Building Machine Learning Model for Predicting Breast Cancer Using different Regression Techniques

K. Devasena¹ and J. Shana²

¹Coimbatore Institute of Technology, Coimbatore, India
²Coimbatore Institute of Technology, Coimbatore, India
shana@cit.edu.in

Abstract. Breast cancer is one of the major life-threatening diseases in the world today. But early diagnosis would ensure timely treatment and even recovery. Apart from the medical procedures that help to identify cancer, techniques such as machine learning can be applied to predict the occurrence of the disease in patients by analyzing their cell data. In this work, an attempt is made to apply three different regression techniques Multiple, Ridge, and Polynomial regression techniques. The performance metrics reveal that Multiple regression has a better performance compared to the other models.

Keywords: Multiple Regression, Ridge regression, Machine Learning, Cancer data, Prediction

1. Introduction

Cancer has been portrayed as a heterogeneous sickness comprising a wide range of subtypes. The early analysis and anticipation of a malignant growth type have become a need in disease research, as it can encourage the resulting clinical administration of patients. The significance of ordering cancer patients into high or low has driven many exploration groups, from the biomedical and the bioinformatics field, to consider the utilization of AI (ML) techniques. In this manner, these methods have been used to show the movement and therapy of carcinogenic conditions. Furthermore, the capacity of machine learning algorithms to recognize key highlights from complex datasets uncovers their significance.

Researchers applied various techniques, for example, screening in the beginning phase, to discover kinds of malignant growth before they cause side effects. Also, they have grown new techniques for the early forecast of disease treatment results. With the appearance of innovations in the field of medication, a lot of malignant growth information has been gathered and is accessible to the clinical examination team.

Machine learning models have become so useful in every domain. It can deliver useful insights from datasets and allow effective decision-making. Linear regression and Logistic regression are the two most prominent techniques in machine learning. Regression analysis is a predictive modeling technique that analyzes the relation between the target or dependent variable and independent variable in a dataset. The different types of regression analysis techniques are used when the target and independent variables show a linear or non-linear relationship between each other. There are many types of regression analysis techniques, and the use of each method depends upon the number of factors.
1. Linear Regression
2. Logistic Regression
3. Ridge Regression
4. Lasso Regression
5. Polynomial Regression
6. Bayesian Linear Regression

The goal of the model is to predict the concavity of the cancer cell using different regression models. 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 tumor potential. It determines how fast this cancer cell divides. If it divides quickly, the tumor 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. Here various regression techniques have been applied and the performance of all these have been compared.

2. Related Work

In [1] the authors have surveyed machine learning techniques applied for cancer data analysis. They include machine learning models such as Naive Bayes, Support Vector Machines, Neural Networks, etc. In [2] the authors have explained Multiple Linear Regression and the steps to do like finding out the missing values. In [3] the authors have explained Ridge regression regularization and its penalty term. The authors in [4] explore the different data mining approaches using classification which can be applied to Breast Cancer data to build deep predictions. Result analysis reveals that among all the classifiers Simple Logistic Regression yields the deep predictions. The authors in [5] have tried seven regression techniques and implemented them on the Breast Cancer dataset to know the best regression technique among those compared. The analysis found that Logistic Regression gives the best results to diagnose Breast Cancer with given attributes among seven regression techniques with least Relative Absolute Error (10.59%) and Least Root Relative Squared Error. (45.84%) using the Weka tool.

In [6] the authors have applied machine learning methods especially logistic regression on the datasets to determine and examine whether there are any signs or possibilities of cancer and if the person is examined as cancerous then the stage of cancer is also determined. And out of all the techniques logistic regression outperformed other techniques. The study in [7] used machine learning techniques to build models for detecting and visualizing significant prognostic indicators of breast cancer survival rate for a Malaysian hospital. Model evaluation using random forest algorithms yielded slightly better accuracy when compared to other algorithms. Nevertheless, accuracies displayed by all the algorithms appeared close.

Study in [8] compared the performance of artificial neural network and multivariable logistic regression models, in prediction of outcomes in head trauma and studied the reproducibility of the findings. Artificial neural network outperformed the other model in term of accuracy.
Authors in [9] investigate their applications in the Breast cancer dataset and compare the results through different algorithms using the Wisconsin breast cancer database (WBCD) which is the benchmark database for comparing the results through different algorithms. In [10], different machine learning and data mining techniques for the detection of breast cancer were proposed. Results obtained with the logistic regression model with all features included showed the highest classification accuracy (98.1%), and the proposed approach revealed the enhancement in accuracy performance. In [11] the authors have explained Polynomial regression, the relationship between the attributes and its evaluating metrics.

3. Regression Techniques

There are many regression techniques apart from linear regression as listed below. There exist linearity and nonlinearity between the variables so in order to model the exact variance of the data Multiple Linear Regression and Polynomial Linear regression both have been implemented and analyzed their performance in terms of variance the model explains (R^2 score) and accuracy of the predictions made by those metrics. In order to use Ridge Regression, VIF (Variance Inflation Factor) has been done to check the multicollinearity between the independent variables and it has multicollinearity.

3.1 Multiple Linear Regression

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, then Multiple linear regression Equation can be written as

\[ y = b_0 + b_1x_1 + b_2x_2 + \ldots + b_nx_n + e \]  

(1)

Where \( y \) = Criterion variable(Independent Variable), \( b_0 = Y\)-Intercept(constant), \( b_1, b_2 = \) Regression coefficients associated with \( x_1, x_2 \)
\( x_1, x_2 = \) Explanatory variables
\( b_n = \) Slope and
\( e = \) Residual value(Error).

The advantage is it is an efficient prediction model with more than one independent variable as it will give a more precise output. The disadvantage is there is too much data and that may have false or incomplete data which is difficult to find.

3.2 Ridge Regression

Ridge Regression is a method that becomes useful when the dataset correlates with the independent variables in other terms when it has multicollinearity. It is an L2 regularization method that reduces the risk of Overfit and improves the model accuracy. As ridge regression is an L2 regularization method, it
determines the parameters of the equation and minimizes the Sum of squared residuals with a penalty term which makes the ridge regression line way more efficient than the least square line.

\[ RSS + \lambda \sum P_i \beta_i^2 \]  

In Ridge Regression, with a small amount of Bias due to a penalty, it reduces the variance of the model. The advantage is it reduces model complexity and provides a long-term solution. The drawback is it will shrink the lambda value close to 0 but never exactly 0. And it becomes computationally expensive.

3.3 Polynomial Linear Regression:
While modeling the data and finding it out as it is performing badly as it needed some curvature line for the data to fit, this is where Polynomial Linear regression comes in handy. It tells the relationship between the predictor and criterion variables as an nth degree polynomial. In Polynomial linear regression, it doesn’t require the relationship between the predictor and criterion variables to be Linear.

\[ y = b_0 + b_1 x_1^2 + b_2 x_2^2 + \cdots + b_n x_n^2 + e \]  

where n is the degree of the polynomial b is the set of coefficients.

The degree depends on the predictor and criterion variables. The polynomial linear regression with degree one, is just a simple linear regression. So, the degree for the polynomial linear regression should be more than one. The best degree for the model, in which it generates the lowest RMSE. In polynomial linear regression, as n takes the higher value it increases, its accuracy but also higher the degree, the higher the chance of overfitting. This regression fits a wide range of curvatures. But it is strongly sensitive to outliers.

4. METHODOLOGY

The process of building the regression model, in general, is depicted in Figure 1. The steps are elaborated in section 4.1. through 4.9.

![Figure 1: Methodology used in Machine Learning](image_url)
4.1 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 center 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** - Perimeter^2/area - 1.0
- **concave points_mean** - Number of concave portions of the contour
- **fractal_dimension_mean**
- **concavity_mean** - Severity of concave portions.

4.2 Importing Data:
Locate the CSV file from the system and import it using the pandas library with the pre-defined function named .read_csv("Filename.csv").

4.3 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.

![Figure 2: Plot between independent and dependent variables](image)

Figure 2 shows the relationship between the concavity_mean and the radius_mean, area_mean, and some features.
4.4 Identifying Dependent and Independent Variables
In simple words, dependent variables are the cause, which means its value is dependent on the other and independent variables are the effect, which means its value changes the other’s value. Here the independent variables are radius_mean, texture_mean, perimeter_mean, area_mean, smoothness_mean, compactness_mean, concave points_mean, fractal_dimension_mean, and the dependent variable is Concavity_mean.

4.5 Data Preprocessing
Data preprocessing refers to the technique of preparing the raw data into data that is suitable to build the model. Without this step, the model's accuracy may reduce.

**Missing values:** Missing values can create problems. It is good if the missing value has been identified and replaced. These can be done with the mean/median method or by simply removing it but it may result in a dosage of data.

**Feature Scaling:** Feature Scaling is a technique used to normalize the range of features of the data. Feature Scaling brings all the values in the dataset to the same magnitude. In this work, Min-max normalization is one of the techniques to perform feature scaling, it rescales a feature value with a distribution value from 0 to 1.

**Removing Outliers:** An outlier is an observation in a dataset that is different from other observations in the feature. It may occur due to input error or data corruption.

![radius_mean](image)

**Figure 3:** Before removing Outliers.

Figure 3 shows how the data will be visualized when it has outliers in the dataset.
4.6 Splitting Dataset
The dataset has been split into 2 separate sets namely the training set and test set. A training set is a subset that is used to train the model on the other hand Test set is the one that is used to test the predictions of the model. Usually, the dataset splitting happens in the ratio of 70:30 or 80:20 which means 70% or 80% of the data is for training the model and 30% or 20% of the data is for testing the model. The splitting can vary from dataset to dataset. In this work, the dataset has been divided into a ratio of 75% for the training dataset and 25% for the testing dataset.

4.7 Model Fitting
The model fitting is a measure of how well a machine learning model generalizes to similar data to that on which it was trained. A well-fitted model produces more accurate outcomes. A model that is overfitted matches the data too closely. A model that is under fitted doesn’t match closely enough.

4.8 Model Evaluation
In Model Evaluation, the model is evaluated using different performance metrics such as Means Square Error (MSE) and Root Mean Square Error (RMSE). A model with good performance metrics is considered the best deployment model. Coefficient of determination (R²) describes how the differences in the Independent variable can be affected by the differences in the Dependent variable. MSE reveals the average residual sum of squares. Mean Absolute Error specifies the average distance between the actual and predicted values. Root Mean Square Error (RMSE) - RMSE tells the Standard Deviation of the residuals. In this work, the model has used the above-mentioned metrics to evaluate the model performance.

4.9 Prediction
The model is used for predicting the disease. Given the new input data, the model will predict the concavity of the breast cancer cell.
5. Experimental Results

Figure 5: Actual values vs predicted values

Figure 5 represents the predictions made by the three different models. This graph shows that the actual and predicted values can differ from model to model. From the above observation, the Multiple Linear Regression model shows less difference in the actual and predicted values accurately than the Ridge Regression model or Polynomial Linear Regression Model.

Figure 6: Prediction Metrics.

Figure 6, depicts the prediction metrics which are used to evaluate the models which are MAE, MSE, RMSE, R^2.

Table 1. Model Performance

| Techniques                | MSE  | MAE   | RMSE  | Coefficient of determination |
|---------------------------|------|-------|-------|------------------------------|
| Multiple Linear Regression| 0.00418 | 0.04794 | 0.06465 | 93.7%                        |
| Ridge Regression          | 0.00489 | 0.05147 | 0.06998 | 92.6%                        |
| Polynomial Linear Regression | 0.00645 | 0.06253 | 0.08036 | 90.3%                        |
Mean Squared value, tells variance of the Residuals (i.e how the data points are located either close or far away from the best fit line), the variance of the residuals is smaller if the data points are held closer to the best fit line of the model and the variance of the residuals is bigger if the data points are held far from the best fit line. Table 1 prediction of MSE tells, Lower the value higher the performance. The performance of Multiple linear regression > Ridge Regression > Polynomial Linear Regression.

Consider the case of Mean Absolute Error, which tells the Mean of the Residuals (i.e absolute distance from each point to the best fit line ). In Table 1 prediction of MAE shows, lower the value higher the performance. The performance of Multiple linear regression > Ridge Regression > Polynomial Linear Regression.

Similarly, Root Mean Squared Error, which tells the standard deviation of errors (root of MSE). The prediction values of RMSE in Table 1 prove that the lower the value higher the performance. The performance of Multiple linear regression > Ridge Regression > Polynomial Linear Regression.

Take the case of the Coefficient of determination, which tells the proportion of variance explained by the model. Table 1, prediction of $R^2$ tells, higher the value higher the performance. The performance of Multiple linear regression > Ridge Regression > Polynomial Linear Regression.

6. Conclusion
In this work, the breast cancer dataset was analyzed using three regression techniques. Machine learning models can predict the target class with high accuracy. According to the performance metrics from the above observation and comparing with each other, Multiple Linear Regression with an accuracy of 93.7% has been performed better than Ridge Regression with an accuracy of 92.6% which performed better than Polynomial Linear Regression with an accuracy of 90.3%.

References
[1] Konstantina Kourou, Themis P.Exarchos Konstantinos P.Exarchos, Michalis V.Karamouzis, Dimitrios I.Fotiadiis 2015 Machine learning applications in cancer prognosis and prediction, Computational and Structural Biotechnology Journal, 13.
[2] Gulden kaya Uyanik, Nese Guler 2013 A study on Multiple Linear Regression Procedia - Social and Behavioral Sciences, 106, 234 – 240.
[3] N S M Shariff, H M B Duzan 2018 An application of proposed Ridge Regression Methods to real data problem, International Journal of engineering and technology, 7,106-108.
[4] Sultana, J., 2018 Predicting Breast Cancer using Logistic Regression and Multi-Class Classifiers, International Journal of Engineering and Technology, 13,109-116.
[5] Ritu Tayal, Anshul Tickoo, 2014, Performance Analysis of Regression Data Mining Techniques Implemented on Breast Cancer Dataset, International Journal of Latest Trends in Engineering and Technology (IJLTET), 4(1), 188-195.
[6] Heena Nankani, Shruti Gupta, Shubham Singh, S. S. Subashka Ramesh, 2019 Detection Analysis of Various Types of Cancer by Logistic Regression using Machine Learning, International Journal of Engineering and Advanced Technology,9(1),45-53.
[7] Ganggayah, M.D., Taib, N.A., Har, Y.C. et al. 2019 Predicting factors for survival of breast cancer patients using machine learning techniques. BMC Med Inform Decis Mak 19, 48.
[8] Eftekhar B, Mohammad K, Ardebili HE, Ghodsi M, Ketabchi E. 2019 Comparison of artificial neural network and logistic regression models for prediction of mortality in head trauma based on initial clinical data. BMC Med Inform Decis Mak.19,48.
[9] Wenbin Yue, Zidong Wang, Hongwei Chen, and Annette Payne and Xiaohui Liu, 2018 Machine Learning with Applications in Breast Cancer Diagnosis and Prognosis, 2(13).
[10] Muhammet Fatih Ak, 2017 *A Comparative Analysis of Breast Cancer Detection and Diagnosis Using Data Visualization and Machine Learning Applications* Healthcare (Basel) 8(2).

[11] Eva Ostertagova 2012 *Modelling on Polynomial regression* Procedia Engineering, 48 500 – 506.

[12] https://www.kaggle.com/uciml/breast-cancer-wisconsin-data