1 Logistic Survey
Write down the following information:
Name: Weining Hu
Student number : 45606134
Faculty : Science
Department/Specialization: Computer science combined with statistics
Degree: BSc
Year of study: 3
Registered/auditting/crashing. NA

2.1 Summary Statistics
1. max_age: 71
   min_age: 14

2. median_female: 25
   median_male: 26

3. 10% 25% 50% 75% 90%
    20   23   26   29   33

2.2 Data Visualization(Attached file at the end)

1.
2.
3.

3 Decision Trees and Cross-Validation
3.1
Plot attached at the end.
Accuracy when depth is 10: 1-0.38 = 0.62

3.2
Given that all the features are binary data, we claim that once a stump uses a feature, this feature will no longer appearing in it's children. For the first pass, we iterate through all the the D features, under each of the feature, we iterate through the n examples. Our goal is to find the feature that 'best' split the data. Even thought not every stump will need to go through every example, but the stump on the same level would go through N examples in total. Under worst case, there could be in total of N*D operations at first stump.Then, at the second level of our decision tree, we iterate through the rest of (D-1) features and N examples, the total operations would be N*(D-1).

Follow this fashion, we could conclude that the cost of operations at each level has relationship with the depth.
depth = 0, cost = N*(D-0)
depth = 1, cost = N* (D-1)
depth = 2, cost = N*(D-2)
…….
depth = m, cost = N*(D-m)
To sum up, N[(D) + (D-1) + (D-2) +….+(D-m)] = (2D-M)(M+1)N/2
That is O(NMD)

Attached file at the end

3.3
From cross-validation, we would choose depth 5.

Depth	Average_Error
1	0.1280
2	0.1278
3	0.1278
4	0.1320
5	0.1188
6	0.1258
7	0.1216
8	0.1218
9	0.1236
10	0.1262
11	0.1262
12	0.1262
13	0.1262
14	0.1262
15	0.1262

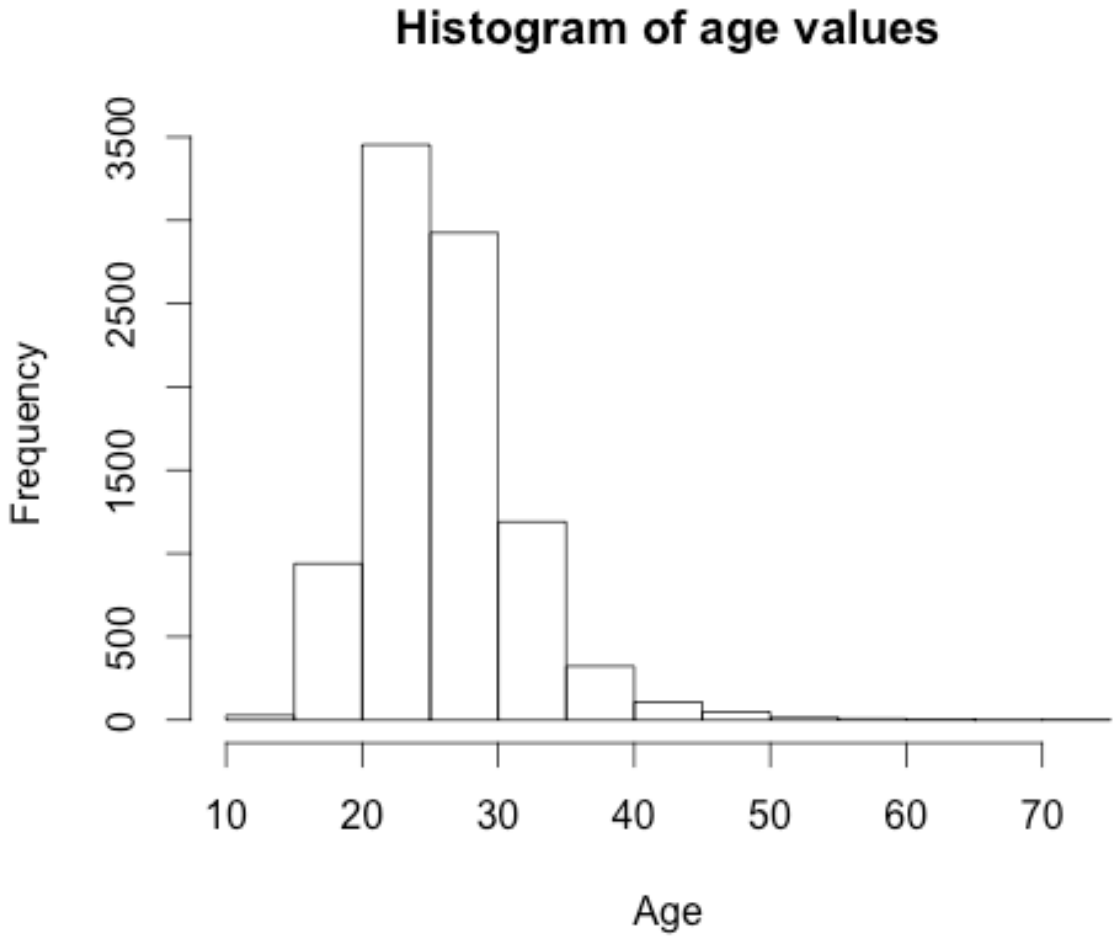
4. Probability excersice
4.1
Apply Bayes Rule:
$P(D=1|T=1) = P(D=1)P(T=1|D=1)/P(T=1)$
$= P(D=1)P(T=1|D=1)/P(T=1|D=1)P(D=1)+P(T=1|D=0)P(D=0)$
$= 0.001*0.99/0.99*0.001+(1-0.99)*(1-0.001)$
$= 0.09$

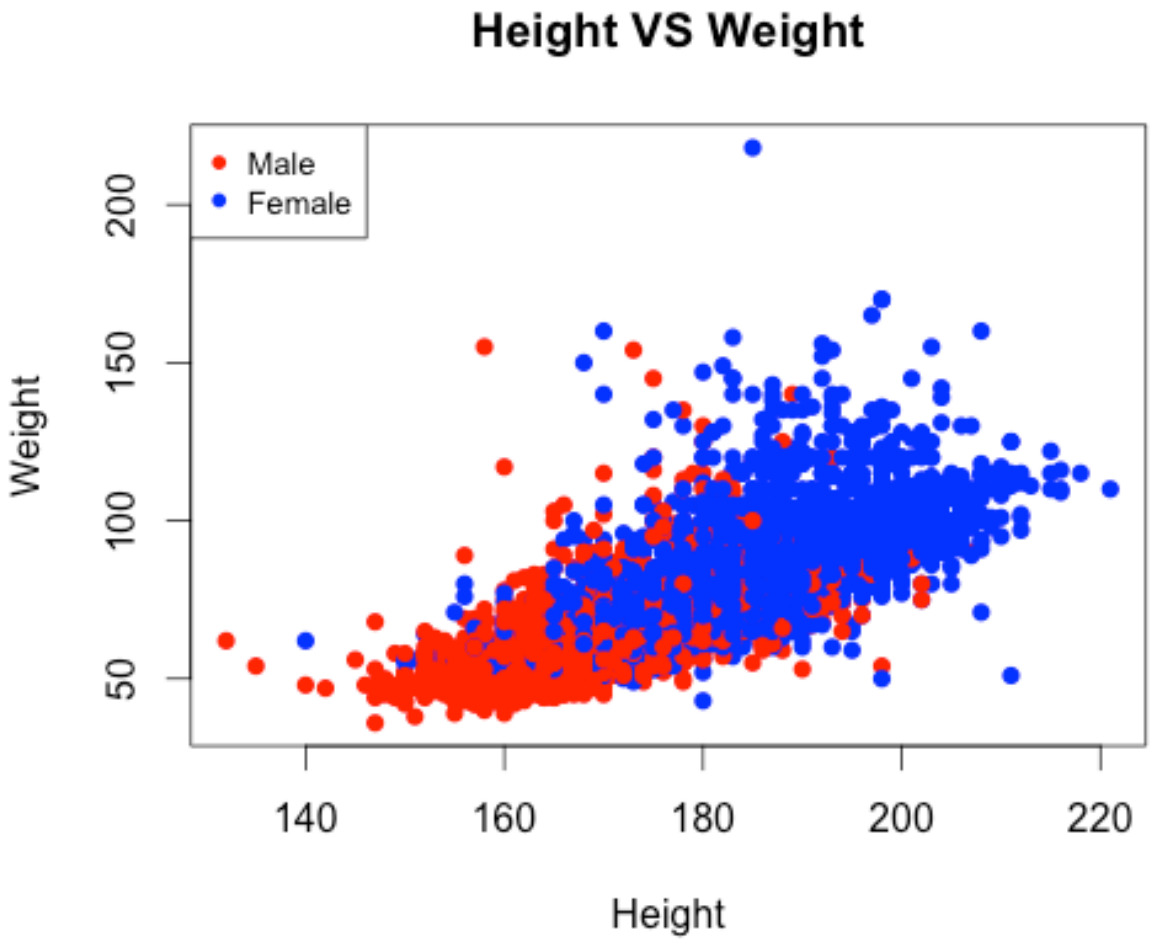
4.2
Two sons
case1: Because the two results have same probability, we could list all the possible results.
1. head, tail
2. head, head
Then the probability is 1/2

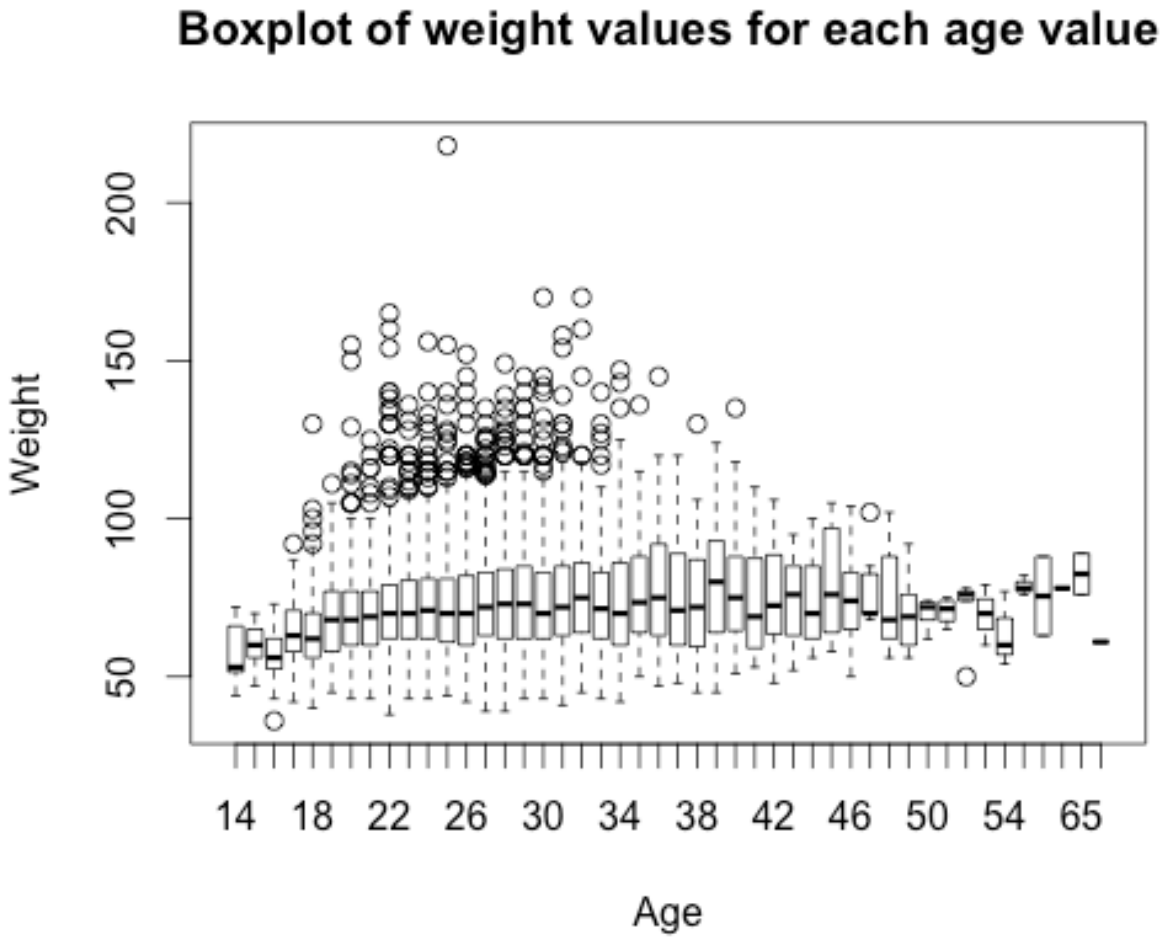
case2: Same as before, list all the possibilities.
1. head tail
2. head head
3. tail head
Then the probability is 1/3

4.3
Prosecutor's fallacy
It is not possible to determine the probability that the person is guilty. We will reason from the info provided in the description.
Let us use G to represent the probability that a random choose person is guilty;
Let us use F to represent the probability that a person out of the innocent people match the footprint of the guilty person;

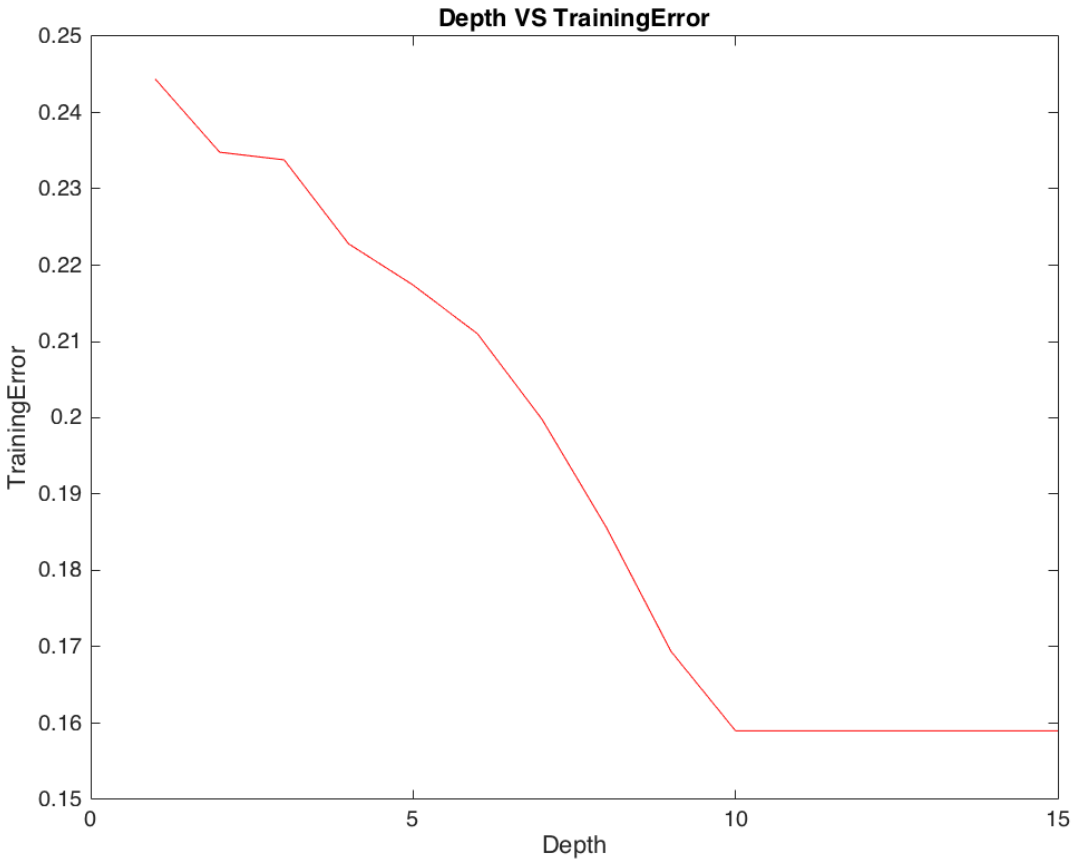
This question ask us to find the probability that given a person's footprint matches with the one at the crime scene, what is the probability that he is guilty?
P(GIF) = P(FIG)*P(G)/P(F)
However, we do not know the probability that a random chosen people is guilty so therefore we could bot determine.

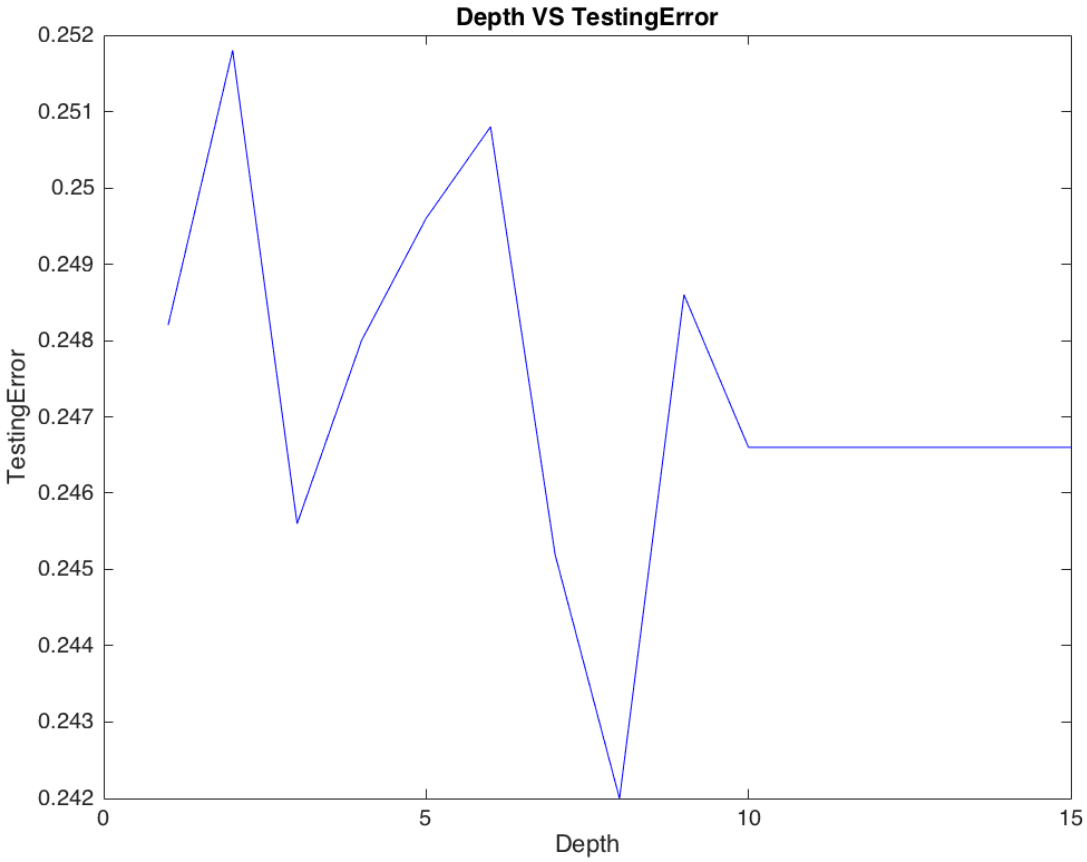

Histogram of age values
Frequency
0
500
1500
2500
3500
10
20
30
40
50
60
70
Age

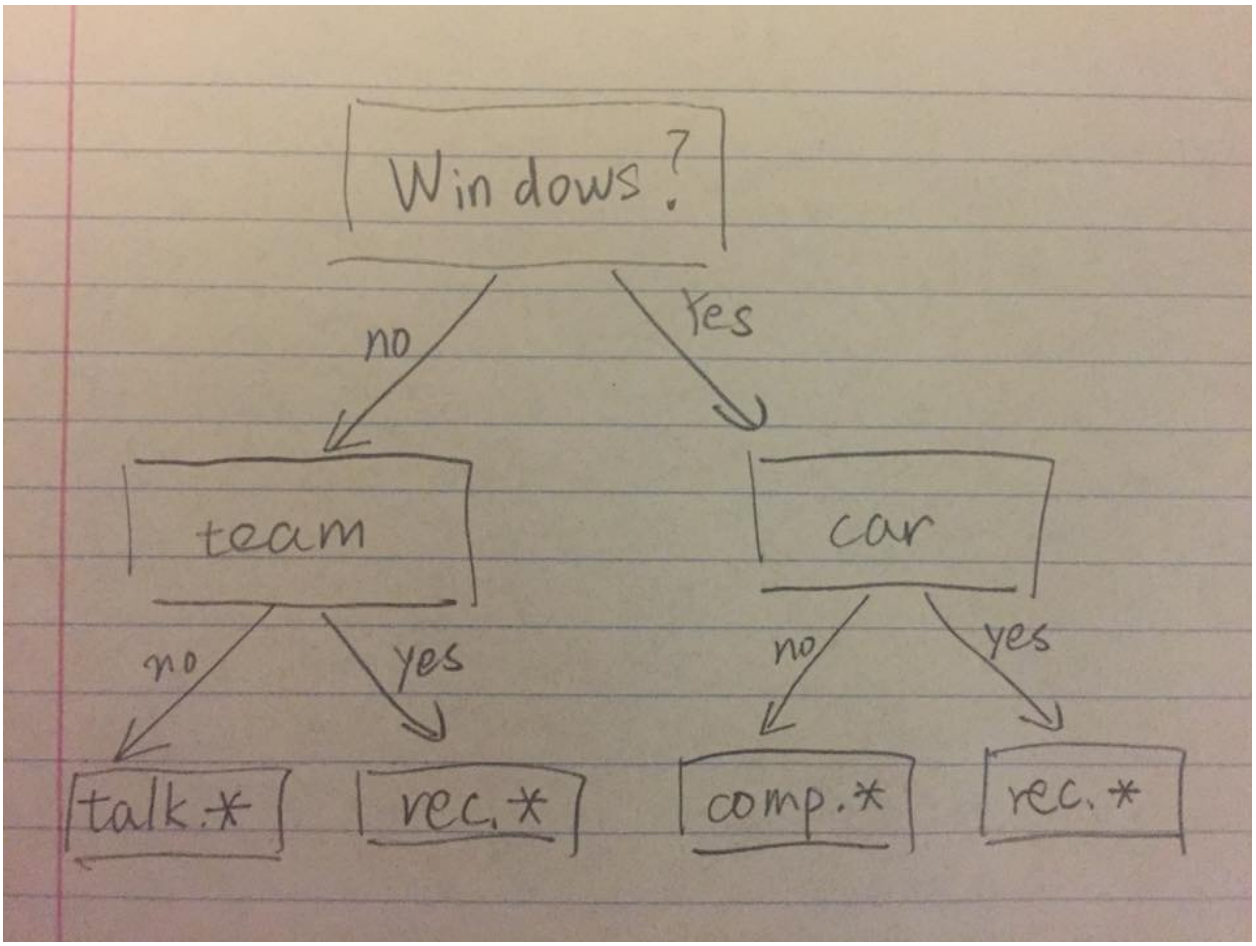

Height VS Weight
Male
Female
Weight
50
100
150
200
140
160
180
200
220
Height


Boxplot of weight values for each age value
Weight
50
100
150
200
14
18
22
26
30
34
38
42
46
50
54
65
Age

3.


Depth VS TrainingError
TrainingError
Depth
0.25
0.24
0.23
0.22
0.21
0.2
0.19
0.18
0.17
0.16
0.15
0
5
10
15


Depth VS TestingError
TestingError
Depth
0.252
0.251
0.25
0.249
0.248
0.247
0.246
0.245
0.244
0.243
0.242
0
5
10
15


Windows?
no
yes
team
car
no
yes
no
yes
talk.*
rec.*
comp.*
rec.*