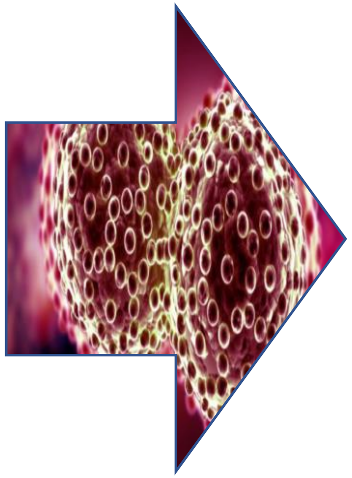
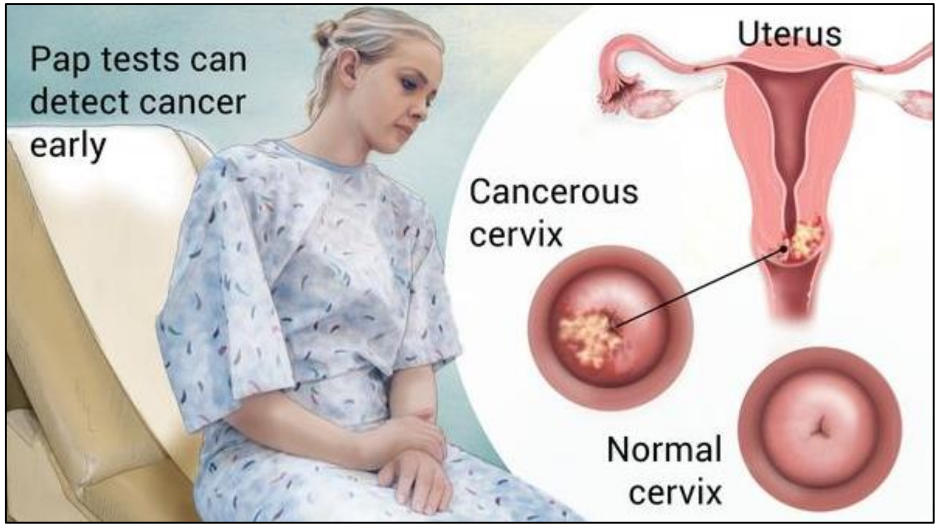
Mentor

A J Sanchez

Mustafa KADIOGLU
Cervical Cancer Risk Factors
Logistic Regression Capstone Project
Springboard Data Science Career Track April-2018 Cohort
github.com/mustafakadioglu

Problem Definition





Affecting over 500,000 women and resulting in approximately 275,000 deaths every year

Data Information

Instituto Autónomo
HOSPITAL UNIVERSITARIO DE CARACAS
El Hospital es del pueblo.

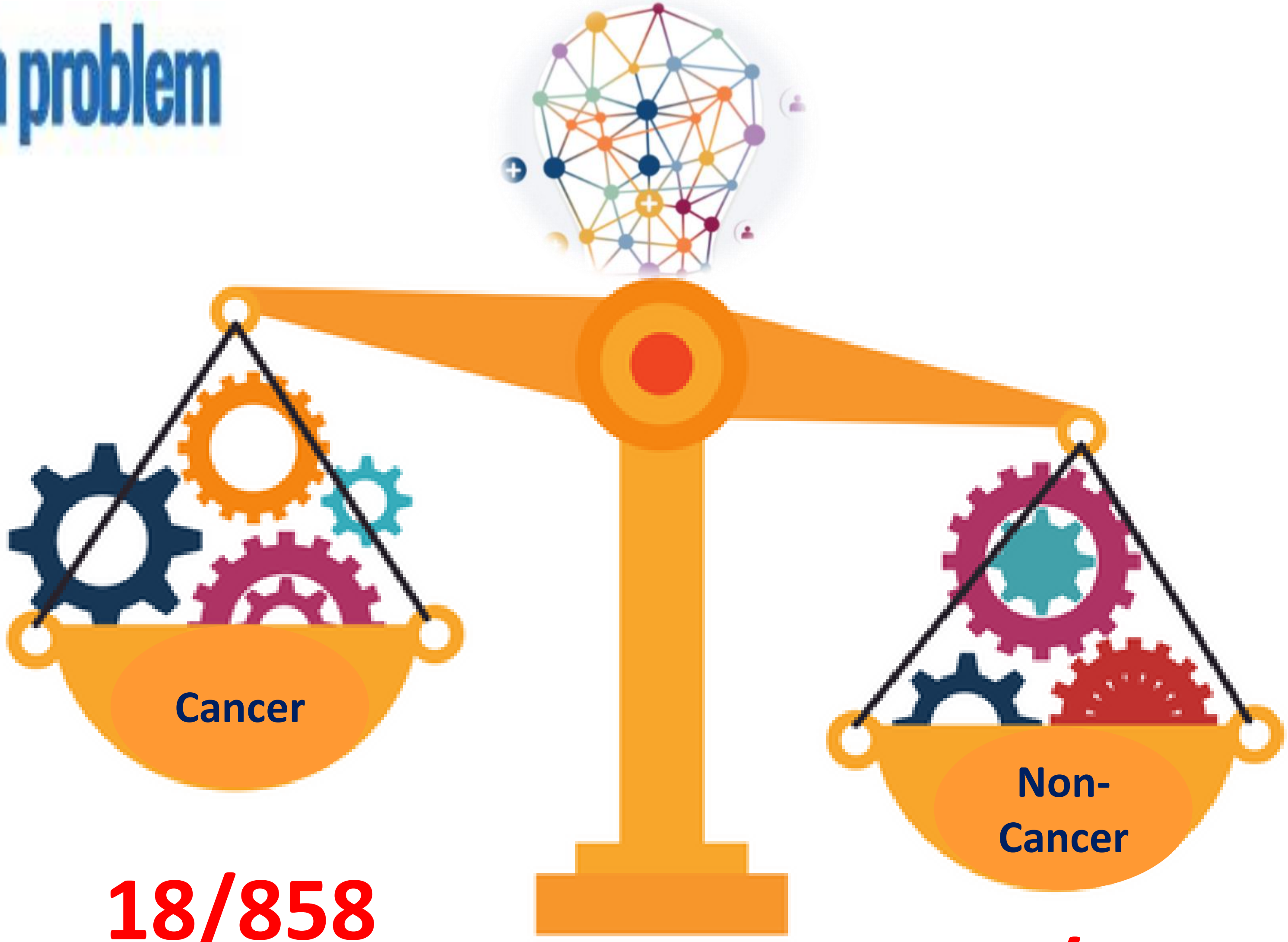
UCI
Machine Learning Repository
Center for Machine Learning and Intelligent Systems

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

858 data points and 36 features

Data Information

Imbalanced classification problem



18/858

840/858

Data Information

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

5 rows × 36 columns

Dx:HPV	Dx	Hinselmann	Schiller	Citology	Biopsy
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
1	0	0	0	0	0
0	0	0	0	0	0

Feature Engineering

Reading the data set

Replacing '?' with 'NaN'

```
df =pd.read_csv('risk_factors_cervical_cancer.csv', na_values = ['?'])
```

```
df.head(5)
```

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

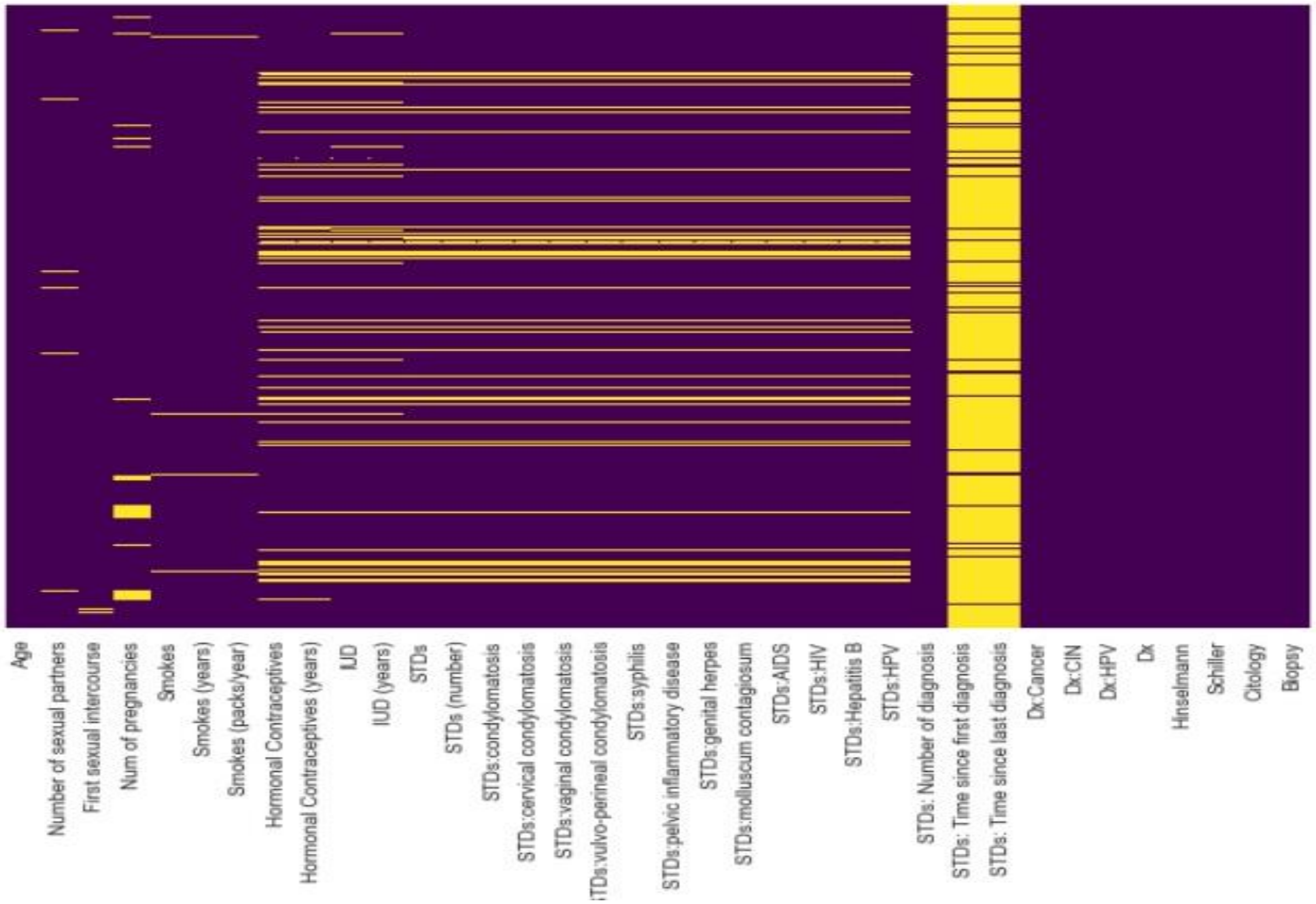
5 rows × 36 columns

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 858 entries, 0 to 857
Data columns (total 36 columns):
```

Feature Engineering

Data Set:
Features & Missing Values



Missing Values

- Categorical Data
  - Dummy Variables (Pd.get_dummies)
- Numeric Data
  - Imputation (Drop/Mean)

Feature Engineering

Dropping

26 out of 36 columns have missing values but since 'STDs: Time since first diagnosis' and 'STDs: Time since last diagnosis' columns have % 91.7 missing values, we are going to drop off these columns.

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

The rest columns have less than %15 missing values. For the numerical missing values, we will use imputing techniques to replace them. Since most of our columns have boolean type of variables we will implement pd.get_dummies() function to create dummy variables for all 0, 1 and NaN values. Thus we will not loose any data points.

Imputing

Imputing the numeric columns

```
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)
```

Dummy Variables

pd.get_dummies() function for categorical missing values

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

Feature Engineering

```
df.head()
```

	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	...	STDs:molluscum contagiosum_1.0	STDs:molluscum contagiosum_nan	STDs:
0	18	4.0	15.0	1.0	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	
2	34	1.0	17.0	1.0	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	...	0	0	
4	46	3.0	21.0	4.0	0.0	0.0	15.0	0.0	0.0	0	...	0	0	

5 rows × 64 columns

Data set consists of numeric data points solely.

```
df.info()
```

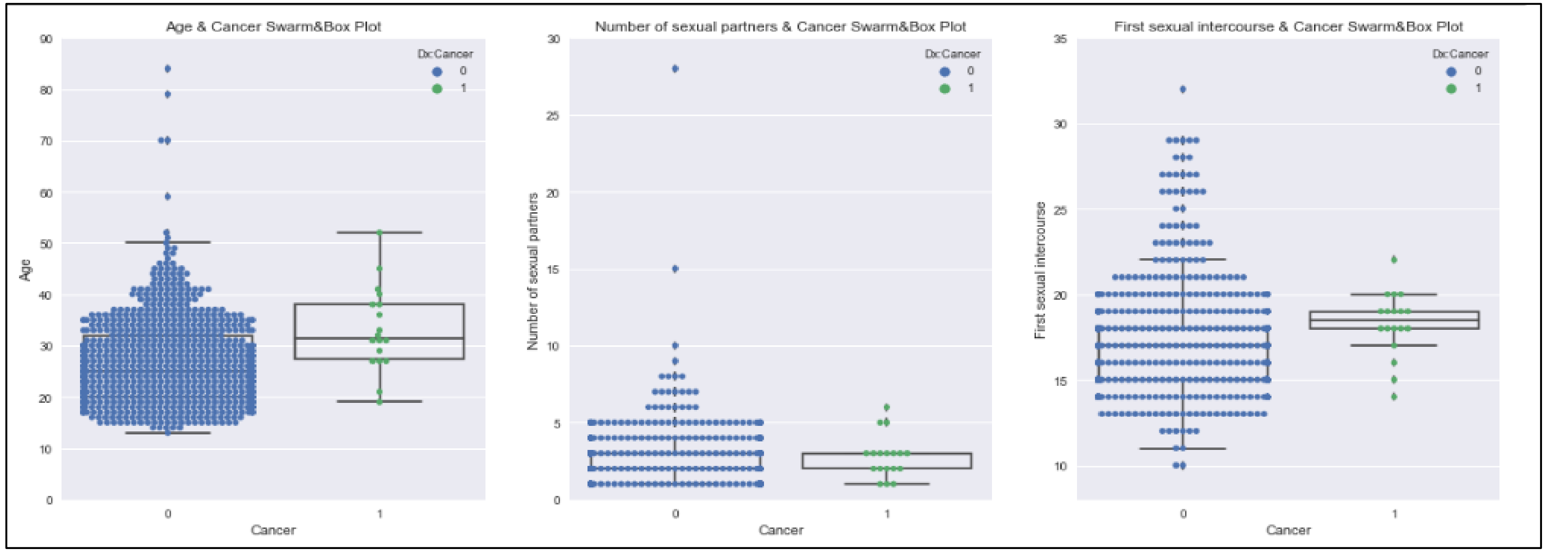
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 858 entries, 0 to 857
Data columns (total 64 columns):
```

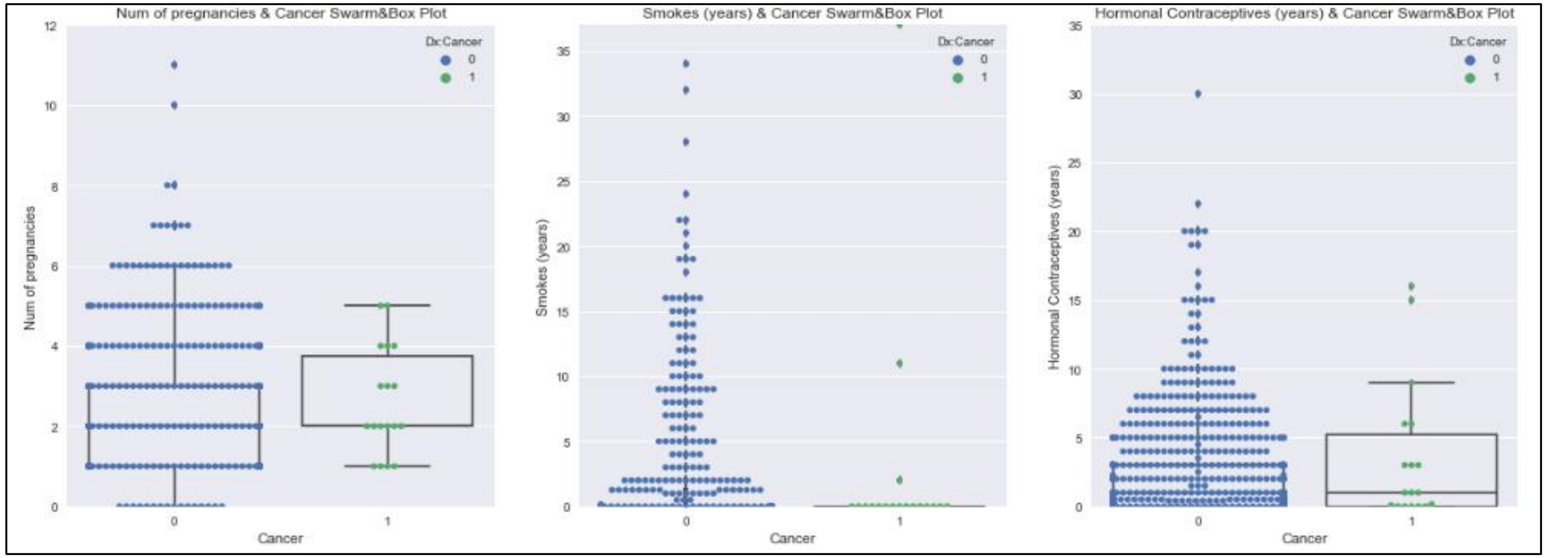
To save it as a csv file.**

```
df.to_csv('Cervical_Cancer_Risk_Cleaned.csv')
```

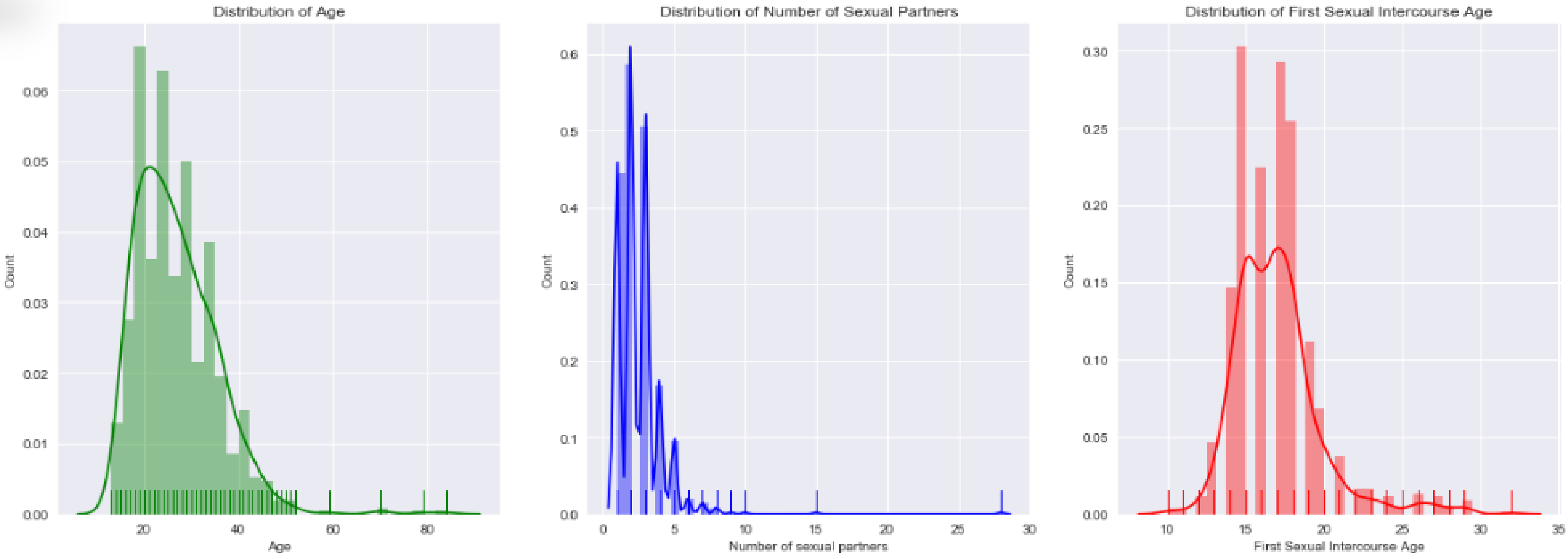
Data Exploration Analysis

- ✓ Cancer diagnosed patient's age are cumulated between 27 to 42. Cancer patient's median age is higher than non-cancers.
- ✓ 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.
- ✓ 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.
- ✓ Cancer patient's median number of pregnancies is higher than non-cancer.
- ✓ Most of the patients are not cancer smoke more than 3 years as well.
- ✓ Most of the non-cancer patients use hormonal contraceptives.



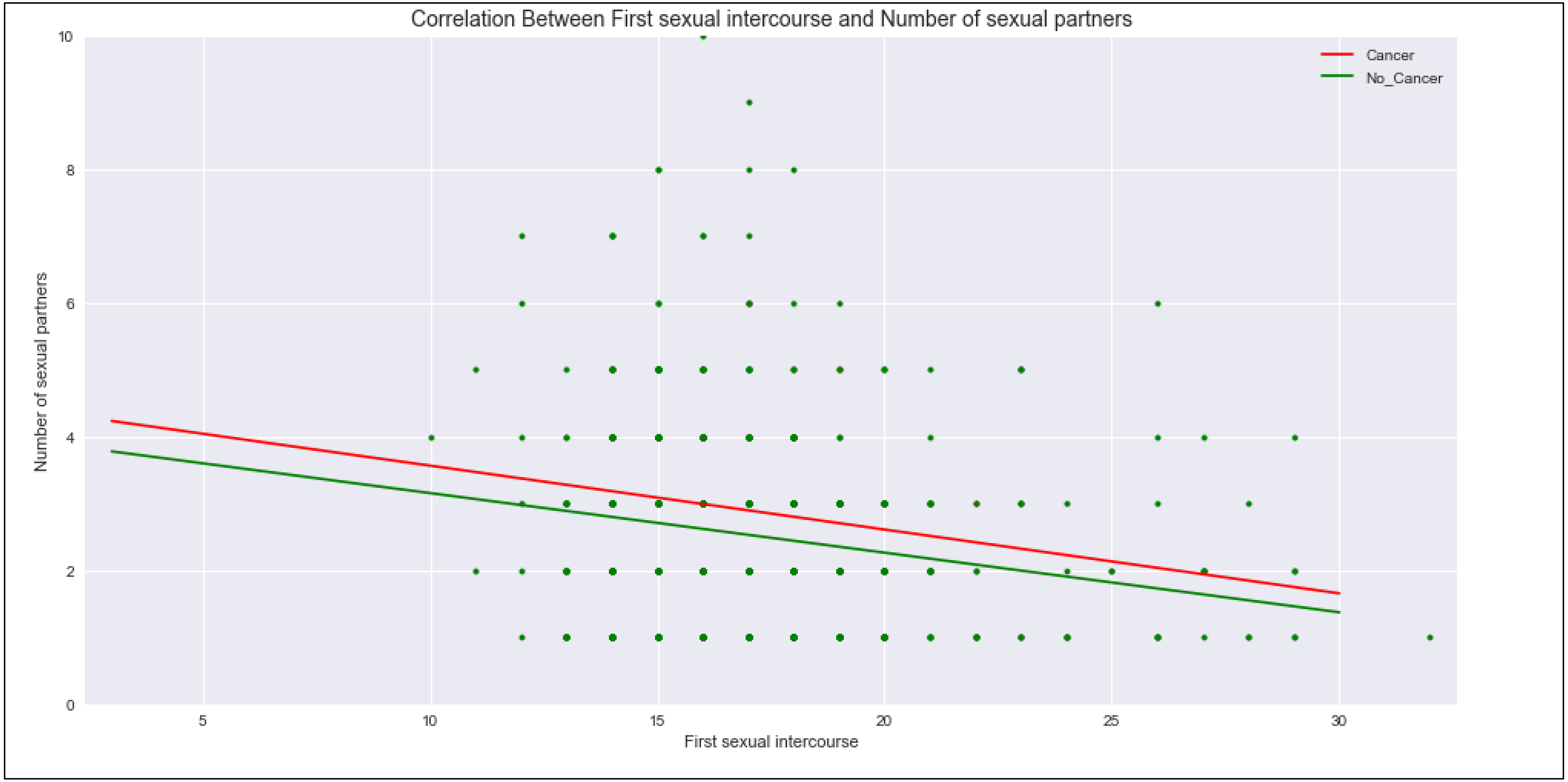


Data Exploration Analysis



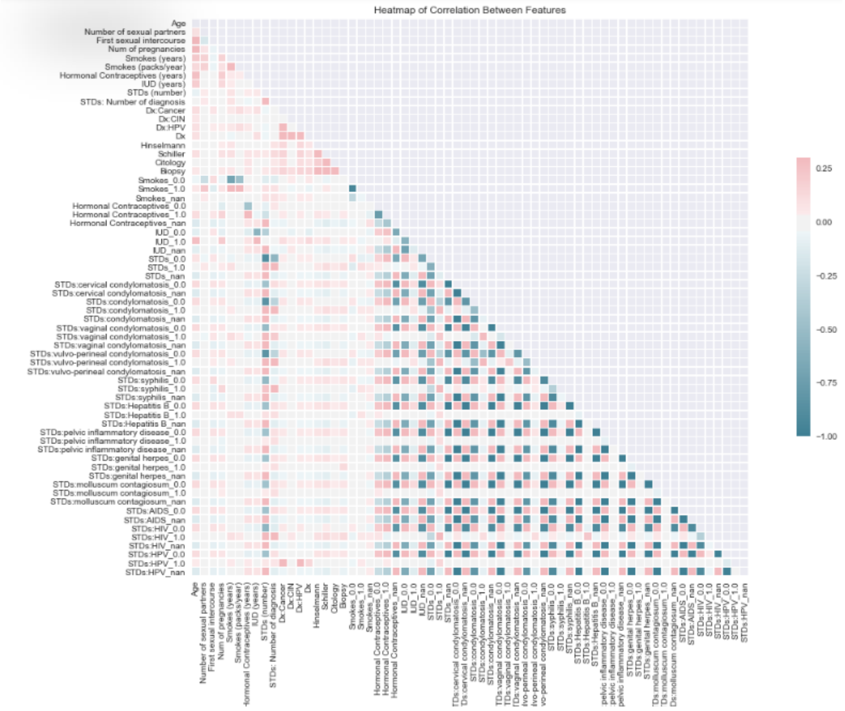
All three features look like normally distributed but skewed to right and there are some outliers in all of them

Data Exploration Analysis

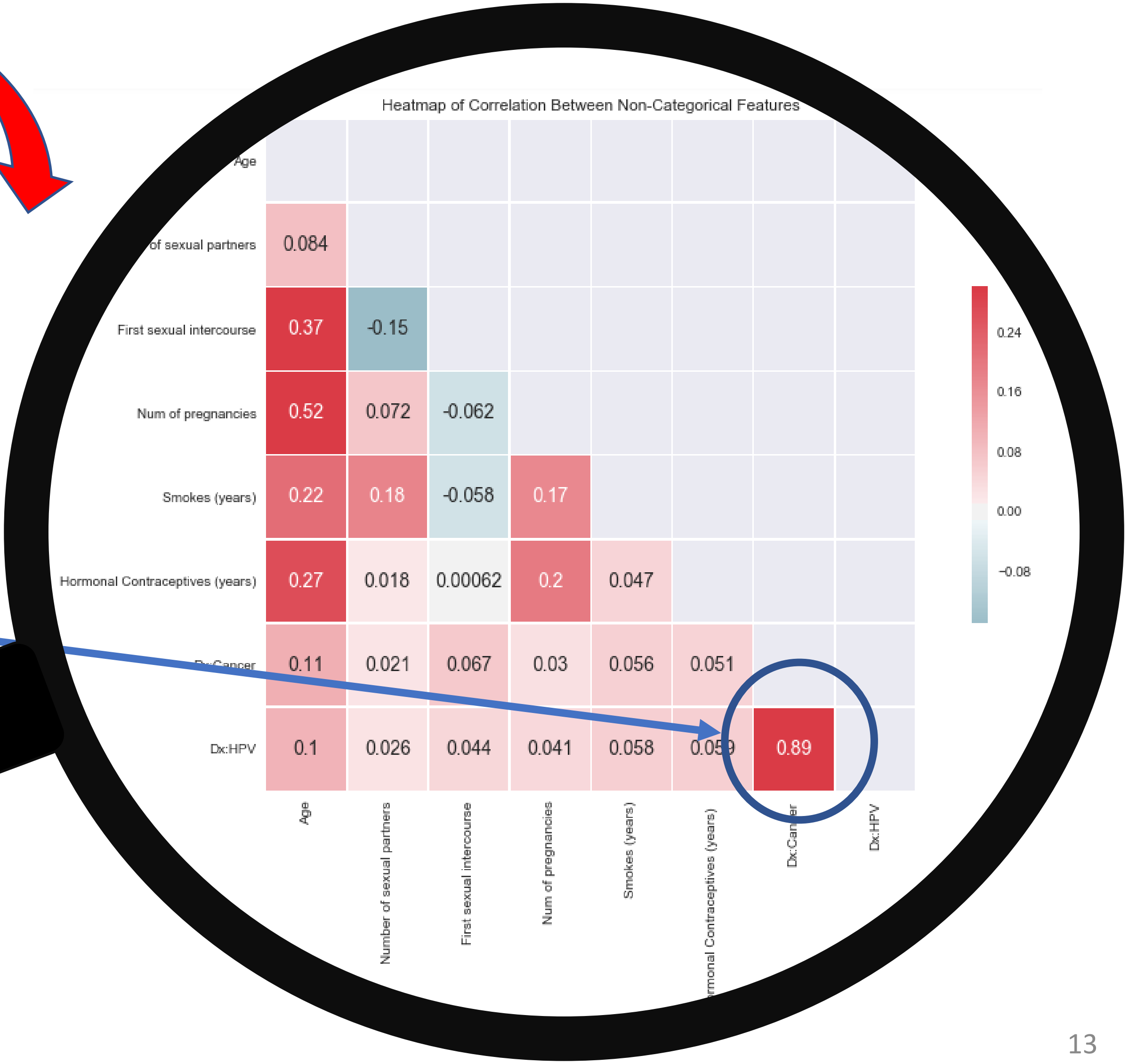


There is negative regression between First sexual intercourse and Number of sexual partners for both Cancer and Non-cancer diagnosed patients.

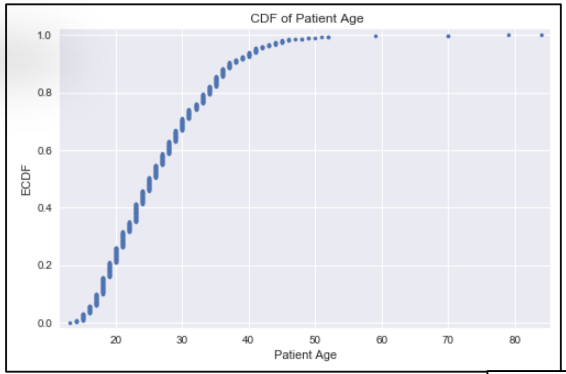
Data Exploration Analysis



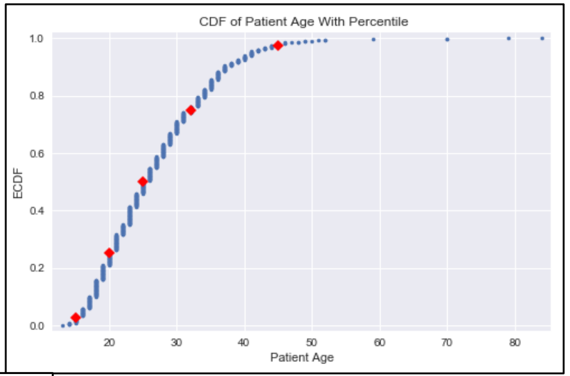
Age
First sexual intercourse
Number of Pregnancy
Smokes (year)
Hormonal Contras
Dx:HPV

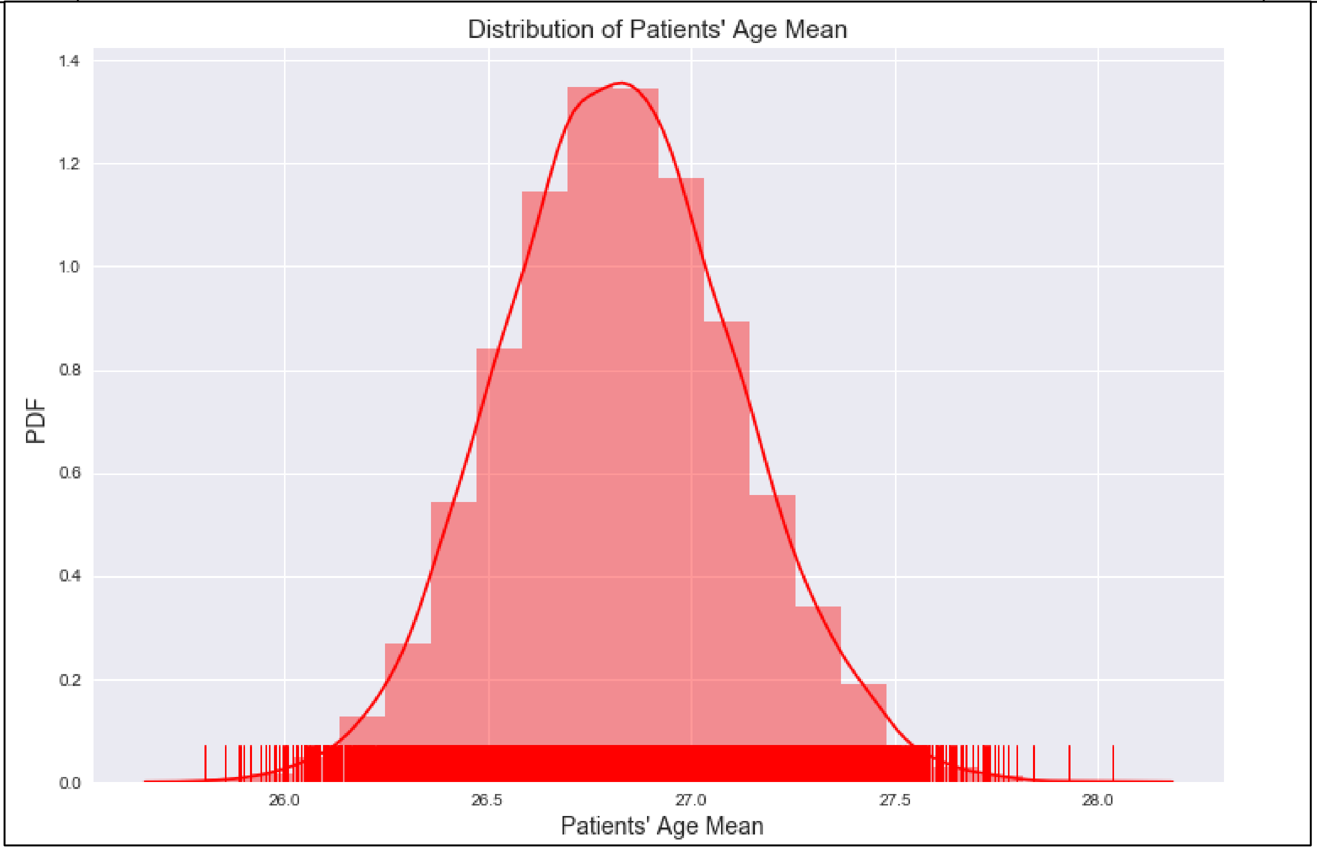


Data Exploration Analysis



Normally Distributed





Based on the 95% confidence interval, ages between 26.6 and 27.9 are considered normal.

There is not enough evidence that there is a significant difference between Cancer and Non-Cancer in average number of sexual partners.

Machine Learning

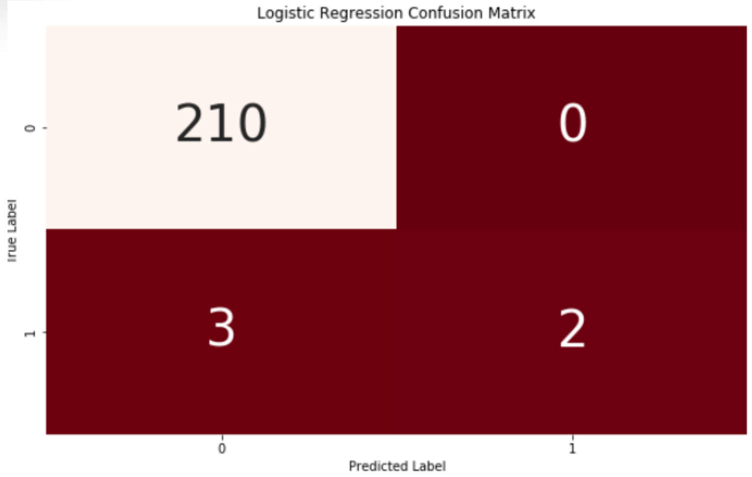
Train/Test = 0.75/0.25

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
							1	1.00	0.77	0.87	13
					Test	0.986	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
							1	0.92	0.85	0.88	13
					Test		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
							1	1.00	1.00	1.00	13
					Test		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	-	-	-	-
							1	-	-	-	-
					Test		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				
							1				
					Test	0.991	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	-	-	-	-
							1	-	-	-	-
					Test		0	1.00	1.00	1.00	210
							1	1.00	0.80	0.89	5

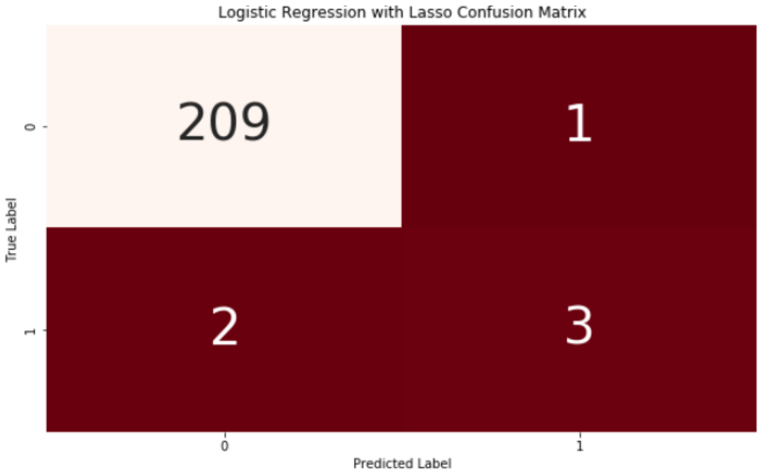
Machine Learning

BEFORE SMOTE

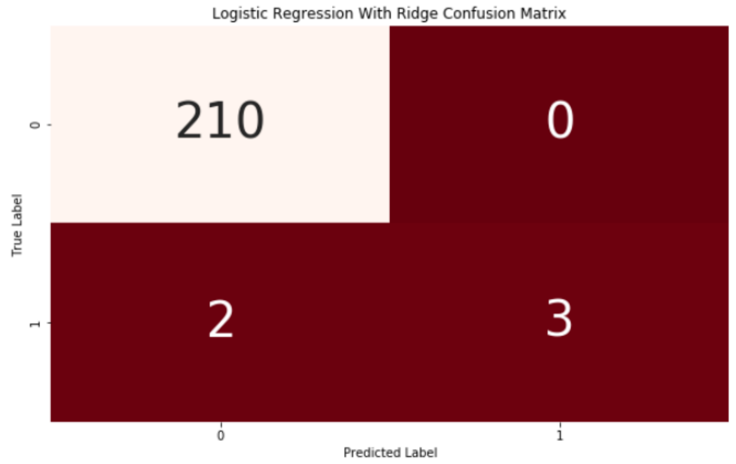
LogReg



Lasso

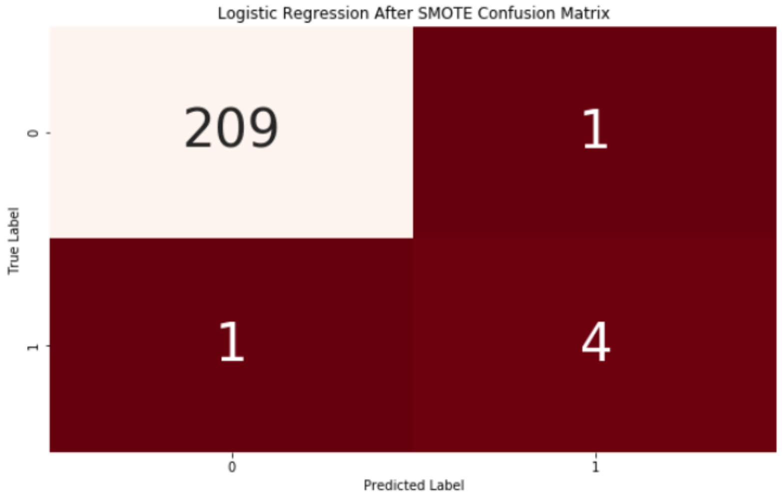


Ridge

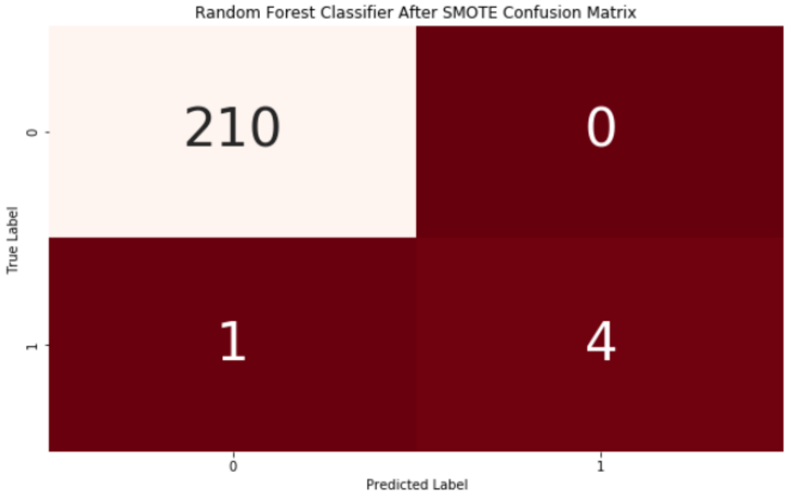


AFTER SMOTE

LogReg



Random Forest



THE BEST

Machine Learning

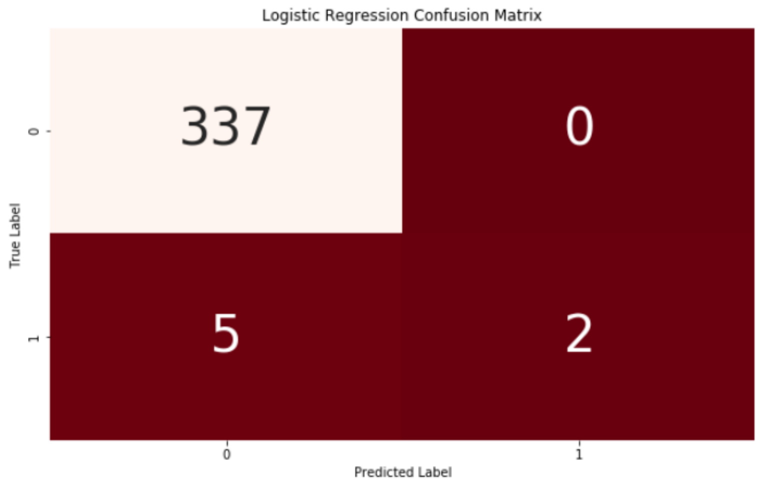
Train/Test = 0.60/0.40

							Classification Report				
Proportion (Train/Test)	SMOTE	DATA SET	Model/Application	Accuracy Score			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	-	-	-	-
							1	-	-	-	-
					Test	0.985	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
							1	0.90	0.82	0.86	11
					Test		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
							1	1.00	1.00	1.00	13
					Test		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	-	-	-	-
							1	-	-	-	-
					Test		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				
							1				
					Test	0.994	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	-	-	-	-
							1	-	-	-	-
					Test		0	1.00	1.00	1.00	337
							1	1.00	0.86	0.92	7

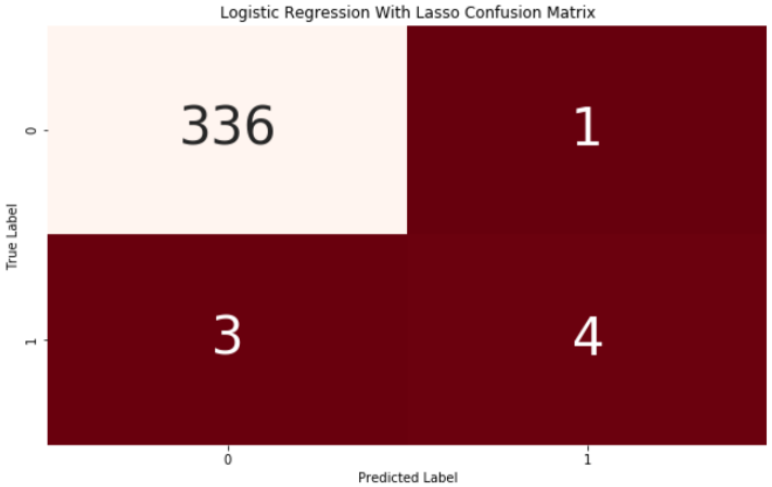
Machine Learning

BEFORE SMOTE

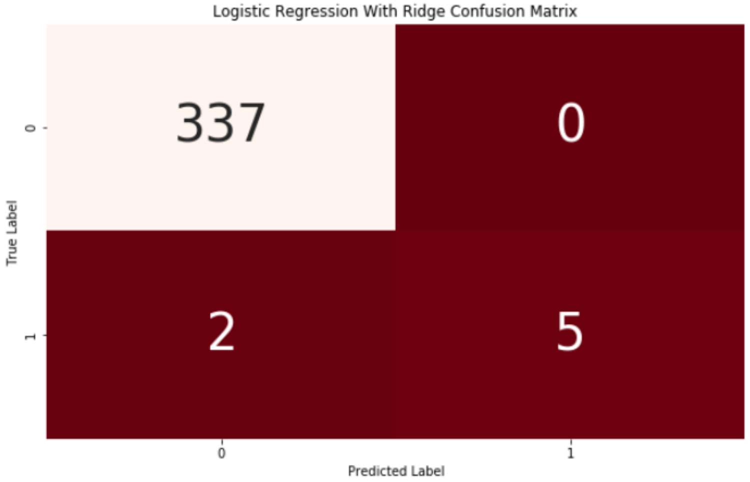
LogReg



Lasso

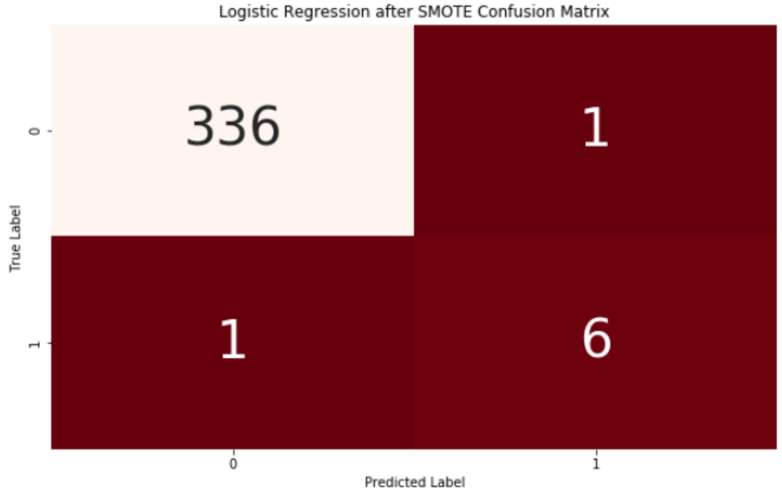


Ridge

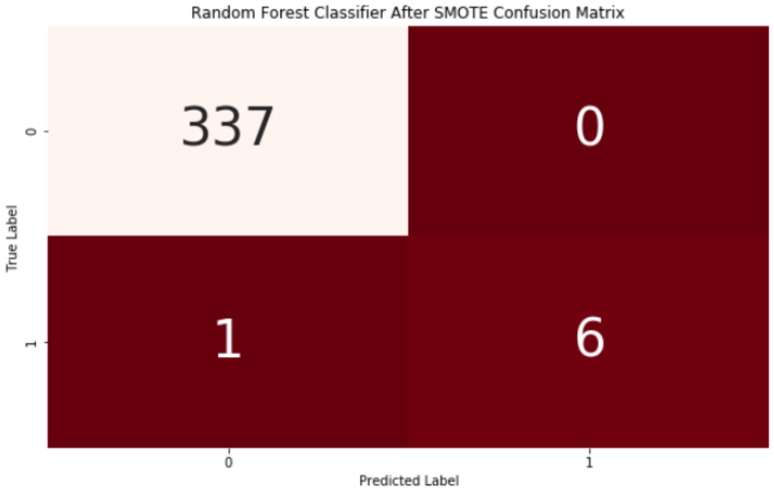


AFTER SMOTE

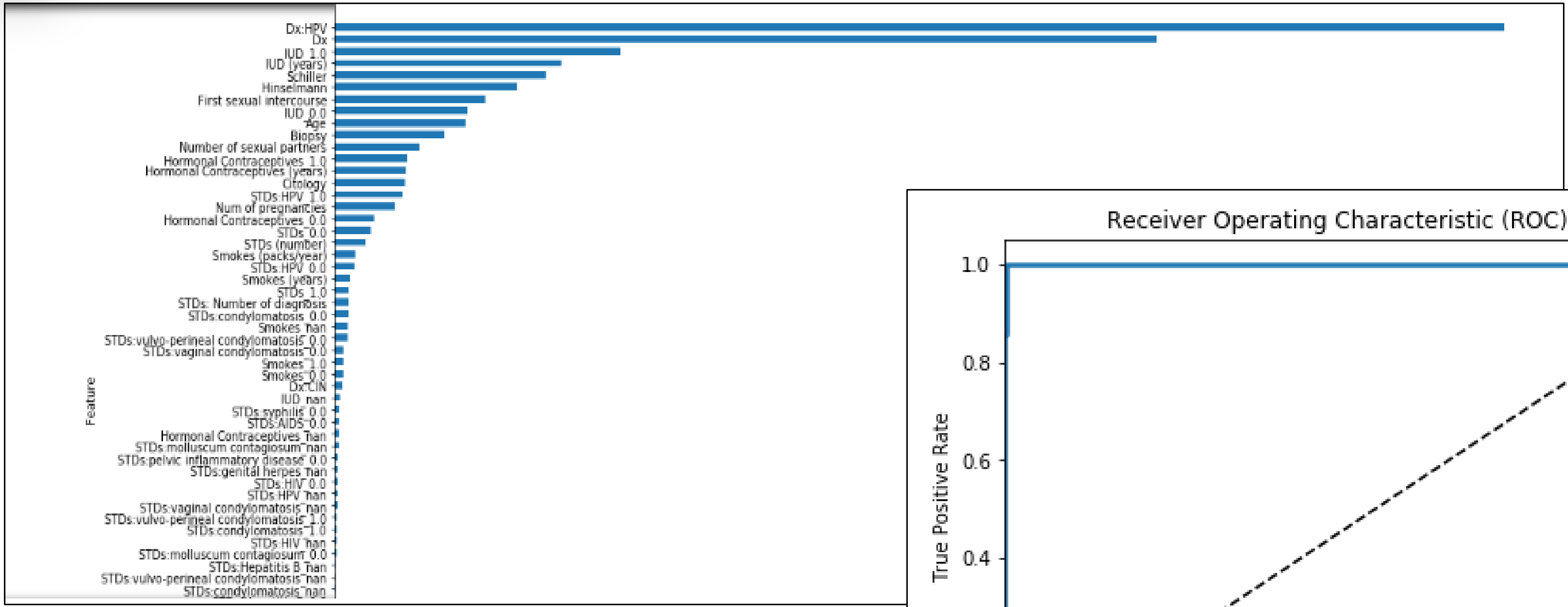
LogReg

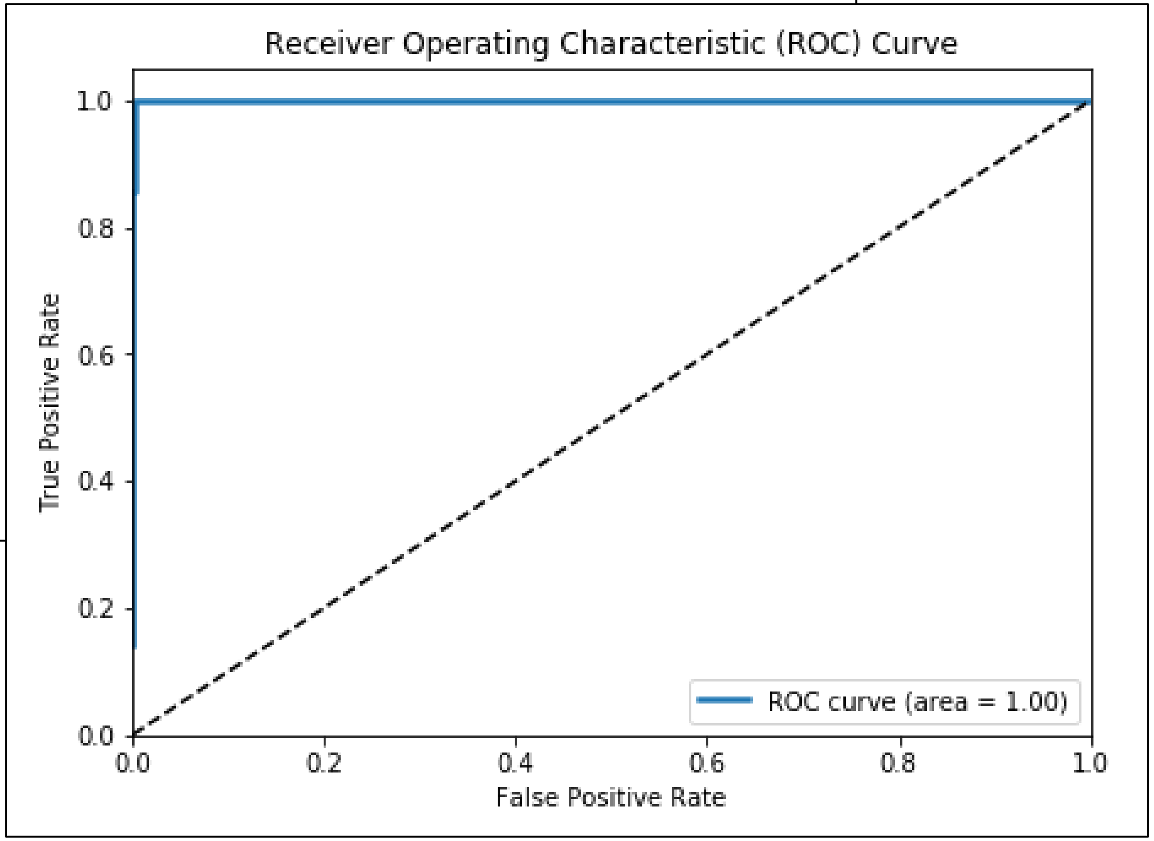


Random Forest



Machine Learning





Conclusion

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.