Text

Description automatically generated

**Building Machine learning Model for Predicting Breast Cancer Using Multi Linear Regression Technique**

**Aashish Aggarwal – 2020CH70151**

**Sanskar Soni – 2020CH10126**

**Ananya Damesha – 2020CH70154**

**Vaibhav Shivran – 2020CH70201**

**Abstract**

**Content:** Breast cancer is a dominant cancer in women worldwide and is increasing in developing countries where most cases are diagnosed in late stages. Regression analysis has become popular among several fields of research and standard tools in analyzing data. This study aims to numerically predict breast cancer using multilinear regression. Breast cancer is one of the major life-threatening diseases in the world today. But early diagnosis would ensure timely treatment and even recovery. We are using several factors which can be responsible for breast cancer. This paper involves multilinear regression equations developed from statistical determination of mean square error and root mean square error etc.**We also designed a web app to predict whether the cancer is benign or malignant with the algorithm we created to do so. Click** [**this**](https://boiling-taiga-43683.herokuapp.com/) **for web app.**

**URL: https://boiling-taiga-43683.herokuapp.com/**

**Key words:** Multi linear regression, Breast Dancer Data, prediction, mean square error, Root mean square error, benign, malignant, concavity of cells.

**Contents:**

1.Introduction

2.Problem Formulation

3.Numerical Analysis

4.Results and Discussions

5.Conclusion

6.Self-Assessment

7.Path Forward

8.References

**INTRODUCTION:**

Breast Cancer is one of a major issue that many of the women worldwide are facing today. Earlier detection of cancer by performing detailed analysis based on the existing records can assist the physicians in providing better treatment to their patients. In 2020, there were 2.3 million women diagnosed with breast cancer and 685 000 deaths globally. As of the end of 2020, there were 7.8 million women alive who were diagnosed with breast cancer in the past 5 years, making it the world’s most prevalent cancer. Breast cancer is hormone related, and the factors that modify the risk of this cancer when diagnosed premenopausally and when diagnosed (much more commonly) postmenopausally are not the same. The 1 in 28 women in India was likely to develop breast cancer during her lifetime. A few decades back, breast cancer was seen only after fifty years of age and the number of young women suffering from this disease was lesser. Almost 65-70 per cent patients were above 50 years and only 30 to 35 per cent women were below 50 years. However, presently, breast cancer was more common in the younger age group and 50 per cent of all cases are in the 25 to 50 years of age group. More than 70 per cent of the cases present in the advanced stage were accounting for poor survival and high mortality. At the current rate, 13 million breast cancer deaths around the world will occur in the next 25 years because of lack of medical awareness.

Incidence' means the number of women who were newly detected with breast cancer in that year. For the year 2018, 162,468 women were newly detected with breast Cancer in India. Also, Breast Cancer accounted for 27.7% of all newly detected cancers in women. Which means, roughly, one in four newly detected cancer in women in India was breast cancer. (Chart.1)

'Mortality' means the number of women who died due to breast cancer in that year. For the year 2018, 87,090 women died due to breast Cancer in India. Breast Cancer accounted for about 23.5% of all cancer related deaths in women in India. Which means that, almost about one in four deaths due to cancer, in women in India, was due to breast cancer. (Chart.2)

**CHART-1**

Chart, pie chart

Description automatically generated

**CHART-2**

Chart, pie chart

Description automatically generated

**How cancer grows: -**

Normal human cells have a finite ability to undergo mitosis due to the end replication problem. This is largely due in part to the ends of chromosomes(telomeres) shortening after each mitotic division.  Once normal human cells reach Hayflick's limit (The concept states that a normal human cell can only replicate and divide forty to sixty times before it cannot divide anymore and will break down by programmed cell death or apoptosis.), cells can go into cellular senescence. Cancer cells exceed Hayflick's limit and undergo mitosis. they can elongate telomeres using enzyme telomerase. They can have an abnormal number of chromosomes per cell and can bear mutations in DNA. Genes that are mutated or lost are TSG (tumor suppressor genes). Over-expressed genes now are oncogenes, which cause cells to proliferate uncontrollably

**Problem Formulation**

The goal of the model is to predict the concavity of the cancer cell using Multi Linear regression model. The larger the concavity, the larger the pressure on surrounding cells, since the cancer cell competes with adjacent normal cells for nutrients. The concavity of the cancer cell tells how much the cell can expand. The concavity of the cancer cell determines its tumour potential. It determines how fast this cancer cell divides. If it divides quickly, the tumour mass grows rapidly and thus it can invade other tissues and can also metastasize. If it divides slowly, mostly it will be benign and does not metastasize. Thus, by predicting the concavity of the cancer cells will help us in deciding how fast cancer would get severe in the future.

**Numerical analysis**

Multiple Linear regression (MLR) is a lot like simple linear regression in which it depends on one predictor variable to predict the criterion value whereas MLR depends on more than one predictor value to predict the criterion value. While doing MLR, the assumptions are as follows:

• Both predictor and criterion variables are in a linear relationship

• Predictor variable is not highly correlated with each other.

• Residuals should be the same at every explanatory level.

• Residuals should be normally distributed. Suppose the model has ‘n’ independent variables

A picture containing diagram

Description automatically generatedText, letter

Description automatically generatedA picture containing table

Description automatically generatedDiagram, text

Description automatically generated

**Methodology**

The Process of Building regression model in general is depicted below

Diagram

Description automatically generated

**Results and discussions**

**Collecting Data:**

Collecting data can be of Primary Source or Secondary Source. In primary sources, data is collected directly without any Third-party whereas, in Secondary sources, it takes the data from the primary source. The dataset for this study has been downloaded from the Kaggle website [12] and falls under secondary data. As it is a classification dataset, the dataset’s features required for regression analysis have been separated from the original dataset. The attributes of the dataset are as follows: ● radius mean - Mean of the distances from the centre to the points on the perimeter of the cancer cell

● texture mean - Standard deviation of greyscale values

● perimeter mean

● area mean

● smoothness mean - Local variation in radius lengths

● compactness mean

● concave points mean - Number of concave portions of the contour

● fractal dimension mean

● concavity mean - Severity of concave portions.

**Data Visualization**

The main goal of data visualization is to make it easier to visualize the information and to identify patterns in the dataset. Charts and Graphs make the new data findings easier than the normal dataset.

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated**Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated**

**Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated**

**Analysing Each Independent Variable**

**Mean Radius**

Text

Description automatically generated with medium confidence

Graphical user interface, application, Word

Description automatically generated

**Conclusion -** 1. Y and X relationship R Square (R ) equals 0.458. It means that 45.8% of the variability of Y is explained by X. correlation (R) equals 0.6768. It means that there is a strong direct relationship between X and Y.

**Mean TextureText

Description automatically generated with low confidence**

Table

Description automatically generated

**Conclusion -** 1. Y and X relationship R Square (R ) equals 0.09146. It means that 9.1% of the variability of Y is explained by X. correlation (R) equals 0.3024. It means that there is a weak direct relationship between X and Y.

**Mean Perimeter**Graphical user interface, text

Description automatically generated with medium confidence

**Table, Word

Description automatically generated**

**Conclusion -** 1. Y and X relationship R Square (R ) equals 0.5129. It means that 51.3% of the variability of Y is explained by X. correlation (R) equals 0.7161. It means that there is a strong direct relationship between X and Y.

**Mean Area**

**Graphical user interface, text

Description automatically generated with medium confidenceTable, Word

Description automatically generated**

**Conclusion-** 1. Y and X relationship R Square (R ) equals 0.4706. It means that 47.1% of the variability of Y is explained by X. correlation (R) equals 0.686. It means that there is a strong direct relationship between X and Y.

**Mean SmoothnessText, letter

Description automatically generatedTable

Description automatically generated**

**Conclusion -** 1. Y and X relationship R Square (R ) equals 0.2725. It means that 27.2% of the variability of Y is explained by X. correlation (R) equals 0.522. It means that there is a moderate direct relationship between X and Y.

**Mean CompactnessText

Description automatically generatedTable

Description automatically generated**

**Conclusion -** 1. Y and X relationship R Square (R ) equals 0.7799. It means that 78% of the variability of Y is explained by X. correlation (R) equals 0.8831. It means that there is a very strong direct relationship between X and Y.

**Mean Concave PointGraphical user interface, application

Description automatically generated**Table

Description automatically generated

**Conclusion -** 1. Y and X relationship R Square (R ) equals 0.849. It means that 84.9% of the variability of Y is explained by X. correlation (R) equals 0.9214. It means that there is a very strong direct relationship between X and Y.

**Mean Fractal Dimension**Text

Description automatically generated with medium confidenceTable

Description automatically generated

**Conclusion -** 1. Y and X relationship R Square (R ) equals 0.1134. It means that 11.3% of the variability of Y is explained by X. correlation (R) equals 0.3368. It means that there is a weak direct relationship between X and Y.

**Data Pre-processing**

1. **Correlation Factors**

**Graphical user interface, application

Description automatically generated**

**As we can see strong correlations among various independent variables and some independent variable have very less correlation with dependent variable. So we run regression by assuming different models and select the best out of them**

**Proposed Models**

1. Taking All Independent Variables
2. Removing fractal dimension and texture mean
3. Removing fractal dimension, texture mean , radius mean
4. Removing fractal dimension, texture mean, perimeter mean , area mean
5. Removing fractal dimension mean , texture mean , perimeter mean , area mean , concave point mean
6. Removing fractal dimension mean , texture mean , perimeter mean , area mean , compactness mean

**We will run regression on all these modals and select the best one out**

**Model -1a**

**Taking all Independent variables**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| SUMMARY OUTPUT | |  |  |  |
|  |  |  |  |  |
| *Regression Statistics* | |  |  |  |
| Multiple R | 0.956367018 |  |  |  |
| R Square | 0.914637873 |  |  |  |
| Adjusted R Square | 0.913418414 |  |  |  |
| Standard Error | 0.023457354 |  |  |  |
| Observations | 569 |  |  |  |
|  |  |  |  |  |
| ANOVA |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* |
| Regression | 8 | 3.301642239 | 0.41270528 | 750.0357961 |
| Residual | 560 | 0.308138569 | 0.000550247 |  |
| Total | 568 | 3.609780807 |  |  |
|  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* |
| Intercept | 0.079357839 | 0.034335681 | 2.311235341 | 0.021182056 |
| radius\_mean | -0.053228566 | 0.010949967 | -4.861070923 | 1.51846E-06 |
| texture\_mean | 0.000427227 | 0.000248663 | 1.718098782 | 0.086331362 |
| perimeter\_mean | 0.006928361 | 0.001762672 | 3.930601863 | 9.53416E-05 |
| area\_mean | 5.46242E-05 | 2.08381E-05 | 2.621360818 | 0.008995972 |
| smoothness\_mean | -0.874074471 | 0.113392279 | -7.708412593 | 5.84167E-14 |
| compactness\_mean | 0.28420385 | 0.08643876 | 3.287921429 | 0.001072592 |
| concave points\_mean | 1.581547438 | 0.096800201 | 16.33826611 | 2.35321E-49 |
| fractal\_dimension\_mean | 0.914246961 | 0.352222042 | 2.595655157 | 0.009688739 |

**As we can see some independent variable have p value >0.001 which is red flag for us.**

**Model -2a**

**Removing fractal dimension and texture Mean as correlation coefficient is very small with dependent variable**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| SUMMARY OUTPUT |  |  |  |  |
|  |  |  |  |  |
| *Regression Statistics* | |  |  |  |
| Multiple R | 0.955644963 |  |  |  |
| R Square | 0.913257296 |  |  |  |
| Adjusted R Square | 0.912331217 |  |  |  |
| Standard Error | 0.02360417 |  |  |  |
| Observations | 569 |  |  |  |
|  |  |  |  |  |
| ANOVA |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* |
| Regression | 6 | 3.29665866 | 0.54944311 | 986.1551799 |
| Residual | 562 | 0.313122148 | 0.000557157 |  |
| Total | 568 | 3.609780807 |  |  |
|  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* |
| Intercept | 0.161656558 | 0.018645094 | 8.670192657 | 4.6252E-17 |
| radius mean | -0.051110472 | 0.010981977 | -4.654031908 | 4.06447E-06 |
| perimeter mean | 0.006083265 | 0.001740757 | 3.494608617 | 0.000512186 |
| area mean | 7.63855E-05 | 1.92609E-05 | 3.965831155 | 8.25926E-05 |
| smoothness mean | -0.857198093 | 0.110588667 | -7.751229075 | 4.28856E-14 |
| concave points\_mean | 1.58642892 | 0.097353559 | 16.29554109 | 3.53672E-49 |
| compactness\_mean | 0.428330257 | 0.068563906 | 6.247168301 | 8.25765E-10 |

**As we can see R square and standard error does not get affected much and p values also decrease for some Independent variable therefore it is a right step to remove fractal dimension and texture mean**

**Model – 3a**

**As perimeter, radius, area is strongly correlated we will first take 2 of them then 1 of them to analyse**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| SUMMARY OUTPUT | |  |  |  |
| Removing radius |  |  |  |  |
| *Regression Statistics* | |  |  |  |
| Multiple R | 0.953894203 |  |  |  |
| R Square | 0.90991415 |  |  |  |
| Adjusted R Square | 0.909114098 |  |  |  |
| Standard Error | 0.024033361 |  |  |  |
| Observations | 569 |  |  |  |
|  |  |  |  |  |
| ANOVA |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* |
| Regression | 5 | 3.284590635 | 0.656918127 | 1137.318828 |
| Residual | 563 | 0.325190172 | 0.000577602 |  |
| Total | 568 | 3.609780807 |  |  |
|  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* |
| Intercept | 0.146835714 | 0.018705168 | 7.850007611 | 2.11057E-14 |
| perimeter mean | -0.001907903 | 0.000291568 | -6.543594659 | 1.35186E-10 |
| area mean | 9.06491E-05 | 1.93613E-05 | 4.681982699 | 3.56433E-06 |
| smoothness mean | -0.973726106 | 0.109675581 | -8.878239744 | 9.06218E-18 |
| compactness mean | 0.675584331 | 0.044130287 | 15.30885884 | 1.74888E-44 |
| concave points\_mean | 1.664048253 | 0.097658354 | 17.03948706 | 8.23705E-53 |

**As we can see removing 1 correlated factor makes p value < 0.001 so it is right to remove them we will do same for area and perimeter**

**Model – 4a**

**As perimeter, radius, area is strongly corelated we will first take 2 of them then 1 of them to analyse**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| SUMMARY OUTPUT |  |  |  |  |
|  |  |  |  |  |
| *Regression Statistics* | |  |  |  |
| Multiple R | 0.952383775 |  |  |  |
| R Square | 0.907034855 |  |  |  |
| Adjusted R Square | 0.906375528 |  |  |  |
| Standard Error | 0.02439276 |  |  |  |
| Observations | 569 |  |  |  |
|  |  |  |  |  |
| ANOVA |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* |
| Regression | 4 | 3.274197013 | 0.818549253 | 1375.697476 |
| Residual | 564 | 0.335583794 | 0.000595007 |  |
| Total | 568 | 3.609780807 |  |  |
|  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* |
| Intercept | 0.098095988 | 0.015106978 | 6.493422491 | 1.84242E-10 |
| radius\_mean | -0.004386274 | 0.000704313 | -6.227733414 | 9.25352E-10 |
| smoothness\_mean | -0.969879709 | 0.110466676 | -8.779839746 | 1.95446E-17 |
| compactness\_mean | 0.570859004 | 0.042523946 | 13.42441275 | 7.47256E-36 |
| concave points\_mean | 1.76951979 | 0.096034428 | 18.42588985 | 1.05811E-59 |

**As we can see removing area and perimeter which is highly correlated with R square and standard error does not get affected much therefore it is a right step similarly, we can do for area and perimeter**

**Model – 5a**

**As compactness and concave point mean are also correlated and concave point mean is also related to radius area and perimeter therefore we will see how results get affected by removing concave point mean**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| SUMMARY OUTPUT | |  |  |  |
|  |  |  |  |  |
| *Regression Statistics* | |  |  |  |
| Multiple R | 0.922535767 |  |  |  |
| R Square | 0.851072242 |  |  |  |
| Adjusted R Square | 0.850281475 |  |  |  |
| Standard Error | 0.030846367 |  |  |  |
| Observations | 569 |  |  |  |
|  |  |  |  |  |
| ANOVA |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* |
| Regression | 3 | 3.072184246 | 1.024061415 | 1076.261905 |
| Residual | 565 | 0.537596561 | 0.000951498 |  |
| Total | 568 | 3.609780807 |  |  |
|  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* |
| Intercept | -0.115296099 | 0.012266913 | -9.39894938 | 1.35596E-19 |
| radius\_mean | 0.0068974 | 0.000439965 | 15.67716072 | 2.96994E-46 |
| smoothness\_mean | -0.10420808 | 0.126429534 | -0.82423843 | 0.410151906 |
| compactness\_mean | 1.118403761 | 0.038465445 | 29.0755444 | 2.5232E-114 |

**As we can see, p value of smoothness mean shoots up therefore it is not a right decision to remove concavity mean**

**Model – 5b**

**Let's remove compactness mean and analyse the results**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| SUMMARY OUTPUT | |  |  |  |
|  |  |  |  |  |
| *Regression Statistics* | |  |  |  |
| Multiple R | 0.936658803 |  |  |  |
| R Square | 0.877329713 |  |  |  |
| Adjusted R Square | 0.876678366 |  |  |  |
| Standard Error | 0.027995349 |  |  |  |
| Observations | 569 |  |  |  |
|  |  |  |  |  |
| ANOVA |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* |
| Regression | 3 | 3.16696796 | 1.055655987 | 1346.947444 |
| Residual | 565 | 0.442812847 | 0.00078374 |  |
| Total | 568 | 3.609780807 |  |  |
|  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* |
| Intercept | 0.149787275 | 0.016765485 | 8.934264092 | 5.77069E-18 |
| radius\_mean | -0.008356099 | 0.000733633 | -11.39003132 | 3.36502E-27 |
| smoothness\_mean | -0.763529272 | 0.125548207 | -6.081562538 | 2.19356E-09 |
| concave points\_mean | 2.670426549 | 0.078839892 | 33.87151434 | 3.8119E-138 |

**As we can see, all p values now are very small but R square decreased by 0.02 and standard error increased by 0.03 so this model can be considered in final decision**

**Final Decision for best model**

Running regression on all the models and analysing them we can conclude that Model 4 is a best model as this model is 90.7% accurate and average error is just 0.024 moreover p value for all the Independent variables are small enough i.e <0.0001 and in regression it is always good to select the simplest model (which have less no of independent variable) and model 4 also follows that so we will take the result of model 4 to make a web app which a normal user can use without getting into technicalities of ML regression

**Conclusion:**

This paper is based on prediction of breast cancer by machine learning model using multi linear regression, it’s an advanced and most accurate method for statistical analysis of dependent variable with several independent variables. we used dependent variable as concavity mean which is the focus of our analysis cause the concavity of the cancer cell tells how much the cell can expand. The concavity of the cancer cell determines its tumour potential. It determines how fast this cancer cell divides. If it divides quickly, the tumour mass grows rapidly and thus it can invade other tissues and can also metastasize. If it divides slowly, mostly it will be benign and does not metastasize. Thus, by predicting the concavity of the cancer cells will help us in deciding how fast cancer would get severe in the future and independent variables are texture mean (Standard deviation of greyscale values), perimeter mean, area mean, smoothness mean (Local variation in radius lengths), compactness mean , concave points mean, fractal dimension mean, We use data from Kaggle for our variables. Our paper could create a fruitful impact on the research related to diagnosis of breast cancer, cause if the cancer is diagnosed earlier, we have a high chance of cure it using medications or surgery.

**Self-Assessment:**

**This is a level-2 term paper,** Codesare working to produce all the results via the proposed method of multi linear regression. All the parameters involved, including RMSE, p-values, correlation factors, standard errors & R2  , the code is based on C++ language.

We take further the research, searching for relations in variables we analyse each independent variable as well as correlations between the independent variables. In the result and discussions section we perform the analysis using graphical visualization of the variables.

In continuation to reach the need of today’s time we also created a web app using JavaScript and other web languages to perform the task. The webapp uses the machine learning algorithm we get from our calculations. It talks about the stage of cancer. If it is benign or malignant.

We also taken account the other possible models which can give result. We analysed them on the basis of various factors and picked up the best possible simplest and accurate model. So, from all this work we have performed for this term paper we can say that this is of level 2 term paper. The models and visualization were not in original research paper.

.

**Path Forward:**

This study can further be extended using computational genomics. A genome is an organism's complete set of DNAS, including all its genes as well as its hierarchical, three-dimensional structural configuration. In contrast to genetics, which refers to the study of individual genes and their roles in inheritance, genomics aims at the collective characterization and quantification of all an organism's genes, their interrelations and influence on the organism. Further heading ahead computational genomics use computational and statistical analysis to decipher biology from genome sequences and related data, including both DNA and RNA sequence. so, we can get more variables and predict the more accurate result through our study.

**References**

**1.** Applied Mathematical Sciences, Vol. 9, 2015, no. 56, 2749 - 2759 HIKARI Ltd, www.m-hikari.com http://dx.doi.org/10.12988/ams.2015.5175 The Use of Fuzzy Linear Regression Models for Tumor Size in Colorectal Cancer in Hospital of Malaysia.

2. Building Machine Learning Model for Predicting Breast Cancer Using different Regression Techniques To cite this article: K. Devasena and J. Shana 2021 IOP Conf. Ser.: Mater. Sci. Eng. 1166 012029.

3. Classification and Prediction of Breast Cancer using Linear Regression, Decision Tree and Random Forest 978-1-5386-3243-7/17/$31.00 ©2017 IEEE I.

4. On Breast Cancer Detection: An Application of Machine Learning Algorithms on the Wisconsin Diagnostic Dataset Abien Fred M. Agarap abienfred.agarap@gmail.co

5. International Journal of Engineering Research & Technology (IJERT) http://www.ijert.org ISSN: 2278-0181 IJERTV9IS020280 (This work is licensed under a Creative Commons Attribution 4.0 International License.) Published by : www.ijert.org Vol. 9 Issue 02, February-2020 “Breast Cancer Classification and Prediction using Machine Learning”

6.Latest Statistics of Breast Cancer in India 2020 | Breast Cancer India

7. lecture notes,CLL113 prof.Jayati sarkar .

8. Kaggle.com