International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 2 Volume No.7 March - April – 2023
DOI:10.46647/ijetms.2023.v07i02.014 ISSN: 2581-4621

Prediction Of Technical Feasibility For Building Bridges In Nepal By Using Data Mining Technique

Mr. Om B Khadka[1], Ayan Gupta[2] , Ayushmaan Paul[3], Sayantan Adhikary[4], Dr.Avijit Kumar chaudhari[5], Prof.Dilip K.Banerjee[6], Dr.Anirban Das[7]

[1] Research Scholar, Department Of IT & Knowledge Management, Seacom Skills University,
[2,3,4]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]Professor, University Of Engineering & Management.
[1]https://orcid.org/0000-0003-3100-1295,[2]https://orcid.org/0000-0001-7769-7545,
[3]https://orcid.org/0000-0002-3463-5931,[4]https://orcid.org/0000-0001-8735-4845,
[5]https://orcid.org/0000-0002-5310-3180

ABSTRACT

A bridge is a structure that spans a valley, road, river, body of water, or any other physical barrier. The architecture of a bridge may vary depending on the nature of the area where it will be built. The first bridge was created by nature itself, with a log falling across a stream or rocks in a river. The first man-made bridges were probably constructed using wood or timbers, and later, stones were used to support a crossbar system. Some early bridges used trees or bamboo poles to cross small tunnels or wells to get from one place to another. Lashing sticks, logs, and a common form of bridge construction involved the use of temporary branches woven together with long trunks or other fibers to form a connecting rope capable of withstanding weight. Mountainous areas pose a unique problem for constructing bridges due to varying climates, geological features, and hydrology parameters. Based on bridge sites and various constraints, types of bridges and construction methods are carefully selected for safe, profitable, and successful completion of the bridge. Human habitation and plants have spread up to 14,000 to 16,000 feet above mean sea level. An innovative idea would involve the installation of an adjunct bridge to provide services such as education, medical facilities, and workplaces. Data has been analyzed using the Random Forest method. In this research, we predicted the technical feasibility of building a trail bridge depending on factors such as total population, total households, river type, total DAG, DDH, DRH, etc. The model was evaluated using 50-50%, 66-34%, and 80-20% train-test splits, and 10-fold cross-validation, with an accuracy of approximately 97%.In this study, the authors collected data by mobilizing local NGOs and informing the public through local radio. A comprehensive study was conducted to collect nationwide bridge demand. With the project management information system developed by Om B Khadka, Head of IT and Knowledge Management, serves as the secondary source.

Keywords- Random Forest, Technical Feasibility

1.INTRODUCTION

From ancient times, safe river crossing was a major challenge in Nepal, and even today, it remains a significant issue, There are still traditionally built bridges with local materials, lack of proper maintenance, and many bridges are not safe for crossing. For instance, a recent study has identified that there are over 9,000 bridges in Nepal, and approximately 8.2% of these bridges are either structurally not safe to cross, obsolete or defunct.

The geographical condition presents a unique challenge for bridge construction in Nepal. To address this challenge, authorities in Nepal suggested using of random forest algorithms to maintain these structures and ensure the safe operation of bridges. Using the RF model, it is relatively easy to predict the technical feasibility of constructing bridges in Nepal.

The challenges in constructing bridges are numerous and varied.

1. Geographical extremes, building bridges over deep gorges and in high mountains.
2. Transportation of construction materials due to limited and seasonal road connectivity.

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 2 Volume No.7 March - April – 2023
DOI:10.46647/ijetms.2023.v07i02.014 ISSN: 2581-4621

3. Constructing long-span trail bridges over wide rivers in the southern belt of Nepal.
4. Unavailability of steel cables and fabricated materials.
5. Lack of required financial and human resources.
6. Various socio-economic factors.

To address these challenges, in this study we used Random Forest models to predict ratings for various components of bridges, including the bridge deck, superstructure, and substructure. The proposed Random Forest models showed a high correlation between predicted and actual values, indicating that the models were effective in predicting the construction of bridges.

The proposed model serves as an information-based state prediction method for bridge components, providing an example of prediction accuracy. Additionally, the model was established as a data-based modeling method for condition ratings, which can improve decision-making processes using the Random Forest algorithm.

2. LITERATURE REVIEW

Cutler et al, 2012 stated that random forests are an extension of bagging idea and were developed as a competitor to boosting. Random Forests can be used for either a categorical response variable, referred to in as "classification," or a continuous response referred to as regression. Similarly, the predictor variables can be either categorical or continuous.[10] This paper helped us to gain knowledge about Random Forest.

A model trained on biased data will fail to classify a spam sample. An approach based on undersampling is presented in this paper to tackle imbalanced training data. It involves synthetically creating new minority class samples from the existing ones until the balance in data is achieved. The enhanced data is then passed through various classifiers for which the performance is recorded.[12]

Berrar, 2019 stated that Cross-validation is one of the most widely used data resampling methods to estimate the true prediction error of models and to tune model parameters. This article provides an introduction to the most common types of cross-validation and their related data resampling methods.[8].With the help of this paper, we understood the logic of cross-validation and K-fold and how to use it in our paper.

Pal et al., in their research article, expressed the uses of data-mining methods, i.e., Decision Tree (DT) and K-Nearest Neighbor (KNN) via the k-fold cross-validation (CV) technique.[9]

3.METHODOLOGY

3.1 The Proposed Approach:

3.1.1 Data Cleaning:

Data cleaning is the process of correcting or removing incorrect, corrupt, malformatted, duplicate, or incomplete data within a dataset. When combining multiple data sources, there are many opportunities to duplicate or mislabel data If the data is wrong, the results and algorithms are unreliable, even if they look correct. In our dataset some data field contains incomplete data, they are causing faulty results so we have to remove them for better results.

3.1.2 Overcome from Oversampling Issue:

The training dataset initially contains an uneven distribution of classes. We propose oversampling the minority class using the random oversampling Technique to reduce the variance present in the training dataset. The technique Random oversampling is the simplest oversampling technique to balance the imbalanced nature of the dataset. It balances the data by replicating the minority class samples. This does not cause any loss of information. We implement the technique on training data such that the number of minority and majority class samples becomes equal. In our dataset, we encountered oversampling in the Technical feasibility field. The field technical feasibility only has two types of data 0 and 1. In the field, at first, the no of 1 is more than zero's it's like 5000:400 so we proposed random oversampling technique to overcome this problem.

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 2 Volume No.7 March - April – 2023
DOI:10.46647/ijetms.2023.v07i02.014 ISSN: 2581-4621

3.1.3 Random Forest :

Random Forest is a machine learning algorithm that is commonly used for classification and regression tasks. It is an ensemble method that combines multiple decision trees and generates a more accurate and robust prediction. The algorithm works by creating a large number of decision trees (also called "forest"), each of which is trained on a random subset of the features and a random sample of the training data. This process is called "bagging" (short for "bootstrap aggregating"). Each decision tree is constructed by recursively splitting the data based on the feature that provides the most information gain or the best split according to some criterion (e.g., Gini impurity or information gain). Once all the decision trees have been constructed, the prediction of the random forest is obtained by averaging the output of all the individual trees (in the case of regression) or by taking a majority vote (in the case of classification). Random Forest has several advantages over individual decision trees. First, it reduces overfitting, which is a common problem with decision trees, by combining the predictions of multiple trees. Second, it is relatively insensitive to the choice of hyperparameters, such as the maximum depth of the trees, because the ensemble approach smooths out the noise in each individual tree. Third, it can handle a large number of features and a large training set, making it suitable for high-dimensional datasets. Random Forest is widely used in various applications, such as image classification, speech recognition, credit scoring, and drug discovery, to name a few. Its flexibility, accuracy, and ease of use make it one of the most popular machine-learning algorithms.

3.2 DATA PROCESSING:

3.2.1 Dataset Description:

In this paper, data was taken from mobilizing local NGOs and informing the public through local radios, for a comprehensive study to gather demand for bridges nationwide. In this dataset, there is a total of 37 columns. Out of this 10 columns are taken (one dependent variable and 9 independent variables). 10 columns are-“Technical Feasibility” i.e. the possibility of making bridges or not if the value of technical feasibility is 1 then it is feasible to build the bridge in the area if the value is 0 then we can’t build the bridge on that area, “www” i.e. it is the width of the bridge, “Risk Factor ” i.e. the possibility of death for not having a bridge a “DRH” i.e. it is the distance to nearest road in hours, “DDH” i.e. Distance to district headquarter in the day, “Total Household” i.e. a household includes the related 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”, Total DAG is sum of total dalit, total janajati, total minor and total otherpoor of a area , “Dist. Gained” i.e. this is distance gained, “River type”.

Table 1: For attributes mean and standard deviation:

ATTRIBUTES	MEAN	STANDARD DEVIATION
Technical Feasibility	0.584432	0.49282
Total Population	2646.187	4254.327
Total DAG	2231.225	3556.377
Total Household	1028.446	54400.55
www	92.523	21.0912
DRH	1.285	1.848
DDH	1.123	1.273
Dist. Gained	2.959	21.781
River Type	5.894	3.376

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 2 Volume No.7 March - April – 2023
DOI:10.46647/ijetms.2023.v07i02.014 ISSN: 2581-4621

Risk Factor	0.251	1.112

3.2.2 Research Method:

Random forest is a supervised learning algorithm. The forest it builds is an ensemble of decision trees, it is usually trained with the bagging method. The general idea of the bagging method is that a combination of learning models increases the overall result.

STEP 1-DEVELOPING EQUATION OF RANDOM FOREST

Random forest is a commonly-used machine learning algorithm trademarked by Leo Breiman and Adele Cutler, which combines the output of multiple decision trees to reach a single result. Its ease of use and flexibility have fueled its adoption, as it handles both classification and regression problems.

The final feature importance, at the Random Forest level, is it's average over all the trees. The sum of the feature's importance value on each trees is calculated and divided by the total number of trees:

$$RFfi_i = \frac{\Sigma j \epsilon all\ trees\ \times normfi_{ij}}{T}$$

- $RFfi_i$= the importance of feature i calculated from all trees in the Random Forest model
- $normfi_{ij}$= the normalized feature importance for i in tree j
- T = total number of trees

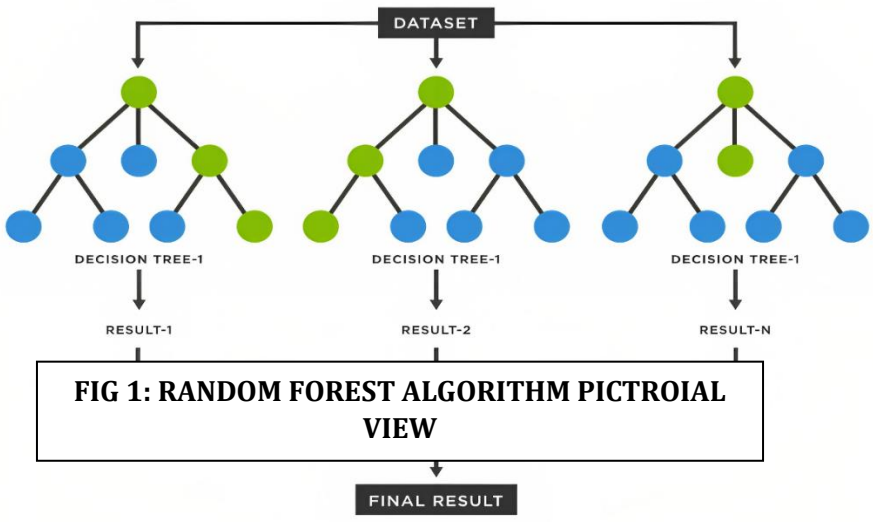

FIG 1: RANDOM FOREST ALGORITHM PICTROIAL VIEW

STEP 2- Performance Metric

Cross-validation is a technique for evaluating ML models by training several ML models on subsets of the available input data and evaluating them on the complementary subset of the data. Use cross-validation to detect overfitting, ie, failing to generalize a pattern.

The confusion matrix visualizes the accuracy of a classifier by comparing the actual and predicted classes. The binary confusion matrix is composed of squares. 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. A 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 variables– TP, TN, FP, and FN. TP stands for 'true positive' which shows the number of positive data classified accurately. 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

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 2 Volume No.7 March - April – 2023
DOI:10.46647/ijetms.2023.v07i02.014 ISSN: 2581-4621

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.

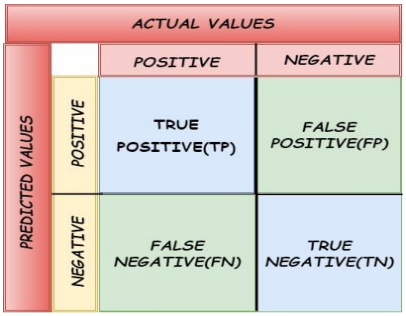

FIG 2: CONFUSION MATRIX

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 correct predictions. In our model, it is the number of times it has predicted that the value of the response variable (risk factor) will be equals 1 to the number of times it has assumed that the value will be greater than equals 0.3.

Mathematically, sensitivity or true positive rate can be calculated as the following:

$$\textbf{Sensitivity} = \frac{(\textbf{True Positive})}{(\textbf{True Positive + False Negative})}$$

SPECIFICITY: When sensitivity is used to evaluate model performance, it is often compared to specificity. Specificity measures the proportion of true negatives that are correctly identified by the model. This implies that there will be another proportion of actual negative which got predicted as positive and could be termed as false positives. This proportion could also be called a True Negative Rate (TNR). The sum of specificity (true negative rate) and false positive rate would always be 1

Mathematically, specificity can be calculated as the following:

$$\textbf{Specificity} = \frac{(\textbf{True Negative})}{(\textbf{True Negative + False Positive})}$$

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

F1-Score: The F1 score is the evaluation matrix that combines two matrices: Precision and Recall, into a single metric by taking their harmonic mean.F1 score Formula

KAPPA STATISTICS: Cohen's kappa coefficient is a statistic used to measure the reliability of qualitative items. In general, κ is considered a more robust measure than a simple percentage agreement calculation because it takes into account the likelihood of agreement occurring by chance. Like most correlation statistics, the kappa can range from −1 to +1

DEVELOPING EQUATION OF CONFUSION MATRIX:

Let's take

TP= TRUE POSITIVE

TN= TRUE NEGATIVE

FP= FALSE POSITIVE

FN= FALSE NEGATIVE

Now,

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

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

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 2 Volume No.7 March - April – 2023
DOI:10.46647/ijetms.2023.v07i02.014 ISSN: 2581-4621

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

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

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

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

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

Where,

Observed Agreement = %(Overall Accuracy)

Expected Agreement = $\frac{(TP+FP)*(TP+FN)+(FN+TN)*(FP+TN)}{100}$

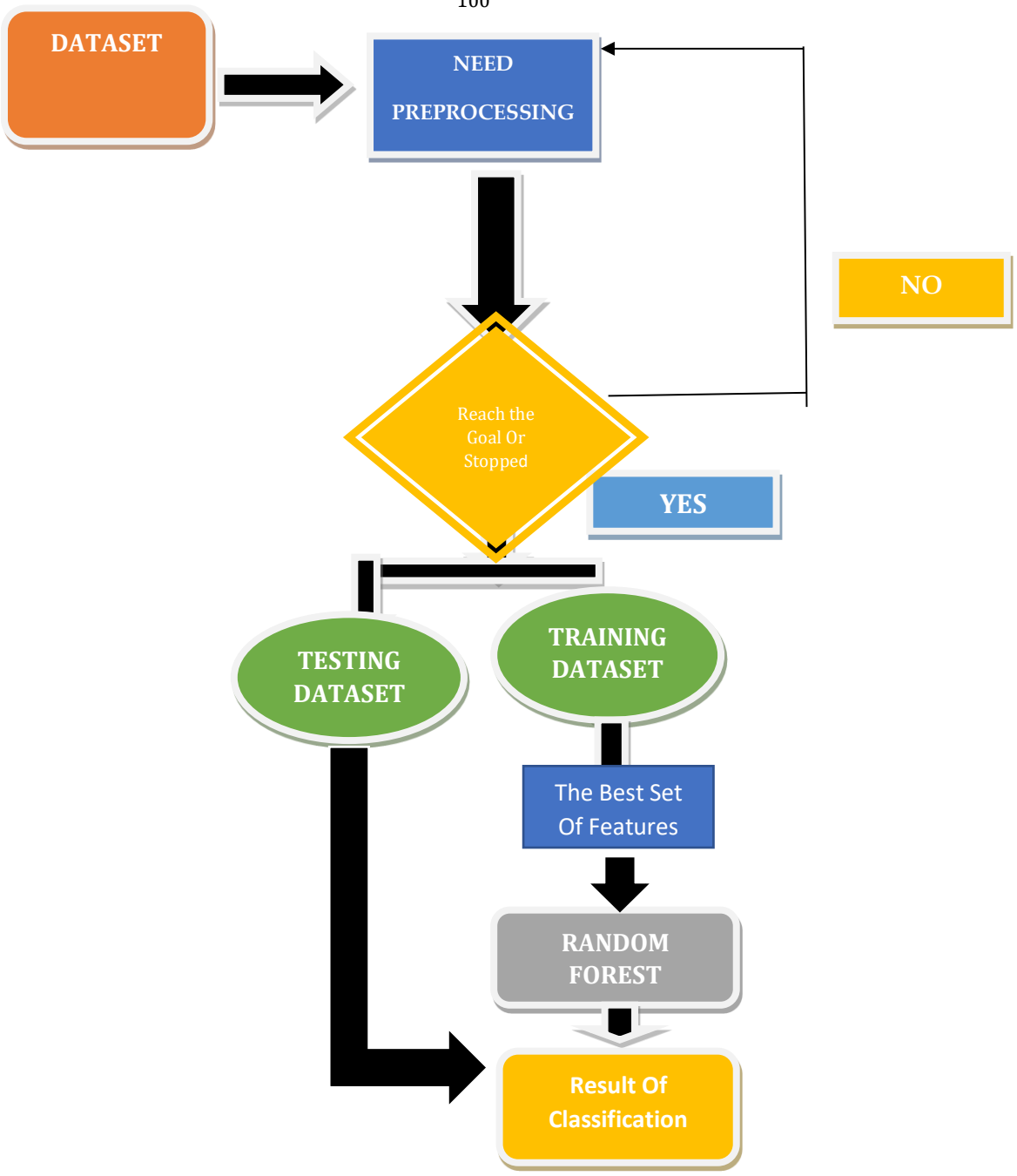

Fig. 2: The framework of the proposed RF procedure

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 2 Volume No.7 March - April – 2023
DOI:10.46647/ijetms.2023.v07i02.014 ISSN: 2581-4621

4.RESULTS AND DISCUSSION

4.1 Train-Test Split:

When it comes to data analysis field, we have to split our data into training and testing sets. A train test split is when we split our data into a training set and a testing set. The training set is used for creating training the model, and the testing set is used to create test our model. This allows us to train our models on the training set, and then test the accuracy on the unseen & unused testing set. There are a few different ways to train test split, but the most common way is to simply split the data into two sets. For example 66% for training and 34% for testing. This ensures that both sets as representative of the entire dataset, and gives us a good way to measure the accuracy of our models. Here we used 3 types of splitting techniques that are,

A) **50-50% train-test split**
B) **66-34% train-test split**
C) **80-20% train-test split**

Table 2: For 50–50% train-test splits:

ATTRIBUTES	VALUE RANGE
Confusion Matrix	TP= 2558 FP= 137 FN=0 TN=1969
Training Accuracy (%)	99.89%
Testing Accuracy (%)	97.06%
Sensitivity	1.00
Specificity	0.93
Precision	0.93
Recall	1.00
F1_Score	0.97
Kappa Test	0.94
ROC_AUC	0.97

Table 3: For 66–34% of train-test splits:

ATTRIBUTES	VALUE RANGE
Confusion Matrix	TP= 1783 FP= 60 FN=0 TN=1329
Training Accuracy (%)	99.89%
Testing Accuracy (%)	98.11%
Sensitivity	1.00
Specificity	0.96
Precision	0.96
Recall	1.00
F1_Score	0.98
Kappa Test	0.96
ROC_AUC	0.98

Table 4: For 80–20% train-test splits:

ATTRIBUTES	VALUE RANGE
Confusion Matrix	TP= 1068 FP= 23 FN=0 TN=775
Training Accuracy (%)	99.99%

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 2 Volume No.7 March - April – 2023
DOI:10.46647/ijetms.2023.v07i02.014 ISSN: 2581-4621

Testing Accuracy (%)	98.77%
Sensitivity	1.00
Specificity	0.97
Precision	0.97
Recall	1.00
F1_Score	0.99
Kappa Test	0.97
ROC_AUC	0.98

4.2 k-Fold Cross Validation Technique:

With this method, we have a data set which we randomly divide into K parts. We use K-1 of those parts for training and reserve a tenth for testing. We repeat this procedure K times each time reserving another Kth element for testing.

Here we used K as 10.

10 Fold CROSS VALIDATION RANDOM FOREST MODEL:

ATTRIBUTES	VALUE RANGE
Average Accuracy (%)	97.588%
Sensitivity	0.97
Specificity	0.97
Precision	0.97
Recall	0.97
F1_Score	0.97
Kappa Test	0.95
ROC_AUC	0.99

CONCLUSIONS

A Random Forest Algorithm is a supervised machine learning algorithm which is extremely popular and is used for Classification and Regression problems in Machine Learning. We know that a forest comprises numerous trees, and the more trees more it will be robust. “Technical Feasibility” i.e. the possibility of making bridges or not if the value of technical feasibility is 1 then it is feasible to build the bridge on the area if the value is 0 then we can’t build the bridge on that area.

In this study, Random Forest Algorithm has been utilized and formulated to perform predictions about the technical feasibility. 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 .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 random forest 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.

REFERENCE

[1] International Journal of Advances in Engineering and Management (IJAEM) Volume 2, Issue 1, pp: 655-682 www.ijaem.net ISSN: 2395-5252(A Study on Construction Challenges of Bridges in The Hilly Areas by Rahul Yadav)

DOI:10.46647/ijetms.2023.v07i02.014 ISSN: 2581-4621

[2] 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.
[3] Althaqafi E, Chou E. Developing Bridge Deterioration Models Using an Artificial Neural Network. *Infrastructures*. 2022; 7(8):101. https://doi.org/10.3390/infrastructures7080101
[4] Oluwaseun Priscilla Olawale, Fezile Ozdamli, Kamil Dimililer, "Data Mining Techniques for the Classification of Medical Cases: A Survey", *2021 5th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, pp.68-73, 2021.
[5] Einstein, A., B. Podolsky, and N. Rosen, 1935, Phys. Rev. **47**, 777-780.
[6] Aziz Saaidia, Ahmed Belaadi, Abdelkrim Haddad. (2022) Moisture Absorption of cork-based Biosandwich Material Extracted from *Quercussuber L*. Plant: ANN and Fick's Modelling. *Journal of Natural Fibers* 19:16, pages 12486-12503.
[7] Hu, J., & Dong, F. (2022). Finite Element of Nanoscale Carbon Fiber-Reinforced Concrete Bridge Engineering Monitoring Based on Data Mining Technology. *Advances in Materials Science and Engineering*, *2022*.
[8] D. Petkov, O. Petkova, T. Andrew, T. Nepal,
Mixing Multiple Criteria Decision Making with soft systems thinking techniques for decision support in complex situations,Decision Support Systems,Volume 43, Issue 4,2007,
[9] Pal, S. C., Islam, A. R. M. T., Chakrabortty, R., Islam, M. S., Saha, A. , & Shit, M. (2022). Application of data-mining technique and hydro-chemical data for evaluating vulnerability of groundwater in Indo-Gangetic Plain. *Journal of Environmental Management*, *318*, 115582.
[10] Cutler, A., Cutler, D. R., & Stevens, J. R. (2012). Random forests. *Ensemble machine learning: Methods and applications*, 157-175.
[11] Baruah, R. D., Roy, S., Bhagat, R. M., & Sethi, L. N. (2016, December). Use of data mining technique for prediction of tea yield in the face of climate change of Assam, India. In *2016 international conference on information technology (ICIT)* (pp. 265-269). IEEE.
[12] Pennington, Deana, et al. "Bridging sustainability science, earth science, and data science through interdisciplinary education." *Sustainability Science* 15 (2020): 647-661.
[13] Kumar, Akshay, Alok Bhushan Mukherjee, and Akhouri Pramod Krishna. "Application of conventional data mining techniques and web mining to aid disaster management." *Environmental Information Systems: Concepts, Methodologies, Tools, and Applications*. IGI Global, 2019. 369-398.
[14] Ratadiya, P., & Moorthy, R. (2019). Spam filtering on forums: A synthetic oversampling based approach for imbalanced data classification. *arXiv preprint arXiv:1909.04826*.
[15] Machine Learning Mastery, Random Oversampling and Undersampling for Imbalanced Classification.
[16] Ron Zacharski, Author of A Programmer's Guide to Data Mining, Training Sets, Test Sets, and 10-fold Cross-validation.
[17] Wikipedia, Random Forest
[18] Jehad Ali, Rehanullah Khan, Nasir Ahmad, Random Forests and Decision Trees
[19] Guide To Data Cleaning: Definition, Benefits, Components, And How To Clean Your Data, tableau.
[20] Satyam Kumar,7 Over Sampling techniques to handle Imbalanced Data
[21] Animesh Samanta, Akash Chowdhury , Dip Das , Arup Kumar Dey , Mrs. Sulekha Das,Prediction through machine learning on the dependence of job prospects in the Afro-American community on proficiency in English