

Cervical Cancer Analysis Using Machine Learning

line 1: 2nd Given Name Surname
line 2: dept. name of organization
(of Affiliation)
line 3: name of organization
(of Affiliation)
line 4: City, Country
line 5: email address or ORCID

Abstract-- Cervical cancer is a common tumor in the female reproductive system and is responsible for death. Identifying cervical cancer is very tough because sometimes on time we are not able to predict it so machine learning plays an important role in performing the analysis on the cervical cancer dataset. Papillomavirus and some other factors are responsible for the cervical cancer. In this research analysis, our main goal is to perform the analysis on the cervical data with the help of machine learning algorithms to find accurate results and understand the behavior of features to treat cervical cancer on time. Logistic Regression, Random Forest, and decision tree algorithms will be implemented in research to predict cervical cancer. The result shows that the logistic regression model achieves an accuracy of 99% which helps in the future to treat cervical cancer on time.

I. INTRODUCTION

Cervical cancer is a disease that occurs when the cervix cell, the uterus's lower part, and some abnormal changes and then create precancerous cells which directly turn into cancer cells. The treatment of cervical cancer contains chemotherapy and surgical resection. There are various risk factors for cervical cancer such as smoking, sexual intercourse, infection with genital herpes, and poor genital hygiene. One of the most important symptoms of cervical cancer is irregular bleeding which contains brownish discharge, bleeding during sexual intercourse, and spotting. According to research analysis, cervical analysis is the 4th most frequent cancer in women by the World Health Organization. There is the main cause of cervical cancer is the lack of early treatment.

In most of the developed countries, cervical cancer disease treatment taking a good path because it decreases the death rate by 4% to 5% every year. Screening test helps clinicians to treat the precancerous lesions in the initial stage only so that can prevent malignant tumors. However, in developing countries, cervical cancer continues at risk because of minimal resources, services not available, proper preventive methods not available, and even people who are also not aware of cervical cancer.

Machine Learning shows a proper involvement in the research field day by day which helps doctors to analyze and predict the diagnosis on time and provide proper treatment. In this research analysis, we also implemented the supervised learning algorithm to predict cervical cancer based on given features present in the dataset so that proper diagnosis can be identified. It is helpful for early detection of cervical cancer. To perform the data analysis, the first step is data collection, the second step is data cleaning, the third step is data visualization, and the last step is model evaluation. We have performed all the steps to predict the symptoms of cervical cancer using machine learning for perfect diagnosis which helps to reduce the death in developing countries as well.

II. LITERATURE REVIEW

According to Tanimu, They have implemented the supervised learning algorithms and in that, logistic regression and decision tree provide better performance scores as compared to the other 6 implemented models. They achieve performance scores between 80% to 90% which is helpful for doctors to use the approach and treat the patient on time.

III. CONSIDERATION OF ETHICAL, LEGAL, AND SOCIAL CONCERNS

- It is very important to ensure that the implementation of cervical cancer does not affect the existing health disparities. Patients should have a full right to be informed properly. Healthcare organizations should have to build trust regarding patient data misuse by applying machine learning and artificial intelligence.
- Whatever artificial models we are implementing, directly affect the role of professionals of healthcare so we have to ensure to be managed properly.
- We have to follow proper legal factors such as HIPAA and GDPR for patient privacy and security.

IV. DATA COLLECTION

Cervical cancer analysis is a very important analysis nowadays for better treatment. We have collected the data from open-source Kaggle and read it using the Python library.

```
In [227]: #read dataset
import pandas as pd
cancer = pd.read_csv('cancer_cervical.csv')
cancer.head()
```

Out[227]:

	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
0	18	4	15	1	0	0	0	0	0	0	...	?
1	15	1	14	1	0	0	0	0	0	0	...	?
2	34	1	?	1	0	0	0	0	0	0	...	?
3	52	5	16	4	1	37	37	1	3	0	...	?
4	48	3	21	4	0	0	0	1	15	0	...	?

5 rows x 36 columns

Figure 1: Cervical Cancer Data

There are 968 patient data present with 36 symptoms and attributes.

```
In [228]: #check rows and columns
cancer.shape
```

Out[228]: (968, 36)

Figure 2: Rows and Columns

We have checked the types of data columns using the Python method and according to it, most of the columns are of object type but contain some integer value.

```
#check data types
cancer.dtypes
```

]: Age int64
Number of sexual partners object
First sexual intercourse object
Num of pregnancies object
Smokes object
Smokes (years) object
Smokes (packs/year) object
Hormonal Contraceptives object
Hormonal Contraceptives (years) object
IUD object
IUD (years) object
STDs object
STDs (number) object
STDs:condylomatosis object
STDs:cervical condylomatosis object
STDs:vaginal condylomatosis object
STDs:vulvo-perineal condylomatosis object
STDs:syphilis object
STDs:pelvic inflammatory disease object
STDs:genital herpes object
STDs:molluscum contagiosum object
STDs:AIDS object
STDs:HIV object
STDs:Hepatitis B object
STDs:HPV object
STDs: Number of diagnosis int64
STDs: Time since first diagnosis object
STDs: Time since last diagnosis object
Dx:Cancer int64
Dx:CIN int64
Dx:HPV int64
Dx int64
Hinselmann int64
Schiller int64
Citology int64
Biopsy int64
dtype: object

Figure 3: Data Types

We have checked the information of the data and it gives non-null count, data types, total column number, and entry details.

```

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 968 entries, 0 to 967
Data columns (total 36 columns):
#   Column                                     Non-Null Count  Dtype
---  -
0   Age                                         968 non-null    int64
1   Number of sexual partners                 968 non-null    object
2   First sexual intercourse                  968 non-null    object
3   Num of pregnancies                        968 non-null    object
4   Smokes                                    968 non-null    object
5   Smokes (years)                           968 non-null    object
6   Smokes (packs/year)                      968 non-null    object
7   Hormonal Contraceptives                  968 non-null    object
8   Hormonal Contraceptives (years)          968 non-null    object
9   IUD                                       968 non-null    object
10  IUD (years)                              968 non-null    object
11  STDs                                     968 non-null    object
12  STDs (number)                           968 non-null    object
13  STDs:condylomatosis                      968 non-null    object
14  STDs:cervical condylomatosis             968 non-null    object
15  STDs:vaginal condylomatosis              968 non-null    object
16  STDs:vulvo-perineal condylomatosis       968 non-null    object
17  STDs:syphilis                            968 non-null    object
18  STDs:pelvic inflammatory disease         968 non-null    object
19  STDs:genital herpes                      968 non-null    object
20  STDs:molluscum contagiosum               968 non-null    object
21  STDs:AIDS                                968 non-null    object
22  STDs:HIV                                 968 non-null    object
23  STDs:Hepatitis B                         968 non-null    object
24  STDs:HPV                                 968 non-null    object
25  STDs: Number of diagnosis                968 non-null    int64
26  STDs: Time since first diagnosis          968 non-null    object
27  STDs: Time since last diagnosis          968 non-null    object
28  Dx:Cancer                                968 non-null    int64
29  Dx:CIN                                   968 non-null    int64
30  Dx:HPV                                   968 non-null    int64
31  Dx                                        968 non-null    int64
32  Hinselmann                              968 non-null    int64
33  Schiller                                 968 non-null    int64
34  Citology                                 968 non-null    int64
35  Biopsy                                   968 non-null    int64

```

Figure 4: Data Information

In data information, we check values are present in integer format but its data type is object. So, we calculated the unique values in each value like which type of data sets are present.

	N-Unique	Unique
Age	44	[18, 15, 34, 52, 46, 42, 51, 26, 45, 44, 27, 4...
Number of sexual partners	13	[4, 1, 5, 3, 2, 6, ?, 7, 15, 8, 10, 28, 9]
First sexual intercourse	22	[15, 14, ?, 16, 21, 23, 17, 26, 20, 25, 18, 27...
Num of pregnancies	12	[1, 4, 2, 6, 3, 5, ?, 8, 7, 0, 11, 10]
Smokes	3	[0, 1, ?]
Smokes (years)	31	[0, 37, 34, 1.266872909, 3, 12, ?, 18, 7, 19, ...
Smokes (packs/year)	63	[0, 37, 3.4, 2.8, 0.04, 0.513202128, 2.4, 6, ?...
Hormonal Contraceptives	3	[0, 1, ?]
Hormonal Contraceptives (years)	41	[0, 3, 15, 2, 8, 10, 5, 0.25, 7, 22, 19, 0.5, ...
IUD	3	[0, 1, ?]
IUD (years)	27	[0, 7, ?, 5, 8, 6, 1, 0.58, 2, 19, 0.5, 17, 0...
STDs	3	[0, 1, ?]
STDs (number)	6	[0, 2, 1, ?, 3, 4]
STDs:condylomatosis	3	[0, 1, ?]
STDs:cervical condylomatosis	2	[0, ?]
STDs:vaginal condylomatosis	3	[0, ?, 1]
STDs:vulvo-perineal condylomatosis	3	[0, 1, ?]
STDs:syphilis	3	[0, 1, ?]
STDs:pelvic inflammatory disease	3	[0, ?, 1]
STDs:genital herpes	3	[0, ?, 1]
STDs:molluscum contagiosum	3	[0, ?, 1]
STDs:AIDS	2	[0, ?]
STDs:HIV	3	[0, 1, ?]
STDs:Hepatitis B	3	[0, ?, 1]
STDs:HPV	3	[0, ?, 1]
STDs: Number of diagnosis	4	[0, 1, 3, 2]

Figure 5: Data Uniqueness

There are various symptoms are present according to the age gap and symptom details are ['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)', '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', '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']

```
In [234]: #check columns
cancer.columns

Out[234]: Index(['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)', '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',
               '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'],
              dtype='object')
```

Figure 6: Column in Cancer Data

V. DATA CLEANING

Data cleaning is a part of data analysis. So, we checked the null values here using the Python function and it provides there are no missing values. At the time of unique value analysis, we identified that in most of the columns and rows, the question mark is present in place of the value which cannot come under the missing values but we have to remove it. To handle the question mark, for those columns, we replace the question mark with the NAN value. Now it will show the missing values in some columns.

```
In [235]: # Identify columns containing "?" values
replace_column = ['Number of sexual partners', 'First sexual intercourse',
                  'Num of pregnancies', 'Smokes', 'Smokes (years)',
                  'Hormonal Contraceptives', 'Hormonal Contraceptives (years)', 'IUD',
                  'IUD (years)', 'STDs', 'STDs (number)', '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',
                  'STDs:Hepatitis B', 'STDs:HPV', 'STDs: Number of diagnosis',
                  'STDs: Time since first diagnosis', 'STDs: Time since last diagnosis']

# Replace "?" with NaN (missing value)
cancer[replace_column] = cancer[replace_column].replace('?', pd.NA)

In [236]: #again check missing values
cancer.isnull().sum()

Out[236]: Age                                0
Number of sexual partners                    29
First sexual intercourse                     10
Num of pregnancies                           70
Smokes                                       16
Smokes (years)                             16
Smokes (packs/year)                         16
Hormonal Contraceptives                     122
Hormonal Contraceptives (years)             122
IUD                                           133
IUD (years)                                 133
STDs                                          122
STDs (number)                              122
STDs:condylomatosis                         122
STDs:cervical condylomatosis                122
STDs:vaginal condylomatosis                 122
STDs:vulvo-perineal condylomatosis          122
STDs:syphilis                              122
STDs:pelvic inflammatory disease            122
```

Figure 7: Data Cleaning

2 columns contain only one type of value and there are other 2 columns which contain the maximum amount of missing values. So we decided to drop these types of columns from the dataset using the drop function.

```
In [237]: #drop unwanted columns
cancer = cancer.drop(['STDs: Time since first diagnosis', 'STDs: Time since last diagnosis',
                     'STDs:cervical condylomatosis', 'STDs:AIDS'],axis=1)
cancer.head()

Out[237]:
```

	Age	Number of sexual partners	First sexual intercourse	Num of pregnancies	Smokes	Smokes (years)	Smokes (packs/year)	Hormonal Contraceptives	H
0	18	4	15	1	0	0	0	0	
1	15	1	14	1	0	0	0	0	
2	34	1	<NA>	1	0	0	0	0	
3	52	5	18	4	1	37	37	1	
4	46	3	21	4	0	0	0	1	

5 rows x 32 columns

Figure 8: Drop Columns

In other columns, some missing values are there so we dropped the missing rows from the dataset.

```
In [238]: #drop missing rows
cancer.dropna(inplace=True)
```

```
In [239]: cancer.info()

<class 'pandas.core.frame.DataFrame'>
Int64Index: 743 entries, 0 to 967
Data columns (total 32 columns):
```

Figure 9: Drop Rows

Duplicate data are not present in our dataset.

```
In [240]: #check duplicated data
cancer.duplicated().count()

Out[240]: 743
```

Figure 10: Check Duplicates

There are most of the columns are of object type so we convert the data type into float and integer according to the unique value analysis.

```
In [241]: #data transformation
cancer[['Number of sexual partners', 'First sexual intercourse', 'Num of pregnan
'STDs:vaginal condylomatosis', 'STDs:vulvo-perineal condylomatosis', 'STDs:
'STDs:pelvic inflammatory disease', 'STDs:genital herpes', 'STDs:mollusc
'Num of pregnancies', 'Smokes','Hormonal Contraceptives', 'IUD', 'STDs', 'S
'STDs:vaginal condylomatosis', 'STDs:vulvo-perineal condylomatosis', 'STDs:
'STDs:pelvic inflammatory disease', 'STDs:genital herpes', 'STDs:molluscum
'STDs:Hepatitis B', 'STDs:HPV']].astype(int)
cancer[['Smokes (years)', 'Smokes (packs/year)', 'Hormonal Contraceptives (years)
['Smokes (years)', 'Smokes (packs/year)',
'Hormonal Contraceptives (years)', 'IUD (years)']].astype(float)

In [242]: cancer.info()

<class 'pandas.core.frame.DataFrame'>
Int64Index: 743 entries, 0 to 967
Data columns (total 32 columns):
#   Column                                     Non-Null Count  Dtype
---  -
0   Age                                         743 non-null    int64
1   Number of sexual partners                 743 non-null    int32
2   First sexual intercourse                  743 non-null    int32
3   Num of pregnancies                        743 non-null    int32
4   Smokes                                    743 non-null    int32
5   Smokes (years)                           743 non-null    float64
6   Smokes (packs/year)                     743 non-null    float64
7   Hormonal Contraceptives                  743 non-null    int32
8   Hormonal Contraceptives (years)         743 non-null    float64
9   IUD                                       743 non-null    int32
10  IUD (years)                              743 non-null    float64
11  STDs                                     743 non-null    int32
12  STDs (number)                           743 non-null    int32
13  STDs:condylomatosis                     743 non-null    int32
14  STDs:vaginal condylomatosis             743 non-null    int32
15  STDs:vulvo-perineal condylomatosis      743 non-null    int32
16  STDs:syphilis                           743 non-null    int32
17  STDs:pelvic inflammatory disease        743 non-null    int32
18  STDs:genital herpes                     743 non-null    int32
```

Figure 11: Data Transformation

Further, we have a plot heat map so according to it, some columns are not highly relatable to each other and are not contributing to increasing the accuracy score. So, we drop those columns as well from the data.

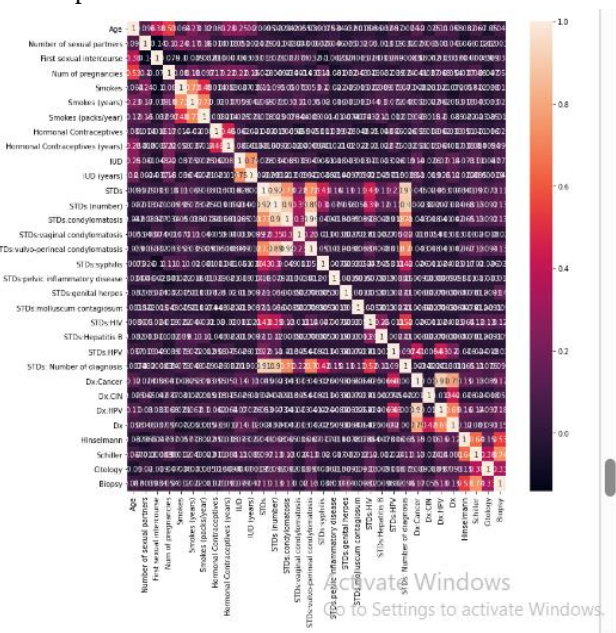


Figure 12: Heat map

```
In [245]: #drop unwanted columns
cancer = cancer.drop(['Dx:CIN', 'STDs:condylomatosis',
'STDs:vaginal condylomatosis',
'STDs:vulvo-perineal condylomatosis', 'STDs:syphilis',
'STDs:pelvic inflammatory disease', 'STDs:genital herpes',
'STDs:molluscum contagiosum'],axis=1)
cancer.head()
```

Out[245]:

	Age	Number of sexual partners	First sexual intercourse	Num of pregnancies	Smokes	Smokes (years)	Smokes (packs/year)	Hormonal Contraceptives	H
0	18	4	15	1	0	0.0	0.0	0	
1	15	1	14	1	0	0.0	0.0	0	
3	52	5	16	4	1	37.0	37.0	1	
4	46	3	21	4	0	0.0	0.0	1	
5	42	3	23	2	0	0.0	0.0	0	

5 rows x 24 columns

Figure 13: Data Preprocessing

VI. DATA VISUALIZATION

Data visualization is a type of exploratory data analysis to represent the information in charts and plots.

We have plotted the line chart to show the mean value of cervical cancer based on age.

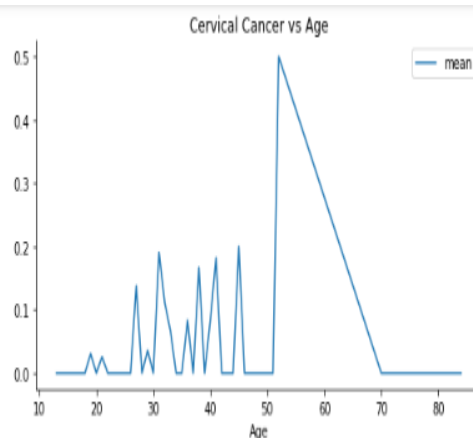


Figure 14: Cervical Cancer VS Age

We have plotted the line chart to show the mean value of cervical cancer based on a number of sexual partners.

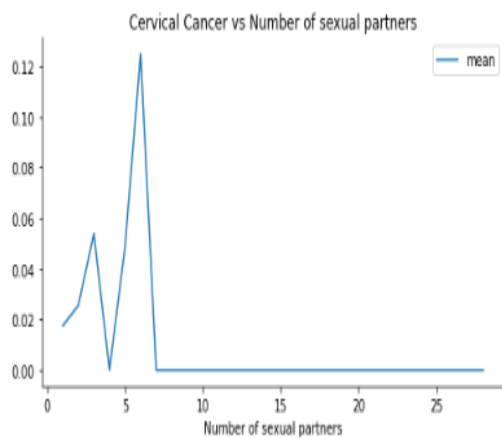


Figure 15: Cervical Cancer vs. Number of Sexual Partners

We have plotted the line chart to show the mean value of cervical cancer based on first sexual intercourse.

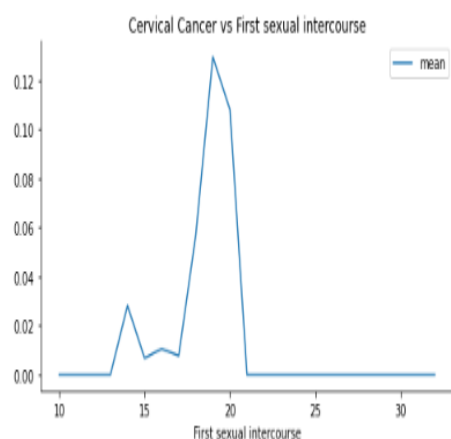


Figure 16: Cervical Cancer VS First Sexual Intercourse

We have plotted the line chart to show the mean value of cervical cancer based on number of pregnancies.

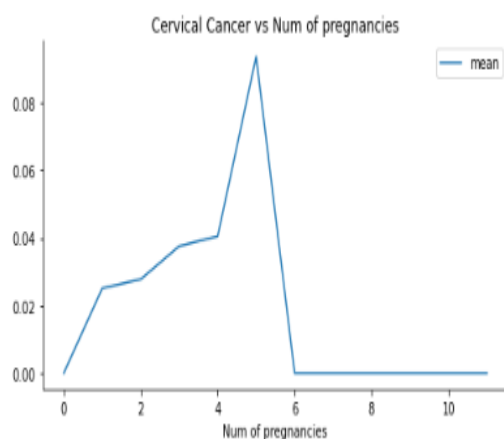


Figure 17: Cervical Cancer VS Number of Pregnancies

We have plotted the line chart to show the mean value of cervical cancer based on smoking.

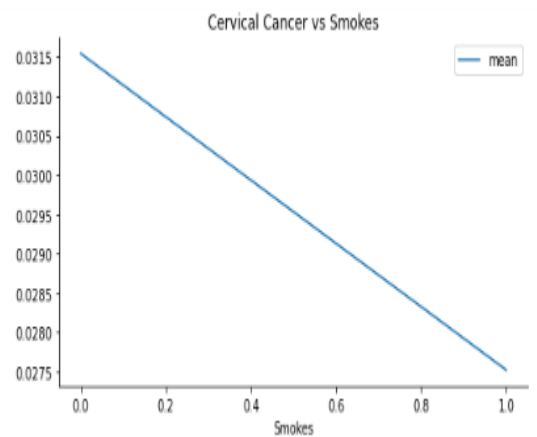


Figure 18: Cervical Cancer VS Smokes

We've generated a line chart illustrating the average cervical cancer incidence corresponding to annual smoking rates.

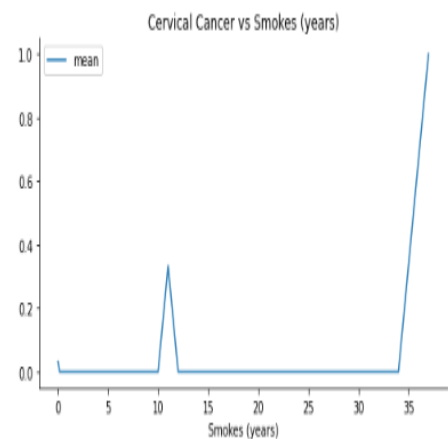


Figure 19: Cervical Cancer VS Smokes per year

We've generated a line chart illustrating the average cervical cancer incidence corresponding to smoke packs per year.

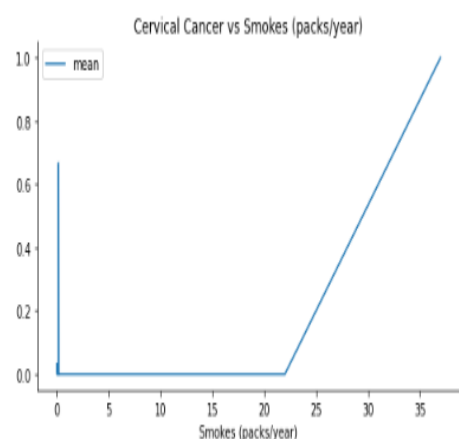


Figure 20: Cervical Cancer VS smokes pack per year

We've generated a line chart illustrating the average cervical cancer incidence corresponding to hormonal contraceptives.

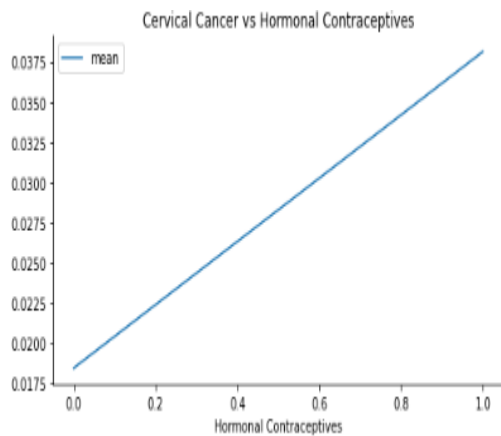


Figure 21: Cervical Cancer VS Hormonal Contraceptive

We've generated a line chart illustrating the average cervical cancer incidence corresponding to hormonal contraceptives per year.

We've generated a line chart illustrating the average cervical cancer incidence corresponding to IUD per year.

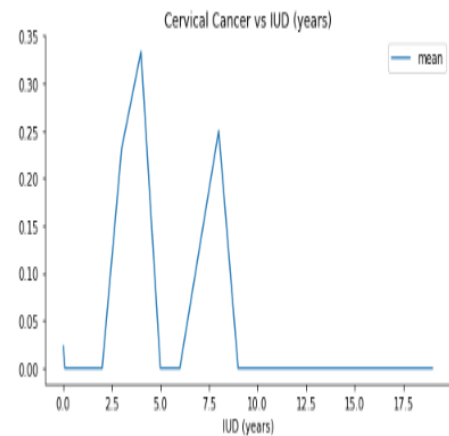


Figure 24: Cervical Cancer VS IUD per year

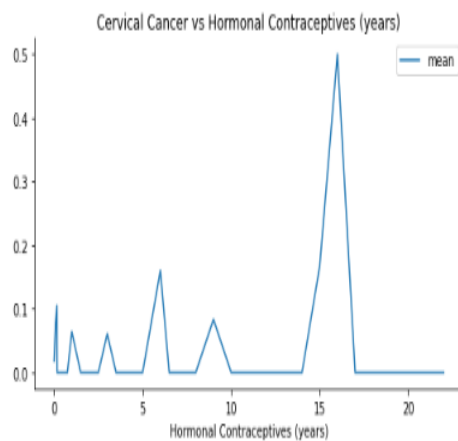


Figure 22: Cervical Cancer VS Hormonal Contraceptive per year

We've generated a line chart illustrating the average cervical cancer incidence corresponding to IUD.

We've generated a line chart illustrating the average cervical cancer incidence corresponding to STDs.

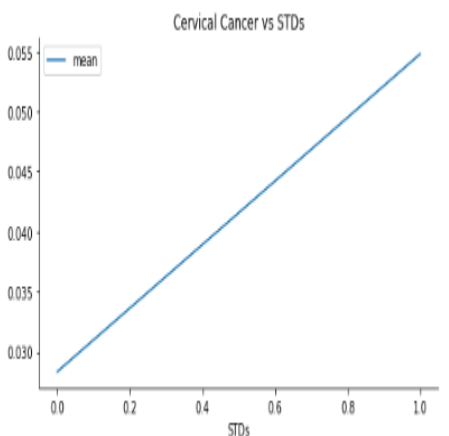


Figure 25: Cervical Cancer VS STD

We've generated a line chart illustrating the average cervical cancer incidence corresponding to STDs per year.

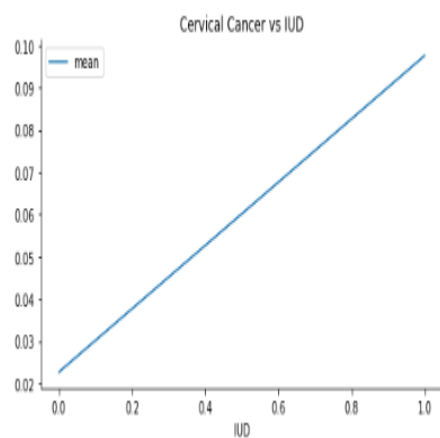


Figure 23: Cervical Cancer VS IUD

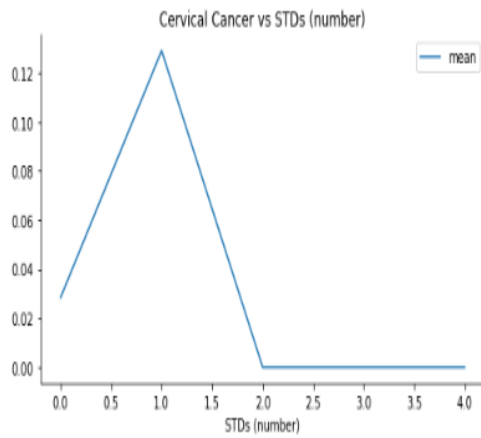


Figure 26: Cervical Cancer VS STD per year

We've generated a line chart illustrating the average cervical cancer incidence corresponding to HIV.

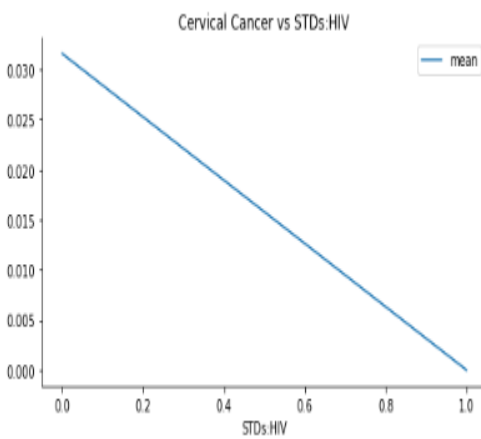


Figure 27: Cervical Cancer VS HIV

We've generated a line chart illustrating the average cervical cancer incidence corresponding to Hepatitis B.

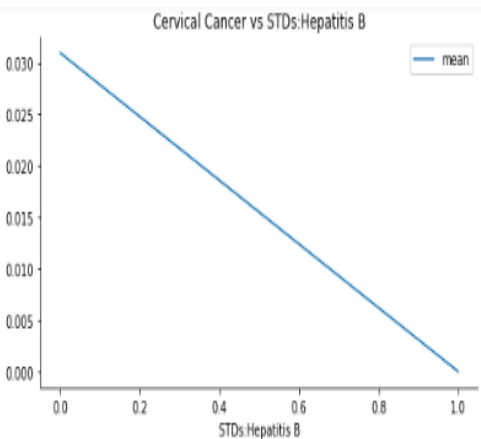


Figure 28: Cervical Cancer VS Hepatitis B

We've generated a line chart illustrating the average cervical cancer incidence corresponding to HPV.

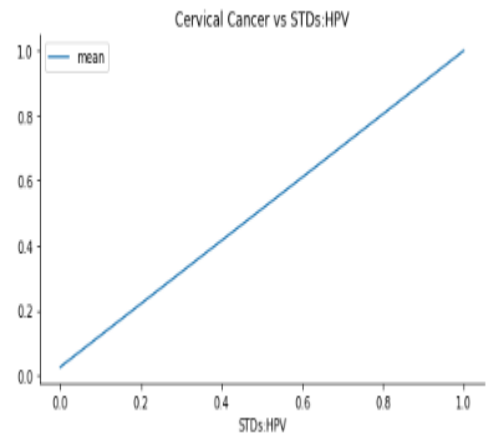


Figure 29: Cervical Cancer VS HPV

We've generated a line chart illustrating the average cervical cancer incidence corresponding to diagnoses.

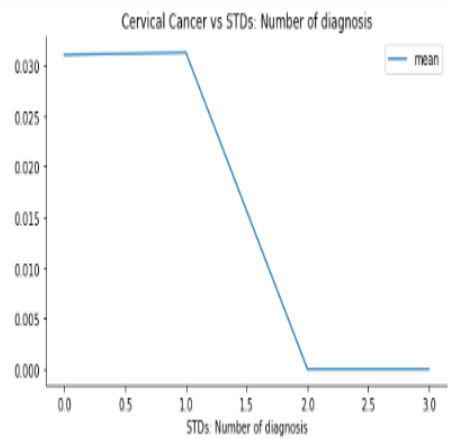


Figure 30: Cervical Cancer VS Number of Diagnoses

We've generated a line chart illustrating the average cervical cancer incidence corresponding to DX HPV.

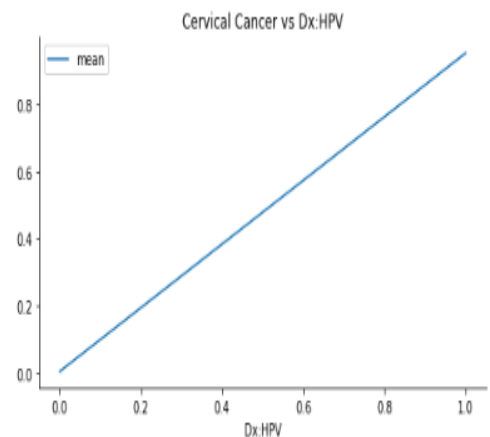


Figure 31: Cervical Cancer VS DX HPV

We've generated a line chart illustrating the average cervical cancer incidence corresponding to hinselmann.

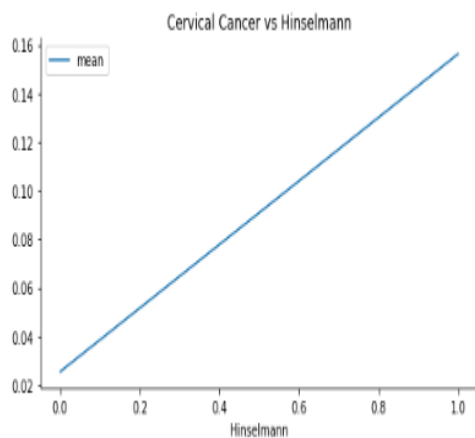


Figure 32: Cervical Cancer VS Hinselmann

We've generated a line chart illustrating the average cervical cancer incidence corresponding to schiller.

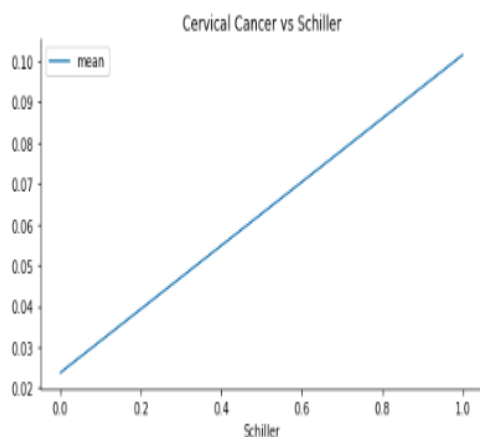


Figure 33: Cervical Cancer VS Schiller

We've generated a line chart illustrating the average cervical cancer incidence corresponding to citology.

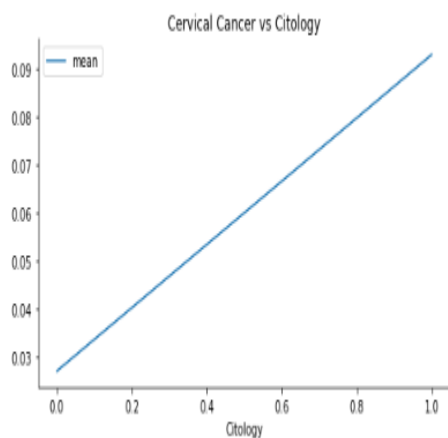


Figure 34: Cervical Cancer VS Citology

We've generated a line chart illustrating the average cervical cancer incidence corresponding to biopsy.

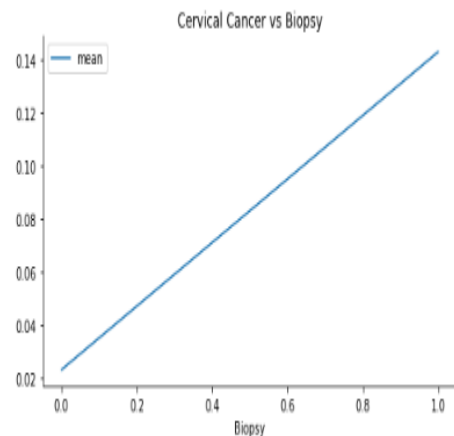


Figure 35: Cervical Cancer VS Biopsy

We've showcased a pie chart representing the age distribution.

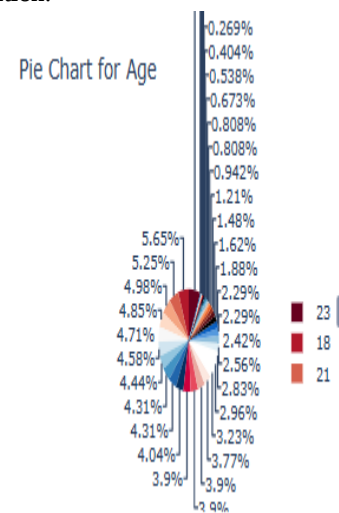


Figure 36: Pie Chart for age

We've showcased a pie chart representing the sexual partner distribution.

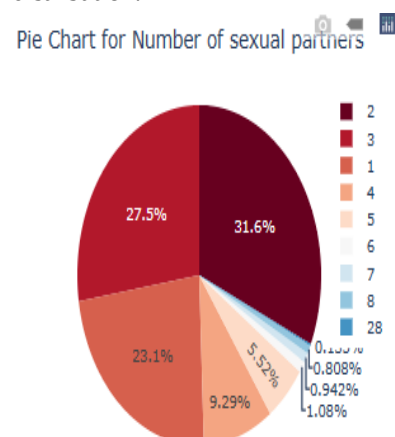


Figure 37: Pie chart for the number of sexual partners

We've showcased a pie chart representing the number of pregnancies distribution.

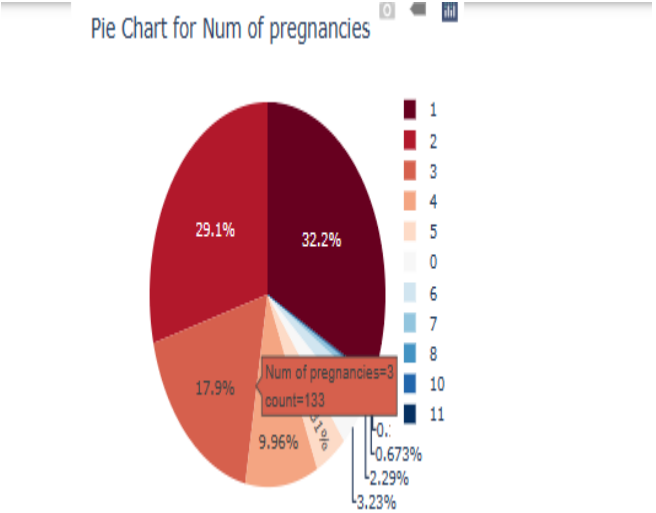


Figure 38: Pie chart for the number of pregnancies

We've showcased a pie chart representing the smokers.

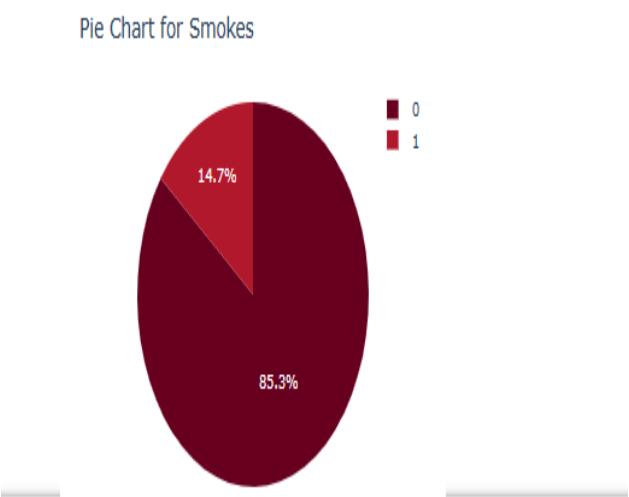


Figure 39: Pie chart for the smokers

We've showcased a pie chart representing the smokes per year distribution.

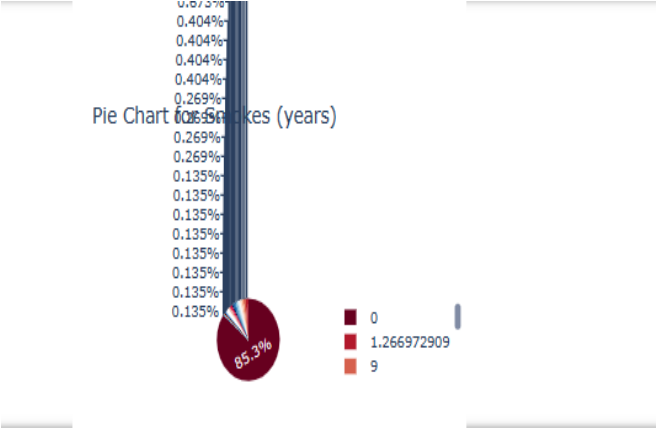


Figure 40: Pie chart for the smokes per year

We've showcased a pie chart representing the smoke packs per year distribution.

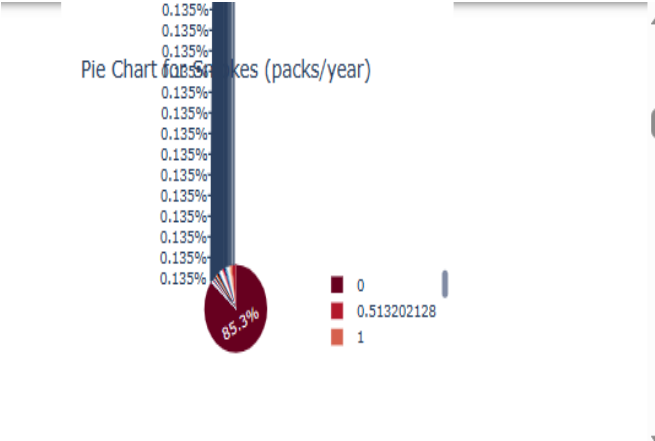


Figure 41: Pie chart for the smoke packs per year

We've showcased a pie chart representing the hormonal contraceptive distribution.

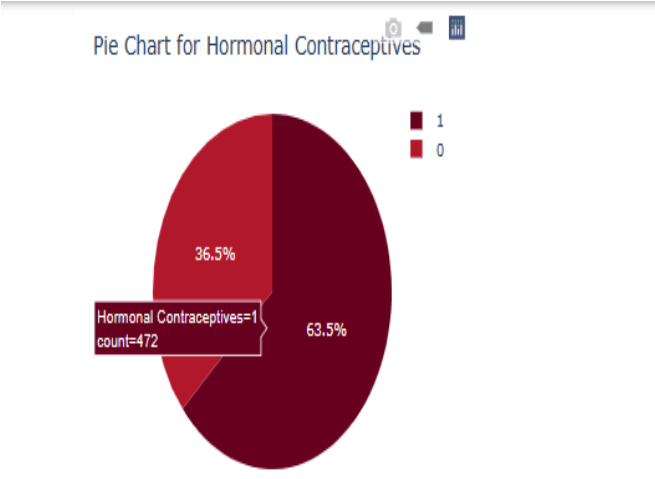


Figure 42: Pie chart for the hormonal contraceptive

We've showcased a pie chart representing the first sexual intercourse distribution.

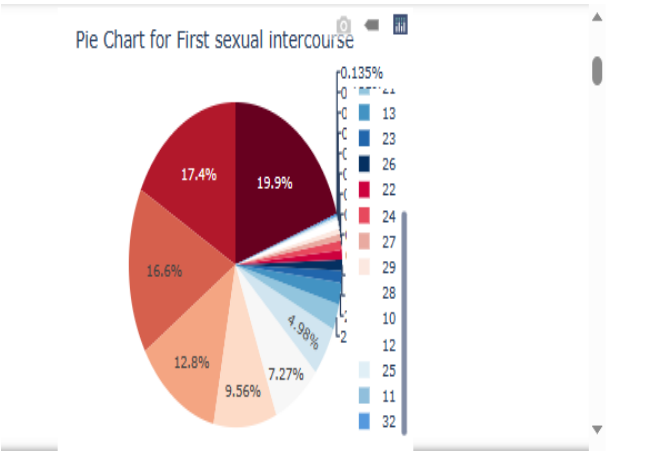


Figure 43: Pie chart for the first sexual intercourse

We have displayed the pie chart for the IUD attribute.

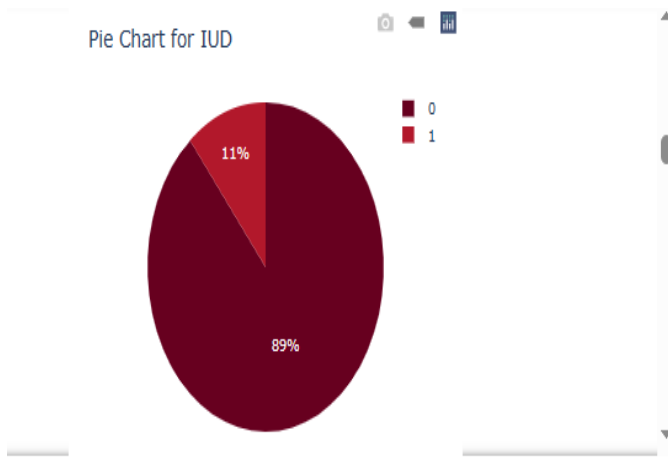


Figure 44: Pie Chart for IUD

We've showcased a pie chart representing the STD distribution.

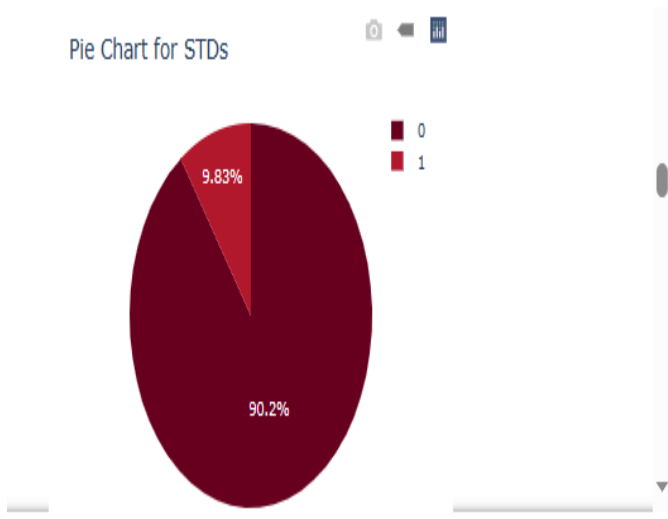


Figure 45: Pie Chart for STD

We've showcased a pie chart representing the biopsy distribution.

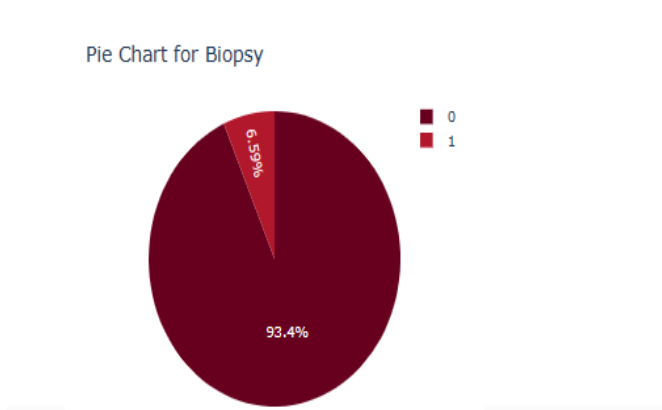


Figure 46: Pie Chart for Biopsy

We've showcased a pie chart representing the cytology distribution.

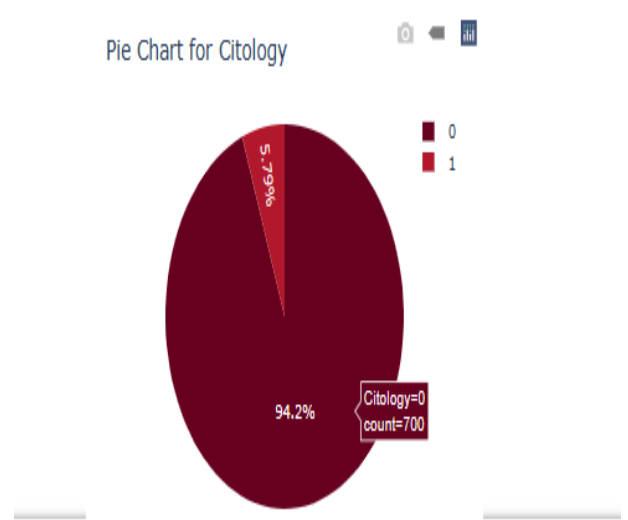


Figure 47: Pie Chart for Cytology

We've showcased a pie chart representing the schiller distribution.

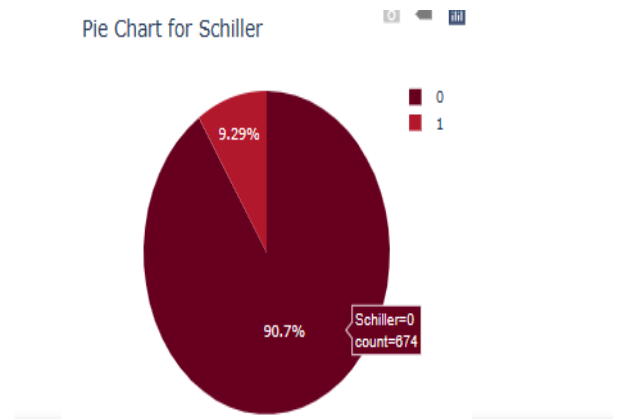


Figure 48: Pie Chart for Schiller

We've showcased a pie chart representing the hinselmann distribution.

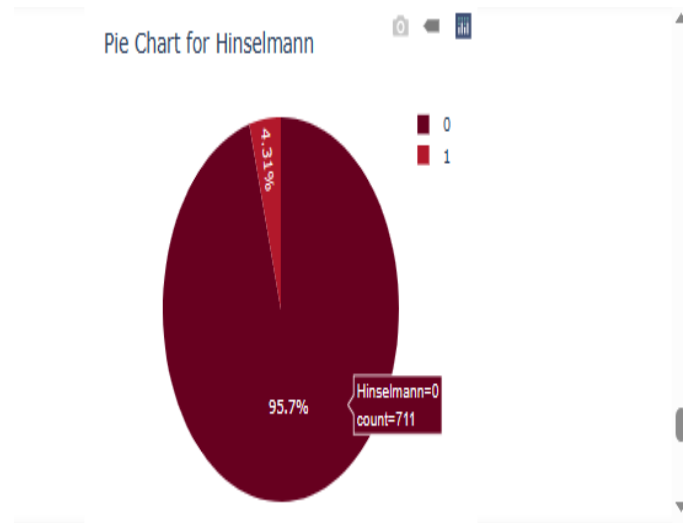


Figure 49: Pie Chart for Hinselmann

We've showcased a pie chart representing the cancer distribution.

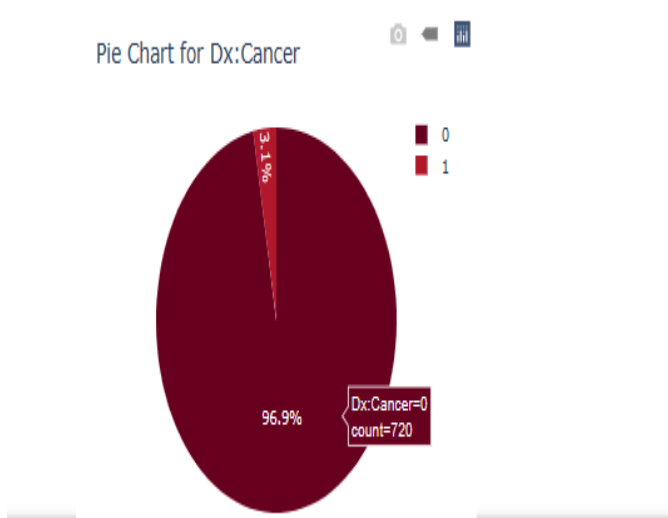


Figure 50: Pie Chart for DX Cancer

We have displayed the pie chart for the number of diagnoses attributes.

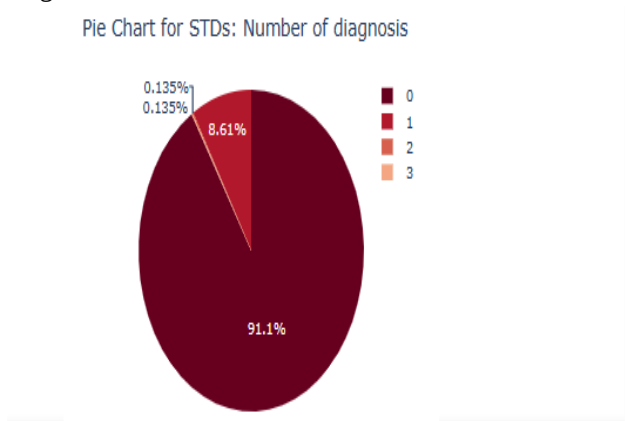


Figure 51: Pie chart for the number of diagnoses

We have displayed the pie chart for the hepatitis B attribute.

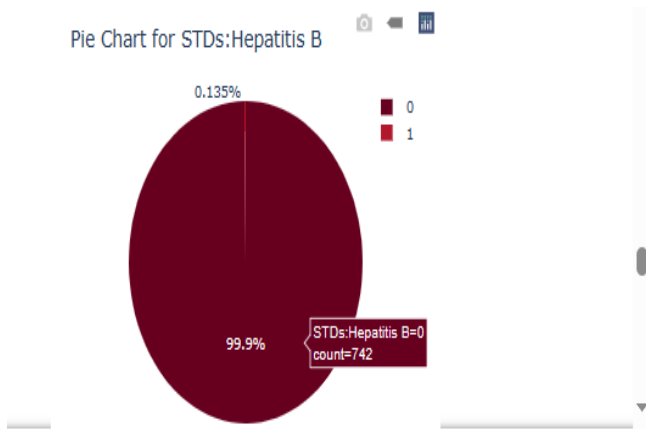


Figure 52: Pie chart for Hepatitis

VII. MODEL EVALUATION

A. Logistic Regression

Logistic regression serves as a supervised model suitable for both binary and multi-classification tasks. In the context of our dataset, we employed this model specifically for multi-classification purposes. Renowned for its accuracy and computational efficiency, logistic regression aids in predicting outcomes, particularly beneficial for telecom companies seeking to discern the pivotal factors influencing telecom churn. Leveraging this tool, they can systematically monitor churn risks among customers and promptly implement strategies to mitigate them.

Logistic Regression

```
In [250]: from sklearn.linear_model import LogisticRegression
classifier = LogisticRegression(random_state = 0)
classifier.fit(x_train, y_train)
y_pred1 = classifier.predict(x_test)
```

```
In [251]: accuracy_score(y_test, y_pred1)
```

```
Out[251]: 0.9932885906040269
```

Figure 53: Logistic Regression

B. Support Vector Machine:

The main aim of the support vector machine is to find the better hyperplane that separates the data points into various classes. This method provides effectiveness and robustness to predict the better accurate result.

Support Vector MACHINE

```
In [255]: from sklearn.svm import SVC
svr = SVC()
svr.fit(x_train, y_train)
y_pred2 = svr.predict(x_test)
```

```
In [256]: accuracy_score(y_test, y_pred2)
```

```
Out[256]: 0.9932885906040269
```

```
In [ ]:
```

Figure 54: Support Vector Machine

C. Random Forest:

Random forest classification, a technique within supervised learning, stands out as an efficient model that delivers highly accurate results. Particularly effective with linear datasets, the random forest approach offers robust classification capabilities.

Random Forest Classification

```
In [252]: from sklearn.ensemble import RandomForestClassifier  
clf = RandomForestClassifier(max_depth=2, random_state=0)  
clf.fit(x_train,y_train)
```

```
Out[252]: RandomForestClassifier(max_depth=2, random_state=0)
```

```
In [253]: y_pred = clf.predict(x_test)
```

```
In [254]: accuracy_score(y_test, y_pred)
```

```
Out[254]: 0.9865771812080537
```

Figure 55: Random Forest Classification

VIII.CONCLUSION

Cervical cancer disease is increasing day by day in most women due to hormonal imbalances. In this research analysis, machine learning and artificial intelligence play a very important role in predicting cervical cancer so that they can decrease the death rate. Here machine learning techniques provide a performance score of approx. 99% which means if we use the techniques by considering the mentioned symptoms then we can provide proper diagnoses to the women.

IX. FUTURE SCOPE

In the future, we can develop a pipeline system to implement deep learning algorithms to identify cervical cancer more quickly and treat the disease in the early stage because in developing countries, treatments are still in progress so we can use deep learning methods. this research work can also be implemented in other gynecological cancer prediction.

REFERENCES

1. Camgoz Akdag, H., Menekse, A. and Sahin, F., 2024. Cervical cancer risk assessment using machine learning integrated fuzzy MCDM methodology. *Journal of Intelligent & Fuzzy Systems*, (Preprint), pp.1-20.
2. Kaushik, K., Bhardwaj, A., Bharany, S., Alsharabi, N., Rehman, A.U., Eldin, E.T. and Ghamry, N.A., 2022. A machine learning-based framework for the prediction of cervical cancer risk in women. *Sustainability*, 14(19), p.11947.
3. Lilhore, U.K., Poongodi, M., Kaur, A., Simaiya, S., Algarni, A.D., Elmannai, H., Vijayakumar, V., Tunze, G.B. and Hamdi, M., 2022. Hybrid model for detection of cervical cancer using causal analysis and machine learning techniques. *Computational and Mathematical Methods in Medicine*, 2022.
4. Ratul, I.J., Al-Monsur, A., Tabassum, B., Ar-Rafi, A.M., Nishat, M.M. and Faisal, F., 2022, May. Early risk prediction of cervical cancer: A machine learning approach. In *2022 19th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)* (pp. 1-4). IEEE.
5. Alsmariy, R., Healy, G. and Abdelhafez, H., 2020. Predicting cervical cancer using machine learning methods. *International Journal of Advanced Computer Science and Applications*, 11(7).
6. Kumawat, G., Vishwakarma, S.K., Chakrabarti, P., Chittora, P., Chakrabarti, T. and Lin, J.C.W., 2023. Prognosis of Cervical

Cancer Disease by Applying Machine Learning Techniques.

Journal of Circuits, Systems and Computers, 32(01), p.2350019.

7. Akter, L., Islam, M.M., Al-Rakhami, M.S. and Haque, M.R., 2021. Prediction of cervical cancer from behavior risk using machine learning techniques. *SN Computer Science*, 2(3), p.177.
8. Tanimu, J.J., Hamada, M., Hassan, M., Kakudi, H. and Abiodun, J.O., 2022. A machine learning method for classification of cervical cancer. *Electronics*, 11(3), p.463.
9. Ramzan, Z., Hassan, M.A., Asif, H.M. and Farooq, A., 2021. A machine learning-based self-risk assessment technique for cervical cancer. *Current Bioinformatics*, 16(2), pp.315-332.
10. Ali, M.M., Ahmed, K., Bui, F.M., Paul, B.K., Ibrahim, S.M., Quinn, J.M. and Moni, M.A., 2021. Machine learning-based statistical analysis for early stage detection of cervical cancer. *Computers in biology and medicine*, 139, p.104985.
11. Asadi, F., Salehnasab, C. and Ajori, L., 2020. Supervised algorithms of machine learning for the prediction of cervical cancer. *Journal of biomedical physics & engineering*, 10(4), p.513.