International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 2 Volume No.7 March - April – 2023
DOI:10.46647/ijetms.2023.v07i02.012 ISSN: 2581-4621

# Predict The Risk Factor of The Possibility of Death For Not Having A Bridge In Nepal By Using Multiple Regression Analysis

**Mr. Om B Khadka[1], Akash Modak[2], Dipanwita Sahoo[3], Viktor Sarkar[4], Mrs. Sulekha Das[5], Prof. Dilip k. Banerjee[6], Dr. Anirban Das[7]**
[1]*Research Scholar, Department of Computer Application, SEACOM SKILLS UNIVERSITY*
[2]*UG - Information Technology, Techno Engineering College Banipur*
[3]*UG - Computer Science and Engineering, Techno Engineering College Banipur*
[4]*UG - Computer Science and Engineering, Techno Engineering College Banipur*
[5]*Assistant Professor, Computer Science and Engineering, Techno Engineering College Banipur*
[6] *Professor, Department of Computer Application, SEACOM SKILLS UNIVERSITY*
[7]*University of Engineering & Management*
*Corresponding Author Orcid ID:*
[1]*https://orcid.org/0000-0003-3100-1295*, [2]*https://orcid.org/0000-0001-6267-0761*,
[3]*https://orcid.org/0000-0002-7268-9105*, [4]*https://orcid.org/0000-0002-4711-2064*,
[5]*https://orcid.org/0000-0002-6641-3268*

**ABSTRACT**
Communication and transportation are important aspects of human life, but in Nepal, the geographically challenging topography and disconnected communities by more than 6000 rivers and rivulets, present great challenges in their daily lives for accessing basic health services, education, and household chores. Bridges are one of the major means of connecting the rural population in Nepal, but the lack of appropriate and safe means of crossing torrential rivers has resulted in increased suffering for local communities and developmental challenges for the actors involved in this sub-sector.
The focus of this research is to technically analyse the potential bridge sites based on the major prioritization indices and to determine risk factors related to particular locations leading to solutions for the permanent transportation problem. The government has been trying its' best in collaboration with various development actors to address the problem and help to reduce the risks related to potential loss of lives while crossing the mountainous terrain to various destinations.
Many practices proposed or implemented have been proven to be risky, especially for children, men, women, and the elderly population.
This research aims to establish a proven module to construct a trail bridge as a safer means of transport across the hills that would accelerate access to basic services such as education, healthcare, and provide people with new development prospects.
The innovative idea involves supporting the construction of a trail bridge to support services such as education, medical facilities, and household chores. The data has been analysed using Multiple Regression Analysis (MRA). In this research, the authors predicted the risk factor of the possibility of death due to the lack of a bridge, which depends on the total population, total households, river type, width of the bridge, etc. The model was evaluated using 50-50%, 66-34%, and 80-20% train-test splits and 10-fold cross-validation and an accuracy of approximately 70% was achieved.
The authors collected data by mobilizing local NGOs and informing the public through local radio to conduct a comprehensive study of nationwide bridge demand. The secondary source of data for post-bridge building assessment is extracted through the project management information system.
**Keywords- Multiple Regression Analysis, Cross-validation, Confusion Matrix**

## 1. INTRODUCTION

Statistics is an application of research methodology that we need in every kind of field in our daily life such as in science, medicine, crime, employed in professional sources, and also in social sciences [1]. We can analyze any kind of data using statistical analysis. Regression is one of the most useful

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 2 Volume No.7 March - April – 2023
DOI:10.46647/ijetms.2023.v07i02.012 ISSN: 2581-4621

statistical methods. There are six types of linear regression analyses: simple linear regression, multiple linear regression, logistic regression, ordinal regression, multinomial regression, and discriminant regression. Multiple linear regression (MLR) uses two or more independent variables to predict something. The main objective of this paper is to predict the possibility of death when people cross the river without a bridge in Nepal. Post selecting the control and response variables, creating the formula, and dividing the dataset into training and testing sets, creating the confusion matrix, different tests were performed and the results were noted down. All are done in the 50-50%, 66-34%, 80-20% train-test splits, and 10-fold cross-validation system.
In Nepal, the evolution of road interconnected systems and bridges begin around 60 years ago [2]. Nepal has highly survived a range of natural hazards, particularly floods, landslides, and earthquakes. In the early years, most of the main works were done with financial and technical assistance from donor countries. Afterward, the Department of Local Infrastructure Development (DoLI) is leading the sub-sector for standardization of technology, formulation of required strategies, and implementation of bridge construction. Currently, around 500 trail bridges are constructed yearly, and of a related range by the regional and local governments.
Researchers have widely used machine learning (ML) methods and Multiple Linear Regression (MLR) methods to predict the risk factor. Studies show high accuracy of prediction of risk factor data analysis using MLR methods. However, the researchers have gotten the most precise results about accuracy.
Researchers proved that prediction improves with the choice of the right features. Therefore, there is a need to choose a subset of many features that best suits the task i.e. also called trained data and test data. Among some of the several methods to select features, researchers use the MLR method because of the linear relationship between all the variables. And the result shows that the method is very much effective.

## 2. LITERATURE REVIEW

Multiple regression analysis is an applied math technique, the target of multiple regression analysis is to use the independent variables whose worths square measure acknowledged to predict the worth of the only dependent value. In this study, we used the MLR technique to predict the risk factor of the possibility of death for not having a bridge in Nepal.
Pan Lu, Shiling Pei, and Denver Tolliver (2016) demonstrated an objective, data-based approach for regression model forecasting ability evaluation [13]. If the model is selected based on an apparent evaluation only, then the forecasting outcome may not be accurate, especially for long-term planning, maintenance, rehabilitation, and replacement decisions.
Ying-Ming Wang and Taha M.S. Elhag (2007) conducted a comparative analysis of three alternative approaches for modeling bridge risks, which are the neural network, the evidential reasoning approach, and multiple regression analysis [14].
Mr. M. S. BARTLETT (1938) gave detailed information about multiple linear regression in his paper named 'FURTHER ASPECTS OF THE THEORY OF MULTIPLE REGRESSION' [15].
After studying some more papers about multiple linear regression, researchers have got a clear idea about the MLR method. That's why in this paper MLR method is used and the full research paper is depending on this method.

## 3. METHODOLOGY

• **DATA PROCESSING:** In this study, data have been collected from mobilizing local NGOs and informing the public people through local radios, which was for a comprehensive study to collect nationwide bridge demand. In this dataset, there is a total of 37 different fields, out of these fields 13 fields are taken for analysis. From the dataset "Risk Factor" i.e. consider as a dependent variable which is the possibility of death for not having a bridge. Other independent variables are "www" i.e. it is the width of the bridge, "DRH" which is the distance to the nearest road in hours, "DDH" i.e. Distance to district headquarters in the day, "Total Household" i.e. a household includes the related

DOI:10.46647/ijetms.2023.v07i02.012 ISSN: 2581-4621

family members and all the unrelated people, if any, such as lodgers, foster children, wards, or employees who share the housing unit, "Total Population" i.e. the whole number of people or inhabitants in a country or region, " Total DAG" which is the addition of "Total Dalit", "Total Janajati", "Total Minor", and "Total Other poor". "Dist. Gained" which is distance gained, and "River type".

**Table 1-** For attributes mean and standard deviation:

| ATTRIBUTES | MEAN | STANDARD DEVIATION |
|---|---|---|
| Risk Factor | 0.27 | 1.29 |
| Total Population | 2957.53 | 4045.18 |
| Total DAG | 2466.19 | 3364.78 |
| Total Household | 1388.51 | 68983.27 |
| Www | 92.69 | 20.41 |
| DRH | 1.48 | 2.06 |
| DDH | 1.16 | 1.37 |
| Dist. Gained | 3.22 | 27.57 |
| River Type | 5.95 | 3.35 |

• **RESEARCH METHOD:** Multiple linear regression (MLR) is a type of statistical regression method. It is used to analyze the relationship between a dependent variable and two or more independent variables [10]. This method was taken for this paper because there are more than two independent controlled variables. In this research paper, the dependent variable(y) is Risk factor while the first independent variable is Total Population(x1), the second independent variable is Total DAG(x2), the third independent variable is the total household(x3), $4^{th}$ independent variable is www(x4), $5^{th}$ independent variable is DRH(x5), $6^{th}$ independent variable is DDH(x6), $7^{th}$ independent variable is Dist. Gained(x7) and the last independent variable is river type(x8).

**STEP 1:**

**PRIMARY WORK:** MLR is used to predict something or find the relationship between the dependent and independent variables. At first, the co-efficient of each independent variable is calculated. Then the dataset has been split into 50-50%, 66-34%, and 80-20% as trained data and test data to find the accuracy in each split.

**CROSS VALIDATION:** Cross-validation is a very useful technique in MLR. Cross-validation is a method in which data resampling is done to assess models' generalization ability and prevent overfitting [4]. Overfitting is one the insidious problem because no one can detect it unless cross-validations are carefully implemented. So cross-validation is one of the most vital steps for a research paper or verifying the results of this paper correctly. We have done cross-validation by dividing the total dataset into 10 sub-dataset using a random module. The training data used in the model is split, into k number of smaller sets, to be used to validate the model. The model is then trained on (k-1) folds of the training set. The remaining fold is then used as a validation set to evaluate the model.

**CONFUSION MATRIX:** A confusion matrix i.e. also called an error matrix, is a type of matrix or a table where we put the results or the performance of the MLR model i.e. the test data [3]. A

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 2 Volume No.7 March - April – 2023
DOI:10.46647/ijetms.2023.v07i02.012 ISSN: 2581-4621

confusion matrix is the shortest way to see and understand the result of the model. In the confusion matrix, there is a total of four parameters– TP, TN, FP, and FN. TP stands for 'True Positive' which shows the number of positive data classified accurately [7]. TN stands for 'True Negative' which shows the number of negative data classified accurately. FP stands for 'False Positive' which indicates the actual value is negative but predicted as positive. FP is also called a TYPE 1 ERROR. FN stands for 'False Negative' which indicates the actual value is positive but predicted as negative. FN is also called a TYPE 2 ERROR.

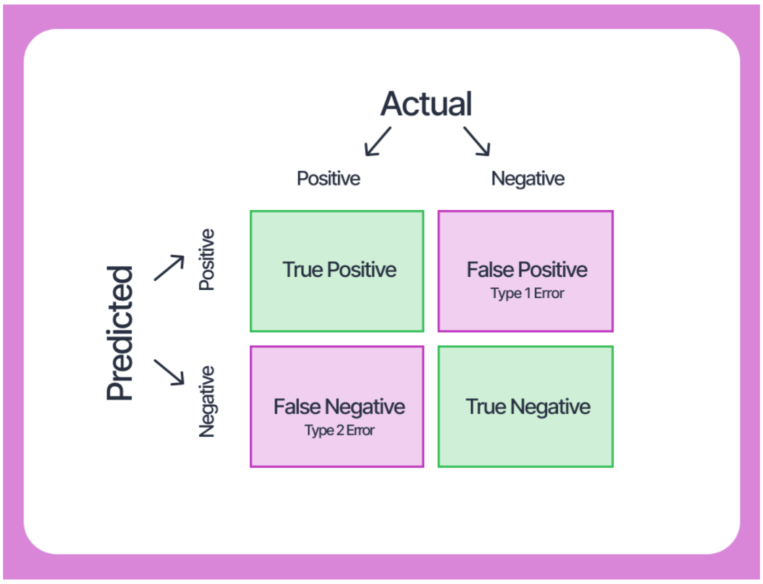


**Fig. 1:** Confusion Matrix

• **Accuracy:** In any model, it represents the ratio of the number of times the model can make the correct prediction to the total number of predictions.

• **Sensitivity:** It is defined as the ratio of the number of times a model can make a positive prediction to the total number of predictions. In this model, it is the number of times it has predicted the value of the response variable.

• **Specificity:** It is defined as the ratio of several times a model can predict that the result will be negative to the total number of times it has made the correct prediction.

• **Precision**: Precision is the method by which way one can say how correctly predicted cases turned positive.

• **F1_SCORE:** F1 score is the measurement of accuracy and it is the harmonic mean of precision and recall. Its maximum value can be 1 and the minimum value can be 0.

• **Kappa statistic:** Kappa is a way to measure agreement between three or more raters. The kappa statistic is frequently used to test interrater reliability. Like most correlation statistics, the kappa can range from −1 to +1. When two measurements agree only at the chance level, the value of kappa is zero. When the two measurements agree perfectly, the value of kappa is 1.0.

**STEP 2 – REQUIRED EQUATIONS FOR MLR:**

The model of multiple linear regression can be represented as:

$$Y = a+b_1x_1+b_2x_2+b3x3\ldots bnxn$$

Where, $a= y – (b_1x_1+b_2x_2+b3x3\ldots bnxn)$

$$bi = \frac{\sum_{i=1}^{n}(x-x)(y-\bar{y})}{\sum_{i=1}^{n}(x-\bar{x})^2}$$

Where i =1,2,3…..n

Y=risk factor
a=constant variable
b1= correlation coefficient of the first controlled variable
b2 = correlation coefficient of the second controlled variable

DOI:10.46647/ijetms.2023.v07i02.012 ISSN: 2581-4621

b3 = correlation coefficient of the third controlled variable
b4 = correlation coefficient of the fourth controlled variable
b5= correlation coefficient of $5^{th}$ controlled variable
b6 = correlation coefficient of $6^{th}$ controlled variable
b7= correlation coefficient of $7^{th}$ controlled variable
b8= correlation coefficient of $8^{th}$ controlled variable
$x_1$= Total Population
$x_2$=Total DAG
$x_3$= Total Household
$x_4$= www
$x_5$=DRH
$x_6$=DDH
$x_{7=}$ Dist. Gained
$x_{8=}$River Type

**CONTENTS OF CONFUSION MATRIX:**

Let's take-
TP= TRUE POSITIVE (High possibility of risk)
TN= TRUE NEGATIVE (Low possibility of risk)
FP= FALSE POSITIVE (Error, while predicting High possibility of risk)
FN= FALSE NEGATIVE (Error, while predicting Low possibility of risk)
Now,

Accuracy = $\frac{TP+TN}{TP+TN+FP+FN}$

Sensitivity = $\frac{TP}{TP+FN}$

Specificity= $\frac{TN}{TN+FP}$

Precision= $\frac{TP}{TP+FP}$

Recall = $\frac{TP}{TP+FN}$

F1_score= $\frac{2*Recall*Precision}{Recall+Precision}$

Kappa Test = $\frac{Observed\ Agreement-Expected\ Agreement}{100-\text{Expected Agreement}}$

Where
Observed Agreement = % (Overall Accuracy)
Expected Agreement = $\frac{(TP+FP)*(TP+FN)+(FN+TN)*(FP+TN)}{100}$

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 2 Volume No.7 March - April – 2023
DOI:10.46647/ijetms.2023.v07i02.012 ISSN: 2581-4621

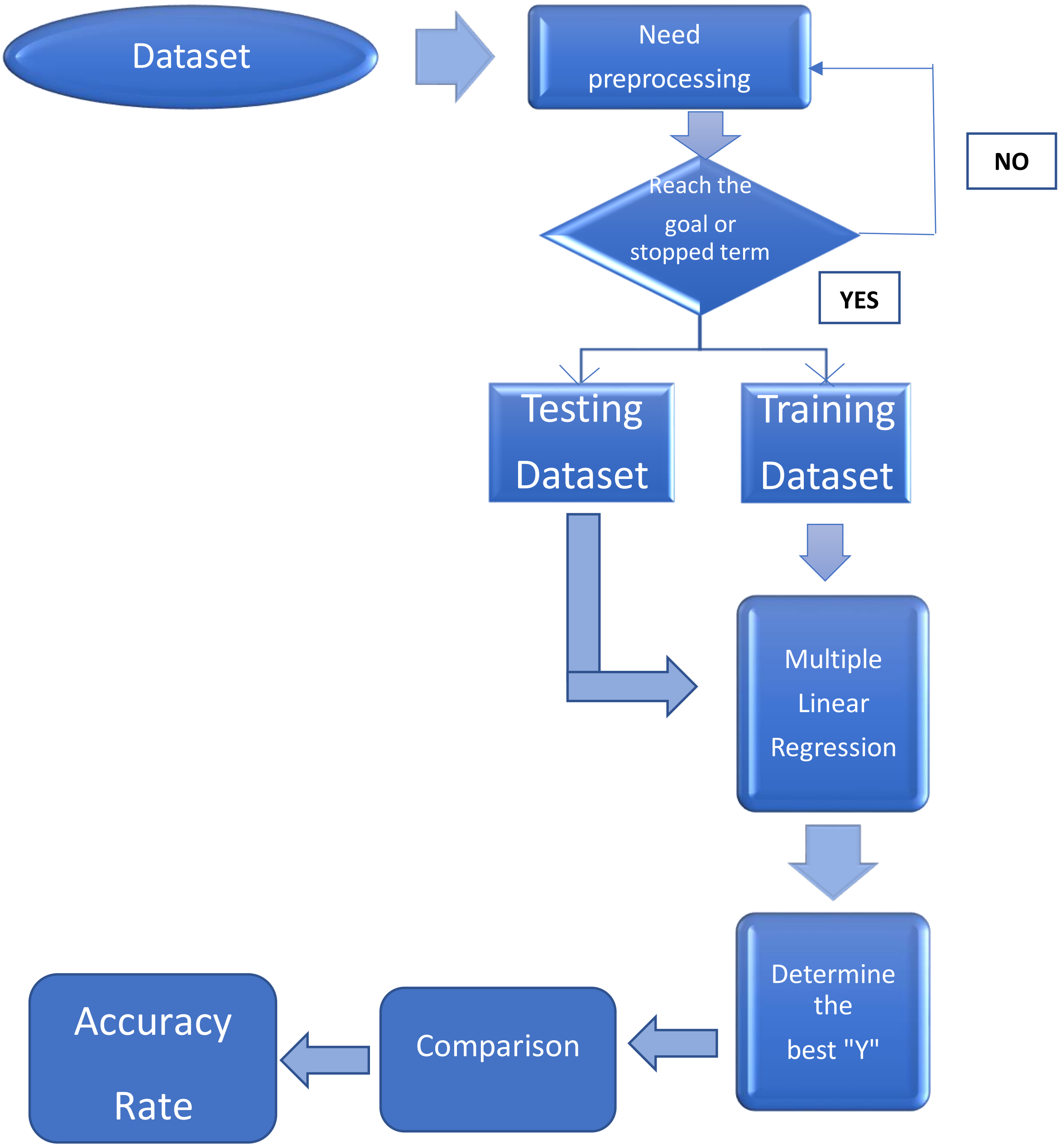


**Fig. 2:** The framework of the proposed MLR procedure

## 4. RESULTS AND DISCUSSION

**Table 1:** For 50–50% train-test splits:

| ATTRIBUTES | VALUE RANGE |
|---|---|
| Confusion Matrix | TP= 418 FP= 395<br>FN=746 TN=1339 |
| Accuracy (%) | 60.63 |
| Sensitivity | 0.36 |
| Specificity | 0.77 |
| Precision | 0.51 |
| Recall | 0.36 |

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 2 Volume No.7 March - April – 2023
DOI:10.46647/ijetms.2023.v07i02.012 ISSN: 2581-4621

| | |
|---|---|
| F1_Score | 0.42 |
| Kappa Test | 1.0 |

**Table 2:** For 66–34% of train-test splits:

| ATTRIBUTES | VALUE RANGE |
|---|---|
| Confusion Matrix | TP= 303 FP= 305<br>FN=464 TN=898 |
| Accuracy (%) | 60.96 |
| Sensitivity | 0.40 |
| Specificity | 0.75 |
| Precision | 0.50 |
| Recall | 0.40 |
| F1_Score | 0.44 |
| Kappa Test | 1.0 |

**Table 3:** For 80–20% train-test splits:

| ATTRIBUTES | VALUE RANGE |
|---|---|
| Confusion Matrix | TP= 203 FP= 214<br>FN=267 TN=474 |
| Accuracy (%) | 58.46 |
| Sensitivity | 0.43 |
| Specificity | 0.69 |
| Precision | 0.49 |
| Recall | 0.43 |
| F1_Score | 0.46 |
| Kappa Test | 1.0 |

**Table 4:** For 10-fold cross-validation:

| TEST CASES | CONFUSION MATRIX | ACCURACY (%) | SENSITIVITY | SPECIFICITY | PRECISION | F1_SCORE | KAPPA TEST |
|---|---|---|---|---|---|---|---|
| 01 | TP=68<br>FP=132<br>FN=81<br>TN=299 | 63.28 | 0.46 | 0.69 | 0.34 | 0.39 | 1.0 |
| 02 | TP=100<br>FP=107<br>FN=146<br>TN=227 | 56.38 | 0.41 | 0.68 | 0.48 | 0.44 | 1.0 |
| 03 | TP=94<br>FP=108<br>FN=84<br>TN=294 | 66.90 | 0.53 | 0.73 | 0.47 | 0.49 | 1.0 |
| 04 | TP=96<br>FP=143 | 50.69 | 0.40 | 0.58 | 0.40 | 0.40 | 1.0 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | FN=143<br>TN=198 | | | | | | |
| 05 | TP=109<br>FP=147<br>FN=110<br>TN=214 | 55.69 | 0.50 | 0.59 | 0.43 | 0.46 | 1.0 |
| 06 | TP=54<br>FP=38<br>FN=141<br>TN=347 | 69.14 | 0.28 | 0.90 | 0.59 | 0.38 | 1.0 |
| 07 | TP=82<br>FP=124<br>FN=87<br>TN=287 | 63.62 | 0.49 | 0.70 | 0.40 | 0.44 | 1.0 |
| 08 | TP=30<br>FP=68<br>FN=122<br>TN=360 | 67.24 | 0.20 | 0.84 | 0.31 | 0.24 | 1.0 |
| 09 | TP=135<br>FP=114<br>FN=122<br>TN=208 | 59.24 | 0.53 | 0.65 | 0.54 | 0.53 | 1.0 |
| 10 | TP=101<br>FP=68<br>FN=167<br>TN=244 | 59.48 | 0.38 | 0.78 | 0.60 | 0.46 | 1.0 |

## 5. CONCLUSIONS

Multiple regression techniques in AI can incorporate multiple parameters into a model. In terms of bridge engineering, a multiple regression model can be trained and tested on data available in the database in order to predict the risk factor of the possibility of death for not having a bridge in Nepal. In this study, multiple regression has been utilized and formulated to perform predictions about the risk factor. The proposed model considers a comprehensive set of geometric and functional parameters of the bridge structure to enhance prediction accuracy. In addition, many standardized approaches are adopted in the proposed model to improve its performance, including the evaluation of the most optimal set of model inputs, pre-processing and dividing the data, selecting internal parameters for control optimization, and model validation [16]. The results were then evaluated in terms of accuracy, sensitivity, specificity, precision, recall, Kappa test, and F1-Score. The model's accuracy is dependent on a number of factors such as excessive training on larger and more diverse datasets and configurations of multiple regression models; therefore, the proposed model can be improved by considering the most optimal combination of these factors. This type of project may help in the future to find any kind of prediction from any data field.

## 6. REFERENCE

**1.** Pal, S. S., Paul, S., Dey, R., Das, S., & Chaudhuri, A. K. Determining the probability of poverty levels of the Indigenous Americans and Black Americans in the US using Multiple Regression.
**2.** https://thehimalayantimes.com/nepal/bridge-failures-in-nepal-diagnosis-and-solutions
**3.** https://www.javatpoint.com/confusion-matrix-in-machine-learning
**4**. Berrar, D. (2019). Cross-Validation.

DOI:10.46647/ijetms.2023.v07i02.012 ISSN: 2581-4621

**5.** Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., ... & Dormann, C. F. (2017). Crossvalidation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography, 40(8), 913929.
**6.** Stone, M. (1978). Cross-validation: A review. Statistics: A Journal of Theoretical and Applied Statistics, 9(1), 127139.
**7.** Visa, S., Ramsay, B., Ralescu, A. L., & Van Der Knaap, E. (2011). Confusion matrix-based feature selection. MAICS, 710, 120-127.
**8.** Yerpude, P. (2020). Predictive modelling of crime data set using data mining. International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol, 7.
**9.** Cho, Y. H. (1972). A multiple regression model for the measurement of the public policy impact on big city crime. Policy Sciences, 3(4), 435-455.
**10.** Chaudhuri, A. K., Banerjee, D. K., & Das, A. (2021). A Dataset Centric Feature Selection and Stacked Model to Detect Breast Cancer. International Journal of Intelligent Systems and Applications (IJISA), 13(4), 24-37.
**11.**https://digitalcommons.wou.edu/cgi/viewcontent.cgi?referer=&httpsredir=1&article=1011&context=maurice
**12.** https://link.springer.com/chapter/10.1007/978-981-13-7446-3_14
**13.** Lu, P., Pei, S., & Tolliver, D. (2016, April). Regression model evaluation for highway bridge component deterioration using national bridge inventory data. In *Journal of the Transportation Research Forum* (Vol. 55, No. 1424-2017-1737, pp. 5-16).
**14.** Wang, Y. M., & Elhag, T. M. (2007). A comparison of neural network, evidential reasoning and multiple regression analysis in modelling bridge risks. *Expert Systems with Applications*, *32*(2), 336-348.
**15.** Bartlett, M. S. (1938, January). Further aspects of the theory of multiple regression. In Mathematical Proceedings of the Cambridge Philosophical Society (Vol. 34, No. 1, pp. 33-40). Cambridge University Press.
**16.** Althaqafi, E., & Chou, E. (2022). Developing Bridge Deterioration Models Using an Artificial Neural Network. *Infrastructures*, *7*(8), 101.
**17.** Addy, M., Chaudhuri, A. K., & Das, A. (2020, March). Role of Data Mining techniques and MCDM model in detection and severity monitoring to serve as precautionary methodologies against 'Dengue'. In *2020 international conference on Computer Science, Engineering and Applications (ICCSEA)* (pp. 1-6). IEEE.
**18.** Chaudhuri, A. K., Das, A., Sinha, D., & Banerjee, D. K. (2021). Early prediction of heart disease using the most significant features of diabetes by machine learning techniques. *Asian Journal For Convergence In Technology (AJCT) ISSN-2350-1146*, *7*(1), 168-178.
**19.** Saha, S., Mondal, J., Arnam Ghosh, M., Das, S., & Chaudhuri, A. K. Prediction on the Combine Effect of Population, Education and Unemployment on Criminal Activity Using Machine Learning.
**20.** Samanta, A., Chowdhury, A., Das, D., Dey, A. K., & Das, M. S. Prediction through machine learning on the dependence of job prospects in the Afro-American community on proficiency in English.