Analyzing Mental Health Data

FARHAD DE SOUSA

MATH 550, FALL 2022

Data Set

- Data collected for 2019 by the Substance Abuse and Mental Health Data Archive (SAMHDA); **6,362,044 data points, 40 features**

Variable	Type	Length	Label
ADHDFLG	Numeric	8	Attention deficit/hyperactivity disorder reported
AGE	Numeric	8	Age (recoded)
ALCSUBFLG	Numeric	8	Alcohol or substance-related disorder reported
ANXIETYFLG	Numeric	8	Anxiety disorder reported
BIPOLARFLG	Numeric	8	Bipolar disorder reported
CASEID	Numeric	8	Case identification number
CMPSERVICE	Numeric	8	SMHA-funded/operated community-based program
CONDUCTFLG	Numeric	8	Conduct disorder reported
DELIRDEMFLG	Numeric	8	Delirium/dementia disorder reported
DEPRESSFLG	Numeric	8	Depressive disorder reported
DETNLF	Numeric	8	Detailed 'not in labor force' category
DIVISION	Numeric	8	Census division
EDUC	Numeric	8	Education
EMPLOY	Numeric	8	Competitive employment status (aged 16 years and older) at discharge or end of the reporting period
ETHNIC	Numeric	8	Hispanic or Latino origin (ethnicity)
GENDER	Numeric	8	Sex
IJSSERVICE	Numeric	8	Institutions under the justice system
LIVARAG	Numeric	8	Residential status — at discharge or end of reporting period
MARSTAT	Numeric	8	Marital status
MH1	Numeric	8	Mental health diagnosis one
MH2	Numeric	8	Mental health diagnosis two
MH3	Numeric	8	Mental health diagnosis three
NUMMHS	Numeric	8	Number of mental health diagnoses reported
ODDFLG	Numeric	8	Oppositional defiant disorder reported
OPISERVICE	Numeric	8	Other psychiatric inpatient
OTHERDISFLG	Numeric	8	Other mental disorder reported
PDDFLG	Numeric	8	Pervasive developmental disorder reported
PERSONFLG	Numeric	8	Personality disorder reported
RACE	Numeric	8	Race
REGION	Numeric	8	Census region
RTCSERVICE	Numeric	8	Residential treatment center
SAP	Numeric	8	Substance use problem
SCHIZOFLG	Numeric	8	Schizophrenia or other psychotic disorder reported
SMISED	Numeric	8	SMI/SED status
SPHSERVICE	Numeric	8	State psychiatric hospital services
STATEFIP	Numeric	8	Reporting state code
SUB	Numeric	8	Substance use diagnosis
TRAUSTREFLG	Numeric	8	Trauma- and stressor-related disorder reported
VETERAN	Numeric	8	Veteran status
YEAR	Numeric	8	Reporting period

Goal of Analysis

- Predict whether a new incoming patient is likely to be classified as Seriously Mentally Ill (SMI) or Seriously Emotionally Disturbed (SED)

SMISED: SMI/SED status

Indicates whether the client has serious mental illness (SMI) or serious emotional disturbance (SED) using the state definition. Use the most recent available status at the end of the reporting period.

Value	Label	Frequency	%
1	SMI	3,195,631	50.2%
2	SED and/or at risk for SED	1,320,906	20.8%
3	Not SMI/SED	1,463,372	23.0%
-9	Missing/unknown/not collected/invalid	382,135	6.0%
	Total	*6,362,044*	*100%*

- SMI and SED status used to determine access to treatment, amount of financial aid, etc.

Other Possible Goals

- Connections between substance use, homelessness, education, and the various types of disorders below (focus on inference, not prediction)

MH1:

Value	Label	Frequency	%
1	Trauma- and stressor-related disorders	876,463	13.8%
2	Anxiety disorders	663,918	10.4%
3	Attention deficit/hyperactivity disorder (ADHD)	415,456	6.5%
4	Conduct disorders	84,862	1.3%
5	Delirium, dementia	16,596	0.3%
6	Bipolar disorders	587,793	9.2%
7	Depressive disorders	1,442,729	22.7%
8	Oppositional defiant disorders	105,453	1.7%
9	Pervasive developmental disorders	57,062	0.9%
10	Personality disorders	47,349	0.7%
11	Schizophrenia or other psychotic disorders	663,037	10.4%
12	Alcohol or substance use disorders	186,071	2.9%
13	Other disorders/conditions	488,153	7.7%
-9	Missing/unknown/not collected/invalid/no or deferred diagnosis	727,102	11.4%
	Total	*6,362,044*	*100%*

Other Possible Goals

- How are different disorders distributed across different regions of the country (or by individual states)?
  - Division:

Value	Label	Frequency	%
0	Other jurisdictions	4,473	0.1%
1	New England	184,364	2.9%
2	Middle Atlantic	1,069,996	16.8%
3	East North Central	1,082,519	17.0%
4	West North Central	426,070	6.7%
5	South Atlantic	821,994	12.9%
6	East South Central	469,034	7.4%
7	West South Central	635,386	10.0%
8	Mountain	521,427	8.2%
9	Pacific	1,146,781	18.0%
	Total	*6,362,044*	*100%*

  - Region:

Value	Label	Frequency	%
0	Other jurisdictions	4,473	0.1%
1	Northeast	1,254,360	19.7%
2	Midwest	1,508,589	23.7%
3	South	1,926,414	30.3%
4	West	1,668,208	26.2%
	Total	*6,362,044*	*100%*

Other Possible Goals

- Look at change in diagnosis based on treatments received

MH1: Mental health diagnosis one

Specifies the client's current first mental health diagnosis during the reporting period.

-9	Missing/unknown/not collected/invalid/no or deferred diagnosis	727,102	11.4%
	Total	*6,362,044*	*100%*

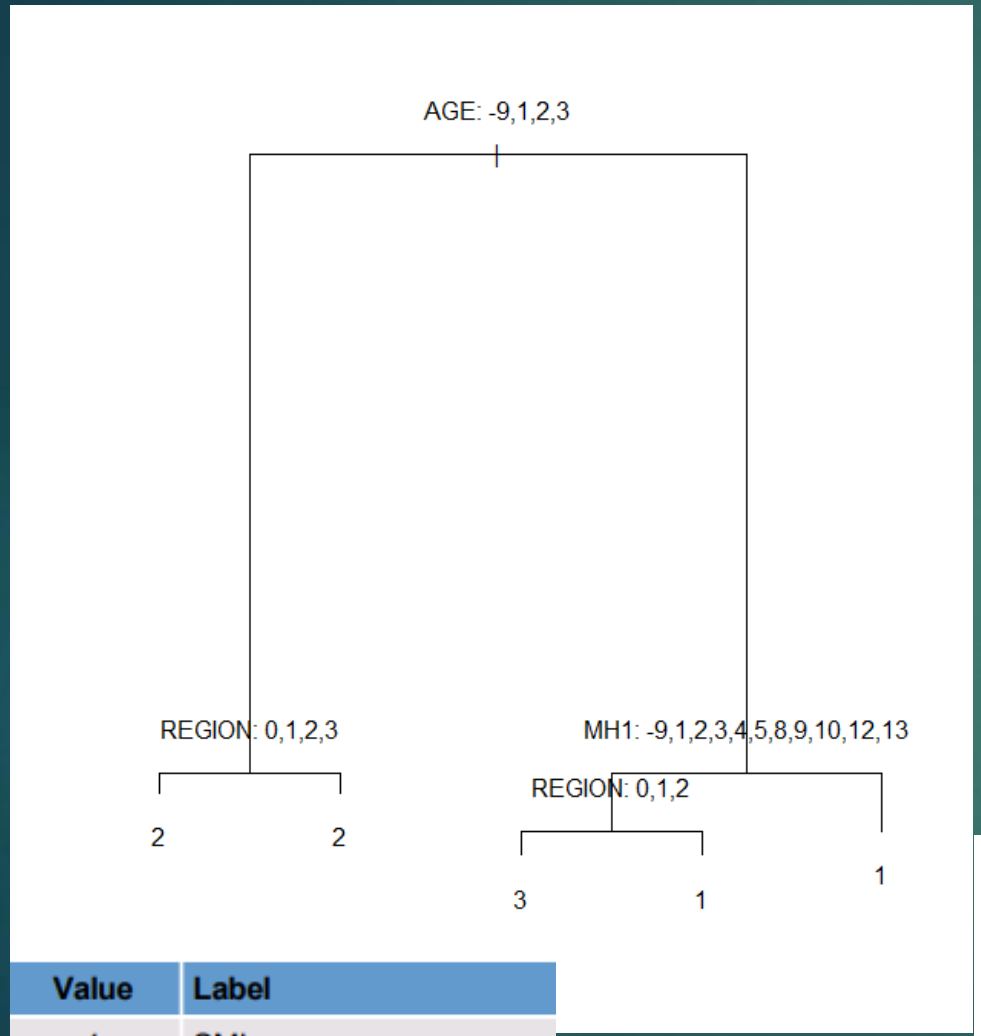
MH2: Mental health diagnosis two

-9	Missing/unknown/not collected/invalid/no or deferred diagnosis	4,404,666	69.2%
	Total	*6,362,044*	*100%*

MH3: Mental health diagnosis three

-9	Missing/unknown/not collected/invalid/no or deferred diagnosis	5,920,406	93.1%
	Total	*6,362,044*	*100%*

Method I: Simple Classification Tree



Value	Label
1	SMI
2	SED and/or at risk for SED
3	Not SMI/SED

Age:

Value	Label
1	0–11 years
2	12–14 years
3	15–17 years
4	18–20 years
5	21–24 years
6	25–29 years
7	30–34 years
8	35–39 years
9	40–44 years
10	45–49 years
11	50–54 years
12	55–59 years
13	60–64 years
14	65 years and older
-9	Missing/unknown/not collected/invalid

MH1:

Value	Label
1	Trauma- and stressor-related disorders
2	Anxiety disorders
3	Attention deficit/hyperactivity disorder (ADHD)
4	Conduct disorders
5	Delirium, dementia
6	Bipolar disorders
7	Depressive disorders
8	Oppositional defiant disorders
9	Pervasive developmental disorders
10	Personality disorders
11	Schizophrenia or other psychotic disorders
12	Alcohol or substance use disorders
13	Other disorders/conditions
-9	Missing/unknown/not collected/invalid/no or deferred diagnosis

Region:

Value	Label
0	Other jurisdictions
1	Northeast
2	Midwest
3	South
4	West

Unexpectedly small tree!

Method I: Classification Tree

```
Classification tree:
tree(formula = SMISED ~ ., data = train)
Variables actually used in tree construction:
[1] "AGE"    "REGION" "MH1"
Number of terminal nodes:  5
Residual mean deviance:  0.965 = 4040000 / 4187000
Misclassification error rate: 0.22 = 920936 / 4186691
```

- Training Accuracy: 78%
- Only three variables used for classification

```
> table (tree.pred,test$SMISED)

tree.pred      1      2      3
        1 842717      0 157175
        2     14 395979 122617
        3 115288      0 159428
> 100* (842717 + 395979 + 159428)/nrow(test)
[1] 77.96732
```

- Testing Accuracy: 77.96%
- Lots of non-seriously ill people classified as SMI or SED

Value	Label
1	SMI
2	SED and/or at risk for SED
3	Not SMI/SED

Ensemble Methods

- Random Forest:

```
> #Random Forest
> library(randomForest)
randomForest 4.7-1.1
Type rfNews() to see new features/changes/bug fixes.
> forest.SMISED = randomForest(SMISED ~ ., data = train, mtry = 4, importance = TRUE)
Error: cannot allocate vector of size 31.2 Gb
```

Boosting: Tried with Stumps. Ran for 30 mins, R terminated because multiclass classification with boosting is currently not fully functioning for some reason

```
> #Boosting
> library(gbm)
Loaded gbm 2.1.8.1
> boost.SMISED = gbm(SMISED ~., data=train, distribution="multinomial",
+                    n.trees=500,
+                    interaction.depth=1)
Warning message:
Setting `distribution = "multinomial"` is ill-advised as it is currently broken. It exists only for backwards compatibility. Use at your own risk.
```

Method II: Random Forest

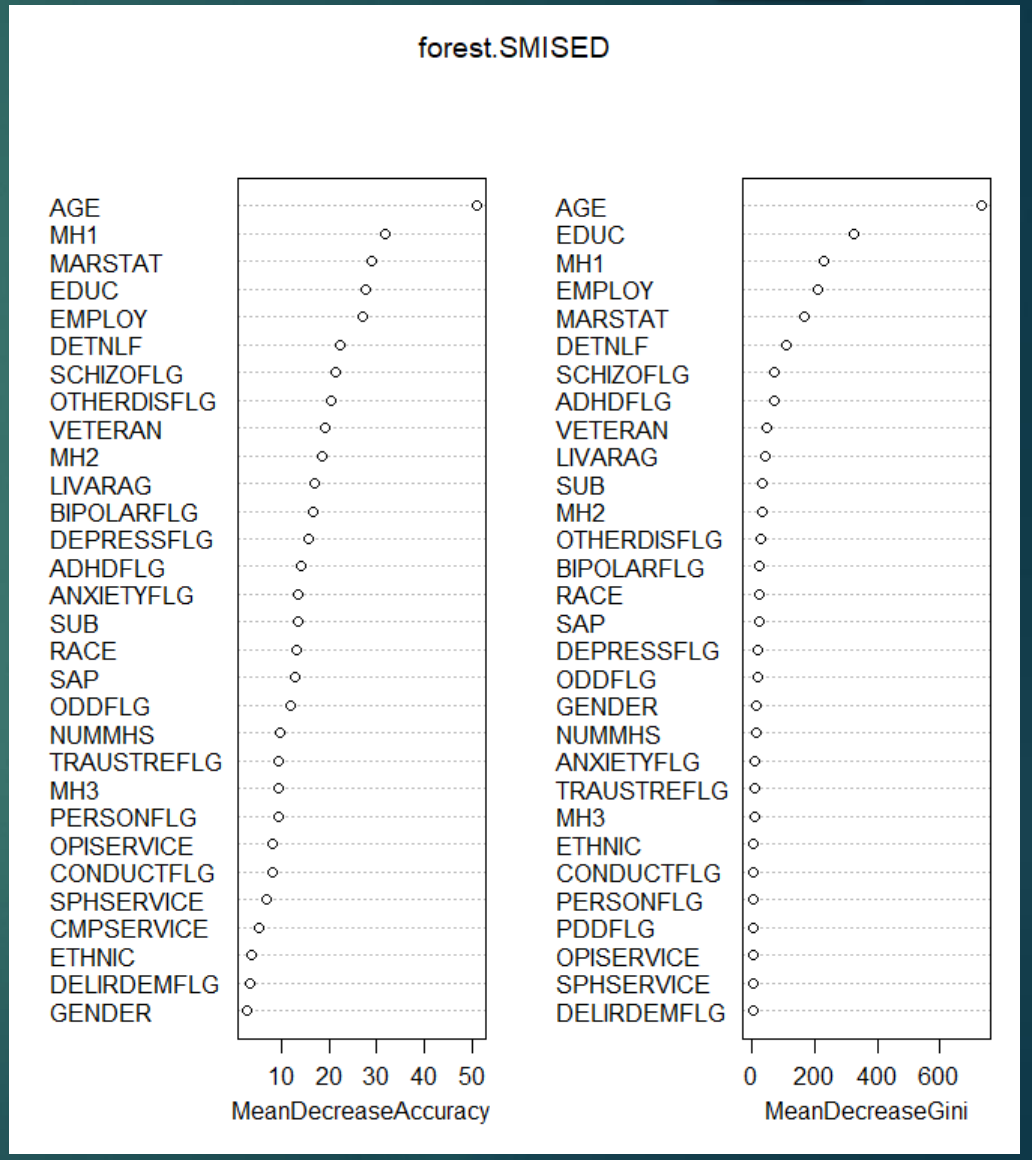
- 5000 data points in training set, 5000 in test set
- 500 trees, 4 variables in each split
- Training accuracy ~ 95.32%
- Out-Of-Bag accuracy ~ 92.94%
- Testing Accuracy ~ 93.32%

```
call:
```

DETNLF: Detailed 'not in labor force' category

This field gives more detailed information about those clients who are coded as 'not in labor force' for employment status (EMPLOY).

Value	Label	Frequency	%
1	Retired, disabled	514,409	8.1%
2	Student	190,398	3.0%
3	Homemaker	41,743	0.7%
4	Sheltered/non-competitive employment	12,301	0.2%
5	Other	697,609	11.0%
-9	Missing/unknown/not collected/invalid	4,905,584	77.1%
	Total	*6,362,044*	*100%*



Method III: SVM

- Polynomial Kernel, with soft SVM (Slack allowed)
- Two hyper-parameters to tune:
  - Degree of polynomial
  - Cost (Slack)

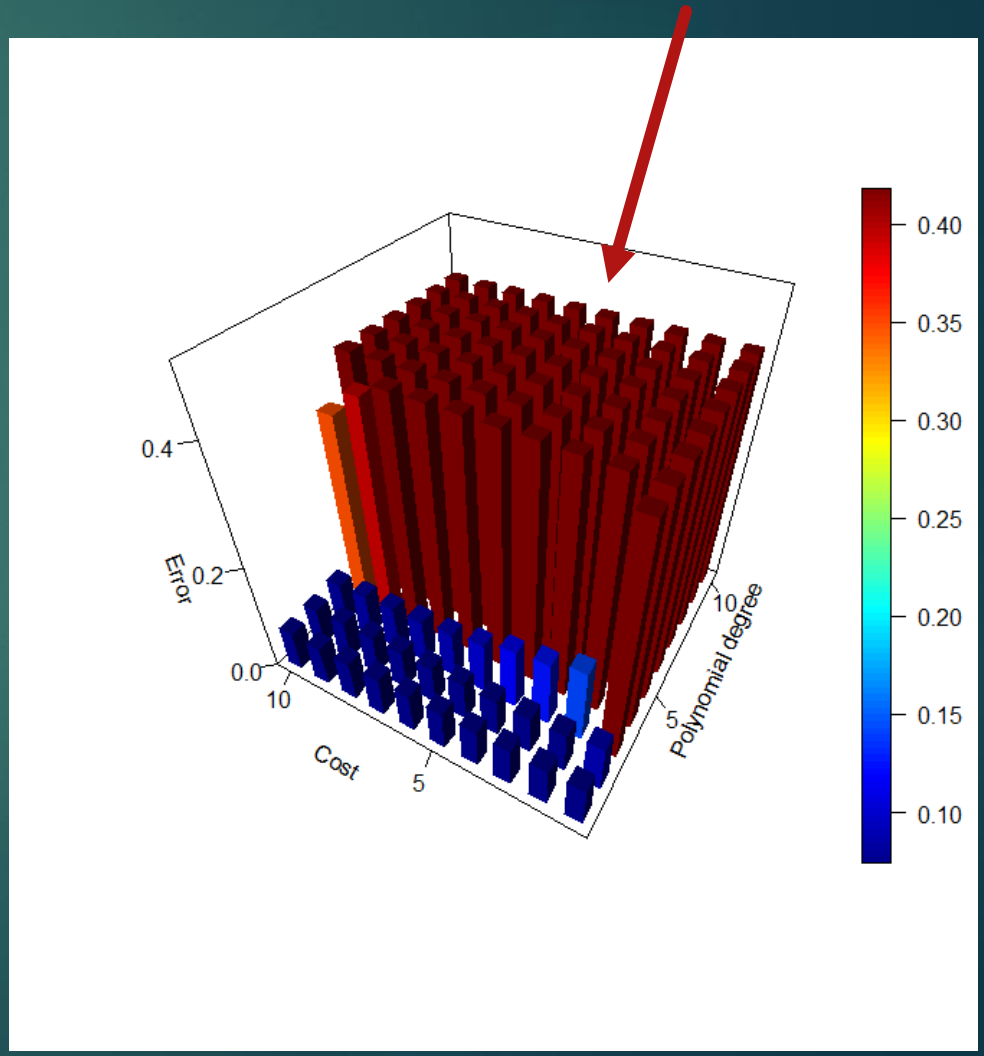
Cross-Validation!!

```
#Cross validation on degree of polynomial and cost
cv.error= matrix(nrow = 10, ncol = 10)
for (i in 1:10) {
  for (j in 1:10) {
    svm.SMISED = svm(SMISED ~., data = small_train, kernel = "polynomial", degree = i, cost = j)
    svm.pred = as.factor(predict (svm.SMISED, small_test))
    cv.error[i,j] = 1 - sum(svm.pred == small_test$SMISED)/length(svm.pred)
  }

}
}
```

```
> cv.error
       [,1]   [,2]   [,3]   [,4]   [,5]   [,6]   [,7]   [,8]   [,9]  [,10]
 [1,] 0.0754 0.0744 0.0740 0.0740 0.0740 0.0740 0.0740 0.0740 0.0740 0.0740
 [2,] 0.0884 0.0766 0.0752 0.0748 0.0746 0.0746 0.0746 0.0746 0.0744 0.0740
 [3,] 0.4182 0.1428 0.1252 0.1162 0.0988 0.0906 0.0872 0.0840 0.0826 0.0796
 [4,] 0.4182 0.4182 0.4182 0.4182 0.4182 0.4182 0.4182 0.4182 0.3950 0.3512
 [5,] 0.4182 0.4182 0.4182 0.4182 0.4182 0.4182 0.4182 0.4182 0.4182 0.4182
 [6,] 0.4182 0.4182 0.4182 0.4182 0.4182 0.4182 0.4182 0.4182 0.4182 0.4182
 [7,] 0.4182 0.4182 0.4182 0.4182 0.4182 0.4182 0.4182 0.4182 0.4182 0.4182
 [8,] 0.4182 0.4182 0.4182 0.4182 0.4182 0.4182 0.4182 0.4182 0.4182 0.4182
 [9,] 0.4182 0.4182 0.4182 0.4182 0.4182 0.4182 0.4182 0.4182 0.4182 0.4182
[10,] 0.4182 0.4182 0.4182 0.4182 0.4182 0.4182 0.4182 0.4182 0.4182 0.4182
>
```

Overfitting!



Method III: SVM

- Tried training on full training set (4 million data points), ran for an hour, then gave up

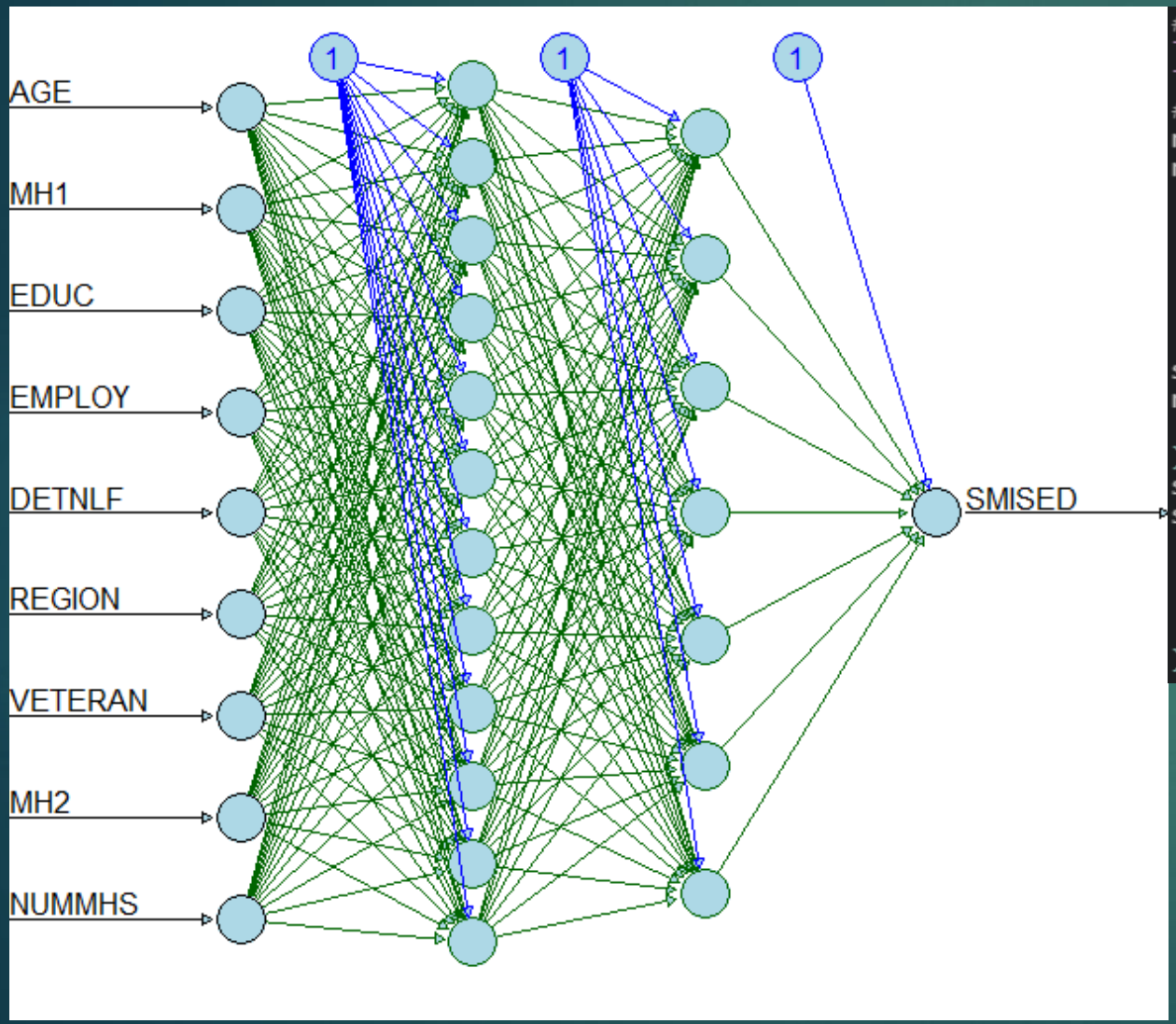
- Trained on 5000 data points: tested on 1.8 million
- Training Accuracy: 92.54%
- Testing Accuracy : 72.37%
- Predicts everyone to have SMI/be SED – zero in third category

	truth		
predict	1	2	3
1	958317	56129	324374
2	0	339363	114918
3	0	0	0

- Trained on 100,000 data points, tested on 30,000:
- Training Accuracy: 83.7%
- Testing Accuracy: 91.97%
- A few people in third category

	truth		
predict	1	2	3
1	17094	0	1861
2	0	10165	435
3	113	0	332

Method IV: Deep Learning *Attempt**




```
#Neural Network
library(keras)
library(tensorflow)
#install_tensorflow(package_url = "https://pypi.python.org/packages/b8/d6/a
neural.net.SMISED = keras_model_sequential()
neural.net.SMISED %>%
  layer_dense (units = 12, activation = "relu",
               input_shape = c (9)) %>%
  layer_dropout (rate = 0.1) %>%
  layer_dense (units = 7, activation = "relu") %>%
  layer_dropout (rate = 0.1) %>%
  layer_dense (units = 3, activation = "softmax")
summary(neural.net.SMISED)
neural.net.SMISED %>% compile(loss = "categorical_crossentropy",
                  optimizer = optimizer_rmsprop(), metrics = "accuracy"
)
small_train %<>% mutate_if(is.factor, as.numeric)
system.time(
  history <- neural.net.SMISED %>%
    #     fit(x_train, y_train, epochs = 30, batch_size = 128,
    fit(small_train, small_train$SMISED, epochs = 15, batch_size = 128,
        validation_split = 0.2)
)
```

- 9 features (picked from random forest ranking)
- 12 neurons in 1st hidden layer, 7 in 2nd
- 609 parameters to learn
- ReLU activation, softmax to convert to probabilities
- One-Hot Encoding for 3 output neurons (diagram slightly wrong)

Co

- Age, r… …nt in determ…
- Ensem…
  - Ran…
  - Onl…
- Comp…