Cervical Cancer Risk Factors Assessment – Final Report

Capstone 1 Project Report – Springboard Data Science Career Track

Notebook by Soumyadip Kundu

1. Acknowledgements

- Mentor – Kenneth Gil-Pasquel
- Springboard team
- Cover Image – Google Images

2. Background

Cervical Cancer is one of the top gynecological cancers affecting the cervical cells of the uterus. According to current WHO data, cervical cancer is the second most frequent cancer occurring in women. In 2018, an estimated 570000 new cases were reported which accounted for nearly 14.7% of all female cancers. In the United States itself, around 11000 new cases of cervical cancer are diagnosed out of which about 4000 women die. It has also been noticed that about 90% of the fatality caused due to cervical cancer occurs in low- and middle-income countries. However, the mortality rate can be fairly rate of cervical cancer can be significantly reduced though a

comprehensive program that includes early diagnosis, prevention, effective screening, and treatment.

Some of the major factors associated with cervical cancer include age, socio-economic factors, usage of birth-control pills, STD infections and HPV infection among many others. Women with a weakened immune system are also more susceptible to getting HPV infections which will significantly increase the individual's chances of getting diagnosed with cervical cancer.

3. Problem Statement

The major objective of this study is to build a predictive machine learning model which can assess all the risk factors and will lead us to ascertain the major risk factors associated with cervical cancer

4. Data

The dataset has been obtained from the UCI data repository and downloaded through Kaggle. The dataset contains a survey from 857 women and contains 33 questions (features) pertaining to cervical cancer diagnosis. The features include age, first sexual intercourse, diagnosis of different STDs, smoking habits, hormonal contraceptives used among many others.

5. Executive Summary

To assess the risk factors, 35 features were considered directly from the dataset or derived from it. This was a classification problem and the following classification models were used:

- Logistic Regression
- Random Forest Classifier
- K-Nearest Neighbor (KNN)
- Support Vector Machines (SVM)

The model evaluation was done after splitting the data into a 80-20 train-test split. Cross-validation and hyperparameter tuning was done in order to tune the Random Forest Classifier. The Random Forest Classifier after tuning gave a very high recall score of 0.994. Since this is clinical/medical data, we would be more concerned with recall scores than accuracy scores.

6. Data Cleaning & preprocessing

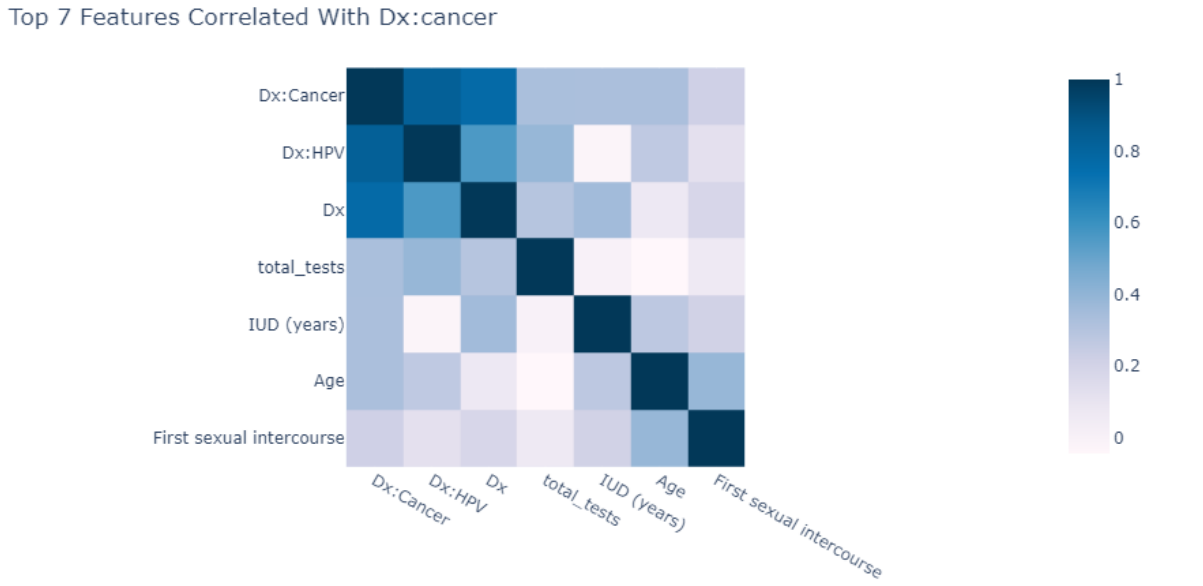
The dataset contained 858 rows and 35 columns. Out of the 35 columns, there were 2 columns with more than 90% missing values. All the other columns had moderate number of missing values. The simple imputer was used to impute the missing NaN values with the median. 3 new columns viz. age category, total STDs and total tests have been created in addition. The age category column is an ordinal category which divides the women into different age categories viz. children, teens, 20s, 30s, 40s ,50s and seniors. Previous medical history data has shown that diagnosis of cervical cancer is more prevalent in women above their 20s than in children.

Class Imbalance – There was a major class imbalance in the target variable, “Dx: Cancer”. Only ~2.1% of the respondents were classified as diagnosed with cancer out of the total respondents. This problem was solved by using oversampling methods like ADASYN.

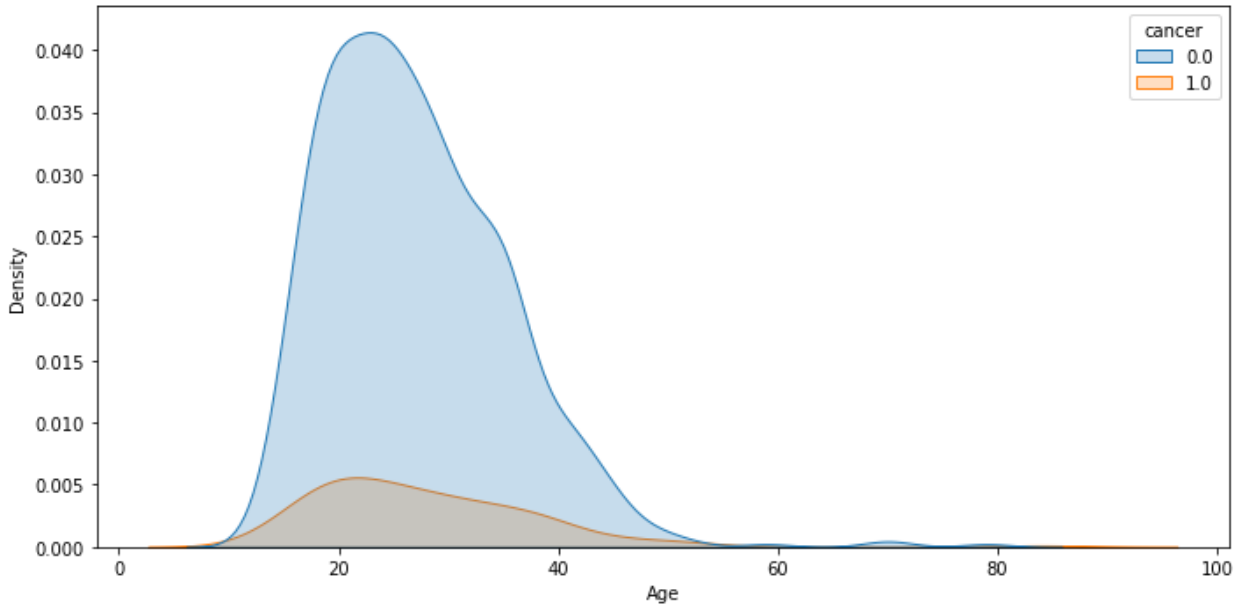
7. EDA

EDA was done before modeling on several features which were the key risk factors for cervical cancer. Python’s Matplotlib, Seaborn and Plotly were used to visualize the preliminary relationship between the features.

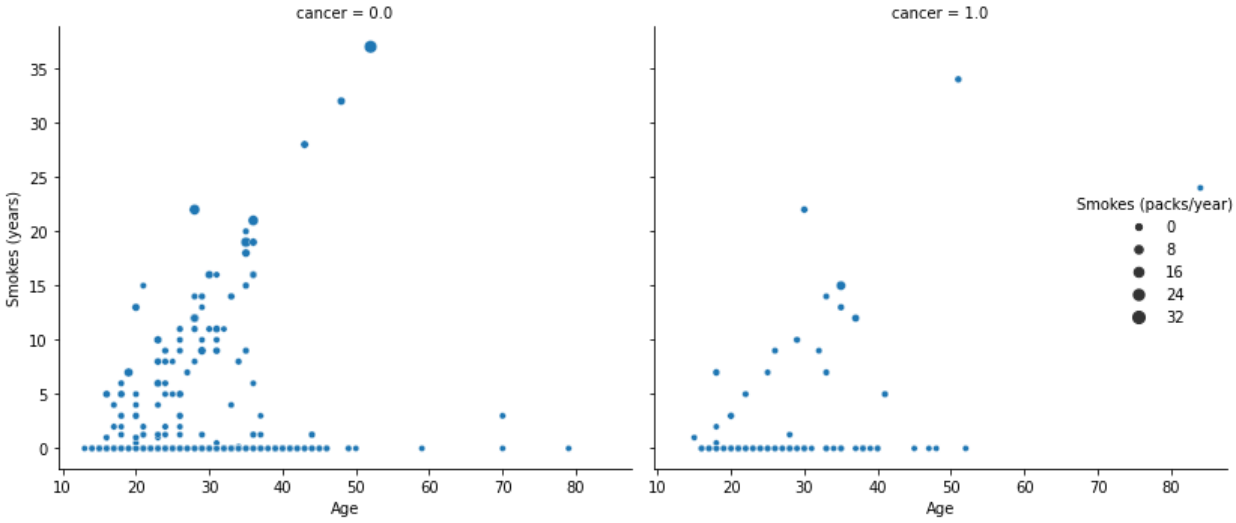
Top 7 features – I did a preliminary EDA to create a heatmap for the top 7 features that were associated with the diagnosis of cancer. From the heatmap it can be observed that the diagnosis of cancer is strongly related to the diagnosis of HPV or CIN and moderately related with age.



Age vs Cancer – I did a Kdeplot to show the density distribution of the diagnosis of cancer across various ages. It can be observed here that majority of the positive diagnoses is seen among women between 20 and 50 years of age.

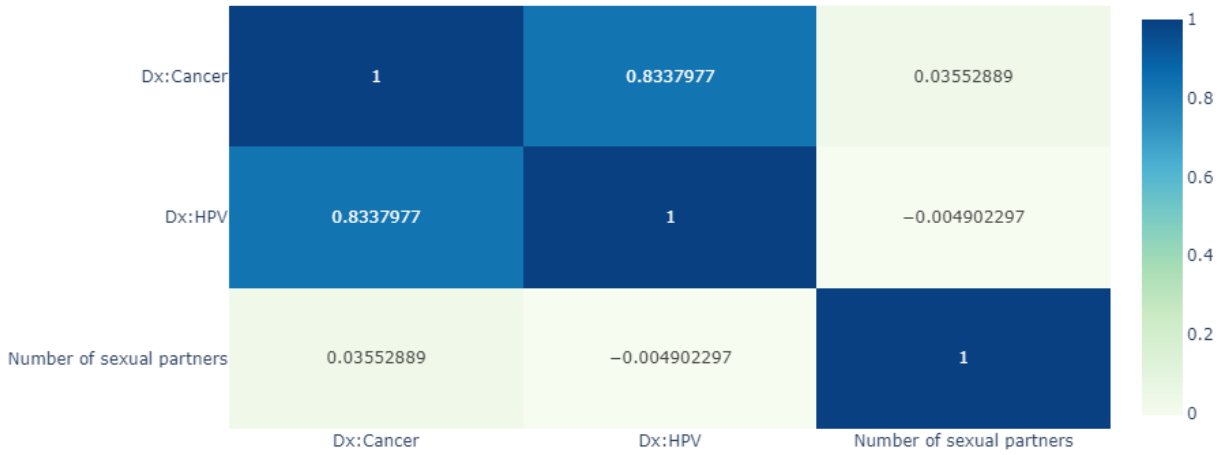


Smokes(packs/yr) vs Age vs Cancer – I did a relplot to do a relational visualization between how may packs smoked per year based on their age groups and how it caused positive diagnosis.

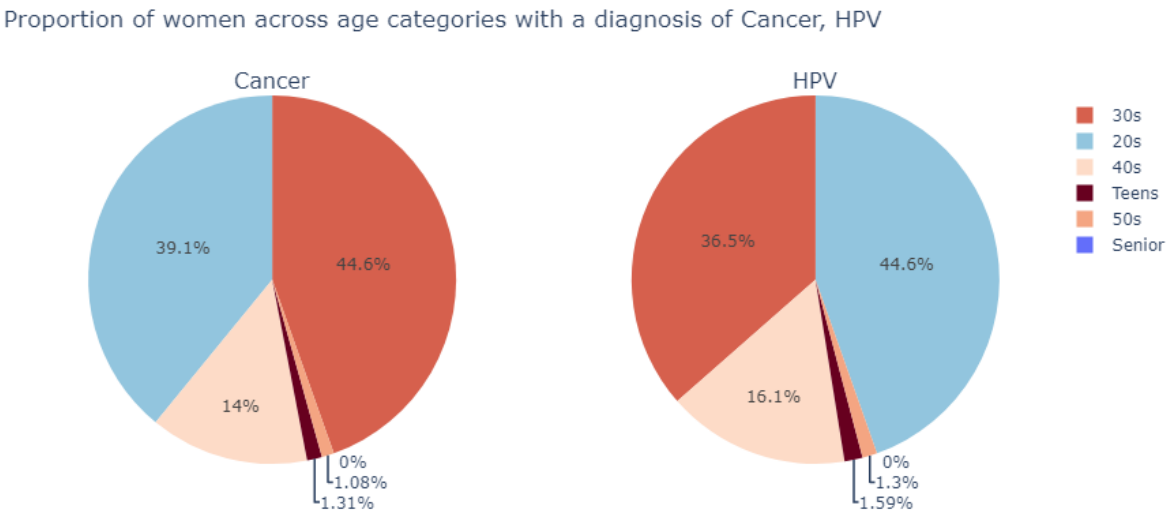


Cancer vs HPV – A correlation matrix and heatmap was done for cancer against the diagnosis of HPV (Human Papilloma Virus) infection. It can be observed that both HPV

and cancer are highly correlated.



Age category vs HPV and Cancer – Two separate pie charts were made to show how age categories varied with the diagnosis of cancer vs diagnosis of HPV.



8. Modeling

This is a classification problem in supervised learning. The following classification models were used.

- Logistic Regression
- Random Forest Classifier
- K-Nearest Neighbor (KNN)

- Support Vector Machines (SVM)

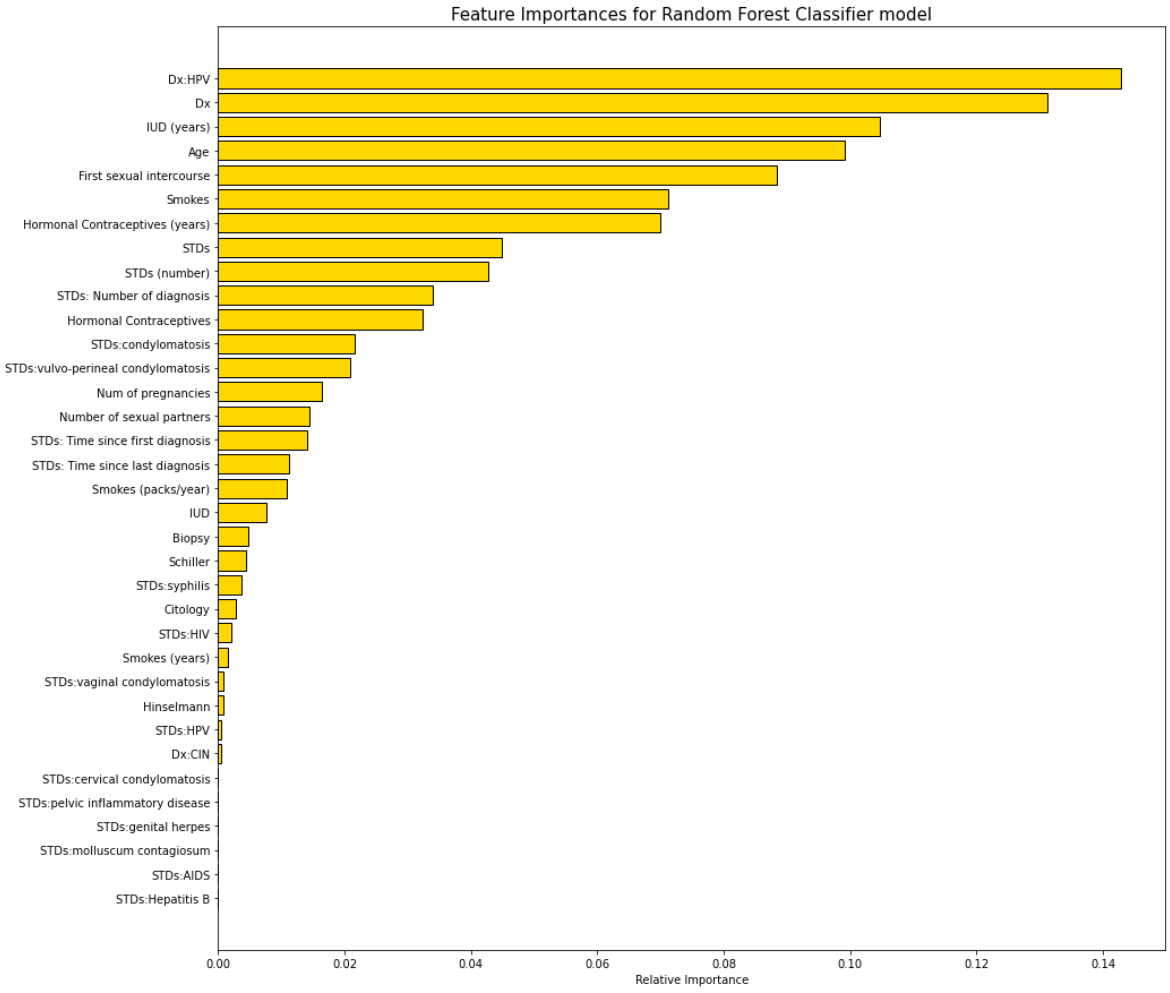
The model performance evaluation by training and validating can lead to overfitting. So the evaluation of the model dataset was split into separate train and test set. K-fold cross validation was used to randomize the data more. A confusion matrix was generated to calculate the accuracy scores, recall scores, precision scores and F1 scores.

	Classifier Name	Accuracy Score	Precision Score	Recall Score	F1 Score
0	LogisticRegression	0.991071	0.991235	0.991071	0.991074
1	RandomForestClassifier	0.991071	0.991091	0.991071	0.991072
2	KNeighborsClassifier	0.961310	0.964200	0.961310	0.961314
3	SupportVectorClassifier	0.997024	0.997042	0.997024	0.997024

From the two figures above, it is evident that except the K-Neighbors Classifier, all the other 3 models did well in predicting the chances of being diagnosed for cervical cancer. Since this is a clinical dataset, we will look at the metric of recall scores. Recall scores measure the correctly positive predicted outcomes of the total number of positive outcomes. This is a metric which should be highly considered while selecting the best models. This is because in the context of diagnosing the presence of cervical cancer, we want to lower the number of false negative cases as much as possible (Actual positive cases labelled as negative). If the false negatives numbers increase, this is concerning as the patient wouldn't be able to know about the disease without receiving proper treatment at the early stages.

9. Hyperparameter Tuning

In Machine learning, hyperparameter tuning refers to the selection of a set of optimal hyperparameters for learning a machine learning model. This is done using GridSearchCV. We tried to run hyperparameter tuning on the Random Forest Classifier model since that model had good recall score. The figure below shows the importance of different features after hyperparameter tuning.



It is evident that the diagnosis of Human Papilloma Virus (HPV) followed by diagnosis of other STDs, Intrauterine devices (IUD) and age are the most important features that need to be considered while looking to detect cervical cancer at early stages

10. Conclusion

This study was aimed at assessing the major risk factors of cervical cancer which is the second most occurring cancer in women. The knowledge of more important risk factors will help in early detection and diagnosis of cervical cancer. This can lead to potential treatment of diagnosed individuals and increase the chances of survival of that

individual. According to our modeling, any one of logistic regression, support vector machines or random forest classifier can be used to assess the potential risk factors. Since this is a clinical dataset, we want to totally reduce the number of people being false diagnosed as negative when actually, they are positive. That is why recall score of the different models should be primarily checked for selection of models. Feature importance scores were generated, and this showed that the diagnosis of Human Papilloma Virus (HPV) is the most important that should give the red flag for being diagnosed by cervical cancer. Other than this, some of the other factors that hold considerable importance as risk factors are diagnosis of other STDs, IUDs, age, smoking, time of first sexual intercourse as well as hormonal contraceptives.

11. Future studies

Cervical cancer is a very important gynecological disease and knowledge of its risk factors should help in its early diagnosis and survival. Future studies should be aimed at looking at the relationship between various socio-economic factors and the diagnosis of cervical cancer. Another area where future studies should be hugely successful is if gene expression and RNA seq data can be used for modeling and predicting the diagnosis of cervical cancer.