

##### **Indeed Inspiring Infotech** innovating ideas

##### A Project Report on­­

‘The Predictive Analysis by using the Wisconsin breast cancer diagnostic data set ’

**UNDER THE GUIDANCE OF PRESENTED BY**

**Kushal Sharma, Aishwarya Shinde Adesh Mekhe.**

# ACKNOWLEDGEMENT

Thanks to the **Indeed Inspiring Infotech** who gave us the golden opportunity to upgrade our skills. We would like to express our special thanks of gratitude to our trainer **Miss. Aishwarya Shinde Ma’am** and **Mr. Kushal Sharma Sir** who gave us the valuable guidance to do this project on the topic **“The Predictive Analysis by using the Wisconsin breast cancer diagnostic data set”**. It helped us in doing a lot of Research and we came to know about a lot of things related to this topic.

|  |  |  |
| --- | --- | --- |
| ***Sr. No*** | ***Title*** | ***Page No.*** |
| ***1*** | ***Introduction*** | ***4-5*** |
| ***2*** | ***Objectives*** | ***6*** |
| ***3*** | ***Terminology Used*** | ***7-8*** |
| ***4*** | ***About Dataset*** | ***9*** |
| ***5*** | ***Graphical Representation*** | ***10-12*** |
| ***6*** | ***Model Defining*** | ***13-17*** |
| ***7*** | ***Conclusion*** | ***18*** |

# INDEX

# INTRODUCTION

### About the breast

The breast is made up of different tissues, ranging from very fatty tissue to very dense tissue. Within this tissue is a network of lobes. Each lobe is made up of small, tube-like structures called lobules that contain milk glands. Small ducts connect the glands, lobules, and lobes, carrying milk from the lobes to the nipple. The nipple is located in the middle of the areola, which is the darker area that surrounds the nipple.  
  
Blood and lymph vessels also run throughout the breast. Blood vessels nourish the cells by delivering oxygen and nutrients and also removing waste and carbon dioxide. Lymph vessels, unlike blood vessels, only carry fluid away from tissues. They connect to lymph nodes and the lymphatic system, which drains bodily waste products. Lymph nodes are the small, bean-shaped organs that help fight infection. Groups of lymph nodes are located in different areas throughout the body, such as in the neck, groin, and abdomen. Regional lymph nodes of the breast are those nearby the breast, such as the lymph nodes under the arm, which are called axillary lymph nodes.

### About breast cancer

Cancer begins when healthy cells in the breast change and grow out of control, forming a mass or sheet of cells called a tumor. A tumor can be cancerous or benign. A cancerous tumor is malignant, meaning it can grow and spread to other parts of the body. A benign tumor means the tumor can grow but has not spread.

This guide covers both non-invasive (stage 0) as well as early-stage and locally advanced invasive breast cancer, which includes stages I, II, and III. The stage of breast cancer describes how much the cancer has grown, and if or where it has spread.

Although breast cancer most commonly spreads to nearby lymph nodes, in which case the breast cancer is still considered a local or regional disease, it can also spread further through the body through the blood vessels and/or lymph nodes to areas such as the bones, lungs, liver, and brain. This is called metastatic or stage IV breast cancer and is the most advanced stage of the disease. However, the involvement of nearby lymph nodes alone is generally not stage IV breast cancer.

### Types of breast cancer

Breast cancer can be invasive or non-invasive. Invasive breast cancer is cancer that spreads into surrounding tissues and/or distant organs. Non-invasive breast cancer does not go beyond the milk ducts or lobules in the breast. There are multiple types of breast cancers, which are classified based on how they look under a microscope.

* **Ductal carcinoma.** This is the most common type of breast cancer.
  + **Ductal carcinoma in situ (DCIS).** This is a non-invasive cancer (stage 0) that is located only in the duct and has not spread outside the duct.
  + **Invasive or infiltrating ductal carcinoma.** This is cancer that has spread outside of the ducts or lobules.
* **Invasive lobular carcinoma.** This is a less commonly occurring type of breast cancer that has spread outside of the ducts or lobules.

### 

# OBJECTIVE

* To Learn concept of machine learning using Wisconsin Breast Cancer Dataset
* To extract pattern and insights from the data.
* The goal is to increase the proportion of breast cancers identified at an early stage
* Understand the Dataset & cleanup
* Also fine-tune the hyperparameters & compare the evaluation metrics of various classification algorithm

# MACHINE LEARNING TECHNIQUES AND SOFTWARES

### MACHINE LEARNING TOOLS

* EDA (Exploratory Data Analysis)
* Logistic Regression
* Decision Tree
* RF (Random Forest Regression)
* Extra Tree Classifier

### SOFTWARES

* PYTHON (GOOGLE COLAB)

# TERMINOLOGIES

**Model Fitting**

Model fitting is the measure of how well a machine learning model generalizes data similar to that with which it was trained. A good model fit refers to a model that accurately approximates the output when it is provided with unseen inputs. Fitting refers to adjusting the parameters in the model to improve accuracy. The process involves running an algorithm on data for which the target variable (“labelled” data) is known to produce a machine learning model. Then, the model’s outcomes are compared to the real, observed values of the target variable to determine the accuracy. The next step involves adjusting the algorithm’s standard parameters in order to reduce the level of error and make the model more accurate when determining the relationship between the features and the target variable.

#### Supervised Machine Learning

Supervised learning is the types of machine learning in which machines are trained using well "labelled" training data, and on basis of that data, machines predict the output. The labelled data means some input data is already tagged with the correct output.

In supervised learning, the training data provided to the machines work as the supervisor that teaches the machines to predict the output correctly. It applies the same concept as a student learns in the supervision of the teacher. Supervised learning is a process of providing input data as well as correct output data to the machine learning model. The aim of a supervised learning algorithm is to find a mapping function to map the input variable(x) with the output variable(y).

#### Regression

Regression algorithms are used if there is a relationship between the input variable and the output variable. It is used for the prediction of continuous variables, such as Weather forecasting, Market Trends, etc.

Below are some popular Regression algorithms which come under supervised learning

* Linear Regression
* Regression Trees
* Non-Linear Regression
* Bayesian Linear Regression
* Polynomial Regression

#### Classification

Classification algorithms are used when the output variable is categorical, which means there are two classes such as Yes-No, Male-Female, True-false, etc.

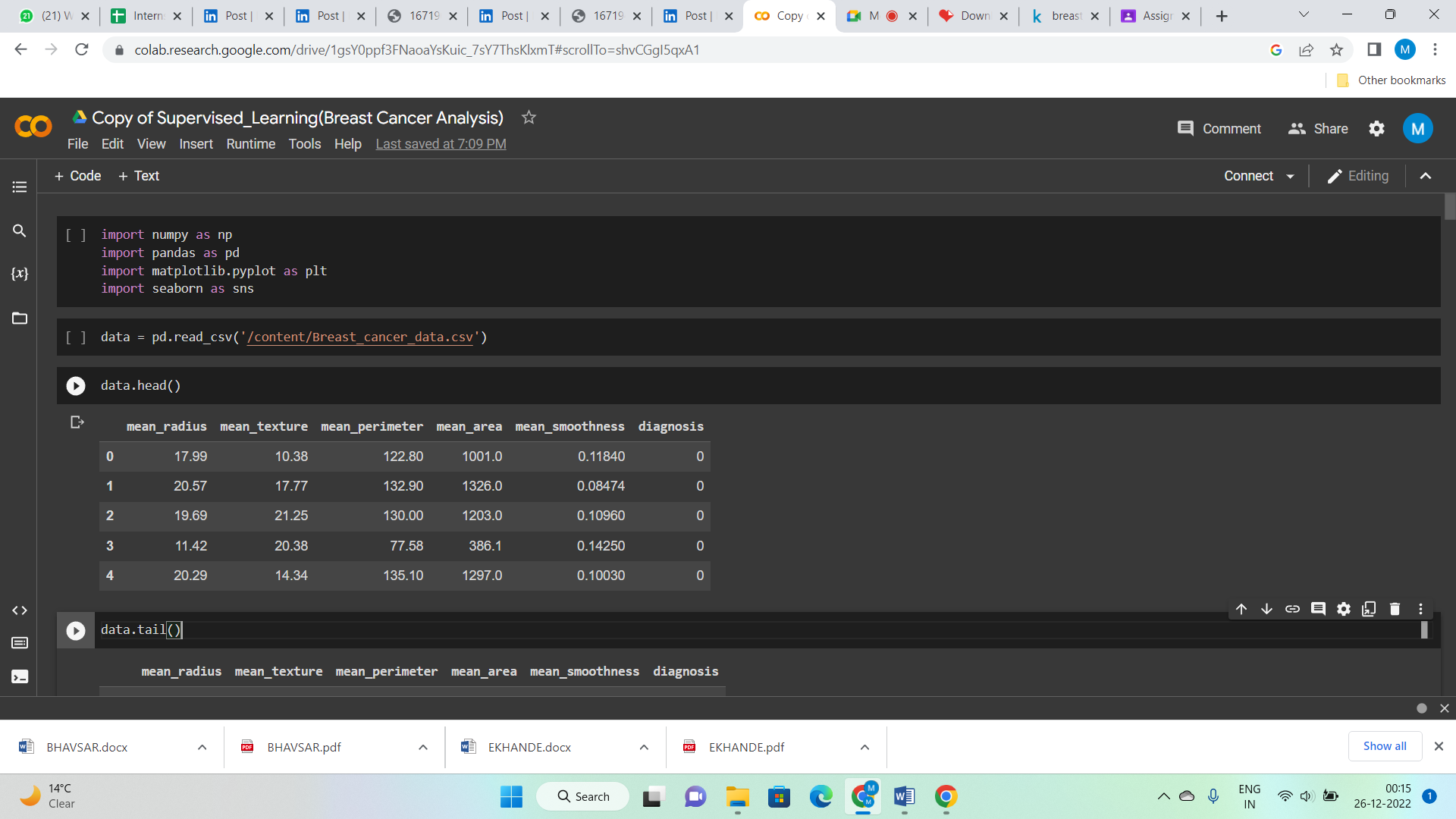
#### Logistic Regression

This is a classification algorithm which uses a logistic function to predict binary outcome (True/False, 0/1, Yes/No) given an independent variable. The aim of this model is to find a relationship between features and probability of particular outcome.

# ABOUT DATASET

**Data Description**: Dataset for AI for Social Good: Women Coders' Bootcamp organized by Artificial Intelligence for Development in collaboration with UNDP Nepal. Breast cancer is a disease in which cells in the breast grow out of control. There are different kinds of breast cancer.

The kind of breast cancer depends on which cells in the breast turn into cancer



In this dataset we have given that the following coloumn

**1. mean\_radius (mean of distances from center to points on the perimeter)**

**2. mean\_texture (standard deviation of gray-sacle values)**

**3. mean\_perimeter**

**4. mean\_area**

**5. mean\_smoothness (local variation in radius lengths)**

# EXPLORATORY DATA ANALYSIS

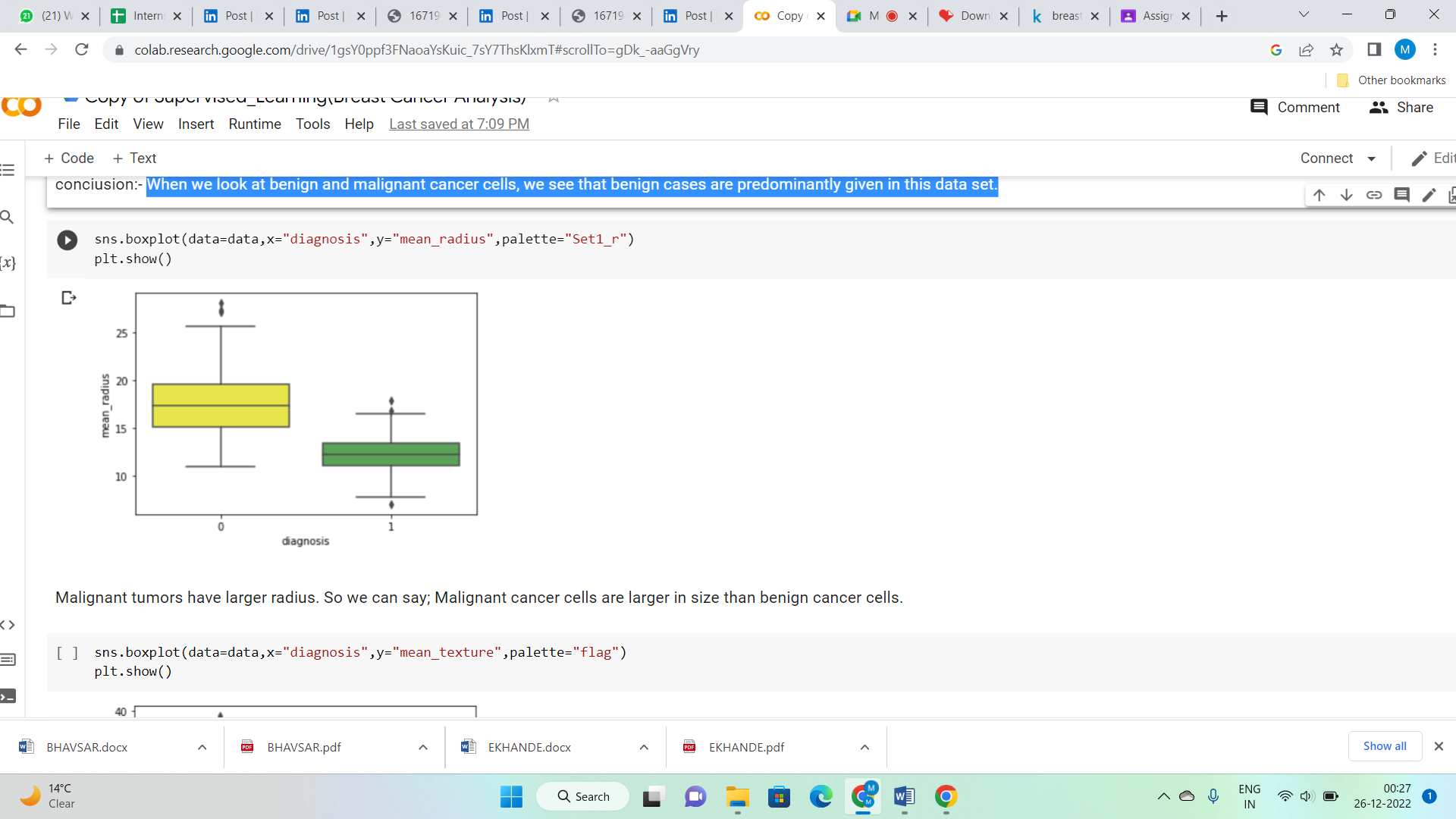
### Data Visualization

### PIE CHART

### 

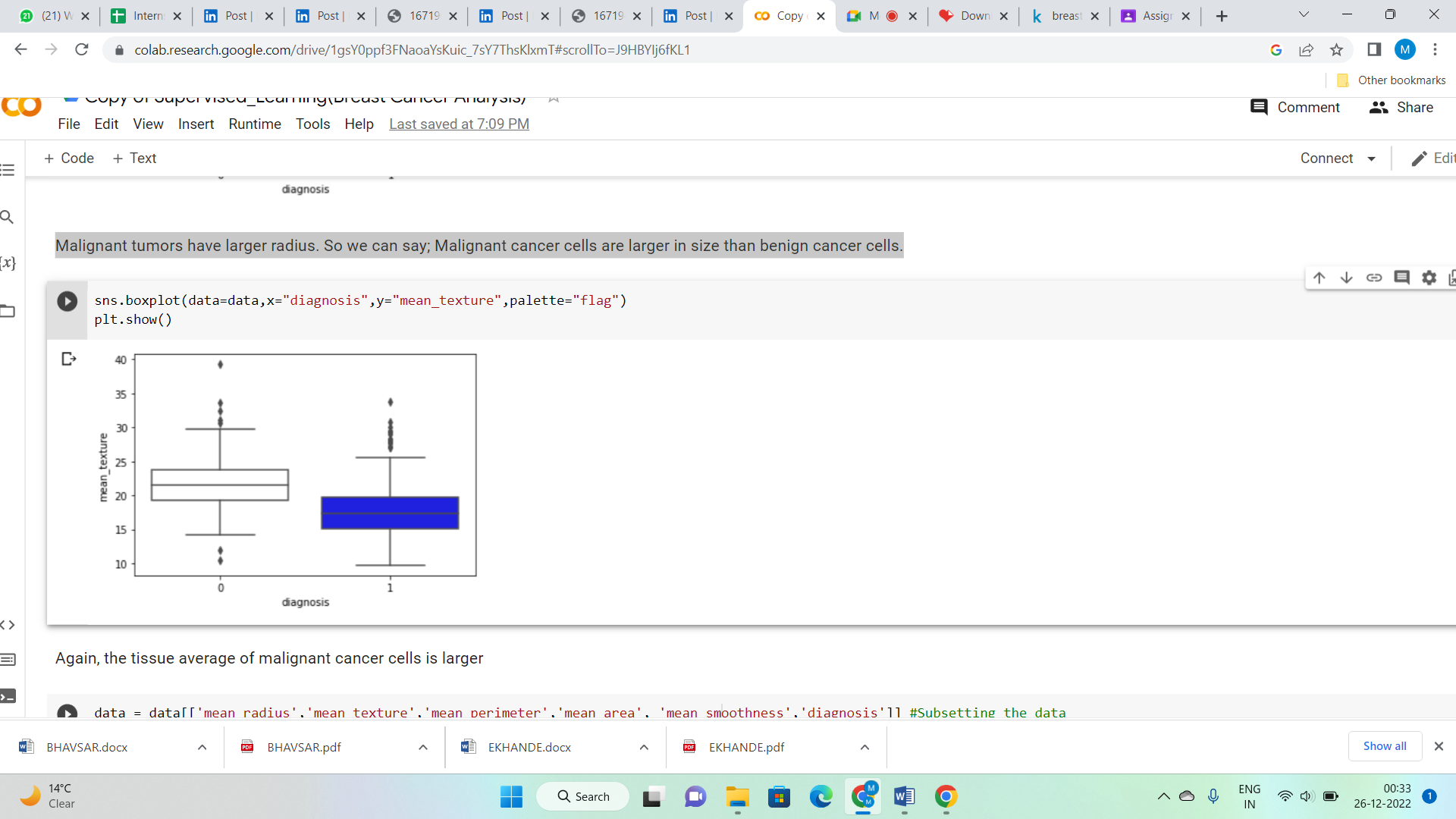
**Conclusion:- When we look at benign and malignant cancer cells, we see that benign cases are predominantly given in this data set**.

1. ***SUB-PLOT***



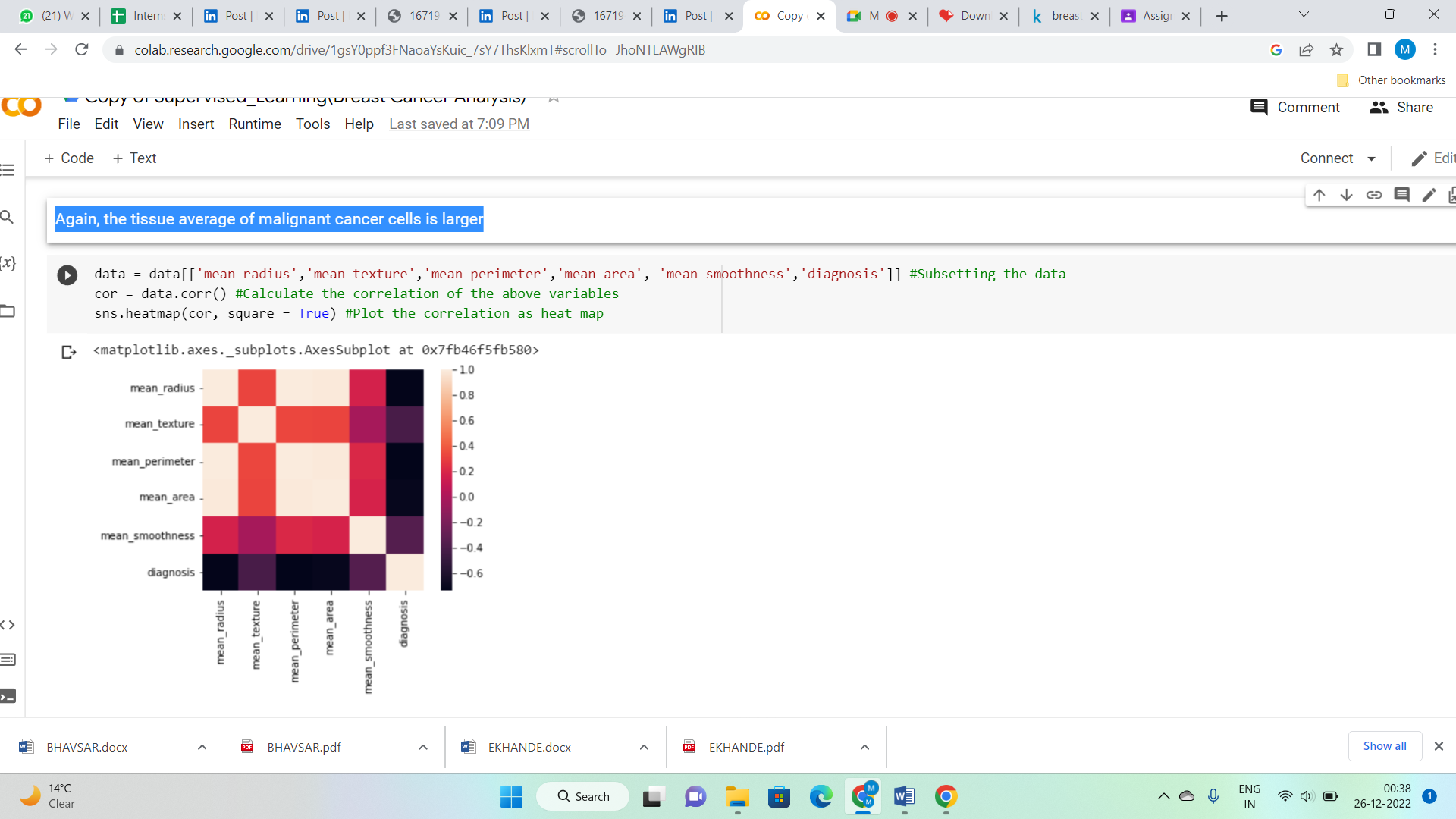
**Conclusion:- Malignant tumors have larger radius. So we can say; Malignant cancer cells are larger in size than benign cancer cells.**

***3***. ***SUB-PLOT***



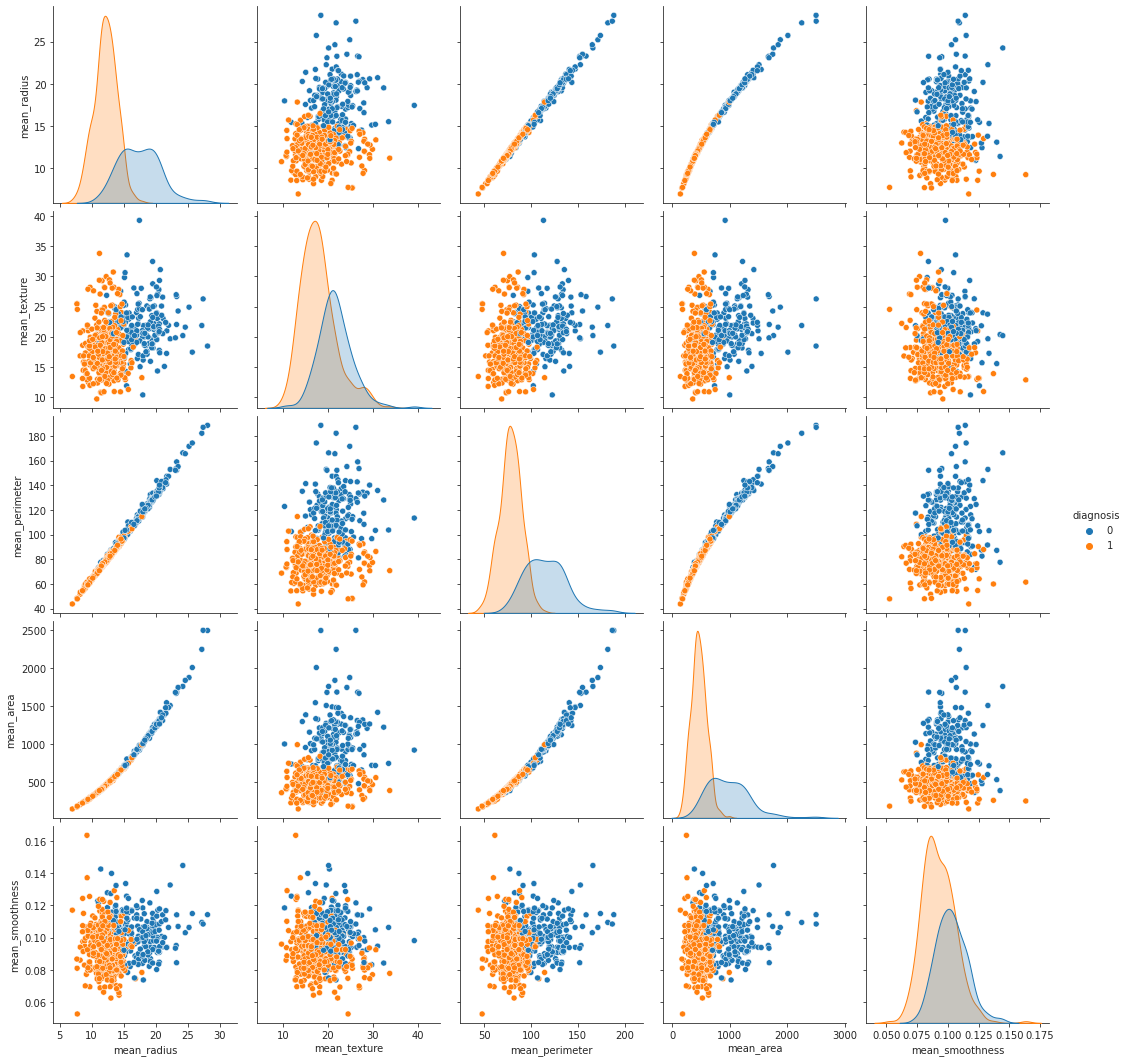
**Conclusion:-** **Again, the tissue average of malignant cancer cells is larger**

***4.CORRELATION MATRIX***



**Conclusion:- As you can see above, we obtain the heatmap of correlation among the variables. The color palette in the side represents the amount of correlation among the variables. The lighter shade represents a high correlation.**

***5.PAIR PLOT***

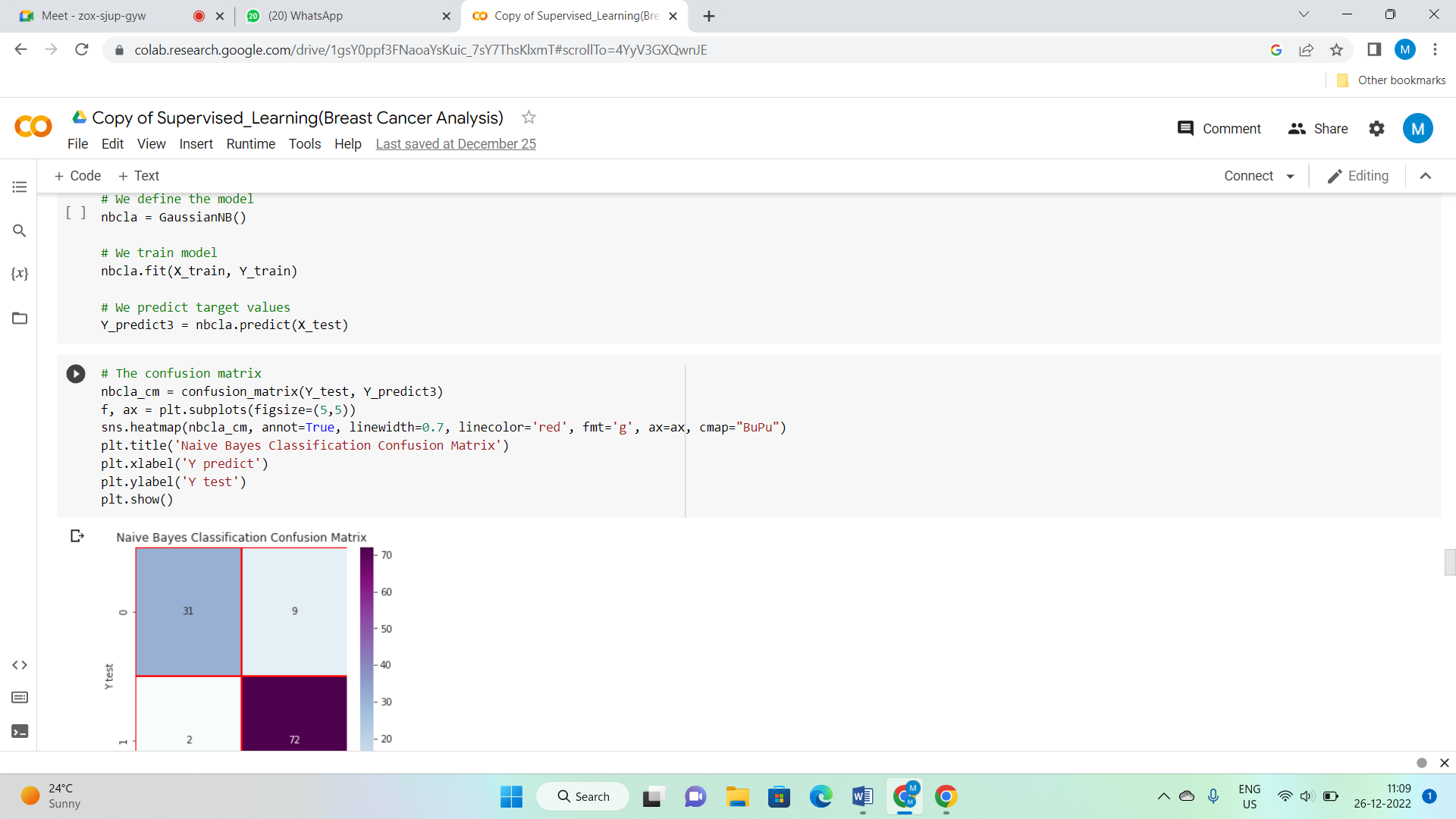


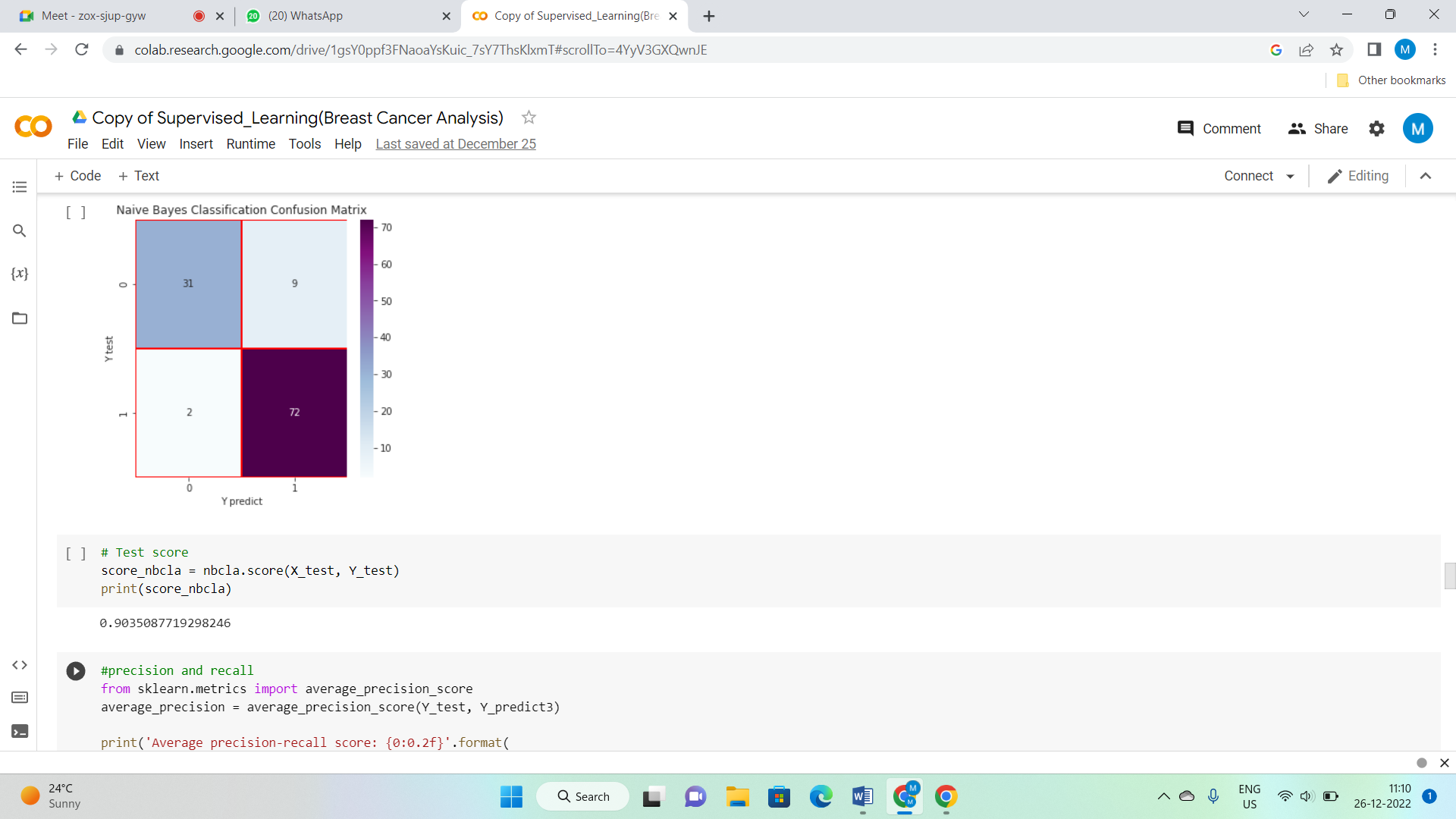
**Conclusion:- Data for training and testing To select a set of training data that will be input in the Machine Learning algorithm, to ensure that the classification algorithm training can be generalized well to new data. For this study using a sample size of 20%, assumed it ideal ratio between training and testing**

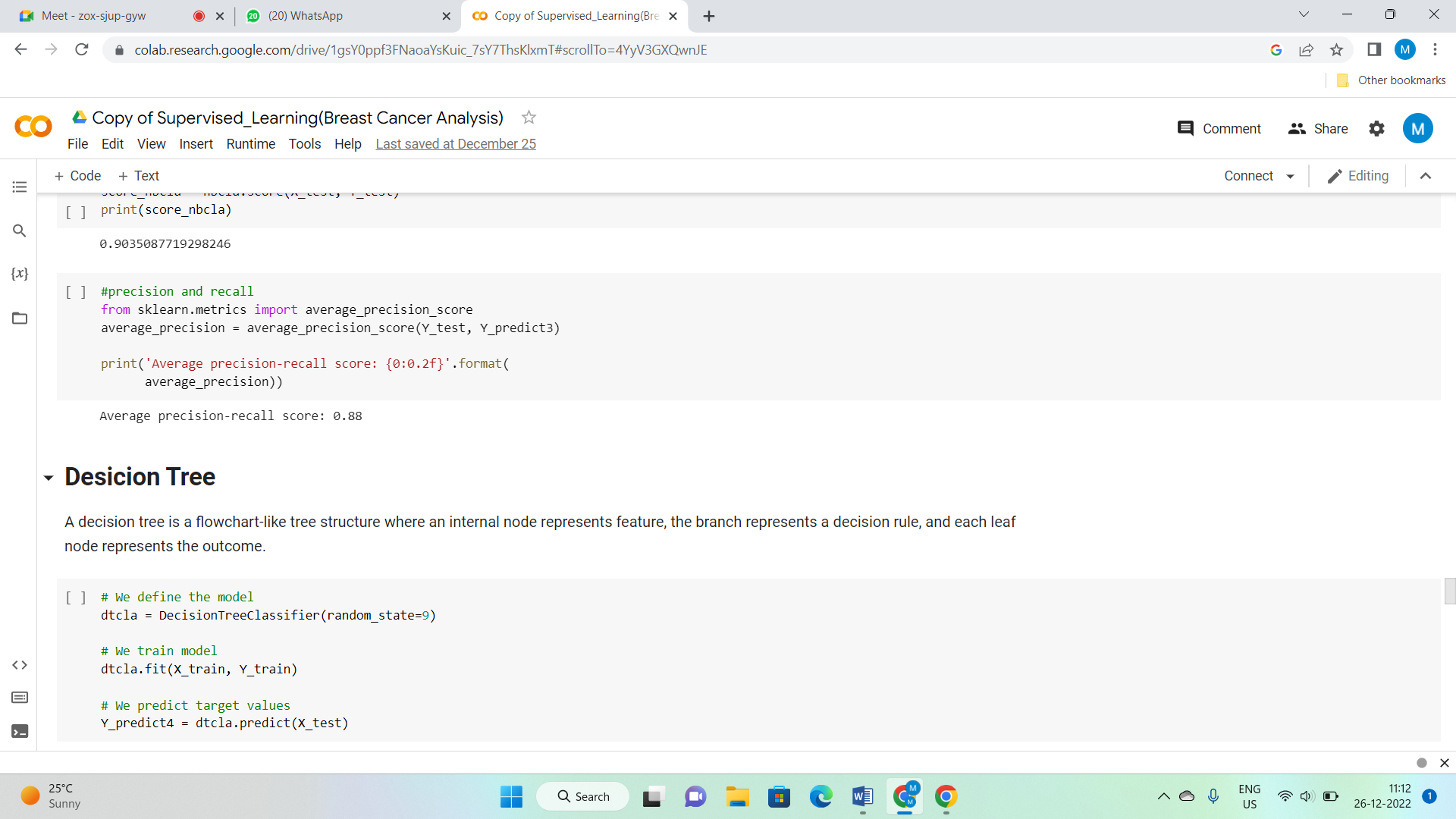
# MODEL FITTING

# LOGISTIC REGRSSION

# Naïve Baiyes







# Decision Tree

# 

# 

# 

# Randon Forest Classification

# 

# 

# 

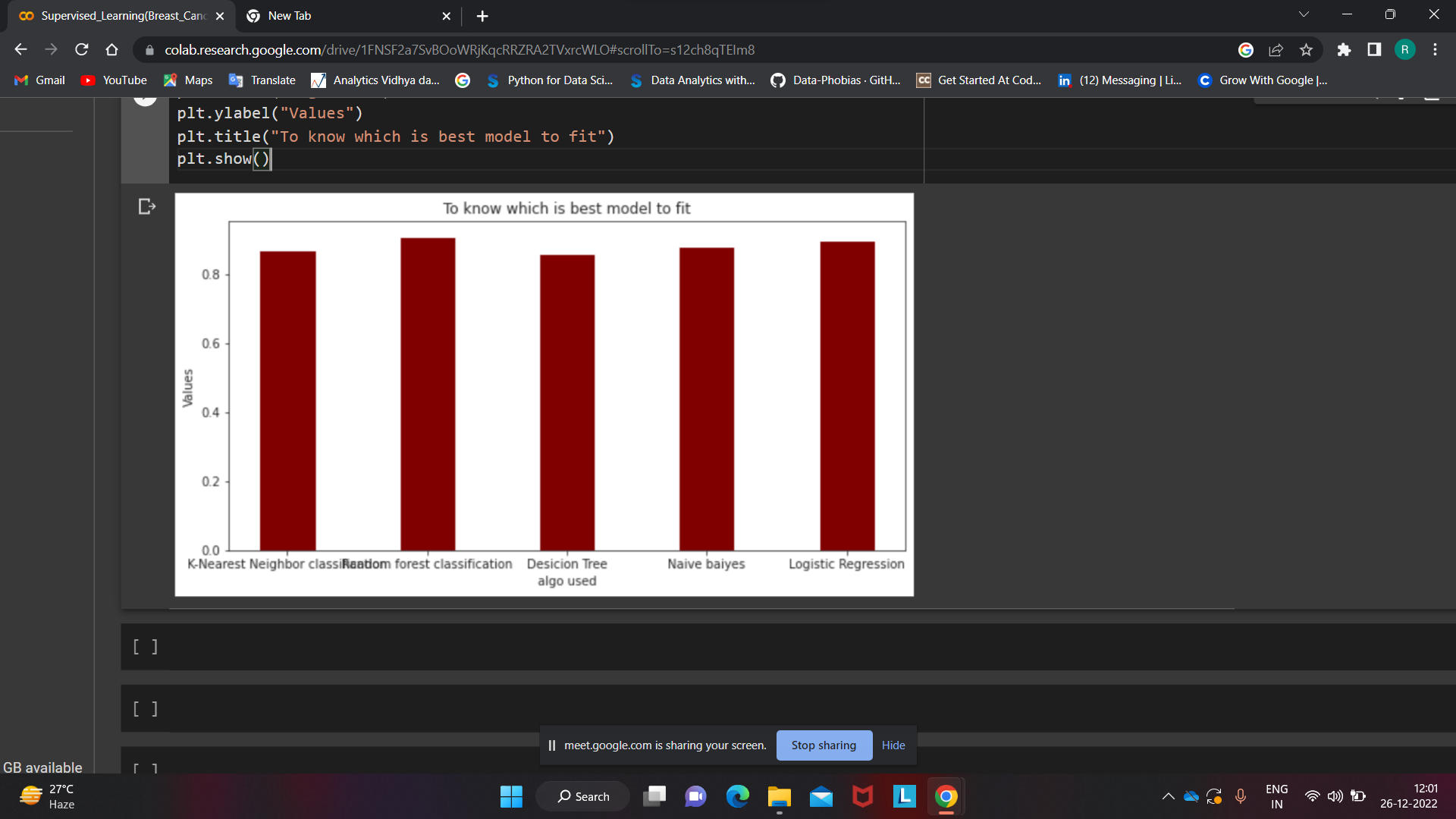
# K-Nearest Neighbor Classification

# 

# 

# 

# conclusion



**Comparison of classification techniques, we can evaluate that here Logistic Regression Classification and Random Forest have the most optimal result of the accuracy.**