UNIVERSITY OF LEEDS

School of Mathematics

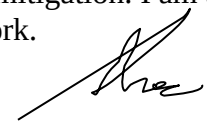
Declaration of Academic Integrity
for Individual Pieces of Work

I am aware that the University defines plagiarism as presenting someone else's work as your own. Work means any intellectual output, and typically includes text, data, images, sound or performance.

I promise that in the attached submission I have not presented anyone else's work as my own and I have not colluded with others in the preparation of this work. Where I have taken advantage of the work of others, I have given full acknowledgement. I have read and understood the University's published rules on plagiarism and also any more detailed rules specified at School or module level. I know that if I commit plagiarism I can be expelled from the University and that it is my responsibility to be aware of the University's regulations on plagiarism and their importance.

I re-confirm my consent to the University copying and distributing any or all of my work in any form and using third parties (who may be based outside the EU/EEA) to monitor breaches of regulations, to verify whether my work contains plagiarised material, and for quality assurance purposes.

I confirm that I have declared all mitigating circumstances that may be relevant to the assessment of this piece of work and that I wish to have taken into account. I am aware of the School's policy on mitigation and procedures for the submission of statements and evidence of mitigation. I am aware of the penalties imposed for the late submission of coursework.



Student Signature _____ Date __**6 May 2020**__

Student Name _____**Nico Septianus**_____ Student Number ____**201380903**_____

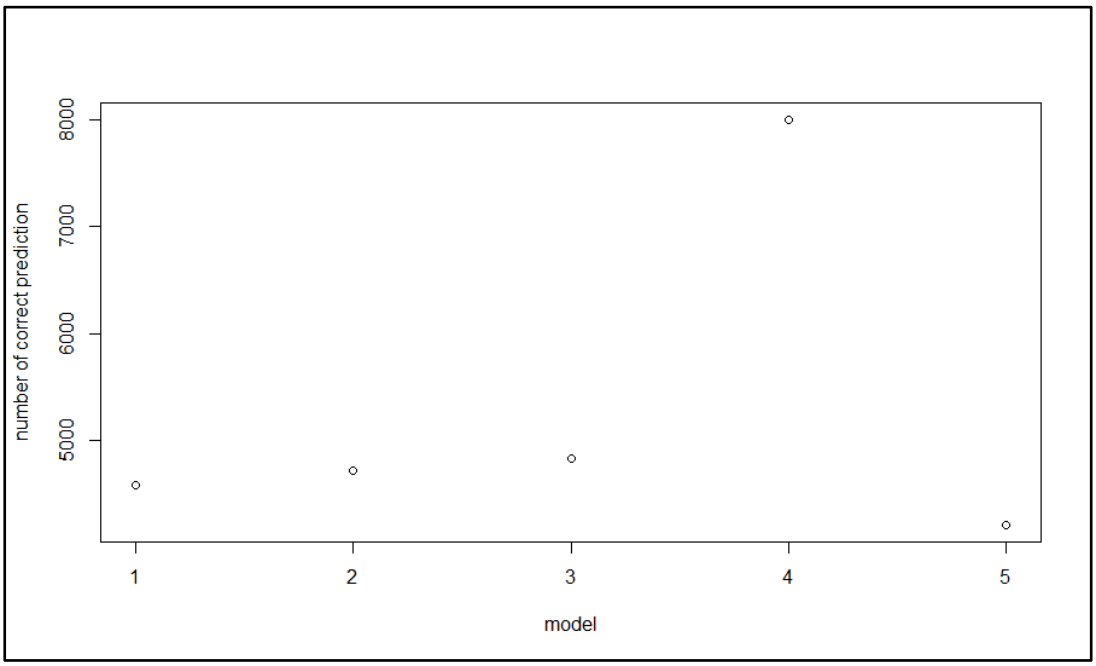
Please note. When you become a registered student of the University at first and any subsequent registration you sign the following authorisation and declaration: "I confirm that the information I have given on this form is correct. I agree to observe the provisions of the University's Charter, Statutes, Ordinances, Regulations and Codes of Practice for the time being in force. I know that it is my responsibility to be aware of their contents and that I can read them on the University web site. I acknowledge my obligation under the Payment of Fees Section in the Handbook to pay all charges to the University on demand. I agree to the University processing my personal data (including sensitive data) in accordance with its Code of Practice on Data Protection http://www.leeds.ac.uk/dpa . I consent to the University making available to third parties (who may be based outside the European Economic Area) any of my work in any form for standards and monitoring purposes including verifying the absence of plagiarised material. I agree that third parties may retain copies of my work for these purposes on the understanding that the third party will not disclose my identity."'

1.) Firstly, random forest fit is method where many decision tree classifiers are being fitted from the dataset. The random forest fit will give us improved predictive accuracy and avoid overfitting.

For finding which input does have the best fit, we can loop over the model for each input with corresponding prediction and find the correct factor by comparing total correct prediction between mushroom data and prediction.

	Models	Total.correct.prediction
model 1	CapShape	4584
model 2	CapSurface	4716
model 3	CapColor	4836
model 4	Odor	8004
model 5	Height	4208

Additionally, by plotting each output correspond to every model result on below figure,



This shows the model 4 which is **odor** has the highest total correct prediction which can be implied as the best fit for edibility. Moreover, to strengthen our belief, we can see from confusion matrix correspond to each model below,

Model 1:

```
> model1$confusion
          Edible Poisonous class.error
Edible      3980       228  0.05418251
Poisonous   3312       604  0.84576098
```

Model 2:

```
> model2$confusion
          Edible Poisonous class.error
Edible      1560      2648   0.6292776
Poisonous    760      3156   0.1940756
```

Model 3:

```
> model3$confusion
          Edible Poisonous class.error
Edible      3080      1128   0.2680608
Poisonous   2160      1756   0.5515832
```

Model 4:

```
> model4$confusion
          Edible Poisonous class.error
Edible      4208         0  0.00000000
Poisonous    120      3796  0.03064351
```

Model 5:

```
> model5$confusion
          Edible Poisonous class.error
Edible      4208         0           0
Poisonous   3916         0           1
```

To see correct results, we need to follow the diagonal matrix from the confusion matrix. As we can see, model 4 has the highest total number of diagonal matrices which sum up to 8004 like we stated before. Therefore, **model 4 with odor as the input has the best fit among others.**

2.) From the allcombs() function correspond to our 5 different input give us 32 different combinations. Since one of them are empty so it is omitted, remain us with only 31 combinations.

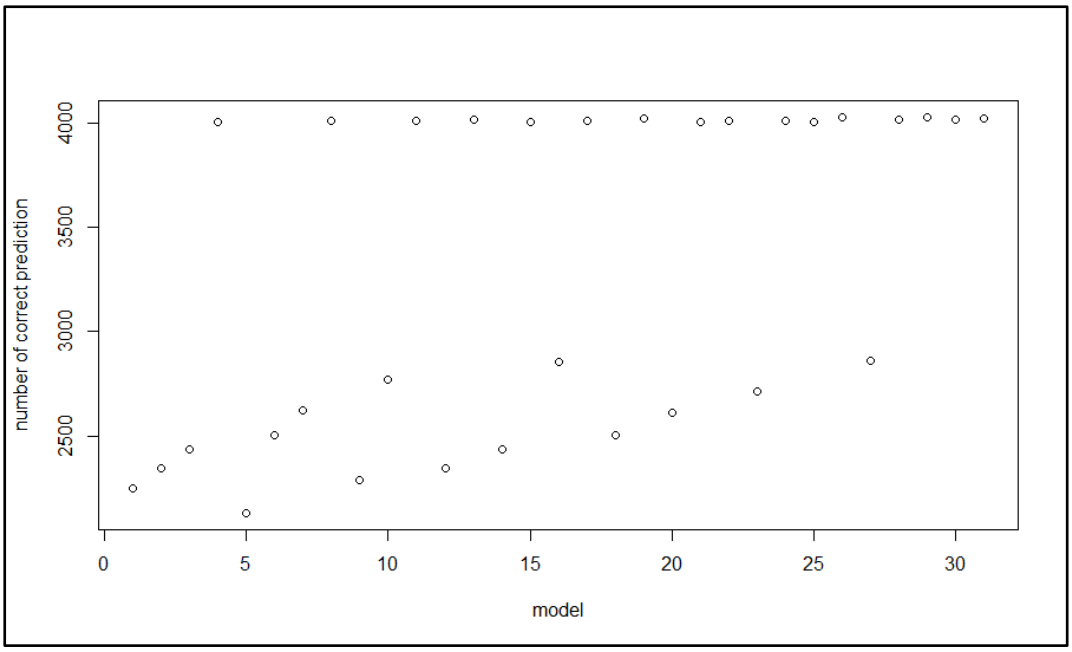
```
 [1] "CapShape"                                  "CapSurface"
 [3] "CapColor"                                  "Odor"
 [5] "Height"                                    "CapShape+CapSurface"
 [7] "CapShape+CapColor"                         "CapShape+Odor"
 [9] "CapShape+Height"                           "CapSurface+CapColor"
[11] "CapSurface+Odor"                           "CapSurface+Height"
[13] "CapColor+Odor"                             "CapColor+Height"
[15] "Odor+Height"                               "CapShape+CapSurface+CapColor"
[17] "CapShape+CapSurface+Odor"                  "CapShape+CapSurface+Height"
[19] "CapShape+CapColor+Odor"                    "CapShape+CapColor+Height"
[21] "CapShape+Odor+Height"                      "CapSurface+CapColor+Odor"
[23] "CapSurface+CapColor+Height"                "CapSurface+Odor+Height"
[25] "CapColor+Odor+Height"                      "CapShape+CapSurface+CapColor+Odor"
[27] "CapShape+CapSurface+CapColor+Height"       "CapShape+CapSurface+Odor+Height"
[29] "CapShape+CapColor+Odor+Height"             "CapSurface+CapColor+Odor+Height"
[31] "CapShape+CapSurface+CapColor+Odor+Height"
```

Some validation test is going to be brought after. Using ratio of 50:50 for training and test dataset, we obtained the single validation test for 31 combinations below,

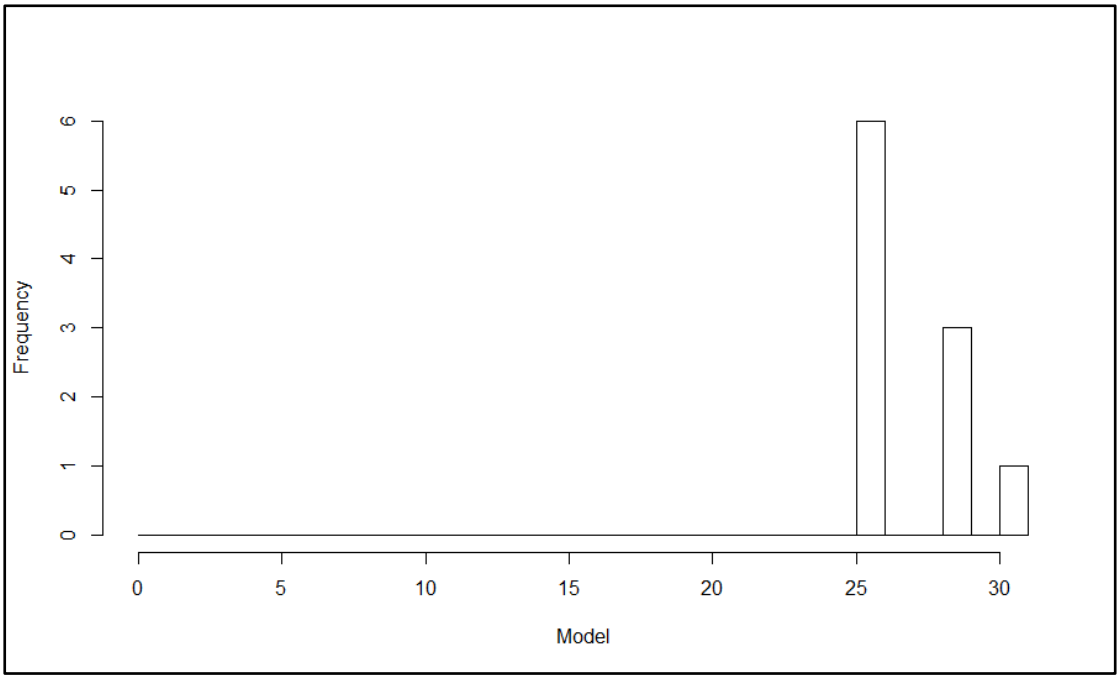
	formula	Total.correct.prediction
1	CapShape	2246
2	CapSurface	2331
3	CapColor	2431
4	Odor	4006
5	Height	2000
6	CapShape+CapSurface	2472
7	CapShape+CapColor	2574
8	CapShape+Odor	4007
9	CapShape+Height	2284
10	CapSurface+CapColor	2731
11	CapSurface+Odor	4009
12	CapSurface+Height	2331
13	CapColor+Odor	4016
14	CapColor+Height	2431
15	Odor+Height	4006

	formula	Total.correct.prediction
16	CapShape+CapSurface+CapColor	2797
17	CapShape+CapSurface+Odor	4007
18	CapShape+CapSurface+Height	2472
19	CapShape+CapColor+Odor	4022
20	CapShape+CapColor+Height	2552
21	CapShape+Odor+Height	4006
22	CapSurface+CapColor+Odor	4006
23	CapSurface+CapColor+Height	2683
24	CapSurface+Odor+Height	4007
25	CapColor+Odor+Height	4006
26	CapShape+CapSurface+CapColor+Odor	4033
27	CapShape+CapSurface+CapColor+Height	2806
28	CapShape+CapSurface+Odor+Height	4022
29	CapShape+CapColor+Odor+Height	4024
30	CapSurface+CapColor+Odor+Height	4017
31	CapShape+CapSurface+CapColor+Odor+Height	4030

The result is quite convincing since the formulas with correct number prediction > 4000 are all contain **odor** in it. This might support the finding from question 1. Moreover, we will plot it to see it more clearly,



Since the result is very stagnant no interesting trend we can point out. Therefore, further research by looping for 10 times among the combinations need to be carried. By finding the maximum from each loop out of 10, we get histogram below,



This shows model 26 which is **CapShape + CapSurface + CapColor + Odor** is the best formula to explain the correct prediction by reaching the highest correct prediction on 6 out of 10 loops compare to the others formula. Which followed by model 29 and 31 respectively.

3.) If I was being asked for using such classifier for foraging mushrooms, the answer is **YES**. As I said in the question 1, this analysis is using decision tree classifier and this random forest method is finding the most optimized result with lower overfitting compare to glm method. Additionally, it is perfect for such cases which all 6 input are factors.

It is obvious after received result from question 1 and 2, using the best fit estimator and best formula. **Odor** is the most important factor for foraging mushrooms.

If we refer back to the confusion matrix, we can find the probability of being mistakenly poison due to this classifier such as 120/8124 = 0.0148 overall error rate. The 120 is wrongly classified mushroom which believe as edible in fact it is poisonous. Nevertheless, there is no wrongly classified for wrongly believed edible. Therefore, we would say 0.0148 rate of us being poisoned by following random forest method classifier.

```
> model4$confusion
          Edible Poisonous class.error
Edible      4208         0  0.00000000
Poisonous    120      3796  0.03064351
```