Cervical Cancer Risk Classification

prediction of cancer indicators

About Dataset

Cervical Cancer Risk Factors for Biopsy: This Dataset is Obtained from UCI Repository and kindly acknowledged!

This file contains a List of Risk Factors for Cervical Cancer leading to a Biopsy Examination!

About 11,000 new cases of invasive cervical cancer are diagnosed each year in the U.S. However, the number of new cervical cancer cases has been declining steadily over the past decades. Although it is the most preventable type of cancer, each year cervical cancer kills about 4,000 women in the U.S. and about 300,000 women worldwide. In the United States, cervical cancer mortality rates plunged by 74% from 1955 - 1992 thanks to increased screening and early detection with the Pap test. AGE Fifty percent of cervical cancer diagnoses occur in women ages 35 - 54, and about 20% occur in women over 65 years of age. The median age of diagnosis is 48 years. About 15% of women develop cervical cancer between the ages of 20 - 30. Cervical cancer is extremely rare in women younger than age 20. However, many young women become infected with multiple types of human papilloma virus, which then can increase their risk of getting cervical cancer in the future. Young women with early abnormal changes who do not have regular examinations are at high risk for localized cancer by the time they are age 40, and for invasive cancer by age 50. SOCIOECONOMIC AND ETHNIC FACTORS Although the rate of cervical cancer has declined among both Caucasian and African-American women over the past decades, it remains much more prevalent in African-Americans -- whose death rates are twice as high as Caucasian women. Hispanic American women have more than twice the risk of invasive cervical cancer as Caucasian women, also due to a lower rate of screening. These differences, however, are almost certainly due to social and economic differences. Numerous studies report that high poverty levels are linked with low screening rates. In addition, lack of health insurance, limited transportation, and language difficulties hinder a poor woman’s access to screening services. HIGH SEXUAL ACTIVITY Human papilloma virus (HPV) is the main risk factor for cervical cancer. In adults, the most important risk factor for HPV is sexual activity with an infected person. Women most at risk for cervical cancer are those with a history of multiple sexual partners, sexual intercourse at age 17 years or younger, or both. A woman who has never been sexually active has a very low risk for developing cervical cancer. Sexual activity with multiple partners increases the likelihood of many other sexually transmitted infections (chlamydia, gonorrhea, syphilis).Studies have found an association between chlamydia and cervical cancer risk, including the possibility that chlamydia may prolong HPV infection. FAMILY HISTORY Women have a higher risk of cervical cancer if they have a first-degree relative (mother, sister) who has had cervical cancer. USE OF ORAL CONTRACEPTIVES Studies have reported a strong association between cervical cancer and long-term use of oral contraception (OC). Women who take birth control pills for more than 5 - 10 years appear to have a much higher risk HPV infection (up to four times higher) than those who do not use OCs. (Women taking OCs for fewer than 5 years do not have a significantly higher risk.) The reasons for this risk from OC use are not entirely clear. Women who use OCs may be less likely to use a diaphragm, condoms, or other methods that offer some protection against sexual transmitted diseases, including HPV. Some research also suggests that the hormones in OCs might help the virus enter the genetic material of cervical cells. HAVING MANY CHILDREN Studies indicate that having many children increases the risk for developing cervical cancer, particularly in women infected with HPV. SMOKING Smoking is associated with a higher risk for precancerous changes (dysplasia) in the cervix and for progression to invasive cervical cancer, especially for women infected with HPV. IMMUNOSUPPRESSION Women with weak immune systems, (such as those with HIV / AIDS), are more susceptible to acquiring HPV. Immunocompromised patients are also at higher risk for having cervical precancer develop rapidly into invasive cancer. DIETHYLSTILBESTROL (DES) From 1938 - 1971, diethylstilbestrol (DES), an estrogen-related drug, was widely prescribed to pregnant women to help prevent miscarriages. The daughters of these women face a higher risk for cervical cancer. DES is no longer prsecribed.

DATASET INFERENCE

Cervical cancer is a type of cancer that occurs in the cells of the cervix — the lower part of the uterus that connects to the vagina.

Various strains of the human papillomavirus (HPV), a sexually transmitted infection, play a role in causing most cervical cancer.

When exposed to HPV, the body's immune system typically prevents the virus from doing harm. In a small percentage of people, however, the virus survives for years, contributing to the process that causes some cervical cells to become cancer cells.

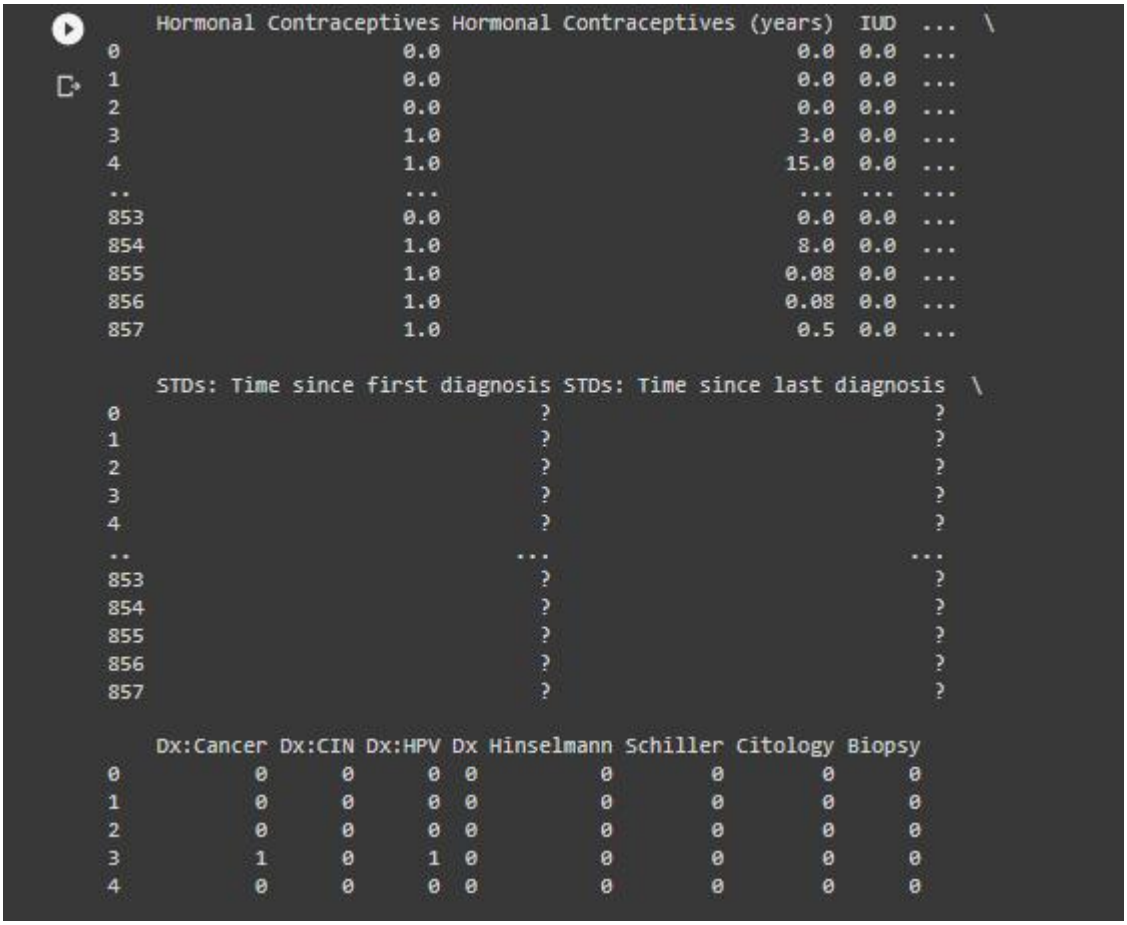
You can reduce your risk of developing cervical cancer by having screening tests and receiving a vaccine that protects against HPV infection.

Risk factors for cervical cancer include

- Many sexual partners. The greater your number of sexual partners — and the greater your partner's number of sexual partners — the greater your chance of acquiring HPV.
- Early sexual activity. Having sex at an early age increases your risk of HPV.
- Other sexually transmitted infections (STIs). Having other STIs — such as chlamydia, gonorrhea, syphilis and HIV/AIDS — increases your risk of HPV.
- A weakened immune system. You may be more likely to develop cervical cancer if your immune system is weakened by another health condition and you have HPV.
- Smoking. Smoking is associated with squamous cell cervical cancer.
- Exposure to miscarriage prevention drug. If your mother took a drug called diethylstilbestrol (DES) while pregnant in the 1950s, you may have an increased risk of a certain type of cervical cancer called clear cell adenocarcinoma.

PRINGING DATASET

```
1.import pandas as pd
import io
df=pd.read_csv(io.BytesIO(uploaded['kag_risk_factors_cervical_cancer.csv']))
print(df)
```



	Hormonal Contraceptives	Hormonal Contraceptives (years)	IUD	...	\
0	0.0	0.0	0.0	...	
1	0.0	0.0	0.0	...	
2	0.0	0.0	0.0	...	
3	1.0	3.0	0.0	...	
4	1.0	15.0	0.0	...	
..	...	...	...	...	
853	0.0	0.0	0.0	...	
854	1.0	8.0	0.0	...	
855	1.0	0.08	0.0	...	
856	1.0	0.08	0.0	...	
857	1.0	0.5	0.0	...	

	STDs: Time since first diagnosis	STDs: Time since last diagnosis	\
0	?	?	
1	?	?	
2	?	?	
3	?	?	
4	?	?	
..	...	...	
853	?	?	
854	?	?	
855	?	?	
856	?	?	
857	?	?	

	Dx:Cancer	Dx:CIN	Dx:HPV	Dx	Hinselmann	Schiller	Citology	Biopsy
0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0
3	1	0	1	0	0	0	0	0
4	0	0	0	0	0	0	0	0

EDA and PREPROCESSING

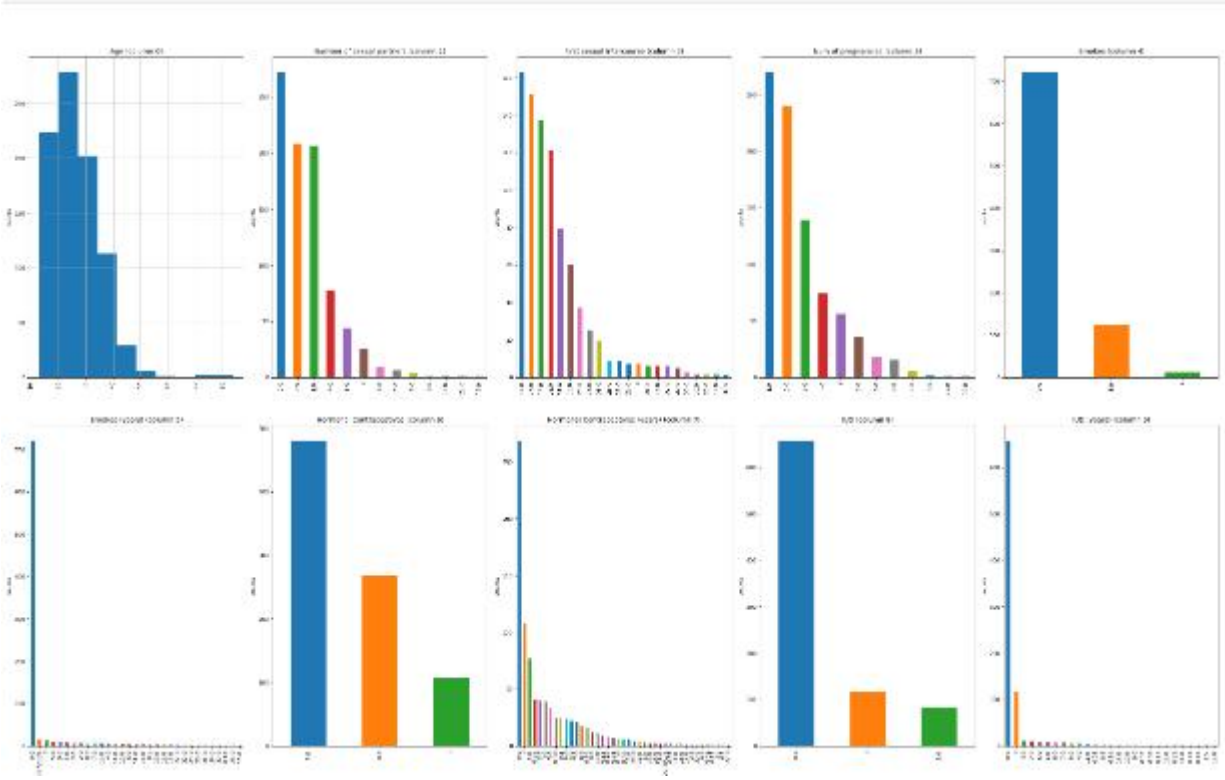
1.
We could find that our dataset has some '?' . So let us replace all the '?' with 0 and then replace that 0 with the median.

```
for feature in df.columns:
    df[feature].replace('?',np.nan,inplace=True )
    df[feature].fillna(value=0,inplace=True)
for feature in df.columns:
    df[feature].replace(0,df[feature].median(),inplace=True)
df.head()
```

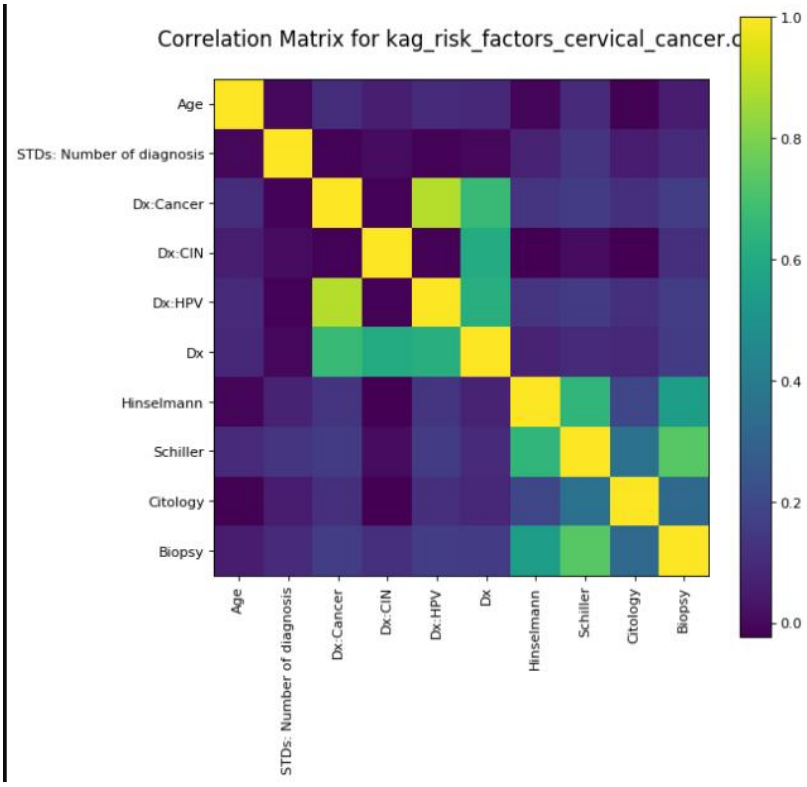
2

```
# Remove columns with all na values
all_na = df.columns[df.isna().all()]
df.drop(all_na, axis = 1, inplace = True)
# percentage of na values in all columns
df.isna().sum()/df.shape[0]*100
```

Distribution graphs (histogram/bar graph) of sampled columns:



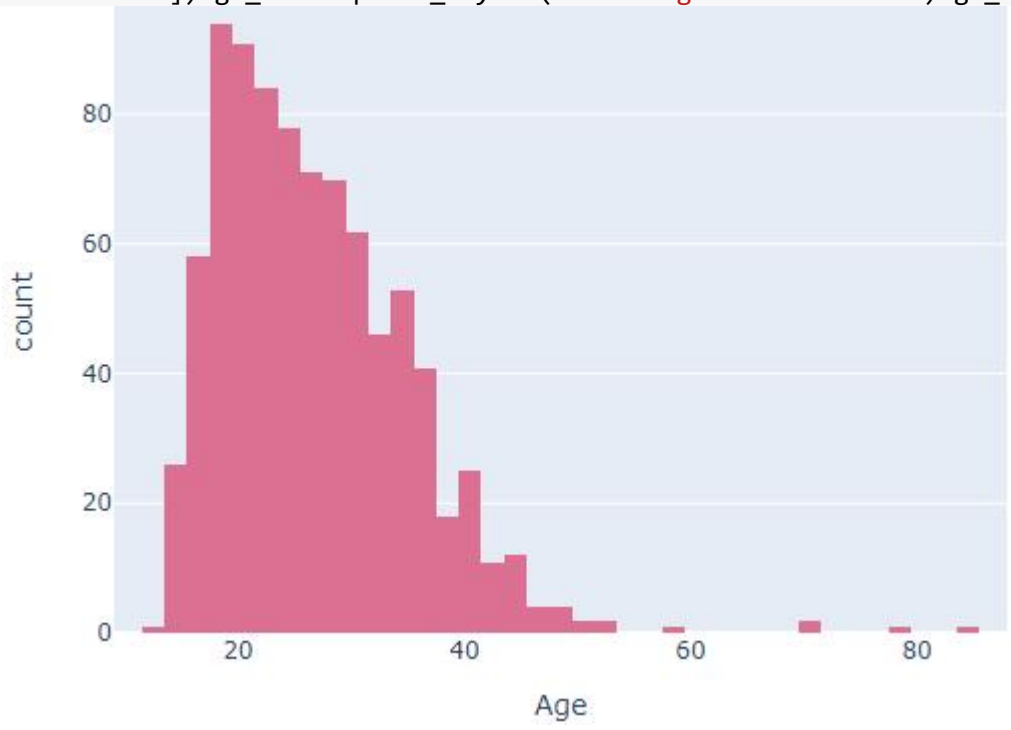
Correlation matrix for entire dataset



INFERNCE:
From the heatmap, we can see that there a correlation coefficent very close to 0, this indicates that, from the data, the number of sexual partners does not have any linear relationship with any of the respective diagnoses. However, we also visually knew that the number of sexual partners remained fairly consistent across age ranges and therefore there are more likely causes of HPV and Cervical Cancer than number of sexual partners with respect to the data.

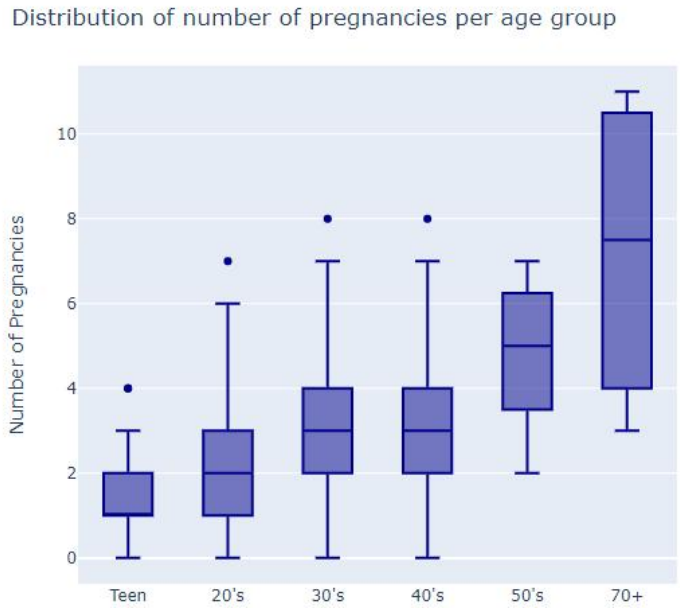
AGE DISTRUBUTION

```
age_dist = px.histogram(risk_factor_df, x="Age", marginal="box", color_discrete_sequence=
["palevioletred"])age_dist.update_layout(title="Age distribution")age_dist.show()
```



Pregnancy Distribution by Age

```
age_preg_bar = px.box(risk_factor_df.sort_values(by="Age",ascending=True), x="age_cat", y=
"Num of pregnancies",
                      color_discrete_sequence=["darkblue"], points="outliers",
                      category_orders=["Teenager", "Twenties", "Thirties", "Forties", "Fif
ties",
                                       "Seventy and over"])age_preg_bar.update_xaxes(title
="Age Category")age_preg_bar.update_yaxes(title="Number of Pregnancies")age_preg_bar.updat
e_layout(title="Distribution of number of pregnancies per age group")age_preg_bar.show()
```



Tests used

Here we observe the number of tests done by patients to determine if they have Cerivcal Cancer / HPV.

The tests used were:

Hinselmann

A colposcopy is a type of cervical cancer test. It lets your doctor or nurse get a close-up look at your cervix — the opening to your uterus. It's used to find abnormal cells in your cervix. Source

Citology

Cytology is the exam of a single cell type, as often found in fluid specimens. It's mainly used to diagnose or screen for cancer. It's also used to screen for fetal abnormalities, for pap smears, to diagnose infectious organisms, and in other screening and diagnostic areas. Source

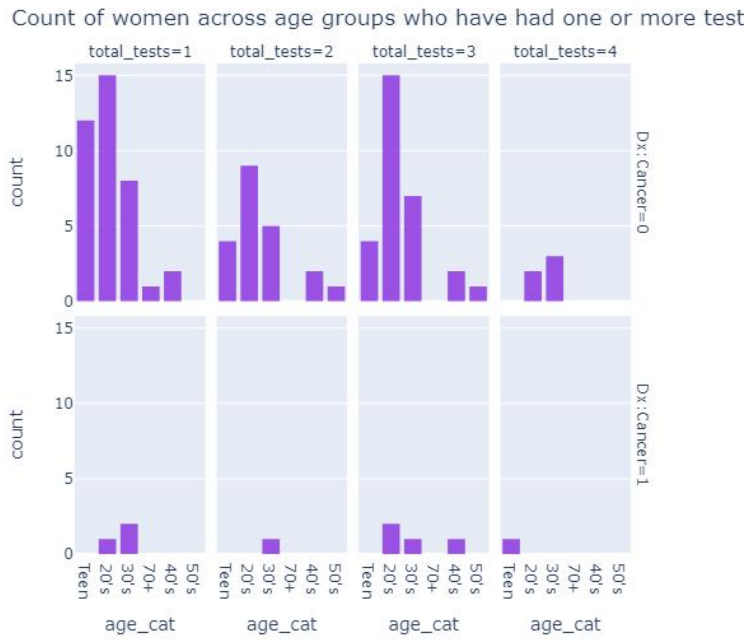
Biopsy

A cervical biopsy is a procedure to remove tissue from the cervix to test for abnormal or precancerous conditions, or cervical cancer. Source

Schiller

A test in which iodine is applied to the cervix. The iodine colors healthy cells brown; abnormal cells remain unstained, usually appearing white or yellow.

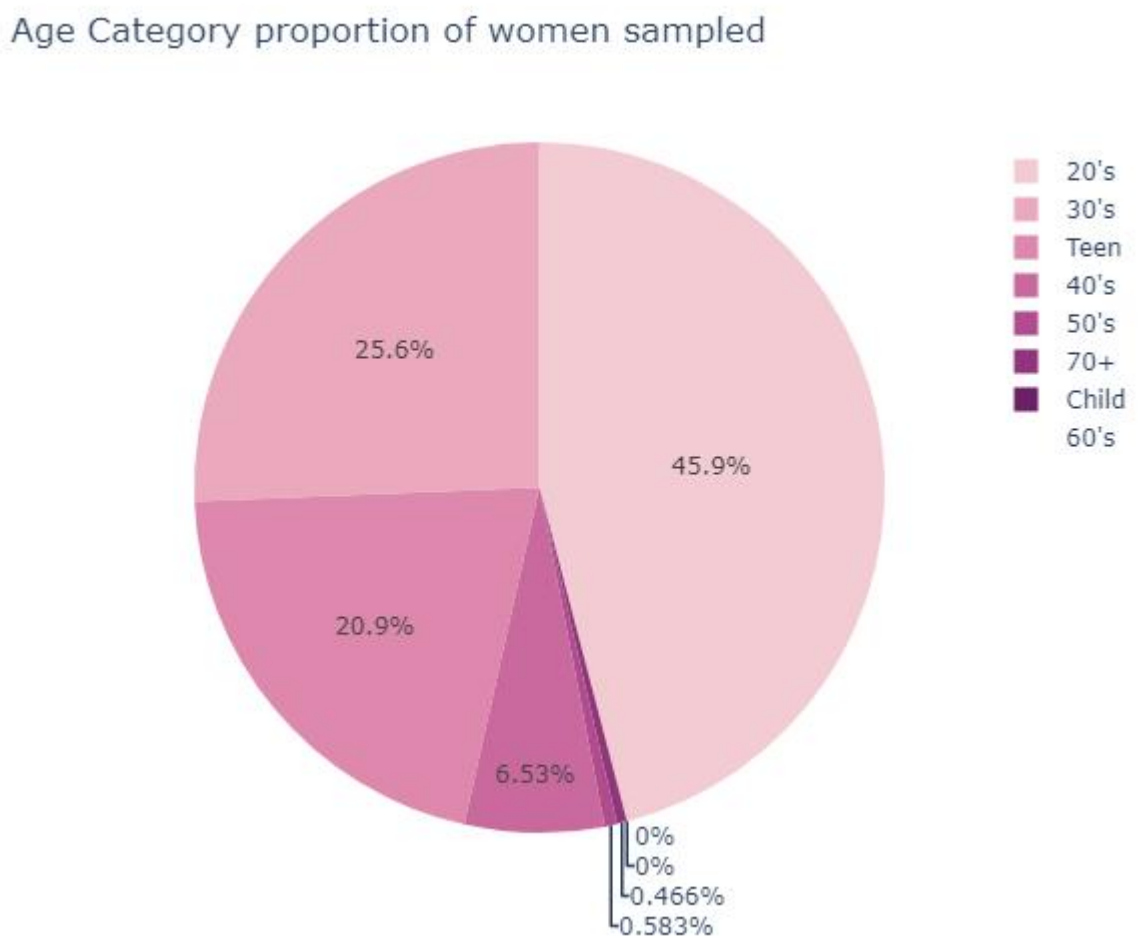
```
fig = px.histogram(risk_factor_df.query("total_tests>0").sort_values(by="total_tests", asc
ending=True),
                   x="age_cat",
                   facet_col="total_tests",
                   facet_row="Dx:Cancer",
                   color_discrete_sequence=["blueviolet"],
                   opacity=0.8)fig.update_layout(title="Count of women across age groups w
ho have had one or more test")
fig.show()
```



Proportions of women who have Cervical Cancer / HPV

This represents the proportion of women by age category who were diagnosed with Cervical Cancer/ HPV. It is seen that women in their 30's have the most prevalance of Cervical Cancer and HPV, followed by women in their 20's.

It is also seen that of all the samples taken, approximately 26% are of women in their 30's. With respect to the women who have cervical cancer, approximately 44% of cases are women in their 30's, also, out of the women who have HPV, approximately 39% of women are in their 30's. This is contrasted with 45% of all samples being women in their 20's and only 28% of the women have cancer are in their 20's, HPV is more comparable at 33%.



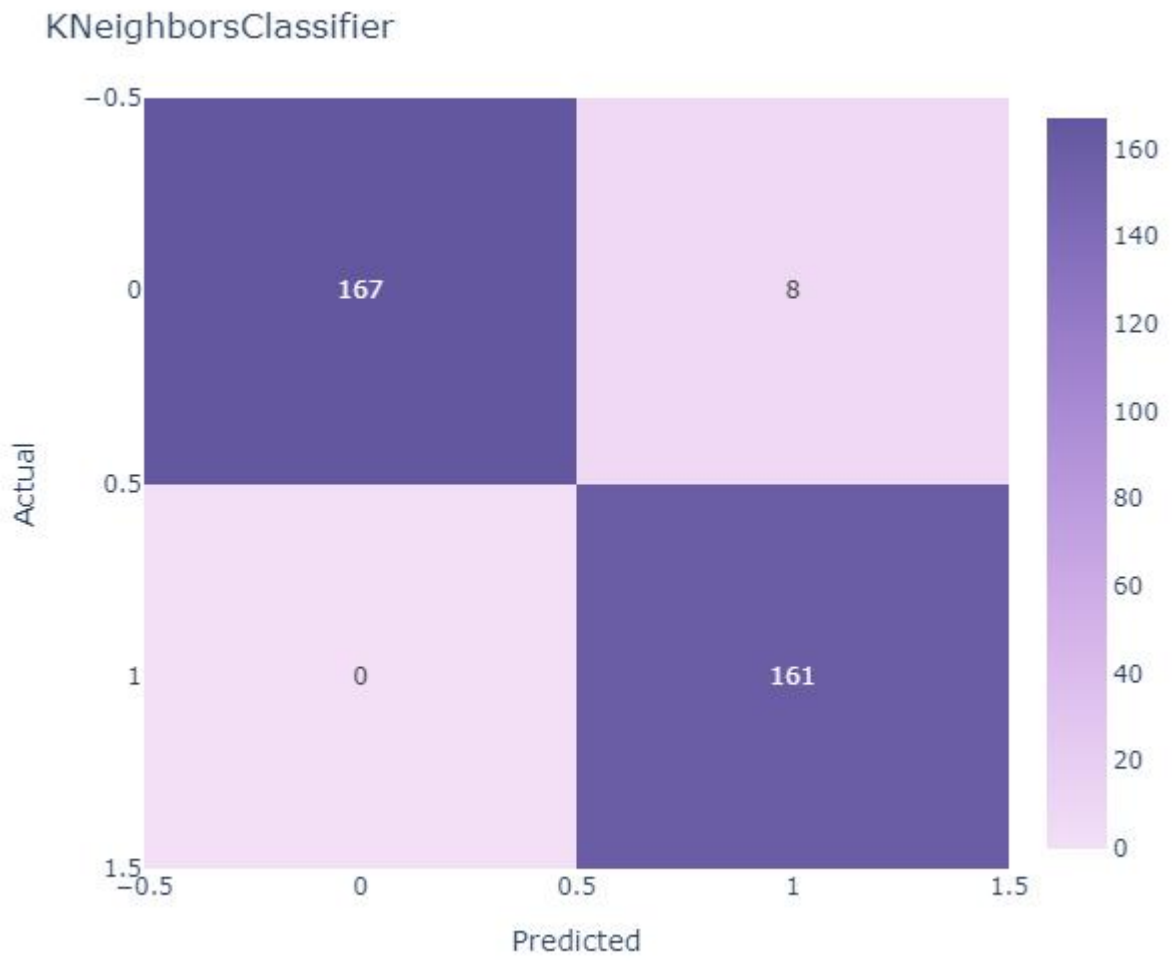
Train-Test Split¶

Data split was stratified on **Age Category**

```
train_set = Nonetest_set = None split = StratifiedShuffleSplit(n_splits=1, test_size=0.2,
random_state=42)for train_idx, test_idx in split.split(risk_factor_df, risk_factor_df["age
_cat"]):
    train_set = risk_factor_df.loc[train_idx]
    test_set = risk_factor_df.loc[test_idx]cols_to_drop = ["age_cat","total_std","total_te
sts"]for set_ in (train_set, test_set):
    for col in cols_to_drop:
        set_.drop(col, axis=1, inplace=True)
X_train = train_set.drop("Dx:Cancer", axis=1)y_train = train_set["Dx:Cancer"].copy()
X_test = test_set.drop("Dx:Cancer", axis=1)y_test = test_set["Dx:Cancer"].copy()
```

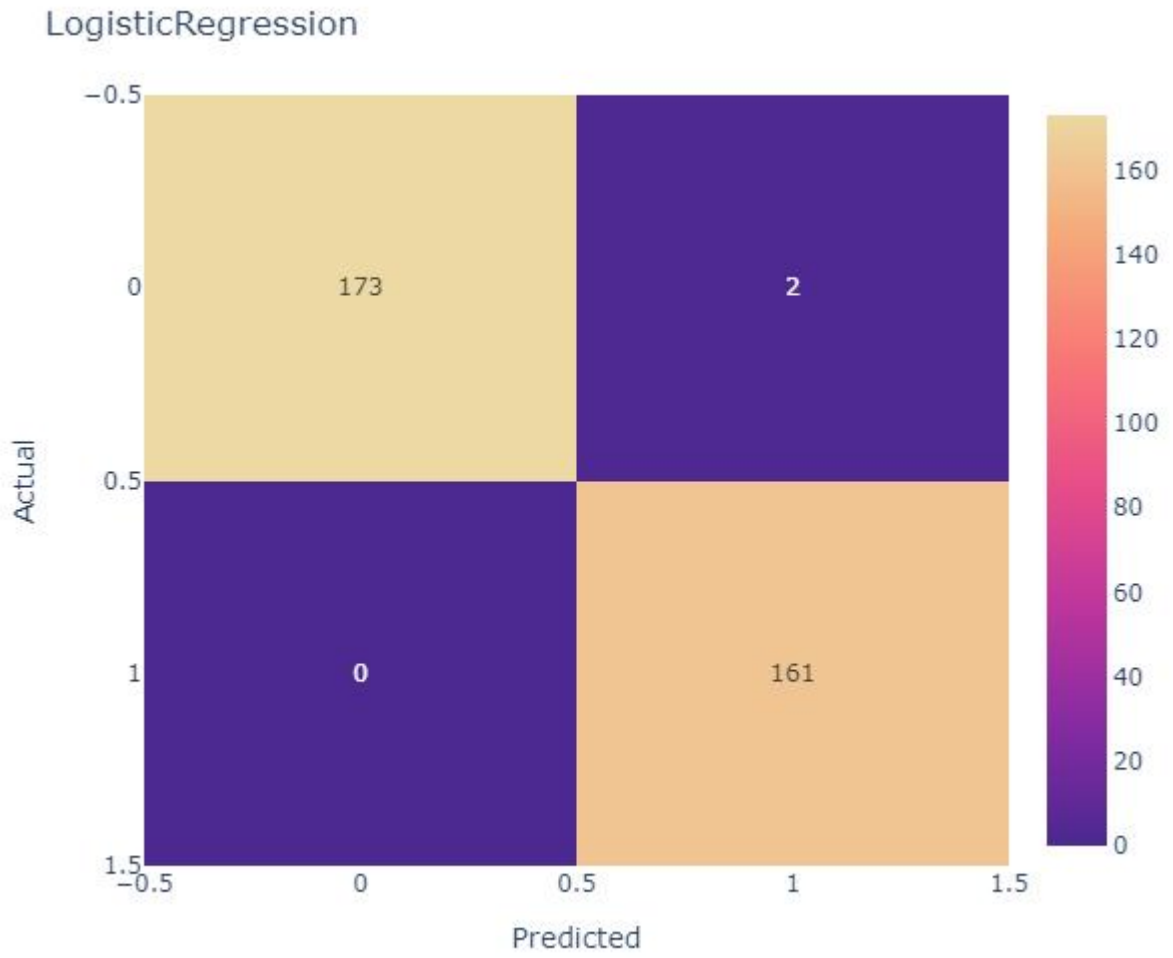
KNN

```
knn_clf = KNeighborsClassifier()knn_param_grid = {"n_neighbors": list(np.arange(1, 100,
2))}knn_clf_cv = GridSearchCV(knn_clf, knn_param_grid, cv=10,refit=True).fit(X_train,y_tra
in)knn_clf_cv = KNeighborsClassifier(**knn_clf_cv.best_params_)
```



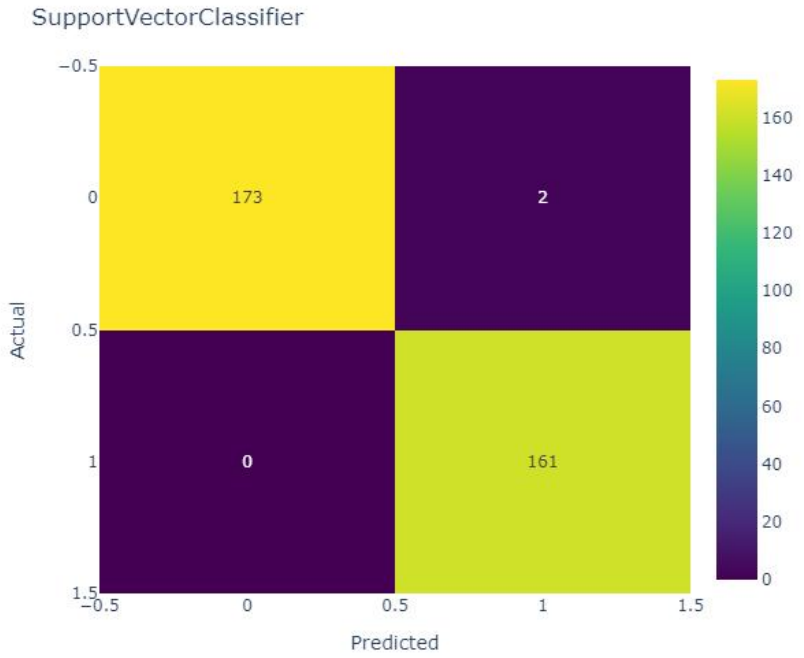
LOGISTIC REGRESSION:

```
param_grid = {'C': np.logspace(-5, 8, 15)}logreg = LogisticRegression()logreg_cv = GridSea
rchCV(logreg, param_grid, cv=10,refit=True).fit(X_train,y_train)logreg_cv = LogisticRegres
sion(**logreg_cv.best_params_)
```



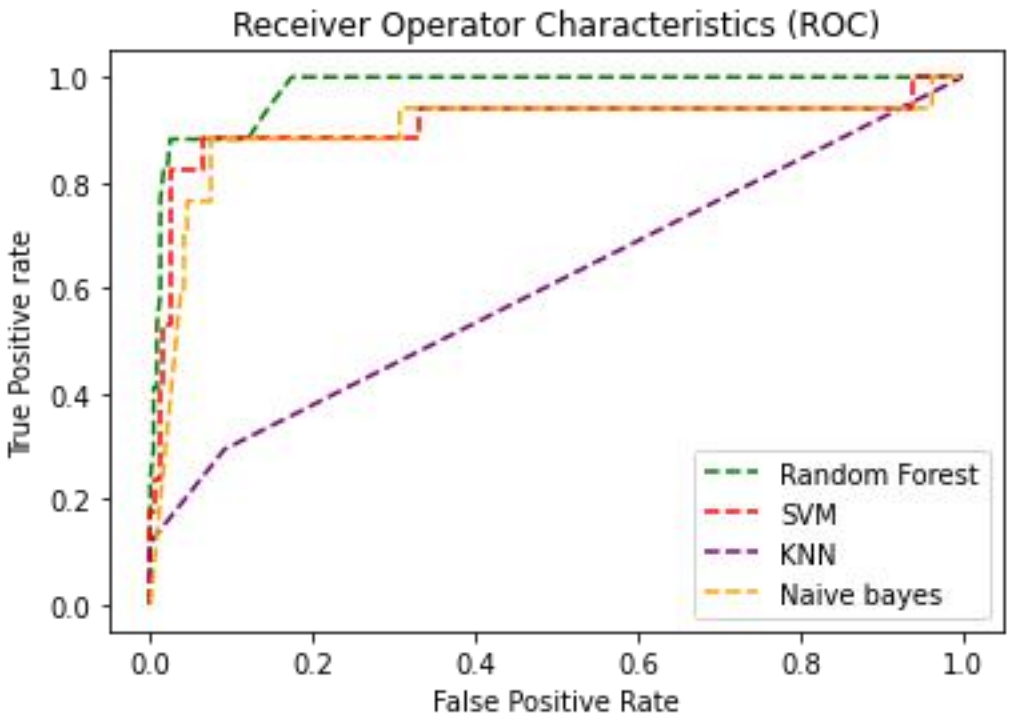
SVM:

```
svm_clf = SVC()svc_param_grid = {'C': np.logspace(-3, 2, 6), 'gamma': np.logspace(-3, 2,
6), }svm_clf_cv = GridSearchCV(svm_clf, svc_param_grid, cv=5)
```



NAIVE BAYES CLASSIFIER

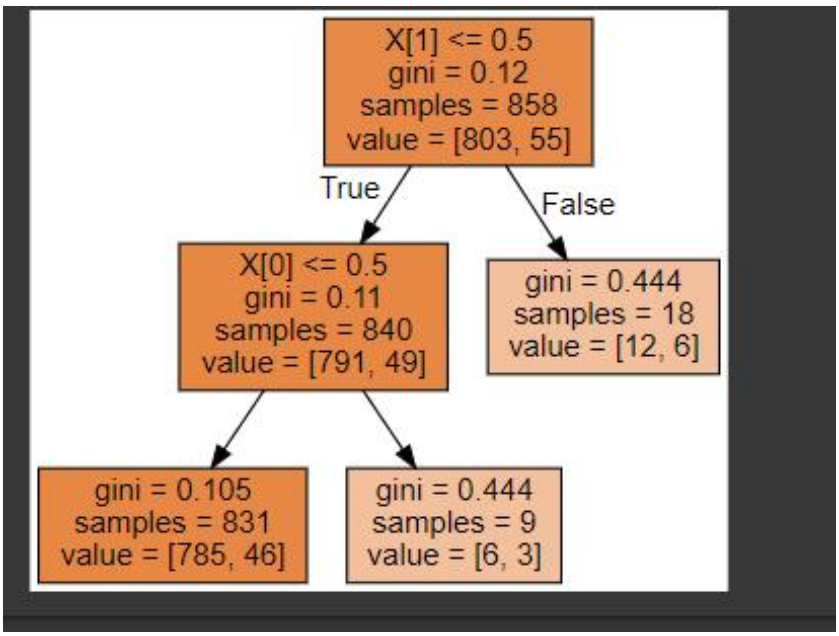
	precision	recall	f1-score	support
0.0	0.99	0.82	0.90	241
1.0	0.26	0.88	0.40	17
accuracy			0.83	258
macro avg	0.62	0.85	0.65	258
weighted avg	0.94	0.83	0.87	258



DECISION TREE :

```
features = ['Dx:CIN', 'Dx:HPV']
xx=features
X = df[features]
y = df['Biopsy']
print(X)
print(y)
```

A PATINENT WILL HAVE TO UNDERGO BIOPSY TEST IF HE IS BOTH POSITIVE IN DX:CIN AND DX:HPV



Interpretation of the results

-

TP: True Positive, these are the values that are positive and were predicted positive

-
-

FP: False Positive, The values which are negative but were wrongly predicted as positive

-
-

TN: True Negative, these are the values that are negative and were predicted negative

-
-

FN: False Negative, The values which are positive but were wrongly predicted as negative

-

Precision:This metric measures the actual positive outcomes out of the total predicted positive outcomes. It attempts to identify the proportion of positive identifications that were correct. The Logistic Regression model and Support Vector Classifer model performed equally well with a precision score of 99.41%.

In the context of diagnoising cervical cancer, this is metric would not be the most ideal to measure performance, as a negative case being labelled as a positive case is easily solved with confirmatory tests. However, one has to also consider the emotional and mental issues brought upon by being diagnosed with cervical cancer, as this can have a lingering effect even after having confirmatory tests. These tests should be done as soon as possible, as there may be another underlying illness that brought them to see a healthcare professional in the first place.

Recall

This metirc measures the correctly positive predicted outocmes of the total number of positive outcomes. It answers the question of what proportions of actual positives were identified correctly. The Logistic Regression model and Support Vector Classifer model performed equally well with a recall score of 99.4%. In terms of measuring performance of the model, this is the metric that should be highly considered.

In the context of diagnosing cervical cancer, we want to reduce the number of false negatives (Actual positive cases labelled as negative cases) as much possible. If an actual positive case is labelled as negative, this has serious consequecnes as the patient would go about their life without actually receiving potentially life saving treatment.

There are many reasons why a cancer can go misdiagnosed, these include:

- The symptoms,especially in the early stages being mistaken for some other type of less serious illness.
- The actual test adminstered by a healthcare professional may give the wrong diagnosis

The 5-year survival rate tells you what percent of people live at least 5 years after the cancer is found. Percent means how many out of 100. The 5-year survival rate for all people with cervical cancer is 66%. Source

Survival rates also depend on the stage of cervical cancer that is diagnosed. When detected at an early stage, the 5-year survival rate for people with invasive cervical cancer is 92%. About 44% of people with cervical cancer are diagnosed at an early stage. If cervical cancer has spread to surrounding tissues or organs and/or the regional lymph nodes, the 5-year survival rate is 58%. If the cancer has spread to a distant part of the body, the 5-year survival rate is 18%. Source

It is clearly important and evident that a correct diagnosis and early treatment is the best possible way to ensure that a patient has a high chance of surviving.

F1 Score

The F1 score is defined as the harmonic mean of precision and recall. Therefore, a high F1 score means both a high precision and recall, same for low and a medium score if one score is high and the other is low.

The Logistic Regression model and Support Vector Classifer model performed equally well with an accuracy score of 99.4%

Accuracy

The Logistic Regression model and Support Vector Classifer model performed equally well with an accuracy score of 99.4%