IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, December 2022
Impact Factor: 6.252

Identification of Significant Features and Data Mining Techniques in Predicting Heart Stroke

Vikram Gude[1], Saravanan V[2], Iswarya RJ[3], Sathya M[4]
Assistant Professor, Department of Computer Science and Engineering[1,2,3,4.]
Malla Reddy Institute of Technology and Science, Hyderabad, India[1]
Mount Zion College of Engineering and Technology, Lena Vilakku, Pudukkottai, Tamil Nadu, India[2,3,4]

Abstract: *The health-care industry generates a large amount of data, which is processed using certain methodologies. One technique that is frequently utilized is data mining. Heart disease is the leading cause of death on a global scale. This system foresees the risk of heart disease developing. The results of this system give you a % likelihood of getting heart disease. Medical parameters are utilized to categories the datasets. This system uses a data mining classification algorithm to analyze such parameters. The datasets are processed in Python programming using two main Machine Learning Algorithms: Decision Tree Approach and Naïve Bayes Algorithm, with the latter showing to be the best algorithm in terms of heart disease accuracy.*

Keywords: Market Equilibrium, Fisher Market, Algorithmic game theory, Edge computing, Fog computing

I. INTRODUCTION

Data mining is the computing process of discovering patterns in large datasets involving methods at the intersection of machine learning, statistics and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Data mining is the analysis step of the "knowledge discovery in databases" process, or KDD.Data mining is about finding new information in a lot of data. The information obtained from data mining is hopefully both new and useful.

Data mining is the process of extracting previously unknown information from large databases or data warehouses and using it to make crucial business decisions. Data mining tools find patterns in the data and infer rules from them. The extracted information can be used to form a prediction or classification model, identify relations between database records, or provide a summary of the database(s) being mined. Those patterns and rules can be used to guide decision-making and forecast the effect of those decisions, and data mining can speed analysis by focusing attention on the most important variables.

Data mining is taking off for several reasons: organizations are gathering more data about their businesses, the enormous drop in storage costs, competitive business pressures, a desire to leverage existing information technology investments, and the dramatic drop in the cost/performance ratio of computer systems. Another reason is the rise of data warehousing. In the past, it was often necessary to gather the data, cleanse it, and merge it. Now, in many cases, the data is already sitting in a data warehouse ready to be used.

There are four basic mining operations supported by numerous mining techniques: predictive model creation supported by supervised induction techniques; link analysis supported by association discovery and sequence discovery techniques; database segmentation supported by clustering techniques; and deviation detection supported by statistical techniques

In sequences, events are linked over time. Classification is probably the most common data mining activity today. It recognizes patterns that describe the group to which an item belongs. It does this by examining existing items that already have been classified and inferring a set of rules from them. Clustering is related to classification, but differs in that no groups have yet been defined. Using clustering, the data mining tool discovers different groupings within the

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, December 2022
Impact Factor: 6.252

data. All of these applications may involve predictions. The fifth application type, forecasting, is a different form of prediction. It estimates the future value of continuous variables based on patterns within the data. A number of tools are used in data mining. These include, but are not limited to, neural networks, decision trees, rule induction, factor analysis, genetic algorithms, and data visualization

Using the described data mining tools, an organization can access and analyse the 10 percent of its information that is structured. To access the rest, a different technique is required – document mining

The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use.

II. RELATED WORKS

Bo Jin, Chao Che et al. (2018) proposed a "Predicting the Risk of Heart Failure with EHR Sequential Data Modeling" model designed by applying neural network. This paper used the electronic health record (EHR) data from real-world datasets related to congestive heart disease to perform the experiment and predict the heart disease before itself. We tend to used one-hot encryption and word vectors to model the diagnosing events and foretold coronary failure events victimization the essential principles of an extended memory network model. By analyzing the results, we tend to reveal the importance of respecting the sequential nature of clinical records [1].

Aakash Chauhan et al. (2018) presented "Heart Disease Prediction using Evolutionary Rule Learning". This study eliminates the manual task that additionally helps in extracting the information (data) directly from the electronic records. To generate strong association rules, we have applied frequent pattern growth association mining on patient's dataset. This will facilitate (help) in decreasing the number of services and shown that overwhelming majority of the rules helps within the best prediction of coronary sickness [2].

Ashir Javeed, Shijie Zhou et al. (2017) designed "An Intelligent Learning System based on Random Search Algorithm and Optimized Random Forest Model for Improved Heart Disease Detection". This paper uses random search algorithm (RSA) for factor selection and random forest model for diagnosing the cardiovascular disease. This model is principally optimized for using grid search algorithmic program.

Two forms of experiments are used for cardiovascular disease prediction. In the first form, only random forest model is developed and within the second experiment the proposed Random Search Algorithm based random forest model is developed. This methodology is efficient and less complex than conventional random forest model. Comparing to conventional random forest it produces 3.3% higher accuracy. The proposed learning system can help the physicians to improve the quality of heart failure detection [3].

"Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques" proposed by Senthilkumar Mohan, Chandrasegar Thirumalai et al. (2019) was efficient technique using hybrid machine learning methodology. The hybrid approach is combination of random forest and linear method. The dataset and subsets of attributes were collected for prediction. The subset of some attributes was chosen from the pre-processed knowledge(data) set of cardiovascular disease. After prep-processing, the hybrid techniques were applied and disgnosis the cardiovascular disease [4].

Prasanna Lakshmi, Dr. C.R.K.Reddy (2015) designed "Fast Rule-Based Heart Disease Prediction using Associative Classification Mining". In the proposed Stream Associative Classification Heart Disease Prediction (SACHDP), we used associative classification mining over landmark window of data streams. This paper contains two phases: one is generating rules from associative classification mining and next one is pruning the rules using chi-square testing and arranging the rules in an order to form a classifier. Using these phases to predict the heart disease easily [5].

M.Satish, et al. (2015) used different Data Mining techniques like Rule based, Decision Tree, Naive Bayes, and Artificial Neural Network. An efficient approach called pruning classification association rule (PCAR) was used to generate association rules from cardiovascular disease warehouse for prediction of heart disease. Heart attack data warehouse was used for pre-processing for mining. All the above discussed data mining technique were described [6].

Lokanath Sarangi, Mihir Narayan Mohanty, Srikanta Pattnaik (2015) "An Intelligent Decision Support System for Cardiac Disease Detection", designed a cost-efficient model by using genetic algorithm optimizer technique. The

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, December 2022
Impact Factor: 6.252

weights were optimized and fed as an input to the given network. The accuracy achieved was 90% by using the hybrid technique of GA and neural networks [7].

"Prediction and Diagnosis of Heart Disease by Data Mining Techniques" designed by Boshra Bahrami, Mirsaeid Hosseini Shirvani. This paper uses various classification methodology for diagnosing cardiovascular disease. Classifiers like KNN, SVO classifier and Decision Tree are used to divide the datasets. Once the classification and performance evaluation the Decision tree is examined as the best one for cardiovascular disease prediction from the dataset [8].

Mamatha Alex P and Shaicy P Shaji (2019) designed "Prediction and Diagnosis of Heart Disease Patients using Data Mining Technique". This paper uses techniques of Artificial Neural Network, KNN, Random Forest and Support Vector Machine. Comparing with the above-mentioned classification techniques in data mining to predict the higher accuracy for diagnosing the heart disease is Artificial Neural Network [9].

III. EXISTING SYSTEM

The purpose of this work is to dissect information mining processes that are required for medicinal data mining, specifically to locate locally visit ailments like as heart disease, lung cancer, and breast disease, among others. Information mining is a method of extracting data with the purpose of locating inactive examples, which Vembandasamy et al. used to assess and detect cardiac disease. The Naive Bayes technique was utilized in this case. The Bayes theorem was employed in the Nave Bayes algorithm. As a result, Naive Bayes has a lot of strength when it comes to making assumptions on its own. The data set for this study was gathered from a top diabetes research facility in Chennai, Tamandu. Weka is the tool used, and the classification is done with a 70% Percentage Split.

IV. PROPOSED SYSTEM

The suggested approach is being offered to address all of the flaws in the current system. We used logistic regression to predict the user's stroke in this dataset. The logistic regression algorithm's coefficients (Beta values b) must be estimated using the training data. Maximum-likelihood estimate is used for this. Although it makes assumptions about the distribution of our data, maximum-likelihood estimation is a typical learning process utilized by a number of machine learning algorithms.

V. SYSTEM MODEL

Fig. 1 depicts a generic network architecture that consists of four layers including the traditional cloud layer, the EC platform, the aggregation layer, and the end-device layer. Besides local execution and remote processing at cloud DCs, data and requests from end-devices (e.g., smartphones, set top-boxes, sensors) can be handled by the EC platform. Note that some data and computing need to be done in the local to keep data privacy. A request typically first goes to a Point of Aggregation (PoA) (e.g., switches/routers, BSs, APs), then it will be routed to an EN for processing. In the EC environment, various sources (e.g., smartphones, PCs, servers in a lab, under-utilized small/mediumdata centers in schools/hospitals/malls/enterprises, BSs, telecom central offices) can act as ENs. Indeed, service/content/application providers like Google, Netflix, and Facebook can proactively install their content and services onto ENs to serve better their customers. Additionally, enterprises, factories, organizations (e.g., hospitals, universities, museums), commercial buildings (shopping malls, hotels, airports), and other third parties (e.g., sensor networks) can also outsource their services and computation to the intelligent edge network.

Each service has a budget for resource procurement and wants to offload as many requests as possible to the edge network. The value of an EN to a service is measured in terms of the maximum revenue that it can generate by using the EN's resource. An EN may have different values to different ser- vices. Since some ENs (e.g., ones with powerful servers) can be over-demanded while some others are under-demanded, it is desirable to harmonize the interests of the services so that each service is happy with its allotment while ensuring high resource utilization. An intuitive solution is to assign prices to ENs and let each service choose its favorite re- source bundle. We assume that there is a platform lying between the services and the ENs. Based on the information collected from the ENs (e.g., computing capacity)

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, December 2022
Impact Factor: 6.252

and the services (e.g., budgets, preferences), the platform computes an ME solution including resource prices and allocation, which not only maximizes the satisfaction of every service but also fully allocates the ENs' resources.

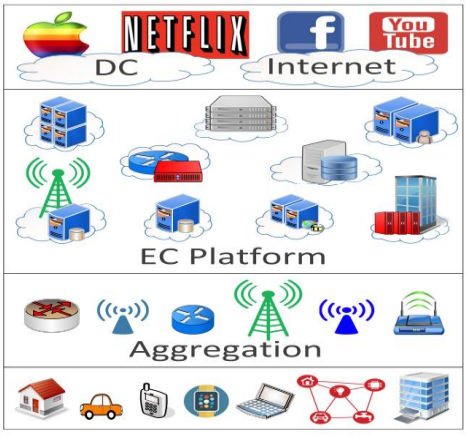

Fig. 1: An EC platform consists of geographically distributed ENs with various configurations. User/service requests are first aggregated at the aggregation layer, then routed to the ENs for processing. Requests that are not handled by the EC platform will be redirected to remote cloud.

In the first model, each service seeks solely to maximize its revenue under the budget constraint, without concerning about the money surplus after purchasing resources. This can be the case where the services and ENs belong to the same entity, and each service is assigned a virtual budget representing the service's priority. In the second model, the remaining money does have intrinsic value to the services. In this case, each service aims to maximize its net profit.

For example, this can be the case where services and Ens owned by different entities,and each SP(e.g.,Google Facebook, enterprises) has a certain budget for buying re-sources from an infrastructure provider (e.g., a Telco). For simplicity, we assume that the values of ENs to the services are fixed. Our model can be extended to capture time-varying valuation in a multi-period model by considering each pair of an EN and a time slot as an independent EN.

Notation	Meaning
PoA, SP	Point of Aggregation, service provider
EN, EC	Edge node, edge computing, data center
EG, ME	Eisenberg-Gale, market equilibrium
CES	Constant Elasticity of Substitution
MBB, EF	Maximum bang-per-buck, envy-free index
PropDyn	Proportional Response Dynamics
PropBR	Proportional Sharing Best Response (BR)
i, j	Service index and EN index
M, N	Number of ENs and number of services
$\mathcal{M}, \mathcal{N}$	Set of ENs and set of services
$D_i(p)$	Set of ENs giving service i MBB at prices p
B_i	Budget of service i
$a_{i,j}$	Revenue of service i from unit resource of EN j
c_j	Resource capacity of EN j
$d^{\mathrm{n}}_{i,j}$	Network delay between service i and EN j
$d^{\mathrm{p}}_{i,j}$	Processing delay of service i at EN j
$T_i^{\max}$	Maximum delay tolerance of service i
$x_{i,j}$	Resource amount of EN j allocated to service i
x_i	Vector of resources allocated to service i
p_j	Price of one computing unit of EN j
$p, \mathcal{X}$	Resource price vector, resource allocation matrix
$u_i(x_i)$	Revenue function of service i
$U_i(x_i, p)$	Utility function of service i
PR_i	Proportionality ratio of service i

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, December 2022
Impact Factor: 6.252

VI. MODULES

1. **DATA SELECTION AND LOADING:** The data selection is the process of selecting the data for detecting the attacks. In this project, the Heart disease dataset is used for detecting disease. The dataset which contains the information about gender, age, hypertension, heart disease, ever married, work type, Residence type, avg glucose level, bmi, smoking status, stroke**.**
2. **DATA PREPROCESSING:** Data pre-processing is the process of removing the unwanted data from the dataset. Missing data removal: In this process, the null values such as missing values are removed using imputer library. Encoding Categorical data: That categorical data is defined as variables with a finite set of label values. That most machine learning algorithms require numerical input and output variables. That an integer and one hot encoding is used to convert categorical data to integer data
3. **SPLITTING DATASET INTO TRAIN AND TEST DATA:** Data splitting is the act of partitioning available data into two portions, usually for cross-validator purposesOne portion of the data is used to develop a predictive model. And the other to evaluate the model's performance Separating data into training and testing sets is an important part of evaluating data mining models. Typically, when you separate a data set into a training set and testing set, most of the data is used for training, and a smaller portion of the data is used for testing
4. **FEARURE EXTRACTION:** Feature scaling. Feature scaling is a method used to standardize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data pre-processing step Feature Scaling or Standardization: It is a step of Data Pre-Processing which is applied to independent variables or features of data. It basically helps to normalize the data within a particular range. Sometimes, it also helps in speeding up the calculations in an algorithm.
5. **CLASSIFICATION:** Logistic regression is another technique borrowed by machine learning from the field of statistics. It is the go-to method for binary classification problems (problems with two class values). In this post you will discover the logistic regression algorithm for machine learning. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.
6. **PREDICTION:** It's a process of predicting the attacks in the network from the dataset This project will effectively predict the data from dataset by enhancing the performance of the overall prediction results and also find the accuracy of the prediction and generate confusion matrix for the output.

VII. CONCLUSION

In this project enhances the performance of the overall classification and prediction. It finds the results effectively and quickly. It alleviates the sparsity problem and reduces the information loss. The accuracy of the classification result is highly increased

REFERENCES

[1]. M. M. Rahman and D. N. Davis, Addressing the class imbalance problem in medical datasets, International Journal of Machine Learning and Computing, Vol.3, No.2, 2013.

[2]. H. L. Yin and T. Y. Leong, A model driven approach to imbalanced data Sampling in medical decision making, Stud Health Techno Inform. 2010; 160(Pt 2):856-60.

[3]. Q. GU, Z. Cai, L. Zhu & B. Huang, data mining on imbalanced data sets, International Conference on Advanced Computer Theory and Engineering, 2008.

[4]. V. Ganganwar, An overview of classification algorithms for imbalanced Datasets, International Journal of Emerging Technology and Advanced Engineering, vol.2, issue 4 2012

[5]. K. Kumar and Abhishek, Artificial Neural Networks for Diagnosis ofKidney Stones Disease, I.J. Information Technology and Computer Science, 7, 20-25, 2012

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, December 2022
Impact Factor: 6.252

[6]. H. Yan, Y. Jiang, J. Zheng, C. Peng, Q. Li "A multilayer perceptron based Medical decision support system for heart disease diagnosis" Expert Systems with Applications, vol.30, pp.272-281, 2006

[7]. L.S. Thota, S.B. Changalasetty, Optimum learning rate for classification Problem with MLP in data mining, International Journal of Advances in Engineering & Technology, vol.6, issue.1, pp.35-44, March 2013.

[8]. H.M. Nguyen, E.W. Cooper, K. Kamei, A comparative study on sampling Techniques for handling class imbalance in streaming data, SCIS-ISIS 2012

[9]. C.V. Krishna Veni, T. R. Shoba, On the classification of imbalancedDatasets, International Journal of Computer Science & Technology 20112:145-148.

[10]. P. S. Ratnoo a comparative study of instance reduction techniques. IntJournal of Advances in Engineering Sciences 2013; 3(3).

[11]. I.H. Witten, E. Frank, Data Mining: Practical Machine Learning Tool and Technique with Java Implementation. Morgan Kaufmann; 2000

[12]. H. Haibo & E.A. Garcia, Learning from imbalanced data. IEEE Transaction on knowledge and data engineering 2009; 21; 9: 1293-1284.

[13]. R.O. Duda, P.E. Hart & D.G. Stork, Pattern classification, NewYork: Wiley, 2001.

[14]. C.M. Bishop, Neural networks for pattern recognition, OxfordOxford University Press, 1995.

[15]. S. Shigetoshi, F. Toshio & S. Takanori, A neural network architecture forIncremental learning, Neuro computing, 9,111-130, 1995.

[16]. J.S.R. Jang, C.T. Sun & E. Mizutani, Neuro-fuzzy and soft computing USA, Prentice Hall, 1997.

[17]. N. Baba, A new approach for finding the global minimum of error Function for neural networks, Neural Networks, 2,367-373, 1989.

[18]. L. Mena, J.A. Gonzalez Machine learning for imbalanced Datasets: applications in medical diagnostic, AAAI, 2006

[19]. T.D.Noia,V.C.Ostuni,F.Pesce,G.Binetti,D.Naso,F.P.Schena,E.D.SciascoAn end stage kidney disease predictor based on an artificial neuralNetworks ensemble, Expert Systems with Applications, vol.40, pp.4438-4445, 2013

[20]. E.U. Kuçuksille, R. Selbas, A. Şencan, Prediction of thermo dynamic Properties of refrigerants using data mining. Energy conversion and Management 2011; 52: 836-848.

[21]. Jaganathan, M., Sabari, A. An heuristic cloud based segmentation technique using edge and texture based two dimensional entropy. Cluster Computing Vol 22, PP 12767–12776(2019). https://doi.org/10.1007/s10586-018-1757-3

[22]. Senthil kumar, V., Prasanth, K. Weighted Rendezvous Planning on Q-Learning Based Adaptive Zone Partition with PSO Based Optimal Path Selection. Wireless Personal Communications 110, 153–167 (2020). https://doi-org.libproxy.viko.lt /10.1007/s11277-019-06717-z.

[23]. Vignesh Janarthanan, A.Viswanathan, M. Umamaheswari, "Neural Network and Cuckoo Optimization Algorithm for Remote Sensing Image Classification ", International Journal of Recent Technology and Engineering., vol. 8, no. 4, pp. 1630-1634, Jun. 2019.

[24]. Dr. V. Senthil kumar, Mr. P. Jeevanantham, Dr. A. Viswanathan, Dr. Vignesh Janarthanan, Dr. M. Umamaheswari, Dr. S. Sivaprakash Emperor Journal of Applied Scientific Research "Improve Design and Analysis of Friend-to-Friend Content Dissemination System "Volume - 3 Issue - 3 2021

[25]. V.Senthilkumar , K.Prashanth" A Survey of Rendezvous planning Algorithms for Wireless Sensor Networks International Journal of communication and computer Technologies, Vol 4 Issue No 1 (2016)

[26]. Dr.Vignesh Janarthanan, Dr.Venkata Reddy Medikonda,. Er. Dr. G.Manoj Someswar Proposal of a Novel Approach for Stabilization of the Image from Omni-Directional System in the case of Human Detection & Tracking "American Journal of Engineering Research (AJER)" vol 6 issue 11 2017

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, December 2022
Impact Factor: 6.252

[27]. Sowmitha, V., and Mr V. Senthilkumar. "A Cluster Based Weighted Rendezvous Planning for Efficient Mobile-Sink Path Selection in WSN." International Journal for Scientific Research & Development Vol 2 Issue 11 2015

[28]. Viswanathan, A., Arunachalam, V. P., & Karthik, S. (2012). Geographical division traceback for distributed denial of service. Journal of Computer Science, 8(2), 216.

[29]. Anurekha, R., K. Duraiswamy, A. Viswanathan, V.P. Arunachalam and K.G. Kumar et al., 2012. Dynamic approach to defend against distributed denial of service attacks using an adaptive spin lock rate control mechanism. J. Comput. Sci., 8: 632-636.

[30]. Umamaheswari, M., & Rengarajan, N. (2020). Intelligent exhaustion rate and stability control on underwater wsn with fuzzy based clustering for efficient cost management strategies. Information Systems and e-Business Management, 18(3), 283-294.

[31]. Babu, G., & Maheswari, M. U. (2014). Bandwidth Scheduling for Content Delivery in VANET. International Journal of Innovative Research in Computer and Communication Engineering IJIRCCE, 2(1), 1000-1007.

[32]. Viswanathan, A., Kannan, A. R., & Kumar, K. G. (2010). A Dynamic Approach to defend against anonymous DDoS flooding Attacks. International Journal of Computer Science & Information Security.

[33]. Kalaivani, R., & Viswanathan, A. HYBRID CLOUD SERVICE COMPOSITION MECHANISM WITH SECURITY AND PRIVACY FOR BIG DATA PROCESS., International Journal of Advanced Research in Biology Engineering Science and Technology, Vol. 2,Special Issue 10, ISSN 2395-695X.

[34]. Ardra, S., & Viswanathan, A. (2012). A Survey On Detection And Mitigation Of Misbehavior In Disruption Tolerant Networks. IRACST-International Journal of Computer Networks and Wireless Communications (IJCNWC), 2(6).