# Constructing a risk model for cervical cancer

Tom Ertman

8/24/2020

### Introduction

This dataset concists of 858 samples of 36 health features collected from female patients in regards to possible risk factors for cervical cancer. The object of this paper is to construct a risk model that includes factors from various diagnostic test so that a subject's risk may be ranked from 0, indicating low or no risk, to 4, indicating a high risk of cervical cancer. Note that this dataset does not indicate whether a subject has cervical cancer.

### A complete feature list:

```
##  [1] "Age"                                "Number.of.sexual.partners"
##  [3] "First.sexual.intercourse"           "Num.of.pregnancies"
##  [5] "Smokes"                             "Smokes..years."
##  [7] "Smokes..packs.year."                "Hormonal.Contraceptives"
##  [9] "Hormonal.Contraceptives..years."    "IUD"
## [11] "IUD..years."                        "STDs"
## [13] "STDs..number."                      "STDs.condylomatosis"
## [15] "STDs.cervical.condylomatosis"       "STDs.vaginal.condylomatosis"
## [17] "STDs.vulvo.perineal.condylomatosis" "STDs.syphilis"
## [19] "STDs.pelvic.inflammatory.disease"   "STDs.genital.herpes"
## [21] "STDs.molluscum.contagiosum"         "STDs.AIDS"
## [23] "STDs.HIV"                           "STDs.Hepatitis.B"
## [25] "STDs.HPV"                           "STDs..Number.of.diagnosis"
## [27] "STDs..Time.since.first.diagnosis"   "STDs..Time.since.last.diagnosis"
## [29] "Dx.Cancer"                          "Dx.CIN"
## [31] "Dx.HPV"                             "Dx"
## [33] "Hinselmann"                         "Schiller"
## [35] "Citology"                           "Biopsy"
```

The features - Schiller - Hinselmann - Citology - Biopsy

Are all medical tests designed to detect cancerous cells on the cervix.

- Dx.CIN indicated a diagnoses of Cervical intraepithelial neoplasia
- Dx.HPV indicates a diagnoses of Human Pappiloma Virus
- Dx.Cancer indicates a previous diagnoses of cancer
- Dx is unknown and dropped from the study

**Analysis:**

There are a number of challenges with this dataset, namely the unbalanced nature of the postive results in the diagnostic tests which make constructing an accurate model difficult.

```
table(dfile$Schiller)
```

```
##
##   0   1
## 784  74
```

```
table(dfile$Hinselmann)
```

```
##
##   0   1
## 823  35
```

```
table(dfile$Citology)
```

```
##
##   0   1
## 814  44
```

```
table(dfile$Biopsy)
```

```
##
##   0   1
## 803  55
```

There are columns with an significant amount of missing data as illustrated here

```
z<-sapply(dfile,function(x){
  sum(is.na(x))
})

#features missing more than half of data
names(z[which(unname(z)>400)])
```

```
## [1] "STDs..Time.since.first.diagnosis" "STDs..Time.since.last.diagnosis"
```

We will deal with these issues by

1) for columns missing less than 25% of data, we will use imputation methods to assign values to missing features. For continuous data, we will substitute the median value for that column, for factors, the mode.
2) we will discard features with a large amount of missing data >25%

For data modeling, we will chose decision tree and randomforest algorithms using cross validation and feature tuning.

We begin by dropping our Dx feature, then dropping our two columns that have a very high amount of NA

```r
#drop dx
dfile<-dfile[,-32]
#drop the two highest NA features
dfile<-dfile[,-28]
dfile<-dfile[,-27]
```

Now we'll impute values on our dataset with the impute function to assign values to missing data

```r
df2<-imputeMissings::impute(dfile)
#names(df2[nearZeroVar(dfile,freqCut = 99/1)])
#df2<-df2[,-nearZeroVar(dfile)]
```

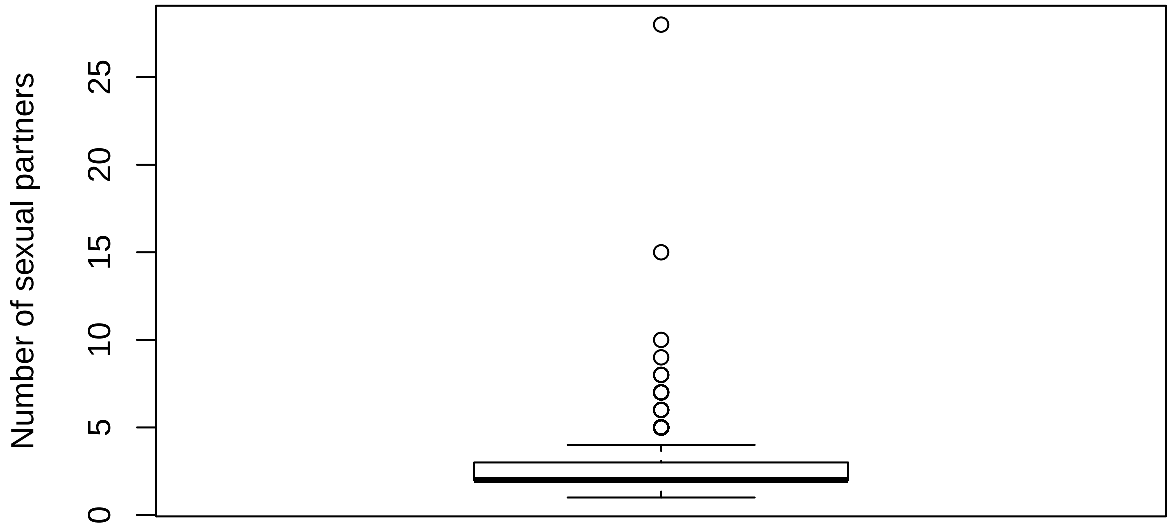
Our dataset is summarized here

```r
summary(df2)
```

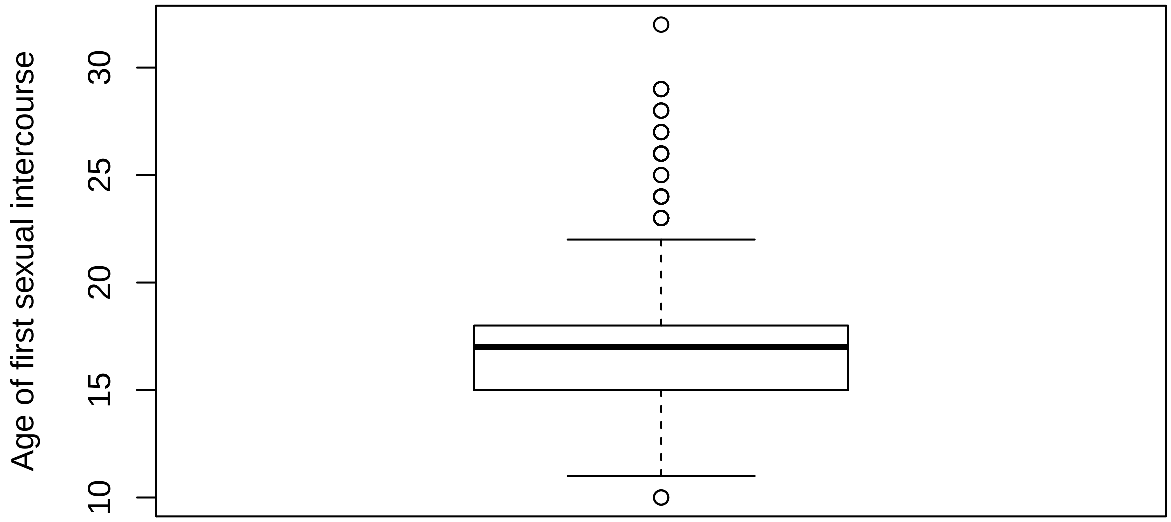
```
##       Age        Number.of.sexual.partners First.sexual.intercourse
##  Min.   :13.00   Min.   : 1.000            Min.   :10
##  1st Qu.:20.00   1st Qu.: 2.000            1st Qu.:15
##  Median :25.00   Median : 2.000            Median :17
##  Mean   :26.82   Mean   : 2.512            Mean   :17
##  3rd Qu.:32.00   3rd Qu.: 3.000            3rd Qu.:18
##  Max.   :84.00   Max.   :28.000            Max.   :32
##  Num.of.pregnancies     Smokes       Smokes..years.   Smokes..packs.year.
##  Min.   : 0.000     Min.   :0.0000   Min.   : 0.000   Min.   : 0.0000
##  1st Qu.: 1.000     1st Qu.:0.0000   1st Qu.: 0.000   1st Qu.: 0.0000
##  Median : 2.000     Median :0.0000   Median : 0.000   Median : 0.0000
##  Mean   : 2.258     Mean   :0.1434   Mean   : 1.201   Mean   : 0.4463
##  3rd Qu.: 3.000     3rd Qu.:0.0000   3rd Qu.: 0.000   3rd Qu.: 0.0000
##  Max.   :11.000     Max.   :1.0000   Max.   :37.000   Max.   :37.0000
##  Hormonal.Contraceptives Hormonal.Contraceptives..years.      IUD
##  Min.   :0.0000          Min.   : 0.000                  Min.   :0.00000
##  1st Qu.:0.0000          1st Qu.: 0.000                  1st Qu.:0.00000
##  Median :1.0000          Median : 0.500                  Median :0.00000
##  Mean   :0.6865          Mean   : 2.035                  Mean   :0.09674
##  3rd Qu.:1.0000          3rd Qu.: 2.000                  3rd Qu.:0.00000
##  Max.   :1.0000          Max.   :30.000                  Max.   :1.00000
##   IUD..years.           STDs          STDs..number.   STDs.condylomatosis
##  Min.   : 0.0000   Min.   :0.00000   Min.   :0.000   Min.   :0.00000
##  1st Qu.: 0.0000   1st Qu.:0.00000   1st Qu.:0.000   1st Qu.:0.00000
##  Median : 0.0000   Median :0.00000   Median :0.000   Median :0.00000
##  Mean   : 0.4446   Mean   :0.09207   Mean   :0.155   Mean   :0.05128
##  3rd Qu.: 0.0000   3rd Qu.:0.00000   3rd Qu.:0.000   3rd Qu.:0.00000
##  Max.   :19.0000   Max.   :1.00000   Max.   :4.000   Max.   :1.00000
##  STDs.cervical.condylomatosis STDs.vaginal.condylomatosis
##  Min.   :0                    Min.   :0.000000
##  1st Qu.:0                    1st Qu.:0.000000
##  Median :0                    Median :0.000000
##  Mean   :0                    Mean   :0.004662
##  3rd Qu.:0                    3rd Qu.:0.000000
##  Max.   :0                    Max.   :1.000000
##  STDs.vulvo.perineal.condylomatosis STDs.syphilis
##  Min.   :0.00000                    Min.   :0.00000
##  1st Qu.:0.00000                    1st Qu.:0.00000
##  Median :0.00000                    Median :0.00000
##  Mean   :0.05012                    Mean   :0.02098
##  3rd Qu.:0.00000                    3rd Qu.:0.00000
##  Max.   :1.00000                    Max.   :1.00000
##  STDs.pelvic.inflammatory.disease STDs.genital.herpes
##  Min.   :0.000000                 Min.   :0.000000
##  1st Qu.:0.000000                 1st Qu.:0.000000
##  Median :0.000000                 Median :0.000000
##  Mean   :0.001166                 Mean   :0.001166
##  3rd Qu.:0.000000                 3rd Qu.:0.000000
##  Max.   :1.000000                 Max.   :1.000000
##  STDs.molluscum.contagiosum   STDs.AIDS    STDs.HIV       STDs.Hepatitis.B
##  Min.   :0.000000           Min.   :0   Min.   :0.00000   Min.   :0.000000
##  1st Qu.:0.000000           1st Qu.:0   1st Qu.:0.00000   1st Qu.:0.000000
##  Median :0.000000           Median :0   Median :0.00000   Median :0.000000
##  Mean   :0.001166           Mean   :0   Mean   :0.02098   Mean   :0.001166
##  3rd Qu.:0.000000           3rd Qu.:0   3rd Qu.:0.00000   3rd Qu.:0.000000
##  Max.   :1.000000           Max.   :0   Max.   :1.00000   Max.   :1.000000
##     STDs.HPV        STDs..Number.of.diagnosis   Dx.Cancer
##  Min.   :0.000000   Min.   :0.00000           Min.   :0.00000
##  1st Qu.:0.000000   1st Qu.:0.00000           1st Qu.:0.00000
##  Median :0.000000   Median :0.00000           Median :0.00000
##  Mean   :0.002331   Mean   :0.08741           Mean   :0.02098
##  3rd Qu.:0.000000   3rd Qu.:0.00000           3rd Qu.:0.00000
##  Max.   :1.000000   Max.   :3.00000           Max.   :1.00000
##      Dx.CIN            Dx.HPV          Hinselmann         Schiller
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000
##  Median :0.00000   Median :0.00000   Median :0.00000   Median :0.00000
##  Mean   :0.01049   Mean   :0.02098   Mean   :0.04079   Mean   :0.08625
##  3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.00000   Max.   :1.00000
##     Citology           Biopsy
##  Min.   :0.00000   Min.   :0.0000
##  1st Qu.:0.00000   1st Qu.:0.0000
##  Median :0.00000   Median :0.0000
##  Mean   :0.05128   Mean   :0.0641
##  3rd Qu.:0.00000   3rd Qu.:0.0000
##  Max.   :1.00000   Max.   :1.0000
```

Here we examine possible outliers to our dataset

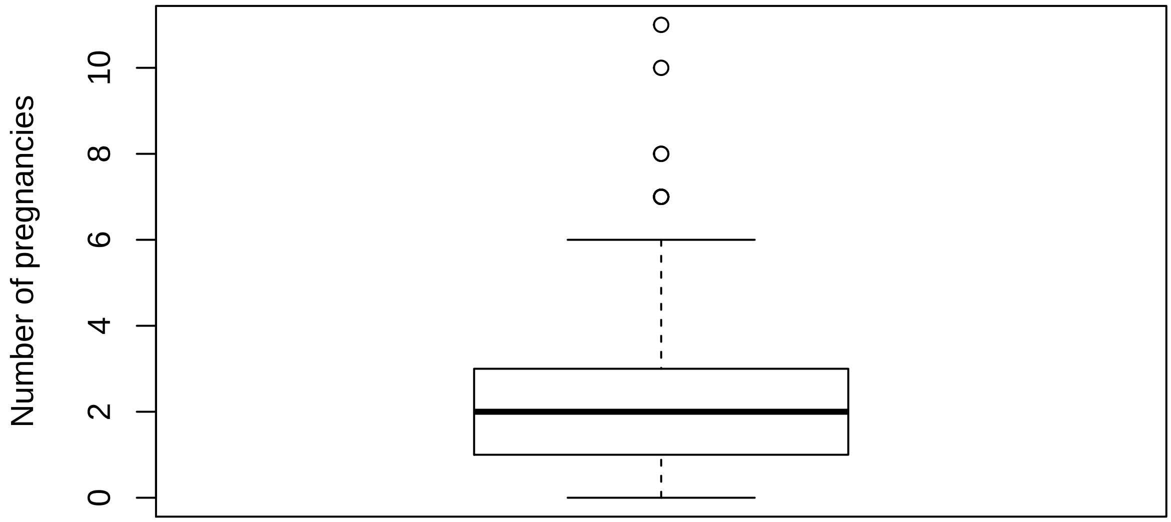
```r
boxplot(df2$Number.of.sexual.partners,ylab="Number of sexual partners")
```



```r
boxplot(df2$First.sexual.intercourse,ylab="Age of first sexual intercourse")
```



```r
boxplot(df2$Num.of.pregnancies,ylab="Number of pregnancies")
```



Upon inspecting the data, we find an outlier to remove hence

```
out_sp<-outliers::outlier(df2$Number.of.sexual.partners)
df2[which(df2$Number.of.sexual.partners==out_sp),]
```

```
##     Age Number.of.sexual.partners First.sexual.intercourse Num.of.pregnancies
## 468  16                        28                       10                  1
##     Smokes Smokes..years. Smokes..packs.year. Hormonal.Contraceptives
## 468      1              5                   5                       0
##     Hormonal.Contraceptives..years. IUD IUD..years. STDs STDs..number.
## 468                               0   0           0    0             0
##     STDs.condylomatosis STDs.cervical.condylomatosis
## 468                   0                            0
##     STDs.vaginal.condylomatosis STDs.vulvo.perineal.condylomatosis
## 468                           0                                  0
##     STDs.syphilis STDs.pelvic.inflammatory.disease STDs.genital.herpes
## 468             0                                0                   0
##     STDs.molluscum.contagiosum STDs.AIDS STDs.HIV STDs.Hepatitis.B STDs.HPV
## 468                          0         0        0                0        0
##     STDs..Number.of.diagnosis Dx.Cancer Dx.CIN Dx.HPV Hinselmann Schiller
## 468                         0         0      0      0          0        0
##     Citology Biopsy
## 468        0      0
```

```
df2<-df2[!df2$Number.of.sexual.partners==out_sp,]
```

Finally, we create our aggragate risk factor:

```r
data_set<-df2%>%mutate(risk_level=Hinselmann+Schiller+Citology+Biopsy)
data_set$risk_level<-factor(data_set$risk_level)
```

Now we seperate into train and test sets

```r
#draw a sample from our completed dataset
set.seed(1999)
index<-createDataPartition(data_set$risk_level,p=0.8)

#sepeate into train and test
training_set<-data_set[index$Resample1,]
test_set<-data_set[-index$Resample1,]
```

For our random forest model, we use 10 cross validations with a grid search for optimal parameters.

```r
contrl=trainControl(method="cv",number=10,search="grid")

#execute the model on our training set
rf_model<-train(risk_level~.,data=training_set,method="rf",trControl=contrl)

#produce the confusion matrix for our results
confusionMatrix(predict(rf_model),training_set$risk_level)
```

```
## Confusion Matrix and Statistics
##
##            Reference
## Prediction   0   1   2   3   4
##          0 604   0   0   0   0
##          1   0  33   0   0   0
##          2   0   0  18   0   0
##          3   0   0   0  27   0
##          4   0   0   0   0   5
##
## Overall Statistics
##
##                Accuracy : 1
##                  95% CI : (0.9946, 1)
##     No Information Rate : 0.8792
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 1
##
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: 0 Class: 1 Class: 2 Class: 3 Class: 4
## Sensitivity            1.0000  1.00000   1.0000   1.0000 1.000000
## Specificity            1.0000  1.00000   1.0000   1.0000 1.000000
## Pos Pred Value         1.0000  1.00000   1.0000   1.0000 1.000000
## Neg Pred Value         1.0000  1.00000   1.0000   1.0000 1.000000
## Prevalence             0.8792  0.04803   0.0262   0.0393 0.007278
## Detection Rate         0.8792  0.04803   0.0262   0.0393 0.007278
## Detection Prevalence   0.8792  0.04803   0.0262   0.0393 0.007278
## Balanced Accuracy      1.0000  1.00000   1.0000   1.0000 1.000000
```

Now we run our model on our test test:

```
final_rf<-predict(rf_model,newdata=test_set)
final_cf<-confusionMatrix(final_rf,test_set$risk_level)
final_cf
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1   2   3   4
##          0 151   0   0   0   0
##          1   0   8   0   0   0
##          2   0   0   4   0   0
##          3   0   0   0   6   0
##          4   0   0   0   0   1
##
## Overall Statistics
##
##                Accuracy : 1
##                  95% CI : (0.9785, 1)
##     No Information Rate : 0.8882
##     P-Value [Acc > NIR] : 1.777e-09
##
##                   Kappa : 1
##
##  Mcnemar’s Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: 0 Class: 1 Class: 2 Class: 3 Class: 4
## Sensitivity            1.0000  1.00000  1.00000  1.00000 1.000000
## Specificity            1.0000  1.00000  1.00000  1.00000 1.000000
## Pos Pred Value         1.0000  1.00000  1.00000  1.00000 1.000000
## Neg Pred Value         1.0000  1.00000  1.00000  1.00000 1.000000
## Prevalence             0.8882  0.04706  0.02353  0.03529 0.005882
## Detection Rate         0.8882  0.04706  0.02353  0.03529 0.005882
## Detection Prevalence   0.8882  0.04706  0.02353  0.03529 0.005882
## Balanced Accuracy      1.0000  1.00000  1.00000  1.00000 1.000000
```

For decision tree model we set up a parameter tuning grid that varies the split, complexity parameter, max depth

```
split<-seq(1,20,2)
cp=seq(.001,.02,.002)
mdepth=seq(20,30,5)



parameters=as.matrix(expand.grid(msplit=split,pval=cp,mxdepth=mdepth))
```

We construct a loop that manually applies each parameter and records the accuracy

```
rpart_test<-function(msplit,p,mxdepth){

  contrl=rpart.control(minsplit=msplit,cp=p,maxdepth=mxdepth)
  dtree=rpart(data=training_set,risk_level~.,control=contrl)
  confusionMatrix(predict(dtree,type="class"),training_set$risk_level)$overall["Accuracy"]

}

acc<-matrix()
for ( i in seq(1,nrow(parameters))){


   acc[i]<-rpart_test(parameters[i,1],parameters[i,2],parameters[i,3])

}
```

Now we apply the optimal parameters to our model

```
index<-first(which(acc==max(acc)))
contrl<-rpart.control(minsplit=parameters[index,1],cp=parameters[index,2],maxdepth=parameters[index,3])

dtree<-rpart(data=training_set,risk_level~.,control=contrl)
confusionMatrix(predict(dtree,type="class"),training_set$risk_level)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1   2   3   4
##          0 604   0   0   0   0
##          1   0  33   0   0   0
##          2   0   0  18   0   0
##          3   0   0   0  27   0
##          4   0   0   0   0   5
##
## Overall Statistics
##
##                Accuracy : 1
##                  95% CI : (0.9946, 1)
##     No Information Rate : 0.8792
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 1
##
##  Mcnemar’s Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: 0 Class: 1 Class: 2 Class: 3 Class: 4
## Sensitivity            1.0000  1.00000   1.0000   1.0000 1.000000
## Specificity            1.0000  1.00000   1.0000   1.0000 1.000000
## Pos Pred Value         1.0000  1.00000   1.0000   1.0000 1.000000
## Neg Pred Value         1.0000  1.00000   1.0000   1.0000 1.000000
## Prevalence             0.8792  0.04803   0.0262   0.0393 0.007278
## Detection Rate         0.8792  0.04803   0.0262   0.0393 0.007278
## Detection Prevalence   0.8792  0.04803   0.0262   0.0393 0.007278
## Balanced Accuracy      1.0000  1.00000   1.0000   1.0000 1.000000
```

Using the optimized hyperparameters for our model we get

```
confusionMatrix(predict(dtree,newdata =test_set,type=’class’),test_set$risk_level)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1   2   3   4
##          0 151   0   0   0   0
##          1   0   8   0   0   0
##          2   0   0   4   0   0
##          3   0   0   0   6   0
##          4   0   0   0   0   1
##
## Overall Statistics
##
##                Accuracy : 1
##                  95% CI : (0.9785, 1)
##     No Information Rate : 0.8882
##     P-Value [Acc > NIR] : 1.777e-09
##
##                   Kappa : 1
##
##  Mcnemar’s Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: 0 Class: 1 Class: 2 Class: 3 Class: 4
## Sensitivity            1.0000  1.00000  1.00000  1.00000 1.000000
## Specificity            1.0000  1.00000  1.00000  1.00000 1.000000
## Pos Pred Value         1.0000  1.00000  1.00000  1.00000 1.000000
## Neg Pred Value         1.0000  1.00000  1.00000  1.00000 1.000000
## Prevalence             0.8882  0.04706  0.02353  0.03529 0.005882
## Detection Rate         0.8882  0.04706  0.02353  0.03529 0.005882
## Detection Prevalence   0.8882  0.04706  0.02353  0.03529 0.005882
## Balanced Accuracy      1.0000  1.00000  1.00000  1.00000 1.000000
```

### Summary

Two models were run on our data and were both accurate in predicting aggragate risk levels associated with cervical cancer. However, we note that the variables used to contruct each respective model differ in importance

```
rforest<-head(arrange(varImp(rf_model)$importance,desc(Overall)),10)
dctree<-head(arrange(varImp(dtree),desc(Overall)),10)
```

For our Random Forest model

```
rforest
```

```
##                                     Overall
## Schiller                        100.0000000
## Citology                         63.2970844
## Biopsy                           34.9116756
## Hinselmann                       16.9386855
## Age                               3.1996203
## Number.of.sexual.partners         1.6711225
## Hormonal.Contraceptives..years.   1.2961201
## First.sexual.intercourse          1.1688094
## Num.of.pregnancies                1.1453710
## IUD..years.                       0.7835994
```

and our Decision tree

```
dctree
```

```
##                              Overall
## Citology                  105.151529
## Biopsy                     84.341873
## Schiller                   66.952077
## Hinselmann                 51.930232
## Age                         8.145089
## Number.of.sexual.partners   6.690218
## IUD..years.                 6.675480
## Dx.Cancer                   3.821026
## STDs.genital.herpes         3.813718
## First.sexual.intercourse    3.318503
```

Using the specificity and sensitiviy data for each test in combination with other tests we should be able to caculate how many subjects develop cancer and if all four diagnostic exams are necessary to make a diagnoses. With image samples from diagnostics test we should be able to identify cancerous cells via machine learning and possibly elminate the need for painful tests such as a biopsy.