Analysis of Cervical Cancer Risk Factors

Mariana Canelas Pais

2024-03-05

Introduction

Cervical cancer is one of the leading causes of cancer-related deaths among women worldwide. The early identification of associated risk factors can significantly contribute to the prevention and effective treatment of this disease. This report focuses on analyzing the dataset concerning cervical cancer risk factors made available by the UCI repository, collected at 'Hospital Universitario de Caracas' in Caracas, Venezuela. The dataset comprises demographic information, habits, and historic medical records of 858 patients. Some patients chose not to answer certain questions due to privacy concerns, leading to missing values in the dataset.

Dataset Characteristics

The dataset is multivariate, covering the health and medicine domain, specifically aimed at classification tasks. It includes both integer and real feature types across various variables.

Variables Description

The dataset contains the following variables, among others, providing a comprehensive overview of each patient's demographic background, habits, and medical history:

- **Age** (int)
- **Number of sexual partners** (int)
- **First sexual intercourse** (int)
- **Number of pregnancies** (int)
- **Smokes** (bool)
- **Smokes (years)** (bool)
- **Smokes (packs/year)** (bool)
- **Hormonal Contraceptives** (bool)
- **Hormonal Contraceptives (years)** (int) contraceptives.
- **IUD** (bool)
- **IUD (years)** (int)
- **STDs** (bool)
- **STDs (number)** (int)
- **STDs: condylomatosis** (bool)
- **STDs: cervical condylomatosis** (bool)
- **STDs: vaginal condylomatosis** (bool)
- **STDs: vulvo-perineal condylomatosis** (bool)
- **STDs: syphilis** (bool)
- **STDs: pelvic inflammatory disease** (bool)
- **STDs: genital herpes** (bool)
- **STDs: molluscum contagiosum** (bool)
- **STDs: AIDS** (bool)
- **STDs: HIV** (bool)

- **STDs: Hepatitis B** (bool)
- **STDs: HPV** (bool)
- **STDs: Number of diagnosis** (int)
- **STDs: Time since first diagnosis** (int)
- **STDs: Time since last diagnosis** (int)
- **Dx:Cancer** (bool) - target variable
- **Dx:CIN** (bool) - target variable
- **Dx:HPV** (bool) - target variable
- **Dx** (bool) - target variable
- **Hinselmann** (bool) - target variable
- **Schiller** (bool) - target variable
- **Cytology** (bool) - target variable
- **Biopsy** (bool) - target variable

Methods

This analysis aims to accurately classify the presence of cervical cancer (`Dx:Cancer`), comparing the performance of single decision trees and random forests. The evaluation of derived models will follow a correct methodology, comparing different estimates of generalization error, such as holdout, cross-validation, and bootstrap methods.

Imports and Data Preparation

To conduct this analysis, we will utilize several key packages within R, which are instrumental in building decision tree models, handling various aspects of data preparation, model training and evaluation, as well as visualization. Below is a brief overview of each package and its role in our analysis: - **rpart**: This package is used for creating decision tree models. It provides functions for building and plotting classification and regression trees. - **rpart.plot**: A companion to rpart, this package offers enhanced functionalities for visualizing decision trees, making it easier to interpret the model structure and decisions. - **caret**: The 'Classification And REgression Training' package is a comprehensive solution for model training. It offers a streamlined workflow for model tuning, training, and performance assessment across a wide range of predictive modeling techniques, including decision trees and random forests. - **ROSE**: 'Random OverSampling Examples' - ROSE is a tool for dealing with imbalanced dataset problems. . **randomForest**: An ensemble method for classification and regression. The randomForest package allows us to build more robust models by creating a 'forest' of decision trees and aggregating their predictions. - **pROC**: Stands for 'Probabilistic ROC' which is essential for evaluating model performance, especially for binary classification problems. The pROC package provides tools for analyzing the performance of predictive models by calculating the area under the ROC (Receiver Operating Characteristic) curve among other functionalities.

```
library(rpart)
library(rpart.plot)
library(caret)
```

```
## Loading required package: ggplot2

## Loading required package: lattice
```

```
library(ROSE)
```

```
## Loaded ROSE 0.0-4
```

```
library(randomForest)
```

```
## randomForest 4.7-1.1

## Type rfNews() to see new features/changes/bug fixes.
```

```
##
## Attaching package: 'randomForest'

## The following object is masked from 'package:ggplot2':
##
##     margin
```

```r
library(pROC)
```

```
## Type 'citation("pROC")' for a citation.

##
## Attaching package: 'pROC'

## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

Reading the Dataset

The dataset is loaded from a CSV file into an R dataframe for manipulation and analysis.

```r
fp <- "assignment_01/cervical.csv"
ds <- read.csv(fp)
```

Data Cleaning

To ensure the quality and usability of the data, we perform several cleaning steps. This includes replacing "?" with NA for missing values, converting certain columns to factors to reflect their categorical nature, and transforming other columns to numeric types to enable quantitative analysis.

Handling Missing Values and Data Types

First, we replace any "?" values with NA across specified columns, acknowledging the presence of missing data in both categorical and numeric fields.

```r
# List of columns to convert to factors
col_to_factor <- c(
  "Smokes",
  "Hormonal.Contraceptives",
  "IUD",
  "STDs",
  "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",
  "Dx.Cancer",
  "Dx.CIN",
  "Dx.HPV",
  "Dx",
  "Hinselmann",
  "Schiller",
  "Citology",
  "Biopsy"
)

# List of columns to convert to numeric
col_to_numeric <- c(
  "Number.of.sexual.partners",
  "First.sexual.intercourse",
  "Num.of.pregnancies",
  "Smokes..years.",
  "Smokes..packs.year.",
  "Hormonal.Contraceptives..years.",
  "STDs..Number.of.diagnosis",
  "STDs..Time.since.first.diagnosis",
  "STDs..Time.since.last.diagnosis",
  "IUD..years.",
  "STDs..number."
)

# Replace "?" with NA and convert data types
ds[c(col_to_factor, col_to_numeric)] <- lapply(ds[c(col_to_factor, col_to_numeric)], function(x) {
  x[x == "?"] <- NA
  return(x)
})
```

Converting Data Types

After handling missing values, we convert the specified columns to their appropriate data types: factors for categorical variables and numeric for continuous variables.

```
# Converting columns to factors
ds[col_to_factor] <- lapply(ds[col_to_factor], factor)

# Converting columns to numeric
ds[col_to_numeric] <- lapply(ds[col_to_numeric], as.numeric)
```

Adjusting Output Variable Category names

```
ds$Dx.Cancer <- factor(ds$Dx.Cancer,
                       levels = levels(ds$Dx.Cancer),
                       labels = make.names(levels(ds$Dx.Cancer), unique = TRUE))
```

Renaming Columns

To improve readability and simplify future references to the dataset columns, we rename them using a more consistent naming convention.

```
# Renaming the columns for better readability
names(ds) <- c("Age", "NumSexualPartners", "FirstSexualIntercoarse", "NumPregnancies", "Smokes", "Smokes
```

Variable Exclusion

To ensure the integrity of our analysis and prevent data leakage, we will remove variables that could introduce bias because they represent outcomes of diagnostics or are directly related to the target variable 'Dx.Cancer'.

```
# Excluding target variables from the dataset
ds <- ds[ , !(names(ds) %in% c("DxCIN", "DxHPV", "Dx", "Hinselmann", "Schiller", "Citology", "Biopsy"))]
```

In refining our dataset for the current exercise, we also undertook further variable selectionto ensure our models rely on variables that provide meaningful, independent insights into the risk factors without being encumbered by redundant or potentially confusing information.

- **Smokes (years)** and **Smokes (packs/year)**: Detailed smoking history variables were excluded in favor of a simpler binary indicator of smoking status. This decision is based on the aim to reduce model complexity and avoid multicollinearity, focusing on whether the patient smokes as the primary smoking-related risk factor.
- **Hormonal Contraceptives (bool)** vs. **Hormonal Contraceptives (years)**: We chose to retain the duration of hormonal contraceptive use over the binary indicator of usage.
- **IUD (bool)** vs. **IUD (years)**: Similarly, the duration of IUD use was retained over the binary indicator.
- **STDs (bool)**, **STDs (number)** and **STDs: Number of diagnosis**: These variables were removed in favor of keeping the individual diagnosis status.

Additionally, preliminary analysis revealed that certain variables, specifically **STDsCervicalCondylomatosis** and **STDsAIDS**, exhibited limited variability within our dataset, having only one category.

```
# Removing other variables from the dataset
ds <- ds[ , !(names(ds) %in% c("SmokesYears", "SmokesPacksYear", "HormonalContraceptives", "IUD", "STDs
```

Results

In this analysis, we aimed to balance our dataset to address the significant class imbalance present and subsequently trained decision tree models using both Gini impurity and information gain as splitting criteria. Here we present the process and findings of our modeling efforts.

Balancing the Dataset

First, we balanced the dataset using the ROSE package's over-sampling method to provide an equal representation of classes. This was crucial for improving our models' ability to learn from the minority class.

```
##
##  X0  X1
## 817 783
```

Creating a Holdout Validation Set

For the correct evaluation of our models, we will hold out a portion of the dataset as a validation set. This allows us to assess the model's performance on unseen data, providing a more accurate estimate of its generalization error. We partition the balanced dataset into training and validation subsets, ensuring that both sets maintain a similar class distribution.

```
# Create indices for a stratified training set, holding out 20% of the data for validation
training.index <- createDataPartition(ds_balanced$DxCancer, p=0.8, list=FALSE)

# Define the validation set
validation <- ds_balanced[-training.index, ]
```

```
# Define the training set
training <- ds_balanced[training.index, ]

# Check the distribution of the target variable in both sets to ensure stratification
cat("Training set distribution of DxCancer:\n")
```

```
## Training set distribution of DxCancer:
```

```
table(training$DxCancer)
```

```
##
##  X0  X1
## 654 627
```

```
cat("\nValidation set distribution of DxCancer:\n")
```

```
##
## Validation set distribution of DxCancer:
```

```
table(validation$DxCancer)
```

```
##
##  X0  X1
## 163 156
```

Model Training

Decision Tree with Gini Impurity

We trained a decision tree model using Gini impurity as the criterion for making splits.

```
tree.gini <- rpart(DxCancer ~ ., data = training, method = "class",
                   parms = list(split = "gini"),
                   control = rpart.control(cp = 0.001, minsplit = 1, maxdepth = 30))
```

Decision Tree with Information Gain

Similarly, we trained another model using information gain (entropy) as the split criterion.

```
tree.information <- rpart(DxCancer ~ ., data=training,
                          parms = list(split = "information"))
```

Random Forest Model

Next, we trained a Random Forest model on the same training set.

```
rf.model <- randomForest(DxCancer ~ ., data=training, method="class", ntree=500, mtry=2, importance=TRU
```

Model Summaries

After fitting our decision tree models using the training data, we can summarize the results to understand the model's complexity, variable importance, and the decision-making process. Below, we present the summaries for both models trained with Gini impurity and information gain as splitting criteria.

Decision Tree Using Gini Impurity

```
summary(tree.gini)
```

```
## Call:
## rpart(formula = DxCancer ~ ., data = training, method = "class",
##     parms = list(split = "gini"), control = rpart.control(cp = 0.001,
##         minsplit = 1, maxdepth = 30))
##   n= 1281
## 
##            CP nsplit rel error   xerror       xstd
## 1 0.015629984      0 1.0000000 1.000000 0.02853517
## 2 0.010366826      5 0.9218501 1.068581 0.02851123
## 3 0.007974482      7 0.9011164 1.038278 0.02853768
## 4 0.005316321      8 0.8931419 1.035088 0.02853900
## 5 0.003189793     14 0.8564593 1.020734 0.02854151
## 6 0.001594896     15 0.8532695 1.028708 0.02854081
## 7 0.001196172     16 0.8516746 1.035088 0.02853900
## 8 0.001000000     20 0.8468900 1.036683 0.02853838
## 
## Variable importance
##                         Age              NumPregnancies 
##                          16                          15 
##           NumSexualPartners                     STDsHPV 
##                          12                          11 
##      FirstSexualIntercoarse HormonalContraceptivesYears 
##                          10                          10 
##                    IUDYears STDsTimeSinceFirstDiagnosis 
##                           8                           7 
##  STDsTimeSinceLastDiagnosis                      Smokes 
##                           6                           3 
##              STDsHepatitisB                     STDsHIV 
##                           1                           1 
##                STDsSyphilis 
##                           1 
## 
## Node number 1: 1281 observations,    complexity param=0.01562998
##   predicted class=X0  expected loss=0.4894614  P(node) =1
##     class counts:   654   627
##    probabilities: 0.511 0.489 
##   left son=2 (12 obs) right son=3 (1269 obs)
##   Primary splits:
##       NumPregnancies                < 0.5   to the left,  improve=5.804109, (0 missing)
##       STDsPelvicInflammatoryDisease splits as  RL,        improve=5.804109, (0 missing)
##       STDsHPV                       splits as  LR,        improve=4.724891, (0 missing)
##       IUDYears                      < 3.5   to the right, improve=2.631457, (0 missing)
##       FirstSexualIntercoarse        < 19.5  to the left,  improve=1.947618, (0 missing)
## 
## Node number 2: 12 observations
##   predicted class=X0  expected loss=0  P(node) =0.009367681
##     class counts:    12     0
##    probabilities: 1.000 0.000 
## 
## Node number 3: 1269 observations,    complexity param=0.01562998
##   predicted class=X0  expected loss=0.4940898  P(node) =0.9906323
##     class counts:   642   627
##    probabilities: 0.506 0.494 
##   left son=6 (1260 obs) right son=7 (9 obs)
##   Primary splits:
##       STDsHPV                     splits as  LR,        improve=4.639919, (0 missing)
##       IUDYears                    < 1.5   to the right, improve=2.904745, (0 missing)
##       FirstSexualIntercoarse      < 19.5  to the left,  improve=1.689383, (0 missing)
##       STDsTimeSinceLastDiagnosis  < 2.5   to the right, improve=1.380550, (0 missing)
##       STDsTimeSinceFirstDiagnosis < 2.5   to the right, improve=1.377681, (0 missing)
## 
## Node number 6: 1260 observations,    complexity param=0.01562998
##   predicted class=X0  expected loss=0.4904762  P(node) =0.9836066
##     class counts:   642   618
##    probabilities: 0.510 0.490 
##   left son=12 (125 obs) right son=13 (1135 obs)
##   Primary splits:
##       IUDYears                    < 3.5   to the right, improve=2.691393, (0 missing)
##       STDsTimeSinceLastDiagnosis  < 2.5   to the right, improve=1.759924, (0 missing)
##       STDsTimeSinceFirstDiagnosis < 2.5   to the right, improve=1.732873, (0 missing)
##       FirstSexualIntercoarse      < 13.5  to the right, improve=1.214542, (0 missing)
##       NumSexualPartners           < 6.5   to the right, improve=1.185420, (0 missing)
##   Surrogate splits:
##       STDsTimeSinceFirstDiagnosis < 21.5  to the right, agree=0.918, adj=0.176, (0 split)
##       STDsTimeSinceLastDiagnosis  < 21.5  to the right, agree=0.918, adj=0.176, (0 split)
## 
## Node number 7: 9 observations
##   predicted class=X1  expected loss=0  P(node) =0.007025761
##     class counts:     0     9
##    probabilities: 0.000 1.000 
## 
## Node number 12: 125 observations
##   predicted class=X0  expected loss=0.392  P(node) =0.09758002
##     class counts:    76    49
##    probabilities: 0.608 0.392 
## 
## Node number 13: 1135 observations,    complexity param=0.01562998
##   predicted class=X1  expected loss=0.4986784  P(node) =0.8860265
##     class counts:   566   569
##    probabilities: 0.499 0.501 
##   left son=26 (45 obs) right son=27 (1090 obs)
##   Primary splits:
##       NumSexualPartners           < 6.5   to the right, improve=1.4303880, (0 missing)
##       HormonalContraceptivesYears < 0.205 to the left,  improve=1.4273440, (0 missing)
##       NumPregnancies              < 3.5   to the left,  improve=1.1362950, (0 missing)
##       FirstSexualIntercoarse      < 13.5  to the right, improve=0.8978331, (0 missing)
##       STDsTimeSinceFirstDiagnosis < 2.5   to the right, improve=0.8588112, (0 missing)
## 
## Node number 26: 45 observations
##   predicted class=X0  expected loss=0.3777778  P(node) =0.03512881
##     class counts:    28    17
##    probabilities: 0.622 0.378 
## 
## Node number 27: 1090 observations,    complexity param=0.01562998
##   predicted class=X1  expected loss=0.493578  P(node) =0.8508977
##     class counts:   538   552
##    probabilities: 0.494 0.506 
##   left son=54 (492 obs) right son=55 (598 obs)
##   Primary splits:
##       HormonalContraceptivesYears < 0.08  to the left,  improve=1.9348760, (0 missing)
##       NumSexualPartners           < 4.5   to the left,  improve=1.4358890, (0 missing)
##       NumPregnancies              < 3.5   to the left,  improve=0.8636377, (0 missing)
##       FirstSexualIntercoarse      < 13.5  to the right, improve=0.7641316, (0 missing)
##       STDsTimeSinceFirstDiagnosis < 2.5   to the right, improve=0.5370910, (0 missing)
##   Surrogate splits:
##       Age                       < 22    to the left,  agree=0.661, adj=0.248, (0 split)
##       Smokes                    splits as  RL,        agree=0.616, adj=0.148, (0 split)
##       FirstSexualIntercoarse    < 21.5  to the right, agree=0.594, adj=0.100, (0 split)
##       STDsVaginalCondylomatosis splits as  RL,        agree=0.590, adj=0.091, (0 split)
##       NumSexualPartners         < 1.5   to the left,  agree=0.588, adj=0.087, (0 split)
##
## Node number 54: 492 observations,    complexity param=0.01036683
##   predicted class=X0  expected loss=0.4735772  P(node) =0.3840749
##     class counts:   259   233
##    probabilities: 0.526 0.474
##   left son=108 (8 obs) right son=109 (484 obs)
##   Primary splits:
##       Age                       < 38    to the right, improve=3.6477190, (0 missing)
##       FirstSexualIntercoarse    < 11.5  to the right, improve=1.1040280, (0 missing)
##       NumPregnancies            < 4.5   to the left,  improve=1.1040280, (0 missing)
##       NumSexualPartners         < 1.5   to the right, improve=0.8775041, (0 missing)
##       STDsVaginalCondylomatosis splits as  LR,        improve=0.7806195, (0 missing)
##
## Node number 55: 598 observations,    complexity param=0.005316321
##   predicted class=X1  expected loss=0.4665552  P(node) =0.4668228
##     class counts:   279   319
##    probabilities: 0.467 0.533
##   left son=110 (539 obs) right son=111 (59 obs)
##   Primary splits:
##       NumSexualPartners           < 4     to the left,  improve=2.734375, (0 missing)
##       FirstSexualIntercoarse      < 20.5  to the left,  improve=1.256944, (0 missing)
##       HormonalContraceptivesYears < 11    to the right, improve=1.161205, (0 missing)
##       STDsTimeSinceFirstDiagnosis < 6     to the right, improve=1.096976, (0 missing)
##       STDsTimeSinceLastDiagnosis  < 6     to the right, improve=1.096976, (0 missing)
##
## Node number 108: 8 observations
##   predicted class=X0  expected loss=0  P(node) =0.006245121
##     class counts:     8     0
##    probabilities: 1.000 0.000
##
## Node number 109: 484 observations,    complexity param=0.01036683
##   predicted class=X0  expected loss=0.481405  P(node) =0.3778298
##     class counts:   251   233
##    probabilities: 0.519 0.481
##   left son=218 (351 obs) right son=219 (133 obs)
##   Primary splits:
##       STDsTimeSinceFirstDiagnosis < 5.5   to the left,  improve=1.6695730, (0 missing)
##       STDsTimeSinceLastDiagnosis  < 5.5   to the left,  improve=1.6695730, (0 missing)
##       FirstSexualIntercoarse      < 11.5  to the right, improve=0.9864179, (0 missing)
##       NumPregnancies              < 4.5   to the left,  improve=0.9864179, (0 missing)
##       Age                         < 28.5  to the left,  improve=0.7501268, (0 missing)
##   Surrogate splits:
```

```
##       STDsTimeSinceLastDiagnosis < 5.5   to the left,  agree=1.000, adj=1.000, (0 split)
##       FirstSexualIntercoarse     < 13.5  to the right, agree=0.835, adj=0.398, (0 split)
##       NumPregnancies             < 3.5   to the left,  agree=0.798, adj=0.263, (0 split)
##       STDsHepatitisB             splits as  LR,        agree=0.793, adj=0.248, (0 split)
##       Age                        < 28.5  to the left,  agree=0.736, adj=0.038, (0 split)
##
## Node number 110: 539 observations,    complexity param=0.005316321
##   predicted class=X1  expected loss=0.4823748  P(node) =0.420765
##     class counts:   260   279
##    probabilities: 0.482 0.518
##   left son=220 (501 obs) right son=221 (38 obs)
##   Primary splits:
##       FirstSexualIntercoarse      < 20.5  to the left,  improve=1.6087600, (0 missing)
##       Smokes                      splits as  RL,        improve=1.1079500, (0 missing)
##       HormonalContraceptivesYears < 11    to the right, improve=0.9195237, (0 missing)
##       STDsTimeSinceLastDiagnosis  < 2.5   to the right, improve=0.5107700, (0 missing)
##       NumSexualPartners           < 2.5   to the right, improve=0.4605854, (0 missing)
##
## Node number 111: 59 observations
##   predicted class=X1  expected loss=0.3220339  P(node) =0.04605777
##     class counts:    19    40
##    probabilities: 0.322 0.678
##
## Node number 218: 351 observations,    complexity param=0.007974482
##   predicted class=X0  expected loss=0.4558405  P(node) =0.2740047
##     class counts:   191   160
##    probabilities: 0.544 0.456
##   left son=436 (246 obs) right son=437 (105 obs)
##   Primary splits:
##       Age                         < 19    to the right, improve=1.3842480, (0 missing)
##       STDsTimeSinceFirstDiagnosis < 2.5   to the right, improve=1.3018700, (0 missing)
##       STDsTimeSinceLastDiagnosis  < 4.5   to the right, improve=1.2325930, (0 missing)
##       NumSexualPartners           < 3.5   to the right, improve=1.0699060, (0 missing)
##       FirstSexualIntercoarse      < 17    to the right, improve=0.3097524, (0 missing)
##   Surrogate splits:
##       NumSexualPartners < 1.5   to the right, agree=0.758, adj=0.19, (0 split)
##
## Node number 219: 133 observations
##   predicted class=X1  expected loss=0.4511278  P(node) =0.1038251
##     class counts:    60    73
##    probabilities: 0.451 0.549
##
## Node number 220: 501 observations,    complexity param=0.005316321
##   predicted class=X1  expected loss=0.493014  P(node) =0.3911007
##     class counts:   247   254
##    probabilities: 0.493 0.507
##   left son=440 (106 obs) right son=441 (395 obs)
##   Primary splits:
##       Smokes                      splits as  RL,        improve=0.7886187, (0 missing)
##       HormonalContraceptivesYears < 11    to the right, improve=0.7726019, (0 missing)
##       FirstSexualIntercoarse      < 19.5  to the right, improve=0.7583701, (0 missing)
##       Age                         < 30.5  to the right, improve=0.5782623, (0 missing)
##       IUDYears                    < 1.5   to the right, improve=0.2733823, (0 missing)
##   Surrogate splits:
```

```
##       FirstSexualIntercoarse      < 19.5  to the right, agree=0.826, adj=0.179, (0 split)
##       STDsTimeSinceFirstDiagnosis < 10.5  to the right, agree=0.802, adj=0.066, (0 split)
##       STDsTimeSinceLastDiagnosis  < 10.5  to the right, agree=0.802, adj=0.066, (0 split)
##       STDsHIV                     splits as  RL,        agree=0.792, adj=0.019, (0 split)
##       NumPregnancies              < 5     to the right, agree=0.790, adj=0.009, (0 split)
##
## Node number 221: 38 observations
##   predicted class=X1  expected loss=0.3421053  P(node) =0.02966432
##     class counts:    13    25
##    probabilities: 0.342 0.658
##
## Node number 436: 246 observations
##   predicted class=X0  expected loss=0.4268293  P(node) =0.1920375
##     class counts:   141   105
##    probabilities: 0.573 0.427
##
## Node number 437: 105 observations,    complexity param=0.003189793
##   predicted class=X1  expected loss=0.4761905  P(node) =0.08196721
##     class counts:    50    55
##    probabilities: 0.476 0.524
##   left son=874 (66 obs) right son=875 (39 obs)
##   Primary splits:
##       Age                         < 17.5  to the left,  improve=0.5394605, (0 missing)
##       STDsSyphilis                splits as  RL,        improve=0.5092946, (0 missing)
##       STDsHIV                     splits as  RL,        improve=0.5092946, (0 missing)
##       STDsTimeSinceFirstDiagnosis < 2.5   to the right, improve=0.5092946, (0 missing)
##       NumSexualPartners           < 1.5   to the right, improve=0.4907085, (0 missing)
##   Surrogate splits:
##       NumPregnancies         < 1.5   to the right, agree=0.819, adj=0.513, (0 split)
##       Smokes                 splits as  LR,        agree=0.819, adj=0.513, (0 split)
##       NumSexualPartners      < 1.5   to the right, agree=0.762, adj=0.359, (0 split)
##       FirstSexualIntercoarse < 15.5  to the left,  agree=0.762, adj=0.359, (0 split)
##
## Node number 440: 106 observations
##   predicted class=X0  expected loss=0.4528302  P(node) =0.08274785
##     class counts:    58    48
##    probabilities: 0.547 0.453
##
## Node number 441: 395 observations,    complexity param=0.005316321
##   predicted class=X1  expected loss=0.478481  P(node) =0.3083528
##     class counts:   189   206
##    probabilities: 0.478 0.522
##   left son=882 (28 obs) right son=883 (367 obs)
##   Primary splits:
##       HormonalContraceptivesYears < 11    to the right, improve=0.9977428, (0 missing)
##       STDsHIV                     splits as  LR,        improve=0.7123776, (0 missing)
##       STDsTimeSinceFirstDiagnosis < 8     to the left,  improve=0.5118396, (0 missing)
##       STDsTimeSinceLastDiagnosis  < 8     to the left,  improve=0.5118396, (0 missing)
##       IUDYears                    < 1.5   to the right, improve=0.4492993, (0 missing)
##   Surrogate splits:
##       IUDYears               < 1.5   to the right, agree=0.965, adj=0.500, (0 split)
##       FirstSexualIntercoarse < 14.5  to the left,  agree=0.937, adj=0.107, (0 split)
##
## Node number 874: 66 observations,    complexity param=0.001594896
##   predicted class=X0  expected loss=0.4848485  P(node) =0.05152225
##     class counts:    34    32
##    probabilities: 0.515 0.485
##   left son=1748 (17 obs) right son=1749 (49 obs)
##   Primary splits:
##       STDsSyphilis                splits as  RL,        improve=0.24460690, (0 missing)
##       STDsHIV                     splits as  RL,        improve=0.24460690, (0 missing)
##       STDsTimeSinceFirstDiagnosis < 2.5   to the right, improve=0.24460690, (0 missing)
##       NumSexualPartners           < 1.5   to the right, improve=0.09945307, (0 missing)
##       FirstSexualIntercoarse      < 15.5  to the left,  improve=0.09945307, (0 missing)
##   Surrogate splits:
##       STDsTimeSinceFirstDiagnosis < 2.5   to the right, agree=1, adj=1, (0 split)
##
## Node number 875: 39 observations
##   predicted class=X1  expected loss=0.4102564  P(node) =0.03044496
##     class counts:    16    23
##    probabilities: 0.410 0.590
##
## Node number 882: 28 observations
##   predicted class=X0  expected loss=0.3928571  P(node) =0.02185792
##     class counts:    17    11
##    probabilities: 0.607 0.393
##
## Node number 883: 367 observations,    complexity param=0.005316321
##   predicted class=X1  expected loss=0.4686649  P(node) =0.2864949
##     class counts:   172   195
##    probabilities: 0.469 0.531
##   left son=1766 (109 obs) right son=1767 (258 obs)
##   Primary splits:
##       HormonalContraceptivesYears < 0.335 to the left,  improve=0.6306535, (0 missing)
##       FirstSexualIntercoarse      < 15.5  to the right, improve=0.5948316, (0 missing)
##       STDsHIV                     splits as  LR,        improve=0.5677099, (0 missing)
##       NumPregnancies              < 1.5   to the left,  improve=0.3916568, (0 missing)
##       STDsTimeSinceFirstDiagnosis < 8     to the left,  improve=0.3318033, (0 missing)
##   Surrogate splits:
##       NumSexualPartners < 1.5   to the left,  agree=0.823, adj=0.404, (0 split)
##       Age               < 18    to the left,  agree=0.760, adj=0.193, (0 split)
##       NumPregnancies    < 1.5   to the left,  agree=0.760, adj=0.193, (0 split)
##       STDsGenitalHerpes splits as  RL,        agree=0.757, adj=0.183, (0 split)
##
## Node number 1748: 17 observations
##   predicted class=X0  expected loss=0.4117647  P(node) =0.01327088
##     class counts:    10     7
##    probabilities: 0.588 0.412
##
## Node number 1749: 49 observations
##   predicted class=X1  expected loss=0.4897959  P(node) =0.03825137
##     class counts:    24    25
##    probabilities: 0.490 0.510
##
## Node number 1766: 109 observations,    complexity param=0.005316321
##   predicted class=X0  expected loss=0.4862385  P(node) =0.08508977
##     class counts:    56    53
##    probabilities: 0.514 0.486
##   left son=3532 (87 obs) right son=3533 (22 obs)
##   Primary splits:
##       FirstSexualIntercoarse      < 15.5  to the right, improve=0.60396120, (0 missing)
##       HormonalContraceptivesYears < 0.205 to the right, improve=0.60396120, (0 missing)
##       Age                         < 33    to the right, improve=0.36433360, (0 missing)
##       NumPregnancies              < 1.5   to the left,  improve=0.33996560, (0 missing)
##       NumSexualPartners           < 2     to the right, improve=0.05446028, (0 missing)
##   Surrogate splits:
##       HormonalContraceptivesYears     < 0.205 to the right, agree=1.000, adj=1.000, (0 split)
##       STDsCondylomatosis              splits as  LR,        agree=0.817, adj=0.091, (0 split)
##       STDsVulvoPerinealCondylomatosis splits as  LR,        agree=0.817, adj=0.091, (0 split)
##
## Node number 1767: 258 observations,    complexity param=0.001196172
##   predicted class=X1  expected loss=0.4496124  P(node) =0.2014052
##     class counts:   116   142
##    probabilities: 0.450 0.550
##   left son=3534 (217 obs) right son=3535 (41 obs)
##   Primary splits:
##       STDsHIV                     splits as  LR,        improve=0.3436260, (0 missing)
##       FirstSexualIntercoarse      < 17.5  to the left,  improve=0.3200771, (0 missing)
##       Age                         < 23.5  to the left,  improve=0.2367061, (0 missing)
##       STDsTimeSinceFirstDiagnosis < 4     to the left,  improve=0.1955751, (0 missing)
##       HormonalContraceptivesYears < 8     to the left,  improve=0.1742046, (0 missing)
##
## Node number 3532: 87 observations
##   predicted class=X0  expected loss=0.4597701  P(node) =0.06791569
##     class counts:    47    40
##    probabilities: 0.540 0.460
##
## Node number 3533: 22 observations
##   predicted class=X1  expected loss=0.4090909  P(node) =0.01717408
##     class counts:     9    13
##    probabilities: 0.409 0.591
##
## Node number 3534: 217 observations,    complexity param=0.001196172
##   predicted class=X1  expected loss=0.4608295  P(node) =0.1693989
##     class counts:   100   117
##    probabilities: 0.461 0.539
##   left son=7068 (153 obs) right son=7069 (64 obs)
##   Primary splits:
##       FirstSexualIntercoarse      < 17.5  to the left,  improve=0.5408007, (0 missing)
##       STDsTimeSinceFirstDiagnosis < 5     to the left,  improve=0.3525629, (0 missing)
##       STDsTimeSinceLastDiagnosis  < 5     to the left,  improve=0.3525629, (0 missing)
##       HormonalContraceptivesYears < 8     to the left,  improve=0.2954782, (0 missing)
##       NumSexualPartners           < 2.5   to the left,  improve=0.2667305, (0 missing)
##   Surrogate splits:
##       STDsTimeSinceFirstDiagnosis < 5     to the left,  agree=0.834, adj=0.438, (0 split)
##       STDsTimeSinceLastDiagnosis  < 5     to the left,  agree=0.834, adj=0.438, (0 split)
##       NumSexualPartners           < 2.5   to the left,  agree=0.770, adj=0.219, (0 split)
##       HormonalContraceptivesYears < 0.71  to the right, agree=0.770, adj=0.219, (0 split)
##       IUDYears                    < 2     to the left,  agree=0.770, adj=0.219, (0 split)
##
## Node number 3535: 41 observations
##   predicted class=X1  expected loss=0.3902439  P(node) =0.03200625
##     class counts:    16    25
##    probabilities: 0.390 0.610 
## 
## Node number 7068: 153 observations,    complexity param=0.001196172
##   predicted class=X1  expected loss=0.4836601  P(node) =0.1194379
##     class counts:    74    79
##    probabilities: 0.484 0.516 
##   left son=14136 (128 obs) right son=14137 (25 obs)
##   Primary splits:
##       Age                             < 17    to the right, improve=0.4183007, (0 missing)
##       FirstSexualIntercoarse          < 14    to the right, improve=0.4183007, (0 missing)
##       HormonalContraceptivesYears     < 1.5   to the right, improve=0.2448839, (0 missing)
##       STDsVulvoPerinealCondylomatosis splits as  RL,        improve=0.2150748, (0 missing)
##       STDsCondylomatosis              splits as  RL,        improve=0.2141177, (0 missing)
##   Surrogate splits:
##       FirstSexualIntercoarse      < 14    to the right, agree=1.000, adj=1.00, (0 split)
##       NumPregnancies              < 1.5   to the right, agree=0.882, adj=0.28, (0 split)
##       STDsTimeSinceFirstDiagnosis < 1.5   to the right, agree=0.863, adj=0.16, (0 split)
##       STDsTimeSinceLastDiagnosis  < 1.5   to the right, agree=0.863, adj=0.16, (0 split)
## 
## Node number 7069: 64 observations
##   predicted class=X1  expected loss=0.40625  P(node) =0.04996097
##     class counts:    26    38
##    probabilities: 0.406 0.594 
## 
## Node number 14136: 128 observations,    complexity param=0.001196172
##   predicted class=X0  expected loss=0.5  P(node) =0.09992194
##     class counts:    64    64
##    probabilities: 0.500 0.500 
##   left son=28272 (21 obs) right son=28273 (107 obs)
##   Primary splits:
##       Age                         < 21    to the left,  improve=0.25634180, (0 missing)
##       STDsTimeSinceFirstDiagnosis < 1.5   to the left,  improve=0.25634180, (0 missing)
##       STDsTimeSinceLastDiagnosis  < 1.5   to the left,  improve=0.25634180, (0 missing)
##       NumSexualPartners           < 2.5   to the left,  improve=0.07729469, (0 missing)
##       IUDYears                    < 0.04  to the left,  improve=0.06477733, (0 missing)
##   Surrogate splits:
##       STDsTimeSinceFirstDiagnosis < 1.5   to the left,  agree=1, adj=1, (0 split)
##       STDsTimeSinceLastDiagnosis  < 1.5   to the left,  agree=1, adj=1, (0 split)
## 
## Node number 14137: 25 observations
##   predicted class=X1  expected loss=0.4  P(node) =0.019516
##     class counts:    10    15
##    probabilities: 0.400 0.600 
## 
## Node number 28272: 21 observations
##   predicted class=X0  expected loss=0.4285714  P(node) =0.01639344
##     class counts:    12     9
##    probabilities: 0.571 0.429 
## 
## Node number 28273: 107 observations
##   predicted class=X1  expected loss=0.4859813  P(node) =0.08352849
##     class counts:    52    55
##    probabilities: 0.486 0.514
```

The Gini model summary indicates the complexity parameter (CP) used at each split, the number of splits (nsplit), the relative error of the model, and the cross-validated relative error (xerror). The most important variables for splitting in the Gini model include Age, FirstSexualIntercoarse, NumPregnancies, and NumSexualPartners, among others. The model attempted several splits, optimizing the CP to reduce overfitting while attempting to capture the complexity of the data.

Decision Tree Using Information Gain

```
summary(tree.information)
```

```
## Call:
## rpart(formula = DxCancer ~ ., data = training, parms = list(split = "information"))
##   n= 1281
##
##           CP nsplit rel error   xerror       xstd
## 1 0.01562998      0 1.0000000 1.000000 0.02853517
## 2 0.01036683      5 0.9218501 1.028708 0.02854081
## 3 0.01000000      7 0.9011164 1.043062 0.02853517
##
## Variable importance
##               NumPregnancies                      STDsHPV
##                           26                           19
##                          Age                     IUDYears
##                           17                            8
## STDsTimeSinceFirstDiagnosis  STDsTimeSinceLastDiagnosis
##                            7                            7
## HormonalContraceptivesYears           NumSexualPartners
##                            6                            5
##      FirstSexualIntercoarse               STDsHepatitisB
##                            3                            1
##                       Smokes   STDsVaginalCondylomatosis
##                            1                            1
##
## Node number 1: 1281 observations,    complexity param=0.01562998
##   predicted class=X0  expected loss=0.4894614  P(node) =1
##     class counts:   654   627
##    probabilities: 0.511 0.489
##   left son=2 (12 obs) right son=3 (1269 obs)
##   Primary splits:
##       NumPregnancies                < 0.5   to the left,  improve=8.121856, (0 missing)
##       STDsPelvicInflammatoryDisease splits as  RL,        improve=8.121856, (0 missing)
##       STDsHPV                       splits as  LR,        improve=6.463262, (0 missing)
##       IUDYears                      < 3.5   to the right, improve=2.654544, (0 missing)
##       FirstSexualIntercoarse        < 19.5  to the left,  improve=1.950426, (0 missing)
##
## Node number 2: 12 observations
##   predicted class=X0  expected loss=0  P(node) =0.009367681
##     class counts:    12     0
##    probabilities: 1.000 0.000
##
## Node number 3: 1269 observations,    complexity param=0.01562998
##   predicted class=X0  expected loss=0.4940898  P(node) =0.9906323
##     class counts:   642   627
##    probabilities: 0.506 0.494
##   left son=6 (1260 obs) right son=7 (9 obs)
##   Primary splits:
##       STDsHPV                     splits as  LR,        improve=6.378255, (0 missing)
##       IUDYears                    < 1.5   to the right, improve=2.921504, (0 missing)
##       FirstSexualIntercoarse      < 19.5  to the left,  improve=1.691982, (0 missing)
##       STDsTimeSinceLastDiagnosis  < 2.5   to the right, improve=1.381120, (0 missing)
##       STDsTimeSinceFirstDiagnosis < 2.5   to the right, improve=1.378244, (0 missing)
## 
## Node number 6: 1260 observations,    complexity param=0.01562998
##   predicted class=X0  expected loss=0.4904762  P(node) =0.9836066
##     class counts:   642   618
##    probabilities: 0.510 0.490 
##   left son=12 (125 obs) right son=13 (1135 obs)
##   Primary splits:
##       IUDYears                    < 3.5   to the right, improve=2.714488, (0 missing)
##       STDsTimeSinceLastDiagnosis  < 2.5   to the right, improve=1.761043, (0 missing)
##       STDsTimeSinceFirstDiagnosis < 2.5   to the right, improve=1.733912, (0 missing)
##       FirstSexualIntercoarse      < 13.5  to the right, improve=1.217098, (0 missing)
##       NumSexualPartners           < 6.5   to the right, improve=1.199128, (0 missing)
##   Surrogate splits:
##       STDsTimeSinceFirstDiagnosis < 21.5  to the right, agree=0.918, adj=0.176, (0 split)
##       STDsTimeSinceLastDiagnosis  < 21.5  to the right, agree=0.918, adj=0.176, (0 split)
## 
## Node number 7: 9 observations
##   predicted class=X1  expected loss=0  P(node) =0.007025761
##     class counts:     0     9
##    probabilities: 0.000 1.000 
## 
## Node number 12: 125 observations
##   predicted class=X0  expected loss=0.392  P(node) =0.09758002
##     class counts:    76    49
##    probabilities: 0.608 0.392 
## 
## Node number 13: 1135 observations,    complexity param=0.01562998
##   predicted class=X1  expected loss=0.4986784  P(node) =0.8860265
##     class counts:   566   569
##    probabilities: 0.499 0.501 
##   left son=26 (45 obs) right son=27 (1090 obs)
##   Primary splits:
##       NumSexualPartners           < 6.5   to the right, improve=1.4441100, (0 missing)
##       HormonalContraceptivesYears < 0.205 to the left,  improve=1.4279460, (0 missing)
##       NumPregnancies              < 3.5   to the left,  improve=1.1372470, (0 missing)
##       FirstSexualIntercoarse      < 13.5  to the right, improve=0.9003039, (0 missing)
##       STDsTimeSinceFirstDiagnosis < 2.5   to the right, improve=0.8590801, (0 missing)
## 
## Node number 26: 45 observations
##   predicted class=X0  expected loss=0.3777778  P(node) =0.03512881
##     class counts:    28    17
##    probabilities: 0.622 0.378 
## 
## Node number 27: 1090 observations,    complexity param=0.01562998
##   predicted class=X1  expected loss=0.493578  P(node) =0.8508977
##     class counts:   538   552
##    probabilities: 0.494 0.506 
##   left son=54 (492 obs) right son=55 (598 obs)
##   Primary splits:
##       HormonalContraceptivesYears < 0.08  to the left,  improve=1.9361930, (0 missing)
##       NumSexualPartners           < 4.5   to the left,  improve=1.4452380, (0 missing)
##       NumPregnancies              < 3.5   to the left,  improve=0.8645657, (0 missing)
##       FirstSexualIntercoarse      < 13.5  to the right, improve=0.7665992, (0 missing)
##       STDsTimeSinceFirstDiagnosis < 2.5   to the right, improve=0.5373142, (0 missing)
##   Surrogate splits:
##       Age                       < 22    to the left,  agree=0.661, adj=0.248, (0 split)
##       Smokes                    splits as  RL,        agree=0.616, adj=0.148, (0 split)
##       FirstSexualIntercoarse    < 21.5  to the right, agree=0.594, adj=0.100, (0 split)
##       STDsVaginalCondylomatosis splits as  RL,        agree=0.590, adj=0.091, (0 split)
##       NumSexualPartners         < 1.5   to the left,  agree=0.588, adj=0.087, (0 split)
## 
## Node number 54: 492 observations,    complexity param=0.01036683
##   predicted class=X0  expected loss=0.4735772  P(node) =0.3840749
##     class counts:   259   233
##    probabilities: 0.526 0.474 
##   left son=108 (8 obs) right son=109 (484 obs)
##   Primary splits:
##       Age                       < 38    to the right, improve=5.1926530, (0 missing)
##       FirstSexualIntercoarse    < 11.5  to the right, improve=1.1108530, (0 missing)
##       NumPregnancies            < 4.5   to the left,  improve=1.1108530, (0 missing)
##       NumSexualPartners         < 1.5   to the right, improve=0.8791395, (0 missing)
##       STDsVaginalCondylomatosis splits as  LR,        improve=0.7817627, (0 missing)
## 
## Node number 55: 598 observations
##   predicted class=X1  expected loss=0.4665552  P(node) =0.4668228
##     class counts:   279   319
##    probabilities: 0.467 0.533 
## 
## Node number 108: 8 observations
##   predicted class=X0  expected loss=0  P(node) =0.006245121
##     class counts:     8     0
##    probabilities: 1.000 0.000 
## 
## Node number 109: 484 observations,    complexity param=0.01036683
##   predicted class=X0  expected loss=0.481405  P(node) =0.3778298
##     class counts:   251   233
##    probabilities: 0.519 0.481 
##   left son=218 (351 obs) right son=219 (133 obs)
##   Primary splits:
##       STDsTimeSinceFirstDiagnosis < 5.5   to the left,  improve=1.6722970, (0 missing)
##       STDsTimeSinceLastDiagnosis  < 5.5   to the left,  improve=1.6722970, (0 missing)
##       FirstSexualIntercoarse      < 11.5  to the right, improve=0.9929083, (0 missing)
##       NumPregnancies              < 4.5   to the left,  improve=0.9929083, (0 missing)
##       Age                         < 28.5  to the left,  improve=0.7509295, (0 missing)
##   Surrogate splits:
##       STDsTimeSinceLastDiagnosis < 5.5   to the left,  agree=1.000, adj=1.000, (0 split)
##       FirstSexualIntercoarse     < 13.5  to the right, agree=0.835, adj=0.398, (0 split)
##       NumPregnancies             < 3.5   to the left,  agree=0.798, adj=0.263, (0 split)
##       STDsHepatitisB             splits as  LR,        agree=0.793, adj=0.248, (0 split)
##       Age                        < 28.5  to the left,  agree=0.736, adj=0.038, (0 split)
## 
```

```
## Node number 218: 351 observations
##   predicted class=X0  expected loss=0.4558405  P(node) =0.2740047
##     class counts:   191   160
##    probabilities: 0.544 0.456
##
## Node number 219: 133 observations
##   predicted class=X1  expected loss=0.4511278  P(node) =0.1038251
##     class counts:    60    73
##    probabilities: 0.451 0.549
```

The summary of the tree trained using information gain shows a similar structure, with the model highlighting the variable importance based on the information gain criterion. Variables like NumPregnancies, STDsHPV, Age, and IUDYears played significant roles in the model's decision-making process. This model also details the nodes, splits, and the primary and surrogate splits at each node, demonstrating how the model decides the predicted class.

Random Forest Model Summary

The following code obtains a summary pertraining the Random Forest model.

```
summary(rf.model)
```

```
##                 Length Class  Mode
## call               7   -none- call
## type               1   -none- character
## predicted       1281   factor numeric
## err.rate        1500   -none- numeric
## confusion          6   -none- numeric
## votes           2562   matrix numeric
## oob.times       1281   -none- numeric
## classes            2   -none- character
## importance        76   -none- numeric
## importanceSD      57   -none- numeric
## localImportance    0   -none- NULL
## proximity          0   -none- NULL
## ntree              1   -none- numeric
## mtry               1   -none- numeric
## forest            14   -none- list
## y               1281   factor numeric
## test               0   -none- NULL
## inbag              0   -none- NULL
## terms              3   terms  call
```

The summary of the Random Forest model provides a comprehensive overview of the model's training process and its outcomes. Notably, the model was trained with a large number of decision trees (ntree), which collectively contribute to the final prediction through a majority vote mechanism.

From the output, the importance measures stand out as particularly insightful, highlighting the variables that play significant roles in the model's predictions.

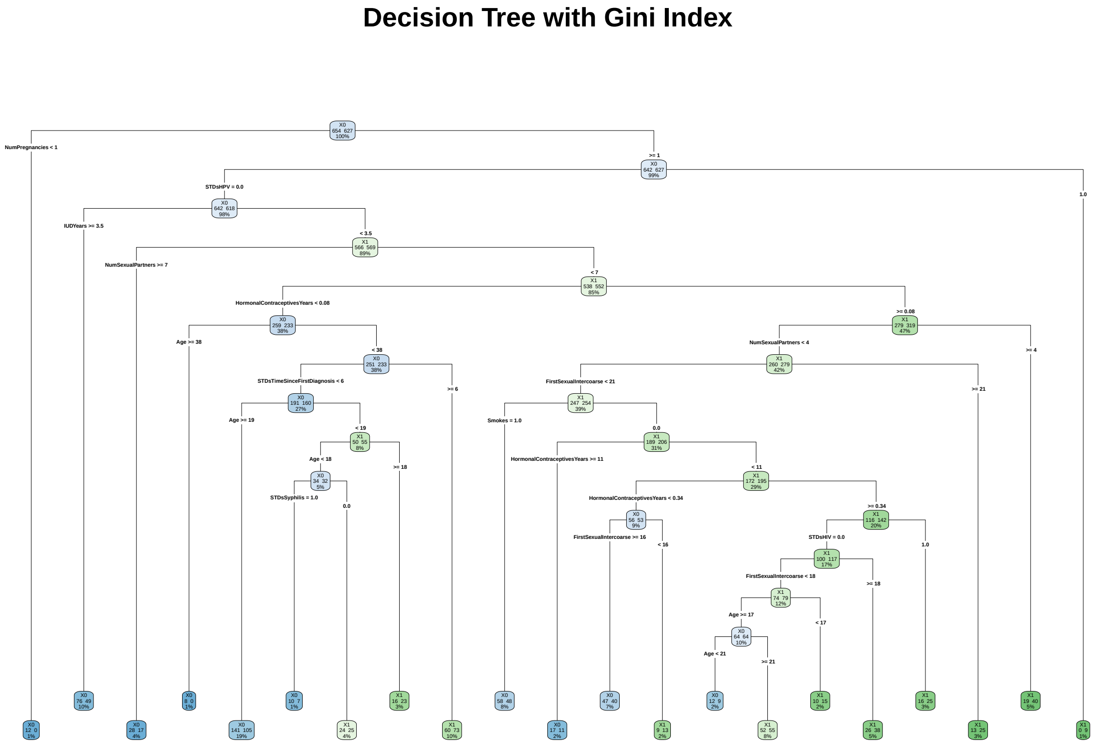
Visualization of Decision Trees

Next, we visualize the structure of the trained decision trees to interpret the splits and leaf nodes.

Decision Tree Using Gini Impurity

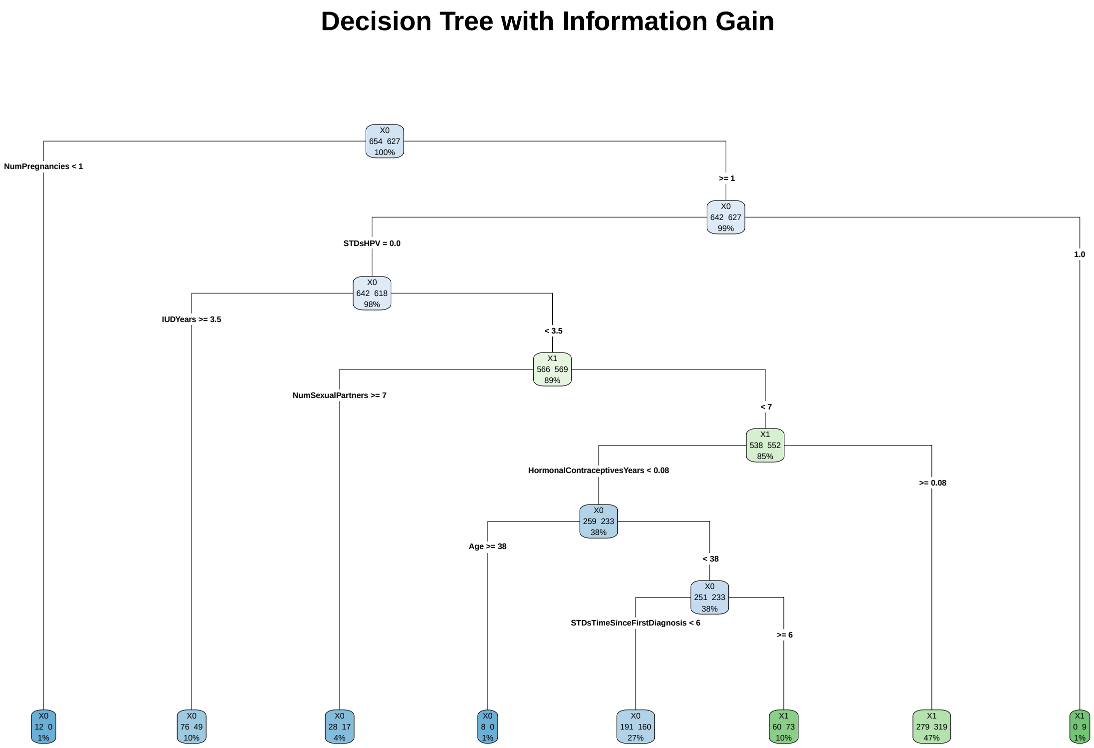
```r
rpart.plot(tree.gini, main = "Decision Tree with Gini Index", type = 4, extra = 101)
```

```
## Warning: labs do not fit even at cex 0.15, there may be some overplotting
```



Decision Tree Using Information Gain

```r
rpart.plot(tree.information, main = "Decision Tree with Information Gain", type = 4, extra = 101)
```



Model Evaluation

Model evaluation will enable us to understand the performance and generalizability of our predictive models. We employ repeated 10-fold cross-validation to evaluate the models, focusing on the ROC AUC as our primary metric.

Setting Up Cross-Validation

```
metric <- "ROC"
control <- trainControl(method="repeatedcv", number=10,
                        summaryFunction=twoClassSummary,
                        classProbs=TRUE,
                        savePredictions = "final", repeats = 3)
```

Evaluating Decision Tree Model with Gini Impurity

```
fit.tree.gini.rcv <- train(DxCancer ~ ., data=training, method="rpart", metric=metric, trControl=control
```

Evaluating Random Forest Model

```
fit.rf.rcv <- train(DxCancer ~ ., data=training, method="rf", metric=metric, trControl=control)
```

Comparing Model Performance

Now we summarize the accuracy of the models and visualize their ROC curves to compare their performance.

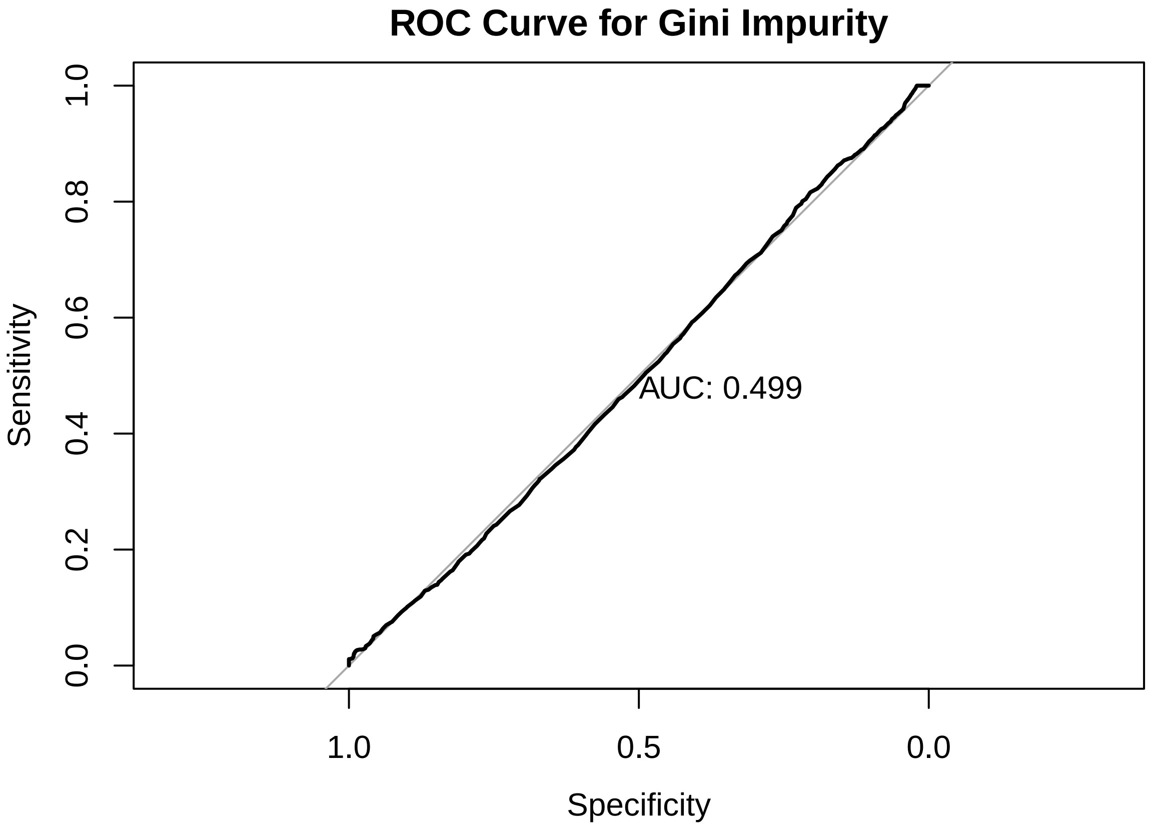
```
fit.models <- list(Tree.Gini=fit.tree.gini.rcv, RF=fit.rf.rcv)
results <- resamples(fit.models)
summary(results)
```

```
##
## Call:
## summary.resamples(object = results)
##
## Models: Tree.Gini, RF
## Number of resamples: 30
##
## ROC
##                Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## Tree.Gini 0.3819144 0.4711764 0.5167148 0.5043105 0.5384376 0.5905983    0
## RF        0.3953824 0.4805039 0.5012210 0.5101214 0.5546093 0.5839346    0
##
## Sens
##                Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## Tree.Gini 0.2575758 0.3869464 0.4772727 0.4893784 0.5961538 0.7076923    0
## RF        0.4242424 0.4942308 0.5266900 0.5403108 0.5692308 0.6969697    0
##
## Spec
##                Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## Tree.Gini 0.2222222 0.4343318 0.5000000 0.5031746 0.5775730 0.7142857    0
## RF        0.3225806 0.3968254 0.4480287 0.4454856 0.4880952 0.6507937    0
```

```
# Plotting ROC curve for model1
numeric_obs_gini <- ifelse(as.character(fit.models$Tree.Gini$pred$obs) == "X1", 1, 0)
roc(numeric_obs_gini, fit.models$Tree.Gini$pred$X1, plot = TRUE, main="ROC Curve for Gini Impurity", pri
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

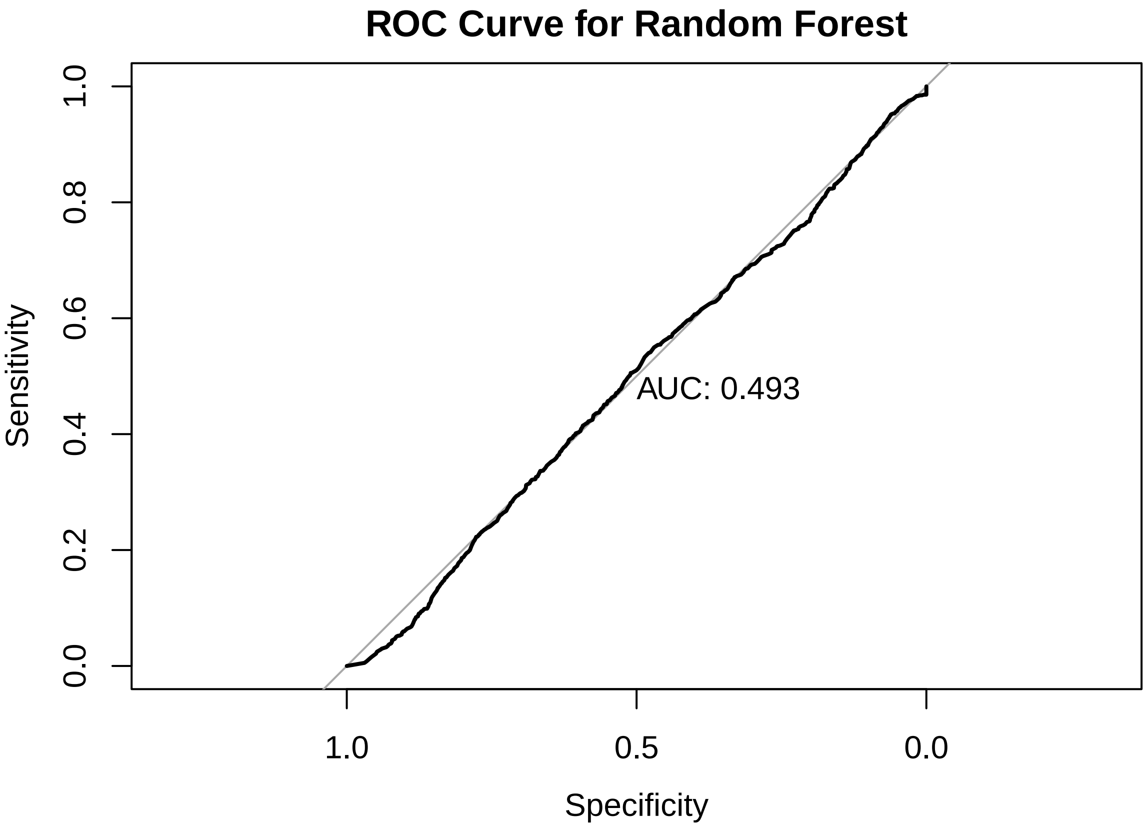


```
##
## Call:
## roc.default(response = numeric_obs_gini, predictor = fit.models$Tree.Gini$pred$X1,     plot = TRUE, m
##
## Data: fit.models$Tree.Gini$pred$X1 in 1962 controls (numeric_obs_gini 0) < 1881 cases (numeric_obs_gi
## Area under the curve: 0.4992
```

```
# Plotting ROC curve for model2
numeric_obs_rf <- ifelse(as.character(fit.models$RF$pred$obs) == "X1", 1, 0)
roc(numeric_obs_rf, fit.models$RF$pred$X1, plot = TRUE, main="ROC Curve for Random Forest", print.auc=TR
```

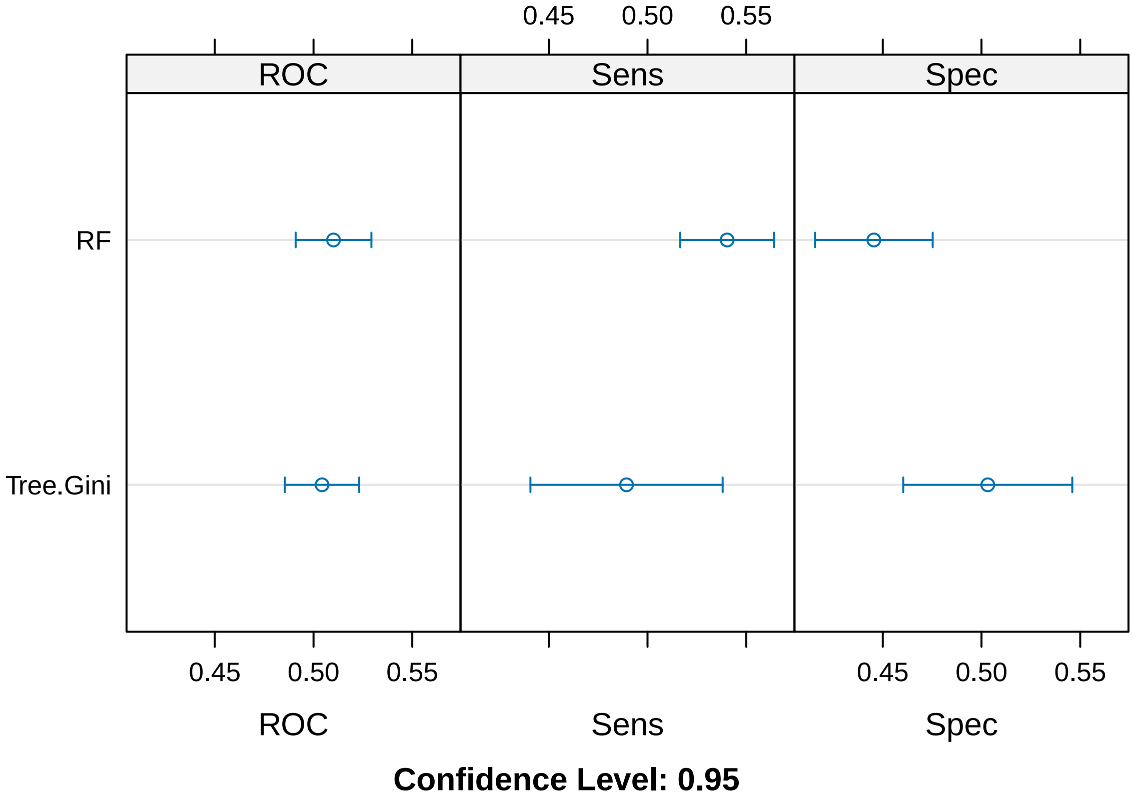
```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls > cases
```



```
##
## Call:
## roc.default(response = numeric_obs_rf, predictor = fit.models$RF$pred$X1,     plot = TRUE, main = "RO
##
## Data: fit.models$RF$pred$X1 in 1962 controls (numeric_obs_rf 0) > 1881 cases (numeric_obs_rf 1).
## Area under the curve: 0.4932
```

```
# Compare accuracy of models using dotplot
dotplot(results)
```



Discussion and Conclusion

The analysis revealed that both the Gini-based decision tree and random forest models performed poorly (AUC of 0.499 and 0.493, respectively), with low sensitivity and specificity, indicating an inability to effectively discriminate between the classes. Further investigation into data quality, feature engineering, and alternative modeling approaches would be necessary to improve performance. However, due to contextual and time limitations, a deeper exploration was not feasible within the scope of this exercise.