Cervical Cancer Risk Factors in Low-Income Countries.

Prepared for: Institute of Data
Prepared by: Doron Man
15 September 2020

Background

Cervical cancer is the second most common cancer in women worldwide.
Although it has been relatively well controlled in many high-income countries, it remains the no 1 cause of cancer-related death in low-income countries.
The reason is that for a successful cytology tests few things needs to happen.

1. Highly trained health care professionals need to be involved in the screening.
2. There is a need fo quality assurance control.
3. Smears need to be transported rapidly to the laboratory.
4. Cytology tests require repeat visits and tests.

All of the above requirements are very hard to achieve in low-income settings.

Problem statment

The problem this project is trying to investigate is the reasons for high mortality rate from cervical cancer in low-income countries compare to high - income countries.

This is a valuable problem to address because solving that problem can save lives.

Industry / Domain

Healthcare.

Stack-holders

Healthcare providers.

Governments.

Research centres.

Business question

The question this project will try to answer is if machine learning can help implement a more simple screening tests by providing a quality assurance tool that can be used by a wide range of health care providers (physicians, nurse, midwives and technicians).

Data question

The data question this project will try to answer is if by using a basic historical information and a simple visual inspection test result we can predict the risk of a cervical cancer in high accuracy using machine learning tools.

Data

The data was obtained from University Hospital of Caracas and it comprises of demographic information, habits, and historic medical records of 858 patients. The data contains 33 features that include:

IUD - Intra Uterine Device (a form of contraceptive)
STDs - Sexually Transmitted Diseases
HPV - Human Papilloma Virus
HIV - Human Immunodeficiency Virus
AIDS - Acquired Immunodeficiency Syndrome (caused by HIV)
CIN - Cervical Intraepithelial Neoplasia
Dx - Medical Abbreviation for Diagnosis
Dx:Cancer (person had previous cervical cancer diagnostic)
Dx:CIN (person had previous diagnostic of Cervical intraepithelial neoplasia)

Data science process

EDA

The average age of women in the dataset is 27.
55 women from 858 records have been diagnosed with cervical cancer.
From the Age distribution of women that diagnosed with positive biopsy, we learn that women age 25-35 face the highest risk of cervical cancer.
We also learn from the Pearson correlation heat-map that the strongest predictors are the Schiller's test, the Hinselmann test and cytology.

Modelling

The main features that were used in the machine learning model were:
1. Schiller's test,
2. DX:Cin - cervical dysplasia diagnosis
3. Age
4. Hormonal Contraceptives
4. First sexual intercourse.

Using SHAP, a state of the art game theoretic approach to explain the out of machine learning models, I analysed an xgboost model that used all the features of the dataset. I learned that the model only used the above 5 features to predict the target value.

Models used in this project:
Logistic regression
Random Forest
Support Vector Machine
MLP Classifier (artificial neural network)
Train time was in average 2 minutes
The model performance metrics were:
Accuracy
Precision
Recall
ROC AUC
LOG LOSS

Selected Model

After comparing all the models was the MLP Classifier with the fallowing metrics:

Accuracy = 0.95
Precision = 0.69
Recall = 0.89
ROC AUC = 0.93
Log Loss = 0.08

Outcomes

The Artificial Neural Network model performed very well and can
improve, complement and enhance the visual screening process in number of ways.
It can serve as a valuable quality assurance control tool.
It can reduce reliance on infrastructure.
It can raise the level of confidence of healthcare providers to refer to biopsy.
It can reduce number of false negative and false positive due to the level of skill of the healthcare provider.

Data answer

The data question was answered satisfactorily.

The confidence level is 89% (Recall)

Business answer

The business question was answered satisfactorily.

Response to stake-holders

This is a very useful tool to deploy in low-income setting that have little to none access to infrastructure.
This tool will enable to provide a wide range of healthcare providers a much needed quality control tool that will help them reach a conclusion regarding their observation and intuitions.

End-to-end solution

The end to end solution will be to create an API that healthcare provider can log-in patient information and tests results and get a prediction of cervical cancer risk.