**PROJECT**

On

Breast Cancer Prediction

Submitted as a part of Curriculum of

**Bachelor of Technology**

in

**Computer Science and Engineering**

By

Under the guidance of

**Mr. D. KRISHNA**

**Associate professor**

Department of Computer Science and Engineering



**ACE Engineering College (An Autonomous Institution)**

NBA ACCREDITED B. TECH COURSES:

EEE, ECE, CSE & MECH

Ankushapur(V), Ghatkesar(M), Medchal.Dist.–501301

**(Affiliated to Jawaharlal Nehru Technological University Hyderabad 2022-2025)**



**CERTIFICATE**

This is to certify that the project work entitled BREAST CANCER PREDICTION is being submitted by Kalagotla suhas (21AG1A0525), Surrender Adunuri (21AG1A0502), Mende Puja (22AG5A0501), Addepalli Pavanteja (21AG1A0501), Aithagoni Akshitha (21AG1A0503), Gunji Vamshi (21AG1A0522), Kakkera Venkat Sai (21AG1A0524), Gangani Sai Rithvik Goud(21AG1A0520), Gugoloth lohitha (21AG1A0521) as a part of Curriculum of Degree of Bachelor of Technology in Computer Science and Engineering to the ACE Engineering College during the academic year 2021- 2025 is a record of bonafide work carried out by them under our guidance and supervision **.**

**Guide Mr.D.Krishna**

**ASSOCIATE PROFESSORHead of Department**

**Dr .M.V.VIJAYA SARADHI**

**Professor and Head of the Dept CSE**

**ACKNOWLEDGEMENT**

We would like to express our gratitude to all the people behind the screen who have helped us to transform an idea into a real time application. We would like to express our heart-felt gratitude to our parents without whom we would not have been privileged to achieve and fulfil our dreams.

A special thanks to our Secretary, Prof. Y. V. GOPALA KRISHNA MURTHY, for having founded such an esteemed institution. We are also grateful to our beloved principal,

Dr.B. L. RAJU for permitting us to carry out this project. We profoundly thank

Dr.M.V.VIJAYA SARADHI, Head of the Department of Computer Science & Engineering.

We are very thankful to our guide D. KRISHNA, Associate Professor who has been excellent and also given continuous support for the completion of our project work. The satisfaction and euphoria that accompany the successful completion of the task would be great, but incomplete without the mention of the people who made it possible, whose guidance and encouragement crown all the efforts with success. In this context, we would like to thank all the other staff members, both teaching and non-teaching, who have extended their timely help and made our task.

22AG5A0501-MENDE PUJA

21AG1A0501-ADDEPALLI PAVANTEJA

21AG1A0502-ADUNURI SURENDER

21AG1A0503-AITHAGONI AKSHITHA

21AG1A0522-GUNJI VAMSHI

21AG1A0524-KAKKERA VENKAT SAI

21AG1A0525-KALAGOTLA SUHAS

21AG1A0520-GANAGANI SAI RITHVIK GOUD

21AG1A0521-GUGULOTH LOHITHA

**DECLARATION**

We hereby declare that project entitled “Breast Cancer Prediction” submitted as a

part of Curriculum of Bachelor of Technology in Computer Science and Engineering. This dissertation is our original work and the project has not formed the basis for the award of any degree, associate ship, fellowship or any other similar titles and no part of it has been published or sent for publication at the time of submission.

Kalagotla suhas (21AG1A0525)

Surrender Adunuri (21AG1A0502)

Mende Puja (22AG5A0501)

Addepalli Pavanteja (21AG1A0501)

Aithagoni Akshitha (21AG1A0503)

Gunji Vamshi (21AG1A0522)

Kakkera Venkat Sai (21AG1A0524)

Gangani Sai Rithvik Goud(21AG1A0520)

Gugoloth lohitha (21AG1A0521)

**ABSTRACT**

The abstract for a breast cancer prediction project documentation could provide a succinct overview of the project. It may touch upon the use of Python and machine learning techniques for developing a predictive model. The abstract could emphasize the significance of early breast cancer detection, the dataset employed for training and evaluation, and the key performance metrics achieved by the model. Additionally, it might mention the potential impact on healthcare outcomes and the overall goal of contributing to advancements in breast cancer diagnosis.

The main objective for predicting breast cancer is early detection,enabling timely intervention and treatment to improve outcomes and increase the chances of successful recovery.

INDEX

1.Introduction - 1

1.1 Logistic Regression

1.2 Correlation

1.3 Regression Model

2.DataSet

3.Exploratory Analysis

4.Fitting the model

5.Visualising

6.Accuracy

7.Conclusion

8.References

1 Introduction

1.1 Logistic Regression:

Logistic regression estimates the probability of an event occurring, such as voted or didn’t vote, based on a given dataset of independent variables. Since the outcome is a probability, the dependent variable is bounded between 0 and 1. In logistic regression, a logit transformation is applied on the odds— that is, the probability of success divided by the probability of failure. This is also commonly known as the log odds, or the natural logarithm of odds, and this logistic function is represented by the following formulas:

Logit(pi) = 1/(1+ exp(-pi))

ln(pi/(1-pi)) = Beta\_0 + Beta\_1\*X\_1 + … + B\_k\*K\_k

In this logistic regression equation,

logit(pi) is the dependent or response variable

x is the independent variable

The beta parameter, or coefficient, in this model is commonly estimated via maximum likelihood estimation (MLE).

Types of logistic regression

Binary logistic regression: In this approach, the response or dependent variable is dichotomous in nature—i.e. it has only two possible outcomes (e.g. 0 or 1). Some popular examples of its use include predicting if an e-mail is spam or not spam

Multinomial logistic regression: In this type of logistic regression model, the dependent variable has three or more possible outcomes; however, these values have no specified order. For example, movie studios want to predict what genre of film a moviegoer is likely to see to market films more effectively.

Ordinal logistic regression: This type of logistic regression model is leveraged when the response variable has three or more possible outcome, but in this case, these values do have a defined order. Examples of ordinal responses include grading scales from A to F or rating scales from 1 to 5.

1.2 Correlation

Correlation quantifies the direction and strength of the relationship between two numeric variables, X and Y, and always lies between -1.0 and 1.0.

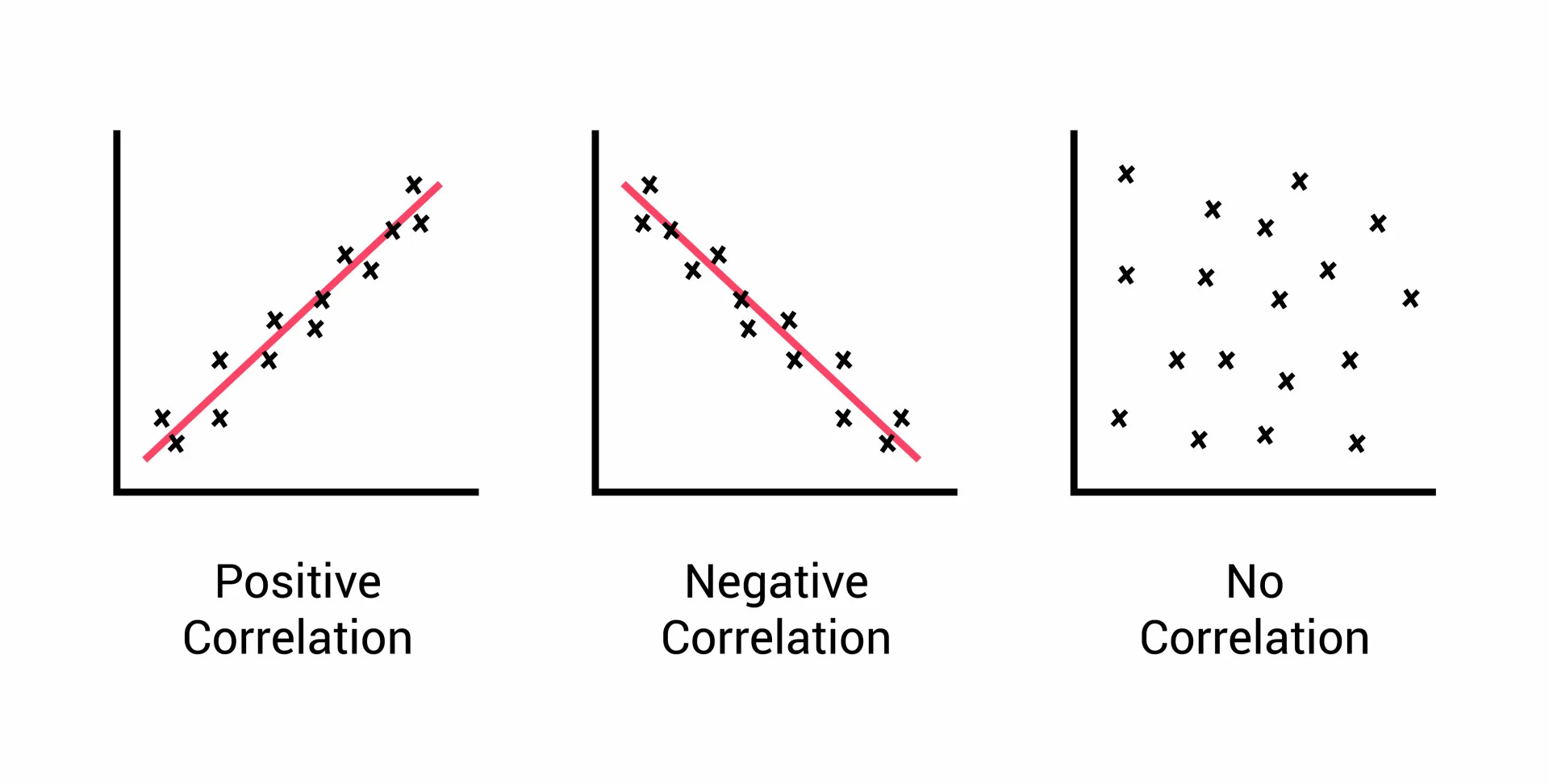
Correlation is simply normalized covariation, and covariation measures how 2 random variables co-variate, that is, how change in one variable is related to change in another one.

For example, there exists a correlation between two variables X and Y, which means the value of one variable is found to change in one direction, the value of the other variable is found to change either in the same direction (i.e. positive change) or in the opposite direction (i.e. negative change). Furthermore, if the correlation exists, it is linear, i.e. we can represent the relative movement of the two variables by drawing a straight line on graph paper.

The below figure depicts the correlation :

 When the correlation (r) is negative, the regression slope (b) will be negative.

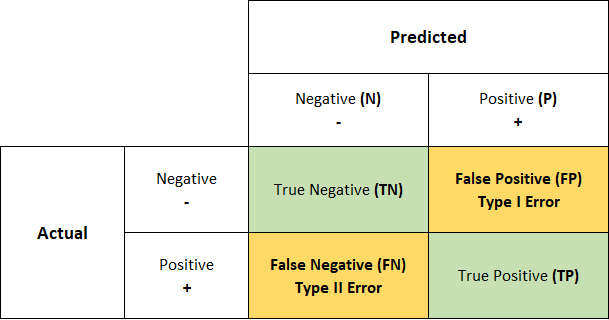
 When the correlation is positive, the regression slope will be positive.



1.3 Regression Model :

Regression analysis is a statistical method to model the relationship between a dependent (target) and independent (predictor) variables with one or more independent variables. More specifically, Regression analysis helps us to understand how the value of the dependent variable is changing corresponding to an independent variable when other independent variables are held fixed. It predicts continuous/real values such as temperature, age, salary, price, etc.

CONFUSION MATRIX:



In [machine learning](https://www.geeksforgeeks.org/ml-machine-learning/amp/), classification is the process of categorizing a given set of data into different categories. In machine learning, to measure the performance of the classification model, we use the **confusion matrix**. Through this tutorial, understand the significance of the confusion matrix.

**What is a Confusion Matrix?**

A **confusion matrix** is a matrix that summarizes the performance of a machine learning model on a set of test data. It is a means of displaying the number of accurate and inaccurate instances based on the model’s predictions. It is often used to measure the performance of classification models, which aim to predict a categorical label for each input instance.The matrix displays the number of instances produced by the model on the test data.

**True positives (TP):** occur when the model accurately predicts a positive data point.

**True negatives (TN)**: occur when the model accurately predicts a negative data point.

**True positives (FP)**: occur when the model predicts a positive data point incorrectly.

**False** **negatives (FN)**: occur when the model mispredicts a negative data point.

### Why do we need a Confusion Matrix?

When assessing a classification model’s performance, a confusion matrix is essential. It offers a thorough analysis of true positive, true negative, false positive, and false negative predictions, facilitating a more profound comprehension of a model’s **recall, accuracy, precision,** and overall effectivenessin class distinction. When there is an uneven class distribution in a dataset, this matrix is especially helpful in evaluating a model’s performance beyond basic accuracy metrics.

2. Data Set

We have considered the Breast Cancer Wisconsin (Diagnostic) Data Set.It contains the information about various women and details about their condition is given among various columns.This dataset is hosted on [Kaggle] https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data. Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image.

n the 3-dimensional space is that described in: [K. P. Bennett and O. L. Mangasarian: "Robust Linear Programming Discrimination of Two Linearly Inseparable Sets", Optimization Methods and Software 1, 1992, 23-34].

Attribute Information:

1) ID number

2) Diagnosis (M = malignant, B = benign)

3-32)

Ten real-valued features are computed for each cell nucleus:

a) radius (mean of distances from center to points on the perimeter)

b) texture (standard deviation of gray-scale values)

c) perimeter

d) area

e) smoothness (local variation in radius lengths)

f) compactness (perimeter^2 / area - 1.0)

g) concavity (severity of concave portions of the contour)

h) concave points (number of concave portions of the contour) i) symmetry

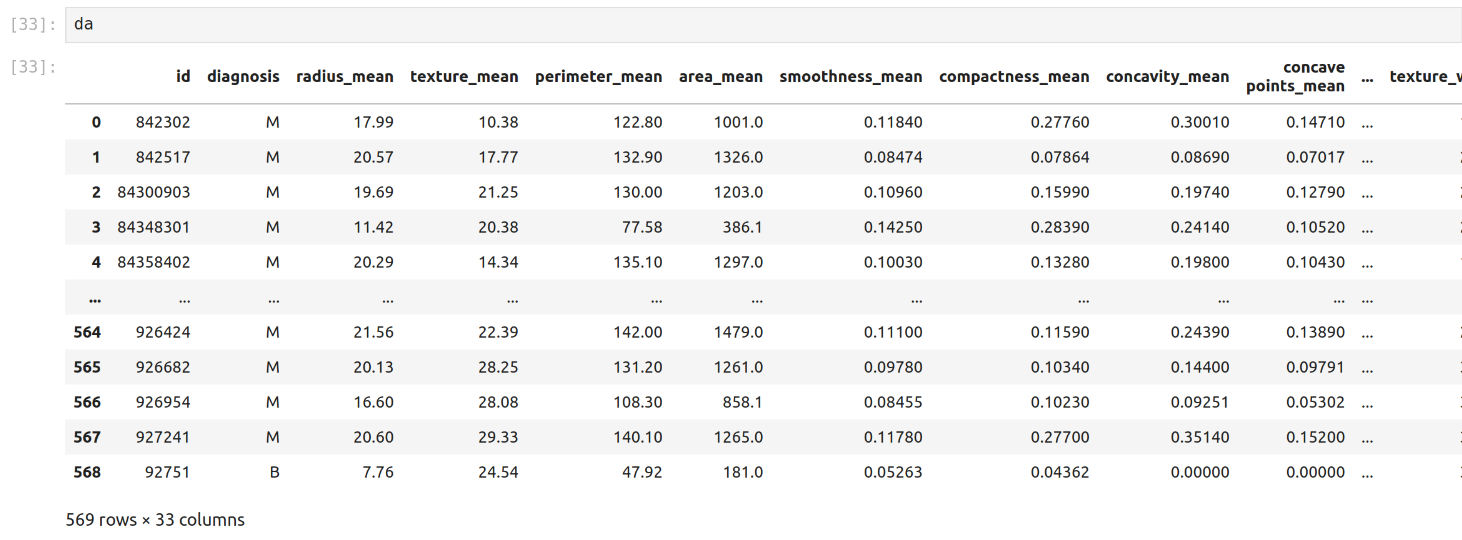
j) fractal dimension ("coastline approximation" - 1)

The mean, standard error and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. For instance, field 3 is Mean Radius, field 13 is Radius SE, field 23 is Worst Radius. All feature values are recoded with four significant digits.

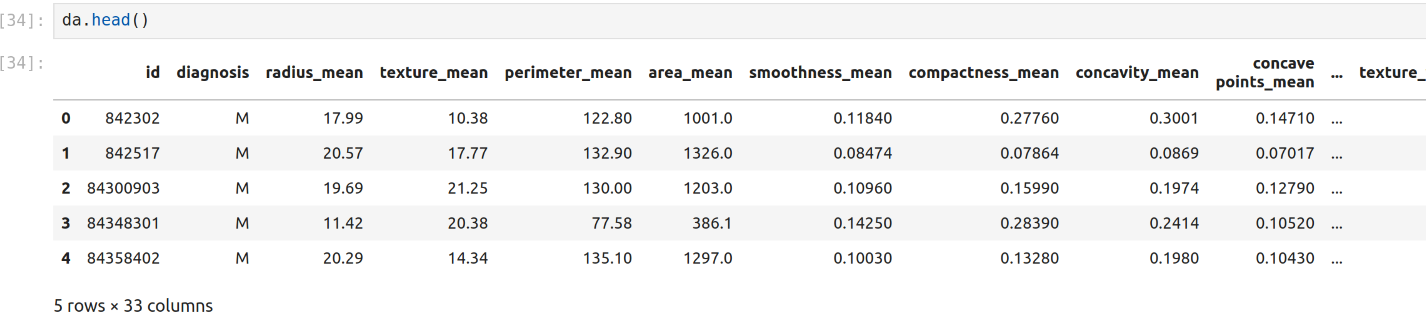
Missing attribute values: none

Class distribution: 357 benign, 212 malignant

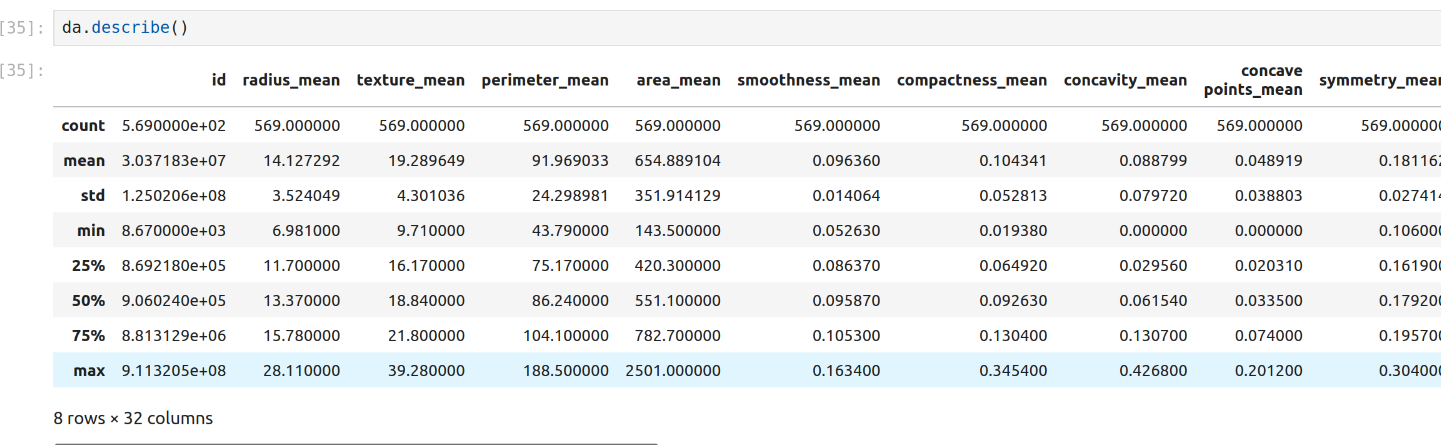
**STRUCTURE OF DATA**

****

**View Dataset**

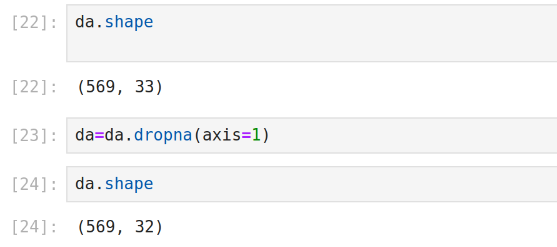
****

describe() summarizes statistical measures

****

3 . Exploratory Analysis :

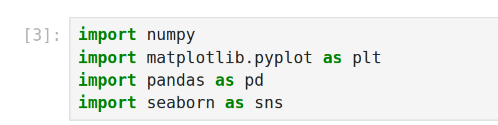
Data Cleaning and Transformation:



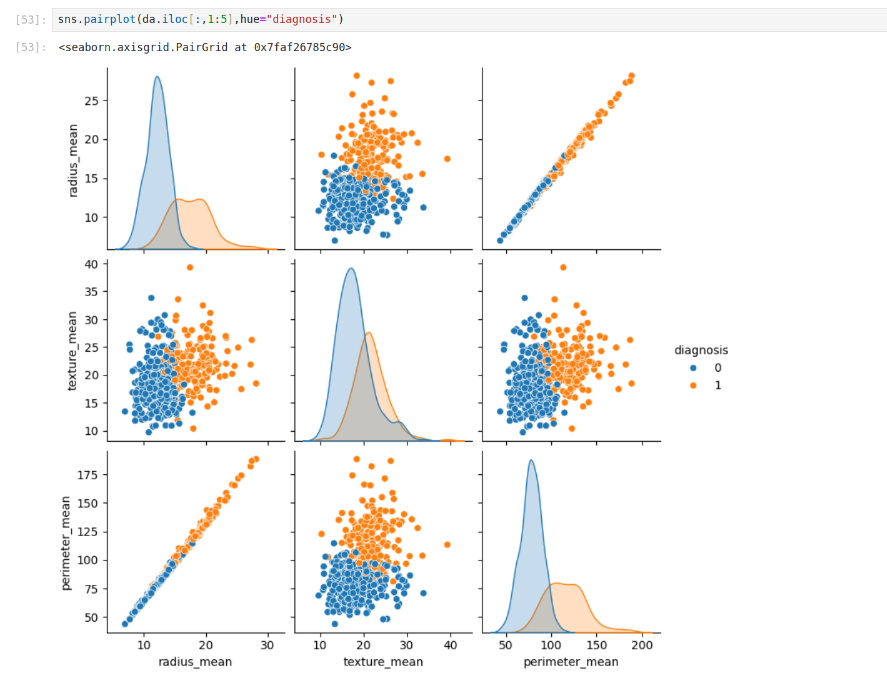
Data cleaning is the process that removes data that does not belong in your dataset. Data transformation is the process of converting data from one format or structure into another.

As the data doesn’t contain any NA values and the columns in the dataset are relevant to our analysis, there is no requirement of data cleaning and transformations.

Loading the libraries:

**PUT LIBRARIES INFO**

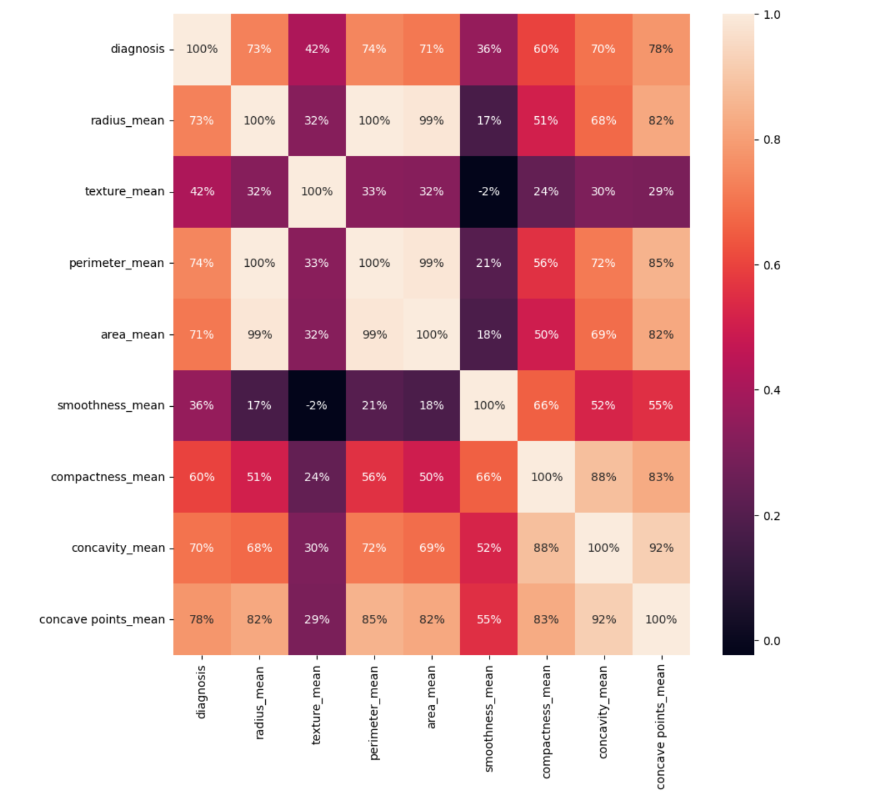
Let’s start by visualizing the distribution of area\_mean for both cancerous and not cancerous tumor cells using boxplots. This will help us see if the means are consistently different between cancerous and non-cancerous tumors.



Fitting the Model

Here’s where we fit our model! Remember, our input variable is the area\_mean and our output variable is the diagnosis.

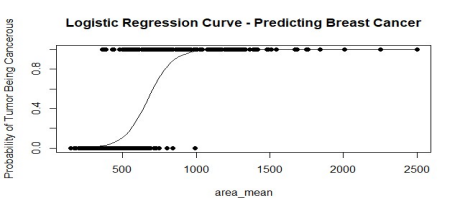
Correlation Matrix



Visualizing the Model

To visualize this simple logistic regression we could make the following plot. plot(diagnosis\_0\_1 ~ area\_mean, data=train,

main="Logistic Regression Curve - Predicting Breast Cancer", ylab='Probability of Tumor Being Cancerous', pch=16)

curve(exp(breast\_cancer\_glm$coef[1]+breast\_cancer\_glm$coef[2]\*x)/ (1+exp(breast\_cancer\_glm$coef[1]+breast\_cancer\_glm$coef[2]\*x)), add=TRUE)

Let’s Make Sure The Model is Valid

Regression model



Accuracy

