Cervical Cancer Prediction

COSC 2669 Legal, Ethical and Policy Issues in Machine Learning

Charles Galea (s3688570)

August 2018

0.1 Load packages and set random seed value

```
Sys.setenv(JAVA_HOME='C:\\Program Files\\Java\\jre-9.0.4')
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(readr)
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------
```

```
## v ggplot2 3.0.0.9000     v purrr   0.2.5
## v tibble  1.4.2          v stringr 1.3.1
## v tidyr   0.8.1          v forcats 0.3.0
```

```
## -- Conflicts ------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(mlr)
```

```
## Loading required package: ParamHelpers
```

```
library(tidyverse) # for ggplot and data wrangly
library(rJava)
library(FSelector)
library(ggplot2)
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
library(gmodels)
library(GGally)
```

```
##
```

```
## Attaching package: 'GGally'

## The following object is masked from 'package:dplyr':
##
##     nasa
```

```
library(cowplot)
```

```
##
##
## ************************************************************

## Note: cowplot does not change the default ggplot2 theme

## anymore. To recover the previous behavior, execute:
##   theme_set(theme_cowplot())

## ************************************************************
```

```
library(tidyr)
library(magrittr)
```

```
##
## Attaching package: 'magrittr'

## The following object is masked from 'package:purrr':
##
##     set_names

## The following object is masked from 'package:tidyr':
##
##     extract
```

```
library(moments)
library(purrr)
library(data.table)
```

```
##
## Attaching package: 'data.table'

## The following object is masked from 'package:purrr':
##
##     transpose

## The following objects are masked from 'package:dplyr':
##
##     between, first, last
```

```
library(latex2exp)
library(caret)
```

```
## Loading required package: lattice

##
## Attaching package: 'caret'

## The following object is masked from 'package:mlr':
##
##     train

## The following object is masked from 'package:purrr':
##
```

```
##     lift
```

```
library(robustHD)
```

```
## Loading required package: perry
```

```
## Loading required package: parallel
```

```
## Loading required package: robustbase
```

```
library(spFSR)
```

```
## Loading required package: parallelMap
```

```
## Loading required package: tictoc
```

```
library(rjson)
library(party)
```

```
## Loading required package: grid
```

```
## Loading required package: mvtnorm
```

```
## Loading required package: modeltools
```

```
## Loading required package: stats4
```

```
##
## Attaching package: 'modeltools'
```

```
## The following object is masked from 'package:rJava':
##
##     clone
```

```
## Loading required package: strucchange
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
## Loading required package: sandwich
```

```
##
## Attaching package: 'strucchange'
```

```
## The following object is masked from 'package:stringr':
##
##     boundary
```

```
library(knitr)
library(kableExtra)
library(stringr)
library(mlbench)
library(e1071)
```

```
##
## Attaching package: 'e1071'
```

```
## The following objects are masked from 'package:moments':
##
```

```
##     kurtosis, moment, skewness
## The following object is masked from 'package:mlr':
##
##     impute
```

```
library(MASS)
```

```
##
## Attaching package: 'MASS'
## The following object is masked from 'package:dplyr':
##
##     select
```

```
set.seed(999)
```

1 Background

There are approximately 500,000 new cases of invasive cervical cancer diagnosed globally and kills about 300,000 women each year. Although this nnumber is steadily decreasing with recent advances in screening technologies.

1.1 Age

Cervical cancer diagnoses occur predominately in women between the ages of 35 and 54 years (50%) with 20% for those aged over 65 and 15% for younger women (20-30 years). Women youger than 20 rarely develop cervical cancer however they can become infected with the human papilloma virus (HPV) which increases their chances of developing this disease later in life.

1.2 Sexual Activity

The human papilloma virus (HPV), which is sexually transmitted, is a major risk factor in developing cervical cancer. Studeies have shown that women wo have taken oral contraceptives for longer than 5-10 years appear to have a higher risk of HPV infection. Other STDs have also been associated with a high risk of developing cervical cancer.

1.3 Oral Contraceptives

Previous studies have suggested a strong correlation between the long term use of oral contraceptives and cervical cance3r.

1.4 Pregnancies

Women who have had a number of children have also been shown to have an increased risk of developing cervical cancer. Particularly for those women infected with HPV.

Table 1: Table 1. Cervical Cancer Dataset

Age	Number of sexual partners	First sexual intercourse	Num of pregnancies	Smokes	Smokes (years)	Smokes (packs/year)	Hormonal Contraceptives	Hormonal Contraceptives (years)	IUD	IUD (years)	STDs	STDs (number)	STDs:condylomatosis	STDs:cervical condylomatosis	STDs:vaginal condylomatosis	STDs:vulvo-perineal condylomatosis	STDs:syphilis	STDs:pelvic inflammatory disease	STDs:genital herpes	STDs:molluscum contagiosum	STDs:AIDS	STDs:HIV	STDs:Hepatitis B	STDs:HPV	STDs: Number of diagnosis	STDs: Time since first diagnosis	STDs: Time since last diagnosis	Dx:Cancer	Dx:CIN	Dx:HPV	Dx	Hinselmann	Schiller	Citology	Biopsy
18	4	15	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	NA	NA	0	0	0	0	0	0	0	0
15	1	14	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	NA	NA	0	0	0	0	0	0	0	0
34	1	NA	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	NA	NA	0	0	0	0	0	0	0	0
52	5	16	4	1	37	37	1	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	NA	NA	1	0	1	0	0	0	0	0
46	3	21	4	0	0	0	1	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	NA	NA	0	0	0	0	0	0	0	0
42	3	23	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	NA	NA	0	0	0	0	0	0	0	0

1.5 Smoking

Smoking has been shown to be associated with a higher risk for precancerous changes (dysplasia) in the cervix and subsequent prgression to cancer, especially for those infected with HPV.

1.6 Preparation

Read in dataset.

```
## Parsed with column specification:
## cols(
##   .default = col_integer(),
##   Smokes_years = col_double(),
##   Smokes_packs_year = col_double(),
##   Hormonal_Contraceptives_years = col_double(),
##   IUD_years = col_double()
## )
```

```
## See spec(...) for full column specifications.
```

1.7 Summary statistics for data set

```
dim(dataCancer)
```

```
## [1] 858  36
```

1.8 Missing values

Removed columns for STDs_cervical_condylomatosis, STDs_vaginal_condylomatosis, STDs_pelvic_inflam_disease, STDs_genital_herpes, STDs_molluscum_contagiosum, STDs_Hepatitis_B, STDs_HPV and STD_AIDS since there 4 or less patient results for these features. Also removed features STDs_Time_since_first_diagnosis and STDs_Time_since_last_diagnosis which contained greater than 60% missing values (787 of 858).

```
dataCancer <- subset(dataCancer, select = -STDs_cervical_condylomatosis)
dataCancer <- subset(dataCancer, select = -STDs_vaginal_condylomatosis)
dataCancer <- subset(dataCancer, select = -STDs_pelvic_inflam_disease)
dataCancer <- subset(dataCancer, select = -STDs_genital_herpes)
dataCancer <- subset(dataCancer, select = -STDs_molluscum_contagiosum)
dataCancer <- subset(dataCancer, select = -STDs_Hepatitis_B)
dataCancer <- subset(dataCancer, select = -STDs_HPV)
dataCancer <- subset(dataCancer, select = -STDs_AIDS)
```

Table 2: Table 2. Statistical Summary

name	type	na	mean	disp	median	mad	min	max	nlevs
Age	integer	0	26.82051288.4979481		25.0	8.1543	13	84	0
Num_sexual_partners	integer	26	2.5276442	1.6677605	2.0	1.4826	1	28	0
First_sexual_intercourse	integer	7	16.99529962.8033554		17.0	2.9652	10	32	0
Num_pregnancies	integer	56	2.2755611	1.4474141	2.0	1.4826	0	11	0
Smokes	integer	13	0.1455621	0.3528756	0.0	0.0000	0	1	0
Smokes_years	numeric	13	1.2197214	4.0890169	0.0	0.0000	0	37	0
Smokes_packs_year	numeric	13	0.4531440	2.2266098	0.0	0.0000	0	37	0
Hormonal_Contraceptives	integer	108	0.6413333	0.4799292	1.0	0.0000	0	1	0
Hormonal_Contraceptives_years	numeric	108	2.2564192	3.7642535	0.5	0.7413	0	30	0
IUD	integer	117	0.1120108	0.3155928	0.0	0.0000	0	1	0
IUD_years	numeric	117	0.5148043	1.9430885	0.0	0.0000	0	19	0
STDs	integer	105	0.1049137	0.3066458	0.0	0.0000	0	1	0
STDs_number	integer	105	0.1766268	0.5619928	0.0	0.0000	0	4	0
STDs_condylomatosis	integer	105	0.0584329	0.2347162	0.0	0.0000	0	1	0
STDs_cervical_condylomatosis	integer	105	0.0000000	0.0000000	0.0	0.0000	0	0	0
STDs_vaginal_condylomatosis	integer	105	0.0053121	0.0727385	0.0	0.0000	0	1	0
STDs_vulvo_perineal_condylomatosis	integer	105	0.0571049	0.2321972	0.0	0.0000	0	1	0
STDs_syphilis	integer	105	0.0239044	0.1528528	0.0	0.0000	0	1	0
STDs_pelvic_inflam_disease	integer	105	0.0013280	0.0364420	0.0	0.0000	0	1	0
STDs_genital_herpes	integer	105	0.0013280	0.0364420	0.0	0.0000	0	1	0
STDs_molluscum_contagiosum	integer	105	0.0013280	0.0364420	0.0	0.0000	0	1	0
STDs_AIDS	integer	105	0.0000000	0.0000000	0.0	0.0000	0	0	0
STDs_HIV	integer	105	0.0239044	0.1528528	0.0	0.0000	0	1	0
STDs_Hepatitis_B	integer	105	0.0013280	0.0364420	0.0	0.0000	0	1	0
STDs_HPV	integer	105	0.0026560	0.0515025	0.0	0.0000	0	1	0
STDs_Number_diagnosis	integer	0	0.0874126	0.3025447	0.0	0.0000	0	3	0
STDs_Time_since_first_diagnosis	integer	787	6.1408451	5.8950240	4.0	4.4478	1	22	0
STDs_Time_since_last_diagnosis	integer	787	5.8169014	5.7552705	3.0	2.9652	1	22	0
Dx_Cancer	integer	0	0.0209790	0.1433976	0.0	0.0000	0	1	0
Dx_CIN	integer	0	0.0104895	0.1019392	0.0	0.0000	0	1	0
Dx_HPV	integer	0	0.0209790	0.1433976	0.0	0.0000	0	1	0
Dx	integer	0	0.0279720	0.1649888	0.0	0.0000	0	1	0
Hinselmann	integer	0	0.0407925	0.1979246	0.0	0.0000	0	1	0
Schiller	integer	0	0.0862471	0.2808923	0.0	0.0000	0	1	0
Citology	integer	0	0.0512821	0.2207011	0.0	0.0000	0	1	0
Biopsy	integer	0	0.0641026	0.2450784	0.0	0.0000	0	1	0

Table 3: Table 3. Number of Missing Values in Each Column

	x
Age	0
Num_sexual_partners	26
First_sexual_intercourse	7
Num_pregnancies	56
Smokes	13
Smokes_years	13
Smokes_packs_year	13
Hormonal_Contraceptives	108
Hormonal_Contraceptives_years	108
IUD	117
IUD_years	117
STDs	105
STDs_number	105
STDs_condylomatosis	105
STDs_vulvo_perineal_condylomatosis	105
STDs_syphilis	105
STDs_HIV	105
STDs_Num_diagnosis	0
Dx_Cancer	0
Dx_CIN	0
Dx_HPV	0
Dx	0
Hinselmann	0
Schiller	0
Citology	0
Biopsy	0

```
dataCancer <- subset(dataCancer, select = -STDs_Time_since_first_diagnosis)
dataCancer <- subset(dataCancer, select = -STDs_Time_since_last_diagnosis)
```

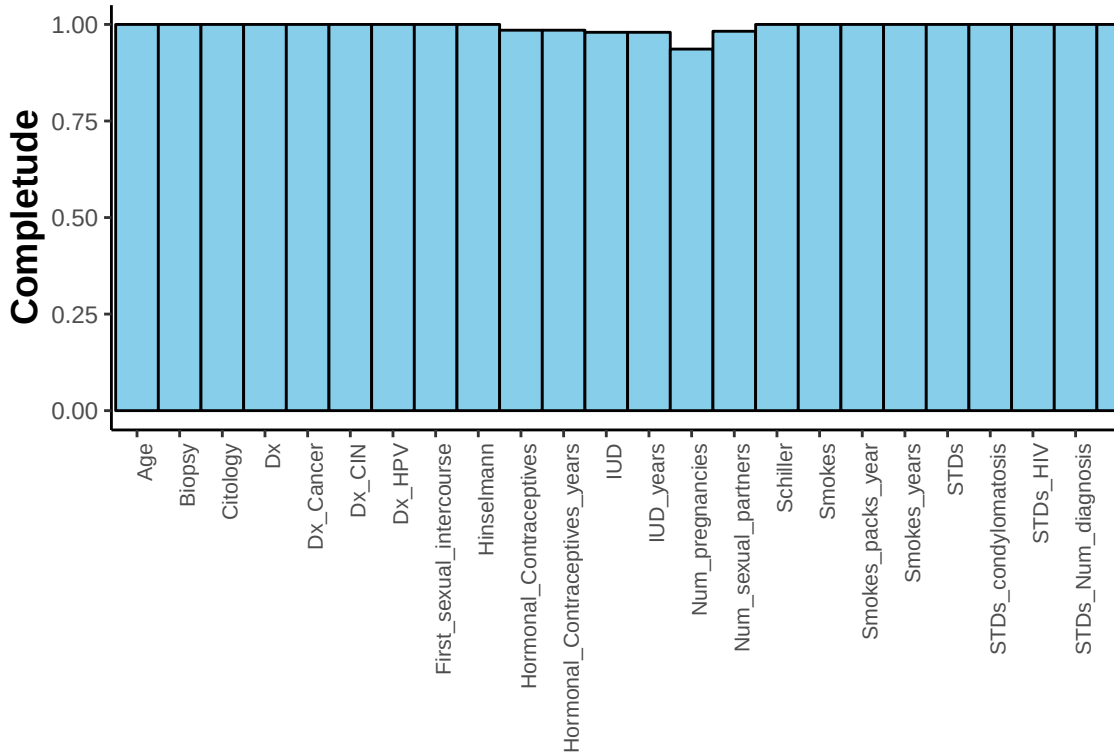
```
# How many NAs are in the data
length(which(is.na(dataCancer)))
```

```
## [1] 1208
```

1.9 Replace missing values

Data rows containing missing values for smokes were removed since these only comprised 13 of the 858 patients. Also removed rows of First_Sexual_Intercourse with missing values (7 of 858 patients). Rows missing values for STDs were also removed since they were missing in all features.

1.10 Dealing with missing values



Proportion of missing values
Figure 1. Proportion of missing values for each feature in the data set.

Assume missing values for number of pregnancies is 0

1.11 Create a target feature

The values for Biopsy, Hinselmann, Schiller, Citology represent the results for cervical cancer exams. Therefore, the data for these columns were combined to produce a target feature called Cancer. Higher values for this feature represent an increased likelihood of cervical cancer. Finally, the Biopsy, Hinselmann, Schiller, Citology features were removed to avoid intoducing bias during the training of the various ML models.

1.12 Exploratory data visualization

1.12.1 Target feature (Cancer)

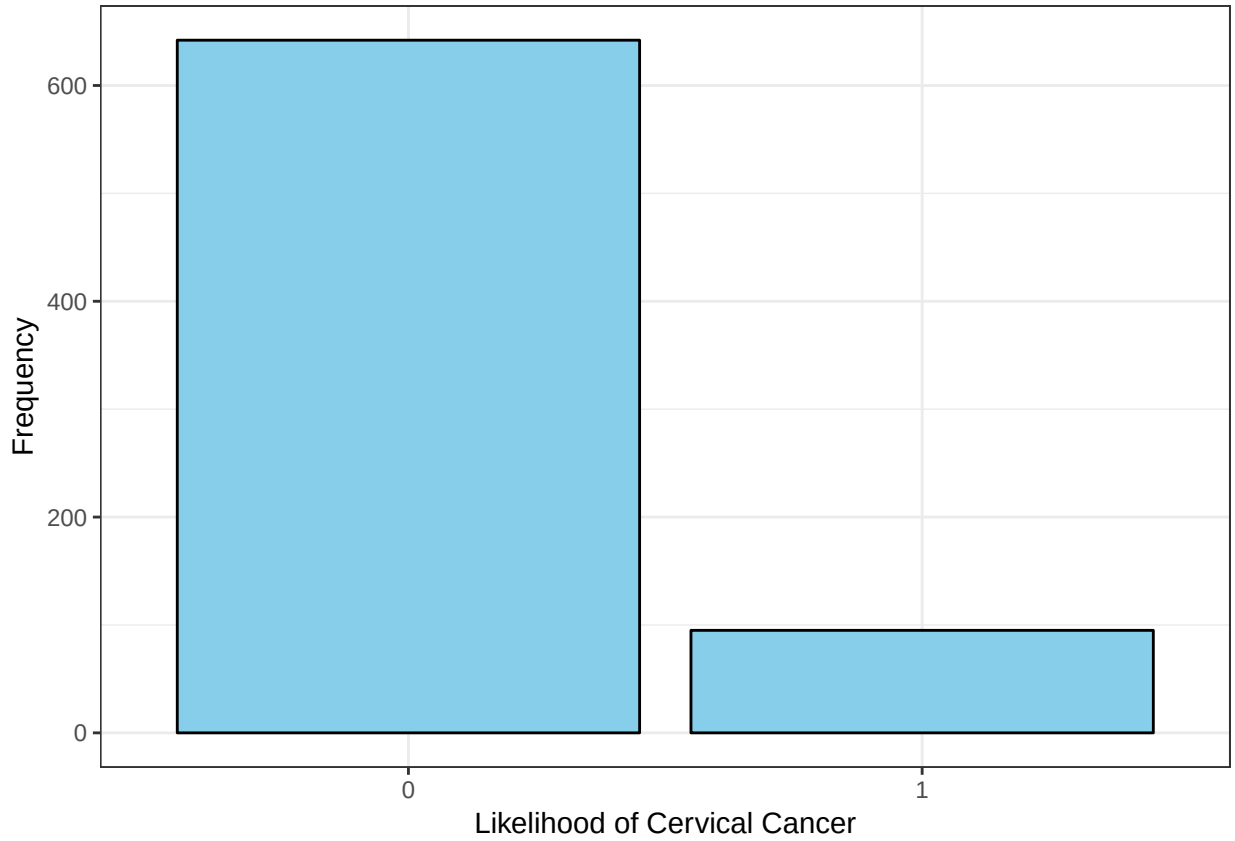


Figure 2. Frequency of the likelihood of cervical cancer based on the results of several cinical exams.

The plot in Fig. 2 shows that the data set is highly unbalanced toward those diagnosed as having no clinical evidence for cervical cancer.

1.12.2 Cervical cancer diagnosis with age of patient

```
p1 <- ggplot(dataCancer, aes(x=Age, fill = Cancer)) + geom_histogram(binwidth = 2, color = "black") + xl
p2 <- ggplot(dataCancer, aes(Cancer, Age)) + geom_boxplot(fill="skyblue") + xlab("Likelihood of Cervical
grid.arrange(p1, p2, nrow = 1)
```

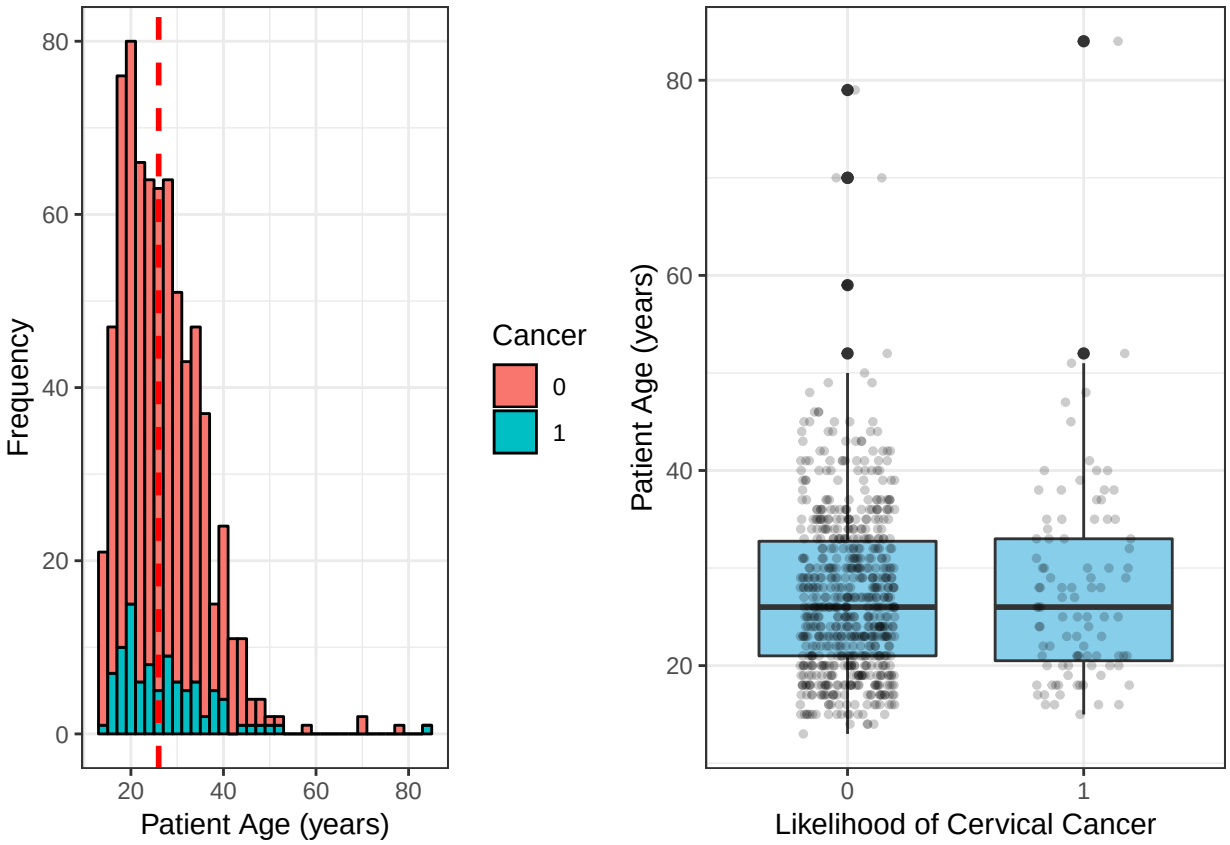


Figure 3. Distribution of Patient ages. Left - histogram. Right - boxplot.

1.12.2.1 Treatment of outliers (capping upper values)

We capped the data by replacing those observations above the upper limit, with the value of 95th %ile.

```
qnt <- quantile(dataCancer$Age, probs=c(.25, .75), na.rm = TRUE)
caps <- quantile(dataCancer$Age, probs=c(.05, .95), na.rm = TRUE)
H <- 1.5 * IQR(dataCancer$Age, na.rm = TRUE)
dataCancer$Age[dataCancer$Age > (qnt[2] + H)] <- caps[2]

p1 <- ggplot(dataCancer, aes(x=Age, fill = Cancer)) + geom_histogram(binwidth = 2, color = "black") + xl
p2 <- ggplot(dataCancer, aes(Cancer, Age)) + geom_boxplot(fill="skyblue") + xlab("Likelihood of Cervical
grid.arrange(p1, p2, nrow = 1)
```

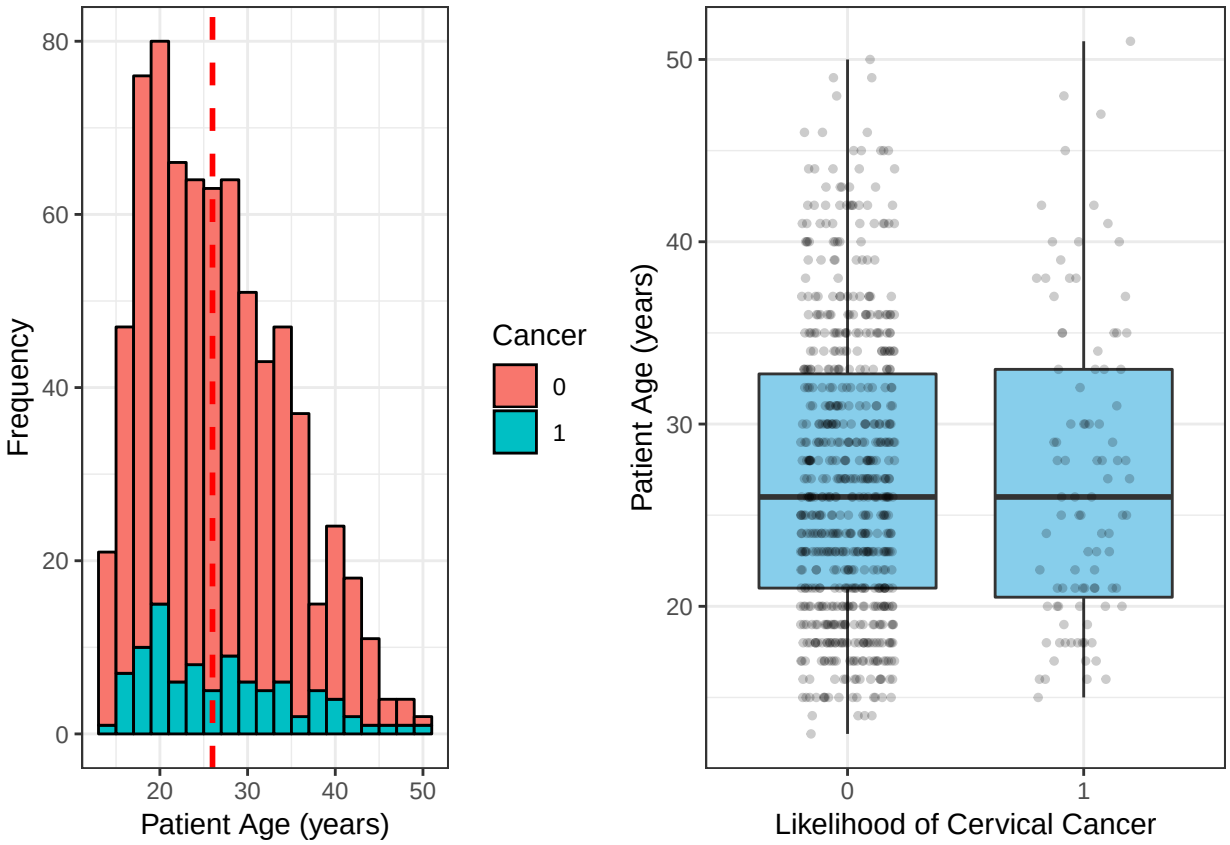


Figure 4. Distribution of likelihood for cervical cancer with patient age. Left - histogram. Right - boxplot.

As illustrated in figs. 3 and 4 the likelihood to have cervical cancer does not appear to vary greatly with age (Fig. 3).

1.12.3 Number of Sexual Partners

```
p1 <- ggplot(dataCancer, aes(x=Num_sexual_partners, fill = Cancer)) + geom_histogram(binwidth = 1, colo
p2 <- ggplot(dataCancer, aes(Cancer, Num_sexual_partners)) + geom_boxplot(fill="skyblue") + xlab("Likel
grid.arrange(p1, p2, nrow = 1)
```

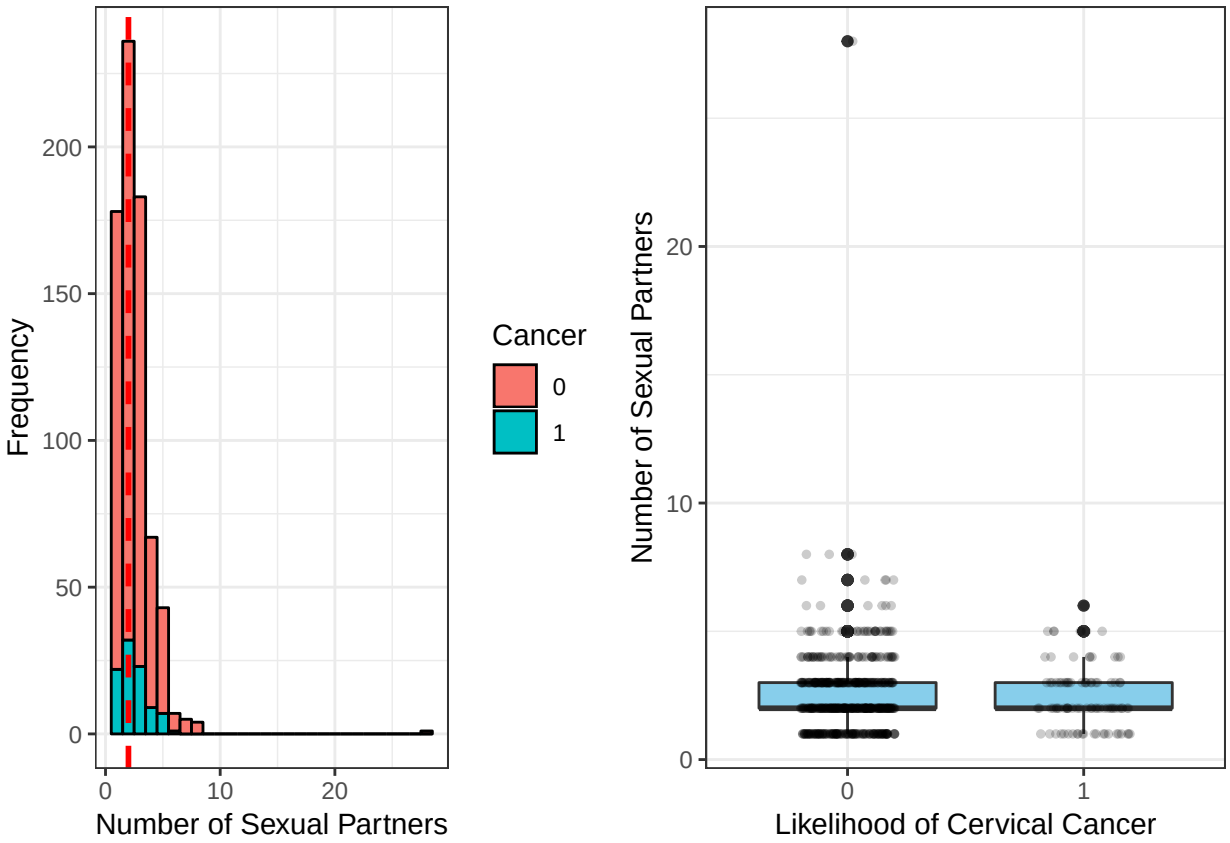


Figure 5. Distribution of likelihood for cervical cancer with number of sexual partners. Left - histogram. Right - boxplot.

The majority of patients had less than 5 sexual partners. The data is highly positively skewed with a maximum value of 28 and a median of 2. Removed outlier Num_sexual_partners = 28 since this patient exhibited no clinical sign of cervical cancer.

```
dataCancer$Num_sexual_partners[(dataCancer$Num_sexual_partners == 28)] <- 0
p1 <- ggplot(dataCancer, aes(x=Num_sexual_partners, fill = Cancer)) + geom_histogram(binwidth = 1, colo
p2 <- ggplot(dataCancer, aes(Cancer, Num_sexual_partners)) + geom_boxplot(fill="skyblue") + xlab("Likel
grid.arrange(p1, p2, nrow = 1)
```

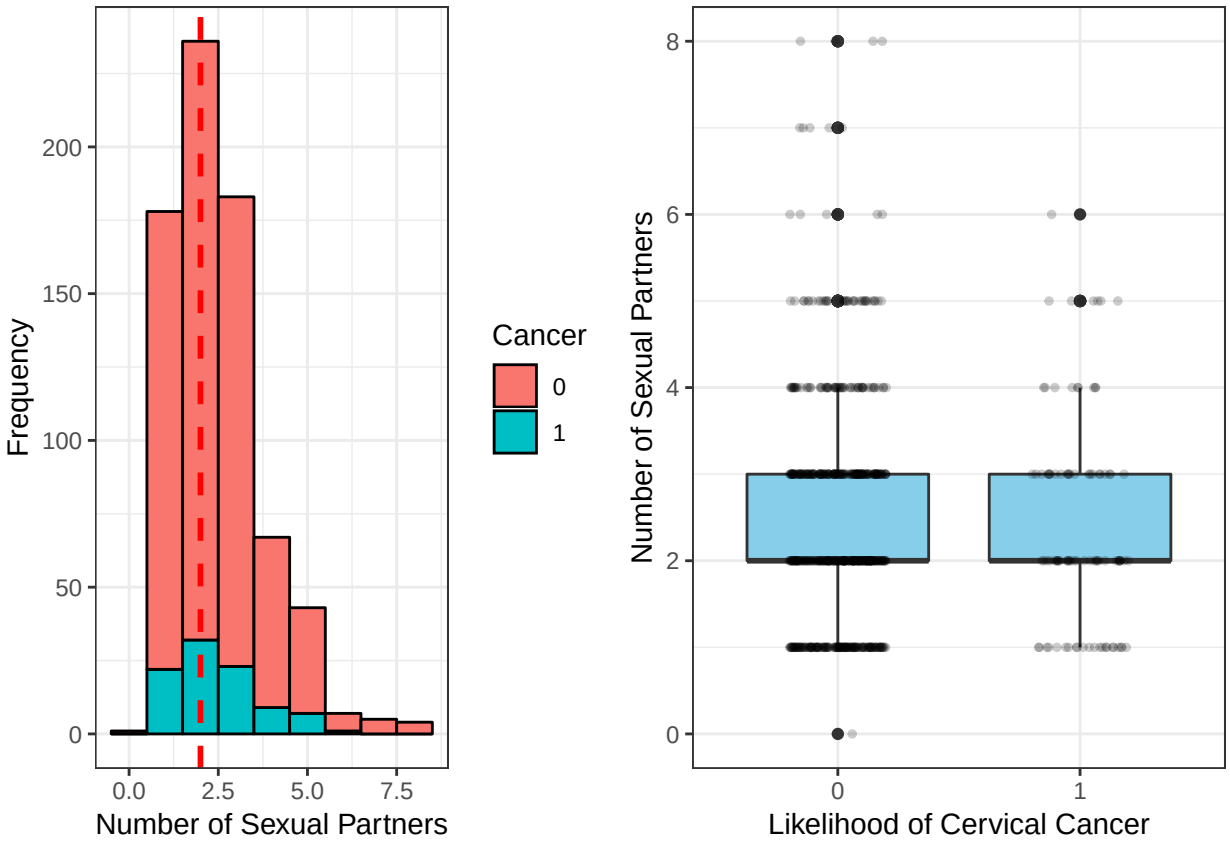


Figure 6. Distribution of the Number of Sexual Partners for each partner. Left - histogram. Right - Boxplot. The number of sexual partners exibited no apparent correlation with the likelihood of having cervical cancer.

1.12.4 Time of first intercourse

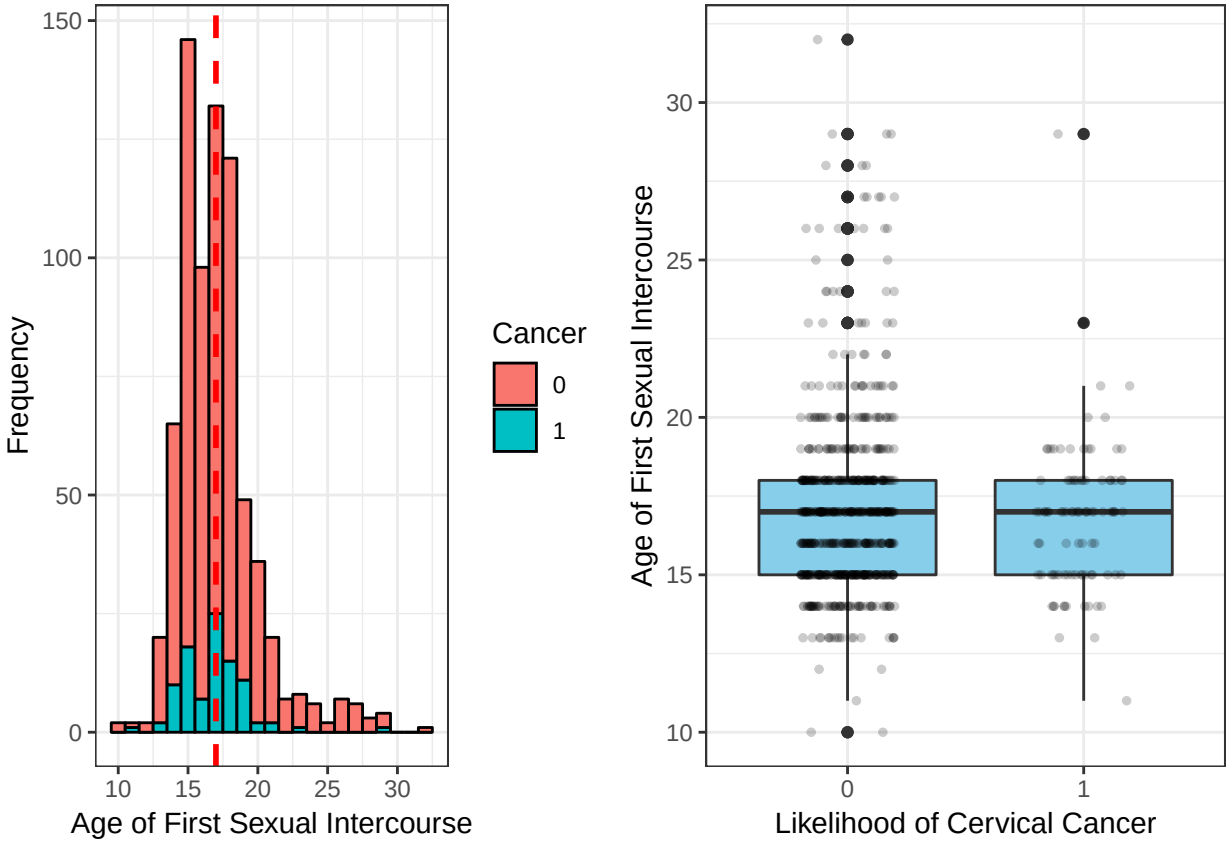
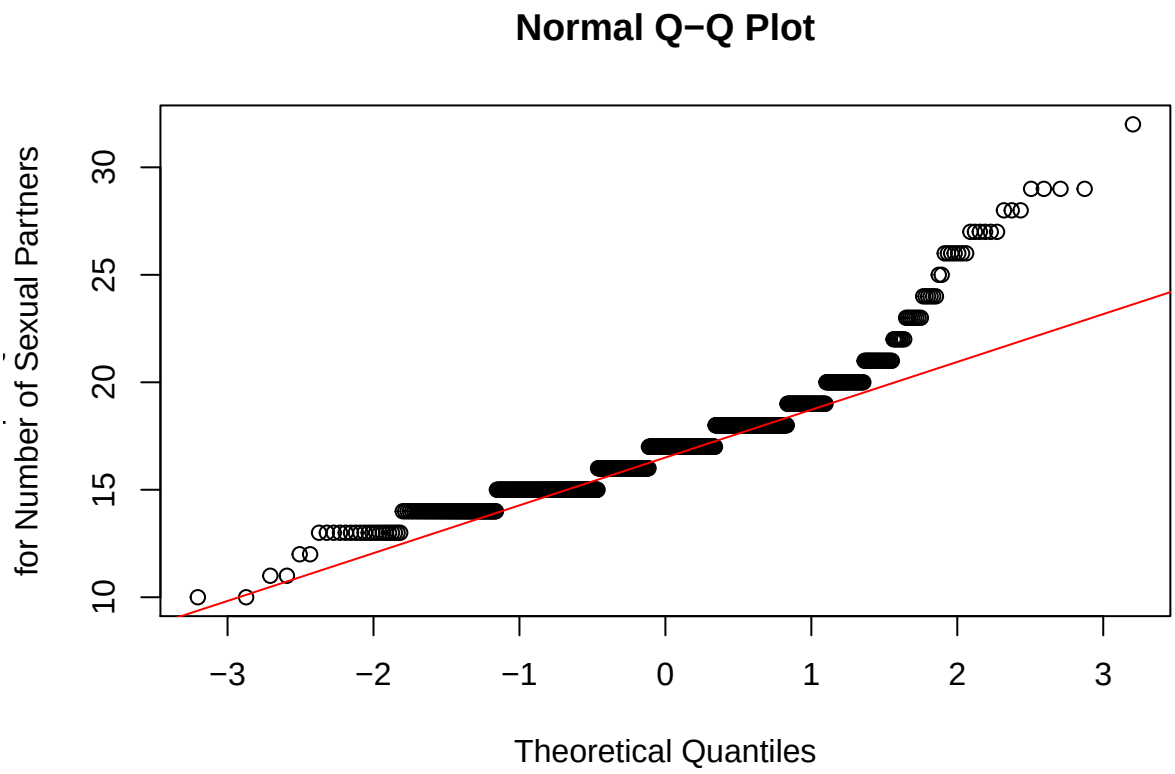


Figure 7. Distribution of time of first sexual intercourse for patients. Left - histogram. Right - boxplot.

Patients varied in the age of first sexual encounter ranging from 10 to 32 years with a largely normal distribution and a median value of 17 years old (Fig. 7).




```
## [1] 1.556652
```

Figure 8. The qq-plot for Number of Sexual Partners feature.

The qqplot for the Number of Sexual Partners indicates that the data is slightly positively skewed (skewness = 1.56).

1.12.5 Number of Pregnacies

```
p5 <- ggplot(dataCancer, aes(x=Num_pregnancies, fill = Cancer)) + geom_histogram(binwidth = 1, color =
p6 <- ggplot(dataCancer, aes(Cancer, Num_pregnancies)) + geom_boxplot(fill="skyblue") + xlab("Likelihoo
grid.arrange(p5, p6, nrow = 1)
```

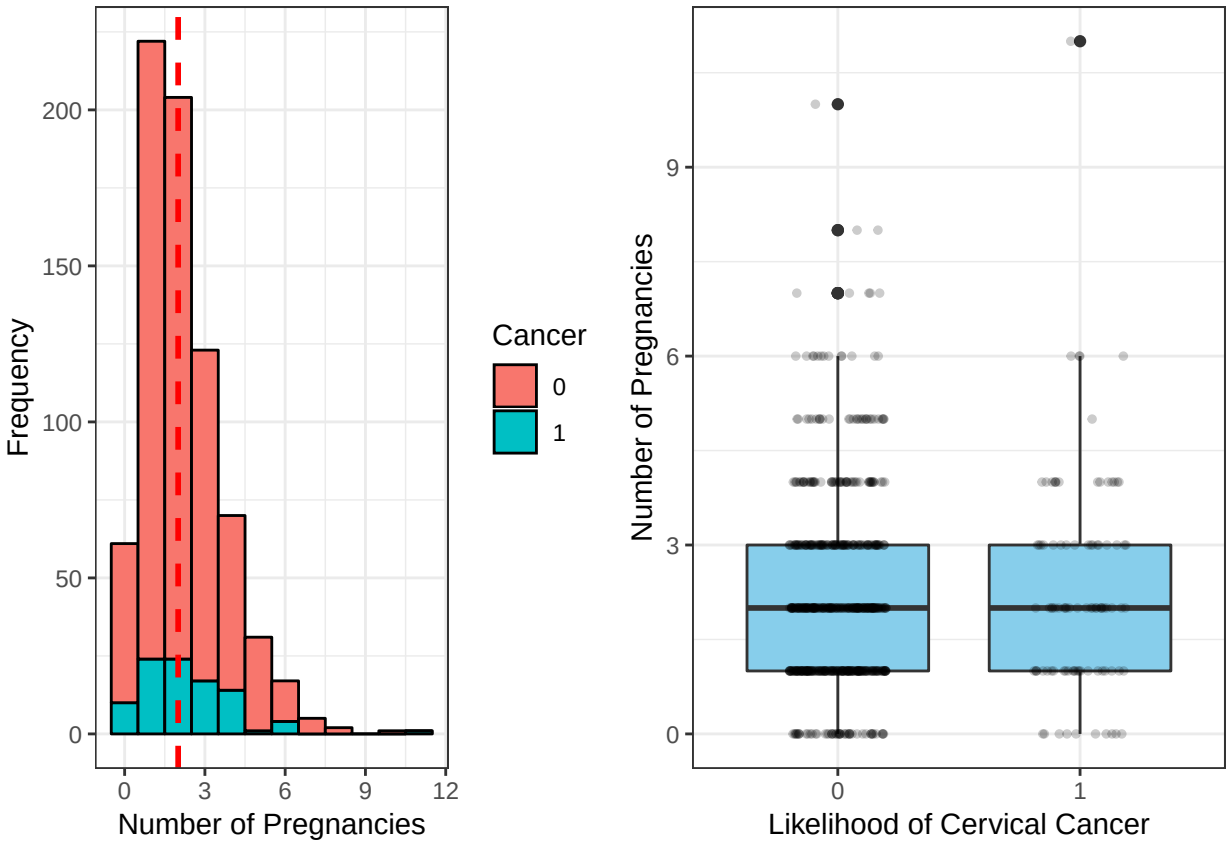
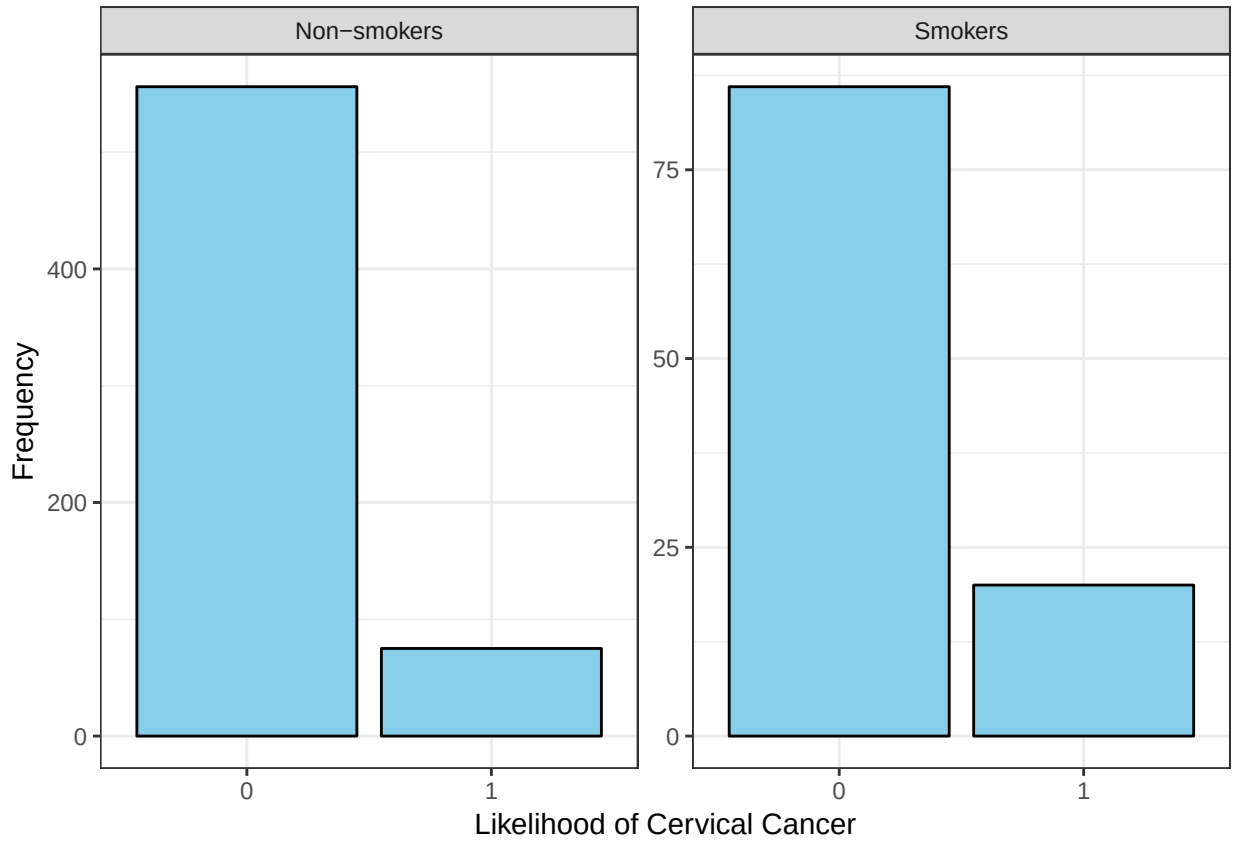


Figure 9. Distribution of the number of pregnancies for each patient. Left - histogram. Right - boxplot.

The number of pregnancies for each patient varied from 0 to 11 with a median value of 2 (Fig. 9). The number of pregnancies did not appear to account for the likelihood to have cervical cancer.

1.12.6 Smoking status

```
dataCancer$Smokes <- factor(dataCancer$Smokes, levels = c("0", "1"))
labels <- c("0" = "Non-smokers", "1" = "Smokers")
ggplot(data = subset(dataCancer, !is.na(Smokes)), aes(x = Cancer)) + geom_bar(color = "black", fill = "
```




```
CrossTable(dataCancer$Cancer, dataCancer$Smokes)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  737
##
##
##                    | dataCancer$Smokes
## dataCancer$Cancer |         0 |         1 | Row Total |
## ------------------|-----------|-----------|-----------|
##                 0 |       556 |        86 |       642 |
##                   |     0.073 |     0.435 |           |
##                   |     0.866 |     0.134 |     0.871 |
##                   |     0.881 |     0.811 |           |
##                   |     0.754 |     0.117 |           |
## ------------------|-----------|-----------|-----------|
```

```
##                  1 |        75 |        20 |        95 |
##                    |     0.494 |     2.939 |           |
##                    |     0.789 |     0.211 |     0.129 |
##                    |     0.119 |     0.189 |           |
##                    |     0.102 |     0.027 |           |
## -------------------|-----------|-----------|-----------|
##       Column Total |       631 |       106 |       737 |
##                    |     0.856 |     0.144 |           |
## -------------------|-----------|-----------|-----------|
##
##
```

Figure 10. Smoking status for patients.

The patients are predominantly non-smokers (85%, Fig. 10).The table shows that there is a slightly higher chance of patients who are smokers to also have cervical cancer.

1.12.7 Years smoking for smokers only in the entire patient cohort

```
Smokes_years_nz <- dataCancer %>% filter(Smokes_years > 0)
p7 <- ggplot(Smokes_years_nz, aes(x=Smokes_years, fill = Cancer)) + geom_histogram(binwidth = 2, color =

p8 <- ggplot(Smokes_years_nz, aes(Cancer, Smokes_years)) + geom_boxplot(fill="skyblue") + xlab("Likelih
grid.arrange(p7, p8, nrow = 1)
```

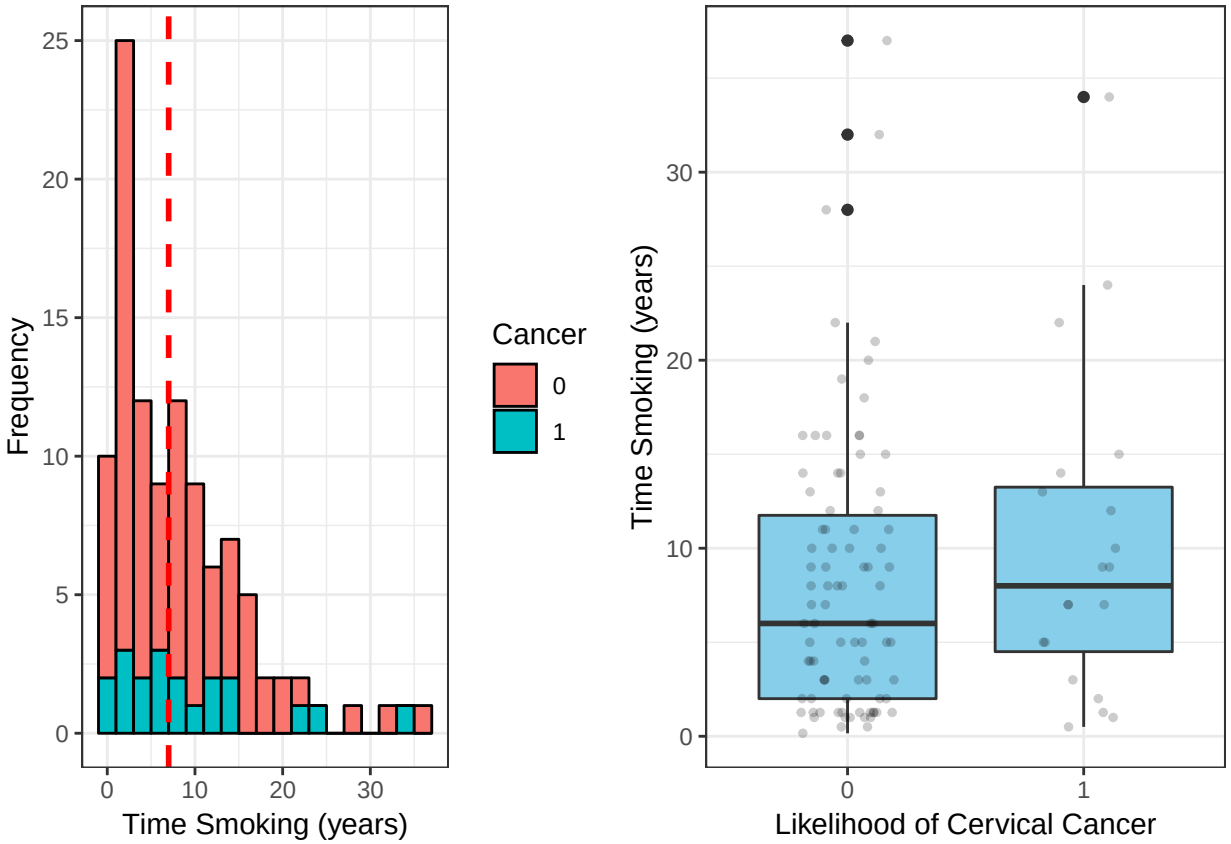


Figure 11.. Distribution of number of years spent smoking for each patient. Left - histogram. Right - boxplot.

The period spent smoking by smokers within the patient cohort varied between 0.16 and 37 years with a median value of 7 years (Fig. 11). The data was positively skewed with several patients who had been smoking for longer than 30 years. Several patients with no clinical signs of cervical cancer had smoked for long periods of their lives (i.e. $>$ 10 years) however these tended to be outliers. Interestingly, smokers with clinical evidence for cervical cancer had smoked for longer periods.

1.12.8 Quantity of cigarettes

```
Smokes_packs_years_nz <- dataCancer %>% filter(Smokes_packs_year > 0)
p9 <- ggplot(Smokes_packs_years_nz, aes(x=Smokes_packs_year, fill = Cancer)) + geom_histogram(binwidth =

p10 <- ggplot(Smokes_packs_years_nz, aes(Cancer, Smokes_packs_year)) + geom_boxplot(fill="skyblue") + xl
grid.arrange(p9, p10, nrow = 1)
```

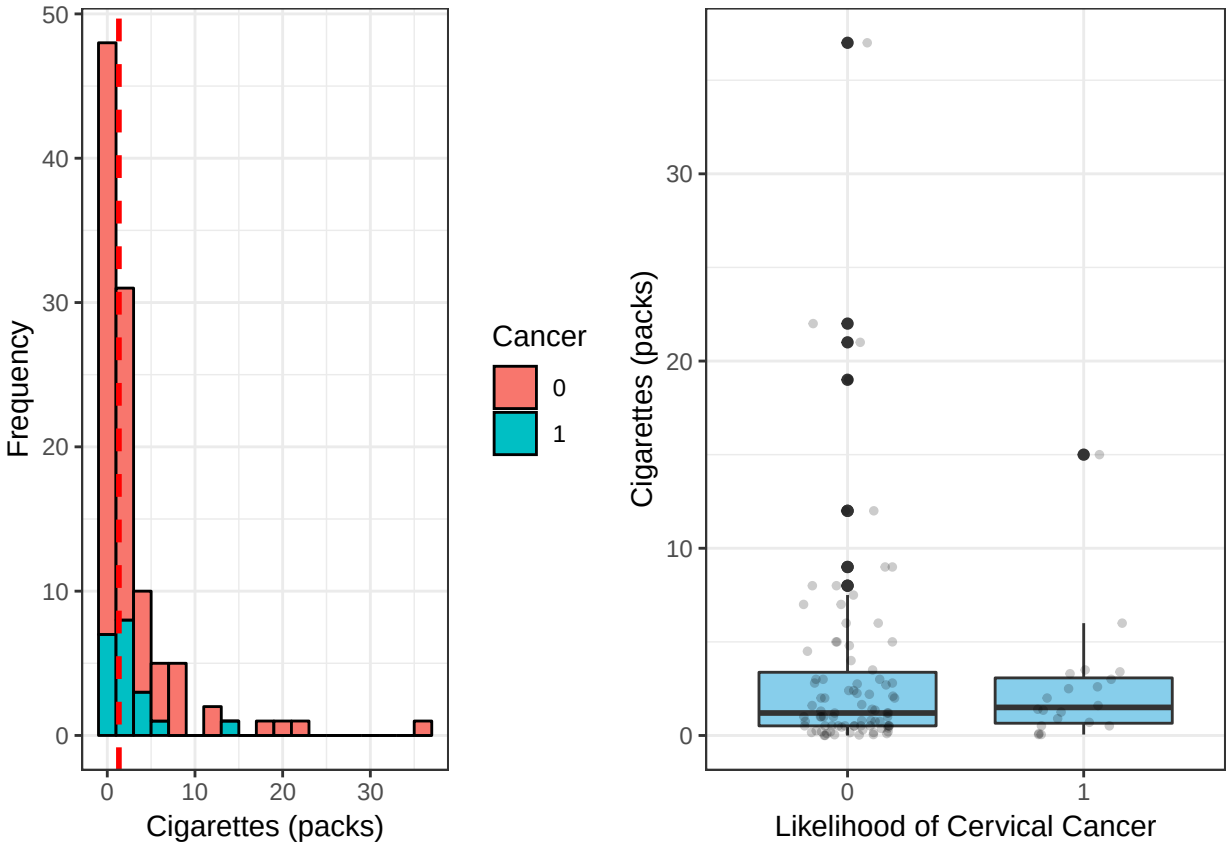
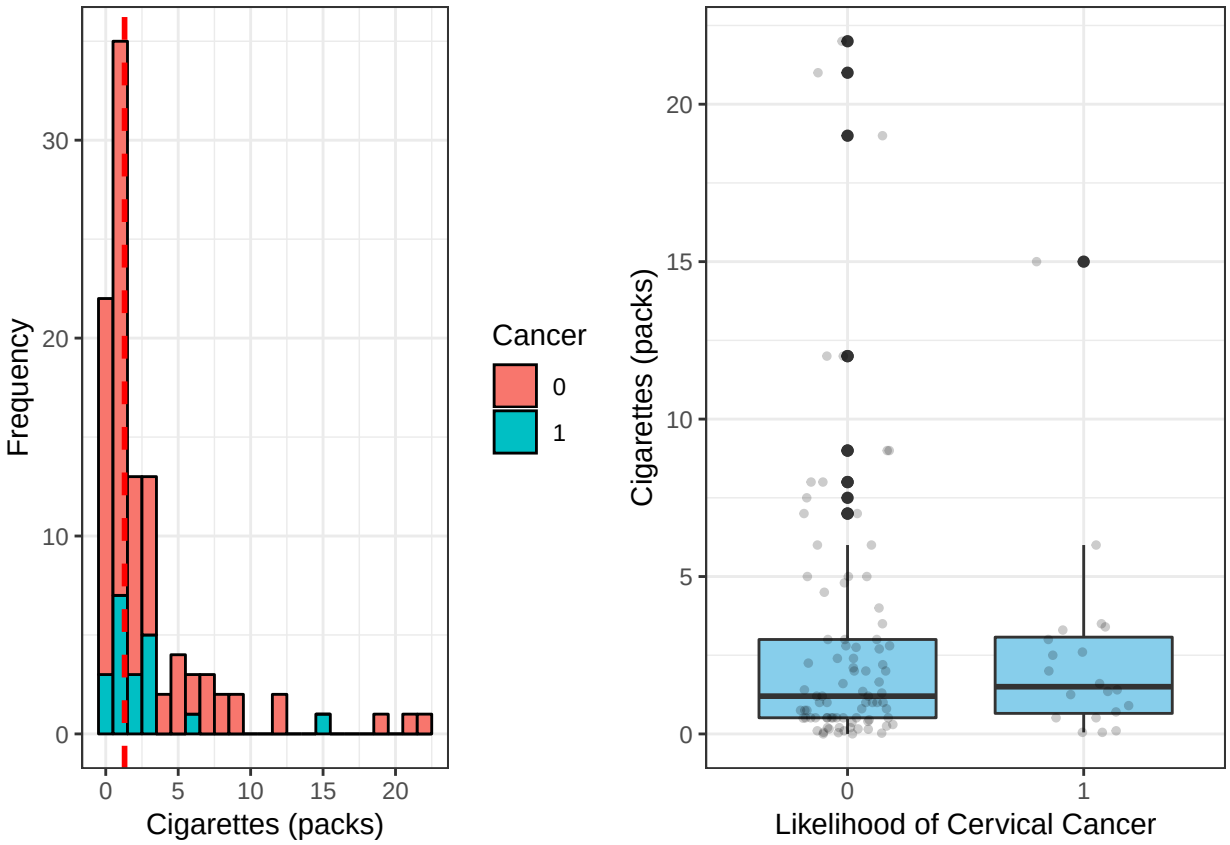


Figure 12. Quantity of cigarettes smoked each year. Left - histogram. Right - boxplot.

An outlier (Smokes_packs_year = 37) was removed since it did not correlate with an increased likelihood for cervical cancer (Fig. 12).

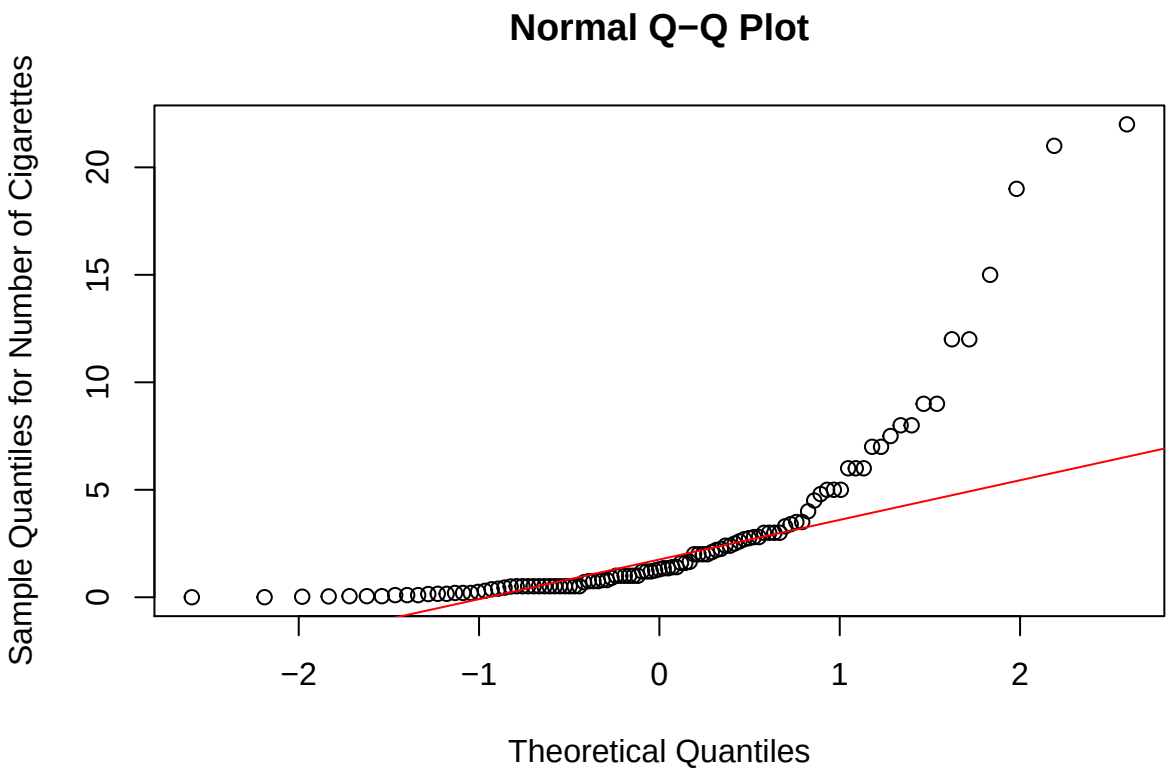
```
dataCancer$Smokes_packs_year[(dataCancer$Smokes_packs_year == 37)] <- 0
Smokes_packs_years_nz <- dataCancer %>% filter(Smokes_packs_year > 0)
p9 <- ggplot(Smokes_packs_years_nz, aes(x=Smokes_packs_year, fill = Cancer)) + geom_histogram(binwidth =

p10 <- ggplot(Smokes_packs_years_nz, aes(Cancer, Smokes_packs_year)) + geom_boxplot(fill="skyblue") + xl
grid.arrange(p9, p10, nrow = 1)
```




```
skewness(dataCancer$Smokes_packs_year)
```

```
## [1] 7.452461
```

Figure 13. Distribution of packs of cigarettes smoked by cervical cancer patients. Left - histogram. Right - boxplot.




```
## [1] 2.69972
```

Figure 14. The qq-plot for the Smokes_packs_year feature.

The qqplot for the Smokes_packs_year feature are highly skewed (skewness = 2.69972). The data for the Smokes_packs_year feature were cube-root transformed in order to obtain a more gaussian distribution.

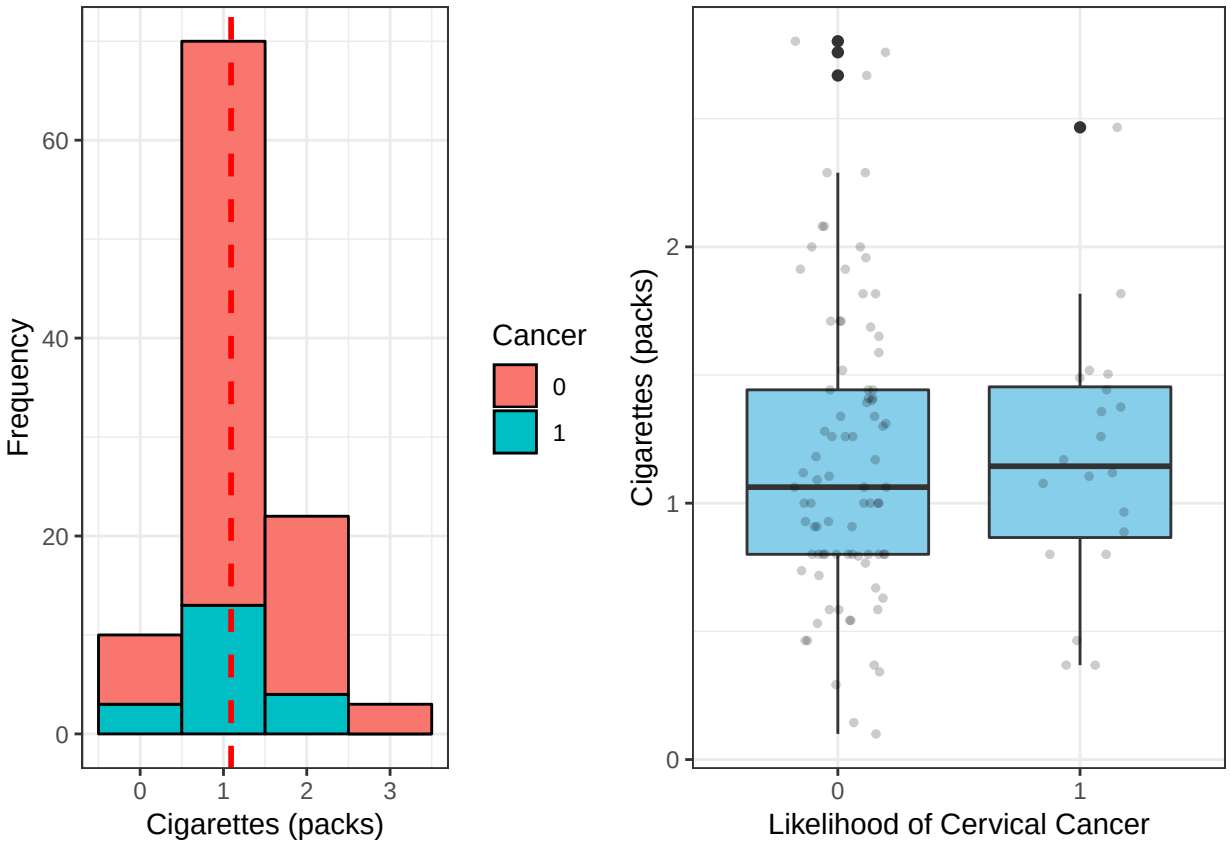
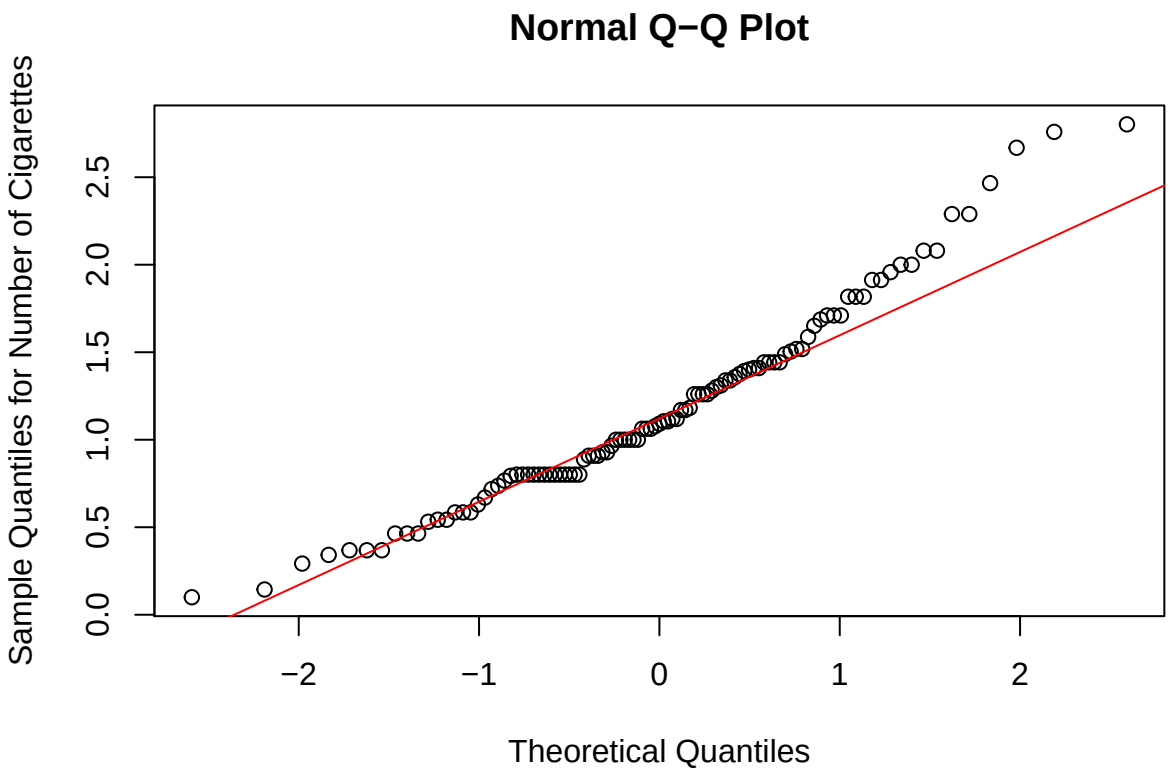


Figure 15. Data for the Smokes_packs_year feature following cubed root transformation. Left - histogram. Right - boxplot.

The boxplots in fig. 15 indicate that smokers who consumed more cigarettes each year had a higher likelihood of exhibiting the clinical signs of cervical cancer.

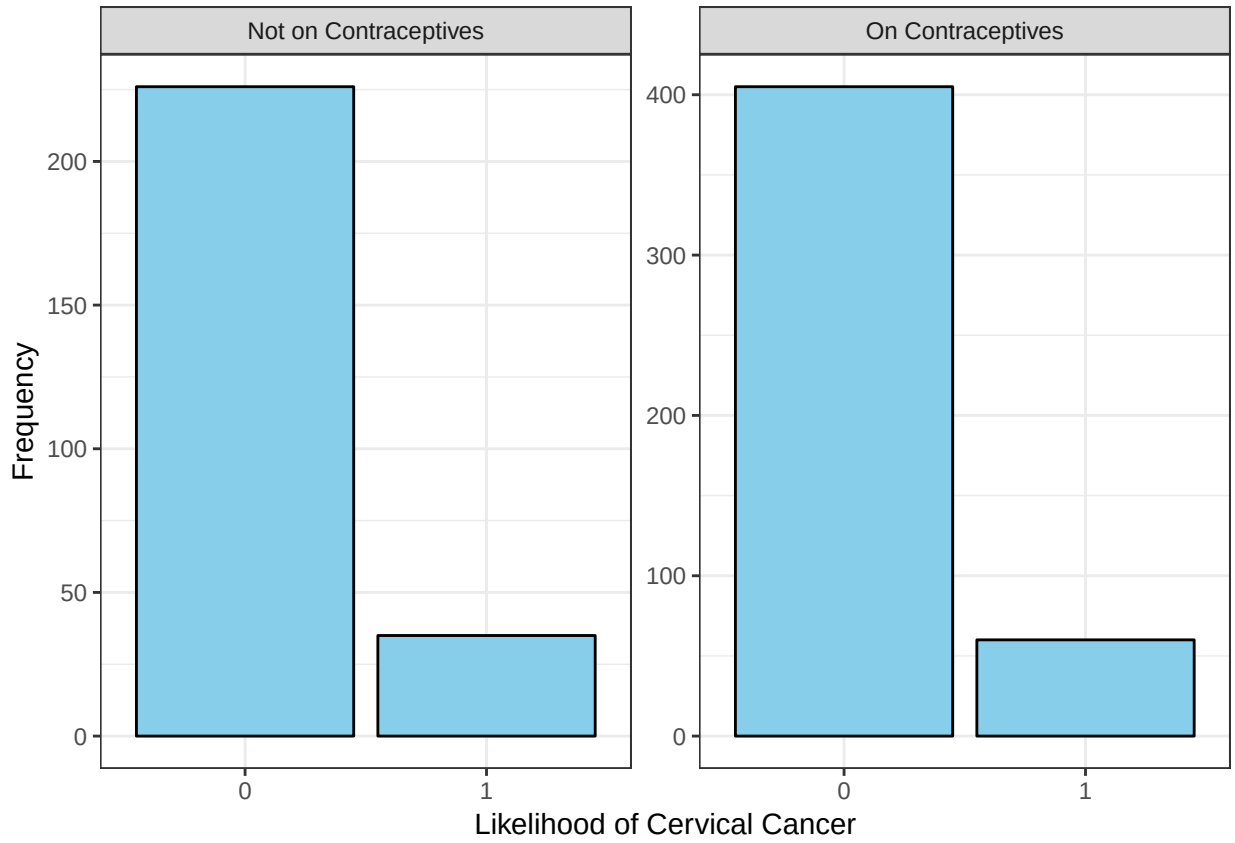



```
## [1] 0.6762561
```

Figure 16. The qq-plot following cubed root transformation of the Smokes_packs_year feature.

1.12.9 Hormone Contraceptives

```
dataCancer$Hormonal_Contraceptives <- factor(dataCancer$Hormonal_Contraceptives, levels = c("0", "1"))
labels <- c("0" = "Not on Contraceptives", "1" = "On Contraceptives")
ggplot(data = subset(dataCancer, !is.na(Hormonal_Contraceptives)), aes(x = Cancer)) + geom_bar(color =
```




```
CrossTable(dataCancer$Cancer, dataCancer$Hormonal_Contraceptives)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  726
##
##
##                    | dataCancer$Hormonal_Contraceptives
## dataCancer$Cancer  |         0 |         1 | Row Total |
## -------------------|-----------|-----------|-----------|
##                  0 |       226 |       405 |       631 |
##                    |     0.003 |     0.002 |           |
##                    |     0.358 |     0.642 |     0.869 |
##                    |     0.866 |     0.871 |           |
##                    |     0.311 |     0.558 |           |
## -------------------|-----------|-----------|-----------|
```

```
##                   1 |        35 |        60 |        95 |
##                     |     0.021 |     0.012 |           |
##                     |     0.368 |     0.632 |     0.131 |
##                     |     0.134 |     0.129 |           |
##                     |     0.048 |     0.083 |           |
## -------------------|-----------|-----------|-----------|
##       Column Total |       261 |       465 |       726 |
##                     |     0.360 |     0.640 |           |
## -------------------|-----------|-----------|-----------|
##
##
```

Figure 17. Patients on hormonal contraceptives by likelihood of cervical cancer.

Approximately two thirds (417 out 650) of the patient cohort were taking hormonal contraceptives (Fig. 17).

1.12.10 Years on hormonal contraceptives

```
Hormonal_Contraceptives_years_nz <- dataCancer %>% filter(Hormonal_Contraceptives_years > 0)
p11 <- ggplot(Hormonal_Contraceptives_years_nz, aes(x=Hormonal_Contraceptives_years, fill = Cancer)) + g
p12 <- ggplot(Hormonal_Contraceptives_years_nz, aes(Cancer, Hormonal_Contraceptives_years)) + geom_boxpl
grid.arrange(p11, p12, nrow = 1)
```

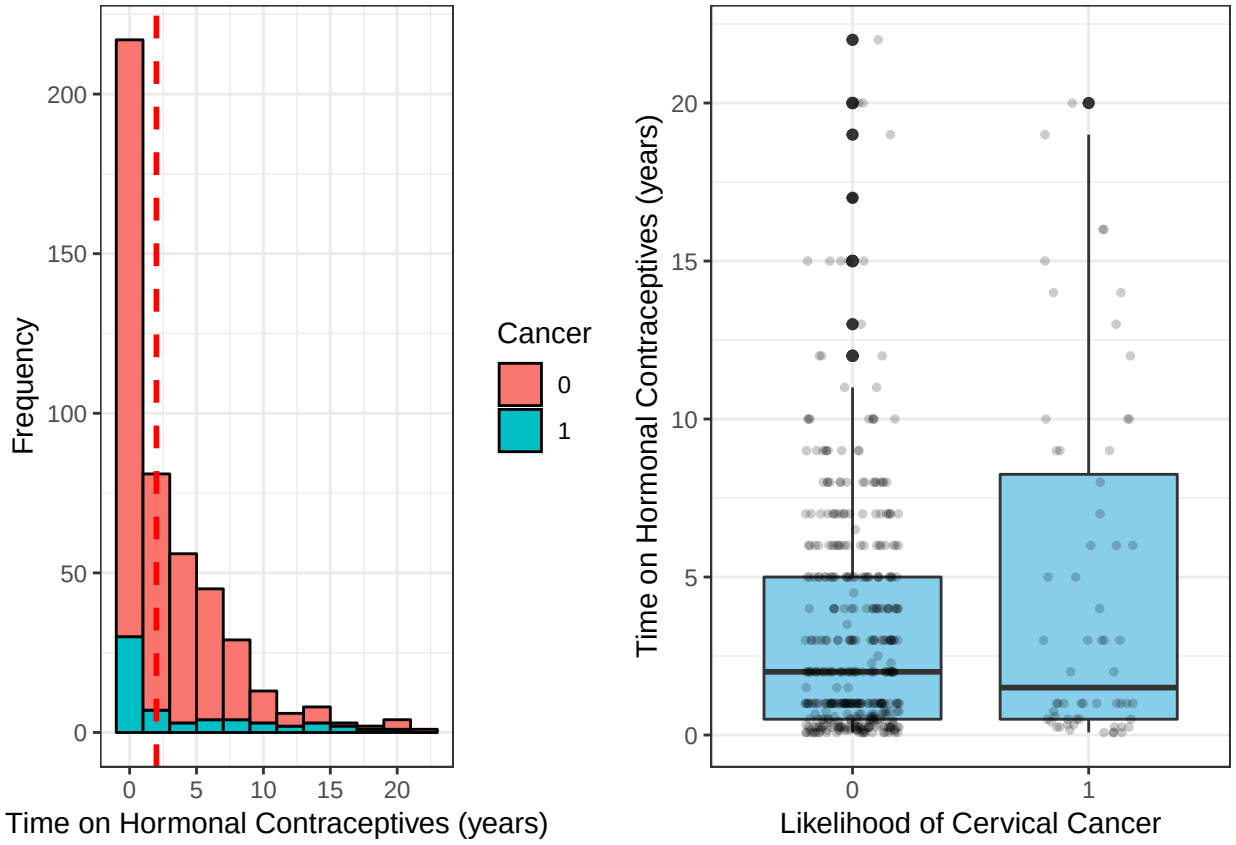
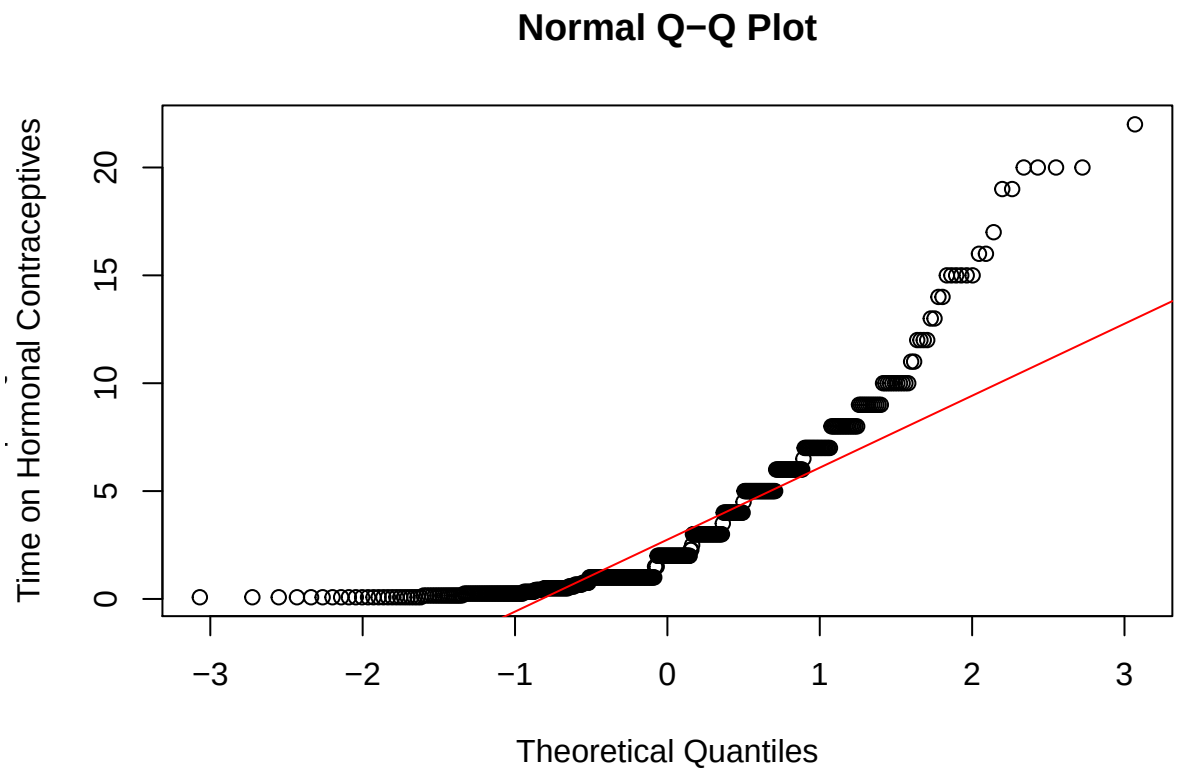


Figure 18. Distribution of length of period on hormonal contraceptives. Left - histogram. Right - boxplot.

Patients on hormonal contraceptives had been taking them for varying lengths of time from 0.08 to 30 years (Fig. 18). The majority of the patient cohort were on contraceptives for less than 10 years.




```
## [1] 1.818768
```

Figure 19. The qq-plot for the Hormonal_Contraceptives_years feature.

The qq-plot for the data for the Hormonal_Contraceptives_years is highly skewed (Fig. 19). The data for the Hormonal_Contraceptives_years feature were cube-root transformed in order to obtain a more gaussian distribution.

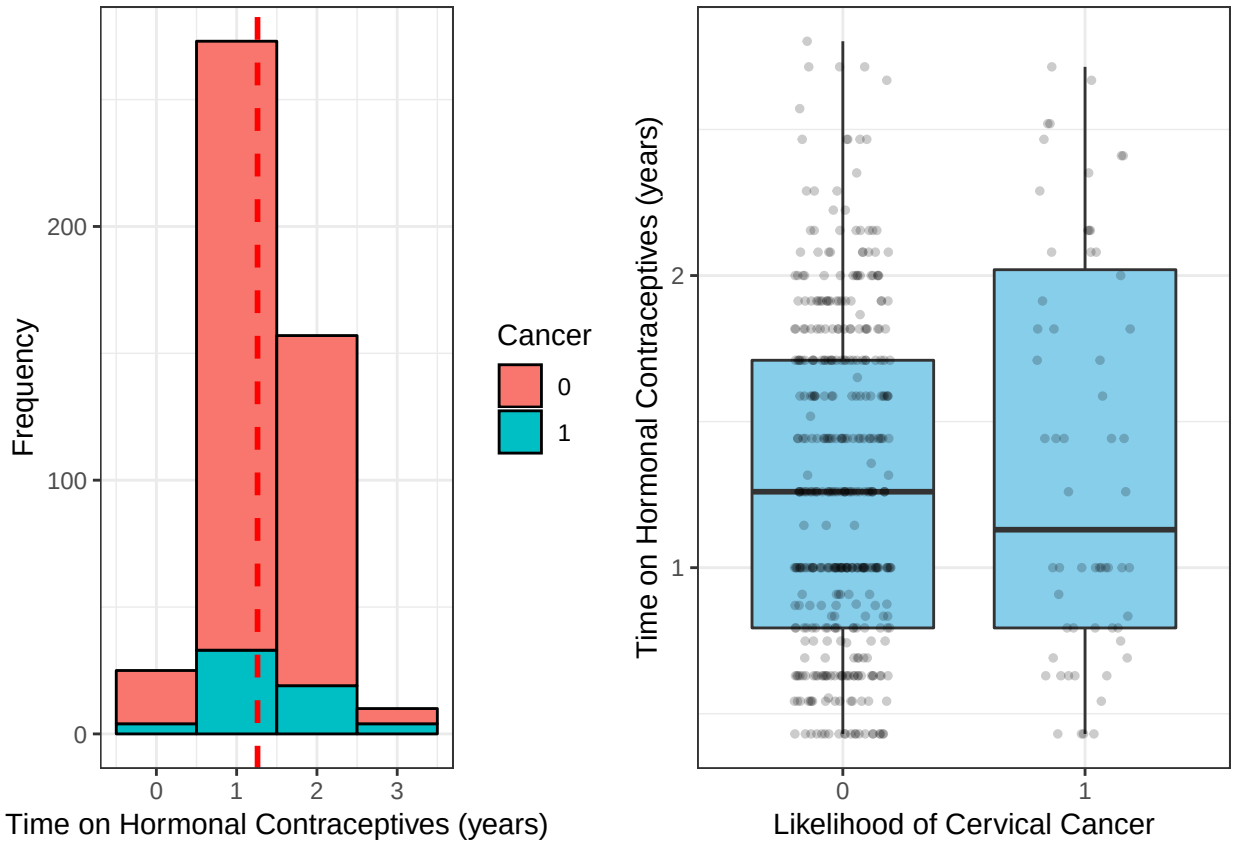
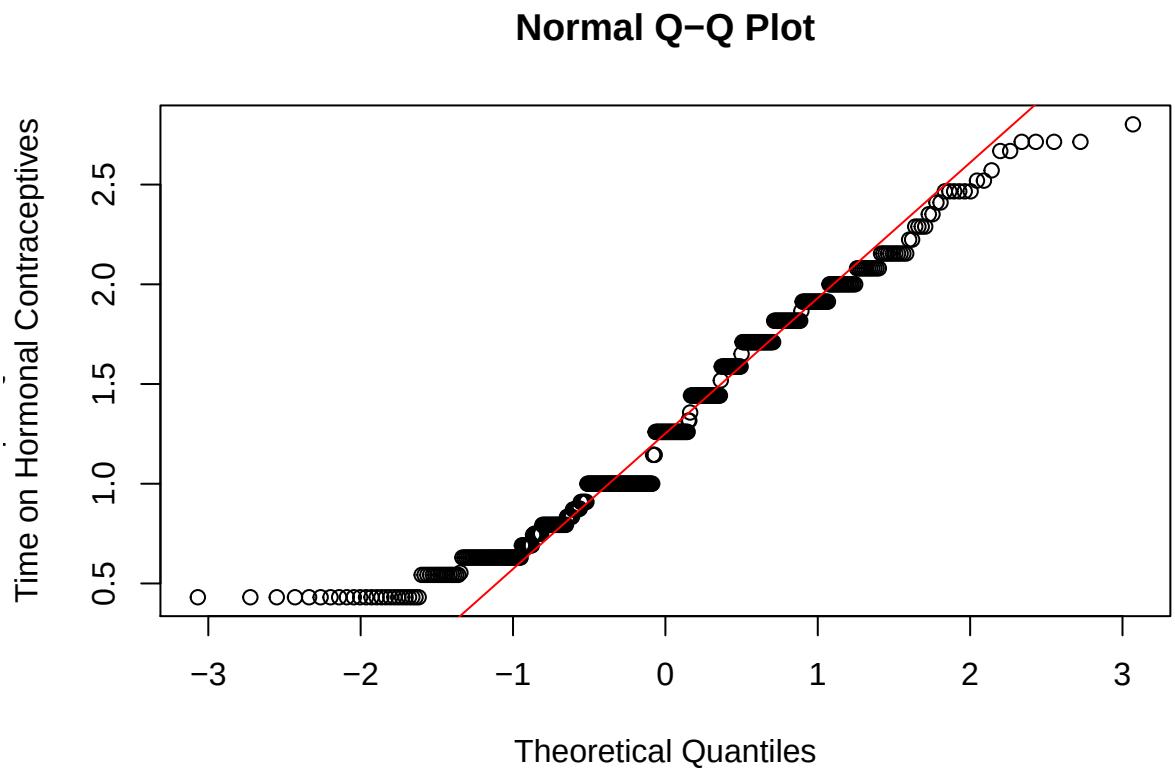


Figure 20. Distribution of the data for the Hormonal_Contraceptives_years feature. Left - histogram. Right - boxplot.

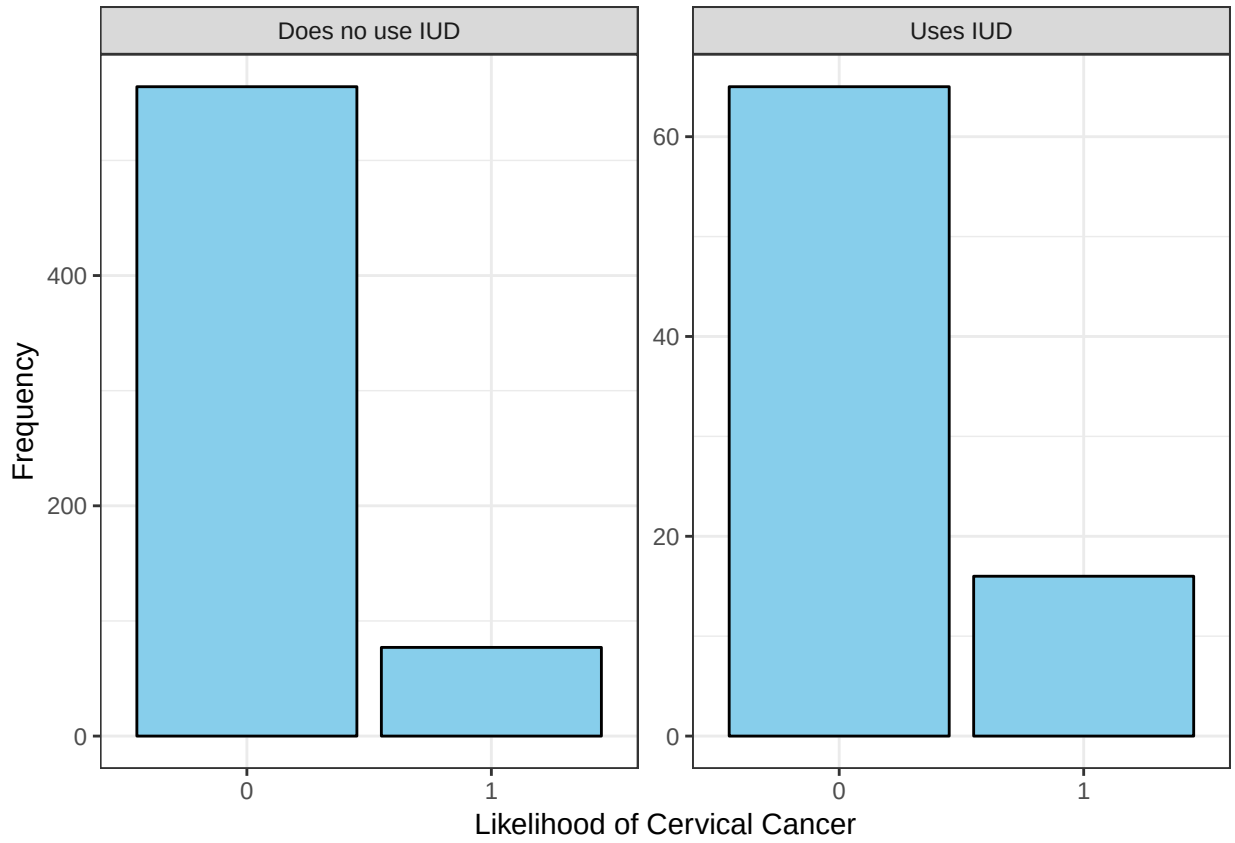



```
## [1] 0.384565
```

Figure 21. The qq-plot of the Hormonal_Contraceptives_years feature following cubed root transformation of the data.

1.12.11 IUDs

```
dataCancer$IUD <- factor(dataCancer$IUD, levels = c("0", "1"))
labels <- c("0" = "Does no use IUD", "1" = "Uses IUD")
ggplot(data = subset(dataCancer, !is.na(IUD)), aes(x = Cancer)) + geom_bar(color = "black", fill = "skyl
```




```
CrossTable(dataCancer$Cancer, dataCancer$IUD)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  722
##
##
##                    | dataCancer$IUD
## dataCancer$Cancer |         0 |         1 | Row Total |
## ------------------|-----------|-----------|-----------|
##                 0 |       564 |        65 |       629 |
##                   |     0.055 |     0.439 |           |
##                   |     0.897 |     0.103 |     0.871 |
##                   |     0.880 |     0.802 |           |
##                   |     0.781 |     0.090 |           |
## ------------------|-----------|-----------|-----------|
```

```
##                   1 |        77 |        16 |        93 |
##                     |     0.375 |     2.970 |           |
##                     |     0.828 |     0.172 |     0.129 |
##                     |     0.120 |     0.198 |           |
##                     |     0.107 |     0.022 |           |
## --------------------|-----------|-----------|-----------|
##       Column Total  |       641 |        81 |       722 |
##                     |     0.888 |     0.112 |           |
## --------------------|-----------|-----------|-----------|
##
##
```

Figure 22. Barchart showing the proportion of patients using an Intra Uterine Device (IUD) by likelihood of cervical cancer.

Approximately 10 percent of the patient cohort were using an Intra Uterine Device (IUD) (Fig. 22).

1.12.12 Years using an IUD

```
IUD_years_nz <- dataCancer %>% filter(IUD_years > 0)
p13 <- ggplot(IUD_years_nz, aes(x=IUD_years, fill = Cancer)) + geom_histogram(binwidth = 2, color = "bla
p14 <- ggplot(IUD_years_nz, aes(Cancer, IUD_years)) + geom_boxplot(fill="skyblue") + xlab("Likelihood o
grid.arrange(p13, p14, nrow = 1)
```

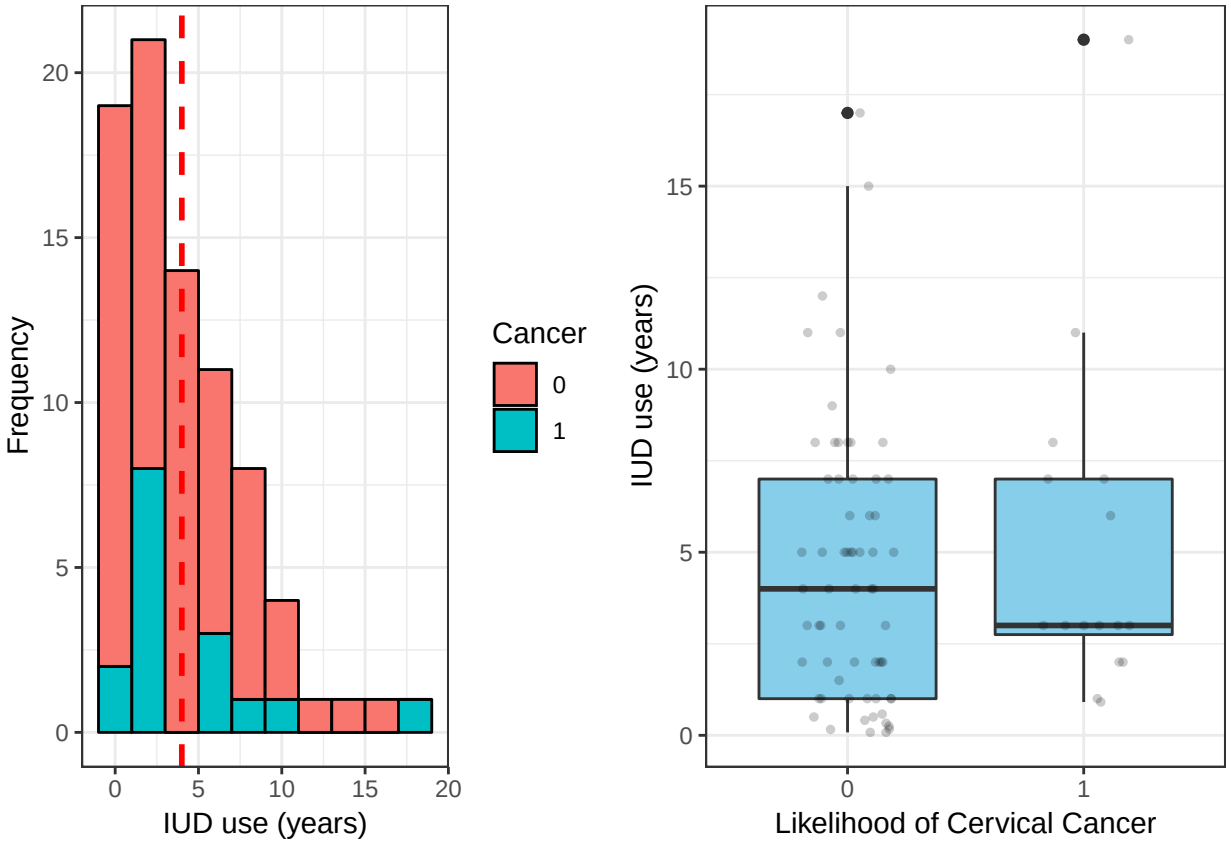
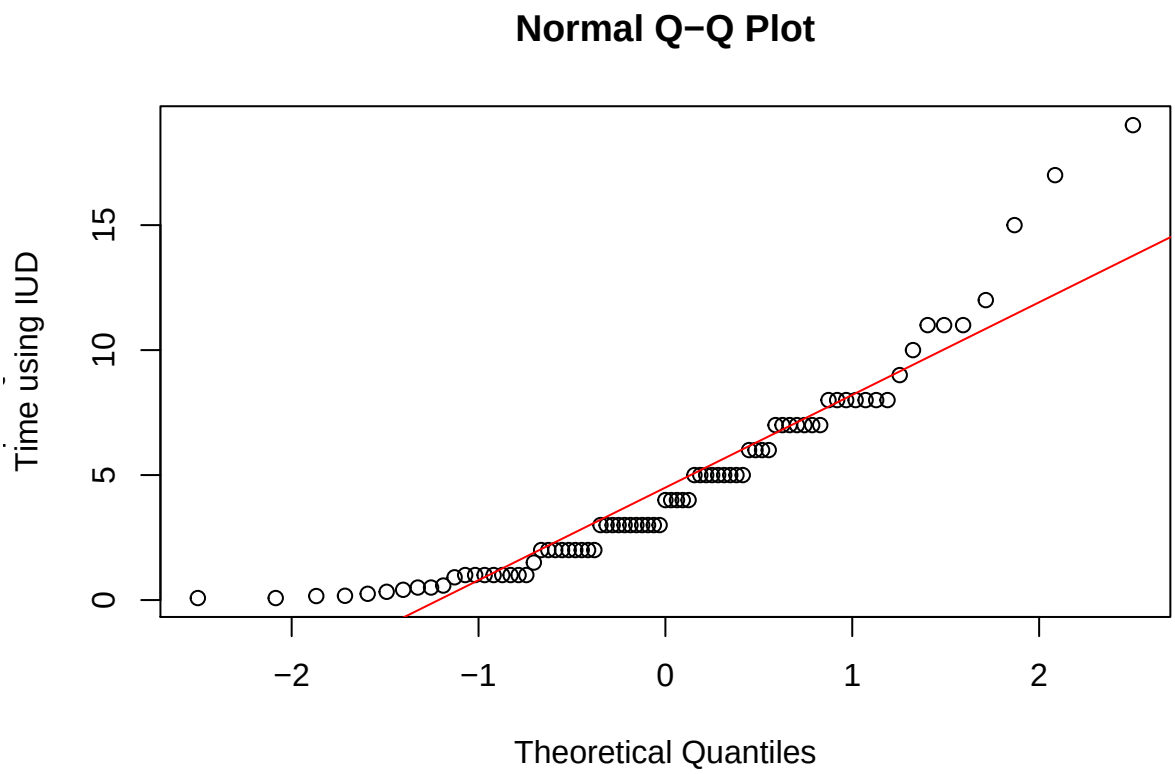


Figure 23. Distribution of the number of years using an IUD. Left - histogram. Right - boxplot.

The period of time that patients had been using an IUD varied between 0.08 and 19 years with a median

value of 4 years (Fig. 23).




```
## [1] 1.301422
```

Figure 24. The qq-plot of the IUD_years feature.

The qq-plot (Fig. 24) indicates that data for IUD_years is highly skewed. Therefore the data were transformed with a cubed root function to obtain a more gaussian distribution.

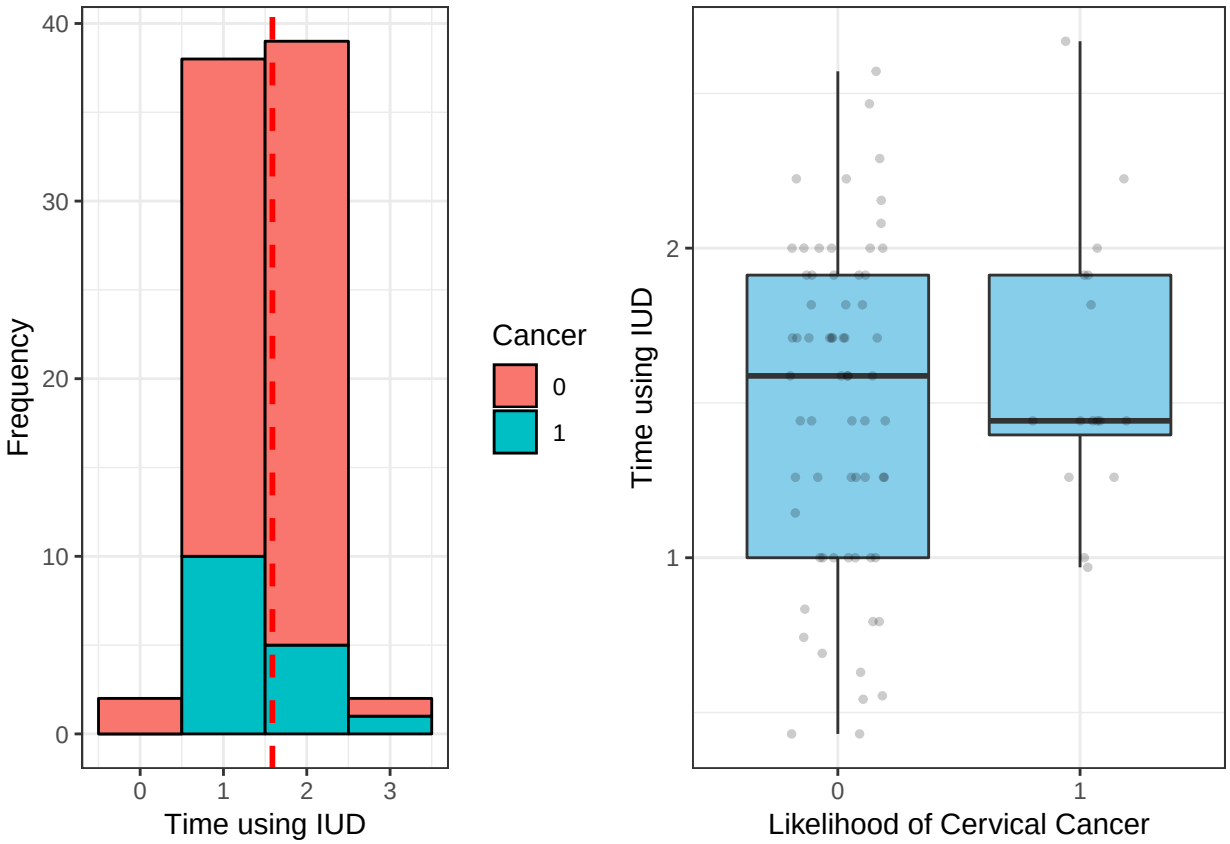
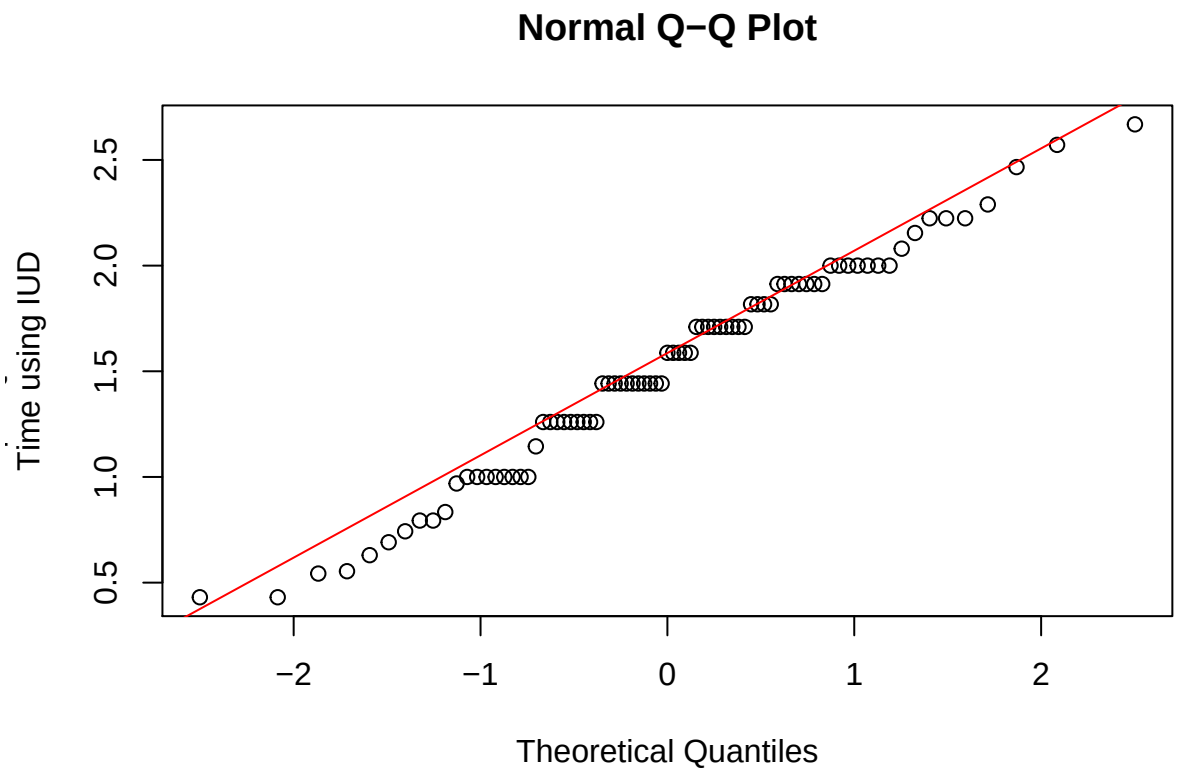


Figure 25. Distribution of cube root transformed data for the IUD_years feature. Left - histogram. Right - boxplot.

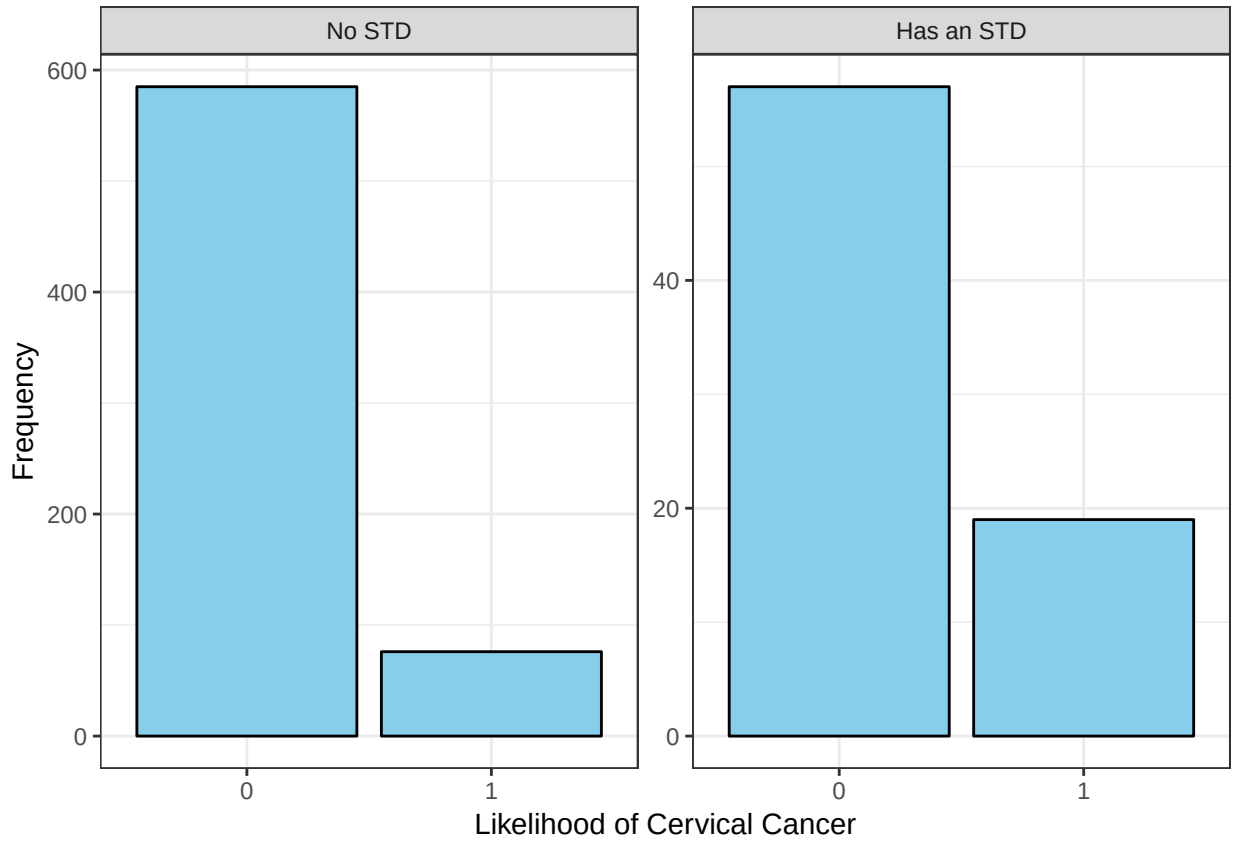



```
## [1] -0.1482602
```

Figure 26. The qq-plot for the cube root transformed data for the IUD_years feature.

1.12.13 STDs

```
dataCancer$STDs <- factor(dataCancer$STDs, levels = c("0", "1"))
labels <- c("0" = "No STD", "1" = "Has an STD")
ggplot(data = subset(dataCancer, !is.na(STDs)), aes(x = Cancer)) + geom_bar(color = "black", fill = "sky
```




```
CrossTable(dataCancer$Cancer, dataCancer$STDs)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  737
##
##
##                    | dataCancer$STDs
## dataCancer$Cancer |         0 |         1 | Row Total |
## ------------------|-----------|-----------|-----------|
##                 0 |       585 |        57 |       642 |
##                   |     0.147 |     1.279 |           |
##                   |     0.911 |     0.089 |     0.871 |
##                   |     0.885 |     0.750 |           |
##                   |     0.794 |     0.077 |           |
## ------------------|-----------|-----------|-----------|
```

```
##                   1 |        76 |        19 |        95 |
##                     |     0.994 |     8.646 |           |
##                     |     0.800 |     0.200 |     0.129 |
##                     |     0.115 |     0.250 |           |
##                     |     0.103 |     0.026 |           |
## -------------------|-----------|-----------|-----------|
##       Column Total |       661 |        76 |       737 |
##                    |     0.897 |     0.103 |           |
## -------------------|-----------|-----------|-----------|
##
##
```

Figure 27. Proportion of patients having an STD by likelihood of cervical cancer.

Less than 10 percent (60 out of 656) of the patient cohort suffered from a sexually transmitted disease (STD) (Fig. 27). The data indicate that a higher proportion of those having a STD also exhibited clinical signs of cervical cancer.

1.12.14 Number of STDs

```
STDs_number_nz <- dataCancer %>% filter(STDs_number > 0)
p15 <- ggplot(STDs_number_nz, aes(x=STDs_number, fill = Cancer)) + geom_histogram(binwidth = 1, color =
p16 <- ggplot(STDs_number_nz, aes(Cancer, STDs_number)) + geom_boxplot(fill="skyblue") + xlab("Likeliho
grid.arrange(p15, p16, nrow = 1)
```

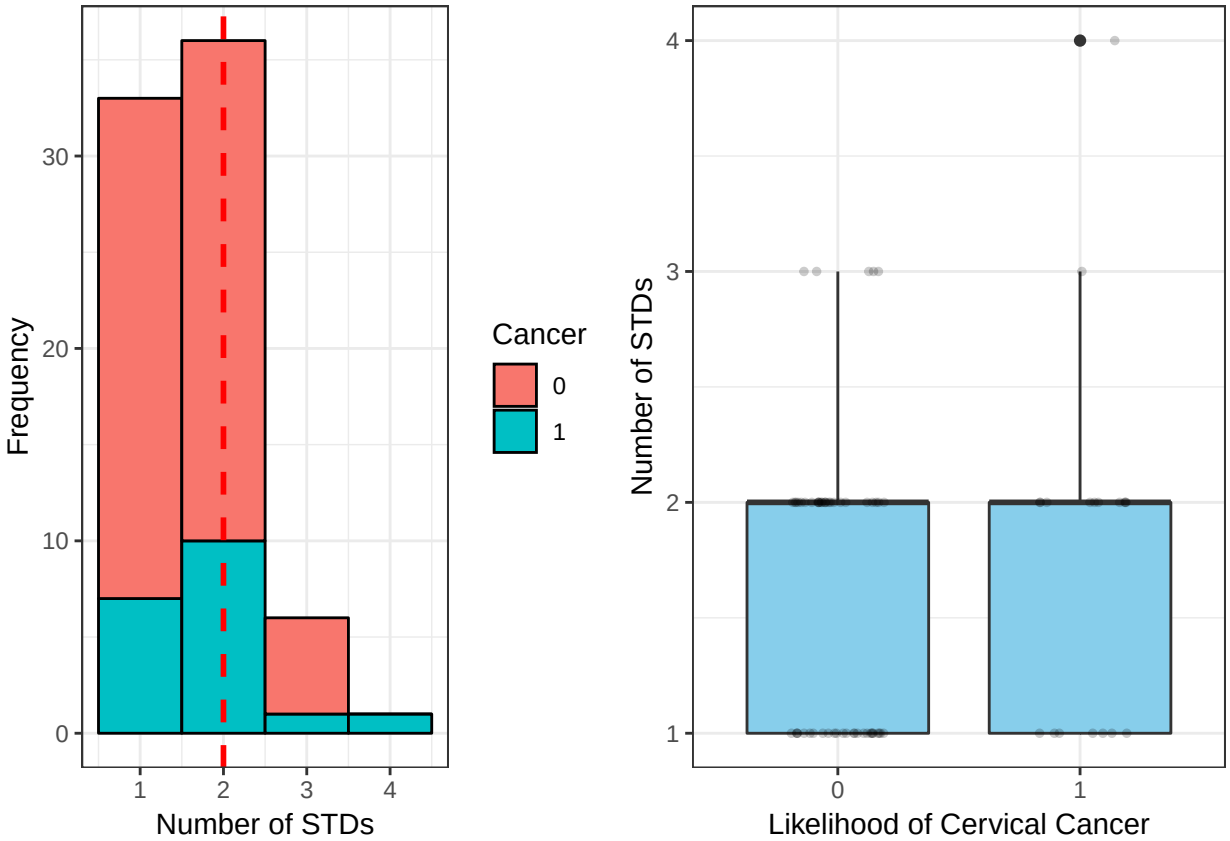


Figure 28. Relationship between number of STDs and the likelihood of exhibiting clinical signs of cervical cancer. Left - histogram. Right - boxplot.

Table 4: Table 4. Feature Summary before Data Preprocessing

name	type	na	mean	disp	median	mad	min	max	nlevs
Age	numeric	0	26.99050207	8.8647650	26.0000000	8.8956	13	51.000000	0
Num_sexual_partners	numeric	0	2.4301221	1.3375083	2.0000000	1.4826	0	8.000000	0
First_sexual_intercourse	integer	0	17.08955222	8.8499832	17.0000000	2.9652	10	32.000000	0
Num_pregnancies	numeric	0	2.1818182	1.5365282	2.0000000	1.4826	0	11.000000	0
Smokes	factor	0	NA	0.1438263	NA	NA	106	631.000000	2
Smokes_years	numeric	0	1.2223388	4.1403498	0.0000000	0.0000	0	37.000000	0
Smokes_packs_year	numeric	0	0.1678524	0.4638158	0.0000000	0.0000	0	2.802039	0
Hormonal_Contraceptives	factor	0	NA	0.3690638	NA	NA	272	465.000000	2
Hormonal_Contraceptives_years	numeric	0	0.8151810	0.7711324	0.7488872	1.1103	0	2.802039	0
IUD	factor	0	NA	0.1099050	NA	NA	81	656.000000	2
IUD_years	numeric	0	0.1664169	0.5026574	0.0000000	0.0000	0	2.668402	0
STDs	factor	0	NA	0.1031208	NA	NA	76	661.000000	2
STDs_number	integer	0	0.1723202	0.5530490	0.0000000	0.0000	0	4.000000	0
STDs_condylomatosis	factor	0	NA	0.0569878	NA	NA	42	695.000000	2
STDs_vulvo_perineal_condylomatosis	factor	0	NA	0.0556309	NA	NA	41	696.000000	2
STDs_syphilis	factor	0	NA	0.0230665	NA	NA	17	720.000000	2
STDs_HIV	factor	0	NA	0.0230665	NA	NA	17	720.000000	2
STDs_Number_diagnosis	integer	0	0.0963365	0.3131192	0.0000000	0.0000	0	3.000000	0
Dx_Cancer	factor	0	NA	0.0230665	NA	NA	17	720.000000	2
Dx_CIN	factor	0	NA	0.0108548	NA	NA	8	729.000000	2
Dx_HPV	factor	0	NA	0.0230665	NA	NA	17	720.000000	2
Dx	factor	0	NA	0.0298507	NA	NA	22	715.000000	2
Cancer	factor	0	NA	0.1289009	NA	NA	95	642.000000	2

Patients in the cohort with STDs had between 1 and 4 different STDs with the median being 2 (Fig. 28).

```
dataCancer$STDs_condylomatosis <- factor(dataCancer$STDs_condylomatosis, levels = c("0", "1"))
dataCancer$STDs_vulvo_perineal_condylomatosis <- factor(dataCancer$STDs_vulvo_perineal_condylomatosis,
dataCancer$STDs_syphilis <- factor(dataCancer$STDs_syphilis, levels = c("0", "1"))
dataCancer$STDs_HIV <- factor(dataCancer$STDs_HIV, levels = c("0", "1"))
dataCancer$Dx_Cancer <- factor(dataCancer$Dx_Cancer, levels = c("0", "1"))
dataCancer$Dx_CIN <- factor(dataCancer$Dx_CIN, levels = c("0", "1"))
dataCancer$IUD <- factor(dataCancer$IUD, levels = c("0", "1"))
dataCancer$Dx_HPV <- factor(dataCancer$Dx_HPV, levels = c("0", "1"))
dataCancer$Dx <- factor(dataCancer$Dx, levels = c("0", "1"))
```

Missing data in features referring to contraceptives, IUDs and Num_sexual_partners were replaced by 0.

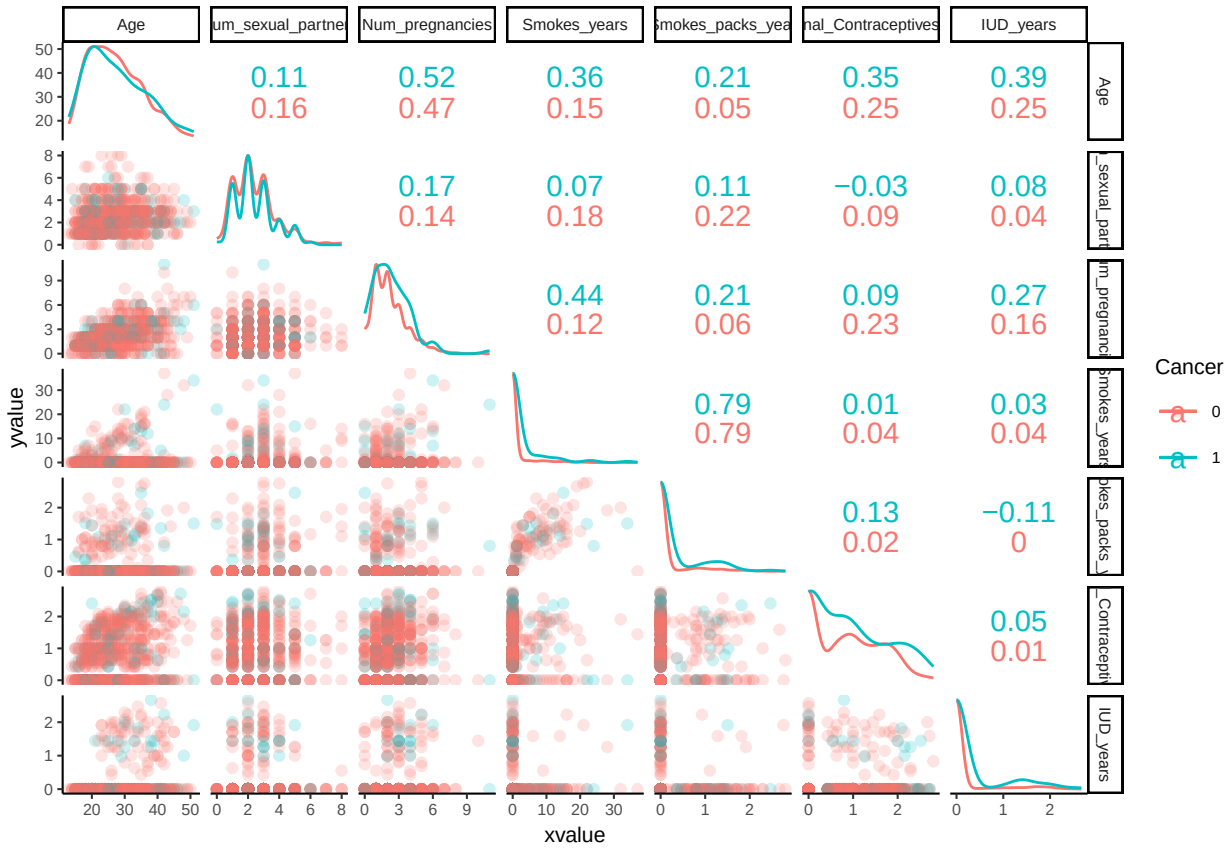


Figure 29. Scatter matrix for various variables in the cervical cancer dataset.

The scatter matrix shown in Fig. 29 indicates correlation between (1) Number of Years Smoking (Smokes_years) and Number of Packs of Cigarettes Smoked per Year (Smokes_packs_year), (2) Patient Age (Age) and Number of Pregnancies (Num_pregnancies), (3) Patient Age (Age) and Number of Years Smoking (Smokes_years), (4) Patient Age (Age) and Number of Years on Contraceptives (Hormonal_Contraceptives_years), (5) Patient Age (Age) and Number of Years on IUDs (IUD_years) and (6) Number of Pregnancies (Num_pregnancies) and Number of Years Smoking (Smokes_years).

1.13 Standardisation

The data were standardized to compensate for differences in the scales used for various features.

```
## [1] -1.175171e-17
```

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

1.14 Shuffle rows prior to splitting dataset

The rows in the dataset were randomized prior to splitting into training and test sets.

1.15 Split data into training and test data (70:30).

1.16 Determine the proportion of disease category in the test and training datasets.

```
##
##         0         1
## 0.8796117 0.1203883
```

```
##
##         0         1
## 0.8513514 0.1486486
```

The datasets were reasonably balanced and representative of the full dataset. We shall use training data for modeling and test data for model evaluation.

2 Modeling with all features

The data were initially modeled using all 33 features.

2.1 Model configuration

2.1.1 Configure classification task

2.1.2 Configure learners with probability type

```
## Loading required package: kknn
```

```
##
## Attaching package: 'kknn'
```

```
## The following object is masked from 'package:caret':
##
##     contr.dummy
```

2.1.3 Model fine-tuning

```
##            Type len Def   Constr Req Tunable Trafo
## laplace numeric   -   0 0 to Inf   -    TRUE     -
```

```
##                      Type  len   Def   Constr Req Tunable Trafo
## ntree             integer    -   500 1 to Inf   -    TRUE     -
## mtry              integer    -     - 1 to Inf   -    TRUE     -
## replace           logical    -  TRUE        -   -    TRUE     -
## classwt     numericvector <NA>     - 0 to Inf   -    TRUE     -
## cutoff      numericvector <NA>     -   0 to 1   -    TRUE     -
## strata            untyped    -     -        -   -   FALSE     -
## sampsize    integervector <NA>     - 1 to Inf   -    TRUE     -
## nodesize          integer    -     1 1 to Inf   -    TRUE     -
## maxnodes          integer    -     - 1 to Inf   -    TRUE     -
## importance        logical    - FALSE        -   -    TRUE     -
## localImp          logical    - FALSE        -   -    TRUE     -
## proximity         logical    - FALSE        -   -   FALSE     -
```

```
## oob.prox          logical    -     -        -   Y   FALSE     -
## norm.votes        logical    -  TRUE        -   -   FALSE     -
## do.trace          logical    - FALSE        -   -   FALSE     -
## keep.forest       logical    -  TRUE        -   -   FALSE     -
## keep.inbag        logical    - FALSE        -   -   FALSE     -
```

```
##              Type len     Def                                    Constr Req
## k         integer   -       7                                  1 to Inf   -
## distance  numeric   -       2                                  0 to Inf   -
## kernel   discrete   - optimal rectangular,triangular,epanechnikov,b...   -
## scale     logical   -    TRUE                                          -   -
##          Tunable Trafo
## k           TRUE     -
## distance    TRUE     -
## kernel      TRUE     -
## scale       TRUE     -
```

For naive bayes we fine-tuned the Laplace parameter testing values between 0 and 30.

For random forest we fine-tuned mtry testing values of 2, 3, 4 (the number of variables randomly sampled as candidates at each split). Breiman, L. (2001), Random Forests, Machine Learning 45(1), 5-32.

2.1.4 Configure the tune control search and a 5-CV stratified sampling scheme

2.1.5 Configure the tune wrapper with tune-tuning settings

2.2 Model Training

2.2.1 Train the tune wrappers

```
## [Tune] Started tuning learner classif.naiveBayes for parameter set:
```

```
##            Type len Def  Constr Req Tunable Trafo
## laplace numeric   -   - 0 to 30   -    TRUE     -
```

```
## With control class: TuneControlGrid
```

```
## Imputation value: 1
```

```
## [Tune-x] 1: laplace=0
```

```
## [Tune-y] 1: mmce.test.mean=0.1923355; time: 0.0 min
```

```
## [Tune-x] 2: laplace=3.33
```

```
## [Tune-y] 2: mmce.test.mean=0.2077758; time: 0.0 min
```

```
## [Tune-x] 3: laplace=6.67
```

```
## [Tune-y] 3: mmce.test.mean=0.2039113; time: 0.0 min
```

```
## [Tune-x] 4: laplace=10
```

```
## [Tune-y] 4: mmce.test.mean=0.2039113; time: 0.0 min
```

```
## [Tune-x] 5: laplace=13.3
```

```
## [Tune-y] 5: mmce.test.mean=0.2019505; time: 0.0 min
```

```
## [Tune-x] 6: laplace=16.7
```

```
## [Tune-y] 6: mmce.test.mean=0.2038923; time: 0.0 min
```

```
## [Tune-x] 7: laplace=20

## [Tune-y] 7: mmce.test.mean=0.2019879; time: 0.0 min

## [Tune-x] 8: laplace=23.3

## [Tune-y] 8: mmce.test.mean=0.2019879; time: 0.0 min

## [Tune-x] 9: laplace=26.7

## [Tune-y] 9: mmce.test.mean=0.2000461; time: 0.0 min

## [Tune-x] 10: laplace=30

## [Tune-y] 10: mmce.test.mean=0.2000461; time: 0.0 min

## [Tune] Result: laplace=0 : mmce.test.mean=0.1923355

## [Tune] Started tuning learner classif.randomForest for parameter set:

##          Type len Def Constr Req Tunable Trafo
## mtry discrete   -   -  2,3,4   -    TRUE     -

## With control class: TuneControlGrid

## Imputation value: 1

## [Tune-x] 1: mtry=2

## [Tune-y] 1: mmce.test.mean=0.1242576; time: 0.0 min

## [Tune-x] 2: mtry=3

## [Tune-y] 2: mmce.test.mean=0.1262184; time: 0.0 min

## [Tune-x] 3: mtry=4

## [Tune-y] 3: mmce.test.mean=0.1281605; time: 0.0 min

## [Tune] Result: mtry=2 : mmce.test.mean=0.1242576

## [Tune] Started tuning learner classif.kknn for parameter set:

##       Type len Def                                    Constr Req Tunable
## k discrete   -   - 2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,...   -    TRUE
##   Trafo
## k     -

## With control class: TuneControlGrid

## Imputation value: 1

## [Tune-x] 1: k=2

## [Tune-y] 1: mmce.test.mean=0.1902605; time: 0.0 min

## [Tune-x] 2: k=3

## [Tune-y] 2: mmce.test.mean=0.1844162; time: 0.0 min

## [Tune-x] 3: k=4

## [Tune-y] 3: mmce.test.mean=0.1863580; time: 0.0 min

## [Tune-x] 4: k=5

## [Tune-y] 4: mmce.test.mean=0.1457120; time: 0.0 min

## [Tune-x] 5: k=6
```

```
## [Tune-y] 5: mmce.test.mean=0.1418281; time: 0.0 min
## [Tune-x] 6: k=7
## [Tune-y] 6: mmce.test.mean=0.1398864; time: 0.0 min
## [Tune-x] 7: k=8
## [Tune-y] 7: mmce.test.mean=0.1321384; time: 0.0 min
## [Tune-x] 8: k=9
## [Tune-y] 8: mmce.test.mean=0.1282549; time: 0.0 min
## [Tune-x] 9: k=10
## [Tune-y] 9: mmce.test.mean=0.1223916; time: 0.0 min
## [Tune-x] 10: k=11
## [Tune-y] 10: mmce.test.mean=0.1223726; time: 0.0 min
## [Tune-x] 11: k=12
## [Tune-y] 11: mmce.test.mean=0.1243143; time: 0.0 min
## [Tune-x] 12: k=13
## [Tune-y] 12: mmce.test.mean=0.1204118; time: 0.0 min
## [Tune-x] 13: k=14
## [Tune-y] 13: mmce.test.mean=0.1223535; time: 0.0 min
## [Tune-x] 14: k=15
## [Tune-y] 14: mmce.test.mean=0.1223158; time: 0.0 min
## [Tune-x] 15: k=16
## [Tune-y] 15: mmce.test.mean=0.1203927; time: 0.0 min
## [Tune-x] 16: k=17
## [Tune-y] 16: mmce.test.mean=0.1203927; time: 0.0 min
## [Tune-x] 17: k=18
## [Tune-y] 17: mmce.test.mean=0.1203927; time: 0.0 min
## [Tune-x] 18: k=19
## [Tune-y] 18: mmce.test.mean=0.1203927; time: 0.0 min
## [Tune-x] 19: k=20
## [Tune-y] 19: mmce.test.mean=0.1223345; time: 0.0 min
## [Tune] Result: k=17 : mmce.test.mean=0.1203927
```

2.3 Model Prediction

2.3.1 Predict on training data

2.4 Model Evaluation

2.4.1 Obtain threshold values for each learner

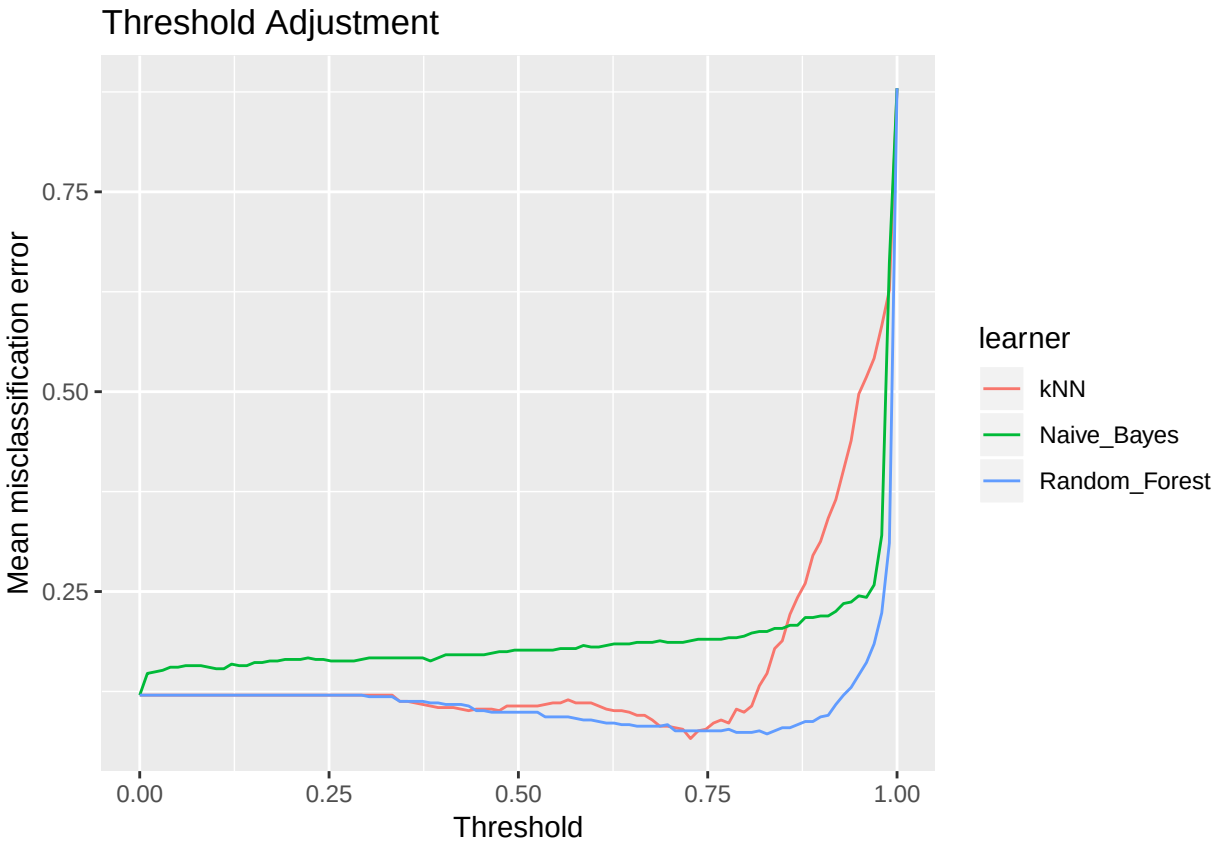


Figure 30. Plot for the optimization of the threshold for the kNn, Naive Bayes and Random Forest classifiers trained on all 23 features.

Table 5: Table 5. Performance for Naive Bayes, Random Forest and kNN classifiers

	Naive Bayes	Random Forest	kNN
mmce	0.2117117	0.1486486	0.1711712
auc	0.5661376	0.6110309	0.5957993

2.5 Evaluation on test data

2.5.1 AUC plots for Naive Bayes, Random Forest and kNN models

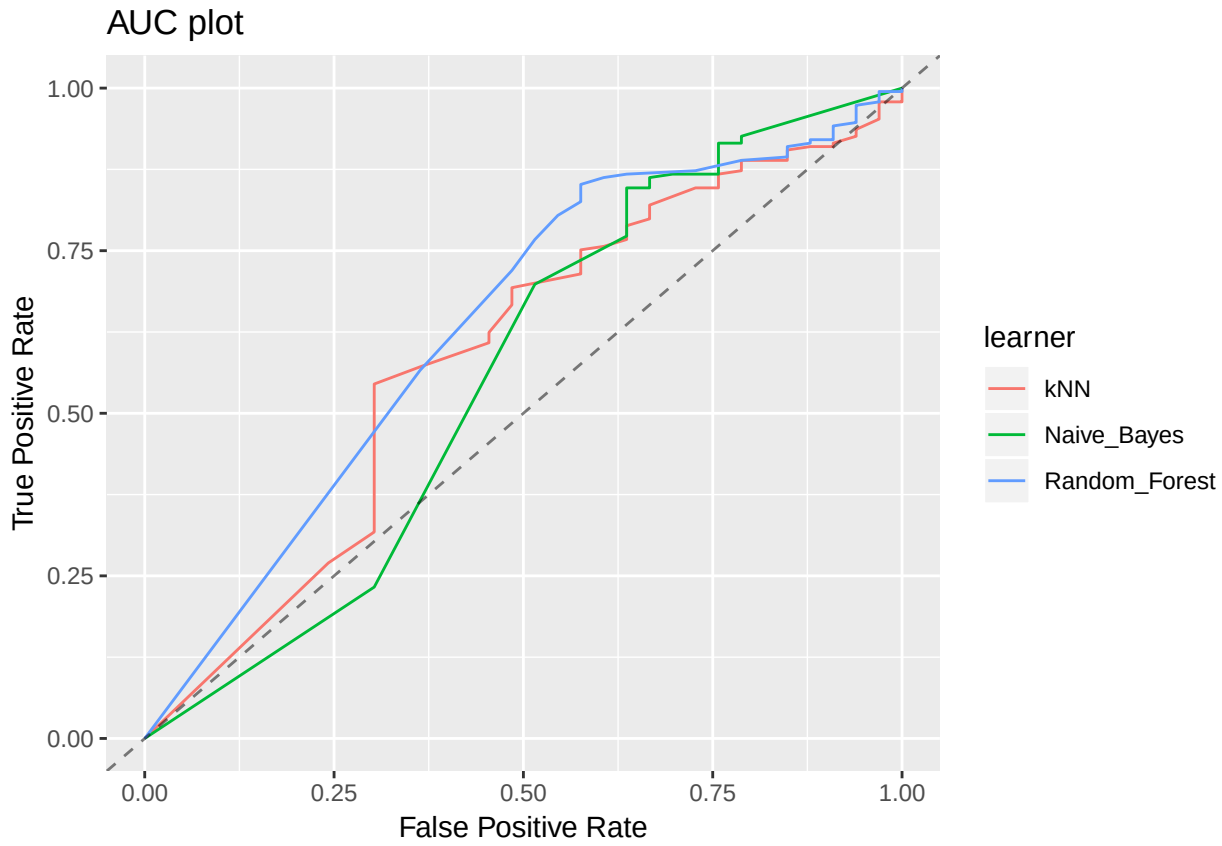


Figure 31. AUC plot for the kNN, Naive Bayes and Random Forest classifers trained on all 23 features.

The AUC plots were similar for the kNN and Random Forest classifiers trained on all 23 features and were significantly better than the Naive Bayes classifier (Fig. 31).

2.5.2 Performance for Naive Bayes model

2.5.2.1 Misclassification Error and AUC value

The Raandom Forest classifier performed the best, when the models were trained on all 23 features, with a mean misclassification error of 0.149 and auv value of 0.611 (Table 5). While the KNN classifier performed slightly better than the Naive Bayes model.

2.5.2.2 Confusion Matrix, Precision and Recall for Naive Bayes

```
##     predicted
## true 0         1
##    0 167       22        tpr: 0.88 fnr: 0.12
##    1 25        8         fpr: 0.76 tnr: 0.24
##      ppv: 0.87 for: 0.73 lrp: 1.17 acc: 0.79
##      fdr: 0.13 npv: 0.27 lrm: 0.48 dor: 2.43
##
##
## Abbreviations:
## tpr - True positive rate (Sensitivity, Recall)
## fpr - False positive rate (Fall-out)
## fnr - False negative rate (Miss rate)
## tnr - True negative rate (Specificity)
## ppv - Positive predictive value (Precision)
## for - False omission rate
## lrp - Positive likelihood ratio (LR+)
## fdr - False discovery rate
## npv - Negative predictive value
## acc - Accuracy
## lrm - Negative likelihood ratio (LR-)
## dor - Diagnostic odds ratio
```

2.5.2.3 Confusion Matrix, Precision and Recall for Random Forest model

```
##     predicted
## true 0         1
##    0 188       1         tpr: 0.99 fnr: 0.01
##    1 32        1         fpr: 0.97 tnr: 0.03
##      ppv: 0.85 for: 0.5 lrp: 1.03 acc: 0.85
##      fdr: 0.15 npv: 0.5 lrm: 0.17 dor: 5.88
##
##
## Abbreviations:
## tpr - True positive rate (Sensitivity, Recall)
## fpr - False positive rate (Fall-out)
## fnr - False negative rate (Miss rate)
## tnr - True negative rate (Specificity)
## ppv - Positive predictive value (Precision)
## for - False omission rate
## lrp - Positive likelihood ratio (LR+)
## fdr - False discovery rate
## npv - Negative predictive value
## acc - Accuracy
## lrm - Negative likelihood ratio (LR-)
## dor - Diagnostic odds ratio
```

2.5.2.4 Confusion Matrix, Precision and Recall for k-Nearest Neighbours model

```
##     predicted
## true 0         1
##    0 183       6         tpr: 0.97 fnr: 0.03
##    1 32        1         fpr: 0.97 tnr: 0.03
##      ppv: 0.85 for: 0.86 lrp: 1    acc: 0.83
##      fdr: 0.15 npv: 0.14 lrm: 1.05 dor: 0.95
```

```
##
##
## Abbreviations:
## tpr - True positive rate (Sensitivity, Recall)
## fpr - False positive rate (Fall-out)
## fnr - False negative rate (Miss rate)
## tnr - True negative rate (Specificity)
## ppv - Positive predictive value (Precision)
## for - False omission rate
## lrp - Positive likelihood ratio (LR+)
## fdr - False discovery rate
## npv - Negative predictive value
## acc - Accuracy
## lrm - Negative likelihood ratio (LR-)
## dor - Diagnostic odds ratio
```

3 Modeling with spFSR Feature Selection

Various feature selection algorithms were then used to determine whether classifier performance could be improved by using a subset of relevant features. Features were initially selected using the spFRS feature selection algorithm (https://cran.r-project.org/web/packages/spFSR/vignettes/spFSR.html). (**???**)

3.1 spFSR Feature Selection for k-Nearest Neighbours

```
## [1] 23
```

3.1.1 Determine if hyperparameters can be optimized further

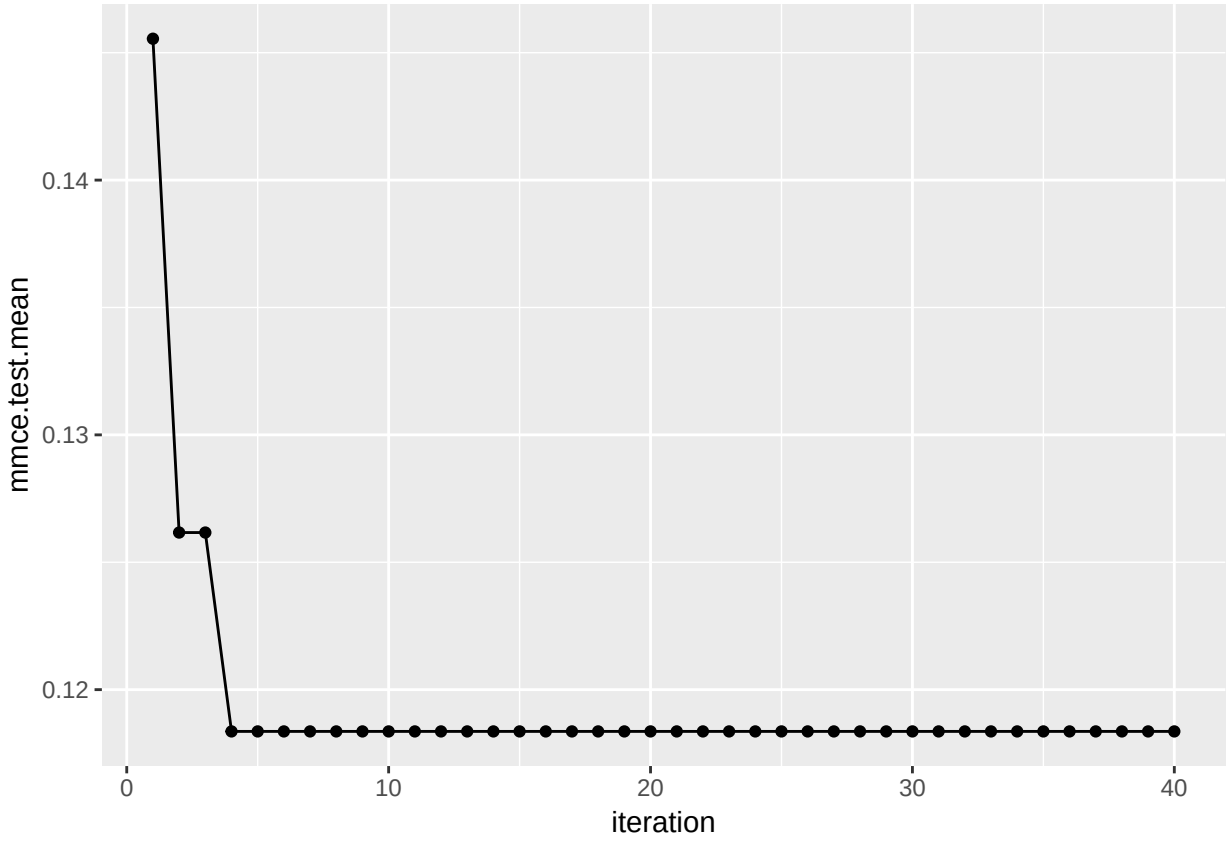


Figure 32. Measurement of kNN classifier performance with each iteration.

The plot shows mmce is dcreasing suggesting that the parameters can be optimized further.

3.1.2 Best hyperparameters

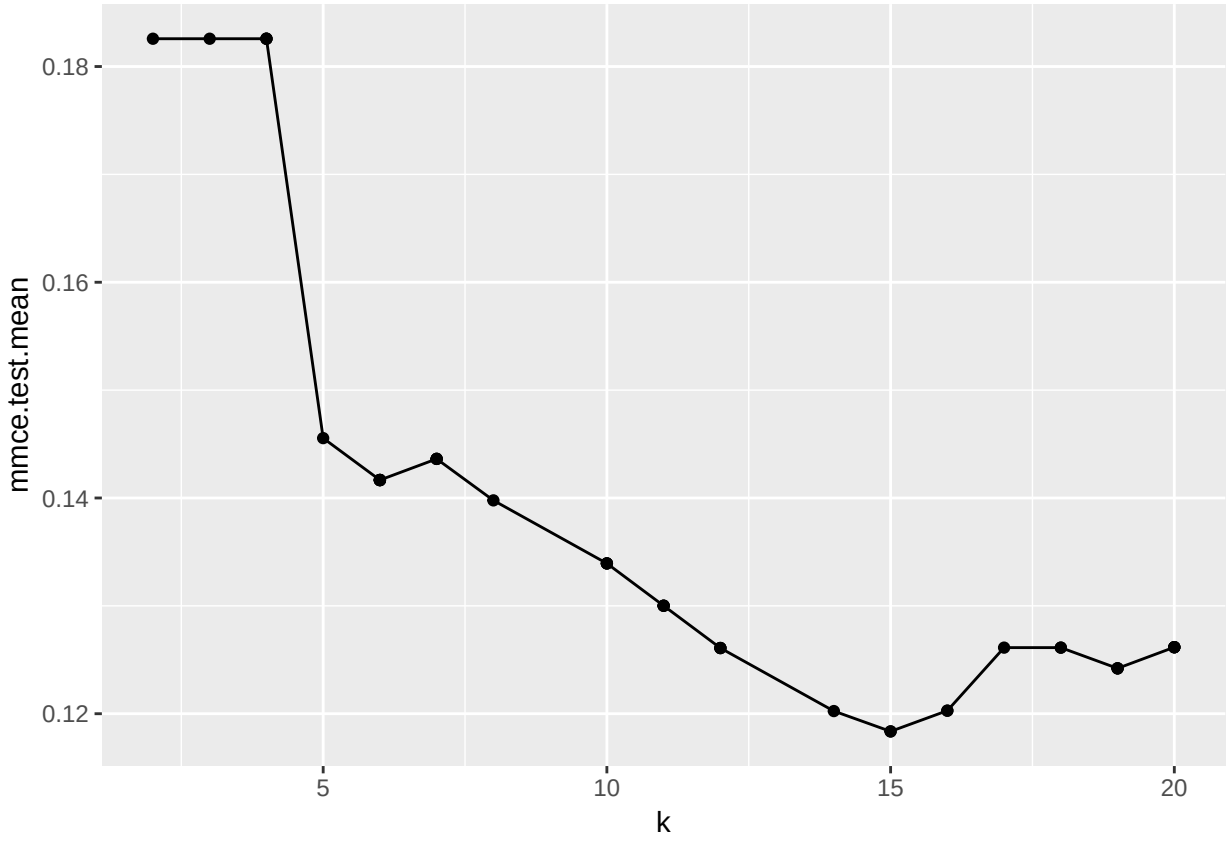


Figure 33. kNN classifier performance for various values of k (nearest neighbours value).

The plot shows the global optimum and indicates that the best performance was obtained in about 15 iterations.

```
## $k
## [1] 15
```

```
## mmce.test.mean
##      0.1183613
```

The optimal value of k for k-Nearest Neighbours was 15 with a mean misclassification error of 0.118.

3.1.3 Construct a learner with the tuned parameters

3.1.4 Best hyperparameters

```
## Learner classif.kknn from package kknn
## Type: classif
## Name: k-Nearest Neighbor; Short name: kknn
## Class: classif.kknn
## Properties: twoclass,multiclass,numerics,factors,prob
## Predict-Type: prob
## Hyperparameters: k=15
```

3.1.5 Run spsa on tuned and not-tuned learner

```
## SPSA-FSR begins:

## Wrapper = kknn

## Measure = auc

## Number of selected features = 0

##
## iter  value   st.dev  num.ft  best.value
## 1     0.66652 0.08724 13      0.66652 *
## 2     0.67873 0.07876 15      0.67873 *
## 3     0.65068 0.0959  11      0.67873
## 4     0.69214 0.0815  12      0.69214 *
## 5     0.69351 0.05994 12      0.69351 *
## 6     0.70517 0.06063 14      0.70517 *
## 7     0.69378 0.09697 14      0.70517
## 8     0.70983 0.08844 14      0.70983 *
## 9     0.72798 0.07736 16      0.72798 *
## 10    0.69858 0.09898 17      0.72798
## 11    0.70915 0.09041 17      0.72798
## 12    0.6678  0.06979 16      0.72798
## 13    0.71376 0.07237 14      0.72798
## 14    0.69546 0.09024 16      0.72798
## 15    0.71057 0.11709 16      0.72798
## 16    0.72438 0.07684 15      0.72798
## 17    0.71411 0.08948 15      0.72798
## 18    0.71586 0.07034 15      0.72798
## 19    0.71477 0.07681 13      0.72798
## 20    0.67661 0.07696 14      0.72798
## 21    0.69783 0.0881  13      0.72798
## 22    0.70833 0.05328 13      0.72798
## 23    0.70509 0.0787  13      0.72798
## 24    0.70275 0.07714 13      0.72798
## 25    0.69688 0.0619  13      0.72798
## 26    0.69833 0.05764 13      0.72798
## 27    0.70386 0.04867 14      0.72798
## 28    0.70727 0.08117 14      0.72798
## 29    0.71183 0.05808 16      0.72798

##
## Best iteration = 9

## Number of selected features = 16

## Best measure value = 0.72798

## Std. dev. of best measure = 0.07736

## Run time = 0.71 minutes.

## SPSA-FSR begins:

## Wrapper = kknn

## Measure = auc

## Number of selected features = 0
```

```
##
## iter  value   st.dev  num.ft  best.value
## 1     0.68515 0.04869 14      0.68515 *
## 2     0.69978 0.07572 16      0.69978 *
## 3     0.67652 0.08389 15      0.69978
## 4     0.64775 0.0809  16      0.69978
## 5     0.67645 0.08247 15      0.69978
## 6     0.6575  0.05673 14      0.69978
## 7     0.62584 0.0885  16      0.69978
## 8     0.61483 0.08257 15      0.69978
## 9     0.70581 0.07028 14      0.70581 *
## 10    0.7031  0.06375 16      0.70581
## 11    0.69397 0.05313 15      0.70581
## 12    0.68069 0.0735  17      0.70581
## 13    0.69909 0.06205 12      0.70581
## 14    0.6537  0.102   17      0.70581
## 15    0.61696 0.09682 17      0.70581
## 16    0.66788 0.06852 17      0.70581
## 17    0.7006  0.06328 17      0.70581
## 18    0.68899 0.05968 17      0.70581
## 19    0.68768 0.06092 18      0.70581
## 20    0.69026 0.06691 16      0.70581
## 21    0.70981 0.06813 17      0.70981 *
## 22    0.71647 0.06136 17      0.71647 *
## 23    0.68643 0.07839 15      0.71647
## 24    0.71347 0.06054 15      0.71647
## 25    0.64659 0.07911 17      0.71647
## 26    0.6628  0.07949 17      0.71647
## 27    0.70956 0.08768 15      0.71647
## 28    0.7131  0.08943 14      0.71647
## 29    0.73374 0.05887 12      0.73374 *
## 30    0.71075 0.05638 12      0.73374
## 31    0.71202 0.06512 13      0.73374
## 32    0.71627 0.05382 12      0.73374
## 33    0.69943 0.09905 12      0.73374
## 34    0.71001 0.07922 11      0.73374
## 35    0.73724 0.04251 12      0.73724 *
## 36    0.73142 0.05276 12      0.73724
## 37    0.70352 0.06059 12      0.73724
## 38    0.71149 0.04781 12      0.73724
## 39    0.7304  0.06965 12      0.73724
## 40    0.72522 0.06263 12      0.73724
## 41    0.67496 0.08656 13      0.73724
## 42    0.69286 0.07266 14      0.73724
## 43    0.70409 0.07992 14      0.73724
## 44    0.67695 0.06096 15      0.73724
## 45    0.69201 0.07691 14      0.73724
## 46    0.71119 0.09074 14      0.73724
## 47    0.7261  0.05982 14      0.73724
## 48    0.71517 0.07737 14      0.73724
## 49    0.69535 0.08876 14      0.73724
## 50    0.70618 0.07714 14      0.73724
## 51    0.72119 0.08025 14      0.73724
## 52    0.70999 0.07289 15      0.73724
```

```
## 53    0.70839 0.06258 15      0.73724
## 54    0.71485 0.03701 16      0.73724
## 55    0.70488 0.08203 15      0.73724
##
## Best iteration = 35
## Number of selected features = 12
## Best measure value = 0.73724
## Std. dev. of best measure = 0.04251
## Run time = 1.2 minutes.
```

3.1.6 Plot the measure values across iterations

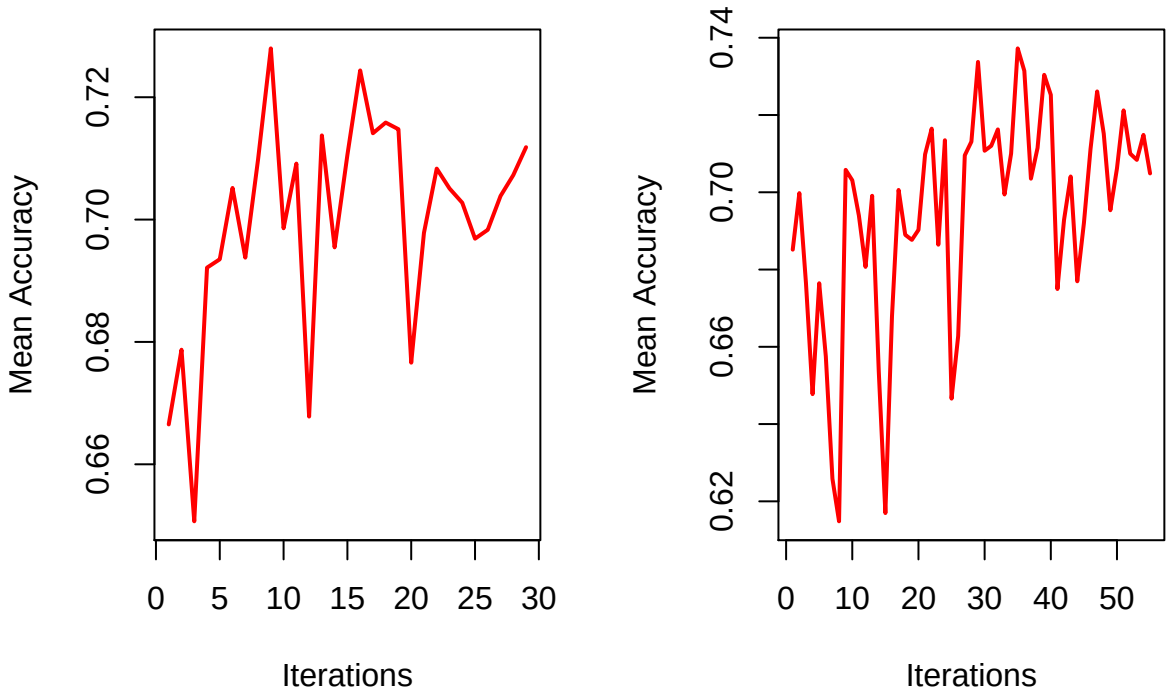


Figure 34. Scatter plot of mean accuracy rate for each iteration for the kNN classifier (left panel) with and (right panel) without hyperparameter tuning.

3.1.7 Get feature importances for kNN

3.1.8 Get feature importances for kNN with tuned parameters

The top 10 features selected by the spFSR algorithm for the kNN classified with and without tuned parameters differed significantly indicating that tuning the hyperparameters for the kNN classifier made a significant difference.

Table 6: Table 6. Selected Features for k-Nearest Neighbours calssifier with and without spFSR feature selection

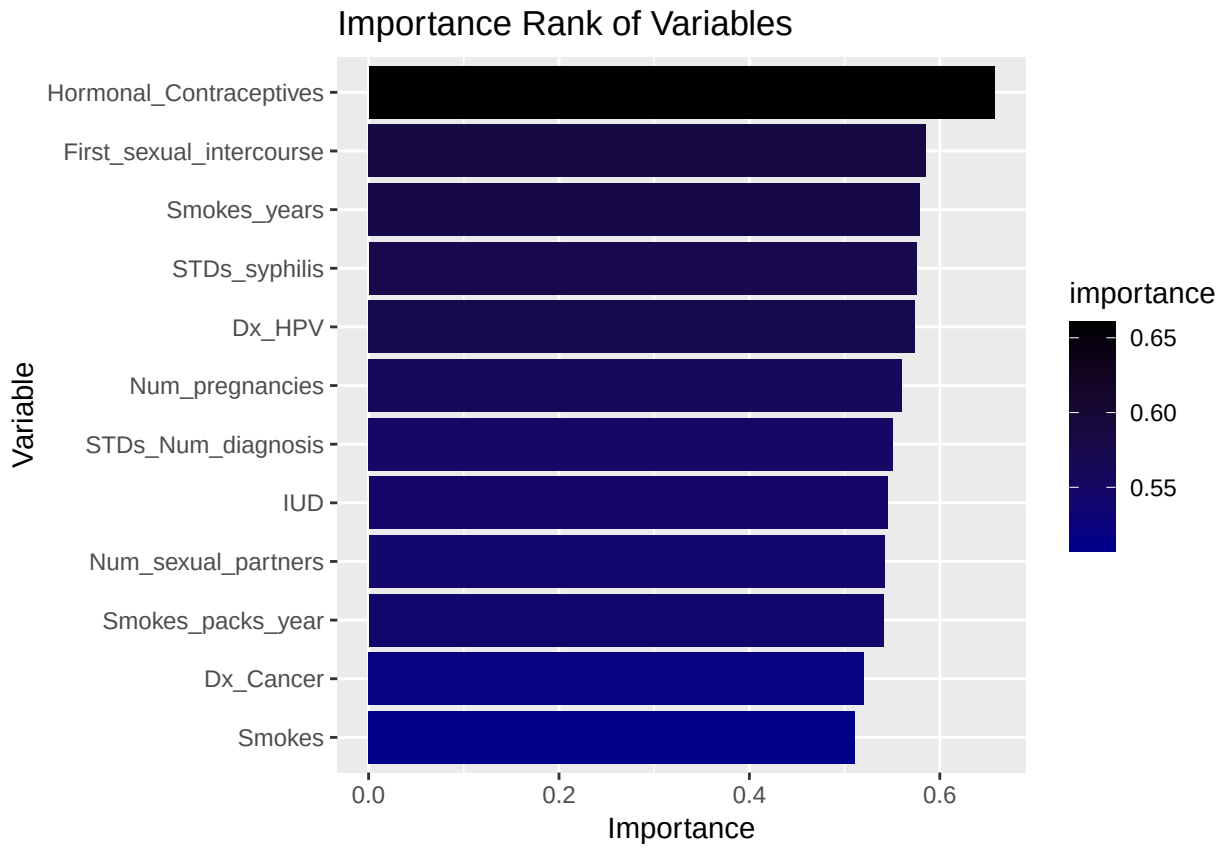
kNN + spFSR		kNN + spFSR + Tuned	
Features	Importance	Features	Importance
Hormonal_Contraceptives	0.65736	STDs_HIV	1.00000
First_sexual_intercourse	0.58572	Dx_HPV	1.00000
Smokes_years	0.57904	Num_sexual_partners	0.97502
STDs_syphilis	0.57561	STDs_syphilis	0.95608
Dx_HPV	0.57364	Dx	0.87179
Num_pregnancies	0.56038	Dx_CIN	0.86142
STDs_Num_diagnosis	0.55085	Dx_Cancer	0.81635
IUD	0.54519	Age	0.81357
Num_sexual_partners	0.54189	STDs	0.78772
Smokes_packs_year	0.54100	STDs_Num_diagnosis	0.76885

3.1.9 Importance Plotting

3.1.9.1 kNN classifier and spFRS feature selection

```
##                    features importance
## 1                  STDs_HIV    1.00000
## 2                    Dx_HPV    1.00000
## 3       Num_sexual_partners    0.97502
## 4             STDs_syphilis    0.95608
## 5                        Dx    0.87179
## 6                    Dx_CIN    0.86142
## 7                 Dx_Cancer    0.81635
## 8                       Age    0.81357
## 9                      STDs    0.78772
## 10       STDs_Num_diagnosis    0.76885
## 11  Hormonal_Contraceptives    0.74904
## 12 First_sexual_intercourse    0.71598
## 13        Smokes_packs_year    0.69250
## 14          Num_pregnancies    0.65648
## 15                   Smokes    0.64179
## 16                IUD_years    0.57251
```

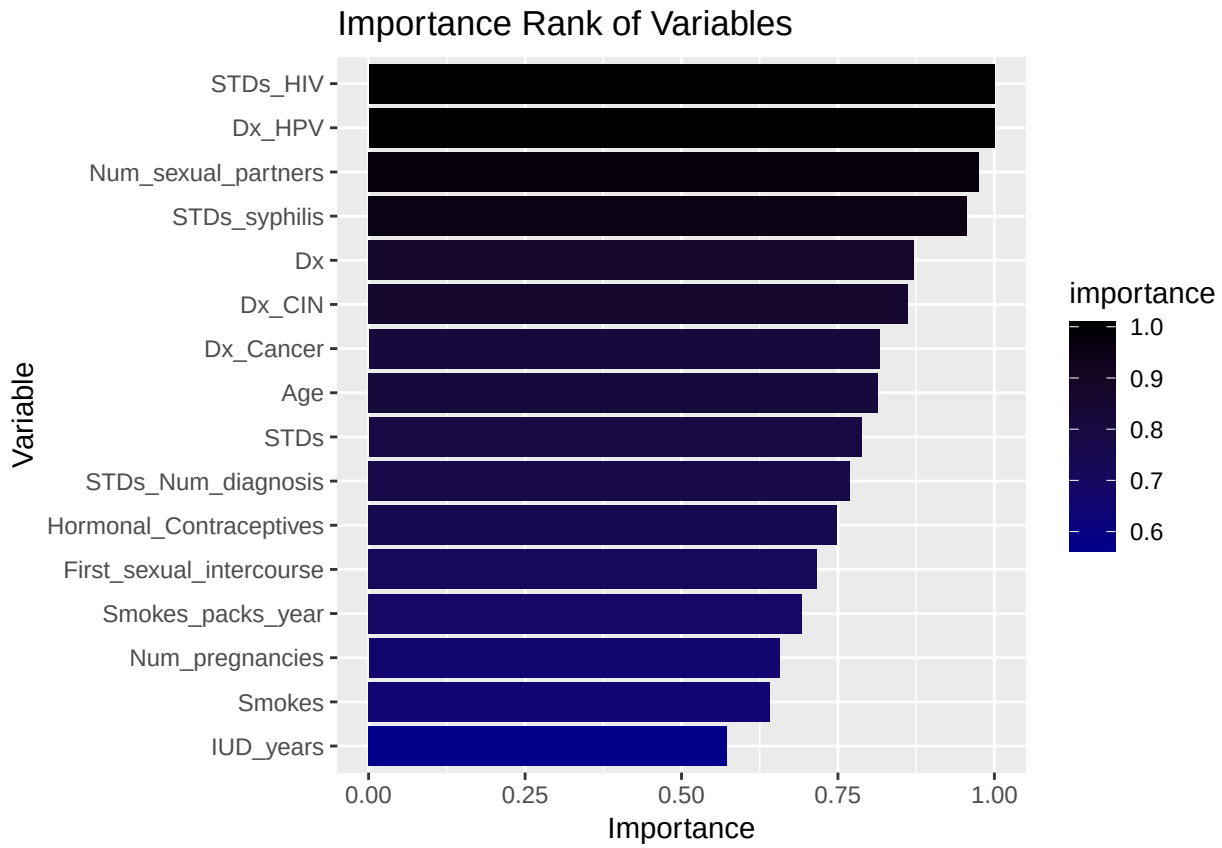
```
##                    features importance
## 1   Hormonal_Contraceptives    0.65736
## 2  First_sexual_intercourse    0.58572
## 3               Smokes_years    0.57904
## 4             STDs_syphilis    0.57561
## 5                    Dx_HPV    0.57364
## 6           Num_pregnancies    0.56038
## 7        STDs_Num_diagnosis    0.55085
## 8                       IUD    0.54519
## 9       Num_sexual_partners    0.54189
## 10        Smokes_packs_year    0.54100
## 11                Dx_Cancer    0.52025
## 12                   Smokes    0.51099
```



Figures 35. Features selected by the spFRS algorithm fused with the kNN classifier and ranked according to importance.

3.1.9.2 kNN classifier with tuned hyperparameters and spFRS feature selection

```
spsa_kNN_1 <- spFSR::plotImportance(spsaMod_kNN_tuned)
```



Figures 36. Features selected by the spFRS algorithm fused with the kNN classifier with tuned hyperparameters and ranked according to importance.

3.2 Train models

3.2.1 Define the test data

3.3 Prediction

3.4 Evaluation

3.4.1 AUC plots for k-nearest neighbours

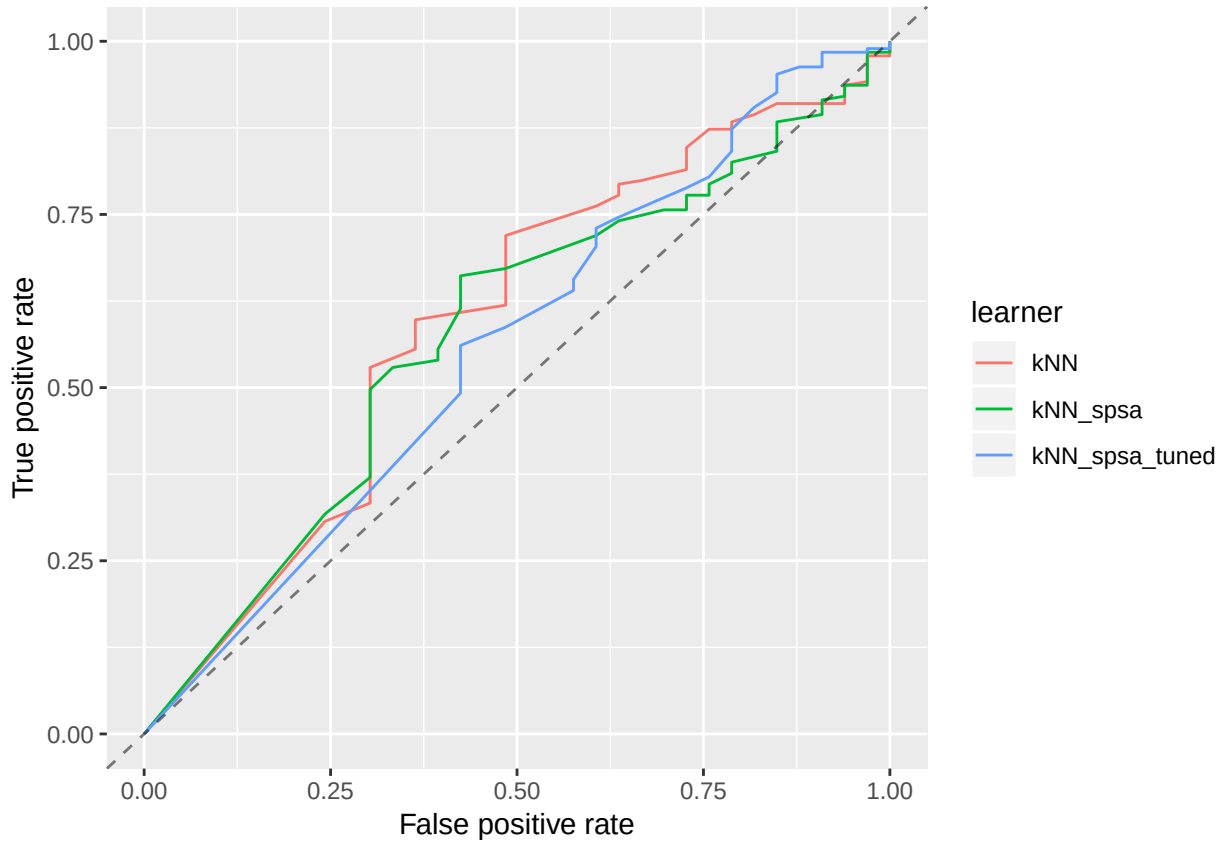


Figure 37. AUC plot for the kNN classifer. kNN - kNN trained on all 23 features; kNN_spsa - kNN classifier trained on features selected using the spFRS algorithm; kNN_spsa_tuned - kNN classifier with tuned hyperparameters and trained on features selected using the spFRS algorithm.

The AUC plots were similar for the kNN classifier without optimization of the hyperparameters and with or without spFRS feature selection while these were both better than the kNN classifier without optimization of the hyperparameters or feature selection (Fig. 37).

3.4.2 Performance for kNN model

3.4.2.1 Misclassification Error and AUC value

The kNN classifier appeared to performed better with hyperparameter optimization and spFRS feature selection with a mean misclassification error of 0.167 and AUC value of 0.560 (Table 6).

Table 7: Table 6. Performance for kNN classifier with and without tuned hyperparameters and spFRS feature selection

	kNN	kNN_spsa	kNN_spsa_tuned
mmce	0.1711712	0.1756757	0.1666667
auc	0.5910694	0.5715889	0.5603656

3.4.2.2 Confusion Matrix, Precision and Recall for kNN model

```
##     predicted
## true 0         1
##    0 183       6         tpr: 0.97 fnr: 0.03
##    1 32        1         fpr: 0.97 tnr: 0.03
##      ppv: 0.85 for: 0.86 lrp: 1    acc: 0.83
##      fdr: 0.15 npv: 0.14 lrm: 1.05 dor: 0.95
##
##
## Abbreviations:
## tpr - True positive rate (Sensitivity, Recall)
## fpr - False positive rate (Fall-out)
## fnr - False negative rate (Miss rate)
## tnr - True negative rate (Specificity)
## ppv - Positive predictive value (Precision)
## for - False omission rate
## lrp - Positive likelihood ratio (LR+)
## fdr - False discovery rate
## npv - Negative predictive value
## acc - Accuracy
## lrm - Negative likelihood ratio (LR-)
## dor - Diagnostic odds ratio
```

3.4.3 Confusion Matrix, Precision and Recall for kNN with spsa feature selection

```
##     predicted
## true 0         1
##    0 182       7         tpr: 0.96 fnr: 0.04
##    1 32        1         fpr: 0.97 tnr: 0.03
##      ppv: 0.85 for: 0.88 lrp: 0.99 acc: 0.82
##      fdr: 0.15 npv: 0.12 lrm: 1.22 dor: 0.81
##
##
## Abbreviations:
## tpr - True positive rate (Sensitivity, Recall)
## fpr - False positive rate (Fall-out)
## fnr - False negative rate (Miss rate)
## tnr - True negative rate (Specificity)
## ppv - Positive predictive value (Precision)
## for - False omission rate
## lrp - Positive likelihood ratio (LR+)
## fdr - False discovery rate
## npv - Negative predictive value
## acc - Accuracy
```

```
## lrm - Negative likelihood ratio (LR-)
## dor - Diagnostic odds ratio
```

3.4.4 Confusion Matrix, Precision and Recall for kNN with tuned parameters and spsa feature selection

```
##     predicted
## true 0         1
##    0 182       7        tpr: 0.96 fnr: 0.04
##    1 30        3        fpr: 0.91 tnr: 0.09
##      ppv: 0.86 for: 0.7 lrp: 1.06 acc: 0.83
##      fdr: 0.14 npv: 0.3 lrm: 0.41 dor: 2.6
##
##
## Abbreviations:
## tpr - True positive rate (Sensitivity, Recall)
## fpr - False positive rate (Fall-out)
## fnr - False negative rate (Miss rate)
## tnr - True negative rate (Specificity)
## ppv - Positive predictive value (Precision)
## for - False omission rate
## lrp - Positive likelihood ratio (LR+)
## fdr - False discovery rate
## npv - Negative predictive value
## acc - Accuracy
## lrm - Negative likelihood ratio (LR-)
## dor - Diagnostic odds ratio
```

3.5 spFSR Feature Selection for Random Forest

```
## [1] 23
```

3.5.1 Determine if hyperparameters can be optimized further

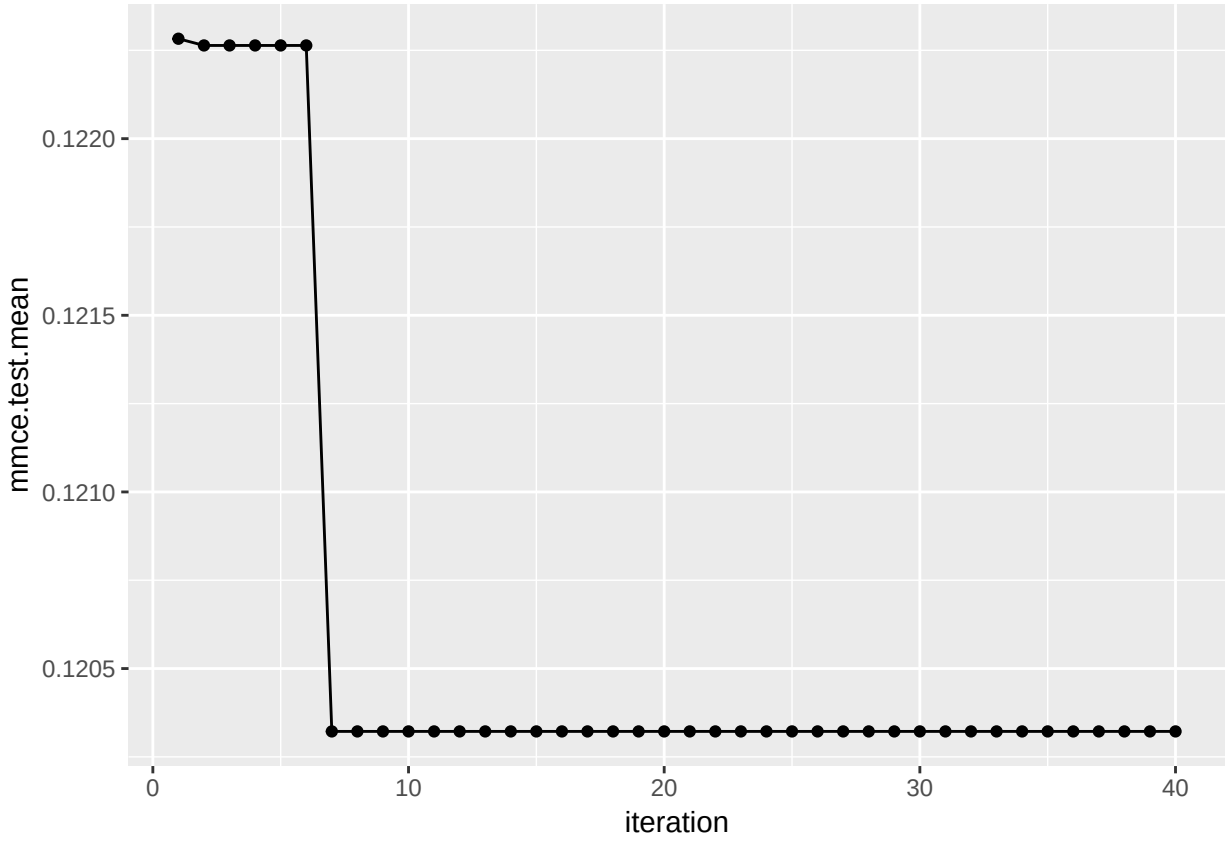


Figure 38. The plot shows that the hyperparameters can be optimized further.

3.5.2 Best hyperparameters

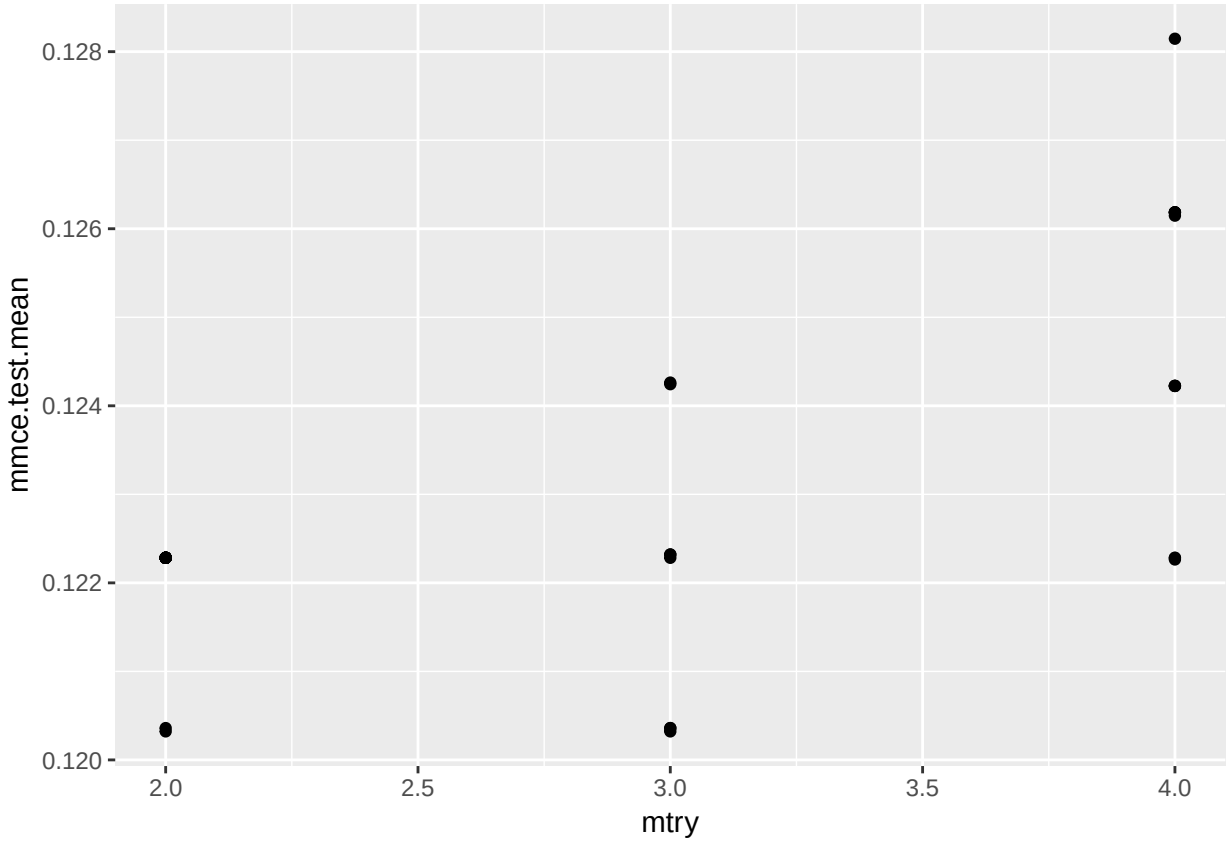


Figure 39. Optimisation of mtry hyperparameter for the random forest classifier.

3.5.3 Best hyperparameters

```
## $mtry
## [1] 3
```

```
## mmce.test.mean  fpr.test.mean  tpr.test.mean
##      0.1203221      0.9371795      0.9912088
```

The optimal of mtry was 3.0 with mean misclassification error of 0.120.

3.5.4 Construct a learner with the tuned parameters

3.5.5 Run spsa on tuned and not-tuned learner

```
## SPSA-FSR begins:
```

```
## Wrapper = rf
```

```
## Measure = auc
```

```
## Number of selected features = 0
```

```
##
## iter  value   st.dev  num.ft  best.value
```

```
## 1     0.64527 0.08256 15      0.64527 *
## 2     0.63736 0.09589 15      0.64527
## 3     0.63    0.07449 15      0.64527
## 4     0.64972 0.05346 13      0.64972 *
## 5     0.63543 0.09206 14      0.64972
## 6     0.65182 0.0954  18      0.65182 *
## 7     0.65542 0.08729 18      0.65542 *
## 8     0.66478 0.08388 18      0.66478 *
## 9     0.64104 0.08914 17      0.66478
## 10    0.62139 0.07835 15      0.66478
## 11    0.65105 0.07116 19      0.66478
## 12    0.64953 0.07725 19      0.66478
## 13    0.64979 0.07474 15      0.66478
## 14    0.6903  0.07046 15      0.6903 *
## 15    0.66876 0.06026 15      0.6903
## 16    0.64628 0.05725 14      0.6903
## 17    0.64109 0.09358 13      0.6903
## 18    0.6492  0.08691 13      0.6903
## 19    0.66394 0.09972 12      0.6903
## 20    0.64763 0.05817 13      0.6903
## 21    0.6818  0.07834 16      0.6903
## 22    0.68017 0.10172 15      0.6903
## 23    0.6868  0.07259 17      0.6903
## 24    0.68493 0.07434 16      0.6903
## 25    0.68191 0.08741 18      0.6903
## 26    0.67666 0.06748 18      0.6903
## 27    0.70749 0.07646 17      0.70749 *
## 28    0.68335 0.07488 17      0.70749
## 29    0.68383 0.07831 17      0.70749
## 30    0.70585 0.04369 17      0.70749
## 31    0.68071 0.07106 16      0.70749
## 32    0.699   0.08312 17      0.70749
## 33    0.71162 0.07812 16      0.71162 *
## 34    0.68929 0.06466 16      0.71162
## 35    0.68584 0.05823 15      0.71162
## 36    0.70445 0.07646 15      0.71162
## 37    0.6999  0.04269 15      0.71162
## 38    0.68615 0.06074 15      0.71162
## 39    0.6781  0.0502  16      0.71162
## 40    0.71194 0.05457 16      0.71194 *
## 41    0.68673 0.0801  14      0.71194
## 42    0.67679 0.05745 15      0.71194
## 43    0.67355 0.0987  15      0.71194
## 44    0.67839 0.06948 16      0.71194
## 45    0.70746 0.09866 15      0.71194
## 46    0.67896 0.08791 15      0.71194
## 47    0.69796 0.07236 15      0.71194
## 48    0.70098 0.07518 15      0.71194
## 49    0.70375 0.07421 12      0.71194
## 50    0.69185 0.06111 12      0.71194
## 51    0.70308 0.08851 13      0.71194
## 52    0.67296 0.08072 14      0.71194
## 53    0.69155 0.08828 14      0.71194
## 54    0.69753 0.05135 14      0.71194
```

```
## 55    0.69872 0.06875 13      0.71194
## 56    0.70262 0.0942  15      0.71194
## 57    0.66506 0.06711 13      0.71194
## 58    0.66741 0.08574 13      0.71194
## 59    0.66759 0.07055 14      0.71194
## 60    0.64173 0.08891 9       0.71194
##
## Best iteration = 40
## Number of selected features = 16
## Best measure value = 0.71194
## Std. dev. of best measure = 0.05457
## Run time = 7.58 minutes.
## SPSA-FSR begins:
## Wrapper = rf
## Measure = auc
## Number of selected features = 0
##
## iter  value   st.dev  num.ft  best.value
## 1     0.64136 0.05998 13      0.64136 *
## 2     0.65978 0.09156 14      0.65978 *
## 3     0.63474 0.06441 13      0.65978
## 4     0.61493 0.05219 12      0.65978
## 5     0.65597 0.09038 13      0.65978
## 6     0.64112 0.07636 13      0.65978
## 7     0.65312 0.0895  14      0.65978
## 8     0.65561 0.07775 16      0.65978
## 9     0.6586  0.08977 11      0.65978
## 10    0.6407  0.10869 13      0.65978
## 11    0.69109 0.05848 13      0.69109 *
## 12    0.6823  0.10371 14      0.69109
## 13    0.65051 0.07801 14      0.69109
## 14    0.66418 0.06428 14      0.69109
## 15    0.68531 0.07193 13      0.69109
## 16    0.67668 0.09573 12      0.69109
## 17    0.64311 0.06947 13      0.69109
## 18    0.67443 0.07573 12      0.69109
## 19    0.6726  0.07219 12      0.69109
## 20    0.65868 0.12066 13      0.69109
## 21    0.6703  0.05989 13      0.69109
## 22    0.67901 0.07662 13      0.69109
## 23    0.67069 0.06384 13      0.69109
## 24    0.67469 0.06642 13      0.69109
## 25    0.69074 0.06232 13      0.69109
## 26    0.67373 0.05437 13      0.69109
## 27    0.67484 0.07171 13      0.69109
## 28    0.65416 0.07359 13      0.69109
## 29    0.6699  0.08349 13      0.69109
## 30    0.67646 0.08579 13      0.69109
## 31    0.68671 0.07337 13      0.69109
```

```
##
## Best iteration = 11
## Number of selected features = 13
## Best measure value = 0.69109
## Std. dev. of best measure = 0.05848
## Run time = 3.84 minutes.
```

3.5.6 Extract the spsa task (with the set of reduced features)

3.5.7 Plot the measure values across iterations

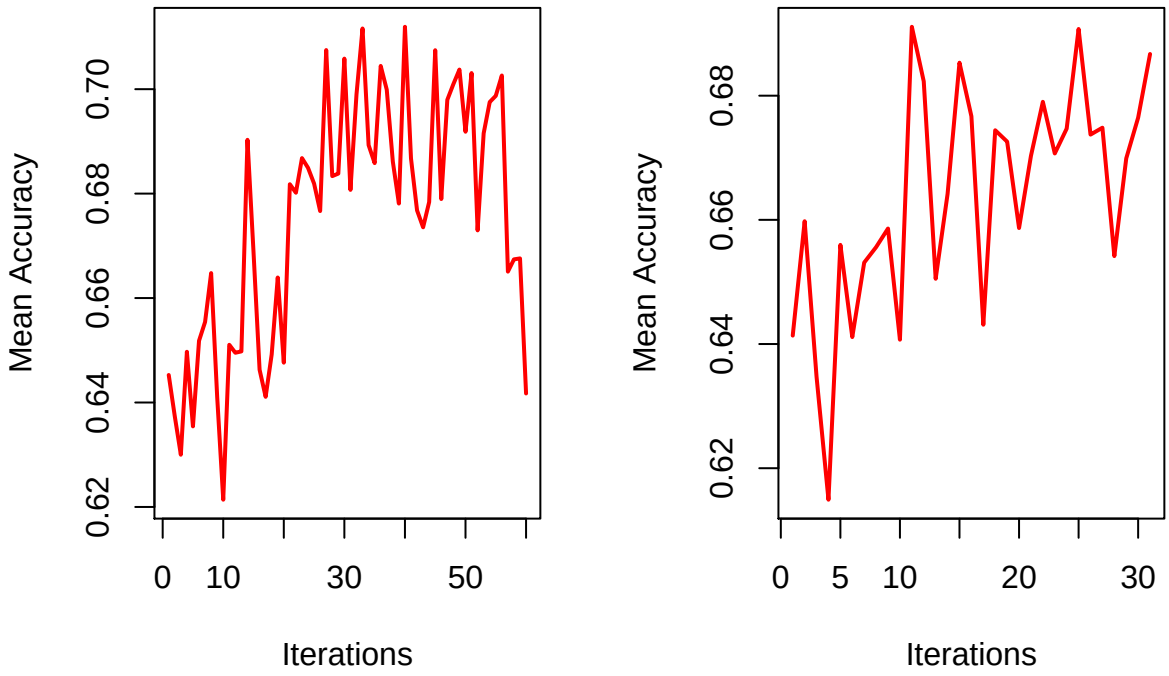


Figure 39. Scatter plots of mean accuracy rate for each iteration for the random forest classifier (left panel) with and (right panel) without hyperparameter tuning.

3.5.8 Get feature importances for random forest

3.5.9 Get feature importances for random forest with tuned parameters

The top 10 features selected using the spFRS algorithm differed significantly for the random forest classifier with and without hyperparameter optimization and features selection (Table 7).

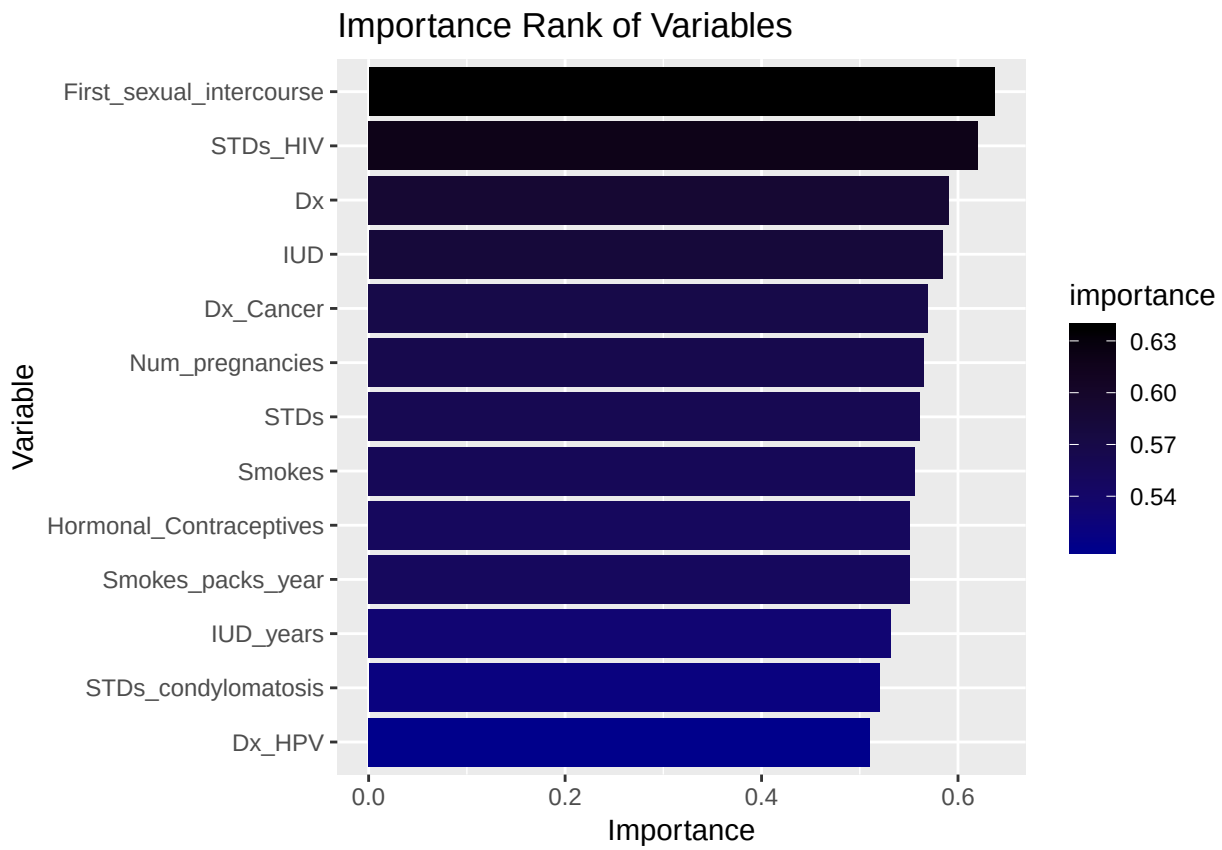
Table 8: Table 7. Selected Features for Random Forest calssifier with and without spFSR feature selection

RF + spFSR		RF + spFSR + Tuned	
Features	Importance	Features	Importance
First_sexual_intercourse	0.63703	Dx_HPV	0.60410
STDs_HIV	0.62002	Hormonal_Contraceptives	0.59345
Dx	0.59105	IUD	0.59153
IUD	0.58433	Smokes_years	0.58493
Dx_Cancer	0.56929	Num_pregnancies	0.58170
Num_pregnancies	0.56564	Num_sexual_partners	0.57620
STDs	0.56093	STDs_vulvo_perineal_condylomatosis	0.56873
Smokes	0.55601	Dx	0.56136
Hormonal_Contraceptives	0.55128	STDs_Num_diagnosis	0.56116
Smokes_packs_year	0.55090	First_sexual_intercourse	0.55971

3.5.10 Importance Plotting

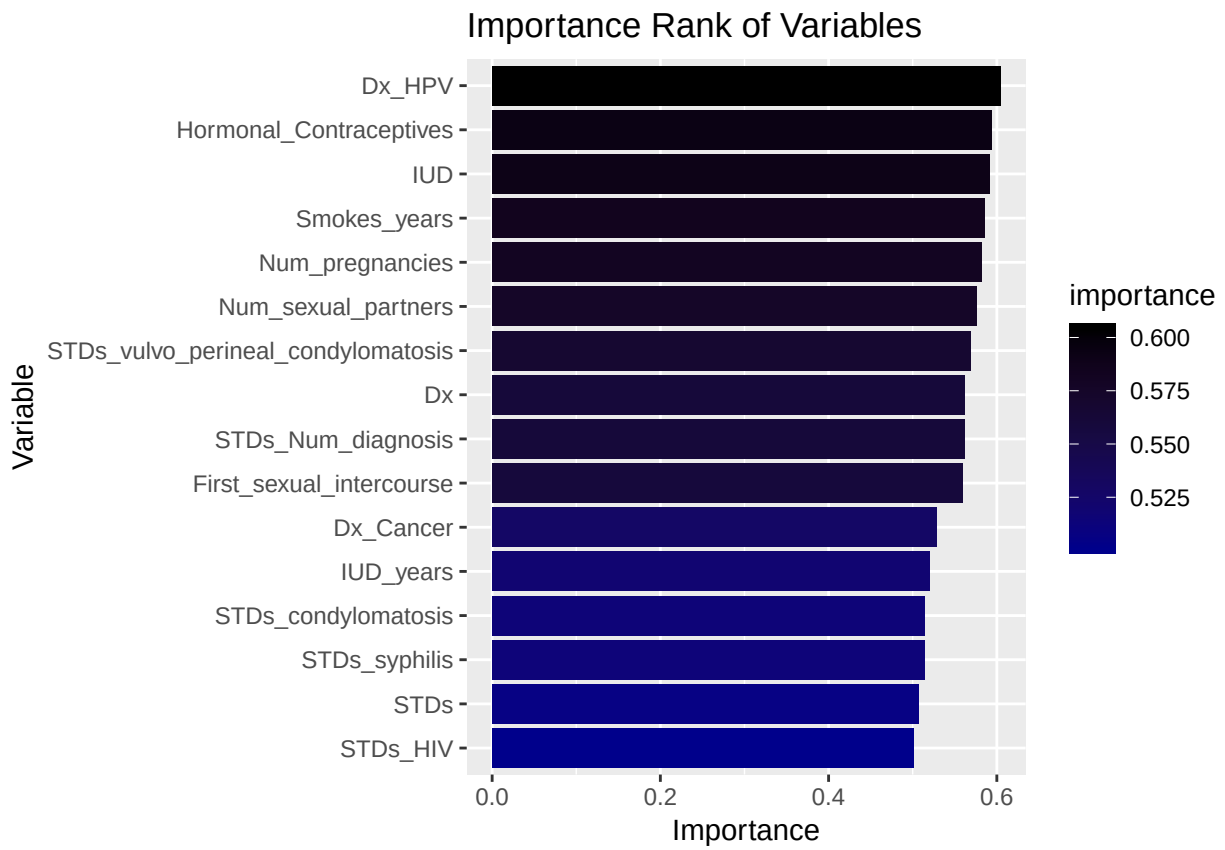
```
##                              features importance
## 1                              Dx_HPV    0.60410
## 2             Hormonal_Contraceptives    0.59345
## 3                                 IUD    0.59153
## 4                        Smokes_years    0.58493
## 5                     Num_pregnancies    0.58170
## 6                  Num_sexual_partners    0.57620
## 7  STDs_vulvo_perineal_condylomatosis    0.56873
## 8                                  Dx    0.56136
## 9                  STDs_Num_diagnosis    0.56116
## 10           First_sexual_intercourse    0.55971
## 11                          Dx_Cancer    0.52804
## 12                          IUD_years    0.51980
## 13                STDs_condylomatosis    0.51459
## 14                      STDs_syphilis    0.51397
## 15                               STDs    0.50654
## 16                           STDs_HIV    0.50134
```

```
##                    features importance
## 1  First_sexual_intercourse    0.63703
## 2                   STDs_HIV    0.62002
## 3                         Dx    0.59105
## 4                        IUD    0.58433
## 5                  Dx_Cancer    0.56929
## 6           Num_pregnancies    0.56564
## 7                      STDs    0.56093
## 8                    Smokes    0.55601
## 9   Hormonal_Contraceptives    0.55128
## 10        Smokes_packs_year    0.55090
## 11                IUD_years    0.53187
## 12      STDs_condylomatosis    0.52019
## 13                   Dx_HPV    0.51022
```



Figures 40. Features selected by the spFRS algorithm fused with the random forest classifier and ranked according to importance.

```
spFSR::plotImportance(spsaMod_rf_tuned)
```



Figures 41. Features selected by the spFRS algorithm fused with the random forest classifier with optimized hyperparameters and ranked according to importance.

3.6 Train models

3.6.1 Define the test data

3.7 Prediction

3.8 Evaluation

3.8.1 AUC plots for random forest

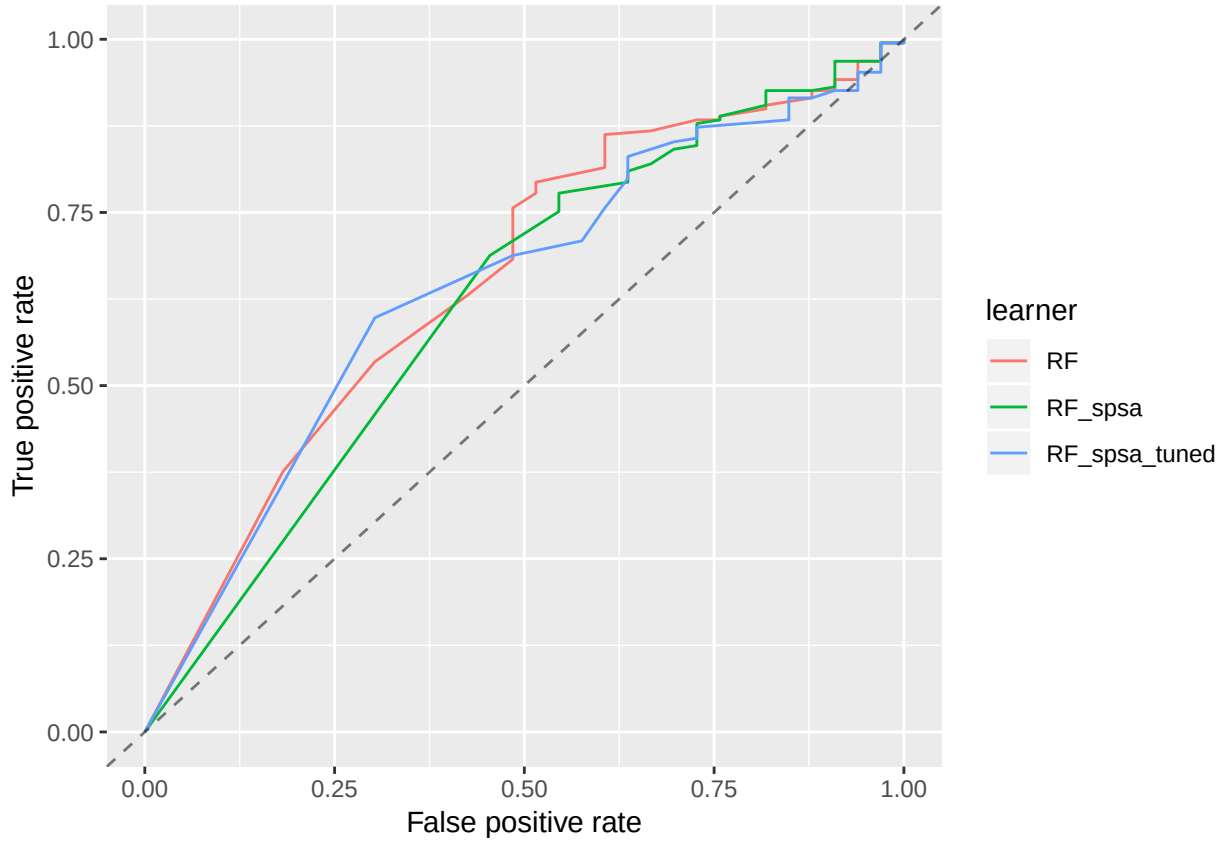


Figure 42. AUC plot for the random forest classifer. RF - Random forest classifier trained on all 23 features; RF_spsa - Random forest classifier trained on features selected using the spFRS algorithm; RF_spsa_tuned - Random forest classifier with tuned hyperparameters and trained on features selected using the spFRS algorithm.

3.8.2 Performance for random forest classifier with and without spSFRS feature selection

```
RF <- performance(pred_on_test_rf, measures = list(mmce, auc))
RF_spsa <- performance(pred_on_test_spsa_rf, measures = list(mmce, auc))
RF_spsa_tuned <- performance(pred_on_test_spsa_rf_tuned, measures = list(mmce, auc))

data_RF <- data.frame(RF, RF_spsa, RF_spsa_tuned)

kable(data_RF, caption = "Table 8. Performance for random forest classifier with and without tuned hype
```

Table 9: Table 8. Performance for random forest classifier with and without tuned hyperparameters and spFRS feature selection

	RF	RF_spsa	RF_spsa_tuned
mmce	0.1621622	0.1621622	0.1621622
auc	0.6537598	0.6418951	0.6361231

The random forest classifier with spFRS feature selection (but nmo hyperparameter optimization) performed the best with a mean misclassification error of 0.153 and an AUC value of 0.601 (Table 8).

3.8.3 Confusion Matrix, Precision and Recall for random forest classifier

```
##     predicted
## true 0         1
##    0 185       4        tpr: 0.98 fnr: 0.02
##    1 32        1        fpr: 0.97 tnr: 0.03
##      ppv: 0.85 for: 0.8 lrp: 1.01 acc: 0.84
##      fdr: 0.15 npv: 0.2 lrm: 0.7  dor: 1.45
##
##
## Abbreviations:
## tpr - True positive rate (Sensitivity, Recall)
## fpr - False positive rate (Fall-out)
## fnr - False negative rate (Miss rate)
## tnr - True negative rate (Specificity)
## ppv - Positive predictive value (Precision)
## for - False omission rate
## lrp - Positive likelihood ratio (LR+)
## fdr - False discovery rate
## npv - Negative predictive value
## acc - Accuracy
## lrm - Negative likelihood ratio (LR-)
## dor - Diagnostic odds ratio
```

3.8.4 Confusion Matrix, Precision and Recall for random forest classifier with features selected using the spFRS algorithm.

```
##     predicted
## true 0         1
##    0 185       4        tpr: 0.98 fnr: 0.02
##    1 32        1        fpr: 0.97 tnr: 0.03
##      ppv: 0.85 for: 0.8 lrp: 1.01 acc: 0.84
##      fdr: 0.15 npv: 0.2 lrm: 0.7  dor: 1.45
##
##
## Abbreviations:
## tpr - True positive rate (Sensitivity, Recall)
## fpr - False positive rate (Fall-out)
## fnr - False negative rate (Miss rate)
## tnr - True negative rate (Specificity)
## ppv - Positive predictive value (Precision)
```

```
## for - False omission rate
## lrp - Positive likelihood ratio (LR+)
## fdr - False discovery rate
## npv - Negative predictive value
## acc - Accuracy
## lrm - Negative likelihood ratio (LR-)
## dor - Diagnostic odds ratio
```

3.8.5 Confusion Matrix, Precision and Recall for random forest classifier with optimized hyperparameters and features selected using the spFRS algorithm.

```
##     predicted
## true 0         1
##    0 185       4        tpr: 0.98 fnr: 0.02
##    1 32        1        fpr: 0.97 tnr: 0.03
##      ppv: 0.85 for: 0.8 lrp: 1.01 acc: 0.84
##      fdr: 0.15 npv: 0.2 lrm: 0.7  dor: 1.45
##
##
## Abbreviations:
## tpr - True positive rate (Sensitivity, Recall)
## fpr - False positive rate (Fall-out)
## fnr - False negative rate (Miss rate)
## tnr - True negative rate (Specificity)
## ppv - Positive predictive value (Precision)
## for - False omission rate
## lrp - Positive likelihood ratio (LR+)
## fdr - False discovery rate
## npv - Negative predictive value
## acc - Accuracy
## lrm - Negative likelihood ratio (LR-)
## dor - Diagnostic odds ratio
```

4 Compare spFSR to other Feature Selection methods with 10 features

4.1 Random forest filter method for feature selection

Filter methods assign an importance to each feature. The feature is ranked according to importance value resulting in a feature subset. Create an object named mfv by calling generateFilterValuesData from mlr on classif.task and using the filter method randomForest.importance.

```
## Supervised task: dataCervicalCancer
## Type: regr
## Target: Cancer
## Observations: 737
## Features:
##    numerics     factors     ordered functionals
##          10          12           0           0
## Missings: FALSE
## Has weights: FALSE
```

```
## Has blocking: FALSE
## Has coordinates: FALSE
```

4.1.1 Plot filtered features obtained random forest method

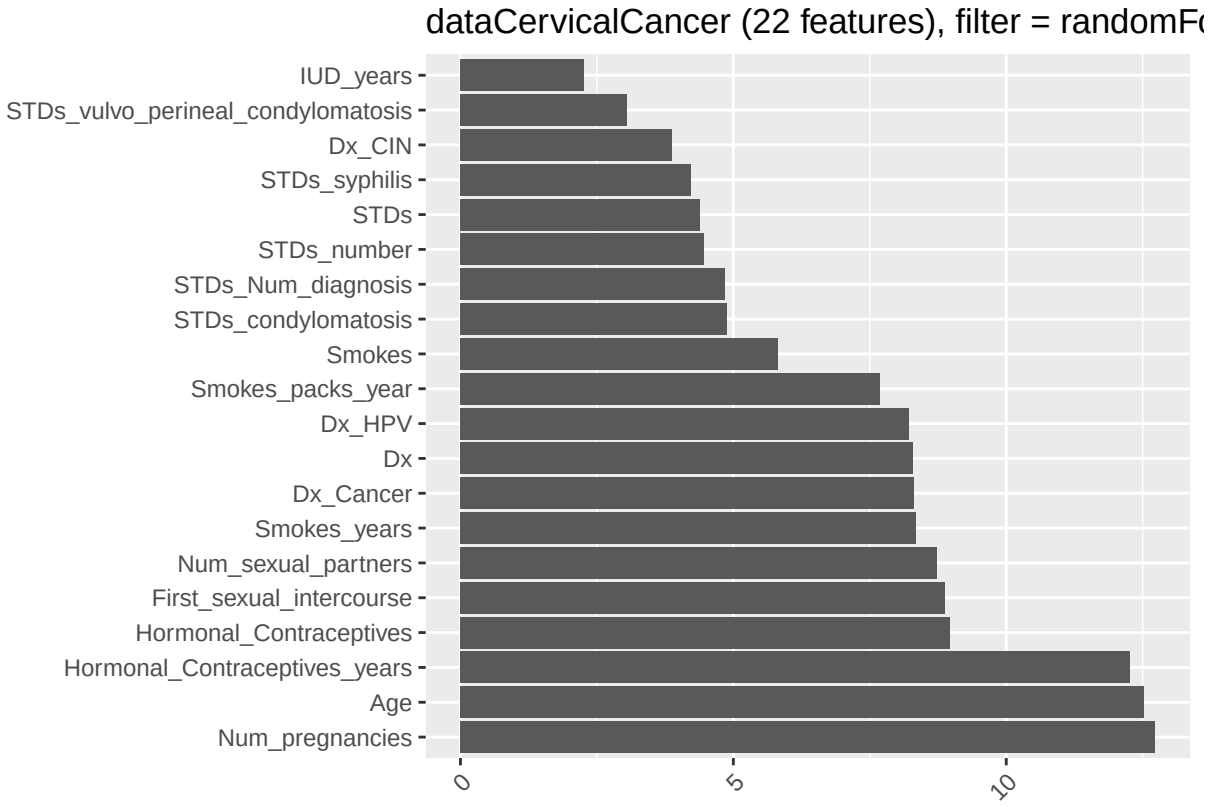


Figure 43. Features selectered sing a filter selection algorithm in mlr (https://mlr-org.github.io/mlr/articles/tutorial/devel/feature_selection.html).(**???**)

We can also compare this with other filter methods to identify consistencies in the selected features (i.e. information.gain and chi.squared) using the interactive mode plotFilterValuesGGVIS(FV) (Fig. 43).

4.1.2 Plot filtered features obtained information gain and chi squared methods

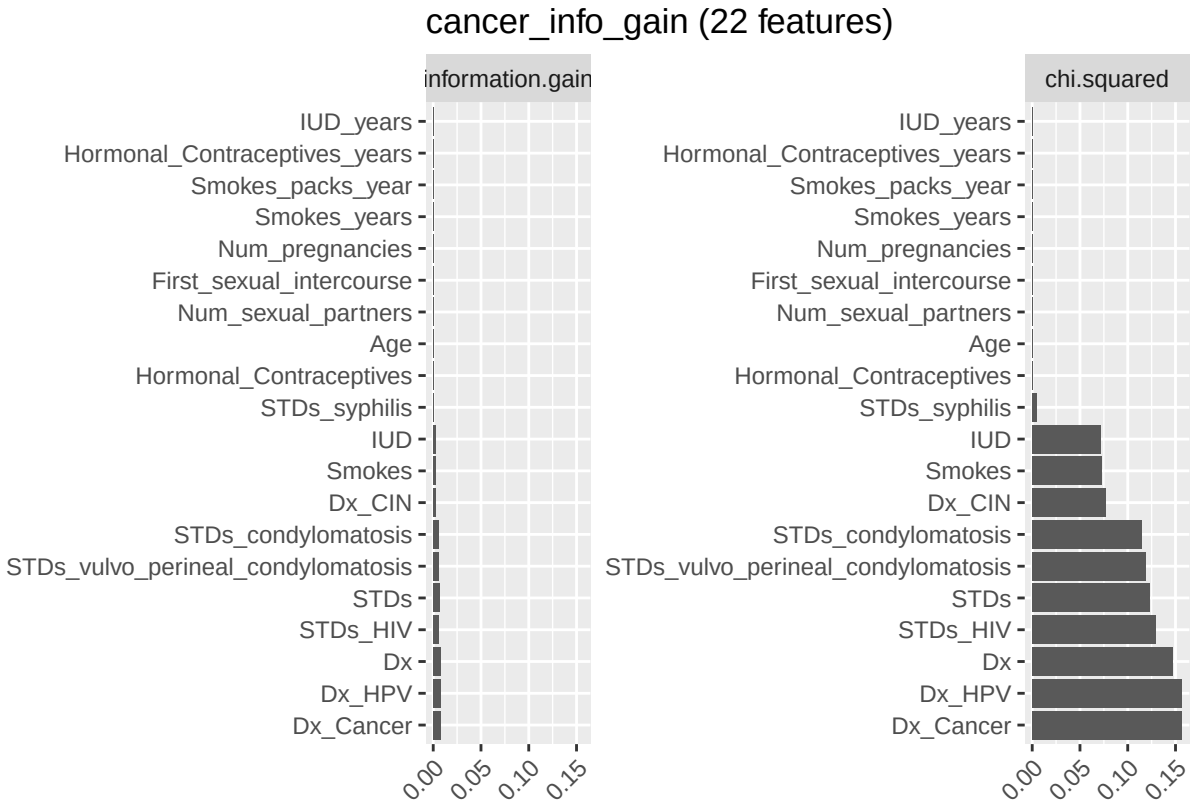


Figure 44. Features selectered using a filter selection algorithm in mlr based on (Left panel) information gain and (right panel) chi squared (https://mlr-org.github.io/mlr/articles/tutorial/devel/feature_selection.html).(**???**)

Using these 2 filters DX_Cancer was consistently the most important feature while IUD_years was the least important.

4.1.3 Fuse the random forest learner with the information gain filter

We now 'fused' the random forest classification learner with the information.gain filter to train the model.

4.1.4 Determine optimal number of features to keep

The optimal percentage of features to keep was determined by 5-fold cross-validation. We use 'information gain' as an importance measure and select the 10 features with highest importance. In each resampling iteration feature selection is carried out on the corresponding training data set before fitting the learner.

```
## [Tune] Started tuning learner classif.randomForest.filtered for parameter set:

##            Type len Def Constr Req Tunable Trafo
## fw.abs discrete   -   -     10   -    TRUE     -

## With control class: TuneControlGrid

## Imputation value: 1
```

```
## [Tune-x] 1: fw.abs=10
## [Tune-y] 1: mmce.test.mean=0.1315775; time: 0.0 min
## [Tune] Result: fw.abs=10 : mmce.test.mean=0.1315775
```

4.1.5 Performance (misclassification error)

The optimal percentage and corresponding misclassification error are:

```
## $fw.abs
## [1] 10
## mmce.test.mean
##      0.1315775
```

4.1.6 Fuse learner with feature selection

We can now fuse it with fw percentage by "wrapper" the random forest learner with the chi-squared method before training the model:

4.1.7 View selected features

Now applied getFilteredFeatures on the trained model to view the selected features.

4.2 Wrapper Methods

4.2.1 Select optimal features to use

Used a random search with ten iterations on the random forest classifier and classif.task.

4.2.2 Performance (misclassification error)

```
## mmce.test.mean
##      0.1184437
```

4.2.3 View the important features

The wrapper method selected fewer features and 2 of these were also selected by the filter method (Table 9).

4.2.4 Wrap feature selection method with learner

By comparing the misclassification error rates, a random search wrapper method out performed the chi squared (filtered) method. We then fused the wrapper method in a learnerusing makeFeatSelWrapper together with makeFeatSelControlRandom and makeResampleDesc objects.

Table 10: Table 9. Selected Features for Random Forest classifier with mlr Feature Selection (selectFeatures)

Filter	Wrapper
IUD	Age
STDs	First_sexual_intercourse
STDs_condylomatosis	STDs_number
STDs_vulvo_perineal_condylomatosis	IUD
STDs_syphilis	STDs
STDs_HIV	STDs_HIV
Dx_Cancer	Dx_CIN
Dx_CIN	Dx_HPV
Dx_HPV	NA
Dx	NA

4.3 Prediction

4.4 Evaluation

Obtain the confusion matrix by running calculateConfusionMatrix(pred_on_test) and get the ROC.

4.4.1 Confusion Matrix

```
##         predicted
## true       0 1 -err.-
##   0      186 3      3
##   1       32 1     32
##   -err.-  32 3     35
```

Table 11: Table 10. Classifier Performance

	x
mmce	0.1576577
auc	0.6289081

4.4.2 AUC plot

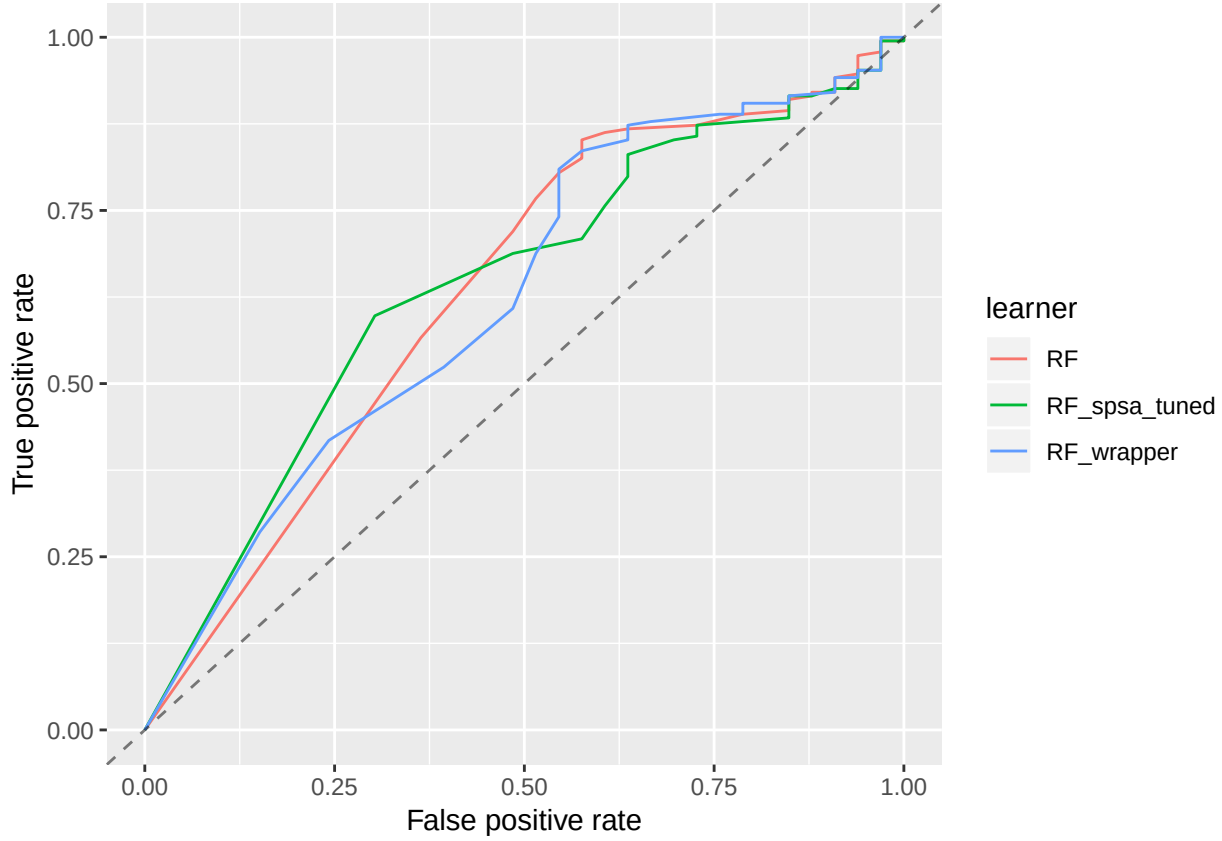


Figure 22. AUC plot for the random forest classifer. RF - Random forest classifier trained on all 23 features; RF_spsa_tuned - Random forest classifier with tuned hyperparameters and trained on features selected using the spFRS algorithm; RF_wrapper - Random forest classifier with tned parameters and trained on features selected using the selectFeatures mlr algorithm.

The AUC plot show that the random forest classifier with spFRS feature selection performed the best.

4.4.3 Performance for random forest with mlr feature selection

4.4.3.1 Misclassification Error and AUC value

The random forest classifier using the wrapper method for feature selection had a mean misclassification error of 0.158 and AUC value of 0.644 (Table 10).

4.4.4 Performance for random forest with mlr feature selection

```
##     predicted
```

```
## true 0         1
##    0 186       3         tpr: 0.98 fnr: 0.02
##    1 32        1         fpr: 0.97 tnr: 0.03
##      ppv: 0.85 for: 0.75 lrp: 1.01 acc: 0.84
##      fdr: 0.15 npv: 0.25 lrm: 0.52 dor: 1.94
##
##
## Abbreviations:
## tpr - True positive rate (Sensitivity, Recall)
## fpr - False positive rate (Fall-out)
## fnr - False negative rate (Miss rate)
## tnr - True negative rate (Specificity)
## ppv - Positive predictive value (Precision)
## for - False omission rate
## lrp - Positive likelihood ratio (LR+)
## fdr - False discovery rate
## npv - Negative predictive value
## acc - Accuracy
## lrm - Negative likelihood ratio (LR-)
## dor - Diagnostic odds ratio
```

5 Discussion

The best results were obtained for the Random Forest classifier with spFRS feature selection. This classifier had a precision of 0.85 and recall of 0.99 with a mmce of 0.153 and AUC value of 0.601. The next best performer was also Random Forest classifier with the feature selection algorithm provided by the CARAT R package. Features assocated with IUDs and STDs were consistently choses as beingof importance by the various feature selector algorithms. Overall the Random Forest classifier has performed reasonably well for the identifying women exhibiting clinical sign of cervical cancer. Although the AUC values suggest that the model can be improved.

In future studies other supervised machine learning algorithms such as Support Vector Machines (SVM) or ensemble methods could be tested to determine whether they may perform better than those trialed in these studies.

6 Conclusion

The random forest classifiers performed reasonably well at identifying cervical cancer patients. However, some improvement is required before these predictors can be used in a clinical setting, as several diseased patients were not identified in these studies.

7 References