CAPSTONE PROJECT 1 PROPOSAL

By: Mustafa KADIOGLU
Date: 06/20/2018

Project Title: Cervical Cancer Risk Factors

Problem: Cervical cancer is the third most common cancer in women worldwide, affecting over 500,000 women and resulting in approximately 275,000 deaths every year. After reading these statistics, you may be surprised to hear that cervical cancer is potentially preventable and curable.

The goal of this project is to explore machine learning models to estimate the probability associated with cervical cancer, with respect to features to be identified.

Who is our focus of interest?

Our focus of interest is medical staff, patients, females

Data: The dataset was collected at 'Hospital Universitario de Caracas' in Caracas, Venezuela. ,and it is available from the link below.
[https://archive.ics.uci.edu/ml/datasets/Cervical+cancer+%28Risk+Factors%29#]The dataset comprises demographic information, habits, and historic medical records of 858 patients, with 36 features per patient. Several patients decided not to answer some of the questions because of privacy concerns (missing values).

Feature Information:

(int) Age
(int) Number of sexual partners
(int) First sexual intercourse (age)
(int) Num of pregnancies
(bool) Smokes
(bool) Smokes (years)
(bool) Smokes (packs/year)
(bool) Hormonal Contraceptives
(int) Hormonal Contraceptives (years)
(bool) IUD
(int) IUD (years)
(bool) STDs
(int) STDs (number)
(bool) 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
(int) STDs: Number of diagnosis
(int) STDs: Time since first diagnosis
(int) STDs: Time since last diagnosis
(bool) Dx:Cancer
(bool) Dx:CIN
(bool) Dx:HPV
(bool) Dx
(bool) Hinselmann: target variable
(bool) Schiller: target variable
(bool) Cytology: target variable
(bool) Biopsy: target variable

Data set has missing values. Missing values of each feature and the percentage are stated below:

	Missing Values	% of Total Values
STDs: Time since last diagnosis	787	91.7
STDs: Time since first diagnosis	787	91.7
IUD	117	13.6
IUD (years)	117	13.6
Hormonal Contraceptives	108	12.6
Hormonal Contraceptives (years)	108	12.6
STDs:vulvo-perineal condylomatosis	105	12.2
STDs:HPV	105	12.2
STDs:Hepatitis B	105	12.2
STDs:HIV	105	12.2
STDs:AIDS	105	12.2
STDs:molluscum contagiosum	105	12.2
STDs:genital herpes	105	12.2
STDs:pelvic inflammatory disease	105	12.2
STDs:syphilis	105	12.2
STDs:cervical condylomatosis	105	12.2
STDs:vaginal condylomatosis	105	12.2
STDs:condylomatosis	105	12.2
STDs (number)	105	12.2
STDs	105	12.2
Num of pregnancies	56	6.5
Number of sexual partners	26	3.0

Smokes (packs/year)	13	1.5
Smokes (years)	13	1.5
Smokes	13	1.5
First sexual intercourse	7	0.8

Modeling approach: In this study, we will mainly focus on machine learning classification algorithms (e.g., Logistic Regression and Random Forests, with hyper-parameter tuning) to predict the Cervical Cancer Risks based on the features listed above.

Possible limitations:

Since two features (STDs: Time since last diagnosis and STDs: Time since first diagnosis) have more than %90 missing values, we might need to drop these two features. Missing values for other columns will be considered on a case-by-case basis.

We will use 'Dx:Cancer' feature as our class target. We have already observed that the ratio of data points characterizing cancer patients to those that characterize non-cancer patients is 17/651, and thus the dataset is imbalanced. Therefore, we might need to implement over sampling and/or under sampling approaches in combination with classification algorithms.

Deliverables:

1. Proposal
2. Data Wrangling
3. Exploratory Data Analysis
4. Data Storytelling
5. Applications of Inferential Statistics
6. Milestone Report
7. Machine Learning Models (Logistic Regression, Random Forests)
8. Capstone Project Report
9. Presentation Slide Deck