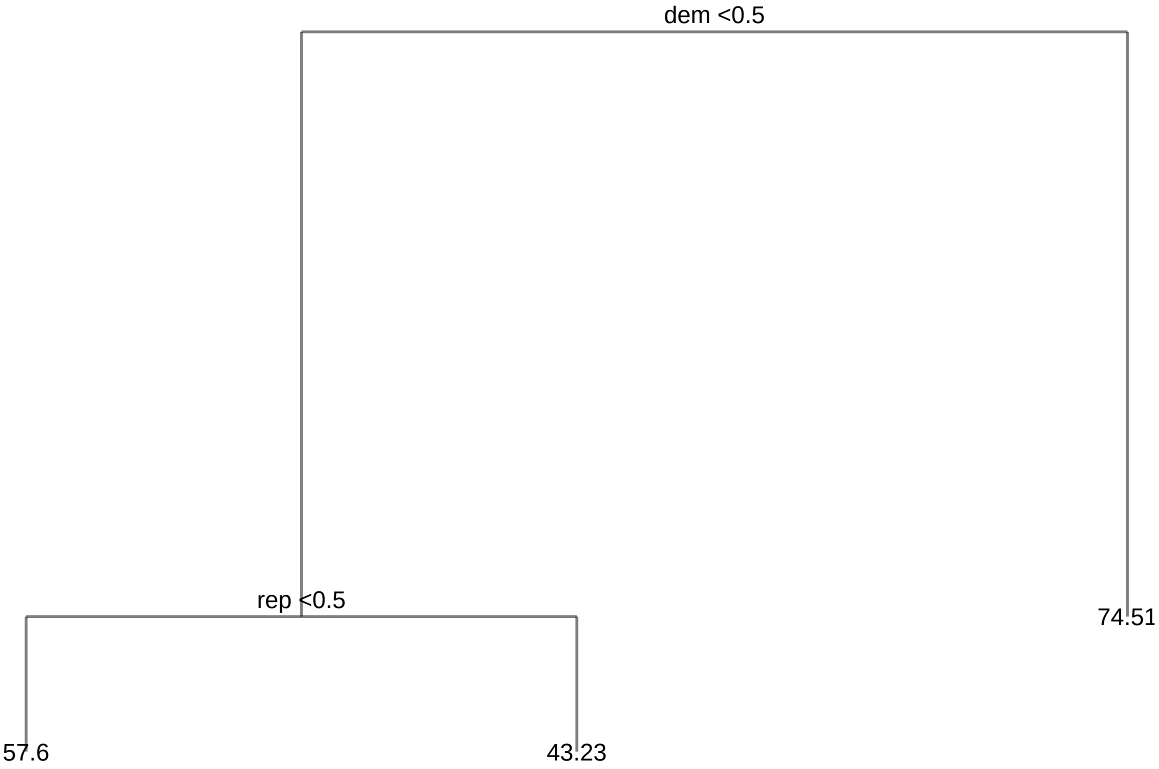
Problem Set 8

MACS 30100 - Perspectives on Computational Modeling Luxi Han 10449918

Problem 1

1.

2.

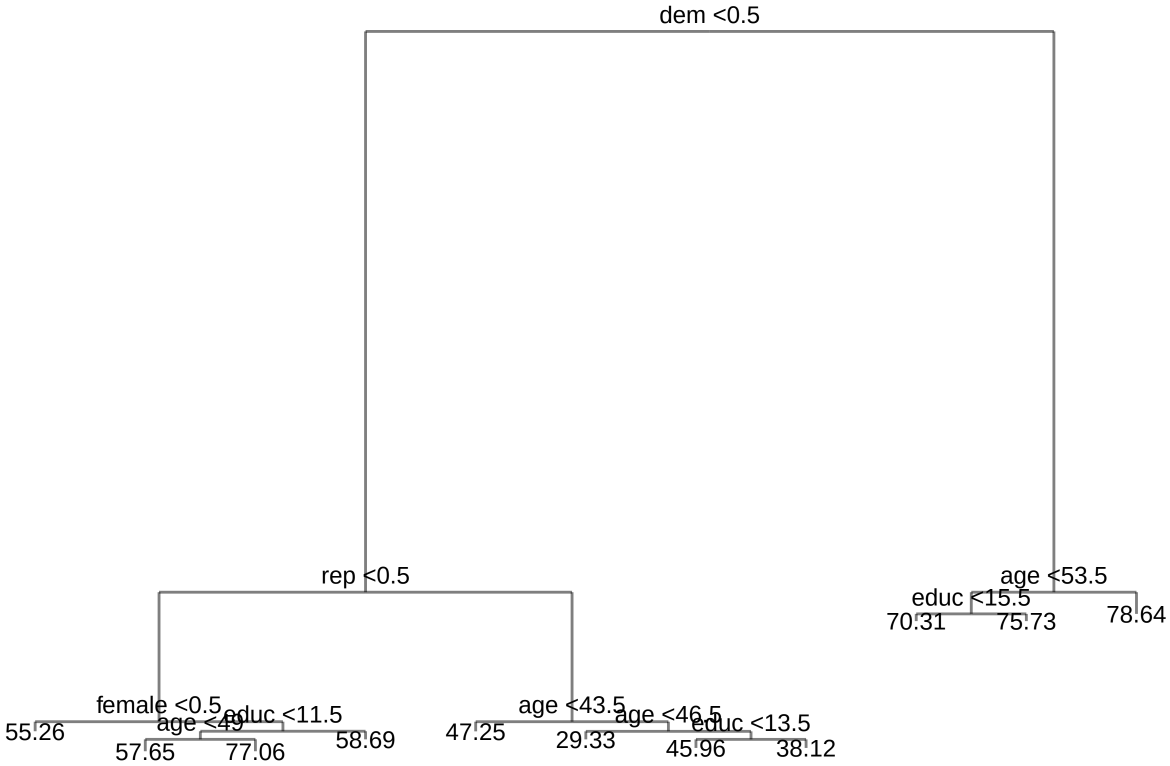


```
## [1] "The MSE for tree method is:"
```

```
## [1] 406.4167
```

We have the tree having two nodes. When the interviwed person is democrat, then their Biden warmth is about 74.51. If this person is republican then her Biden warmth is about 43.23. If he or she is independent, then their Biden warmth is estimated to be 57.6.

3.



```
## [1] "Best complexity is:"

## [1] 11

## [1] "Best MSE is:"

## [1] 401.0746
```

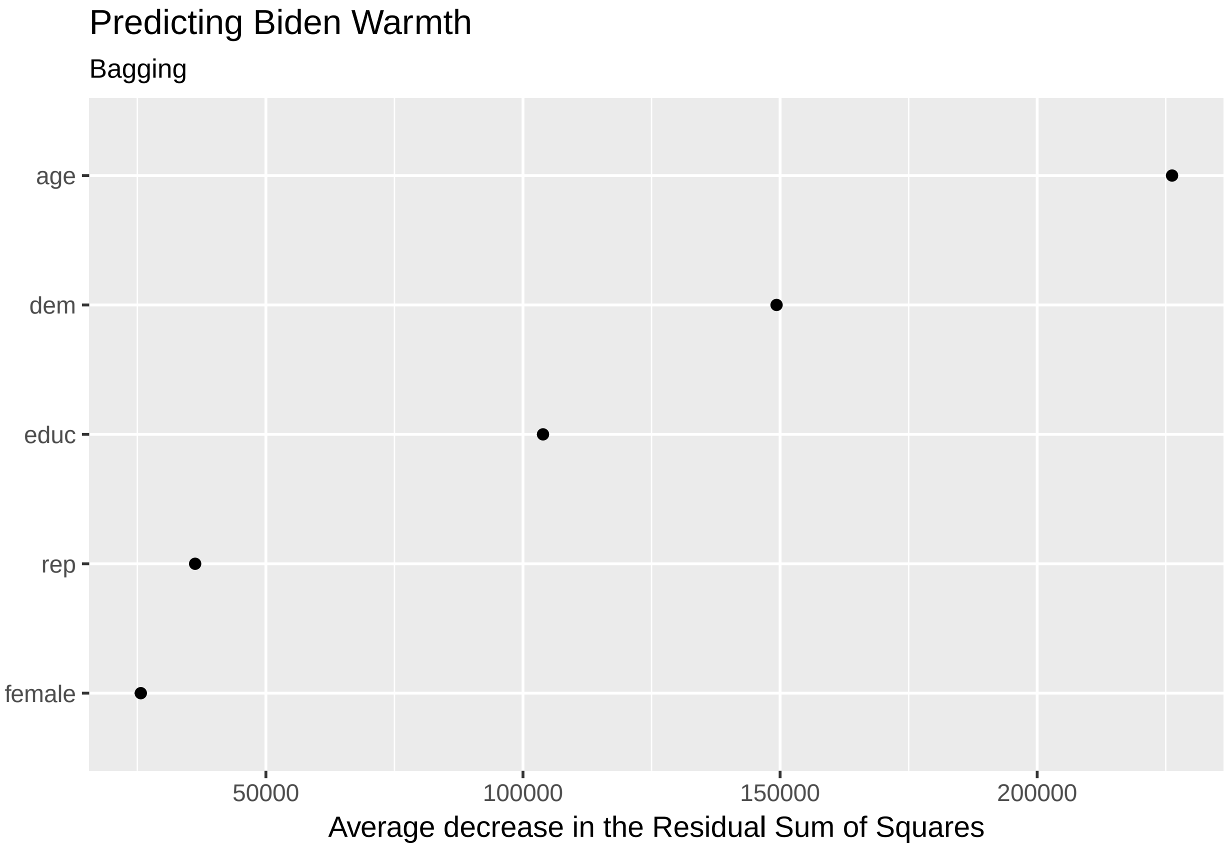
The best model is a tree that has 11 terminal nodes. We first see whether one is democrat or not. If yes, then we will look at their age with a threshold of 53.5. If age is higher than 53.5, then we can have a prediction value of 78.64. If not then, we will look at their education level. If education year is less than 15.5, wewill have a prediction value of 70.31; if higher then we will have a prediction of about 75.73.

If the person is not democrat, we will then look at whether he or she is republican or not. If yes, then we will look at their age. We use a threshold of 43.5 years old. If the age is higher than this value, the nwe will further look at whtether their age is below or higher than 46.5. If yes ,then we will look at their education to give a predictino.

For independent people, gender and education matters for their prediciton value.

In general ,we can see than democrat has the highest Biden warmth while republicans have the lowest. Education affects demovrats differently. Democrats with higher education level

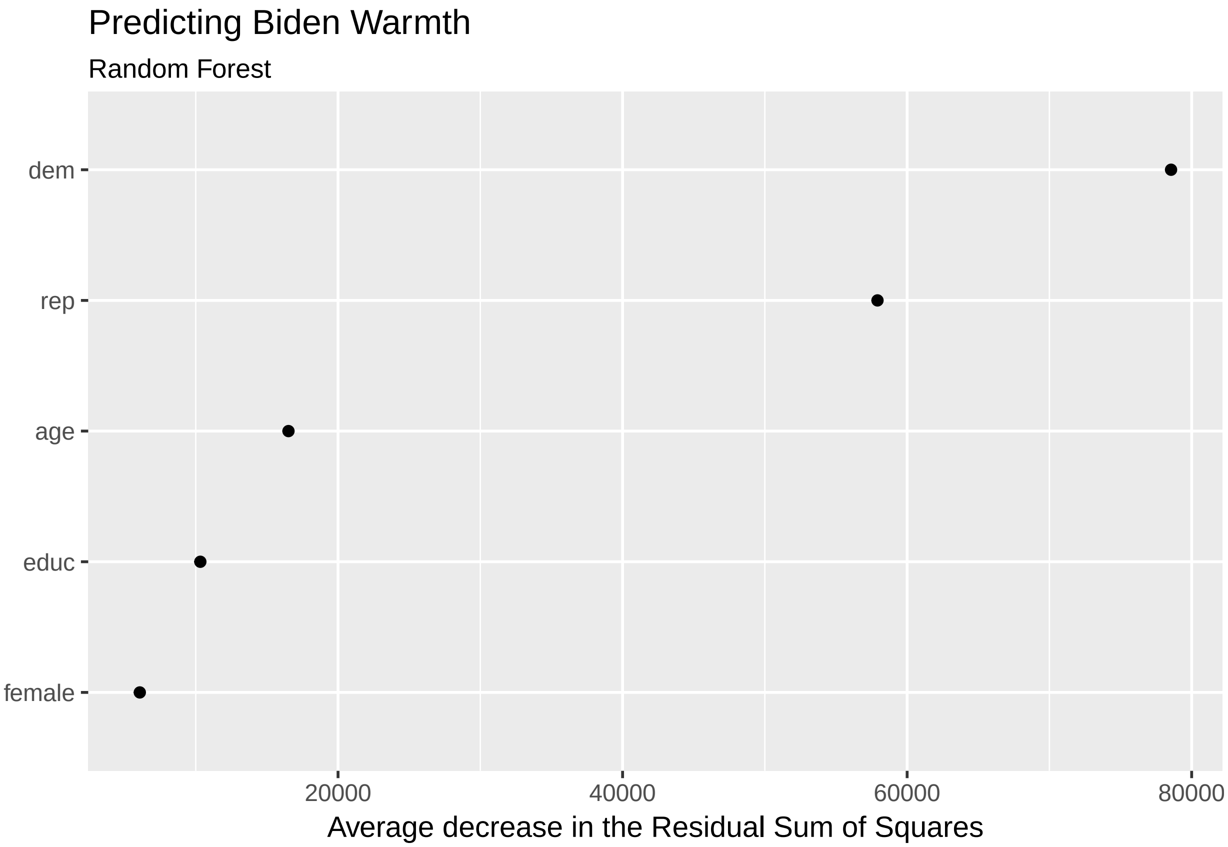
4.



```
## [1] 482.4561
```

We conduct bagging approach. We use residual sum of square as our importance measure. Then we can see that education and wage are the two most important varaibles for the model. The MSE for the bagging method is about 482.45.

5.



```
## [1] 410.3338
```

Using random forest method, we can see that whether a person is democrat and whether the person is a republican are the two most important factors. Adding these two variables decrease the residual sum of squares the most. The MSE for the random foest method is 410.33. This is better than the bagging method.

6.

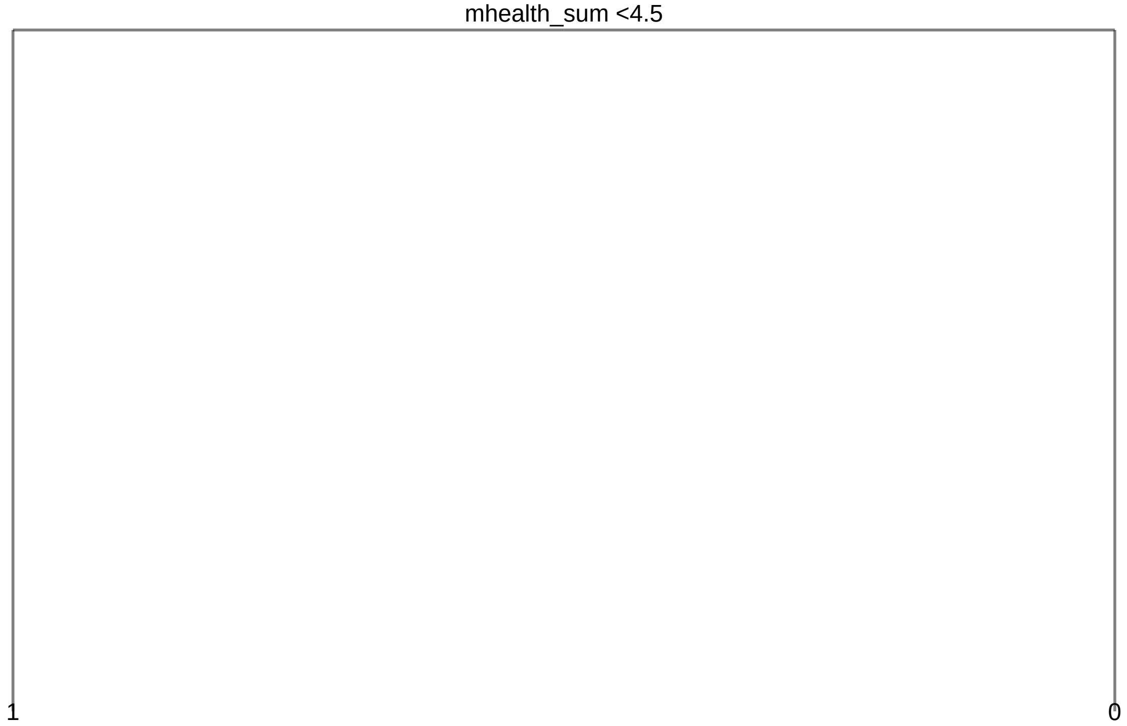
```
## Distribution not specified, assuming gaussian ...
```

```
## [1] 694.7527
```

Problem 2

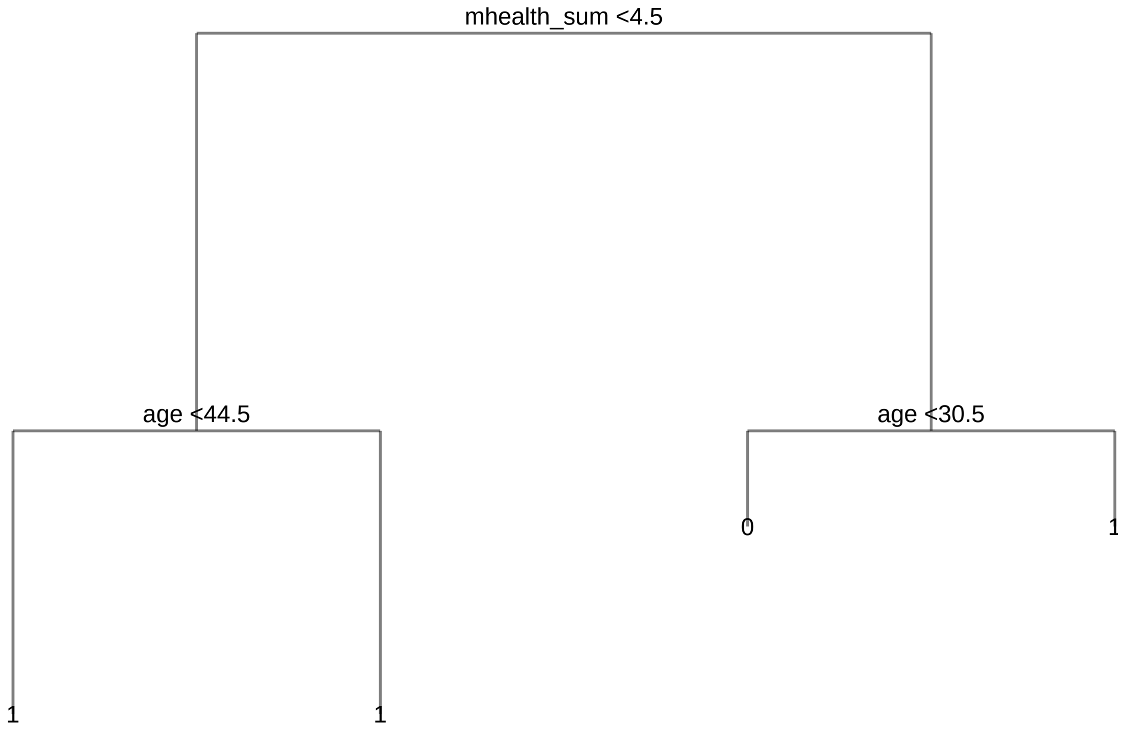
1.

Model 1



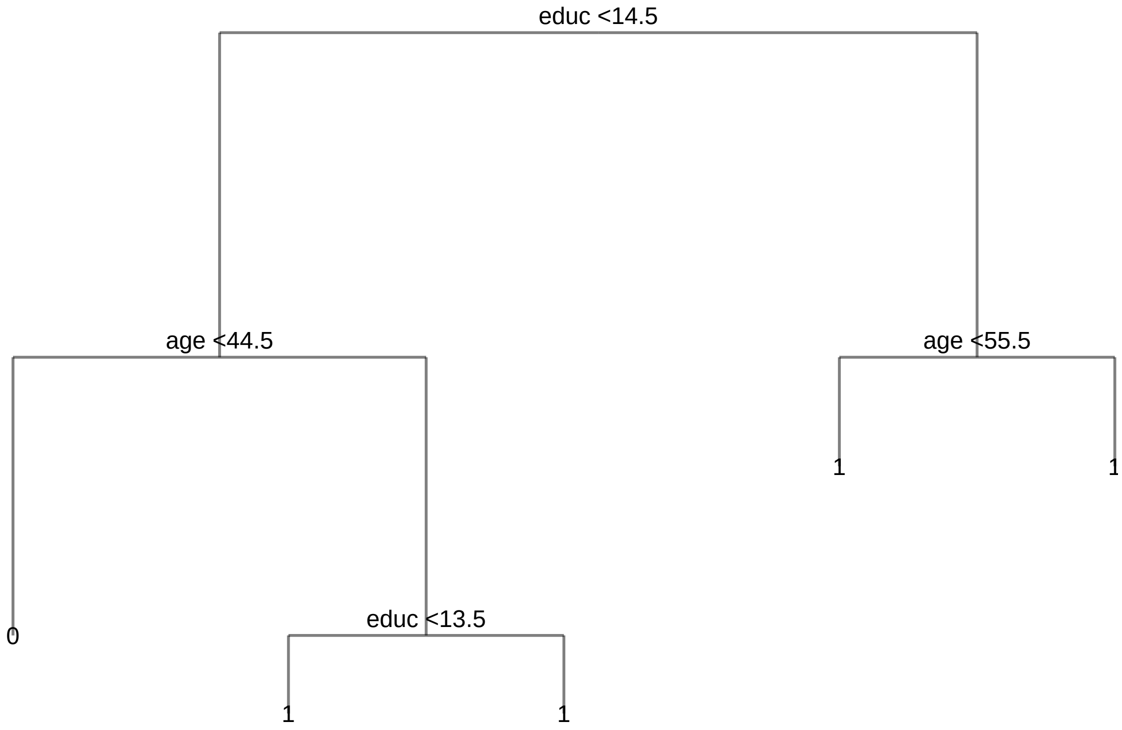
```
## [1] "The MSE for model 2 is:"

## [1] 0.3352436
```

Model 2



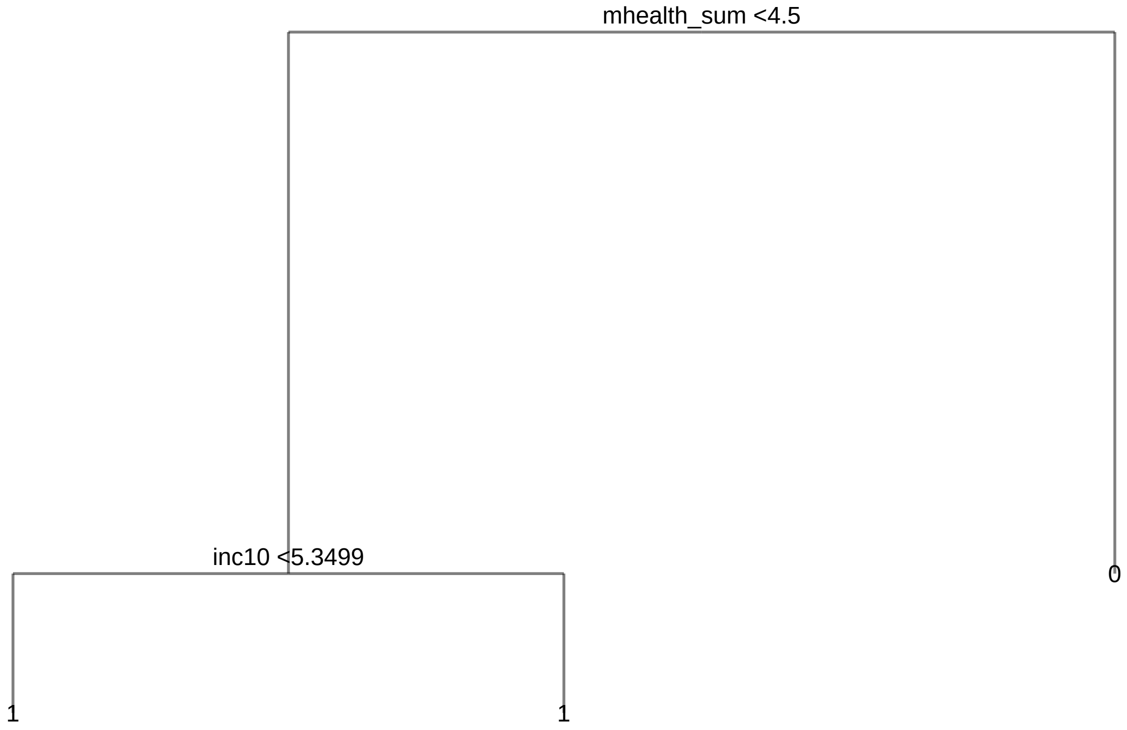
```
## [1] "The MSE for model 2 is:"

## [1] 0.3037249
```

Model 2



```
## [1] "The MSE for model 2 is:"
## [1] 0.3065903
```

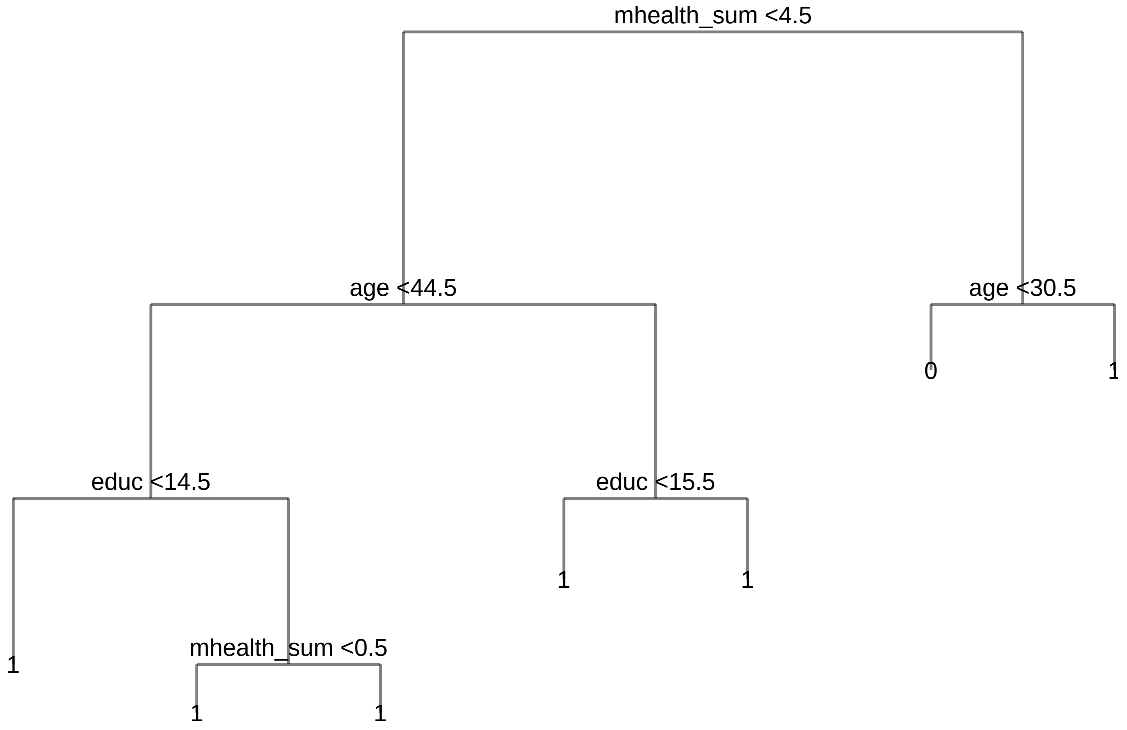
Model 2



```
## [1] "The MSE for model 2 is:"
```

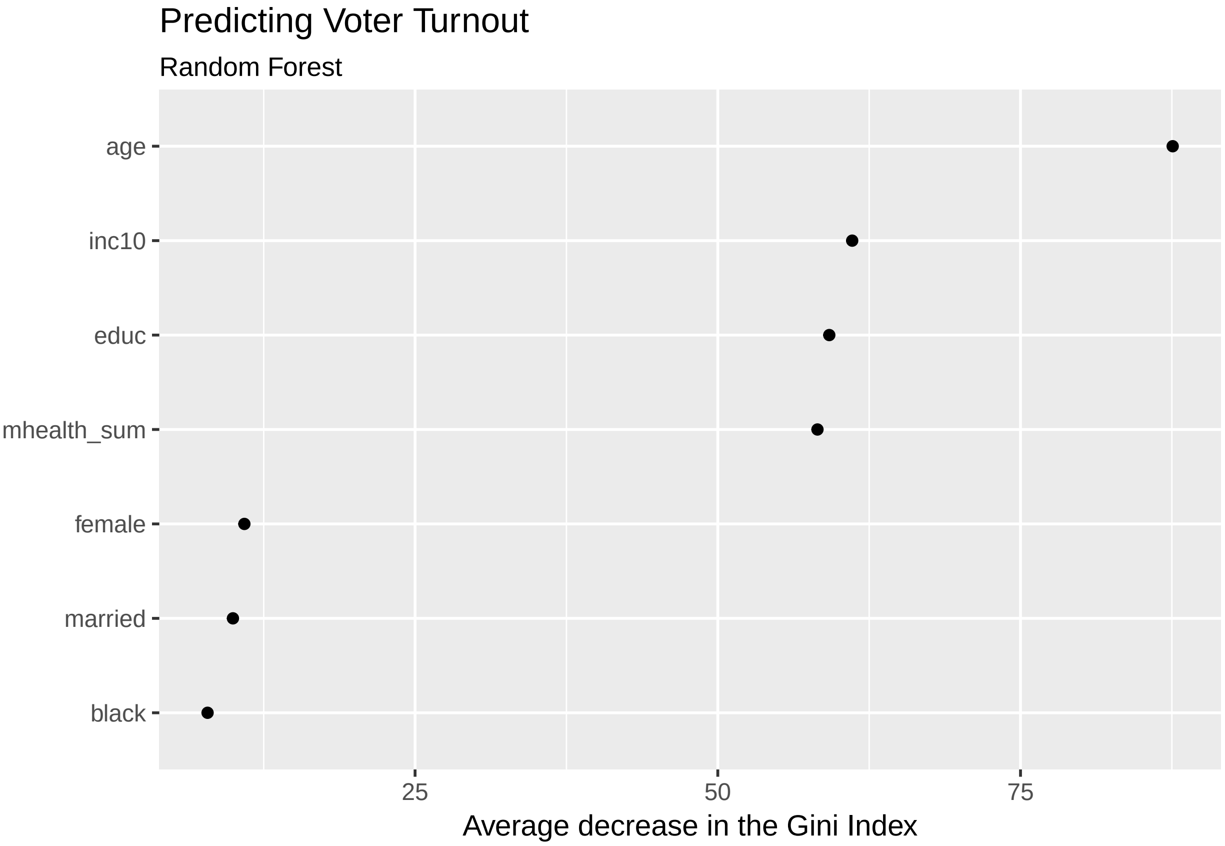
```
## [1] 0.3352436
```

Model 5



```
## [1] "The MSE for model 2 is:"

## [1] 0.3037249
```

The above are the models that I estimated. As we can see, there is little variation in the predict power of the models. The full model does perform better than most of the models. But the model with only age and educatoin performs even better than the full model. We plot the decrease in Gini index using random forest method. This is somewhat contrary to our cross validation method. The random forest method does confirm the importance age but it also indicates that the health index is an important factor.



2.

Model 1 Linear Kernel 1

```
##
## Call:
## best.tune(method = svm, train.x = vote96 ~ mhealth_sum + age +
##     educ, data = as_tibble(voter_split$train), ranges = list(cost = c(0.001,
##     0.01, 0.1, 1, 5, 10, 100)), kernel = "linear")
##
##
## Parameters:
##    SVM-Type:  C-classification
##  SVM-Kernel:  linear
##        cost:  5
##       gamma:  0.3333333
##
```

```
## Number of Support Vectors:  510
##
##  ( 255 255 )
##
##
## Number of Classes:  2
##
## Levels:
##  0 1
```

```
## [1] "The AUC value for model 1 is:"
```

```
## Area under the curve: 0.7365
```

I use education, mental health index and age for predictors. We use a linear kernel to fit the model. We can see that we have a large cost parameter. This means that the model allows the support vectors to lie on both the wrong side of the hyperplane and the margin. This may indicate that a proportion of the observatoin is not predicted well.

Model 2 Linear Kernel 2

```
##
## Call:
## best.tune(method = svm, train.x = vote96 ~ mhealth_sum + married +
##     inc10, data = as_tibble(voter_split$train), ranges = list(cost = c(0.001,
##     0.01, 0.1, 1, 5, 10, 100)), kernel = "linear")
##
##
## Parameters:
##    SVM-Type:  C-classification
##  SVM-Kernel:  linear
##        cost:  0.001
##       gamma:  0.25
##
## Number of Support Vectors:  524
##
##  ( 263 261 )
##
##
## Number of Classes:  2
##
## Levels:
##  0 1
```

```
## [1] "The AUC for model 2 is:"
```

```
## Area under the curve: 0.647
```

In this model we fit a linear kernel on mental health index, marriage stauts and income.

Model 3 Linear Kernel: All

```
##
## Call:
## best.tune(method = svm, train.x = vote96 ~ ., data = as_tibble(voter_split$train),
##     ranges = list(cost = c(0.001, 0.01, 0.1, 1, 5, 10, 100)),
##     kernel = "linear")
##
##
## Parameters:
##    SVM-Type:  C-classification
##  SVM-Kernel:  linear
##        cost:  0.1
##       gamma:  0.125
##
## Number of Support Vectors:  511
##
##  ( 256 255 )
##
##
## Number of Classes:  2
##
## Levels:
##  0 1
```

```
## [1] "The AUC for model 3 is:"
```

```
## Area under the curve: 0.7423
```

Model 4: Polynomial Kernel

```
##
## Call:
## best.tune(method = svm, train.x = vote96 ~ ., data = as_tibble(voter_split$train),
##     ranges = list(cost = c(0.001, 0.01, 0.1, 1, 5, 10, 100)),
##     kernel = "polynomial")
##
##
## Parameters:
##    SVM-Type:  C-classification
##  SVM-Kernel:  polynomial
##        cost:  5
##      degree:  3
##       gamma:  0.125
##      coef.0:  0
##
## Number of Support Vectors:  495
##
##  ( 258 237 )
##
##
## Number of Classes:  2
##
## Levels:
##  0 1
```

```
## [1] "The AUC for model 4 is:"
```

```
## Area under the curve: 0.7413
```

Now we fit a polynomial kernel on all of the variables

Model 5: Radial Kernel

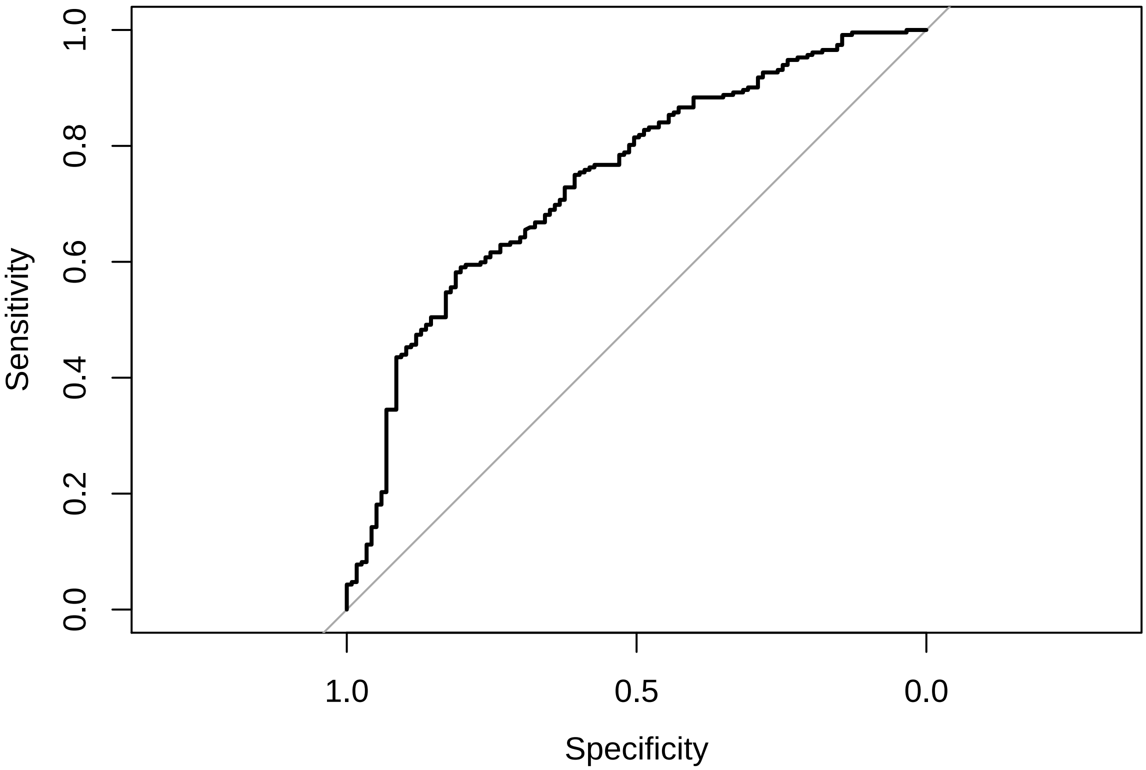
```
##
## Call:
## best.tune(method = svm, train.x = vote96 ~ ., data = as_tibble(voter_split$train),
##     ranges = list(cost = c(0.001, 0.01, 0.1, 1, 5, 10, 100)),
##     kernel = "radial")
##
##
## Parameters:
##    SVM-Type:  C-classification
##  SVM-Kernel:  radial
##        cost:  1
##       gamma:  0.125
##
## Number of Support Vectors:  509
##
##  ( 265 244 )
##
##
## Number of Classes:  2
##
## Levels:
##  0 1
```

```
## [1] "The AUC for model 5 is:"
```

```
## Area under the curve: 0.7349
```

To sum up, we can see that the linear kernel using all of the varaibles generates the most ideal result. On one hand, the area under the curve is the largest for the linear kernel, on the other hand, the cost parameter for lienar kernel is the smallest and thus the model can be imposed under a more restircted constraint. This may indicate a better prediction accuracy.

The following is the ROC curve:



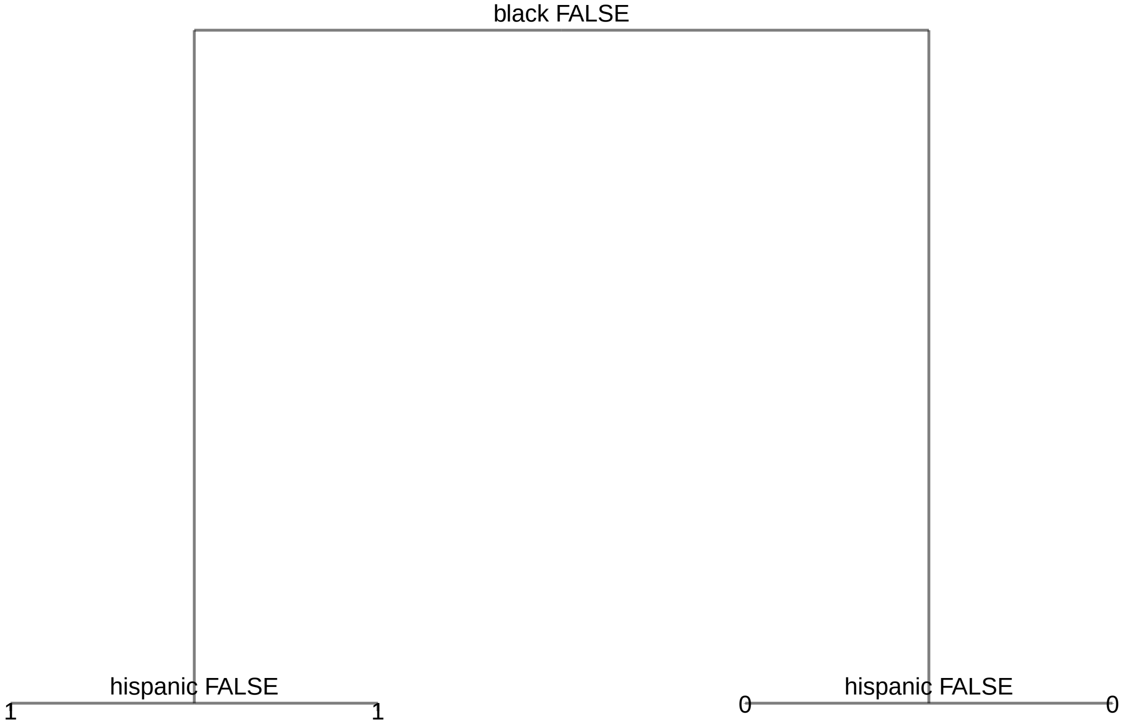
Problem 3

1.

Table 1: Logit Model

	Dependent variable:
	guilt
black	-3.053^{***}
	(0.217)
hispanic	-0.539^{*}
	(0.285)
Constant	1.459^{***}
	(0.094)
Observations	987
Log Likelihood	−478.414
Akaike Inf. Crit.	962.828
Note:	*p<0.1; **p<0.05; ***p<0.01

Tree Method



```
## [1] "The test error rate for Tree Based Method is:"

## [1] 0.1701632
```

I used logit and support vector machine method to study the relationship between their preception of the guilty of OJ simpson. As we can see that For a logit model, compared to non black and non hispanic people, black people tend to think OJ simpson is not guilty. They have a 3.053 points lower log odds ratio. The difference between hispanic and the reference group (all other races) are not clear. Then we can examine the tree based method. The tree based method gives us a clearer classfication. The tree method doesn't classify hispanic as different from other non-black people. As long as the person is not black, then the tree based method will classify him or her as thinking OJ Simpson as guilty.

2.

In this section, we use cross validation method to test the robustness of a logit model and a SVM model Below is how I use logit and SVM method to predict the voter turnout. Again, the logit model confirms our hypothesis that race matters for voter turnout. In this case, education also matters in the sense that people with the lowest degree has the lowest turnout rate. The logit model will generate a AUC of 0.8232.

Compared to the logit model, the SVM model performs not as well. Ihas an AUC of 0.770. The test error rate is the same as the test error rate when we are using the tree based method.

Then we know that the logit model performs better and race variable does give a strong prediction power.

[1] "AUC is:" Area under the curve: 0.7423

```
##
## Call:
## best.tune(method = svm, train.x = guilt ~ dem + rep + educ +
##     age + female + black + hispanic + income, data = oj_fac,
##     ranges = list(cost = c(0.001, 0.01, 0.1, 1, 5, 10, 100)),
```

Table 2: Logit Model

	Dependent variable:
	guilt
dem	0.120
	(0.270)
rep	0.528^{*}
	(0.282)
educHIGH SCHOOL GRAD	−0.339
	(0.230)
educNOT A HIGH SCHOOL GRAD	-1.276^{***}
	(0.352)
educREFUSED	13.674
	(572.422)
educSOME COLLEGE(TRADE OR BUSINESS)	−0.252
	(0.237)
age	0.018^{***}
	(0.005)
female	−0.197
	(0.176)
black	-2.888^{***}
	(0.228)
hispanic	−0.373
	(0.297)
50,000	−0.292
	(0.231)
75,000	−0.018
	(0.306)
75,000	0.375
	(0.374)
incomeREFUSED/NO ANSWER	-1.303^{***}
	(0.371)
15,000	−0.272
	(0.293)
Constant	0.978^{**}
	(0.427)
Observations	987
Log Likelihood	−450.198
Akaike Inf. Crit.	932.397
Note:	*p<0.1; **p<0.05; ***p<0.01

```
##     kernel = "linear")
##
##
## Parameters:
##    SVM-Type:  C-classification
##  SVM-Kernel:  linear
##        cost:  0.01
##       gamma:  0.0625
##
## Number of Support Vectors:  641
##
##  ( 328 313 )
##
##
## Number of Classes:  2
##
## Levels:
##  0 1

## [1] "The test error rate for the SVM is"

## [1] 0.1701632

## [1] "AUC is :"

## Area under the curve: 0.7696
```