IARJSET

ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Vol. 8, Issue 8, August 2021

DOI: 10.17148/IARJSET.2021.8817

Selection of Variables in Logistic Regression Model with Genetic Algorithm for Stroke Prediction

Avijit Kumar Chaudhuri[1], Arkadip Ray[2], Prof. Dilip K. Banerjee[3], Dr. Anirban Das[4]

[1]Research Scholar, Department of Computer Application, SEACOM SKILLS UNIVERSITY, Kendradangal, Bolpur, Birbhum, 731236, West Bengal, India

[2]Department of Information Technology, Government College of Engineering and Ceramic Technology, Kolkata, West Bengal, 700010, India

[3]Professor, Department of Computer Application, SEACOM SKILLS UNIVERSITY, Kendradangal, Bolpur, Birbhum, 731236, West Bengal, India

[4]University of Engineering & Management, Kolkata, West Bengal, India

ABSTRACT: The most important issue for avoiding and preventing the progression of various diseases is earlier risk assessment and identification. To estimate disease risk factors, the researchers typically used the statistical comparative analysis or step-by-step methods of feature selection using regression techniques. The results of these methods focused on individual risk factors separately. However, rather than just one factor, a combination of factors is more likely to influence disease development. Genetic algorithms (GA) can be beneficial and efficient for finding a combination of factors for the fastest diagnosis with the highest accuracies, especially when dealing with a large number of complicated and poorly understood components, as in diseases prediction. Our proposed model demonstrates the potential for using GA to diagnose disease and predict accuracy. Our proposed ensemble model revealed that combining a limited selection of input features gives better results than using all of the single significant features individually. This model not only forecasts the optimal feature sets and accuracy but also overcomes the dataset's missing values problem. Variables more commonly picked by LR may be more relevant for disease development prediction and accuracy by GA.

Keywords: Data Mining, Logistic Regression (LR), Genetic Algorithm (GA), Feature Selection (FS), Decision Tree (DT), Random Forest (RF)

1. INTRODUCTION

Developing nations, such as India, have a double burden of communicable and non-communicable illnesses. Stroke is one of the most common causes of death and disability in India. The estimated adjusted stroke prevalence rate ranges from 84-262/100,000 in rural regions to 334-424/100,000 in urban areas [1]. Stroke risk prediction can contribute significantly to forbade it and begin early treatment. Multiple medical studies and data analyses have been conducted to understand effective prophesiers of stroke. The Framingham Study [2,3] reported multiple factors which increase the risk of stroke which include age, systolic blood pressure, use of antihypertensive therapy, diabetes mellitus, smoking cigarette, prior cardiovascular disease, atrial fibrillation, and left ventricular hypertrophy by electrocardiogram. Furthermore, other numerous studies in the past decade [4,5,6,7] have made discoveries of factors that are riskier like creatinine level, the time required to walk 15 feet, and more. Most preceding prophecy models have adopted features of risk factors that are confirmed by clinical trials or selected manually by medical specialists. For example, Lumley et al. [6] built a 5-year stroke prophecy model based on Cardio-vascular Health Study [8] dataset using a set of 16 manually chosen features (given in [9]) from a total of approximately one thousand features. With numerous features in recent medical datasets, it is an unmanageable task to identify and confirm each risk factor manually. On the other hand, machine learning algorithms can identify features highly related to stroke occurrence dexterously from the huge

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Vol. 8, Issue 8, August 2021

DOI: 10.17148/IARJSET.2021.8817

collection of features; therefore, we believe machine learning can be used to (i) improve the prophecy precision of stroke risk, and (ii) discover new risk factors.

Lumley et al.'s [6] 5-year stroke prediction model adopted the Cox proportional hazards model, one of the most commonly used statistical methods in medical research [9]. It has been considerably studied [9,10] and applied to the prophecy of various diseases including stroke [6,11,12]. However, the performance of the actual Cox model depends mostly on the quality of the pre-selected features. To address this problem, several applications have been suggested recently [13,14]. Thus far, there have been very few studies on differentiating the Cox regression with machine learning methods in making prophecy on censored data. Kattan [15] compared Cox proportional hazards regression with several machine learning methods (neural networks and tree-based methods) based on three urological datasets. However, Kattan's study emphasized datasets with only five features, while machine learning algorithms are thought to effectively handle many more features. In inclusion, the paper considered only some relatively simple machine learning algorithms, and high-performance machine learning algorithms such as SVM and NB were not inspected.

This paper shows a combined machine learning approach for stroke risk prophesy. We explored machine learning algorithms to improve the prophecy correctness and conducted considerable differences between our results and those with the Cox proportional hazards model. Our approach examines the problems of data imputation, feature selection, and prophecy in medical datasets. We suggest a novel automatic feature selection algorithm that selects robust features on our suggested heuristic: conservative mean. In inclusion, our work has also identified potential risk factors that have not been explored by traditional applications. Last, we note that this method can be applied to the clinical prophecy of other diseases, where missing data are usual and risk factors are not well recognized.

2. RELEVANT LITERATURE

There can be single or multiple subgroups of characteristics that are equally useful for a particular objective, for multiple redundant or outdated characteristics. A single subset of features does not always result in FS. FS removes unnecessary characteristics to get the optimum subset of features for discriminating across classes.

Clinical and diagnostic evidence has led to the development of several effective early detection services and other health-related technologies in both the data mining and healthcare sectors. Artificial intelligence (AI) is increasingly accepted in research and health care. Classification, often known as predictive analytics, is an important component of AI in machine learning (ML). The current evaluations of novel prediction models based on ML techniques show promise in the field of scientific research [16]. Chaudhuri et al. estimate the diseases using the recursive feature elimination (RFE) technique, which picks an optimal selection of characteristics, and an ensemble algorithm, the enhanced decision tree (EDT). The results obtained in their study demonstrate that the accuracy level of EDT is not affected by the elimination of less relevant characteristics, allowing decision-makers to focus on a few features to decrease treatment time and error. EDT achieves a good level of consistency in forecasting the illness, with or without feature selection [17]. Chaudhuri et al. compared proven approaches and proposed a framework for integrating findings from various DMT to avoid Type 2 and Type 1 errors. To predict the disease, two sets of data were used: disease and treatment datasets, as well as features identified as significant by the ensemble method – the random forest. The results show that traditional methods, such as LR, outperformed RF in terms of significant features. This approach, however, fails when the data dichotomy (i.e., disease or no disease) is not distinct. The DT analysis was performed consistently across all variants of the dataset used in this paper [18]. Nahar et al. [20] conducted a study in which they tested several classifiers for extracting heart disease (HD). When absolute accuracy is employed as the output metric, they proved that SVM has promise. The experiment findings indicated that applying the medical knowledge-driven feature selection (MFS) approach considerably improved the output for most classifiers for most datasets, especially in terms of accuracy.

Chen et al. [21] presented a unique Convolutional Neural Network-based multimodal disease risk prediction approach in hospitals that employs both structured and unstructured data. Three circumstances were anticipated for diseases such as diabetes, cerebral infarction, and heart disease. Machine learning algorithms such as NB, decision trees, and k-nearest neighbours (KNN) are used to predict heart disease, diabetes, and brain abnormalities. The findings of

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Vol. 8, Issue 8, August 2021

DOI: 10.17148/IARJSET.2021.8817

algorithm DT outperform those of algorithms NB and KNN. Alizadehsani et al. [22] gave a systematic and multifaceted review of all relevant studies for ML-based Coronary artery disease diagnosis published between 1992 and 2019. The impacts of several aspects such as dataset characteristics (geographic location, sample size, features, and severity of each coronary artery) and applied ML methods (feature selection, performance metrics, and process) are fully investigated. Based on a huge dataset of structured and unstructured data collected from hospitals, the researcher [23] created a novel machine learning algorithm CNN-MDRP to anticipate the likelihood of various illnesses. To do the prediction analysis on the structured datasets, he used an existing NB-based ML algorithm CNNUDRP. According to the findings of this study, CNN-MDRP is far more powerful than the next algorithm and can be applied to both structured and unstructured datasets. In their study of an HD dataset from Andhra Pradesh, the authors [24] utilized the Lazy Association classification approach to calculate accuracy. They discovered a 10.26 % gain in accuracy over the J4.8 and an 8.6 percent improvement over the NB ML algorithms.

In research, Amin et al. [25] proposed a new approach for forecasting heart illness based on hybrid neural networks and distinct characteristics of GA. Taking into account the risk variables such as age, family history of HD, blood sugar, smoking, obesity, and many more. On the training set, the suggested classifier achieved 96 percent accuracy and 89 percent accuracy on the testing set. The authors of [26] proposed a computational intelligence strategy to predict HD by employing a variety of FS approaches to improve the accuracy provided by the NB classifier. Using NB as the classification technique and one-R as the FS algorithm, they improved accuracy to 86.29 %. The Fuzzy K-NN test was proposed by Krishnaiah et al. [27] as part of a hybrid approach to HD prediction. This method removed uncertainty from the traditional K-NN classifier and calculated appropriate values for building a new model. The proposed system revealed that people in their 50s and 60s had similar symptoms of HD to those in their 40s and 45s.

The authors analyzed the HD data set using AdaBoost for the feature subset and the PCA technique for feature selection [28]. Using this ensemble technique, prediction accuracies improved by 2.11% over J4.8 and 7.33% over 10-fold cross validations. Masethe et al. [29] tested and analyzed several precise HD machine learning techniques. The analysis indicated that NB produced 99% accuracy compared to other classifiers. For HD prediction, Karoalis et al. [30] utilized the DT algorithm and C4.5 and tested both for PCI and CABG models. They achieved maximum precision of 75%. Samples of 620 patients from Paphos district were examined by the authors. However, the accuracy achieved by this technique is poor compared with other algorithms. The researcher [31] implements the DTRS clustering technique to forecast fuzzy sets and heart disease. The author indicates that DTRFCS is more effective than DTRS algorithms in this investigation. To improve accuracy, the authors [32] devised a unique strategy for predicting HD using KNN and NB Classification algorithms. By evaluating accumulated HD records, they were able to obtain an accuracy of 82.6%. Because not all variables are useful predictors, the authors are attempting to determine the optimal subset. Genetic algorithms provide a workaround for this restriction because they are known to produce virtually optimal outcomes when searching for only a tiny portion of a search area.

3. METHODOLOGY

When creating a multivariable regression model, it is necessary to choose a parameter or function. The main objective of the selection of features is to include clinically important and statistically significant features in the model with noise/redundant features being exempted [1,2]. A range of approaches, such as targeted selection, the best subset, step-by-step regression, and association rules, are commonly employed in this regard. Neither technique is an ideal answer, especially for a wide variety of characteristics during the era of Big Data. For the proper selection of variables, a univariate screening to track variables is typically required. These variables are then used to create a regression template. While some important variables are working together, a system that does not have statistical significance when evaluated independently can ignore them. Because all possible candidate variables are tested, the best subset approach can solve this problem [3]. Furthermore, while this method may be useful in a data set with a small number of variables, it may not be the best option for a large dataset with a large number of features. Another popular method is the step-by-step approach, which is, at its core, a local research process [4]. GA [5,6] is a search heuristic that mimics biological and natural selection methods. According to Darwin's theory, the population changes as a result of selection, crossover, and mutation. The fittest individuals survive and reproduce, while the weakest are wiped out. GA generates artificial random populations (called chromosomes in the terminology of GA) that are assessed using a mathematical

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Vol. 8, Issue 8, August 2021

DOI: 10.17148/IARJSET.2021.8817

fitness method. Both discrete and continuous functions have been successfully used to solve optimization problems when picking, replicating, crossing, and mutating. The mathematical fitness function is best served by these artificially designed chromosomes.

This paper explains how to use GA to achieve the best classification precision for variables chosen by LR. The LR classification model is the most popular approach to clinical research because most dependent variables (e.g., diagnosis) are categorical. This framework can make use of some types of generalized linear models and data sets. Traditional feature selection (FS) techniques contribute a small amount to classification, but machine learning (ML) techniques and feature selection (FS) techniques together can help make better decisions and classify many diseases with high accuracy levels. Due to the exponential growth of the IT infrastructure in health services and the increased dissemination of medical databases, medical institutions produce and collect large volumes of health data. It is critical to investigate the appropriate feature, which is becoming available as a result of the rapidly expanding corpus of medical data, to improve patient care efficiency and, as a result, reduce treatment costs and time. LR classification is a popular, very common, and efficient technique in the healthcare sector. Several studies on treating different diseases involving LR models have already been published. Some of these models were planned to prospectively be used on previously unknown instances. Traditionally, this problem has been resolved by step-by-step techniques of variable selection for logistic regression models. This sequential method significantly restricts the number of models being examined. Another way would be to look at all existing models. The number of possible models is 2n, given n number of variables to choose from, making this exhaustive method for variables other than small numbers impracticable. Some periodic system provides this method e.g., SAS but for only 10 variables or less.

The authors in this article compare an LR dependent variable selection method for GA models with randomly generated values to standard linear variable selection approaches using patients (attacked in stroke) data set in the proposed approach. In this section, the authors go over the fundamental concepts of LR and show how it can be used to choose variables. The goal of this study was to find the best-performing feature combinations for early diagnosis and development of stroke and other diseases. In this study, LR was used to select one or more sets of diagnostic test results (features) that can accurately predict disease progression, and GA was used to build prediction models. The algorithm architecture and mechanism that combined LR and GA to predict disease status are shown in Fig. 1 and 2. As the GA input, the user-selected features for the LR were used. To simplify and identify the best set of features for the various sets, the GA used LR tests.

3.1 Architecture of The System

In the system design, our suggested classifier has been coupled with GA. There were three major phases in the search process. LR is used separately to construct a predictive model for each instance within potential GA solutions, together with the set of selected characteristics from RF (RF-with or without contradicting features), DT, and GA. All anticipated variables from the LR's future outputs will be created in a new data set. Variables with values between the minimum and maximum of these variables will be created at random. At the last step of the search, GA would discover an improved solution that would replace previously discovered less relevant solutions. GA follows the iterative learning theory, which was first introduced by Holland. This methodology operates on a principle close to that of natural-system genetic simulations (Fig. 2 represents the architecture of the GA technique). This algorithm was initially used to identify individuals with a permanent population using a space snapshot. The role of exercise is designed for individual evaluation. Any operations are carried out to develop new generations [19]. Genes consist of chromosomes. A clinical variable is considered as a gene (Table 1) and a set of clinical variables can be considered as chromosomes to construct a regression model. A chromosome size determines the number of variables on a model. For example, if each model has 6 different variables the chromosome size will be 6. The GA begins by compiling a random selection of models. A fitness function is used to evaluate each model in the population. When the objective is achieved, the model is picked and the process of evolution ends. If the objective is not met in the present generation, the model population will keep evolving. The present population acts as a parent, while the following generation is a descendant. Models with increased fitness in parents have a better probability to reproduce, as is the case with biological evolution. The reproduced chromosome is being crossed and mutated to generate variety, which gives the offspring greater opportunity to achieve the fitness objective. The cycle goes on to the fitness goal or the maximum number of generations (Fig. 2).

IARJSET

ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Vol. 8, Issue 8, August 2021

DOI: 10.17148/IARJSET.2021.8817

3.2 Stroke Dataset and Attribute Description

The authors gather information from a population of 5110 people are involved in this study with 2995 females and 2115 males. The dataset for this study is extracted from Kaggle data repositories (https://www.kaggle.com/datasets) to predict whether a patient is likely to get stroke based on the following attribute information.

4. RESULTS AND DISCUSSION

In the space of all possible subsets of predictor variables, a GA is defined, which searches for the best LR model with only significant variables. The approach was demonstrated to be effective with a large data set containing thousands of records and a large number of variables using the stroke data set with different independent variables. While statistical techniques are used in LR to select meaningful variables, GA statistical techniques cannot be used to select variables because it is a black-box model. Step-by-step LR will be used to select variables in the proposed method. Explanatory variables derived from LR and projections can also be used as GA inputs. As a result, the proposed method takes advantage of both the LR and GA. When the results in Table 2 were compared to other methods, it was found that the proposed method produced the best results in terms of criteria for various disease data. As shown in Table 2 we found from different analyses on stroke data-set that the classification accuracy of the LR is highest, so variables chosen for GA analysis are age, hypertension and avg_glucose_level. With randomly generated values and LR equation,

$$Y = -7.5788 + 0.0706 * \text{age} + 0.3845 * \text{hypertension} + 0.0044 * \text{glucose\_level}$$

the prediction accuracy for our proposed model is 99% which outperforms other popular ML classifiers. Our suggested model not only provides the best accuracies in various disease diagnostics, but it effectively handles the missing value problem in datasets. Noise, missing values, and inconsistency are common features of medical datasets found in various repositories. Researchers use various pre-processing steps to solve these problems, such as data cleaning, data integration, data transformation, data reduction, and so on. For all factors, the proposed classification model generates numbers at random between the minimum and maximum values. This helps to test any possible combination of values from various factors, and this technique also overcomes the problem of noise, missing value, and inconsistency.

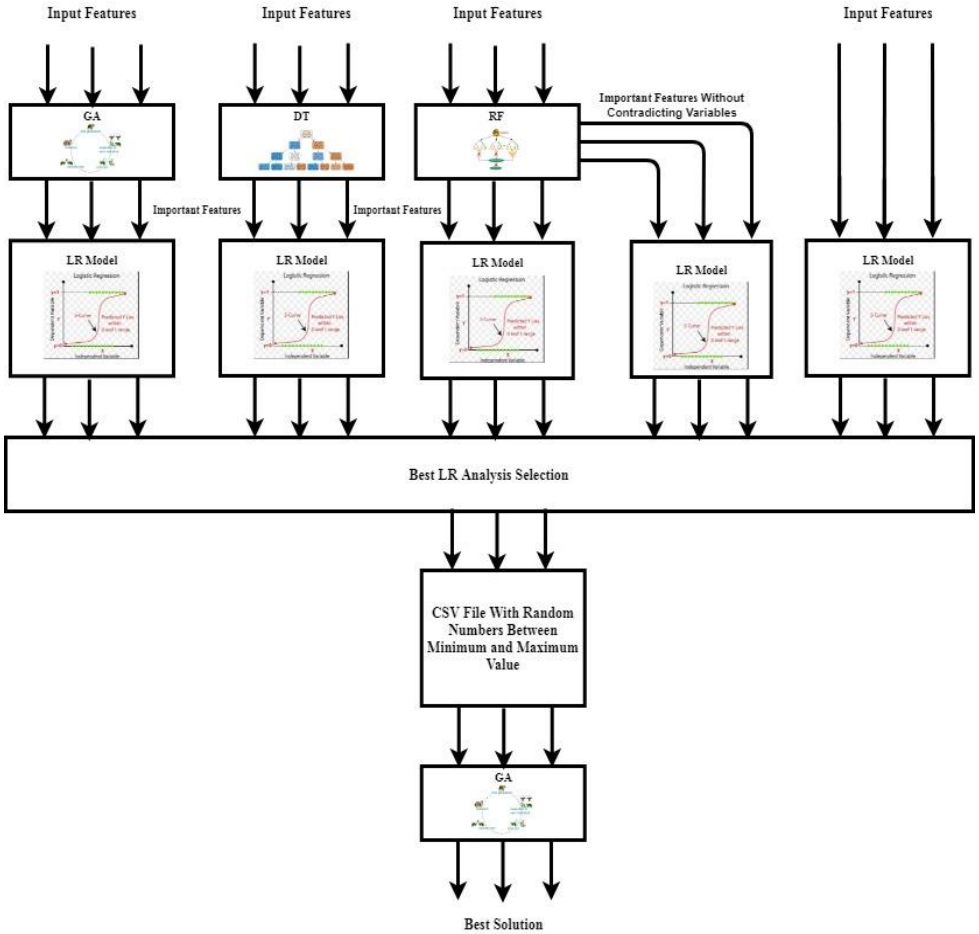

Fig. 1. Architecture of The System

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Vol. 8, Issue 8, August 2021

DOI: 10.17148/IARJSET.2021.8817

CONCLUSION

This study used GA to predict disease progression by combining the impacts of a wide set of independent variables from datasets (n number of potential combinations between the lowest and highest values) with the best LR formula used to compute the best accuracy. The classification result of our proposed classifier shows that using a combination of variables to predict disease progression is superior in quality to using a single significant variable or a finite set of variables. To improve the prediction models, the developed algorithm will be tested and modified with more data. The GA method proposed in this research was implemented as a general solution that may be used for a different disease or non-disease datasets. GA's technique is intended to discover solutions that need a successful search of a subset of characteristics to identify nearly optimal combinations to solve vast, difficult, or poorly understood solution spaces. Clinical diagnosis and prognostic outcomes can be considered by selecting the appropriate characteristics for categorization and reliability. Today, LR is a more commonly utilized classification approach in clinical science. This research shows how to use GA to estimate accuracies and choose features in combination with LR. Patient survey findings, observational retrospective research, and clinical trials can also enable us to evaluate the model's efficacy, quality, and cost. Thus, the authors conclude: (1) The selection of features is important for the choice of whether the features to be retained or discarded. (2) Various machine learning techniques (MLTs), given certain requirements, such as dataset size, feature count, and records per category, can predict with better accuracy. As a result, the authors suggest a combination of feature selection and machine learning approaches for disease prediction that reduces mistakes. But trying to optimize the accuracy of the prediction by addressing the specificity and sensitivity limitation through an FS and MLT simulation might get better results. The authors propose such an attempt as future work.

Table. 1. Description of Stroke Dataset

	Attributes	Description	Values
1	id	unique identifier	Continuous
2	gender	"Male", "Female" or "Other"	
3	age	age of the patient	Continuous
4	hypertension	0 if the patient doesn't have hypertension, 1 if the patient has hypertension	
5	heart_disease	0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease	
6	ever_married	"No" or "Yes"	
7	work_type	"children", "Govt_jov", "Never_worked", "Private" or "Self-employed"	
8	Residence_type	"Rural" or "Urban"	
9	avg_glucose_level	average glucose level in blood	Continuous
10	bmi	body mass index	
11	smoking_status	"formerly smoked", "never smoked", "smokes" or "Unknown"	
12	stroke	1 if the patient had a stroke, 0 the patient do not have a stroke	

Table. 2. Result of Analysis on Stroke Dataset

Method	Accuracy	Selected Features
LR	0.9510	age, hypertension, avg_glucose_level
LR after DT	0.9491	age, avg_glucose_level
LR after RF	0.9491	age, avg_glucose_level
LR after GA	0.9491	Hypertension, heart_disease, avg_glucose_level, smoking_status
Our Proposed Model	**0.9937**	**age, hypertension**

IARJSET

ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Vol. 8, Issue 8, August 2021

DOI: 10.17148/IARJSET.2021.8817

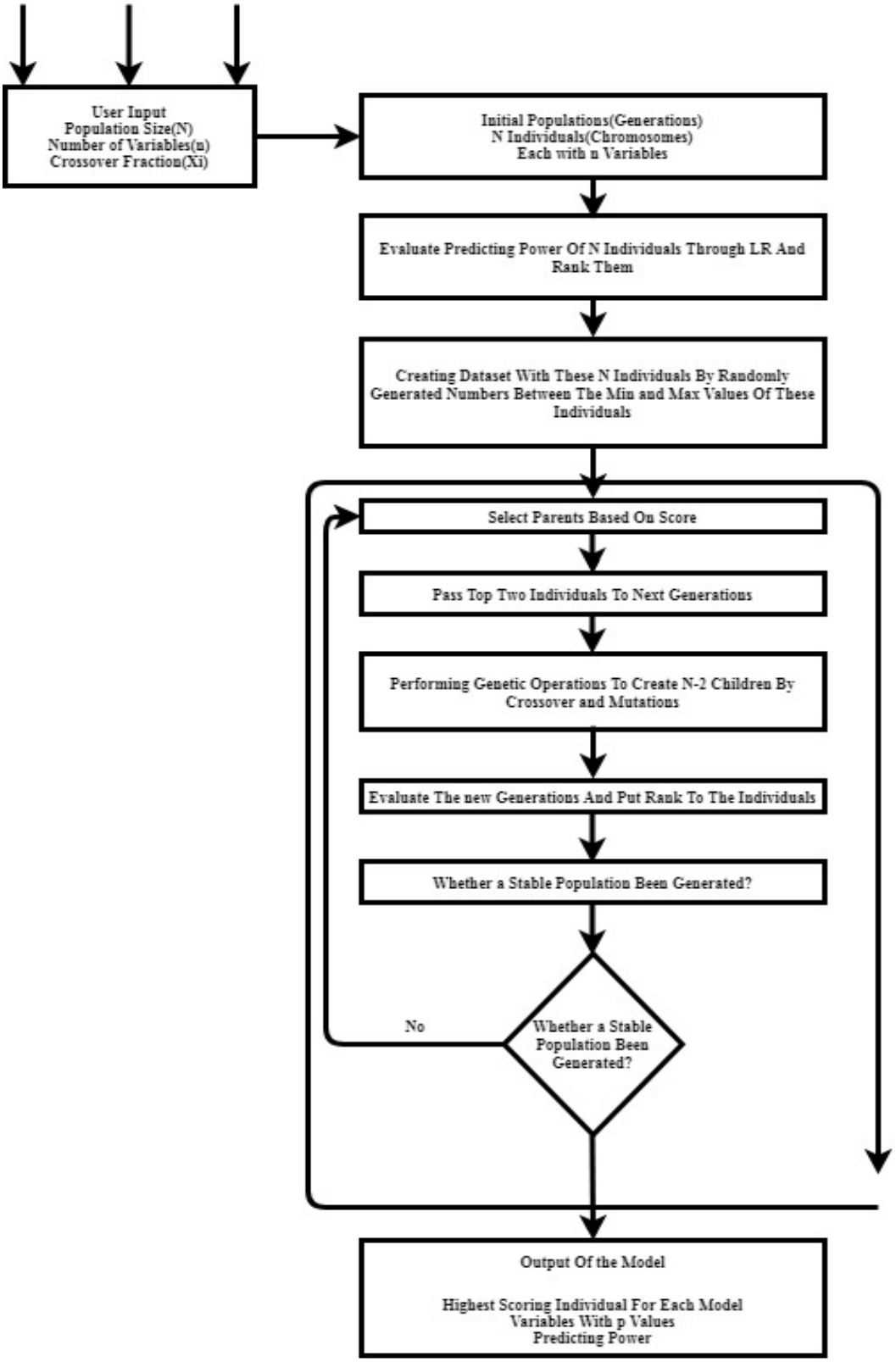

Fig. 2. Genetic Algorithm for Disease Outcome Modelling

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Vol. 8, Issue 8, August 2021

DOI: 10.17148/IARJSET.2021.8817

REFERENCES

1. Pandian, J. D., & Sudhan, P. (2013). Stroke epidemiology and stroke care services in India. Journal of stroke, 15(3), 128.
2. Dawber, T. R., Meadors, G. F., & Moore Jr, F. E. (1951). Epidemiological approaches to heart disease: the Framingham Study. American Journal of Public Health and the Nations Health, 41(3), 279-286.
3. Wolf, P. A., D'Agostino, R. B., Belanger, A. J., & Kannel, W. B. (1991). Probability of stroke: a risk profile from the Framingham Study. Stroke, 22(3), 312-318.
4. Manolio, T. A., Kronmal, R. A., Burke, G. L., O'Leary, D. H., & Price, T. R. (1996). Short-term predictors of incident stroke in older adults: the Cardiovascular Health Study. Stroke, 27(9), 1479-1486.
5. Longstreth, W. T., Bernick, C., Fitzpatrick, A., Cushman, M., Knepper, L., Lima, J., & Furberg, C. D. (2001). Frequency and predictors of stroke death in 5,888 participants in the Cardiovascular Health Study. Neurology, 56(3), 368-375.
6. Lumley, T., Kronmal, R. A., Cushman, M., Manolio, T. A., & Goldstein, S. (2002). A stroke prediction score in the elderly: validation and Web-based application. Journal of clinical epidemiology, 55(2), 129-136.
7. McGinn, A. P., Kaplan, R. C., Verghese, J., Rosenbaum, D. M., Psaty, B. M., Baird, A. E., ... & Wassertheil-Smoller, S. (2008). Walking speed and risk of incident ischemic stroke among postmenopausal women. Stroke, 39(4), 1233-1239.
8. Fried, L. P., Borhani, N. O., Enright, P., Furberg, C. D., Gardin, J. M., Kronmal, R. A., ... & MPH for the Cardiovascular Health Study Research Group. (1991). The cardiovascular health study: design and rationale. Annals of epidemiology, 1(3), 263-276.
9. Bender, R., Augustin, T., & Blettner, M. (2005). Generating survival times to simulate Cox proportional hazards models. Statistics in medicine, 24(11), 1713-1723.
10. Akazawa, K., Nakamura, T., Moriguchi, S., Shimada, M., & Nose, Y. (1991). Simulation program for estimating statistical power of Cox's proportional hazards model assuming no specific distribution for the survival time. Computer methods and programs in biomedicine, 35(3), 203-212.
11. Ikeda, K., Kumada, H., Saitoh, S., Arase, Y., & Chayama, K. (1991). Effect of repeated transcatheter arterial embolization on the survival time in patients with hepatocellular carcinoma. An analysis by the Cox proportional hazard model. Cancer, 68(10), 2150-2154.
12. Liang, K. Y., Self, S. G., & Liu, X. (1990). The Cox proportional hazards model with change point: An epidemiologic application. Biometrics, 783-793.
13. Goeman, J. J. (2010). L1 penalized estimation in the Cox proportional hazards model. Biometrical journal, 52(1), 70-84.
14. Park, M. Y., & Hastie, T. (2007). L1-regularization path algorithm for generalized linear models. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 69(4), 659-677.
15. Kattan, M. W. (2003). Comparison of Cox regression with other methods for determining prediction models and nomograms. The Journal of urology, 170(6S), S6-S10.
16. Ray, A., & Chaudhuri, A. K. (2021). Smart healthcare disease diagnosis and patient management: Innovation, improvement and skill development. Machine Learning with Applications, 3, 100011.
17. Chaudhuri, A. K., Sinha, D., Banerjee, D. K., & Das, A. (2021). A novel enhanced decision tree model for detecting chronic kidney disease. Network Modeling Analysis in Health Informatics and Bioinformatics, 10(1), 1-22.
18. Chaudhuri, A. K., Sinha, D., Bhattacharya, K., & Das, A. An Integrated Strategy for Data Mining Based on Identifying Important and Contradicting Variables for Breast Cancer Recurrence Research.
19. Zhang, Z., Trevino, V., Hoseini, S. S., Belciug, S., Boopathi, A. M., Zhang, P., ... & Dai, S. (2018). Variable selection in Logistic regression model with genetic algorithm. Annals of translational medicine, 6(3).
20. Nahar, J., Imam, T., Tickle, K. S., & Chen, Y. P. P. (2013). Computational intelligence for heart disease diagnosis: A medical knowledge driven approach. Expert Systems with Applications, 40(1), 96-104.
21. Chen, M., Hao, Y., Hwang, K., Wang, L., & Wang, L. (2017). Disease prediction by machine learning over big data from healthcare communities. Ieee Access, 5, 8869-8879.
22. Alizadehsani, R., Abdar, M., Roshanzamir, M., Khosravi, A., Kebria, P. M., Khozeimeh, F., ... & Acharya, U. R. (2019). Machine learning-based coronary artery disease diagnosis: A comprehensive review. Computers in biology and medicine, 111, 103346.
23. Shirsath, S. S., & Patil, S. (2018). Disease prediction using machine learning over big data. International Journal of Innovative Research in Science, Engineering and Technology, 7(6), 6752-6757.
24. Jabbar, M. A., Deekshatulu, B. L., & Chandra, P. (2013, March). Heart disease prediction using lazy associative classification. In 2013 International Mutli-Conference on Automation, Computing, Communication, Control and Compressed Sensing (iMac4s) (pp. 40-46). IEEE.
25. Amin, S. U., Agarwal, K., & Beg, R. (2013, April). Genetic neural network based data mining in prediction of heart disease using risk factors. In 2013 IEEE Conference on Information & Communication Technologies (pp. 1227-1231). IEEE.
26. Jabbar, M. A., Deekshatulu, B. L., & Chandra, P. (2015, March). Computational intelligence technique for early diagnosis of heart disease. In 2015 IEEE International Conference on Engineering and Technology (ICETECH) (pp. 1-6). IEEE.
27. Krishnaiah, V., Srinivas, M., Narsimha, G., & Chandra, N. S. Diagnosis of heart disease patients using fuzzy classification technique. In international conference on computer and communications technologies (iccct) 2014 (pp. 1-7).
28. Jabbar, M. A., Deekshatulu, B. L., & Chndra, P. (2014, November). Alternating decision trees for early diagnosis of heart disease. In International Conference on Circuits, Communication, Control and Computing (pp. 322-328). IEEE.
29. Masethe, H. D., & Masethe, M. A. (2014, October). Prediction of heart disease using classification algorithms. In Proceedings of the world Congress on Engineering and computer Science (Vol. 2, No. 1, pp. 25-29).
30. Karaolis, M., Moutiris, J. A., & Pattichis, C. S. (2008, October). Assessment of the risk of coronary heart event based on data mining. In 2008 8th IEEE International Conference on BioInformatics and BioEngineering (pp. 1-5). IEEE.
31. Agrawal, S., & Tripathy, B. K. (2015, November). A decision theoretic rough fuzzy c-means algorithm. In 2015 IEEE International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN) (pp. 192-196). IEEE.
32. Ferdousy, E. Z., Islam, M. M., & Matin, M. A. (2013). Combination of naive bayes classifier and K-Nearest Neighbor (cNK) in the classification based predictive models. Computer and information science, 6(3), 48

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Vol. 8, Issue 8, August 2021

DOI: 10.17148/IARJSET.2021.8817

33.Kowshik B; Savitha V; Nimosh madhav M; Karpagam G; Sangeetha K. "Plant Disease Detection Using Deep Learning". International Research Journal on Advanced Science Hub, 3, Special Issue ICARD-2021 3S, 2021, 30-33. doi: 10.47392/irjash.2021.057

34.Salini Suresh; Suneetha V; Niharika Sinha; Sabyasachi Prusty; Sriranga H.A. "Machine Learning: An Intuitive Approach In Healthcare". International Research Journal on Advanced Science Hub, 2, 7, 2020, 67-74. doi: 10.47392/irjash.2020.67

35.Kowshik B; Savitha V; Nimosh madhav M; Karpagam G; Sangeetha K. "Plant Disease Detection Using Deep Learning". International Research Journal on Advanced Science Hub, 3, Special Issue ICARD-2021 3S, 2021, 30-33. doi: 10.47392/irjash.2021.057

36..Maneesha M; Savitha V; Jeevika S; Nithiskumar G; Sangeetha K. "Deep Learning Approach For Intelligent Intrusion Detection System". International Research Journal on Advanced Science Hub, 3, Special Issue ICARD-2021 3S, 2021, 45-48. doi: 10.47392/irjash.2021.061

37.Sona Solanki; Asha D Solanki. "Review of Deployment of Machine Learning in Blockchain Methodology". International Research Journal on Advanced Science Hub, 2, 9, 2020, 14-20. doi: 10.47392/irjash.2020.141

38.Mohd. Akbar; Prasadu Peddi; Balachandrudu K E. "Inauguration in Development for Data Deduplication Under Neural Network Circumstances". International Research Journal *on Advanced Science Hub*, 2, 6, 2020, 154-156. doi: 10.47392/irjash.2020.55

