Report on Risk Classification of Cervical Cancer

1. Introduction

1.1 Topic Chosen for Software Engineer, ML code challenge:

I have Chosen the topic of Cervical cancer risk classification because there has been a lot of research going on in this area and doctors are learning about the tumor, finding a lot of ways to detect and treat the women in a better way who are diagnosed with this type of tumor [1].

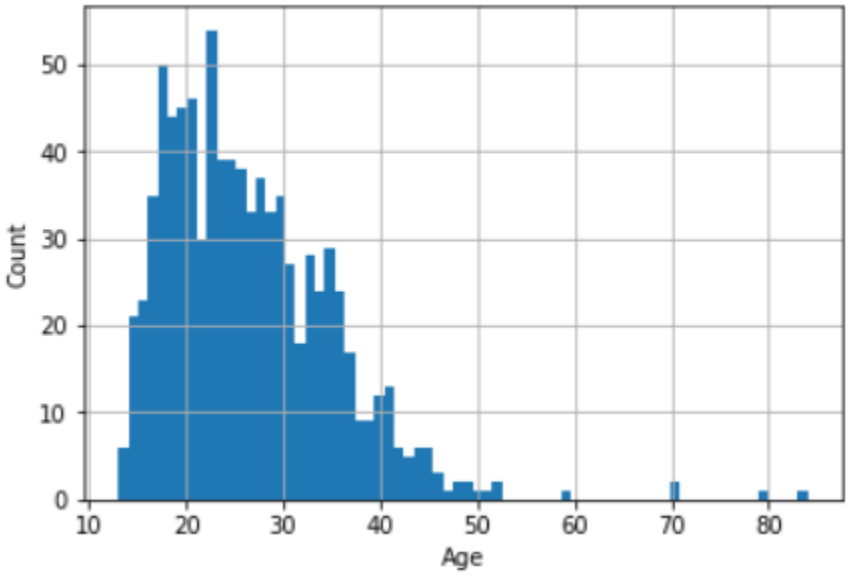
1.2 Dataset Collection: The dataset obtained for this challenge is from Kaggle [2]. The cervical cancer dataset consists of 32 risk factors and 4 target variables (Hinselmann, Schiller, Citology, Biopsy) which are the tests to detect the cervical cancer.

2. Data Analysis

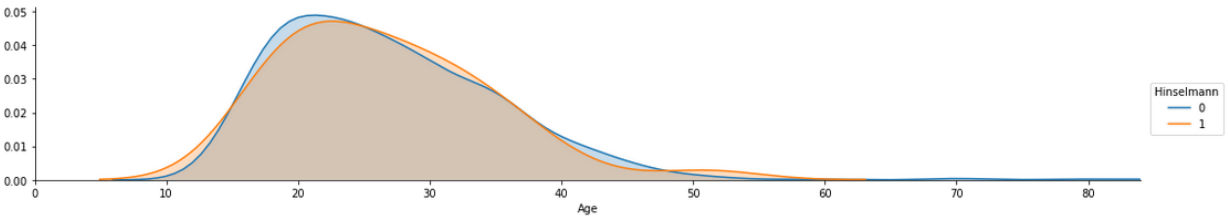
Many data visualizations are done to understand the statistics of the data. Following are some of the statistics of the dataset.

2.1 Age of the women facing the risk:

The minimum age of the women developing the risk of cervical cancer is 26. Distribution of ages is as shown below:

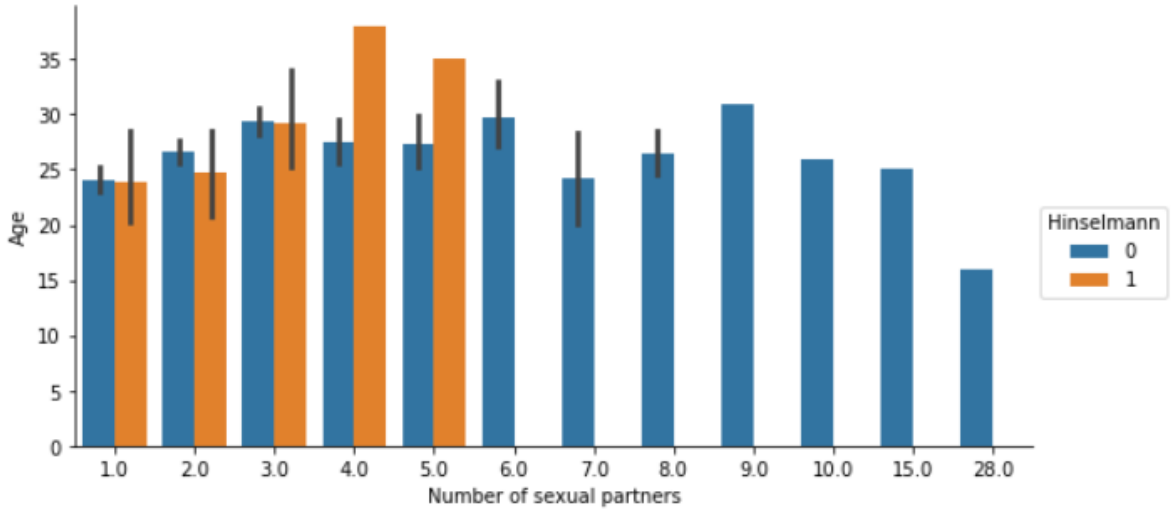


The density plot against the Age and the test indicate that the women with the age in range of 20 to 35 have the highest chances of developing the risk of cervical cancer. The peaks at age of 50 and the further extension of the density plot indicate that some of the women face the risk of cervical cancer even at that age. (Below is the graph only for Hinselmann, please refer code file for the remaining test plots).



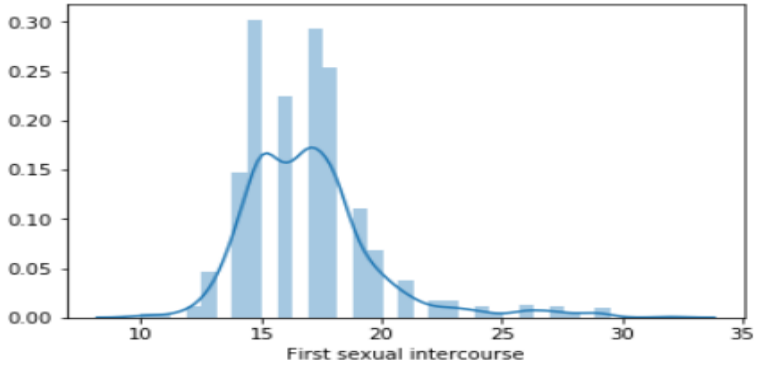
2.2 Age and Number of sexual partners:

As age of women increases, the number of sexual partners might increase which in turn increases the possibility of facing the risk. Even the resources here: https://www.mayoclinic.org/diseases-conditions/cervical-cancer/symptoms-causes/syc-20352501 supports this statement. From the below we can see that the women at the age of 30 to 35 having the number of sexual partners of 4 or 5 is having the higher risk of developing cervical cancer.

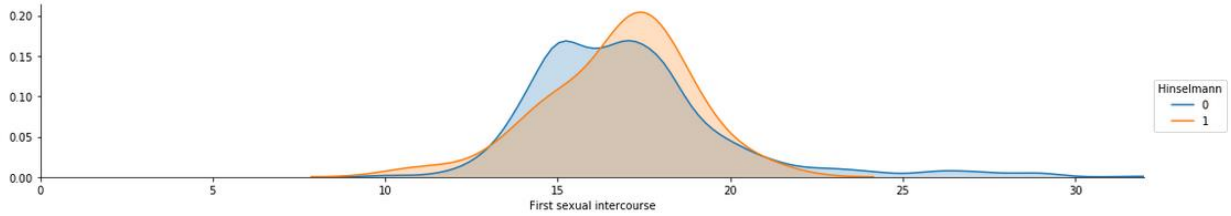


2.3 First Sexual intercourse:

Mean age of the women whose first sexual intercourse is 16. The plot below gives an intuition that most of the women started first sexual intercourse between the age 15 to 20.

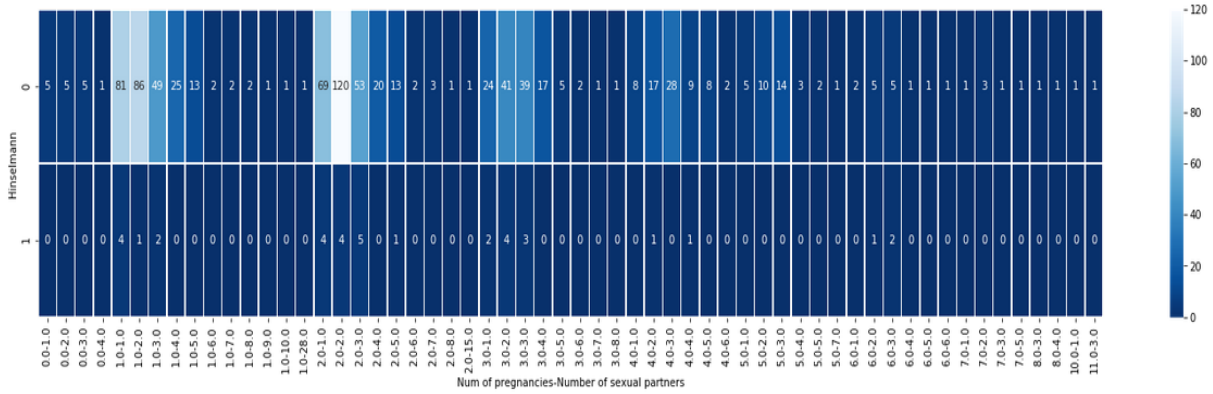


The density plots show that the women who started the Sexual intercourse at the age of 15 to 20 are having the high chances of developing the risk of cancer. Some of the resources online https://www.mayoclinic.org/diseases-conditions/cervical-cancer/symptoms-causes/syc-20352501 says that the early sexual activity increases the risk of developing cancer.



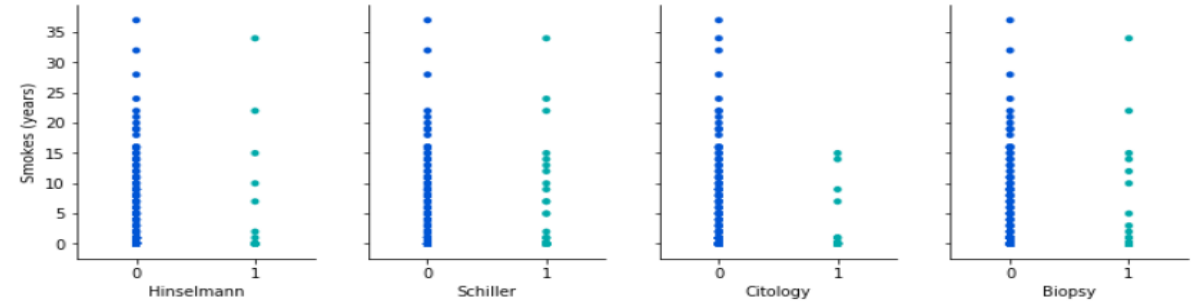
2.4 Number of Pregnancies:

Data is analyzed against the number of pregnancies and the number of sexual partners. It is found that 4% of the women having number of sexual partners greater than 2 with number of pregnancies greater than 3 are facing the risk of cancer. The below is the chart for number of women facing the risk.



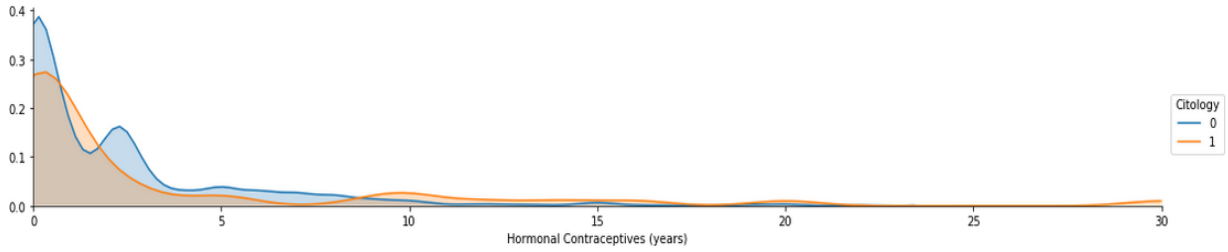
2.5 Risk for Women Who Smokes:

8% of women are smoking for more than 5 years and among them 16 % are the developing the risk of facing the cancer. The tests like schiller and biopsy can detect these types of factors causing cancer. The studies like https://www.mayoclinic.org/diseases-conditions/cervical-cancer/symptoms-causes/syc-20352501 shows that smoking is one the key risk factor of developing cervical cancer.



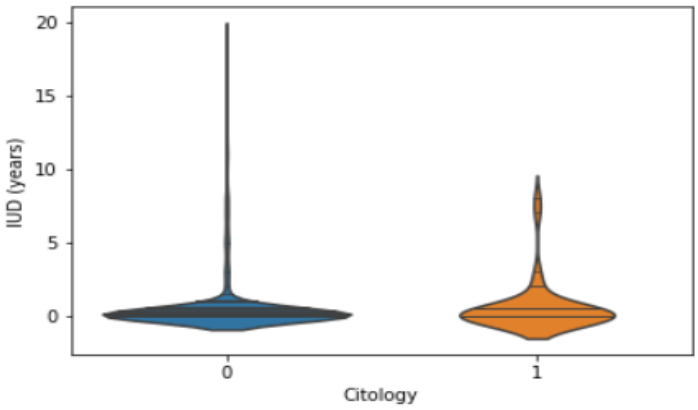
2.6 Women using Hormonal Contraceptives:

Many of the women reported that they are using the Hormonal contraceptives from minimum of 0 to 30 years. Women who have taken the harmonal contraceptives for more than five years have the higher risk of cervical cancer. The small peaks at the age 10 & 15 and also more than 15 shows even higher risk than women who have not used the contraceptives. The resources here: https://www.cancer.gov/about-cancer/causes-prevention/risk/hormones/oral-contraceptives-fact-sheet#q6 shows the studies supporting the statement that women face higher risk of cervical cancer if they used hormonal contraceptives.



2.7 Women using IUD:

The distribution of zero (negative test result) shows that the IUD used for any number of years has less chance of facing risk of cancer. Latest studies say that Women Who Use IUDs May Have Lower Risk of Cervical Cancer. Resource: https://www.livescience.com/60881-iuds-may-lower-risk-of-cervical-cancer.html.



2.8 Final statistics of Cervical Cancer Dataset:

a. Mean age of the women facing the risk of cancer is 26. Women in age groups of 20 to 35 have the higher risk of developing the risk.
b. As the number of sexual partners increase, the chances of developing the risk of cancer are higher.

c. Women started the early sexual activity between the age of 15 to 20 have the higher chances of facing the risk of cancer. Mean of the women with the first sexual activity is 16.
d. Women who have full term pregnancies greater than three have the higher risk of getting the cancer.
e. 16% of the women who are smoking for more than 5 years are facing the problem of cervical cancer.
f. 18% of the women stated that they are using hormonal contraceptives. Women using hormonal contraceptives for more than 5 years are having the higher probability of getting the cancer.
g. 8% of the women who used hormonal contraceptives are facing the problem of cervical cancer while 5% of women who used IUD are having the problem of cervical cancer.

3. Building a Predictive Model:

3.1 Features:

After the analysis, I have dropped some of the feature columns because of their correlation with other feature columns. Please refer Table 2 for final Set of features used for training and building a model.

3.1 Target Variable:

There are 4 target variables in the dataset. For building the model, I have considered a new variable cervical_cancer which is the sum of the 4 target variables. That is

Cervical_cancer = Hinselmann + Schiller + Citology + Biopsy

Which gives five different values ranging from 0 to 4. These values represent the level of risk i.e. 0 represent no risk and 4 represents higher level of risk. So, the final class/target variables are 0, 1, 2, 3, 4.

Table 1. Class Imbalance:

Class/Target Variable	Total Count of Class/Target Variable
0	756
1	41
2	22
3	33
4	6

As there are only 6 data points in the class 4, it is difficult to train and test the model. So, I have used the oversampling technique to overcome the problem of imbalance using SMOTE.

3.3 Results and Discussion:

I have created two different models, one with all the 30 features and the other with just 26 features obtained after data analysis. In this section, I am going to discuss about the results of both the models and compare them. In both the models, I have used **SMOTE** for oversampling and a **5-StratifiedKFold cross validation** for finding the best model.

Table 2. Features considered in both models

Baseline Model Features	**26-Features Model**
Age	Age
Number of sexual partners	Number of sexual partners
First sexual intercourse	First sexual intercourse
Num of pregnancies	Num of pregnancies
Smokes	
Smokes (years)	Smokes (years)
Smokes (packs/year)	Smokes (packs/year)
Hormonal Contraceptives	
Hormonal Contraceptives (years)	Hormonal Contraceptives (years)
IUD	
IUD (years)	IUD (years)
STDs	
STDs (number)	STDs (number)
STDs:condylomatosis	STDs:condylomatosis
STDs:cervical condylomatosis	STDs:cervical condylomatosis
STDs:vaginal condylomatosis	STDs:vaginal condylomatosis
STDs:vulvo-perineal condylomatosis	STDs:vulvo-perineal condylomatosis
STDs:syphilis	STDs:syphilis
STDs:pelvic inflammatory disease	STDs:pelvic inflammatory disease
STDs:genital herpes	STDs:genital herpes
STDs:molluscum contagiosum	STDs:molluscum contagiosum
STDs:AIDS	STDs:AIDS
STDs:HIV	STDs:HIV
STDs:Hepatitis B	STDs:Hepatitis B
STDs:HPV	STDs:HPV
STDs: Number of diagnosis	STDs: Number of diagnosis
Dx:Cancer	Dx:Cancer
Dx:CIN	Dx:CIN
Dx:HPV	Dx:HPV

Table 3. Results of the oversampled dataset:

	Baseline Model		26-Feature Model	
Classifiers Used	SVM	Random Forest	SVM	Random Forest
Accuracy	0.84	0.95	0.84	0.95
Precision	0.86	0.94	0.86	0.95
Recall	0.85	0.95	0.84	0.94
F1-Score	0.83	0.95	0.84	0.95

Table 4. Results of the non-oversampling dataset:

	Baseline Model		26-Feature Model	
Classifiers Used	SVM	Random Forest	SVM	Random Forest
Accuracy	0.88	0.87	0.88	0.87
Precision	0.77	0.8	0.78	0.81
Recall	0.88	0.87	0.88	0.86
F1-Score	0.83	0.83	0.83	0.83

Table 3 shows that, there is not much change in the results even with lesser number of features. But with oversampling, there is an increase in accuracy and F1 score of the models as shown in Table 4.

3.4 Predictions on Raw Data:

Finally, Random Forest is chosen for training and predicting the raw data. Some of the examples are mentioned in Table 5.

Table 5. Prediction on Raw Data.

Raw Data	Target Prediction
→Women with an age of 35 →Number of sexual partners:5 →First sexual intercourse:11 →Num of pregnancies:2 →Smokes (years):15, →Smokes (packs/year):15 →Hormonal Contraceptives (years): 0 and →all the other features are 0	[**4**] According to our findings, Women with more number of sexual partners, early sexual activity, smoking for more than 5 years all indicate the strong factors for developing the risk of cervical cancer
→Women with an age of 19 →Number of sexual partners:1 →First sexual intercourse:17 →Num of pregnancies:1 →Smokes (years):1, →Smokes (packs/year):3.4 →Hormonal Contraceptives (years): 0 and →all the other features are 0	[**0**] According to the data analysis, the women under age 20 are less like to develop the risk of cervical cancer.
→Women with an age of 48 →Number of sexual partners:2 →First sexual intercourse:15 →Num of pregnancies:2 →Smokes (years):0 →Smokes (packs/year):0, →Hormonal Contraceptives (years): 0.5 →IUD (years):19 →STDs:genital herpes:1 and →all the other features are 0	[**1**] This level of risk is acceptable because this women is of age 48, had early sexual activity. Despite used IUD for 19 years, the risk is very low. Because, based on the findings from the data analysis, women who used IUD has less risk than women who use hormonal contraceptives.

References:

1. Link to latest research on cervical cancer: https://www.cancer.net/cancer-types/cervical-cancer/latest-research
2. Link to Kaggle cervical cancer dataset: https://www.kaggle.com/loveall/cervical-cancer-risk-classification