CERVICAL CANCER EPIDEMIOLOGY

Insights into Hormonal Contraceptives, IUD, Smoking and Age

GROUP-8

- TEAM MEMBERS:

Kalepu Srinath

Veronica Angelina Itta

Adarsh Viswanath

Mahitha Vontimitta

Arun Kumar Soora

Ritheesh Miridoodi

- GUIDED BY:

JOHN BURNS

INTRODUCTION

- Every year, cervical cancer is diagnosed in over half a million women, and the disease leads to more than 300,000 fatalities globally (Cohen et al., 2019).
- As per the latest data, cervical cancer holds the fourteenth position among all cancer types and is the fourth most prevalent cancer among women worldwide (Fowler et al., 2022).
- The risk of cervical cancer is significantly influenced by factors such as early initiation of sexual activity and having multiple sexual partners. Additionally, variations in the incidence of cervical cancer among different countries can be attributed to the implementation of screening programs. While the overall trend suggests a reduction in both incidence and mortality, there are indications of an elevated risk of cervical cancer, likely associated with shifts in sexual behavior. Smoking and human papillomavirus (HPV) infection are currently recognized as significant factors in the multifaceted and progressive process of cervical carcinogenesis (Zhang et al., 2020).
- Interventions for cervical cancer primarily center on primary and secondary prevention. The most effective approach to reduce the impact of cervical cancer and lower mortality rates is through primary prevention and regular screening (Fowler et al., 2022. We utilized this dataset to develop a model for the identification of individuals with a high risk of cervical cancer.

AIM:

The dataset titled "Cervical Cancer Epidemiology: Insights into Hormonal Contraceptives, IUD, Age, and Smoking" aims to provide important insights and support research on the connection between cervical cancer and four significant factors. By gathering and examining data related to these elements, the dataset seeks to offer a comprehensive understanding of cervical cancer epidemiology concerning hormonal contraceptives, IUD usage, and age. This valuable information can empower researchers and healthcare professionals to make informed decisions, develop effective prevention strategies, and enhance the quality of patient care.

PURPOSE:

The purpose of this project is to gain insights into cervical cancer patterns linked to age, number of sexual partners, number of pregnancies, smoking, hormonal contraceptives, IUD usage and STDs.

To correlate the above mentioned factors with cervical cancer using machine learning models to gain perspective on their role in the cervical cancer incidence.

RESEARCH QUESTION:

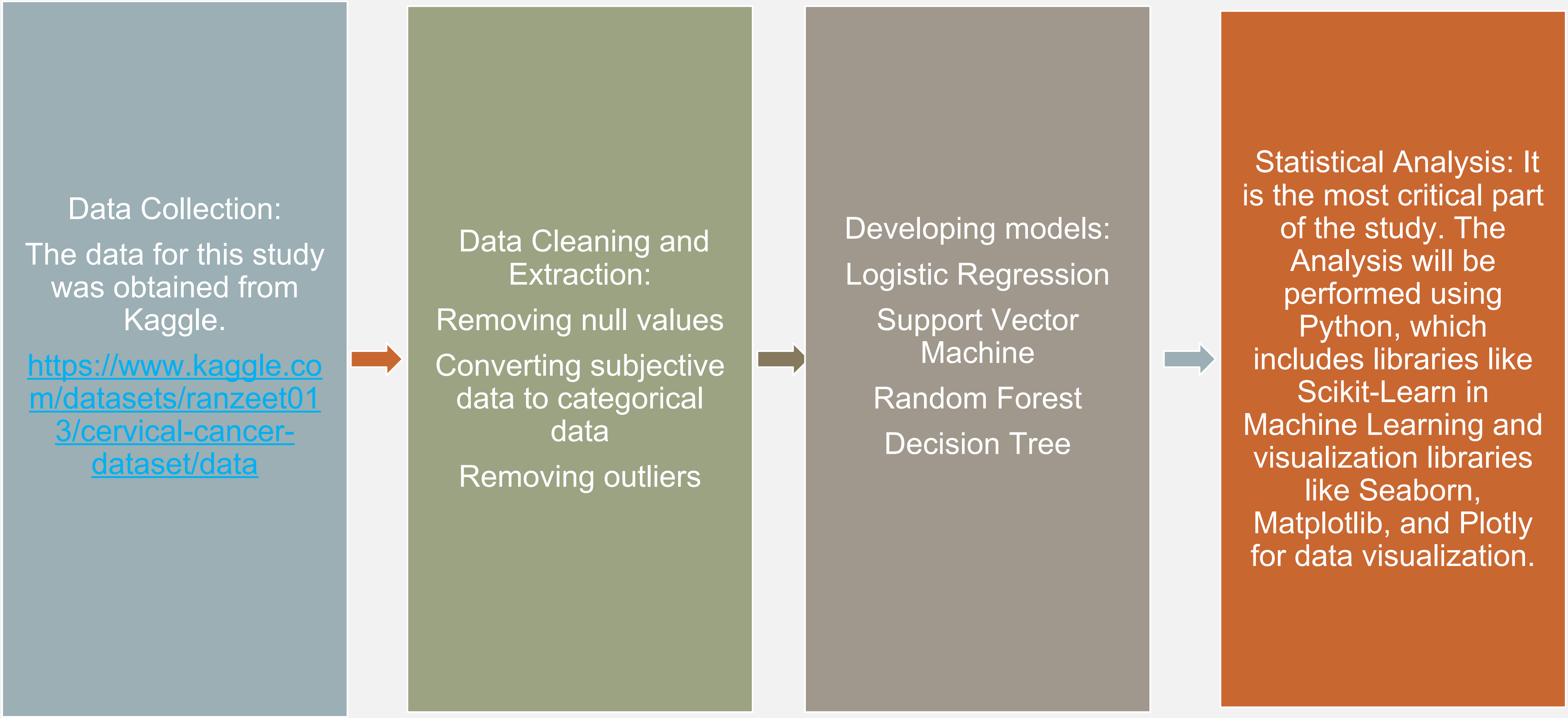
How do hormonal contraceptives, IUD usage, age, and smoking contribute to the epidemiology of cervical cancer? Specifically, what are the interconnections, correlations, or trends among these factors in relation to the occurrence and prevalence of cervical cancer?

HYPOTHESIS

Null Hypothesis: There is no significant interaction between age and the use of hormonal contraceptives in terms of their effect on cervical cancer risk, meaning that the impact of hormonal contraceptives on cervical cancer risk does not differ significantly between various age groups.

Alternative Hypothesis: An interaction effect exists between age and the use of hormonal contraceptives, leading to varying impacts of hormonal contraceptives on cervical cancer risk among different age groups.

METHODOLOGY



Data Collection:

The data for this study was obtained from Kaggle.

https://www.kaggle.com/datasets/ranzeet013/cervical-cancer-dataset/data

Data Cleaning and Extraction:

Removing null values

Converting subjective data to categorical data

Removing outliers

Developing models:

Logistic Regression

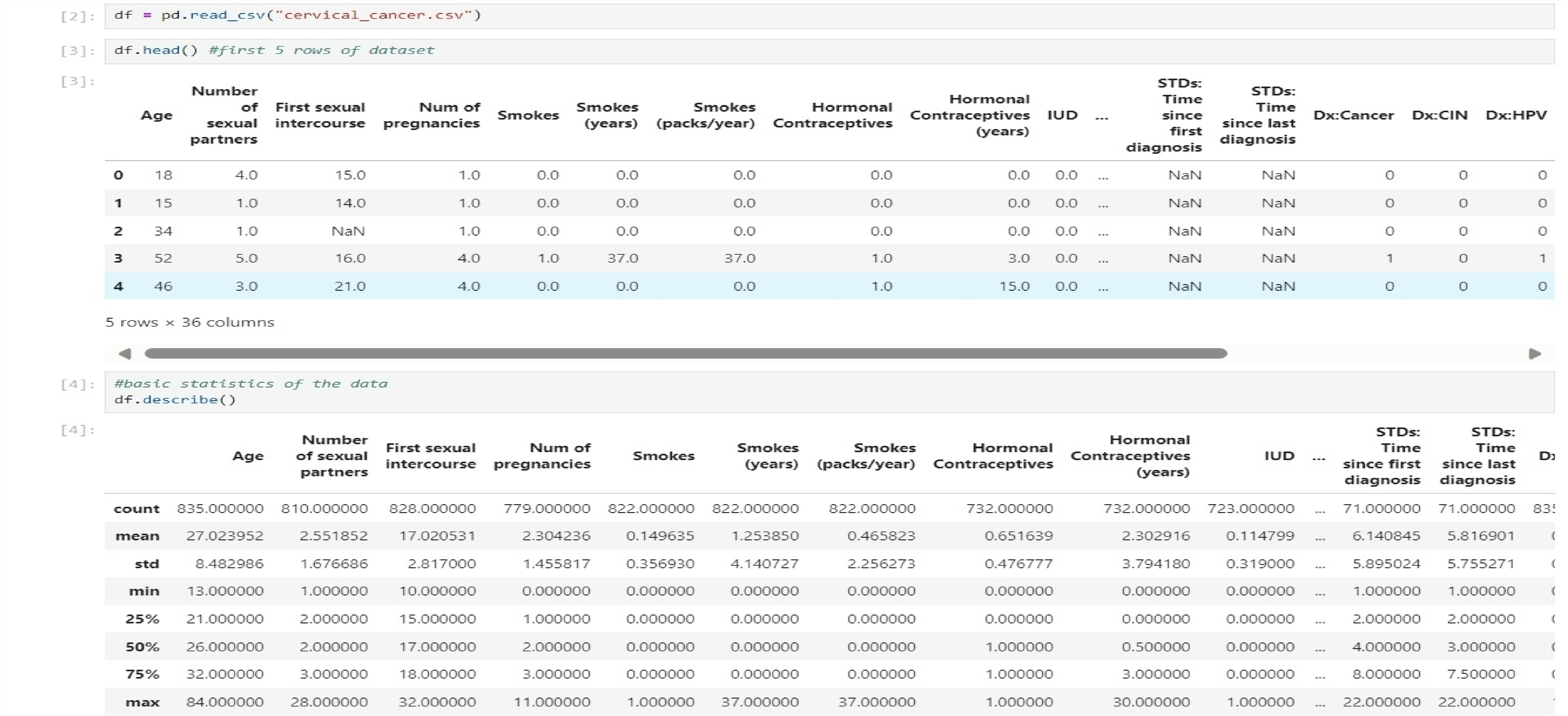
Support Vector Machine

Random Forest

Decision Tree

Statistical Analysis: It is the most critical part of the study. The Analysis will be performed using Python, which includes libraries like Scikit-Learn in Machine Learning and visualization libraries like Seaborn, Matplotlib, and Plotly for data visualization.

READING THE DATASET



```
[2]: df = pd.read_csv("cervical_cancer.csv")
```

```
[3]: df.head() #first 5 rows of dataset
```

[3]:

	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	Dx:HPV
0	18	4.0	15.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	...	NaN	NaN	0	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	0
2	34	1.0	NaN	1.0	0.0	0.0	0.0	0.0	0.0	0.0	...	NaN	NaN	0	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	1
4	46	3.0	21.0	4.0	0.0	0.0	0.0	1.0	15.0	0.0	...	NaN	NaN	0	0	0

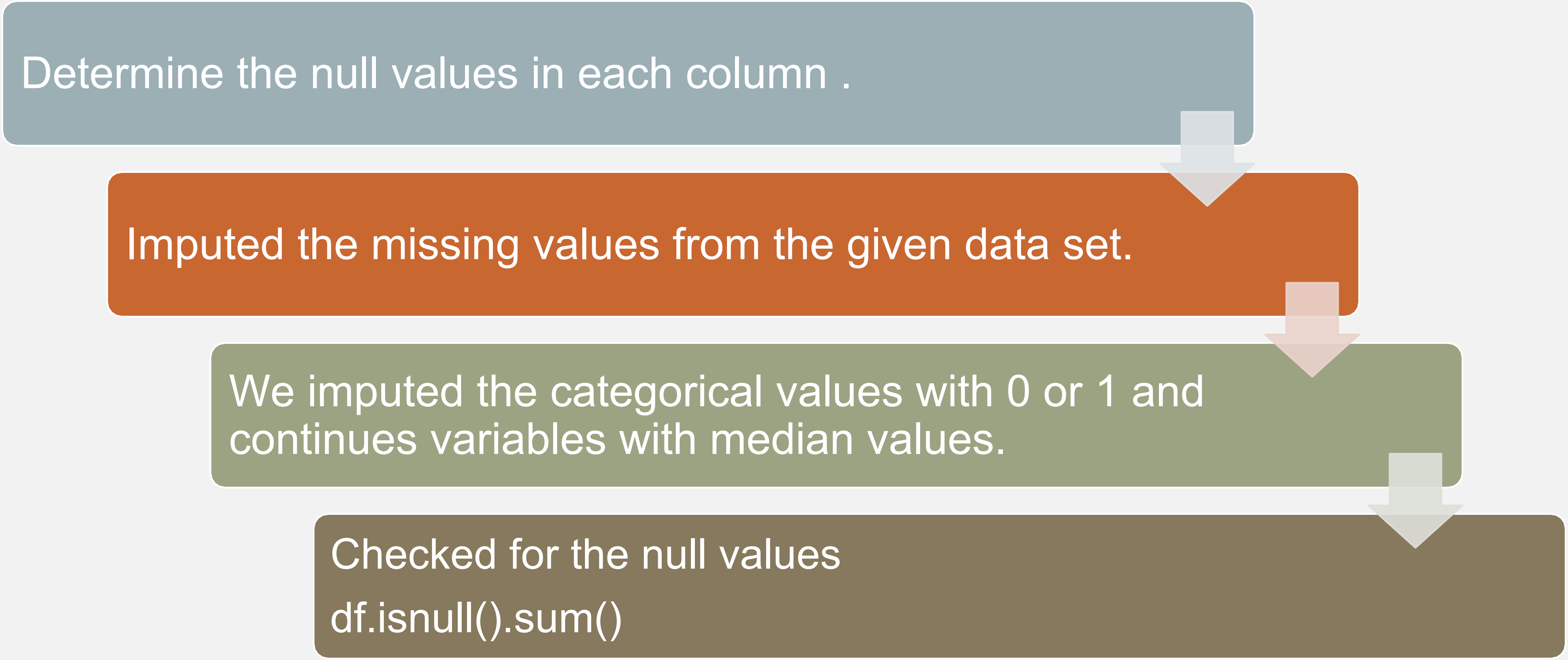
5 rows × 36 columns

```
[4]: #basic statistics of the data
     df.describe()
```

[4]:

	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	D
count	835.000000	810.000000	828.000000	779.000000	822.000000	822.000000	822.000000	732.000000	732.000000	723.000000	...	71.000000	71.000000	83
mean	27.023952	2.551852	17.020531	2.304236	0.149635	1.253850	0.465823	0.651639	2.302916	0.114799	...	6.140845	5.816901	
std	8.482986	1.676686	2.817000	1.455817	0.356930	4.140727	2.256273	0.476777	3.794180	0.319000	...	5.895024	5.755271	
min	13.000000	1.000000	10.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	1.000000	1.000000	
25%	21.000000	2.000000	15.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	2.000000	2.000000	
50%	26.000000	2.000000	17.000000	2.000000	0.000000	0.000000	0.000000	1.000000	0.500000	0.000000	...	4.000000	3.000000	
75%	32.000000	3.000000	18.000000	3.000000	0.000000	0.000000	0.000000	1.000000	3.000000	0.000000	...	8.000000	7.500000	
max	84.000000	28.000000	32.000000	11.000000	1.000000	37.000000	37.000000	1.000000	30.000000	1.000000	...	22.000000	22.000000	

DATA CLEANING



Determine the null values in each column .

Imputed the missing values from the given data set.

We imputed the categorical values with 0 or 1 and continues variables with median values.

Checked for the null values

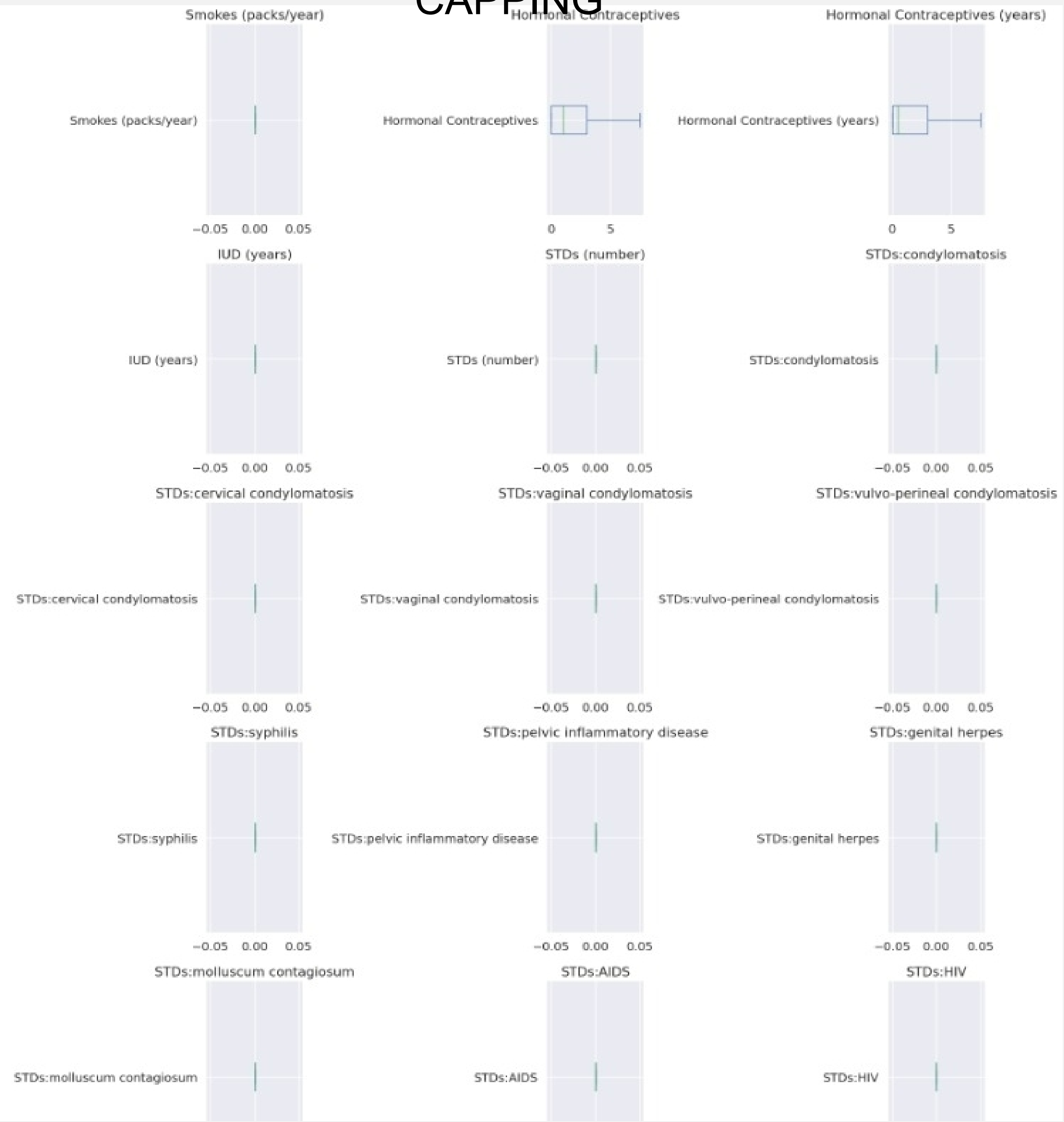
df.isnull().sum()

OUTLIERS

Detected the outliers

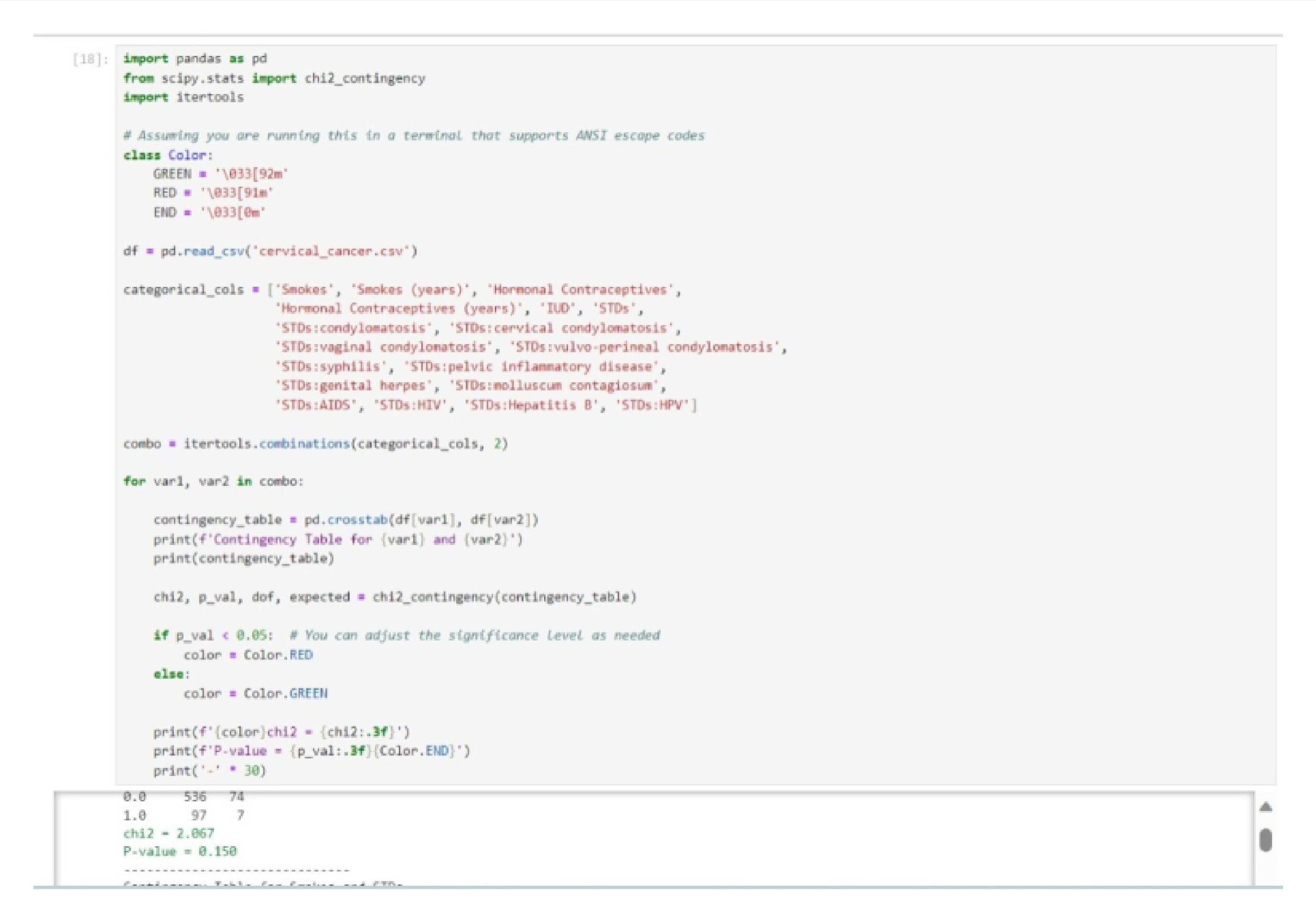
Successfully capped the outliers

AFTER CAPPING



STATISTICAL TESTING

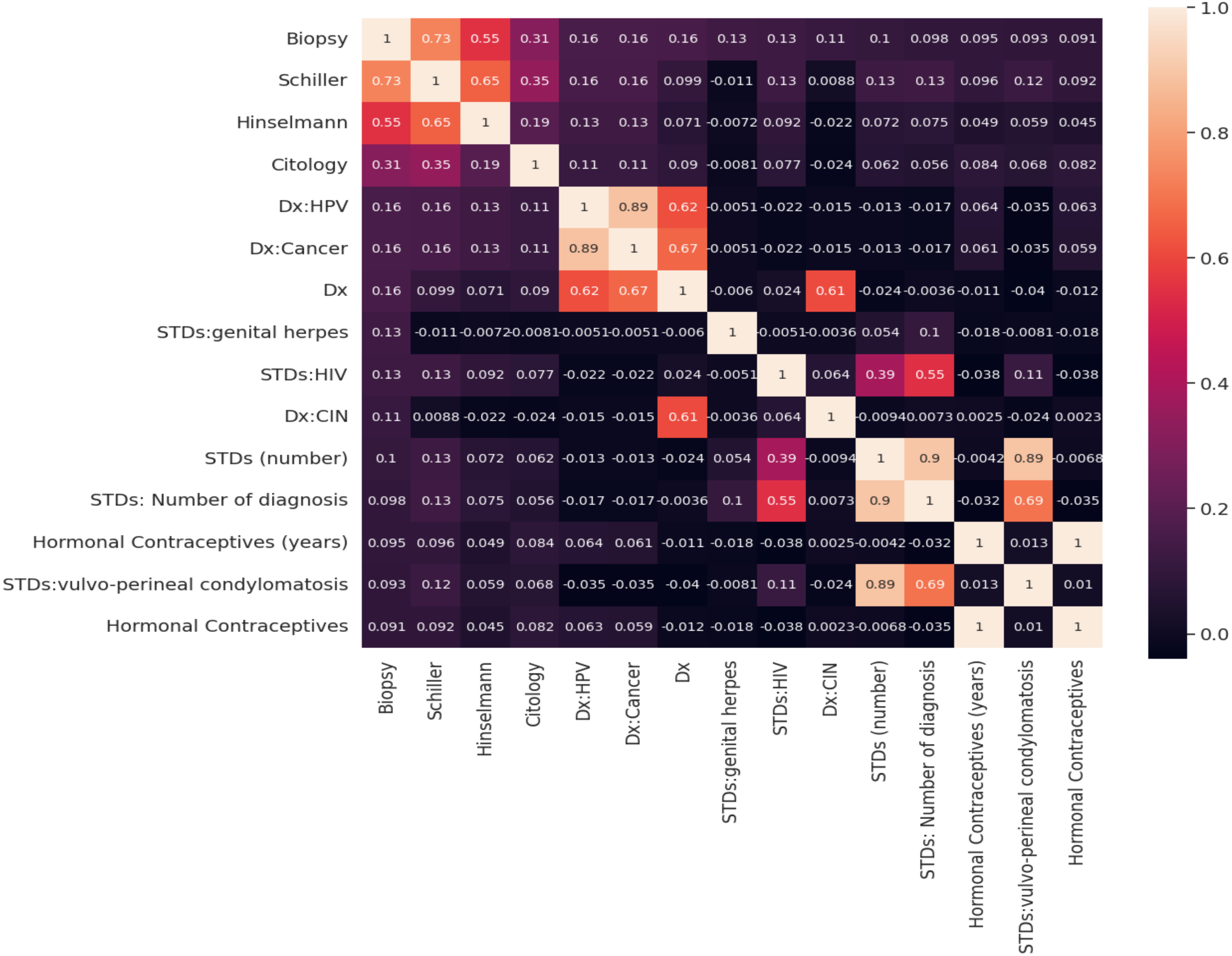
- PERFORMED A CHI-SQUARE TEST



```
[18]: import pandas as pd
      from scipy.stats import chi2_contingency
      import itertools

      # Assuming you are running this in a terminal that supports ANSI escape codes
      class Color:
          GREEN = '\033[92m'
          RED = '\033[91m'
          END = '\033[0m'

      df = pd.read_csv('cervical_cancer.csv')

      categorical_cols = ['Smokes', 'Smokes (years)', 'Hormonal Contraceptives',
                          'Hormonal Contraceptives (years)', 'IUD', 'STDs',
                          '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']

      combo = itertools.combinations(categorical_cols, 2)

      for var1, var2 in combo:

          contingency_table = pd.crosstab(df[var1], df[var2])
          print(f'Contingency Table for {var1} and {var2}')
          print(contingency_table)

          chi2, p_val, dof, expected = chi2_contingency(contingency_table)

          if p_val < 0.05:  # You can adjust the significance level as needed
              color = Color.RED
          else:
              color = Color.GREEN

          print(f'{color}chi2 = {chi2:.3f}')
          print(f'P-value = {p_val:.3f}{Color.END}')
          print('-' * 30)
```

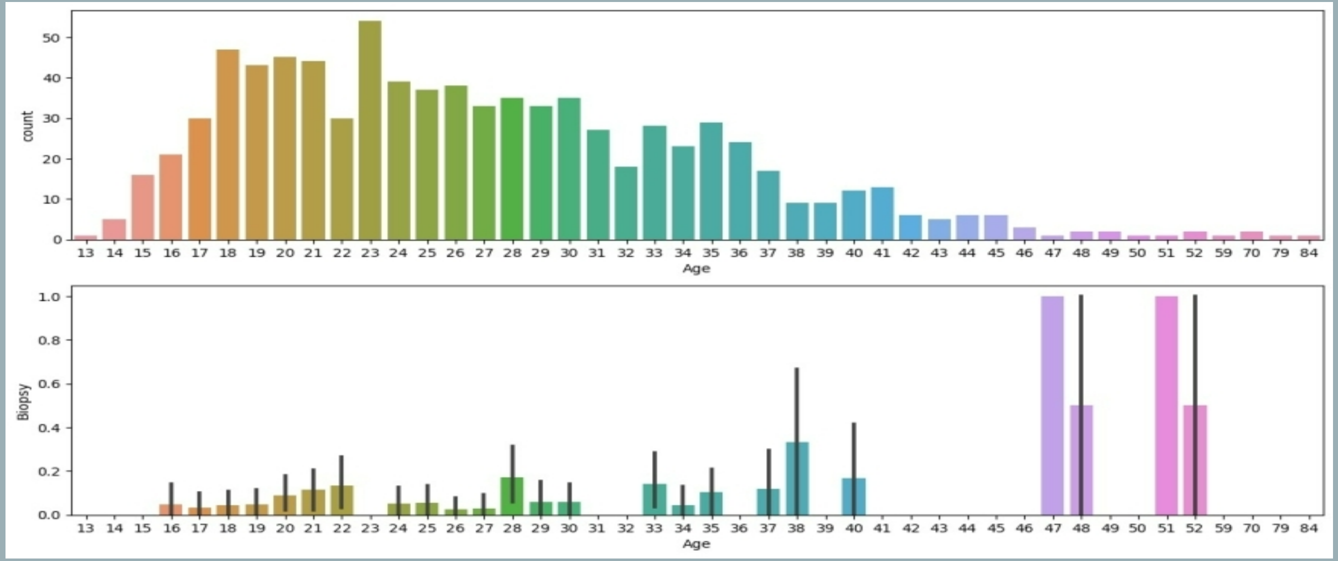
```
0.0     536   74
1.0      97    7
chi2 = 2.067
P-value = 0.150
------------------------------
Contingency Table for Smokes and STDs
```

HEATMAP OF TOP CORRELATION AFTER IMPUTATION

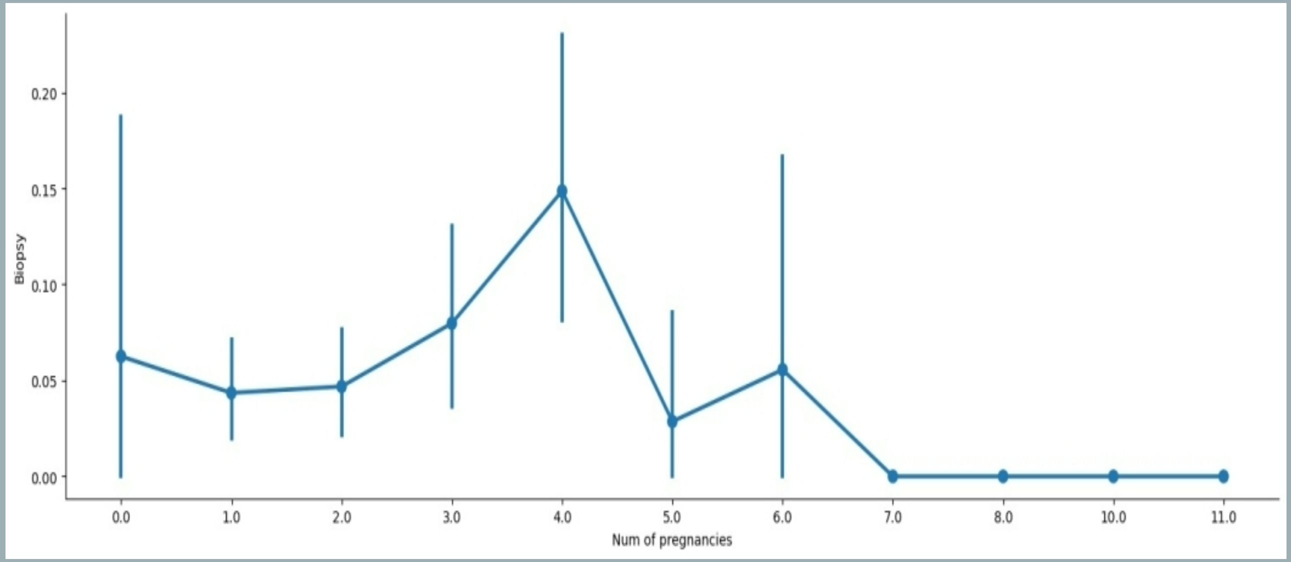


	Biopsy	Schiller	Hinselmann	Citology	Dx:HPV	Dx:Cancer	Dx	STDs:genital herpes	STDs:HIV	Dx:CIN	STDs (number)	STDs: Number of diagnosis	Hormonal Contraceptives (years)	STDs:vulvo-perineal condylomatosis	Hormonal Contraceptives
Biopsy	1	0.73	0.55	0.31	0.16	0.16	0.16	0.13	0.13	0.11	0.1	0.098	0.095	0.093	0.091
Schiller	0.73	1	0.65	0.35	0.16	0.16	0.099	-0.011	0.13	0.0088	0.13	0.13	0.096	0.12	0.092
Hinselmann	0.55	0.65	1	0.19	0.13	0.13	0.071	-0.0072	0.092	-0.022	0.072	0.075	0.049	0.059	0.045
Citology	0.31	0.35	0.19	1	0.11	0.11	0.09	-0.0081	0.077	-0.024	0.062	0.056	0.084	0.068	0.082
Dx:HPV	0.16	0.16	0.13	0.11	1	0.89	0.62	-0.0051	-0.022	-0.015	-0.013	-0.017	0.064	-0.035	0.063
Dx:Cancer	0.16	0.16	0.13	0.11	0.89	1	0.67	-0.0051	-0.022	-0.015	-0.013	-0.017	0.061	-0.035	0.059
Dx	0.16	0.099	0.071	0.09	0.62	0.67	1	-0.006	0.024	0.61	-0.024	-0.0036	-0.011	-0.04	-0.012
STDs:genital herpes	0.13	-0.011	-0.0072	-0.0081	-0.0051	-0.0051	-0.006	1	-0.0051	-0.0036	0.054	0.1	-0.018	-0.0081	-0.018
STDs:HIV	0.13	0.13	0.092	0.077	-0.022	-0.022	0.024	-0.0051	1	0.064	0.39	0.55	-0.038	0.11	-0.038
Dx:CIN	0.11	0.0088	-0.022	-0.024	-0.015	-0.015	0.61	-0.0036	0.064	1	-0.0094	0.0073	0.0025	-0.024	0.0023
STDs (number)	0.1	0.13	0.072	0.062	-0.013	-0.013	-0.024	0.054	0.39	-0.0094	1	0.9	-0.0042	0.89	-0.0068
STDs: Number of diagnosis	0.098	0.13	0.075	0.056	-0.017	-0.017	-0.0036	0.1	0.55	0.0073	0.9	1	-0.032	0.69	-0.035
Hormonal Contraceptives (years)	0.095	0.096	0.049	0.084	0.064	0.061	-0.011	-0.018	-0.038	0.0025	-0.0042	-0.032	1	0.013	1
STDs:vulvo-perineal condylomatosis	0.093	0.12	0.059	0.068	-0.035	-0.035	-0.04	-0.0081	0.11	-0.024	0.89	0.69	0.013	1	0.01
Hormonal Contraceptives	0.091	0.092	0.045	0.082	0.063	0.059	-0.012	-0.018	-0.038	0.0023	-0.0068	-0.035	1	0.01	1

DATA VISUALIZATION

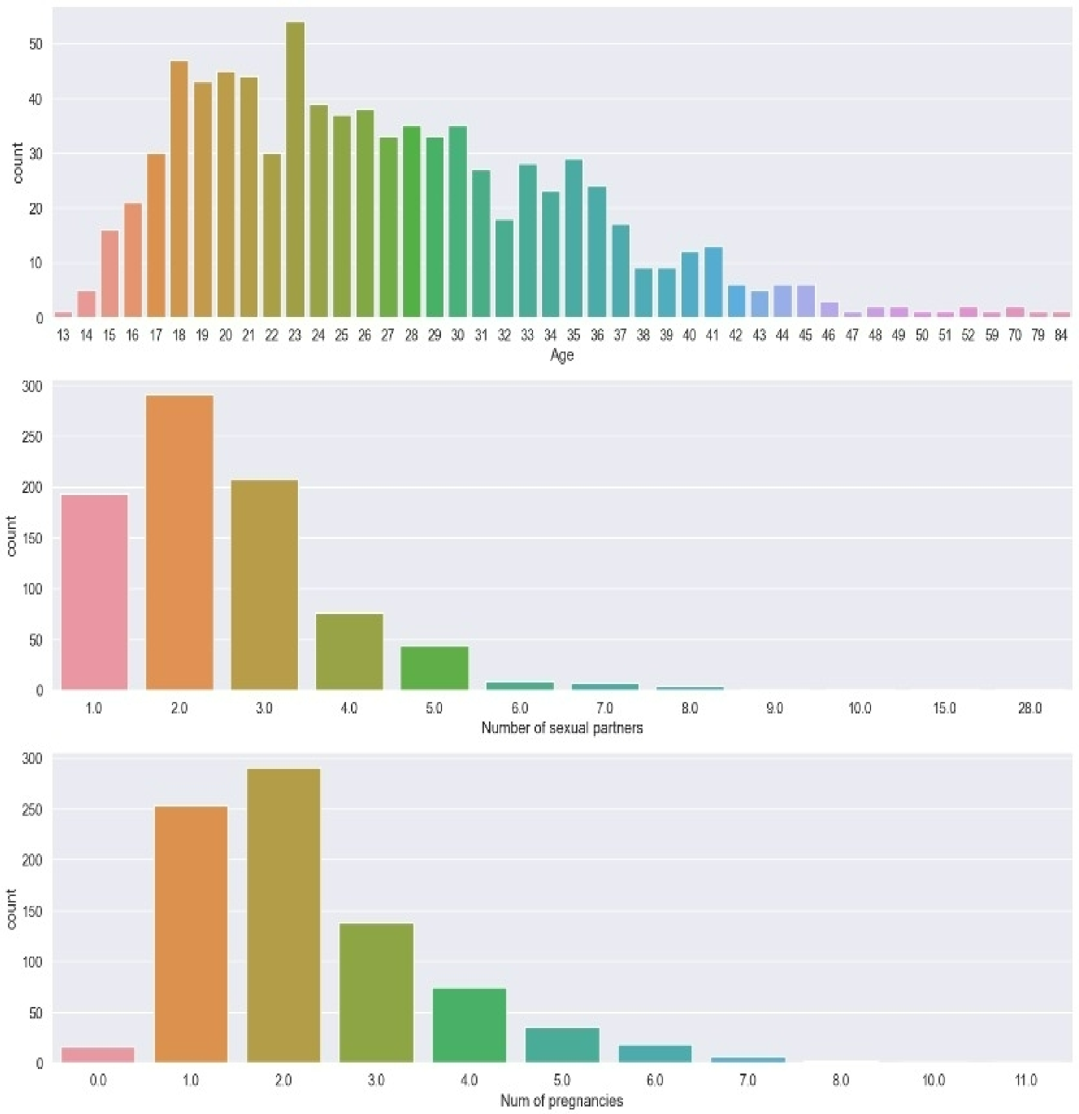


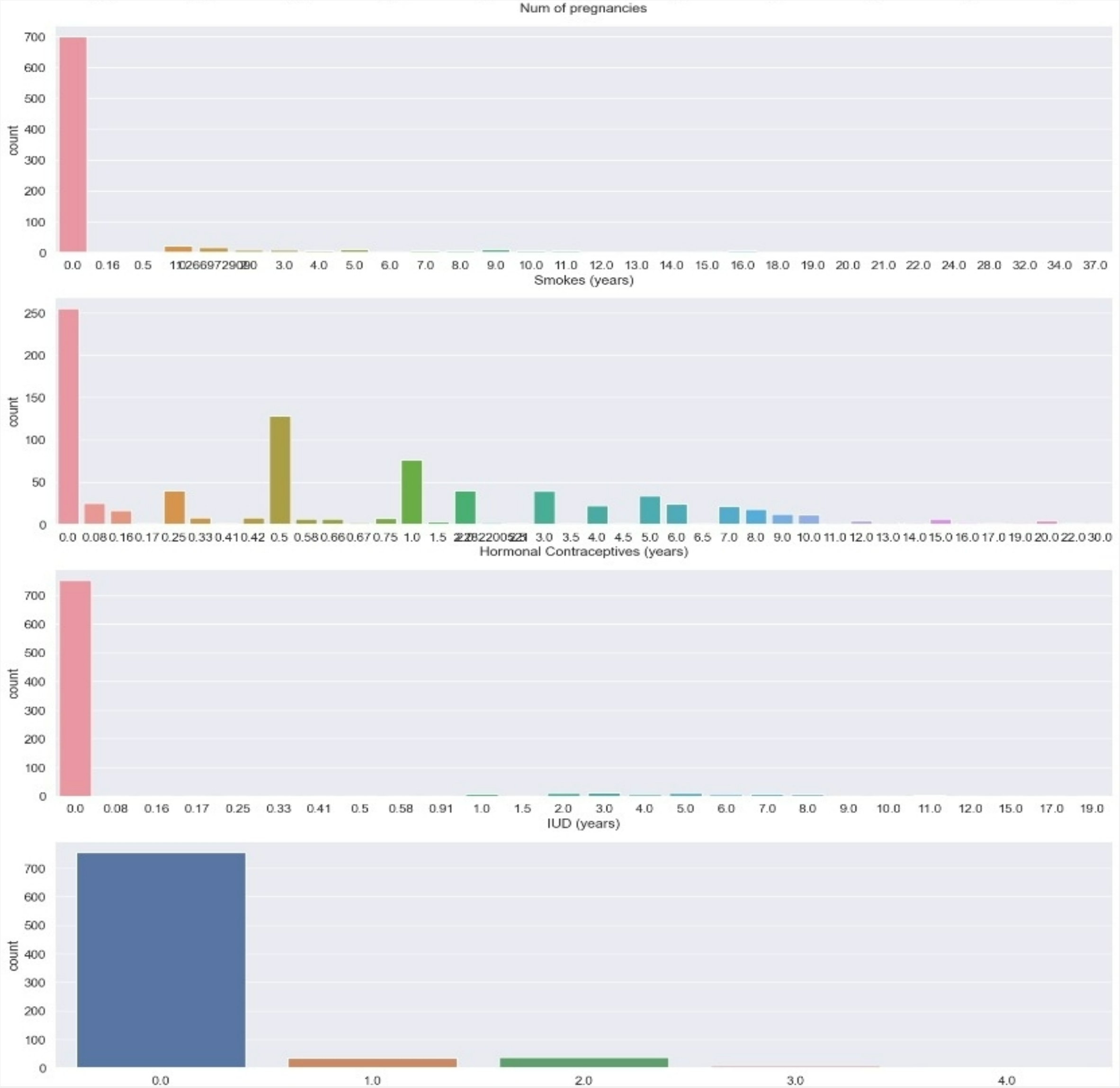
BIOPSY VS NO.OF SEXUAL PARTNERS



BIOPSY VS NO.OF PREGNANCIES WITH FACTOR PLOT

BAR PLOTS FOR AGE, SEXUAL PARTNERS, SMOKE, IUD, HORMONAL CONTRACEPTIVES AND STD'S





MACHINE LEARNING MODELS

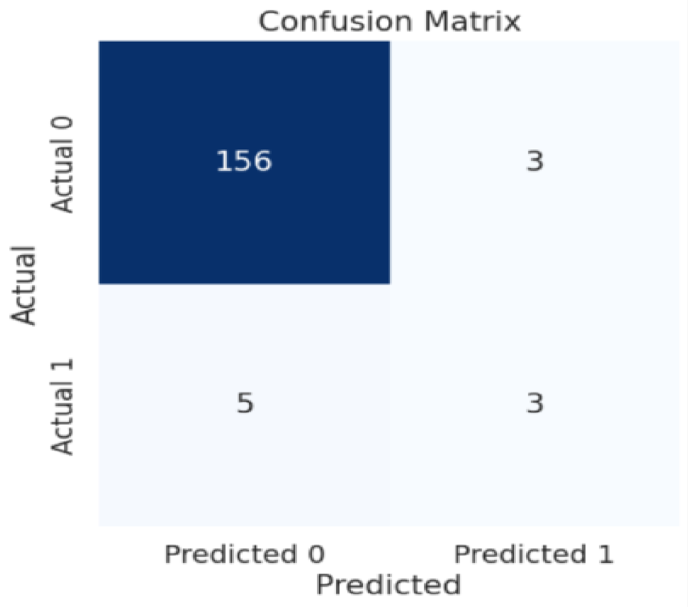
- LOGISTIC REGRESSION
- SUPPORT VECTOR MACHINE
- RANDOM FOREST
- DECISION TREE

LOGISTIC REGRESSION

TESTING ACCURACY:0.96

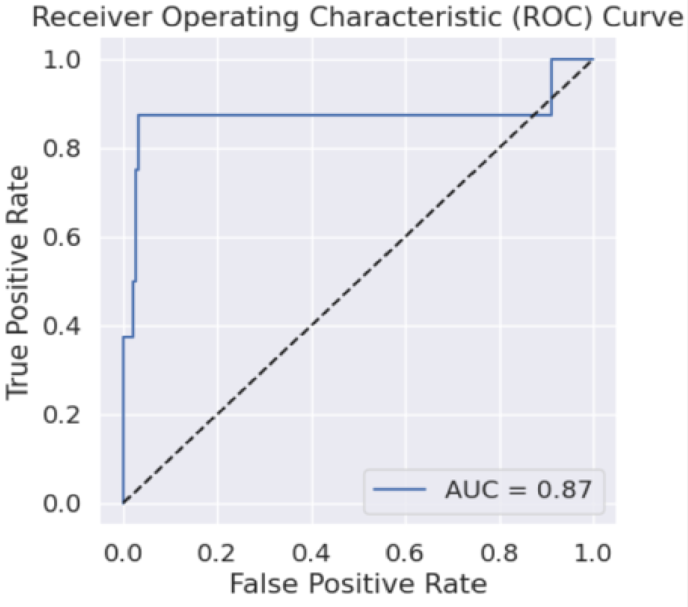
ACCURACY CURVE: 0.87

CONFUSION MATRIX


Confusion Matrix
Actual
Predicted

	Predicted 0	Predicted 1
Actual 0	156	3
Actual 1	5	3

RECEIVER OPERATING CHARACTERISTIC (ROC) CURVE

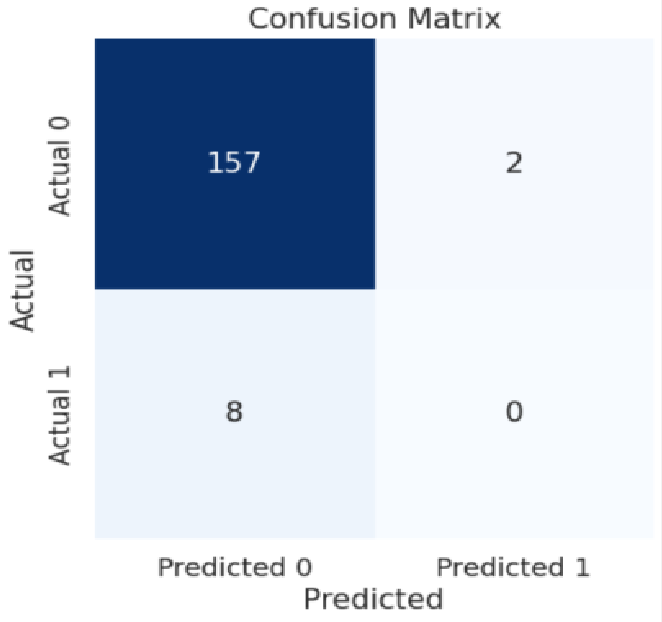


SUPPORT VECTOR MACHINE (SVM)

TESTING ACCURACY:0.97

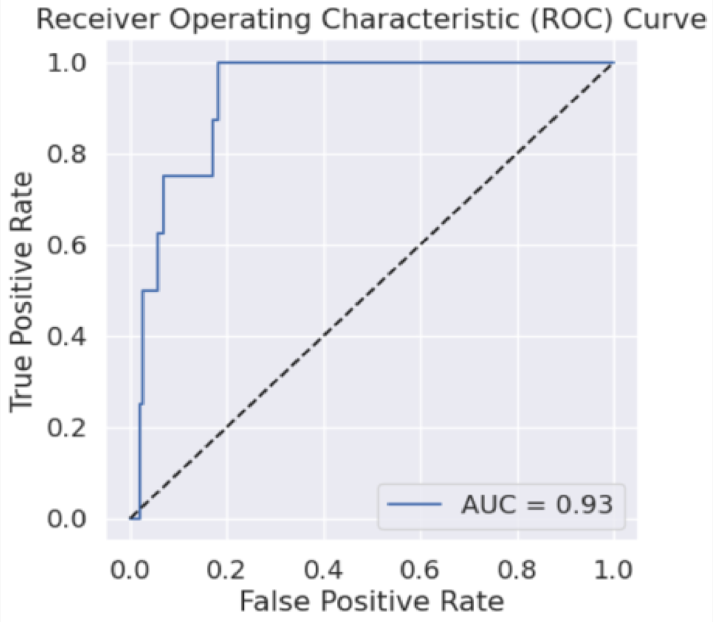
ACCURACY CURVE: 0.87

CONFUSION MATRIX



	Predicted 0	Predicted 1
Actual 0	157	2
Actual 1	8	0

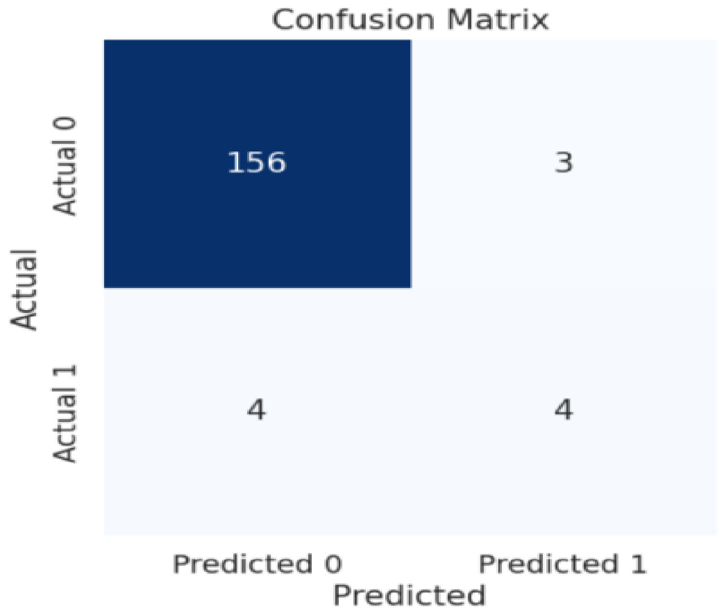
RECEIVER OPERATING CHARACTERISTIC (ROC) CURVE



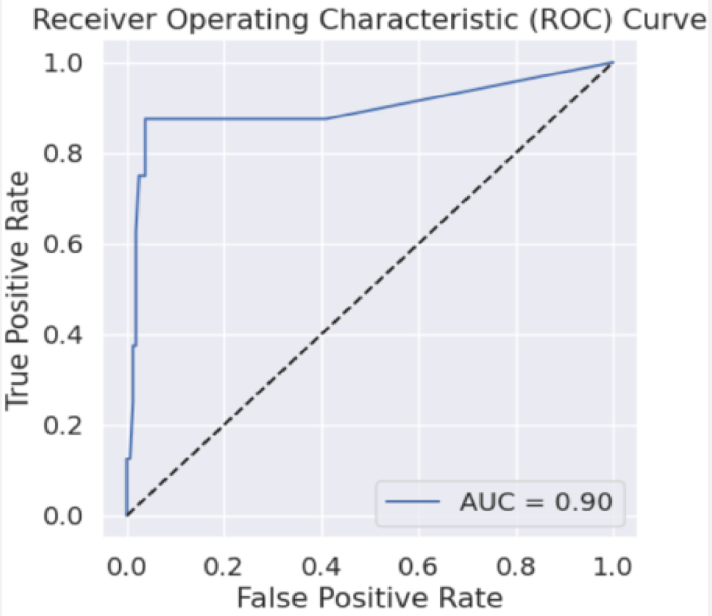
RANDOM FOREST

TESTING ACCURACY:0.94
ACCURACY CURVE: 0.89

CONFUSION MATRIX



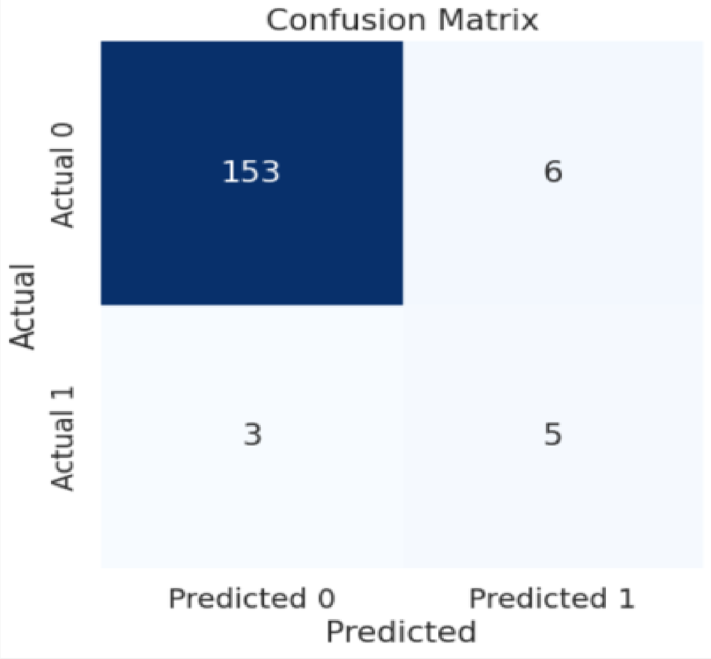
RECEIVER OPERATING CHARACTERISTIC (ROC) CURVE



DECISION TREE

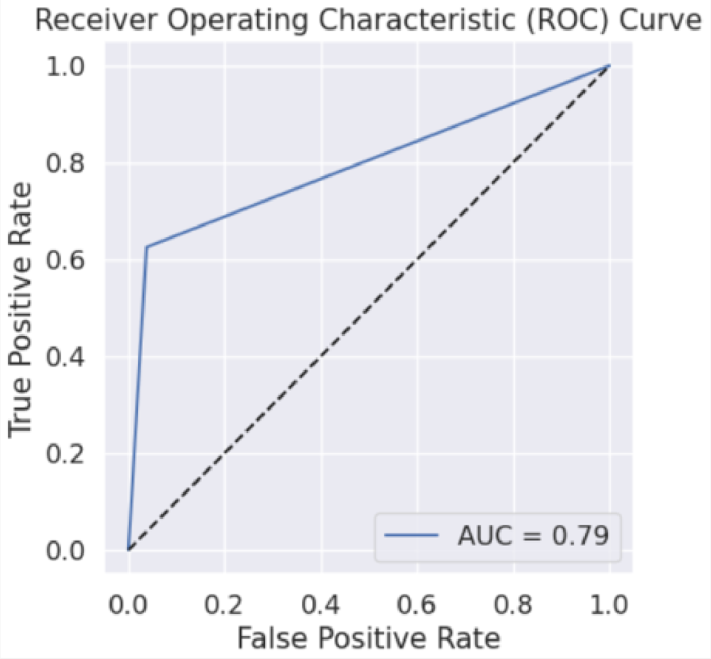
TESTING ACCURACY:0.94
ACCURACY CURVE: 0.73

CONFUSION MATRIX

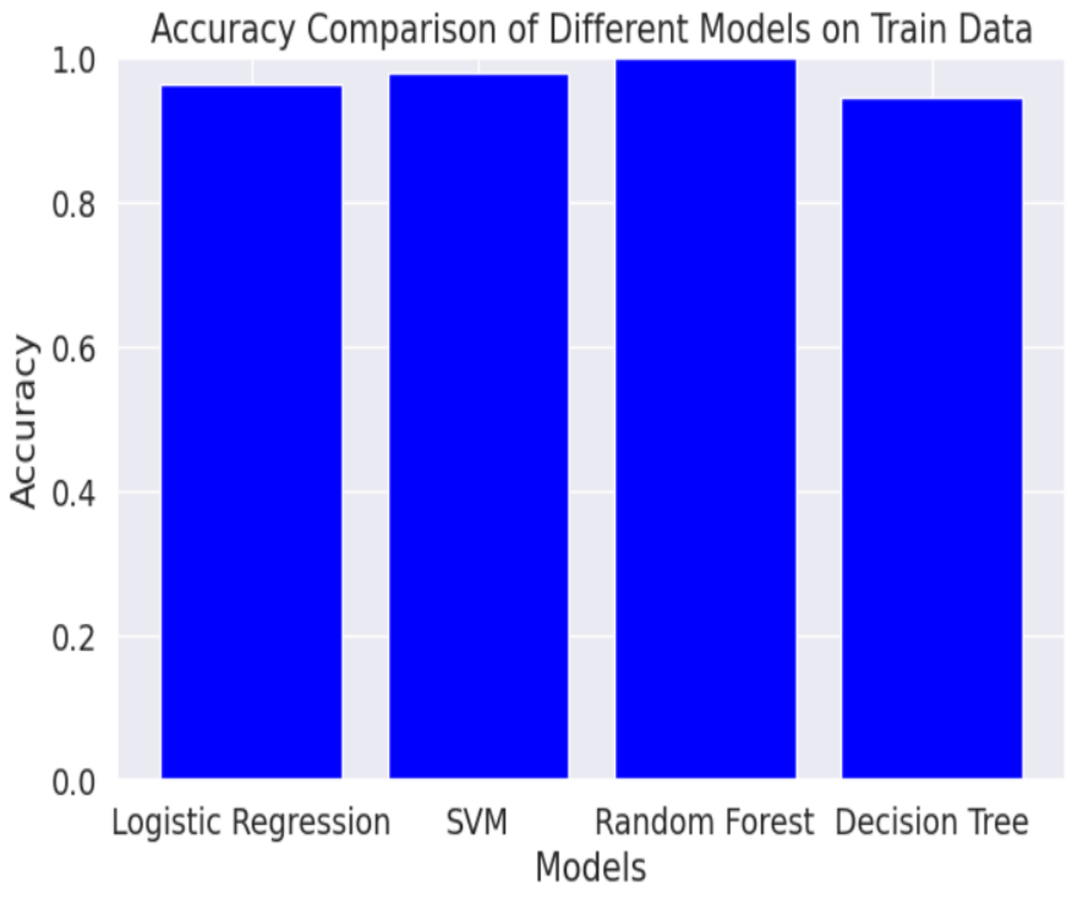


	Predicted 0	Predicted 1
Actual 0	153	6
Actual 1	3	5

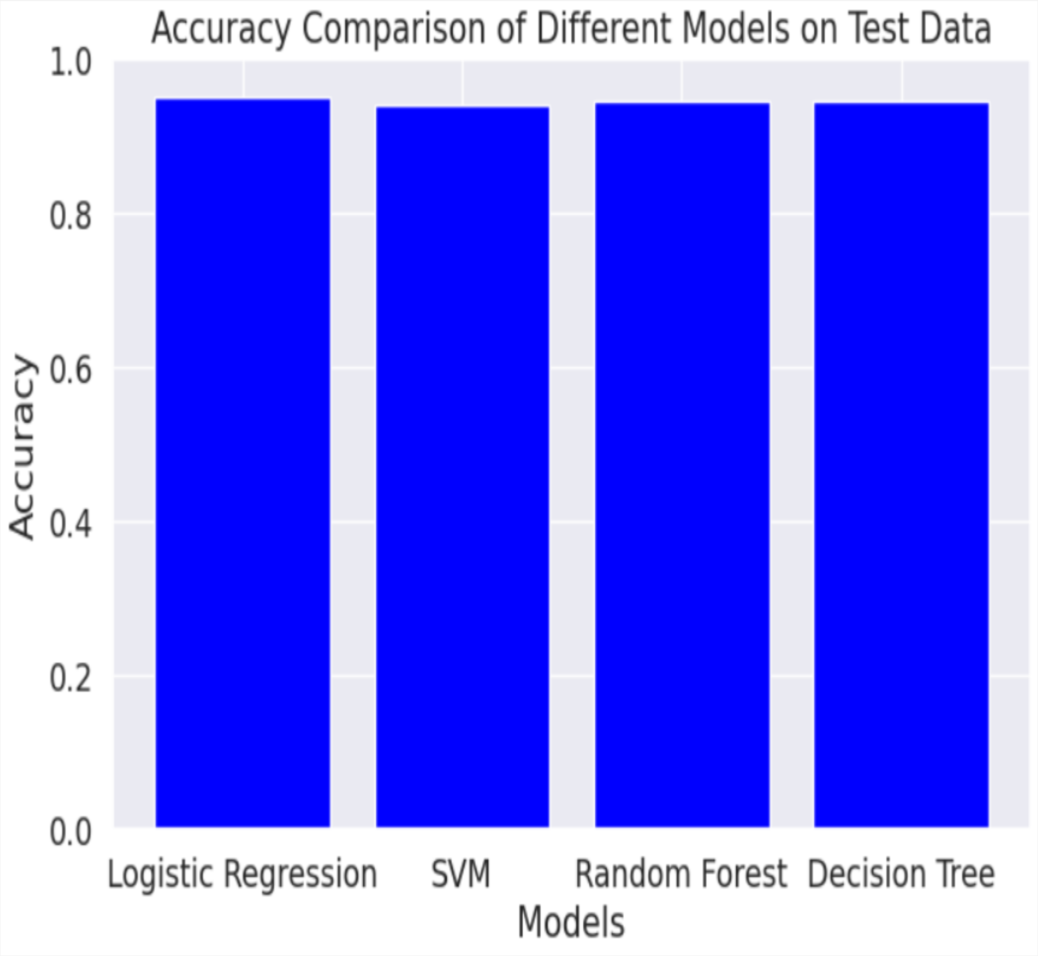
RECEIVER OPERATING CHARACTERISTIC (ROC) CURVE



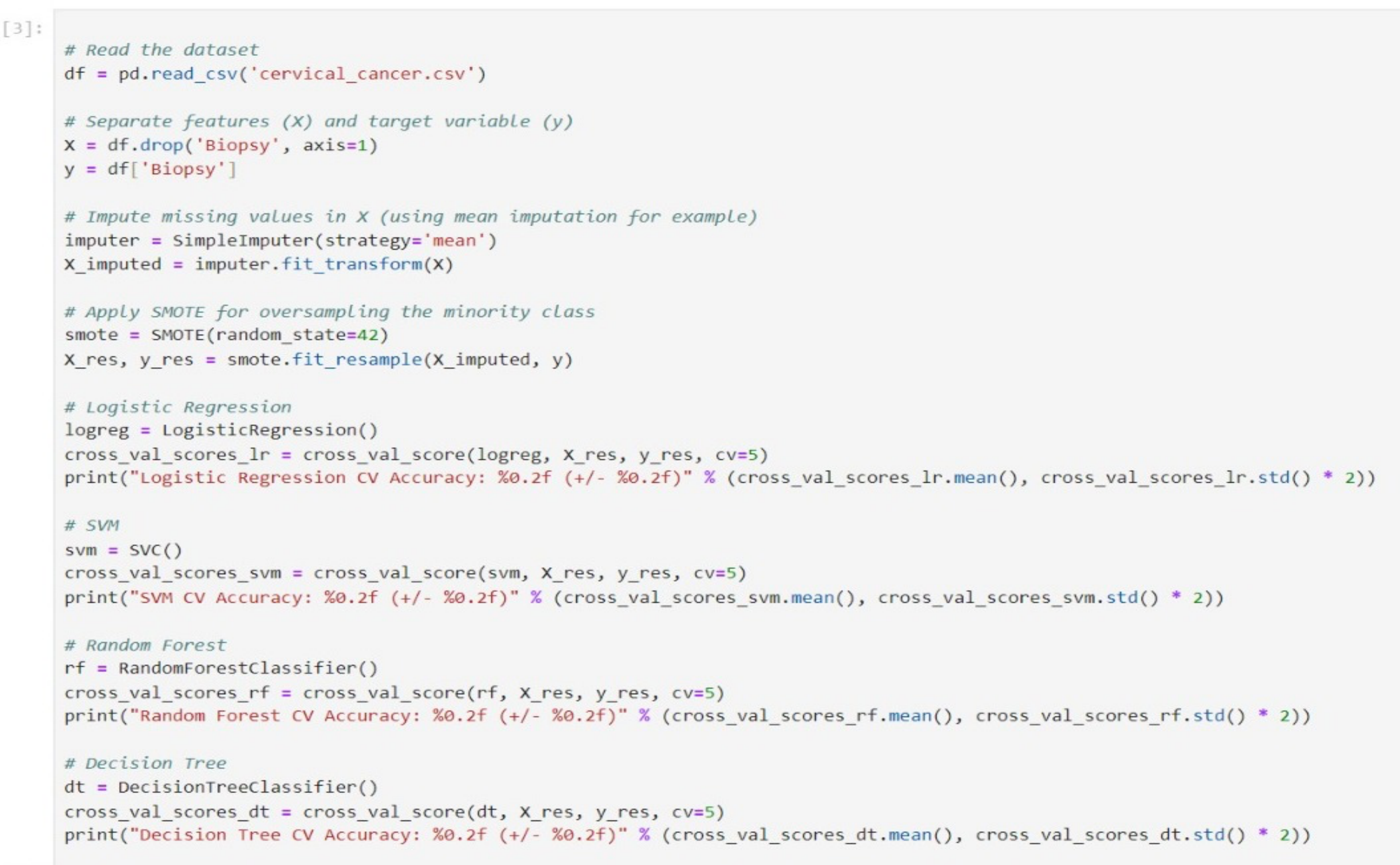
ACCURACY COMPARISON OF DIFFERENT MODELS ON TRAIN DATA



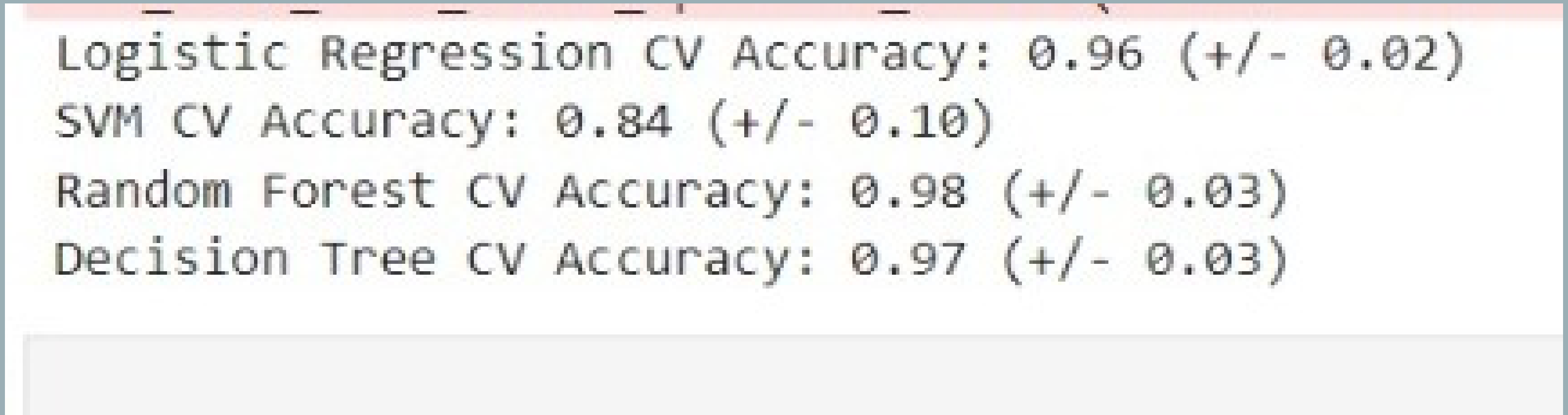
ACCURACY COMPARISON OF DIFFERENT MODELS ON TEST DATA



CROSS VALIDATION



```
# Read the dataset
df = pd.read_csv('cervical_cancer.csv')

# Separate features (X) and target variable (y)
X = df.drop('Biopsy', axis=1)
y = df['Biopsy']

# Impute missing values in X (using mean imputation for example)
imputer = SimpleImputer(strategy='mean')
X_imputed = imputer.fit_transform(X)

# Apply SMOTE for oversampling the minority class
smote = SMOTE(random_state=42)
X_res, y_res = smote.fit_resample(X_imputed, y)

# Logistic Regression
logreg = LogisticRegression()
cross_val_scores_lr = cross_val_score(logreg, X_res, y_res, cv=5)
print("Logistic Regression CV Accuracy: %0.2f (+/- %0.2f)" % (cross_val_scores_lr.mean(), cross_val_scores_lr.std() * 2))

# SVM
svm = SVC()
cross_val_scores_svm = cross_val_score(svm, X_res, y_res, cv=5)
print("SVM CV Accuracy: %0.2f (+/- %0.2f)" % (cross_val_scores_svm.mean(), cross_val_scores_svm.std() * 2))

# Random Forest
rf = RandomForestClassifier()
cross_val_scores_rf = cross_val_score(rf, X_res, y_res, cv=5)
print("Random Forest CV Accuracy: %0.2f (+/- %0.2f)" % (cross_val_scores_rf.mean(), cross_val_scores_rf.std() * 2))

# Decision Tree
dt = DecisionTreeClassifier()
cross_val_scores_dt = cross_val_score(dt, X_res, y_res, cv=5)
print("Decision Tree CV Accuracy: %0.2f (+/- %0.2f)" % (cross_val_scores_dt.mean(), cross_val_scores_dt.std() * 2))
```



```
Logistic Regression CV Accuracy: 0.96 (+/- 0.02)
SVM CV Accuracy: 0.84 (+/- 0.10)
Random Forest CV Accuracy: 0.98 (+/- 0.03)
Decision Tree CV Accuracy: 0.97 (+/- 0.03)
```

CONCLUSION

- In summary, we tried to predict the correlation between the factors and cervical cancer.
- The online dataset was collected from Kaggle.
- Data cleaning was done by replacing the placeholders with standard missing value representation(NaN), dropping the columns, converting the subjective data to categorical, removing the outliers of numerical variables and removing the outliers of numerical variables and removing the null values from the data.
- Data visualization was done by plotting the histograms of the variables.
- We developed the models including Logistic Regression, Support Vector Machine, Random Forest and Decision Tree
- We compared the accuracy of different models on Train Data and Test Data.
- We found that the Decision Tree has comparatively given the prediction better.
- We have not observed enough accuracy to conclude that this dataset helps predict the correlation between those factors and cervical cancer incidence.
- We are accepting the null hypothesis as the observed p-value is greater than 0.05.