MINI PROJECT ON

**BREAST TUMER CLASSIFICATION AS**

**BENIGNE OR MAGLINANT USING NAÏVE BAYES CLASSIFICATION ALGORITHM**

**Submitted by**

Arpit Rohila(J1/44)

2318496

**Submitted to**

Prof. Dr. Ajay Shukla

A logo of a graphic era hill university

Description automatically generated

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

GRAPHIC ERA HILL UNIVERSITY DEHRADUN

JANUARY 2025

Contents:

**Introduction**…………………………………………………………….....................................1

Overview

Background and Methodology

Objective

**Library & Tools Used……………………………………………………………………….2**

**Data set & Statistics .………………………………………………………………………..3**

Correlation

Heatmap

Feature selection

**Splitting and Training………………………………………………………………………4**

Training data

Cross validation data

Estimating the parameters for benign and malignant distributions

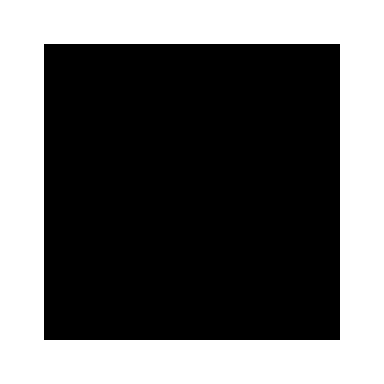
Training the model

**Performance Analysis………………………………………………………………………5**

Confusion matrix

Classification report

Accuracy

 1 Introduction

**Overview**

Diagnosing a breast tumor as cancerous (Malignant) or not (Benign) correctly makes a big difference in life expectancy and quality of life of a patient. Misdiagnosis of a breast tumor can have drastic consequences on the medical condition of a patient. In many cases, quite frequently misdiagnosis of breast tumors happen in the presence of human intervention and medical experts. So, it means that sometimes even medical experts are not able to diagnose breast tumors correctly. Hence, it becomes very important to develop a Machine Learning based system which can diagnose breast tumors correctly at a very high rate, possibly even higher than medical experts. So, given a dataset of breast tumors we must develop a system which will learn to perform correct diagnosis of breast tumors from the same dataset.

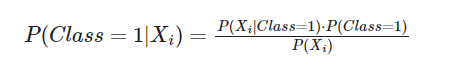
**Background and Motivation**

Diagnosing a breast tumor at high level of accuracy more than a medical expert and without any medical expert by just uploading the picture of the tumor . It can cause a significant difference in diagnosis of tumor at right time .

A tumor can contain around 32 parameters based on that our model try to measure these parameters and by measuring them we are able to classify weather the given tumor is benign or Malignant .

Since we have categorical data in the dimensions of diagnosis as Benign or Malignant so we can use the classification algorithm , for the purpose of tumor classification we are going to use Naïve Bayes Classification algorithm .

Naïve Bayes Classification Algorithm is an extension of conditional probability of an event given that the features .

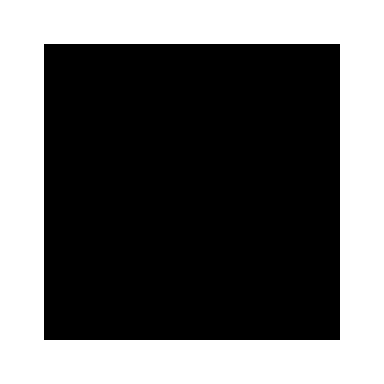


**Objective**

The aim of the model is to classify the class of tumor as benign or malignant by estimating the parameter essential for /which has most classification the power .

For training purpose , the data set <https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic>

Is used to train our model with 75% of data set is used for training purpose and 25% of data set is used for cross validation purposes.

 2 Library & Tools used

**Tools**

For Implementation purpose , an open source tool **python** version 3.12.5 is used .



**Jupyter notebook** version 7.2.2 is used as environment for implementation



**Libraries**

**pandas** for loading and manipulating the Dataframe

**numpy** for matrix calculations and linear algebra

**seaborn** for visualization the statistical results

**sicpy.stats** for calculations of statistical results

**matplotlib.pyplot** for plotting the graphs , histograms etc.

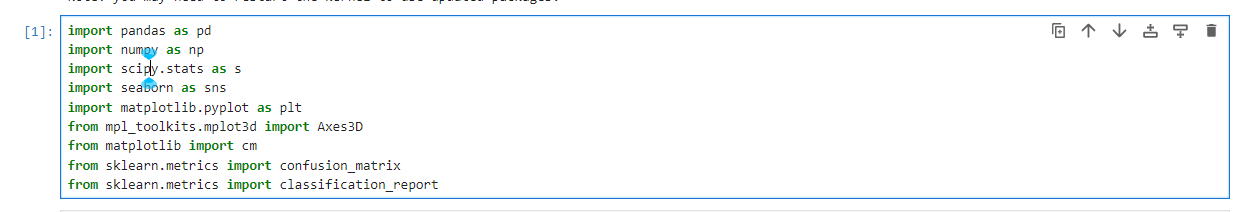
**matplotlib** for plotting the distributions(frequency , p+robablity)

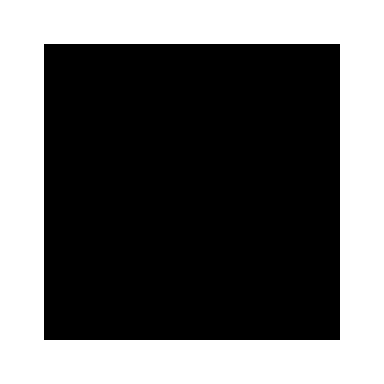
**sklearn.metricies**

confusion\_matrix

classification\_report

for evaluating the performance of model



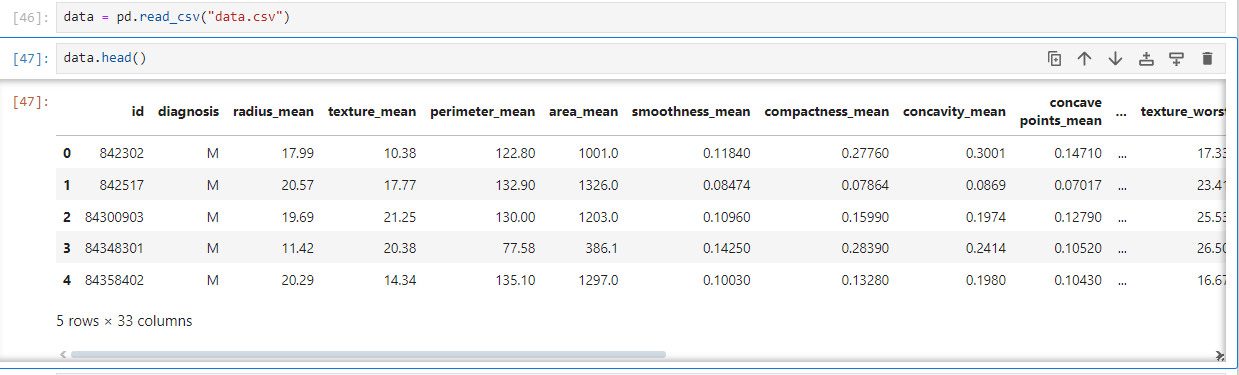
 3 Data set & Statistics

**Overview**

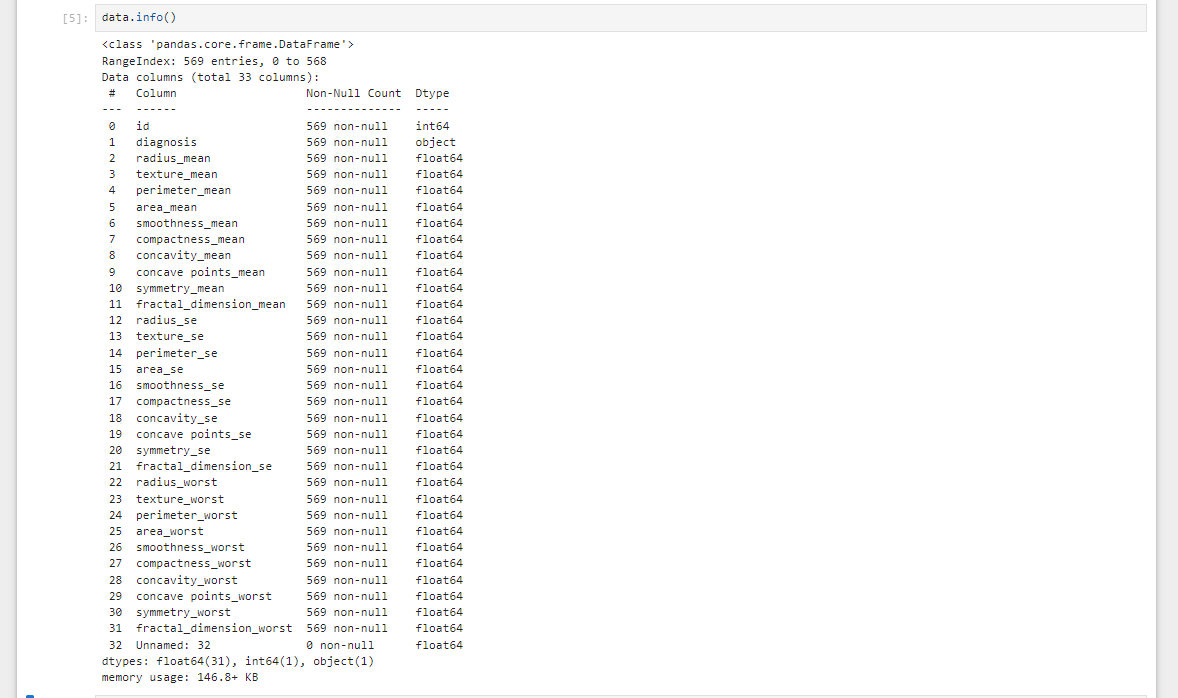
To train our model , the data set <https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic> Is used .

This data set contains 33 columns as the features and 569 row as the different inputs.

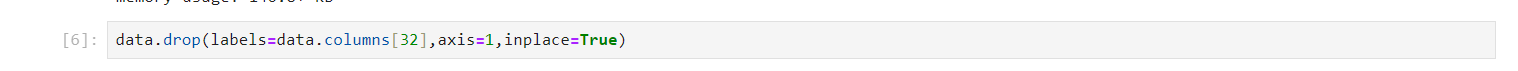
Let’s open the data set and look at that.



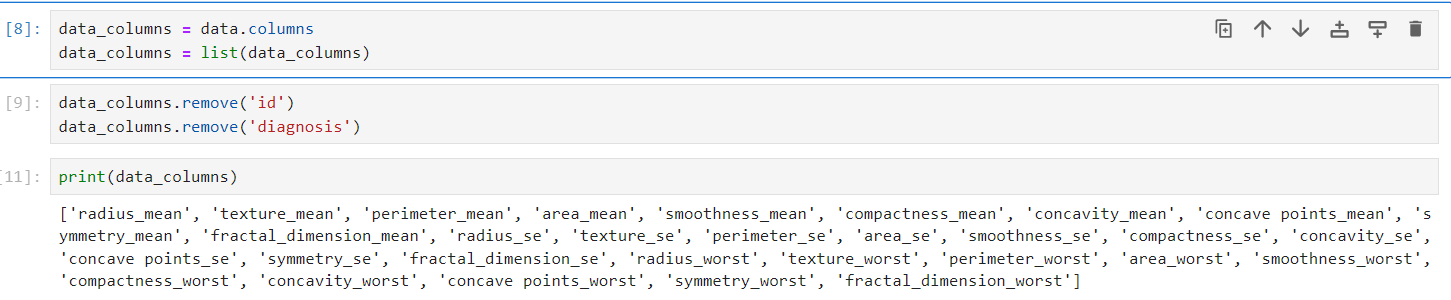
Lets check the information of our data set



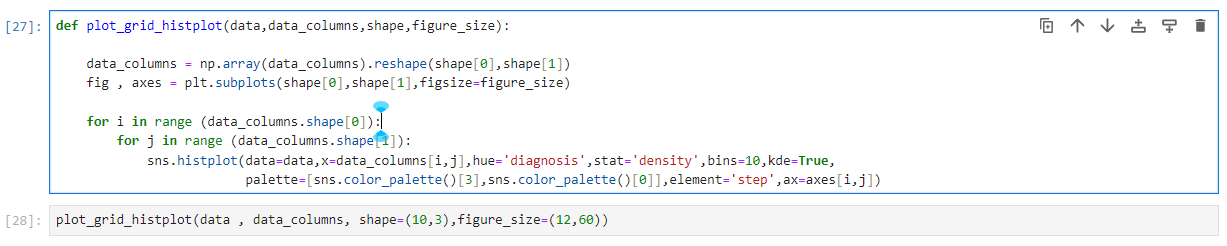
Since the data set contains 33 columns and 33th column having 0 non-null values so we have to drop it.

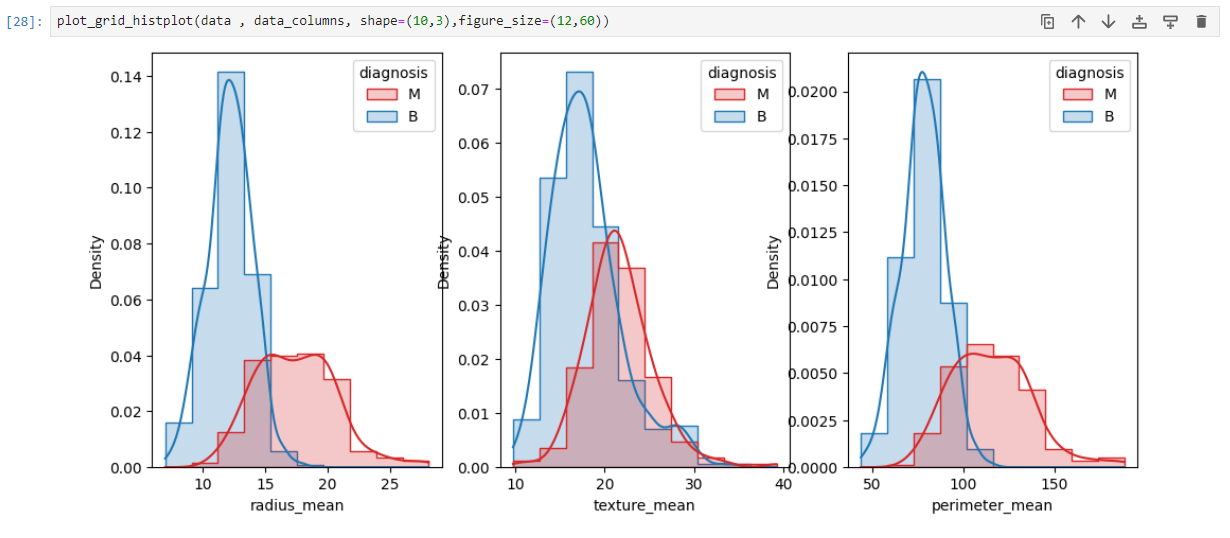


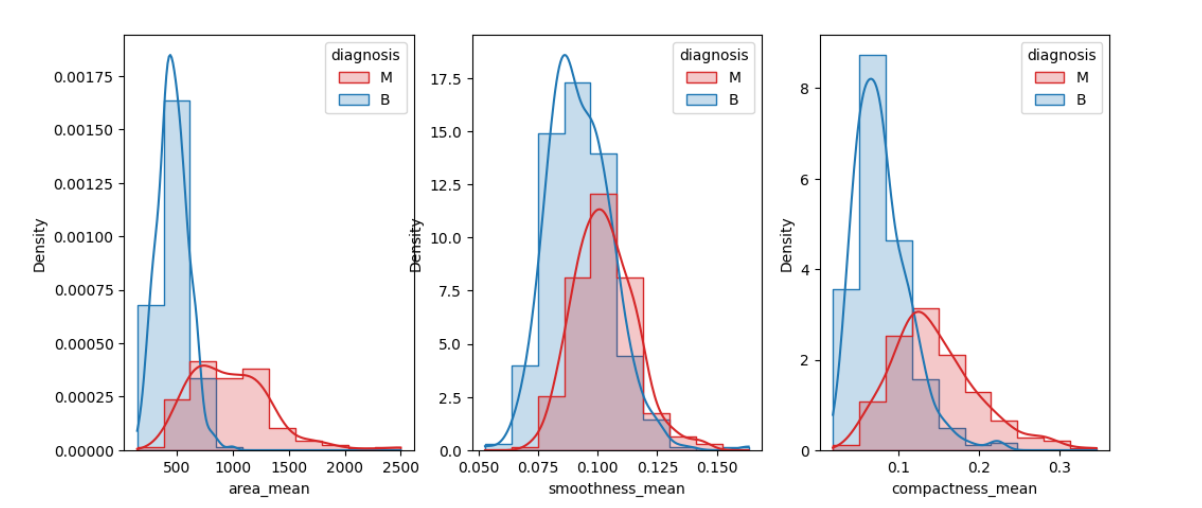
To work independently with the data set and features lets extract all the features into the list and as we can see the parameters ‘id’ and ‘diagnosis’ have different data type among all the data set so we need to remove these parameters also.

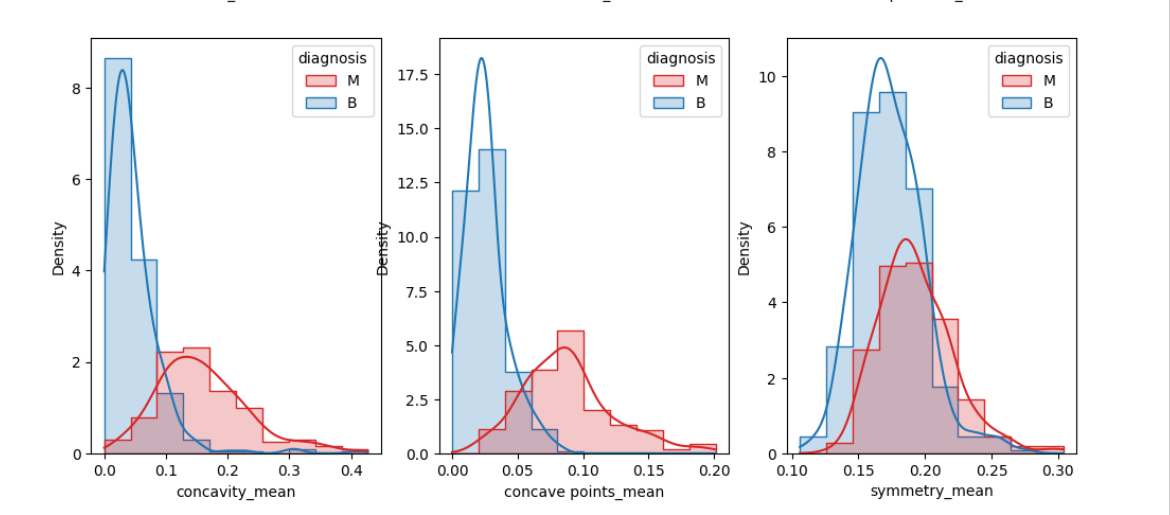
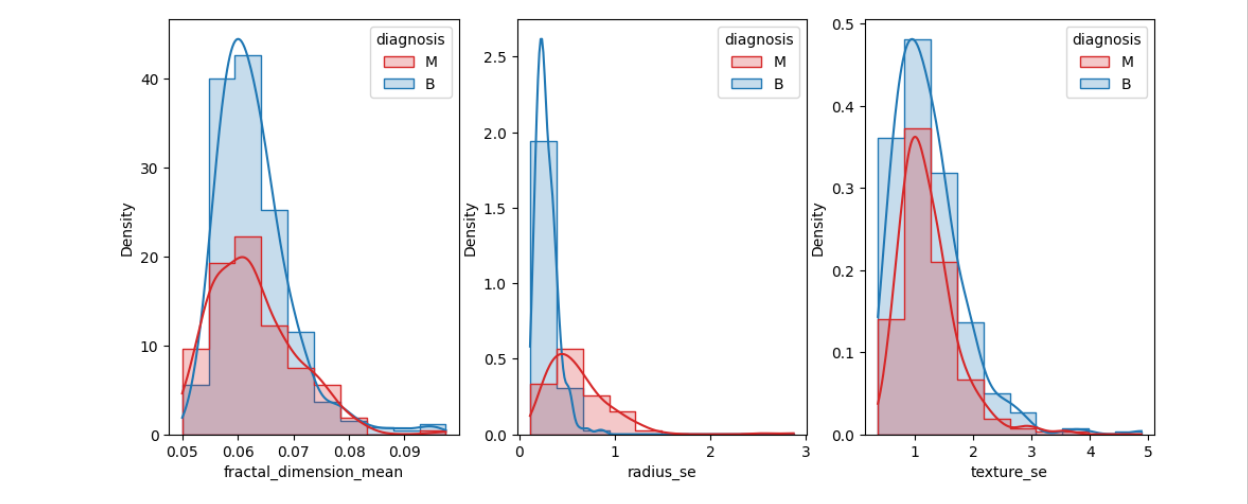


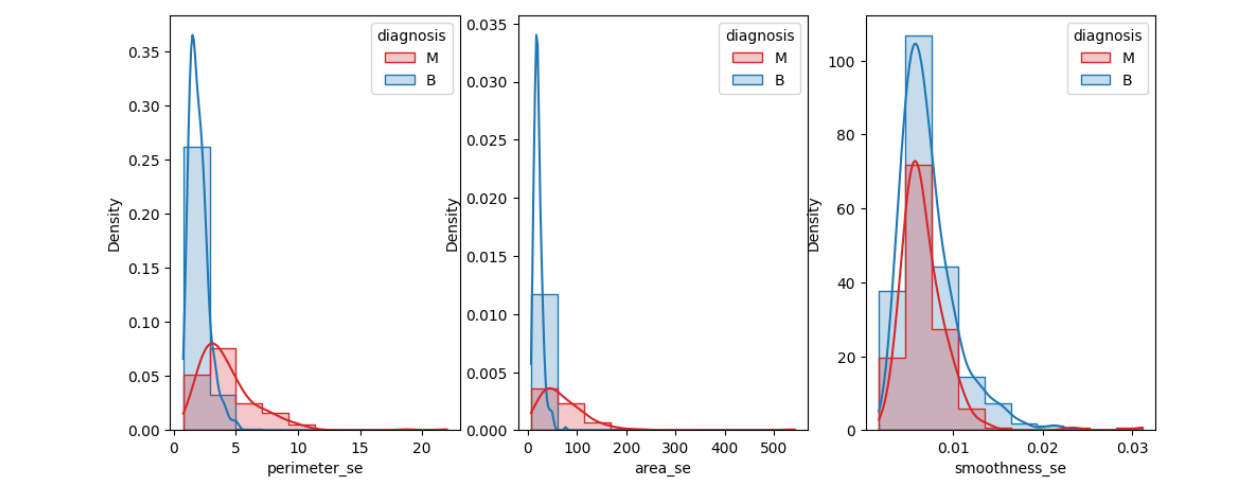
Let's try to plot P(X|Class=Benign) as well as P(X|Class=Malignant) for different features one by one for X equals to different features in the data\_columns list.

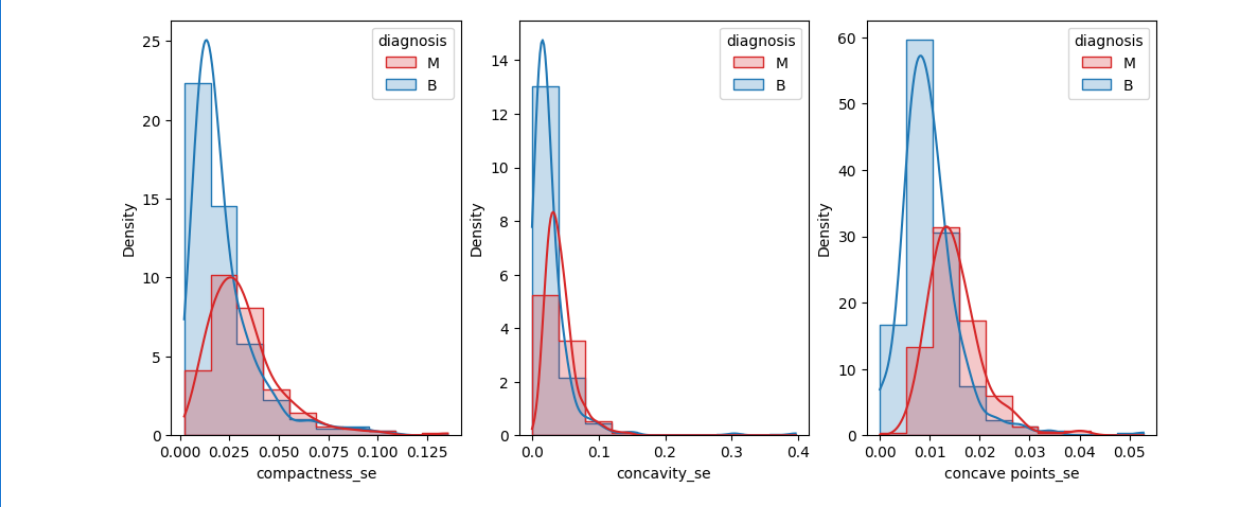


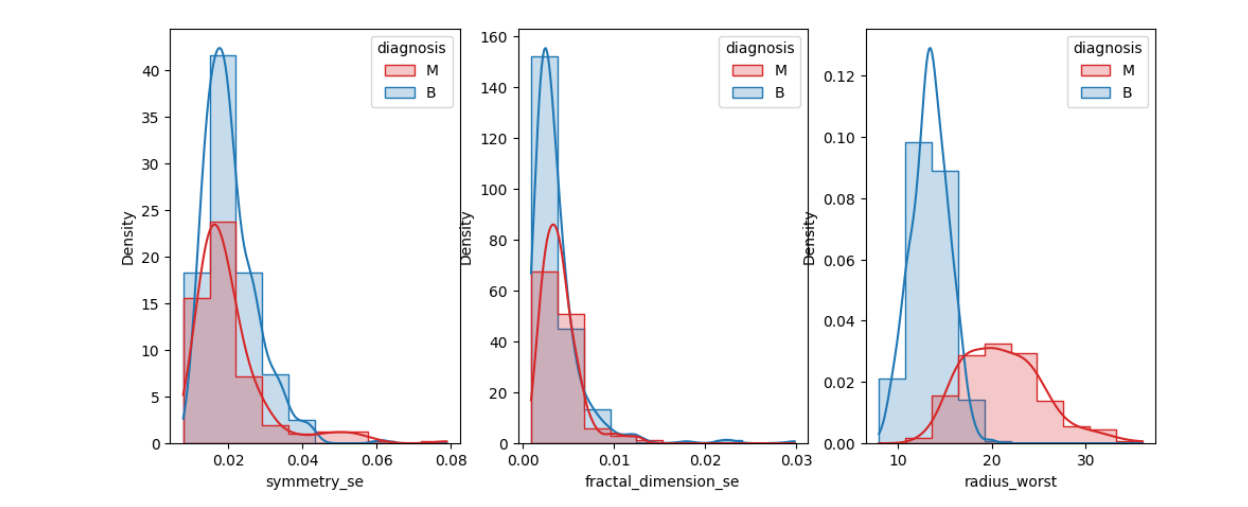


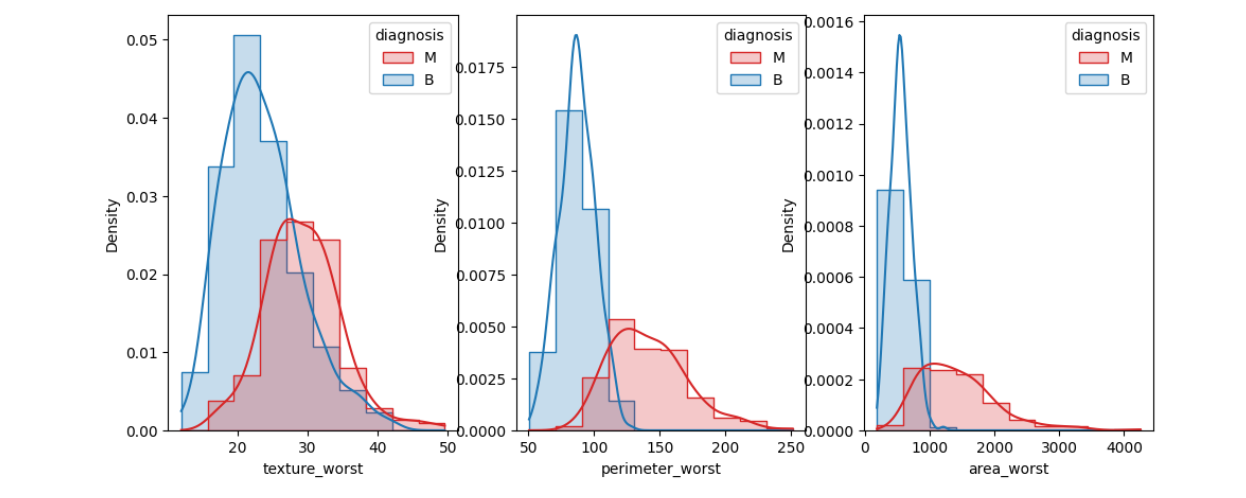


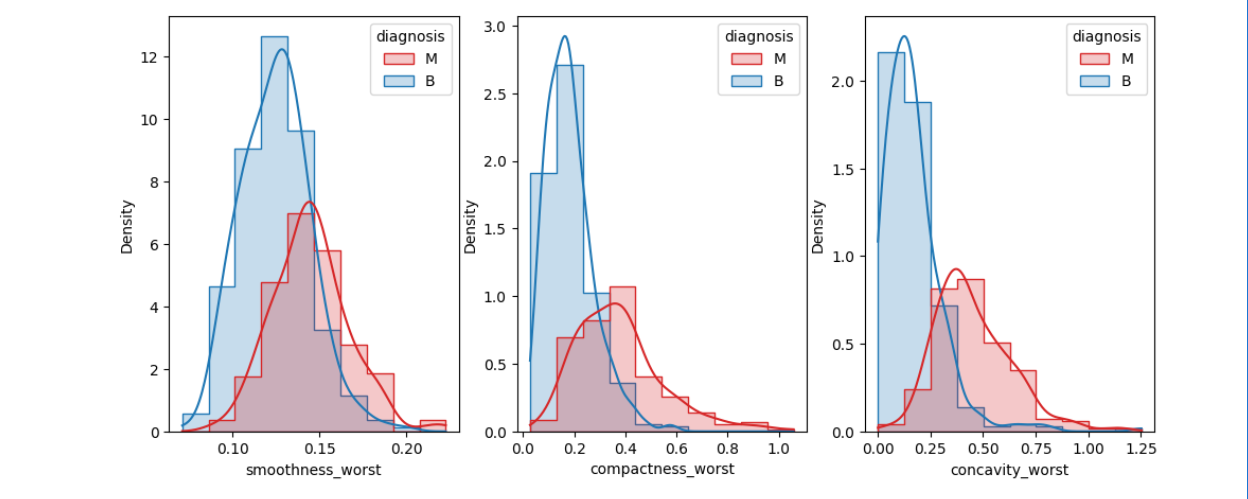


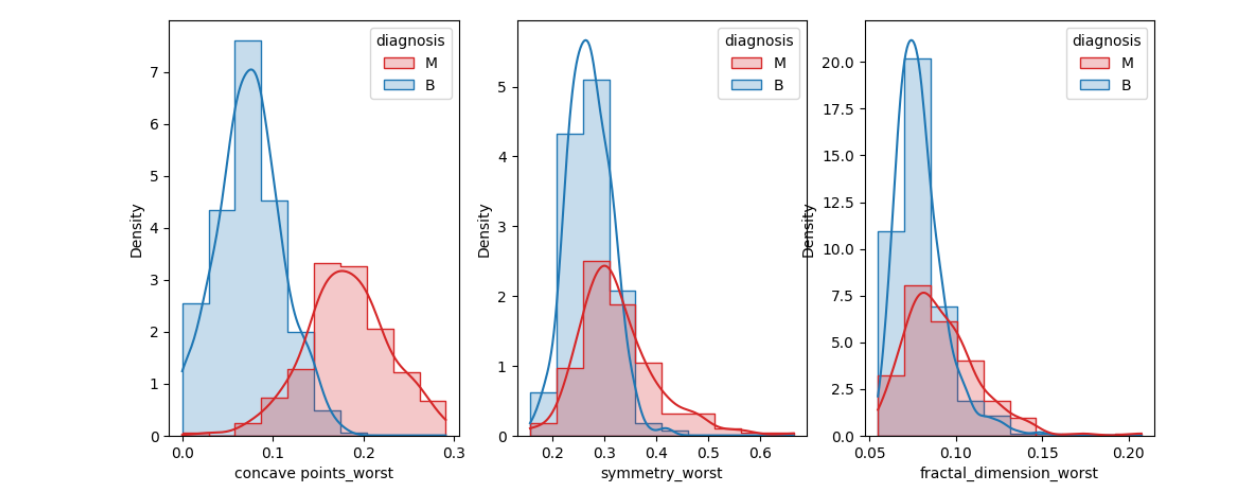












So, we have to select only those features for which the two frequency distributions are overlapping as minimum as possible.

If seen above then the for the following features, two frequency distributions are overlapping as minimum as possible:

**radius\_mean**

**perimeter\_mean**

**area\_mean**

**concavity\_mean**

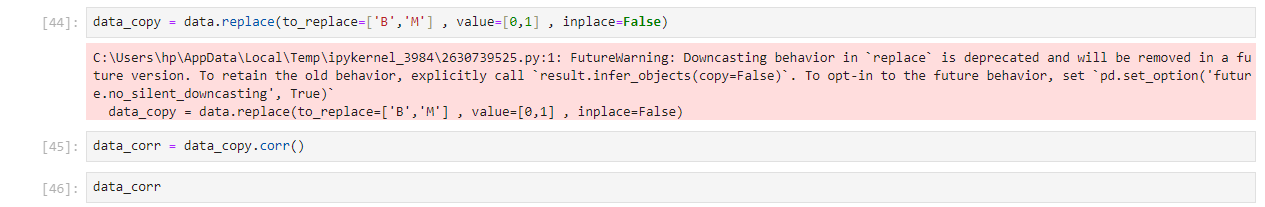
**concave\_points\_mean**

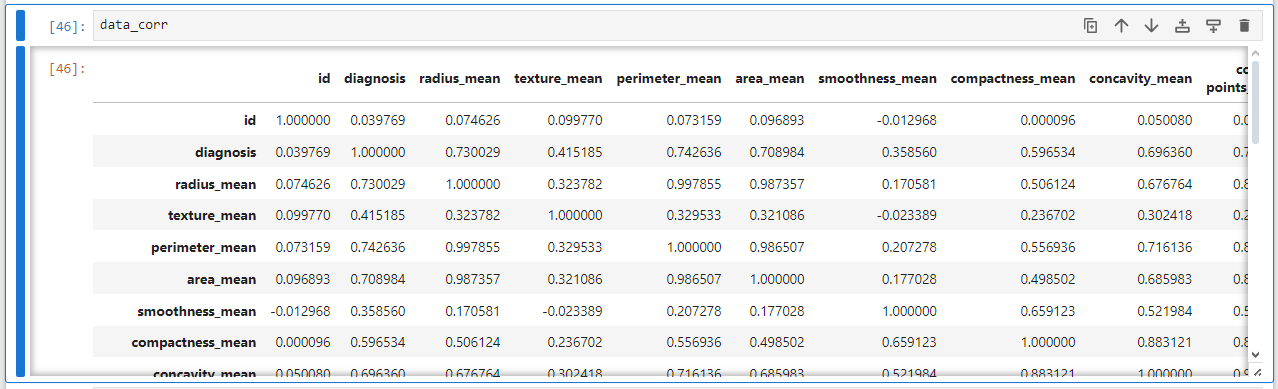
**radius\_worst**

**perimeter\_worst**

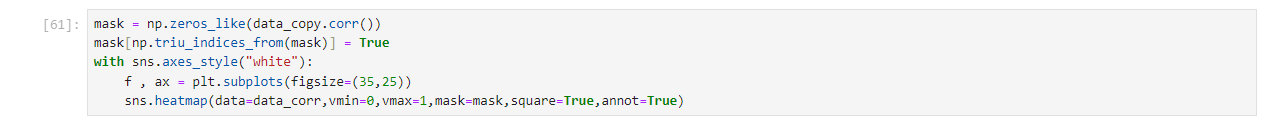
**area\_worst**

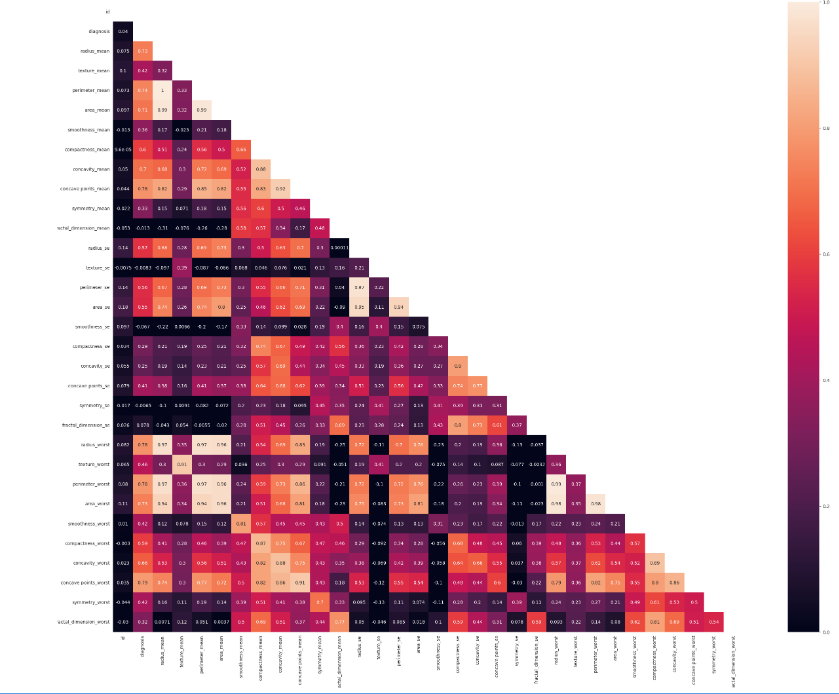
Let's try to visualize the frequency distributions for these individual features:





Let's try to verify our findings with the help of another plot called heatmap:





### The values at each of the cell are given by Pearson Correlation Coefficient. Pearson Correlation Coefficient is computed as:[¶](http://localhost:8888/notebooks/Downloads/Complete_Visualization_Naive_Bayes_Classification_industry_demo.ipynb#The-values-at-each-of-the-cell-are-given-by-Pearson-Correlation-Coefficient.-Pearson-Correlation-Coefficient-is-computed-as:)

### rX1,X2=Cov(X1,X2)σX1⋅σX2

### Where rX1,X2 = Pearson Correlation Coefficient

### Cov(X1,X2) = Covariance between features X1 and X2 which is given by:

### Cov(X1,X2)=∑i=1N(x1i−μX1)⋅(x2i−μX2)N

### σ1 = Standard Deviation of feature X1

### σ2 = Standard Deviation of feature X2

### These values are arranged in the form of a matrix called Pearson Correlation Matrix given as:

### 

### One interesting property regarding Pearson Correlation Coefficient is that:

### −1≤rXm,Xn≤+1

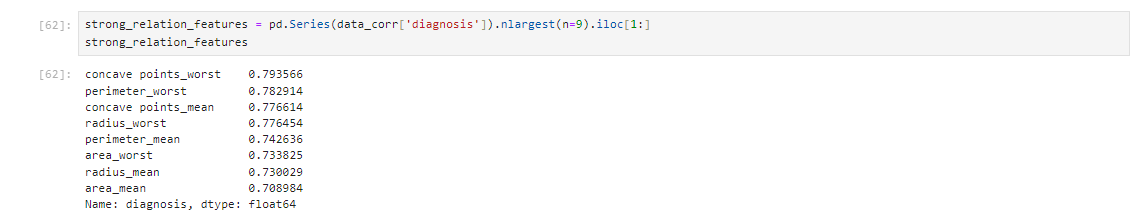
### Where the value of -1 imples that there is strong negative relationship between features Xm and Xn and on the other hand the value of +1 implies that there is a strong positive relationship between them. Whereas, the value close to 0 irrespective of the sign implies that there is no relationship between the two features.

### rXm,Xn is sometimes also called normalized covariance.[¶](http://localhost:8888/notebooks/Downloads/Complete_Visualization_Naive_Bayes_Classification_industry_demo.ipynb#rXm,Xn-is-sometimes-also-called-normalized-covariance.)

### Darker the color in a cell of the heatmap, weaker the relationship between the two features located at the row and column whose intersection is resulting in that cell position. Brighter the color in a cell, stronger the relationship.

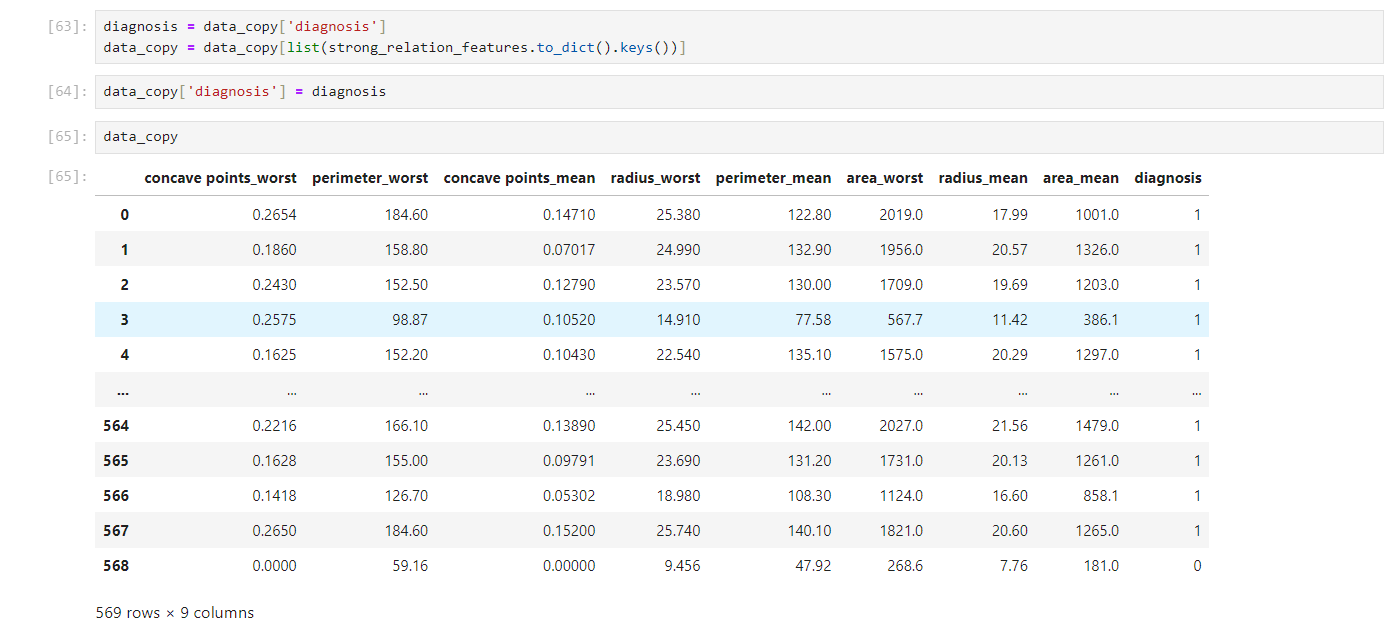
### If carefully observed then the features which were having a good classification power are also the ones which have high magnitude of pearson correlation coefficient with the 'diagnosis' feature so it can be said that they are the ones on whom the decision of a tumor being classified as Benign or Malignant strongly depends.[¶](http://localhost:8888/notebooks/Downloads/Complete_Visualization_Naive_Bayes_Classification_industry_demo.ipynb#If-carefully-observed-then-the-features-which-were-having-a-good-classification-power-are-also-the-ones-which-have-high-magnitude-of-pearson-correlation-coefficient-with-the-'diagnosis'-feature-so-it-can-be-said-that-they-are-the-ones-on-whom-the-decision-of-a-tumor-being-classified-as-Benign-or-Malignant-strongly-depends.)

**Let's fetch the features which have strong relationship with 'diagnosis' feature.**

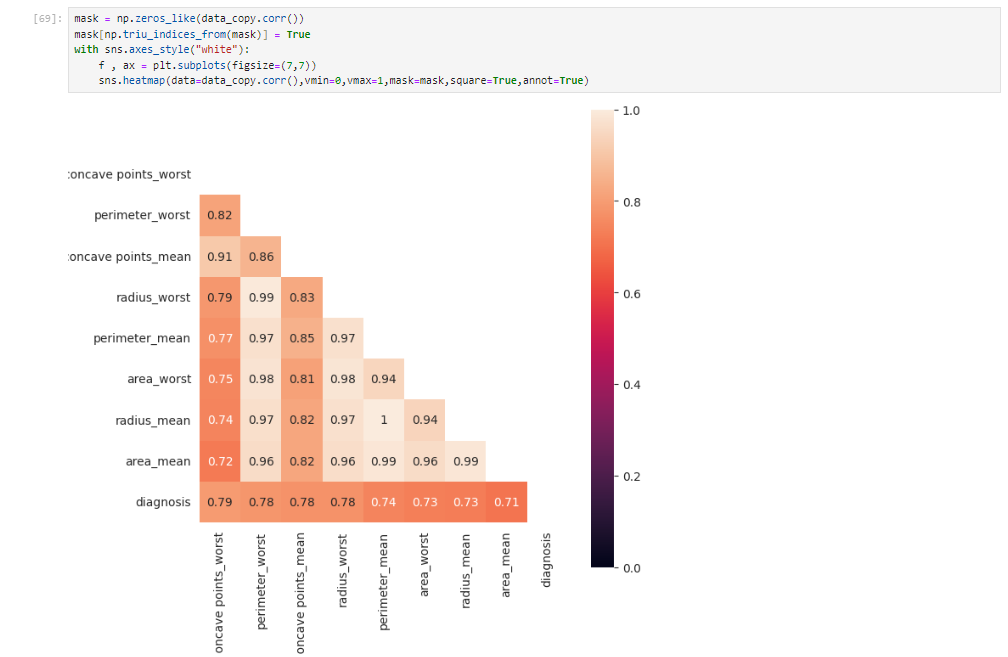
****

It can be observed that most of features fetched here are in common with what we inferred from the probability distribution plots.

Therefore, the top eight features will be selected in our dataset and rest of the features will be discarded from our dataset. Now, let's select these top eight features from our dataset along with the 'diagnosis' column and let's observe the heatmap of them once again.



Now, lets have a look on the heatmap of the updated dataframe having top eight features along with the labels (diagnosis)



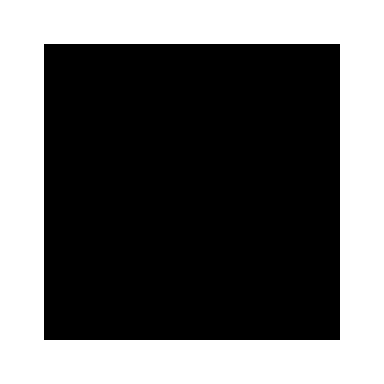
Another interesting thing we can observe here is that definitely these top eight features have a good correlation with diagnosis but these top eight features have even higher correlation among themselves also. And if we compute the covariance matrix or correlation matrix in this case then probably, our covariance matrix is going to be singular.

Let's see whether this matrix is singular or not. So, in order to check whether it is singular or not, we should check it's determinant because the determinant of the singular matrix is always zero.

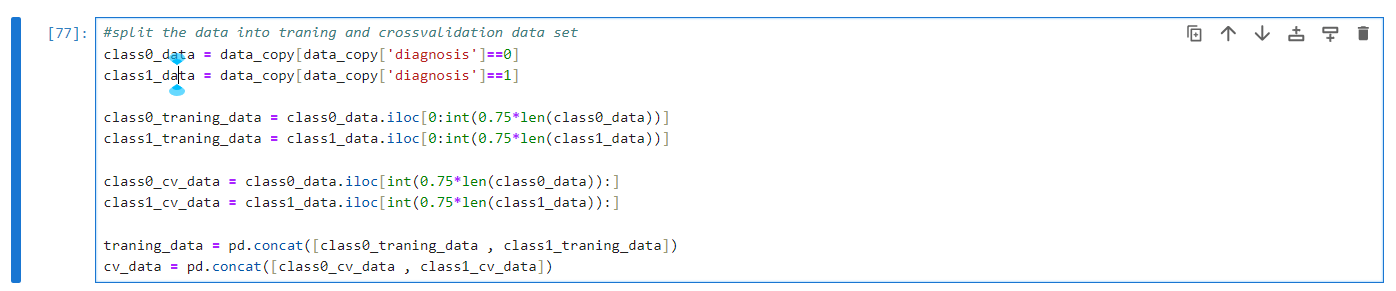


Let's see that how many benign tumors (negative class, 0 class) and malignant tumors (positive class, 1 class) are there in our data ?

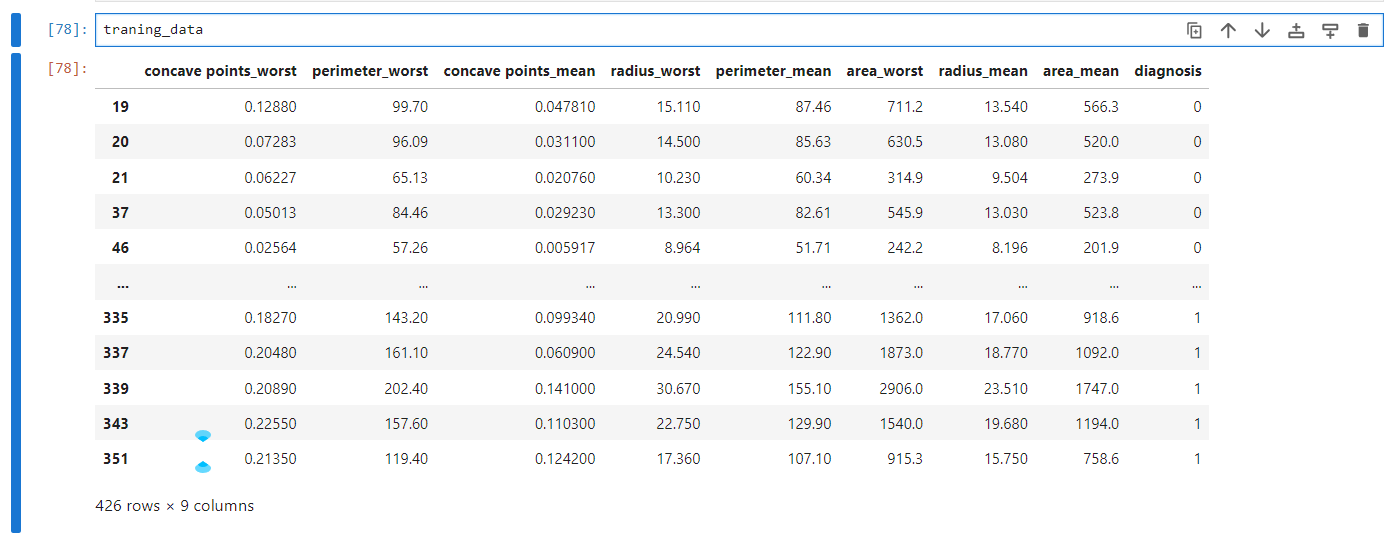


 4 Splitting & Training

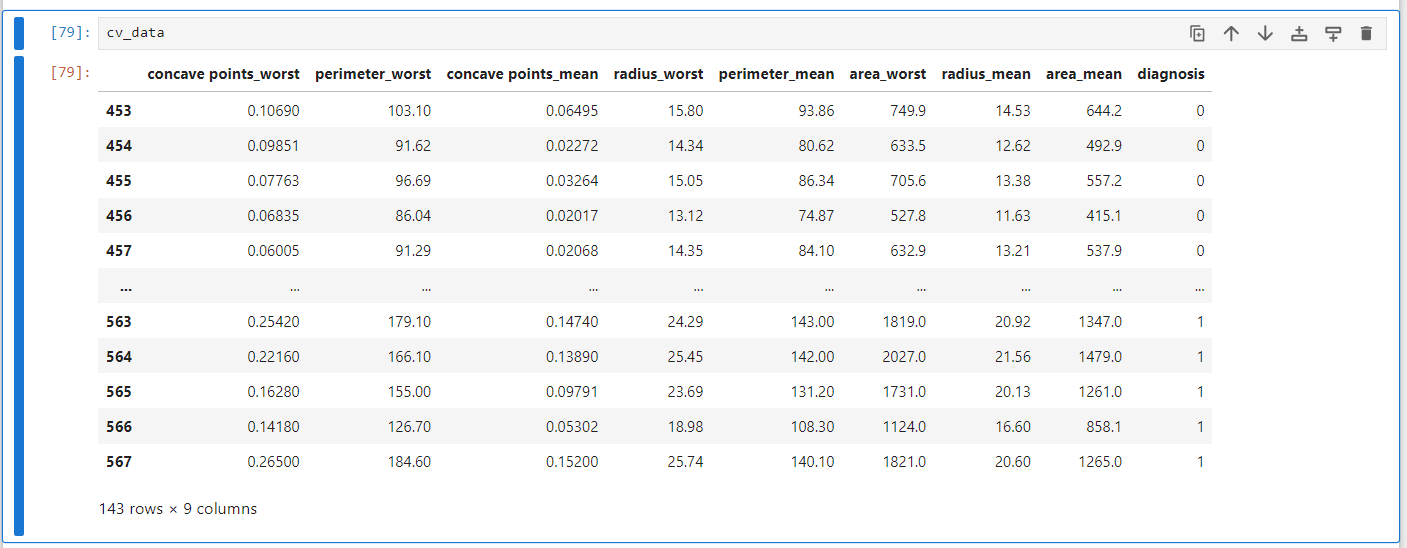
As equal number of both tumors are required for our training data therefore, we will take out 75 % of bening as well as malignant tumors in our training data and put the rest of the tumors inside Cross Validation Data.



Training data



Cross Validation Data



**Now, we will train Naive Bayes Classifier over our training data.**

So, now the question arrises is that how Naive Bayes Classifier works ? Well, one thing is obvious that Naive Bayes Classifier must be working on Bayes Rule which is given as:

P(A|B)=P(B|A)⋅P(A)P(B)

Now, let's first ask ourselves this question that how Naive Bayes Classifier works ?

Here we are trying to solve a classification problem where we want to classifiy the tumor as either belonging to class 0 (benign) or belonging to class 1 (malignant).

So what we are going to do in Naive Bayes Classifier is that first of all we would want to compute this probability:

P(Class=1|Features of that specific tumor) or P(Class=1|Xi) where Xi is the feature vector which will be one or eight dimensional according to our dataset, given as:

Xi=[xi1xi2xi3xi4xi5xi6xi7xi8]

And if P(Class=1|Xi)>0.5 then the classifier will classify the tumor as Malignant (belonging to class 1) otherwise, it will classify the tumor as Benign (belonging to class 0).

But now the question arrises is that how we are going to compute P(Class=1|Xi) for a specific tumor for which we want to perform classification.

The answer to the above question is pretty simple. Let's say that P(Class=1|Xi) is same as P(A|B) in our Bayes Rule. So, if we rewrite Bayes Rule now in context of our classification problem then it will be given as:

P(Class=1|Xi)=P(Xi|Class=1)⋅P(Class=1)P(Xi)

As can be clearly seen that in order to compute P(Class=1|Xi), the three probabilities on the Right Hand Side of the equation need to be computed**.**

Before proceeding further, let's first try to know and understand what these four probabilities are in Bayes Rule:

P(Class=1|Xi) is POSTERIOR PROBABILITY

P(Xi|Class=1) is LIKELIHOOD

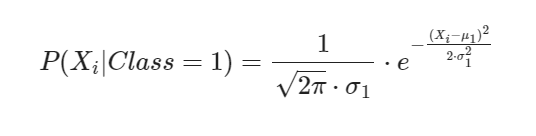
P(Class=1) is PRIOR

P(Xi) is NORMALIZING PROBABILITY or EVIDENCE

**Now, let's compute the Likelihood P(Xi|Class=1). Here, we can have two cases:**

**1. Xi is single dimensional**

As we can observe that all the eight features of any tumor are dense and not sparse (that is continuous and not discrete) and if we consider Xi to be single dimensional (any one of these eight features) then the probability of occurence can be computed using the PDF (Probability Density Function) of Univariate Gaussian Distribution. Hence, the likelihood, P(Xi|Class=1) is given as:



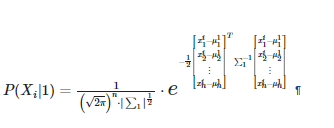
**2. Xi is multidimensional**

Now, if Xi is multidimensional as in eight dimensional then the likelihood, P(Xi|Class=1) can be further written as:

P(xi1∩xi2∩xi3∩xi4∩xi5∩xi6∩xi7∩xi8|Class=1)

As we can observe that all the eight features of any tumor are dense and not sparse (that is continuous and not discrete) and hence the joint probability of their occurence can be computed using the PDF (Probability Density Function) of Gaussian Distribution and it will be specifically Multivariate Gaussian Distribution because there are eight variables (multivariate).

Therefore, the likelihood, P(Xi|Class=1) is given as:

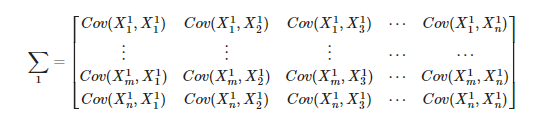


This is the PDF of generic n-dimensional Multivariate Gaussian Distribution for computing the probability of occurence of n dimensional vector observation Xi on a condition that the tumor belongs to Class 1 or Malignant category. The parameters of this distribution are n-dimensional mean vector, given as:

A math equation with numbers and symbols

Description automatically generated with medium confidence

and (n,n) Covariance Matrix ∑1 given as:



Therefore the above PDF can be rewritten as:

A mathematical equation with numbers and symbols

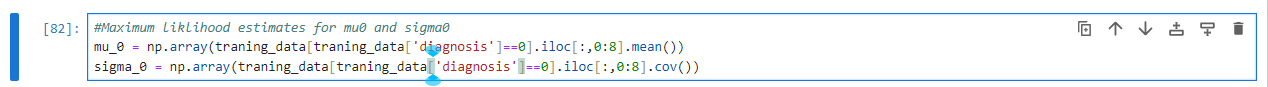
Description automatically generated

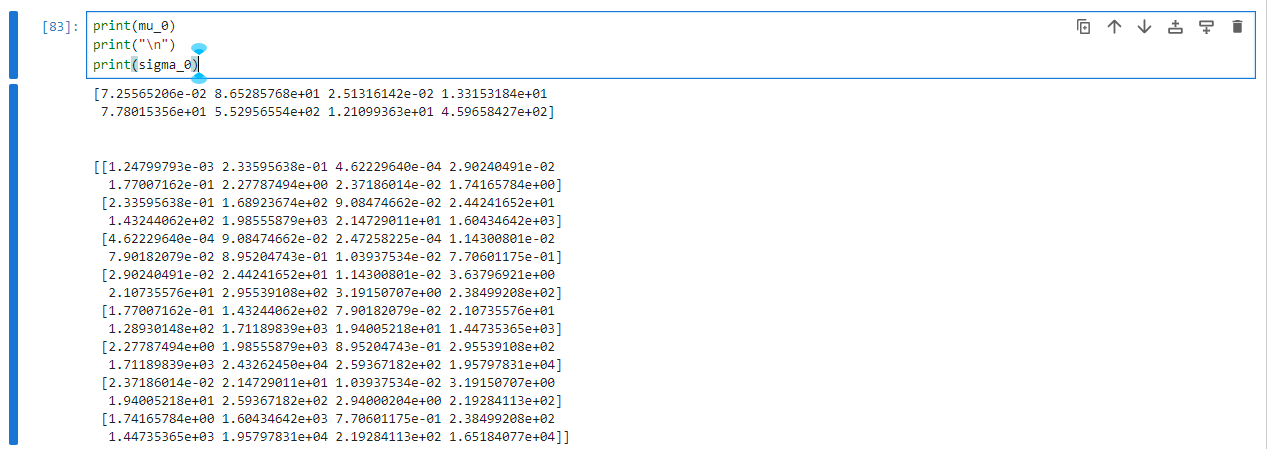
So, now let's train our model (Naive Bayes Classifier) from scratch on training data consisting of eight features, for performing classification of tumors as either Benign or Malignant.

Let's first compute the Maximum Likelihood Estimates of parameters of both conditional likelihoods, P(X|Class=1) and P(X|Class=0)

Let's compute Maximum Likelihood Estimates of:

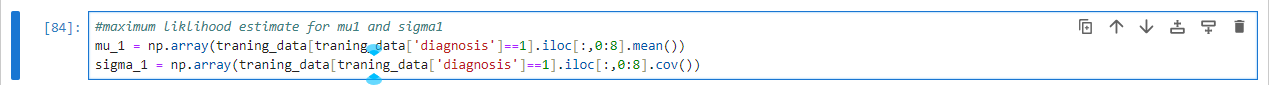
(μ0,∑0) for benign

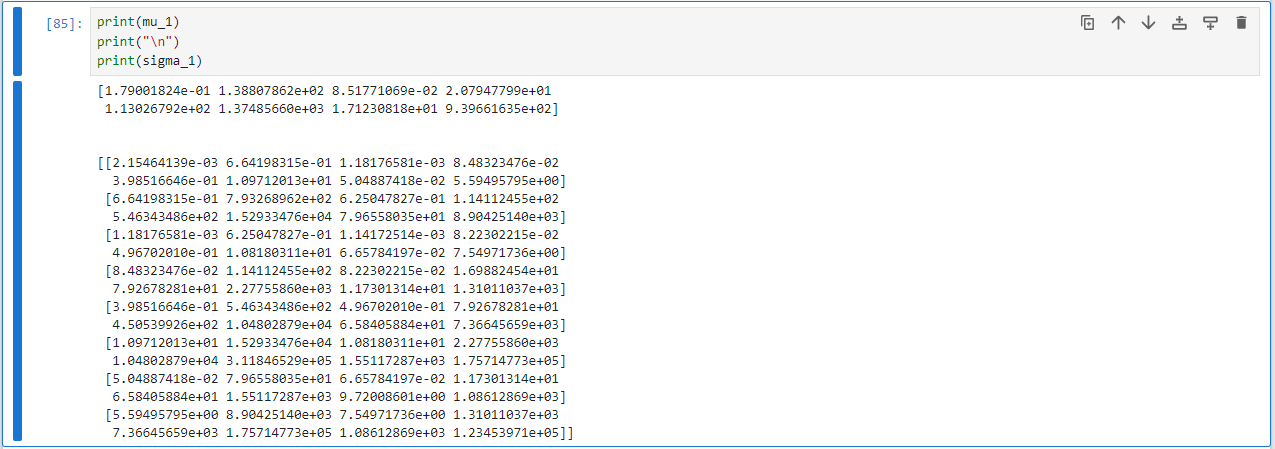




Let's compute Maximum Likelihood Estimates of:

(μ1,∑1) for malignant

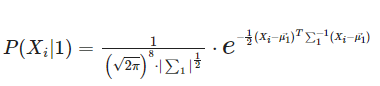




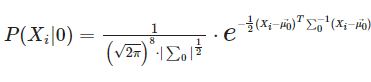
**Now, let's compute P(Class=1|Xi) with the help of Bayes Rule. For this, we are going to define a function called:**

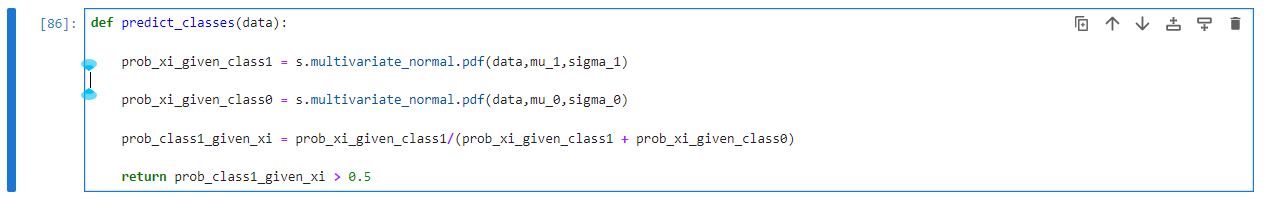
**In this function, we will be implementing two functions:**

1. **s.multivariate\_normal.pdf(x,mu\_1,sigma\_1)** : This function will implement the PDF of multivariate normal distribution with (μ1→,∑1):



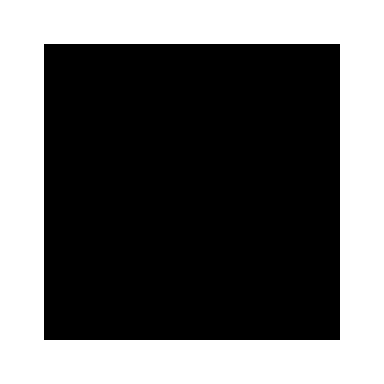
2. **s.multivariate\_normal.pdf(x,mu\_0,sigma\_0)** : This function will implement the PDF of multivariate normal distribution with (μ0→,∑0):

****



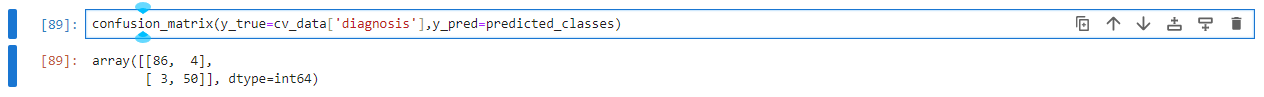




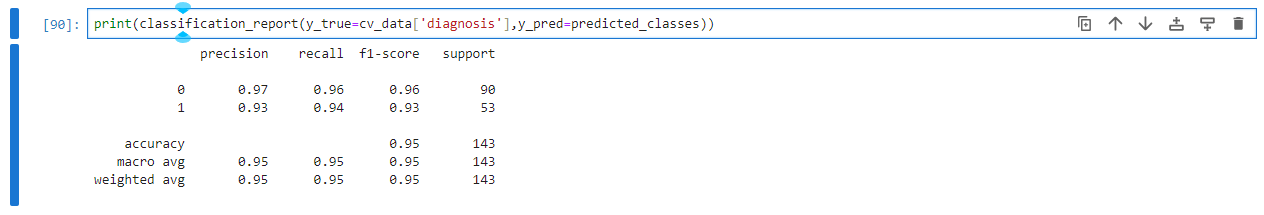
 5 Performance Analysis

After training Naive Bayes Classifier, we are getting predicted categories where False means that we are getting the predicted category as "Benign" or Class 0 and True means that we are getting predicted category as "Malignant" or Class 1.

**Now, let's see the performance of our trained model in the form of confusion matrix.**

****

Looking at the confusion matrix, it can be said that the number of True Positives (Tumors that were actually Malignant as well as classified as Malignant) are 86, the number of False Negatives (Tumors that were actually Malignant but classified as Benign) are 4, the number of False Positives (Tumors that were actually Benign but classified as Malignant) are 3 and the number of True Negatives (Tumors that were actually Benign as well as classified as Benign) are 50. Now, looking at these numbers, it looks like our model is performing very well in terms of performance metrics such as Precision, Recall, Accuracy, F1 Score. Let's try to have a look on the performance metrics now.

****

As, all classification metrics are above or equal to 93 %

The model performs well, achieving high precision, recall, and F1 scores for both classes, with an overall accuracy of 95%.