IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

# “Machine Learning Based Regression Model to Predict Health Insurance Claim”

Name of 1stAuthor: ROHINI H K, Name of 2nd Author: SAHANA G C

[1]Designation of 1st Author: BE ,CSE,RIT HASSAN ,[2]Designation of 2nd Author:Assistant Professor ,Cse Department Rajeev institute of technology

Department of 1st Author: computer Science and engineering Rajeev institute of Technology, Hassan,
Department of 2nd Author: ASSISTANT PROFESSOR at Rajeev institute of technology (computer science) hassan
City: Hassan State:Karnataka,country:india

***Abstract:*** In today’s world, where there are a lot of accidents and disasters happening, there is a need for a huge amount of money for health care as the cost of treatments and medicines requirements are very high. There are many cases where people can’t afford the treatment and lose their life. To prevent this, government along with many private banks tied up with numerous hospitals to set up many insurance agencies. People deposit some amount of money in the bank on a monthly or yearly basis in return for which the insurance agency provides money as health insurance. Costs of health insurance throughout the globe have increased critically in recent times. The increased trend results in the requirement of quick decision-making while providing better accuracy. This paper deals with the prediction regarding claiming health insurance i.e., whether a person claims his/her health insurance based on different factors. Using human brains for prediction agencies faces different errors which often results in extra administrative effort for reworking and finally leads to very little accuracy. Machine Learning algorithms can be the alternatives for the above-said problems as they can handle huge amounts of data. In this paper, Artificial Neural Networks (ANN), Decision Trees (DT), Support Vector Machines (SVM) and Logistic Regression Models (LRM) are applied to the dataset. The DT shows the best accuracy amoung all the applied algorithms.

Index Terms: Insurance, Classification, Machine Learning, ,SVM, ANN, Logistic Regression, Decision Tree;

## INTRODUCTION

For fostering a medical care framework, there is an immersed necessity to go through bearings that are lined up with advancement. Perhaps the main direction is expenses of the life protection Insurance. However, there is very much difficulty in breaking down the expenses of life protection Insurance given countless input factors. Then again, this is a major information issue, which requires a more powerful methodology. Logical and innovative advances progressively are considered better for fitting life care protection plans to particular health hazard profiles [1]. Nonlinear expense partaking in health care coverage energizes transient replacements since patients can decrease their cash-based expenses by focusing on the correct time in the years when they hit the deductible [2]. Policymakers are keen on the effect of health care coverage on people's clinical uses [3]. This paper [4] fostered a model of protection decision and determined the ramifications for insurance plan determination when customers are misfortune disinclined since past work had shown that medical coverage plan exchanging expenses can expand government assistance by lessening unfavorable choices. A costly protection strategy doesn't support interest in loss reduction work as much it ought to [5]. A model of backup plan with value setting and consumer’s health care assistance under hazard change, a strategy usually used to battle wasteful arranging due to unfavorable determination in health care coverage markets was created concurring to the review [6]. In this review [7], it was tracked down that a 10% increment in last merchandise and clinical area usefulness stuns, each has a constructive outcome on total government assistance. In this review [8], the lifetime impacts of exogenous changes in health care coverage included on the dynamic ideal designation (utilization, recreation, and wellbeing use), status (wellbeing and riches), and government assistance and results feature positive impacts of protection on wellbeing, riches, and government assistance, just as midlife replacement away from solid relaxation for more wellbeing costs, brought about by cresting compensation, and speeding up medical problems. The impact of a singular protection order on the interest for private health care coverage in the US was investigated in this review [9] and it was shown that this arrangement essentially affects the general interest for private health care coverage in the US. Since dissecting of the medical care protection costs is presently a major information issue, there is a need to utilize computational insight draws near due to high nonlinearity and an enormous piece of information. To streamline the expectation interaction of the medical services protection costs in this review, a determination strategy is performed to separate the main elements. Artificial Neural Network and some concepts of Adaptive Neuro Fuzzy Inference System (ANFIS) [10- 16], Logistic Regression, Decision Tree Classifier, and Support Vector Machine are utilized for the chosen method.

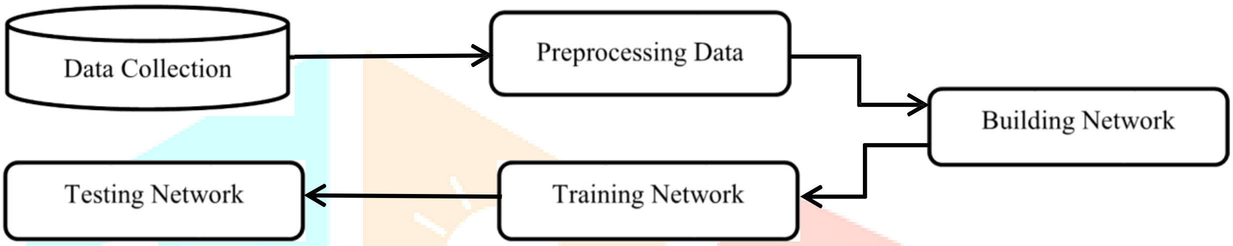

Figure 1: Basic flow diagram

## DATA DESCRIPTION:

The dataset utilized in the paper depends on the Database available on Kaggle [20]. The data is additionally accessible

Over GitHub. The recipient's living regions are the northeast, southeast, southwest, what's more, the northwest in the US. Totally 1338 input/output tests values on the expenses of the singular health protection are present [17]. Output addresses charges on the singular clinical expenses charged by medical coverage. There are five contributions to add up to as given underneath:

A. Current Age: time of the essential recipient, which is at least 18 summers and the most extreme, is 64. Normal is 39.2.

B. Gender: female, male.

C. BMI: The Body mass index is giving comprehension of the body along with weights that are moderately high or low comparative with tallness, target file of this (kg/m2) utilizing the proportion of tallness to mass, preferably over 18.5 to 24.9. 15.96 BMI is insignificant and 53.12 is the greatest. 30.66 is the Normal BMIKids: No of kids covered by medical coverage. Markers reach between 1 and 5.

D. Smoker: Smokes or not. Overall in total 1064 smokes and 274 don’t.

## Models:

### A. ANN MODEL

ANN, a hybrid AI procedure, which utilizes the artificial neuron framework. The association of this technique empowers the framework to learn and to save the learned information. The learned information can additionally be utilized without retraining the ANN. This hybrid learning technique builds an input/output planning dependent on human information and specified input/ output sets [18]. Figure 1 represents the basic workflow of building the ANN model. In the beginning, raw data is collected from a verified source for processing further. The collected data is to fulfill the requirement of the proposed network and it is completely compatible as per the format of the data which the proposed network supports. Otherwise, the network will not process the dataset further. We are using Relu Activation Function (RAF) with 20, 18, 16, and 12 as weights respectively in the hidden layer while the sigmoid for the final output layer having 1 as weight. Where the adam optimizer is used.

## B. LOGISTIC REGRESSION MODEL:

Factual investigation techniques used in anticipating any information consider depending on earlier psychology of an information collection is named as Logistic Regression. The logistic regression somewhat relates to the neural network. Logistic Regression can also be presented as a single-layer neural network. Using a logistic sigmoid function for activation functions in a neural network's hidden layer is very much common

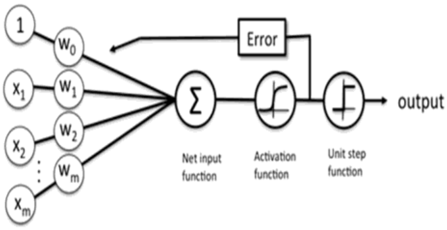

Figure 2: Schematic of a Logistic Regression Classifier

## C.DECISION TREE CLASSIFIER MODEL:

A supervised learning technique is used for both Regression along classification problems, yet generally, the decision tree is liked in supporting Classifications determined on verifiable information. This had turned into an important instrument in this AI discipline. Under Machine Learning, DT is one of the simplest and most versatile structures that can be used very much for classification problems. A DT under Machine Learning is fundamentally a "tree" of decisions that makes up the nodes where "branches" are the split of the tree. Every individual node, as well as its sub-nodes, makes decisions based on predetermined variable values, ultimately leading to the classification of each elements. into specific classes [19]. The initial variable, known as the root, divides the dataset, serving as the starting point for the entire process. Each decision made is referred to as a node, and the connections between these decisions are called branches.These tree classifiers are structured in a hierarchical manner, where internal nodes represent the conditions or features of a dataset, branches represent the decision criteria, and each leaf node represents the outcome or result.

## D.SUPPORT VECTOR MACHINE MODEL:

SVM is the general, most famous Supervised Learning algorithm, which is handled for Classification just as Regression. SVM calculation's main objective is to build the appropriate line or choice limit which can identify n- dimensional space into classes which can undoubtedly allow us to place the new data point in the appropriate classification later on.

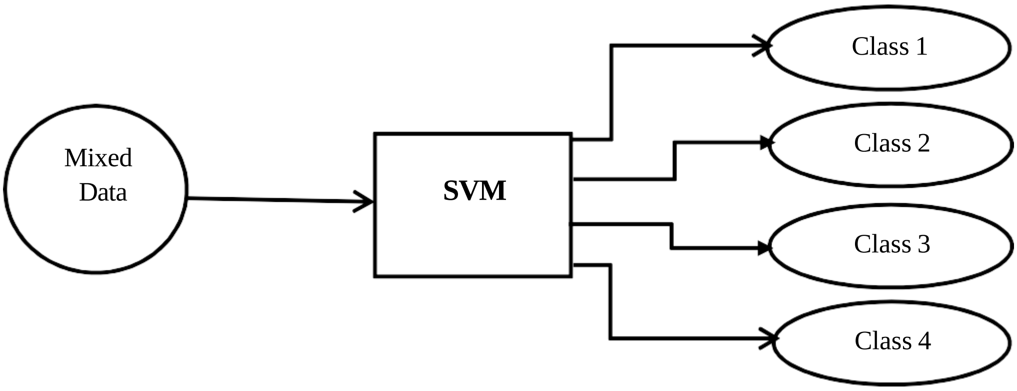

FIGURE 3 : Support Vector Machine Process

This solves different linear and non-linear problems and also works very well in the case of many practical problems. Moreover, fundamentally, In AI Domain for classification SVM is utilized This best choose limit is coined as a hyperplane that separates the data into classes. SVM collects the excessive vectors/focuses that assist in making the hyperplane. Those unusual cases are known as help vectors, and consequently, the calculation is named Support Vector Machine. Data is being applied on 4 different kernels of SVM i.e,sigmoid, linear,poly, and RBF

## METHODOLOGY:

As mentioned the goal of this is to build models that use the company's data on clients to predict claims and their costs more accurately, and to assist the subscribers during the policy renovation process. As is common in the insurance business, two main components need to be considered to estimate the clients' costs: the number of claims and the cost of those claims or, in other words, the frequency and the costs. This data-driven renovation process is still in its first stages in the company, so this report is detailing a pilot project which comprises only a few selected clients (companies) and the data spans over the course of three years, from 2017 to 2019.Both datasets were extracted from the company servers, using SAS Enterprise Guide, by merging a panoply of different tables containing the information deemed necessary to build the proposed models, for example, customers' gender, age, area of residence, claim and diagnosis history, among others.The data was properly handled, treated, and anonymised to protect the clients' identity and to comply with data protection rules. These two events are estimated separately, as they are dissimilar and are often explained by different variables, thus there is a need to have two different data frames: the Cost Dataset and the Frequency Dataset.In the beginning, data is collected from a website named "Kaggle" in the raw form. Then this raw information is preprocessed where all the NAN values were handled and then the standardization of the data is done. After the preprocessing of the data, the required features are extracted as per the need, and then the data splits for training and testing purposes. For training 70% of the data is provided and for testing 30%. Then this data is fitted into the proposed models i.e., LR, DT, SVM, ANN. After the testing of the models, the prediction is taken out and the performance of all the proposed models is compared.

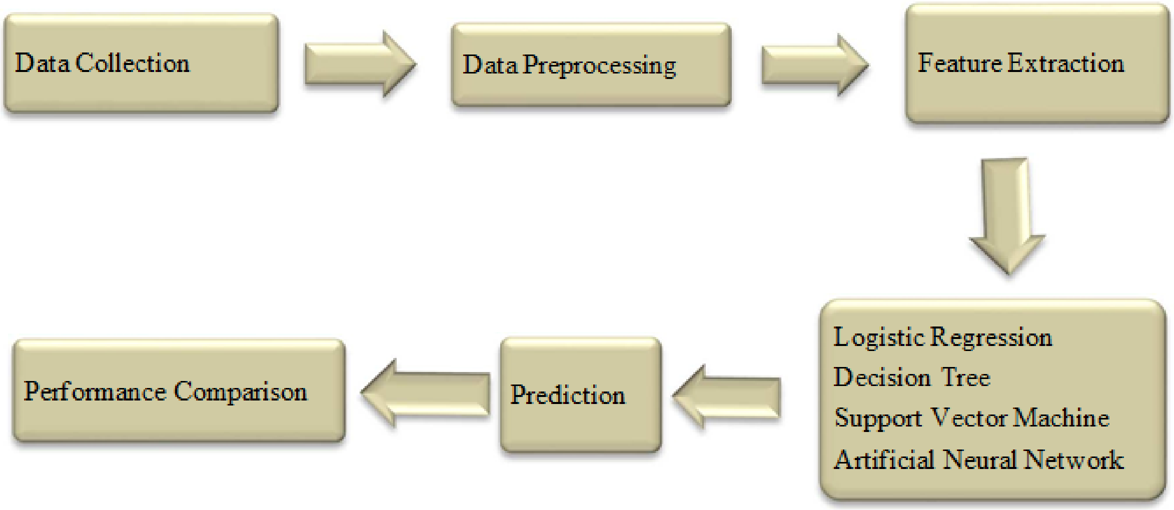

Figure 4: Workflow Diagram of Proposed Methodology

## RESULT ANALYSIS:

Upon comparing the outputs of the processed data using various proposed models, it becomes apparent that the Artificial Neural Network model achieves a high accuracy of 93%, which is considered excellent. Following that, the Logistic Regression model exhibits an accuracy of 87%, which is also commendable. Lastly, the Decision Tree model demonstrates its performance with a slightly lower accuracy compared to the previous two models. with the highest accuracy of 96.76% which depict that our decision tree model can predict more accurately If we talk about the Support vector Machine, with its 4 different kernels the accuracy of the sigmoid kernel is comparatively less with 78.6%. While the linear kernel gives 88% along with the polynomial kernel having 89% whereas out of all 4 kernels the radial basis function kernel gives the highest result with 90.54%. The below-given table below represents the output accuracy of different machine learning methods

| Models | Accuracy |
|---|---|
| Artificial Neural Network | 93.03 % |
| Logistic Regression | 87.06 % |
| Decision Tree Classifier | 96.76 % |
| Support Vector Machine (Kernel = Sigmoid) | 78.60 % |
| Support Vector Machine (Kernel = Linear) | 88.80 % |
| Support Vector Machine (Kernel = Polynomial) | 89.80 % |
| Support Vector Machine (Kernel = Radial Basis Function) | 90.54 % |

TABLE I. DIFFERENT ML MODELS RESULTS

After comparing different bars which have formed when the processed data were applied on the proposed models and their accuracy was taken out we can say that the Decision tree being the highest of all gives the best accuracy with 96.76% closely followed by Artificial Neural Network with 93.03% backing up by Support Vector Machine with Radial Basis Function kernel having 90.54%. Out of all the bars, it is visible that the SVM model with Sigmoid kernel gives the least accuracy with 78.6%.

### ATTRIBUTES CO-RELATION:

The variable selecting process is worked for the choosing of necessary variables in the prediction of the costs of health insurance. For simplification of the process of prediction, it is necessary to extract the most important variables. We are using age, gender, BMI, steps, children, smoker, region, and charges as the independent variable while an insurance claim is a dependent variable. Based on this accuracy we can conclude that our predictive results are acceptable.The Heat Map highlights that as per our data set the feature named steps (i.e. no of steps the person walked daily) and children (i.e. no of kids the person is having in his family) are highly correlated with the target variable named as the insurance claim.

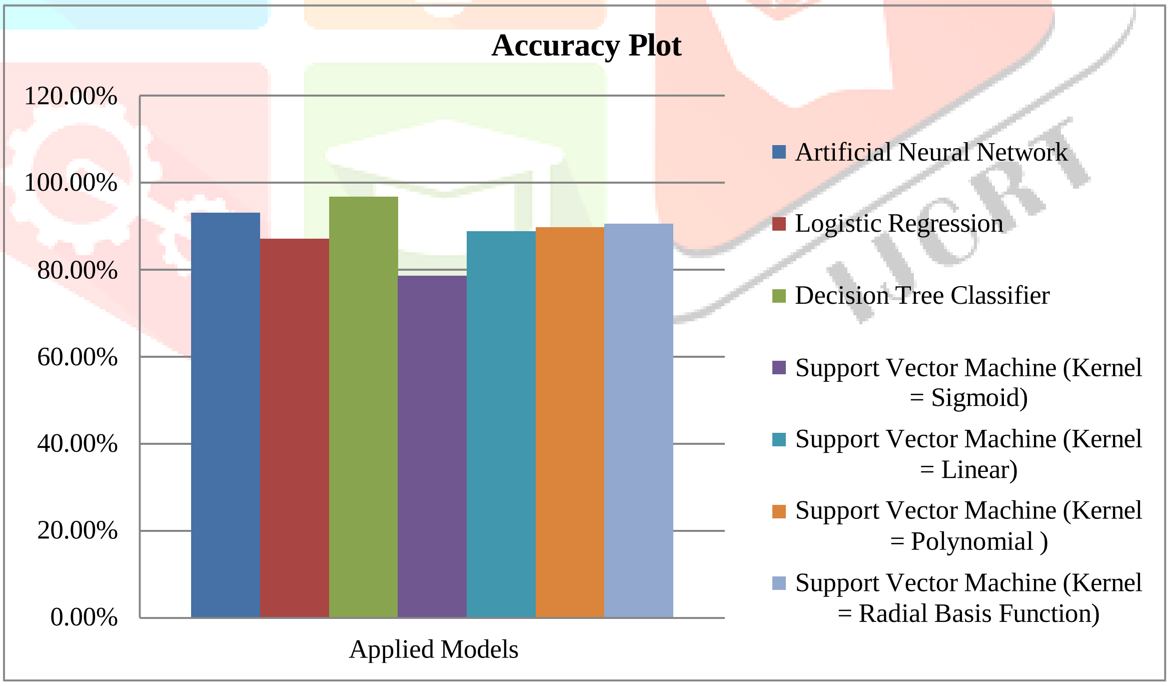

Figure 5: Result Analysis of different ML Models

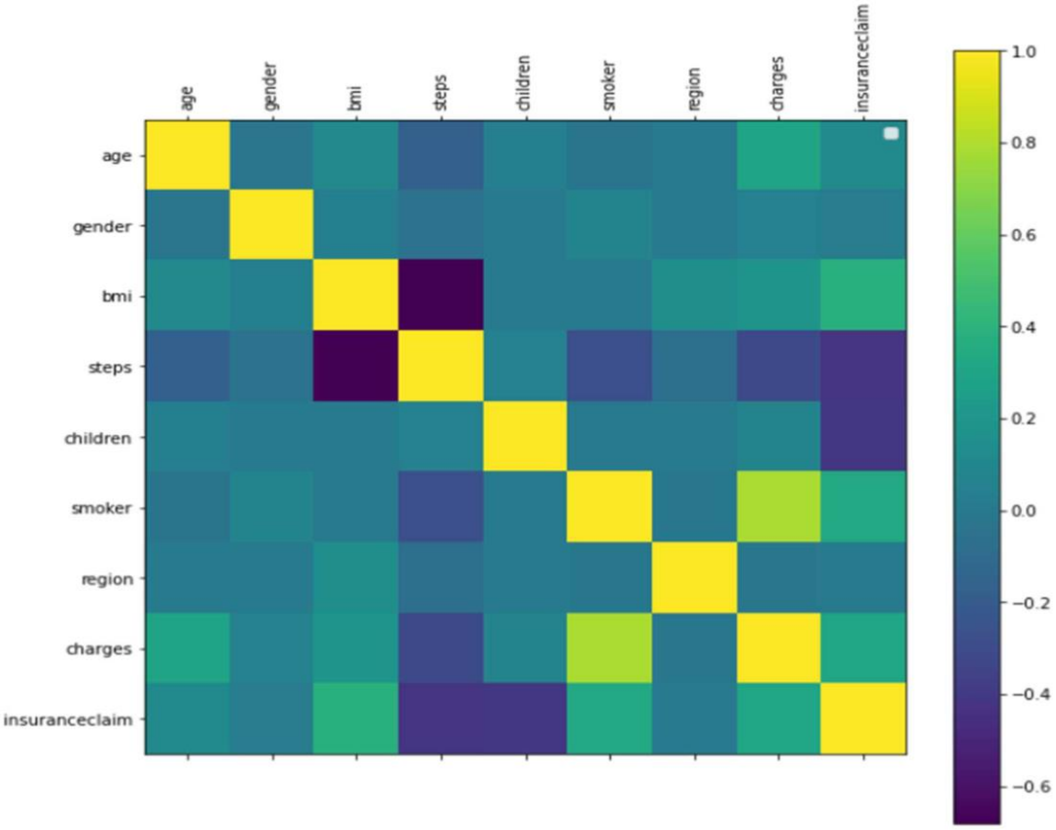

Figure 6: Different Attribute Co-Relation Process

## CONCLUSION:

Medical care framework improvement is a vital errand for any nation to guarantee supportable medical services for each resident. However, the advancement is affected by various bearings, which ought to be followed. Medical coverage costs are quite possibly the main direction for the improvement of the entire medical care framework. Examining and foreseeing the medical care protection costs is an extremely difficult assignment as a result of the enormous number of information where customary relapse techniques couldn't be helpful. To work on examining the medical services protection price in this review, the chosen strategy was performed with the principal objective of separating the main factors for the expectation of the medical services protection costs. The obtained results actually indicate that the decision tree classifier achieved the highest accuracy. The Artificial Neural Network model's accuracy is still satisfactory, but there is potential for improvement by fine-tuning the parameters in future iterations to further increase its accuracy.

## REFERENCES:

[1] Mimra, Wanda, Janina Nemitz, and Christian Waibel. "Voluntary pooling of genetic risk: A health insurance experiment." Journal of Economic Behavior & Organization 180 (2020): 864-882.

[2] Lin, Haizhen, and Daniel W. Sacks. "Intertemporal substitution in health care demand: Evidence from the RAND Health Insurance Experiment." Journal of Public Economics 175 (2019): 29-43.

[3] Chen, Yi, Julie Shi, and Castiel Chen Zhuang. "Income-dependent impacts of health insurance on medical expenditures: Theory and evidence from China." China Economic Review 53 (2019): 290-310.

[4] Cardon, James H. "Loss aversion and health insurance plan switching." Journal of Economic Behavior & Organization 180 (2020): 955-966.

[5] Pannequin, François, Anne Corcos, and Claude Montmarquette. "Are insurance and self-insurance substitutes? An experimental approach." Journal of Economic Behavior & Organization (2019).

[6] Layton, Timothy J. "Imperfect risk adjustment, risk preferences, and sorting in competitive health insurance markets." Journal of health economics 56 (2017): 259-280.

[7] Kelly, Mark. "Health capital accumulation, health insurance, and aggregate outcomes: A neoclassical approach." Journal of Macroeconomics 52 (2017): 1-22.

[8] Pelgrin, Florian, and Pascal St-Amour. "Life cycle responses to health insurance status." Journal of health economics 49 (2016): 76-96.

[9] Stavrunova, Olena, and Oleg Yerokhin. "Tax incentives and the demand for private health insurance." Journal of health economics 34 (2014): 121-130.

[10] Jang, J-SR. "ANFIS: adaptive-network-based fuzzy inference system." IEEE transactions on systems,

man, and cybernetics 23.3 (1993): 665-685.

[11] Gavrilović, Snežana, et al. "Statistical evaluation of mathematics lecture performances by soft computing approach." Computer Applications in Engineering Education 26.4 (2018): 902-905.

[12] Nikolić, Vlastimir, et al. "Selection of the most influential factors on the water-jet assisted underwater laser process by adaptive neuro- fuzzy technique." Infrared Physics & Technology 77 (2016): 45-50.

[13] Petković, Dalibor. "Prediction of laser welding quality by computational intelligence approaches." Optik 140 (2017): 597-600.

[14] Nikolić, Vlastimir, et al. "Wind speed parameters sensitivity analysis based on fractals and neuro-fuzzy selection technique." Knowledge and Information Systems 52.1 (2017): 255-265.

[15] Petković, Dalibor, Nenad T. Pavlović, and Žarko Ćojbašić. "Wind farm efficiency by adaptive neuro-fuzzy strategy." International Journal of Electrical Power & Energy Systems 81 (2016): 215- 221.

[16] Petković, Biljana, et al. "Neuro-fuzzy estimation of reference crop evapotranspiration by neuro fuzzy logic based on weather conditions." Computers and Electronics in Agriculture 173 (2020): 105358.

[17] Mohapatra, S., Satpathy, S., & Paikaray, B. K. (2023). A machine learning approach to assist prediction of Alzheimer's disease with convolutional neural network. International Journal of Bioinformatics Research and Applications, 19(2), 141-150.

[18] Paikaray, B. K., Pramanik, J., & Samal, A. K. (2022, October). An Introductory Approach to Spectral Image Analysis Using Machine Learning Classifiers. In 2022 1st IEEE International Conference on Industrial Electronics: Developments & Applications (ICIDeA) (pp. 198-201). IEEE.

[19] Mohammed Siddique, Tumbanath Samantara, Siba Prasad Mishra,(2021) A hybrid prediction model of kernel principal component analysis, support vector regression and teaching learning based optimization techniques, Current Journal of Applied Science and Technology,Vol:40,Issue:20 pp -17-25

[20] Paikaray, Dr. B. K. (2018, February 21). Medical Cost Personal Datasets. Kaggle. https://www.kaggle.com/mirichoi0218/insurance