1. PROBLEM DEFINITION:

Binary Classification on Predicting Cervical Cancer

a. Client:

Medical staff:

Cervical Cancer one of the most cancer types which females face off, and diagnosing the cancer in early stages are extremely important to cure the disease. Thus, Medical Staff's early detection of Cancer makes huge different on the treatment phase.

b. Data Set:

1) The dataset was collected at 'Hospital Universitario de Caracas' in Caracas, Venezuela.

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

2) The dataset comprises demographic information, habits, and historic medical records of 858 patients.

3) Several patients decided not to answer some of the questions because of privacy concerns (missing values).

4) Data set has 36 features and 858 data points.

5) Since target variable ('Dx:Cancer') consists of 18 positive samples (1) and 840 negatives (0), the data set is extremely **imbalanced.**

2. DATA WRANGLING:

Features:

(int) Age

(int) Number of sexual partners

(int) First sexual intercourse (age)

(int) Num of pregnancies

(bool) Smokes

(bool) Smokes (years)

(bool) Smokes (packs/year)

(bool) Hormonal Contraceptives

(int) Hormonal Contraceptives (years)

(bool) IUD (intrauterine device)

(int) IUD (years)

(bool) STDs (Sexually transmitted disease)

(int) STDs (number)

(bool) STDs:condylomatosis

(bool) STDs:cervical condylomatosis

(bool) STDs:vaginal condylomatosis

(bool) STDs:vulvo-perineal condylomatosis

(bool) STDs:syphilis

(bool) STDs:pelvic inflammatory disease

(bool) STDs:genital herpes

(bool) STDs:molluscum contagiosum

(bool) STDs:AIDS

(bool) STDs:HIV

(bool) STDs:Hepatitis B

(bool) STDs:HPV

(int) STDs: Number of diagnosis

(int) STDs: Time since first diagnosis

(int) STDs: Time since last diagnosis

(bool) Dx:Cancer

(bool) Dx:CIN (Cervical intraepithelial neoplasia)

(bool) Dx:HPV (Human papillomavirus infection)

(bool) Dx

(bool) Hinselmann: target variable

(bool) Schiller: target variable

(bool) Cytology: target variable

(bool) Biopsy: target variable

	Age	Number of sexual partners	First sexual intercourse	Num of pregnancies	Smokes	Smokes (years)	Smokes (packs/year)	Hormonal Contraceptives	Hormonal Contraceptives (years)	IUD	IUD (years)	STDs	STDs (number)
0	18	4.0	15.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	15	1.0	14.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	34	1.0	NaN	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	52	5.0	16.0	4.0	1.0	37.0	37.0	1.0	3.0	0.0	0.0	0.0	0.0
4	46	3.0	21.0	4.0	0.0	0.0	0.0	1.0	15.0	0.0	0.0	0.0	0.0

STDs:condylomatosis	STDs:cervical condylomatosis	STDs:vaginal condylomatosis	STDs:vulvo-perineal condylomatosis	STDs:syphilis	STDs:pelvic inflammatory disease	STDs:genital herpes	STDs:molluscum contagiosum	STDs:AIDS	STDs:HIV
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

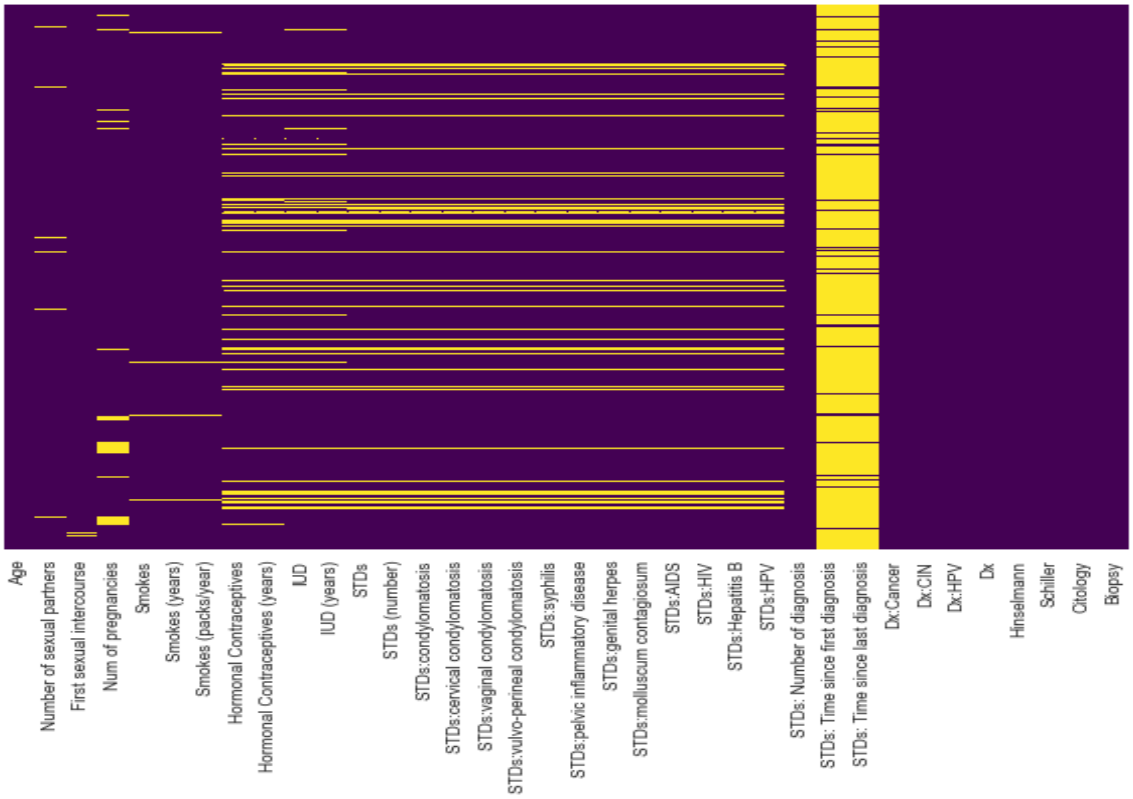
STDs:Hepatitis B	STDs:HPV	STDs: Number of diagnosis	STDs: Time since first diagnosis	STDs: Time since last diagnosis	Dx:Cancer	Dx:CIN	Dx:HPV	Dx	Hinselmann	Schiller	Citology	Biopsy
0.0	0.0	0	NaN	NaN	0	0	0	0	0	0	0	0
0.0	0.0	0	NaN	NaN	0	0	0	0	0	0	0	0
0.0	0.0	0	NaN	NaN	0	0	0	0	0	0	0	0
0.0	0.0	0	NaN	NaN	1	0	1	0	0	0	0	0
0.0	0.0	0	NaN	NaN	0	0	0	0	0	0	0	0

a. Missing Values:

26 out of 36 features have missing values in the data set. Missing values of each feature and the respective percentages are written below:

	Missing Values	**% of Total Values**
STDs: Time since last diagnosis	787	91.7
STDs: Time since first diagnosis	787	91.7
IUD	117	13.6
IUD (years)	117	13.6

Hormonal Contraceptives	108	12.6
Hormonal Contraceptives (years)	108	12.6
STDs:vulvo-perineal condylomatosis	105	12.2
STDs:HPV	105	12.2
STDs:Hepatitis B	105	12.2
STDs:HIV	105	12.2
STDs:AIDS	105	12.2
STDs:molluscum contagiosum	105	12.2
STDs:genital herpes	105	12.2
STDs:pelvic inflammatory disease	105	12.2
STDs:syphilis	105	12.2
STDs:cervical condylomatosis	105	12.2
STDs:vaginal condylomatosis	105	12.2
STDs:condylomatosis	105	12.2
STDs (number)	105	12.2
STDs	105	12.2
Num of pregnancies	56	6.5
Number of sexual partners	26	3.0
Smokes (packs/year)	13	1.5
Smokes (years)	13	1.5
Smokes	13	1.5
First sexual intercourse	7	0.8



Since ‘STDs: Time since last diagnosis’ and ‘STDs: Time since first diagnosis’ features have more than %91 percent missing values, they were dropped off.

```
df.drop(['STDs: Time since first diagnosis', 'STDs: Time since last diagnosis'], axis =1 , inplace = True)
```

For the rest numeric features which had missing values were applied mean statistical method.

```
df['STDs (number)'].fillna(np.ceil(df['STDs (number)'].mean()), inplace=True)
df['IUD (years)'].fillna(np.ceil(df['IUD (years)'].mean()), inplace=True)
df['Hormonal Contraceptives (years)'].fillna(np.ceil(df['Hormonal Contraceptives (years)'].mean()), inplace=True)
df['Smokes (packs/year)'].fillna(np.ceil(df['Smokes (packs/year)'].mean()), inplace=True)
df['Smokes (years)'].fillna(np.ceil(df['Smokes (years)'].mean()), inplace=True)
df['Number of sexual partners'].fillna(np.ceil(df['Number of sexual partners'].mean()), inplace=True)
df['Num of pregnancies'].fillna(np.ceil(df['Num of pregnancies'].mean()), inplace=True)
df['First sexual intercourse'].fillna(np.ceil(df['First sexual intercourse'].mean()), inplace=True)
```

But categorical features which had missing values were applied pd.get_dummies() function to create dummy variables for all categorical values including the missing value (‘NaN’).

Before doing that, we converted the string type of values to categorical ones and then applied the function.

```
for col in ['Smokes', 'Hormonal Contraceptives', 'IUD', 'STDs','STDs:cervical condylomatosis','STDs:condylomatosis',
            'STDs:vaginal condylomatosis','STDs:vulvo-perineal condylomatosis','STDs:syphilis','STDs:Hepatitis B',
            'STDs:pelvic inflammatory disease', 'STDs:genital herpes','STDs:molluscum contagiosum','STDs:AIDS',
            'STDs:HIV','STDs:HPV']:
    df[col] = df[col].astype('category')
```

```
df2 = pd.get_dummies(df[['Smokes', 'Hormonal Contraceptives', 'IUD', 'STDs','STDs:cervical condylomatosis',
                         'STDs:condylomatosis', 'STDs:vaginal condylomatosis',
                         'STDs:vulvo-perineal condylomatosis','STDs:syphilis','STDs:Hepatitis B',
                         'STDs:pelvic inflammatory disease', 'STDs:genital herpes',
                         'STDs:molluscum contagiosum','STDs:AIDS', 'STDs:HIV','STDs:HPV']], dummy_na = True)
df2.head()
```

After concatenating the new data set consisted of dummy features to the main data set, we dropped the features from which we produced the dummy ones from the main data set.

```
df.drop(['Smokes', 'Hormonal Contraceptives', 'IUD', 'STDs','STDs:cervical condylomatosis','STDs:condylomatosis',
         'STDs:vaginal condylomatosis','STDs:vulvo-perineal condylomatosis','STDs:syphilis','STDs:Hepatitis B',
         'STDs:pelvic inflammatory disease', 'STDs:genital herpes','STDs:molluscum contagiosum','STDs:AIDS',
         'STDs:HIV', 'STDs:HPV'], axis = 1, inplace = True)
```

The cleaned data set had 64 features and 848 data points and all the features consisted of numeric values.

	Age	Number of sexual partners	First sexual intercourse	Num of pregnancies	Smokes (years)	Smokes (packs/year)	Hormonal Contraceptives (years)	IUD (years)	STDs (number)	STDs: Number of diagnosis	Dx:Cancer	Dx:CIN	Dx:HPV	Dx	Hinselmann
0	18	4.0	15.0	1.0	0.0	0.0	0.0	0.0	0.0	0	0	0	0	0	0
1	15	1.0	14.0	1.0	0.0	0.0	0.0	0.0	0.0	0	0	0	0	0	0
2	34	1.0	17.0	1.0	0.0	0.0	0.0	0.0	0.0	0	0	0	0	0	0
3	52	5.0	16.0	4.0	37.0	37.0	3.0	0.0	0.0	0	1	0	1	0	0
4	46	3.0	21.0	4.0	0.0	0.0	15.0	0.0	0.0	0	0	0	0	0	0

Schiller	Citology	Biopsy	Smokes_0.0	Smokes_1.0	Smokes_nan	Hormonal Contraceptives_0.0	Hormonal Contraceptives_1.0	Hormonal Contraceptives_nan	IUD_0.0	IUD_1.0	IUD_nan
0	0	0	1	0	0	1	0	0	1	0	0
0	0	0	1	0	0	1	0	0	1	0	0
0	0	0	1	0	0	1	0	0	1	0	0
0	0	0	0	1	0	0	1	0	1	0	0
0	0	0	1	0	0	0	1	0	1	0	0

STDs_0.0	STDs_1.0	STDs_nan	STDs:cervical condylomatosis_0.0	STDs:cervical condylomatosis_nan	STDs:condylomatosis_0.0	STDs:condylomatosis_1.0	STDs:condylomatosis_nan
1	0	0	1	0	1	0	0
1	0	0	1	0	1	0	0
1	0	0	1	0	1	0	0
1	0	0	1	0	1	0	0
1	0	0	1	0	1	0	0

STDs:vaginal condylomatosis_0.0	STDs:vaginal condylomatosis_1.0	STDs:vaginal condylomatosis_nan	STDs:vulvo-perineal condylomatosis_0.0	STDs:vulvo-perineal condylomatosis_1.0	STDs:vulvo-perineal condylomatosis_nan	STDs:syphilis_0.0
1	0	0	1	0	0	1
1	0	0	1	0	0	1
1	0	0	1	0	0	1
1	0	0	1	0	0	1
1	0	0	1	0	0	1

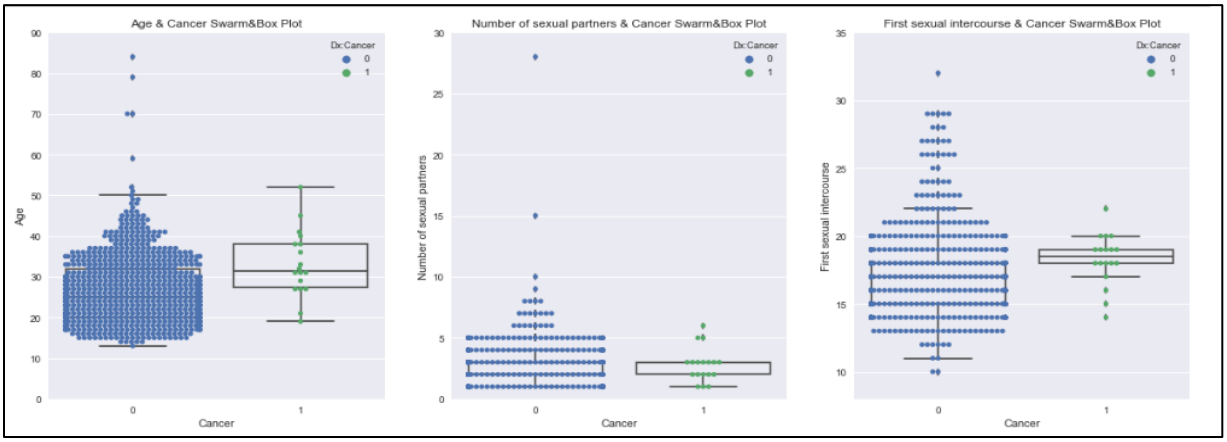
STDs:syphilis_1.0	STDs:syphilis_nan	STDs:Hepatitis B_0.0	STDs:Hepatitis B_1.0	STDs:Hepatitis B_nan	STDs:pelvic inflammatory disease_0.0	STDs:pelvic inflammatory disease_1.0	STDs:pelvic inflammatory disease_nan	STDs:genital herpes_0.0	STDs:genital herpes_1.0
0	0	1	0	0	1	0	0	1	0
0	0	1	0	0	1	0	0	1	0
0	0	1	0	0	1	0	0	1	0
0	0	1	0	0	1	0	0	1	0
0	0	1	0	0	1	0	0	1	0

STDs:molluscum contagiosum_0.0	STDs:molluscum contagiosum_1.0	STDs:molluscum contagiosum_nan	STDs:AIDS_0.0	STDs:AIDS_nan	STDs:HIV_0.0	STDs:HIV_1.0	STDs:HIV_nan	STDs:HPV_0.0
1	0	0	1	0	1	0	0	1
1	0	0	1	0	1	0	0	1
1	0	0	1	0	1	0	0	1
1	0	0	1	0	1	0	0	1
1	0	0	1	0	1	0	0	1

STDs:HPV_1.0	STDs:HPV_nan
0	0
0	0
0	0
0	0
0	0

3. EXPLORATORY DATA ANALYSIS (EDA)-DATA VISUALIZATION

a. Age, Number of Sexual Partners, First Sexual Intercourse vs. Cancer Graph:



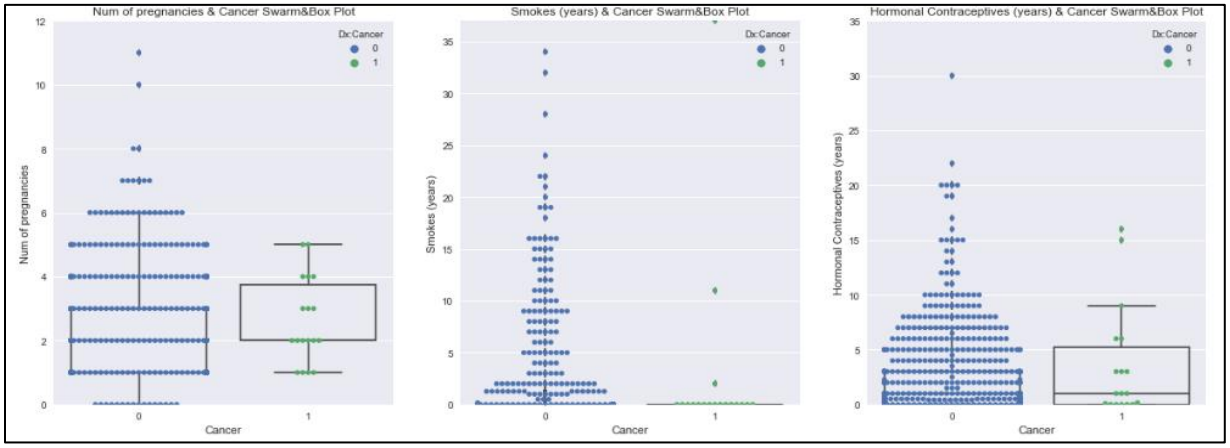
This graphic shows that;

1) Cancer diagnosed patient's age are cumulated between 27 to 42. Cancer patient's median age is higher than non-cancers.

2) Cancer diagnosed patient's number of sexual partners are cumulated between 1 to 5. Most of the patients have had either 5 or less partners.

3) Cancer diagnosed patient's first sexual intercourses are cumulated between 17 to 20. There is outlier even at 10. Cancer patient's median first sexual intercourse age is higher than non-cancer ones.

b. Number of Pregnancies, Smokes (Year), Hormonal Contraceptives vs. Cancer Graph:



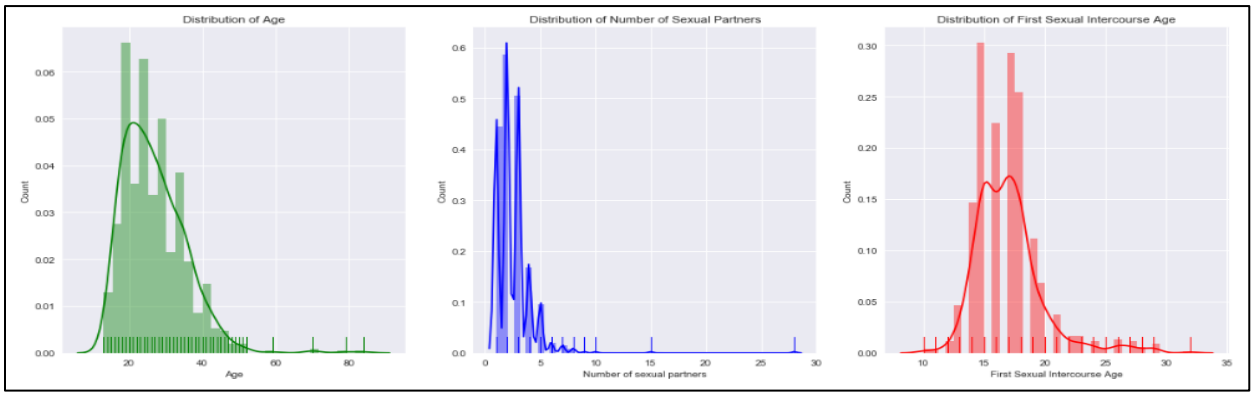
This graphic shows that;

1) Cancer patient's median number of pregnancies is higher than non-cancer.

2) Most of non-cancer patients smoke more than 3 years. Most of the cancer patients do not smoke.

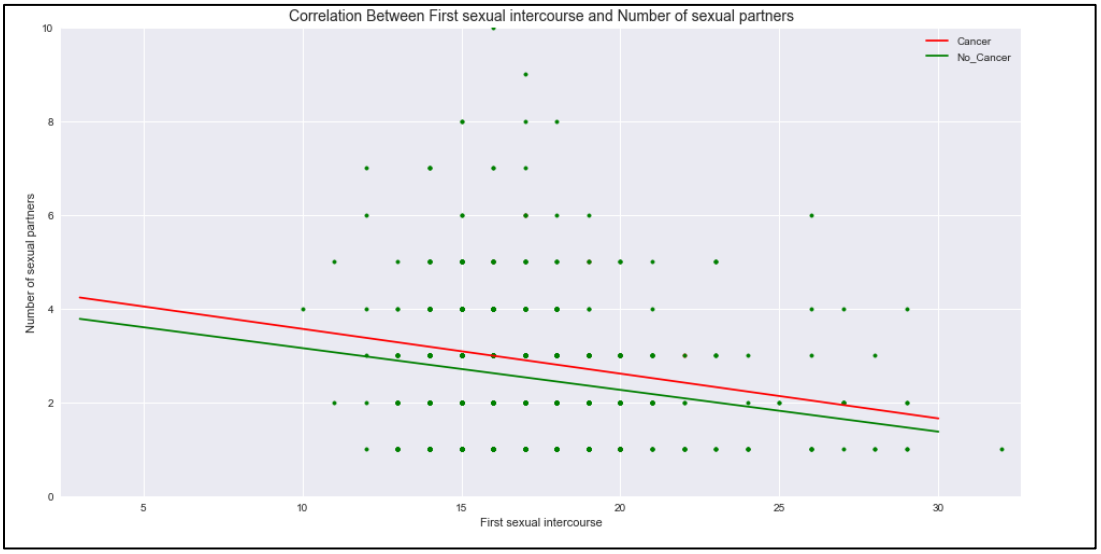
3) Most of the non-cancer patients use hormonal contraceptives. More than %50 of the cancer patients also use hormonal contraceptives.

c. Distribution of Age, Number of Sexual Partners and First Sexual Intercourse Graph:



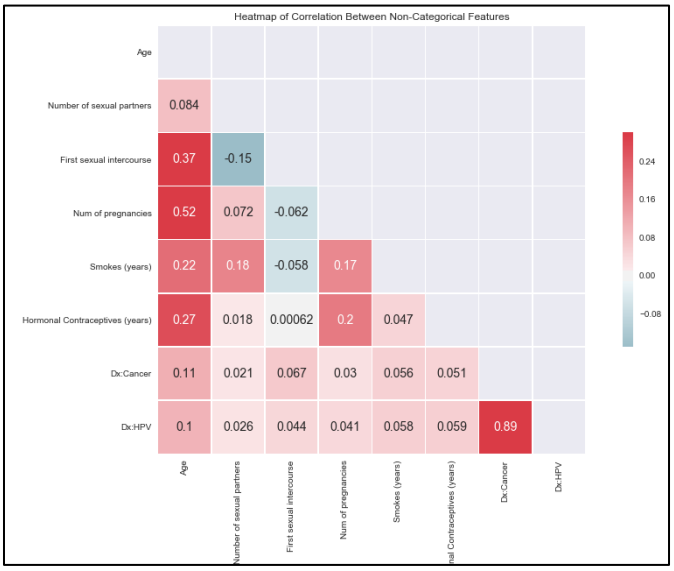
The graphic shows that all three features look like normally distributed but skewed to right. There are some outliers in all three features.

d. Correlation Between First Sexual Intercourse and Number of Sexual Partners Graph:



The graph shows that There is negative regression between First sexual intercourse and Number of sexual partners for both Cancer and Non-cancer diagnosed patients.

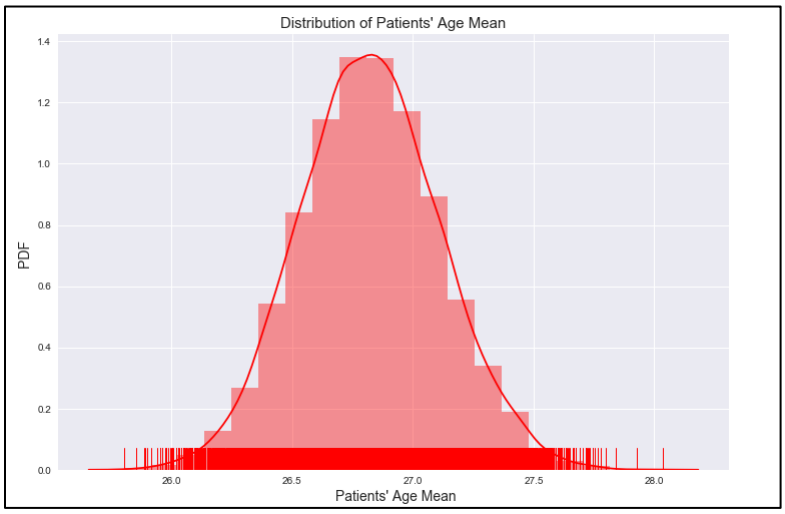
e. Correlation Between Non-Categorical Features Graph:



The graph shows that there is strong correlation between Cervical Cancer and Human papillomavirus infection (HPV) and the other correlations are stated below;

Age	0.1
First sexual intercourse	0.044
Number of Pregnancy	0.041
Smokes (year)	0.058
Hormonal Contras	0.059

f. Patients' Age Distribution Graph:



This graph shows that patients' age are normally distributed and average age of the patients is around 26-27 years old.

4. MACHINE LEARNING MODELS

This is a supervised binary classification problem. We are trying to predict whether a patient is cancer or not. We used Python Scikit Learn libraries to solve our problem. But since our data set is extremely imbalanced, we applied Synthetic Minority Oversampling Technique (SMOTE) to create more data points synthetically.

a. In the first stage, we split our data into training (%75) and test (%25) set then we used Logistic Regression with 5-Fold Cross Validation. To overcome the overfitting problem we used

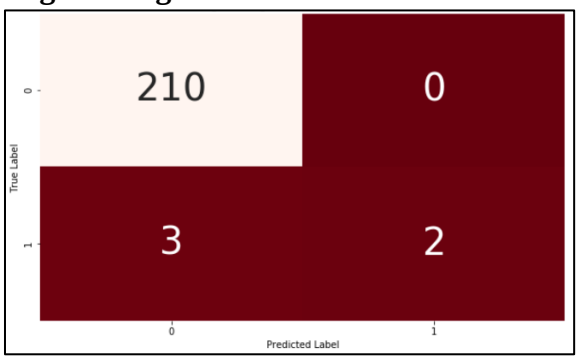
Logistic Regression with Grid Search L1 (Lasso) Hyper Parameter Tuning, Logistic Regression with Grid Search L2 (Ridge) Hyper Parameter Tuning, and finally Random Forest Classifier algorithms.

We applied SMOTE to the training data set and then used Logistic Regression and Random Forest Classifier algorithms to get best prediction.

b. In the second stage, we tweak the proportion of our training and test sets as (0.6/0.4) and applied almost the same methods and algorithms. As an evaluation metric we used Classification reports and Confusion Matrices.

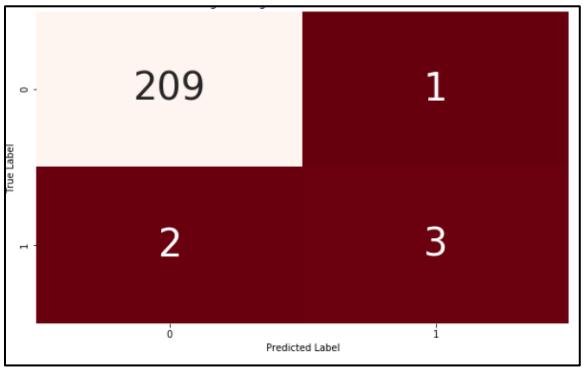
c. Train/Test Set Size Proportion is 0.75/0.25

1) Logistic Regression with 5-Fold Cross-Validation:



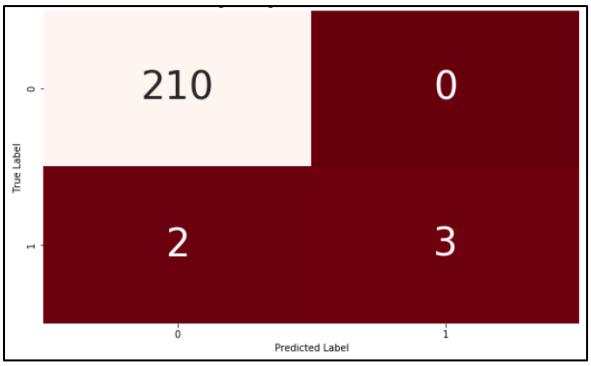
Despite Logistic Regression works well with the Non-Cancer patients, it misclassified 3 Cancer patients as Non-Cancer out of 5 patients with %40 prediction accuracy.

2) Logistic Regression with Grid Search CV (Lasso) Hyper Parameter Tuning:



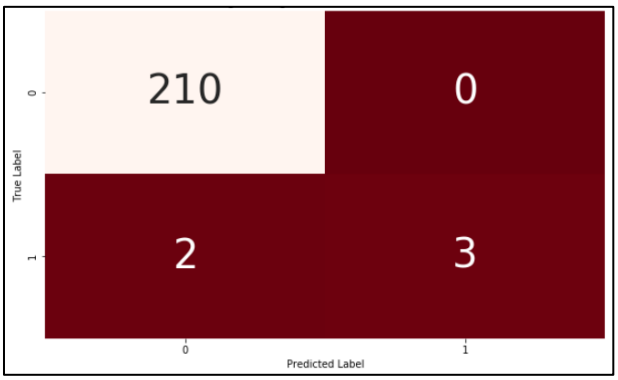
With the Lasso hyper parameter tuning our model worsen on Nan-Cancer patients and misclassified 1 patient but it predicted accurately one patient more than default Logistic Regression algorithm.

3) Logistic Regression with Grid Search CV (Ridge) Hyper Parameter Tuning:



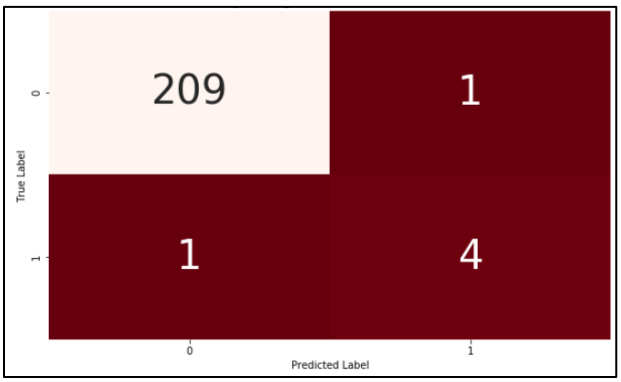
With the Ridge hyper parameter tuning our model outperformed Lasso and Nan-Cancer patients were predicted with %100 accuracy and but there was no changing on the recall, Cancer patients.

4) Random Forest Classifier:



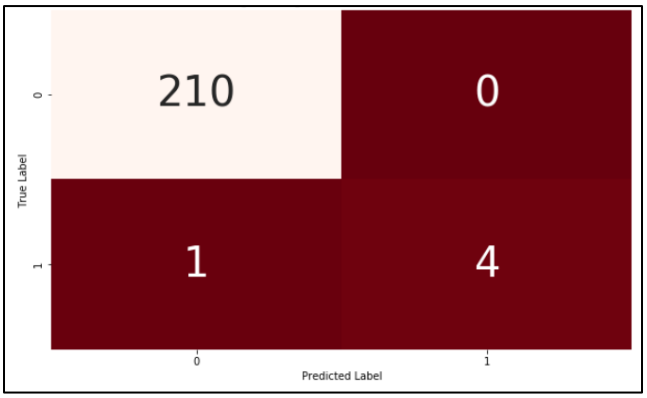
Random Forest Classifier was better than default Logistic Regression but there was no marginal changing comparatively with Lasso and Ridge.

5) Logistic Regression after SMOTE:



After SMOTE application, our default Logistic Regression model got better prediction than the previous models on Cancer patients but missed one non-cancer patients. The accuracy of prediction of Cancer patients increased up to %80.

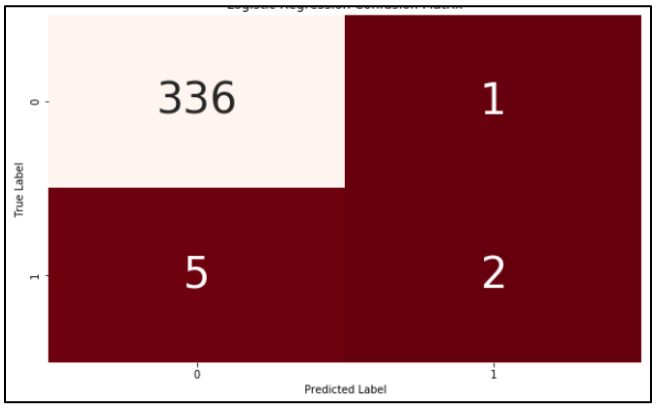
6) Random Forest Classifier after SMOTE:



We got the best prediction results with Random Forest Classifier after SMOTE application. Our model predicted non-cancer patients with %100 accuracy and for the Cancer patients with %80 accuracy.

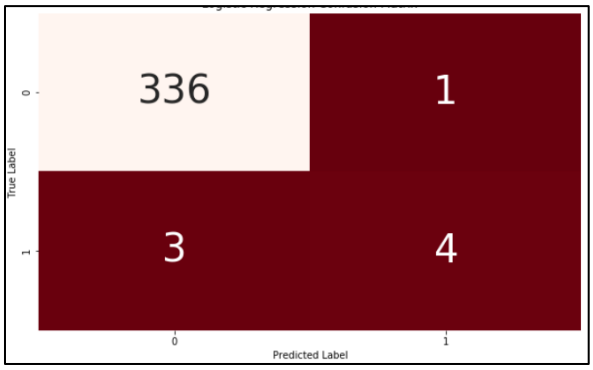
d. Train/Test Set Size Proportion is 0.60/0.40

1) Logistic Regression with 5-Fold Cross-Validation:



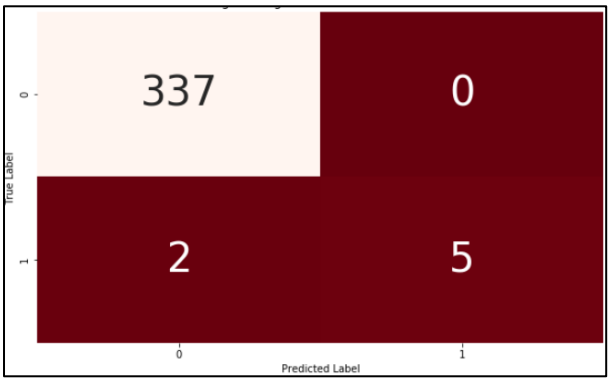
Despite Logistic Regression works well with the Non-Cancer patients, it misclassified 5 Cancer patients as Non-Cancer out of 7 patients with %29 prediction accuracy.

2) Logistic Regression with Grid Search CV (Lasso) Hyper Parameter Tuning:



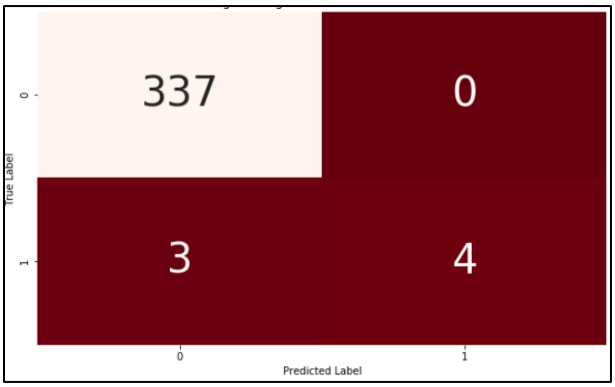
With the Lasso hyper parameter tuning, our model made better prediction and recall increased almost double from %20 to %57 and our model predicted 4 Cancer patients accurately out of 7 patients.

3) Logistic Regression with Grid Search CV (Ridge) Hyper Parameter Tuning:



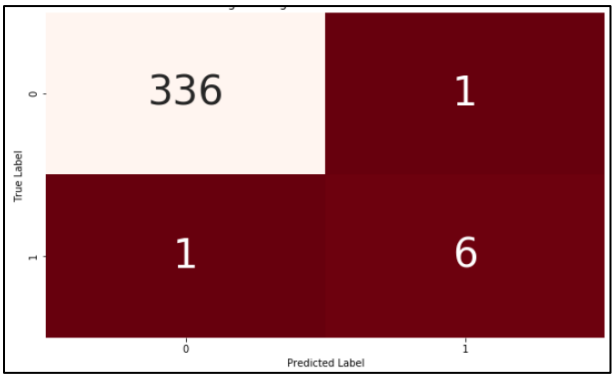
With the Ridge hyper parameter tuning our model outperformed Lasso and Nan-Cancer patients were predicted with %100 accuracy and Cancer patients prediction increased from 4 to 5 patients.

4) Random Forest Classifier:



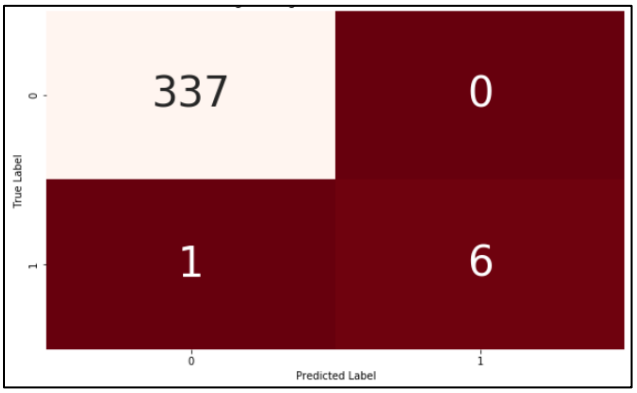
Random Forest Classifier worsen than Lasso and Ridge on Cancer patient but there was no changing on the non-cancer patients.

5) Logistic Regression after SMOTE:



After SMOTE application, our default Logistic Regression model got better prediction than the previous models on Cancer patients but missed one Cancer patient and non-cancer patient. The accuracy of prediction of Cancer patients increased up to %86.

6) Random Forest Classifier after SMOTE:



We got the best prediction results with Random Forest Classifier after SMOTE application. Our model predicted non-cancer patients with %100 accuracy and for the Cancer patients with %86 accuracy.

e. Model Comparison:

Train/Test = 0.75/0.25

Proportion (Train/Test)	SMOTE	DATA SET	Model/Application	Accuracy Score			Classification Report: Category	precision	recall	f1-score	Support
0.75/0.25	NO	848 Data Points 64 Features	Default Logistic Regression	Accuracy Score	Train	0.995	0	1.00	1.00	1.00	630
					Test	0.986	1	1.00	0.77	0.87	13
							0	**0.99**	**1.00**	**0.99**	**210**
							1	**1.00**	**0.40**	**0.57**	**5**
			Logistic Regression with 5 fold Cross Validation	Accuracy Score	Train	-	-	-	-	-	-
							-	-	-	-	-
					Test	0.986	**0**	**0.99**	**1.00**	**0.99**	**210**
							1	**1.00**	**0.40**	**0.57**	**5**
			Logistic Regression with Grid Search CV L1 Penalty	Accuracy Score	Train	0.986	0	1.00	1.00	1.00	630
					Test		1	0.92	0.85	0.88	13
							0	**0.99**	**1.00**	**0.99**	**210**
							1	**0.75**	**0.60**	**0.67**	**5**
			Logistic Regression with Grid Search CV L2 Penalty	Accuracy Score	Train	0.99	0	1.00	1.00	1.00	630
					Test		1	1.00	1.00	1.00	13
							0	**0.99**	**1.00**	**1.00**	**210**
							1	**1.00**	**0.60**	**0.75**	**5**
			Random Forest Classifier N-Estimator = 400	Accuracy Score	Train	-	0	-	-	-	-
					Test		1	-	-	-	-
							0	**0.99**	**1.00**	**0.99**	**210**
							1	**1.00**	**0.40**	**0.57**	**5**
	YES	1260 Data Points 64 Features	SMOTE with Logistic Regression	Accuracy Score	Train	0.999	0				
					Test	0.991	1				
							0	**1.00**	**1.00**	**1.00**	**210**
							1	**0.80**	**0.80**	**0.80**	**5**
			SMOTE with Random Forest Classifier N-Estimator = 400	Accuracy Score	Train	-	0	-	-	-	-
					Test		1	-	-	-	-
							0	**1.00**	**1.00**	**1.00**	**210**
							1	**1.00**	**0.80**	**0.89**	**5**

Random Forest Classifier is the best algorithm after SMOTE and Logistic Regression is the worst one for our problem.

Train/Test = 0.60/0.40

Proportion (Train/Test)	SMOTE	DATA SET	Model/Application	Accuracy Score			Classification Report: Category	precision	recall	f1-score	Support
0.60/0.40	NO	848 Data Points 64 Features	Default Logistic Regression	Accuracy Score	Train	0.994	0	-	-	-	-
					Test	0.985	1	-	-	-	-
							0	**0.99**	**1.00**	**0.99**	**337**
							1	**1.00**	**0.29**	**0.44**	**7**
			Logistic Regression with 5 fold Cross Validation	Accuracy Score	Train	-	-	-	-	-	-
							-	-	-	-	-
					Test	0.984	**0**	**0.99**	**1.00**	**0.99**	**337**
							1	**0.67**	**0.29**	**0.40**	**7**
			Logistic Regression with Grid Search CV L1 Penalty	Accuracy Score	Train	0.988	0	1.00	1.00	1.00	503
					Test		1	0.90	0.82	0.86	11
							0	**0.99**	**1.00**	**0.99**	**337**
							1	**0.80**	**0.57**	**0.67**	**7**
			Logistic Regression with Grid Search CV L2 Penalty	Accuracy Score	Train	0.994	0	1.00	1.00	1.00	630
					Test		1	1.00	1.00	1.00	13
							0	**0.99**	**1.00**	**1.00**	**337**
							1	**1.00**	**0.71**	**0.83**	**7**
			Random Forest Classifier N-Estimator = 400	Accuracy Score	Train	-	0	-	-	-	-
					Test		1	-	-	-	-
							0	**0.99**	**1.00**	**1.00**	**337**
							1	**1.00**	**0.57**	**0.73**	**7**
	YES	1006 Data Points 64 Features	SMOTE with Logistic Regression	Accuracy Score	Train	0.999	0				
					Test	0.994	1				
							0	**1.00**	**1.00**	**1.00**	**337**
							1	**0.86**	**0.86**	**0.86**	**7**
			SMOTE with Random Forest Classifier N-Estimator = 400	Accuracy Score	Train	-	0	-	-	-	-
					Test		1	-	-	-	-
							0	**1.00**	**1.00**	**1.00**	**337**
							1	**1.00**	**0.86**	**0.92**	**7**

17

Again, Random Forest Classifier is the best algorithm after SMOTE and Logistic Regression is the worst one for our problem.

5. CONCLUSIONS:

In our study we have used all necessary features (all the one left after the dropped ones) in our model. In our model, Random Forest Classifier showed the best performance after SMOTE in both proportions. Despite studying with the imbalanced data is very hard, we could manage to catch up %86

accuracy with 18 positive samples total. Hyper parameter tuning also showed us that using proper parameters also increases the accuracy of the algorithm.

6. RECOMMENDATIONS TO CLIENT:

After changing the proportion of Train and Test Set, our model's prediction accuracy almost increased up to %6 and 6 out of 7 patients are also predicted as Cancer correctly. If we had more Cancer patient samples in data set, we would have train our model better and get more accurate predictions. To said that we would recommend to the client to get more Cancer samples to have better predictions.

7. FUTURE WORK:

In this study we focused on proportion selection of train and test set, hyper parameter tuning and SMOTE. As a future study we will concentrate the other algorithms such as Ada Boost Classifier and Gradient Boost Classifier to see their performance with the imbalanced data set.