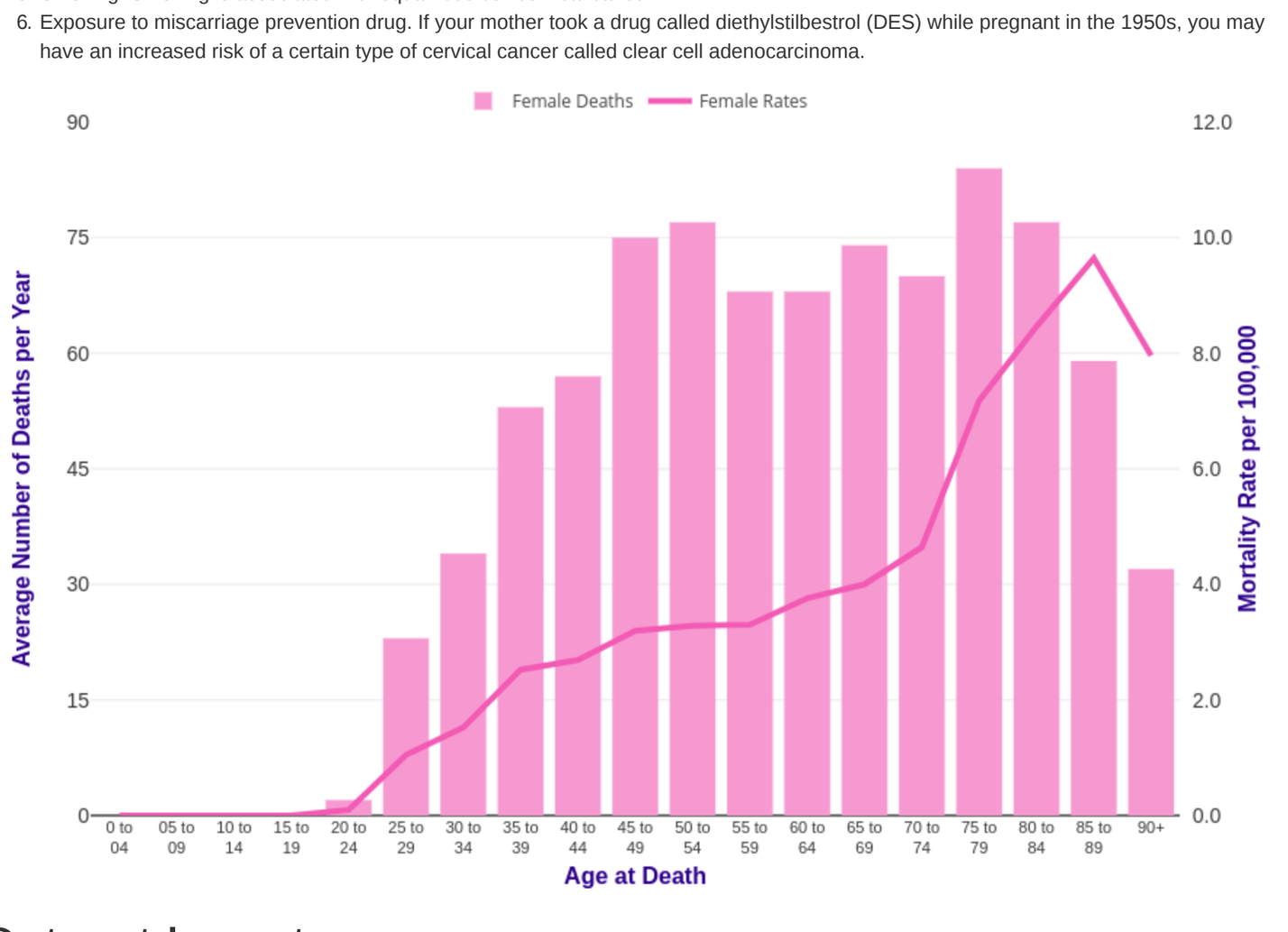
Cervical Cancer Prediction

Testa Luca

Overview

Cervical cancer is a type of cancer that occurs in the cells of the cervix — the lower part of the uterus that connects to the vagina. Various strains of the human papillomavirus (HPV), a sexually transmitted infection, play a role in causing most cervical cancer. Risk factors for cervical cancer include:

1. Many sexual partners. The greater your number of sexual partners — and the greater your partner's number of sexual partners — the greater your chance of acquiring HPV.
2. Early sexual activity. Having sex at an early age increases your risk of HPV.
3. Other sexually transmitted infections (STIs). Having other STIs — such as chlamydia, gonorrhea, syphilis and HIV/AIDS — increases your risk of HPV.
4. A weakened immune system. You may be more likely to develop cervical cancer if your immune system is weakened by another health condition and you have HPV.
5. Smoking. Smoking is associated with squamous cell cervical cancer.
6. Exposure to miscarriage prevention drug. If your mother took a drug called diethylstilbestrol (DES) while pregnant in the 1950s, you may have an increased risk of a certain type of cervical cancer called clear cell adenocarcinoma.

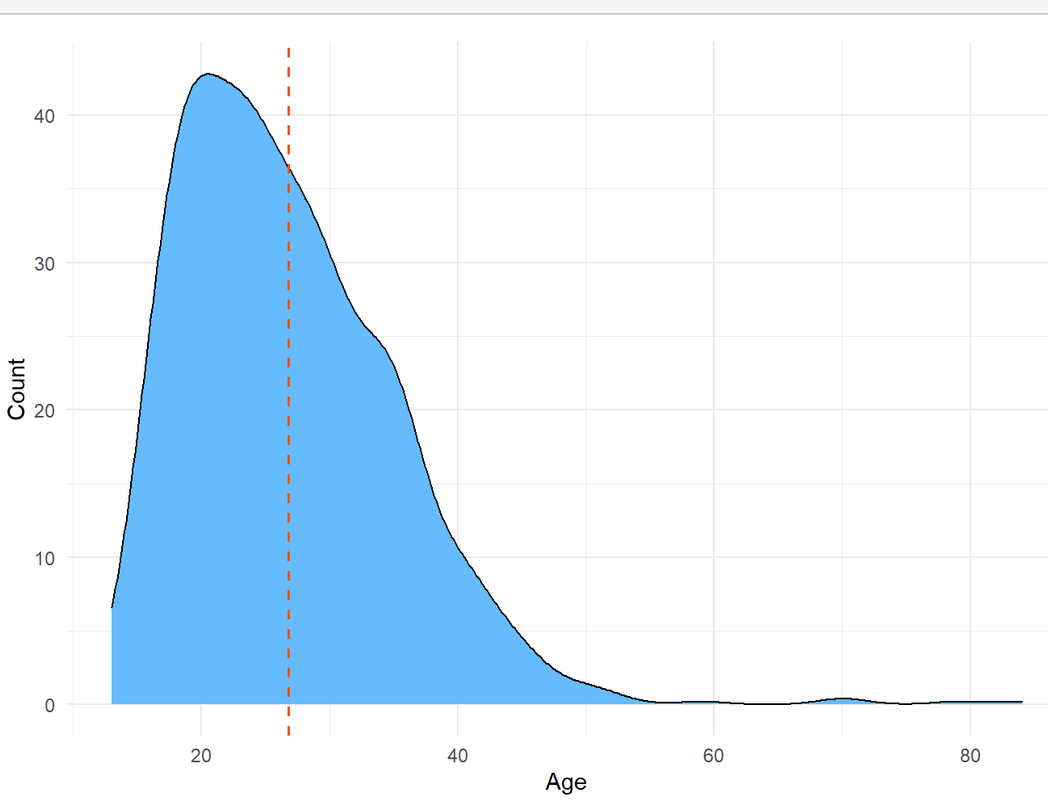


Dataset Import

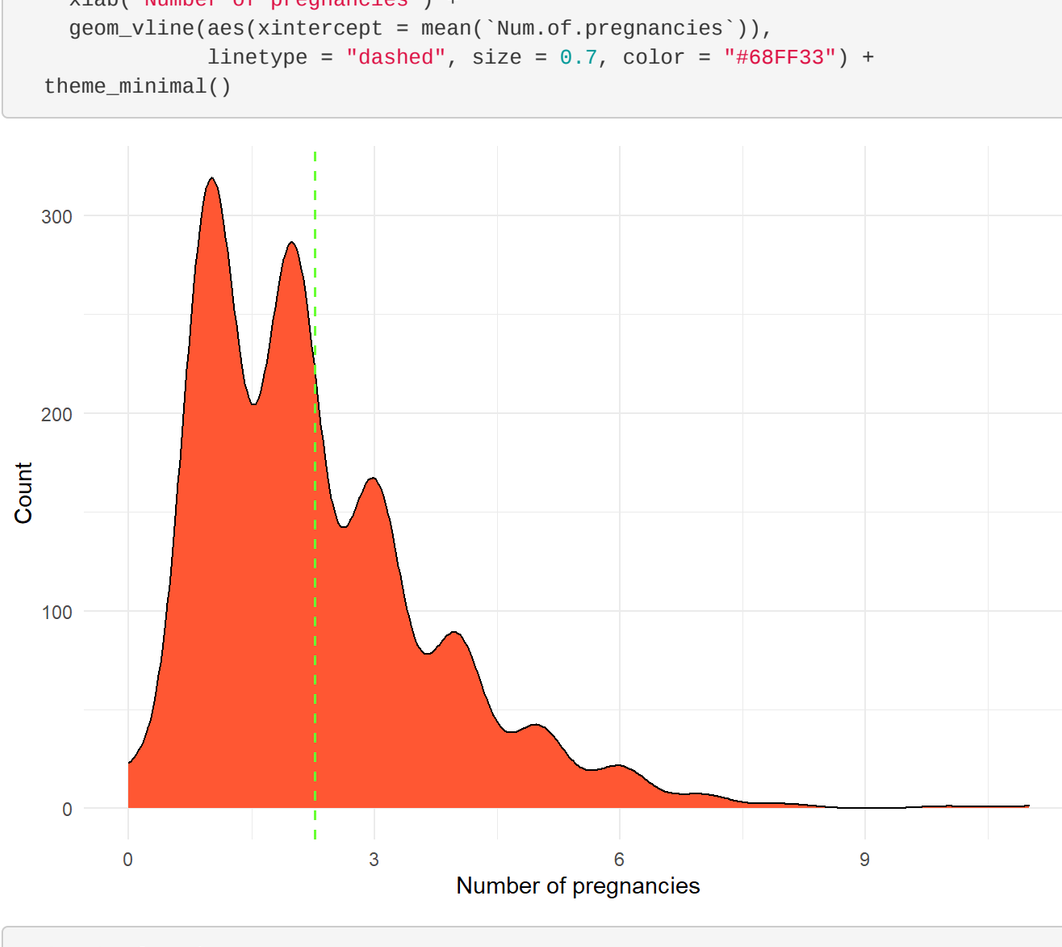
```
# - Age
# - Number of sexual partners
# - Num of pregnancies
# - Smokes
# - Smokes (packs/years)
# - Hormonal Contraceptives (bool)
# - IUD
# - All STDs (not time)
# - Dx:Cancer
# - Dx:HPV
# - Dx:CIN
# - Hinselmann
# - Schiller
# - Citology
# - Biopsy


library(plyr)
library(dplyr)
library(caret)
library(repr)
library(ROSE)
library(caTools)
require(rpart)
library(partykit)
library(corrplot)
library(xgboost)
library(Boruta)
library(ggplot2)
library(class)
library(corrgram)


library(readr)
setwd("E:/UNIVERSITA/MACHINE LEARNING/Cancer_Project")
df <- read_delim("kag_risk_factors_cervical_cancer2.csv",
         ";",
         col_types = cols(Age = col_integer(),
                      Biopsy = col_integer(),
                      Citology = col_integer(),
                      `Number.of.sexual.partners` = col_number(),
                      `First.sexual.intercourse`= col_number(),
                      `Num.of.pregnancies` = col_number(),
                      Smokes = col_factor(levels = c()),
                      `Smokes.(packs/year)` = col_number(),
                      `Hormonal.Contraceptives.years` = col_number(),
                      `IUD.years` = col_number(),
                      `STDs.number`  = col_number(),
                      Dx = col_factor(levels = c()),
                      `Dx.CIN` = col_factor(levels = c()),
                      `Dx.Cancer` = col_factor(levels = c()),
                      `Dx.HPV` = col_factor(levels = c()),
                      Hinselmann = col_integer(),
                      `Hormonal.Contraceptives` = col_factor(levels = c()),
                      IUD = col_factor(levels = c()),
                      STDs = col_factor(levels = c()),
                      `STDs..Number.of.diagnosis` = col_integer(),
                      `STDs.AIDS` = col_factor(levels = c()),
                      `STDs.HIV` = col_factor(levels = c()),
                      `STDs.HPV` = col_factor(levels = c()),
                      `STDs.Hepatitis.B` = col_factor(levels = c()),
                      `STDs.cervical.condylomatosis` = col_factor(levels = c()),
                      `STDs.condylomatosis` = col_factor(levels = c()),
                      `STDs.genital.herpes` = col_factor(levels = c()),
                      `STDs.molluscum.contagiosum` = col_factor(levels = c()),
                      `STDs.pelvic.inflammatory.disease` = col_factor(levels = c()),
                      `STDs.syphilis` = col_factor(levels = c()),
                      `STDs.vaginal.condylomatosis` = col_factor(levels = c()),
                      `STDs.vulvo.perineal.condylomatosis` = col_factor(levels = c()),
                      Schiller = col_integer(),
                      Smokes = col_factor(levels = c())))
```

Descriptive Analysis

```
# Age Distribution with mean of 27
ggplot(df, aes(x = Age), linetype = 0) +
  geom_density(aes(y = ..count..), fill = "#65BBFE") +
  ylab("Count") +
  geom_vline(aes(xintercept = mean(Age)), linetype = "dashed",
           size = 0.7, color = "#FC4E07") +
  theme_minimal()
```

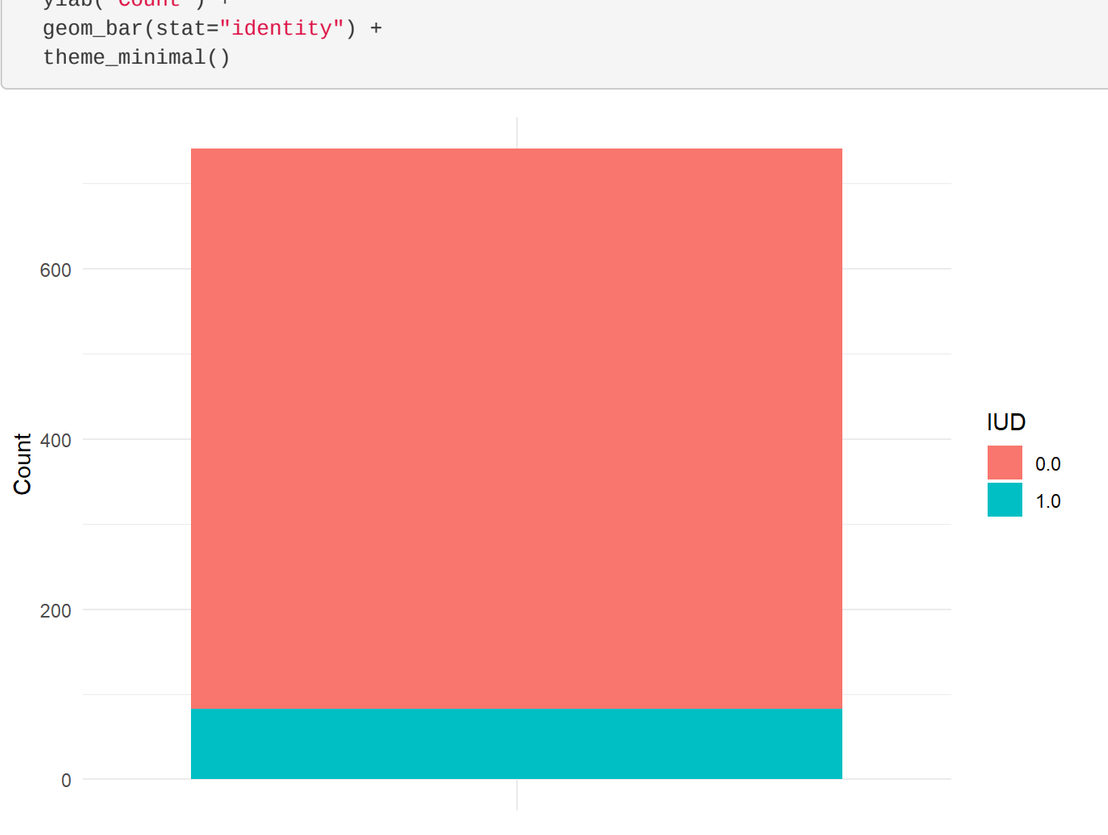



```
# Number of pregnancies Distribution with mean of 3
df %>%
  dplyr::filter(Num.of.pregnancies != -1) %>%
  ggplot(aes(x = Num.of.pregnancies), linetype = 0) +
    geom_density(aes(y = ..count..), fill = "#FF5733") +
    ylab("Count") +
    xlab("Number of pregnancies") +
    geom_vline(aes(xintercept = mean(`Num.of.pregnancies`)),
             linetype = "dashed", size = 0.7, color = "#68FF33") +
  theme_minimal()
```

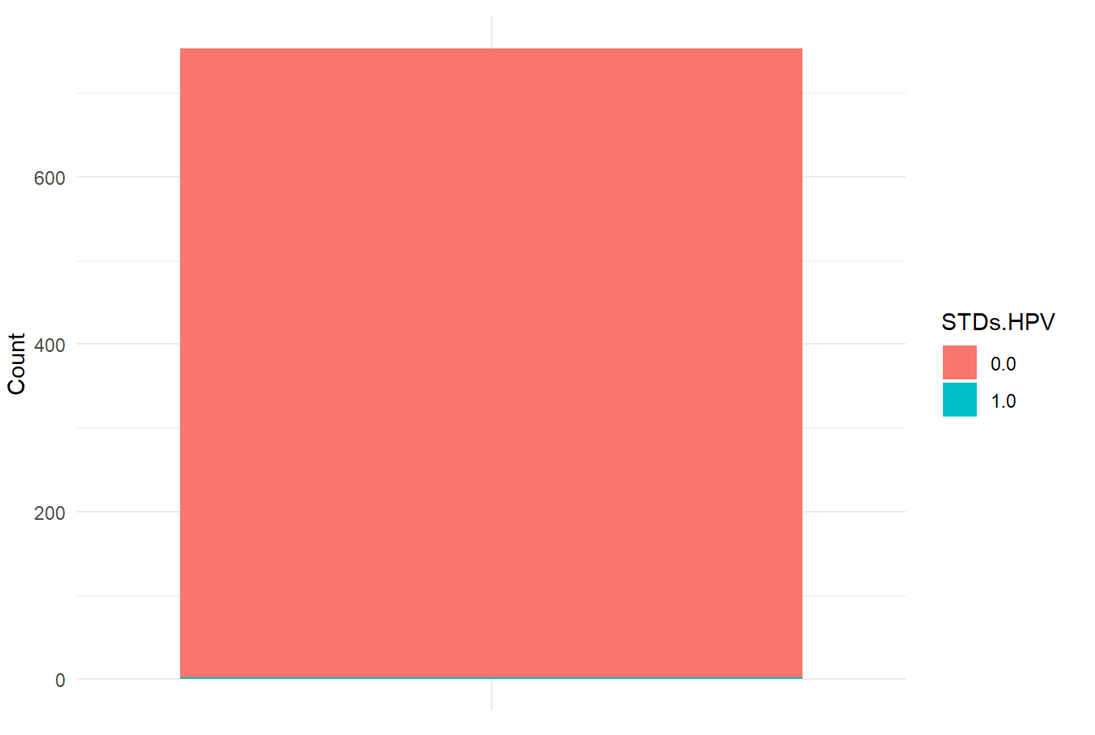



```
# Count of smoker
df %>%
  dplyr::filter(Smokes != -1) %>%
  dplyr::group_by(Smokes) %>%
  dplyr::summarize(count_smokes = n()) %>%
  as.data.frame() %>%
  ggplot(aes(x = "", y = count_smokes, fill = as.factor(Smokes))) +
  xlab("Smokes") +
  ylab("Count") +
  geom_bar(stat="identity") +
  scale_fill_discrete(name = "Smokes") +
  theme_minimal()
```

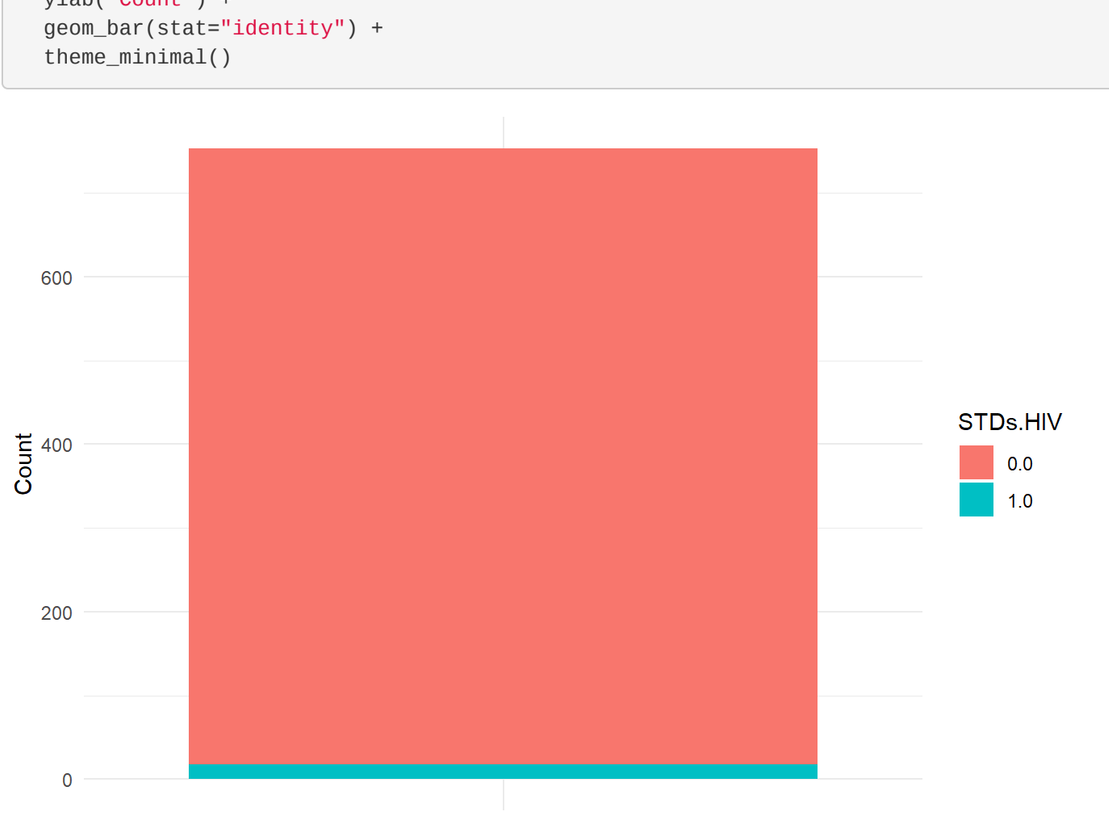

```
# Some research suggests that women who had ever used an intrauterine device (IUD)
# had a lower risk of cervical cancer. The effect on risk was seen even in women
# who had an IUD for less than a year, and the protective effect remained after the
# IUDs were removed
df %>%
  dplyr::filter(IUD != -1) %>%
  dplyr::group_by(IUD) %>%
  dplyr::summarize(count_iud = n()) %>%
  as.data.frame() %>%
  ggplot(aes(x = "", y = count_iud, fill = IUD)) +
  xlab("IUD") +
  ylab("Count") +
  geom_bar(stat="identity") +
  theme_minimal()
```




```
# Human papillomavirus (HPV) is the most common sexually transmitted infection
df %>%
  dplyr::filter(`STDs.HPV` != -1) %>%
  dplyr::group_by(`STDs.HPV`) %>%
  dplyr::summarize(count_hpv = n()) %>%
  as.data.frame() %>%
  ggplot(aes(x = "", y = count_hpv, fill = `STDs.HPV`)) +
  xlab("HPV") +
  ylab("Count") +
  geom_bar(stat="identity") +
  theme_minimal()
```

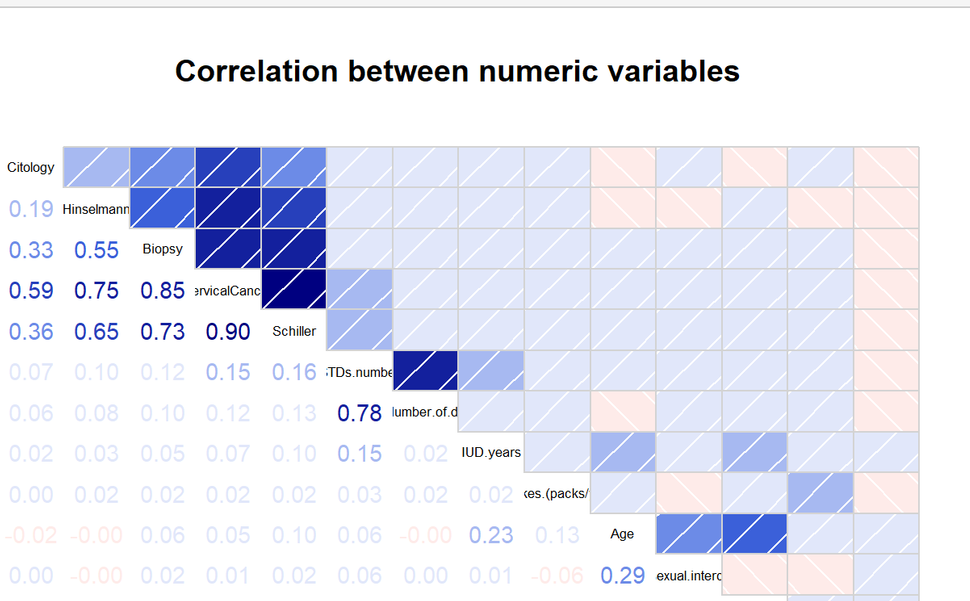



```
# The likelihood that a woman living with HIV will develop invasive cervical
# cancer is up to five times higher than for a woman who is not living with HIV.
df %>%
  dplyr::filter(`STDs.HIV` != -1) %>%
  dplyr::group_by(`STDs.HIV`) %>%
  dplyr::summarize(count_hiv = n()) %>%
  as.data.frame() %>%
  ggplot(aes(x = "", y = count_hiv, fill = `STDs.HIV`)) +
  xlab("HIV") +
  ylab("Count") +
  geom_bar(stat="identity") +
  theme_minimal()
```




```
# They're 4 tests for cancer diagnostics
df$CervicalCancer = df$Hinselmann + df$Schiller + df$Citology + df$Biopsy


df %>%
  dplyr::group_by(CervicalCancer) %>%
  dplyr::summarize(count = n()) %>%
  mutate(Patient_Percent = count/858) %>%
  arrange(desc(Patient_Percent)) %>%
  as.data.frame() %>%
  ggplot(aes(x = CervicalCancer, y = Patient_Percent, fill = as.factor(CervicalCancer))) +
  xlab("Cervical Cancer") +
  ylab("Percent of patient") +
  geom_bar(stat="identity", show.legend = FALSE) +
  theme_minimal()
```


```
numeric.var <- sapply(df, is.numeric)
corr.matrix <- cor(df[, numeric.var])
corrgram(corr.matrix, order = TRUE, main = "Correlation between numeric variables",
        upper.panel = panel.shade, lower.panel = panel.cor)
```



Studying the correlation between numerical variables. Since that "STDs.number" and "STDs..Number.of.diagnosis" are highly correlated, "STDs.number" has been removed

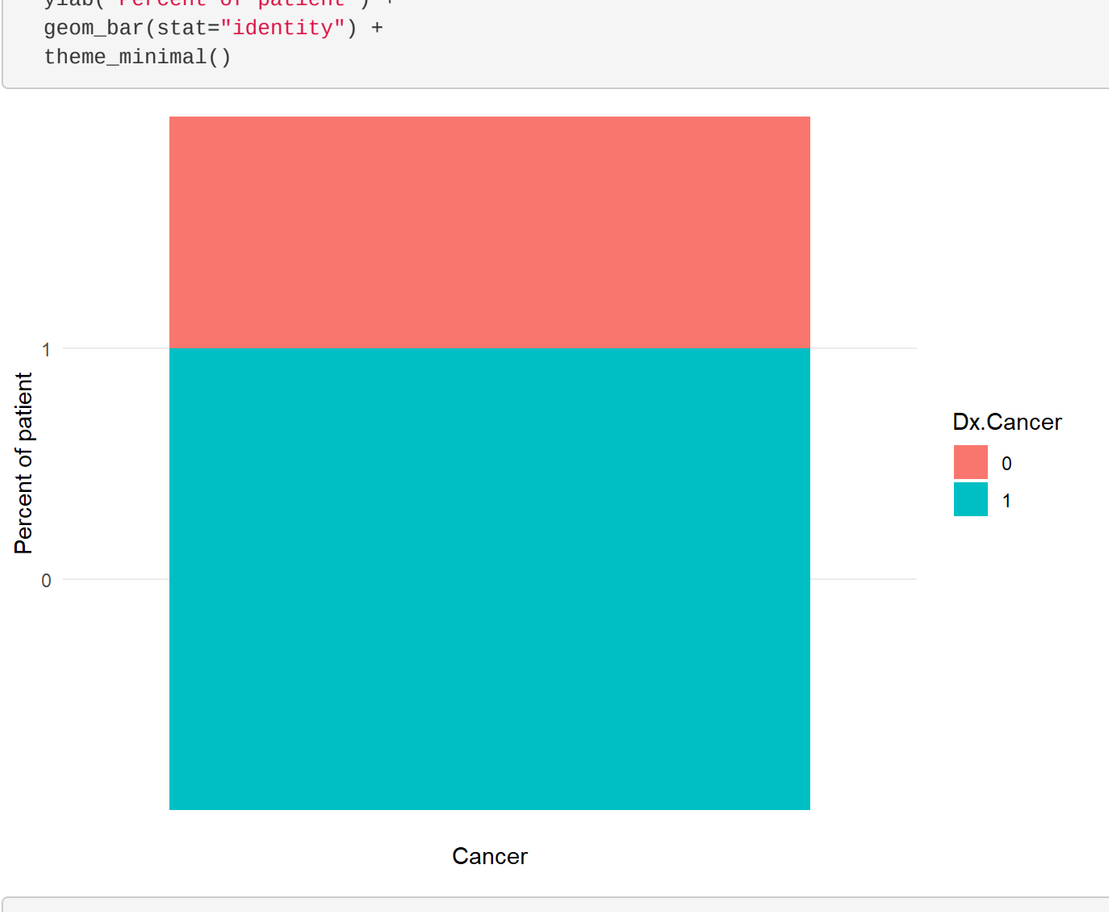
```
df$STDs.number <- NULL
df[c("Hinselmann", "Schiller", "Biopsy", "Citology")] <- NULL
```

Feature Selection with Boruta

```
# Unbalanced Dataset
table(df$Dx.Cancer)
```

```
##
##   0   1
## 840  18
```

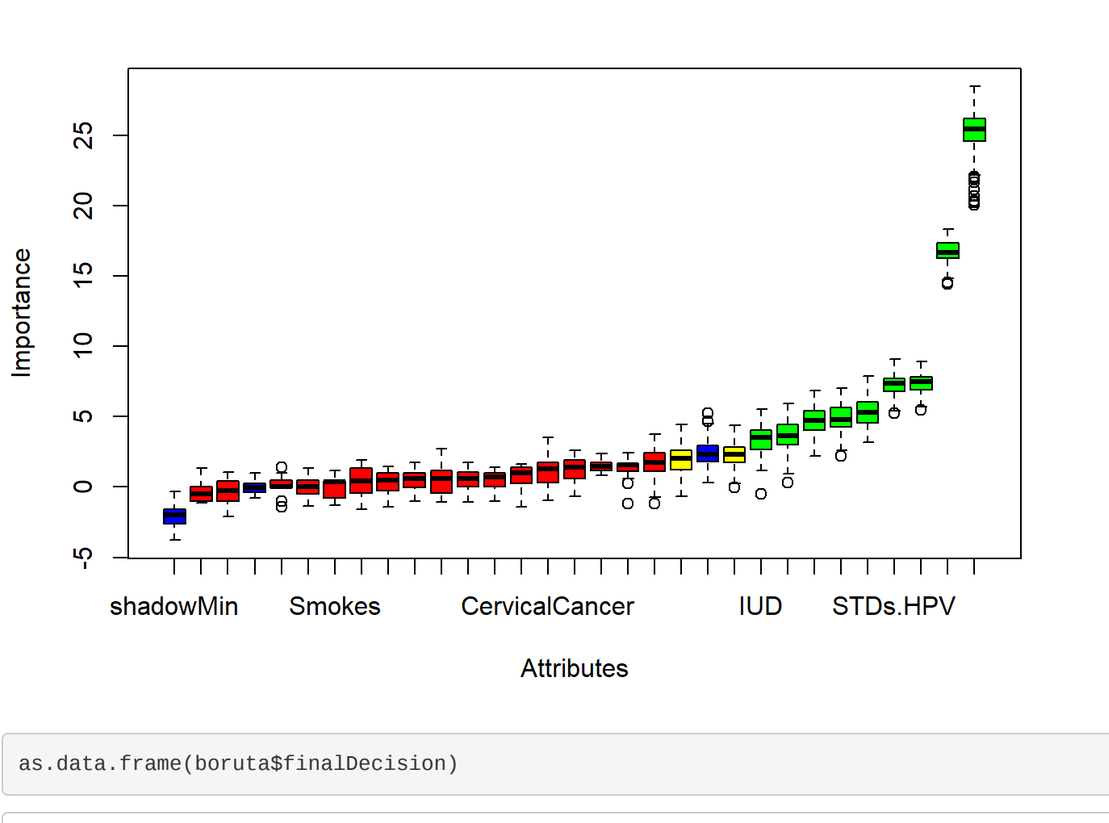
```
df %>%
  dplyr::group_by(Dx.Cancer) %>%
  dplyr::summarize(count = n()) %>%
  as.data.frame() %>%
  ggplot(aes(x = "", y = Dx.Cancer, fill = Dx.Cancer)) +
  xlab("Cancer") +
  ylab("Percent of patient") +
  geom_bar(stat="identity") +
  theme_minimal()
```




```
train = df

# Perform Boruta Analysis on the training set
set.seed(123)
boruta <- Boruta(Dx.Cancer ~., data = train)

# We can see which variables are important for next models
print(boruta)
```

```
## Boruta performed 99 iterations in 2.891159 secs.
##  9 attributes confirmed important: Age, Dx, Dx.CIN, Dx.HPV,
## First.sexual.intercourse and 4 more;
##  17 attributes confirmed unimportant: `Smokes.(packs/year)`,
## CervicalCancer, Hormonal.Contraceptives.years,
## Number.of.sexual.partners, Smokes and 12 more;
##  2 tentative attributes left: Hormonal.Contraceptives,
## STDs.condylomatosis;
```

```
plot(boruta)
```




```
as.data.frame(boruta$finalDecision)
```

	boruta$finalDecision <fctr>
Age	Confirmed
Number.of.sexual.partners	Rejected
First.sexual.intercourse	Confirmed
Num.of.pregnancies	Confirmed
Smokes	Rejected
`Smokes.(packs/year)`	Rejected
Hormonal.Contraceptives	Tentative
Hormonal.Contraceptives.years	Rejected
IUD	Confirmed
IUD.years	Confirmed

1-10 of 28 rows   Previous 1 2 3 Next

```
train[, c("Number.of.sexual.partners", "Smokes", "Smokes.(packs/year)",
         "STDs.cervical.condylomatosis", "STDs.genital.herpes",
         "STDs.vaginal.condylomatosis", "STDs..Number.of.diagnosis",
         "STDs.syphilis", "STDs.pelvic.inflammatory.disease",
         "STDs.genital.herpes", "STDs.molluscum.contagiosum",
         "Hormonal.Contraceptives.years", "STDs", "CervicalCancer",
         "Dx.CIN")] = NULL

# Split the dataset for predictor and balance it
intrain <- ovun.sample(Dx.Cancer ~., train, seed = 123, method = "under")$data
table(intrain$Dx.Cancer)
```

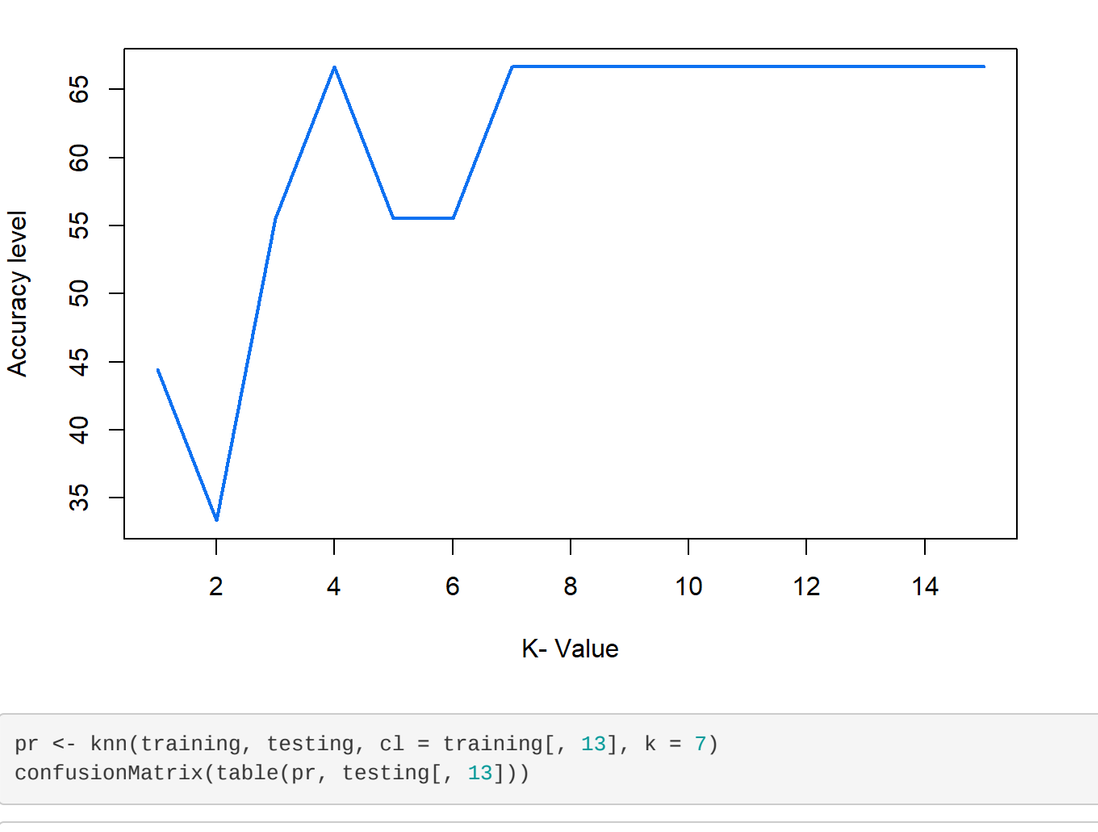
```
##
##  0  1
## 23 18
```

```
set.seed(12345)
split <- sample.split(intrain, SplitRatio = 0.80)
training <- subset(intrain, split == T)
testing <- subset(intrain, split == F)

training$Dx.Cancer = as.factor(training$Dx.Cancer)
testing$Dx.Cancer = as.factor(testing$Dx.Cancer)
```

Classification Model

KNN Model

```
i = 1
k.optm = 1
for (i in 1:15){
  knn.mod <- knn(train = training, test = testing, cl = training[, 13], k = i)
  k.optm[i] <- 100 * sum(testing[, 13] == knn.mod) / NROW(testing[, 13])
  k = i
}

plot(k.optm, type = "l", col = "#0D71F1", lwd = 2, xlab = "K- Value",
     ylab = "Accuracy level")
```




```
pr <- knn(training, testing, cl = training[, 13], k = 7)
confusionMatrix(table(pr, testing[, 13]))
```

```
## Confusion Matrix and Statistics
##
##
## pr  0 1
##   0 3 1
##   1 2 3
##
##                Accuracy : 0.6667
##                  95% CI : (0.2993, 0.9251)
##     No Information Rate : 0.5556
##     P-Value [Acc > NIR] : 0.3743
##
##                   Kappa : 0.3415
##
##  Mcnemar's Test P-Value : 1.0000
##
##             Sensitivity : 0.6000
##             Specificity : 0.7500
##          Pos Pred Value : 0.7500
##          Neg Pred Value : 0.6000
##              Prevalence : 0.5556
##          Detection Rate : 0.3333
##    Detection Prevalence : 0.4444
##       Balanced Accuracy : 0.6750
##
##        'Positive' Class : 0
##
```

Boosted Classification Tree

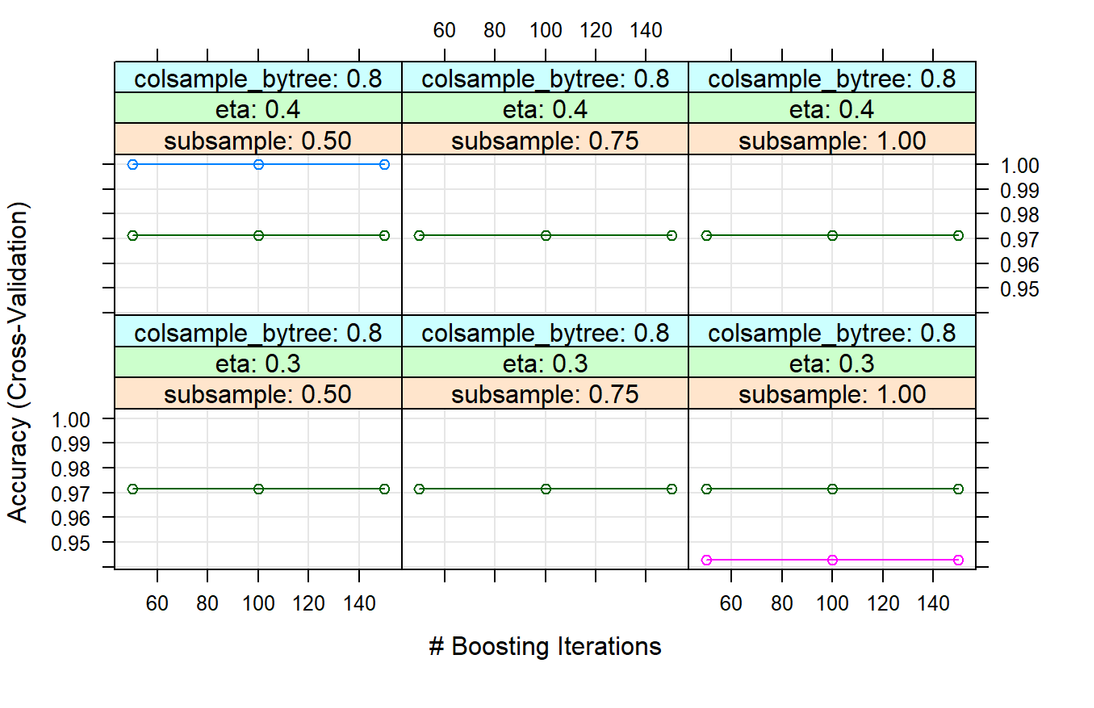
```
set.seed(123)
boosted <- train(
  Dx.Cancer ~., data = training, method = "xgbTree",
  trControl = trainControl("cv", number = 5) # 5 fold - cross validation
)

predicted_boost <- predict(boosted, testing)
caret::confusionMatrix(as.factor(predicted_boost),
                       as.factor(testing$Dx.Cancer))
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction 0 1
##          0 5 1
##          1 0 3
##
##                Accuracy : 0.8889
##                  95% CI : (0.5175, 0.9972)
##     No Information Rate : 0.5556
##     P-Value [Acc > NIR] : 0.04134
##
##                   Kappa : 0.7692
##
##  Mcnemar's Test P-Value : 1.00000
##
##             Sensitivity : 1.0000
##             Specificity : 0.7500
##          Pos Pred Value : 0.8333
##          Neg Pred Value : 1.0000
##              Prevalence : 0.5556
##          Detection Rate : 0.5556
##    Detection Prevalence : 0.6667
##       Balanced Accuracy : 0.8750
##
##        'Positive' Class : 0
##
```

```
varImp(boosted)
```

```
## xgbTree variable importance
##
##   only 20 most important variables shown (out of 21)
##
##                                      Overall
## Dx.HPV1                              100.000
## Dx1                                   29.876
## First.sexual.intercourse               5.281
## Age                                    3.843
## Num.of.pregnancies                     3.710
## STDs.HPV1.0                            0.000
## STDs.vulvo.perineal.condylomatosis1.0  0.000
## STDs.HPV-1                             0.000
## IUD-1                                  0.000
## STDs.Hepatitis.B-1                     0.000
## STDs.Hepatitis.B1.0                    0.000
## STDs.condylomatosis-1                  0.000
## STDs.HIV-1                             0.000
## Hormonal.Contraceptives1.0             0.000
## STDs.HIV1.0                            0.000
## IUD.years                              0.000
## Hormonal.Contraceptives-1              0.000
## STDs.condylomatosis1.0                 0.000
## STDs.AIDS-1                            0.000
## STDs.vulvo.perineal.condylomatosis-1   0.000
```

```
plot(boosted)
```



Conclusion

Having regard to the outcome of accuracy computed, boosted classification tree has to be considered the best model to predict, with a certain probability, the presence of cervical cancer in women. Nevertheless, critical issues have occured using "ovun.sample" function. In fact, data have been drastically reduced in order to solve the unbalanced classes problem, so the results obtained are not greatly reliable.