**DETERMINATION OF CERVICAL CANCER**

**using machine learning algorithms**

Thesis Submitted in fulfillment of the Requirement for the degree

Of

**Bachelor of Technology (B.Tech)**

**in**

**Department of Computer Science and Engineering**

**By**

**Anisha Chakraborty**

**(Roll No.500118021015 Reg. No.181430110012)**

**Barnali Guha Thakurata**

**(Roll No.500118021031 Reg. No.181430110028)**

**Disha Das**

**(Roll No.500118021049 Reg. No.181430110045)**

**Nayanika Nandi**

**(Roll No.500118021063 Reg. No.181430110059)**

**Under the guidance of**

**Prof. Bidyutmala Saha**

**Assistant Professor**

**C:\Users\SONY\Desktop\logo.png**

**Guru Nanak Institute of Technology, Kolkata-700110**

**ACKNOWLEDGEMENT**

We find it as an auspicious occasion to express our deep regard and great admiration to our mentor and project guide Mrs. Bidyutmala Saha Asst. Professor of Department of Computer Science Engineering, Guru Nanak Institute of Technology, for her untiring endeavour throughout the course of the thesis work at every step by valuable guidance and advice, persistent encouragement, constant inducement and supervision which made the completion of this work at right time.

We are very much thankful to all Professor of Department of Computer Science Engineering for their support and the valuable help, suggestion and encouragement extended to us.

Date:20-05-2022

**Anisha Chakraborty**

**Barnali Guha Thakurata**

**Disha Das**

**Nayanika Nandi**

**GURUNANAK INSTITUTE OF TECHNOLOGY**

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

**CERTIFICATE**

We hereby recommend that the thesis prepared under our supervision by**, Anisha Chakraborty(500118021015)Barnali Guha Thakurata(500118021031) Disha Das(500118021049) Nayanika Nandi(500118021063).** Entitled “**DETERMINATION OF CERVICAL CANCER using machine learning algorithms**” be accepted in the partial fulfilment for paper PR892of Bachelor of Technology in Computer Science and Engineering by Guru Nanak Institute Of Technology (Autonomous Institute).

**Supervisor**

----------------------------------------------

**Bidyutmala Saha(Assistant Professor)**

Department of Computer Science and Engineering

GNIT, Kolkata.

**Head of the Department**

**----------------------------------**

**Mrs. Sangeeta Bhattacharya**

Head of the Department

Department of Computer Science and Engineering,

GNIT, Kolkata

**Contents**

**Page No.**

* Abstract 05
* Introduction 06
* Problem Statement 07
* Proposed System 08
* Implementation of System Model 08-11
  1. Modules
* Dataset 12
* Data Manipulation 13-14
* Data Visualization 15-16
* Data Preprocessing 17
* Data Splitting 17
* Experimental Results 18-20
  1. Test Results
* Conclusion 21
* Future Study 22
* References 23

**ABSTRACT**

Cervical cancer is the fourth most frequent cancer in women representing 7.9% of female cancers as stated by the World Health Organization. One of the leading causes of this has been the lack of effective early treatment. Therefore, early detection of cervical cancer is of utmost importance. The asymptomatic nature is the major challenge faced in the diagnosis of cervical cancer in the early stage. Machine learning has been of great help in many medical applications and can be used as a classifier in the early detection of the cancerous cells present in the cervix region of the uterus. In this paper, a survey and analysis of the various machine learning approaches that have been implemented for the diagnosis of cervical cancer is done. The survey paper draws a comparison of the various existing techniques for the prediction of cervical cancer using medical data and points out their advantages and shortcomings.

**INTRODUCTION**

Cervical cancer starts in a woman's cervix, which is the lower, narrow part of the uterus. The cervix connects the lower part of the uterus to the vagina and forms the birth canal. Cervical cancer begins when healthy cells on the surface of the cervix change and grow out of control, forming a mass called a tumor. Cervical cancer begins with abnormal changes in the cervical tissue. Cervical cancer begins with abnormal changes in the cervical tissue. The risk of developing these abnormal changes is associated with infection with human papilla mavirus (HPV). Genetic material that comes from certain forms of HPV (high-risk subtypes) has been found in cervical tissues that show cancerous or precancerous changes. In 2018 an estimated 5,70,000 woman were diagnosed with cervical cancer worldwide and about 3,11,000 woman died from the disease.

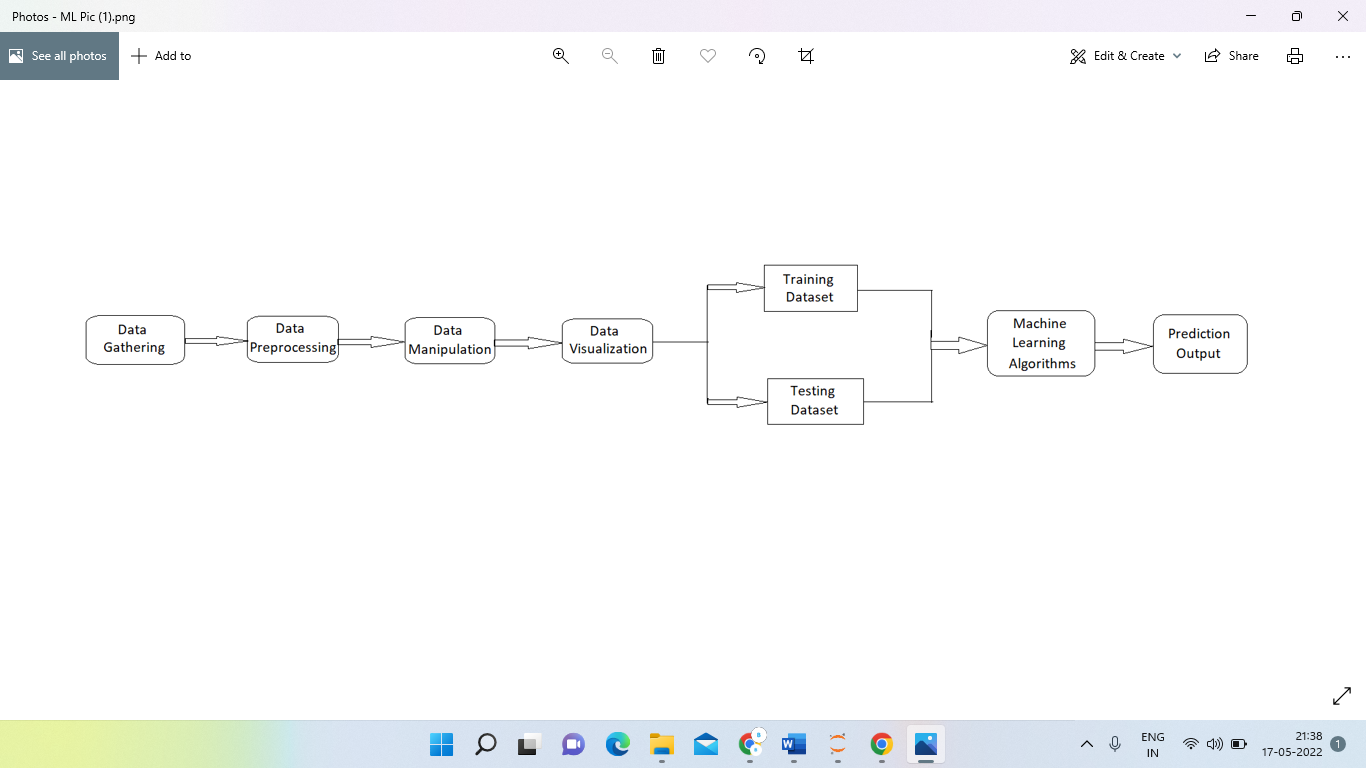
**PROBLEM STATEMENT**

The early prediction of any disease gives a patient better chances of successful treatment than disease discovery at an advanced stage of its development. If we do not know how to treat patients, any treatment we can provide would be useful and would provide a more comfortable life.

In this project our aim is to build some machine learning models to predict the possibility of cervical cancer in women using different machine learning algorithms to train the model capable of diagnosing cervical cancer with high accuracy and sensitivity.

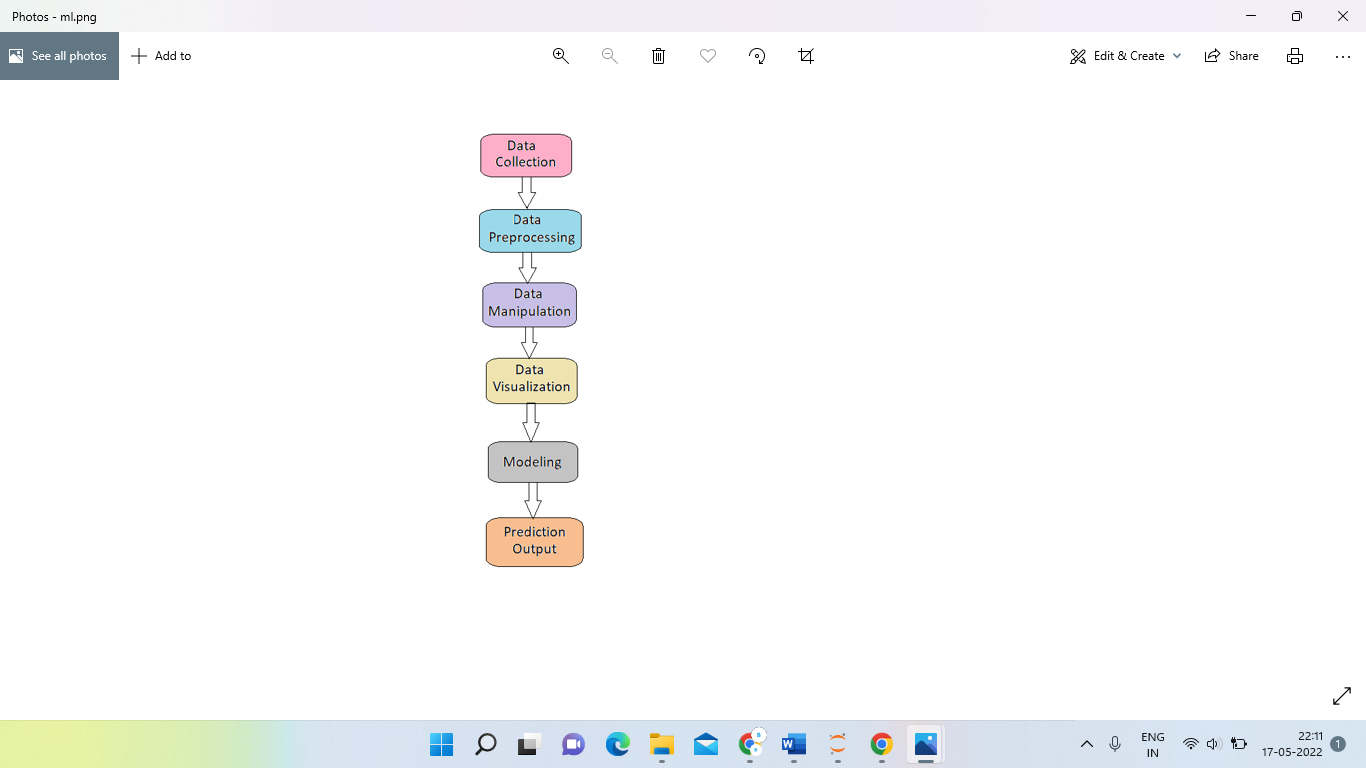
**PROPOSED SYSTEM**

The proposed work sets itself apart by harnessing the powers of Machine Learning. In the project work, a system, with a strong prediction algorithm, which implements powerful classification steps with a comprehensive report generation module.



**IMPLEMENTATION of SYSTEM MODEL**

**MOLULES:**



**DATA COLLECTION:**

Data collection is the first step of the machine learning life cycle. The goal of this step is to identify and obtain all data-related problems.

In this step, we need to identify the different data sources, as data can be collected from various sources such as files, database, internet, or mobile devices. It is one of the most important steps of the life cycle. The quantity and quality of the collected data will determine the efficiency of the output. The more will be the data, the more accurate will be the prediction.

This step includes the below tasks:

1. Identify various data sources
2. Collect data
3. Integrate the data obtained from different sources

**DATA PROCESSING:**

After collection the raw data for ML training, the most important task is data pre-processing. In broad sense, data pre processing will convert the selected data into a form we can work with or can feed to ML algorithms. We always need to pre process our data so that it can be as per the expectation of machine learning algorithm.

**DATA MANIPULATION:**

Data manipulation is **the process of changing data to make it easier to read or be easier to process**. Data manipulation is often used on machine learning before the start of model building as part of the data preprocessing also during the model building to transform the data into a more suitable form for processing.

**DATA VISUALIZATION:**

Data visualization provides an important suite of tools for identifying a qualitative understanding. This can be helpful when we try to explore the dataset and extract some information to know about a dataset and can help with identifying patterns, corrupt data, outliers, and much more.

If we have a little domain knowledge, then data visualizations can be used to express and identify key relationships in plots and charts that are more helpful to yourself and stakeholders than measures of association or significance.

**MODELLING:**

The aim of this step is to build a machine learning model to analyze the data using various analytical techniques and review the outcome. It starts with the determination of the type of the problems, where we select the machine learning techniques such as Classification, Regression, Cluster analysis, Association, etc. then build the model using prepared data, and evaluate the model.

**Machine learning models used : -**

**Decision Tree:**

Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.

**Random Forest:**

Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

**Support Vector Machine:**

Support Vector Regression is a supervised learning algorithm that is used to predict discrete values. Support Vector Regression uses the same principle as the SVMs. The basic idea behind SVR is to find the best fit line. In SVR, the best fit line is the hyperplane that has the maximum number of points.

**Logistic Regression:**

Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.

**AdaBoost Classifier:**

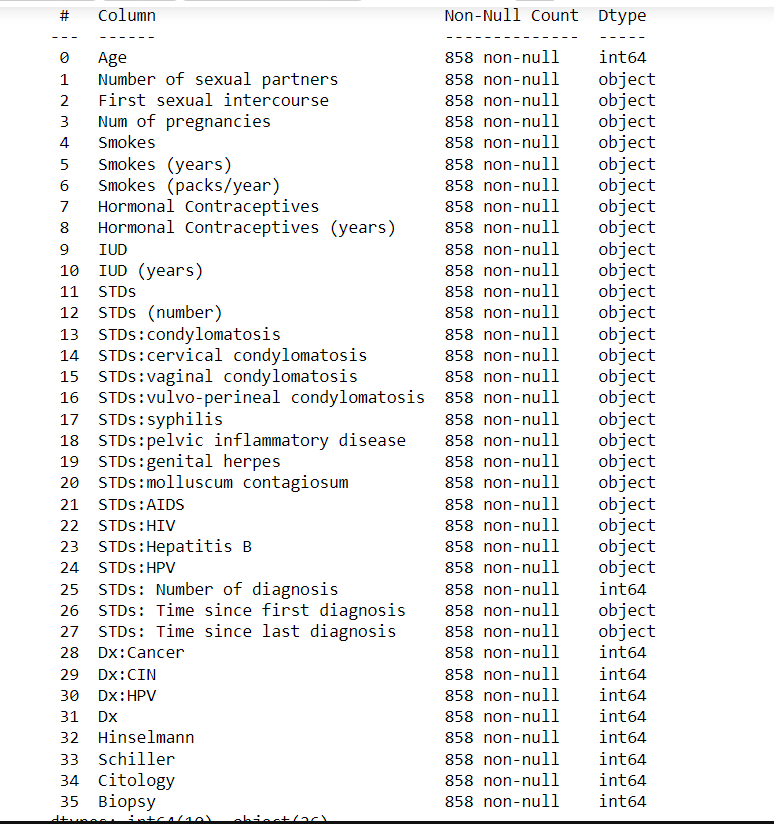
AdaBoost also called Adaptive Boosting is a technique in Machine Learning used as an Ensemble Method. The most common algorithm used with AdaBoost is decision trees with one level that means with Decision trees with only 1 split. These trees are also called Decision Stumps.

**PREDICTION OF OUTPUT:**

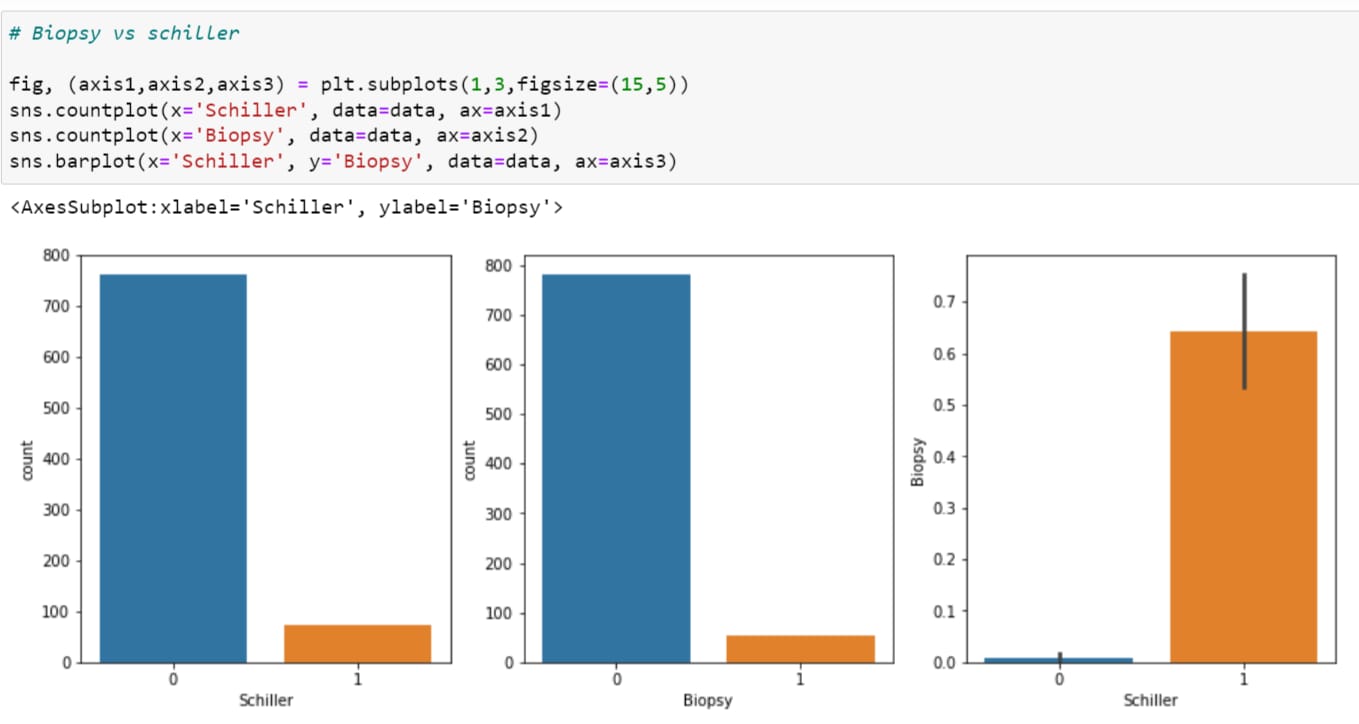
We eventually our model is ready to predict the early detection of cervical Cancer Using Machine Learning algorithms based on the given dataset.

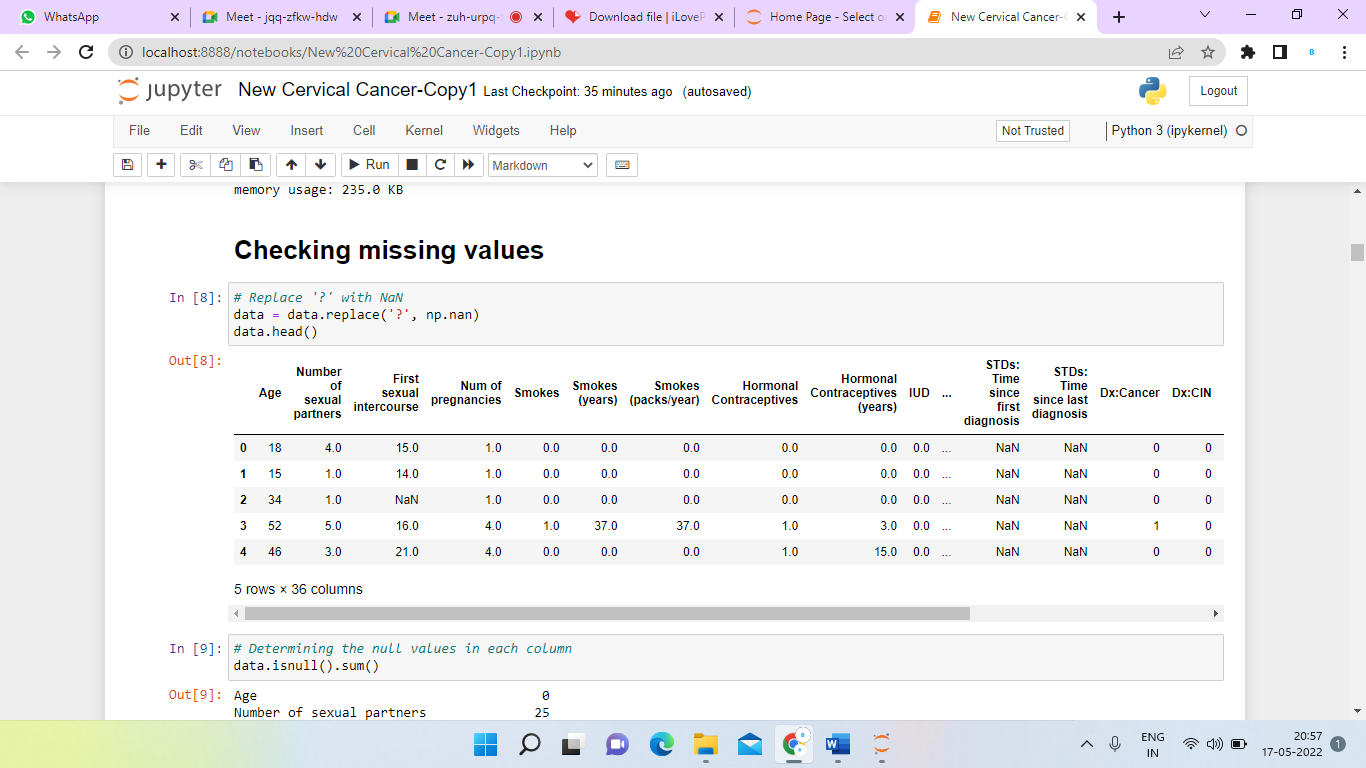
**DATASET**

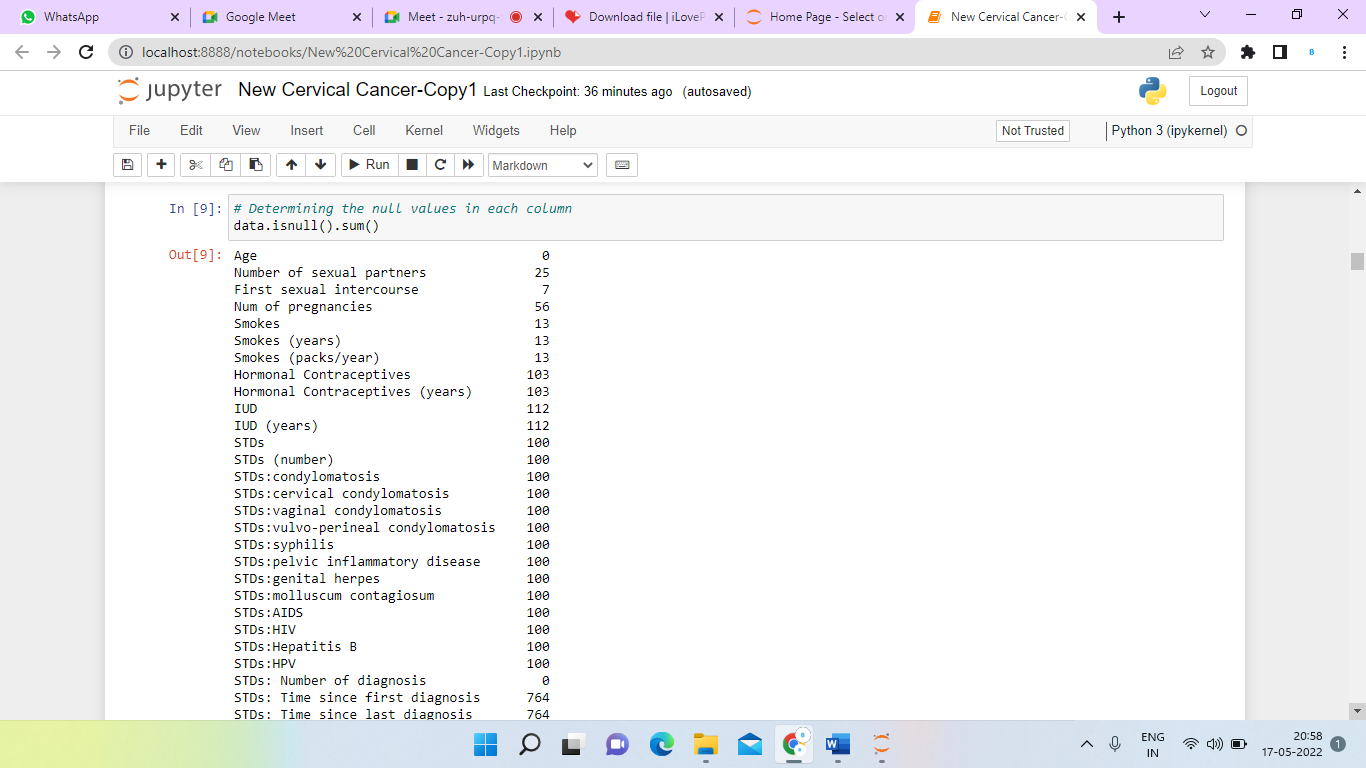
The dataset, “Cervical Cancer Risk Factors for Biopsy” was obtained from the UCI Repository. The dataset contains habits, demographic information, and medical history of 858 patients from the hospital. There are many missing values in this dataset, due to many patients not answering questions because of privacy concerns. The dataset consists of 858 instances, with 36 attributes. The dataset consists of 36 variables and records of 858 women patients. Of the 36 variables 4 variables are the target variables.

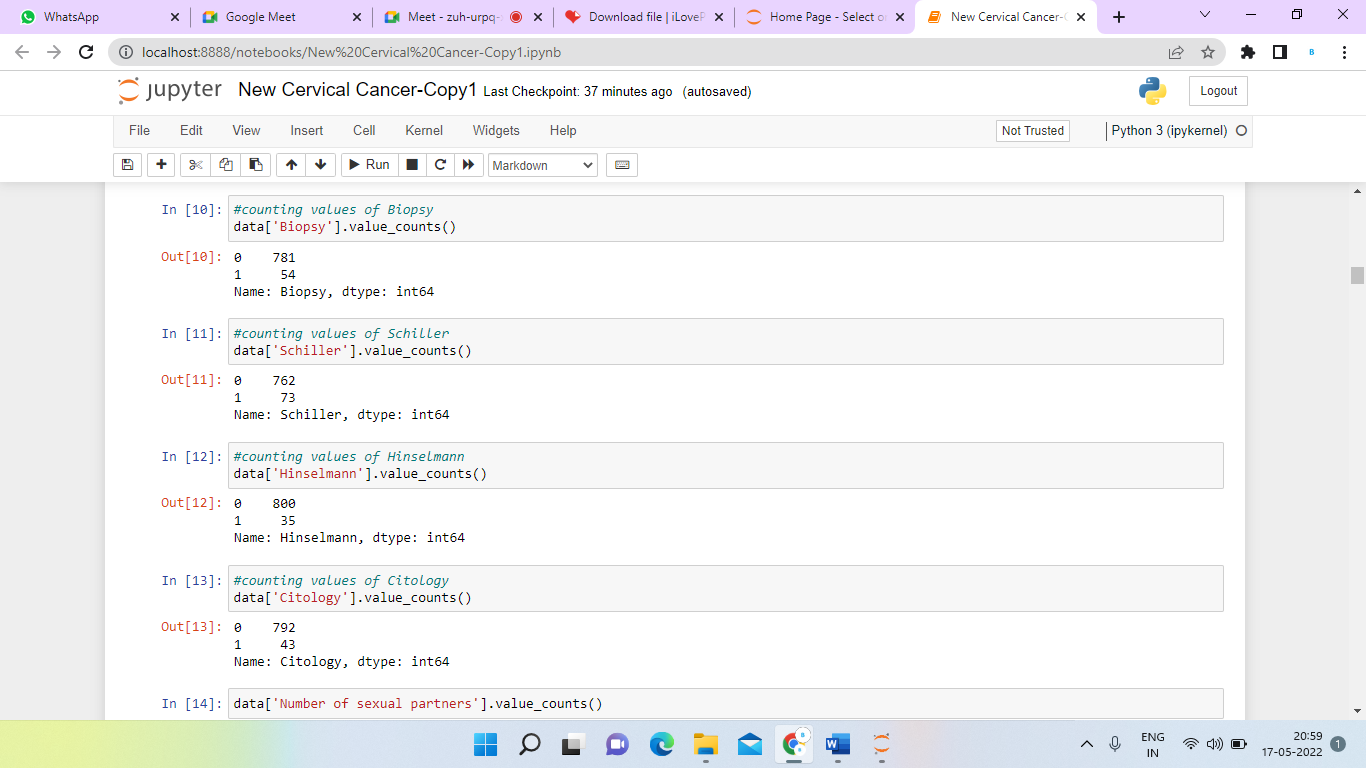


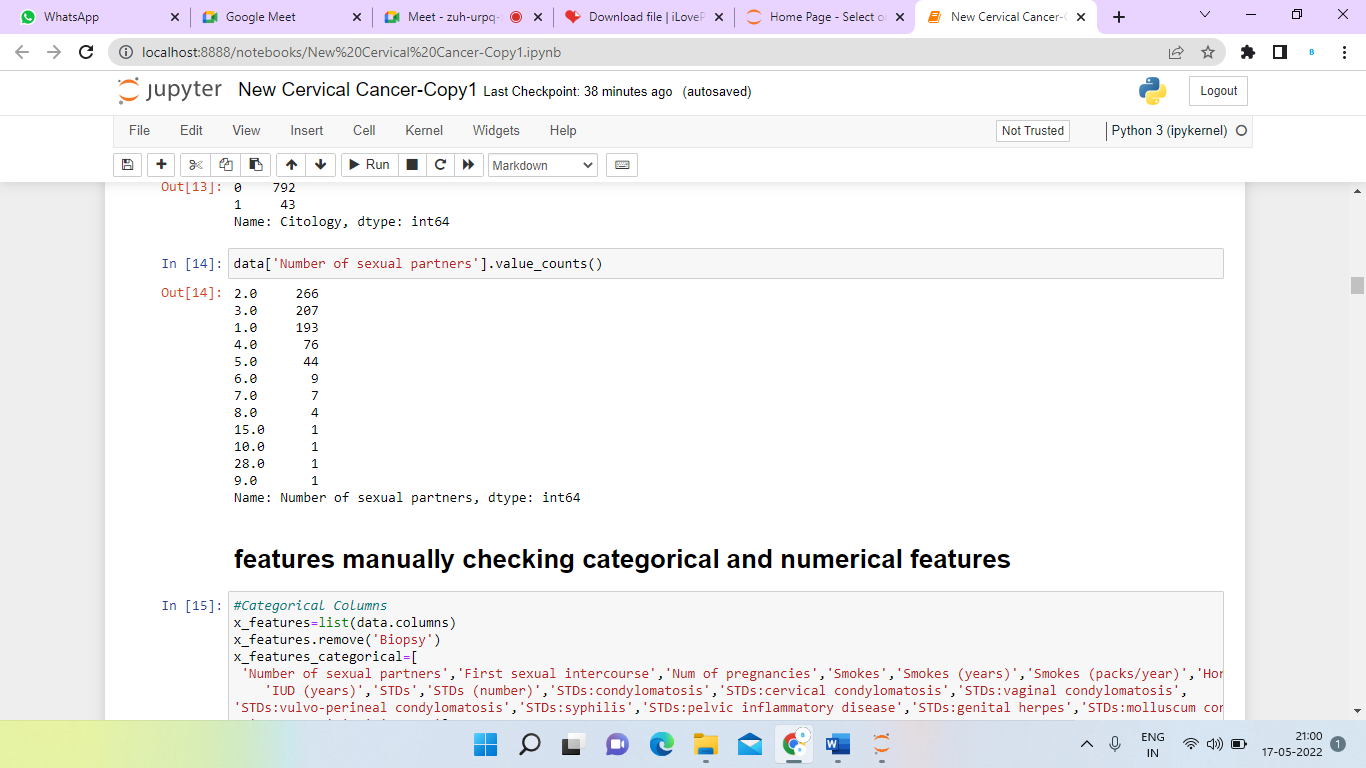
**DATA MANIPULATION**

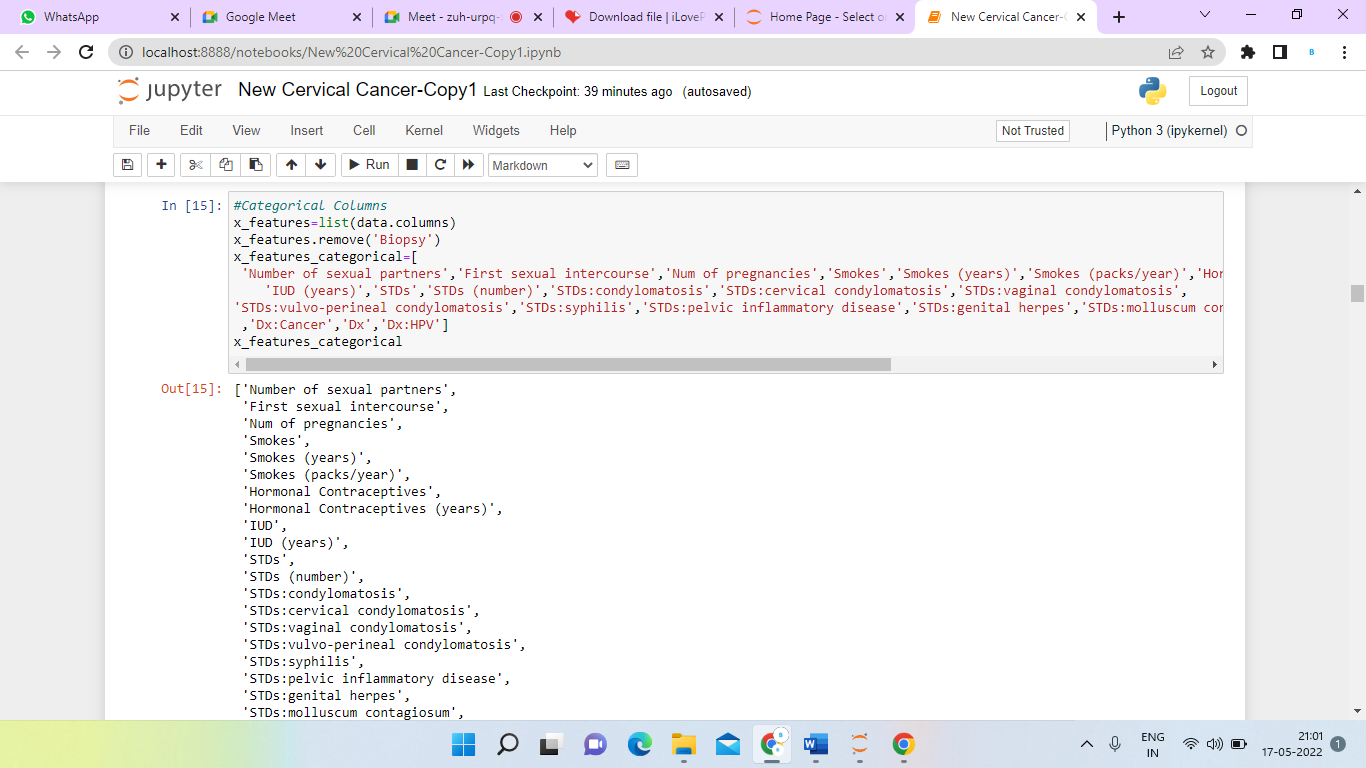
Data manipulation is basically known as EDA is a general approach to exploring datasets by means of simple summary statistics and graphic visualizations in order to gain a deeper understanding of the data. 

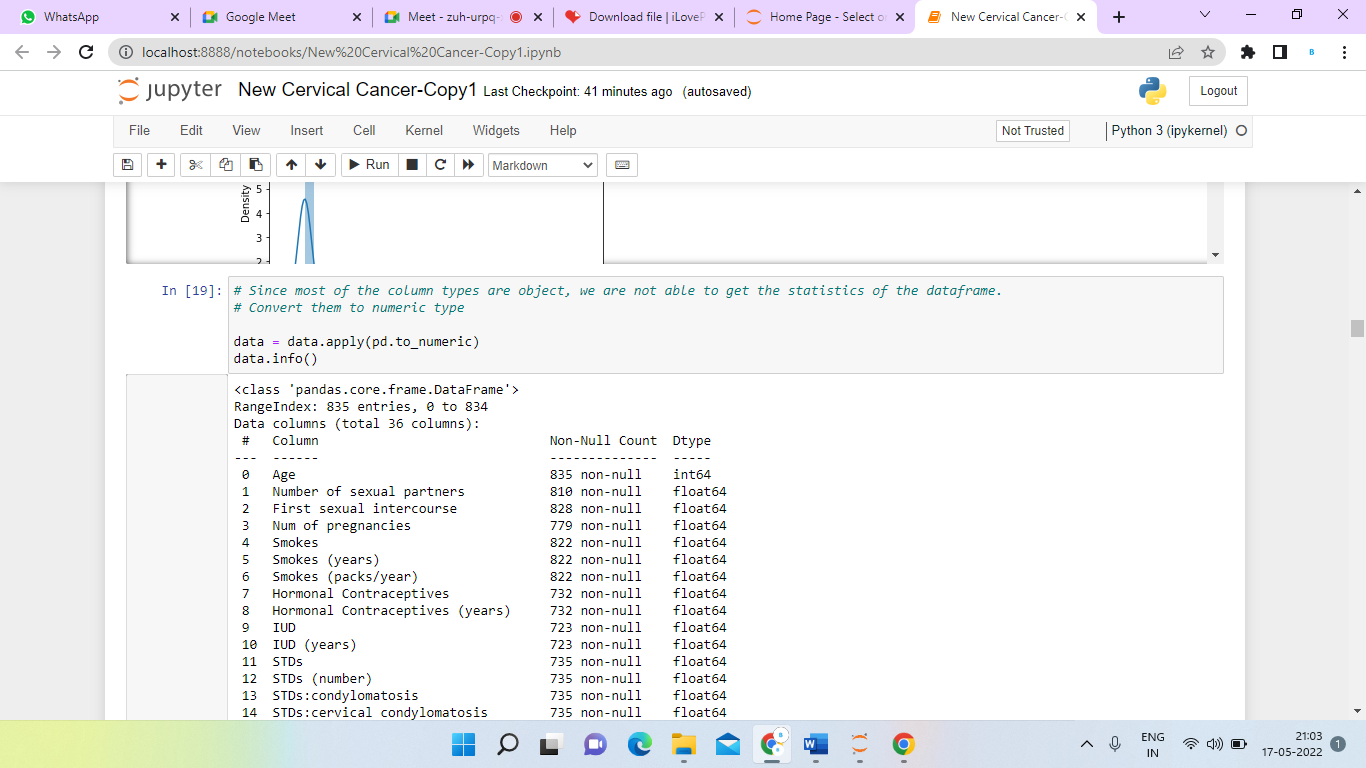






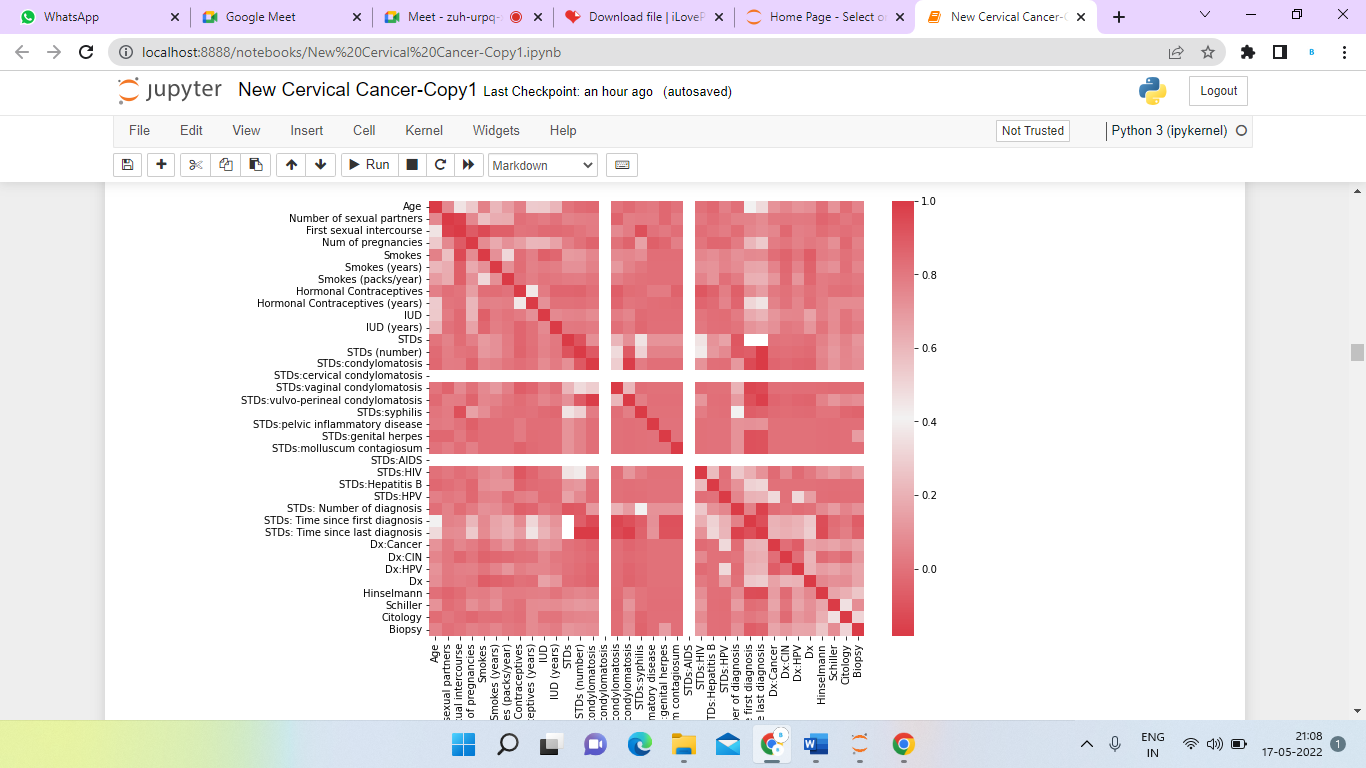


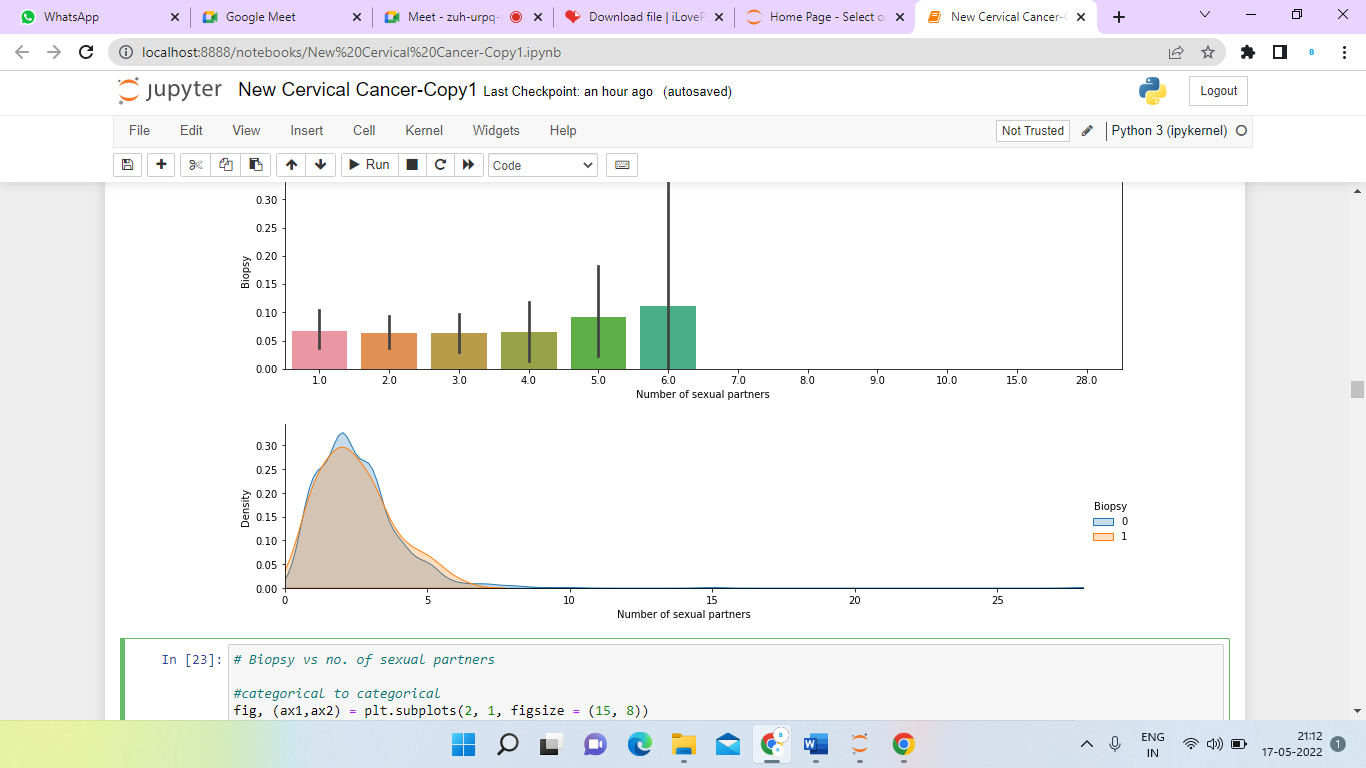


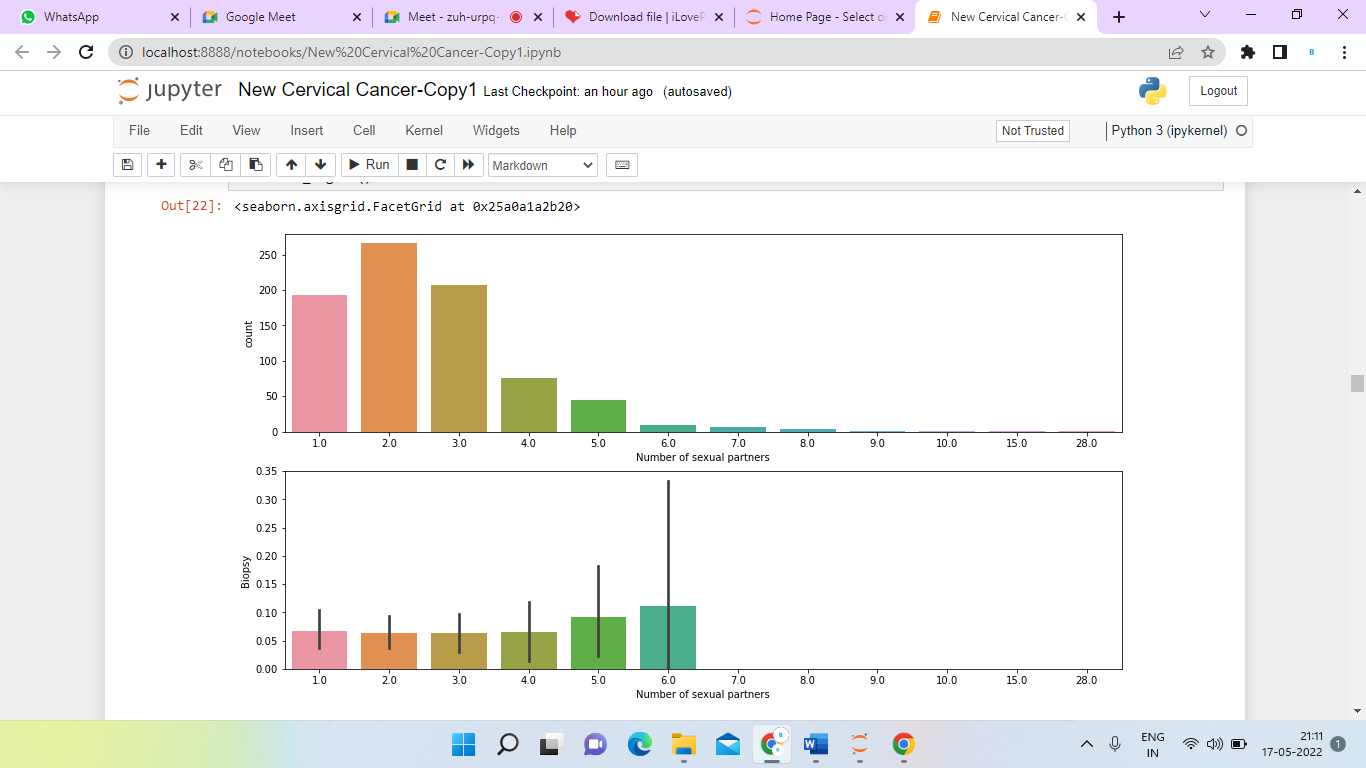


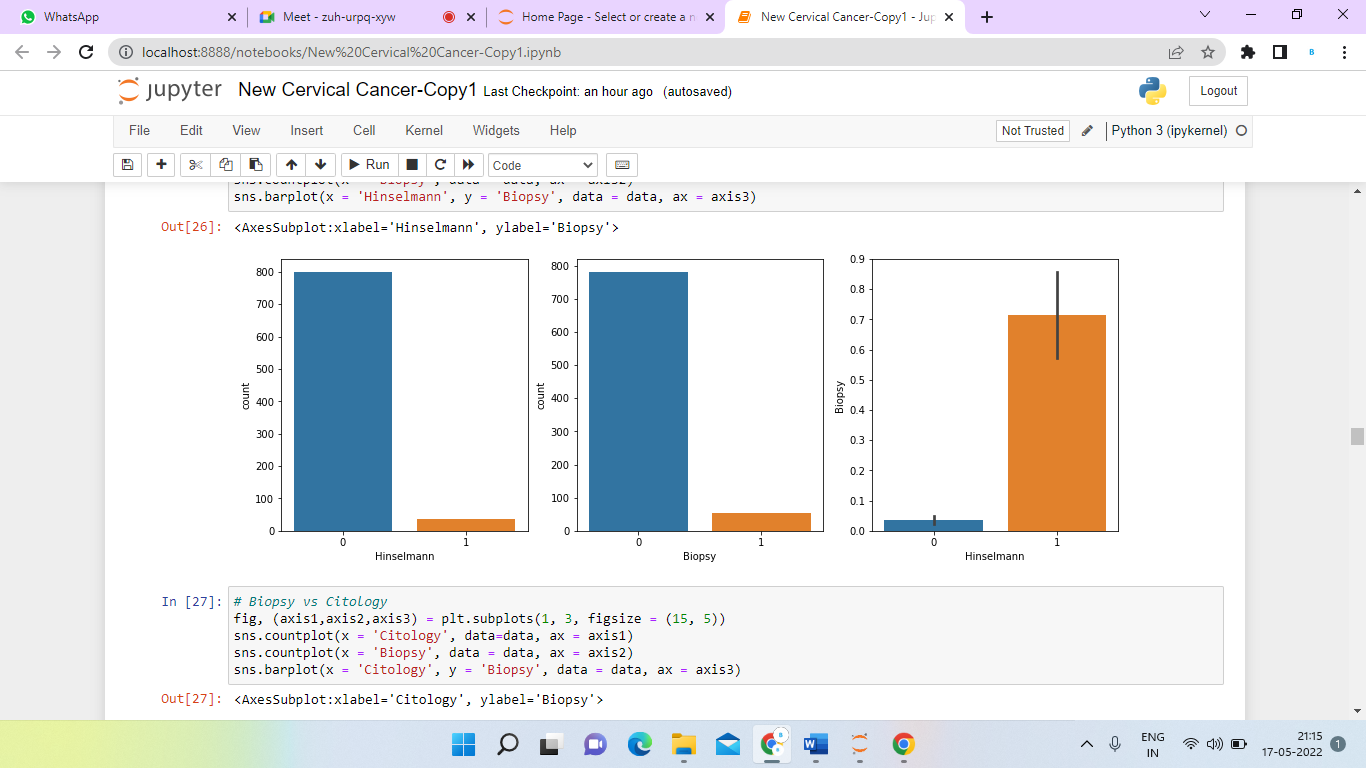
**DATA VISUALIZATION**

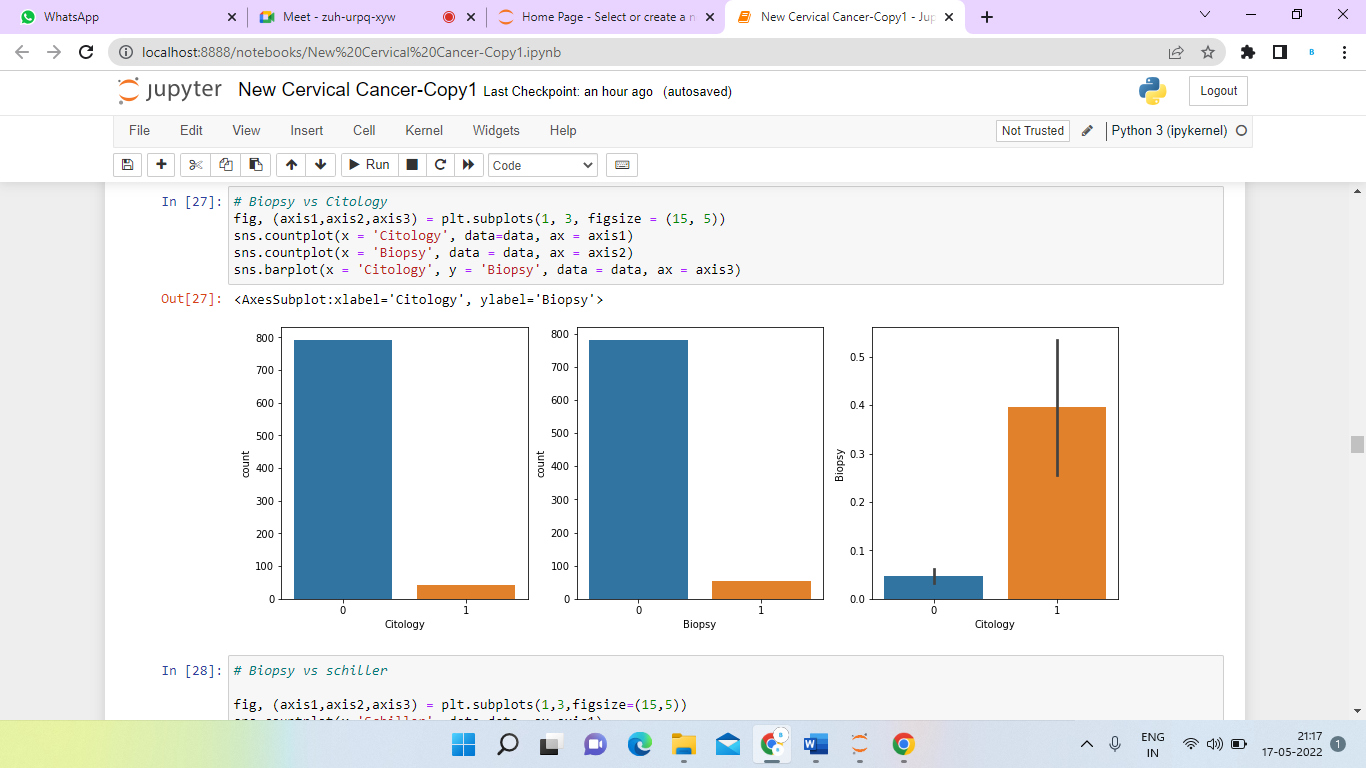
During feature engineering process, data is classified based on the measurement levels as Numerical data and Categorical data. Target variables are classified separately. Numerical data are represented as numbers. Features which cannot be grouped are classified under numerical data. Categorical data describes categories or groups and also answers to yes or no questions.

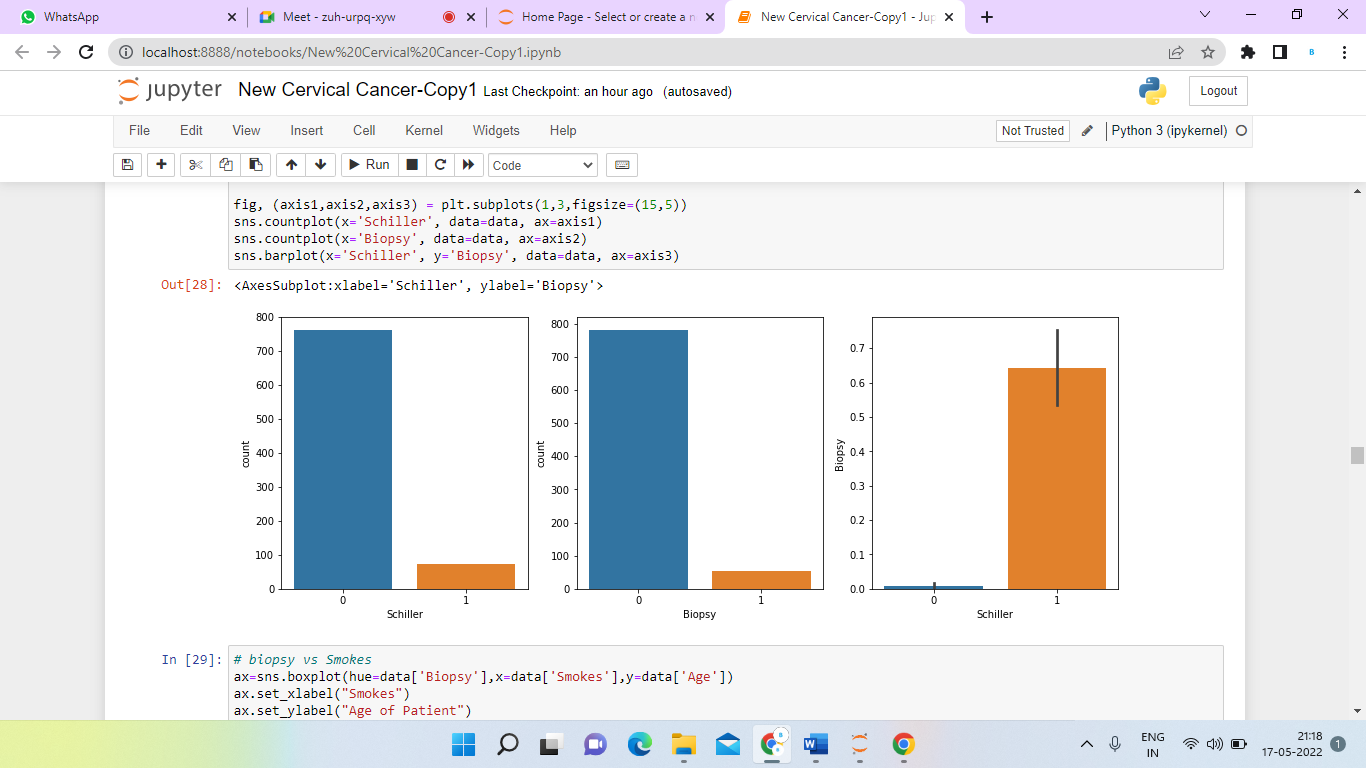






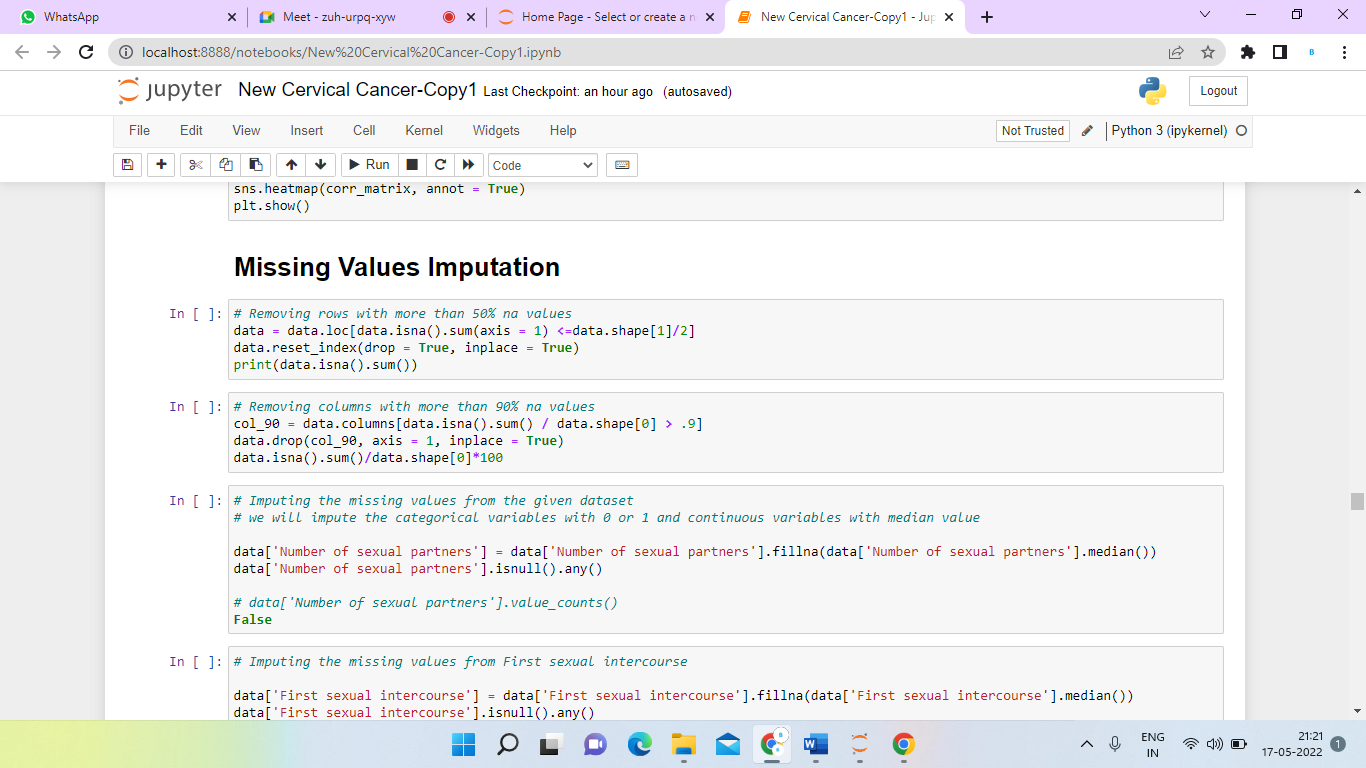






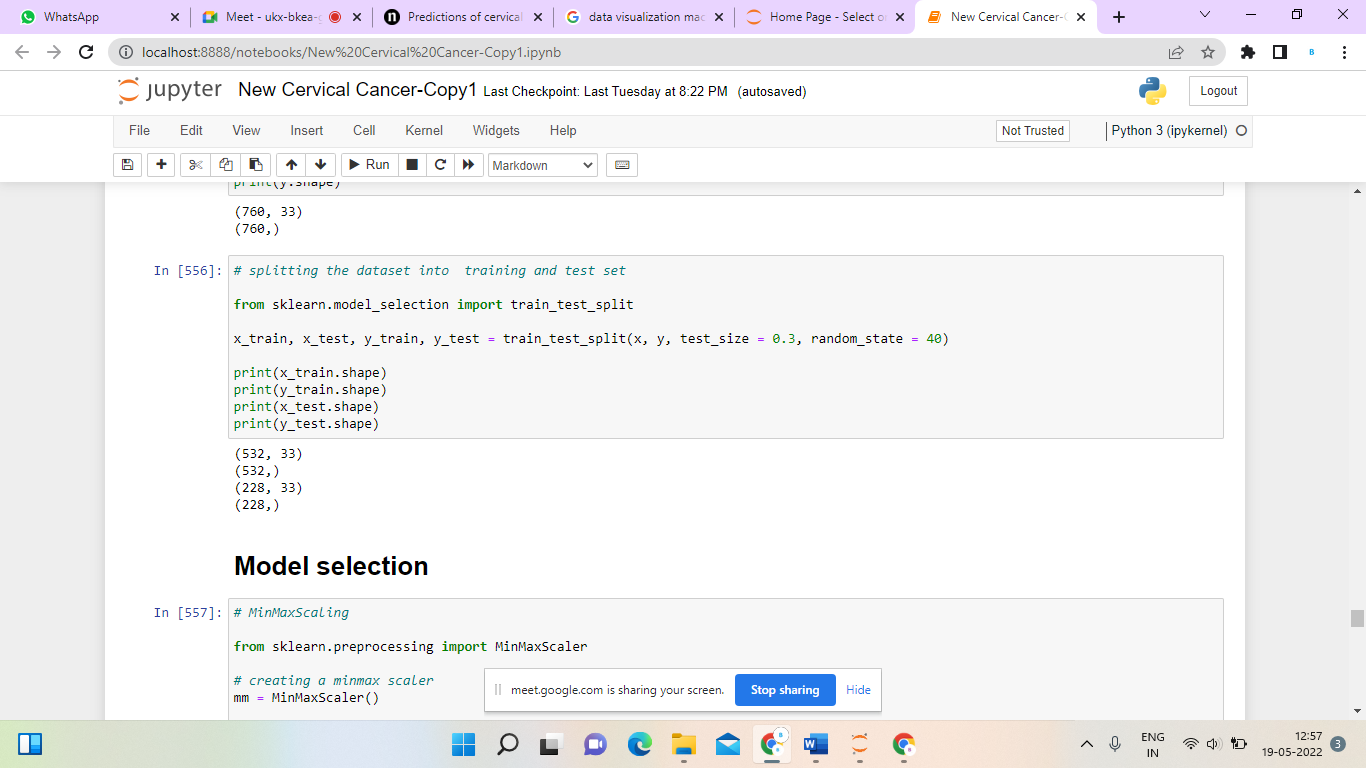
**DATA PREPROCESSING**

The project requires certain fields, which determine the key factors for causing cervical cancer. In this method, the null values are removed from the dataset and the unwanted attributes are dropped so that the model can train with better accuracy. To drop null values from dataset we will impute the categorical variables with 0 or 1 and continuous variables with median value.

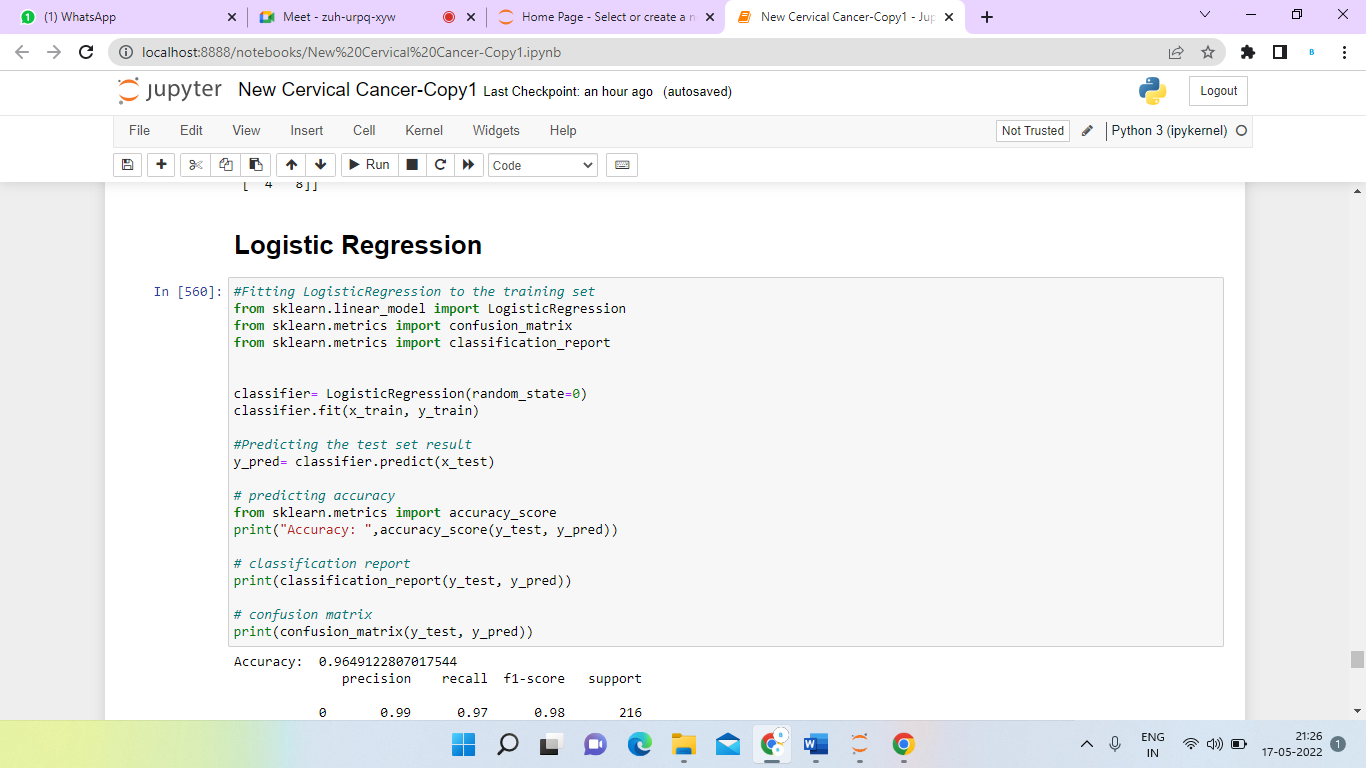


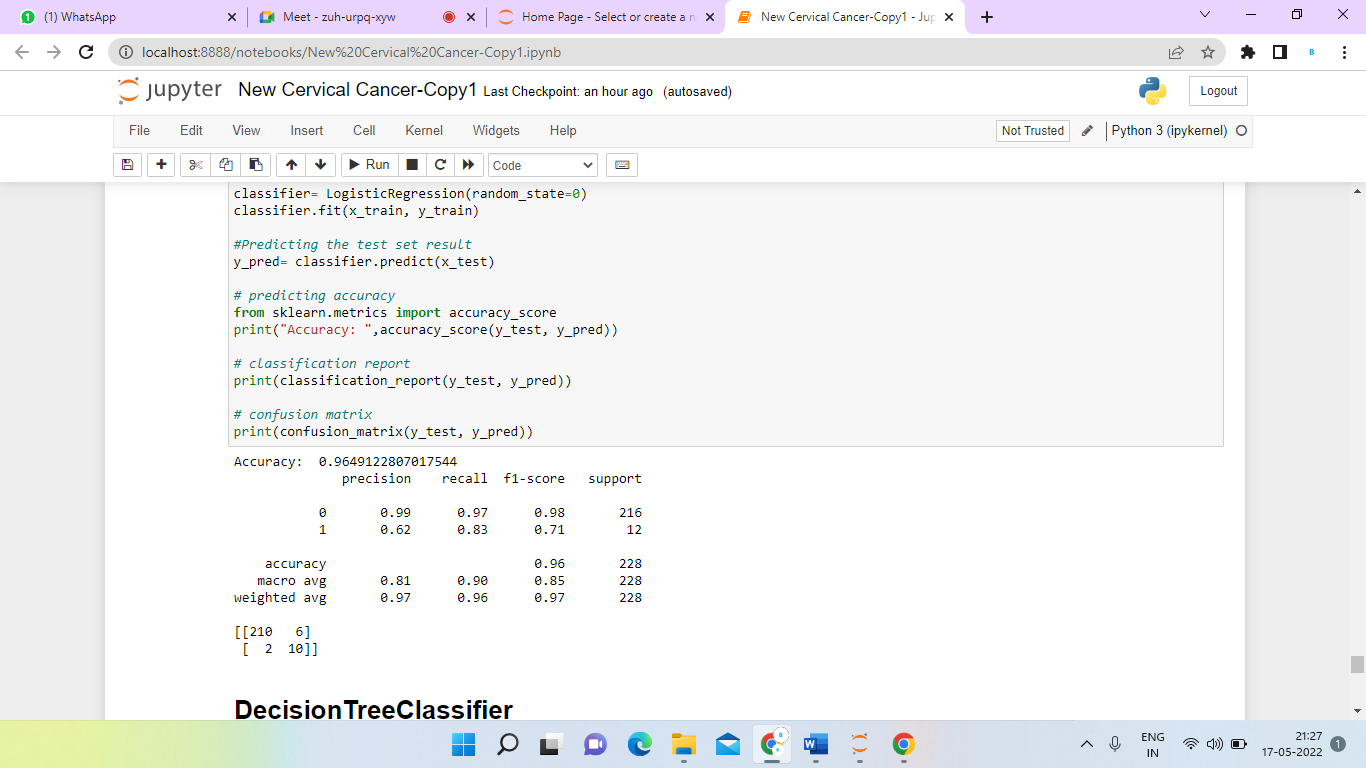
**DATA SPLITTING**

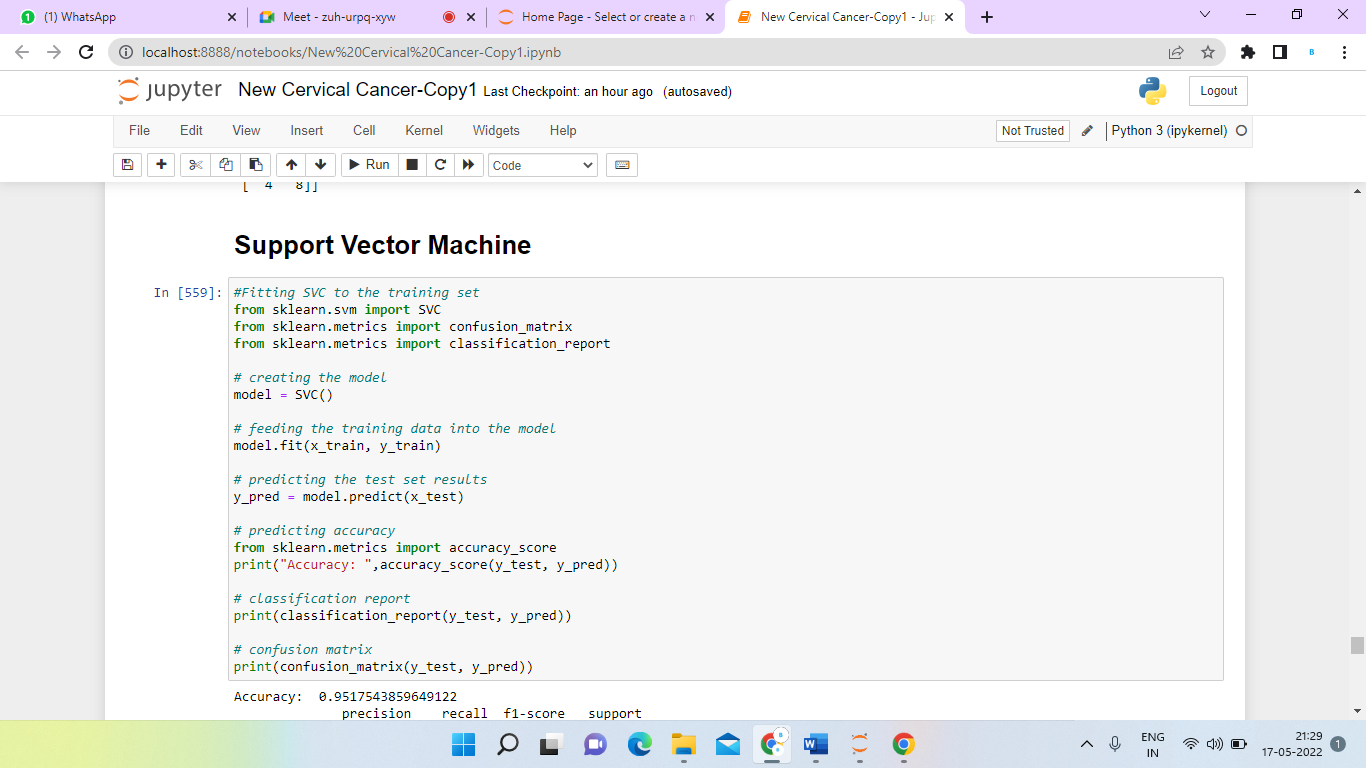
To determine the performance of machine learning algorithms, datasets are split into training and testing sets. During the dataset splitting process, the training set receives the majority of the data, while the testing set receives a smaller portion. The collected dataset is separated into two parts: 80 percent training data, and 20 percent testing data.

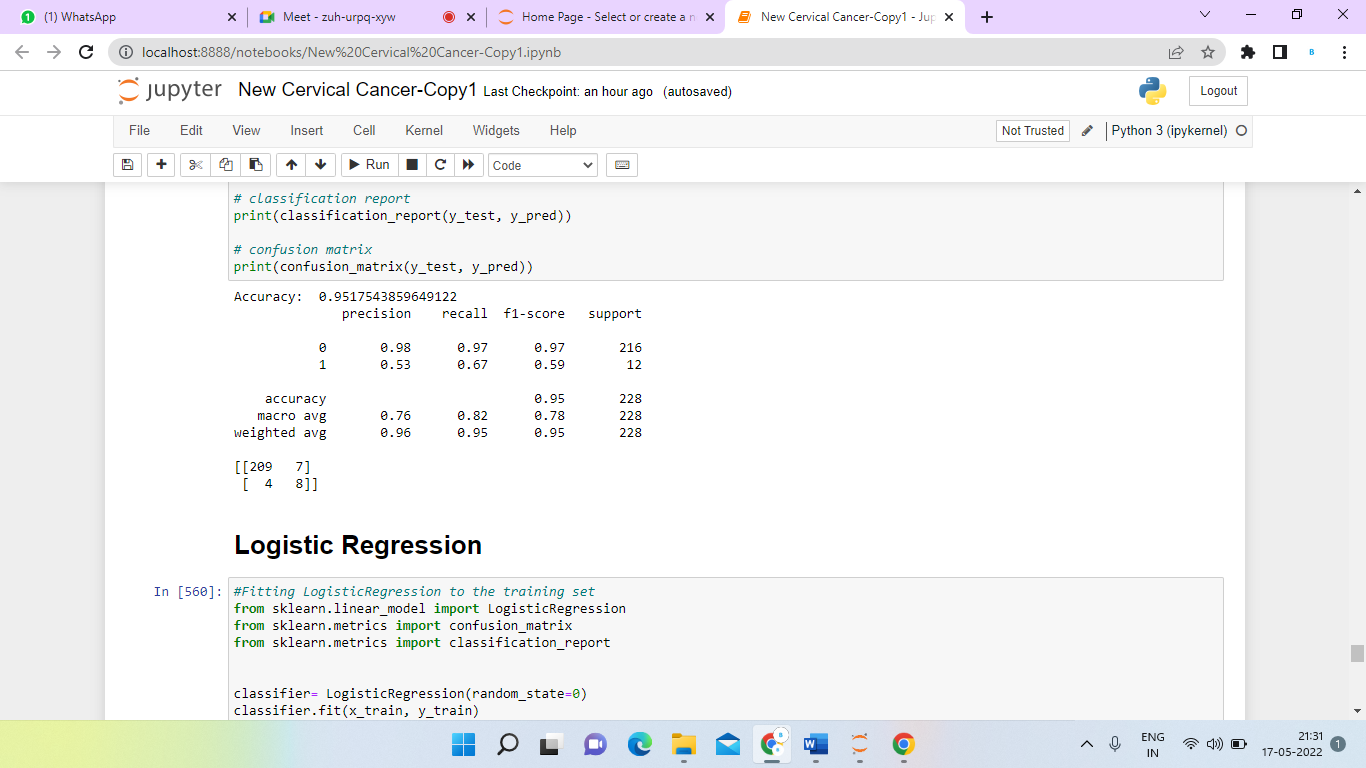


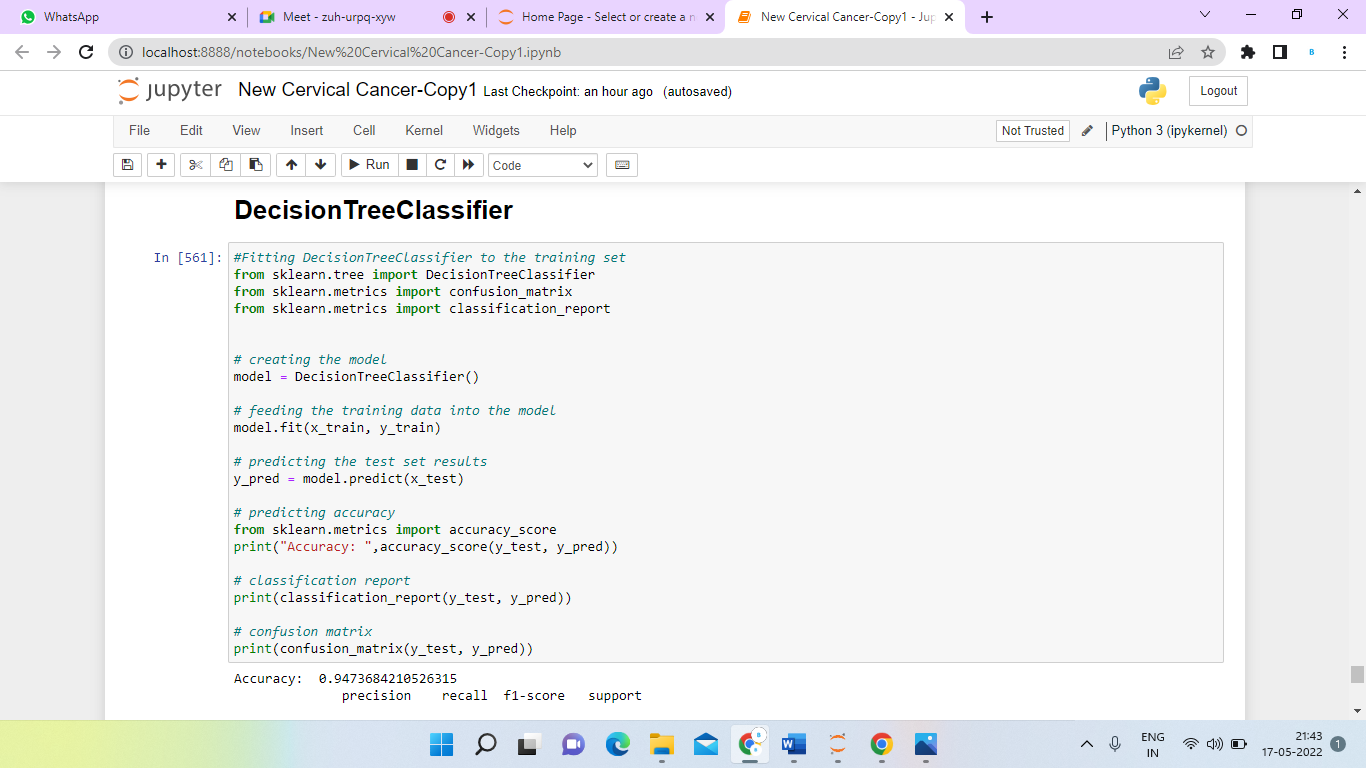
**EXPERIMENTAL RESULTS**

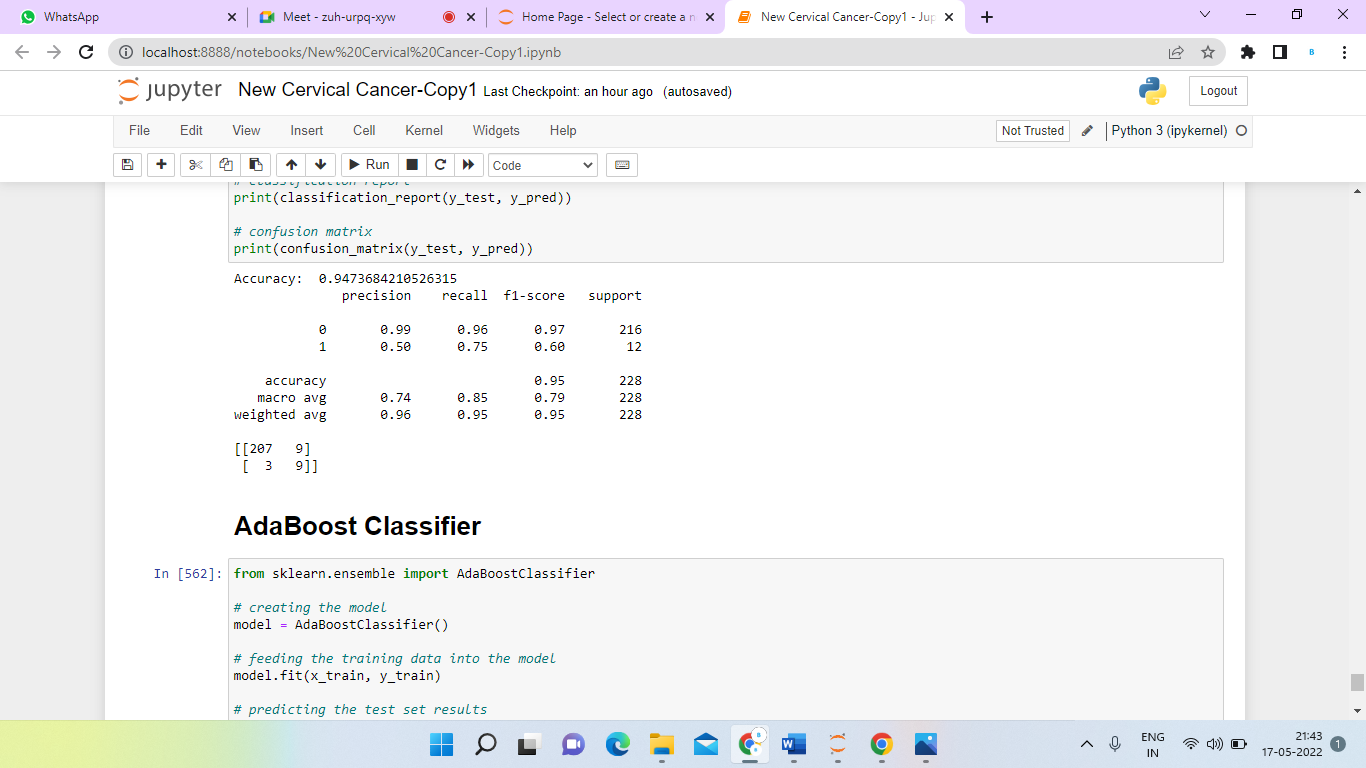


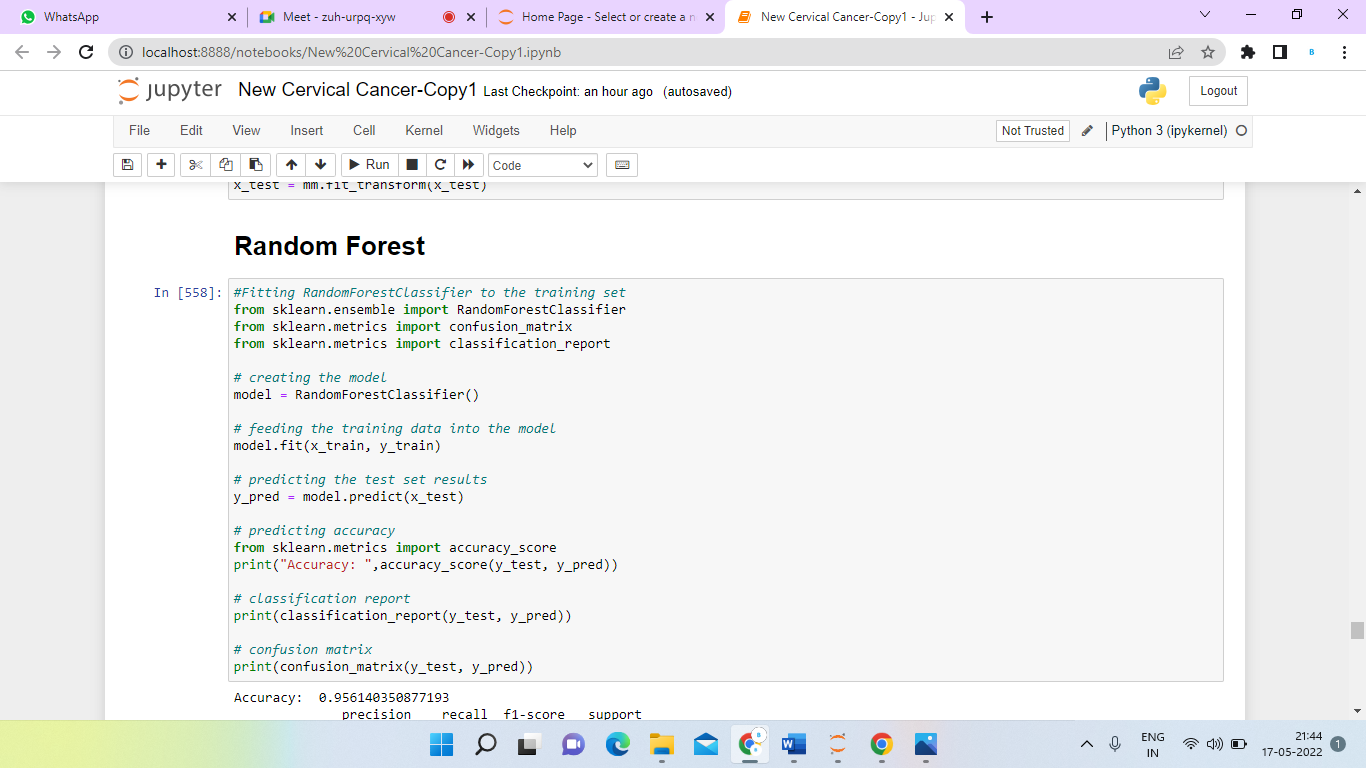




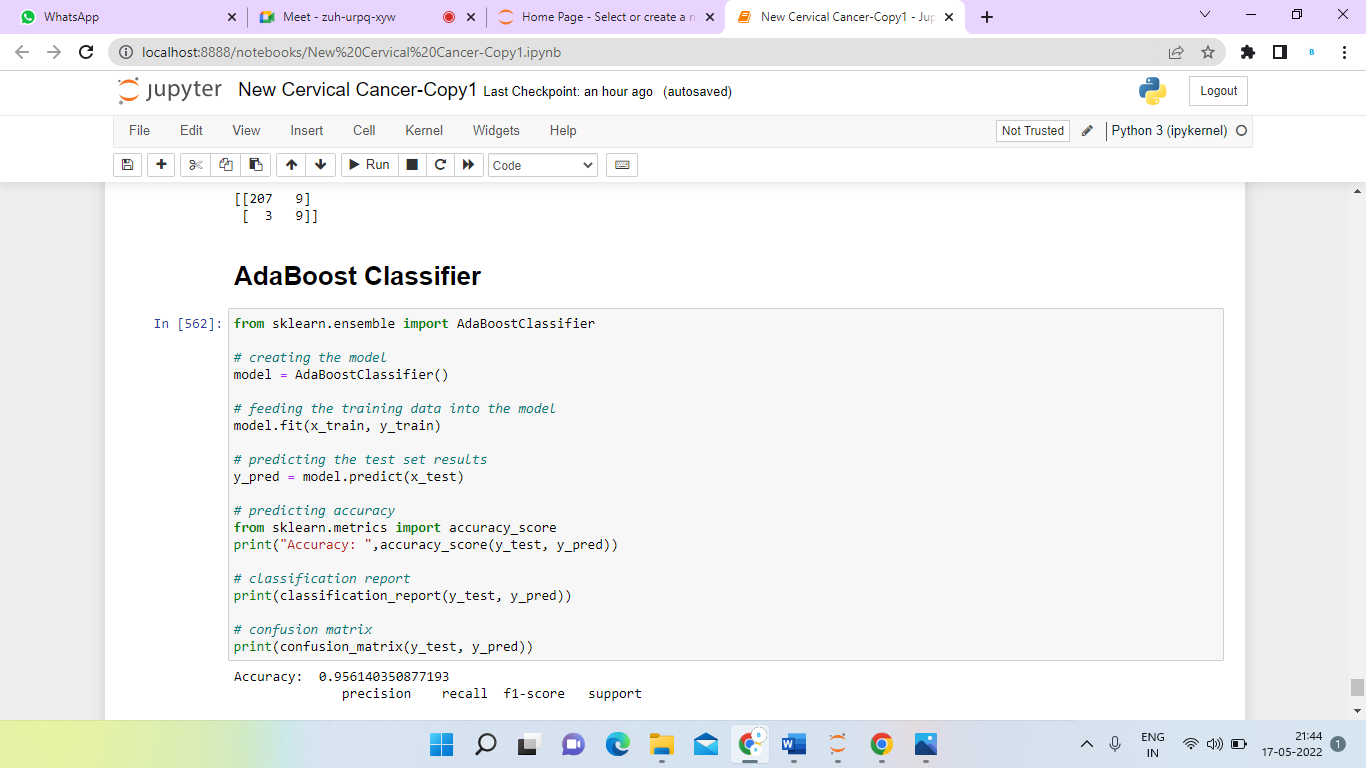


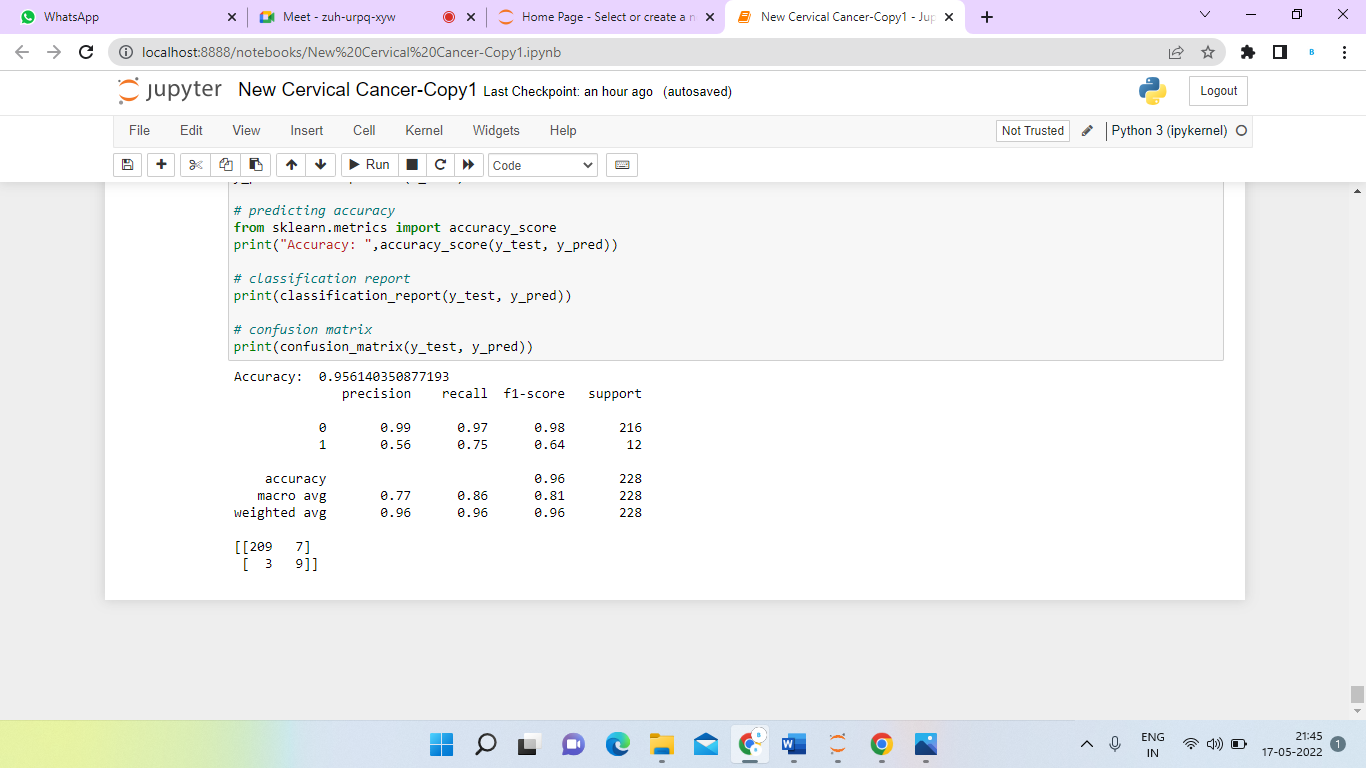












**CONCLUSION**

The aim of the research is to develop a model that can accurately diagnose and analyze cervical cancer data using classification algorithms such as Random Forest , Decision Tree, Logistic Regression, SVC, and AdaBoost. A combination of feature selection, recall, precision, accuracy and f1-score parameters are used to analyze those classifications performance. As a result, when compared to the other four classification algorithms, LOGISTIC REGRESSION has the highest rate of accuracy, precision, recall and f1-score, while Decision Tree has the lowest rate of accuracy, precision, recall and f1-score. Classification models claim highest accuracy on specific top features such as logistic regression was 96% on top 34 features.

**FUTURE STUDY**

Further research work can be carried out towards finding more detail about cervical cancer related parameters. In this dataset we have included only life style and medical history further toit, symptoms can be added. A generalize model can bedesigned to accurate the model by using multiple classification algorithms instead of only decision tree method. Datapre-processing steps could be used to improve training phase too. As discussed earlier, every cancer has unpredictablesymptoms and it is difficult to detect test requirement easily. Still our decision tree classifier with experimental thresholdsprovides acceptable accurate results. The more to the improvement, experimental thresholds can be replaced with dynamicthresholds. The comparison could be made of independent test prediction and combined tests predictions. More to theaccuracy improvement, multiple classification methods can be used and results can be compared for more accuracy. Thesame results can be accepted easily while the different results can be analysed further to conclude with a single result. Theanalysis can be done in more detail with other indirect parameters like Kappa, Sensitivity, Specificity etc.

**REFERENCES**

1. Kulasingam SL, Havrilesky L, Ghebre R, Myers ER. Screening for Cervical Cancer: A Decision Analysis for the US Preventive Services Task Force. AHRQ Publication No 11-05157-EF-1. Rockville, MD: Agency for Healthcare Research and Quality; 2011.
2. Abeloff MD, Armitage JO, Niederhuber JE, Kastan MB, McKenna WG, editors. Clinical Oncology. 3. Philadelphia, PA: Elsevier Churchill Livingstone; 2004.
3. International Agency for Research on Cancer. IARC Handbooks of Cancer Prevention: Cervix Cancer Screening. Vol. 10. Lyon, France: IARC Press; 2005.