### **Cervical cancer risk prediction and identification of key risk factors.**

1. **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.

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. 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.

1. **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 was obtained from UCI repository and 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).

1. **Project objectives:**

The aim of this project is identification of risk factors for cervical cancer through statistical analysis, visualization methods and predictive modeling.

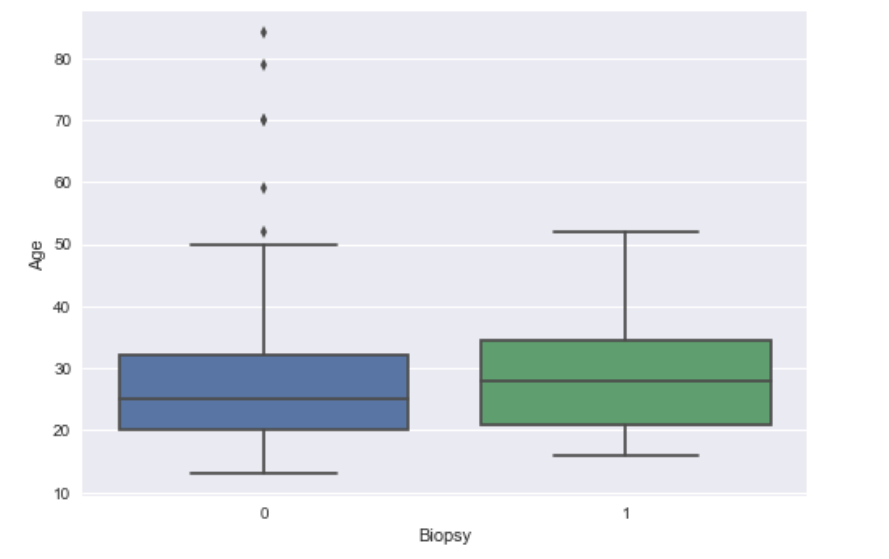
1. **Preparing and cleaning the data:**

Dataset consists of 858 patient’s records, each record with 35 features. There are some missing values named as '?', that were replaced with median value for the column.

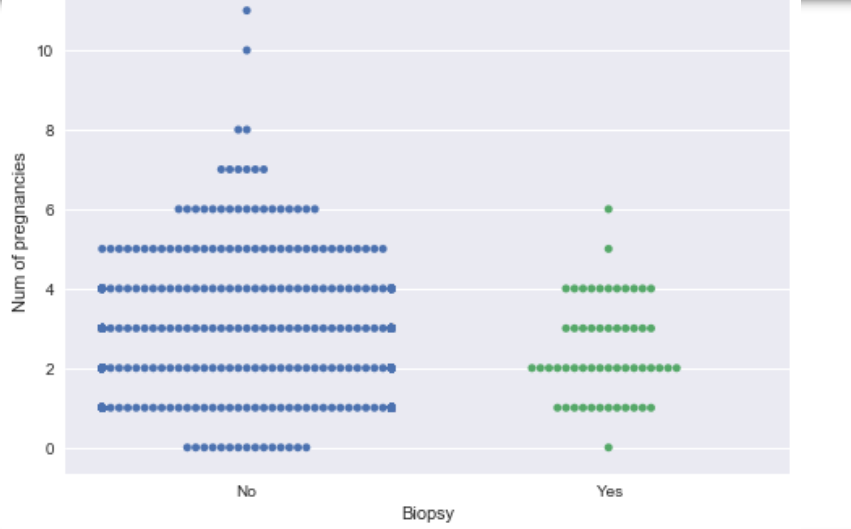
1. **EDA**

For better data understanding some plots and visualizations were done.

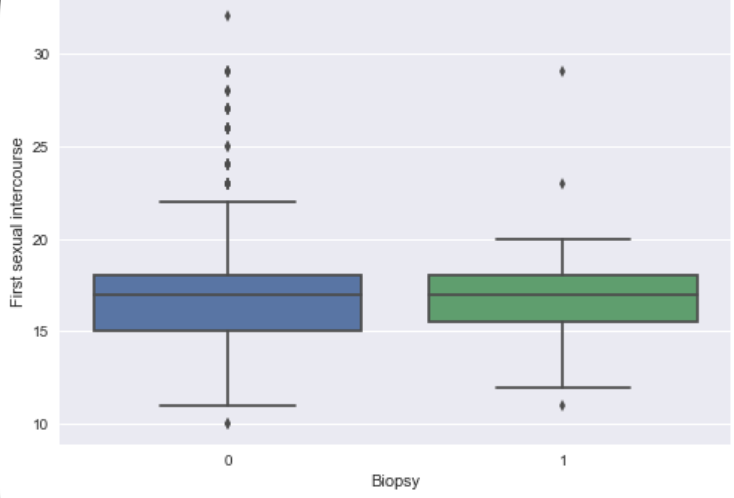
The age of most patients with biopsy(+) is from 21 to 35 years old:



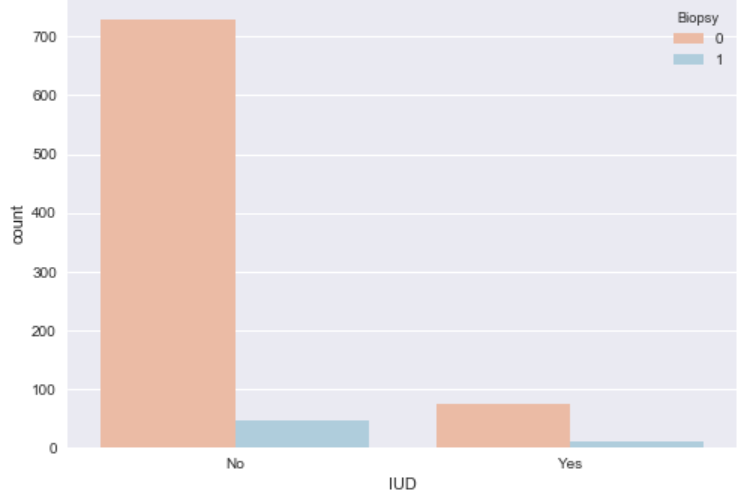
Patients with biopsy(+) have 1 or more pregnancies:



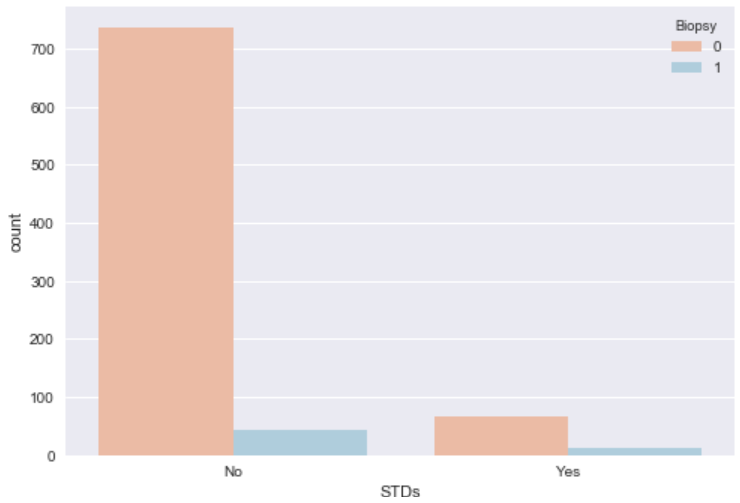
First sexual intercourse was under age of 20:



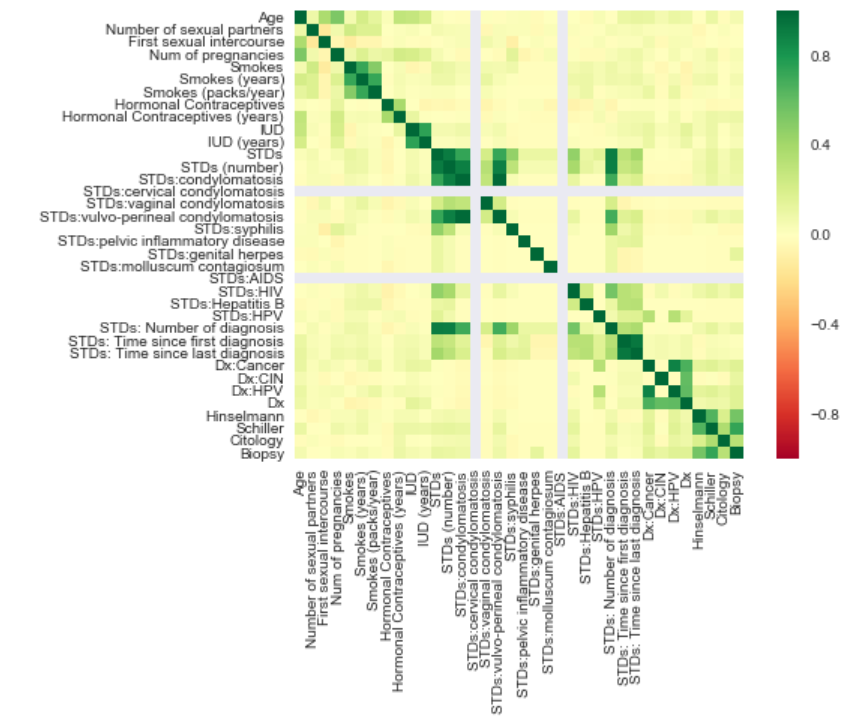
From all the patients with IUD about 20% is also with biopsy(+):

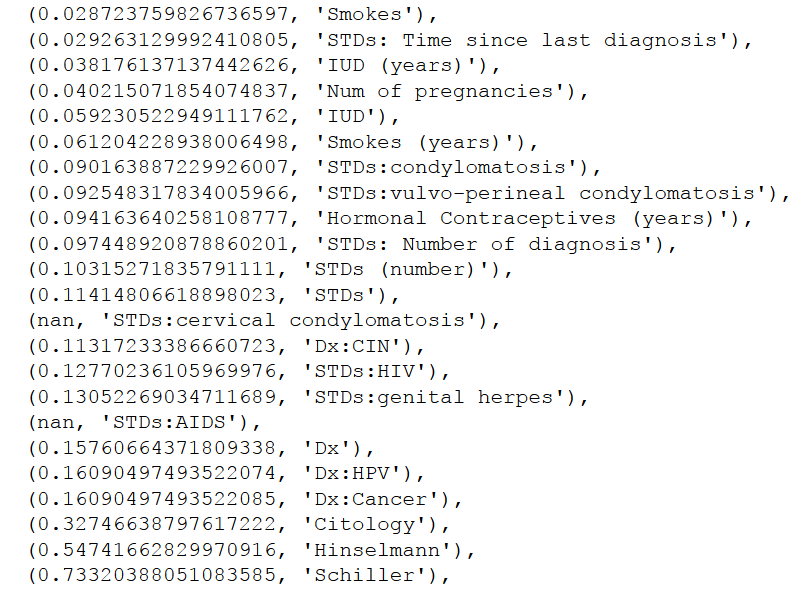


The same is for patients with STDs (sexually transmitted diseases):



From correlation map and after calculation of correlation coefficients 'Schiller', 'Hinselmann' and 'Cytology' have the highest correlation with biopsy. It makes sense from medical point of view.

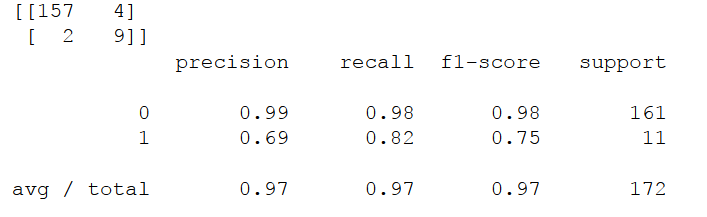




1. **Findings**

I used only highly correlated columns for prediction. I tried several classifiers (KNeighbors, DecisionTree, RandomForest, SVM, LogisticRegression) and it appears that DecisionTreeClassifier performs the better result in predicting if patient has cancer.

Here we can see the results. From 161 patients without cancer the model predicts 157 correctly, and from 11 who has cancer predicts 9 correctly.



7. **Conclusions and recommendations**

The project aimed at analyzing patients data for cervical cancer prediction and to find out what features are significant in determining cervical cancer. According to our findings these 3 features have the strongest impact on cancer predictions: Hinselmann, Schiller and Cytology test results. Factors that may raise a woman's risk of developing cervical cancer: HIV, high number of sexually transmitted diseases, long time use of hormonal contraceptives.

Research continues to look into what factors cause this type of cancer and what women can do to lower their personal risk. There is no proven way to completely prevent this disease, but there may be steps you can take to lower your cancer risk. For future research and analysis one of the most important things is data collection and it’s accuracy.