Yuefu Jiang

Biostatistics Independent project: Cervical Cancer Risk Factors

Assignment 2: **Data Description**

1. **Methods**

This is a clinically collected dataset as the main method of data collection was voluntary self-reveal of past clinical history (like STDs diagnosed before) and clinical diagnosis (like diagnosis of cervical cancer). The dataset was assembled and utilized by Dr. Kelwin Fernandes et al for a machine learning approach in cancer diagnosis.

1. **Study site**

All patient data was collected Hospital Universitario de Caracas in Caracas, Venezuela. Data was collected by doctors from 858 patients with or without cervical cancer in forms of survey and diagnosis. The data was maintained by by Dr. Kelwin Fernandes of INESC TEC & FEUP, Porto, Portugal.

1. **Timeline**

As the dataset is already completed, the source of the data would not halt the progress of the project.

1. **Feasibility**

There are some concerns about the missing values displayed in the dataset as these missing values could be crucial to further analysis of this dataset. For example, this dataset has past STD history of patients, while a number of patients chose to not reveal this data to researcher due to privacy issues. In the case of this project, I think that some preprocessing of this data is needed.

1. **Collaborators**

There is no collaborator working with me on this project.

1. **Reference**

Link of database: https://archive.ics.uci.edu/ml/datasets/Cervical+cancer+%28Risk+Factors%29

Kelwin Fernandes, Jaime S. Cardoso, and Jessica Fernandes. Transfer Learning with Partial Observability Applied to Cervical Cancer Screening. *Iberian Conference on Pattern Recognition and Image Analysis. Springer International Publishing,* 2017.

Yaseen Alwesabi, Avishek Choudhury and Daehan Won. Classification of Cervical Cancer Dataset. *Proceedings of the 2018 IISE Annual Conference At Orlando*, July 2018.