Cervical Cancer Risk Prediction

About Dataset:

Link: <https://www.kaggle.com/datasets/loveall/cervical-cancer-risk-classification>

Cervical Cancer Risk Factors for Biopsy: This Dataset is Obtained from kaggle repository.

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

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 roleincausingmost 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.

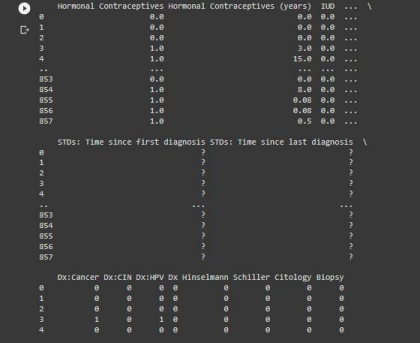
∙ Exposure to miscarriage prevention drug. If your mother took a drug called diethylstilbestrol (DES)while pregnant in the 1950s, you may have an increased risk of a certain type of cervical cancer called clear cell adenocarcinoma.

PRINGING DATASET

1.import pandas as pd

import io

df=pd.read\_csv(io.BytesIO(uploaded['kag\_risk\_factors\_cervical\_cancer.csv'])) print(df)



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

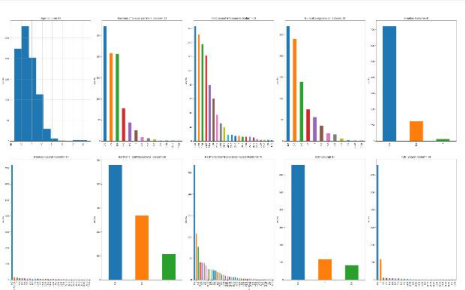
all\_na = df.columns[df.isna().all()]

df.drop(all\_na, axis = 1, inplace = True)

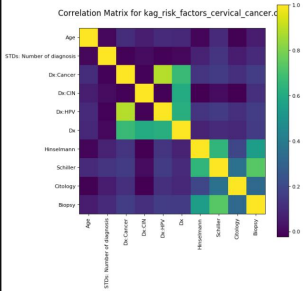
# percentage of na values in all columns

df.isna().sum()/df.shape[0]\*100

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



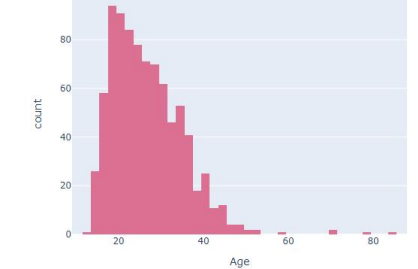
Correlation matrix for entire dataset



INFERENCE:

From the heatmap, we can see that there a correlation coefficent very close to 0, this indicatesthat, fromthe data, the number of sexual partners does not have any linear relationship with any of therespectivediagnoses. However, we also visually knew that the number of sexual partners remainedfairlyconsistentacross age ranges and therefore there are more likely causes of HPV and Cervical Cancer thannumberof 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()

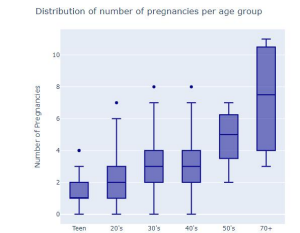
**Pregnancy Distribution by Age**

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.update\_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.

**Cytology**

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.

**Biopsy**

A cervical biopsy is a procedure to remove tissue from the cervix to test for abnormal or pre cancerous 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",ascending=True),

x="age\_cat",

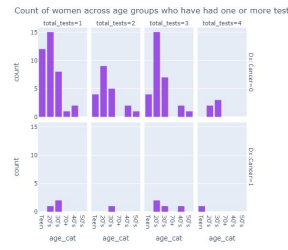
facet\_col="total\_tests",

facet\_row="Dx:Cancer",

color\_discrete\_sequence=["blueviolet"],

opacity=0.8)fig.update\_layout(title="Count of women across agegroupswho have had one or more test")

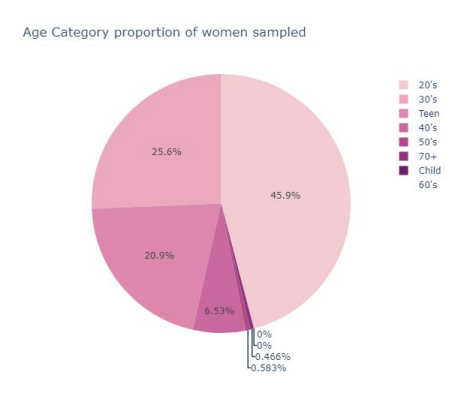
fig.show()



**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 prevalence of Cervical Cancer andHPV, 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. Withrespectto the women who have cervical cancer, approximately 44% of cases are women in their 30's, also, outof the women who have HPV, approximately 39% of women are in their 30's. This is contrastedwith45%of all samples being women in their 20's and only 28% of the women have cancer are in their 20's, HPVismore 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\_tests"]for set\_ **in** (train\_set, test\_set):

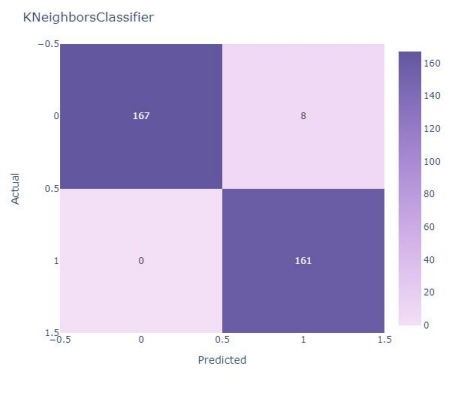
for col **in** cols\_to\_drop:

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()

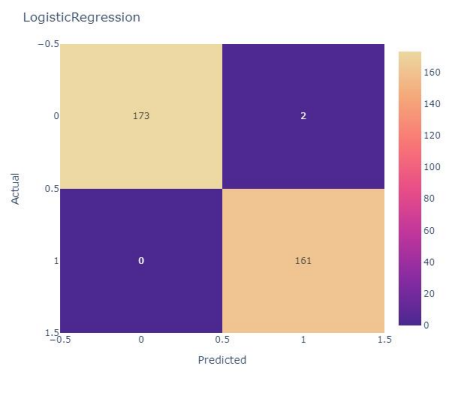
**KNN:**

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\_train)knn\_clf\_cv = KNeighborsClassifier(\*\*knn\_clf\_cv.best\_params\_)



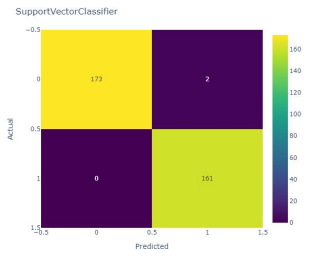
**LOGISTIC REGRESSION:**

param\_grid = {'C': np.logspace(-5, 8, 15)}logreg = LogisticRegression()logreg\_cv=GridSearchCV(logreg, param\_grid, cv=10,refit=True).fit(X\_train,y\_train)logreg\_cv = LogisticRegression(\*\*logreg\_cv.best\_params\_)



SVM:

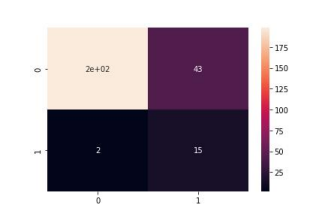
svm\_clf = SVC()svc\_param\_grid = {'C': np.logspace(-3, 2, 6), 'gamma': np.logspace(-3,2,6), }svm\_clf\_cv = GridSearchCV(svm\_clf, svc\_param\_grid, cv=5)

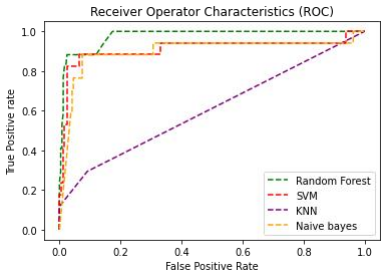


**NAIVE BAYES CLASSIFIER**

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']

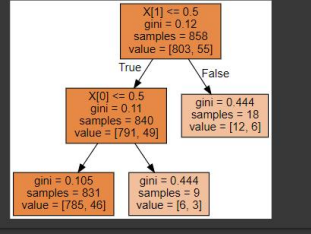
xx=features

X = df[features]

y = df['Biopsy']

print(X)

print(y)

A PATIENT WILL HAVE TO UNDERGO BIOPSY TEST IF HE IS BOTH POSITIVE IN DX:CINANDDX: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. TheLogisticRegression model and Support Vector Classifier model performed equally well with a precision score of 99.41%.

In the context of diagnosing cervical cancer, this is metric would not be the most ideal to measure performance, as a negative case being labeled 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 metric measures the correctly positive predicted outcomes of the total number of positive outcomes.It answers the question of what proportions of actual positives were identified correctly. TheLogisticRegression model and Support Vector Classifier 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 labelledasnegative, this has serious consequecnes as the patient would go about their life without actuallyreceiving 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 administered by a healthcare professional may give the wrong diagnosisThe 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 is66%.

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%.

**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 slow.

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

**Accuracy:**

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