BIOPSY CLASSIFICATION OF CERVICAL CANCER

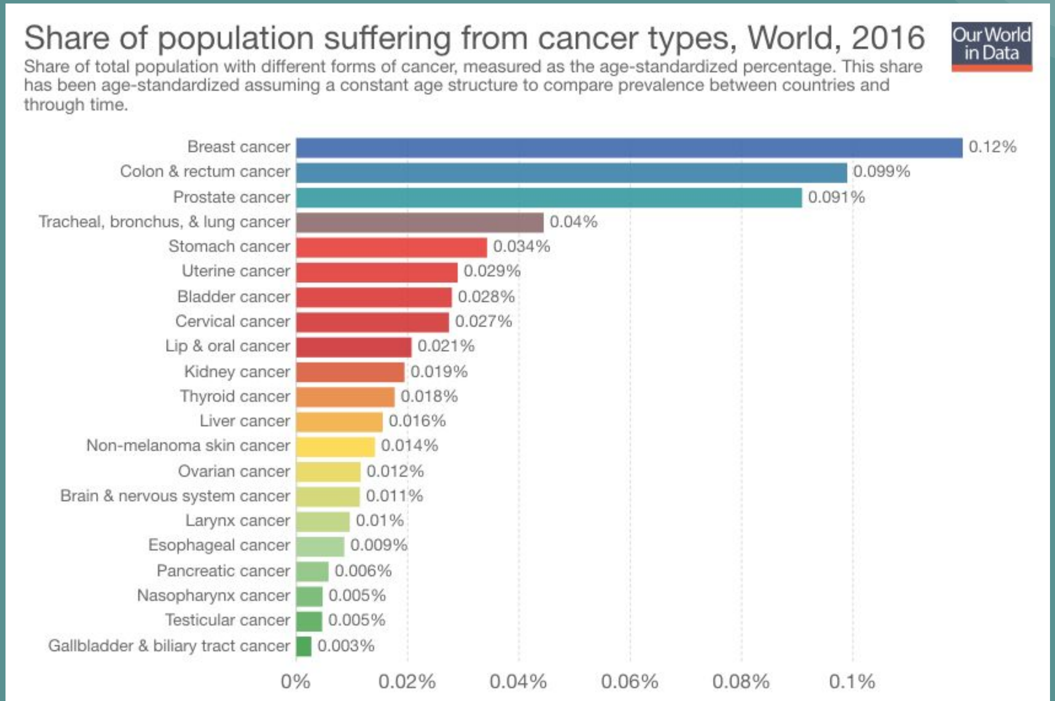
Introduction

Third most common cancer in women worldwide.

500,000 women worldwide die of cervical cancer annually, of which 4,100 deaths are from the USA.

Most common cause of cancer death where proper tests are not available.

Easiest gynecologic cancer to prevent through screening.

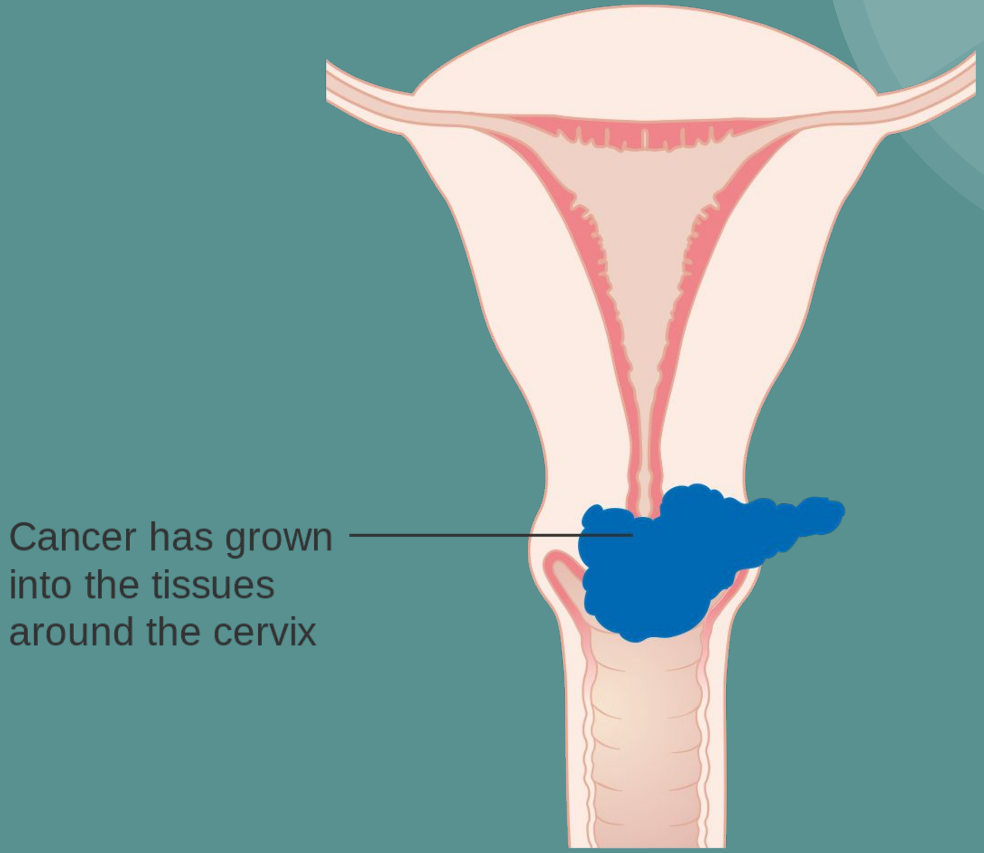


Share of population suffering from cancer types, World, 2016

Share of total population with different forms of cancer, measured as the age-standardized percentage. This share has been age-standardized assuming a constant age structure to compare prevalence between countries and through time.

Cervical Cancer

What causes cervical cancer: HPV (human papillomavirus). It's the most common sexually transmitted infection. **HPV** is usually harmless and goes away by itself, but some types can lead to cancer.



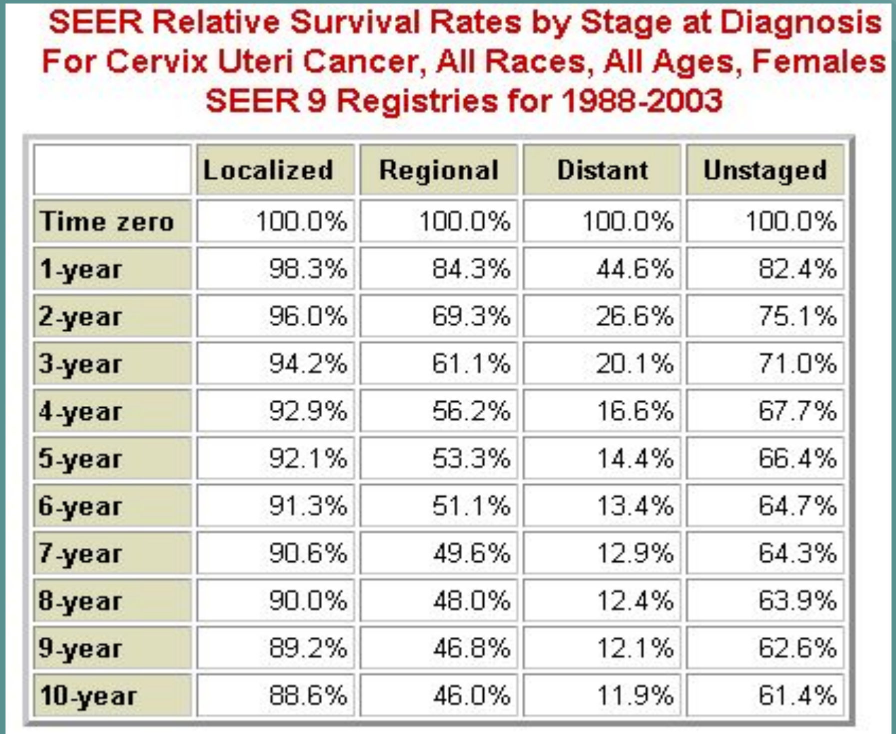
Activities that increase the risk of HPV and for cervical cancer are:

- Numbers of sexual partners
- Use of birth control contraception pills for extended periods of time
- Numbers of Sexually Transmitted Diseases
- Giving birth to three or more children

How it’s diagnosed

Screening: The earlier cervical cancer is detected, the higher the chance of it being treated successfully. Pap tests are used to detect abnormal cells within the cervix.

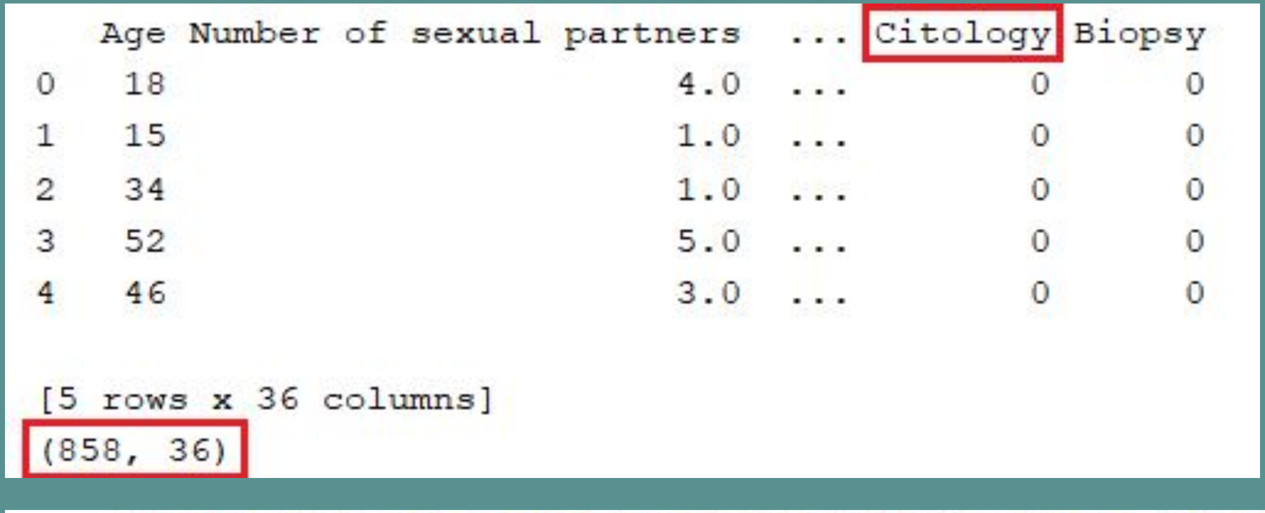
Biopsy: If cancer is suspected, the doctor will take a sample of cervical cells (biopsy) and conduct either a Punch or Cone Biopsy



SEER Relative Survival Rates by Stage at Diagnosis
For Cervix Uteri Cancer, All Races, All Ages, Females
SEER 9 Registries for 1988-2003

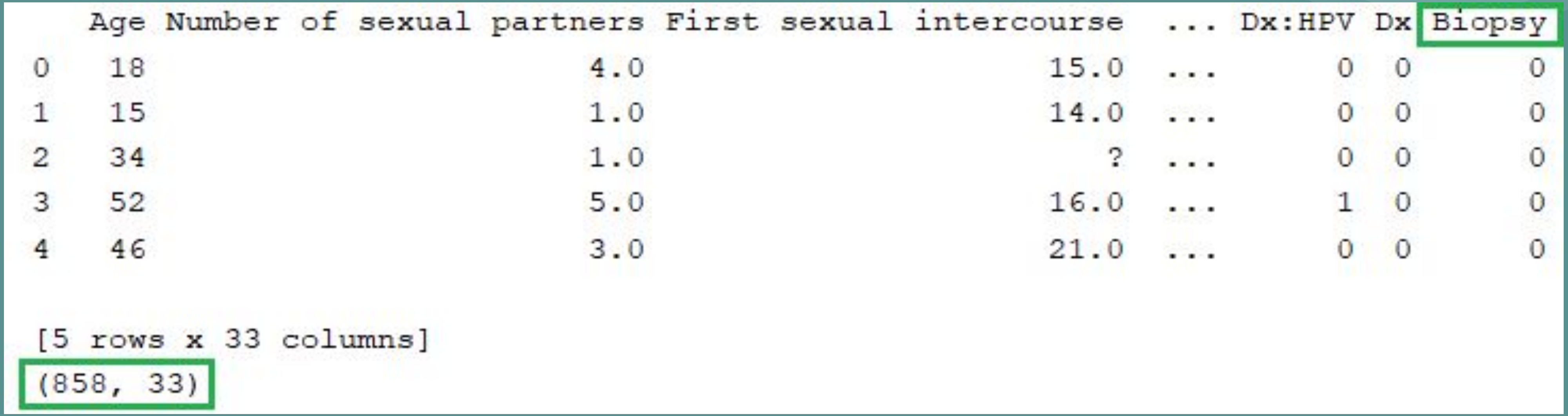
	Localized	Regional	Distant	Unstaged
Time zero	100.0%	100.0%	100.0%	100.0%
1-year	98.3%	84.3%	44.6%	82.4%
2-year	96.0%	69.3%	26.6%	75.1%
3-year	94.2%	61.1%	20.1%	71.0%
4-year	92.9%	56.2%	16.6%	67.7%
5-year	92.1%	53.3%	14.4%	66.4%
6-year	91.3%	51.1%	13.4%	64.7%
7-year	90.6%	49.6%	12.9%	64.3%
8-year	90.0%	48.0%	12.4%	63.9%
9-year	89.2%	46.8%	12.1%	62.6%
10-year	88.6%	46.0%	11.9%	61.4%

Initial Data



```
   Age  Number of sexual partners  ...  Citology  Biopsy
0   18                        4.0  ...         0       0
1   15                        1.0  ...         0       0
2   34                        1.0  ...         0       0
3   52                        5.0  ...         0       0
4   46                        3.0  ...         0       0

[5 rows x 36 columns]
(858, 36)
```



```
   Age  Number of sexual partners  First sexual intercourse  ...  Dx:HPV  Dx  Biopsy
0   18                        4.0                      15.0  ...       0   0       0
1   15                        1.0                      14.0  ...       0   0       0
2   34                        1.0                         ?  ...       0   0       0
3   52                        5.0                      16.0  ...       1   0       0
4   46                        3.0                      21.0  ...       0   0       0

[5 rows x 33 columns]
(858, 33)
```

Prior to the Preprocessing of the Data

Model	Accuracy
Decision Tree	94.44
Random Forest	83.33
Support Vector Machine	88.88
*K-Nearest Neighbors	83.33
Naïve Bayes	83.33

The table below demonstrates the comparison of accuracies of all the models run prior to any cleaning methodology

Prior to the Preprocessing of the Data Continued

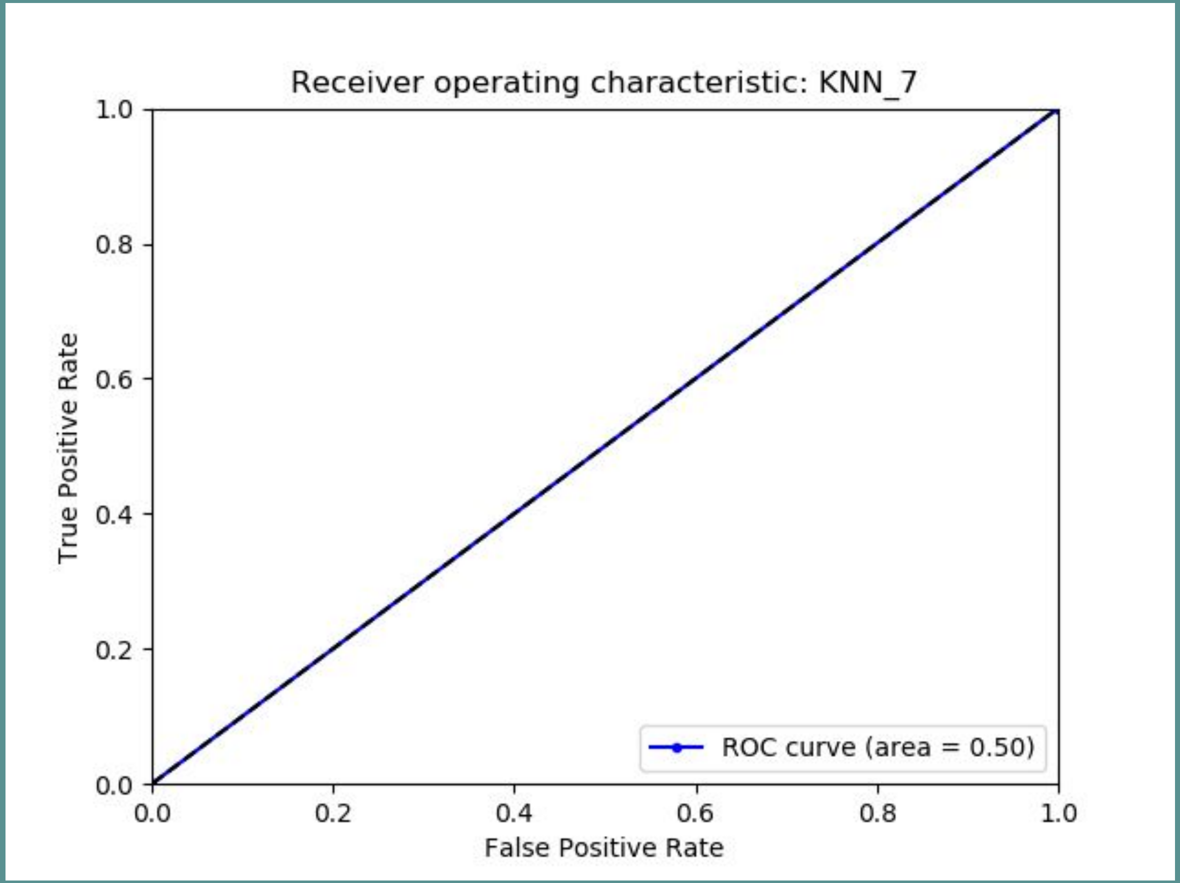
Models	Precision	Recall	F1 Score
Decision Tree	1	0.93	0.97
Random Forest	0.83	1.00	0.91
Support Vector Machine	0.88	1.00	0.94
K-Nearest Neighbors	0.83	1	0.91
Naïve Bayes	0.88	0.93	0.90

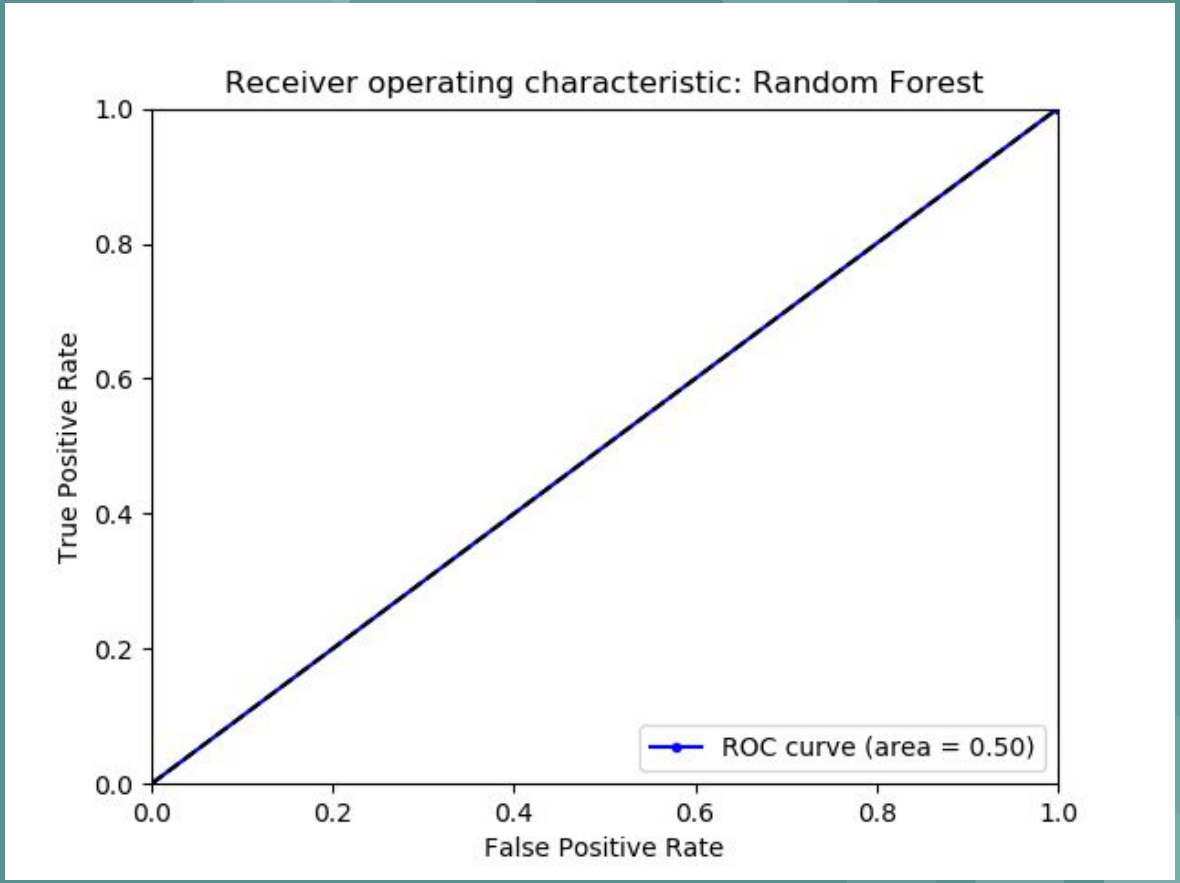
For women without cervical cancer

Models	Precision	Recall	F1 Score
Decision Tree	0.75	1	0.86
Random Forest	0	0	0
Support Vector Machine	1	0.33	0.50
K-Nearest Neighbors	0	0	0
Naïve Bayes	0.50	0.33	0.40

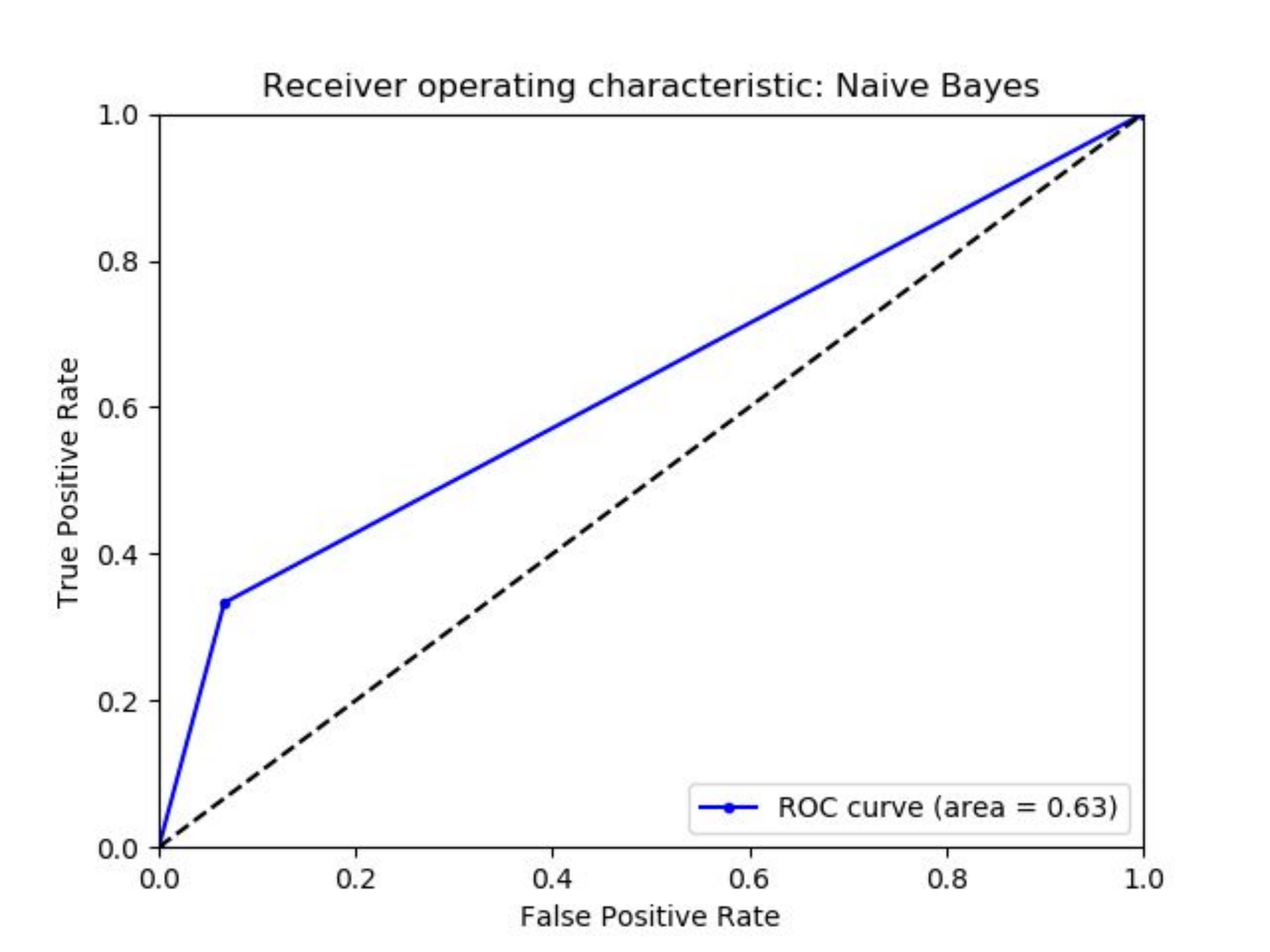
For women with cervical cancer

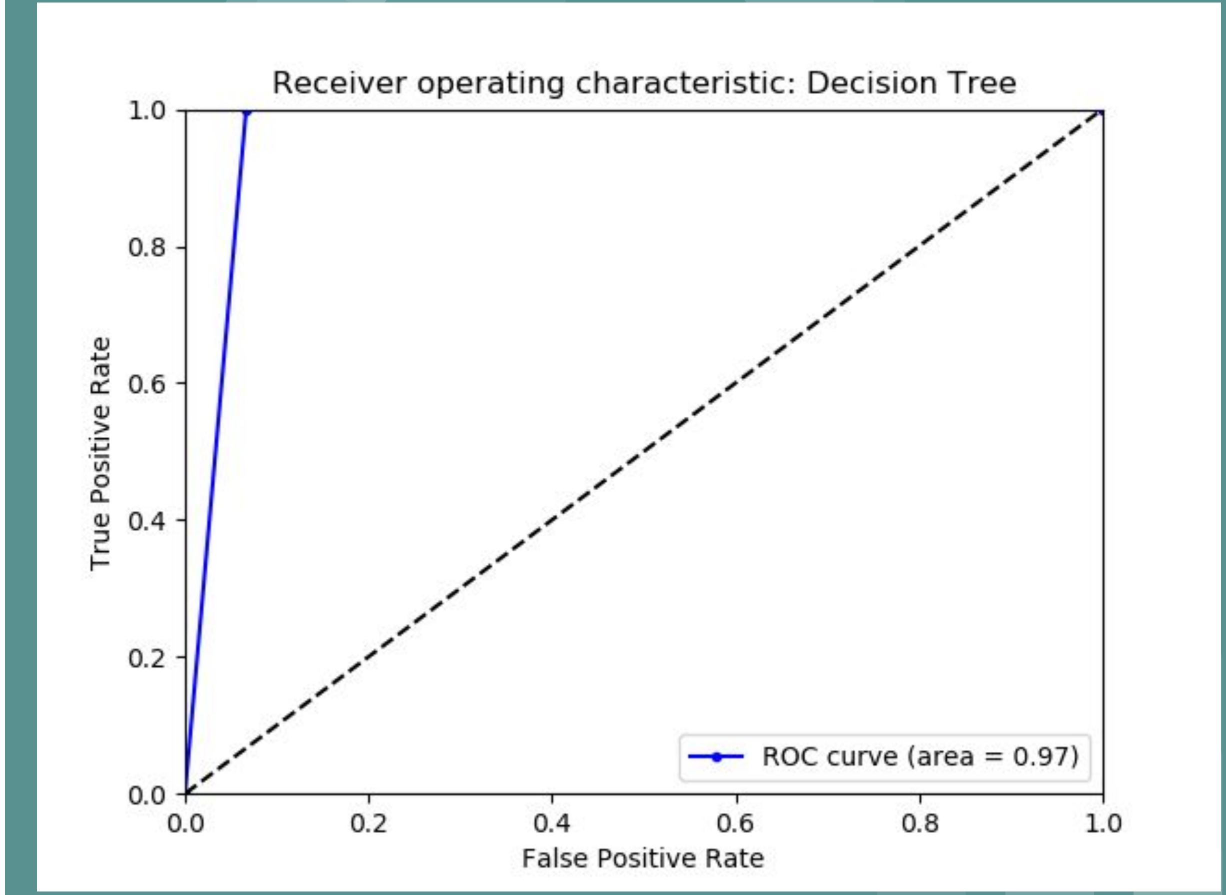
ROC Curves


Receiver operating characteristic: KNN_7
1.0
0.8
0.6
0.4
0.2
0.0
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
ROC curve (area = 0.50)

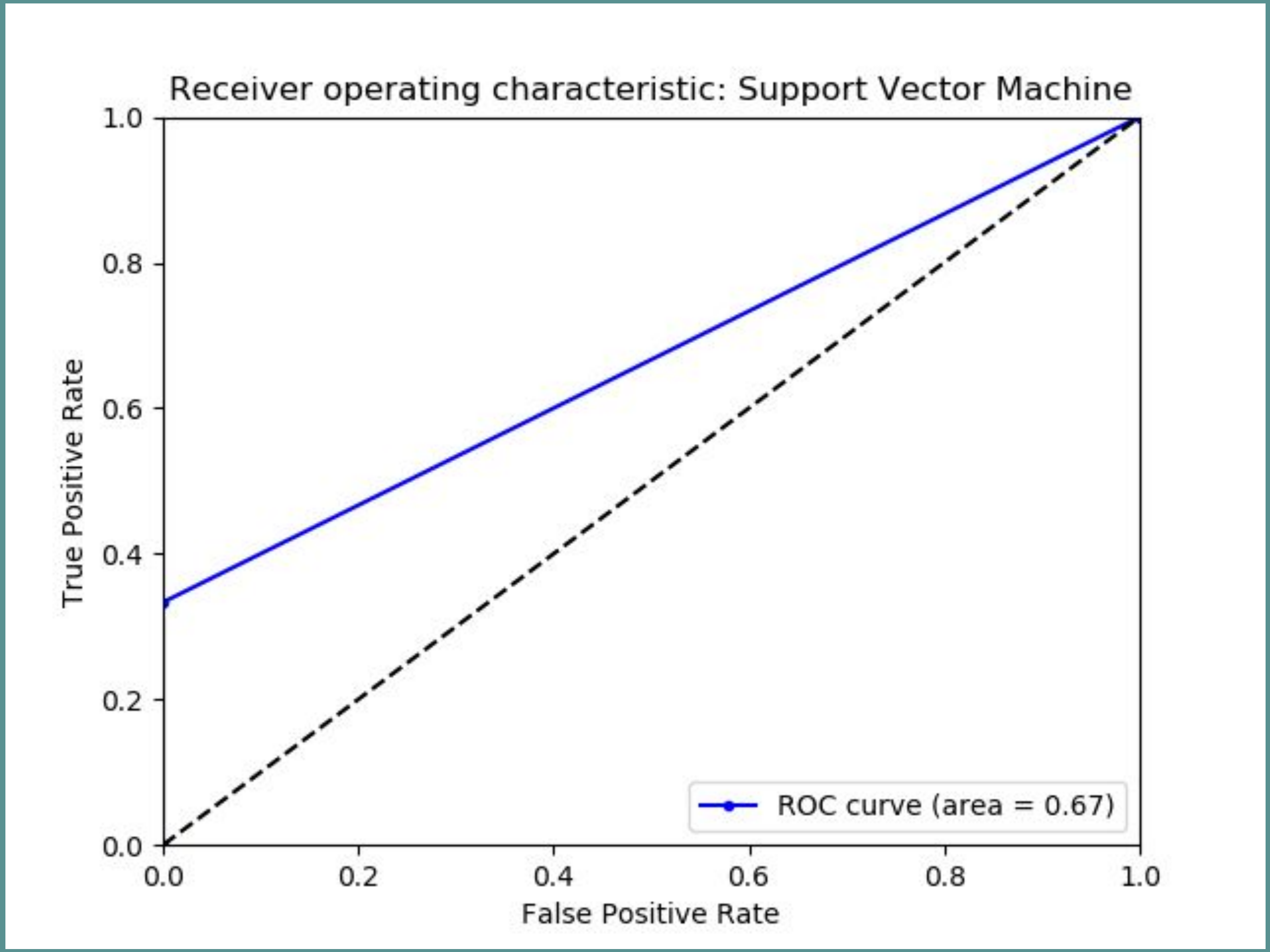

Receiver operating characteristic: Random Forest
1.0
0.8
0.6
0.4
0.2
0.0
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
ROC curve (area = 0.50)

ROC Curves

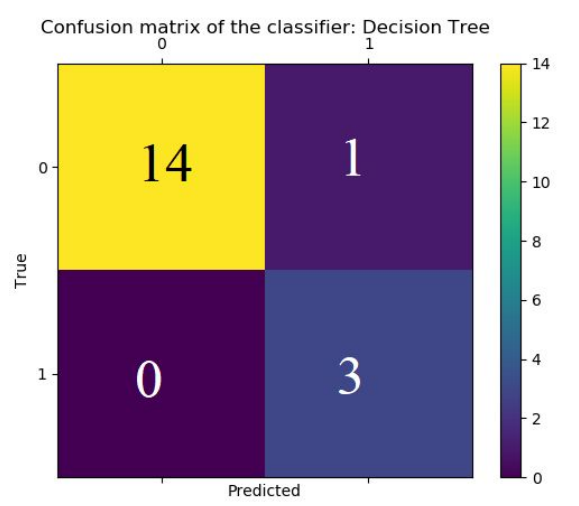

Receiver operating characteristic: Naive Bayes
1.0
0.8
0.6
0.4
0.2
0.0
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
ROC curve (area = 0.63)


Receiver operating characteristic: Decision Tree
1.0
0.8
0.6
0.4
0.2
0.0
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
ROC curve (area = 0.97)

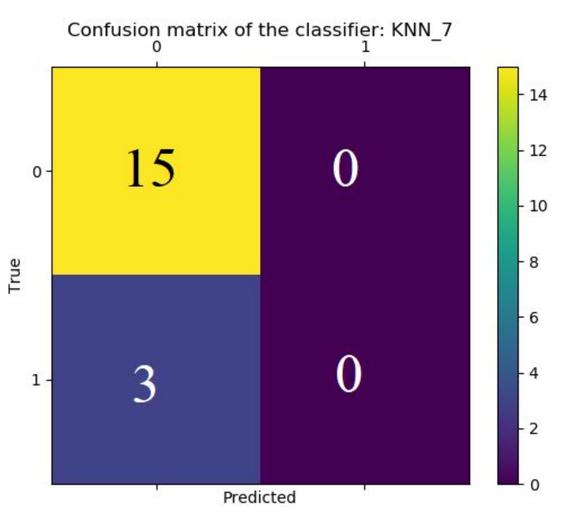
ROC Curves


Receiver operating characteristic: Support Vector Machine
1.0
0.8
0.6
0.4
0.2
0.0
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
ROC curve (area = 0.67)

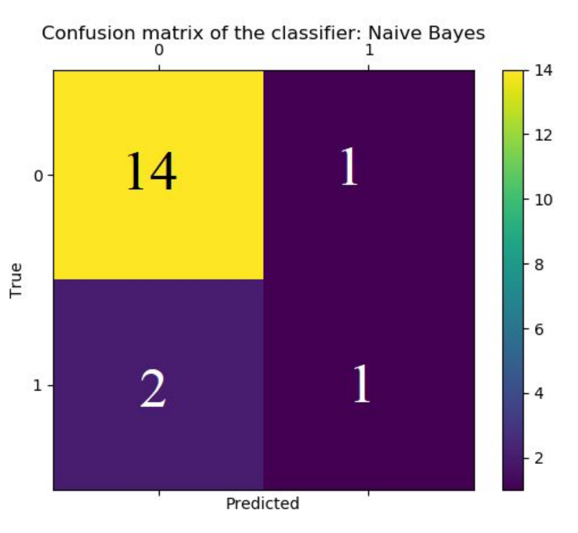
Confusion Matrices



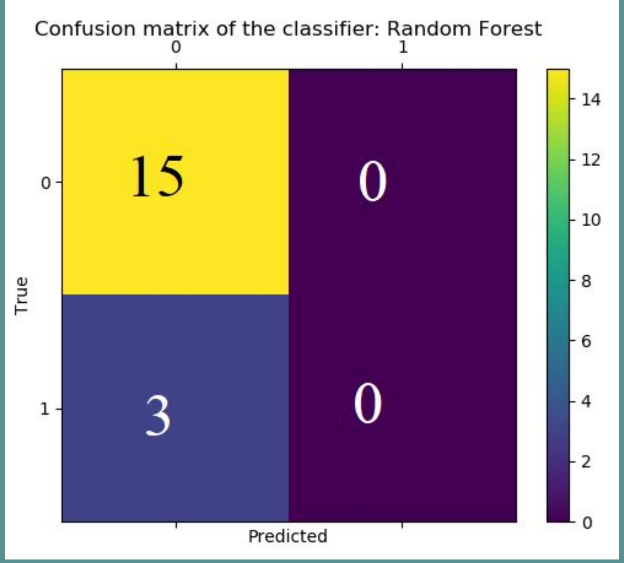
Confusion matrix of the classifier: Decision Tree



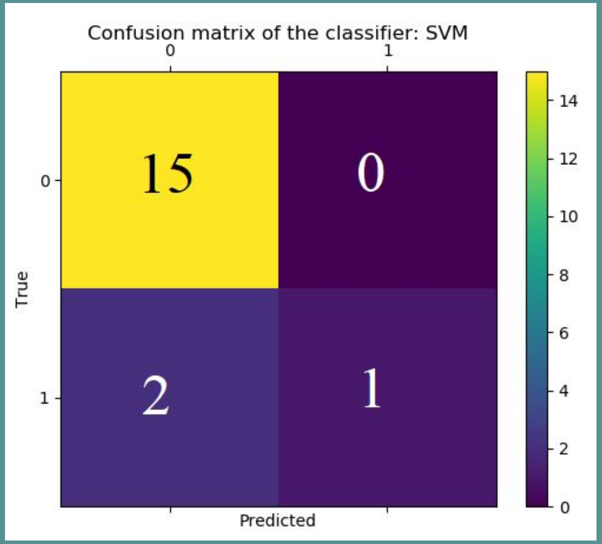
Confusion matrix of the classifier: KNN_7



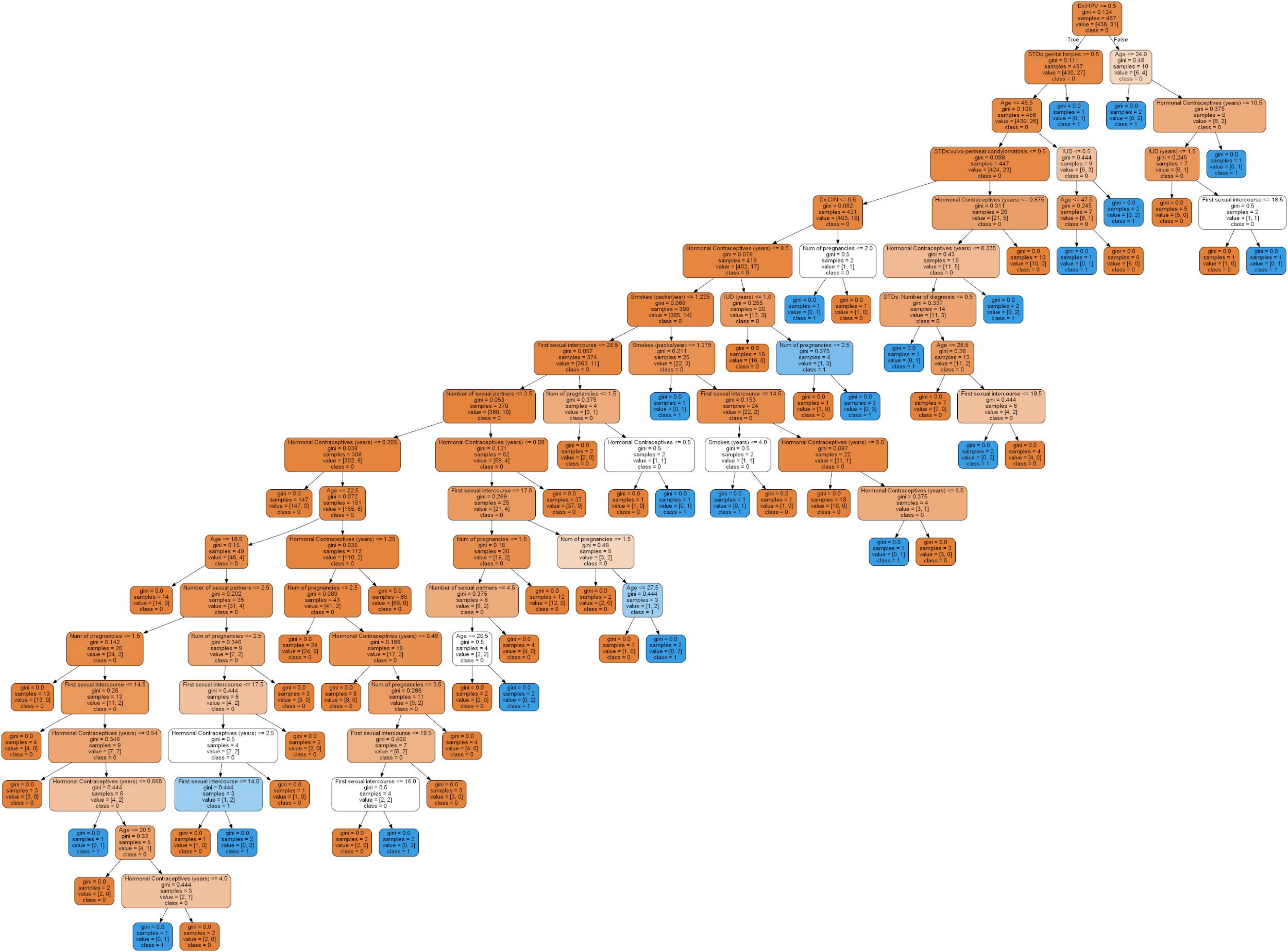
Confusion matrix of the classifier: Naive Bayes

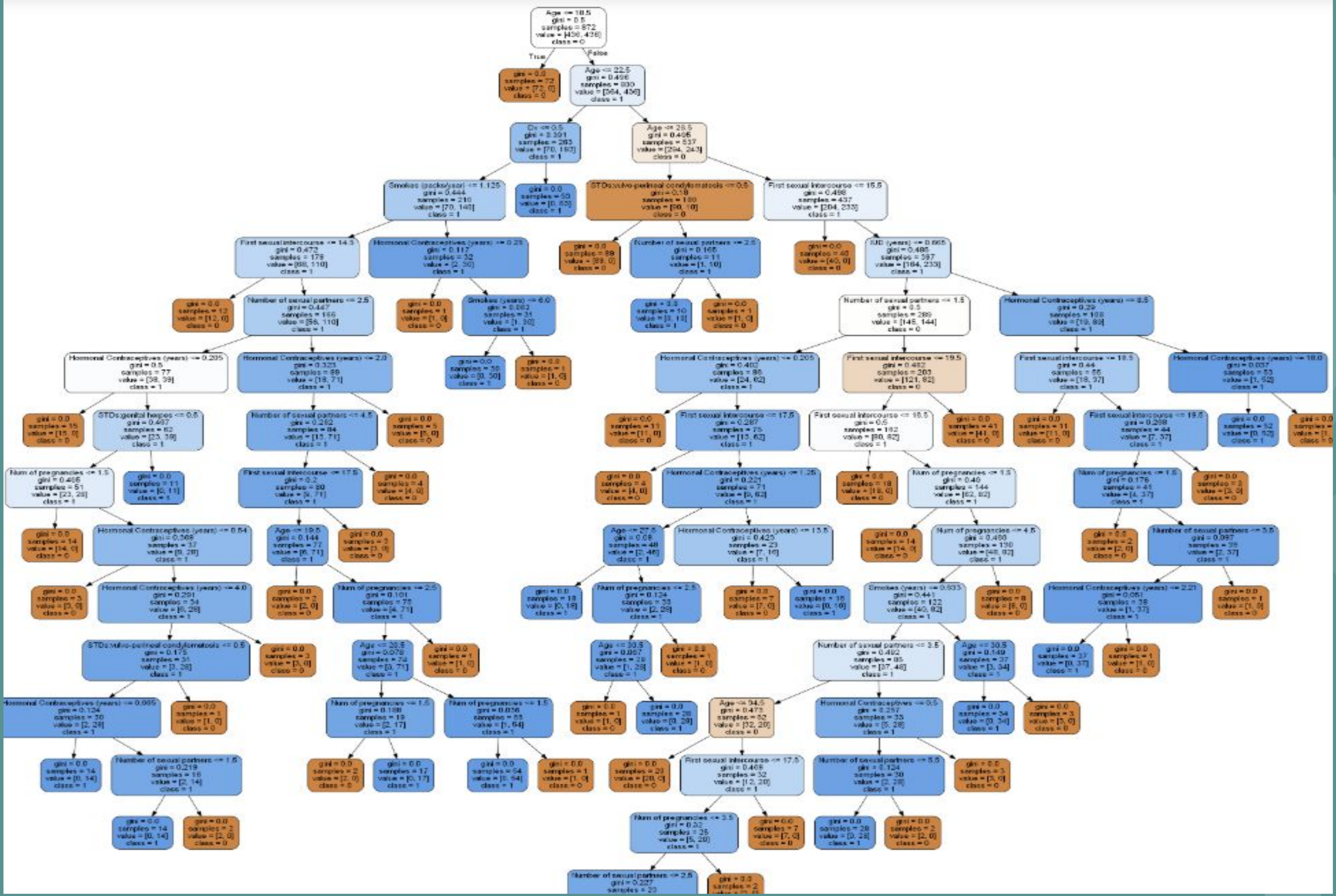


Confusion matrix of the classifier: Random Forest

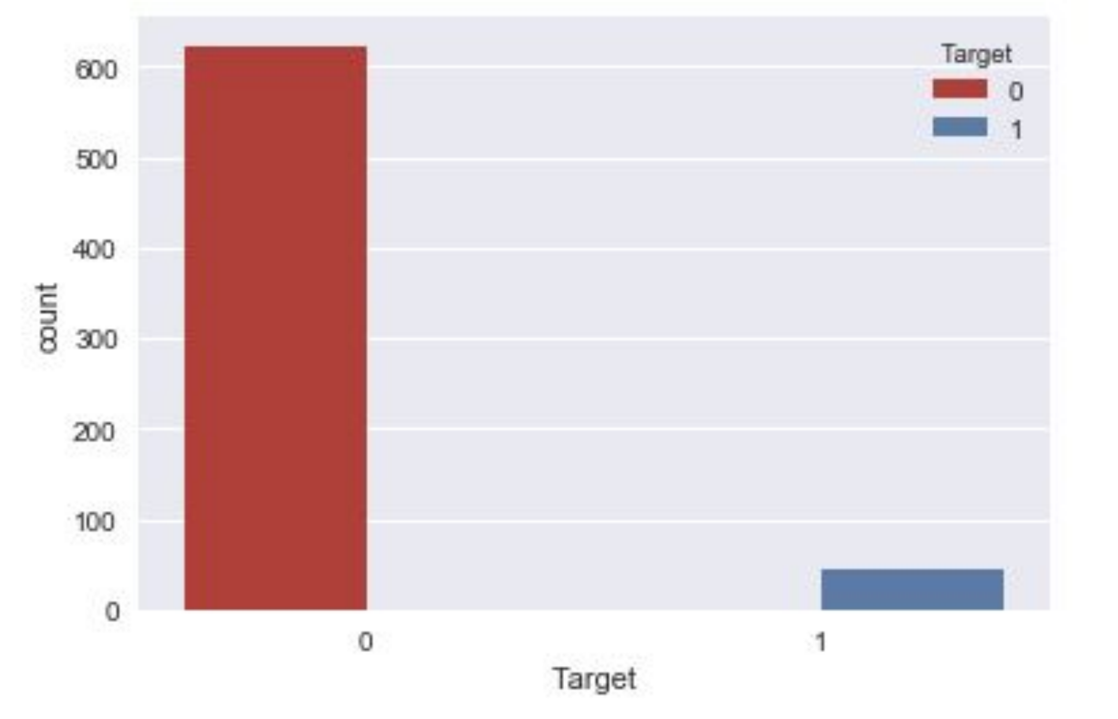


Confusion matrix of the classifier: SVM

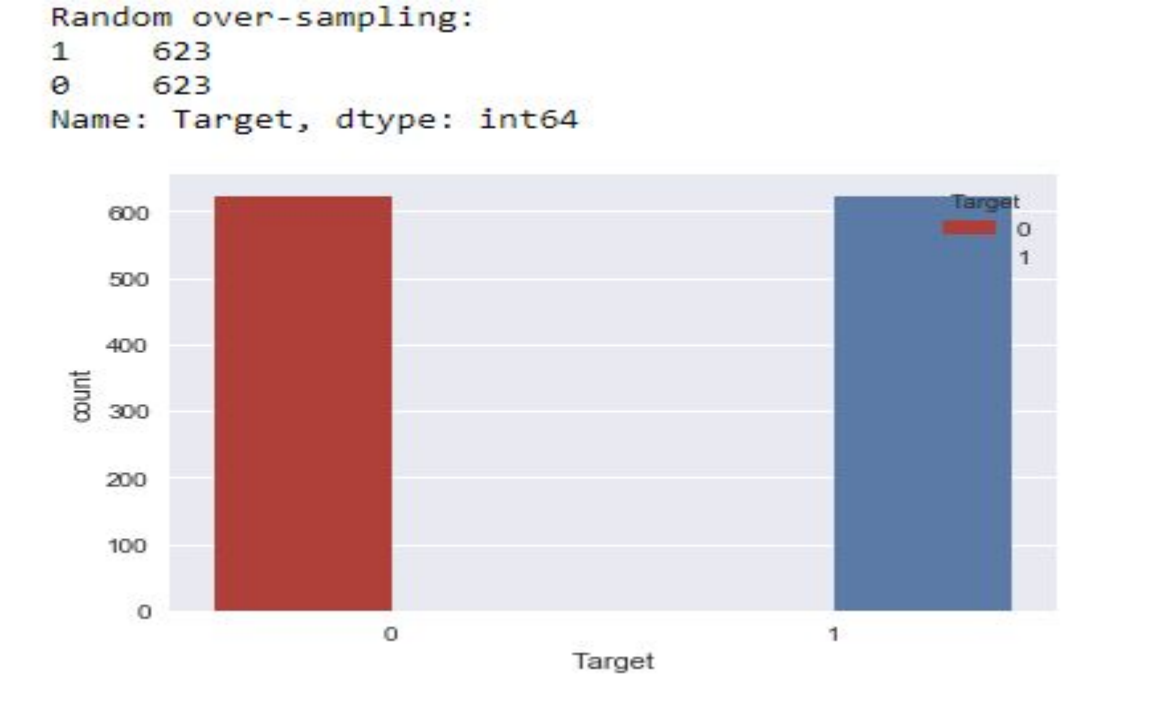

Dx:HPV <= 0.5 gini = 0.124 samples = 467 value = [436, 31] class = 0
True
False
STDs:genital herpes <= 0.5 gini = 0.111 samples = 457 value = [430, 27] class = 0
Age <= 24.0 gini = 0.48 samples = 10 value = [6, 4] class = 0
Age <= 46.5 gini = 0.108 samples = 456 value = [430, 26] class = 0
gini = 0.0 samples = 1 value = [0, 1] class = 1
gini = 0.0 samples = 2 value = [0, 2] class = 1
Hormonal Contraceptives (years) <= 10.5 gini = 0.375 samples = 8 value = [6, 2] class = 0
STDs:vulvo-perineal condylomatosis <= 0.5 gini = 0.098 samples = 447 value = [424, 23] class = 0
IUD <= 0.5 gini = 0.444 samples = 9 value = [6, 3] class = 0
IUD (years) <= 1.5 gini = 0.245 samples = 7 value = [6, 1] class = 0
gini = 0.0 samples = 1 value = [0, 1] class = 1
Dx:CIN <= 0.5 gini = 0.082 samples = 421 value = [403, 18] class = 0
Hormonal Contraceptives (years) <= 0.875 gini = 0.311 samples = 26 value = [21, 5] class = 0
Age <= 47.5 gini = 0.245 samples = 7 value = [6, 1] class = 0
gini = 0.0 samples = 2 value = [0, 2] class = 1
gini = 0.0 samples = 5 value = [5, 0] class = 0
First sexual intercourse <= 18.5 gini = 0.5 samples = 2 value = [1, 1] class = 0
Hormonal Contraceptives (years) <= 8.5 gini = 0.078 samples = 419 value = [402, 17] class = 0
Num of pregnancies <= 2.0 gini = 0.5 samples = 2 value = [1, 1] class = 0
Hormonal Contraceptives (years) <= 0.335 gini = 0.43 samples = 16 value = [11, 5] class = 0
gini = 0.0 samples = 10 value = [10, 0] class = 0
gini = 0.0 samples = 1 value = [0, 1] class = 1
gini = 0.0 samples = 6 value = [6, 0] class = 0
gini = 0.0 samples = 1 value = [1, 0] class = 0
gini = 0.0 samples = 1 value = [0, 1] class = 1
Smokes (packs/year) <= 1.225 gini = 0.068 samples = 399 value = [385, 14] class = 0
IUD (years) <= 1.0 gini = 0.255 samples = 20 value = [17, 3] class = 0
gini = 0.0 samples = 1 value = [0, 1] class = 1
gini = 0.0 samples = 1 value = [1, 0] class = 0
STDs: Number of diagnosis <= 0.5 gini = 0.337 samples = 14 value = [11, 3] class = 0
gini = 0.0 samples = 2 value = [0, 2] class = 1
First sexual intercourse <= 28.5 gini = 0.057 samples = 374 value = [363, 11] class = 0
Smokes (packs/year) <= 1.275 gini = 0.211 samples = 25 value = [22, 3] class = 0
gini = 0.0 samples = 16 value = [16, 0] class = 0
Num of pregnancies <= 2.5 gini = 0.375 samples = 4 value = [1, 3] class = 1
gini = 0.0 samples = 1 value = [0, 1] class = 1
Age <= 26.5 gini = 0.26 samples = 13 value = [11, 2] class = 0
Number of sexual partners <= 3.5 gini = 0.053 samples = 370 value = [360, 10] class = 0
Num of pregnancies <= 1.5 gini = 0.375 samples = 4 value = [3, 1] class = 0
gini = 0.0 samples = 1 value = [0, 1] class = 1
First sexual intercourse <= 14.5 gini = 0.153 samples = 24 value = [22, 2] class = 0
gini = 0.0 samples = 1 value = [1, 0] class = 0
gini = 0.0 samples = 3 value = [0, 3] class = 1
gini = 0.0 samples = 7 value = [7, 0] class = 0
First sexual intercourse <= 19.5 gini = 0.444 samples = 6 value = [4, 2] class = 0
Hormonal Contraceptives (years) <= 0.205 gini = 0.038 samples = 308 value = [302, 6] class = 0
Hormonal Contraceptives (years) <= 0.08 gini = 0.121 samples = 62 value = [58, 4] class = 0
gini = 0.0 samples = 2 value = [2, 0] class = 0
Hormonal Contraceptives <= 0.5 gini = 0.5 samples = 2 value = [1, 1] class = 0
Smokes (years) <= 4.0 gini = 0.5 samples = 2 value = [1, 1] class = 0
Hormonal Contraceptives (years) <= 5.5 gini = 0.087 samples = 22 value = [21, 1] class = 0
gini = 0.0 samples = 2 value = [0, 2] class = 1
gini = 0.0 samples = 4 value = [4, 0] class = 0
gini = 0.0 samples = 147 value = [147, 0] class = 0
Age <= 22.5 gini = 0.072 samples = 161 value = [155, 6] class = 0
First sexual intercourse <= 17.5 gini = 0.269 samples = 25 value = [21, 4] class = 0
gini = 0.0 samples = 37 value = [37, 0] class = 0
gini = 0.0 samples = 1 value = [1, 0] class = 0
gini = 0.0 samples = 1 value = [0, 1] class = 1
gini = 0.0 samples = 1 value = [0, 1] class = 1
gini = 0.0 samples = 1 value = [1, 0] class = 0
gini = 0.0 samples = 18 value = [18, 0] class = 0
Hormonal Contraceptives (years) <= 6.5 gini = 0.375 samples = 4 value = [3, 1] class = 0
Age <= 18.5 gini = 0.15 samples = 49 value = [45, 4] class = 0
Hormonal Contraceptives (years) <= 1.25 gini = 0.035 samples = 112 value = [110, 2] class = 0
Num of pregnancies <= 1.5 gini = 0.18 samples = 20 value = [18, 2] class = 0
Num of pregnancies <= 1.5 gini = 0.48 samples = 5 value = [3, 2] class = 0
gini = 0.0 samples = 1 value = [0, 1] class = 1
gini = 0.0 samples = 3 value = [3, 0] class = 0
gini = 0.0 samples = 14 value = [14, 0] class = 0
Number of sexual partners <= 2.5 gini = 0.202 samples = 35 value = [31, 4] class = 0
Num of pregnancies <= 2.5 gini = 0.089 samples = 43 value = [41, 2] class = 0
gini = 0.0 samples = 69 value = [69, 0] class = 0
Number of sexual partners <= 4.5 gini = 0.375 samples = 8 value = [6, 2] class = 0
gini = 0.0 samples = 12 value = [12, 0] class = 0
gini = 0.0 samples = 2 value = [2, 0] class = 0
Age <= 27.5 gini = 0.444 samples = 3 value = [1, 2] class = 1
Num of pregnancies <= 1.5 gini = 0.142 samples = 26 value = [24, 2] class = 0
Num of pregnancies <= 2.5 gini = 0.346 samples = 9 value = [7, 2] class = 0
gini = 0.0 samples = 24 value = [24, 0] class = 0
Hormonal Contraceptives (years) <= 0.46 gini = 0.188 samples = 19 value = [17, 2] class = 0
Age <= 20.5 gini = 0.5 samples = 4 value = [2, 2] class = 0
gini = 0.0 samples = 4 value = [4, 0] class = 0
gini = 0.0 samples = 1 value = [1, 0] class = 0
gini = 0.0 samples = 2 value = [0, 2] class = 1
gini = 0.0 samples = 13 value = [13, 0] class = 0
First sexual intercourse <= 14.5 gini = 0.26 samples = 13 value = [11, 2] class = 0
First sexual intercourse <= 17.5 gini = 0.444 samples = 6 value = [4, 2] class = 0
gini = 0.0 samples = 3 value = [3, 0] class = 0
gini = 0.0 samples = 8 value = [8, 0] class = 0
Num of pregnancies <= 3.5 gini = 0.298 samples = 11 value = [9, 2] class = 0
gini = 0.0 samples = 2 value = [2, 0] class = 0
gini = 0.0 samples = 2 value = [0, 2] class = 1
gini = 0.0 samples = 4 value = [4, 0] class = 0
Hormonal Contraceptives (years) <= 0.54 gini = 0.346 samples = 9 value = [7, 2] class = 0
Hormonal Contraceptives (years) <= 2.5 gini = 0.5 samples = 4 value = [2, 2] class = 0
gini = 0.0 samples = 2 value = [2, 0] class = 0
First sexual intercourse <= 18.5 gini = 0.408 samples = 7 value = [5, 2] class = 0
gini = 0.0 samples = 4 value = [4, 0] class = 0
gini = 0.0 samples = 3 value = [3, 0] class = 0
Hormonal Contraceptives (years) <= 0.665 gini = 0.444 samples = 6 value = [4, 2] class = 0
First sexual intercourse <= 14.0 gini = 0.444 samples = 3 value = [1, 2] class = 1
gini = 0.0 samples = 1 value = [1, 0] class = 0
First sexual intercourse <= 16.0 gini = 0.5 samples = 4 value = [2, 2] class = 0
gini = 0.0 samples = 3 value = [3, 0] class = 0
gini = 0.0 samples = 1 value = [0, 1] class = 1
Age <= 20.5 gini = 0.32 samples = 5 value = [4, 1] class = 0
gini = 0.0 samples = 1 value = [1, 0] class = 0
gini = 0.0 samples = 2 value = [0, 2] class = 1
gini = 0.0 samples = 2 value = [2, 0] class = 0
gini = 0.0 samples = 2 value = [0, 2] class = 1
gini = 0.0 samples = 2 value = [2, 0] class = 0
Hormonal Contraceptives (years) <= 4.0 gini = 0.444 samples = 3 value = [2, 1] class = 0
gini = 0.0 samples = 1 value = [0, 1] class = 1
gini = 0.0 samples = 2 value = [2, 0] class = 0


Age <= 18.5 gini = 0.5 samples = 872 value = [436, 436] class = 0
True
False
gini = 0.0 samples = 72 value = [72, 0] class = 0
Age <= 22.5 gini = 0.496 samples = 800 value = [364, 436] class = 1
Dx <= 0.5 gini = 0.391 samples = 263 value = [70, 193] class = 1
Age <= 28.5 gini = 0.495 samples = 537 value = [294, 243] class = 0
Smokes (packs/year) <= 1.125 gini = 0.444 samples = 210 value = [70, 140] class = 1
gini = 0.0 samples = 53 value = [0, 53] class = 1
STDs:vulvo-perineal condylomatosis <= 0.5 gini = 0.18 samples = 100 value = [90, 10] class = 0
First sexual intercourse <= 15.5 gini = 0.498 samples = 437 value = [204, 233] class = 1
First sexual intercourse <= 14.5 gini = 0.472 samples = 178 value = [68, 110] class = 1
Hormonal Contraceptives (years) <= 0.25 gini = 0.117 samples = 32 value = [2, 30] class = 1
gini = 0.0 samples = 89 value = [89, 0] class = 0
Number of sexual partners <= 2.5 gini = 0.165 samples = 11 value = [1, 10] class = 1
gini = 0.0 samples = 40 value = [40, 0] class = 0
IUD (years) <= 0.665 gini = 0.485 samples = 397 value = [164, 233] class = 1
gini = 0.0 samples = 12 value = [12, 0] class = 0
Number of sexual partners <= 2.5 gini = 0.447 samples = 166 value = [56, 110] class = 1
gini = 0.0 samples = 1 value = [1, 0] class = 0
Smokes (years) <= 6.0 gini = 0.062 samples = 31 value = [1, 30] class = 1
gini = 0.0 samples = 10 value = [0, 10] class = 1
gini = 0.0 samples = 1 value = [1, 0] class = 0
Number of sexual partners <= 1.5 gini = 0.5 samples = 289 value = [145, 144] class = 0
Hormonal Contraceptives (years) <= 8.5 gini = 0.29 samples = 108 value = [19, 89] class = 1
Hormonal Contraceptives (years) <= 0.205 gini = 0.5 samples = 77 value = [38, 39] class = 1
Hormonal Contraceptives (years) <= 2.0 gini = 0.323 samples = 89 value = [18, 71] class = 1
gini = 0.0 samples = 30 value = [0, 30] class = 1
gini = 0.0 samples = 1 value = [1, 0] class = 0
Hormonal Contraceptives (years) <= 0.205 gini = 0.402 samples = 86 value = [24, 62] class = 1
First sexual intercourse <= 19.5 gini = 0.482 samples = 203 value = [121, 82] class = 0
First sexual intercourse <= 18.5 gini = 0.44 samples = 55 value = [18, 37] class = 1
Hormonal Contraceptives (years) <= 18.0 gini = 0.037 samples = 53 value = [1, 52] class = 1
gini = 0.0 samples = 15 value = [15, 0] class = 0
STDs:genital herpes <= 0.5 gini = 0.467 samples = 62 value = [23, 39] class = 1
Number of sexual partners <= 4.5 gini = 0.262 samples = 84 value = [13, 71] class = 1
gini = 0.0 samples = 5 value = [5, 0] class = 0
gini = 0.0 samples = 11 value = [11, 0] class = 0
First sexual intercourse <= 17.5 gini = 0.287 samples = 75 value = [13, 62] class = 1
First sexual intercourse <= 16.5 gini = 0.5 samples = 162 value = [80, 82] class = 1
gini = 0.0 samples = 41 value = [41, 0] class = 0
gini = 0.0 samples = 11 value = [11, 0] class = 0
First sexual intercourse <= 19.5 gini = 0.268 samples = 44 value = [7, 37] class = 1
gini = 0.0 samples = 52 value = [0, 52] class = 1
gini = 0.0 samples = 1 value = [1, class = 0
Num of pregnancies <= 1.5 gini = 0.495 samples = 51 value = [23, 28] class = 1
gini = 0.0 samples = 11 value = [0, 11] class = 1
First sexual intercourse <= 17.5 gini = 0.2 samples = 80 value = [9, 71] class = 1
gini = 0.0 samples = 4 value = [4, 0] class = 0
gini = 0.0 samples = 4 value = [4, 0] class = 0
Hormonal Contraceptives (years) <= 1.25 gini = 0.221 samples = 71 value = [9, 62] class = 1
gini = 0.0 samples = 18 value = [18, 0] class = 0
Num of pregnancies <= 1.5 gini = 0.49 samples = 144 value = [62, 82] class = 1
Num of pregnancies <= 1.5 gini = 0.176 samples = 41 value = [4, 37] class = 1
gini = 0.0 samples = 3 value = [3, 0] class = 0
gini = 0.0 samples = 14 value = [14, 0] class = 0
Hormonal Contraceptives (years) <= 0.54 gini = 0.369 samples = 37 value = [9, 28] class = 1
Age <= 19.5 gini = 0.144 samples = 77 value = [6, 71] class = 1
gini = 0.0 samples = 3 value = [3, 0] class = 0
Age <= 27.5 gini = 0.08 samples = 48 value = [2, 46] class = 1
Hormonal Contraceptives (years) <= 13.5 gini = 0.425 samples = 23 value = [7, 16] class = 1
gini = 0.0 samples = 14 value = [14, 0] class = 0
Num of pregnancies <= 4.5 gini = 0.466 samples = 130 value = [48, 82] class = 1
gini = 0.0 samples = 2 value = [2, 0] class = 0
Number of sexual partners <= 3.5 gini = 0.097 samples = 39 value = [2, 37] class = 1
gini = 0.0 samples = 3 value = [3, 0] class = 0
Hormonal Contraceptives (years) <= 4.0 gini = 0.291 samples = 34 value = [6, 28] class = 1
gini = 0.0 samples = 2 value = [2, 0] class = 0
Num of pregnancies <= 2.5 gini = 0.101 samples = 75 value = [4, 71] class = 1
gini = 0.0 samples = 18 value = [0, 18] class = 1
Num of pregnancies <= 2.5 gini = 0.124 samples = 30 value = [2, 28] class = 1
gini = 0.0 samples = 7 value = [7, 0] class = 0
gini = 0.0 samples = 16 value = [0, 16] class = 1
Smokes (years) <= 0.833 gini = 0.441 samples = 122 value = [40, 82] class = 1
gini = 0.0 samples = 8 value = [8, 0] class = 0
Hormonal Contraceptives (years) <= 2.21 gini = 0.051 samples = 38 value = [1, 37] class = 1
gini = 0.0 samples = 1 value = [1, 0] class = 0
STDs:vulvo-perineal condylomatosis <= 0.5 gini = 0.175 samples = 31 value = [3, 28] class = 1
gini = 0.0 samples = 3 value = [3, 0] class = 0
Age <= 20.5 gini = 0.078 samples = 74 value = [3, 71] class = 1
gini = 0.0 samples = 1 value = [1, 0] class = 0
Age <= 30.5 gini = 0.067 samples = 29 value = [1, 28] class = 1
gini = 0.0 samples = 1 value = [1, 0] class = 0
Number of sexual partners <= 3.5 gini = 0.492 samples = 85 value = [37, 48] class = 1
Age <= 30.5 gini = 0.149 samples = 37 value = [3, 34] class = 1
gini = 0.0 samples = 37 value = [0, 37] class = 1
gini = 0.0 samples = 1 value = [1, 0] class = 0
Hormonal Contraceptives (years) <= 0.085 gini = 0.124 samples = 30 value = [2, 28] class = 1
gini = 0.0 samples = 1 value = [1, 0] class = 0
Num of pregnancies <= 1.5 gini = 0.188 samples = 19 value = [2, 17] class = 1
Num of pregnancies <= 1.5 gini = 0.036 samples = 55 value = [1, 54] class = 1
gini = 0.0 samples = 1 value = [1, 0] class = 0
gini = 0.0 samples = 28 value = [0, 28] class = 1
Age <= 34.5 gini = 0.473 samples = 52 value = [32, 20] class = 0
Hormonal Contraceptives <= 0.5 gini = 0.257 samples = 33 value = [5, 28] class = 1
gini = 0.0 samples = 34 value = [0, 34] class = 1
gini = 0.0 samples = 3 value = [3, 0] class = 0
gini = 0.0 samples = 14 value = [0, 14] class = 1
Number of sexual partners <= 1.5 gini = 0.219 samples = 16 value = [2, 14] class = 1
gini = 0.0 samples = 2 value = [2, 0] class = 0
gini = 0.0 samples = 17 value = [0, 17] class = 1
gini = 0.0 samples = 54 value = [0, 54] class = 1
gini = 0.0 samples = 1 value = [1, 0] class = 0
gini = 0.0 samples = 20 value = [20, 0] class = 0
First sexual intercourse <= 17.5 gini = 0.469 samples = 32 value = [12, 20] class = 1
Number of sexual partners <= 5.5 gini = 0.124 samples = 30 value = [2, 28] class = 1
gini = 0.0 samples = 3 value = [3, 0] class = 0
gini = 0.0 samples = 14 value = [0, 14] class = 1
gini = 0.0 samples = 2 value = [2, 0] class = 0
Num of pregnancies <= 3.5 gini = 0.32 samples = 25 value = [5, 20] class = 1
gini = 0.0 samples = 7 value = [7, 0] class = 0
gini = 0.0 samples = 28 value = [0, 28] class = 1
gini = 0.0 samples = 2 value = [2, 0] class = 0
Number of sexual partners <= 2.5 gini = 0.227 samples = 23
gini = 0.0 samples = 2

Imbalanced vs. Balanced Class Plots



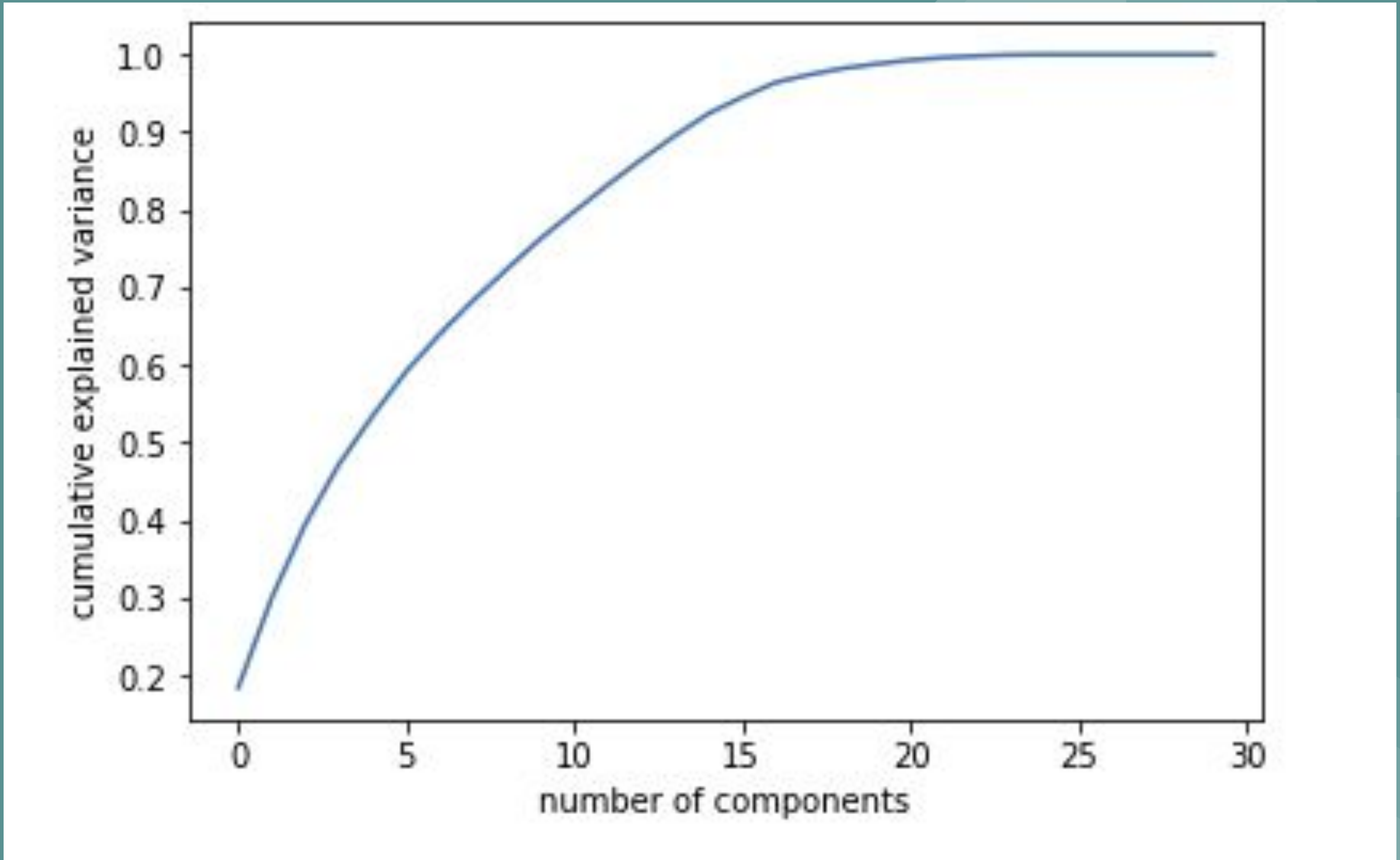
Imbalanced Class Plot



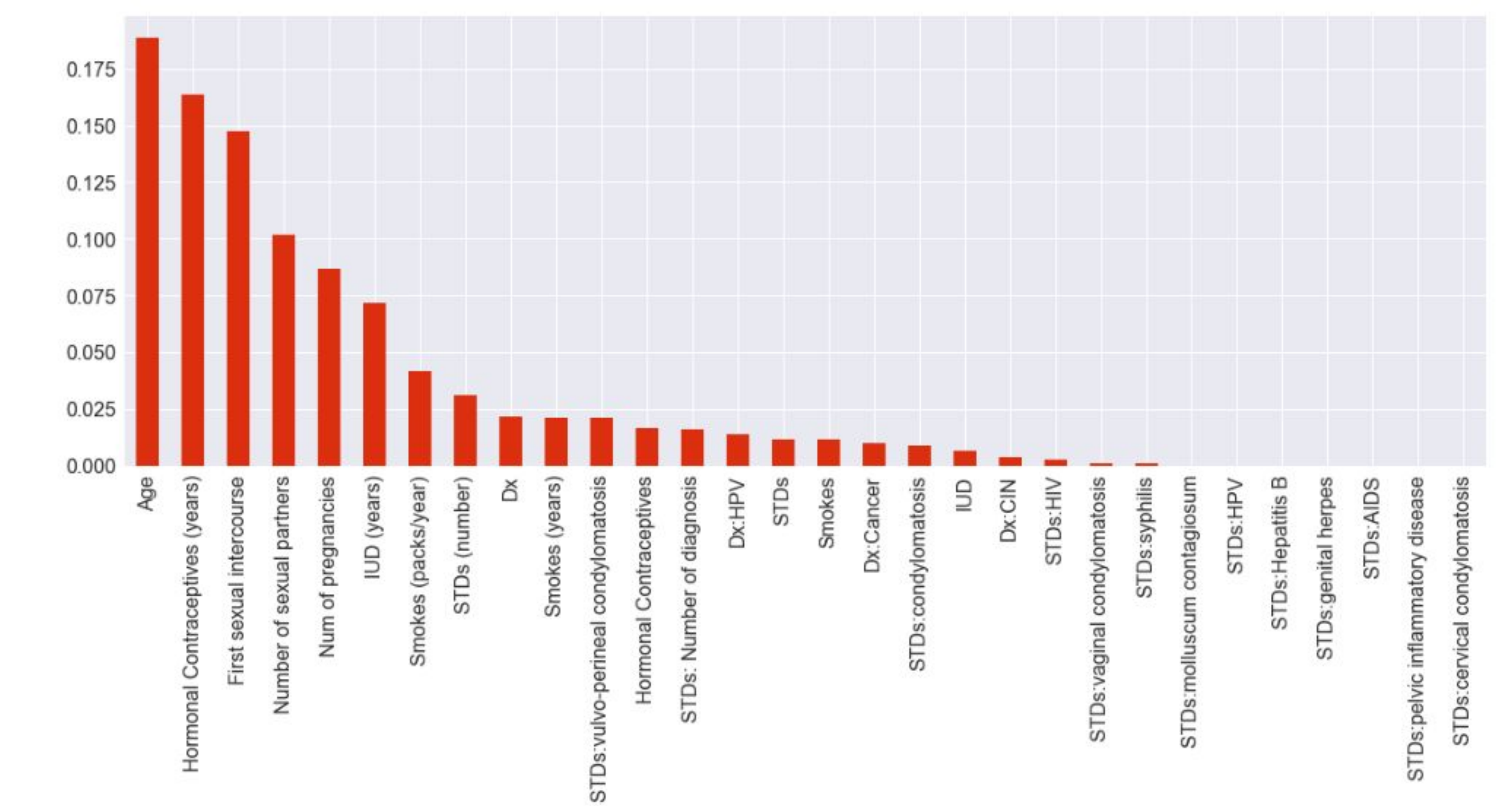
```
Random over-sampling:
1     623
0     623
Name: Target, dtype: int64
```

Balanced Class Plot

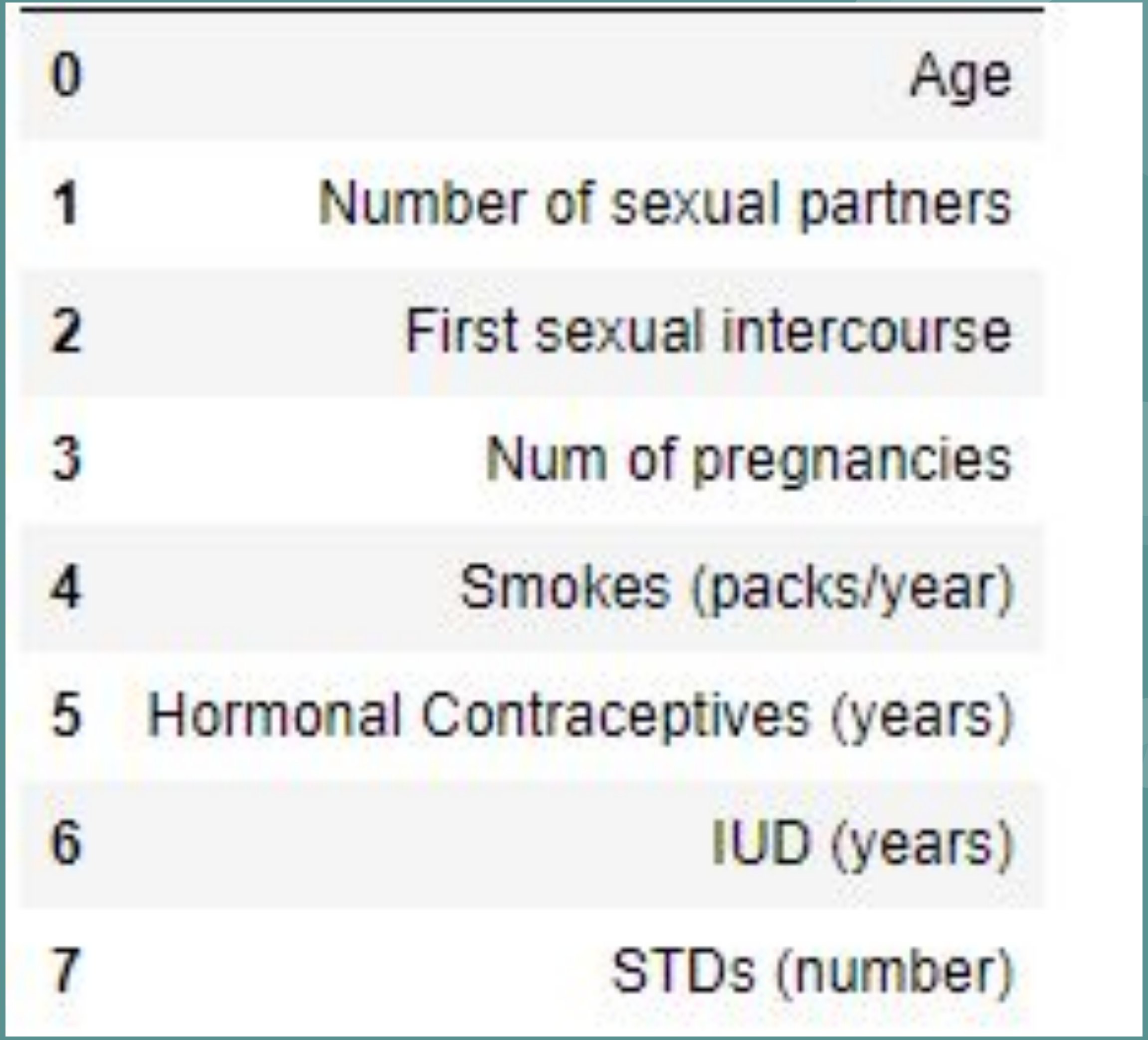
PCA Plot


1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
cumulative explained variance
0
5
10
15
20
25
30
number of components

Feature Selection

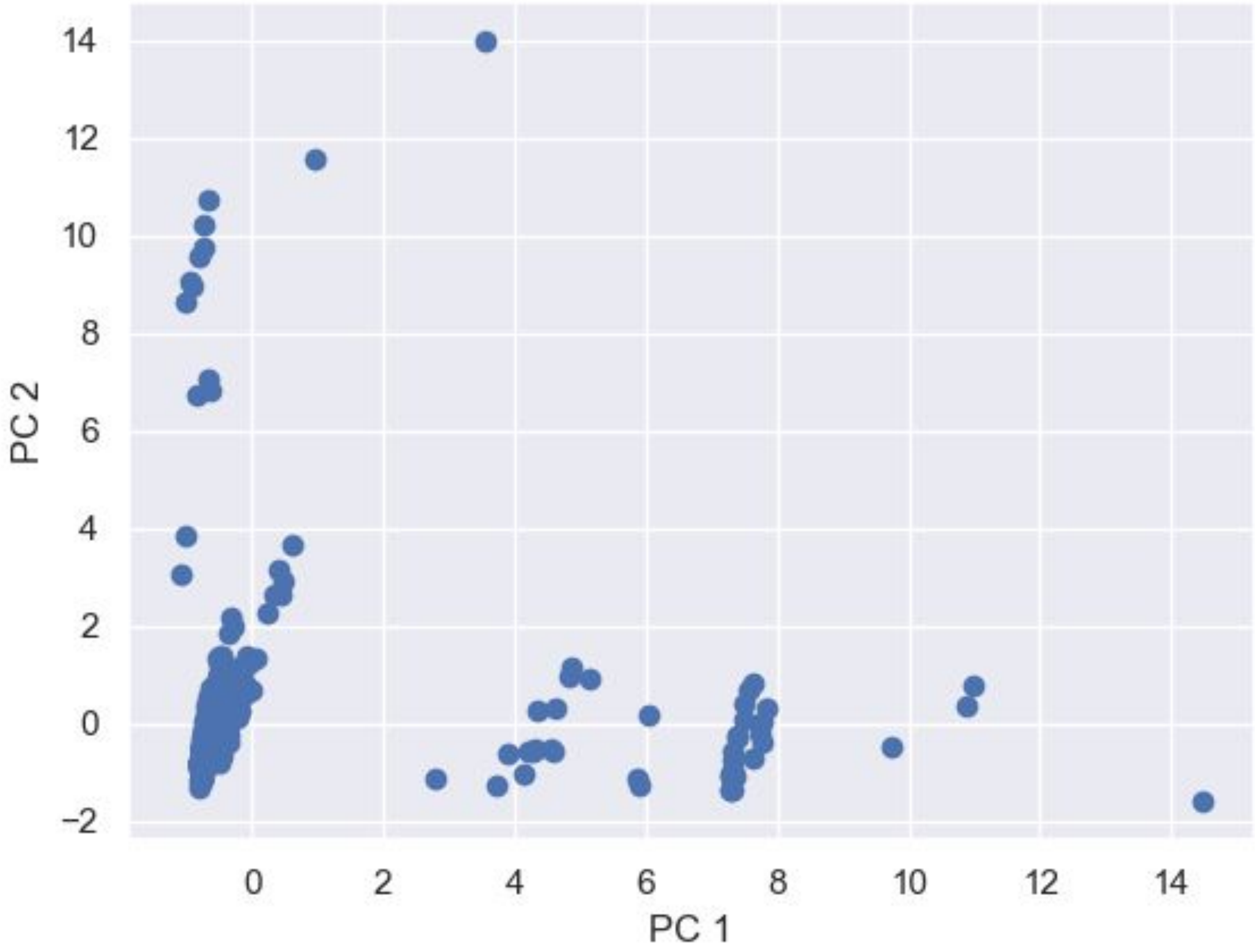


7 Highest Predictive Powers

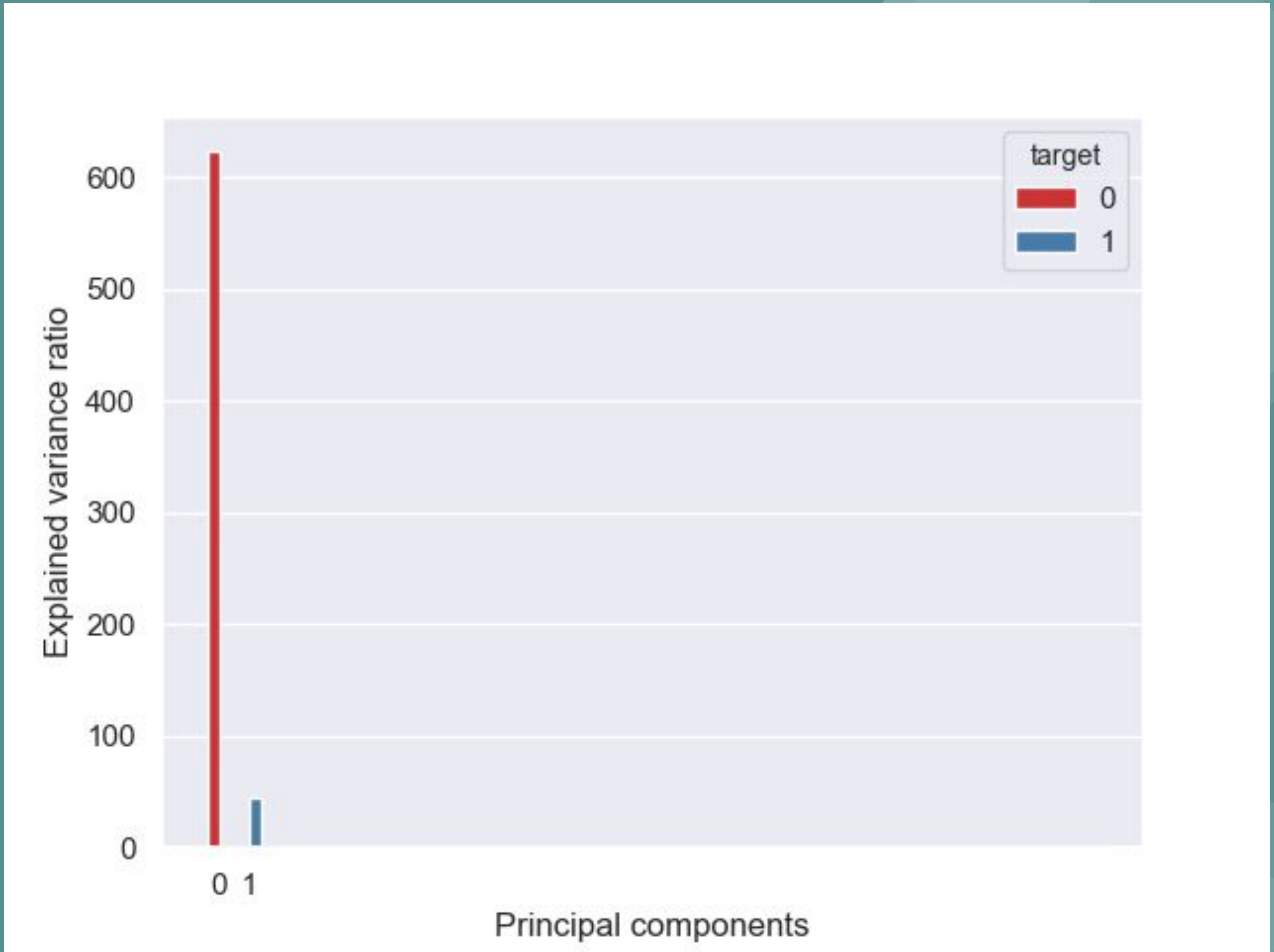


0	Age
1	Number of sexual partners
2	First sexual intercourse
3	Num of pregnancies
4	Smokes (packs/year)
5	Hormonal Contraceptives (years)
6	IUD (years)
7	STDs (number)

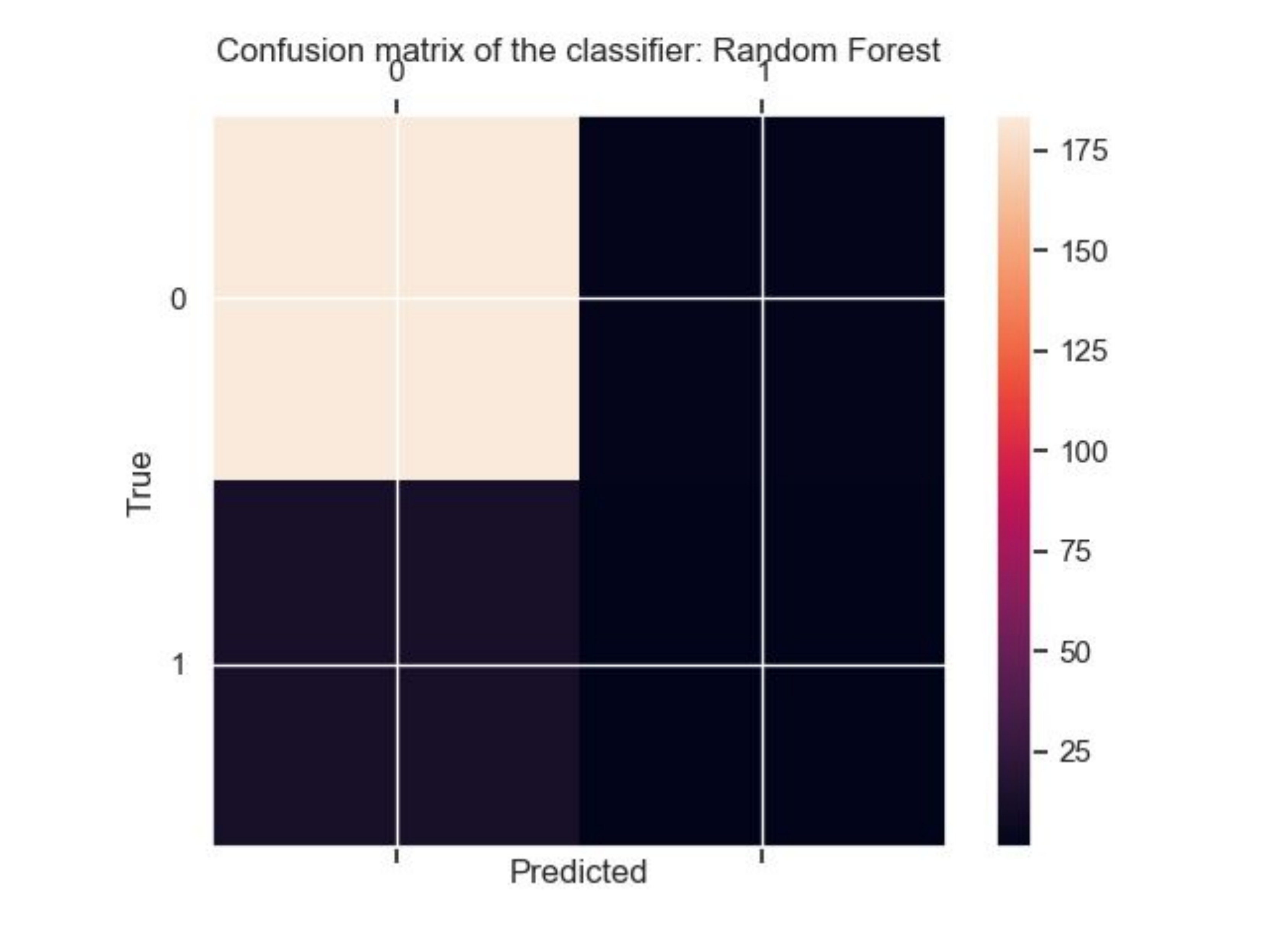
PC 1 vs. PC 2

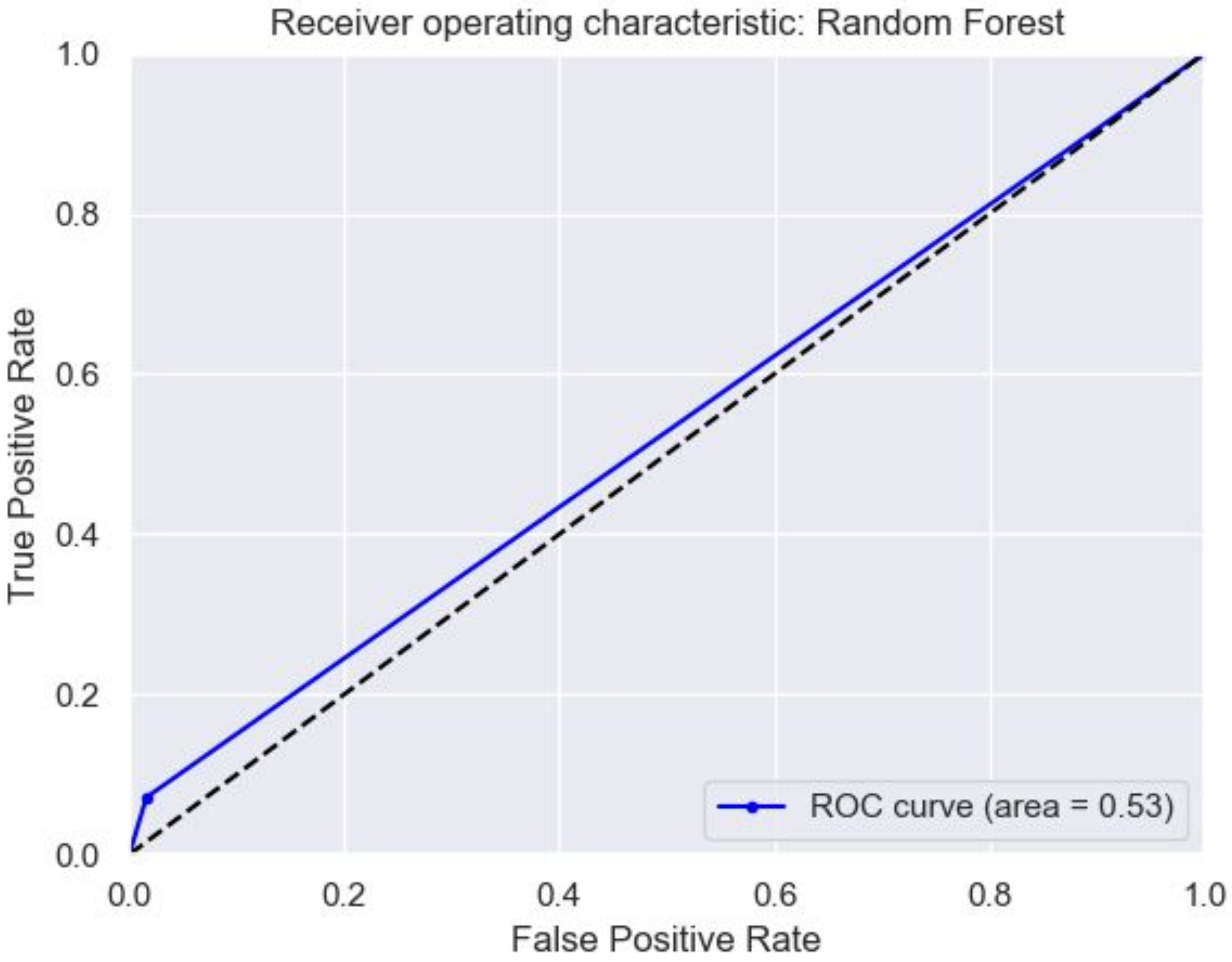


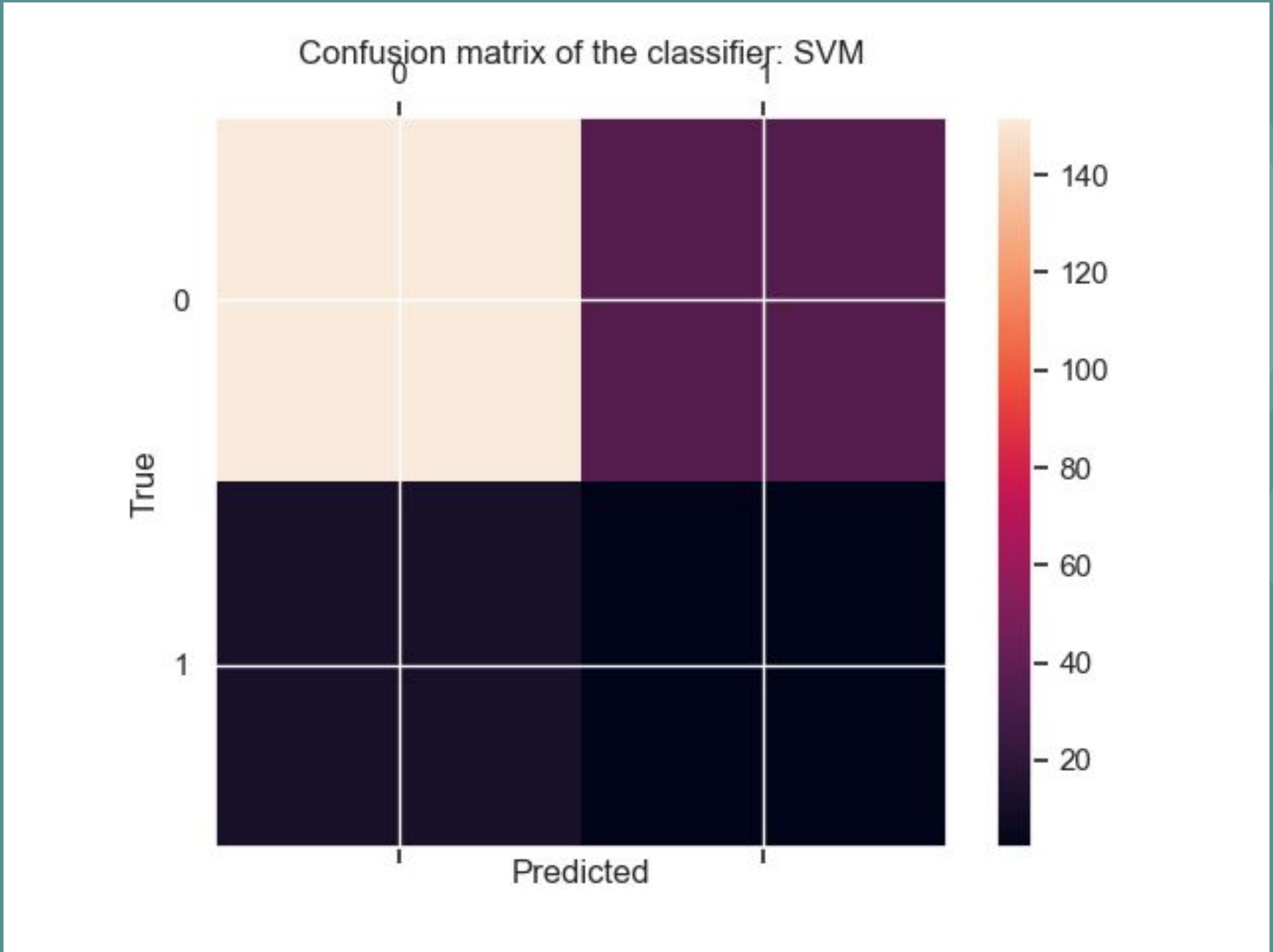
Variance Ratio vs. Principal Components

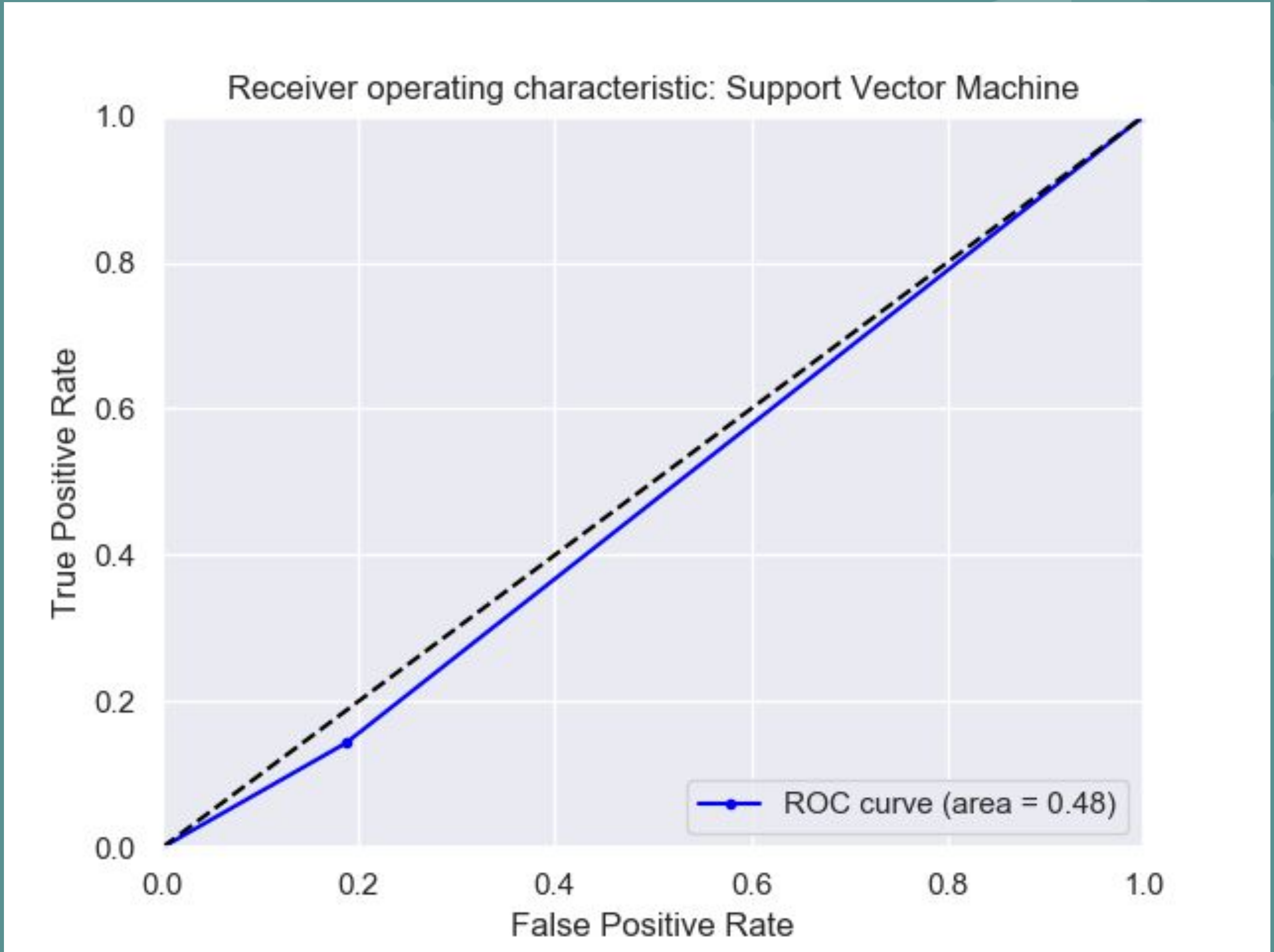


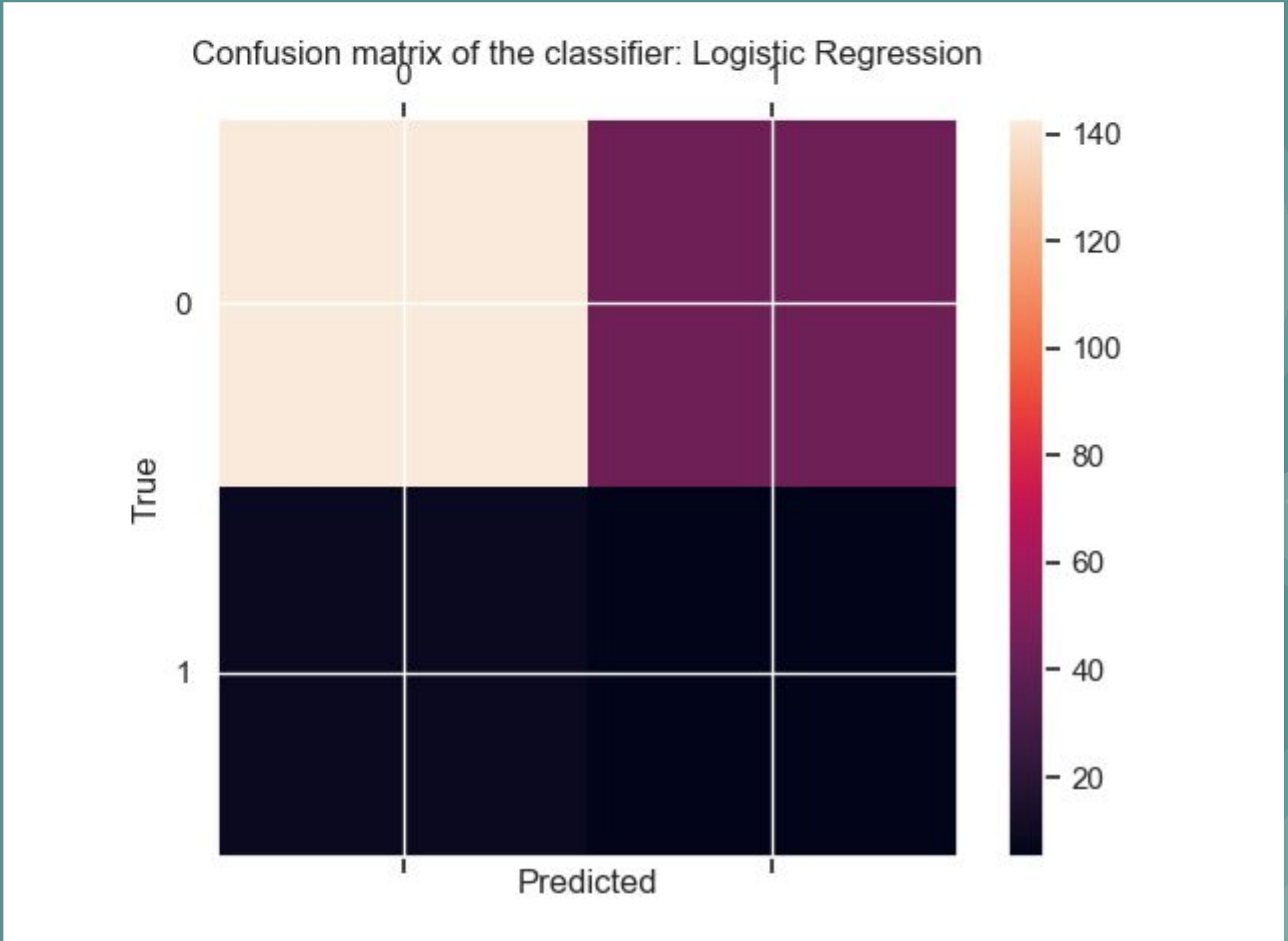
After the Preprocessing of the Data



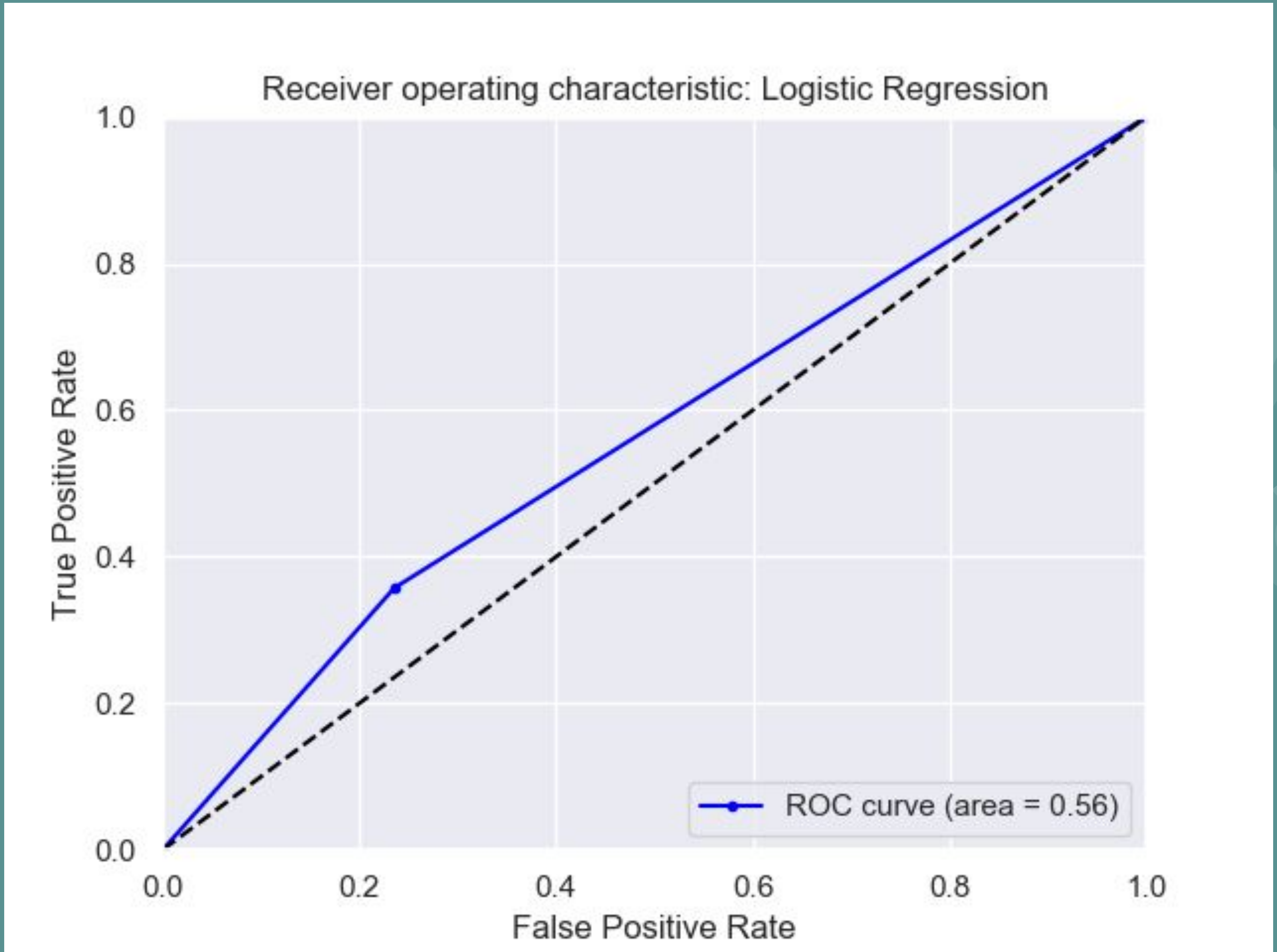

Receiver operating characteristic: Random Forest
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.53)

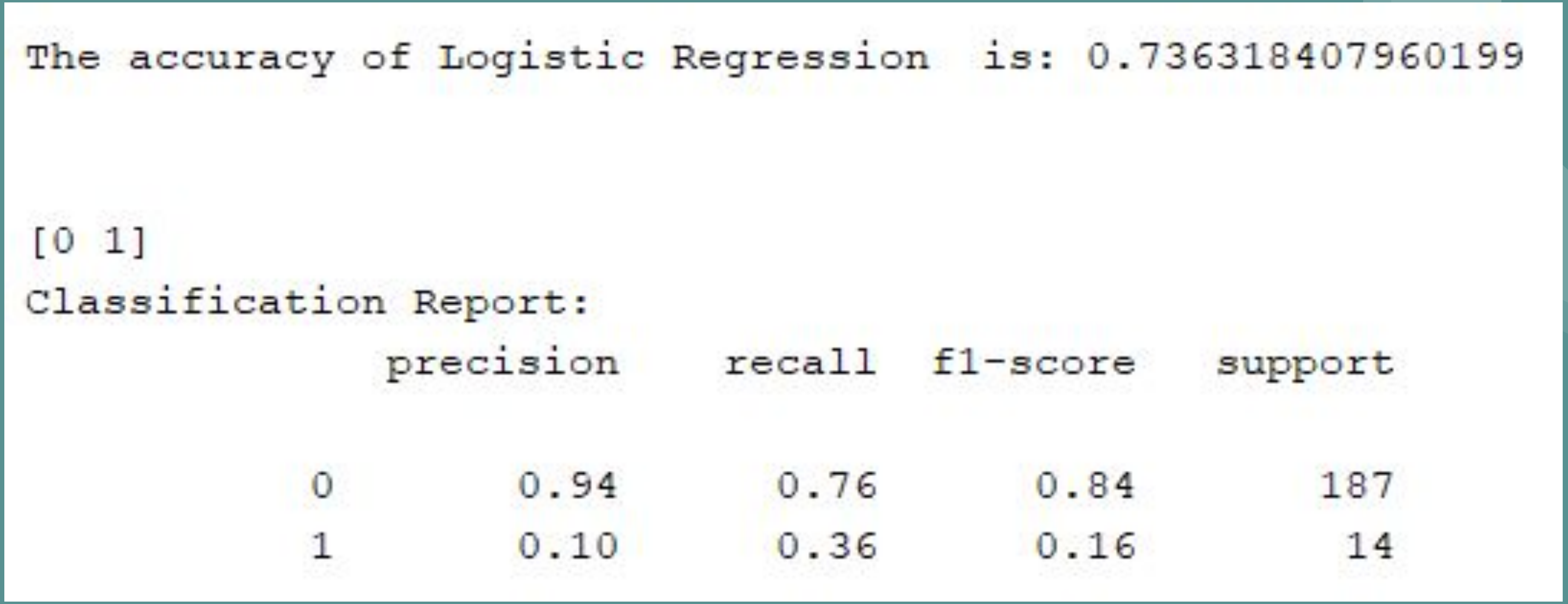

Confusion matrix of the classifier: SVM
0
1
0
1
True
Predicted
140
120
100
80
60
40
20


Receiver operating characteristic: Support Vector Machine
1.0
0.8
0.6
0.4
0.2
0.0
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
ROC curve (area = 0.48)



Confusion matrix of the classifier: Logistic Regression
0
1
0
1
True
Predicted
140
120
100
80
60
40
20


Receiver operating characteristic: Logistic Regression
1.0
0.8
0.6
0.4
0.2
0.0
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
ROC curve (area = 0.56)



```
The accuracy of Logistic Regression  is: 0.736318407960199



[0 1]
Classification Report:
              precision    recall  f1-score   support

           0       0.94      0.76      0.84       187
           1       0.10      0.36      0.16        14
```

Conclusion

- The 10 ten features with the highest predictive power were confirmed by the literature
- Preprocessing had a impact on the performance of our model
- The model could be run on other targets identified in the data
- Other classifiers could be run on this data