

**CSE180 ADVANCE PYTHON PROGRAMMING**

NAME: RUTHVIKA MUCHALA

SIVANE SENTHIL

ANNE ASRITA

UNIQUE ID : E0121036

E0121027

E0121057

PROGRAM: BTECH AI ML

COURSE FACULTY: DR.ANU SINGHA

YEAR OF STUDY: 2021-2022

**BREAST CANCER DATASET**

**Rubrics table:**

|  |  |
| --- | --- |
| **MODULE** | **TOPICS** |
| Module 1 | **DATA PREPARATION, DATA CLEANING, DATA NORMALISATION, DATA STANDARDIZATION** |
| Module 2 | **NUMPY AND PANDA OPERATIONS** |
| Module 3 | **VISUALIZATION USING**  **1.BARPLOT**  **2.DISTPLOT**  **3.VIOLINPLOT**  **4.COUNTPLOT**  **5.SCATTERPLOT**  **6.STRIPPLOT** |
| Module 4 | **MACHINE LEARNING MODEL -**  **CLASSIFICATION** |
| Module 5 | **CONFUSION MATRIX, PREDICTION** |

INTRODUCTION:

Breast cancer (BC) is one of the most common cancers among women worldwide, representing the majority of new cancer cases and cancer-related deaths according to global statistics, making it a significant public health problem in today’s society.

The early diagnosis of BC can improve the prognosis and chance of survival significantly, as it can promote timely clinical treatment to patients. Further accurate classification of benign tumors can prevent patients undergoing unnecessary treatments. Thus, the correct diagnosis of BC and classification of patients into malignant or benign groups is the subject of much research. Because of its unique advantages in critical features detection from complex BC datasets, machine learning (ML) is widely recognized as the methodology of choice in BC pattern classification and forecast modelling.

Classification and data mining methods are an effective way to classify data. Especially in medical field, where those methods are widely used in diagnosis and analysis to make decisions.

OBJECTIVE:

The objective of this project is to observe which features are most helpful in predicting malignant or benign cancer and to see general trends that may aid us in model selection and hyper parameter selection. The goal is to classify whether the breast cancer is benign or malignant. To achieve this I have used machine learning classification methods to fit a function that can predict the discrete class of new input.

**DATA CLEANING:**

Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. When combining multiple data sources, there are many opportunities for data to be duplicated or copied. If data is incorrect, outcomes and algorithms are unreliable, even though they may look correct. There is no one absolute way to prescribe the exact steps in the data cleaning process because the processes will vary from dataset to dataset. Here, the data in the dataset was completely clean, no duplicate data or misspelled data and the data was thoroughly checked for data cleaning.

**DATA PREPARATION:**

**Data preparation** is the act of manipulating (or pre-processing) raw data (which may come from disparate data sources) into a form that can readily and accurately be analysed. Here the dataset is available in CSV format and was converted to data frame for further analysis.

**DATA NORMALISATION:**

Data normalization is the organization of data to appear similar across all records and fields.

**DATA STANDARDIZATION**: Data standardization is a data processing workflow that converts the structure of different datasets into one common format of data. It deals with the transformation of datasets after the data are collected from different sources and before it is loaded into target systems.

**DESCRIPTION OF THE DATA SET:**

The feature form this dataset are computed from a digitized image of a fine needle aspirate (FNA) of a breast tumour. The target feature records the prognosis (i.e., malignant or benign). The dataset contains 286 instances and 10 attributes in which 201 were no-recurrence-events and 85 were recurrence events. In the Breast Cancer dataset, the value of the attribute (node-caps) status was missing in 8 records.

**DESCRIPTION ABOUT DATA SET ACCESS AND INFORMATION**

**The data set accessing methods used in this project are:**

1. .head()
2. .isnull().sum()
3. .isna().sum()

**DATA VISUALIZATION**

Data visualization is the representation of data through use of common graphics, such as charts, plots, infographics, and even animations. These visual displays of information communicate complex data relationships and data-driven insights in a way that is easy to understand.

**Libraries for Python Data Visualizations that are used in this project are:**

* Matplotlib
* Seaborn
* Pandas
* Plotly

**DATA VISUALIZATION USED IN THIS PROJECT:**

**1.DISTPLOT:** It is used basically for univariant set of observations and visualizes it through a histogram.

**2.COUNT PLOT:** To check whether the data is imbalanced or not.

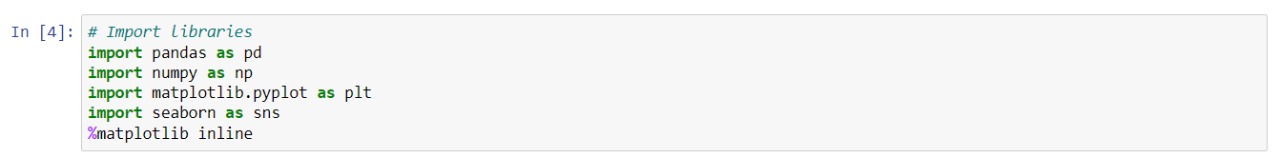
**3.SCATTERPLOT:** A scatter plot is a type of plot or mathematical diagram using Cartesian coordinates to display values for typically two variables for a set of data. If the points are coded, one additional variable can be displayed.

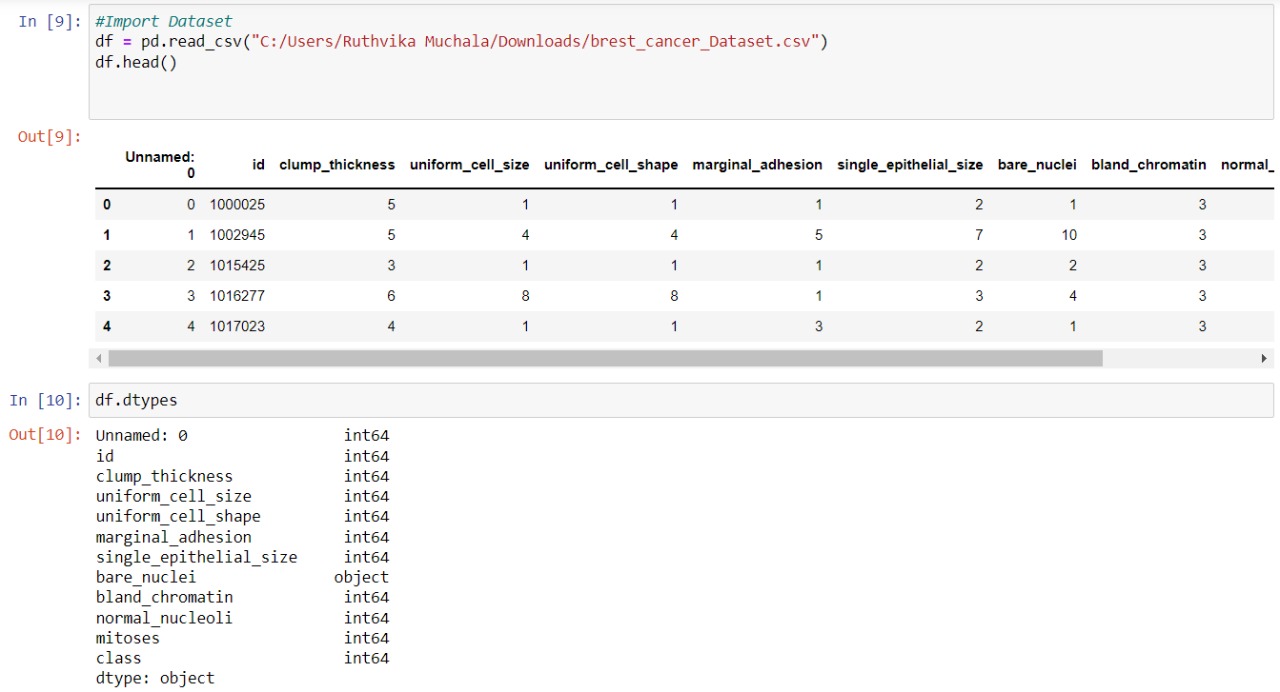
**4.BARPLOT:** A bar plot or bar chart is a graph that represents the category of data with rectangular bars with lengths and heights that is proportional to the values which they represent. The bar plots can be plotted horizontally or vertically. A bar chart describes the comparisons between the discrete categories. One of the axis of the plot represents the specific categories being compared, while the other axis represents the measured values corresponding to those categories.

**5.STRIPLOT:** A strip plot is drawn on its own. It is a good complement to a boxplot or violin plot in cases where all observations are shown along with some representation of the underlying distribution. It is used to draw a scatter plot based on the category.

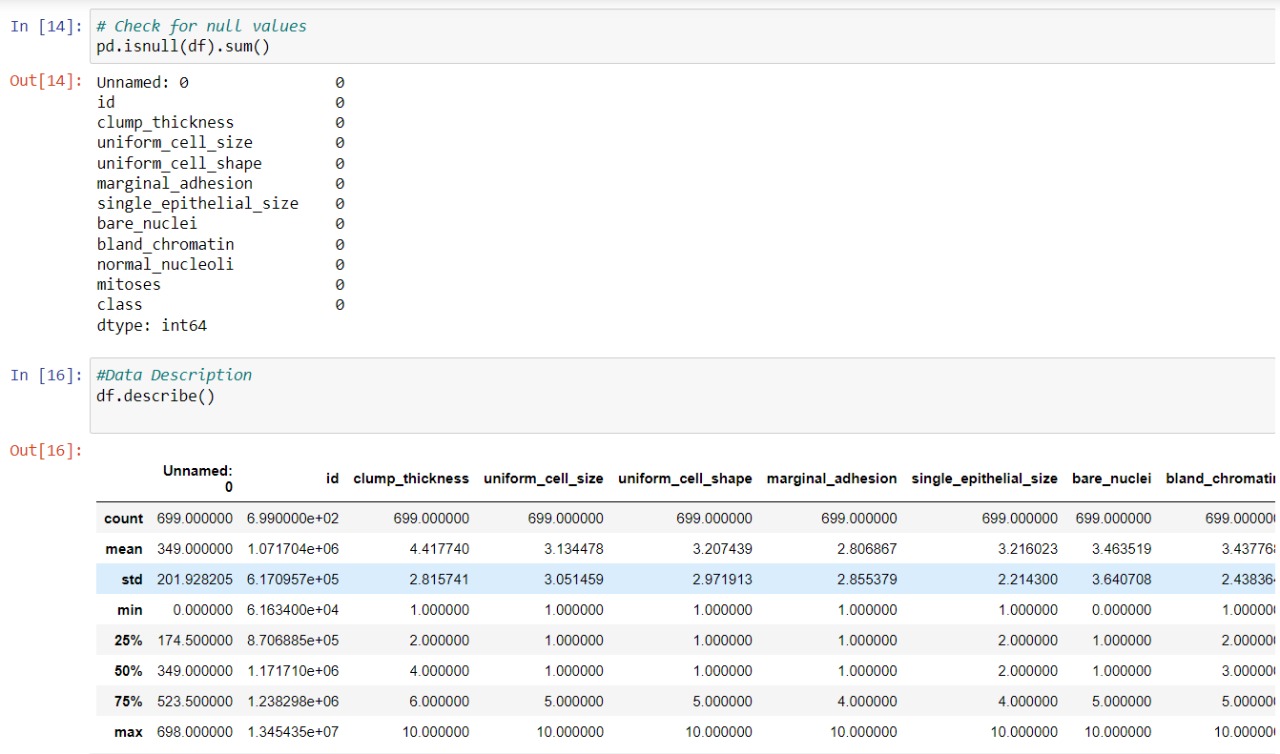
**6.VIOLINPLOT**: A violin plot is a method of plotting numeric data. It is similar to a box plot, with the addition of a rotated kernel density plot on each side.

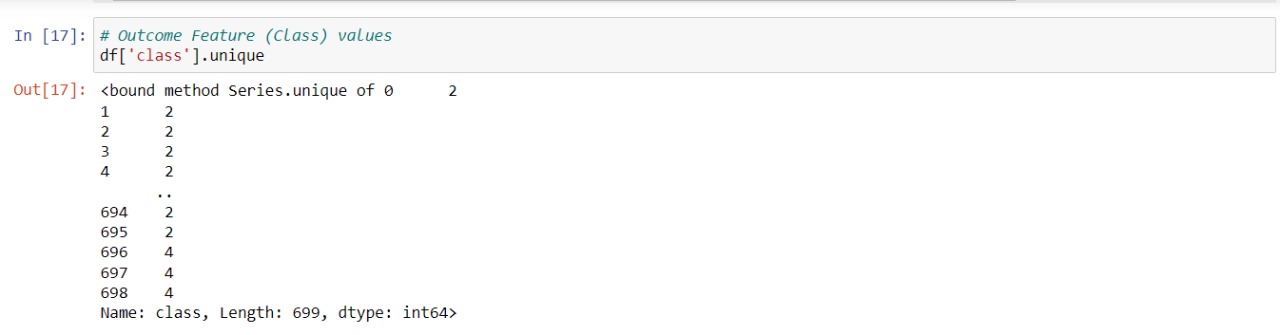
**CODE:**

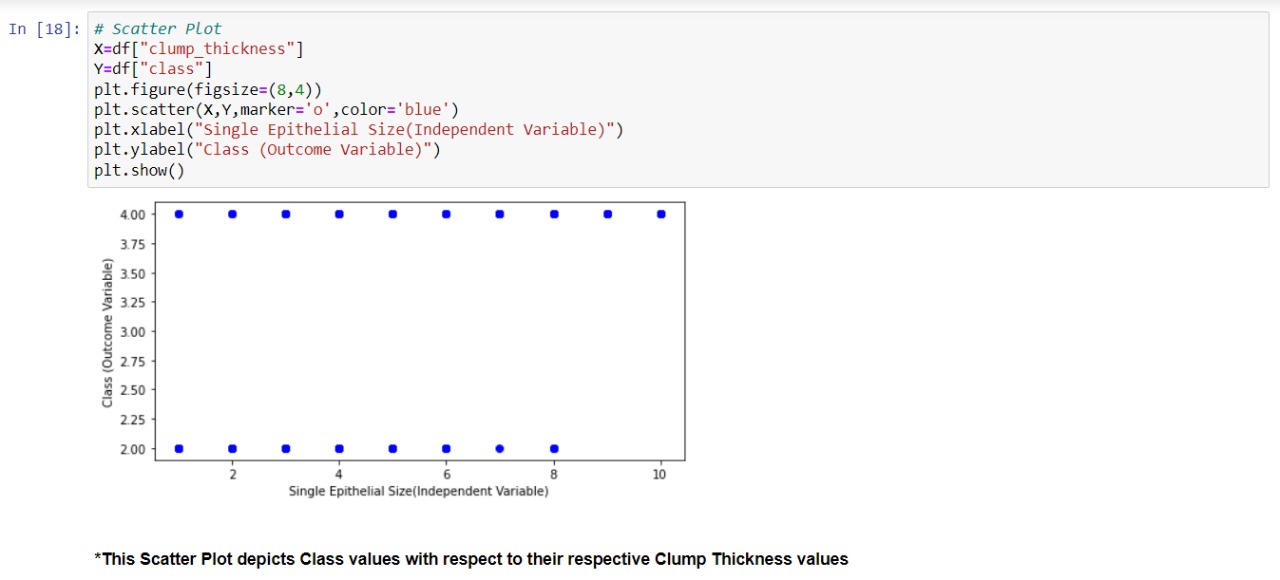


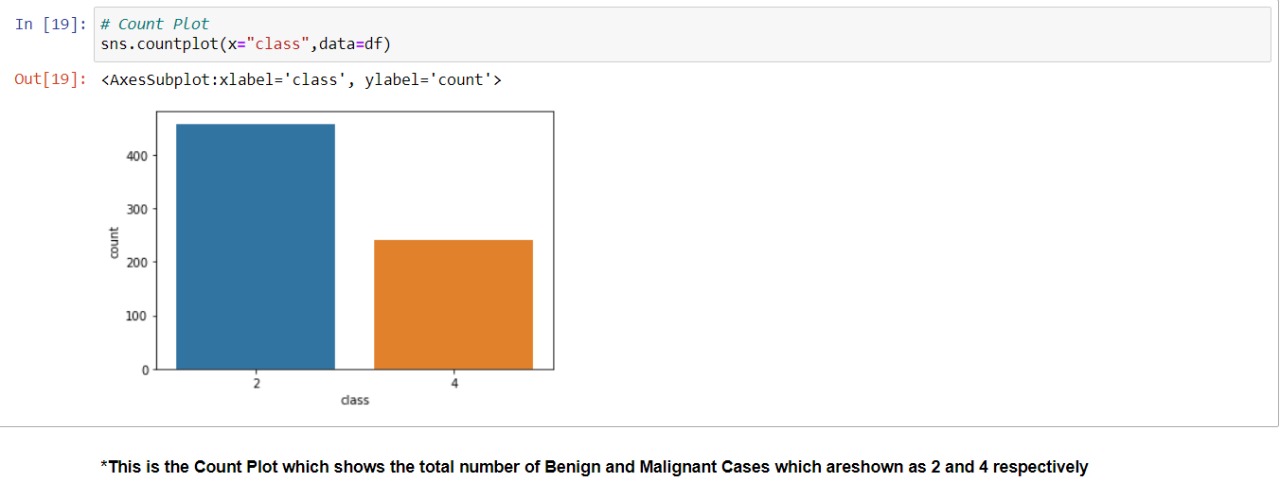


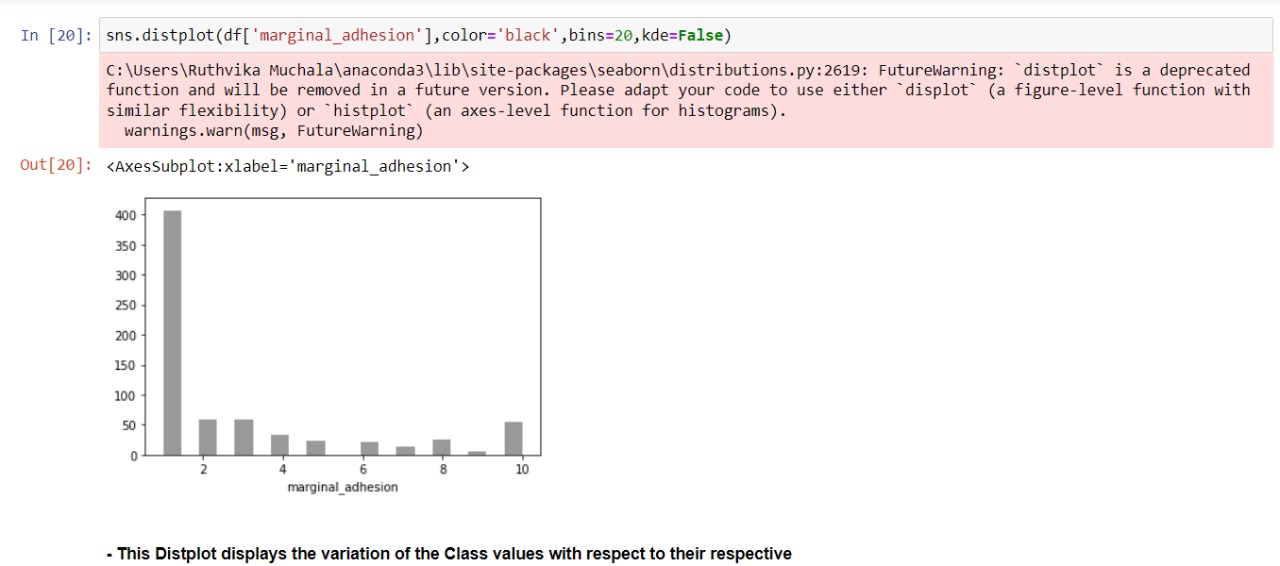


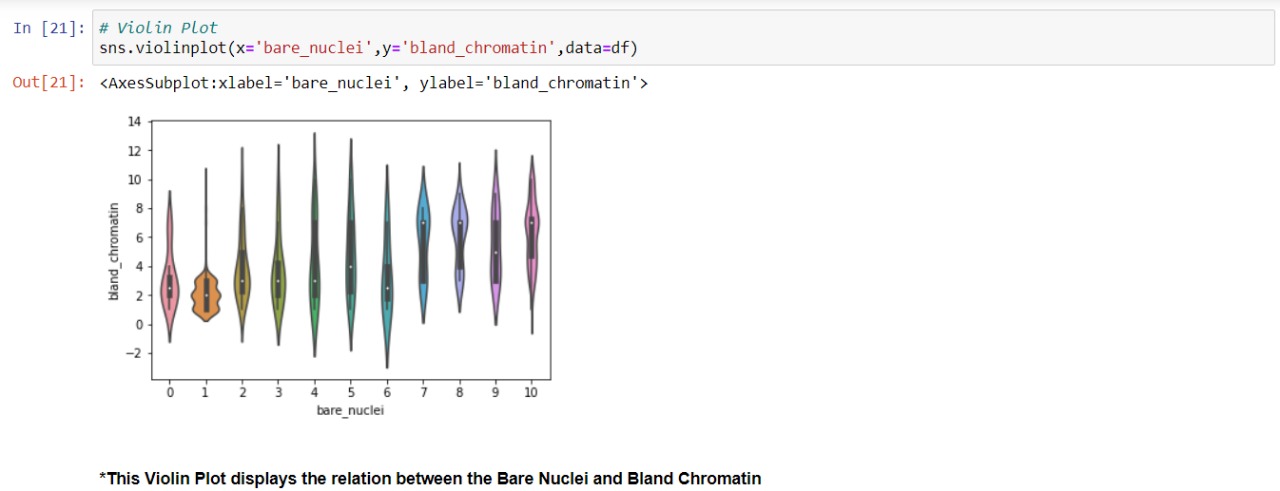


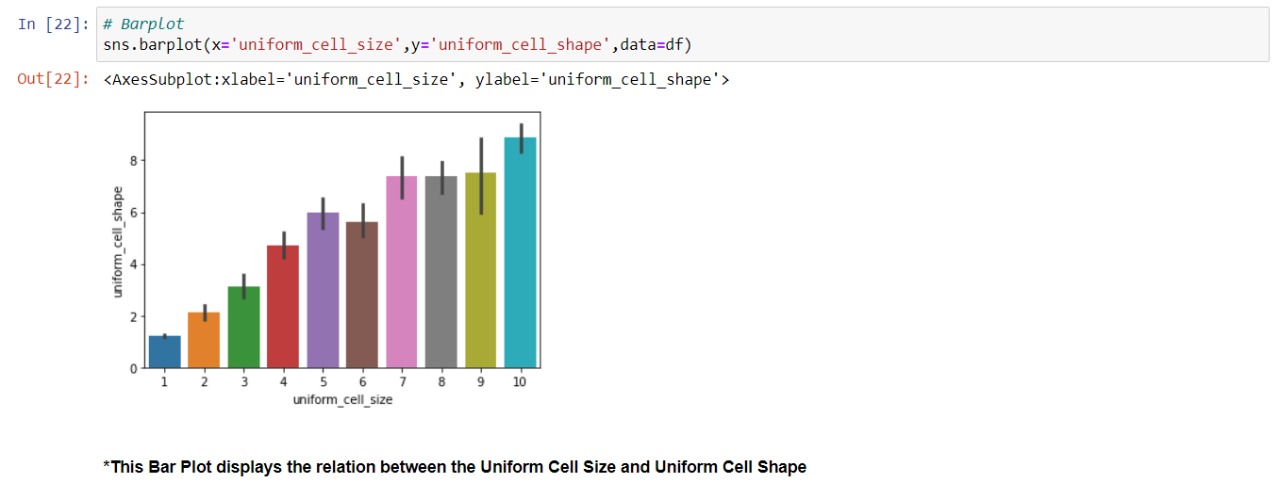


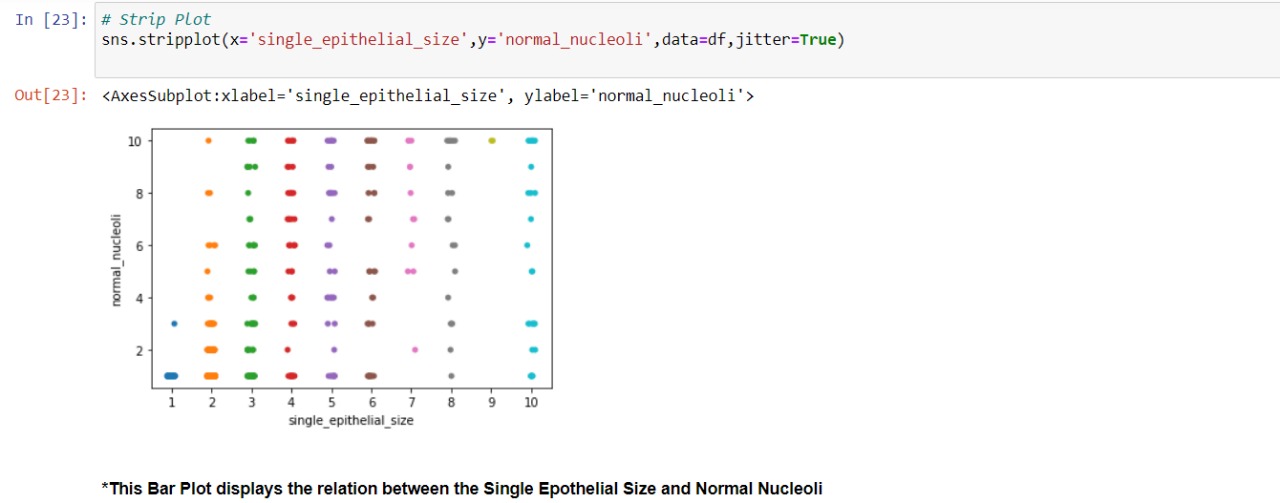


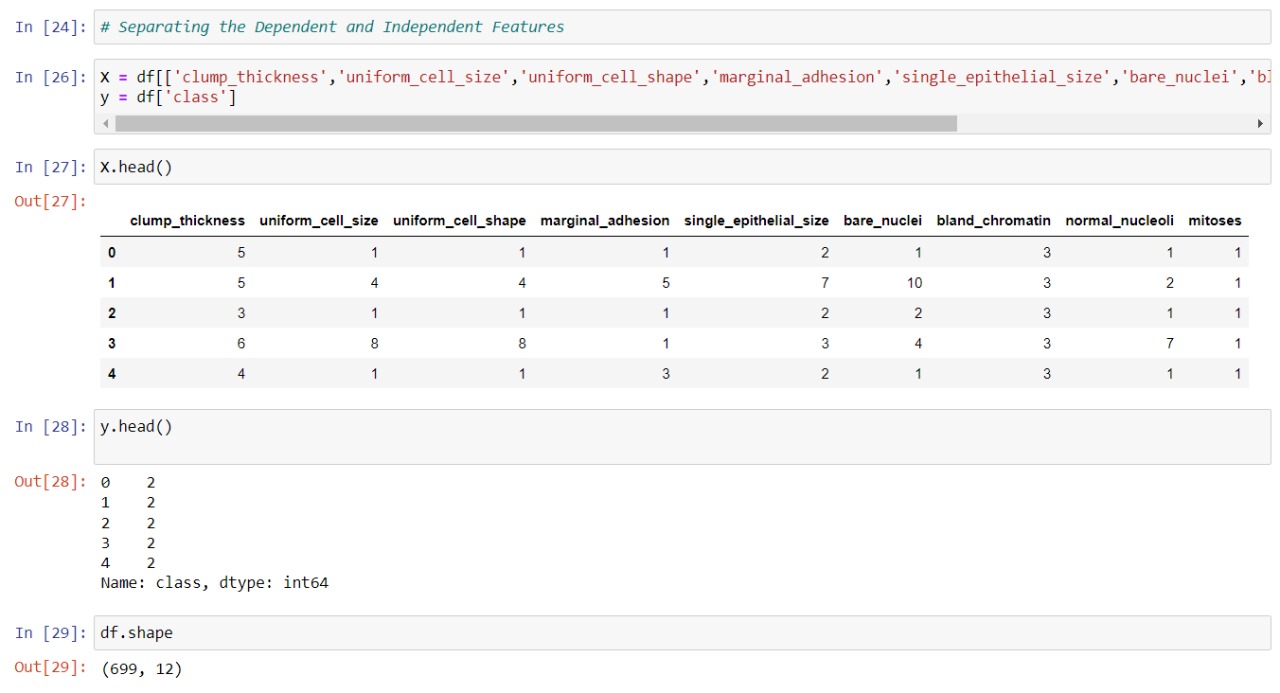




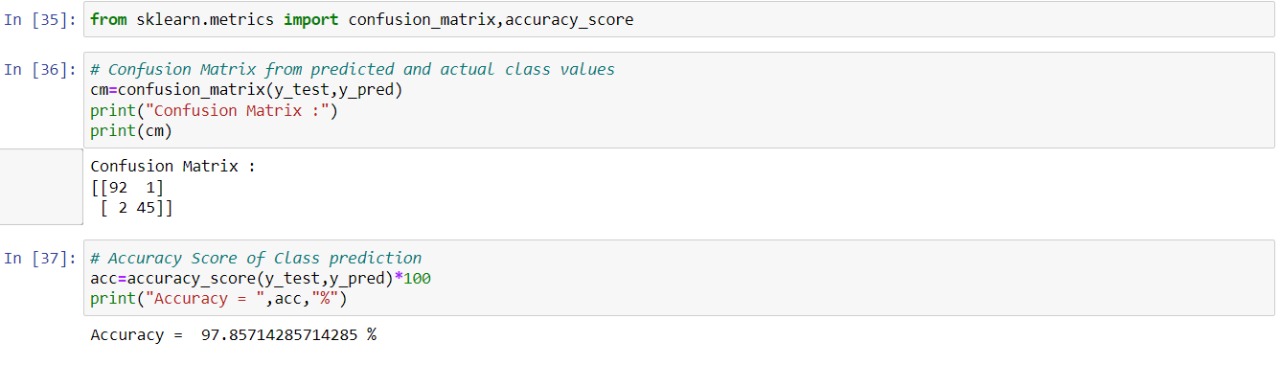


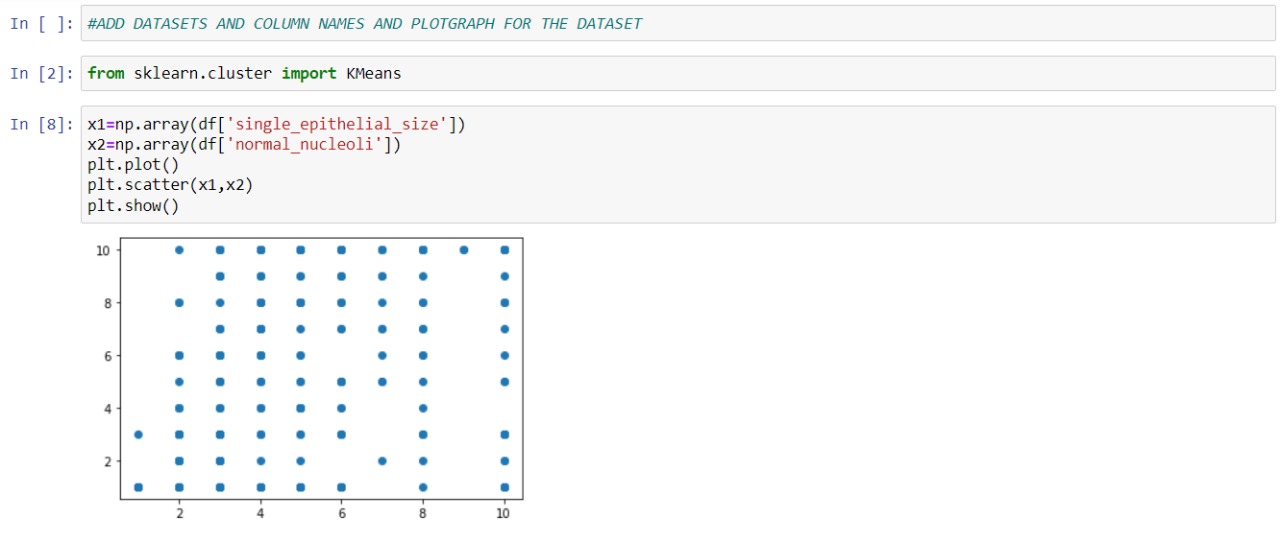


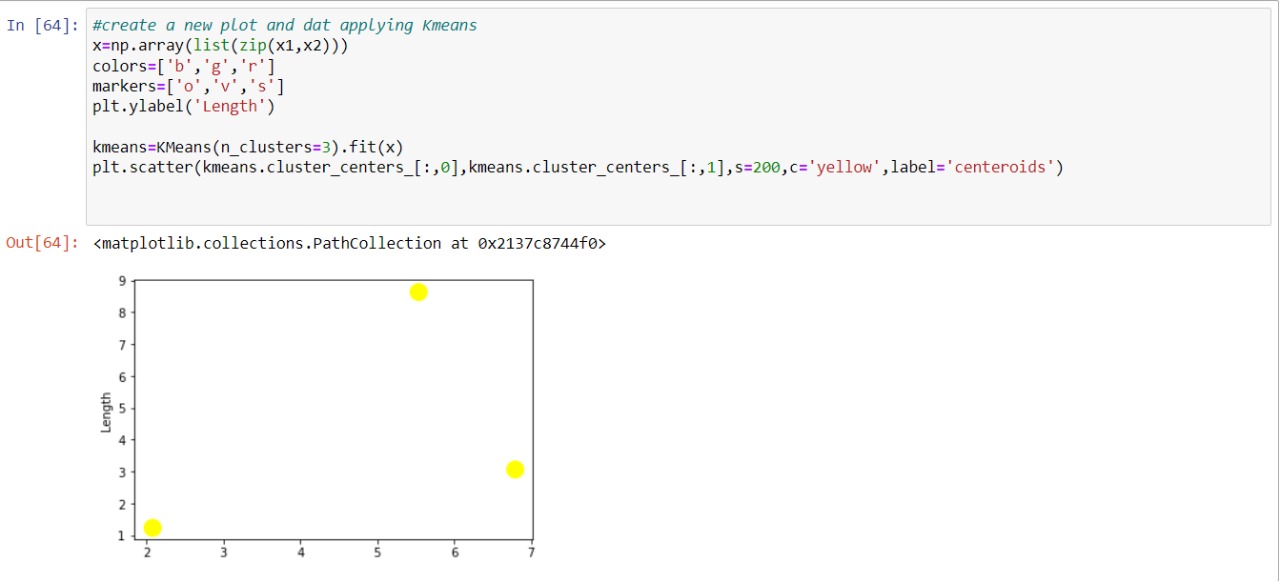


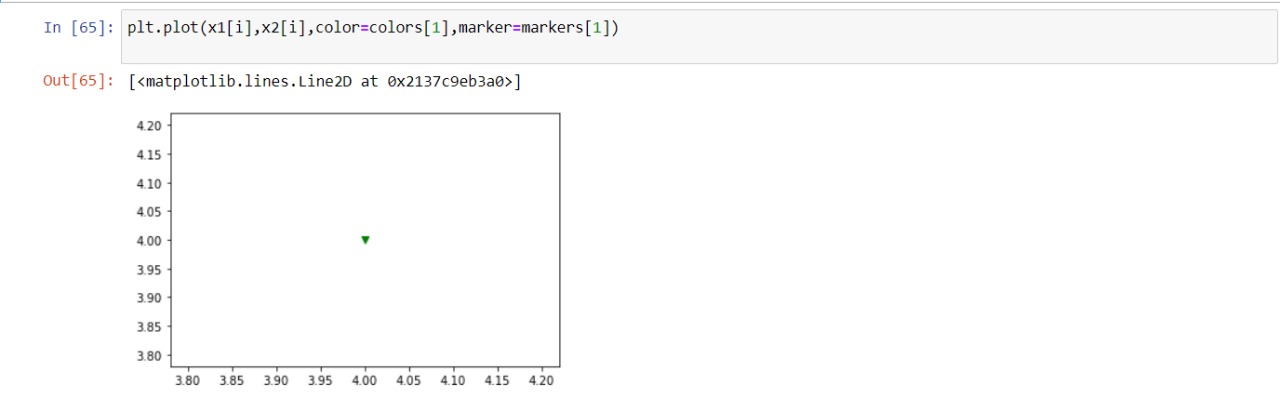












CONCLUSION: Breast cancer is considered to be one of the significant causes of death in women. Early detection of breast cancer plays an essential role to save women’s life. Breast cancer detection can be done with the help of modern machine learning algorithms. In this paper, we focus on how to deal with imbalanced data that have missing values using resampling techniques to enhance the classification accuracy of detecting breast cancer. In our work, three classifiers algorithms J48, NB, and SMO applied on two different breast cancer datasets. Results show that using the resample filter in the preprocessing phase enhances the classifier’s performance. In the future, the same experiments will apply to different classifiers and different datasets.