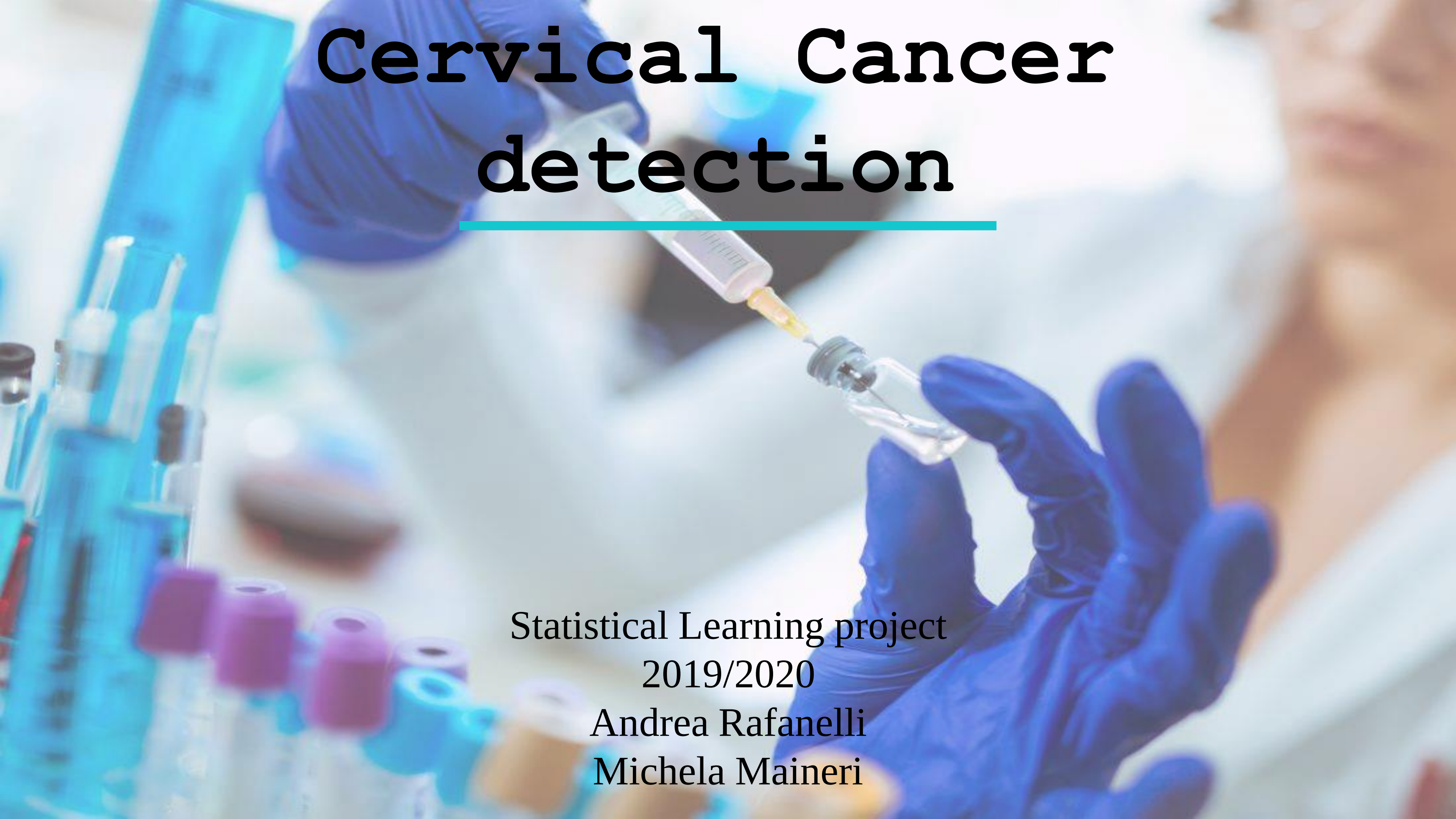


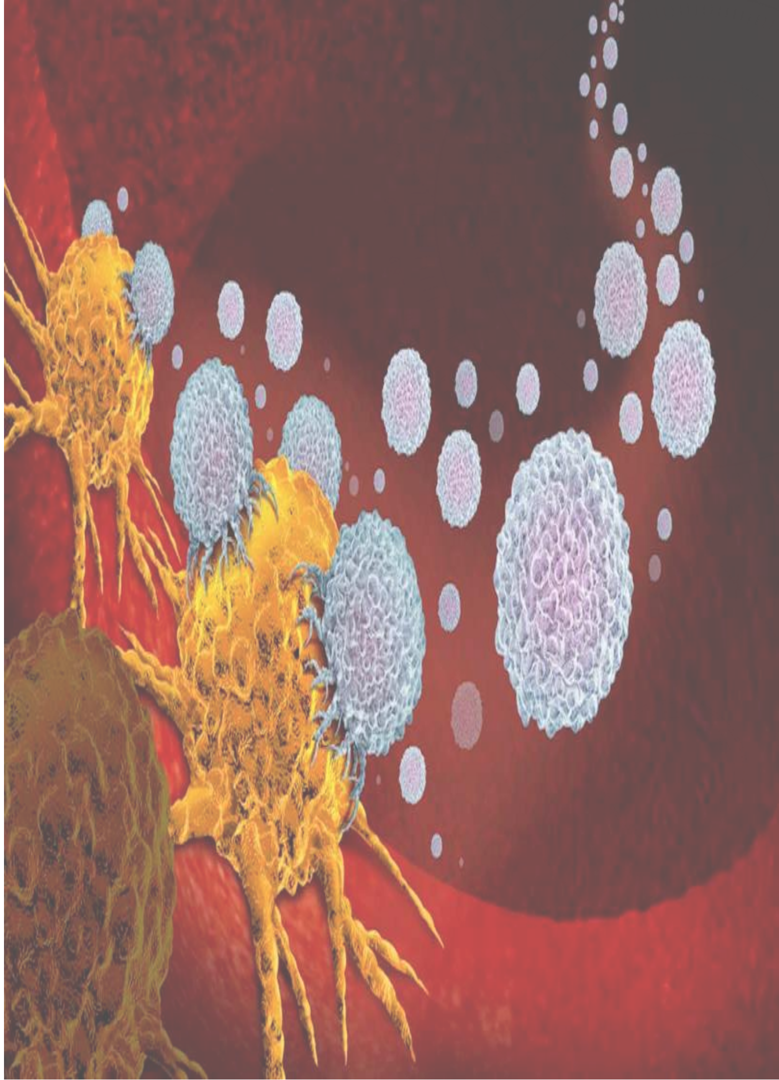
Cervical Cancer detection

Statistical Learning project
2019/2020
Andrea Rafanelli
Michela Maineri

Our goal:

- Cancer mortality can be reduced if cases are detected and treated early
- *Is it possible to predict it ?*
- *Which factors causes cervical cancer?*

- *Hospital agency meeting*



Agenda :

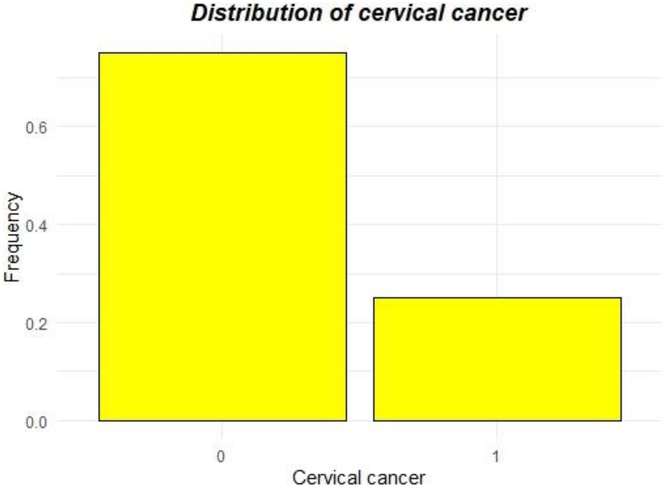
1. ***Data cleaning***
2. *Exploratory Descriptive Analysis*
3. *Variables importance*
4. *Parametric method*
5. *Non-parametric method*

Our dataset

- Data collection from surveys: demographic information, habits and historical medical records

- **Cervical Cancer** target variable: **25%** of the patients diagnosticated with cervical cancer: Healthy patients 660 out of 880

- *Uci Machine Learning Repository*:
  880 observations, 25 variables



Predictors

QUANTITATIVE

1. *Age*
2. *Number of sexual partners*
3. *Years old when first having sex*
4. *Years of smoke*
5. *Years of hormonal contraceptive*
6. *Years of intra-uterine disease*
7. *Number of sexually trasmitted diseases*
8. *Number of diagnosis*
9. *Number of pregnancies*

QUALITATIVE

1. *Smoke (0-1)*
2. *Hormonal contraceptive (0-1)*
3. *Intra-uterine dispositive (0-1)*
4. *Sexually trasmitted disease (0-1)*
5. *Condylomatosis (0-1)*
6. *Syphilis (0-1)*
7. *Genital herpes (0-1)*
8. *Molluscum contagious (0-1)*
9. *HIV (0-1)*
10. *Hepatitis B (0-1)*
11. *HPV (0-1)*
12. *DX test (0-1)*
13. *Cancer (0-1)*
14. *Schiller test (0-1)*
15. *Cervical cytology test (0-1)*
16. *Hinselmann test (0-1)*

Data cleaning (1)

Possible distortive variables:

- **Hinselmann test:** to detect **cervical** tumors when they were relatively small.
- **Schiller test:** iodine solution is applied to the cervix in order to diagnose cervical cancer.
- **Cervical cytology:** to detect abnormal or potentially abnormal cells from the uterine **cervix**

Data cleaning (2)

Missing values:

Replace with median:

Number of sexual partner: 28
Years of first sex: 7
Number of pregnancies: 56
Years of smoke: 13
Years of hormonal contraceptive: 108
Years of IUD: 117

Replace with mode:

Smoke: 13
Hormonal contraceptive: 108
IUD: 117
HIV: 105

Agenda :

1. *Data cleaning*

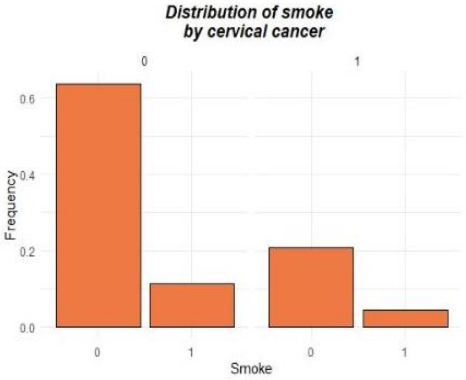
2. ***Exploratory Descriptive Analysis***

3. *Variables importance*

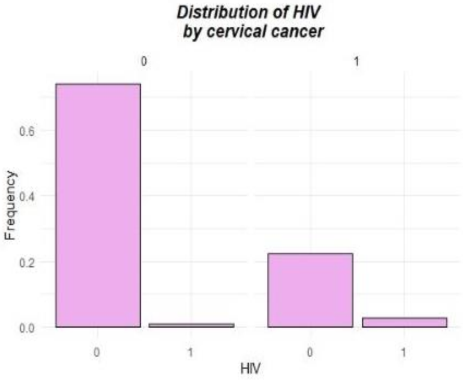
4. *Parametric method*

5. *Non-parametric method*

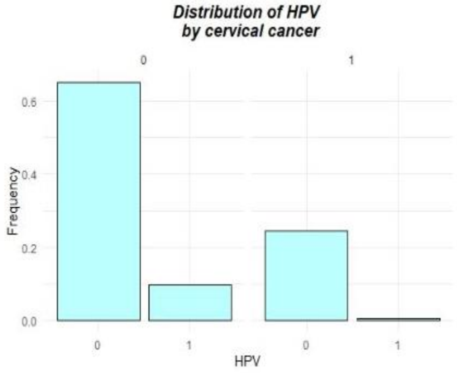
EDA: Qualitative predictors



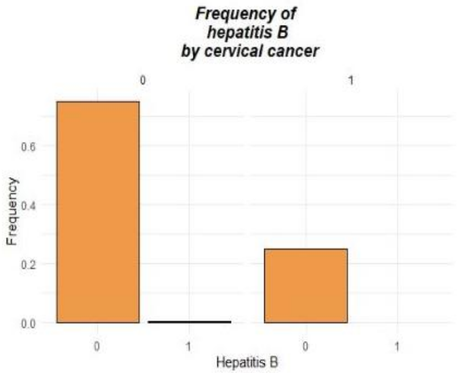
Distribution of smoke by cervical cancer



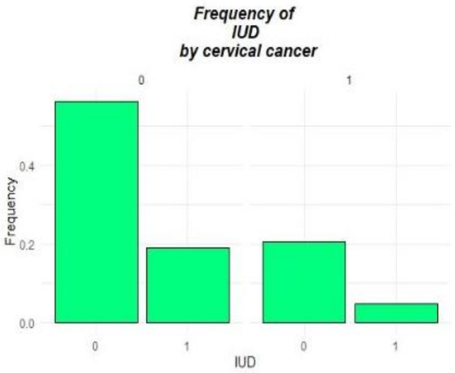
Distribution of HIV by cervical cancer



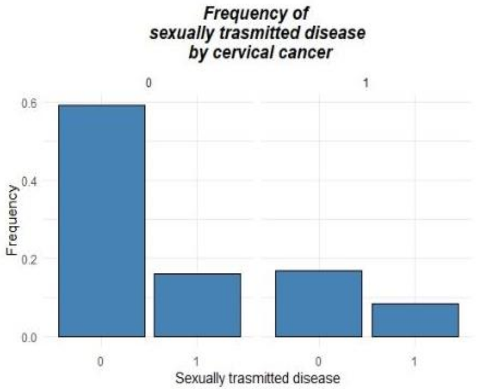
Distribution of HPV by cervical cancer



Frequency of hepatitis B by cervical cancer

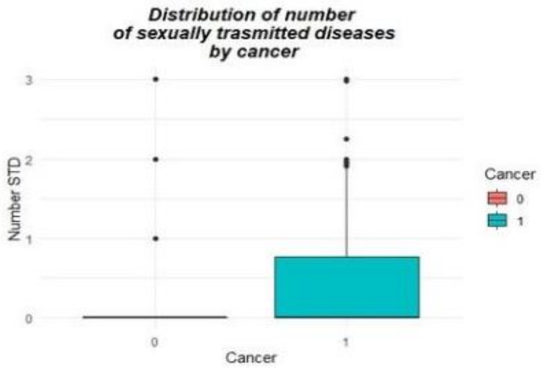


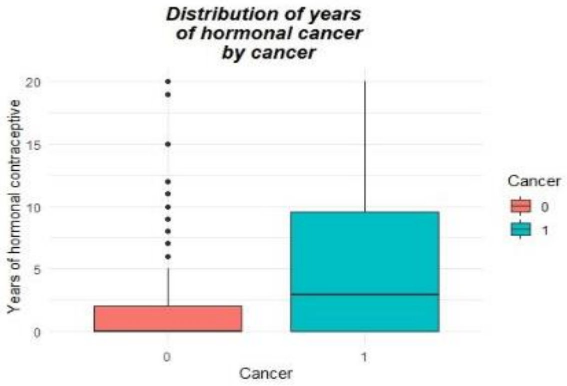
Frequency of IUD by cervical cancer

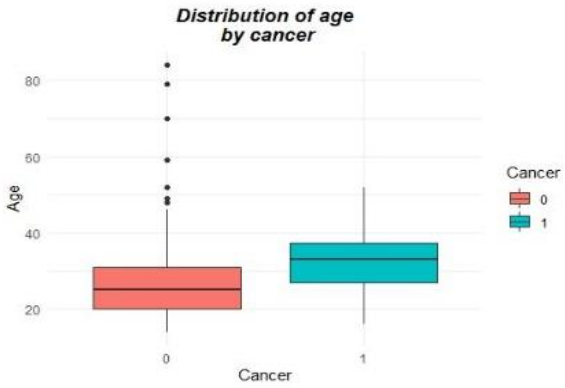


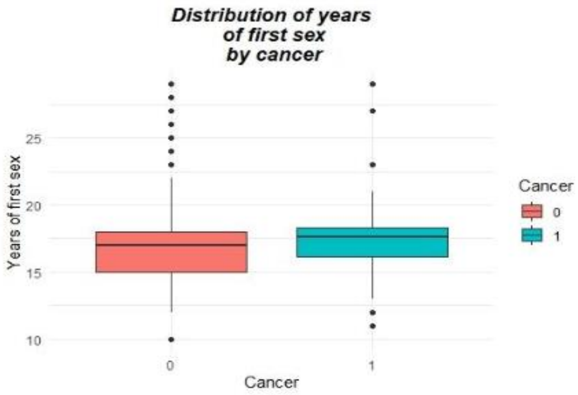
Frequency of sexually trasmitted disease by cervical cancer

EDA: Quantitative predictors



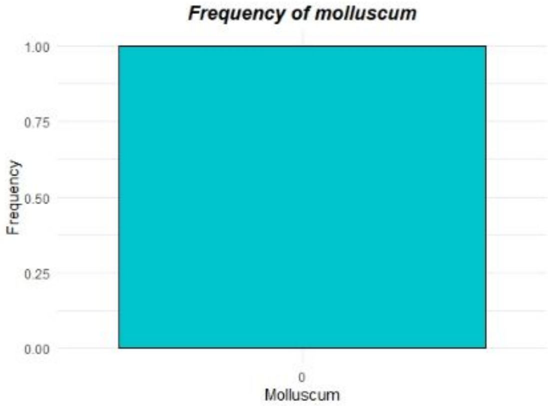


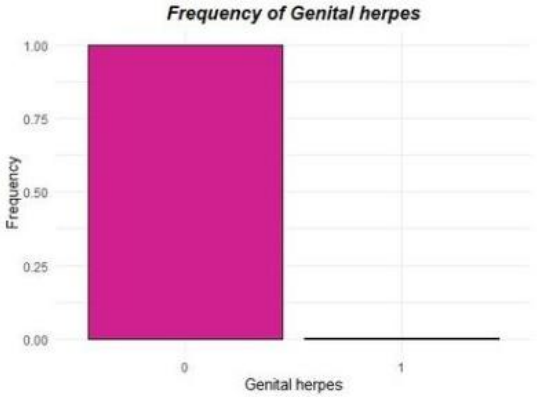


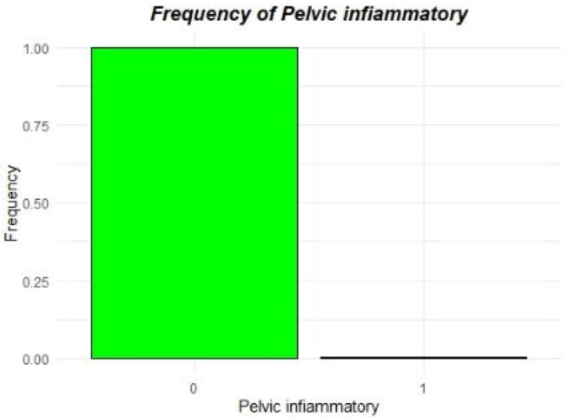


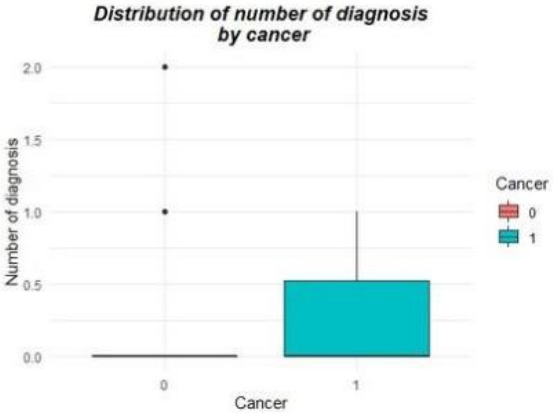
EDA: Not explicative variables

We decide to avoid them:



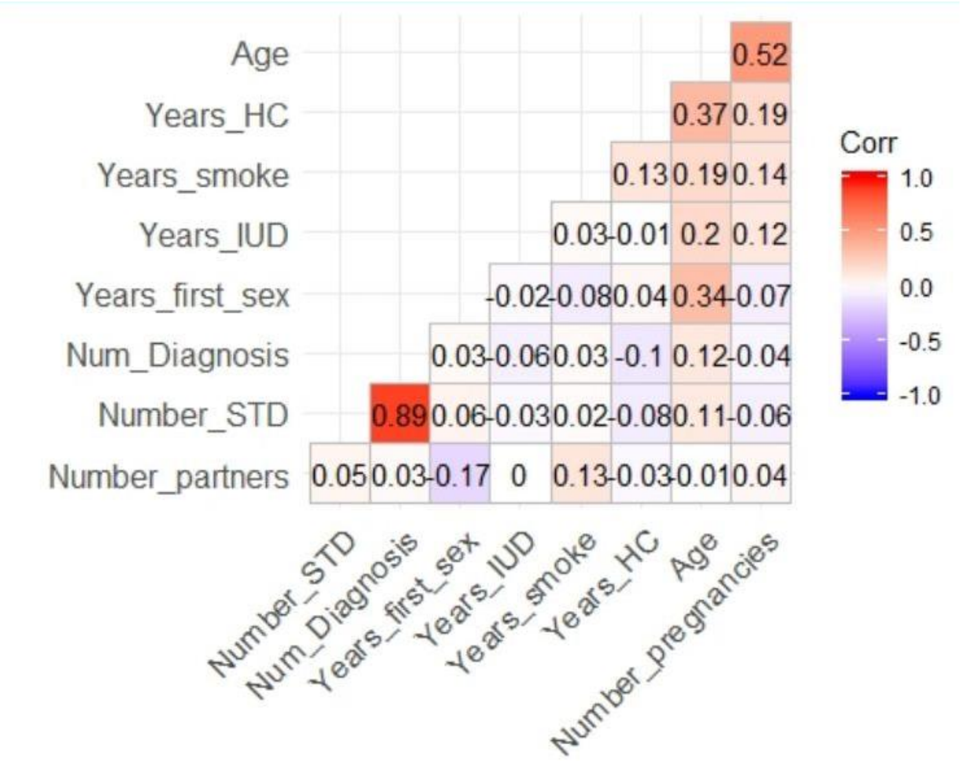






EDA: Correlations

- *Not relevant correlations*
- *Number of sexually trasmitted diseases correlated with number of cancer diagnosis*
- *Age with number of pregnancies*



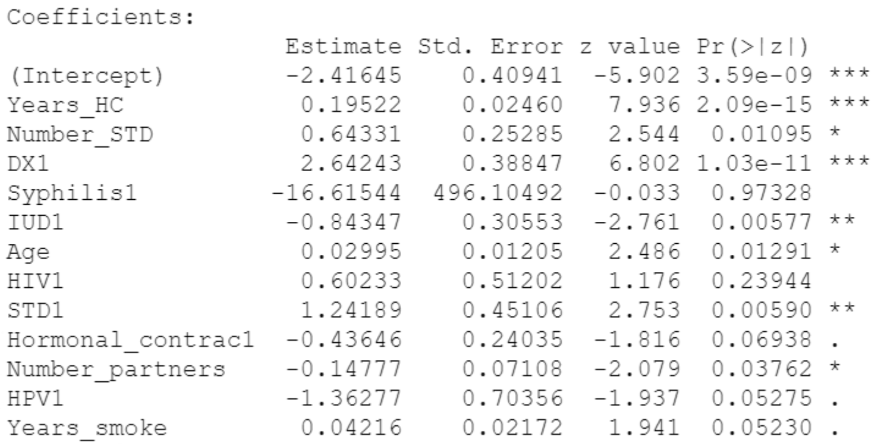
Agenda :

1. *Data cleaning*
2. *Exploratory Descriptive Analysis*
3. ***Variables importance***
4. *Parametric method*
5. *Non-parametric method*

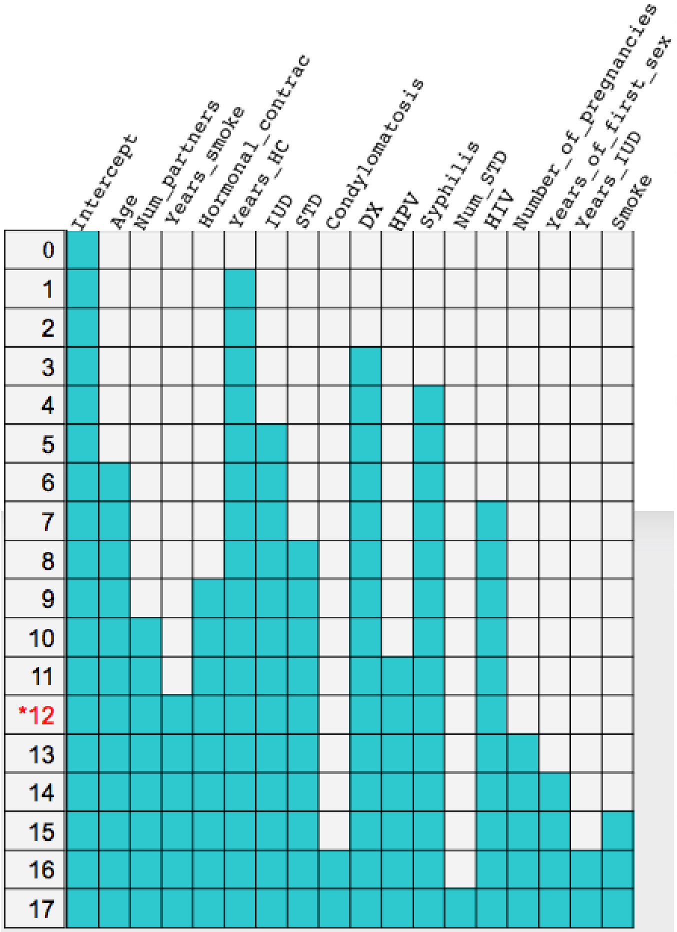
Exploring variables importance(1)

Variables selection:

- ❑ *Try with different method: forward, backward, bothwise*
- ❑ *Best is the forward with AIC: 727.11*



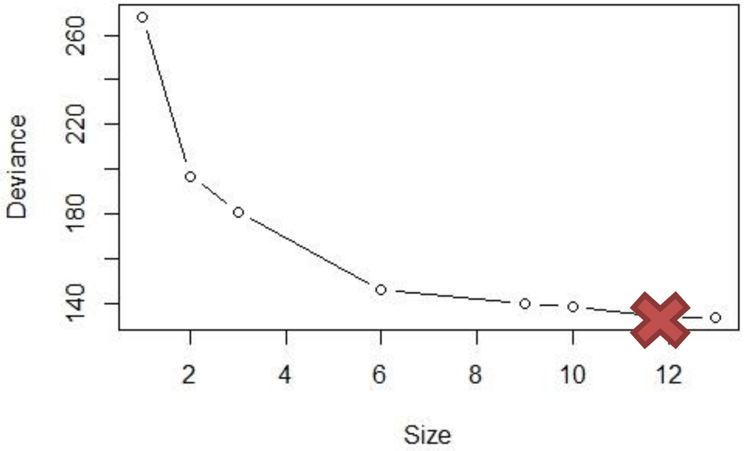
```
Coefficients:
                    Estimate Std. Error z value Pr(>|z|)    
(Intercept)         -2.41645    0.40941  -5.902 3.59e-09 ***
Years_HC             0.19522    0.02460   7.936 2.09e-15 ***
Number_STD           0.64331    0.25285   2.544  0.01095 *  
DX1                  2.64243    0.38847   6.802 1.03e-11 ***
Syphilis1          -16.61544  496.10492  -0.033  0.97328    
IUD1                -0.84347    0.30553  -2.761  0.00577 ** 
Age                  0.02995    0.01205   2.486  0.01291 *  
HIV1                 0.60233    0.51202   1.176  0.23944    
STD1                 1.24189    0.45106   2.753  0.00590 ** 
Hormonal_contrac1   -0.43646    0.24035  -1.816  0.06938 .  
Number_partners     -0.14777    0.07108  -2.079  0.03762 *  
HPV1                -1.36277    0.70356  -1.937  0.05275 .  
Years_smoke          0.04216    0.02172   1.941  0.05230 .  
```



	Intercept	Age	Num_partners	Years_smoke	Hormonal_contrac	Years_HC	IUD	STD	Condylomatosis	DX	HPV	Syphilis	Num_STD	HIV	Number_of_pregnancies	Years_of_first_sex	Years_IUD	Smoke
0	■																	
1	■					■												
2	■					■												
3	■					■				■								
4	■					■				■		■						
5	■					■	■			■		■						
6	■	■				■	■			■		■						
7	■	■				■	■			■		■		■				
8	■	■				■	■	■		■		■		■				
9	■	■			■	■	■	■		■		■		■				
10	■	■	■		■	■	■	■		■		■		■				
11	■	■	■		■	■	■	■		■	■	■		■				
*12	■	■	■	■	■	■	■	■		■	■	■		■				
13	■	■	■	■	■	■	■	■		■	■	■		■	■			
14	■	■	■	■	■	■	■	■		■	■	■		■	■	■		
15	■	■	■	■	■	■	■	■		■	■	■		■	■	■		■
16	■	■	■	■	■	■	■	■	■	■	■	■		■	■	■	■	■
17	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■

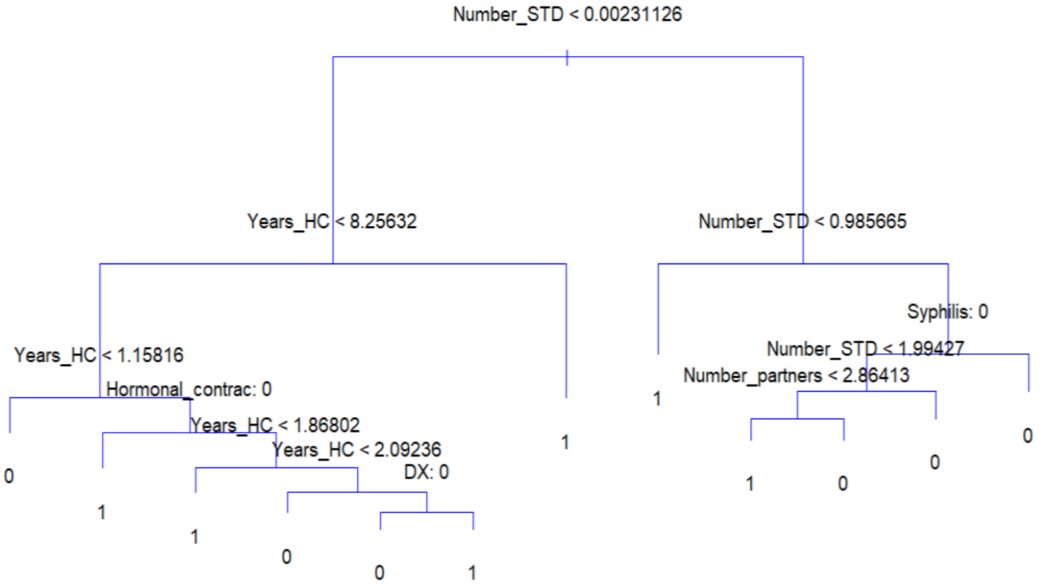
Exploring variables importance(2)

Tree:



Cross Validation to select the best split: 12

- Number of sexually trasmitted diseases;
- Years of hormonal contraceptive;
- Number of partners;
- Syphilis
- DX



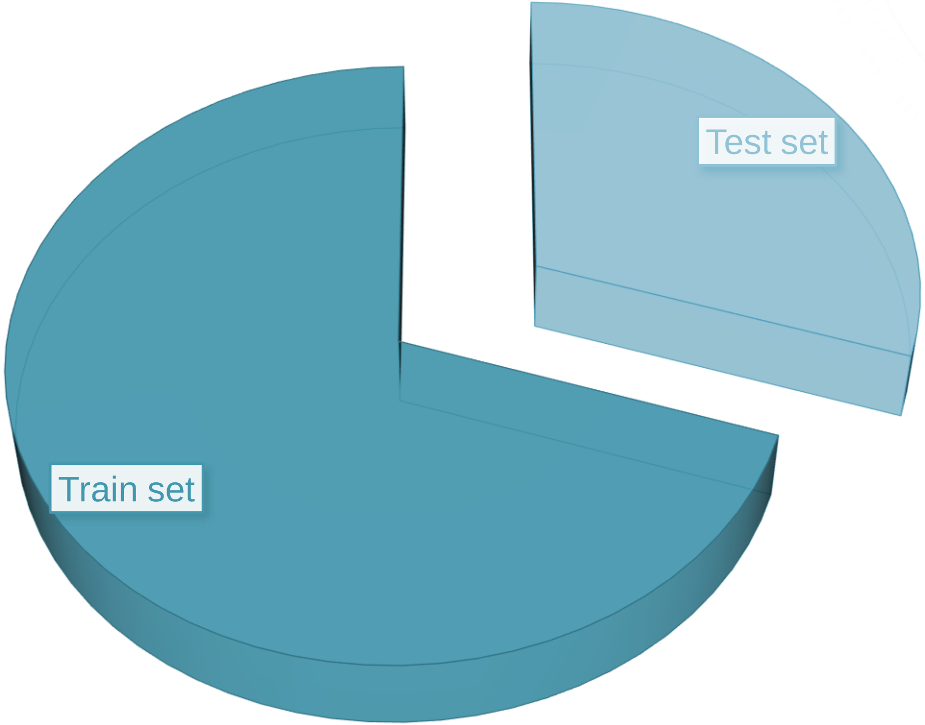
Misclassification error rate 10.68%

Agenda :

1. *Data cleaning*
2. *Exploratory Descriptive Analysis*
3. *Variables importance*
4. ***Parametric method***
5. *Non-parametric method*

TRAIN AND TEST SET

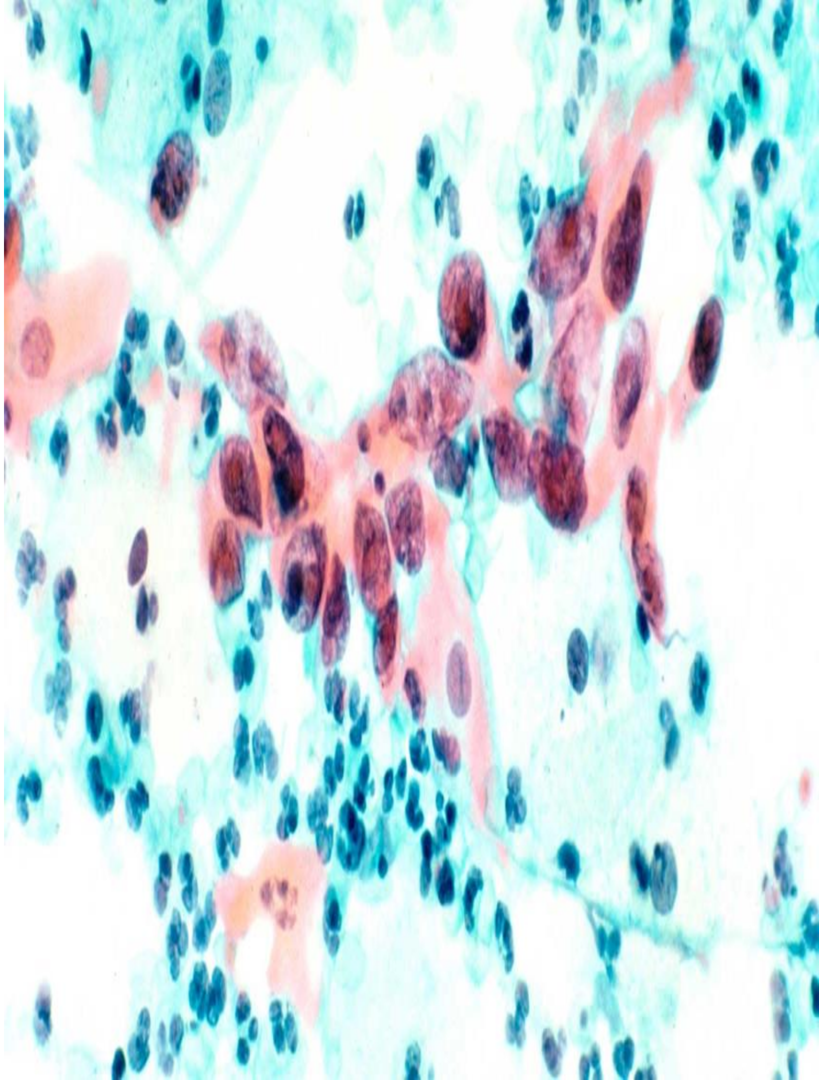
- *70% training set*
- *30% test set*



GLM (1)

Based on forward selection:

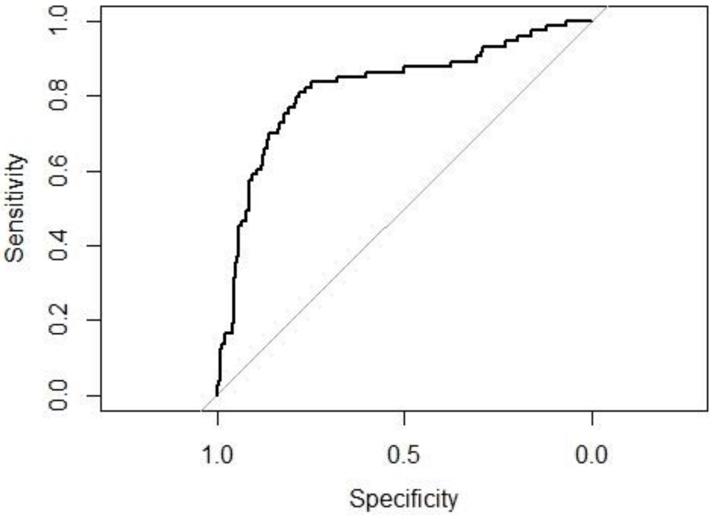
	Exp (coeff.)	*Significativity*
Intercept	*0.156705*	***
Years HC	*1.25446*	***
Number STD	*2.240323*	*
DX1	*18.88446*	***
Syphilis1	*5.623143 e-08*	
IUD1	*0.6178992*	
Age	*1.016901*	
HIV1	*3.494883*	
STD1	*3.386815*	*
HC	*0.4610955*	*
Number_partners	*0.8085199*	*
HPV1	*0.4661583*	
Years_smoke	*1.0365*	



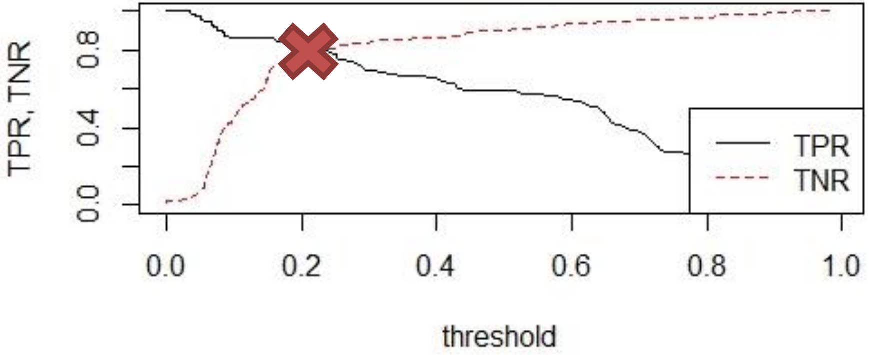
GLM(2)

Choice of the threshold:

AUC: 82.53%



THRESHOLD: 0.1832281



	False	***True***
0	*172*	*49*
1	*13*	*60*

- Specificity: 77.40%
- Sensitivity: 83.56%
- **Accuracy: 79%**

Agenda :

1. *Data cleaning*

2. *Exploratory Descriptive Analysis*

3. *Variables importance*

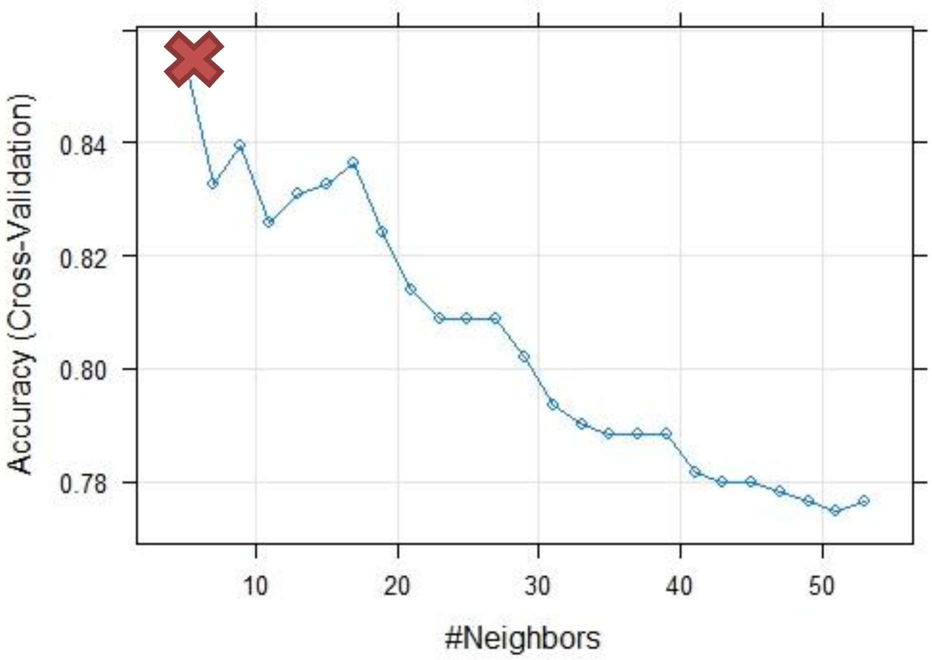
4. *Parametric method*

5. ***Non-parametric method***

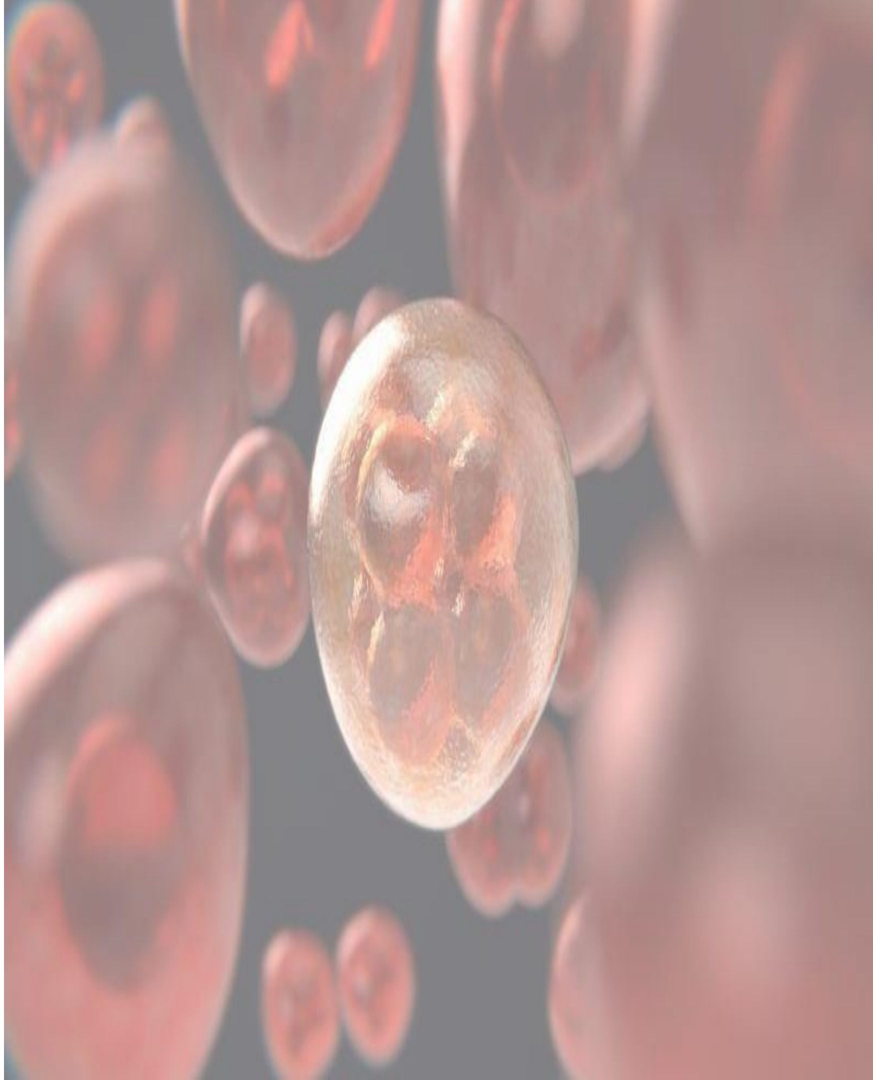
KNN

10-fold cv for the best K:

K=5



Accuracy 85%



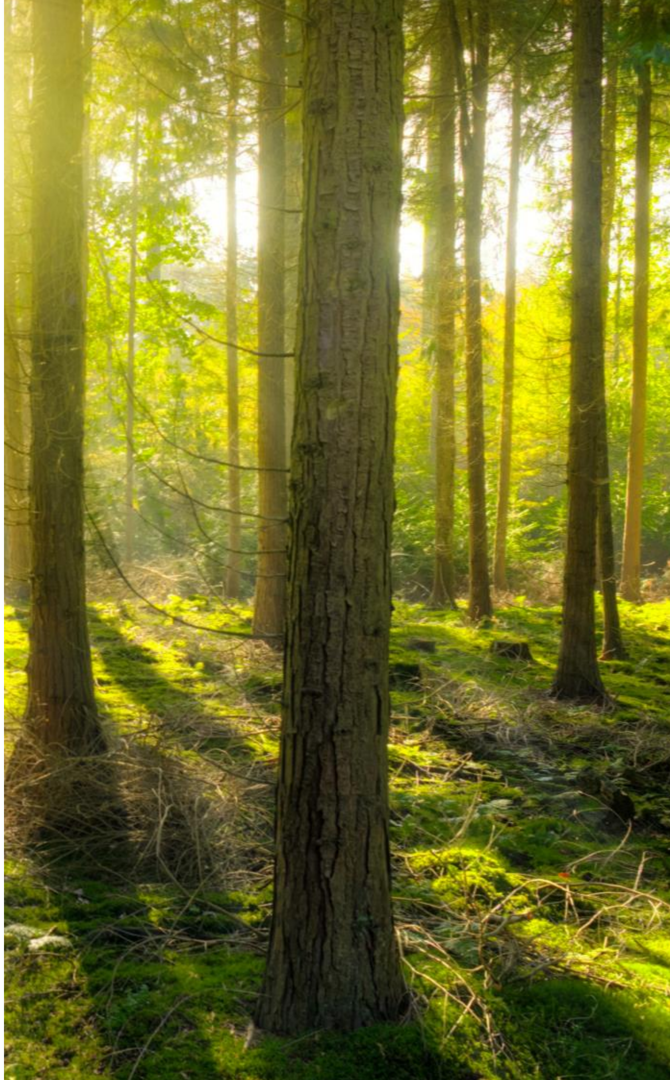
RANDOM FOREST (1)

- *Aggregating many decision trees improves the predictive performance*

- *Bagging reduce variance in trees: m=p*

- *Decorrelating the trees*

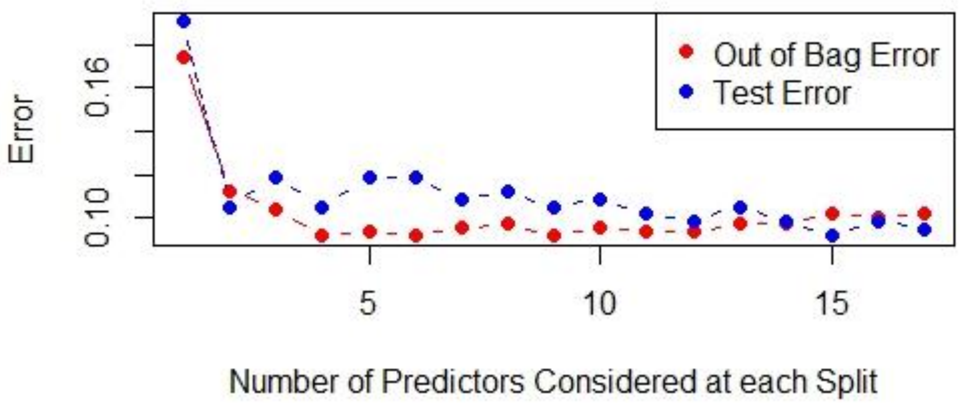
considering only a subset m of the predictors m<p



RANDOM FOREST (2)

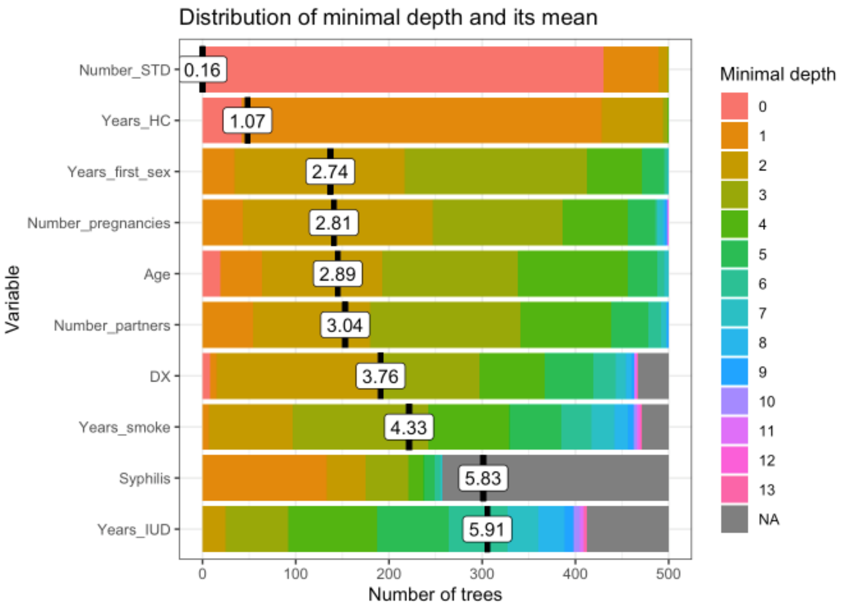
How to select the best number of predictors at each split?

Try with all possible values of "mTry":



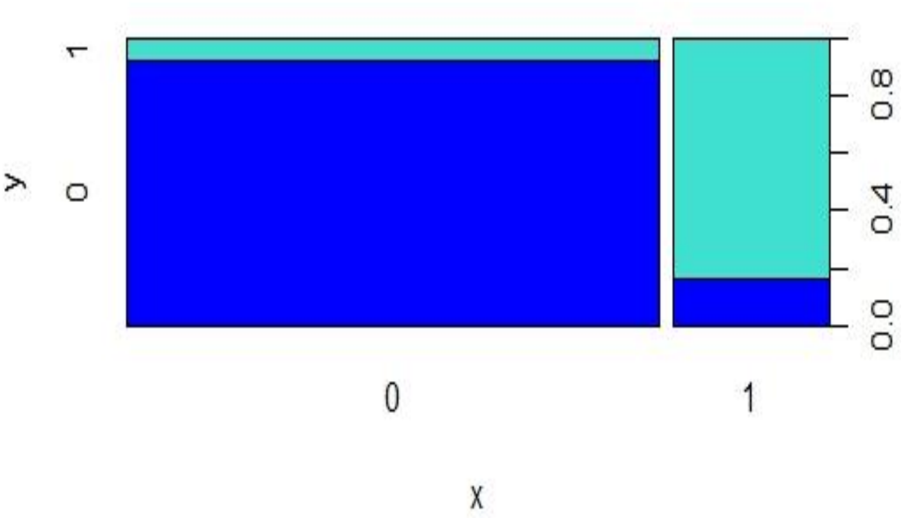
The best is m=15

- *Out of bag error (OOB): 9.22%*
- *Test error: 9.18%*

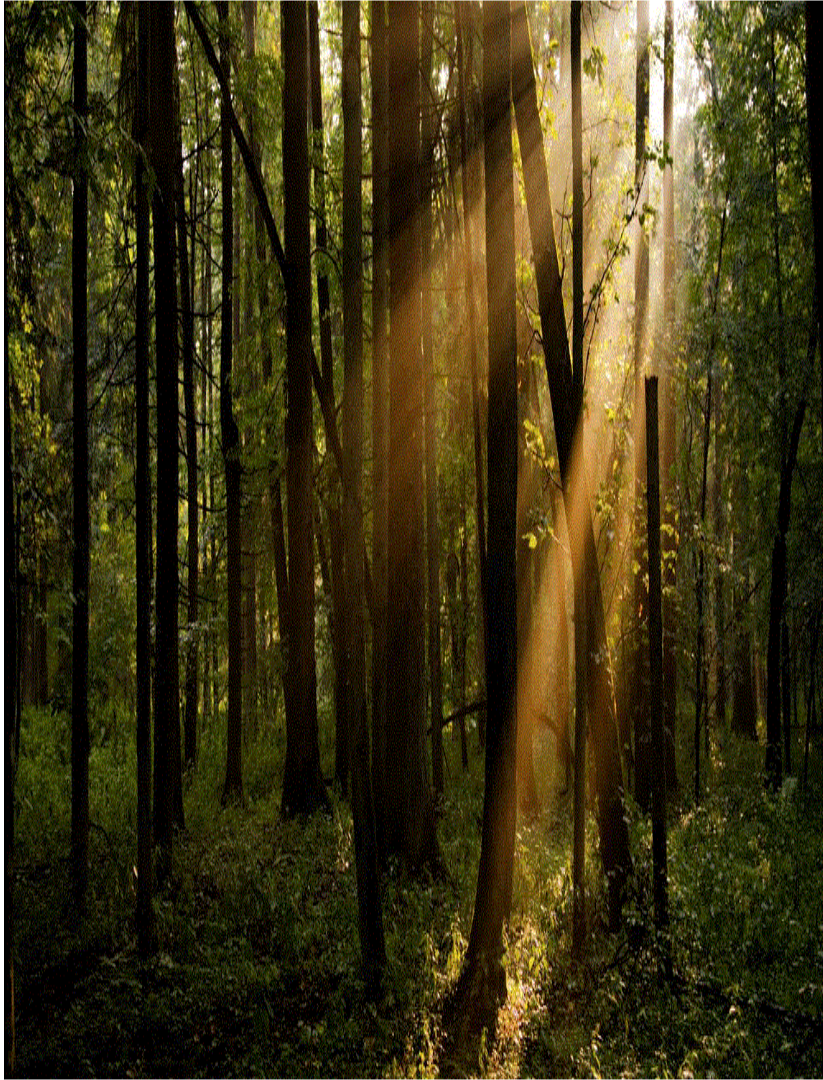


RANDOM FOREST (3)

Predictions m=15



Accuracy: 91%



CONCLUSION AND IMPROVEMENTS

1. Cancer most influencing variables: ***number of sexually transmitted diseases, years of hormonal contraceptive***

2. Best prediction with random forest

GLM	***KNN***	***RANDOM FOREST***
79 %	*85 %*	*91 %*

3. Try different methods for data cleaning

4. Cross validation for random forest and GLM threshold: to improve ‘1’ prediction

5. Try Neural Network

REFERENCES

- Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, *"An Introduction to Statistical Learning" ,* Springer Science+Business Media New York 2013

- Muhammed Fahri Unlersen1, Kadir Sabanci2, Muciz Özcan1 *,"Determining Cervical Cancer Possibility by Using Machine Learning Methods" ,* International Journal of Latest Research in Engineering and Technology, December 2017