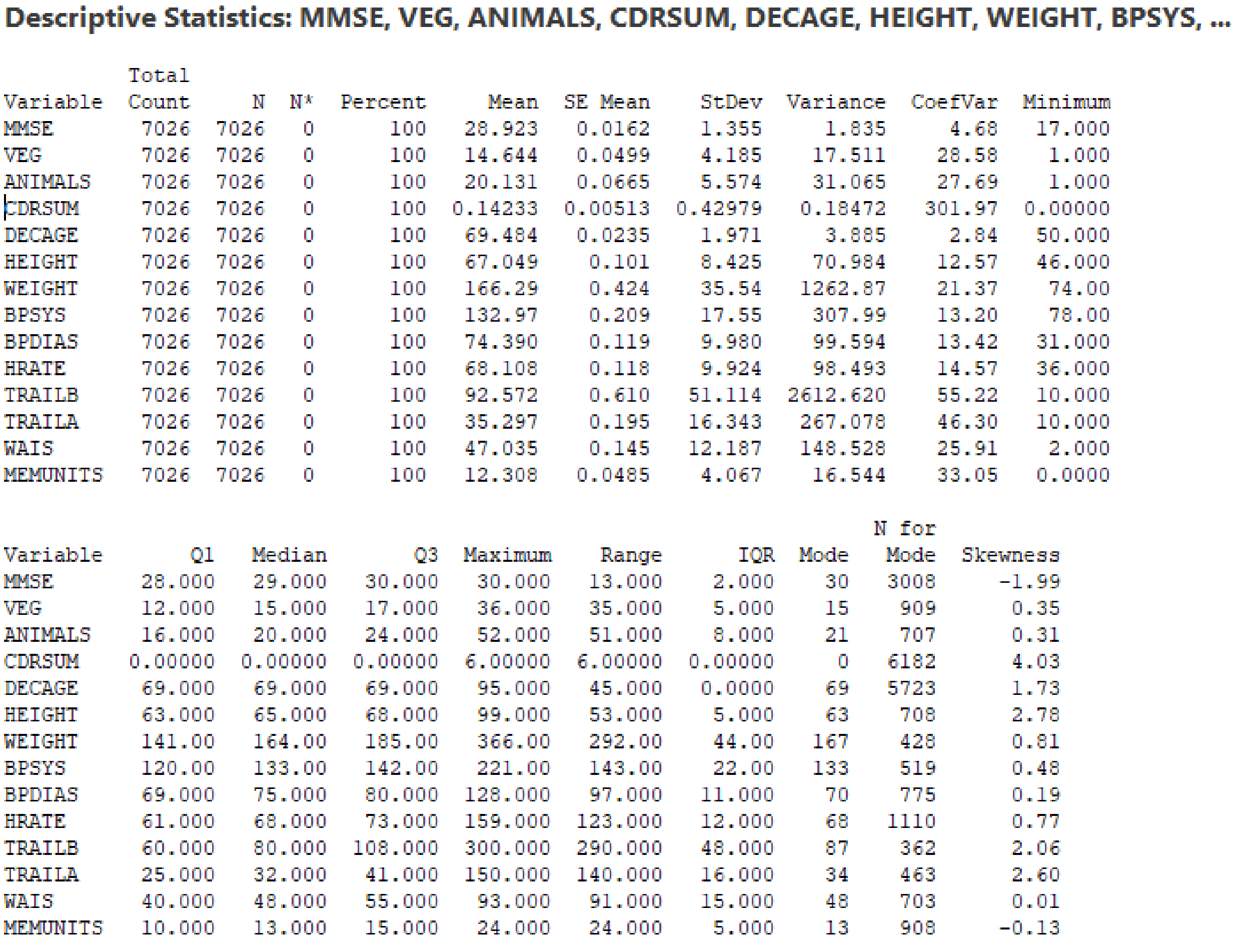


The pooled standard deviation was used to calculate the intervals.

Figure 7- 95% CI for continuous variables

- **Descriptive Analysis:**

Below is the descriptive analysis done on the continuous variables.



```
Descriptive Statistics: MMSE, VEG, ANIMALS, CDRSUM, DECAGE, HEIGHT, WEIGHT, BPSYS, ...

          Total
Variable  Count     N  N*  Percent     Mean  SE Mean     StDev  Variance  CoefVar  Minimum
MMSE       7026  7026   0      100   28.923   0.0162     1.355     1.835     4.68   17.000
VEG        7026  7026   0      100   14.644   0.0499     4.185    17.511    28.58    1.000
ANIMALS    7026  7026   0      100   20.131   0.0665     5.574    31.065    27.69    1.000
CDRSUM     7026  7026   0      100  0.14233  0.00513   0.42979   0.18472   301.97  0.00000
DECAGE     7026  7026   0      100   69.484   0.0235     1.971     3.885     2.84   50.000
HEIGHT     7026  7026   0      100   67.049    0.101     8.425    70.984    12.57   46.000
WEIGHT     7026  7026   0      100   166.29    0.424     35.54   1262.87    21.37    74.00
BPSYS      7026  7026   0      100   132.97    0.209     17.55    307.99    13.20    78.00
BPDIAS     7026  7026   0      100   74.390    0.119     9.980    99.594    13.42   31.000
HRATE      7026  7026   0      100   68.108    0.118     9.924    98.493    14.57   36.000
TRAILB     7026  7026   0      100   92.572    0.610    51.114  2612.620    55.22   10.000
TRAILA     7026  7026   0      100   35.297    0.195    16.343   267.078    46.30   10.000
WAIS       7026  7026   0      100   47.035    0.145    12.187   148.528    25.91    2.000
MEMUNITS   7026  7026   0      100   12.308   0.0485     4.067    16.544    33.05   0.0000

                                                                       N for
Variable       Q1   Median       Q3  Maximum    Range      IQR  Mode   Mode  Skewness
MMSE       28.000   29.000   30.000   30.000   13.000    2.000    30   3008     -1.99
VEG        12.000   15.000   17.000   36.000   35.000    5.000    15    909      0.35
ANIMALS    16.000   20.000   24.000   52.000   51.000    8.000    21    707      0.31
CDRSUM    0.00000  0.00000  0.00000  6.00000  6.00000  0.00000     0   6182      4.03
DECAGE     69.000   69.000   69.000   95.000   45.000   0.0000    69   5723      1.73
HEIGHT     63.000   65.000   68.000   99.000   53.000    5.000    63    708      2.78
WEIGHT     141.00   164.00   185.00   366.00   292.00    44.00   167    428      0.81
BPSYS      120.00   133.00   142.00   221.00   143.00    22.00   133    519      0.48
BPDIAS     69.000   75.000   80.000  128.000   97.000   11.000    70    775      0.19
HRATE      61.000   68.000   73.000  159.000  123.000   12.000    68   1110      0.77
TRAILB     60.000   80.000  108.000  300.000  290.000   48.000    87    362      2.06
TRAILA     25.000   32.000   41.000  150.000  140.000   16.000    34    463      2.60
WAIS       40.000   48.000   55.000   93.000   91.000   15.000    48    703      0.01
MEMUNITS   10.000   13.000   15.000   24.000   24.000    5.000    13    908     -0.13
```

Figure 8 - Descriptive statistics of continuous variables

- **Predictive Analysis:**

We performed binary classification on the NACC dataset, where we have multiple predictor variables and a class which state within 4 years if the subject gets MCI or not. 5 models were trained with l2 regularization. Each model was trained on 8 subset of original dataset. Each subset was derived using a different feature selection or dimensionality reduction technique. The 5 models and subsets of data are listed below:

- Models:
  - Logistic Regression
  - SVM
  - Naive Bayes
  - Decision Tree

- Subsets of dataset:
    - 20 Features:
        - PCA
        - Chi2
        - fclassif
        - mutual info classif
    - 30 Features:
        - PCA
        - Chi2
        - fclassif
        - mutual info classif

INTERPRETATION OF ANALYSIS RESULTS

- **Continuous Data:**

From the Analysis of one way variance (ANOVA), we got a p value of 0. Since our hypothesis was that all the means are same, on this part we conclude that all the columns are statistically different and independent.

Therefore, we can use them in our prediction as they are independent.

A few columns had a very high variance, like columns Trailb, height, etc. Therefore, we decided to use a normalizer to normalize all the continuous valued columns.

We feel we have good results with the subset of continuous columns.

- **Categorical Data:**

For the categorical data we created a heatmap of correlation using Pearson correlation coefficient. We also removed columns which had coefficient of greater than 0.4, we did this because of some columns had spurious correlation. For instance, a column had 2009 as a value for all the subjects with MCI (label 1) and 2008 for all subjects with no MCI (label 0). Therefore, the correlation for this was exactly 1, which is spurious

EVALUATION OF ANALYSIS RESULTS

From the ANOVA on the continuous data we conclude that we have independent set of features which can be used in the prediction model.

Also, as far as our categorical variables are concerned, we had removed columns which have spurious correlation, implying that our new subset of feature have real information instead of just random chance correlation.

Therefore, our analysis results do in fact help us to meet our originally state goals. Thus using the above state analysis techniques, we were successfully able to remove noisy information.

We also needed to go back and forth between the analysis and the updation of our policies for filtering data. For example, we had to experiment with the threshold value of correlation value which gave us optimal subset of feature.

PREDICTIVE MODEL AND PREDICTION RESULTS

- **Cross Validation:**

In all of our predictive models we have used a 5-fold cross validation, where the training data and test data was split into 80% - 20% respectively.

- **Predictive Model:**

We have fit our data to 4 different classification models using different regularization techniques. Below are the list of models we have used.

1. Logistic Regression
2. SVM
3. Naive Bayes
4. Decision Tree

a) Logistic Regression

We choose logistic regression because the prediction of onset of MCI is a logistic regression problem where we predict 1 or 0 for the class/target. We used l2 regularization with this model to reduce the overfitting. We chose this model also because we wanted to compare our results to the study [1].

b) SVM

SVM was chosen because it performs well in case of a binary plane separation. SVM projects the features from lower dimensional space to a higher dimensional space and puts a separating hyperplane which was perfect for this data and for this problem. We chose this model also because we wanted to compare our results to the study [1]

c) Naive Bayes

Naive Bayes with a binomial NB was chosen as we can formulate this problem in terms of probability as well. For example, given X,Y,Z…. features, what is the probability that the given subject will acquire dementia/MCI or not.

d) Decision Tree

Decision tree was chosen because it is well known to generalize the data, specially in case when the dataset contains large number of features. And since we had lot of features we chose decision tree as a classifier.

- **Prediction Results:**
  a. **Results table:**

Below are the results for our predictive analysis:

Our best model is Decision Tree Classifier with dataset of 30 features obtained from fclassif feature selection. This accuracy is higher than achieved in the study [1]. We feel this is a very good accuracy. From the AUC-ROC we conclude that the model classifies well with a reasonable false positives, false negative, true positive and true negative.

	Chi2		fclassif		mutual info classif		PCA	
Model	20	30	20	30	20	30	20	30
Logistic Regression	86	86	85	85	86	86	85	86
SVM	85	85	85	85	85	85	85	85
Naive Bayes	82	82	82	82	82	83	84	85
Decision Trees	88	89	77	90	89	89	78	78

Table 1 - Accuracy table result for different machine learning classifiers

b. AUC - ROC

Below we have ROC for the best model, decision tree with 30 feature set. Area under the ROC curve is **0.8037**

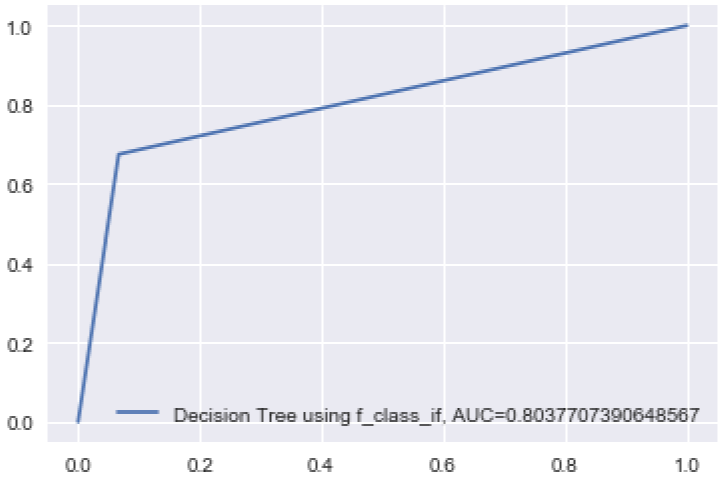


Figure 9 - AUC-ROC curve for decision tree classifier

CONCLUSION AND DISCUSSION

We did a detailed analysis of the NACC dataset and cross interpreted the rules and constraints for every variable from the NACC data dictionary [6].Our analysis on this concludes that the variables are multi-level, meaning that the value of every variable depends on some other variable and the value of that variable in turn depends on another. So, using the dataset as a sparse input the model will not work.

We have taken into account the multi-level dependency of each variable and created a tree hierarchy. We have used this to impute our data accordingly.

The result of this is that we have a significantly higher accuracy compared to other studies.

We also attribute the performance of our models to the feature selection methods used in our study. These methods produced completely different variables than the ones used in the paper[1]. We have used 4 different models each with different feature sets in our project and we have compared the accuracy of each.

Our model helps in detecting the early onset of MCI, which if not treated eventually leads to dementia. Therefore, this model can be used in pre-clinical trials which segregates which subjects are suitable for more detailed clinical trials for dementia study.

FUTURE WORK

One future approach can be to study the efficacy of the Artificial neural network to solve this problem. We can also do feature engineering using one-hot encoding of the categorical variables. We tried this in our current analysis, but due to time limitations, we decided not to explore further. So, we would like to try one hot encoding, as it helps in efficient implementation of machine learning algorithms.

Another important point to note is the further detailed study of the clinical variables form the data dictionary. The dictionary we used did not have rules and description for around 100 of the variables. Therefore, future work would include the analysis of these variables considering their multi-level dependencies.

REFERENCES

[1] Lin, Ming & Gong, Pinghua & Yang, Tao & Ye, Jieping & Albin, Roger & Dodge, Hiroko. (2017). Big Data Analytical Approaches to the NACC Dataset: Aiding Preclinical Trial Enrichment. Alzheimer Disease & Associated Disorders. 32. 1. 10.1097/WAD.0000000000000228.

[2] Di Stefano, Francesca & Epelbaum, Stéphane & Coley, Nicola & Cantet, Christelle & Hampel, Harald & Bakardjian, Hovagim & Lista, Simone & Vellas, Bruno & Dubois, Bruno & Andrieu, Sandrine. (2015). Prediction of Alzheimer's Disease Dementia: Data from the GuidAge Prevention Trial. Journal of Alzheimer's disease : JAD. 48. 10.3233/JAD-150013.

[3] Dale, William, Kotwal, Ashwin A., Shega, Joseph W., Schumm, L. Philip, Kern, David W., Pinto, Jayant M., Pudelek, Kelly M., Waite, Linda J., McClintock, Martha K.: Cognitive Function and its Risk Factors Among Older US Adults Living at Home. Alzheimer Dis Assoc Disord 32(3): 207-213, 2018

[4] Rabin, L. A., Wang, C., Katz, M. J., Derby, C. A., Buschke, H., & Lipton, R. B. (2012). Predicting Alzheimer's disease: neuropsychological tests, self-reports, and informant reports of cognitive difficulties. Journal of the American Geriatrics Society, 60(6), 1128-34.

[5] Pereira, Telma & Lemos, Luís & Cardoso, Sandra & Silva, Dina & Pina Rodrigues, Ana & Santana, Isabel & de Mendonça, Alexandre & Guerreiro, Manuela & C . Madeira, Sara. (2017). Predicting progression of mild cognitive impairment to dementia using neuropsychological data: A supervised learning approach using time windows. BMC Medical Informatics and Decision Making. 17. 10.1186/s12911-017-0497-2.

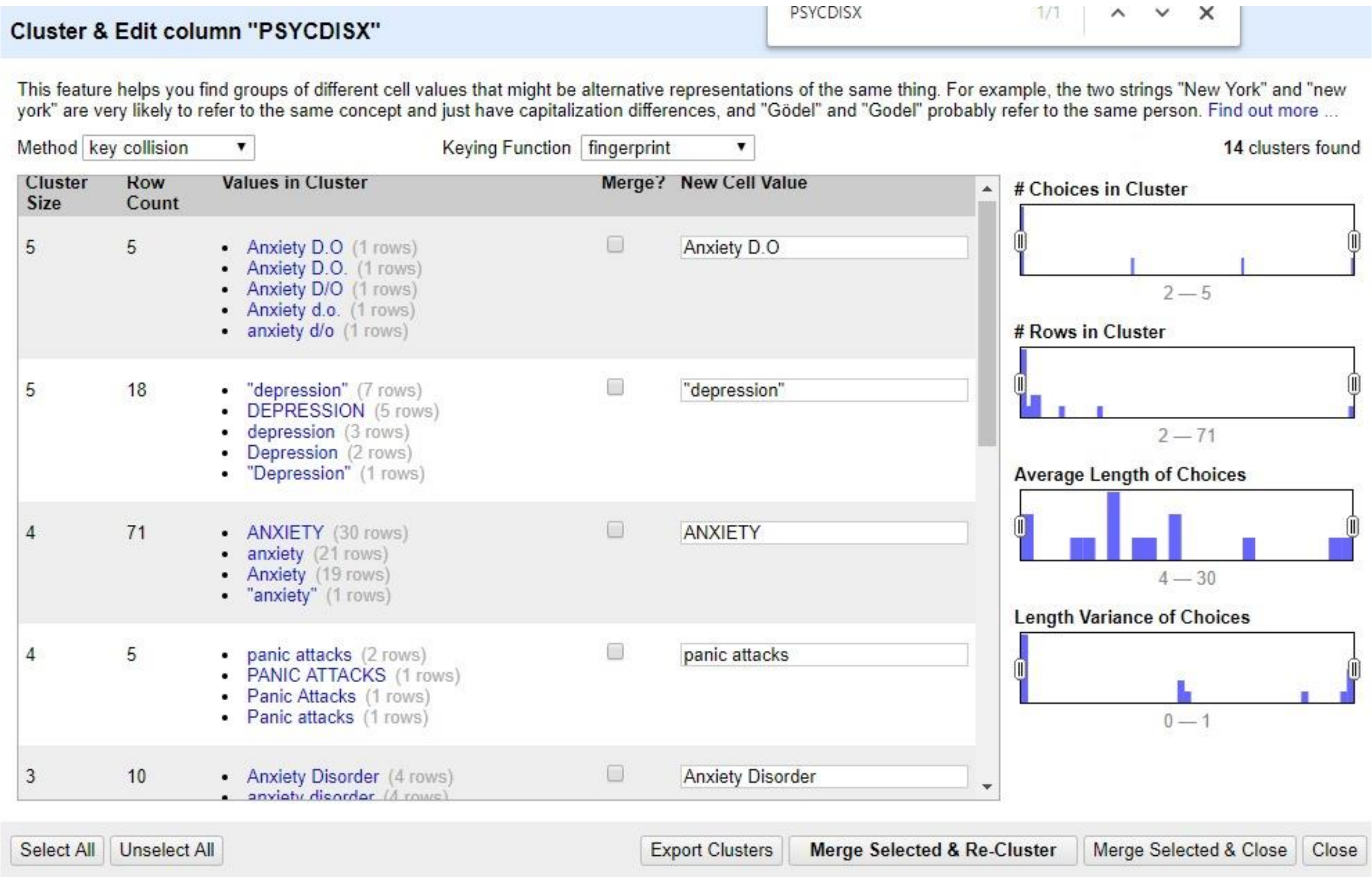
[6] https://www.alz.washington.edu/NONMEMBER/UDS/DOCS/VER1_2/ided.pdf

GROUP PARTICIPATION

- Pratik
  - 80% data analysis, model building and prediction
- Shruti
  - 80% data cleaning and pruning
- Mrudula
  - 80% visualization, refinement and normalization

APPENDICES

1) Clustering of text data



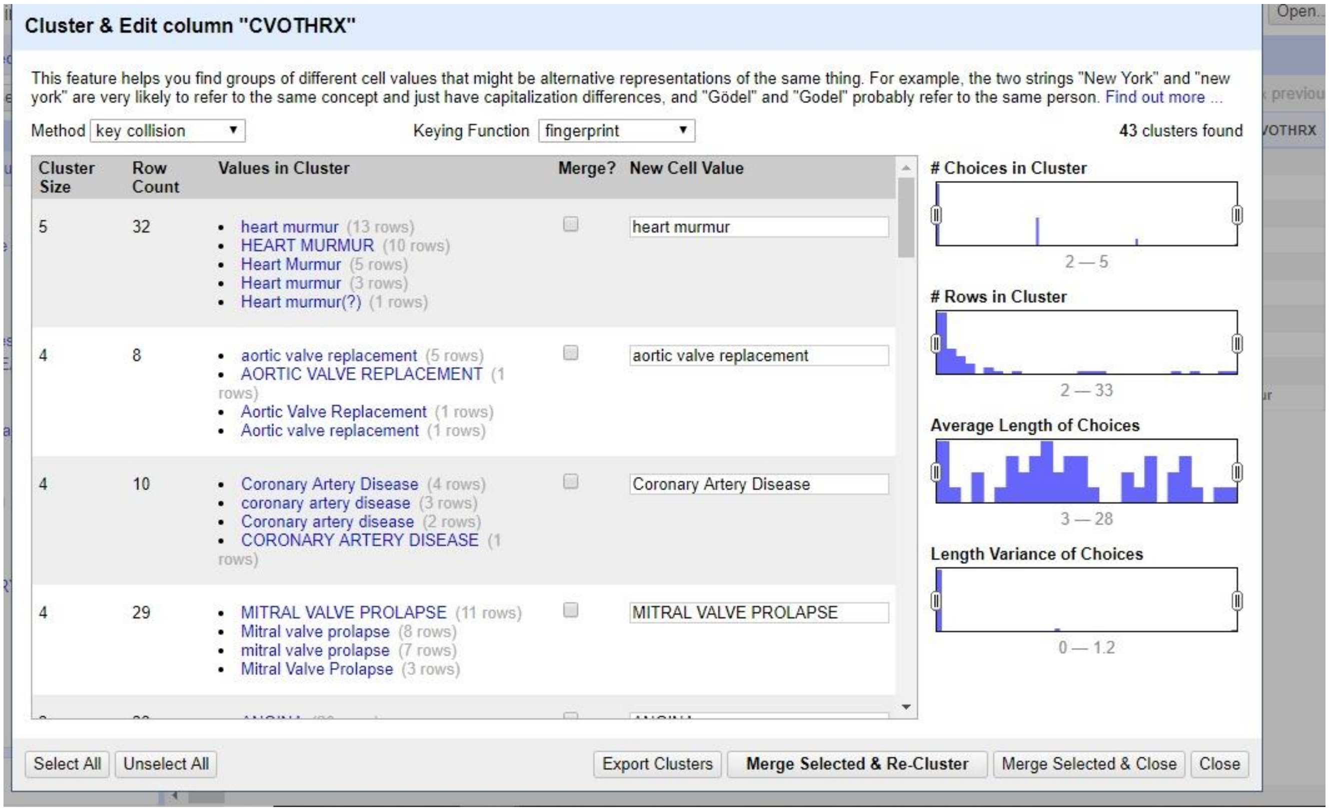
PSYCDISX 1/1

Cluster & Edit column "PSYCDISX"

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. Find out more ...

Method: key collision ▼ Keying Function: fingerprint ▼ **14** clusters found

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
5	5	• Anxiety D.O (1 rows) • Anxiety D.O. (1 rows) • Anxiety D/O (1 rows) • Anxiety d.o. (1 rows) • anxiety d/o (1 rows)	□	Anxiety D.O
5	18	• "depression" (7 rows) • DEPRESSION (5 rows) • depression (3 rows) • Depression (2 rows) • "Depression" (1 rows)	□	"depression"
4	71	• ANXIETY (30 rows) • anxiety (21 rows) • Anxiety (19 rows) • "anxiety" (1 rows)	□	ANXIETY
4	5	• panic attacks (2 rows) • PANIC ATTACKS (1 rows) • Panic Attacks (1 rows) • Panic attacks (1 rows)	□	panic attacks
3	10	• Anxiety Disorder (4 rows) • anxiety disorder (4 rows)	□	Anxiety Disorder



Cluster & Edit column "CVOTHRX"

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. Find out more ...

Method: key collision

Keying Function: fingerprint

43 clusters found

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
5	32	• heart murmur (13 rows) • HEART MURMUR (10 rows) • Heart Murmur (5 rows) • Heart murmur (3 rows) • Heart murmur(?) (1 rows)	□	heart murmur
4	8	• aortic valve replacement (5 rows) • AORTIC VALVE REPLACEMENT (1 rows) • Aortic Valve Replacement (1 rows) • Aortic valve replacement (1 rows)	□	aortic valve replacement
4	10	• Coronary Artery Disease (4 rows) • coronary artery disease (3 rows) • Coronary artery disease (2 rows) • CORONARY ARTERY DISEASE (1 rows)	□	Coronary Artery Disease
4	29	• MITRAL VALVE PROLAPSE (11 rows) • Mitral valve prolapse (8 rows) • mitral valve prolapse (7 rows) • Mitral Valve Prolapse (3 rows)	□	MITRAL VALVE PROLAPSE