Cervical cancer risk prediction and identification of key risk factors.

Introduction: Worldwide, cervical cancer is both the fourth-most common cause of cancer and the fourth-most common cause of death from cancer in women. In 2012, an estimated 528,000 cases of cervical cancer occurred, with 266,000 deaths.This is about 8% of the total cases and total deaths from cancer. About 70% of cervical cancers occur in developing countries. (https://en.wikipedia.org/wiki/Cervical_cancer)

Problem statement: Despite the possibility of prevention with regular cytological screening, cervical cancer remains a significant cause of mortality in low-income countries. As in many other diseases, there are several screening and diagnosis methods. For instance, in the detection of precancerous cervical lesions, screening strategies include cytology, colposcopy and biopsy. In developing countries resources are very limited and patients usually have poor adherence to routine screening due to low problem awareness. Consequently, the prediction of the individual patient's risk and the best screening strategy during her diagnosis becomes a fundamental problem. Most of these screening methods highly depend on the physician expertise and subjective comfort on the decision process, being a key aspect to improve data acquisition using the physician preferences. Identification of key risk factors would improve the collection of necessary patient data and, on their basis, determine whether the patient needs an additional examination.

Data Source: This project will examine data collected at 'Hospital Universitario de Caracas' in Caracas,Venezuela.(http://archive.ics.uci.edu/ml/datasets/Cervical+cancer+%28Risk+Factors%29) The dataset comprises demographic information, habits, and historic medical records of 858 patients. Several patients decided not to answer some of the questions because of privacy concerns (missing values).

Analysis: The aim of this project is to find correlations between different risk factors for cervical cancer and through statistical analysis and visualization methods identify the key ones. For cervical cancer risk prediction predictive modeling will be used.

Deliverables for this project: The code used to perform the analysis and a slide deck explaining key steps will be provided.