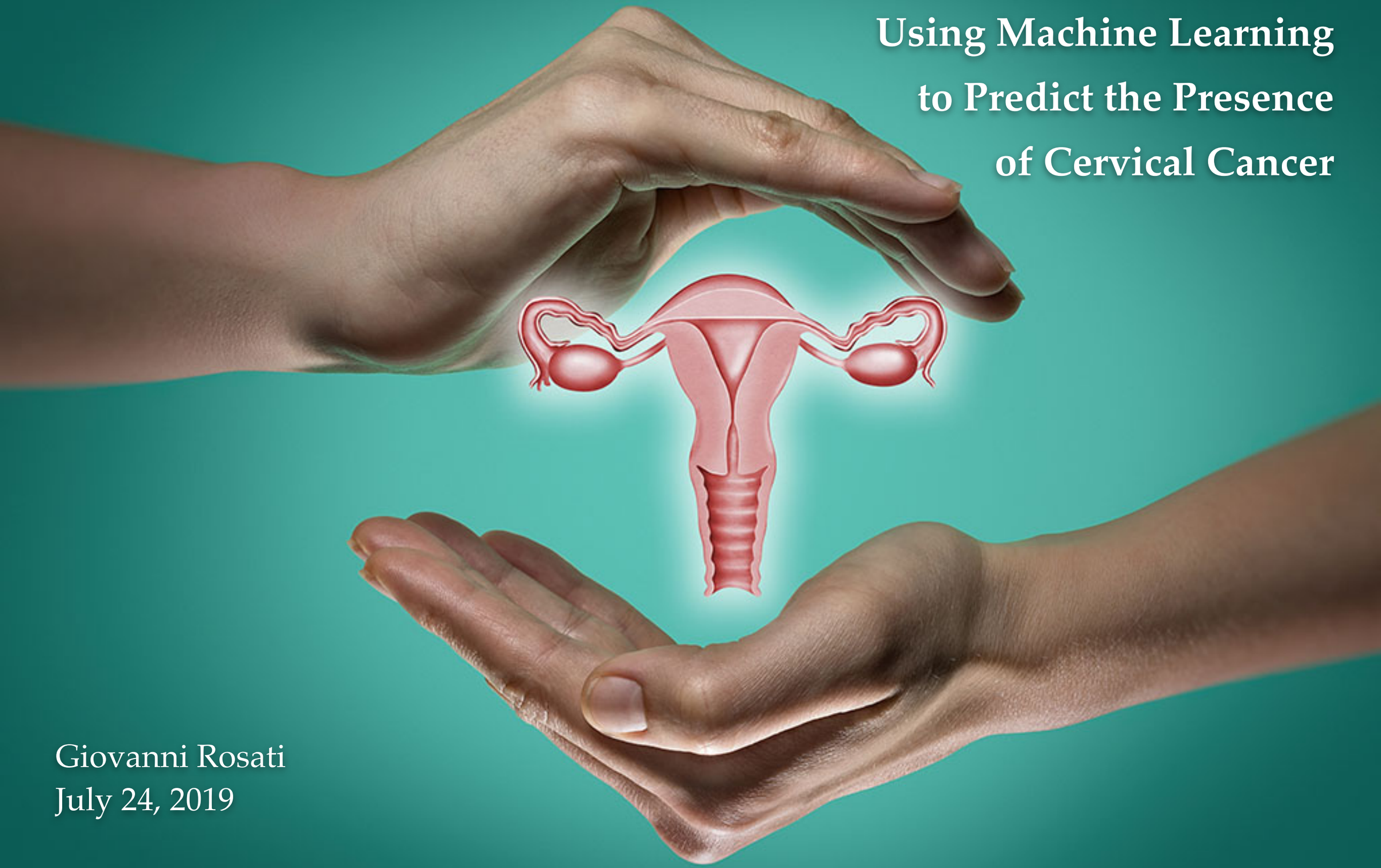
Using Machine Learning to Predict the Presence of Cervical Cancer



Giovanni Rosati
July 24, 2019

Overview

- The Data:
  - 858 women seen at a hospital in Caracas, Venezuela.
  - 35 variables that include demographic information, habits, and historic medical records.
- The Goal:
  - Predict whether or not a patient has cervical cancer, as confirmed by a biopsy.
  - “False negatives” could result in death…
  - “False positives” result in unnecessary medical procedures…
- The Challenges:
  - Many missing values
  - An unbalanced “target” variable (small proportion of actual cancer cases).

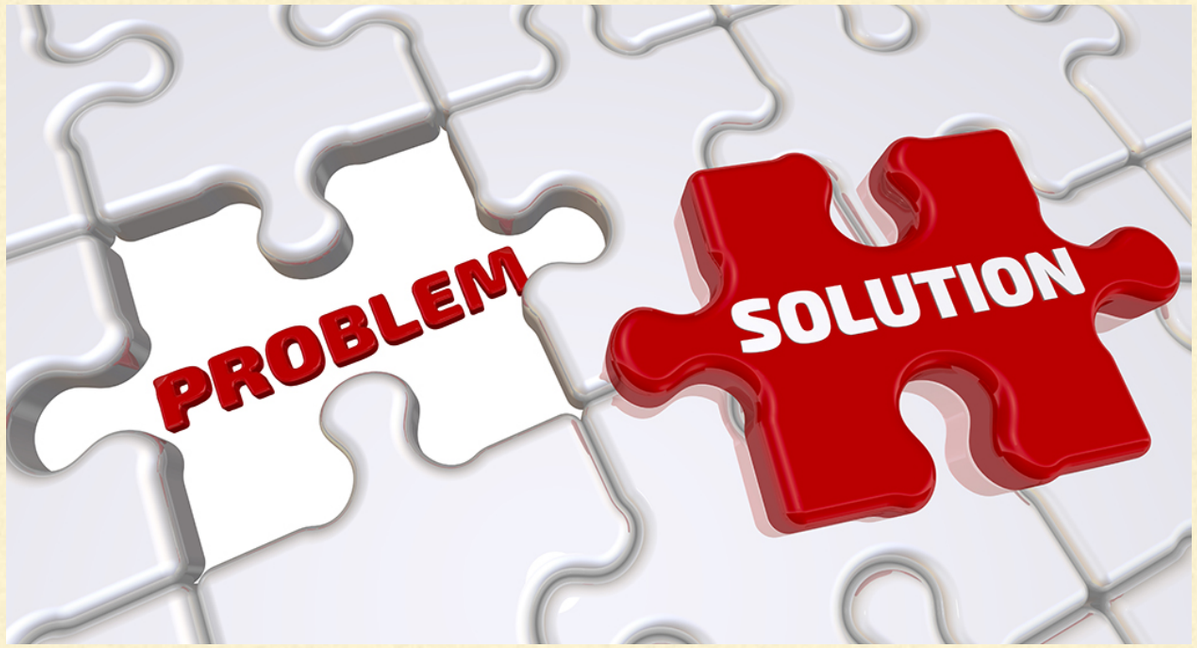
The Data

- Demographics
  - Age
- Personal History
  - # of sexual partners
  - Age at first intercourse
  - Tobacco usage
  - Hormonal Contraceptives
  - IUD
- Medical History
  - STD's
  - Dx: Cancer
  - Dx: CIN (Cervical Intraepithelial Neoplasia)
  - Dx: HPV (Human papillomavirus)
- Cervical Cancer Tests
  - Hinselman
  - Schiller
  - Citology

	Data Type	Nulls	% missing	Low Value	Hi Value	Notes
Age	int64	0	0.00%	13	84	OK
Number of sexual partners	float64	26	3.03%	1	28	could be int
First sexual intercourse	float64	7	0.82%	10	32	
Num of pregnancies	float64	56	6.53%	0	11	could be int
Smokes	float64	13	1.52%	0	1	boolean
Smokes (years)	float64	13	1.52%	0	37	
Smokes (packs/year)	float64	13	1.52%	0	37	
Hormonal Contraceptives	float64	108	12.59%	0	1	boolean
Hormonal Contraceptives (years)	float64	108	12.59%	0	30	
IUD	float64	117	13.64%	0	1	boolean
IUD (years)	float64	117	13.64%	0	19	
STDs	float64	105	12.24%	0	1	boolean
STDs (number)	float64	105	12.24%	0	4	could be int
STDs:condylomatosis	float64	105	12.24%	0	1	boolean
STDs:cervical condylomatosis	float64	105	12.24%	0	1	boolean
STDs:vaginal condylomatosis	float64	105	12.24%	0	1	boolean
STDs:vulvo-perineal condylomatosis	float64	105	12.24%	0	1	boolean
STDs:syphilis	float64	105	12.24%	0	1	boolean
STDs:pelvic inflammatory disease	float64	105	12.24%	0	1	boolean
STDs:genital herpes	float64	105	12.24%	0	1	boolean
STDs:molluscum contagiosum	float64	105	12.24%	0	1	boolean
STDs:AIDS	float64	105	12.24%	0	1	boolean
STDs:HIV	float64	105	12.24%	0	1	boolean
STDs:Hepatitis B	float64	105	12.24%	0	1	boolean
STDs:HPV	float64	105	12.24%	0	1	boolean
STDs: Number of diagnosis	int64	0	0.00%	0	3	
STDs: Time since first diagnosis	float64	787	91.72%	1	22	very high % missing
STDs: Time since last diagnosis	float64	787	91.72%	1	22	very high % missing
Dx:Cancer	int64	0	0.00%	0	1	OK (boolean)
Dx:CIN	int64	0	0.00%	0	1	OK (boolean)
Dx:HPV	int64	0	0.00%	0	1	OK (boolean)
Dx	int64	0	0.00%	0	1	OK (boolean)
Hinselmann	int64	0	0.00%	0	1	OK (boolean)
Schiller	int64	0	0.00%	0	1	OK (boolean)
Citology	int64	0	0.00%	0	1	OK (boolean)
Biopsy (prediction target)	int64	0	0.00%	0	1	OK (boolean)

Data "Cleaning"

- Significant Missing Data
  - About 80% of the variables were missing some information.
  - A few variables were missing data in over 90% of records.
  - Several variables had missing data in over 12% of the records.
- Strategy
  - Because the dataset is small, preserve as much information as possible.
  - Independently evaluate each variable to select the best approach.
- Summary
  - Useless variables were deleted.
  - A combination of "feature engineering" and filling missing values with measures of centrality (mean, median, or mode) was used.

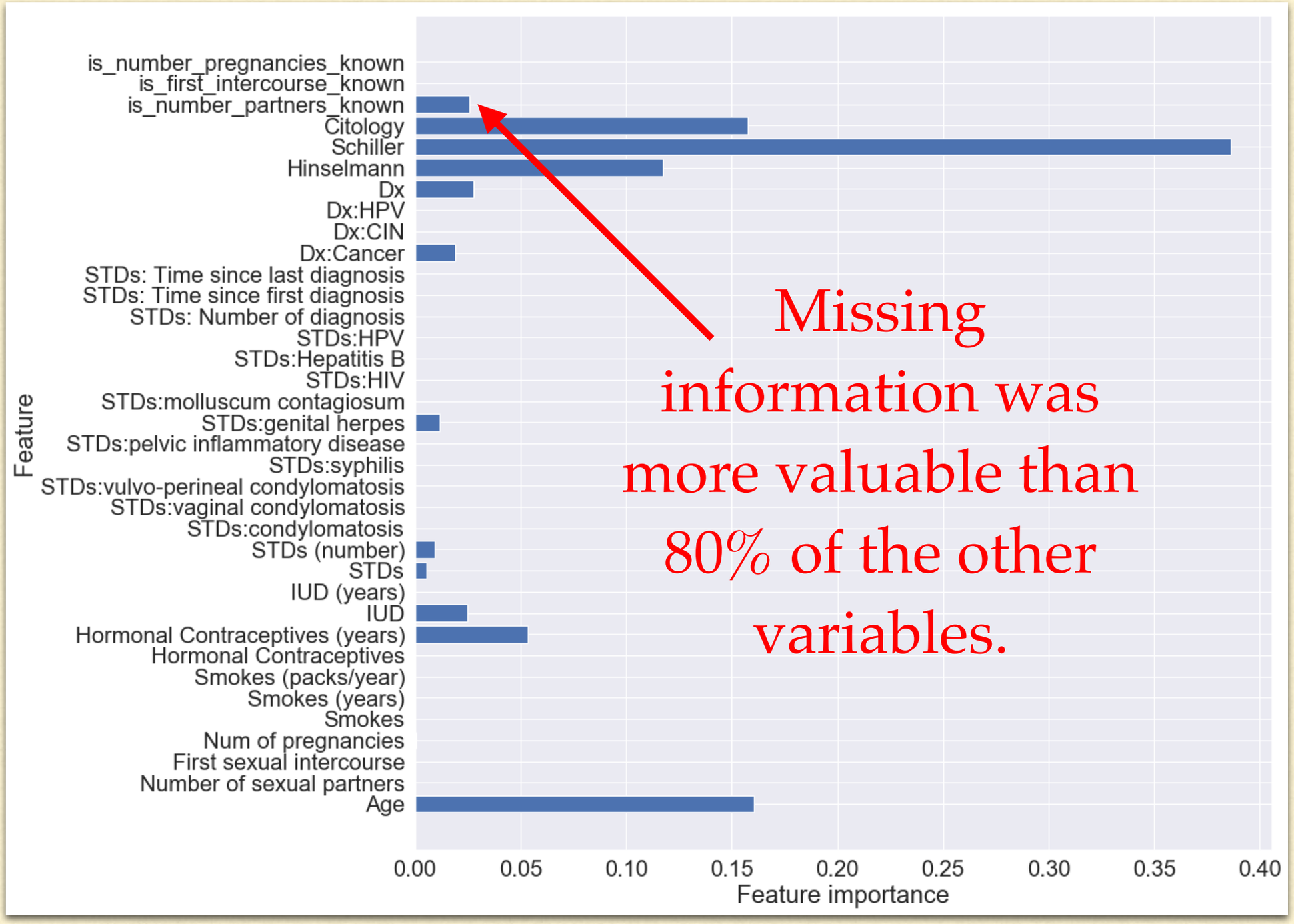


Missing Data

- Missing information may in itself contain information…
  - Example: Number of sexual partners was missing a value in only 3% of the records.
  - It could be that some patients were unwilling to provide this information due to:
    - Embarrassment
    - Fear of being judged
    - Fear that the information may be ‘leaked’ to others
- The data was processed to preserve the distinction between records that had this information and those that did not.
- “Lazy” approaches such as filling the missing information with the mean or deleting the records (only 3%) would not have caught a statistically significant relationship.

This “missing” information ended up being the 7th most important factor.

Missing Data



Predictive Model

- Precision vs. Recall
  - Precision represents the proportion of the models' predictions of cancer where cancer is actually present.
  - Recall represents the proportion of all cases of cancer that the model accurately predicted.
- Prioritize Recall
  - Find the cancer cases!

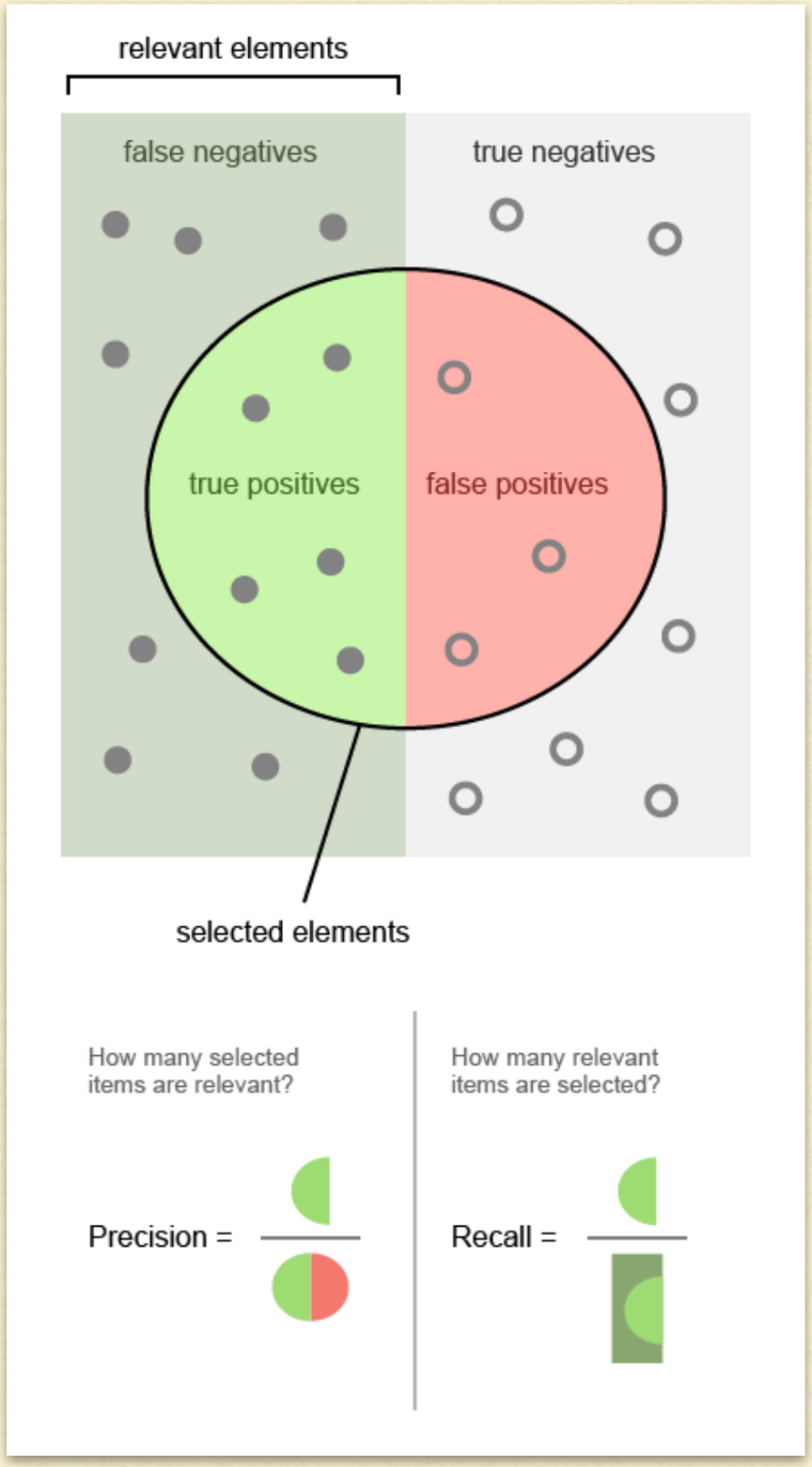
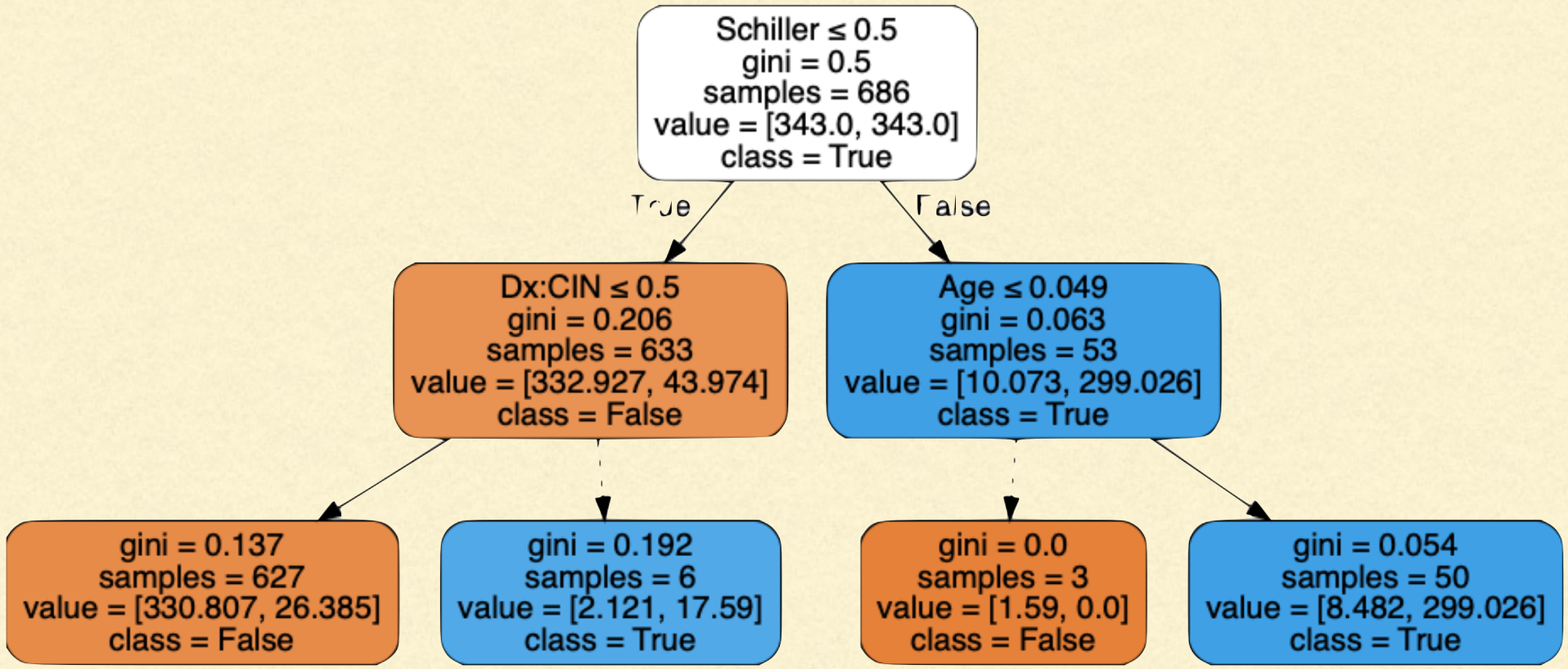


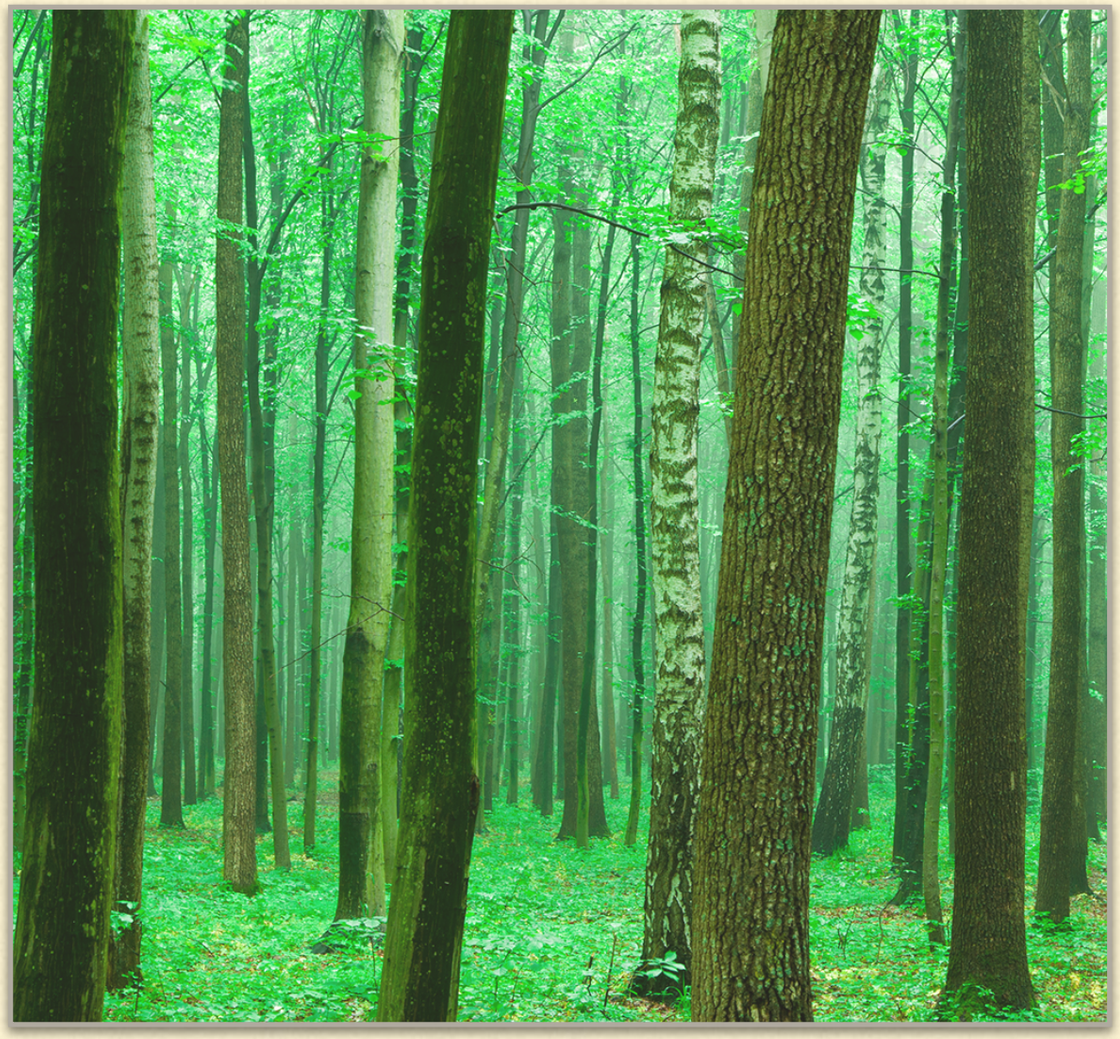
Image By Walber - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=36926283

A Decision Tree

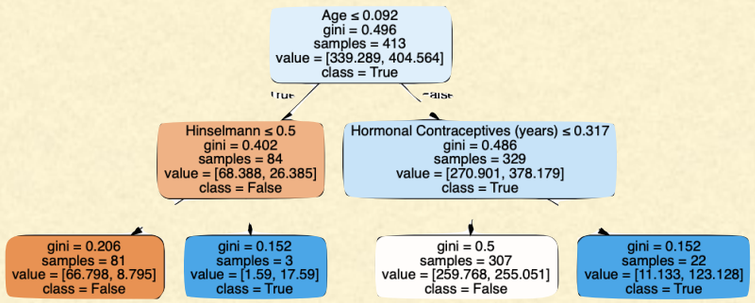


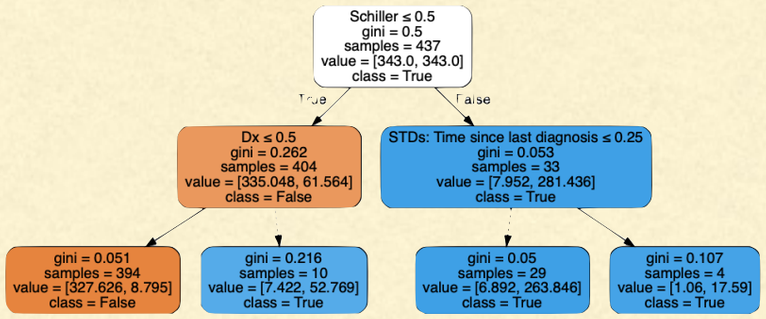
This model "finds" 88% of patients with cervical cancer (recall), but when it predicts a patient has cancer it is only correct 61% of the time (precision).

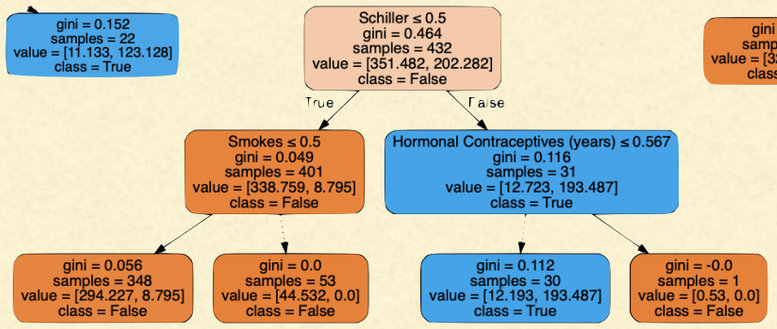
The Predictive Model

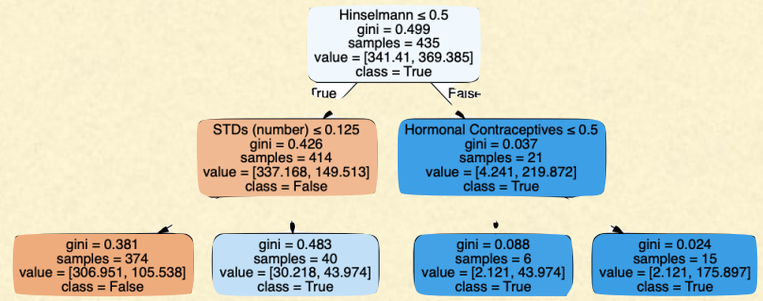


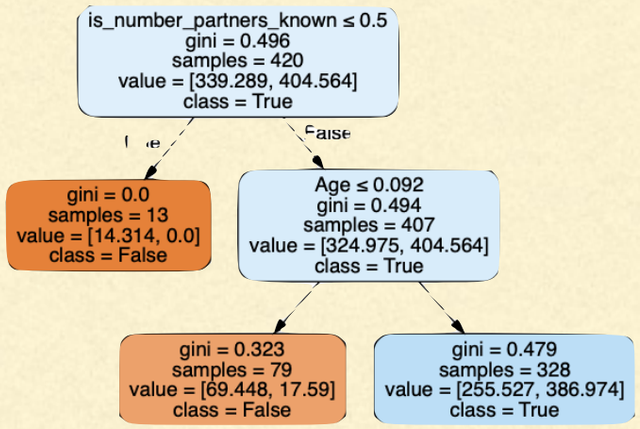
- Random Forest
  - A random forest is a “forest” of different “decision trees” that each make a prediction.
  - Like a democracy, if more “trees” vote for a particular prediction, that prediction wins.





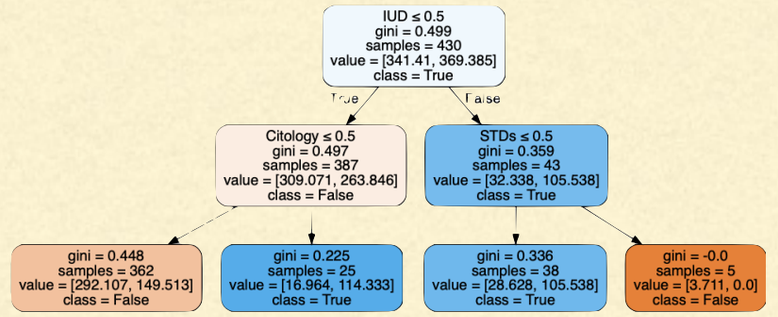


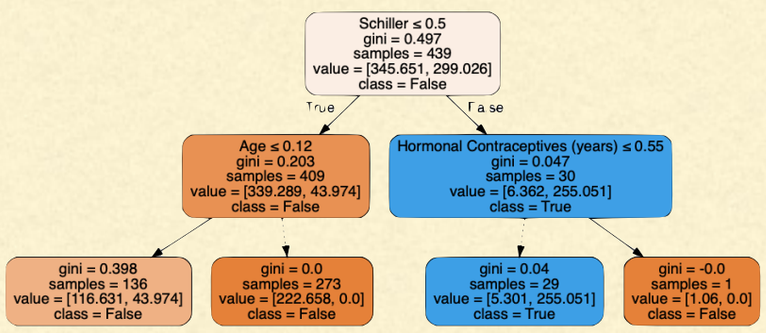


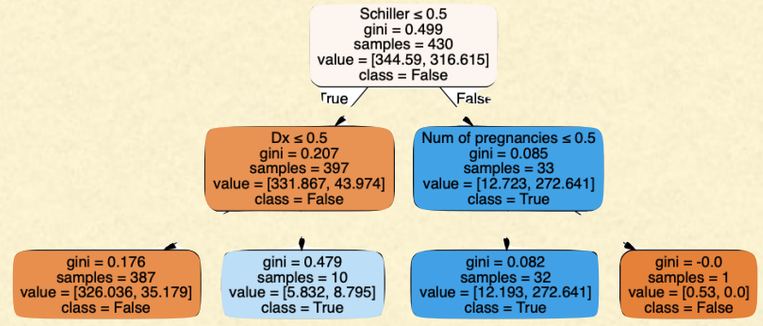


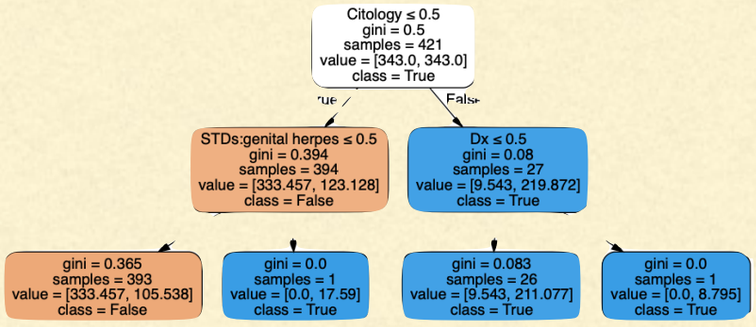
The Forest

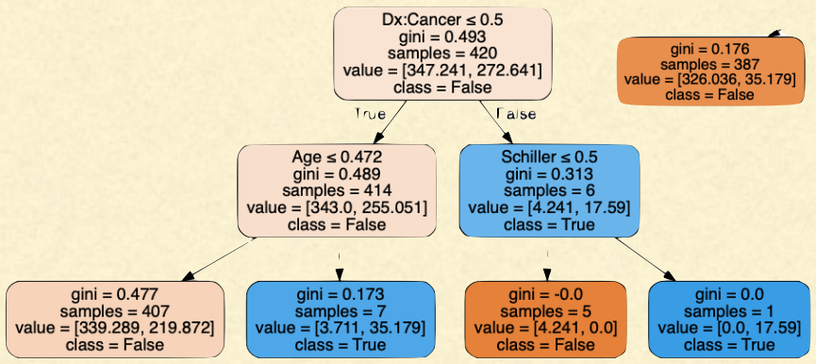
This model "finds" 88% of patients with cervical cancer (recall), and when it predicts a patient has cancer it is correct 67% of the time (precision).

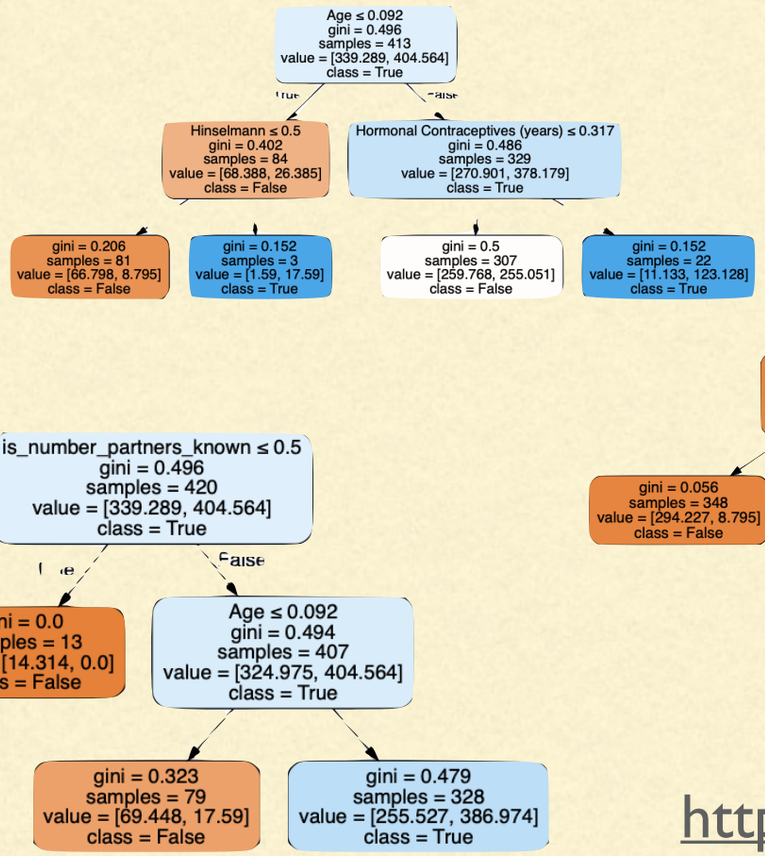


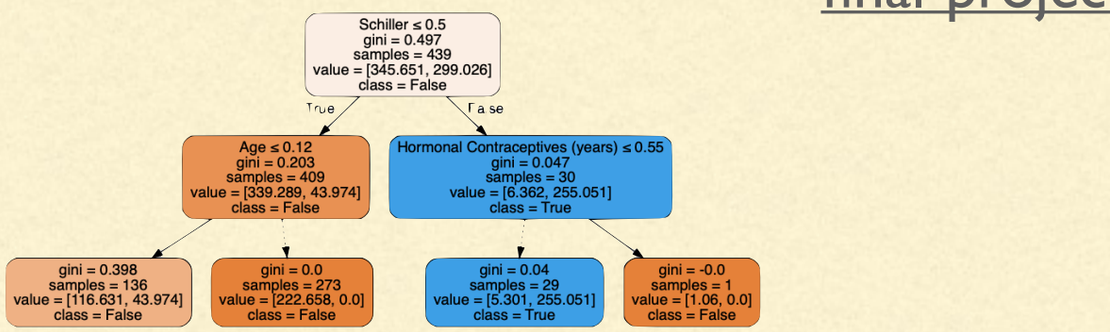


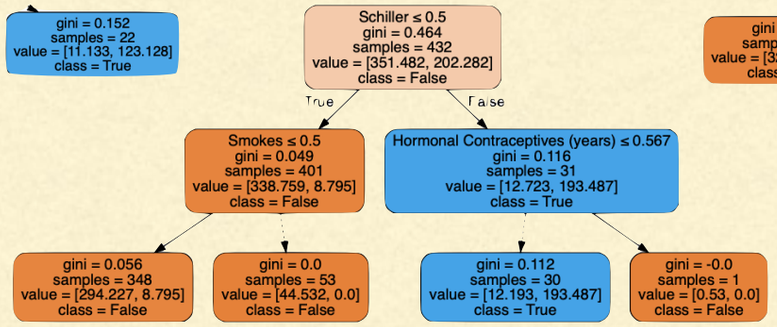


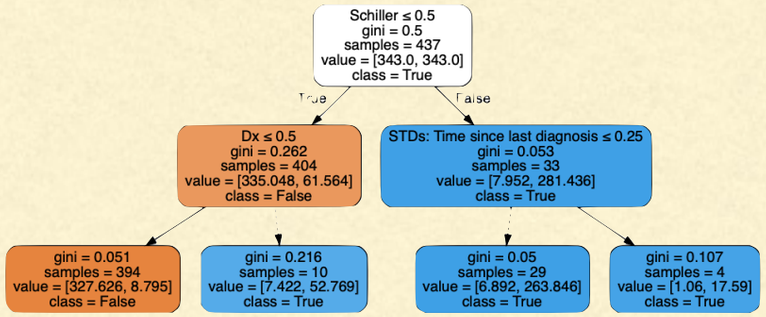


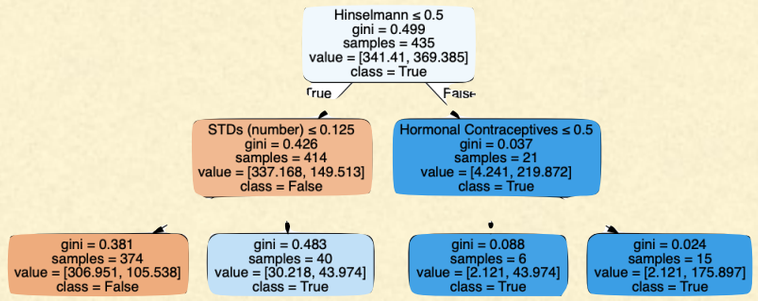












https://github.com/giorosati/dsc-3-final-project-online-ds-pt-100118