IRAQI
Academic Scientific Journals

University of Kirkuk Journal For Administrative and Economic Science. A special issue of the 1st Scientific Conference of the College of Administration and Economics - University of Kirkuk - Information technology, digitization, and their impact on sustainable development - 8-9, Oct- 2025.

UKJAES

University of Kirkuk Journal For Administrative and Economic Science

ISSN:2222-2995 E-ISSN:3079-3521
University of Kirkuk Journal For Administrative and Economic Science

Hadba Wakaa Ali. Evaluation and comparison of different regression models and machine learning algorithms for prediction on the Swiss fertility dataset. ***University of Kirkuk Journal for Administrative and Economic Science*** **Volume (15) Issue (4) Part (2) Supplement (1) A special issue of the 1st Scientific Conference of the College of Administration and Economics - University of Kirkuk - Information technology, digitization, and their impact on sustainable development - 8-9, Oct- 2025, p-p:173-187.**

Evaluation and comparison of different regression models and machine learning algorithms for prediction on the Swiss fertility dataset

Wakaa Ali Hadba

University of Kirkuk - College of Administration and Economics - Department of Statistics, Kirkuk, Iraq

wakaa2017@uokirkuk.edu.iq

Abstract. In this paper, we present a comparative evaluation of several classical regression models and machine learning algorithms applied to the Swiss fertility dataset. The models evaluated include linear regression, ridge regression, Lasso regression, elastic net, principal component regression (PCR), partial least squares (PLS), and machine learning algorithms such as decision tree, k-nearest neighbour, random forest, support vector regression (SVR), and XGBoost. Performance metrics such as RMSE, MAE, R-square, AIC, and BIC were used to evaluate the models. Visual comparisons using bar charts enabled the identification of the most effective models. SVR outperformed other regression models in prediction accuracy, while Lasso regression showed the best model simplicity based on AIC/BIC. The results highlight the power of machine learning models in detecting nonlinear relationships in fertility prediction. datasets.

Keywords: Swiss Dataset, Fertility Prediction, Machine Learning, Statistical Modelling, Support Vector Regression (SVR), Regularization Methods,

تقييم ومقارنة نماذج الانحدار المختلفة وخوارزميات التعلم الآلي للتنبؤ بمجموعة بيانات الخصوبة السويسرية

أ.م.د. وكاع علي هدبه

جامعة كركوك-كلية الإدارة والاقتصاد-قسم الإحصاء، كركوك، العراق

المستخلص: في هذه الورقة البحثية، نُقدّم تقييمًا مقارنًا للعديد من نماذج الانحدار الكلاسيكية وخوارزميات التعلم الآلي المُطبّقة على مجموعة بيانات الخصوبة السويسرية. تشمل النماذج المُقيّمة الانحدار الخطي، وانحدار التلال، وانحدار لاسو، والشبكة المرنة، وانحدار المكونات الرئيسية (PCR)، والمربعات الصغرى الجزئية (PLS)، وخوارزميات التعلم الآلي مثل شجرة القرار، وأقرب جار (k)، والغابة العشوائية، وانحدار متجه الدعم (SVR)، وXGBoost. استُخدمت مقاييس الأداء مثل RMSE، وMAE، ومربع R، وAIC، وBIC لتقييم النماذج. أتاحت المقارنات البصرية باستخدام المخططات الشريطية تحديد النماذج الأكثر فعالية. تفوّق SVR على نماذج الانحدار

الأخرى في دقة التنبؤ، بينما أظهر انحدار لاسو أفضل بساطة للنموذج استنادًا إلى AIC/BIC. تُبرز النتائج قوة نماذج التعلم الآلي في اكتشاف العلاقات غير الخطية في مجموعات بيانات التنبؤ بالخصوبة.

الكلمات المفتاحية: مجموعة بيانات سويسرية، التنبؤ بالخصوبة، التعلم الآلي، النمذجة الإحصائية، انحدار متجه الدعم (SVR)، أساليب التنظيم.

Corresponding Author: E-mail: wakaa2017@uokirkuk.edu.iq

Introdaction:

Fertility forecasting is an essential tool for population and economic planning, helping to anticipate health, education, and labor needs, and to identify population aging and labor shortages. It also provides a deeper understanding of the factors influencing historical population patterns and facilitates effective future policy planning. In the era of data-driven decision-making, accurate forecasting of economic indicators is critical for policymakers, analysts, and researchers. Regression analysis remains one of the most widely used statistical methods for modeling the relationship between dependent and independent variables. However, with the increasing complexity and volume of economic data, traditional linear regression models often suffer from poor prediction accuracy and robustness. For example, life expectancy predictions are based on socioeconomic indicators, fertility rates, house prices, stock market performance, and others. To predict the value of a quantitative outcome using linear and nonlinear regression strategies and model selection methods, advanced regression techniques are being developed, including penalized regression (ridge regression, Lasso models, and elastic nets). Principal component regression (PCR) and partial least squares (PLS) regression are useful when the data contains multiple, interdependent predictor variables. These models improve their performance through regularization, dimensionality reduction, and nonlinear modeling capabilities. By identifying the most effective model using criteria such as root mean square error (RMSE), mean absolute error (MAE), and regression coefficient squared (R^2), these criteria contribute to improving economic forecasting tools and guide model selection in practical applications. Machine learning includes a variety of algorithms designed for different types of data and predictions. Among these methods, regression and classification stand out as two primary approaches to supervised learning. Classification models predict categorical outcomes and are suitable for tasks such as spam detection or disease diagnosis. Supervised machine learning typically involves using a set of features or covariates for prediction. When using the term prediction, where both X and Y (the training data) are observed 80% of the time, the goal is to predict an outcome (Y) on a 20% independent test set based on the observed values of [Y|]. These assumptions are the only objective assumptions necessary for the success of most machine learning methods. A variety of machine learning algorithms, such as support vector regression (SVR), random forests, regression trees, and support vector regression (SVR), are available for constructing robust regression models using nearest neighbor methods and XGBoost. Machine learning methods use data-driven model selection. That is, the analyst is provided with a list of shared variables or features, but the functional model is at least partially specified as a function of the data, rather than a single estimate. Therefore, this approach is best described as an algorithm that estimates multiple alternative models and then selects among them to maximize the criterion. The machine learning literature uses various techniques to balance expressivity and overfitting. Model validation and evaluation techniques are used to measure the performance of the predictive model. A diagnostic model is used to detect and fix potential problems with the predictive model. The most common approach is cross-validation, where the analyst repeatedly estimates the model based on a portion of the data (training) and then evaluates it in the appendix (testing). Model complexity is determined to minimize the mean square error. Prediction (squared difference between model prediction and actual outcome).Other methods used to control overfitting include averaging several different models, There is a lot of research into the topic of machine learning and regression from them García-Gutiérrez,et al 2015[1].A comparison of machine learning regression techniques for LiDAR-derived estimation of forest variables This paper presents a comparison between the classic MLR-based methodology and regression techniques in machine learning (neural networks, support vector machines, nearest neighbour, ensembles such as random forests) with special emphasis on regression trees. The results confirm that classic MLR is outperformed by machine learning techniques and concretely.HusejinovicAdmel, , et al 201٨[2]. Multiple machine learning algorithms have been applied to analyze virtual credit card payment. Using financial institution customer data provided by the UCI Machine Learning Repository, they evaluated and compared the performance of model candidates in order to choose the strongest model and what features were important in the best predictive model JACOB HALLMAN2019[3].Two different machine learning approaches were used; Linear regression and neural network regression. Uddin, Shahadat and et al. 2019.They presented this studyaims to identify key trends among different types of supervised machine learning algorithms, their performance and use for disease risk prediction. The Support Vector Machine (SVM) algorithm was found to be frequently applied (in 29 studies). Stephen HANSEN2020 [5].A chapter is devoted to applications of machine learning algorithms Economic research and policy making. Discusses the quantification of unstructured data and how to retrieve information in a way that is useful to economists. , where the combination of machine learning and new digital data provides the opportunity to develop measures of things like inflation and economic activity. Fernando Fachini2021[6].In his master's thesis he compared the voltage and system load mapping capabilities of a variety of regression algorithms, such as adaptive network-based fuzzy inference system (ANFIS), artificial neural networks (ANN), K-nearest neighbors (KNN), support vector regression (SVR) and decision tree (DT). It was found that ANFIS and KNN have better performance in predicting critical voltage and load compared to the rest. Sihombing and et al.2023[7].Comparison of Regression Analysis with Machine Learning Supervised Predictive Model Techniques This research aims to determine the factors that contribute to people's happiness. In terms of modelling, this study will compare several regressions modelling using machine learning, including regression trees, random forests and Support Vector Regression (SVR). Lorena González-Castroet al 2023 [8].Machine learning algorithms for predicting breast cancer recurrence using structured and unstructured sources from electronic health records. We evaluated the performance of five machine learning algorithms for predicting cancer recurrence within five years, and we selected the best performer to test our hypothesis. The XGB (maximum incremental boosting) model performed best among the five evaluated algorithms. Lu, Xiangning2024[9]Regression methods have been used to predict future trends and consumer behaviour with high accuracy. and using machine learning algorithms to analyze massive amounts of data to identify patterns and trends and make accurate predictions about future demand, product inventory levels, and other important factors that drive business success in the retail industry. Xiangning Lu[٩] ..٢٠٢٤Compared machine learning regression algorithms in retail to accurately predict demand and optimize supply chains. The study aims to quantitatively, experimentally, and comparatively examine regression models and machine learning algorithms for fertility rate prediction in a Swiss dataset. This allows for an assessment of the predictive power of both methods. Hadbaand Naser.2024.[10]They demonstrated that machine learning methods, such as SVM, LR, LLR, and ANN, are effective in accurately predicting breast tumor type, with Lasso logistic regression outperforming them in improving performance and handling multicollinearity and high-dimensionality problems. Abdullah and et al. 2025[11].They provide a comprehensive review of different machine learning approaches and a critical analysis of their performance and challenges based on several applied criteria. K. Ramesh & et al. 2025[12].Present machine learning applications and techniques in improving manufacturing processes within an industrial context. O. Shobayo&et al. 2025[13]They compared the performance of three machine learning models: support vector regression (SVR), recurrent neural networks (RNN), and long short-term memory (LSTM) in predicting market prices. LSTM outperformed. Y. Ghribi .et al .2025[14]They compared traditional statistical and machine learning models in accurately predicting demand for medical devices. The study demonstrated the superiority of deep learning models, especially LSTM.

The rest of the paper is organized as follows. Section 2 provides an overview of regression, principal component regression, and partial least squares (PLS) regression. Regularization methods include Ridge regression, Lasso regression, and elastic net. Section 3 introduces machine learning algorithms. K-nearest neighbour KNN regression, Decision trees, Random Forest, Support vector regression (SVR). Section 4 covers model performance metrics. Section 5 covers practical aspects:
A description of the data used in this work, the data analysis method, and the results tables. Section 6 presents the obtained results, their statistical validation, and the main findings. Section 6 summarizes the conclusions and discusses future work.

1st: Theoretical Framework:

1- Regression models:

A. Fundamentals of Regression Analysis.

A regression model allows us to predict a continuous outcome variable (y) based on the value of a single variable or multiple predictor variables (x). The goal of a regression model is to construct a mathematical equation that defines y as a function of variables x. This equation can then be used to predict the outcome (y) based on the new values of the predictor variables (x).Ordinary least squares (ols) Simple linear regression Multiple linear regression, Linear regression is the most simple and popular method for predicting a continuous variable.[15].When creating a linear regression model, you need to diagnose whether a linear model is a good fit for your data. In some cases, the relationship between the outcome and the expected variables is not linear. In these situations, you need to create nonlinear regression, such as polynomial regression.
The multiple linear regression models can be expressed as:

$$\mathrm{Y_i = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \cdots + \ \beta_k X_k + \varepsilon_i} \qquad (1)$$

The model in terms of the observations, Equation (1), may be written in matrix notation as[16]

$$\mathrm{Y = X\beta + \ \varepsilon} \qquad (2)$$

Y: is called dependent (response) variable.
X: is referred to as the design matrix.
β : represents the parameter vector.
ε : denotes the error vector.
with

$$\begin{aligned}
\mathbf{y} \ &= (y_1, \dots, y_n)^\top \in \mathbb{R}^n, \\
\mathbf{X} \ &= \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix} \in \mathbb{R}^{n\times p}, \\
\boldsymbol{\beta} \ &= (\beta_0, \beta_1, \beta_2, \dots, \beta_k)^\top \in \mathbb{R}^p \\
\boldsymbol{\epsilon} \ &= (\epsilon_1, \dots, \epsilon_n)^\top \in \mathbb{R}^n,
\end{aligned}$$

The objective is to determine the vector least squares estimators denoted as, $\hat{\beta}$, which minimizes
$\mathrm{L = \sum_{i=1}^{n} \varepsilon_i^2 = \varepsilon'\varepsilon = (Y - X\beta)'(Y - X\beta)}$.......(3)

$$\hat{\beta} = \frac{\mathrm{X'Y}}{\mathrm{(X'X)}} = \ (X'\,X)^{\,-1}\mathrm{X'Y} \qquad (4)$$

This part presents regression methods based on dimension reduction techniques, which can be very useful when you have a large data set with multiple correlated predictor variables[15].There are several problems with regression models, such as multicollinearity. This occurs when there is a strong relationship or high correlation between two or more independent variables (X variables). The model cannot determine the true effect of each variable on the dependent variable (Y). This affects the stability of the regression coefficients. Variability increases, interpretation of the results becomes difficult, and confidence in the estimates weakens.Generally, all dimension reduction methods work by first summarizing the original predictors into few new variables called principal components (PCs), which are then used as predictors to fit the linear regression model [15].described two well-known regression methods based on dimension reduction: Principal Component Regression (PCR) and Partial Least Squares (PLS) regression.

B. Principal component regression (PCR)

Principal component regression (PCR) is used to analyze principal components by transforming the original predictors into a smaller set of uncorrelated variables, called principal components, which are linear combinations of the original variables.[15]. PCR is a popular approach for extracting a set of low-dimensional features from a large set of variables. The principal components regression (PCR) approach involves first creating M principal components, Z1,..., ZM, and then using these components as predictors in a linear regression model fitted using least squares. The basic idea is that a small number of principal components is often sufficient to explain most of the variance in the data, as well as the relationship with the response. In other words, we assume that the directions in which X1,..., Xp exhibit the greatest amount of variance are the directions associated with Y.model to Z1,...,ZM will lead to better results than fitting a least squares model to X1,...,Xp, since most or all of the information in the data that relates to the response is contained in Z1,... ,ZM, and by estimating only M , p coefficients we can mitigate overfitting[17].
PCR reduces dimensionality using principal components analysis, producing a small number of principal components that account for most of the variance in the data. These components are then used as predictors in a linear regression to model the relationship with the response variable.[18]

C. Partial least squares (PLS) regression:

A potential drawback of PCR is that there is no guarantee that the selected principal components will be associated with the outcome. Here, the selection of principal components to be incorporated into the model is not supervised by the outcome variable. Partial least squares (PLS) regression is an alternative to multiple regression analysis, identifying new principal components that not only summarize the original predictors but are also associated with the outcome. These components are then used to fit the regression model. Therefore, compared to multiple regression analysis, PLS uses a dimensionality reduction strategy supervised by the outcome.[15].PLS regression reduces predictors to a smaller set of components, like PCR, but unlike PCR, it uses the response variable to identify components that explain both the predictors and the response. This supervised approach helps improve forecasting and resolves multicollinearity by projecting the data into a new regression space. [19][20] Within the new space, basic relationships between two matrices are investigated – the X matrix (predictors) and the Y matrix (responses).
The model attempts to identify the trend within space
The basic form of PLSR is:

$$X = TP^T + E \dots\dots,, (5)$$
$$Y = UQ^T + F \dots\dots\dots (6)$$

where -X and Y are the matrices of predictors (matrix of n × m predictions) and responses (matrix of n × p responses),
respectively;
- T and U represent the projections of X (degrees of X) and Y (degrees of Y).
degrees) and both are n × l matrices;
-P and Q represent orthogonal loading matrices of the predicted X and Y scores; And the,
-E and F are the error terms of the expectation matrix and the response matrix, and are assumed to be independent.
The overall goal is to use fundamental factors to predict population reactions. This is done via
Extracting the T and U factors (expected X and Y scores from the data sample). Extracted factors T (X scores)
These methods are used to predict U(Y) scores, and the predicted Y scores are then used to construct predictions of future responses. To emphasize, PLS considers both the scores (predictors) and Y scores (responses) to extract latent variables, whereas some other extraction methods

consider only X scores or Y scores. Furthermore, the precise number of factors extracted depends on the maximum variance that can be explained by the fewest number of factors.[21].

D. Shrinkage methods (Regularization):

Subset selection methods involve using least squares to fit a linear model containing a subset of predictors. Alternatively, we can fit a model containing all predictors for p using a technique that regularizes parameter estimates through shrinkage regression methods, which are modified versions of ordinary least squares regression that reduce the parameters.[16].Shrinkage is a type of regularization technique that fits a regression model using all p-predictors while applying constraints on the size of the estimated coefficients. This approach helps reduce the variance of the estimates, enhancing model stability. It can also shrink some coefficients to zero, allowing for variable selection. The most common shrinkage techniques include Ridge regression, Lasso regression, and Elastic Net.[22].

(1) Ridge Regression Model:

Hoerl and Kennard (1970) first developed the Ridge Regression model theory to add thepenaltyterm “squared magnitude” of the coefficient to the loss function of linear regression. Theridgeestimation is as follows:

$$\hat{\beta}_{\text{ridge}} = \operatorname{argmin}\left(\sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{i=1}^{p} x_{ij}\beta_j\right)^2\right) + \lambda \sum_{i=1}^{p} \beta_j^2 \quad \ldots\ldots(7)$$

where, N is the sample size, p is the number of variables, y_i is the observed target value, β_j are parameters or coefficients of the variables x_{ij}, λ is the complexity parameter which controls the amount of shrinkage, $\sum_{i=1}^{p}\beta_j^2$ is the L_2 Penalty term .Ridge regression introduces a penalty term and reduces the value of the coefficients without forcing them to zero. By reducing the coefficients, it helps reduce the complexity of the model and improve its generalization ability, making it less susceptible to overfitting.[23].When $\boldsymbol{\lambda}$ is set to 0, the penalty term in ridge regression has no effect, and the solution is equivalent to the ordinary least squares solution used in linear regression. As the value of α increases, the term becomes sharper. Therefore, to reduce the effect of the penalty and reach a balanced solution, it is natural for the regression coefficients to become smaller. The model aims to maintain a balance between fitting the training data well and preventing the coefficients from growing too large over the course of the penalty [23].

(2) Lasso regression

An alternative way to simplify a large multivariate model is to use penalized regression, which penalizes the model for containing too many variables [15].

The Lasso regression model was proposed by Tibshirani (1996), and the Lasso estimate β is defined by [23].

$$\hat{\beta}_{\text{lasso}} = \operatorname{argmin}\left(\sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{i=1}^{p} x_{ij}\beta_j\right)^2\right)\ldots\ldots..(8)$$

$$\text{subject to } \sum_{\mathrm{j}=1}^{\mathrm{p}} \left|\beta_{\mathrm{j}}\right| \leq u_{\text{lasso}} \ldots\ldots\ldots.(9)$$

The symbols in Equation (٨) are identical to that in Equation (٩) except u_{lasso} is a constant. In addition, we could write Equation 2-6 in the equivalent Lagrangian form:

$$\hat{\beta}_{\text{lasso}} = \operatorname{argmin}\left(\sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{i=1}^{p} x_{ij}\beta_j\right)^2\right) + \lambda \sum_{j=1}^{p} \left|\beta_j\right|\ldots\ldots(10)$$

Note that there is a one-to-one correspondence between the parameters λ in Equation 9 and Equation 10, for the penalty L.$\sum_{j=1}^{p}\left|\beta_j\right|$ replaces L_2 Penalty $\sum_{i=1}^{p}\beta_j^2$. Since $\sum_{j=1}^{p}\left|\beta_j\right| \leq u_{\text{lasso}}$ is the constraint

When the Lasso regularization coefficient u_{lasso} is small, some of the βj coefficients become exactly zero. As the value of λ changes from large to small, the importance order of each variable remains relatively stable. Thus, Lasso regression is characterized by its ability to select a subset of predictor variables from a large set and include only these variables in the final model.

Lasso regression reduces some parameter coefficients, βj, to zero to achieve variable selection. Lasso regression can reduce the underlying variables to the target variable. This regularization technique is also suitable when dealing with multicollinearity in the data or when seeking to control model complexity.

(3) Elastic net:

Zhou and Hastie (2005) introduced the elastic network (EN) to address many of the drawbacks of LASSO. When p > n, LASSO can select at most n predictors, even when more are correlated with the response variable. LASSO tends to select only one predictor from a set of highly correlated predictors. When n > p, and if there are high correlations among the predictors, LASSO's prediction performance is dominated by ridge regression. EN is a linear combination of L1 and L2 penalties.[26]:

According to [27], both Ridge and Lasso regression can be said to be special cases of the more general L_q optimization program

$$\hat{\boldsymbol{\beta}}_q = \arg\min_{\boldsymbol{\beta}} \left\{ \mathbf{y}'\mathbf{y} - 2\boldsymbol{\beta}'\mathbf{X}'\mathbf{y} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + \lambda \sum_{j=1}^{k} |\beta_j|^q \right\} \dots\dots\dots (11),$$

where $q = 2$ and $q = 1$ are equivalent to Ridge and Lasso respectively and the parameter $q \geq 0$[16].

which can be interpreted as a weighted split between Ridge ($\alpha = 0$) and Lasso ($\alpha = 1$). The elastic net is then used as a penalizing term to obtain the elastic net estimate

$$\hat{\boldsymbol{\beta}}_{\text{Elastic net}} = \arg\min_{\boldsymbol{\beta}} \{\mathbf{y}'\mathbf{y} - 2\boldsymbol{\beta}'\mathbf{X}'\mathbf{y} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + \lambda \textstyle\sum_{j=1}^{k} \left((1-\alpha)\beta_j^2 + \alpha|\beta_j|\right)\} \dots\dots\dots (12)$$

In this case, one do not need an advanced method to choose a parameter q[16].

2nd: machine learning algorithms:

Machine learning, a branch of artificial intelligence, aims to develop data-driven applications without explicit programming, using algorithms capable of accurately predicting and handling large-scale data. Its methods include classification for categorical variables and regression for continuous variables, with supervised learning being the most prominent, in which mathematical models are built to predict future outcomes.

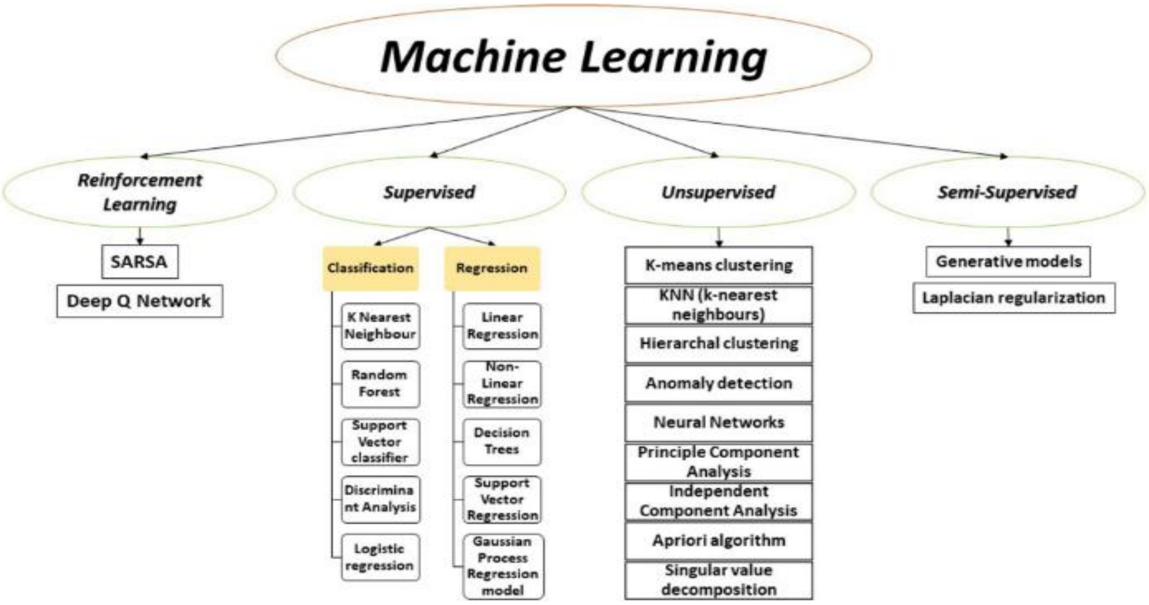

Fig. (1): Description of machine learning algorithms.

Criteria for selecting an algorithm include: the nature of the data, the desired outcome, the constraints of the problem, and the strengths of the algorithm. Analyzing the decision boundary from linearity to nonlinearity helps better understand the performance of models. K-fold cross-validation is an effective method for evaluating models. The data is divided into k sets, the model is trained on k−1 of them, and tested on the remaining set. The process is repeated until all sets have been used for testing.When calculating the average of the recorded errors (k), this is called the validation error, and it is a measure of the model's performance. Overall, the best model is the one with the lowest cross-validation error, which is the RMSE.

1- The nearest neighbor (KNN):

KNN regression is a simple machine learning technique used to predict numerical values. It relies on finding the k closest observations to a new observation x in the training data, then calculating their average values to use as an estimate of the predicted value.[15].Advantages: K-Nearest Neighbors is easy to implement and easy to understand. This algorithm is versatile and is used for regression and classification.It does not require training before making predictions and new data can be added seamlessly.Disadvantages: K-Nearest neighbors do not work well with large datasets due to a large number of dimensions. This algorithm is sensitive to noisy data present in the dataset and missing values[28].The KNN algorithm relies on finding the k-nearest neighbors from the training data to be classified. After identifying the neighbors, the data is assigned to the class that represents the majority among the observations. This method classifies samples based on the distance between them, where the "nearest neighbor" is defined using various distance metrics, the most common of which is the Euclidean distance. The optimal value of k is determined empirically to achieve the best model performance.

2- Decision trees:

The decision tree model is a technique rooted in predictive analytics, and was originally introduced by[30].The decision tree method is a powerful and popular predictive machine learning technique that is used for both classification and regression. Therefore, they are also known as classification and regression trees (CART)[15].The decision tree (DT) is one of the oldest and most popular machine learning algorithms. It represents decision-making logic through tests and outcomes arranged in a tree structure. The tree starts with the root node at the top, while the internal nodes represent tests on the input variables and branch into branches that ultimately lead to nodes representing outcomes or decisions.[4].

Building a decision tree involves two basic steps:

First:Partitioning the prediction space, i.e., the set of possible values for the variables X1, X2,..., Xp, X1, X2,..., Xp, into J separate, non-overlapping regions: R1, R2,..., RJ, R1, R2,..., RJ.

Secondly:Making a prediction for each observation within a given region Rj, such that the prediction is the average of the response values of the training observations within that region.

For example, if the partitioning in the first step yields two regions R1R1 and R2R2, and the average response in R1R1 is 10 and in R2R2 is 20, then for a new observation

X = x X = x:

If x ∈ R1x ∈ R1 we expect a value of 10,

If x ∈ R2x ∈ R2 we expect a value of 20,

$$\sum_{j=1}^{J} \sum_{i \in R_j} \left(y_i - \hat{y}_{R_j} \right)^2 \text{........(13)}$$

where $\hat{y}_{R_j}$ is the mean response for the training observations within the j th box. We apply a top-down approach that is known as recursive binary splitting. This method begins at the top of the tree and then successively splits the predictor space; each split is indicated via two new branches further down on the tree.[31].

To perform such recursive binary splitting, we first select the predictor X_j and the cutpoint s such that splitting the predictor space into the regions $\{X \mid X_j < s\}$ and $X \mid X_j \geq s\}$ leads to the greatest possible reduction in RSS. In details, for any j and s, we define the pair of half-planes
$R_1(j,s) = \{X \mid X_j < s\}$ and $R_2(j,s) = \{X \mid X_j \geq s\}$,……..(14)
and we seek the value of j and s that minimize the equation
$\sum_{i:x_i\in R_1(j,s)} \left(y_i - \hat{y}_{R_1}\right)^2 + \sum_{i:x_i\in R_2(j,s)} \left(y_i - \hat{y}_{R_2}\right)^2$,………(15)
where $\hat{y}_{R_1}$ is the mean response for the training observations in $R_1(j,s)$, and $\hat{y}_{R_2}$ is the mean response for the training observations in $R_2(j,s)$.
Next, we repeat this process, looking for the best predictor and best cutpoint in order to split the data further so as to minimize the RSS within each of the resulting regions. However, instead of splitting the entire predictor spacewe split one of the two previously identified regions[31].

3- Random Forest:

Random forest algorithm is one of the most widely used and powerful machine learning techniques. It is a special type of bagging applied to decision trees.Compared to the standard CART model (decision tree models), random forest provides a robust improvement, which consists of applying bagging to the data and bootstrapping to the predictor variables in each partition[32]. This means that in each partitioning step of the tree algorithm, a random sample of n predictors is selected as partitioned candidates from the full set of predictors[15].Random forest can be used for both classification (predicting a categorical variable) and regression (predicting a continuous variable).
This bagging consists of taking multiple subsets of the training data set, then building multiple independent decision tree models, and then averaging the models allowing for the creation of a predictive model that is superior in performance compared to the classical CART (decision tree models) model[15].Random forests improve ensemble trees by making a simple modification that removes correlation between trees, such as padding the decision trees in the initial training sample. During decision tree construction, a split occurs each time the tree is randomly selected from the full set of p predictors as split candidates. That is, when constructing a random forest, the algorithm is not allowed to consider the majority of the available predictors at each split in the tree. Random forests force each split to consider only a subset of predictors. Therefore, the mean (p − m)/p splits will not consider the strong predictor, and other predictors will have a greater chance. This process, called random forests, can also be thought of as tree splitting, which makes the average of the resulting trees less variable and therefore more reliable. Thus, the main difference between ensemble and random forest is the choice of the size of the subset of predictors, m. [31].
Random Forests is an ensemble learning method that creates multiple decision trees and outputs their prediction patterns for classification tasks or their average predictions for regression tasks. It has several properties:
A. Automatic clustering (bagging): Each tree is trained on a random subset of the training data.
B. Attribute randomization: At each node split, only a random subset of attributes is considered.
C. Voting/averaging: Final predictions are made by aggregating the predictions from all trees.[12].

4- Boosting:

An alternative method is called boosting, which is similar to bagging, except that trees grow sequentially: each successive tree is planted using information from previously planted trees, with the aim of minimizing the error of previous models[17].For example, given the current regression tree model, the procedure is as follows[15]:
Decision tree fit using residual error model as the outcome variable.Add this new decision tree, modified by the lambda shrinkage parameter, to the fitted function in order to update the residuals. Lambda is a small positive value, typically between 0.01 and 0.001[17].This approach improves the fitted model slowly and sequentially leading to a high-performance model. The booster has different tuning parameters including[15]:Number of trees b,Lambda shrinkage parameter,Number

of cracks in each tree.There are different types of boosting, including Adaboost, Gradient Boosting, and Stochastic Gradient Boosting[15].

Algorithm. Boosting for Regression[31].

A. Set $\hat{f}(x) = 0$ and $r_i = y_i$ for all i in the training set,

B. For $b = 1,2, \dots , B$, repeat:

(a) Fit a tree $\hat{f}^b$ with d splits $(d+1$ terminal nodes) to the training data (X,r).

(b) Update $\hat{f}$ by adding in a shrunken version of the new tree:

$\hat{f}(x) \leftarrow \hat{f}(x) + \lambda \hat{f}^b(x)$ ….…….(16)

(c) Update the residuals,

$$r_i \leftarrow r_i - \lambda \hat{f}^b(x_i).$$

C. Output the boosted model,

$\hat{f}(x) = \sum_{b=1}^{B} \lambda \hat{f}^b(x)$ ….…….(17)

5- Support Vector Regression (SVR):

The Support Vector Machine (SVM) is a commonly used classifier that performs classification tasks within a feature space. It is trained by solving a constrained quadratic optimization problem, which allows for determining the optimal class separation. Kernel techniques can also be used to create complex nonlinear classifiers. Kernels transform sample variables from a low-dimensional space to a higher-dimensional space, facilitating the separation of nonlinear data[33].To enable non-linearly separable samples to continue using the SVM model, linear, polynomial, and Gaussian functions are the most common kernels. [29].

It works by finding the hyperplane that best separates different classes in a high-dimensional space.

Key Concepts:

* **Kernel Trick:** Support vector machines can use different kernel functions to transform data into higher dimensions where it becomes linearly separable.

* **Margin Maximization:** Support vector machines aim to find the hyperplane with the maximum margin between classes.

* **Support Vectors:** These are the data points closest to the separating hyperplane, and are essential for determining it.

SVR was developed and trained as a form of support vector machine (SVM) technique specifically used for regression tasks[34]. The architectural components of SVR are the kernel function, the regularization parameter (C), the epsilon (ε), and the support vectors. The model separates higher-dimensional data points using the kernel function. We used the linear kernel for final model training. This is simple, fast, and efficient when the input characteristics and target variables are linearly related. The model balances a wide class margin and low error rates with the regularization parameter (C). The epsilon (ε) defines the region where the model is not penalized for errors. This makes the model less sensitive to small changes in the data and improves performance. The support vectors are the margin data points that define the decision boundary (the line or curve separating the data points).

3rd: Model performance metrics:

To evaluate the prediction performance of the models, this study uses a comprehensive set of evaluation metrics: mean absolute error (MAE), root mean square error (RMSE), R-squared (R^2), In addition, to compare regression models, we use: AIC, BIC.

There are several statistical measures to compare the performance of different models on data, compare them, and then choose the best approach that explains the data well, interprets it, and predicts the results of new test data.[15].

The best model is defined as the one with the lowest prediction error. The most common metrics for comparing regression models include:

The root mean square error, which measures the model's prediction error. It represents the average difference between the known observed values of the outcome and the value predicted by the modelThe lower the RMSE, the better the model.
. The RMSE is calculated using the formula:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum(yi\ -\ f\hat{}(xi))^2}\ ....(18)$$

where $f\hat{}(xi)$is the prediction given by $f\hat{}$for observation i. As the name suggests, a high MSE value indicates that the prediction is far from the true value while a low value means that the prediction is fairly accurate[3].
Both the root mean square error (RMSE) and the mean absolute error (MAE) are regularly used in model evaluation studies.[35].

$$\text{MAE}=\frac{1}{n}\sum\left|yi\ -\ f\hat{}(xi)\right|\quad \ldots\ldots,(19)$$

R-squared (R2), which is the proportion of variance in the outcome explained by the predictor variables. In multiple regression models, R2 corresponds to the squared correlation between observed outcome values and the values predicted by the model. The higher the R-square, the better the model[15].
R2 or coefficient of determination is a way to measure the validity of a regression model. It can be interpreted as a proportion of the variance in the expected outcome. The output of this method ranges from 0 to 1, where a value of 1 indicates that each point on the regression line fits the data perfectly [36].A value of 0.7 means that 70% of the data points fall within the result of the regression line. This value can be increased by including more independent variables, which is why there is a modified version of R2 as well. The following equation determines R2:

$$R^2 = \frac{(\,yi - \bar{y})^2}{(\hat{y}i\ -\ \bar{y})^2}\ldots\ldots,(20)$$

where $\hat{y}$ is the estimated value of the dependent value for the ith observation by the regression equation, y is the observed or baseline value of the dependent variable for the ith observation, and y is the mean of all observations of the dependent variable [3][36]. It should be noted that the above metrics must be calculated based on new test data that was not used to train (i.e., build) the model. If you have a large dataset containing many records, you can randomly split the data into a training set (80% for building the predictive model) and a test or validation set (20% for evaluating model performance).
Akaike's (1973) information criterion (AIC) and Schwartz's (1978) bottom-up information criterion (BIC) are used to select the best possible econometric and statistical model[37]

$$\text{AIC} = 2\text{k} - 2\text{ln}\ldots\ldots\ldots(21)$$

Where:k = number of parameters in the model,L = value of the maximum likelihood function of the Model[38].

$$BIC = kln(n) - 2ln(L)\ldots\ldots(22)$$

Where:n = number of observations in the sample,k = number of parameters,L = value of the maximum likelihood function[39].

4th: The practical side

Data sets used in the R package datasets: swiss[40].
Dependent variable: Fertility - birth rate per 1,000 women aged 15 to 49. Independent variables: Agriculture, examinations, education, Catholicism, infant mortality.
Swiss Data contains demographic and economic statistics for 47 cantons in Switzerland at the end of the 19th century and is used to identify factors affecting fertility rates.
Dependent Variable:Fertility: Fertility rate (number of births per 1,000 women)
Independent Variables,Agriculture: Proportion of the male population employed in agriculture ,Examination: Proportion of males who passed the military entrance exam,Education: Proportion of men with higher than primary education,Catholic: Proportion of the Catholic population

Infant Mortality: Infant mortality rate (per 1,000 live births)
The data were analyzed in the R program using the types of packages.The data was split 80% of the data for training, and 20% for testing.

Table (1): Results of training and testing data analysis for regression models and machine learning algorithms

Model	AIC	BIC	RMSE	MAE	R squared
Linear Regression	274.4575	286.1025	9.6٦	7.٩	0.3448380
Ridge Regression	267.0599	272.0506	9.206313	7.400285	0.4046553
Lasso Regression	266.4972	271.4879	9.615772	7.869695	0.3505206
Elastic Net	266.6585	271.6491	9.609013	7.860843	0.3514333
PCR	283.2500	288.2407	9.80	8.57	0.325
PLS	272.5648	277.5555	9.78	7.78	0.329
Random Forest			8.825639	7.929739	0.4528714
SVR			6.832450	5.550334	0.6720936
XGBoost			8.469301	7.257396	0.4961606
Decision Tree			9.278123	8.309935	0.3953316
K-Nearest Neighbors			8.163345	6.426000	0.5319057

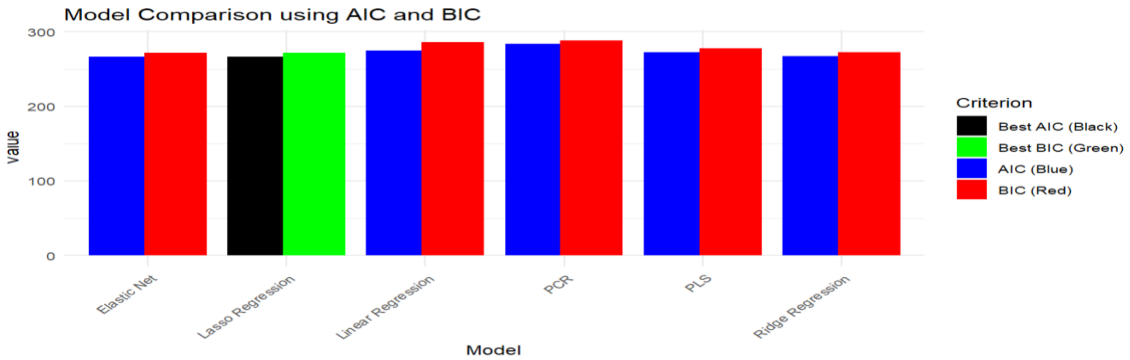

Fig. (2): Regression Model Selection Criteria Comparison between AIC and BIC

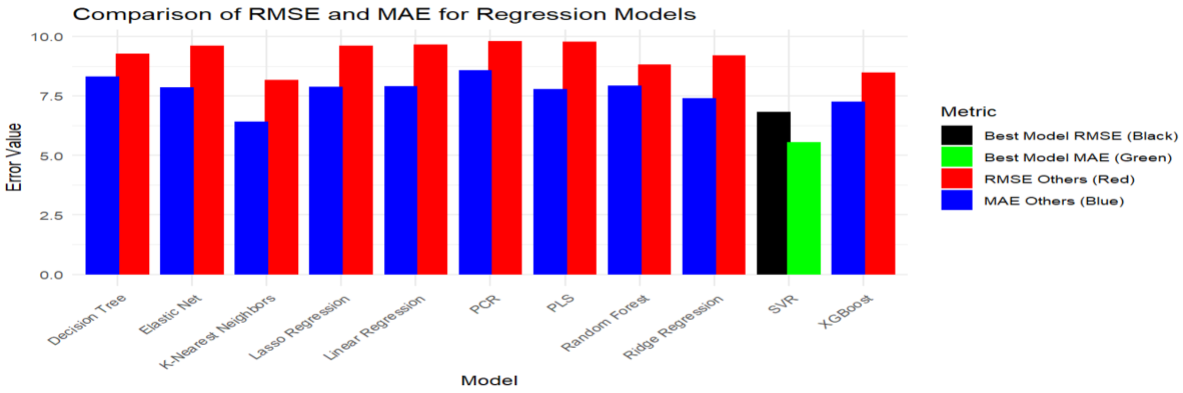

Fig. (3): Comparison of RMSE and MAE across regression models and machine learning algorithms.

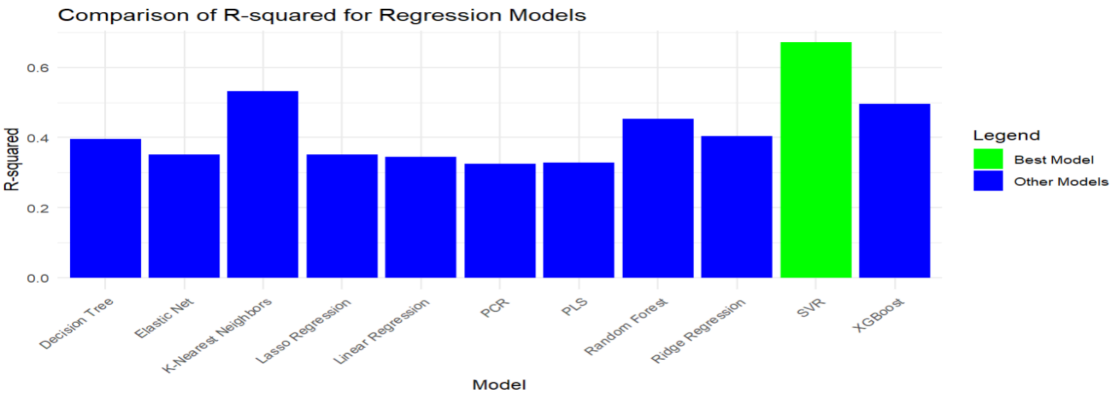

Fig. (4): Compare models based on the coefficient of determination (R²).

5th: Discussion of Results:

This study evaluated the performance of six regression models and five different machine learning algorithms on the Swiss dataset, using three main performance metrics in addition to the AIC/BIC criteria for regression models.

The optimal regression model for comparison using the AIC/BIC criteria was Lasso regression. Its lowest AIC was 266.4972, and its lowest BIC was 271.4879. Its RMSE was 9.6157, and its coefficient of determination (R^2) was 0.3505. The PCR and PLS models performed the worst, showing poor performance with the highest AIC and BIC values, the highest MAE and RMSE values (especially PCR), and the lowest R^2 values.

When comparing the performance of different models on the dataset, it was found that the SVR model outperformed the other models. RMSE = 6.832, MAE = 5.550, and R-squared = 0.672.

This makes SVR offer the best predictive accuracy among all models, with the lowest RMSE and MAE values, and the highest R-squared value, reflecting its superior ability to represent data and reduce prediction error. XGBoo Good = 0.496, RMSE = 8.47

KNN R^2 = 0.5319, very good performance

Balanced Random Forest Performance, R^2 = 0. 4528.The graph () shows two columns for each model. The first column shows the RMSE (root mean square error) and the second column shows the MAE (mean absolute error). The best model (SVR) is shown in black and green because it has the lowest error values. The weaker models are shown in red and blue, indicating that SVR has the best predictive accuracy. The machine learning algorithms KNN, XGBoost, and Random Forest also performed well compared to the regression models, with PCR and PLS showing the weakest.

The graph () represents each model's ability to explain the variance in the data. SVR had the highest R^2 (0.672), meaning it explained more than 67% of the variance in fertility.

It was considered the best in terms of explanatory power. It is highlighted in green.

KNN and XGBoost also had R^2s above 0.5, indicating good performance.

Linear, PCR, and PLS were the weakest models in terms of explanatory power. Machine learning models outperformed regression models on this criterion.

6th: Conclusions and Recommendations:

Conclusions:

1. The SVR (Support Vector Regression) model offered the best predictive performance, with the lowest RMSE and MAE values and the highest R-squared value (0.672).
2. The K-Nearest Neighbors and XGBoost models qualified well after SVR, and they varied their correlations and differed in handling nonlinear data.
3. Regression models (such as Linear, Ridge, Lasso, and Elastic Net) performed less effectively than machine learning models, especially in terms of RMSE and R-squared.
4. Although Lasso Regression did not perform the best predictively, it recorded the lowest AIC and BIC values, making it a good candidate.
5. Models such as PCR and PLS are not suitable for this dataset.

Recommendations

1. SVR is recommended as the first choice for prediction applications on this type of data, especially when accuracy is the primary goal.
2. It is suggested to consider models such as XGBoost and KNN as effective alternatives, especially in cases where SVR parameters may be difficult to adjust or faster prediction processes are required.
3. Lasso Regression represents a good balance between complexity and performance if simplicity and interpretability are more important than accuracy.
4. We recommend comparing other regression types, such as Posson regression, gamma regression, and other machine learning algorithms.

References

1- Abdullah, N. S. Mohammed, M. Khanzadi, and M. Safar, “A Comprehensive Review of Machine Learning Approaches: Techniques, Applications, and Trends Abdulhady Abas Abdullah Artificial Intelligence and Innovation Centre, University of Computer Science, Faculty of Science, Soran University Health Information,” no. February, 2025, doi: 10.36227/techrxiv.174059987.79916266/v1.

2- Akinjole, A.; Shobayo, O.; Popoola, J.; Okoyeigbo, O.; Ogunleye, “Ensemble-Based Machine Learning Algorithm for Loan Default Risk Prediction.,” Mathematics, vol. 12, no. 3423, 2024.

3- and M. B. Astrid Schneider, Gerhard Hommel, “Linear regression analysis: part 14 of a series on evaluation of scientific publications.” In: Deutsches Arzteblatt international 107 44 (2010),., 2010.

4- B. Sun and H. L. Wei, “Machine Learning for Medical and Healthcare Data Analysis and Modelling: Case Studies and Performance Comparisons of Different Methods,” 2022 27th International Conference on Automation and Computing: Smart Systems and Manufacturing, ICAC 2022, no. September 2022, pp. 1–6, 2022, doi: 10.1109/ICAC55051.2022.9911176.

5- C. A. (Greenwood, C. J., Youssef, G. J., Letcher, P., Macdonald, J. A., Hagg, L. J., Sanson, A., Mcintosh, J.,Hutchinson, D. M., Toumbourou, J. W., Fuller-Tyszkiewicz, M., &Olsson, “Acomparison of penalised regression methods for informing the selection of predictive markers,” PLoS ONE, 15 (11), e0242730. https://doi.org/10.1371/journal.pone.0242730, 2020.

6- D. Project et al., “DEGREE PROJECT IN THE FIELD OF TECHNOLOGY Regression Modeling from the Statistical Learning Perspective with an Application to Advertisement Data,” 2018.

7- editors. T. Hastie, R. Tibshirani, and J.H. Friedman, The elements of statistical learning. 2009.

8- F. Andreis, “Shrinkage methods (ridge , lasso , elastic nets) Shrinkage methods and variable selection : Ridge , Lasso , and Elastic Nets,” no. November, 2017.

9- F. Fachini and B. I. L. Fuly, “A Comparison of machine learning regression models for critical bus voltage and load mapping with regards to max reactive power in PV buses,” Electric Power Systems Research, vol. 191, 2021, doi: 10.1016/j.epsr.2020.106883.

10-G. Bhumireddy Venkata, A. Surendra, M. Anala, V. A. Surendra, and Y. Hu, “Comparison of Machine Learning algorithms on detecting the confusion of students while watching MOOCs,” no. February, 2022, [online]. Available: www.bth.se.

11-G. Schwarz, “Estimating the dimension of a model,” The Annals of Statistics, vol. 6, no. 2, pp. 461–464, 1978.

12-Gareth James • Daniela Witten • Trevor Hastie and Robert Tibshirani, An Introduction to Statistical Learning with Applications in R. 2013.

13-H. Akaike, “A new look at the statistical model identification".,” IEEE Transactions on Automatic Control, vol. 19, no. 6, pp. 716–723, 1974.

14-J. García-Gutiérrez, F. Martínez-Álvarez, A. Troncoso, and J. C. Riquelme, “A comparison of machine learning regression techniques for LiDAR-derived estimation of forest variables,” Neurocomputing, vol. 167, pp. 24–31, 2015, doi: 10.1016/j.neucom.2014.09.091.

15-J. Hallman, “A comparative study on Linear Regression and Neural Networks for estimating order quantities of powder blends,” 2019.

16-K. C. and assessment of machine learning approaches in manufacturing applications Ramesh, M. N. Indrajith, Y. S. Prasanna, S. S. Deshmukh, C. Parimi, and T. Ray, “Comparison and assessment of machine learning approaches in manufacturing applications,” Industrial Artificial Intelligence, vol. 3, no. 1, 2025, doi: 10.1007/s44244-025-00023-3.

17-Kassambara, “Machine Learning Essentials. Practical Guide in R.,” Sthda. p. 210, 2017, [Online]. Available: https://www.m-culture.go.th/mculture_th/download/king9/Glossary_about_HM_King_Bhumibol_Adulyadej’s_Funeral.pdf.

18-L. González-Castro et al., “Machine Learning Algorithms to Predict Breast Cancer Recurrence Using Structured and Unstructured Sources from Electronic Health Records,” Cancers, vol. 15, no. 10, 2023, doi: 10.3390/cancers15102741.

19-M. Friendly and D. Meyer, “Data sets used in the R package datasets: swiss,” Base R datasets, 2015.

20-M. L. Sawatsky, M. Clyde, and F. Meek, “Partial least squares regression in the social sciences,” The Quantitative Methods for Psychology, vol. 11, no. 2, pp. 52–62, 2015, doi: 10.20982/tqmp.11.2.p052.

21-M. Z. Husejinovic Admel, Keco Dino, “Application of Machine Learning Algorithms in Credit Card Default Payment Prediction,” International Journal of Scientific Research, vol. 7, no. 10, pp. 425–426, 2018, doi: 10.15373/22778179#husejinovic.

22-O. Shobayo, S. Adeyemi-Longe, O. Popoola, and O. Okoyeigbo, “A Comparative Analysis of Machine Learning and Deep Learning Techniques for Accurate Market Price Forecasting,” Analytics, vol. 4, no. 1, p. 5, 2025, doi: 10.3390/analytics4010005.

23-P. R. Sihombing, S. Budiantono, A. M. Arsani, T. M. Aritonang, and M. A. Kurniawan, “Comparison of Regression Analysis with Machine Learning Supervised Predictive Model Techniques,” Jurnal Ekonomi Dan Statistik Indonesia, vol. 3, no. 2, pp. 113–118, 2023, doi: 10.11594/jesi.03.02.03.

24- P. Rintara et al., “RIDGE-TYPE SHRINKAGE ESTIMATIONS IN SOME STATISTICAL MODELS WITH MULTICOLLINEARITY PROBLEM RIDGE-TYPE SHRINKAGE ESTIMATIONS IN SOME STATISTICAL MODELS WITH MULTICOLLINEARITY PROBLEM,” 2020.

25- P. Sun, “Comparative Study of Machine Learning Methods Applied in Hydrological Models,” 2023.

26- Patle and D. S. Chouhan, “SVM kernel functions for classification,” Jan. . http://dx.doi.org/10.1109/icadte.2013.6524743, 2013.

27- R. James, G., Witten, D., Hastie, T., Tibshirani, An Introduction to Statistical Learning - with Applications in R | Gareth James | Springer. 2013.

28- R. M. Golden, “Statistical Machine Learning,” Statistical Machine Learning, 2020, doi: 10.1201/9781351051507.

29- R. M. Golden, “Statistical Machine Learning,” Statistical Machine Learning, no. September, 2020, doi: 10.1201/9781351051507.

30- R. Tobias, “An introduction to partial least squares regression.,” Proceedings of the Twentieth Annual SAS Users Group International Conference. Cary, NC: SAS Institute Inc., 1995.

31- S. Criterion and M. Z. Hossain, “AIC and BIC – The two competitive information criteria for model selection in economics and statistics,” Time, no. November, pp. 1–3, 2016, [Online]. Available: https://www.researchgate.net/publication/282731304_AIC_and_BIC_-_The_two_competitive_information_criteria_for_model_selection_in_economics_and_statistics.

32- S. Hansen, “Machine Learning for Economic and Policy,” pp. 369–395, 2020.

33- S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, “Comparing different supervised machine learning algorithms for disease prediction,” BMC Medical Informatics and Decision Making, vol. 19, no. 1, pp. 1–16, 2019, doi: 10.1186/s12911-019-1004-8.

34- S. Van Roon, P., Zakizadeh, J., & Chartier, “Partial least squares tutorial for analyzing neuroimaging data.,” The Quantitative Methods for Psychology, vol. 10, pp. 200–215, 2014.

35- T. Chai and R. R. Draxler, “Root mean square error (RMSE) or mean absolute error (MAE)? -Arguments against avoiding RMSE in the literature,” Geoscientific Model Development, vol. 7, no. 3, pp. 1247–1250, 2014, doi: 10.5194/gmd-7-1247-2014.

36- T. et. all. Hastie, “Springer Series in Statistics The Elements of Statistical Learning,” The Mathematical Intelligencer, vol. 27, no. 2, pp. 83–85, 2009, [Online]. Available: http://www.springerlink.com/index/D7X7KX6772HQ2135.pdf.

37- W. A. Belson, “Matching and prediction on the principle of biological classification.,” Journal of the Royal Statistical Society: Series C (Applied Statistics, 8(2), vol. 8, no. 2, pp. 65–75, 1959.

38- W. A. Hadba and H. I. Naser, “Comparison of lasso logistic regression, artificial neural networks, and support vector machine in predicting breast cancer,” International Journal of Statistics and Applied Mathematics, vol. 9, no. 1, pp. 83–89, 2024.

39- X. Lu, “A comparative study of machine learning-based regression models for supply chain management,” vol. 0, pp. 48–55, 2024, doi: 10.54254/2755-2721/53/20241233.

40- Y. Ghribi, E. R. Graha, and H. Wicaksono, “Comparative Analysis of Statistical and Machine Learning Models for Enhancing Demand Forecasting Accuracy in the Medical Device Industry,” Procedia CIRP, vol. 134, pp. 849–854, 2025, doi: 10.1016/j.procir.2025.02.209.