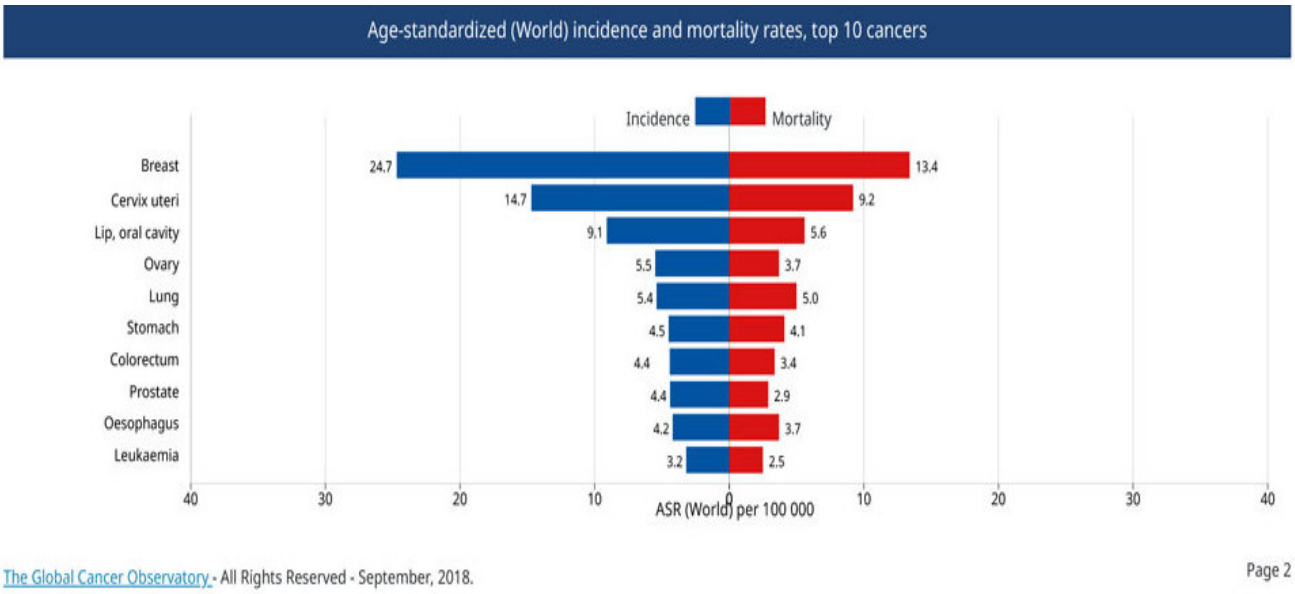
PREDICTION OF INDICATORS OF CERVICAL CANCER USING BIOPSY TEST

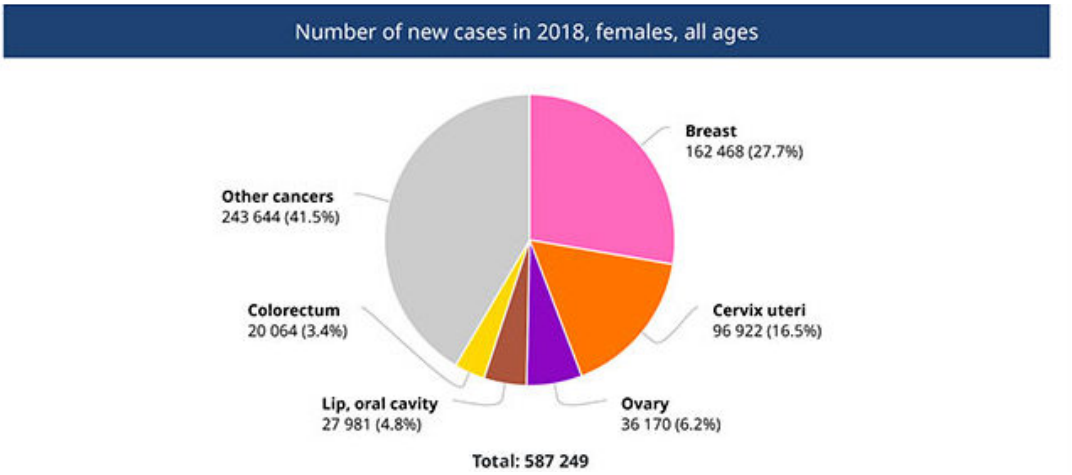
1.INTRODUCTION

Cervical cancer continues to be listed among the top gynaecologic cancers worldwide. According to current data of WHO, Cervical cancer is the fourth most frequent cancer in women with an estimated 570,000 new cases in 2018 representing 6.6% of all female cancers. Approximately 90% of deaths from cervical cancer occurred in low- and middle-income countries. The high mortality rate from cervical cancer globally could be reduced through a comprehensive approach that includes prevention, early diagnosis, effective screening and treatment programmes. There are currently vaccines that protect against common cancer-causing types of human papilloma virus and can significantly reduce the risk of cervical cancer.



In India, cervical cancer is the second most common cancer in women (aged 15–44 years) after breast cancer accounting for almost 14% of all female cancer cases. The higher mortality rate can be attributed largely to the lack of appropriate healthcare infrastructure in India. Cervical cancer in its advanced stage has a dismal outcome in terms of both prognosis and quality of life, registering approximately 67,477 deaths (23.3% of all cancer-related deaths) each year in Indian women.

PREDICTION OF INDICATORS OF CERVICAL CANCER USING BIOPSY TEST



Overview

Cervical cancer is the term used to describe tumours that can grow at the lower end of the womb. These tumours usually develop from abnormal cell changes at the entrance to the womb from the vagina (the opening of the cervix). Abnormal cell changes can be detected through screening and then removed. A vaccine against viruses that cause cancer (HPV vaccine) can reduce the risk of cervical cancer.

Symptoms

As in many cancers, you may have no signs or symptoms of cervical cancer until it has progressed to a dangerous stage. They may include:

- pain in the abdomen and pelvis
- pain when urinating
- abnormal bleeding – between menstruation or after the menopause
- abnormal vaginal discharge that might have an unpleasant odour
- tiredness and weight loss

PREDICTION OF INDICATORS OF CERVICAL CANCER USING BIOPSY TEST

Cervical Cancer Survival Rates by Stage

Stage	5 Year Survival
I	81-96%
II	65-87%
III	35-50%
IV	15-20%

Causes

Cervical cancer is nearly always caused by a long-term infection with particular viruses known as human papillomaviruses (HPV). There are many different types of HPV. They infect skin cells and mucous membrane cells, and are spread through sex or direct contact in the genital area. Infection through body fluids like sperm, blood or saliva is considered to be unlikely. Most women have an HPV infection at some point in their lives, many between the ages of 20 and 30. Men can also become infected and pass the infection on.

Risk factors

Certain groups of women are more likely than others to get cervical cancer. For instance, women who smoke are at higher risk. This might be because their immune systems are less effective at fighting the viruses. Other things that weaken the immune system – such as diseases like AIDS, and medication that is taken after an organ transplant – can increase this risk too.

Because HPV is a sexually transmitted infection, things that generally increase the risk of sexually transmitted diseases are also considered to be risk factors for cervical cancer. So, for example, the more sexual partners someone has, the higher their risk of cervical cancer.

Diagnosis

There may be reason to believe that a woman has precancerous cell changes or cervical cancer if she has certain symptoms, abnormal smear test (Pap test) results, or abnormal findings during a gynaecological examination. In this kind of physical examination, the

doctor feels the area around the womb with their hands, both from the outside (belly) and from the inside (vagina). They also insert a special instrument (speculum) into the vagina, to look at the tissue at the opening of the cervix. If the tissue has changed a lot, it can be examined more closely using a kind of magnifying glass called a colposcope. A sample of tissue can be taken too (biopsy). This is then tested in a lab.

Prevention

Cervical cancer nearly always develops as a rare consequence of a long-term infection with particular types of HPV. So, at least theoretically, there are three ways to help prevent it:

- Avoiding sexual contact or using condoms
- HPV vaccine in girls
- Screening and removing abnormal tissue

Treatment

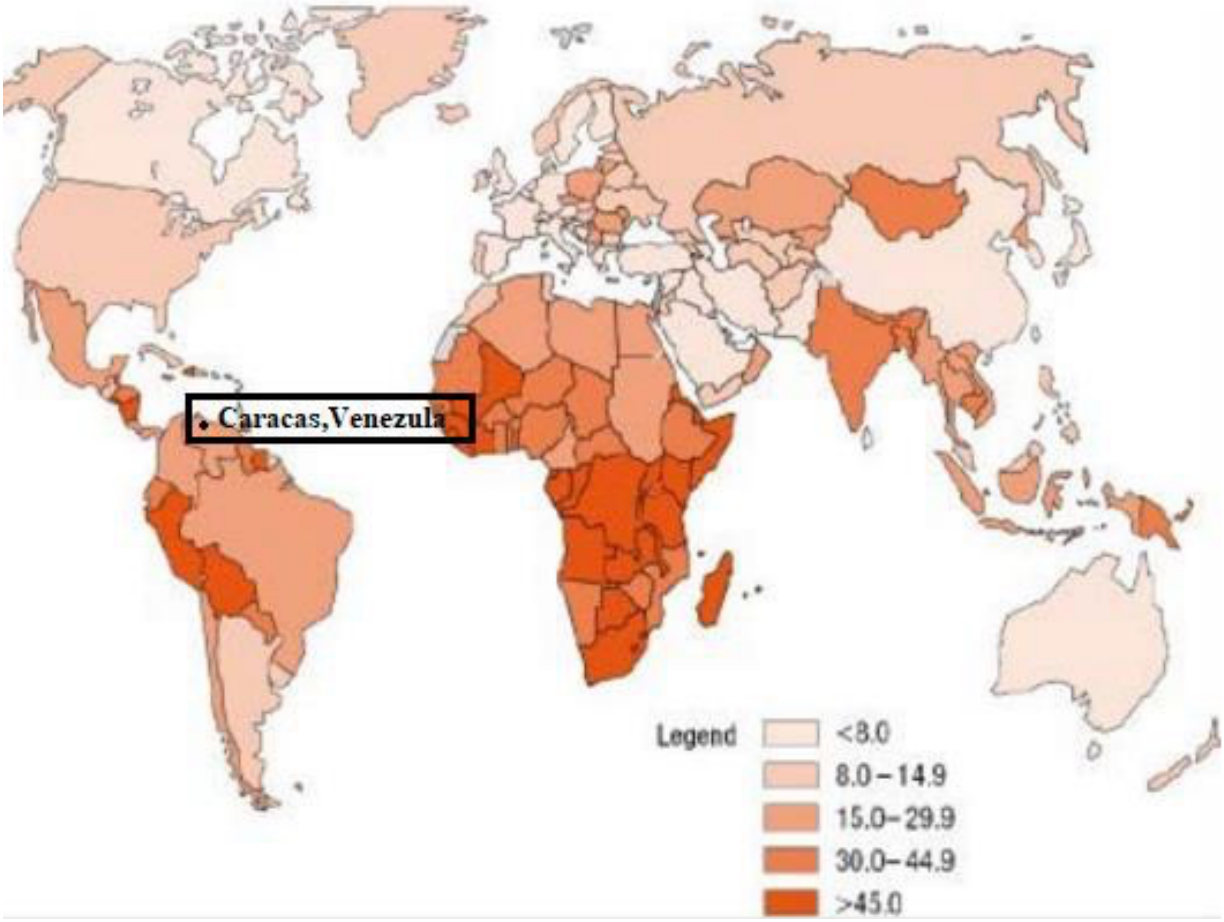
The most appropriate type of treatment will mainly depend on the size of the tumour and whether the cancer has spread. If the tumour is discovered at a very early stage, a small surgical procedure (conization) might be enough. If the tumour has already spread to the surrounding tissue, doctors usually recommend having surgery to remove the entire womb (a hysterectomy). The lymph nodes are removed in a wide area around the womb too. Radiotherapy might also be considered. Radiotherapy is still an option even if the tumour can no longer be removed through surgery. In some patients it can be combined with chemotherapy.

2.DATASET

DATASET INFORMATION

The dataset was collected at 'Hospital Universitario de Caracas' in Caracas, Venezuela. The dataset 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). This dataset focuses on the prediction of indicators/diagnosis of cervical cancer.

PREDICTION OF INDICATORS OF CERVICAL CANCER USING BIOPSY TEST



Abstract:

This dataset focuses on the prediction of indicators/diagnosis of cervical cancer. The features cover demographic information, habits, and historic medical records

Attribute name	Type
Age	int
Number of sexual partners	int
First sexual intercourse (age)	int
Number of pregnancies	int
Smokes	bool
Smokes (years)	int
Smokes (packs/year)	int
Hormonal Contraceptives	bool
Hormonal Contraceptives (years)	int
IUD	bool
IUD (years)	int

Attribute name	Type
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	Bool
STDs: Number of diagnosis	int
STDs: Time since first diagnosis	int
STDs: Time since last diagnosis	int
Dx: Cancer	Bool

PREDICTION OF INDICATORS OF CERVICAL CANCER USING BIOPSY TEST

STDs	bool
STDs (number)	int
STDs: condylomatosis	bool
STDs: cervical condylomatosis	bool
STDs: vaginal condylomatosis	bool
STDs: vulvo-perineal condylomatosis	bool
STDs: syphilis	bool

Dx: CIN	Bool
Dx: HPV	Bool
Dx	Bool
Hinselmann: screening test	Bool
Schiller: screening test	Bool
Cytology: screening test	Bool
Biopsy: target variable	Bool

The following are the description of independent and the dependent attributes:

1. **Age** - It indicates the age of a woman. It is expressed in terms of numerical values
2. **Number of sexual partners** – It indicates the total number of sexual partners encountered. It is expressed in terms of numerical values.
3. **First sexual intercourse**- It indicates the age of a woman when she had her first sexual intercourse. It is expressed in terms of the count.
4. **Number of pregnancies** – It indicates the total number of times the woman got pregnant. It is expressed in terms of the total count.
5. **Smokes**- It indicates whether the person smokes or not. It is expressed in terms of zeros (does not smoke) and ones(smokes).
6. **Smokes (years)-** It indicates the total number of years for which the woman is smoking. It is expressed in terms of total count.
7. **Smokes (packs/year)-** It indicates the total number of packets of cigarettes per year the woman smokes. It is expressed in terms of numbers
8. **Hormonal Contraceptives -** It indicates whether the patient uses hormonal contraceptives or not.
9. **Hormonal Contraceptives (years)** – It indicates that for how many years the contraceptive method was used. It was in expressed in terms of total number of years.
10. **Intra-Uterine Device**- It indicated where the intrauterine contraceptive device was used or not. It was expressed in terms of zeros(did not used IUD) and ones(used IUD).
11. **IUD (years)** – It indicated that for how many years the IUD was used. It is expressed in terms of the total number of years.
12. **STDs** - It indicates the presence of Sexually Transmitted Diseases. It is

expressed in terms of zeroes and ones.

13. **STDs (number)** – It indicates the total number of sexually transmitted disease present with the patient. It is expressed in terms of numbers.
14. **STDs:condylomatosis** – It indicates the presence of Condylomatosis with the patient.
15. **STDs:cervical condylomatosis** –It indicates the presence of Cervical condylomatosis.
16. **STDs:vaginal condylomatosis -** It indicates the presence of Vaginal condylomatosis.
17. **STDs:vulvo-perineal condylomatosis** – It indicates the presence of Vulvo- Perineal condylomatosis.
18. **STDs:syphilis** – It indicates the presence of Syphilis.
19. **STDs:pelvic inflammatory disease**- It indicates the presence of pelvic inflammatory disease.
20. **STDs:genital herpes** – It indicates the presence of Genital Herpes.
21. **STDs:molluscum contagiosum** – It indicates the presence of Molluscum Contagiosum.
22. **STDs:AIDS –** It indicates the presence of AIDS in the patient.
23. **STDs:HIV** – It indicates the presence of HIV in the patient.
24. **STDs:Hepatitis B** – It indicates the presence of Hepatitis B in the patients.
25. **STDs:HPV –** It indicates the presence of HPV in the patients.
26. **STDs: Number of diagnosis –** It indicate the total number of times the STDs have been diagnosed.
27. **STDs: Time since first diagnosis –** It indicates the total number of years since the first diagnose.
28. **STDs: Time since last diagnosis** – It indicates the total number of years elapsed since the last diagnose.
29. **Dx:Cancer** – It indicates the presence of Cancer after the diagnose.
30. **Dx:CIN** – It indicates the presence of Cervical intraepithelial neoplasia.
31. **Dx:HPV**- It indicates the presence of Human papillomaviruses.
32. **Dx** - It indicates the presence any one among cancer, CIN and HPV.
33. **Hinselmann –** also known as colposcopy, is a medical diagnostic procedure to examine an illuminated, magnified view of the cervix as well as the vagina and vulva.

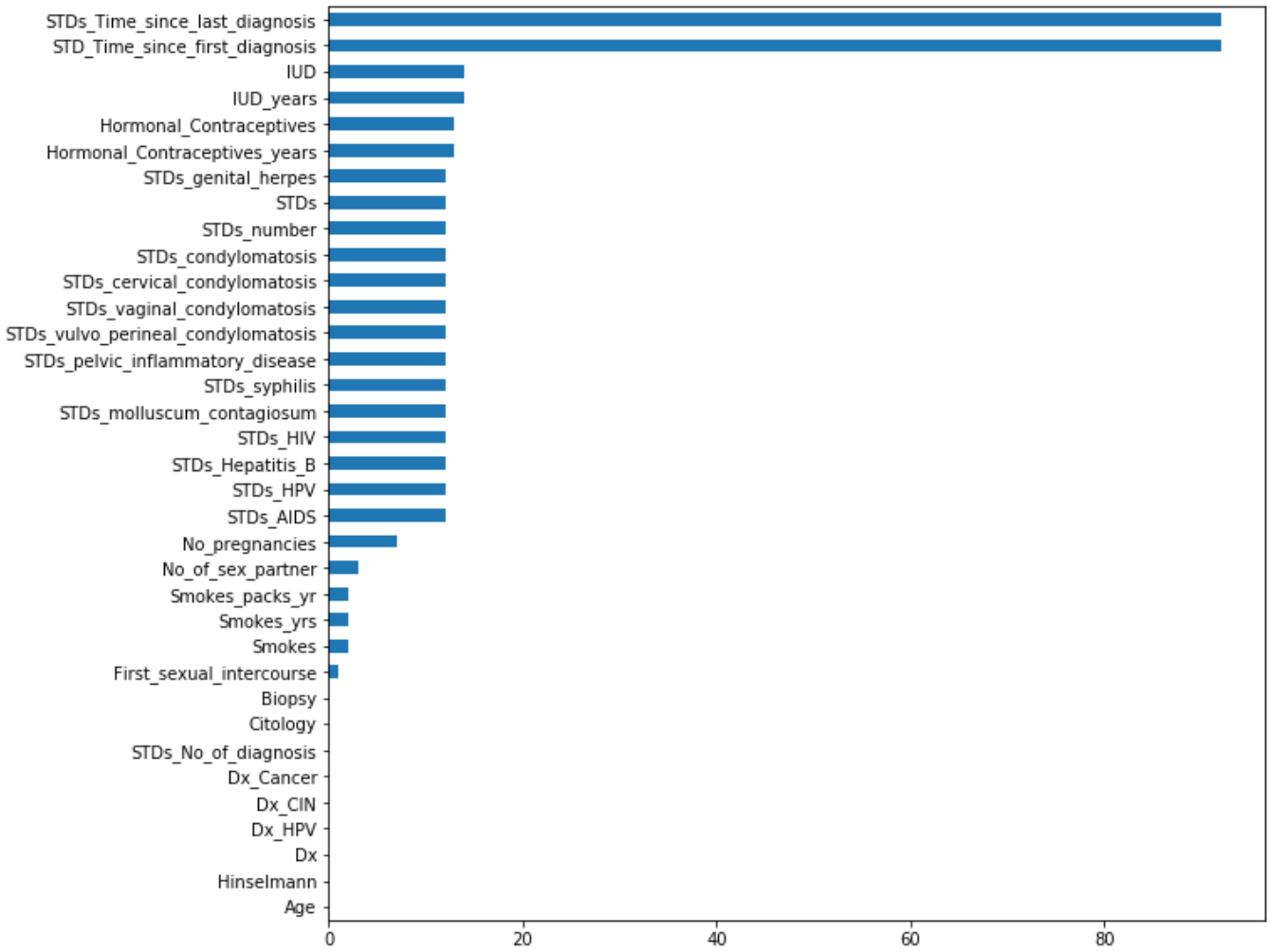
PREDICTION OF INDICATORS OF CERVICAL CANCER USING BIOPSY TEST

34. **Schiller -** Schiller's Iodine test is a medical test in which iodine solution is applied to the cervix in order to diagnose cervical cancer.
35. **Cytology –** also called as PaP smears test, helps detect abnormal cells in the cervix, which can develop into cancer.
36. **Biopsy -** A cervical biopsy is a surgical procedure in which a small amount of tissue is removed from the cervix. A cervical biopsy is usually done after an abnormality has been found during cytology.

3.EXPLORATORY DATA ANALYSIS:

TREATMENT OF MISSING VALUES:

Missing data in the training data set can reduce the power / fit of a model or can lead to a biased model because we have not analysed the behaviour and relationship with other variables correctly. It can lead to wrong prediction or classification.



PREDICTION OF INDICATORS OF CERVICAL CANCER USING BIOPSY TEST

The amount of missing value percentage from the above graph shows that the attributes STDs_Time_since_last_diagnosis and STDs_Time_since_first_diagnosis has more than 80 percent of null values and hence the two columns are removed from the dataset. Also, the columns Smokes and First Sexual Intercourse whose rows having the least missing values are removed from the dataset.

MISSING VALUE IMPUTATION:

Missing values are imputed by two methods:

i) Imputing missing values with mean/median/mode value
ii) Imputing missing values using machine learning model

IMPUTATION WITH MEDIAN AND MODE VALUES:

Generally, the numeric and categorical variables are replaced with Median and Mode values respectively.

Categorical Variables:

Smokes	Hormonal_Contraceptives
STDs_AIDS	STDs_cervical_condylomatosis
STDs_HIV	STDs_vaginal_condylomatosis
STDs_Hepatitis_B	STDs_vulvo_perineal_condylomatosis
STDs_HPV	STDs_syphilis
Dx_Cancer	STDs_pelvic_inflammatory_disease
Dx_CIN	STDs_genital_herpes
Dx_HPV	STDs_molluscum_contagiosum
Dx	STDs_condylomatosis

Numerical Variables:

Age
No_of_sex_partner
First_sexual_intercourse
No_pregnancies

PREDICTION OF INDICATORS OF CERVICAL CANCER USING BIOPSY TEST

Smokes_yrs
Smokes_packs_yr
STDs_No_of_diagnosis.
Hormonal_contraceptive_years

The question marks (?) have been replaced with NAN values. We calculate the median and mode of all the features and then the NANs are replaced with median and mode values respectively for each of the above variables. This is an approximation which can add variance to the data set. But the loss of data can be negated by this method which yields better results compared to removal of rows and columns. Replacing with the above two approximations are a statistical approach of handling the missing vales.

IMPUTATION USING MACHINE LEARNING ALGORITHMS:

Using the features which do not have missing values, we can predict the nulls with the help of machine learning algorithm. We used regression techniques for continuous target variables and classification techniques for categorical target variables. This method may result in better accuracy, unless a missing value is expected to have a very high variance.

We took each of the columns individually and treated them. We first took the variables which had maximum missing values from the remaining variables. The variable with maximum missing value was taken and a model was performed on the column, keeping the other variables as dependent variables. We did one hot encoding for the categorical variables. We predicted the missing values for that dependent column and replaced the missing values with predicted value. This iteration process of modelling the predicting and replacing the missing values was until there was no missing values in any one of the columns.

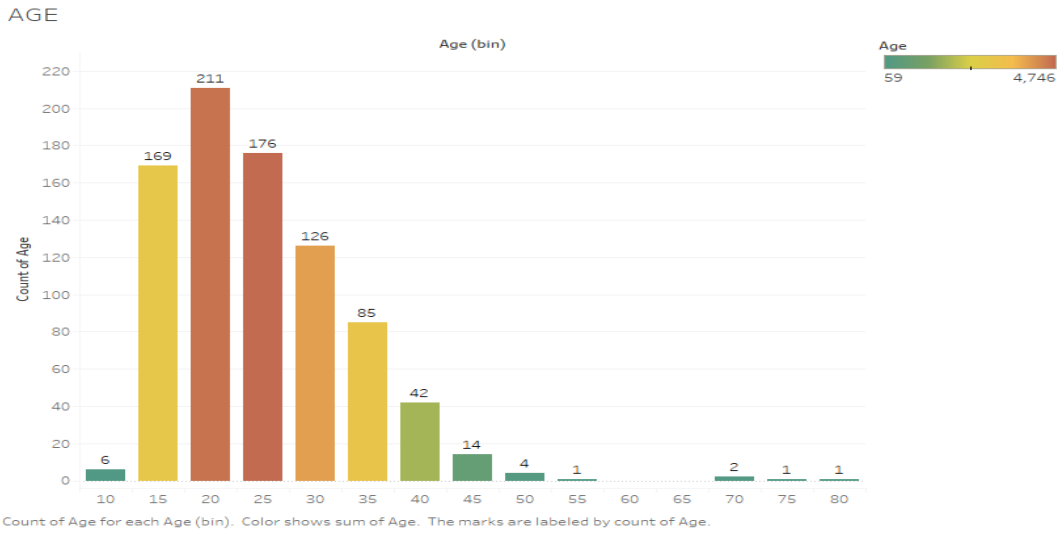
Now we have a dataset where the missing values were replaced by the predicted values from the models.

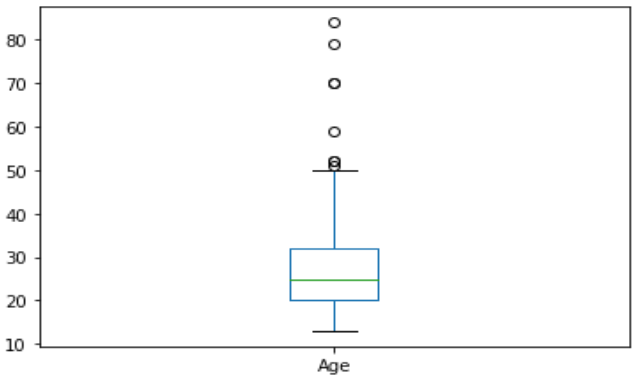
DATASET ATTRIBUTES DESCRIPTION & ANALYSIS

AGE:

DISTRIBUTION:

PREDICTION OF INDICATORS OF CERCIVAL CANCER USING BIOPSY TEST





FIVE POINT SUMMARY

count	838
mean	26.8126
std	8.52921
min	13
25%	20
50%	25
75%	32
max	84

count	54
mean	28.55
std	8.947
min	16
25%	21
50%	28
75%	34.75
max	62

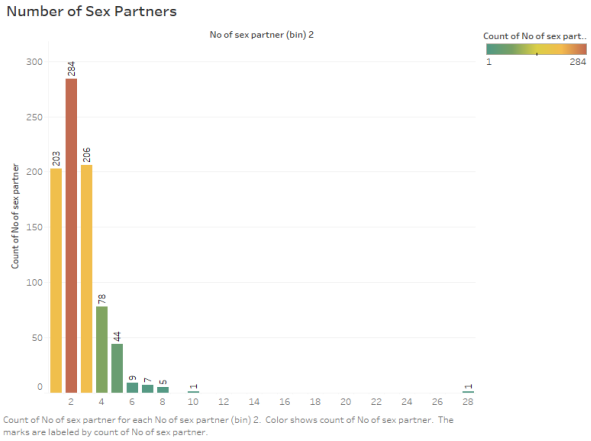
These 54 records are positive biopsy records and the mean and median values of age are 28.5 and 28 respectively. The above age distribution of cervical cancer shows that the Venezuela cervical cancer standardized age is between 15-30.

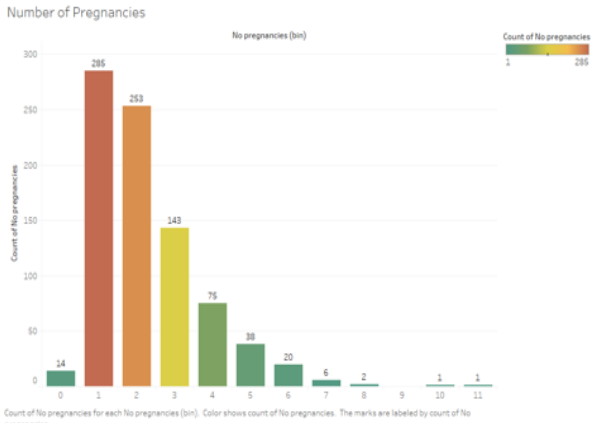
PREDICTION OF INDICATORS OF CERVICAL CANCER USING BIOPSY TEST

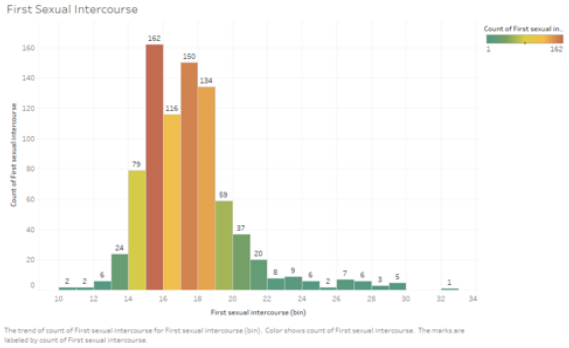
SEXUAL HABIT ATTRIBUTES

- Number of sexual partners – Total number of sexual partners.
- First sexual intercourse (age) – Age at which the patient had first sexual intercourse.
- Number of pregnancies – Number of times patient become pregnant.

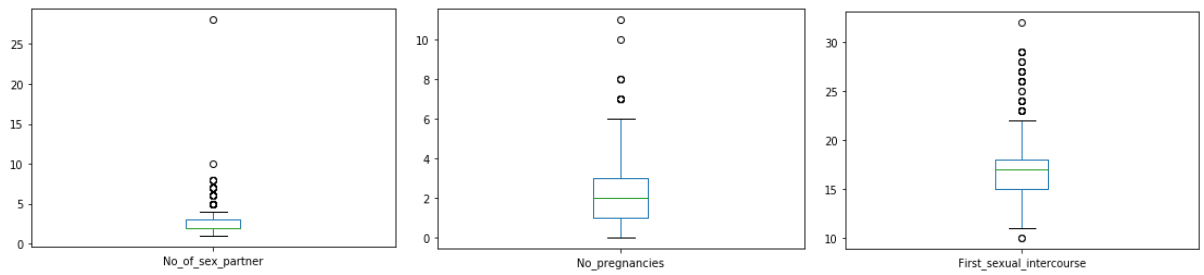
DISTRIBUTION:





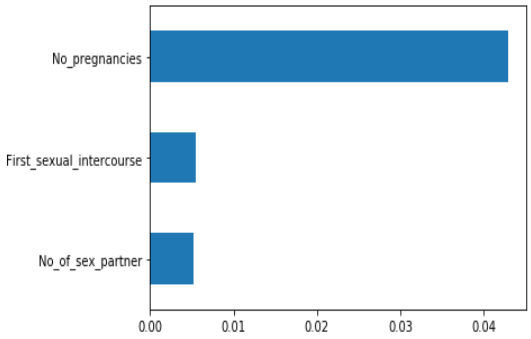


5 POINT SUMMARY:

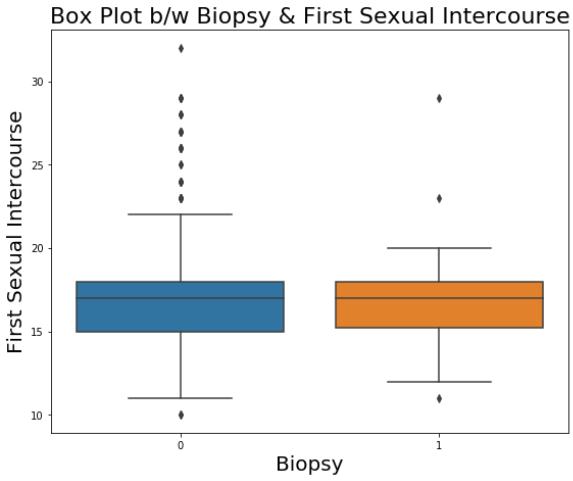


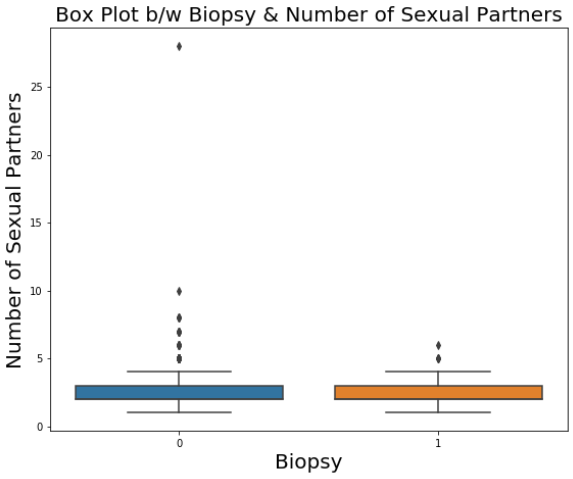
PREDICTION OF INDICATORS OF CERVICAL CANCER USING BIOPSY TEST

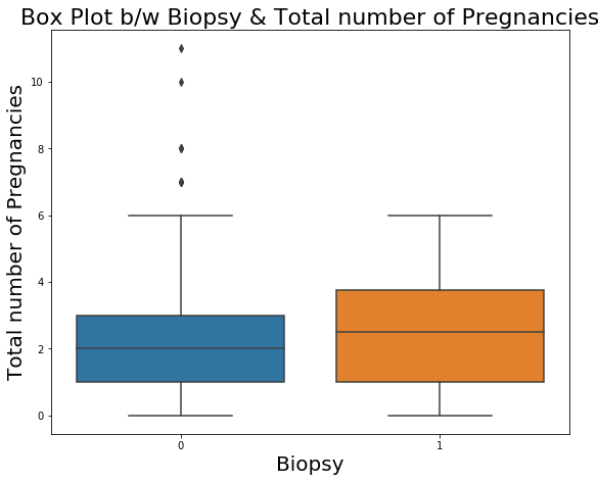
	First_sexual_intercourse	No_of_sex_partner	No_pregnancies
count	838.000000	838.000000	838.000000
mean	16.996420	2.509547	2.270088
std	2.812965	1.589964	1.449391
min	10.000000	1.000000	0.000000
25%	15.000000	2.000000	1.000000
50%	17.000000	2.000000	2.000000
75%	18.000000	3.000000	3.000000
max	32.000000	28.000000	11.000000



RELATION B/W BIOPSY AND SEXUAL ATTRIBUTES:







PREDICTION OF INDICATORS OF CERVICAL CANCER USING BIOPSY TEST

INFERENCES:

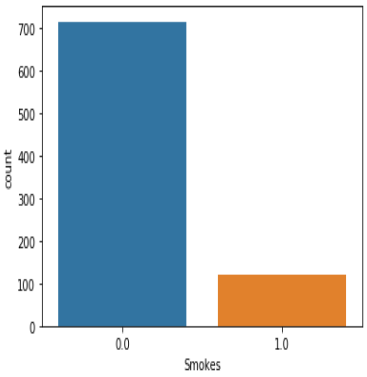
- Data is not normally distributed for all sex attributes and the data is right skewed.
- The maximum sexual partners per person is in the range of 1 to 3 and average age at which the patient had their first sexual intercourse is between 15 years and 19 years.
- Those who had sexual intercourse at very early age, have the highest number of sex partners.
- Those who smoke and the first sexual intercourse age between 15yrs and 18yrs are more prone to be test as Positive in Biopsy test.
- The persons who have sexual partners between 1 and 3 are more prone to be tested as positive in Biopsy test.
- The patients who have sexual partners between 1 and 4 are getting pregnant more often. The number of sexual partners does not have any effect on the age at which they had their first sexual intercourse.

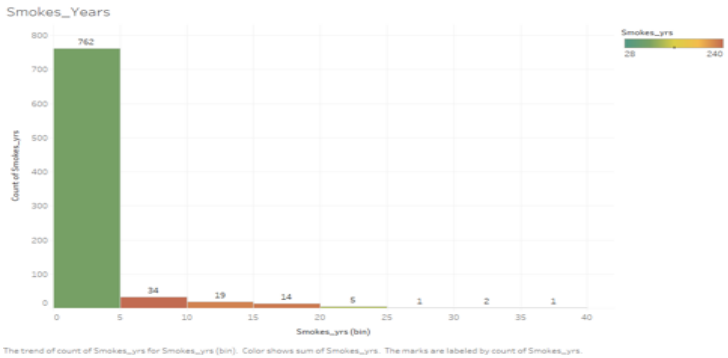
SMOKING HABIT ATTRIBUTES:

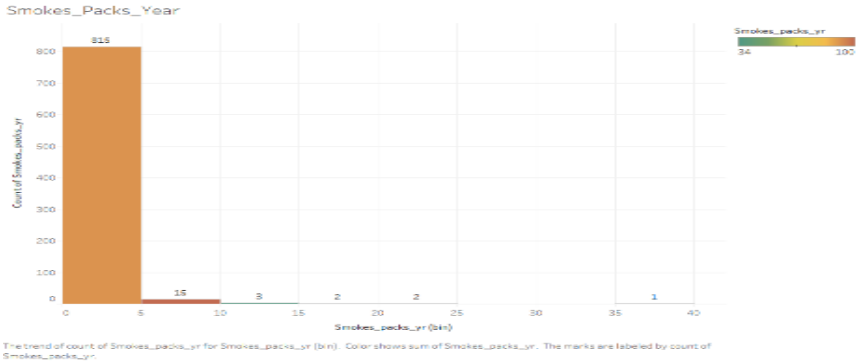
- Smokes – Patient having smoking habit or not.
- Smokes (years) – Number of years having smoking habit (0 for non-smokers).
- Smokes (packs/year) – Number of packets smoking per year (0 for non-smokers).

PREDICTION OF INDICATORS OF CERVICAL CANCER USING BIOPSY TEST

DISTRIBUTION:

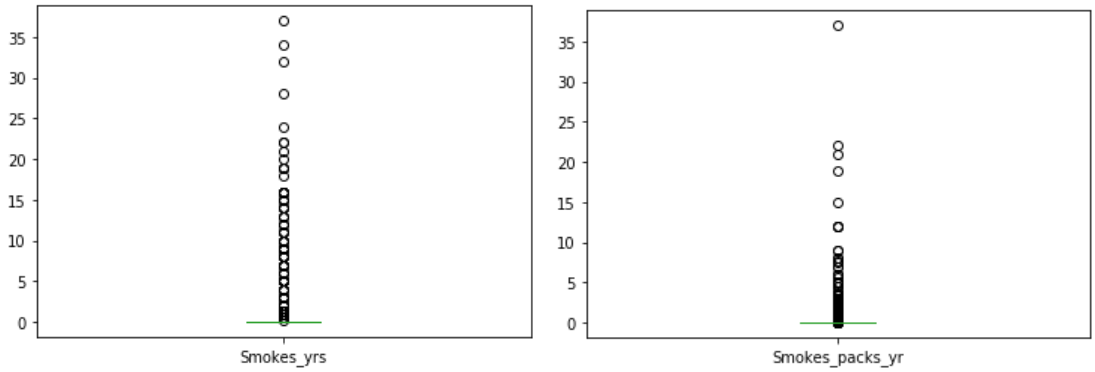




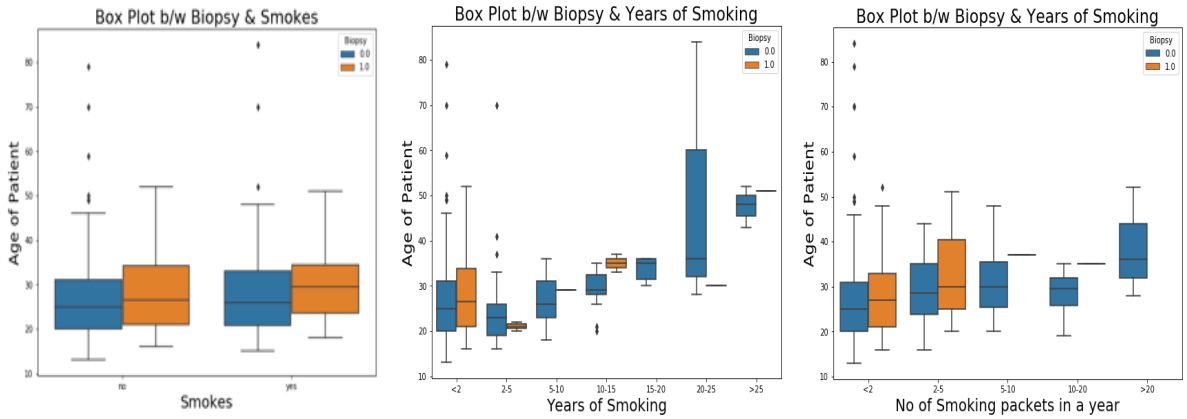


	Smokes_yrs	Smokes_packs_yr
count	838	838
mean	1.216784	0.450366
std	4.090836	2.228754
min	0	0
25%	0	0
50%	0	0
75%	0	0
max	37	37

PREDICTION OF INDICATORS OF CERVICAL CANCER USING BIOPSY TEST



RELATIONSHIP B/W BIOPSY AND SMOKING ATTRIBUTES



INFERENCES:

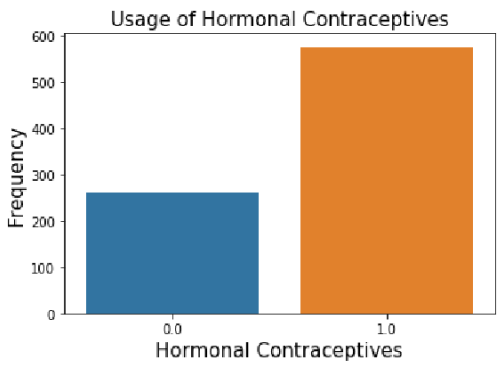
- Cancer can affect to non-smokers as well and even in the young age (around 22 years) the Biopsy test can be tested as positive.
- The person who smoke for at least one year are more prone to be test as positive in Biopsy test.
- The person who smoke a greater number of packets a year with more age are prone to be tested as positive in Biopsy test and the average age is found to be 30 years.

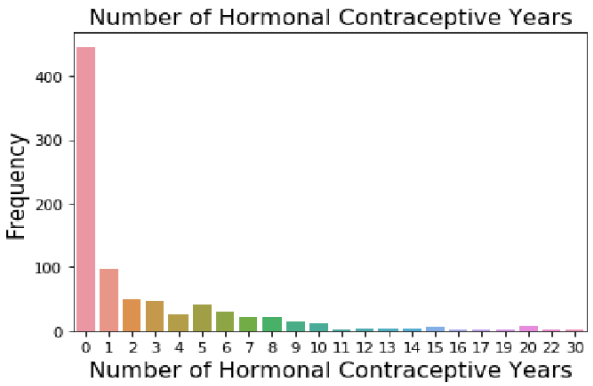
PREDICTION OF INDICATORS OF CERVICAL CANCER USING BIOPSY TEST

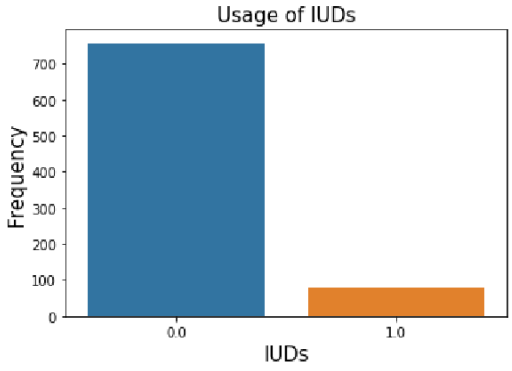
3.1.3. BIRTH CONTROL ATTRIBUTES

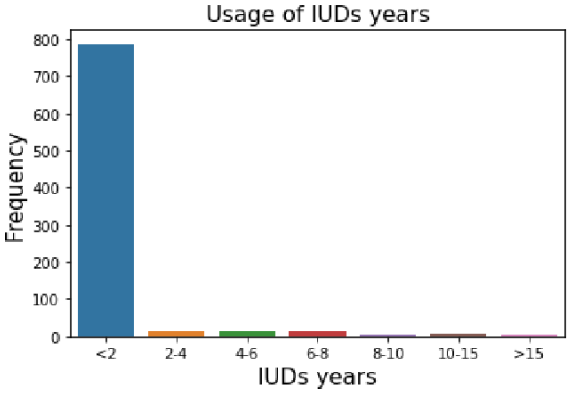
- Hormonal Contraceptives – Patient using Hormonal Contraceptives (Birth control methods that act on the endocrine system. Almost all methods are composed of steroid hormones).
- Hormonal Contraceptives (years) – Number of years patient using Hormonal contraceptives.
- IUD – Patient using IUD (A small, often T-shaped birth control device that is inserted into a woman's uterus to prevent pregnancy).
- IUD (years) – Number of years patient using IUD.

DISTRIBUTION:



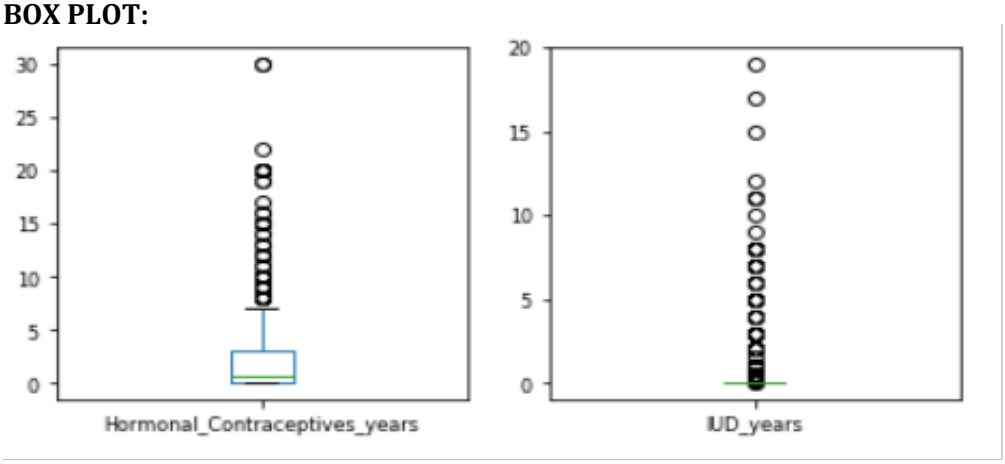




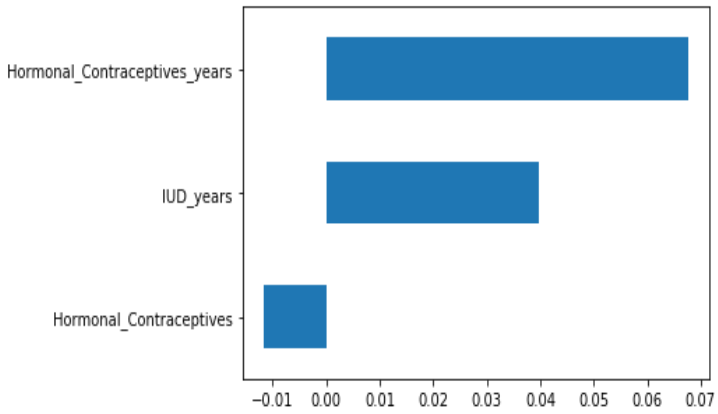


PREDICTION OF INDICATORS OF CERVICAL CANCER USING BIOPSY TEST

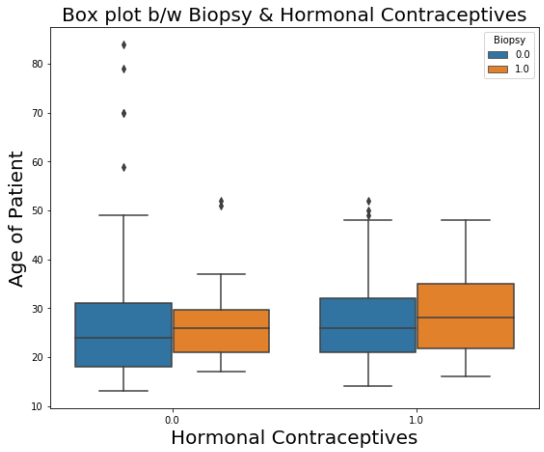
BOX PLOT:

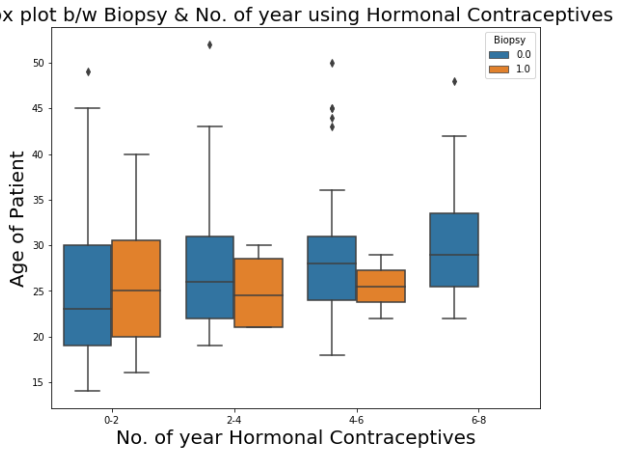


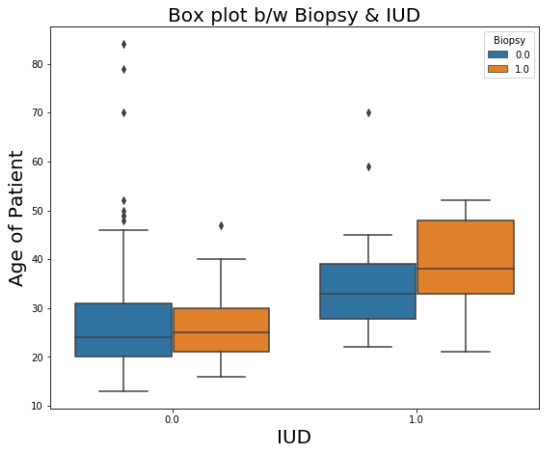
RELATION B/W SCHILLER AND BIRTH CONTROL ATTRIBUTES:

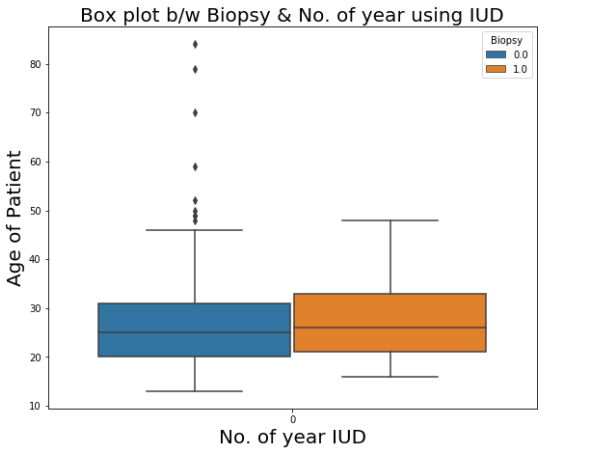


PREDICTION OF INDICATORS OF CERVICAL CANCER USING BIOPSY TEST









INFERENCES:

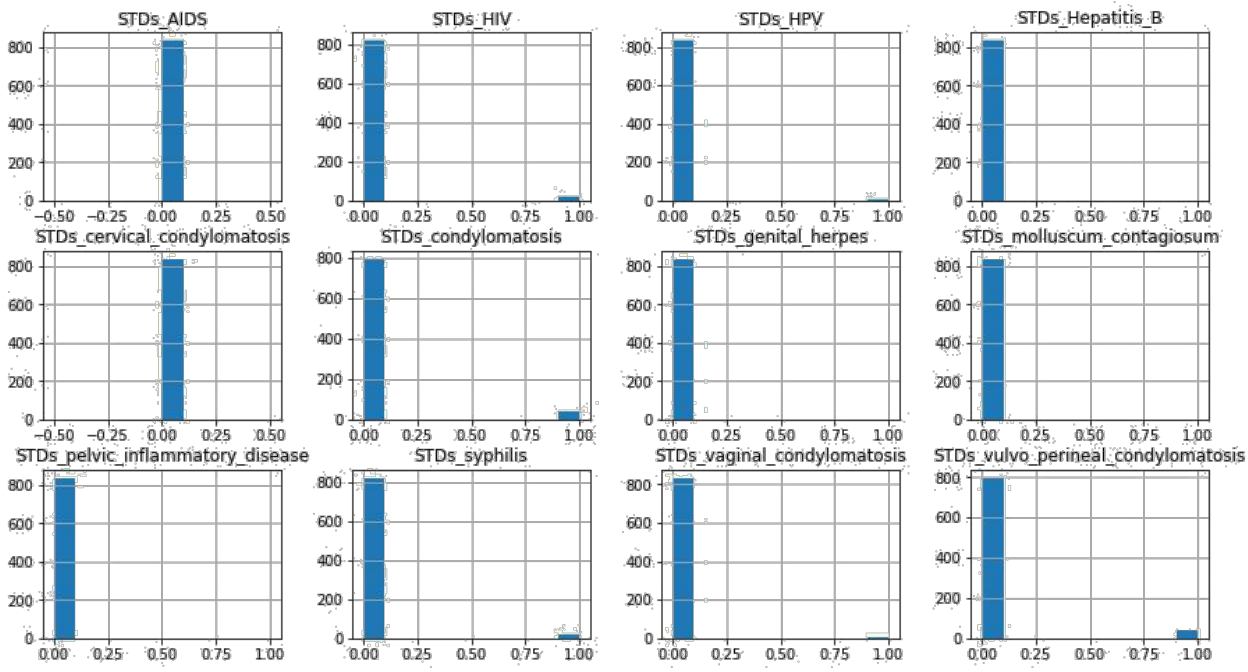
- We see that women prefer Hormonal Contraceptives when compared to the Intra Uterine Contraceptive Devices as it is safer than the latter.
- Most of the women have used the contraceptives in the period of less than 2 years.
- The persons who did not use the hormonal contraceptives and with more age are high in number, who show positive for Biopsy test
- The patients with 2 years of usage in hormonal contraceptives and the average age between 25yrs & 35yrs show positive for Biopsy test.
- The persons who did not use IUD and with age 25 years & 35 years, show positive for Biopsy test.

PREDICTION OF INDICATORS OF CERVICAL CANCER USING BIOPSY TEST

3.1.3. STDs ATTRIBUTES

- Condylomatosis - A disease of sexual transmission caused by the virus of papilloma.
- Cervical condylomatosis - Type of Condylomatosis.
- Vaginal condylomatosis - Type of Condylomatosis.
- Vulvo-perineal condylomatosis - Type of Condylomatosis.
- Syphilis - A bacterial infection usually spread by sexual contact.
- Pelvic inflammatory disease - An infection of the female reproductive organs.
- Genital herpes - An infection marked by genital pain and sores.
- Molluscum contagiosum - A viral skin infection.
- AIDS - A disease which lowers the immunity.
- HIV - HIV causes AIDS and interferes with the body's ability to fight infections.
- Hepatitis B - A serious liver infection caused by the hepatitis B virus.
- HPV - An infection that causes warts in various parts of the body.

DISTRIBUTION:

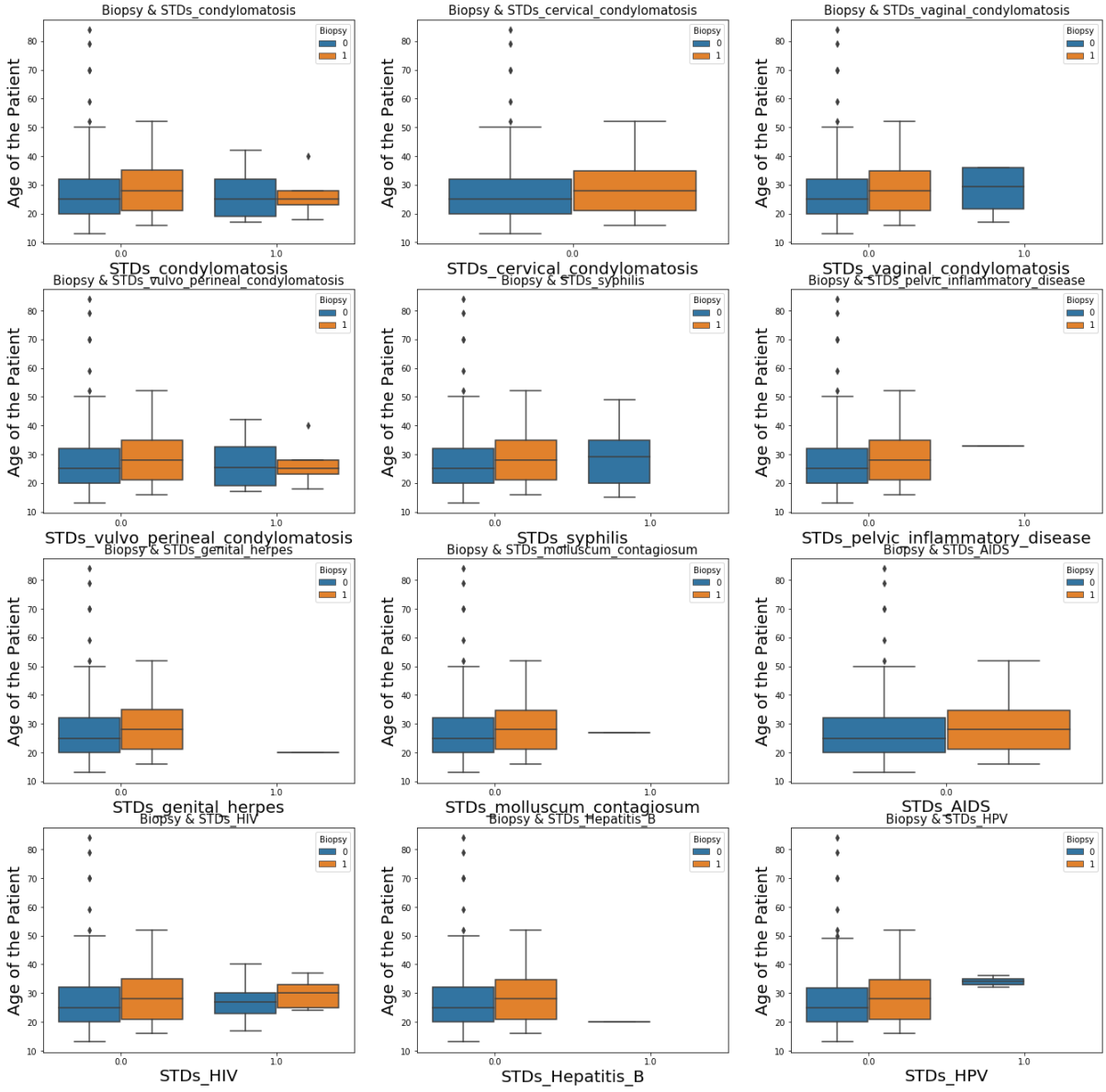


PREDICTION OF INDICATORS OF CERVICAL CANCER USING BIOPSY TEST

Attribute	0 (%)	1(%)
STDs_condylomatosis	0.95	0.05
STDs_cervical_condylomatosis	1	0
STDs_vaginal_condylomatosis	0.995	0.005
STDs_vulvo_perineal_condylomatosis	0.951	0.049
STDs_syphilis	0.98	0.02
STDs_pelvic_inflammatory_disease	0.999	0.001
STDs_genital_herpes	0.999	0.001
STDs_molluscum_contagiosum	0.999	0.001
STDs_AIDS	1	0
STDs_HIV	0.98	0.02
STDs_Hepatitis_B	0.999	0.001
STDs_HPV	0.998	0.002

PREDICTION OF INDICATORS OF CERVICAL CANCER USING BIOPSY TEST

RELATION B/W STDS ATTRIBUTES AND SCHILLER:



INFERENCES:

- We can see from the distribution plot that STD cervical condolymatosis and STD AIDS have no effect on Biopsy Test and hence variables will be of least importance when predicting the Cancer using Biopsy test.

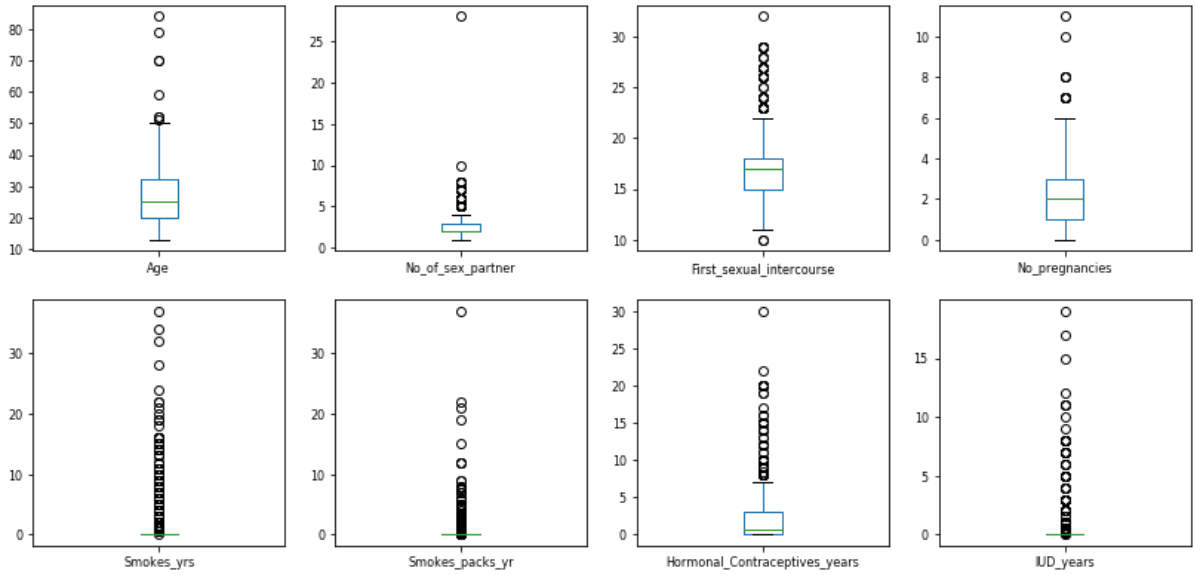
PREDICTION OF INDICATORS OF CERVICAL CANCER USING BIOPSY TEST

- In all the plots we can clearly the person with more age are prone to be tested as Positive in Biopsy test.
- We cannot predict the effect of individual STDs on Biopsy test due to class imbalance in the dataset of each STD attributes.

Outlier Detection and Treatment

OUTLIER DETECTION:

Outlier is an observation that appears far away and diverges from an overall pattern in a sample.



A box plot (box-and-whisker plot) shows the distribution of quantitative data in a way that facilitates comparisons between variables. The box shows the five-point summary of each numerical attributes. The above-mentioned numerical columns outliers are treated by using the IQR(Inter-quartile range).

PREDICTION OF INDICATORS OF CERVICAL CANCER USING BIOPSY TEST

IQR = 75% quartile – 25% quartile.

Range = (25% quartile -1.5*IQR) to (75% quartile +1.5*IQR).

Range of Age b/w 2.0 and 50.0

Range of number of sex partner b/w 0.5 and 4.5

Range of first sexual intercourse age b/w 10.5 and 22.5

Range of number of pregnancies b/w -2.0 and 6.0

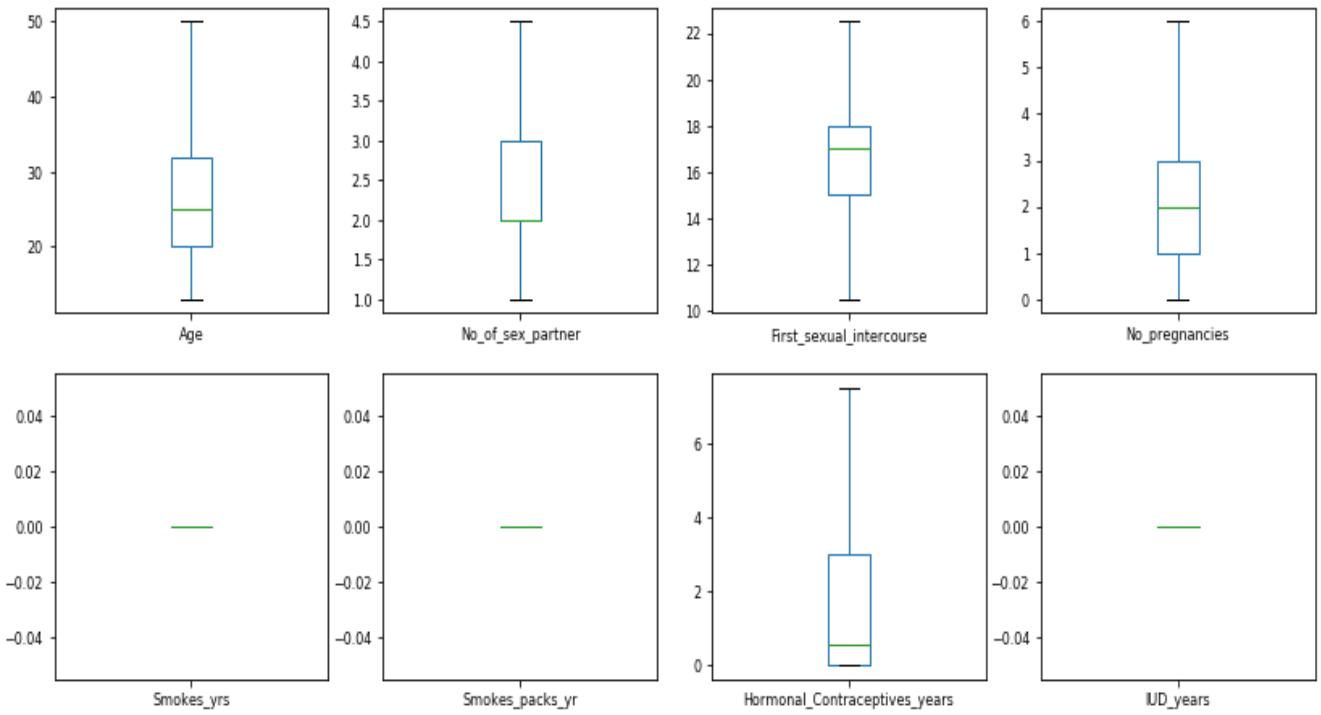
Range of Smoking years b/w 0.0 and 0.0

Range of smoking packets per year b/w 0.0 and 0.0

Range of Hormonal contraceptive usage years b/w -4.5 and 7.5

Range of IUD usage years b/w 0.0 and 0.0

AFTER OUTLIER TREATMENT



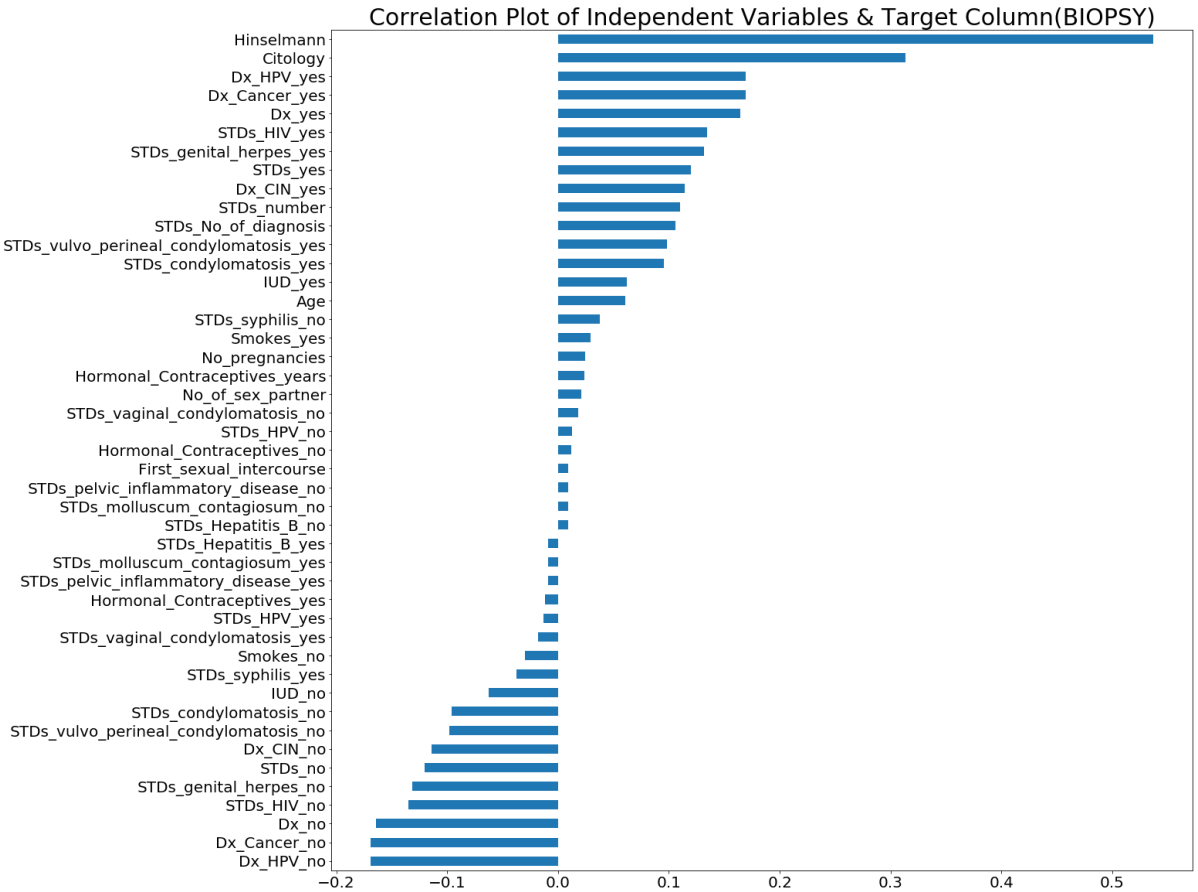
PREDICTION OF INDICATORS OF CERVICAL CANCER USING BIOPSY TEST

EDA - Correlation Plot

Correlation analysis measures the statistical relationship between two different variables. The correlation coefficient is bound between -1 and 1 and tells you the linear relationship between these two variables.

A coefficient close to 1 means a strong and positive association between the two variables.

A coefficient close to -1 means strong negative association between the two variables



The independent attributes correlation with the target column is not very strong and the maximum correlation value is around 0.5. Heinselmann, Cytology are the highly influencing attributes of Biopsy column.

PREDICTION OF INDICATORS OF CERCIVAL CANCER USING BIOPSY TEST

Chi-Square Analysis:

A chi-square test for independence compares two variables in a contingency table to see if they are related.

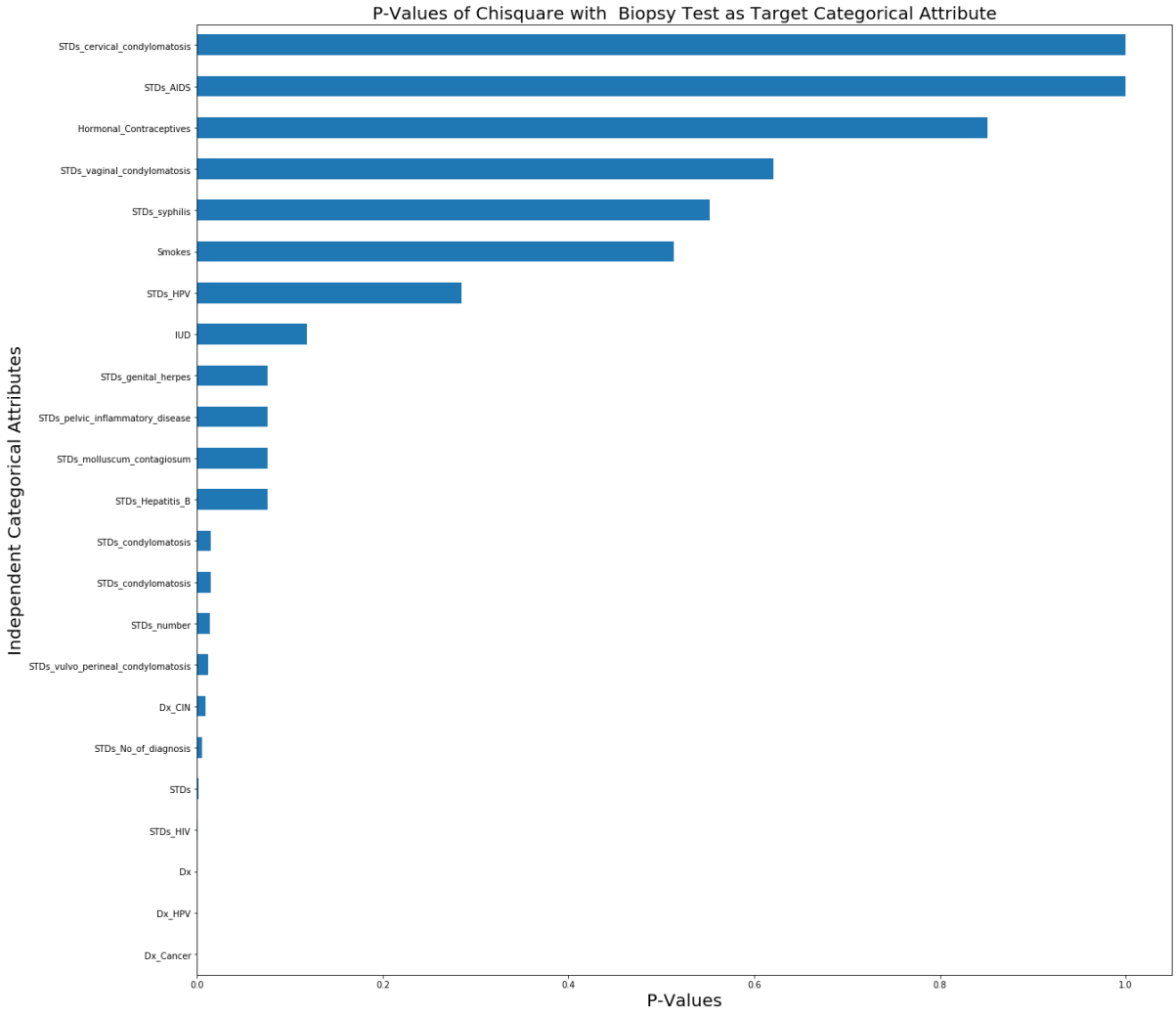
$$c^2 = \sum_{i=1}^{k} \left[\frac{(O_i - E_i)^2}{E_i} \right]$$

here “O” is the Observed value, “E” is the expected value and “i” is the “ith” position in the table.

A chi square test will give you a p-value. The p-value will tell you if your test results are significant or not.

p<0.05 – Two variables are dependent.

p>0.05 – Two variables are independent.



PREDICTION OF INDICATORS OF CERVICAL CANCER USING BIOPSY TEST

MODEL BUILDING:

- The given problem statement is classification problem. So, four base classification algorithms are taken:
  - Logistic Regression
  - Decision Trees (CART)
  - Naïve Bayes.
  - K Nearest Neighbours.

BASE MODEL COMPARISION

The given dataset is divided into train data and test data for each of the mentioned base models. Train dataset is used for fitting the model and test dataset is used to check how the fitted model predicts the test target column.

This is a classification problem, Base Classifier is imported. The model is fitted on the train independent attributes and dependent labels. Then the trained classifier/ model is used to predict the labels of the test attributes.

A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known. The matrix is generated after making predictions on the test data and then obtaining the values of True Positives, True Negatives, False Positives and False Negatives. The F1 score can be interpreted as a weighted average of the precision and recall.

The following table indicates the model performance of the different base models.

	Model	accuracy	f1score	recall	precision	ruc_auc
0	LogisticRegression	0.944444	0.416667	0.333333	0.555556	0.6582281
1	DecisionTree	0.916667	0.461538	0.600000	0.375000	0.7683542
2	RandomForest	0.932540	0.370370	0.333333	0.416667	0.6518993
3	GaussianNB	0.071429	0.113636	1.000000	0.060241	0.5063294
4	KNN	0.940476	0.0000000	0.0000000	0.000000	0.500000