DOI: https://doi.org/10.52756/ijerr.2024.v38.012 *Int. J. Exp. Res. Rev.*, Vol. 38: 132-146(2024)

Original Article | Peer Reviewed | Open Access

International Journal of Experimental Research and Review (IJERR)
© Copyright by International Academic Publishing House (IAPH)
ISSN: 2455-4855 (Online)
www.iaph.in

Breast Cancer Disease Prediction Using Random Forest Regression and Gradient Boosting Regression

Pradeep Yadav[1]*, Chandra Prakash Bhargava[1], Deepak Gupta[1], Jyoti Kumari[1], Archana Acharya[1] and Madhukar Dubey[2]

[1]Department of Computer Science and Engineering, ITM, Gwalior -474001, Madhya Pradesh, India;
[2]Department of Information and Technology, ITM, Gwalior -474001, Madhya Pradesh, India

E-mail/Orcid Id:

PY, pradeep.yadav@itmgoi.in, https://orcid.org/0000-0003-1073-7076; *CPB,* chandra.prakashbhargava@itmgoi.in, https://orcid.org/0009-0004-4751-9149; *DG,* deepak.gupta@itmgoi.in, https://orcid.org/0000-0003-3929-1362; *JK,* jyotikumari.asst@itmgoi.in, https://orcid.org/0009-0004-0533-4306; *AA,* archanaacharya.it@itmgoi.in, https://orcid.org/0009-0007-0032-6122; *MD,* madhukardubey.it@itmgoi.in, https://orcid.org/0000-0002-0874-2039

Article History:

Received: 06th Dec., 2023
Accepted: 23th March., 2024
Published: 30th Apr., 2024

Keywords:

Gradient Boosting Regression, Mean Squared Error, Mean Absolute Error, Random Forest, R-squared.

How to cite this Article:
Pradeep Yadav, Chandra Prakash Bhargava, Deepak Gupta, Jyoti Kumari, Archana Acharya and Madhukar Dubey (2024). Breast Cancer Disease Prediction Using Random Forest Regression and Gradient Boosting Regression. *International Journal of Experimental Research and Review*, *38*, 132-146.
DOI: https://doi.org/10.52756/ijerr.2024.v38.012

Abstract: The current research endeavors to evaluate the efficacy of regression-based machine learning algorithms through an assessment of their performance using diverse metrics. The focus of our study involves the implementation of the breast cancer Wisconsin (Diagnostic) dataset, employing both the random forest and gradient-boosting regression algorithms. In our comprehensive performance analysis, we utilized key metrics such as Mean Squared Error (MSE), R-squared, Mean Absolute Error (MAE), and Coefficient of Determination (COD), supplemented by additional metrics. The evaluation aimed to gauge the algorithms' accuracy and predictive capabilities. Notably, for continuous target variables, the gradient-boosting regression model emerged as particularly noteworthy in terms of performance when compared to other models. The gradient-boosting regression model exhibited remarkable results, highlighting its superiority in handling the breast cancer dataset. The model achieved an impressively low MSE value of 0.05, indicating minimal prediction errors. Furthermore, the R-squared value of 0.89 highlighted the model's ability to explain the variance in the data, affirming its robust predictive power. The Mean Absolute Error (MAE) of 0.14 reinforced the model's accuracy in predicting continuous outcomes. Beyond these core metrics, the study incorporated additional measures to provide a comprehensive understanding of the algorithms' performance. The findings underscore the potential of gradient-boosting regression in enhancing predictive accuracy for datasets with continuous target variables, particularly evident in the context of breast cancer diagnosis. This research contributes valuable insights to the ongoing exploration of machine learning algorithms, providing a basis for informed decision-making in medical and predictive analytics domains.

Introduction

Artificial intelligence is a computer science branch that covers machine learning and deep learning concepts. With several innovations in numerous domains, machine learning has recently grown in importance as a topic of study. The discipline is not without its difficulties and restrictions, though, including the requirement for a lot of data, the possibility of biased algorithms, and the difficulty of deciphering and understanding the behavior of complicated models. Addressing these issues and improving the state of the art in machine learning are the main goals of ongoing research. Machine learning consists of supervised, semi-supervised and unsupervised learning (Mao et al., 2019). The supervised learning machine-learning paradigm uses a collection of paired input-output training samples to discover the connection between a system's input and output. Given that the output is viewed as the input's label or oversight, an

*Corresponding Author: pradeep.yadav@itmgoi.in input-output training sample is also referred to as labelled training data (Verbraeken et al., 2020). Supervised learning consists of two things: regression and classification.

A supervised machine learning technique for predicting continuous values is regression. The final goal of the regression process is to draw the line or curve that best fits the data. Regression models map the input domain into a real-value domain. Classification is another technique of supervised learning used to map the input with predefined classes (Choi and Lim, 2020; Mishra et al., 2004).

Regression is of different types, which are discussed as follows:

Simple Linear Regression

Linear regression (Sudhaman et al., 2022) aims to reveal the relationship between two variables. One variable is supposed to be independent, while the other is supposed to be dependent. Simple regression separates the influence of independent factors from the interface of dependent variables. This linear regression shown in eq.1 is also known as the population regression function.

$s = \beta_0 + \beta_1 t + \varepsilon$ …………………………………(1)

Where β_0 and β_1 are estimates and ε is the error term.

Multivariate Linear Regression

Multivariate linear regression (Maulud and Abdulazeez, 2020) is a supervised learning algorithm that involves multiple input independent variables and one dependent variable.

$s = \beta_0 + \beta_1.t_1 + \beta_2.t_2 + \ldots.. + \beta n.tn + \varepsilon$ …………(2)

It is a technique for simulating the interaction between several independent variables i.e. t_1 to t_n and a dependent variable s, while assuming a linear relationship. It can be applied to both models and anticipate how different variables would affect a dependent variable.

When the relationship between the variables is expected to be approximately linear, multivariate regression is often used, whereas polynomial regression is used when the relationship is expected to be nonlinear. However, the method to use is ultimately determined by the specific problem and the nature of the data.

Polynomial Linear Regression

In this regression, the relationship between the independent variable t and the dependent variable s is handled (Tabelini et al., 2021) as an nth-degree polynomial in t. Polynomial regression (Jie and Zheng, 2019) can fit a nonlinear relationship between the value of t and the associated conditional mean of s.

$s = \beta_0 + \beta_1 t + \beta_2 t\ 2 + \cdots + \beta_h t^{h} + \varepsilon$ ………………(3)

Where h is the polynomial degree

The analyst can determine the degree of the polynomial function based on the complexity of the relationship between the dependent and independent variables. A degree 2 polynomial, for example, would fit a quadratic relationship between the variables, whereas a degree 3 polynomial would fit a cubic relationship.

During training, the polynomial regression model employs an optimization algorithm to determine the coefficient values that best fit the training data. Ordinary least square is the most commonly used algorithm, which minimizes the sum of the squared differences between the dependent variable's actual value and its anticipated value.

It should, however, be used with caution because higher-degree polynomials can overfit the training data, resulting in poor simplification of new data. It is used to model complex nonlinear relationships between variables in many fields, including finance, engineering, and social sciences. After training, the polynomial regression model will be used to predict new data by inputting the values of the independent variable(s) and using the model to compute the corresponding predicted value of the dependent variable.

Logistic Regression

Logistic regression models are commonly used to investigate how various predictors impact categorical outcomes. For binary outcomes, such as the existence or lack of a disease, a binary logistic model is appropriate. If the model includes just one predictor variable, it is known as a logistic regression model. On the other hand, if the model involves multiple interpreters, such as categorical and continuous variables, it is mentioned as a multivariable logistic regression (Khadhouri et al., 2022).

A logistic function (also known as the sigmoid function) is used in the logistic regression model to convert a linear combination of predictor variables into a probability value between 0 and 1. The logistic regression equation is as follows:

T = 1 / (1 + e^(-n))………………………… (4)

Where: The predicted probability of the dependent variable having the value 1 is given by T.

The direct combination of the predictor variables and their coefficients is denoted by n, which can be written as:

n = β0 + β1t1 + β2t2 + ... + βntn……………...(5)

Where: β0 is the intercept or bias term; β1, β2, ..., βn are the coefficients or weights of the predictor variables t1, t2, ..., tn.

DOI: https://doi.org/10.52756/ijerr.2024.v38.012

Formerly trained, the logistic regression model can be used to predict new data by inputting the values of the predictor variables and using the model to compute the corresponding probability of the dependent variable with the value 1. To convert the probability value into a binary classification decision, a threshold value can be set. The threshold value is typically set to 0.5, but it can be adjusted to achieve the desired balance of precision and recall.

Ridge Regression

Ridge regression is a type of regularized linear regression that is commonly used in machine learning and statistical modelling. It is employed when there are many predictor variables (sometimes referred to as features) in comparison to the number of observations or when the predictor variables have a high degree of correlation.

Ridge regression is a system that comprises adding a penalty term to the cost function of ordinary least squares (OLS) regression. This cost function minimizes the squared difference between the actual and predicted values. The added penalty term is based on the L2-norm of the regression coefficients, which encourages the coefficients to be smaller and helps prevent overfitting of the model.

The ridge regression model is formulated as:

$S = t\beta + \varepsilon$……………………………………(6)

Where S is the dependent variable, t represents the predictor variable matrix, β represents the vector of regression coefficients, and ε stands for the error term. The OLS cost function is augmented with a penalty term λ||β||^2, where λ is a hyperparameter that controls the strength of the penalty and ||β||^2 is the L2-norm of the coefficient vector. Ridge regression was first proposed by Arthur Hoerl and Robert Kennard (Hoerl and Kennard, 1970) in 1970. Since then, it has become a popular tool for dealing with high-dimensional data in a variety of fields, including economics, finance, engineering, and bioinformatics.

Lasso regression

Lasso regression is another type of regularized linear regression that is used to address overfitting and feature selection. It stands for "Least Absolute Shrinkage and Selection Operator" and was coined by (Tibshirani, 1996). In this, a penalty term is added to the OLS cost function, like Ridge regression. However, instead of using the L2-norm of the coefficient vector, lasso uses the L1-norm. This leads to a sparse solution where some of the coefficients are exactly zero, effectively performing feature selection.

The lasso regression model is formulated as:

$S = t\beta + \varepsilon$……………………………………(7)

Where S and t is the dependent variable and the matrix of predictor variables respectively, β is the vector of regression coefficients, and ε is the error term. The OLS cost function is augmented with a penalty term λ||β||_1, where λ is a hyperparameter that controls the strength of the penalty and ||β||_1 is the L1-norm of the coefficient vector.

Lasso regression has found applications in various fields, including finance, genetics, and computer vision.

Poisson Regression

To model count data, a type of generalized linear model (GLM) known as Poisson regression is often used. This approach assumes that the response variable follows a Poisson distribution (Joe and Zhu, 2005) with the predictor variables affecting the distribution's mean.

The Poisson regression model can be expressed as:

$\log(E(S \mid T)) = \beta0 + \beta1T1 + \beta2T2 + ... + \beta kTk$….(8)

where E(S | T) is the expected value of S given T and S is the response variable T is a vector of predictor variables, is a vector of coefficients, and The natural logarithm (log) function is the link function used in Poisson regression, which warranties that the predicted values are not negative. The Poisson regression model is frequently used to represent count data, such as the number of events, occurrences, or observations in a particular time or region, in disciplines including epidemiology, biology, and social sciences.

Stepwise Regression

An approach for choosing a selection of predictor variables to include in a linear regression model is stepwise regression. Depending on the significance of each variable, it can be carried out either forwards or backwards, adding or eliminating each one one at a time. By avoiding overfitting, the objective is to determine the most significant predictors. Using a criterion like the F-test or AIC, the forward stepwise regression approach starts with an empty model and adds variables one at a time based on their importance. Starting with a complete model, the backward stepwise regression approach eliminates variables one at a time according to their relevance.

Mathematically, the forward stepwise regression (Chen et al., 2014) method can be expressed as follows:

1. Start with an empty model: $S = \beta0 + \varepsilon$
2. For each predictor variable Ti, fit the model: $S = \beta0 + \beta iTi + \varepsilon$
3. Choose the variable Ti that results in the highest F-statistic or lowest AIC value

DOI: https://doi.org/10.52756/ijerr.2024.v38.012

4. Add the variable Ti to the model: S = β0 + βiTi + βjTj + ε
5. Repeat steps 2-4 until no variable can be added to the model

The backward stepwise regression method can be expressed as follows:

1. Begin with a full model: S = β0 + β1T1 + β2T2 + ... + βkTk + ε
2. For each predictor variable Ti, fit a model without that variable: S = β0 + β1T1 + ... + βi-1Ti-1 + βi+1Ti+1 + ... + βkTk + ε
3. Select the variable Ti that yields the lowest F-statistic or the highest AIC value
4. Remove the variable Ti from the model: Y = β0 + β1T1 + ... + βi-1Ti-1 + βi+1Ti+1 + ... + βkTk + ε
5. Repeat steps 2-4 until no variable can be removed from the model.

Stepwise regression contains constraints and underlying assumptions that should be thoroughly examined before using it to choose significant predictors. Stepwise regression can either be employed in addition to or in place of other variable selection techniques like regularization or model averaging.

Multilevel Regression

Multilevel regression is a statistical method for analyzing data that has an ordered or nested structure, such as students nested within schools, personnel nested within groups or patients nested within hospitals. It is also known as hierarchical linear modelling or mixed-effects modelling. By modelling the variation at each level of the hierarchy and predicting the associations between variables at each level, multilevel regression takes into consideration the hierarchical structure of the data.

When examining data with nested structures, multilevel regression is an effective method that can give important insights into the correlations between variables at different levels of the hierarchy (Bosker and Snijders, 2012).

Quantile Regression

Given the predictor variables, quantile regression calculates the conditional quantile function of the response variable. It is said that the conditional quantile function is:

$Q(s|t) = \inf \{q: P(s <= q \mid t) >= \tau\}$ …………….(9)

Where s is the response variable, t is the predictor variable (s), τ is the quantile of interest (e.g., τ=0.5 for the median), and Q(s|t) is the value of the response variable at the τth quantile given the predictor variables.

The quantile regression (Geraci and Bottai, 2007) estimator minimizes the following objective function:

$\sum_i [\tau - I(s_i <= t_i\beta)](\rho(s_i - t_i\beta))$…………………(10)

Where (I) is the indicator function, is a vector of regression coefficients, and ρ(u)controls how the residuals are weighted. Based on the required characteristics of the estimator, the function ρ(u) can be selected.

Bayesian Regression

A statistical technique for simulating relation-nships between factors is called Bayesian regression. Bayesian regression offers a means to include prior knowledge or beliefs about the variables in the model, unlike conventional regression techniques.

Assuming a linear regression model with a regularly distributed error structure, response variable y, and predictor variable x, we can write:

s_i = β0 + β1*t_i + epsilon_i ……………….(11)

Where s_i is the observed response for the ith observation, t_i is the corresponding predictor value, β0 and β1 are the intercept and slope coefficients to be estimated, and epsilon_i are the error terms assumed to be normally distributed with mean 0 and variance sigma^2.

In Bayesian regression (Emami et al.,2018) we specify prior distributions for the model parameters β0, β1, and sigma^2, and update them based on the observed data using Bayes' theorem.

Specifically, we have:

p(β0, β1, sigma^2 | y, x) = p(y | β0, β1, sigma^2, x) * p(β0, β1, sigma^2) ……………………………..(12)

Where, p(y | β0, β1, sigma^2, x) is the likelihood function of the data, which specifies the probability of observing the data given the model parameters, and p(β0, β1, sigma^2) is the prior distribution of the parameters.

Metrics used in Regression

Explanation of each metric commonly used to assess regression models:

Mean Squared Error and Root Mean Squared Error

The MSE is the mean squared error between the actual number and the predicted value. A smaller MSE (James et al., 2013) indicates a better fit of the model.

$MSE = 1/n * \Sigma (y_i - \bar{y})^2$…………………….(13)

where n signifies the numeral of observations, y_i signifies the expected value for observation i and $\bar{y}$ signifies the average of the actual values.

$RMSE = \sqrt{MSE}$……………………………..(14)

DOI: https://doi.org/10.52756/ijerr.2024.v38.012

Table 1. Shows the detailed comparison of the discussed regression

Regression Model	Dependent Variable	Independent Variables	Type of Relationship	Model Complexity	Assumptions	Suitable for	Strengths	Weaknesses	Suitable data size	References
Linear Regression	Continuous	Continuous	Linear	Low	Normality, Homoscedasticity, Linearity	Correlation Analysis	Easy to interpret, Good for exploring relationships	Assumes linear relations-hip, Not suitable for non-linear data	Large	Azur et al., 2011; Siemsen et al., 2010
Multiple Regression	Continuous	Continuous or Categorical	Linear	Medium	Normality, Homoscedasticity, Linearity, No multicollinearity	Multiple Predictor Analysis	Can include multiple predictors, Improved model accuracy	Assumes linear relationship, Not suitable for non-linear data	Large	Mason and Perreault, 1991; Uyanık and Güler, 2013
Polynomial Regression	Contin-uous	Continuous	Non-linear	Medium	Normality, Homoscedasticity, Linearity	Non-linear Regression	Can model non-linear data, Improved model accuracy	Can lead to overfitting, Requires choosing the correct degree	Medium to large	Ostertagová, 2012; Maulud and Abdulazeez, 2020; Li and Yamamoto, 2016
Logistic Regression	Binary or Categorical	Continuous or Categorical	Login	Low	Independence of observations, Linearity, No multicollinearity	Binary Classification	Good for modeling probabilities, Easily interpretable	Assumes linear relationship, Not suitable for non-linear data	Large	Hosmer et al., 2013; DeMaris et al., 2013; Christodoulou et al., 2019
Ridge Regression	Contin-uous	Continuous	Linear	Medium	Normality, Homoscedasticity, Linearity, No multicollinearity	Reducing Multicollinearity	Reduces multicollinearity, Can improve model stability	Bias-variance tradeoff, Requires tuning parameter	Medium to large	Shigeto et al., 2015; Li and Niu, 2013

DOI: https://doi.org/10.52756/ijerr.2024.v38.012

Regression Model	Dependent Variable	Independent Variables	Type of Relationship	Model Complexity	Assumptions	Suitable for	Strengths	Weaknesses	Suitable data size	References
Lasso Regression	Continuous	Continuous	Linear	Medium	Normality, Homoscedasticity, Linearity, No multicollinearity	Feature Selection	Can select important predictors, Improved model interpretability	Biased towards selecting a small number of predictors, Not well-suited for data with high multicollinearity	Medium to large	Muthukrishnan and Rohini, 2016; Emmert-Streib and Dehmer, 2019
Poisson Regression	Count	Continuous or Categorical	Exponential	Low	Independence of obse-rvations, Linearity, No multicollinearity	Count Data Analysis	Suitable for modeling count data, Easy to interpret	Assumes exponential distribution, Not suitable for under- or over-dispersed data	Large	Jia et al., 2017; El-Gabbas and Dormann, 2018
Stepwise Regression	Continuous	Continuous or Categorical	Linear or Non-linear	High	Normality, Homoscedasticity, Linearity, No multicollinearity	Feature Selection	Can select important predictors, Improved model interpretability	Biased towards selecting a small number of predictors, May not select the best predictors	Medium to large	Shanableh and Assaleh, 2010; Mohsenijam et al., 2017
Multilevel Regression	Continuous	Continuous or Categorical	Linear	High	Normality, Homoscedasticity, Linearity, No multicollinearity	Hierarchical Data Analysis	Can model nested data, Improved model accuracy	Can be computationally intensive, Requires large sample sizes	Medium to large	Rácz et al., 2019; Muthén et al. 2011
Quantile Regression	Continuous	Continuous or Categorical	Non-linear	Medium	Normality, Homoscedasticity	Robust Regression	Can model non-linear data, Resistant to outliers	Requires choosing the correct quantile, May be less interpretable	Medium to large	Romano et al., 2019; Yang et al., 2013

The same units are used to express the dependent variable and the RMSE (Wang and Lu, 2018) which is the square root of the MSE. Both metrics penalize large errors more heavily than small errors.

Mean Absolute Error

The MAE is the average absolute alteration between the expected and real values. Like the MSE and RMSE, a lower MAE (De Myttenaere et al., 2016) indicates a greater fit of the model. Because it does not square the errors, the MAE is less susceptible to outliers than the MSE and RMSE.

$\text{MAE} = 1/n * \Sigma |y_i - \hat{y}_i|$ ……………………..(15)

R-squared (R^2) and Adjusted R-squared (R^2)

How well the model accounts for the deviation in the dependent variable is determined by its R-squared (Gelman et al., 2019) value. Values between 0 and 1 indicate the goodness of fit, with higher values suggesting a better fit. The adjusted R-squared penalizes the model for having too many variables and is useful for relating models with different numbers of predictors.

$R^2 = 1 - (SS_{res} / SS_{rt})$ ……………………….(16)

Where SS_{res} represents the summation of squares of residuals or the difference between anticipated and actual values, and SS_{rt} represents the sum of squares of all the values (the change between the actual values and the mean of the actual values).

To account for the numeral of predictors in a model, a modified version of R-squared known as adjusted R-squared is often used:

Adjusted $R^2 = 1 - [(1 - R^2) * (n - 1) / (n - p - 1)]$…(17)

Where p is the numeral of predictors in the model.

Mean Absolute Percentage Error

The MAPE is the normal of the total percentage differences between the expected and real values (Makridakis et al., 2018; De Myttenaere et al., 2016). It is expressed as a percentage and is useful for evaluating models in which the scale of the variable is important. The MAPE is sensitive to small values and can become undefined if the actual value is zero.

$\text{MAPE} = 100/n * \Sigma |(y_i - \hat{y}_i)/y_i|$ ……………..(18)

y_i is the expected value for the i^{th} observation, and $n\bar{y}$ is the mean of the actual values.

Coefficient of Determination

COD, which represents the square of the correlation coefficient between the predicted and actual data, is a metric of quality of fit. A better fit is indicated by a higher COD value, which ranges from 0 to 1. The COD is commonly employed in industries like banking and economics even if it is less understandable than R-squared (Chicco et al., 2021; Schober et al., 2018).

$\text{COD} = r^2$ ……………………………………(19)

where r is the correlation coefficient.

Akaike Information Criterion and Bayesian Information Criterion

AIC and BIC are measures that compare the quality of a model to that of other models. These metrics consider both the model's goodness of fit and its complexity. A lower value of AIC (Vrieze, 2012) or BIC (Acquah et al., 2010) indicates a better fit, with AIC being more severe in penalizing overfitting.

They are calculated as follows:

$\text{AIC} = -2 \ln(J) + 2r$ …………………………(20)

$\text{BIC} = -2 \ln(J) + r \ln(n)$ ……………………(21)

In the formula, J represents the likelihood of the data given the model, r is the number of parameters in the model, and n is the no. of observations.

Mean Forecast Error

The average of the discrepancies between the predicted and real values is known as the MFE. In contrast to the other metrics, a smaller MFE (De Myttenaere et al., 2016) is not always superior because it ignores the direction of the errors.

$\text{MFE} = 1/n * \Sigma (y_i - \hat{y}_i)$ ……………………(22)

Methodology

Here breast cancer datasets have been used for research some of which may be more accurate or representative of real-world scenarios than others. Here are a few examples:

The SEER Dataset:

The National Cancer Institute's surveillance, epidemiology and end results (SEER) initiative compiles information on cancer patients in the country. The SEER dataset (Ahmed et al., 2023) contains statistics on patient demographics, cancer stage and treatment, as well as survival rates for people with breast cancer.

The TCGA Dataset:

The full form of TCGA is the cancer genome atlas, it is a program that collects genomic data and clinical information from multiple cancer types, including breast cancer. The TCGA (Dehkharghanian et al., 2023) breast cancer dataset includes information on gene expression, DNA mutations, and clinical outcomes for breast cancer patients.

The METABRIC Dataset:

The full form of METABRIC (Chen et al., 2023) is the molecular taxonomy of the breast cancer international consortium. It is a multi-centre study that collected gene expression data, clinical information, and survival

DOI: https://doi.org/10.52756/ijerr.2024.v38.012

outcomes for breast cancer patients. The dataset includes information on over 2,000 patients with primary breast cancer.

The ICGC Dataset:

The International Cancer Genome Consortium (ICGC) is a collaborative project that aims to collect genomic data and clinical information on multiple cancer types, including breast cancer. The ICGC (He et al., 2023) breast cancer dataset includes information on DNA mutations, gene expression, and clinical outcomes for breast cancer patients.

In our work, the breast cancer Wisconsin dataset (Nemade et al., 2023) is taken, which is one of the most commonly used breast cancer datasets. This dataset has twelve features and 569 instances. Other versions of this dataset have additional attributes or slightly different attribute names. Id_number, radius, diagnosis, area, texture, perimeter, compactness, concave points, smoothness, concavity, symmetry, and fractal dimension are the features of this dataset.

Following are the steps for model building:

Preprocess the Data:

Data preprocessing is a vital step in machine learning because it can increase the precision and dependability of the final model. Here the dataset consists of 569 instances and 12 attributes, a detailed explanation of each step in preprocessing the Breast Cancer Wisconsin (Diagnostic) dataset is given below:

Importing Dataset:

Importing the dataset is the first stage. A dataset with 569 instances and 12 columns is obtained from the UCI Machine Learning Repository.

Splitting the Dataset into Labels and Features:

A label (output variable) in the dataset shows whether the mass was malignant or benign, and features (input variables) in the dataset are measurements of various characteristics of breast mass samples. Features will be separated from the label before applying machine-learning algorithms.

Handling Missing Values:

It's essential to figure out whether the dataset contains any missing values. There are different approaches to handling missing values. Here instead of dropping the rows with the missing values are imputed with mean, median, and mode.

Encoding Categorical Data:

Some features are categorical, such as the diagnosis (M or B). These are encoded in numerical values before applying the machine-learning algorithm. One popular method for encoding categorical data is one-hot encoding, which creates a new column for each possible value of the categorical variable.

The general architecture of the preprocessing and model building is shown in Figure 1.

Training and Testing:

Two sets—the training set and the testing set—are produced once the data has been preprocessed. The testing set is used to evaluate the machine learning model's performance, while the training set is used to train the model. Here, 80% of the data are used for training and 20% are used for assessment.

Model Building:

With different machine learning algorithms like decision trees, random forests, support vector machines, and neural networks, the Wisconsin dataset is typically used for classification tasks. But here regression algorithms are used to predict continuous variables (radius and area of the breast mass). These continuous variables are included as features in the dataset and are related to the malignancy of the mass. By Using regression algorithms, the radius or area of a breast mass will be predicted.

Performance Evaluation:

Instead of using classification metrics like accuracy, precision, recall and F1 score, it is important to assess the performance of the regression model using suitable metrics like mean squared error, R-squared, mean absolute error, coefficient of determination and mean forecast error.

Results & Discussion

We conducted a regression analysis on the breast cancer dataset using the different regression algorithms, implemented in Python 3.9.4. The analysis was run on a Dell XPS 13 laptop with an Intel Core i7-1165G7 processor and 16 GB of RAM. perimeter, and compactness. In Table 2 the target variable is radius. Here gradient boosting regression shows less MSE value as shown in Figure 2.

DOI: https://doi.org/10.52756/ijerr.2024.v38.012

Table 2. Comparison between different regressions using radius as a target variable

Target Variable	Regression Type	MSE	R-squared	MAE	COD	MFE
Radius	Linear Regression	0.08	0.78	0.21	0.77	0.00
Radius	Ridge Regression	0.09	0.77	0.21	0.76	-0.001
Radius	Lasso Regression	0.13	0.66	0.27	0.65	0.00
Radius	Elastic Net	0.11	0.71	0.25	0.70	0.00
Radius	Decision Tree Regression	0.15	0.60	0.30	0.59	0.00
Radius	Random Forest Regression	0.07	0.82	0.17	0.82	0.00
Radius	Gradient Boosting Regression	0.05	0.89	0.14	0.88	0.00

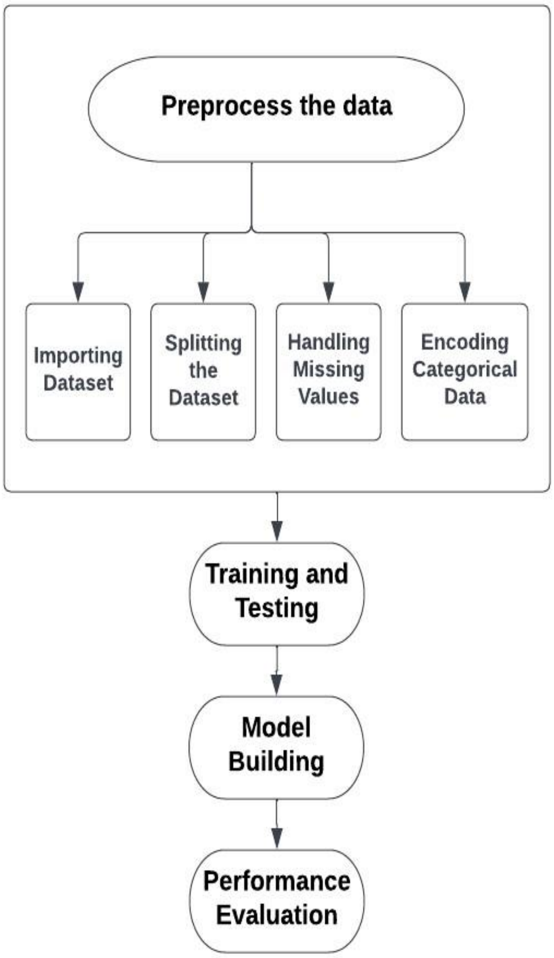

Figure 1. Comparison between different regressions using radius as a target variable.

DOI: https://doi.org/10.52756/ijerr.2024.v38.012

Table 3. Comparison between different regressions using perimeter as a target variable.

Target Variable	Model	MSE	R-squared	MAE	COD	MFE
Perimeter	Linear Regression	0.24	0.54	0.37	0.53	0.00
Perimeter	Ridge Regression	0.25	0.52	0.38	0.52	0.001
Perimeter	Lasso Regression	0.37	0.29	0.45	0.27	0.00
Perimeter	Elastic Net	0.30	0.42	0.39	0.40	0.00
Perimeter	Decision Tree Regression	0.31	0.40	0.40	0.39	0.00
Perimeter	Random Forest Regression	0.15	0.66	0.27	0.65	0.00
Perimeter	Gradient Boosting Regression	0.11	0.73	0.23	0.72	0.00

Table 4. Comparison between different regressions using compactness as a target variable

Target Variable	Model	MSE	R-squared	MAE	COD	MFE
Compactness	Linear Regression	0.28	0.46	0.41	0.45	0.00
Compactness	Ridge Regression	0.28	0.45	0.41	0.44	0.00
Compactness	Lasso Regression	0.38	0.30	0.46	0.28	0.00
Compactness	Elastic Net	0.32	0.38	0.43	0.37	0.00
Compactness	Decision Tree Regression	0.44	0.17	0.50	0.15	0.00
Compactness	Random Forest Regression	0.22	0.63	0.32	0.62	0.00
Compactness	Gradient Boosting Regression	0.18	0.69	0.28	0.69	0.00

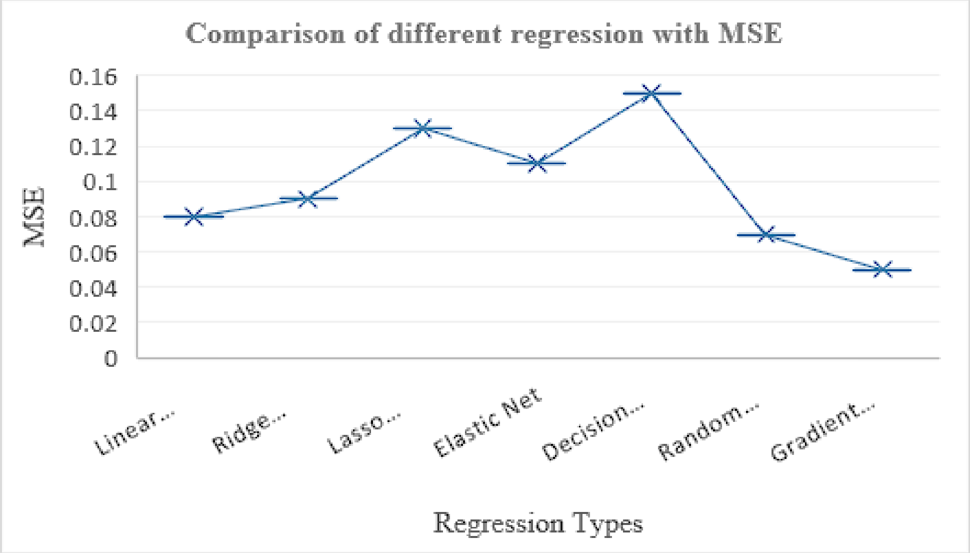

Figure 2. Comparison of different regressions with MSE when the target variable is a radius.

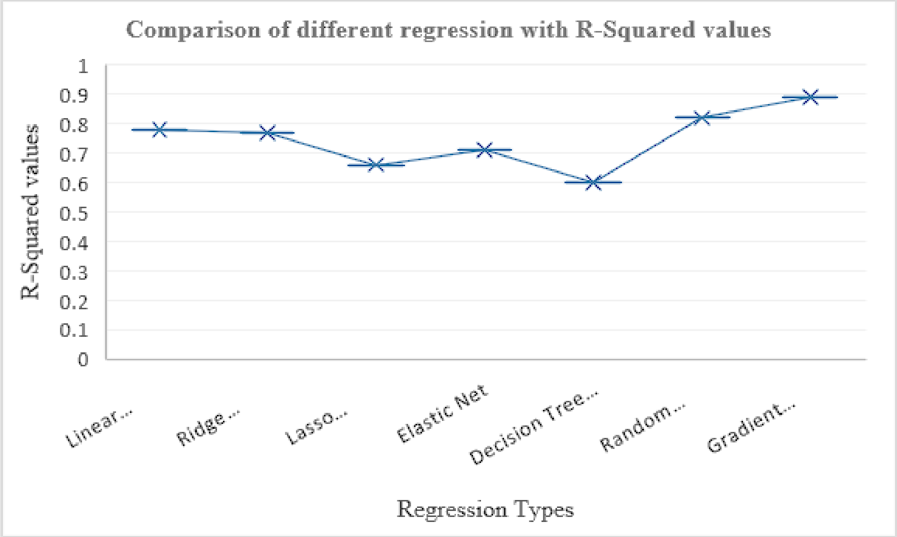

Figure 3. Comparison of different regression with R-squared values when the target variable is a radius

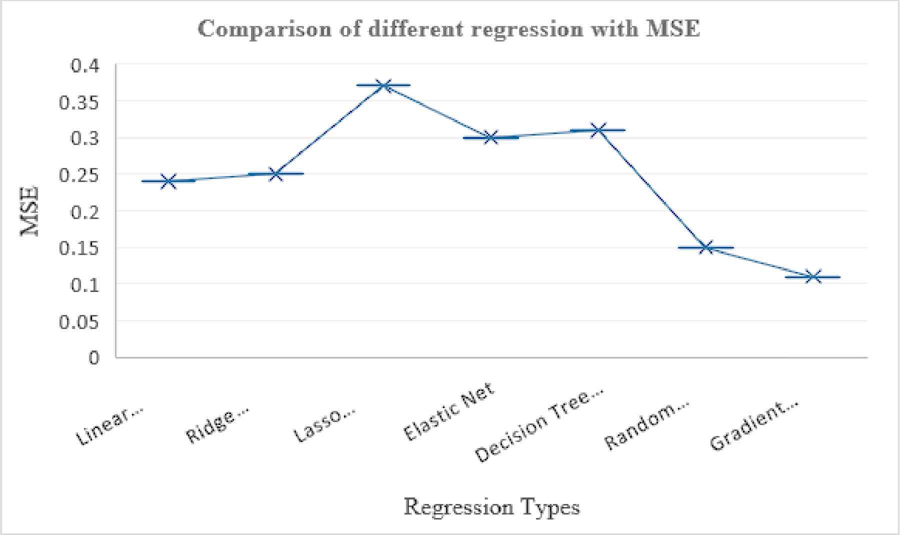

Figure 4. Comparison of different regression with MSE when the target variable is a perimeter.

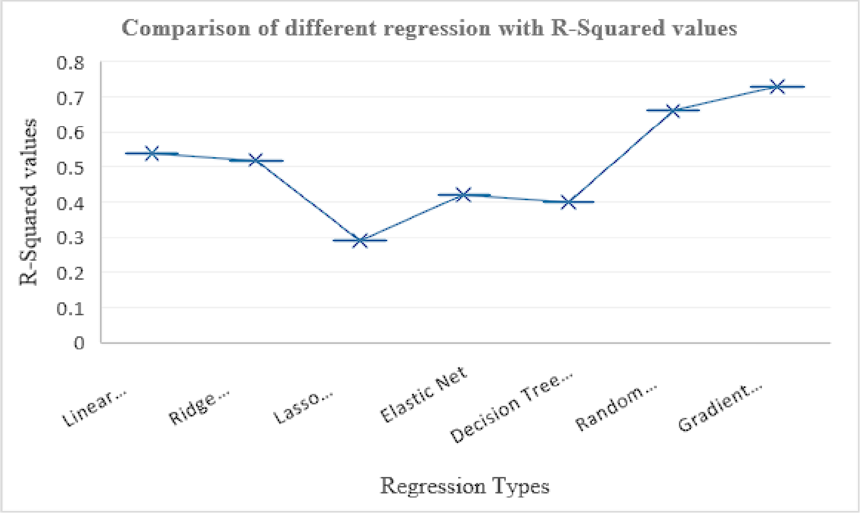

Figure 5. Comparison of different regression with R-Squared values when the target variable is a perimeter.

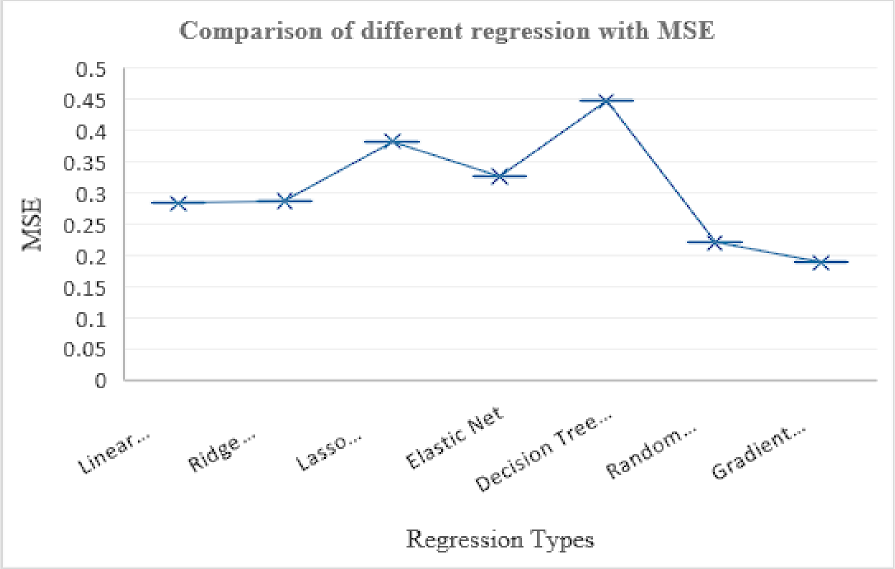

Figure 6. Comparison of different regression with MSE when the target variable is a compactness.

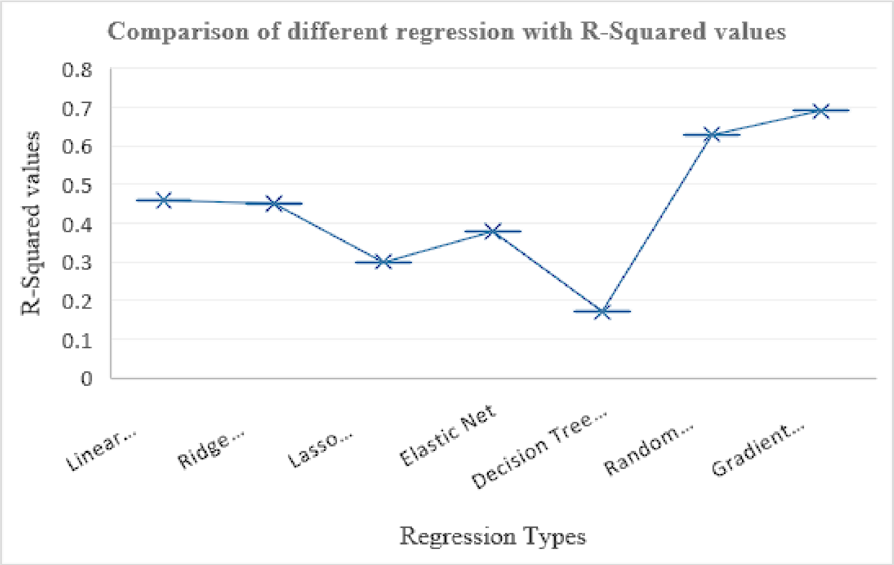

Figure 7. Comparison of different regression with MSE when the target variable is a compactness.

Conclusion

We have applied six different regression models on the breast cancer dataset using various continuous variables as the target variable. The Random Forest and Gradient Boosting Regression models consistently outperformed the other models in terms of their mean squared error, R-squared, and mean absolute error.

For example, when using 'radius' as the target variable, the Random Forest Regression model achieved an MSE of 0.07, R-squared of 0.82, and MAE of 0.17, while the Gradient Boosting Regression model achieved an MSE of 0.05, R-squared of 0.89, and MAE of 0.14. In contrast, the other models achieved higher MSE and lower R-squared values, indicating that they were not as effective at capturing the underlying relationships between the predictors and target variables.

We also calculated the Coefficient of Determination (Adj R-squared) to account for the number of predictors used in each model. This provided a more accurate measure of the model's performance, especially when comparing models with different numbers of predictors. Gradient Boosting Regression models consistently achieved higher Adj R-squared values across multiple target variables, indicating that they can better capture the variation in the data.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

DOI: https://doi.org/10.52756/ijerr.2024.v38.012

Acknowledgements

We would like to thank the management of ITM, Gwalior for providing us the opportunity to carry out this research.

References

Acquah, H. D. G. (2010). Comparison of Akaike information criterion (AIC) and Bayesian information criterion (BIC) in selection of an asymmetric price relationship. *Journal of Development and Agricultural Economics, 2*(1), 001-006.

Ahmed, A., Whittington, J., & Shafaee, Z. (2023). Impact of Commission on Cancer Accreditation on Cancer Survival: A Surveillance, Epidemiology, and End Results (SEER) Database analysis. *Annals of Surgical Oncology*, *31*(4), 2286–2294. https://doi.org/10.1245/s10434-023-14709-4

Azur, M., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Multiple imputation by chained equations: what is it and how does it work? *International Journal of Methods in Psychiatric Research, 20*(1), 40–49. https://doi.org/10.1002/mpr.329

Chen, R., Cai, N., Luo, Z., Wang, H., Liu, X., & Li, J. (2023). Multi-task banded regression model: A novel individual survival analysis model for breast cancer. *Computers in Biology and Medicine, 162*, 107080. https://doi.org/10.1016/j.compbiomed.2023.107080

Chen, S., Goo, Y. J. J., & Shen, Z. D. (2014). A hybrid approach of stepwise regression, logistic regression, support vector machine, and decision tree for forecasting fraudulent financial statements. *The Scientific World Journal, 2014*, 1-9. https://doi.org/10.1155/2014/968712

Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *Peerj Computer Science*, *7*, e623. https://doi.org/10.7717/peerj-cs.623

Choi, J. A., & Lim, K. (2020). Identifying machine learning techniques for classification of target advertising. *ICT Express*, *6*(3), 175-180. https://doi.org/10.1016/j.icte.2020.04.012

Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., & Van Calster, B. (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology*, *110*, 12-22. https://doi.org/10.1016/j.jclinepi.2019.02.004

De Myttenaere, A., Golden, B., Le Grand, B., & Rossi, F. (2016). Mean absolute percentage error for regression models. *Neurocomputing*, *192*, 38-48. https://doi.org/10.1016/j.neucom.2015.12.114

Dehkharghanian, T., Bidgoli, A. A., Riasatian, A., Mazaheri, P., Campbell, C. J., Pantanowitz, L., Tizhoosh, H. R., & Rahnamayan, S. (2023). Biased data, biased AI: deep networks predict the acquisition site of TCGA images. *Diagnostic Pathology*, *18*(1). https://doi.org/10.1186/s13000-023-01355-3

DeMaris, A., & Selman, S. H. (2013). Converting data into evidence. A statistics primer for the medical practitioner. In Springer eBooks, New York. https://doi. org/10.1007/978-1-4614-7792-1.

El-Gabbas, A., & Dormann, C. F. (2018). Improved species-occurrence predictions in data-poor regions: using large-scale data and bias correction with down-weighted Poisson regression and Maxent. *Ecography*, *41*(7), 1161-1172. https://doi.org/10.1111/ecog.03149

Emami, N. P., Degeling, M., Bauer, L., Chow, R., Cranor, L. F., Haghighat, M. R., & Patterson, H. (2018). The influence of friends and experts on privacy decision making in IoT scenarios. *Proceedings of the ACM on Human-Computer Interaction*, *2*(CSCW), 1-26. https://doi.org/10.1145/3274317

Emmert-Streib, F., & Dehmer, M. (2019). High-dimensional LASSO-based computational regression models: regularization, shrinkage, and selection. *Machine Learning and Knowledge Extraction*, *1*(1), 359-383. https://doi.org/10.3390/make1010021

Gelman, A., Goodrich, B., Gabry, J., & Vehtari, A. (2019). R-squared for Bayesian regression models. The American Statistician, 73(3), 307–309. https://doi.org/10.1080/00031305.2018.1549100

Geraci, M., & Bottai, M. (2007). Quantile regression for longitudinal data using the asymmetric Laplace distribution. *Biostatistics*, *8*(1), 140-154. https://doi.org/10.1093/biostatistics/kxj039

He, B., Sun, H., Bao, M., Li, H., He, J., Tian, G., & Wang, B. (2023). A cross-cohort computational framework to trace tumor tissue-of-origin based on RNA sequencing. *Scientific Reports*, *13*(1), 15356. https://doi.org/10.1038/s41598-023-42465-8

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55-67. https://doi.org/10.1080/00401706.1970.10488634

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression. John Wiley & Sons. https://doi.org/10.1002/9781118548387

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning. *In Springer Texts in Statistics, 112*, 18. https://doi.org/10.1007/978-1-0716-1418-1

Jia, Y., Kwong, S., Wu, W., Wang, R., & Gao, W. (2017). Sparse Bayesian learning-based kernel Poisson regression. *IEEE Transactions on Cybernetics*, *49*(1), 56-68. https://doi.org/10.1109/TCYB.2017.2764099

Jie, H., & Zheng, G. (2019). Calibration of Torque Error of Permanent Magnet Synchronous Motor Base on Polynomial Linear Regression Model. In IECON 2019-45th Annual Conference of the IEEE Industrial Electronics Society (Vol. 1, pp. 318-323). IEEE. https://doi.org/10.1109/IECON.2019.8927537

Joe, H., & Zhu, R. (2005). Generalized Poisson Distribution: the Property of Mixture of Poisson and Comparison with Negative Binomial Distribution. *Biometrical Journal*, *47*(2), 219–229. https://doi.org/10.1002/bimj.200410102

Khadhouri, S., Gallagher, K., MacKenzie, K., Shah, T. T., Gao, C., Moore, S., Zimmermann, E., Edison, E., Jefferies, M., Nambiar, A., Anbarasan, T., Mannas, M., Lee, T., Marra, G., Rivas, J. G., Marcq, G., Assmus, M., Uçar, T., Claps, F., . . . Zainuddin, Z. M. (2022). Developing a Diagnostic Multivariable Prediction Model for Urinary Tract Cancer in Patients Referred with Haematuria: Results from the IDENTIFY Collaborative Study. *European Urology Focus*, *8*(6), 1673–1682. https://doi.org/10.1016/j.euf.2022.06.001

Li, G., & Niu, P. (2013). An enhanced extreme learning machine based on ridge regression for regression. Neural Computing and Applications, 22, 803-810. https://doi.org/10.1007/s00521-011-0771-7

Li, H., & Yamamoto, S. (2016). Polynomial regression based model-free predictive control for nonlinear systems. In 2016 55th annual conference of the society of instrument and control engineers of Japan (SICE) (pp. 578-582). IEEE. https://doi.org/10.1109/SICE.2016.7749264

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). The M4 Competition: Results, findings, conclusion and way forward. *International Journal of Forecasting*, *34*(4), 802-808. https://doi.org/10.1016/j.ijforecast.2018.06.001

Mao, X., Yang, H., Huang, S., Liu, Y., & Li, R. (2019). Extractive summarization using supervised and unsupervised learning. *Expert systems with applications*, *133*, 173-181. https://doi.org/10.1016/j.eswa.2019.05.011

Mason, C. H., & Perreault Jr, W. D. (1991). Collinearity, power, and interpretation of multiple regression analysis. *Journal of Marketing Research, 28*(3), 268-280. https://doi.org/10.1177/002224379102800302

Maulud, D., & Abdulazeez, A. M. (2020). A review on linear regression comprehensive in machine learning. *Journal of Applied Science and Technology Trends*, *1*(2), 140-147. https://doi.org/10.38094/jastt1457

Mohsenijam, A., Siu, M. F. F., & Lu, M. (2017). Modified stepwise regression approach to streamlining predictive analytics for construction engineering applications. *Journal of Computing in Civil Engineering*, *31*(3), 04016066. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000636

Muthén, B., & Asparouhov, T. (2011). Beyond multilevel regression modeling: Multilevel analysis in a general latent variable framework. In Handbook of advanced multilevel analysis (pp. 15-40). Routledge. https://doi.org/10.4324/9780203848852

Muthukrishnan, R., & Rohini, R. (2016). LASSO: A feature selection technique in predictive modeling for machine learning. In 2016 IEEE international conference on advances in computer applications (ICACA) (pp. 18-20). IEEE. https://doi.org/10.1109/ICACA.2016.7887916

Nemade, V., & Fegade, V. (2023). Machine learning techniques for breast cancer prediction. Procedia Computer Science, 218, 1314-1320. https://doi.org/10.1016/j.procs.2023.01.110

Ostertagová, E. (2012). Modelling using polynomial regression. *Procedia Engineering*, *48*, 500-506. https://doi.org/10.1016/j.proeng.2012.09.545

Rácz, A., Bajusz, D., & Héberger, K. (2019). Multi-level comparison of machine learning classifiers and their performance metrics. *Molecules*, *24*(15), 2811. https://doi.org/10.3390/molecules24152811

Romano, Y., Patterson, E., & Candès, E. J. (2019). Conformalized Quantile Regression. Neural Information Processing Systems.

Schober, P., Boer, C., & Schwarte, L. A. (2018). Correlation coefficients: appropriate use and interpretation. *Anesthesia & Analgesia*, *126*(5), 1763-1768. https://doi.org/10.1213/ANE.0000000000002864

Shanableh, T., & Assaleh, K. (2010). Feature modeling using polynomial classifiers and stepwise regression. *Neurocomputing*, *73*(10-12), 1752-1759. https://doi.org/10.1016/j.neucom.2009.11.045

Shigeto, Y., Suzuki, I., Hara, K., Shimbo, M., & Matsumoto, Y. (2015). Ridge regression, hubness, and zero-shot learning. In Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2015, Porto, Portugal, September 7-11, 2015, Proceedings, Part I 15 (pp. 135-151). Springer International Publishing. https://doi.org/10.1007/978-3-319-23528-8_9

Siemsen, E., Roth, A., & Oliveira, P. (2010). Common method bias in regression models with linear, quadratic, and interaction effects. *Organizational Research Methods*, *13*(3), 456-476. https://doi.org/10.1177/1094428109351241

Snijders, T. A., & Bosker, R. (2012). Multilevel analysis: An introduction to basic and advanced multilevel modeling. Second edition. London etc.: Sage Publishers, 2012

Sudhaman, K., Akuthota, M., & Chaurasiya, S. K. (2022). A Review on the Different Regression Analysis in Supervised Learning. Bayesian Reasoning and Gaussian Processes for Machine Learning Applications, pp.15-32.

Tabelini, L., Berriel, R., Paixao, T. M., Badue, C., De Souza, A. F., & Oliveira-Santos, T. (2021, January). Polylanenet: Lane estimation via deep polynomial regression. In 2020 25th International Conference on Pattern Recognition (ICPR) (pp. 6150-6156). IEEE. https://doi.org/10.1109/ICPR48806.2021.9412265

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *58*(1), 267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Uyanık, G. K., & Güler, N. (2013). A study on multiple linear regression analysis. *Procedia-Social and Behavioral Sciences*, *106*, 234-240. https://doi.org/10.1016/j.sbspro.2013.12.027

Verbraeken, J., Wolting, M., Katzy, J., Kloppenburg, J., Verbelen, T., & Rellermeyer, J. S. (2020). A survey on distributed machine learning. *ACM Computing Surveys (Esur)*, *53*(2), 1-33. https://doi.org/10.1145/3377454

Vrieze, S. I. (2012). Model selection and psychological theory: a discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychological methods*, *17*(2), 228. https://psycnet.apa.org/doi/10.1037/a0027127

Wang, W., & Lu, Y. (2018, March). Analysis of the mean absolute error (MAE) and the root mean square error (RMSE) in assessing rounding model. In IOP conference series: materials science and engineering (Vol. 324, p. 012049). IOP Publishing. https://doi.org/10.1088/1757-899X/324/1/012049

Yang, J., Meng, X., & Mahoney, M. (2013). Quantile regression for large-scale applications. In International Conference on Machine Learning. *Proceedings of the 30th International Conference on Machine Learning*, PMLR, *28*(3), 881-887.

How to cite this Article:

Pradeep Yadav, Chandra Prakash Bhargava, Deepak Gupta, Jyoti Kumari, Archana Acharya and Madhukar Dubey (2024). Breast Cancer Disease Prediction Using Random Forest Regression and Gradient Boosting Regression. *International Journal of Experimental Research and Review*, *38*, 132-146.

DOI : https://doi.org/10.52756/ijerr.2024.v38.012