Cervical Cancer Risk Factor Analysis

By: Shahir Kottilingal

Abstract

Cervical Cancer Risk Factor Analysis project was on “Cervical cancer (Risk Factors) Data Set” from UCI machine learning repository. Tried to find out the major risk factor for Cervical Cancer and how HPV prevention can reduce the Cervical Cancer cases. Methods used was Pearson’s correlation coefficient method, Diagonal correlation matrix heat map, and Line Plot. Findings from the analysis of dataset was, 1) Major risk factor for Cervical Cancer is HPV.
2) HPV prevention can lead to major reduction in Cervical Cancer.

Motivation

About 11,000 new cases of invasive cervical cancer are diagnosed each year in the U.S. Although it is the most preventable type of cancer, each year cervical cancer kills about 4,000 women in the U.S and about 300,000 women worldwide. Analysis of major risk factors will help us to take precautionary steps to prevent Cervical Cancer.

Dataset

“Cervical cancer (Risk Factors) Data Set” from UCI machine learning repository.
This dataset focuses on the prediction of indicators/diagnosis of cervical cancer.
The features cover demographic information, habits, and historic medical records.
This data set contains 858 number of instances and 36 number of attributes.

Data Preparation and Cleaning

This data set contain 858 instances and 36 attributes OR columns. But only 10 columns were numeric type.26 columns were objective data type. So had to convert all these columns to numeric data type. And, there were “?” characters. So converted them to NaN values. We found that there were two columns with 787 NaN values. So we droped these two columns. Also many columns were with around 100 NaN values, but not more than 117 rows. So decided to drop all rows with NaN values. Finally our cleaned data set contain 668 number of instances and 34 number of attributes.

Research Question(s)

1. What are the major risk factors for cervical cancer?

2. Will HPV prevention help to reduce the cervical cancer risk?

Methods

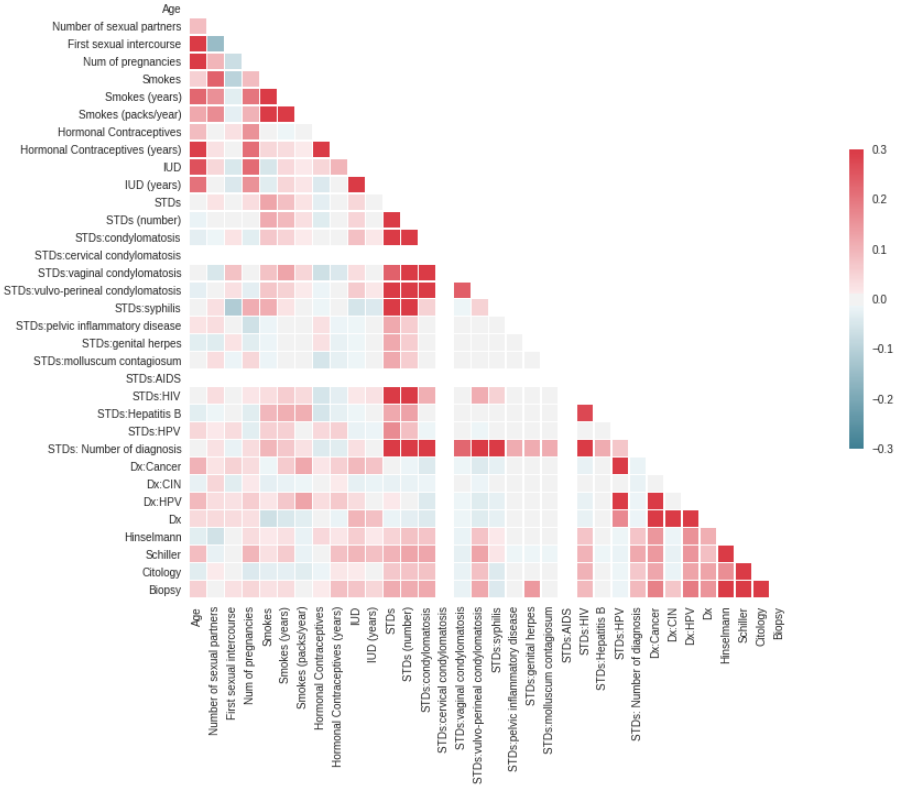
1. Pearson's Correlation Coefficient: To find out the correlation of all other features with Dx:Cancer feature.
2. Heat map of correlation matrix: To visualize the overall correlation between features.
3. Linear line plot : To see the linear relationship between Dx:Cancer and Dx:HPV features.

Findings

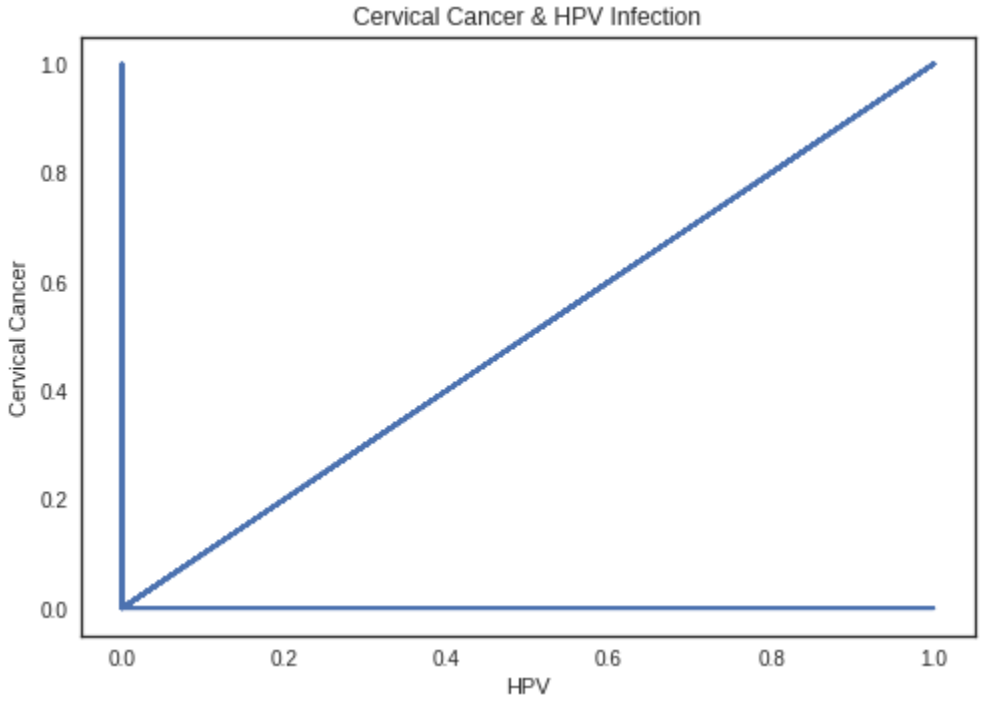
Firstly, Major risk factor for Cervical Cancer is HPV infection. As per Pearson’s correlation coefficient , There is high positive correlation (0.907229) between cervical cancer and HPV positive cases. However higher Age is also contributing to the risk of Cervical Cancer in a slightly manner as per the Pearson’s correlation coefficient (0.105179). Please refer the attached diagonal correlation matrix heat map.

Secondly, There is a strait linear relationship between HPV infection and Cervical Cancer.This is a clear indication that HPV Prevention will help us to reduce the Cervical Cancer cases. Please refer the attached linear line plot graph.

Findings: Diagonal Correlation Matrix



Findings: Linear Line Plot



Limitations

The dataset comprises demographic information, habits, and historic medical records of 858 patients from Venezuela. There may be a little variation on patients from other parts of the word based on their ethnicity and life style.

Conclusions

1. Major risk factor for Cervical Cancer is HPV infection.

2. HPV prevention will reduce the risk of Cervical Cancer.

Acknowledgements

“Cervical cancer (Risk Factors) Data Set” taken from UCI machine learning repository.

No feedbacks received.

References

The work has been done by myself (Shahir Kottilingal)