## *CANCER CELL PREDICTION*

## *BY USING MACHINE LEARNING*

## *CONCEPTS*

**Under the Guidance Of**

**Sofikul Mullick**

Ardent Computech Pvt Ltd (An ISO 9001:2008 Certified) Module 132, SDF Building Sector - 5, Salt Lake City, Kolkata - 700 092

A

Project report

Submitted in Partial Fulfillment of The Requirements For the Award Of the

Bachelor of Technology Project Carried Out At



##### Ardent Computech Pvt Ltd (An ISO 9001:2008 Certified) Module 132, SDF Building Sector- 5, Salt Lake City, Kolkata-700 092

**Submitted by:**

**ACHINTA GORAIN**

**DIBYENDU GHOSH**

**SAGAR DAS**

**RUBEL HALDER**

Kolkata, West Bengal- 700040

1

## Project Responsibility Form

|  |  |  |
| --- | --- | --- |
| Serial Number | Name of Member | Responsibility |
| 1 | ACHINTA GORAIN  DIBYENDU GHOSH  SAGAR DAS  RUBEL HALDER | CODING |
| 2 | ACHINTA GORAIN  DIBYENDU GHOSH  SAGAR DAS  RUBEL HALDER | DOCUMENTATION |

# Cancer Cell

# Prediction

# By Using

# ML Concepts

### TABLE OF CONTENTS

#### S.NO. Name of the Topic Page No.

|  |  |  |
| --- | --- | --- |
| **1.** | **Company’s Profile** | **9** |
| **2.** | **Introduction** | **10** |
| 2.a. | Objective | 11 |
| 2.b. | Scope  Methodology | 12  12 |
| **3.** | **System Analysis** | **13** |
| 3.a | Identification of Need | 14 |
| 3.b | **Study of the System (Algorithm Used)** | **15-19** |
| **4.** | **System Design** | **20** |
| 4.a | Data Flow Diagram (DFD) | 21-27 |
| **5** | **Output Screen (Step by Step Working)** | **28-34** |
| **6. System Security Measures 55** | **System measurement** | **35** |
| 6.a | Limitations/Project Drawbacks | 36 |
| 6.b | Future Work and Possibilities | 36 |
| 6.c | Conclusion /Summary | 37 |
| 6.d | Bibliography/References | 38 |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

1. **ARDENT COMPUTECH PVT.LTD.**

Ardent Computech Private Limited is an ISO 9001-2008 certified Software Development Company in India. It has been operating independently since 2003. It was recently merged with ARDENT TECHNOLOGIES.

#### Ardent Technologies

ARDENT TECHNOLOGIES is a Company successfully providing its services currently in UK, USA, Canada and India. The core line of activity at ARDENT TECHNOLOGIES is to develop customized application software covering the entire responsibility of performing the initial system study, design, development, implementation and training. It also deals with consultancy services and Electronic Security systems. Its primary clientele includes educational institutes, entertainment industries, resorts, theme parks, service industry, telecom operators, media and other business houses working in various capacities.

#### Ardent Collaborations

ARDENT COLLABORATIONS, the Research Training and Development Department of ARDENT COMPUTECH PVT LTD is a professional training Company offering IT enabled services & industrial trainings for B-Tech, MCA, BCA, MSc and MBA fresher’s and experienced developers/programmers in various platforms. Summer Training / Winter Training / Industrial training will be provided for the students of B. TECH, M. TECH, MBA and MCA only.

Deserving candidates may be awarded stipends, scholarships and other benefits, depending on their performance and recommendations of the mentors.

#### Associations

Ardent is an ISO 9001:2008 company. It is affiliated to National Council of Vocational Training (NCVT), Directorate General of Employment & Training (DGET), Ministry of Labor & Employment, and Government of India.

### INTRODUCTION

Machine learning is a branch of artificial intelligence that employs a variety of statistical, probabilistic and optimization techniques that allows computers to “learn” from past examples and to detect hard-to-discern patterns from large, noisy or complex data sets. This capability is particularly well-suited to medical applications, especially those that depend on complex proteomic and genomic measurements. As a result, machine learning is frequently used in cancer diagnosis and detection. More recently machine learning has been applied to cancer prognosis and prediction. This latter approach is particularly interesting as it is part of a growing trend towards personalized, predictive medicine. In assembling this review we conducted a broad survey of the different types of machine learning methods being used, the types of data being integrated and the performance of these methods in cancer cell detection . A number of trends are noted, including a growing dependence on protein biomarkers and microarray data, a strong bias towards applications in prostate and breast cancer, and a heavy reliance on “older” technologies such artificial neural networks (ANNs) instead of more recently developed or more easily interpretable machine learning methods. A number of published studies also appear to lack an appropriate level of validation or testing. Among the better designed and validated studies it is clear that machine learning methods can be used to substantially (15-25%) improve the accuracy of predicting cancer susceptibility, recurrence and mortality. It is also evident that machine learning is also helping to improve our basic understanding OF CANCER CELL DETECTION.

### OBJECTIVE

The objective of cancer cell Prediction is to allow computers to “learn” from past examples and to detect hard-to-discern patterns from large, noisy or complex data sets. This capability is particularly well-suited to medical applications, especially those that depend on complex proteomic and genomic measurements. As a result, machine learning is frequently used in cancer diagnosis and detection.

Machine learning classifiers are very popular for detecting breast cancer. Several research works have been done in this area. Here a classifier algorithm named “Logistic Regression” has been modified to detect the malignancy or benignancy of the tumorous cell more accurately.

The main **Project Goal** is to create a Machine Learning Model with the help of Past records and Data set so that the model can easily Predict any cancer case with reference of the Data sets Provided.

It will be a very useful and beneficial model for the society as it will work very fast in accordance with any other systems available till now.

### SCOPE

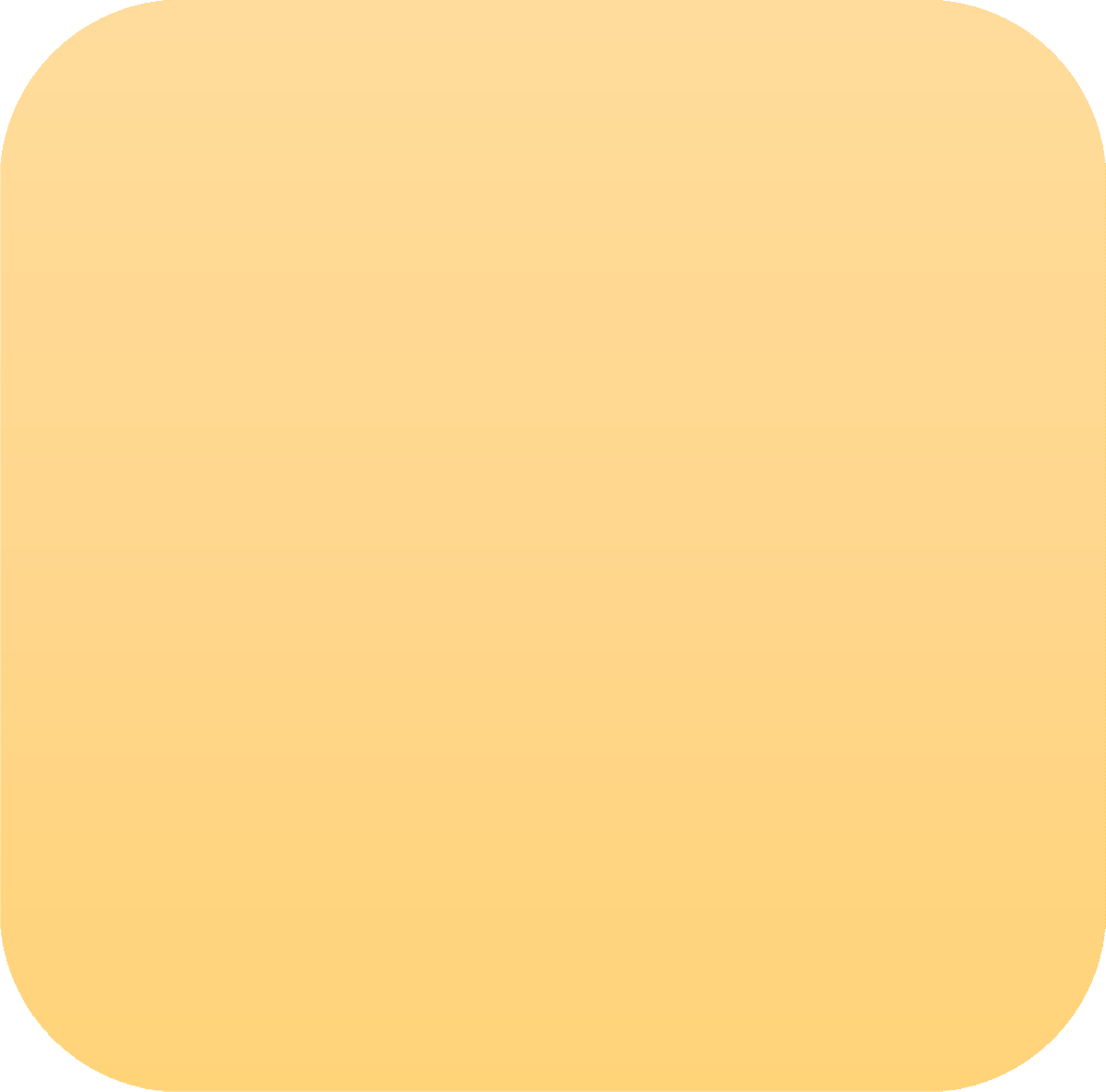
Today we’re at a position where things we require are achieved with just a few taps in handsets. As a result, ML methods have become a popular tool for medical researchers. These techniques can discover and identify patterns and relationships between them, from complex datasets, while they are able to effectively predict future outcomes of a **cancer** type.

**Machine learning** is taking over the world - it is benefiting companies across industries. It is helping organizations create systems that can understand, learn, predict, adapt and operate on their own. Thus, understanding how **machine learning** works is one of the most **valuable** and **useful** things you can do.

**METHODOLOGY**

The various Methods used for making this Project are as follows :-

1. Data Collection
2. Data Analysis
3. Data Cleaning
4. Data Plotting
5. Model Creation
6. Training and Testing the created Model
7. Accuracy checking of the models Prediction and matching with the records provided



# SYSTEM ANALYSIS

### IDENTIFICATION OF NEED

System analysis is a process of gathering and interpreting facts, diagnosing problems and the information to recommend improvements on the system. It is a problem-solving activity that requires intensive communication between the system users and system developers. System analysis or study is an important phase. The system studies the minutest detail and gets analyzed. The system analysist plays the role of the interrogator and dwells deep into the working of the present system. The System is viewed as a whole and the input to the system are identified.

The world is quietly being reshaped by machine learning. We no longer need to teach computers how to perform complex tasks like image recognition or text translation: instead, we build systems that let them learn how to do it themselves.

### STUDY OF THE SYSTEM

***Logistic regression :***

is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (a form of binary regression).

Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

## Logistic Function

Logistic regression is named for the function used at the core of the method, the logistic function.

The [logistic function](https://en.wikipedia.org/wiki/Logistic_function), also called the sigmoid function was developed by statisticians to describe properties of population growth in ecology, rising quickly and maxing out at the carrying capacity of the environment. It’s an S-shaped curve that can take any real-valued number and map it into a value between 0 and 1, but never exactly at those limits.

1 / (1 + e^-value)

Where e is the [base of the natural logarithms](https://en.wikipedia.org/wiki/E_(mathematical_constant)) (Euler’s number or the EXP() function in your spreadsheet) and value is the actual numerical value that you want to transform. Below is a plot of the numbers between -5 and 5 transformed into the range 0 and 1 using the logistic function.

## Representation Used for Logistic Regression

Logistic regression uses an equation as the representation, very much like linear regression.

Input values (x) are combined linearly using weights or coefficient values (referred to as the Greek capital letter Beta) to predict an output value (y). A key difference from linear regression is that the output value being modeled is a binary values (0 or 1) rather than a numeric value.

Below is an example logistic regression equation:

y = e^(b0 + b1\*x) / (1 + e^(b0 + b1\*x))

where y is the predicted output, b0 is the bias or intercept term and b1 is the coefficient for the single input value (x). Each column in your input data has an associated b coefficient (a constant real value)

## Making Predictions with Logistic Regression

Making predictions with a logistic regression model is as simple as plugging in numbers into the logistic regression equation and calculating a result.

Let’s make this concrete with a specific example.

Let’s say we have a model that can predict whether a person is male or female based on their height (completely fictitious). Given a height of 150cm is the person male or female.

We have learned the coefficients of b0 = -100 and b1 = 0.6. Using the equation above we can calculate the probability of male given a height of 150cm or more formally P(male|height=150). We will use EXP() for e, because that is what you can use if you type this example into your spreadsheet:

y = e^(b0 + b1\*X) / (1 + e^(b0 + b1\*X))

y = exp(-100 + 0.6\*150) / (1 + EXP(-100 + 0.6\*X))

y = 0.0000453978687

Or a probability of near zero that the person is a male.

In practice we can use the probabilities directly. Because this is classification and we want a crisp answer, we can snap the probabilities to a binary class value, for example:

0 if p(male) < 0.5

1 if p(male) >= 0.5

Now that we know how to make predictions using logistic regression.

**Decision Tree :-**

**Decision tree** can be used to predict a pattern or to classify the class of a data. It is commonly used in data mining. The goal to use the decision tree Algorithm is to create the model that predict the Target variable based upon the several input variables. In decision tree each leaf represents a value of target variable given the value of input variables represented by the path from the root of the leaf. A tree can be learned by splitting the source set into the subset based on an attribute value test. This process repeated in each derived subset called recursive. The general fashion for tree is top down induction.

**Decision tree used in data mining are of two main types:-**

**1. Classification Tree:-**

When the predicted outcome is the class to which the data belongs.

**2. Regression Tree:-**

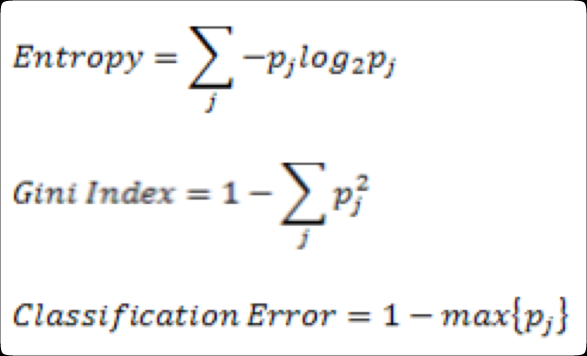
When the predicted outcome can be consider a real number.

The term classification & regression tree(CART) analysis is an umbrella term used to refer to both of the above procedure.

Some other techniques constructs more than one

decision tree like, Bagging, Random Forest, Boosted tree etc. We have used Random Forest decision tree for this project. The algorithm that are used for constructing decision trees usually work top-down by choosing a variable at each step, that is next best variable to use in splitting the set of variables. “Best” is defined by how well the variable splits the set into homogeneous subsets that have the same value as target variable. Different algorithm use

different formulae for measuring “Best”. These are the mathematical function through we measure the Impurity.

tree etc.

# SYSTEM



**DESIGN**

### DATA FLOW DIAGRAM

A **data flow diagram (DFD)** is a graphical representation of the "flow" of data through an information system, modelling its process aspects. A **DFD** is often used as a preliminary step to create an overview of the system, which can later be elaborated.

DFDs can also be used for the visualization of data processing (structured design).

A DFD shows what kind of information will be input to and output from the system, where the data will come from and go to, and where the data will be stored. It does not show information about the timing of process or information about whether processes will operate in sequence or in parallel (which is shown on a flowchart).

diagrams can be used to provide the end user with a physical idea of where the data they input ultimately has an effect upon the structure of the whole system from order to dispatch to report.

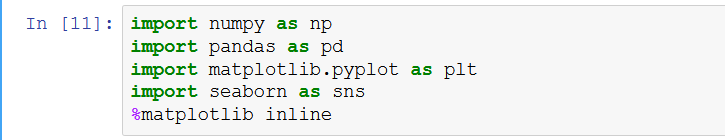
How any system is developed can be determined through a data flow diagram model.

##### DFD EXAMPLE:

Customer

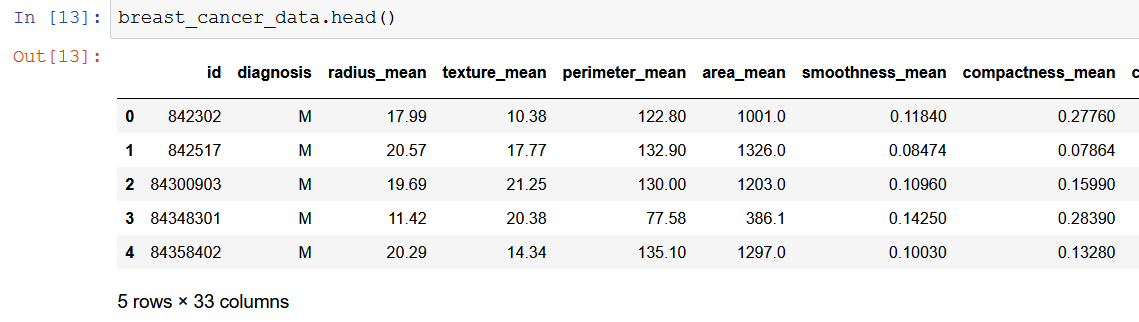
Database Input ML Output

Now **import the packages/libraries** to make it easier to write the program.



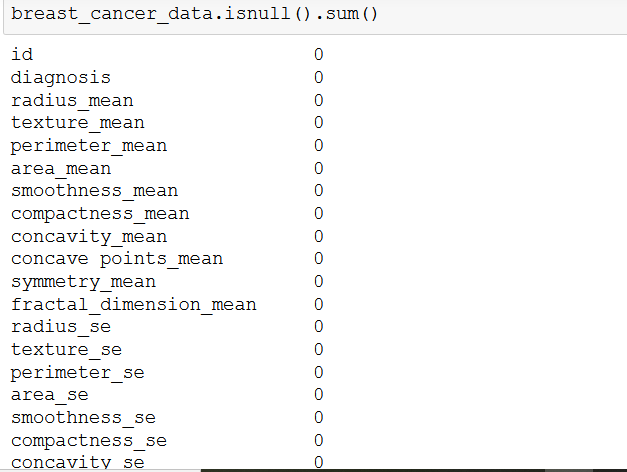
Put the data file

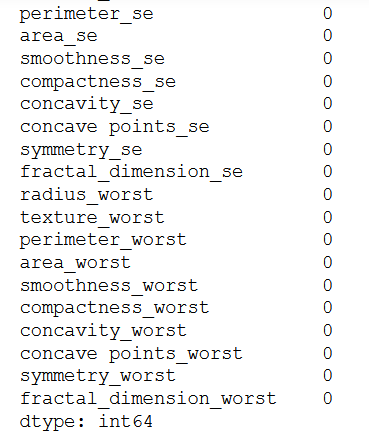




Explore the data and **count the number of rows and columns** in the data set of breast cancer. Their are 5rows of data which means their are 5 patients in this data set, and 33 columns which mean their are 33 features or data points for each patient for breast cancer.

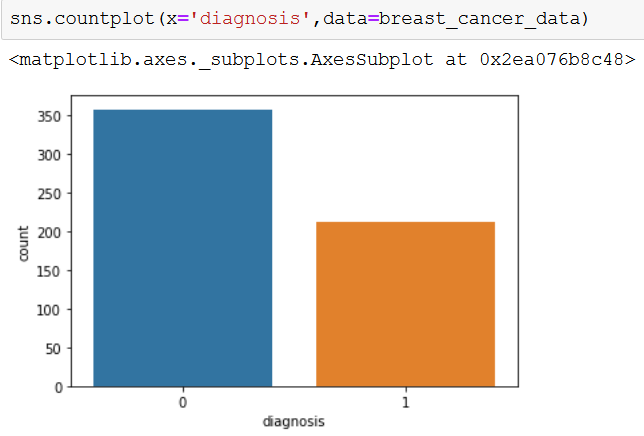
Count of all the empty values per column/feature:





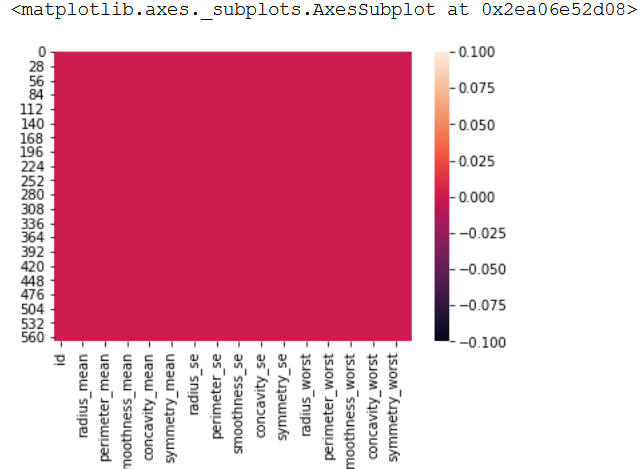
**Visualize the counts**, by creating a count plot.

I can see from the data types that all of the columns/features are numbers except for the column ‘diagnosis’, which is categorical data represented as an object in python.

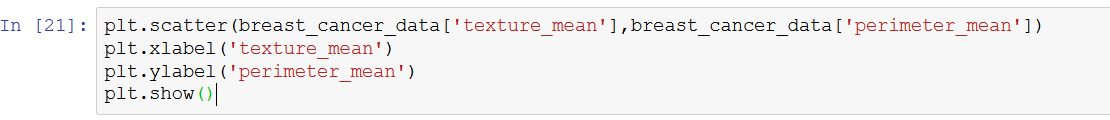


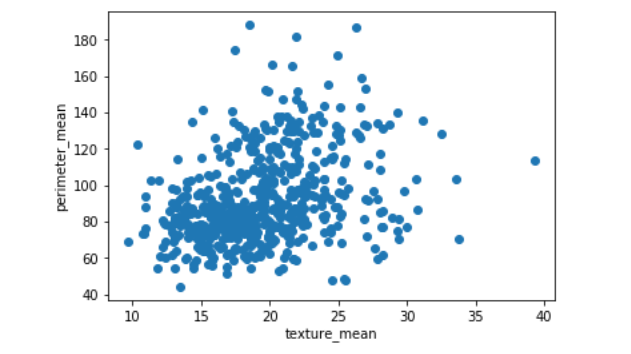
There are some other diagram which explain us cancer cell detection briefly



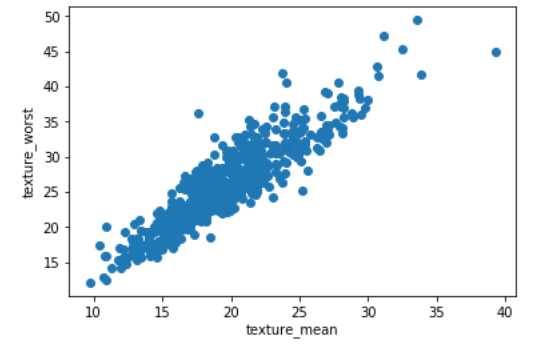


If we plot the breast cancer data as x-label and y-label

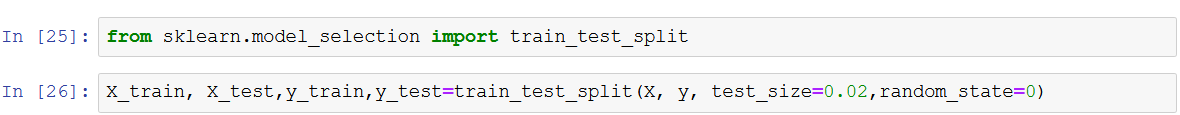


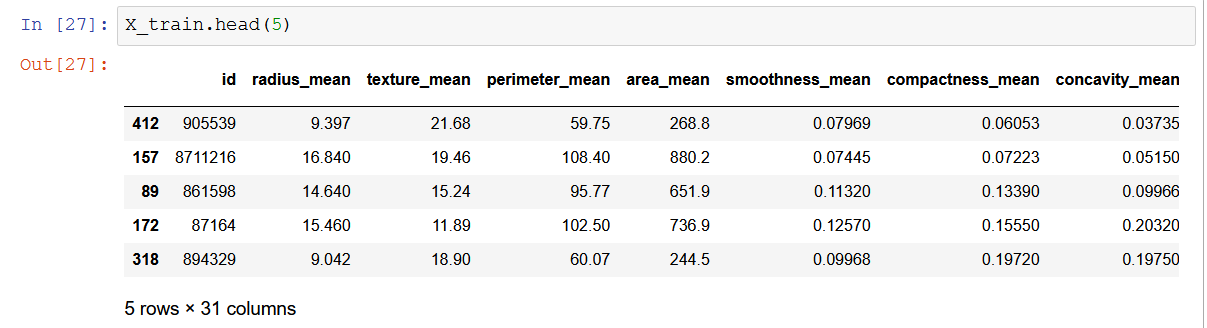


And

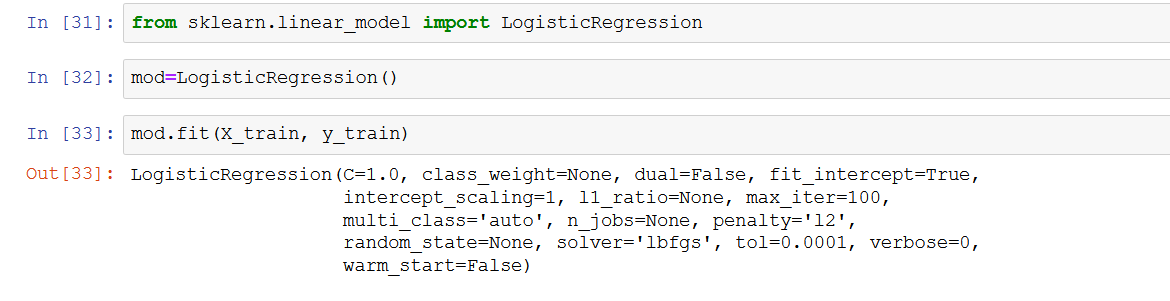
  


**Split the data** again, but this time into 98% training and 2% testing data sets.

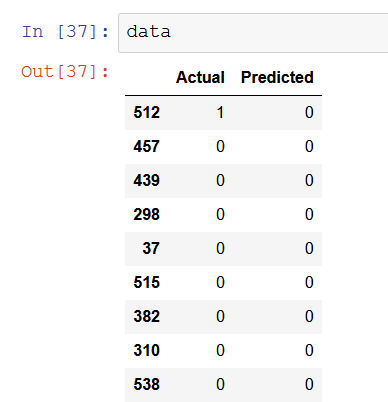




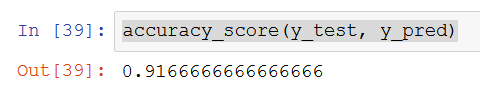
**By** applying Logistic Regression



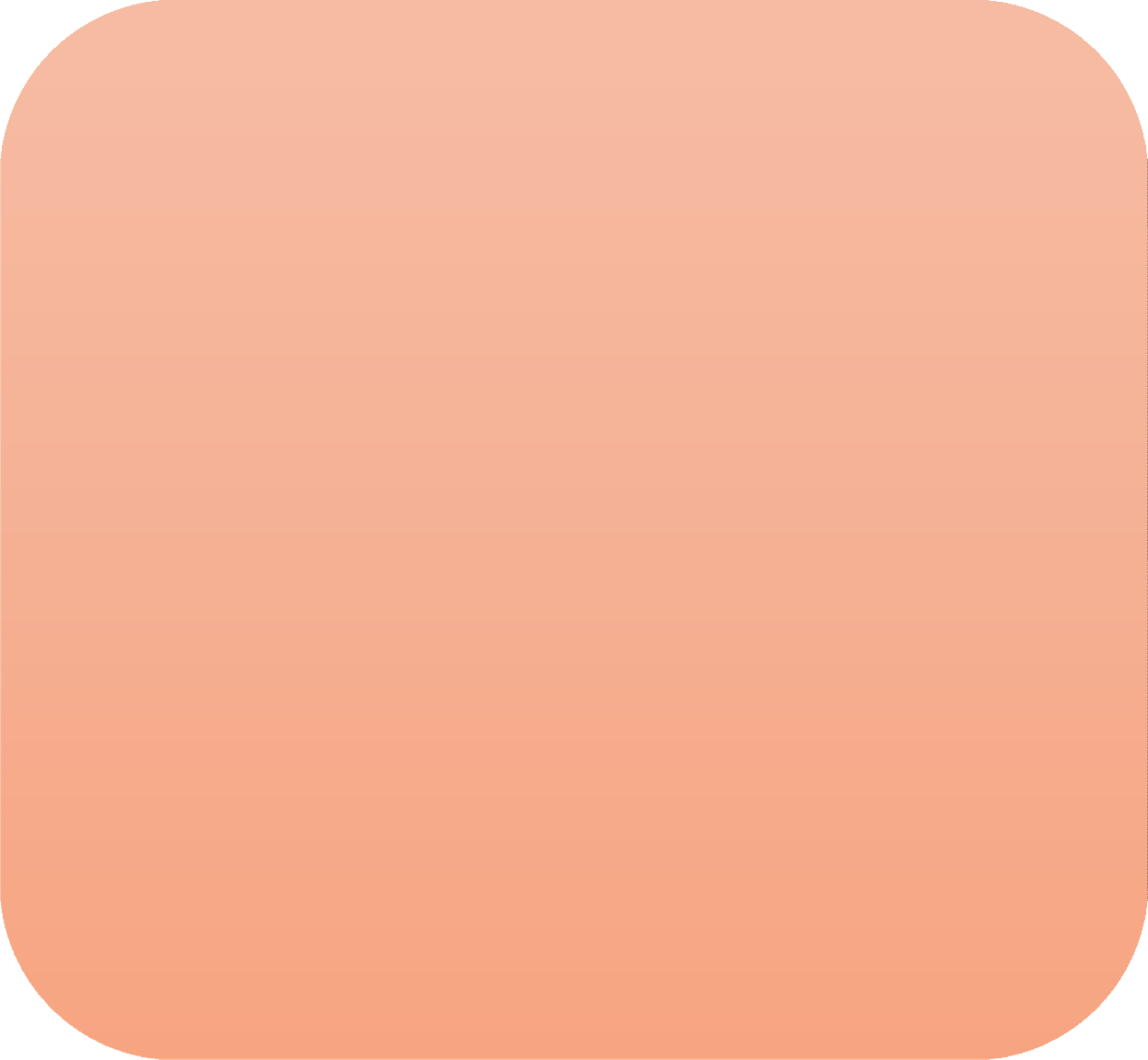
We get actual and predict data:

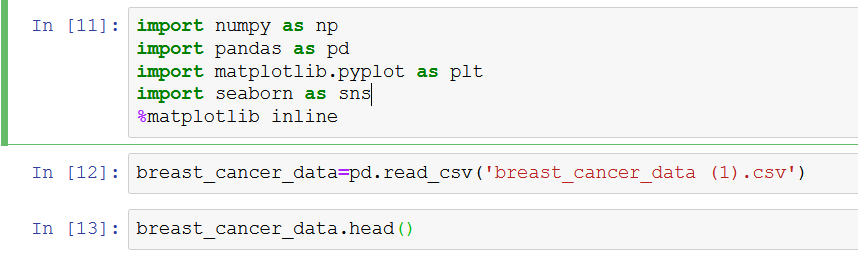


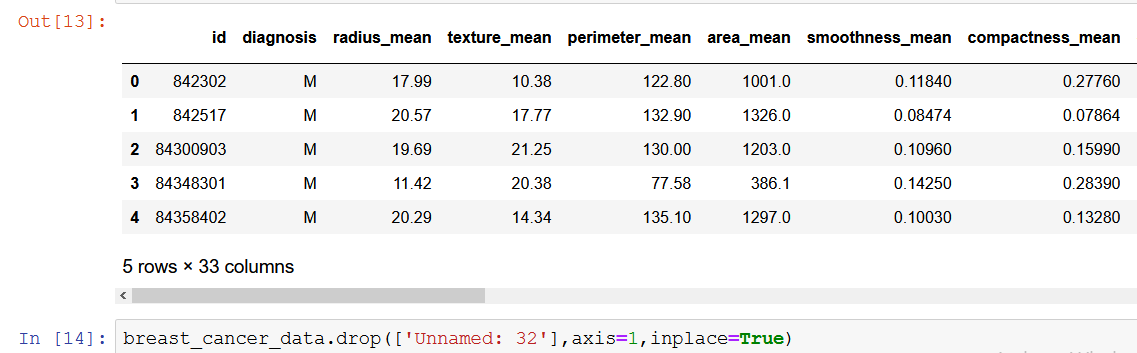
**Accuracy score :**

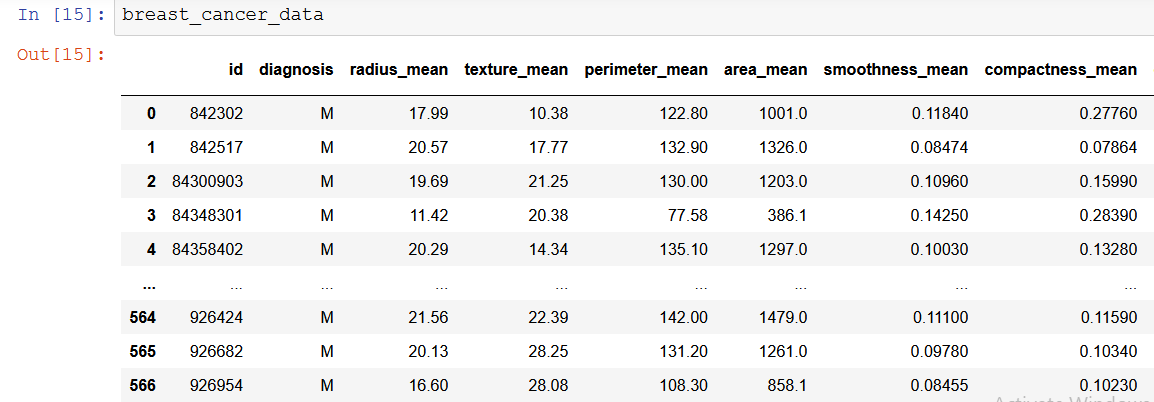
****

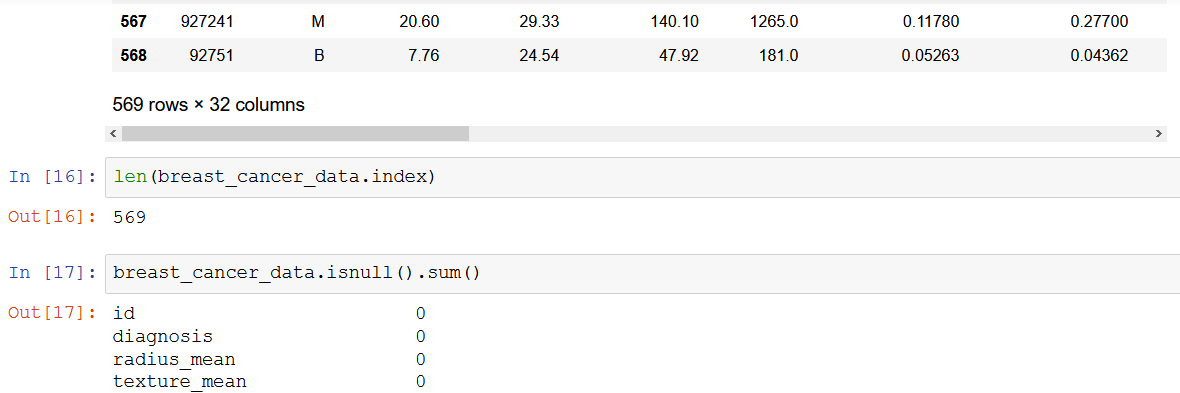
# OUTPUT SCREEN

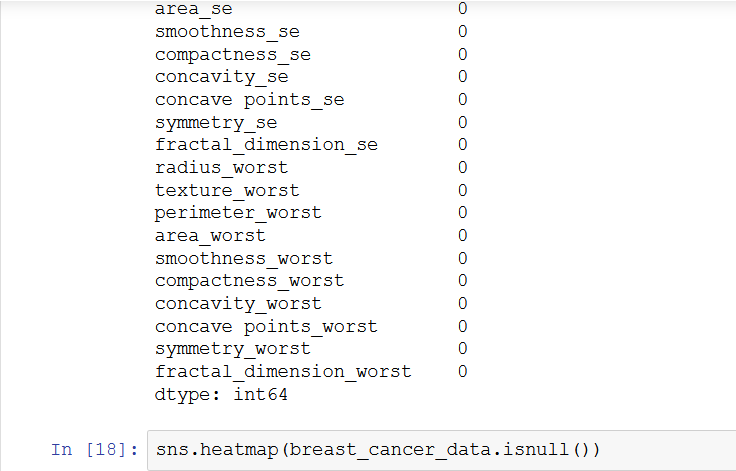


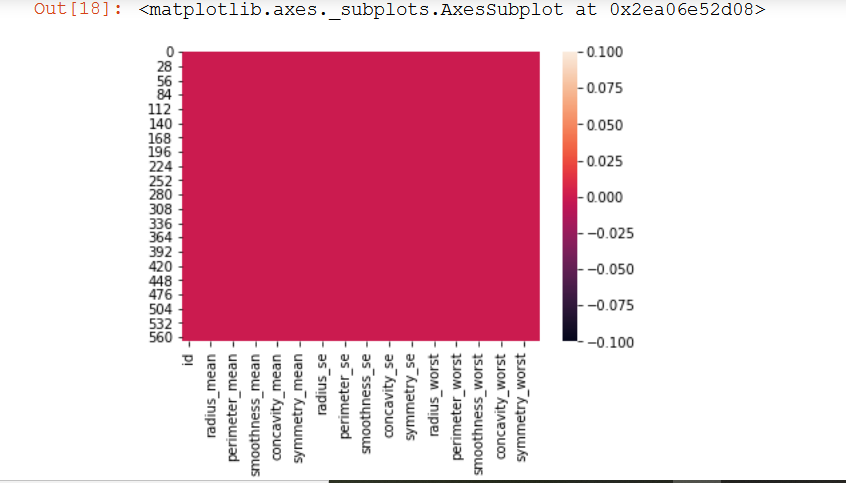


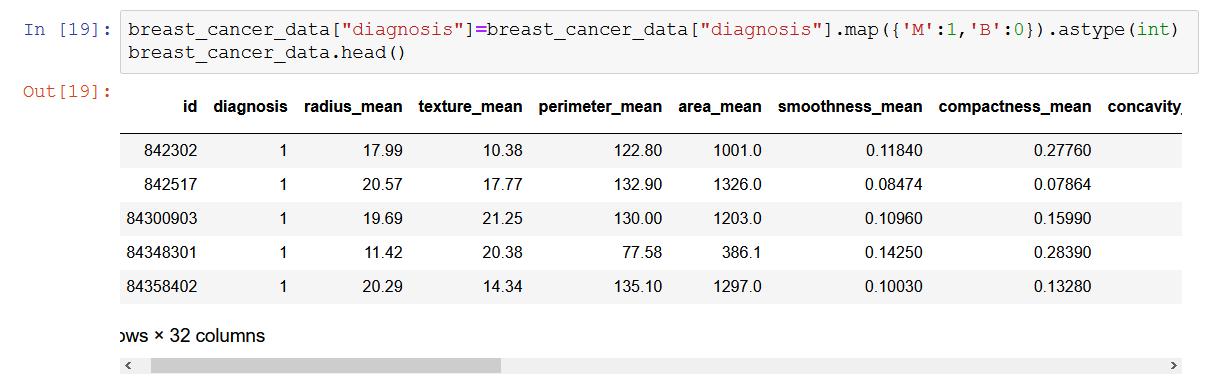


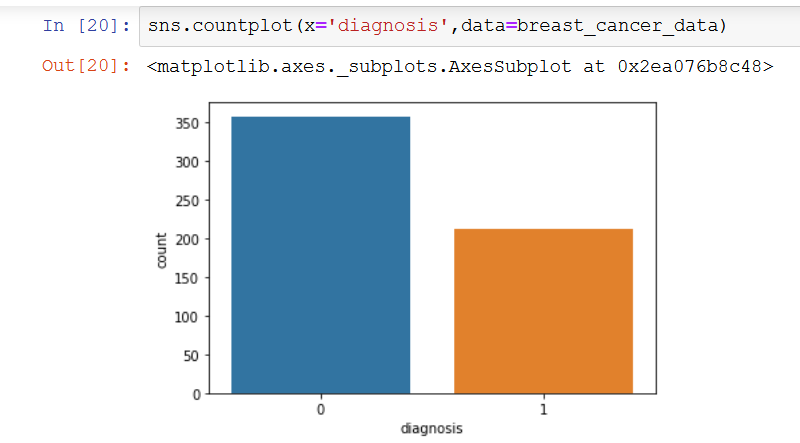


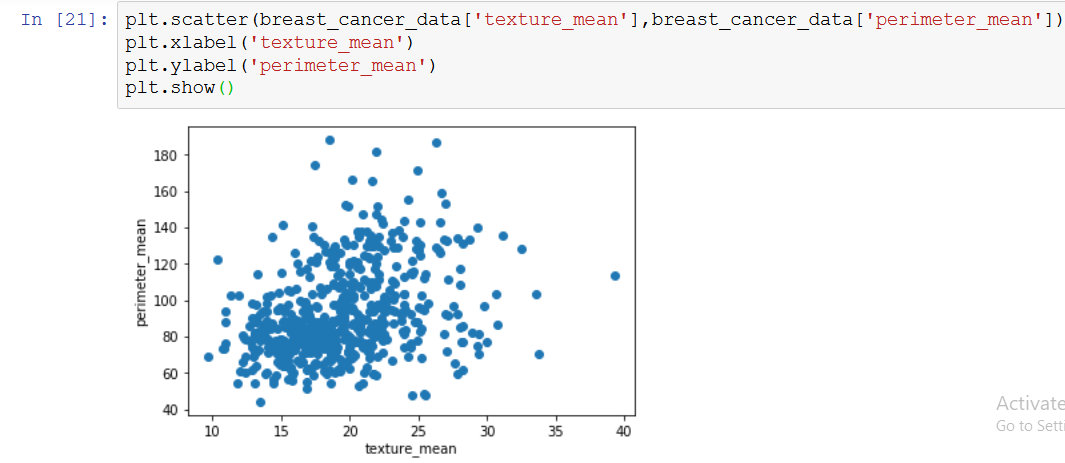


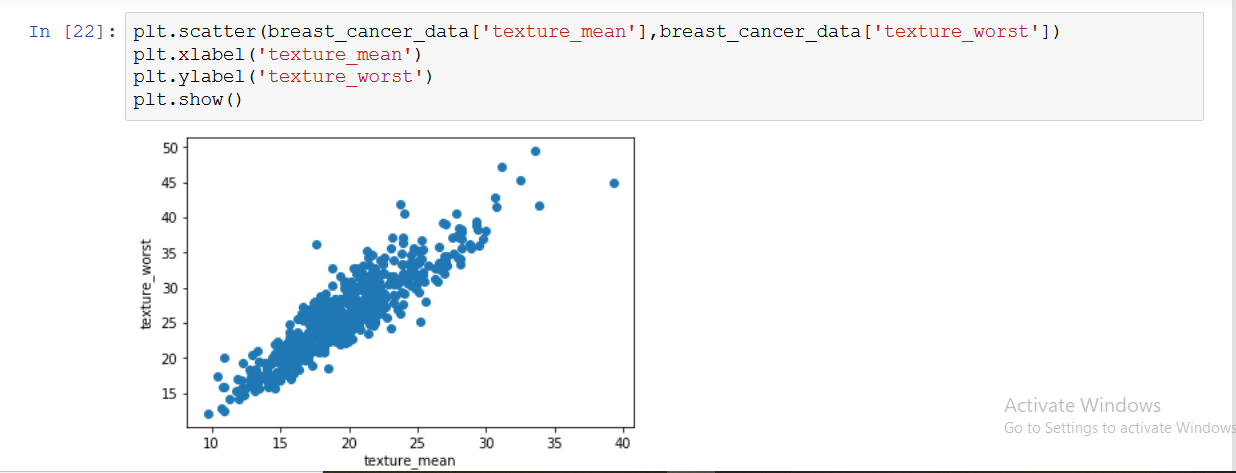


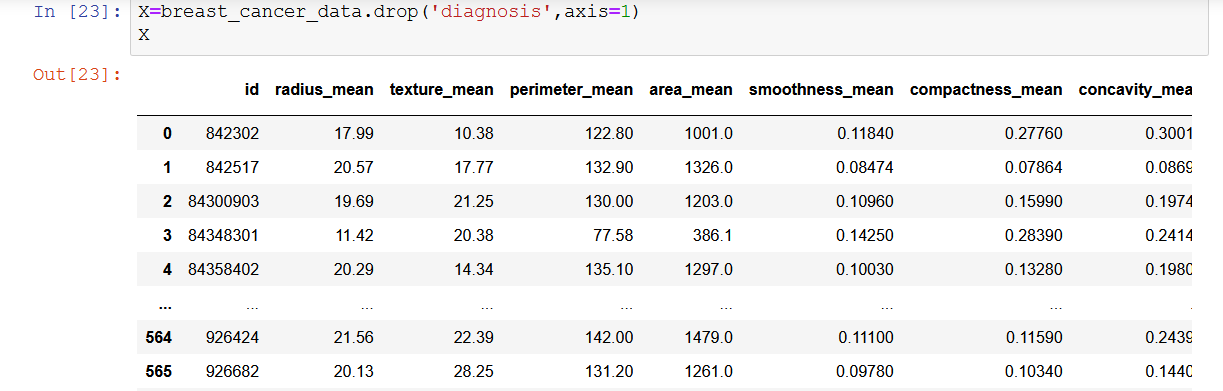


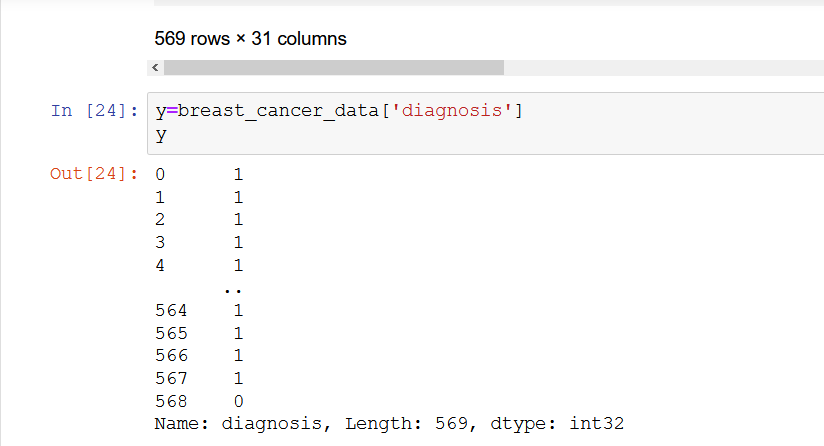


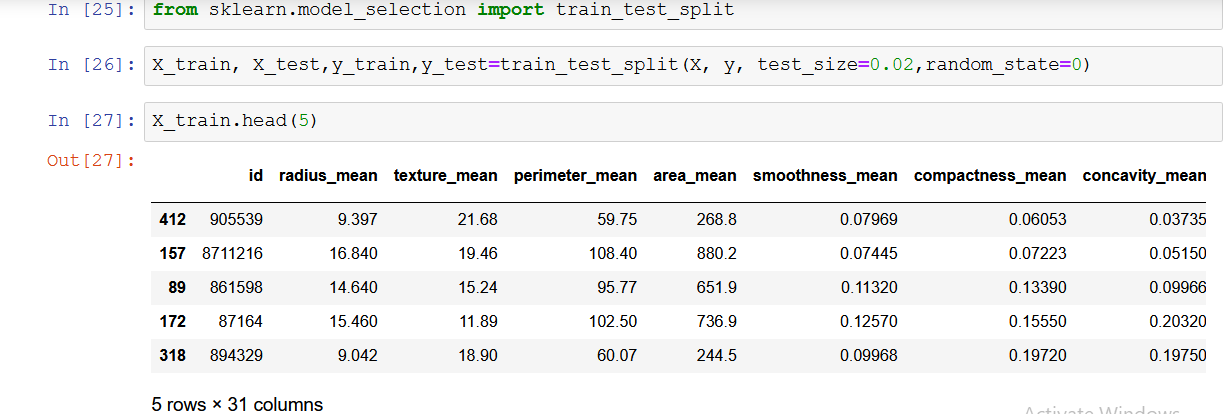


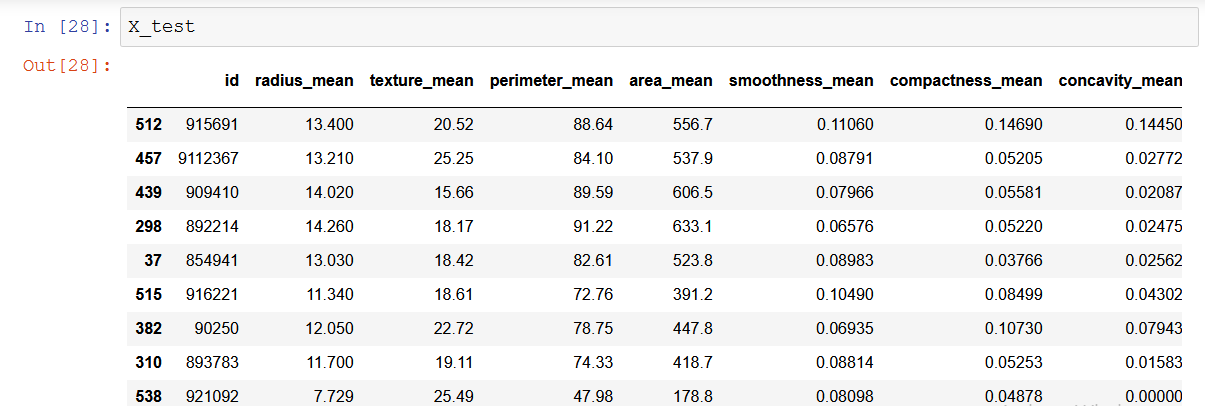


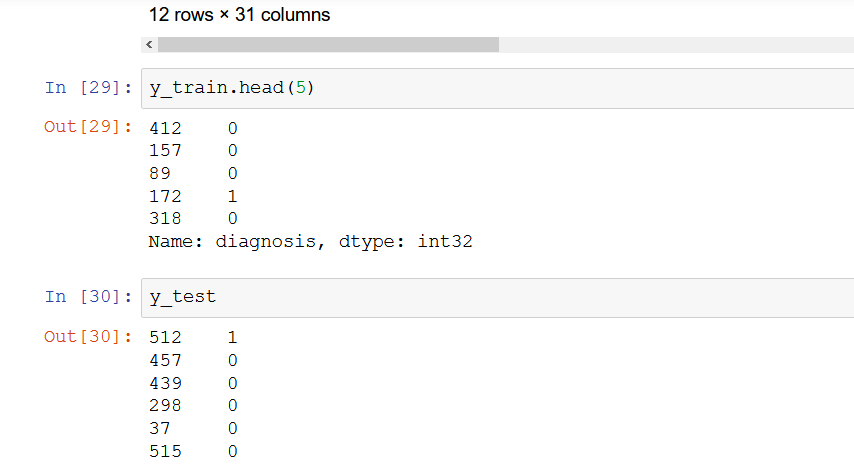


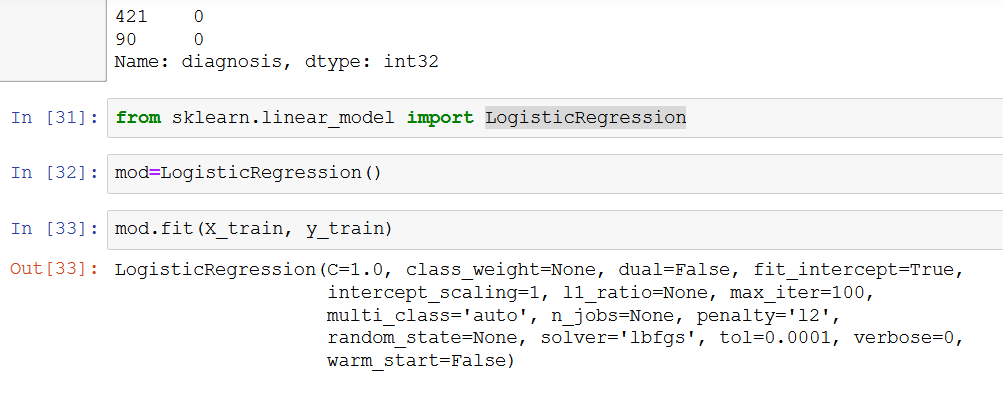


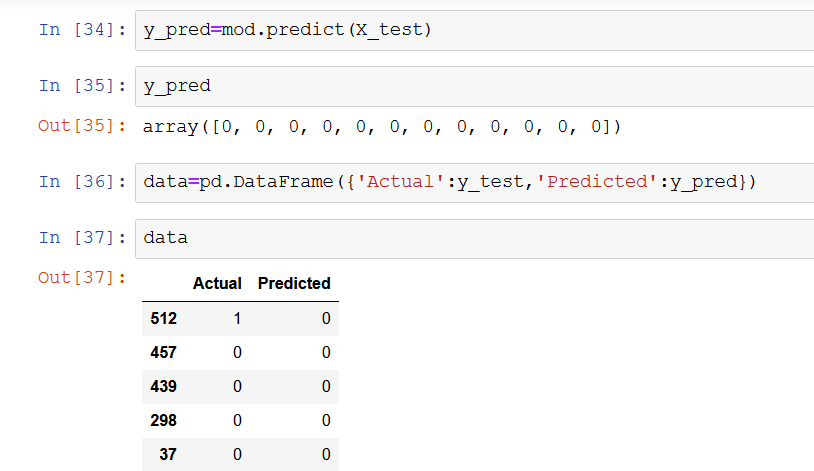


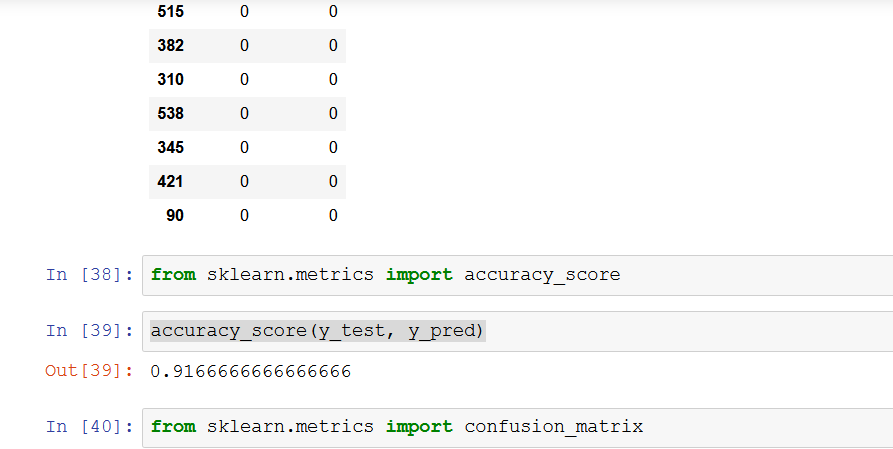


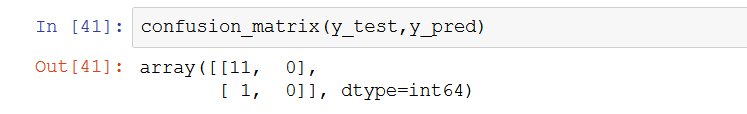














**System**

**Measurement**

### LIMITATIONS

* Since it is an online project, the results depends upon the online data , we have provided in the dataset.
* People who are not familiar with computers can’t create any project of this kind with this data as it requires programming skills along with Machine Learning.

**FUTURE WORK and POSSIBILITIES**

1 ) This types of projects can be very useful for the early detection of Cancers in any individual without any surgical diagnosis .

2) It is a very fast and useful process as it requires very less time so immediate actions can be taken with it.

3) In near future we would like to implement this technology in different hospitals as well as research centers for the advancement of the technology as well as to do the required work faster.

### CONCLUSION

**“*CANCER CELL PREDICTION BY USING MACHINE LEARNING CONCEPTS*”**  is a big success today.

* All the data will be stored in this database. So, the admin can view all the data.
* By this project , all people can know about this survey result.
* It will also help to medical science to collect all types of previous history about cancer cell.

### BIBLIOGRAPHY

* https://www.google.com/
* https://medium.com/@randerson112358/breast-cancer-detection-using-machine-learning-38820fe98982
* <http://localhost:8888/tree>
* Documentation Provided by our Mentor.