Team Titans

- [ ] **Stephen Njuguna- Team Lead**
- [ ] **Ayub Bett**
- [ ] **Harun Rotich**
- [ ] **Mutura Kuria**
- [ ] **Farnadis Kanja**

An Analysis of Risk Factors of Cervical Cancer

1.Business Understanding

Business Overview

Cervical cancer kills an estimated 260,000 women every year and has been said to be the leading cause of cancer deaths in women. At least 85% of these deaths take place in developing countries. In Kenya, it is ranked as the second highest killer cancer after breast cancer among women. WHO (2018) claims that 33 per 100,000 women in Kenya were diagnosed with cervical cancer and unfortunately 22 in 100,000 women will die from the disease. Notably, all women, in the reproductive ages, are at a risk of developing the disease but it is more prevalent for women over the age of 30.

Recent research shows that routine cervical cancer screening and human papillomavirus testing has dramatically reduced cervical cancer deaths especially in developed countries, developed countries still lag behind. Poor infrastructure, medical facilities and insufficient health care workers witnessed in developing countries result in poor prognosis. Research shows that HPV is the leading cause of cervical cancer.

Business Objective

To determine if long term infection of human papillomavirus(HPV) is the main cause of cervical cancer.

Business Success Criteria

To create a model that can predict the likelihood of a woman being diagnosed with cervical cancer from the risk factors being investigated.

Assessing the Situation

1. **Resource Inventory**
    a. Datasets: Cervical Cancer Dataset Link
    b. Software(Github, Google Collaboratory, Python Libraries)
2. **Assumptions**
    a. The data provided is correct and up to date
3. **Constraints**
    a. Some women were not comfortable disclosing personal information and thus the dataset had many missing values.

Data Mining Goals

Our data mining goals for this project are as follows:

- Determine the highest risk factors of cervical cancer
- Create a Working model for prediction
- Reject or fail to reject the null hypothesis

Data Mining Success Criteria

Our success criteria will be measured by the following criteria;

- Model has a high accuracy
- Have enough evidence to either reject or fail to reject the null hypothesis.

2. Data Understanding

Data Understanding Overview

For this project, we used a data set chosen from github. It was conducted by recording the health history, demographic information and lifestyle. This data set was collected from 'Hospital Universitario de Caracas' in Caracas, Venezuela. The dataset had 835 rows and 32 columns.

Data Description

This data set was collected from 'Hospital Universitario de Caracas' in Caracas, Venezuela. The dataset had 835 rows and 32 columns.

Verifying Data Quality

The data was cross checked and was found to be of good quality.

3. Data Preparation

These are the steps followed in preparing the data

1. **Loading Data**

   Loaded the datasets from the CSV and a data frame created.
2. **Importing Libraries**
   We import python libraries that will help in the analysis. Python, Numpy, Seaborn and Matplotlib.
3. Cleaning Data
   Data cleaning was done in the following steps
   1. Dropping unnecessary columns. A total of 12 columns were dropped from the dataset.

2. We checked the data types in the data to see if any needed changing. They were all ok.
3. We checked for missing values and discovered that they were numerous. We replaced them with the mode of the particular column. We chose this over filling with zero in order to maintain the integrity of the data as most of the variables were boolean.
4. We checked for outliers and decided to keep them for their uniqueness
5. Having done the necessary cleaning, we exported the clean data set and did a preview of the same.

4. Data Analysis

There are various types of analysis that were conducted in this section

Univariate Analysis

Once we did univariate analysis, we came to the following conclusions;

- We noticed that many of the respondents in the dataset were between 20 and 30 years old.
- We noticed that most respondents had between one and three sexual partners in their lifetime. However, it was noticed that there was one extreme outlier of 26 sexual partners.
- We noticed that most people had between and three pregnancies.
- Since the kurtosis values are greater than zero, then the distributions of respective columns have heavier tails and thus our data is a leptokurtic distribution.
- Since we have positive values for skewness, the distributions are skewed to the right, the tail on the curve's right-hand side is longer than the tail on the left-hand side, and the means are greater than the modes in the respective columns. This situation is also called positive skewness.

Bivariate Analysis

Once we did bivariate analysis, we came to the following conclusions;

- No direct correlation was found between age and number of sexual partners.
- Most respondents had between 1 and 5 sexual partners.
- Further, the women were most sexually active between the ages of 12 and 50.
- No direct correlation was found between age and number of pregnancies.
- It was noticed that women who had carried one pregnancy also had the most number of each of those diagnoses. This might not be a correlation but instead just a coincidence.
- It was also noticed that the diagnosis grew less as the number of pregnancies increased. However, those without any pregnancies were almost equal to the ones with about six children.
- Heat map summary
    - There is a correlation between HIV and STDs.
    - There is a correlation between STDs:Hepatitis B and STDs.

- There is a very high correlation between HPV Diagnosis and Having another cancer.
- There is a strong correlation between Diagnosis of HPV and diagnosis of cervical cancer.
- Similarly there is a strong correlation between Diagnosis of CIN and diagnosis of cervical cancer.
- Similarly there is a strong correlation between Diagnosis of any other cancer and diagnosis of cervical cancer.
- There is a correlation between the age of a person and the number of pregnancies they have had.
- There is a very high correlation between STDs:vulvo-perineal condylomatosis and STDs:condylomatosis

Multivariate Analysis

After separating the target label and subjecting our features to the LDA model, we discovered that a diagnosis of CIN, any other cancer, HPV, use of IUD had the highest coefficient correlation with the target label i.e. diagnosis of cervical cancer.

Model Prediction

We created a model with a 97.6% accuracy of predicting the chances of a woman being diagnosed with cervical cancer while taking into account the variables under investigation. The model was tested and proven to give accurate results.

Hypothesis Testing.

Null Hypothesis: HPV is one of the leading causes of cervical cancer

Alternative Hypothesis :HPV is not one of the leading causes of cervical cancer

Since our data set is quite large, we decided to use a sample of 200 respondents. We used simple random sampling so that all participants can have an equal and fair chance of being selected and thus the resulting sample is unbiased. It was chosen since the population under study is homogenous ie all female and a cervical cancer diagnosis is beyond their control.

At a significance level of 0.05, and using a two tailed test, the p-value was greater than the significance level and thus, we fail to reject Ho.
The above analysis was done using Google Collaboratory. The full analysis can be found in the following notebook.[Link]
A dashboard summary of the findings has been shown on Tableau as shown Tableau Link

5. Recommendations and Conclusions

Having done the analysis, we made the following conclusions;

- While HPV is a leading risk factor of cervical cancer , there are other higher risk factors that predispose one to cervical cancer;
- A CIN diagnosis has been seen to be the leading risk factor of cervical cancer.
- A diagnosis of any other cancer, apart from cervical cancer, is the second leading risk factor of cervical cancer
- A HPV diagnosis comes in third as a risk factor when it comes to a cervical cancer diagnosis.

In light of this, we recommend that governments across the world be more vigilant about educating the women about early detection of cervical cancer through regular screening. Further, young girls should receive the HPV vaccine before they reach their child bearing years to reduce their risk factors of cervical cancer. In Kenya, in particular, the recently introduced HPV vaccination program of girls between the ages of 9 and 11 should be aggressively followed up. Regular screening for all other cancers is necessary so as to further reduce the risk of developing cervical cancer. Women with several sexual partners should be encouraged to use condoms in order to prevent STDs, many of which go untreated and thus increasing the risk factors. For a cancer that is largely treatable, it is devastating to continually lose so many women to cervical cancer.