

Project 1 Stat - Yiran

The key points are highlight by bold.

Dataset description

Dataset description

This is a **retrospective study**. The disease status is known ahead of time and patients are screened to access for risk factors.

Response DIABBC (Categorical variable):

Only diabetes currently and healthy(never has diabetes) are presented here. relatively few diabetes quit and these observations (around 50 ppl) have been excluded from the data

- Yes(1) : (Diabetes group) Ever told has diabetes mellitus, still current and long term
- No(0) : (Healthy group) Never told has diabetes mellitus

The response is relatively unbalanced (i.e, we have much more observations in the healthy group, compared to the diabetes group): **597 people in diabetes group, 11506 people in health groups**. In total, there are 12103 people participate in this dataset.

Predictors

See the detail in [appendix](#)

'Diabetes', 'Age", 'Sex', 'Income Level", Exercise Minutes in last week', 'BMI", Met Recommended Dietary Guidelines, 'Zinc (mg)', 'Red Met Consumption , Selenium (ug)

Data pro-processing detail

- Modify DIABBC = 1 to DIABBC = Yes; DIABBC = 5 to DIABBC = No
- remove outliers :
 - For the exercise time, if the exercise time > 1500 mins/week, trade as outliers, set the value to 0
- remove children (age < 18)

- remove currently pregnant (sabdym = 4)
- remove the group DIABBC = 3 as too few people in this group

Data used for modelling (Decision Tree + Random Forest)

- remove rows with missing values for the selected columns before modelling
- Randomly split 80% data for training, and 20% data for testing the model performance only.

Variable Selection for Diabetes

Key Variables related to Diabetes

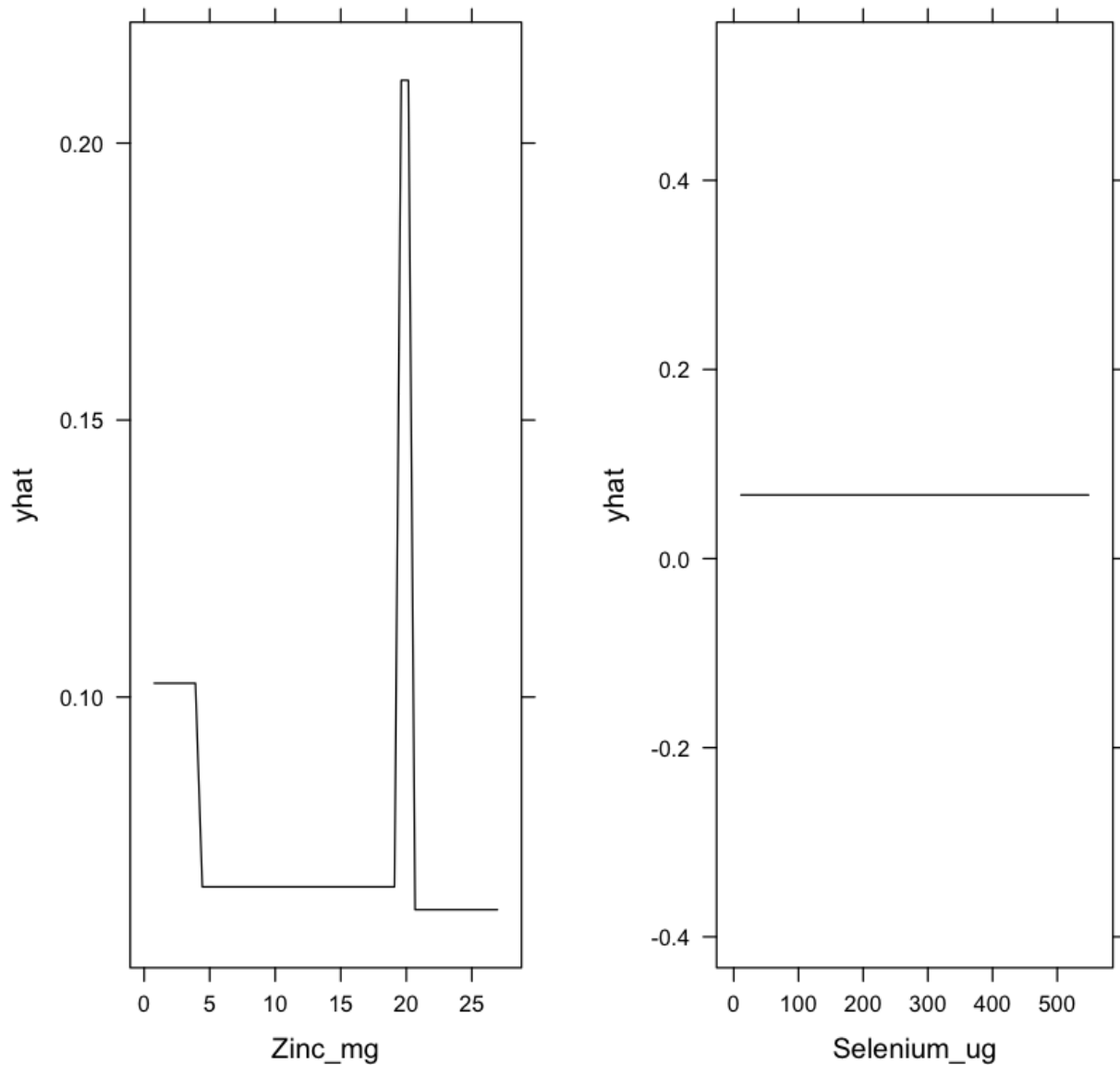
ZINCT

Confirm using ZINCT_MEAN, instead of ZINCT1, ZINCT2 is reasonable: see [appendix](#)

Partial dependence plots between ZINCT_MEAN and Diabetes based on Decision Tree (DT)

To interpret the marginal effects of plots we can use partial dependence plots.

Partial dependence plots of Zinc (mg) and Selenium (ug) based on Decision Tree

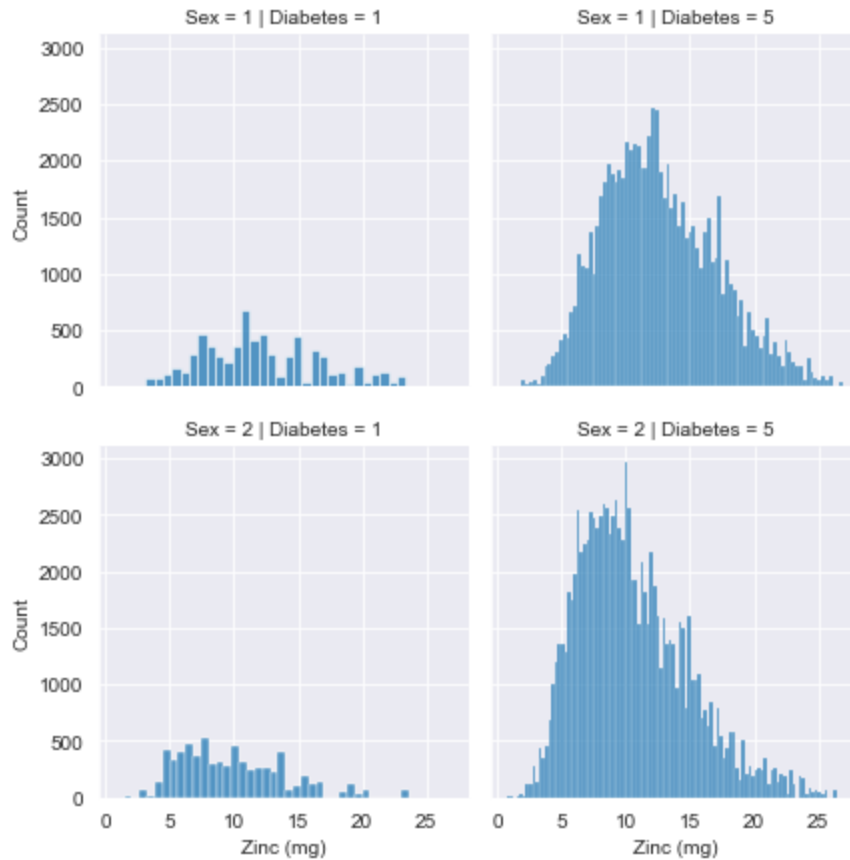


The yhat is the predicted probability of diabetes based on DT model.

From the plot we can see that:

- when $\text{Zinc (mg)} < 5$ or $18 < \text{Zinc (mg)} < 21$, the ppl are more likely (higher probability) get diabetes.
- Based on decision tree model, Selenium has no influence on the result prediction.

The distribution of Zinc (mg) conditional on Diabetes and Sex



From the Distribution of Zinc (mg) conditional on Diabetes and Sex, we can see that:

- **The distribution of Zinc (mg) conditional on Sex is different!**
- The distribution of Male (sex = 1) in both healthy and diabetic group are symmetric, while the distributions of Female (sex = 2) are right skewed, which imply that the mode/mean of Zn in female group is lower than male.
- Thus, there is relationship between Sex, Zn and diabetes (this relation will be further confirmed in the [Markov graph](#) section below)

T test: Check the mean of Zn in groups are different (healthy group VS diabetes group)

Conclusion: Since the $p_value < 0.05$, we can see that the mean of healthy group and diabetes group are statistically significant. And these the mean of Zn in these two group are different.

See the T-test stat details in [appendix](#)

SELT

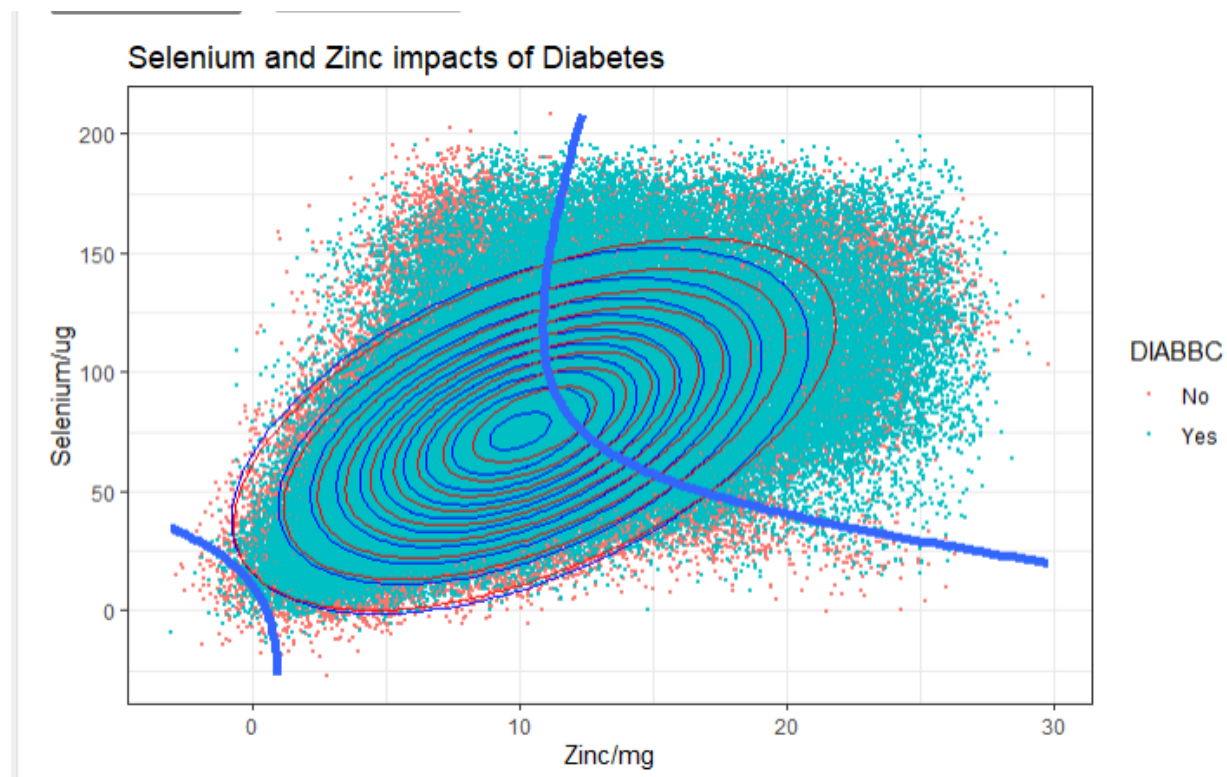
Confirm using SELT_MEAN, instead of Se1, Se2 is reasonable: see [appendix](#)

T test: Check the mean of Se in groups are different (healthy group VS diabetes group)

- Conclusion: Since the $p_value < 0.05$, we can see that the mean of healthy group and diabetes group are statistically significant. And these the mean of Sn in these two group are different.
- see detail in [appendix](#)

Combine Zn and Se

Now we want to know how Zn and Se influent DIABBC together. Then we plot them together, blue dots are diabete group, and orange dots are health group. There are two regression lines. The right top corner and left bottom corner points has higher risk to get diabetes.



ANOVA test to analysis the interaction impact between DIABBC and interaction of Se and Zn

Since the p-value is less than 0.05. So there are correlation between DIABBC and interaction of Se and Zn.

See detail in [appendix](#)

Sex

Risk measurement: Odds ratios (OR) for sex differences on Diabetes

odds ratio is a way to measure risk. The odds of success is the ratio of the chance of success(healthy) to the chance of failures(has diabetes).

	Diabetes group	Healthy group	rowTotal
SEX			
1	2301	191	2492
2	2823	184	3007
All	5124	375	5499

Odds ratio interpretation:

- The estimated odds of diabetes among female is 0.785 times (lower risk) the estimated odds for those diabetes among male
- Since $OR < 1$, implies decreased risk of disease if the ppl is female
- 95% CI for the odds ratio is (0.64, 0.97)

Pearson's chi-squared test for independence between DIABBC and Sex

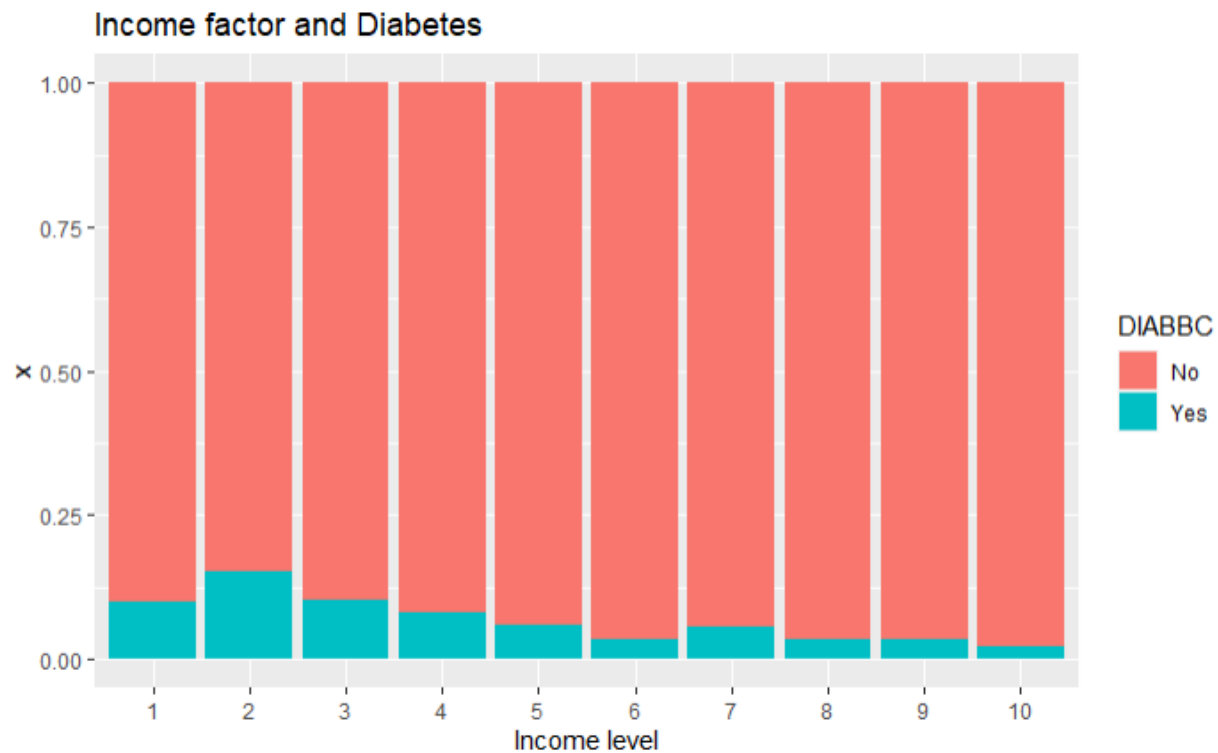
Chi-square test statistic

- Null-hypothesis: DIABBC and sex are independent
- Alternative hypothesis: they are not independent
- Chi-square test stat = 5.122, with $p_value = 0.0136$

Since $p_value < 0.05$, these two variables are dependent. See detail in [appendix](#)

Income

We believe Income and diabetes has correlation, and we plot the percentage graph of it. As we can see, higher income people have lower risk to get diabetes. We also make a chi-squared test to see if there are any correlations.



Pearson's chi-squared test for independence between DIABBC and Income

Conclusion: Since the p-value is extremely low (less than 0.05), so there are dependent.

See detail in [appendix](#)

Build Red Meat indicator

From the research, we believe `Vegan` has less risk suffer from diabetes, but since our group has less than 10 ppl are vegan (too less sample), we use red-meat indicator instead.

- `MEAT1N`: Meat, poultry (serves/day)

For red meat (`MEAT1N(nutr)`)

Where 0 intake = not taken red meat (`RedMeat_indicator = 0`)

	Diabetes group	Healthy group	rowTotal
RedMeat_indicator			
0	504	28	532
1	4620	347	4967
All	5124	375	5499

Risk measurement: Odds ratios (OR) for red meat differences on Diabetes

- The estimated odds ratio of diabetes for red meat indicator is 1.35 times (higher risk) the estimated odds for those not eating red-meat.
- Since $OR > 1$, implies increased risk of disease

Pearson's chi-squared test for independence between DIABBC and RedMeat_indicator

Since $p_value < 0.05$, these two variables are dependent. See detail in [appendix](#)

Markov graph for Zn and Se and Partial Dependency Check

The Markov graph can be used to visualization relationship among continuous variables

We consider following variables

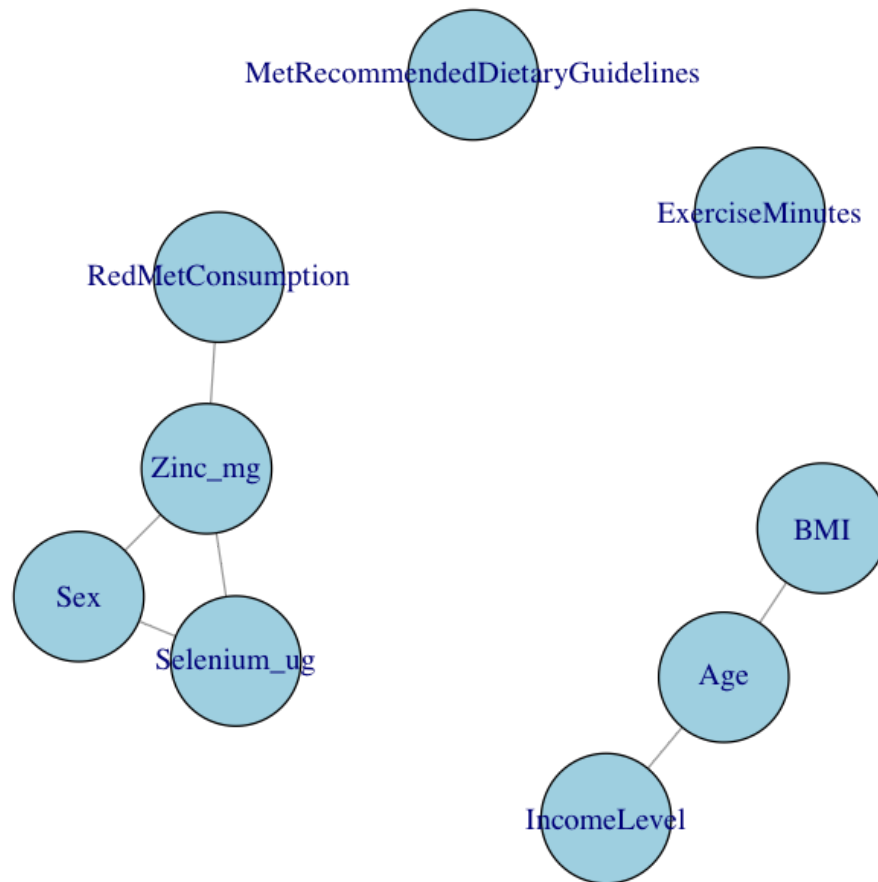
"DIABBC", "SEX", "AGEC", "INCDEC", "EXLWTBC", "BMISC", "ZINCT_MEAN", "SELT_MEAN"

This graph is based on the `partial correlation`

Use the **parital correlation matrix** to plot the corresponding Markov graph where an edge appears in the graph if the corresponding absolute partial correlation is greater than 0.1.

Markov graph overall

Markov graph (the partial correlation) overall



If there is line between nodes, it means these two variables have relationship (partial correlation > 0.1)

- This we find relationship among **Selenium (ug)**, **Zinc (mg)**, **Red Met Consumption** and **'Sex'**.
- **Income Level** is dependent with **Age** and **BMI**
- **Met Recommended Dietary Guidelines** and **Exercise Minutes (per week)** are completely independent with other predictors.

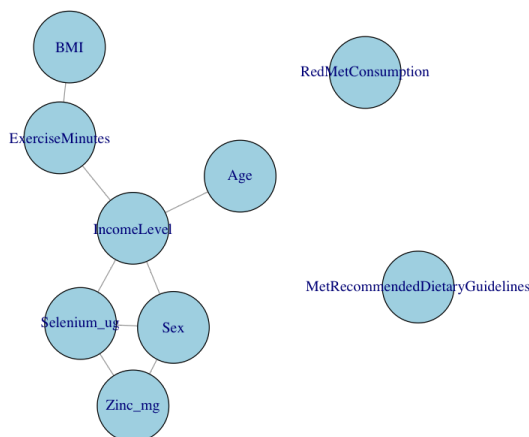
Markov graph conditional on Diabetes

Separate graph based on diabetes, if has line between 2 nodes, it means these two node/variables has relationship

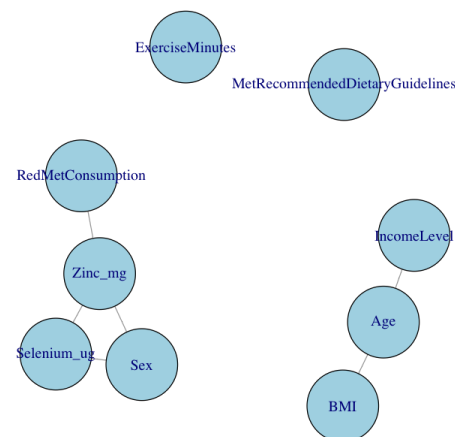
The first graph below is based on the diabetes group, the second graph below is based on the healthy group. So we can see that

1. In the diabetes group, **Selenium (ug)** and **Zinc (mg)** are dependent with **Income Level**, by contrast, while in the healthy group, **Selenium (ug)** and **Zinc (mg)** are independent with **Income Level**, which at least imply that the level of **Selenium (ug)** and **Zinc (mg)** among diabetes are different with the normal group, and thus **Selenium (ug)** and **Zinc (mg)** can be the good indicator for the diabetes detection.
2. Conditional Independence: **Selenium (ug)** and **Zinc (mg)** are conditional independent with **Age**, **exercise minutes** and **BMI** conditional on **Income Level**
3. The **Sex**, **Selenium (ug)** and **Zinc (mg)** are dependent with each other in both group. (there is 3-way interaction)
4. The **Red Met Consumption** has no relation with the level of **Selenium** and **Zinc** in the diabetes group
5. In all graphs, **Met Recommended Dietary Guidelines** is partial independent with other nodes, which might imply that Met dietary guidelines or not is not a good indicator in our case study.

Markov graph (the partial correlation) for diabetes group



Markov graph (the partial correlation) for healthy group



Diabetes Prediction Model

Model 1: Decision Tree

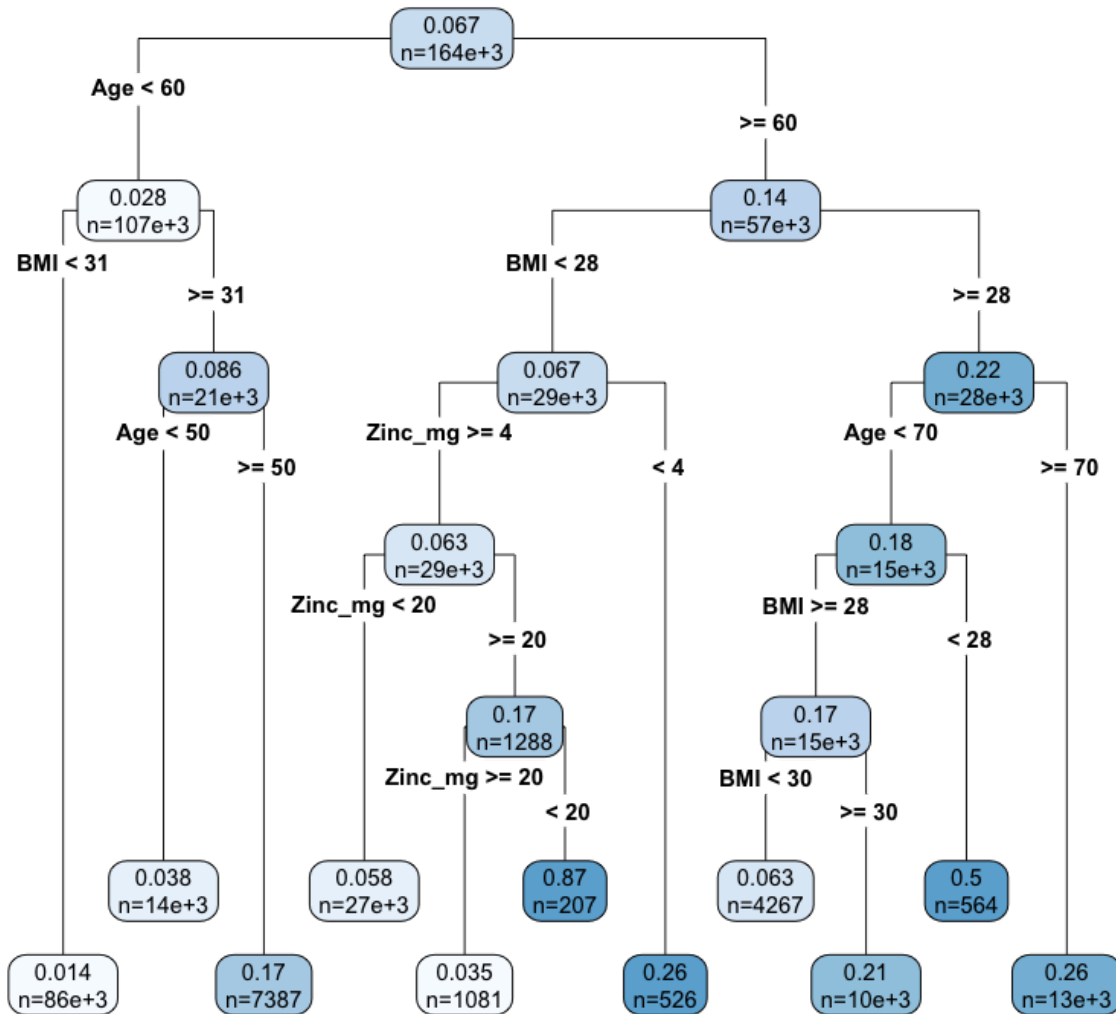
I use cross-entropy loss, combined with 10-fold cross validation method (repeated 10 time) to fit the tree from down to top. We tune the value of complexity parameter (cp) by creating a hyper-parameter grid for cp with range from 0.001 to 0.12.

From plot in [appendix](#), we can see that, the smaller the value of complexity parameter (cp), the higher accuracy, but also more complex the tree. To avoid overfitting and get a simpler tree for interpretation purpose, I choice the pruning ****complexity parameter (cp)**** as 0.005, which means that stop splitting when the total loss ≤ 0.005 . Thus, we cut the full tree (cp = 0) to the new decision tree with optimal value of cp = 0.005

Tree visualisation and interpretation

- At the top, it is the overall probability of Diabetes. It shows the proportion of ppl that get diabetes. (6.7%)-
- On the next level, it ask if the ppl's age < 60, if yes, the proportion of ppl that get diabetes is 2.8%.
- keep the same logic for interpretations. **So for example, if the ppl with age > 60, BMI < 28, ZINCT (mg) ≥ 4 , and ZINCT (mg) < 20, then the ppl has 0.87 predicted probability of diabetes**

Diabetes Study



Model 2: Random Forest

Turning hyper-parameters of Random Forest

Next we will tune the RF by creating a hyperparameter grid with some values of `mtry` (the number of variables to randomly select at each split), the `min.node.size` (minbucket in rpart), whether to sample with or without replacement (replace), and the `sample.fraction`. We access the top 10 models (See appendix)

Thus, the suggested model should choose

`num.trees = n_features * 5,`

```
mtry = 3,  
min.node.size = 5,  
sample.fraction = .50,  
replace = FALSE,
```

We fit random forest model based on the parameters suggested above.

Model 3: Logistic Regression

Mingjie to do

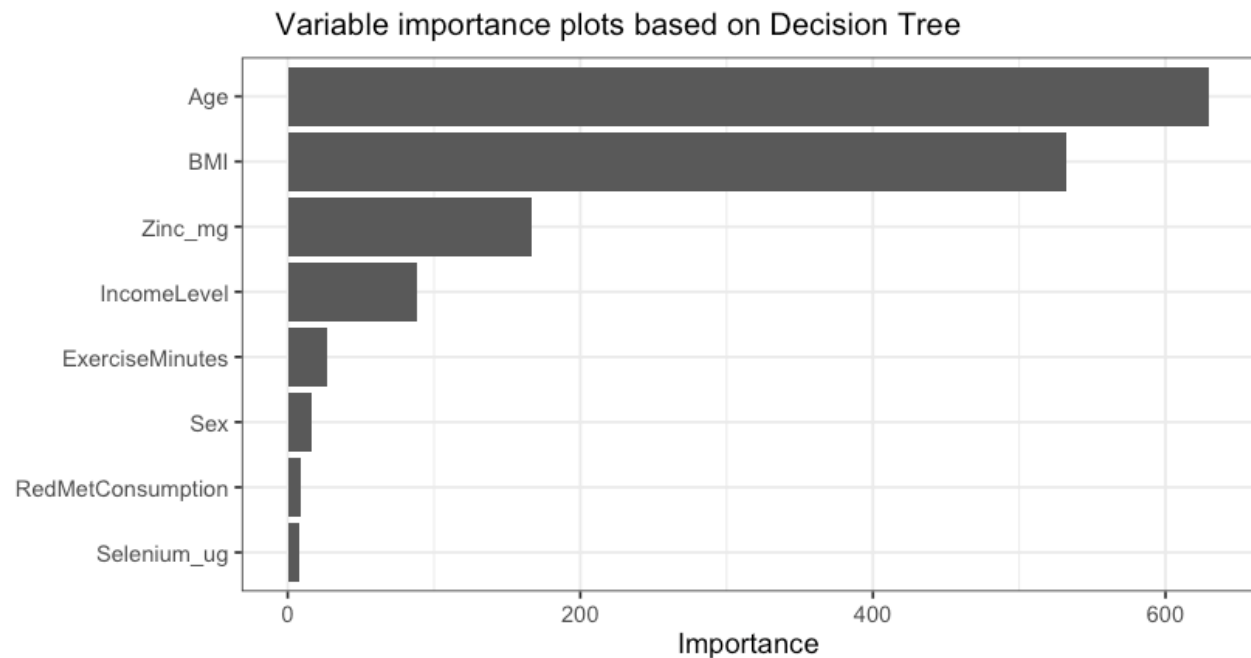
1. select the logistic with best performance
2. Model-based interpretation (given formula)

Which variables are important for classification if the target?

Variable importance based on decision Tree

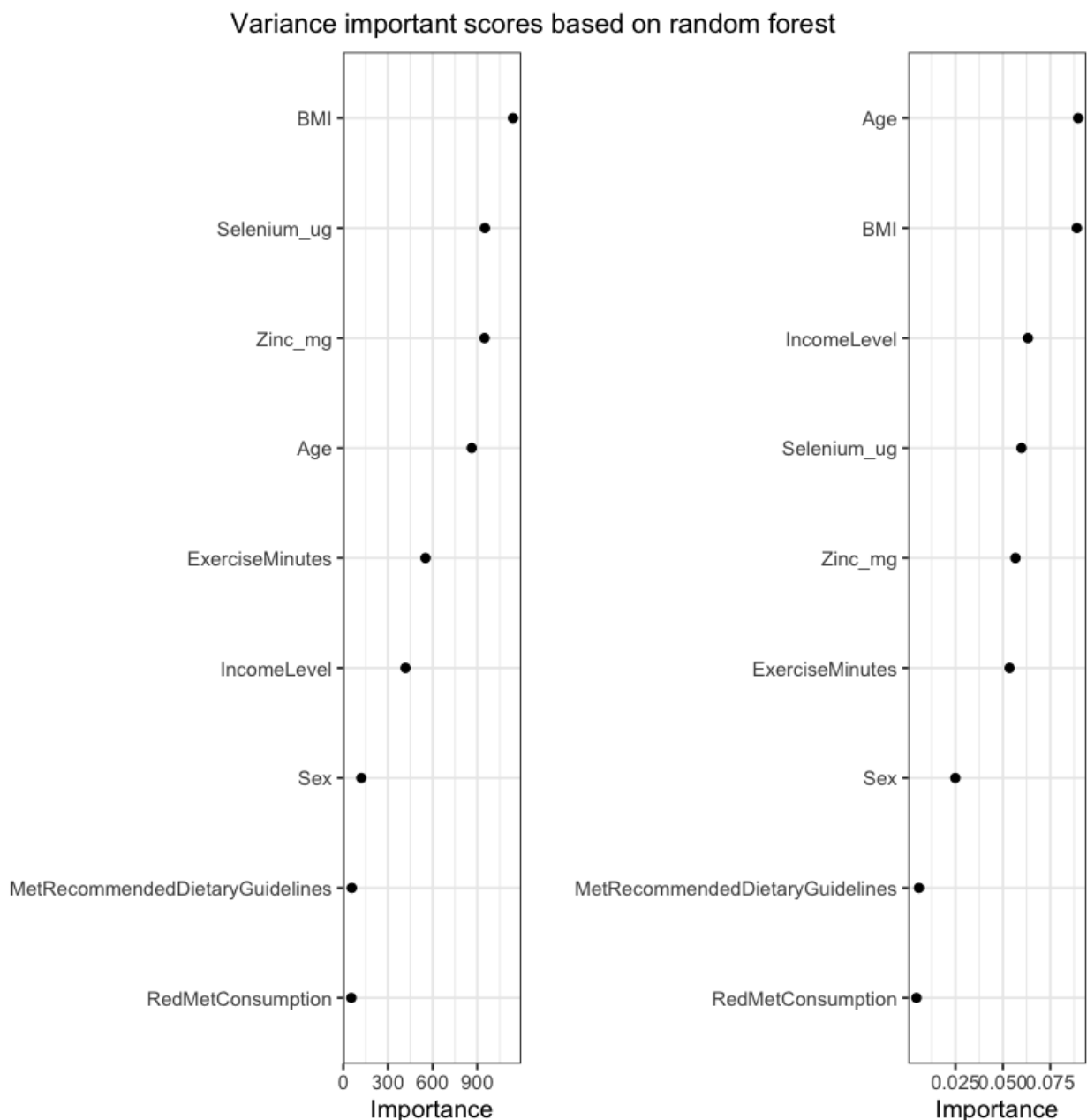
Let's have a look at the variable importance scores for the top variables.

- We can see that the top important variables are Age , BMI, Zinc (mg), and Income Level, the Selenium (ug) is relative less important.



Variance important scores based on Random Forest

We calculate the variance important scores using an `impurity` (left) and a `permutation` (right) approach, based on the turned hyper parameters.



From the plot above, we can see that using Random Forest model, the top 6 most important variables are the same: `Age`, `BMI`, `Zinc (mg)`, `Income Level`, `Selenium (ug)` and `Exercise Minutes` using both methods.

Based on result of selected important variables in both Decision Tree and Random Forest, **the robust important variables are** `Age`, `BMI`, `Zinc (mg)`, `Income Level`,

Selenium (ug) .

What effect they have on the target variable?

Partial dependence on the important variables based on Decision Tree

To interpret the marginal effects of plots we can use partial dependence plots.

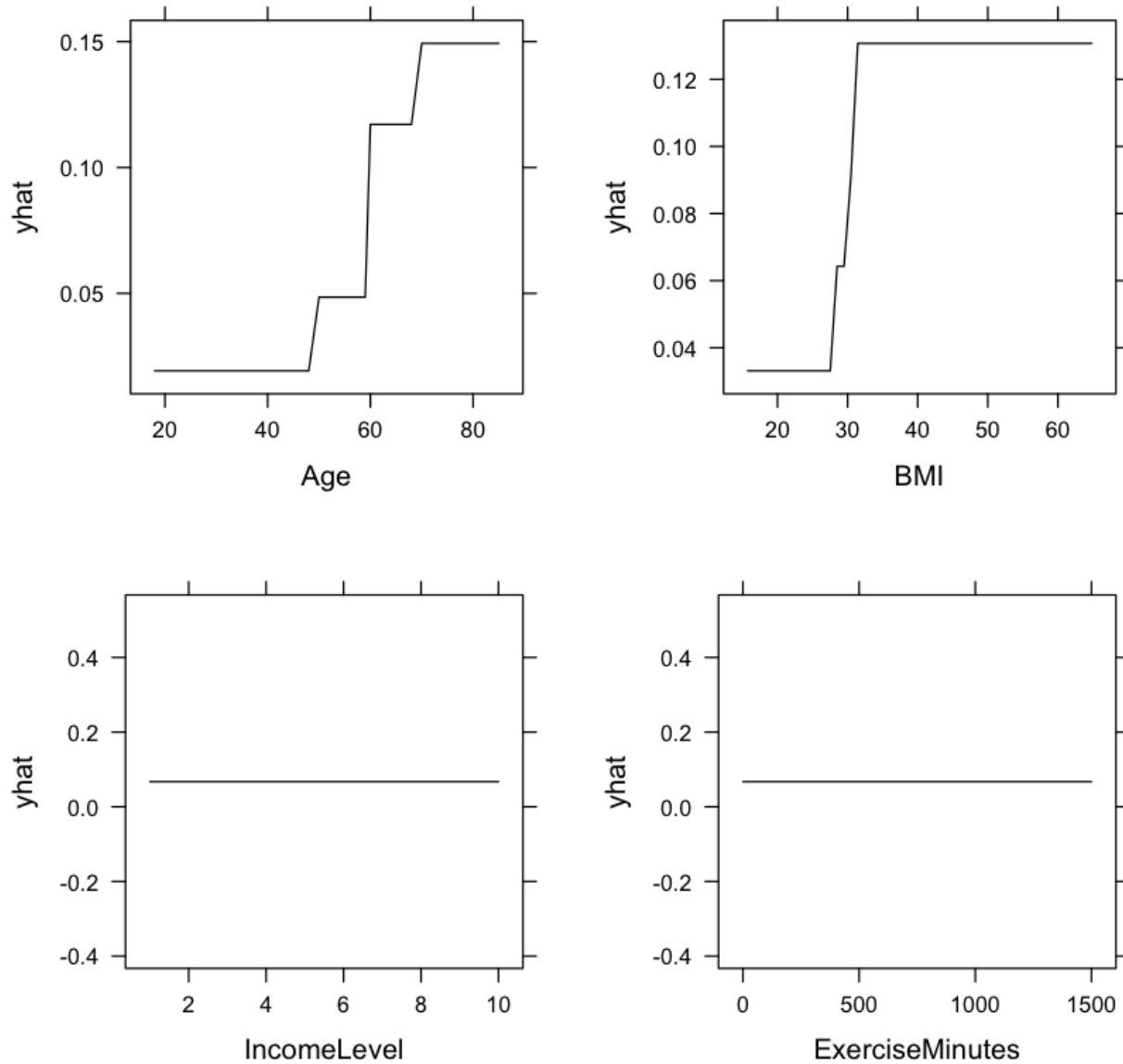
From the plot below we can see that, as the age and BMI increase, the probability of suffering diabetes increase as well

1. Partial dependency result on Zn and Se and See see [section above](#)
2. Partial dependency on Sex and BMI

The elder the people, the higher probability that be classification as diabetes.

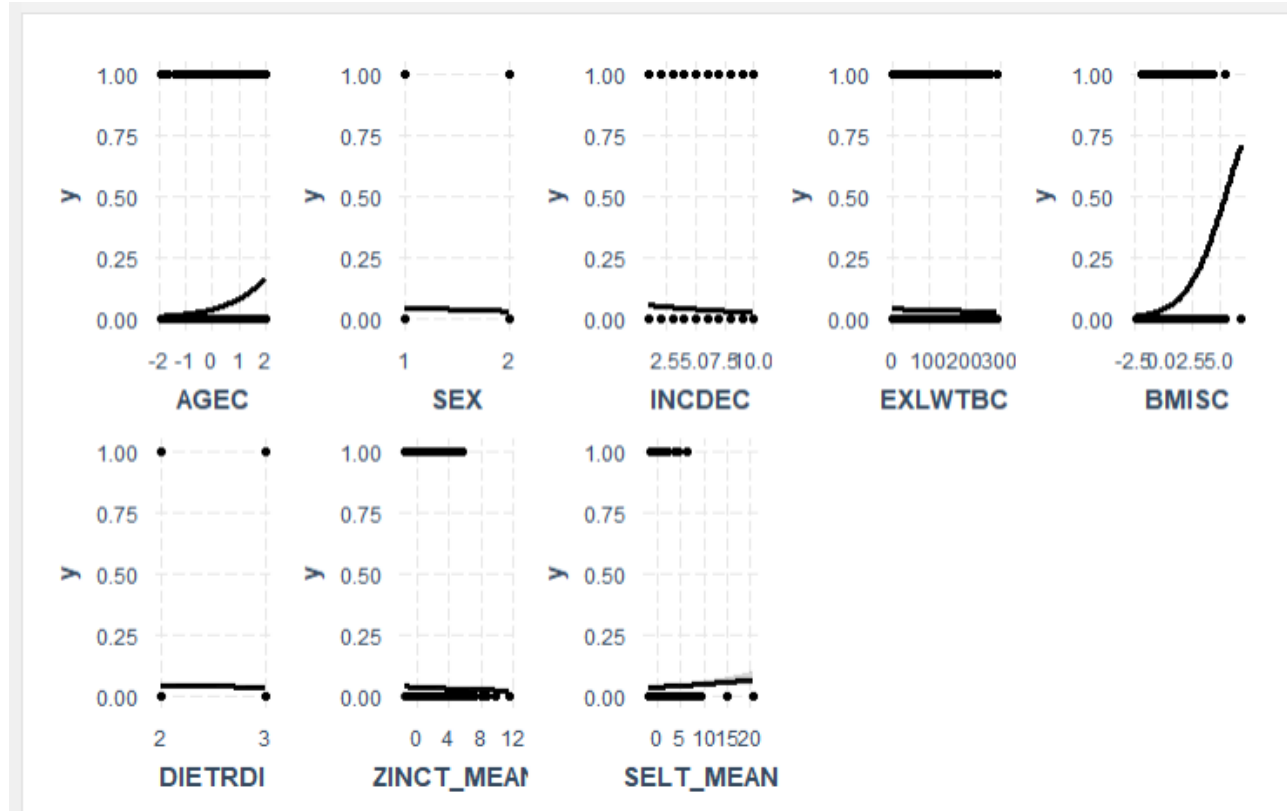
Similarly, the larger the BMI, the higher probability that be classification as diabetes.

Partial dependence plots based on Decision Tree



The effect sizes of predictors based on the best performing logistic regression classifier

mingjie todo. (add title, change variable name, write prediction)



User-uploaded image: group_infact.png

Model assessment

BaseLine - the majority class classifier

We have 597 observations in the diabetes group (1), 12103 observations in the healthy group (0), when the model always predicts the class with the largest number of cases, the misclassification rate = $597 / (12103 + 597) = 0.047$, the accuracy is 0.953.

Multiple Cross Validation









Multiple cross validation error plot

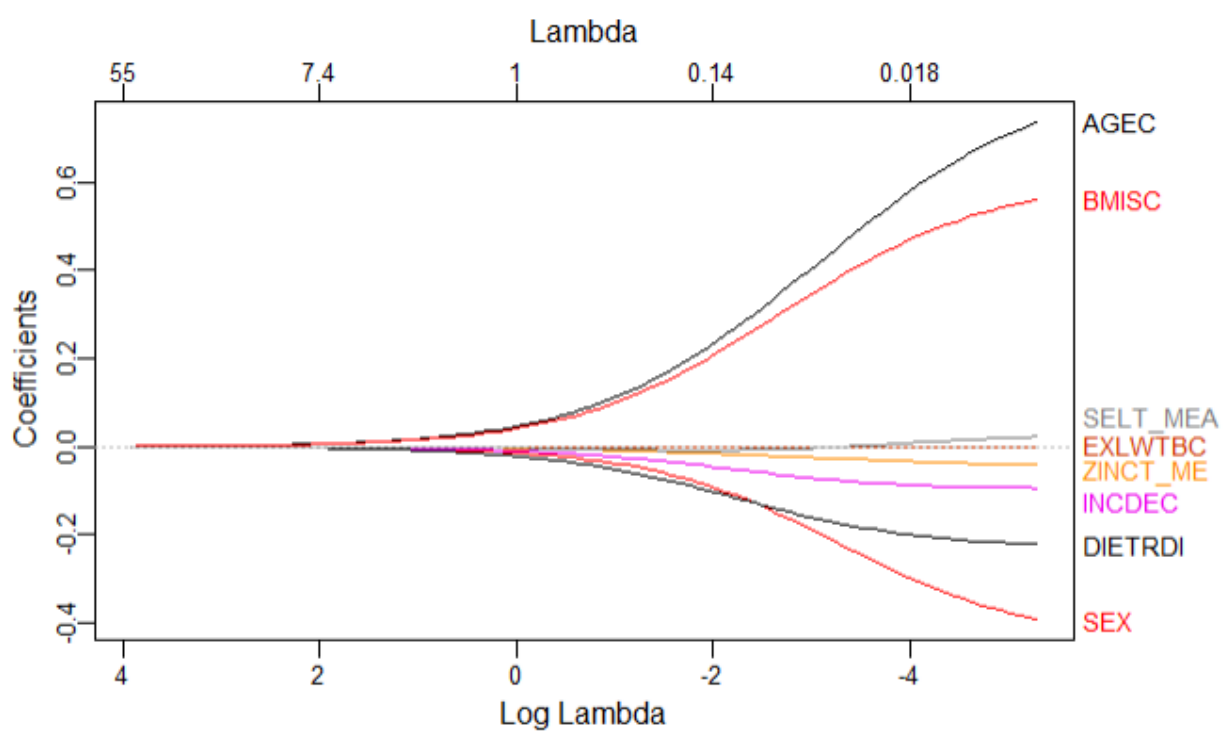
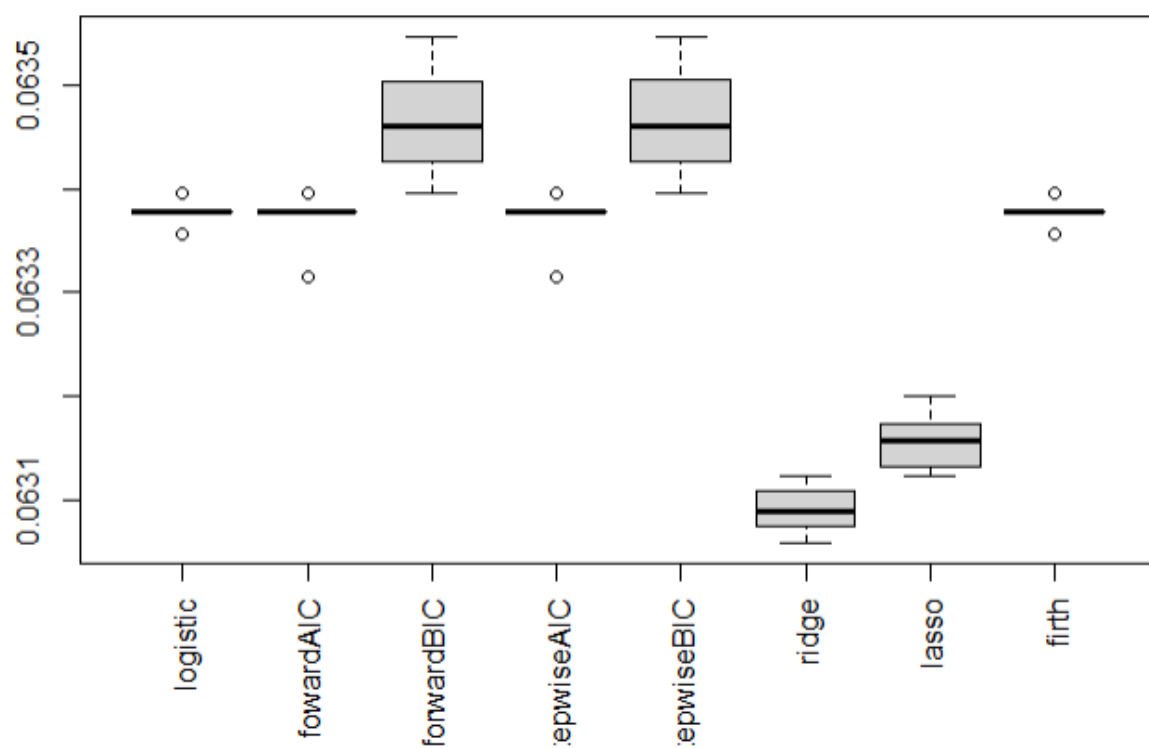
Multiple cross validation error summary table

Un-Interpreted plots given by Mingjie

	logistic full	forward AIC	forward BIC	backward AIC	backward BIC	firth
(Intercept)	-1.348***	-1.348***	-1.355***	-1.348***	-1.355***	-1.347***
	(0.108)	(0.108)	(0.108)	(0.108)	(0.108)	(0.108)
AGEC	0.836***	0.836***	0.836***	0.836***	0.836***	0.836***
	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)
SEX	-0.445***	-0.445***	-0.442***	-0.445***	-0.442***	-0.445***
	(0.019)	(0.019)	(0.019)	(0.019)	(0.019)	(0.019)
INCDEC	-0.094***	-0.094***	-0.094***	-0.094***	-0.094***	-0.094***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
EXLWTBC	-0.001***	-0.001***	-0.001***	-0.001***	-0.001***	-0.001***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
BMISC	0.613***	0.613***	0.614***	0.613***	0.614***	0.612***
	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)
DIETRDI	-0.231***	-0.231***	-0.230***	-0.231***	-0.230***	-0.231***
	(0.034)	(0.034)	(0.034)	(0.034)	(0.034)	(0.034)
ZINCT_MEAN	-0.047***	-0.047***		-0.047***		-0.047***
	(0.011)	(0.011)		(0.011)		(0.011)
SELT_MEAN	0.032***	0.032***		0.032***		0.032***
	(0.011)	(0.011)		(0.011)		(0.011)
Num.Obs.	219429	219429	219429	219429	219429	219429
AIC	86661.6	86661.6	86677.7	86661.6	86677.7	86661.6
BIC	86754.2	86754.2	86749.8	86754.2	86749.8	86754.2
Log.Lik.	-43321.777	-43321.777	-43331.870	-43321.777	-43331.870	-43321.778

* p < 0.1, ** p < 0.05, *** p < 0.01

	Unique (#)	Missing (%)	Mean	SD	Min	Median	Max	
logistic	3	0	6.34	0.00	6.34	6.34	6.34	
fowardAIC	3	0	6.34	0.00	6.33	6.34	6.34	
forwardBIC	16	0	6.35	0.00	6.34	6.35	6.35	
stepwiseAIC	3	0	6.34	0.00	6.33	6.34	6.34	
stepwiseBIC	17	0	6.35	0.00	6.34	6.35	6.35	
ridge	11	0	6.31	0.00	6.31	6.31	6.31	
lasso	9	0	6.32	0.00	6.31	6.32	6.32	
firth	3	0	6.34	0.00	6.34	6.34	6.34	



Limitation of above stat

1. Unbalanced dataset:
2. Modelling limits: Decision tree is sensitive to outliers(large value of BMI for example), so the threshold value is larger than what we expected. (in BMI, we expect 25, 30, rather than 28, 31)
3. Confounding factors;
4. Missing variables: some important factors/variables havenot been included within the model

Appendix

Dataset Description

Variables	Original name in dataset	Meaning
Diabetes (Response)	DIABBC	Whether has diabetes mellitus: <ul style="list-style-type: none">• The diabetes group in our study is Ever told has diabetes mellitus, still current and long term.• The health group is Never told has

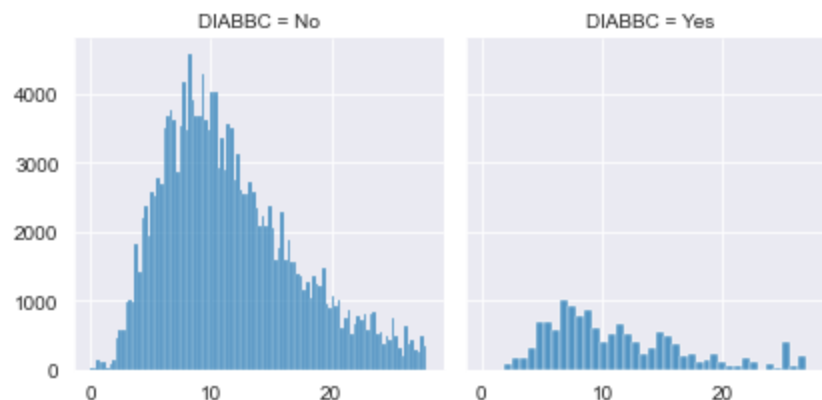
		diabetes mellitus
Age	AGEC	Age of person (y)
Sex	SEX	Sex of person
IncomeLevel	INCDEC	Equivalised income of household: deciles, range from 1 to 10
BMI	BMISC	BMISC
Zinc_mg	ZINCT_MEAN (the mean of ZINCT1 and ZINCT2)	The mean of Zinc (mg) in day 1 and day 2
Selenium_ug	SELT_MEAN (the mean of SELT1 and SELT2)	
RedMetConsumption	-	The mean of Selenium (ug) in day 1 and day 2
ExerciseMinutes	EXLWTBC	Total mins undertaken physical activity in last week
MetRecommended DietaryGuidelines	DIETRDI	Whether vegetable and fruit consumption met recommended

Confirm using ZINCT_MEAN is reasonable

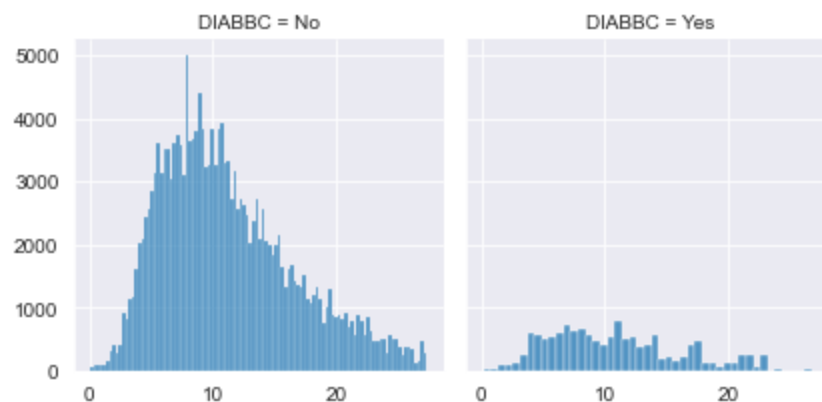
From plots below, we can see that

- The column ZINCT1, ZINCT2, and ZINCT_MEAN has quite similar distribution. So it make sense to use ZINCT_MEAN only in our following analysis.
- We can see that the distributions are right-skewed.
- The distribution of Zn in the healthy group looks quite different from that of diabete group

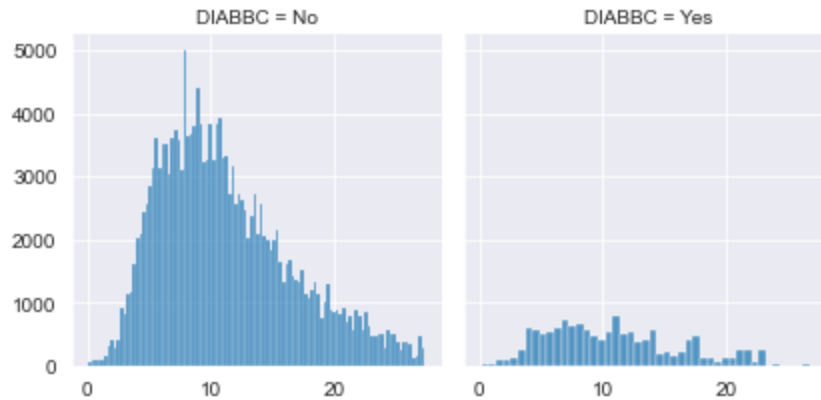
From plot above, we can see that ZINCT1, ZINCT2 and ZINCT_MEAN have quite similar distributions under both healthy group and diabete group, so in the following analysis, we use ZINCT_MEAN as the Zinct measurement.



ZINCT1 VS Diabetes.



ZINCT2 VS Diabetes.



ZINCT_MEAN VS Diabetes.

T test detail: Check the mean of Zn in groups are different

T- Test :- A t-test is a type of inferential statistic which is used to determine if there is a significant difference between the means of two groups which may be related in certain features.

- H0: The mean of Zn in healthy group (Diabbc = 0) and diabete gruop (Diabbc = 1) are the same
- H1: The mean of Zn in two groups are different
- **Our testing is based on the data removing the missing value, and mean of the two columns**
- If the $p_value < 0.05$, it means that the mean of two groups are different
- Since $p_value < 0.05$, we can see that the mean of two groups are statistically significant.

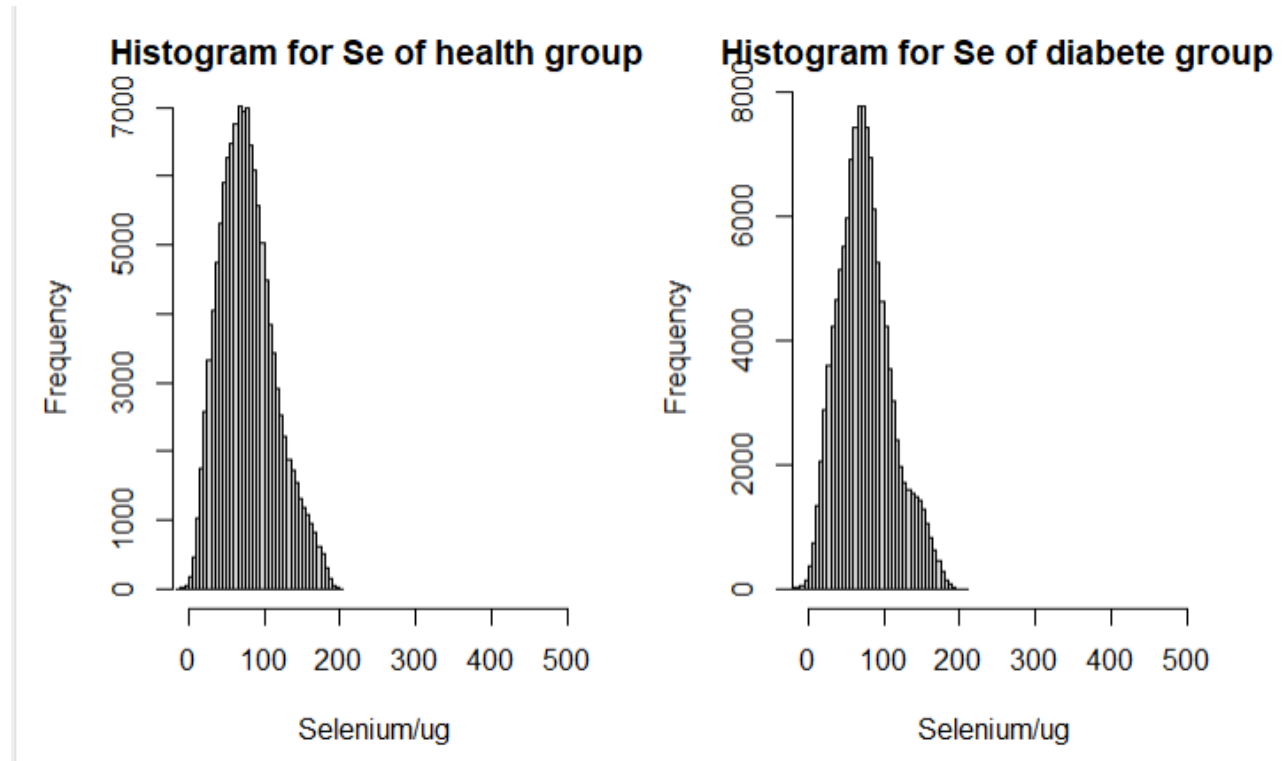
```
Mean of healthy group of ZINCT1: 11.727630477063556
Mean of diabetes group of ZINCT1: 11.189306695624573
T test on the ZINCT1:
Ttest_indResult(statistic=-10.786967658084365, pvalue=4.037367974784837e-27)
```

```
Mean of healthy group of ZINCT2: 11.259002909715415
Mean of diabetes group of ZINCT2: 10.786820399714147
T test on the ZINCT2:
Ttest_indResult(statistic=-9.774824473014103, pvalue=1.4614552395160844e-22)
```

```
Mean of healthy group of ZINCT_MEAN: 11.49331669338949
Mean of diabetes group of ZINCT_MEAN: 10.98806354766936
T test on the ZINCT_MEAN:
Ttest_indResult(statistic=-12.65381780395426, pvalue=1.1017792710573974e-36)
```

Confirm using S_MEAN is reasonable

Based on the graph between diabete and Se intake, there are not very significant difference between health group and diabete group. Because the number of observarions has huge difference of these two group. So we use ROSE package to balance two groups.



T test detail: Check the mean of Se in groups are different

- H_0 : The mean of Se in healthy group ($Diabbc = 0$) and diabete gruop ($Diabbc = 1$) are the same
- H_1 : The mean of Se in two groups are different
- Our testing is based on the data removing the missing value, and mean of the two columns
- If the $p_value < 0.05$, it means that the mean of two groups are different
- Since $p_value < 0.05$, we can see that the mean of two groups are statistically significant.
-

For health group:

```
welch Two Sample t-test

data: as.numeric(test_bfn_health_mean$DIABBC) and test_bfn_health_mean$SELT_MEAN
t = -1039, df = 241227, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -76.55947 -76.27115
sample estimates:
mean of x mean of y
 3.00000 79.41531
```

For diabetes group:

```
welch Two Sample t-test

data: as.numeric(test_bfn_diabetes_mean$DIABBC) and test_bfn_diabetes_mean$SELT_MEAN
t = -283.45, df = 17182, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -76.00879 -74.96479
sample estimates:
mean of x mean of y
 1.00000 76.48679
```

According to the result of these two T test, the mean Se of health group is 79.42, the mean Se of diabetes group is 76.49.

Detail of Pearson's chi-squared test for independence between DIABBC and Income

Here are the result of chi-squared test:

H0: DIABBC and INCDEC are independent

H1: DIABBC and INCDEC are dependent

	1	2	3	4	5	6	7	8	9	10
1	2373	3919	2491	1738	1414	644	1584	796	849	625
3	21971	22054	22077	19737	23358	19293	26406	23401	23584	27083

```
Pearson's Chi-squared test

data: tbl
X-squared = 6023.8, df = 9, p-value < 2.2e-16
```

As we can see the p-value is extremely low (less than 0.05), so there are dependent.

Detail of ANOVA test to analysis the interaction impact between DIABBC and interaction of Se and Zn

Then we can use two way anova to analysis the interaction impact between DIABBC and interaction of Se and Zn.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
ZINCT_MEAN	1	44	44.44	179.08	< 2e-16	***
SELT_MEAN	1	3	2.64	10.66	0.0011	**
ZINCT_MEAN:SELT_MEAN	1	16	16.30	65.68	5.32e-16	***
Residuals	248963	61778	0.25			

signif. codes:	0	'***'	0.001	'**'	0.01	'*'
				0.05	'.'	0.1
					' '	1

The p-value is 5.32e-16 which is less than 0.05. So there are correlation between DIABBC and interaction of Se and Zn.

Detail of Pearson's chi-squared test for independence between DIABBC and Sex

stat is 5.122081540143781

p value is 0.013612595748290643

Dependent (reject H0)

Therefore, these two variables are dependendent

Detail of Pearson's chi-squared test for independence between DIABBC and RedMeat_indicator

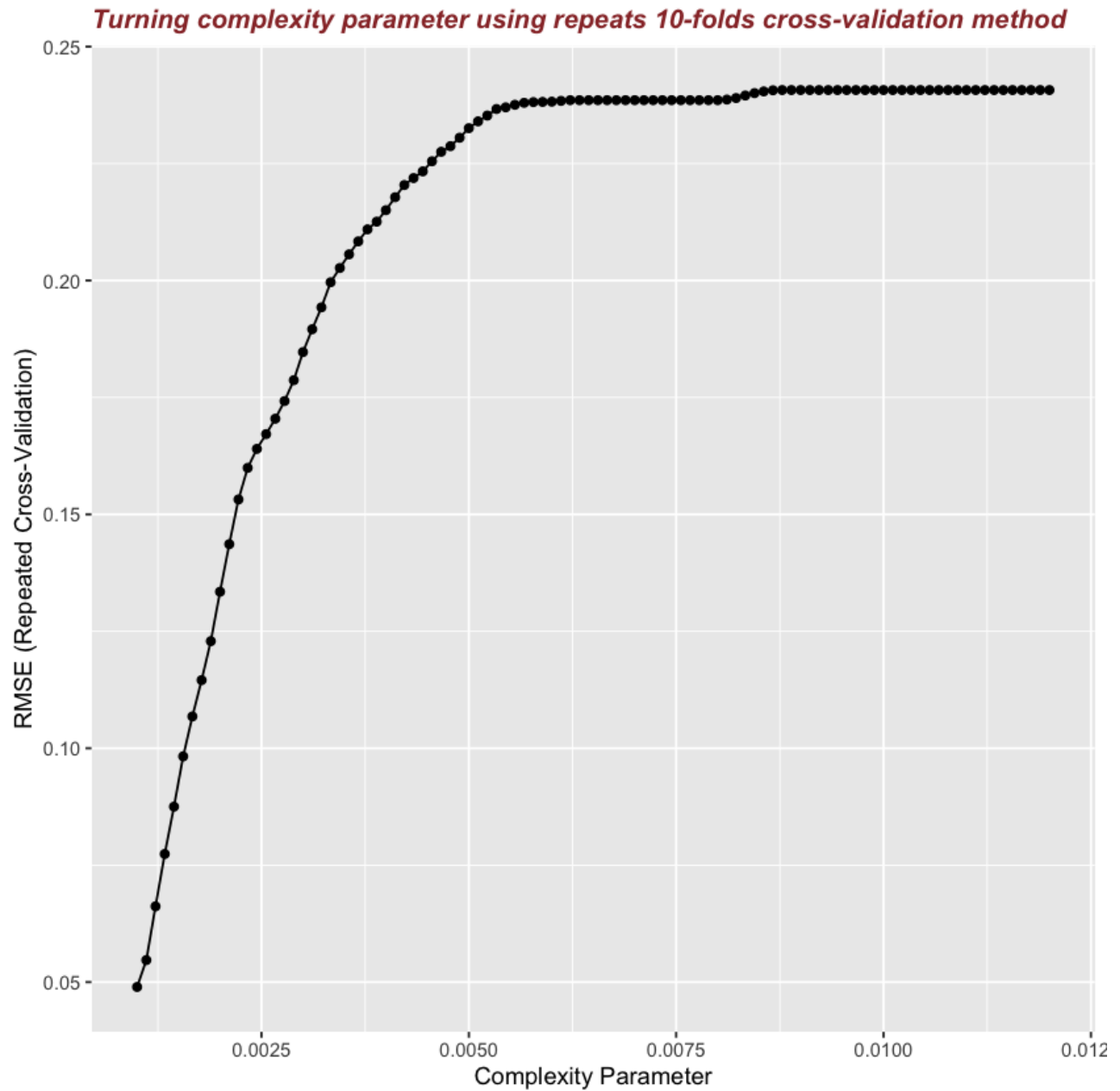
stat is 2.2448880276781025

p value is 0.086664549232786

Dependent (reject H0)

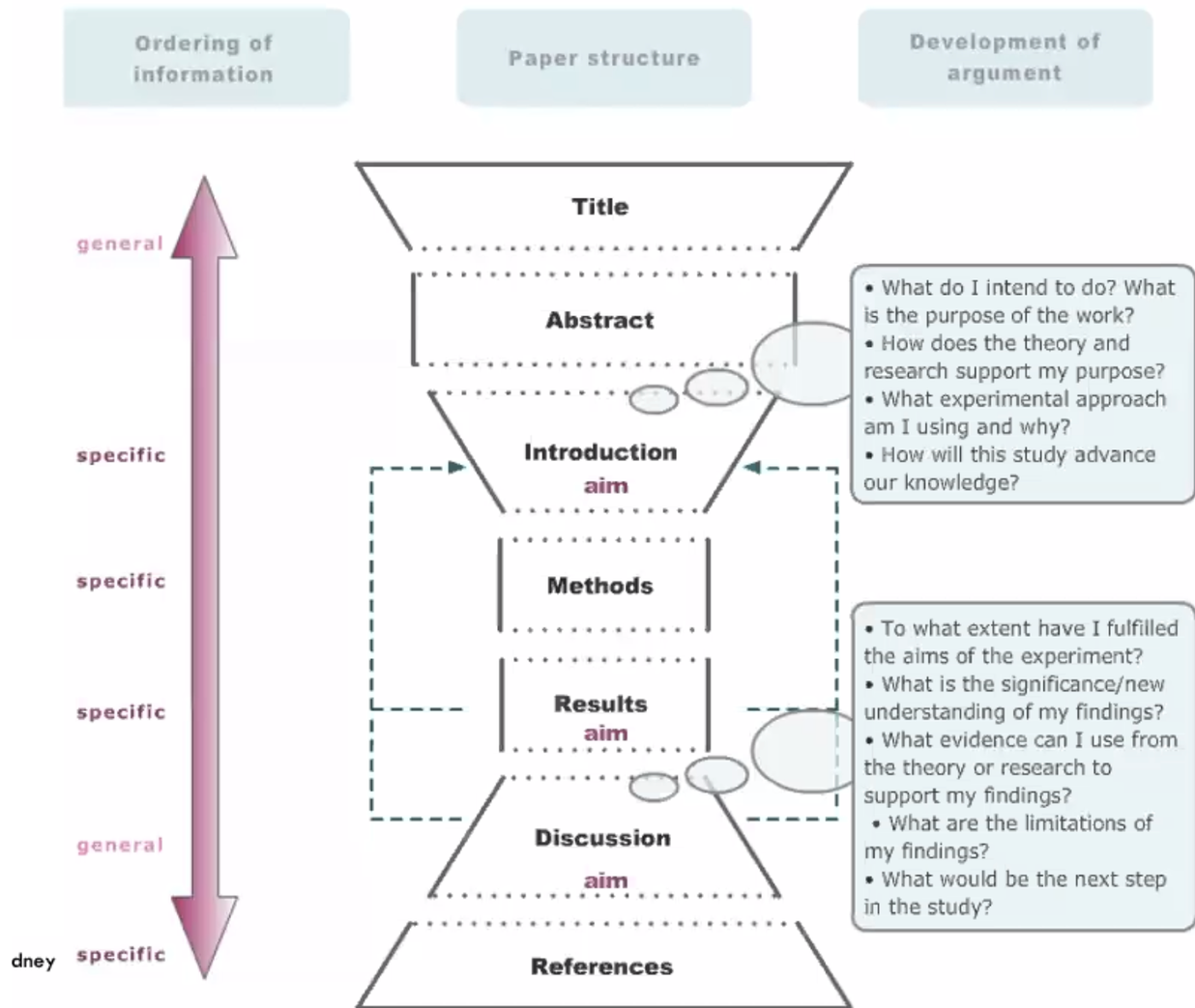
Therefore, these two variables are dependendent

The plot of turning complexity parameter using repeats 10-folds cross-validation method for decision tree



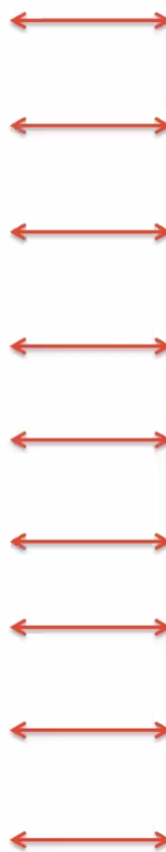
Detail of T test to check the mean Se intake

Report structure



Your article: iScience

Title
Abstract (Summary)
Introduction
Results
Discussion
Limitations
Methods
References
Supplementary material



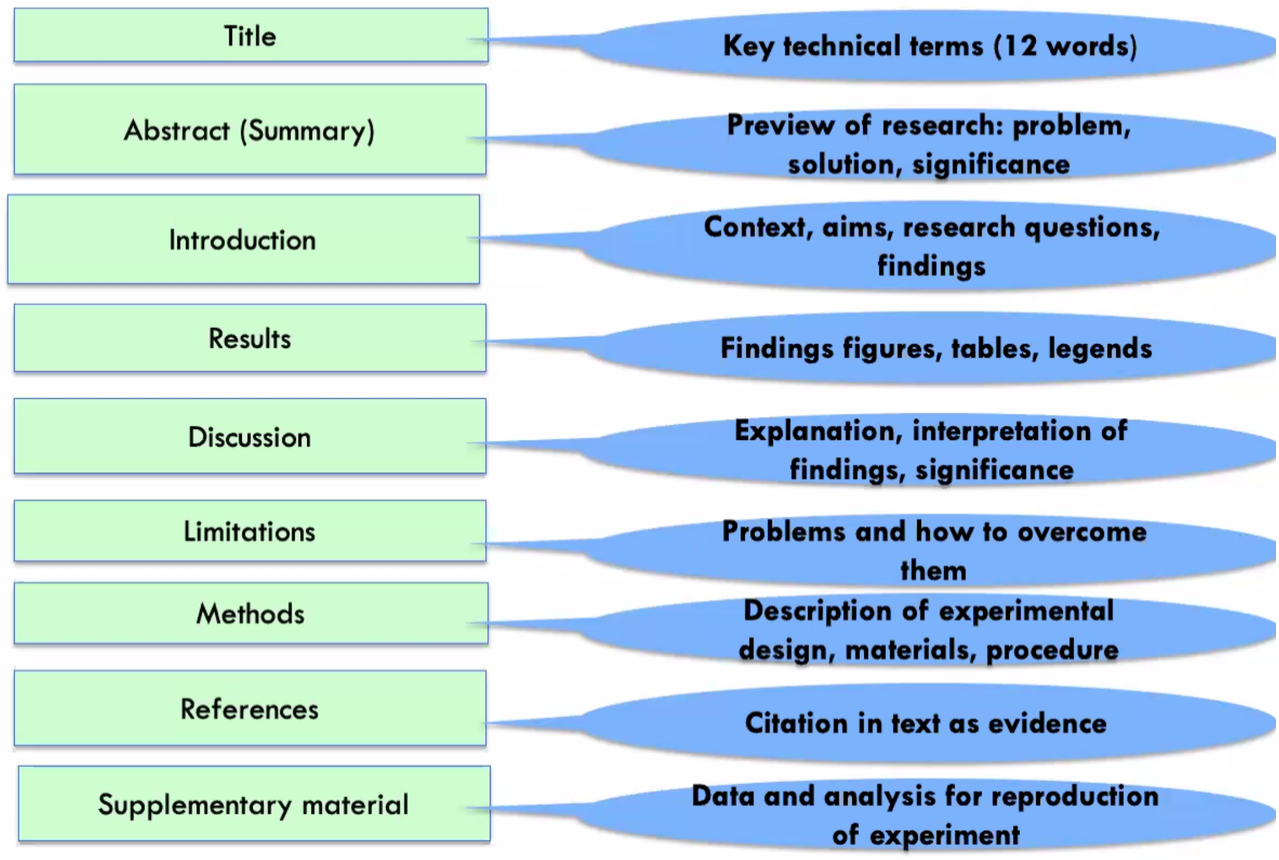
A flow of :

- **argument**
- **evidence and**
- **evaluation**

Linking all sections

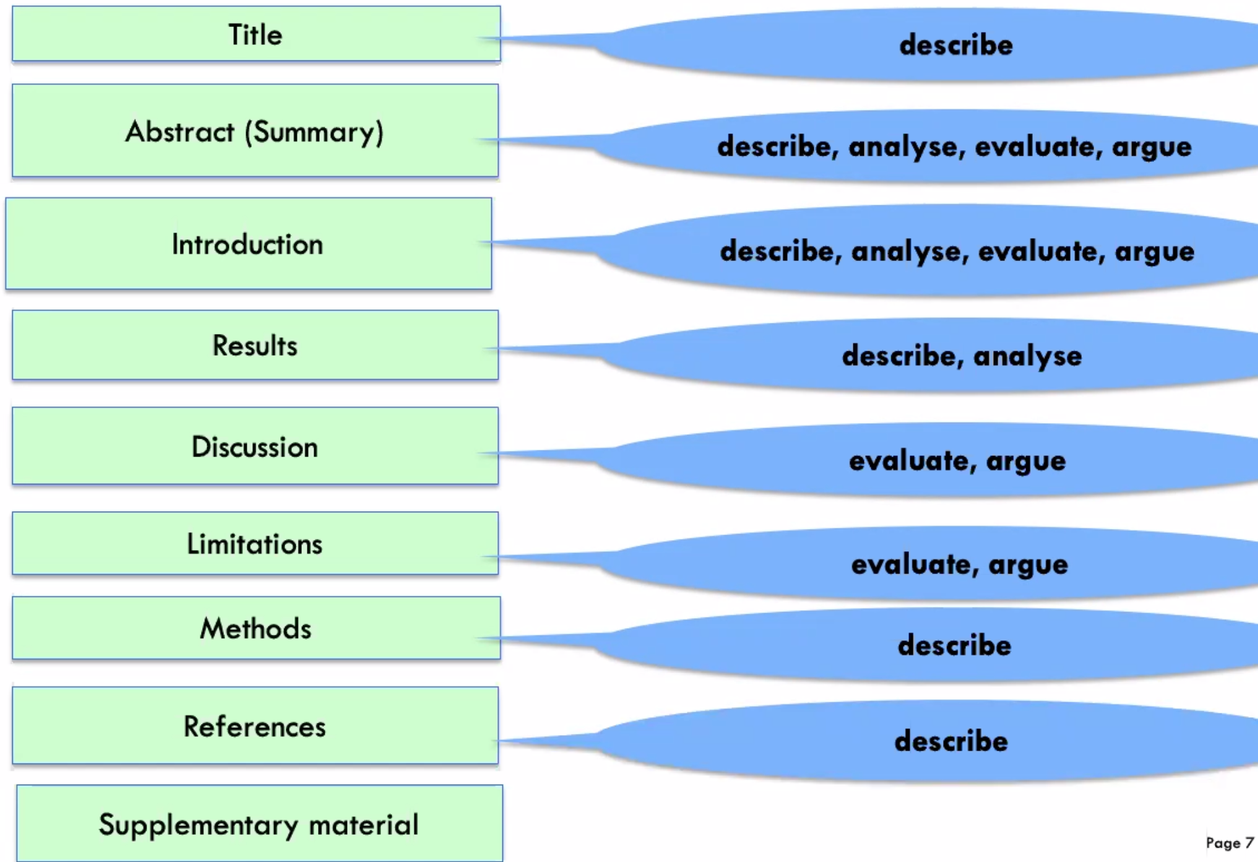
Your article: iScience

Talk



Supplementary material is used for data analysis related to the research topic, and not included in the 4000 words.

Your article: iScience



What is the reviewer seeking?

