Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)
Volume 6- Issue 1, Paper 27,January 2023

Survey on Machine Learning Algorithms for Data Science

DR.SANTHOSH BABU A.V, HAMSAREKA S, VINOTH K ,

ERODE SENGUNTHAR ENGINEERING COLLEGE

Abstract— ***Data science is a multidisciplinary field that uses number of processes, scientific methods, and various algorithms to extract the meaningful data's and insights from many structural and unstructured datasets.it is nothing but extracting useful information from larger datasets. Data science uses many kinds of algorithms. Machine learning and artificial intelligence is a core part of data science.in this paper, we are going to study different kinds of machine learning algorithms for data science and data science applications.***

Keywords—Machine Learning, Data Science

I.INTRODUCTION

Data science involves a number of disciplines and expertise areas to produce a thorough, and holistic refined look into raw data. Data scientists must be skilled in everything from math, statistics, data engineering, advanced computing and visualizations to be able to effectively sift through muddled masses of information and communication that will help drive innovation and its efficiency.
Data scientists also rely heavily machine learning and deep learning, to create models and make predictions using these kind of algorithms and other techniques.

II. OVERVIEW OF METHOD AND RESULTS

In HRFLM, we use a computational approach with the three association rules of mining namely, apriori, predictive and Tertius to the factors of heart disease on the UCI Cleveland dataset. The available information points to the deduction that females have less of a chance for heart disease compared to males. In heart diseases, accurate diagnosis is primary. But, the traditional approaches are inadequate for accurate prediction and diagnosis. HRFLM makes use of ANN with back propagation along with 13 clinical features as the input. The obtained results are comparatively analyzed against traditional methods. The risk levels become very high and a number of attributes are used for accuracy in the diagnosis of the disease. The nature and complexity of heart disease require and cacious treatment plan. Data mining methods help in remedial situations in the medical. The data mining methods are further used considering DT, NN, SVM, and KNN. Among several employed methods, the results from SVM prove to be useful in enhancing accuracy in the prediction of disease [25]. The nonlinear method with a module for monitoring heart function is introduced to detect the arrhythmias like bradycardia, tachycardia, atrial, atrial ventricular utters, and many others. The performance of this method can be estimated from the accuracy in the outcome results based on ECG data.
ANN training is used for the accurate diagnosis of disease and the prediction of possible abnormalities in the patient. Diverse data mining approaches and prediction methods, such as KNN, LR, SVM, NN, and Vote have been rather popular lately to identify and predict heart disease. The novel method Vote in conjunction with a hybrid approach using LR and NB is proposed in this paper. The UCI dataset is used for conducting the experiments of the proposed method, which resulted in 87:4% accuracies in the prediction of heart disease. The Probabilistic Principal Component Analysis (PPCA) method is proposed for evaluation, based on three data sets of Cleveland, Switzerland, and Hungarian in UCI respectively. The method extracts the vectors with high covariance and vector projection used for minimizing the feature dimension. The feature selection with minimizing dimension is provided to a radial basis function, which supports kernel-based SVM. The results of the methods are 82:18%, 85:82% and 91:30% of UCI data sets of Cleveland, Switzerland and Hungarian respectively. The hybrid method combining Linear regression (LR), Multivariate Adaptive Regression Splines (MARS) and ANN is introduced with rough set techniques and is the main novel contribution of this paper. The proposed method effectively reduced the set of critical attributes. The remaining attributes are input for ANN subsequently. The heart disease datasets are used to demonstrate the efficiency of the development of the hybrid approach. The heart disease prediction with multilayer perception of NN is proposed. This method uses 13 clinical attribute features as the input and trained by back propagation are very accurate results in identifying whether
The patient has heart disease or not.

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)

Volume 6- Issue 1, Paper 27,January 2023

III.PROPOSED METHOD HRFLM

In this study, we have used an *R* studio rattle to perform heart disease classification of the Cleveland UCI repository. It provides an easy-to-use visual representation of the dataset, working environment and building the predictive analytics. ML process starts from a pre-processing data phase followed by feature selection based on DT entropy, classification of modeling performance evaluation, and the results with improved accuracy. The feature selection and modeling keep on repeating for various combinations of attributes. Table 1 shows the UCI dataset detailed information with attributes used. Table 2 shows the data type and range of values. The performance of each model generated based on 13 features and ML techniques used for each iteration and performance are recorded. Section A summarizes the data pre-processing, Section B discusses the feature selection using entropy, Section C explains the classification with ML techniques and Section D presented for the performance of the results.

A. DATA PRE-PROCESSING

Heart disease data is pre-processed after collection of various records. The dataset contains a total of 303 patient records, where 6 records are with some missing values. Those 6 records have been removed from the dataset and the remaining 297 patient records are used in pre-processing. The multiclass variable and binary classification are introduced for the attributes of the given dataset. The multi-class variable is used to check the presence or absence of heart disease. In the instance of the patient having heart disease, the value is set to 1, else the value is set to 0 indicating the absence of heart disease in the patient. The pre-processing of data is carried out by converting medical records into diagnosis values. The results of data pre-processing for 297 patient records indicate that 137 records show the value of 1 establishing the presence of heart disease while the remaining 160 rejected the value of 0 indicating the absence of heart disease.

Num.	Code	Feature	Type	Description
1	Age	Age	Continuous	Age in years
2	Sex	Sex	Discrete	sex (1 = male; 0 = female)
3	Cp	Chest pain type	Discrete	1 = typical angina; 2 = atypical angina; 3 = non-angina pain; 4 = asymptomatic
4	Trestbps	Resting boold pressure (mg)	Continuous	At the time of admission in hospital [94, 200]
5	Chol	Serum cholesterol (mg/dl)	Continuous	Multiple values between [Minimum Chol: 126, Maximum Chol: 564]
6	Fbs	Fasting bood sugar > 120 mg/dl	Discrete	1 = yes; 0 = no
7	Restecg	Resting electrocardiographic results	Discrete	0 = normal; 1 = ST-T wave abnormal; 2 = left ventricular hypertrophy
8	Thalach	Maximum heart rate achieved	Continuous	Maximum heart rate achieved [71, 202]
9	Exang	Exercise induced angina	Discrete	1 = yes; 0 = no
10	Oldpeak	ST depression induced by exercise relative to rest	Continuous	Multiple real number values between 0 and 6.2.
11	Slope	The slope of the peak exercise ST segment	Discrete	1 = upsloping; 2 = flat; 3 = downsloping
12	Ca	Number of major vessels (0–3) colored by fluoroscopy	Discrete	Number of major vessels coloured by fluoroscopy (values 0–3)
13	Thal	Exercise thallium scintigraphy	Discrete	3 = normal; 6 = fixed defect; 7 = reversible defect
14	Class (Target)	The predicted attribute	Discrete	0 = no presence; 1 = presence

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)

Volume 6- Issue 1, Paper 27,January 2023

Table 1: UCI data set detailed information

B. FEATURE SELECTION AND REDUCTION

From among the 13 attributes of the data set, two attributes pertaining to age and sex are used to identify the personal information of the patient.

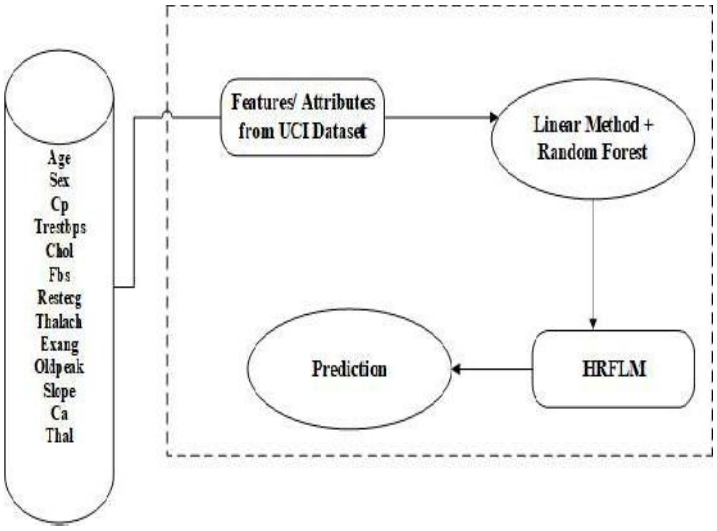

Fig 1: Prediction of heart disease with HRFLM.

The remaining 11 attributes are considered important as they contain vital clinical records. Clinical records are vital to diagnosis and learning the severity of heart disease. As previously mentioned in this experiment, several (ML) techniques are used namely, NB, GLM, LR, DL, DT, RF, GBT and SVM. The experiment was repeated with all the ML techniques using all 13 attributes. Figure 1 shows the prediction method of HRFLM.

C. CLASSIFICATION MODELLING

The clustering of datasets is done on the basis of the variables and criteria of Decision Tree (DT) features. Then, the classifiers are applied to each clustered dataset in order to estimate its performance. The best performing models are identified from the above results based on their low rate of error. The performance is further optimized by choosing the DT cluster with a high rate of error and extraction of its corresponding classifier features. The performance of the classifier is evaluated for error optimization on this data set.

1) DECISION TREES

For training samples of data *D*, the trees are constructed based on high entropy inputs. These trees are simple and fast constructed in a top down recursive divide and conquer (DAC) approach. Tree pruning is performed to remove the irrelevant samples on *D*.

Entropy = -(p(0) * log(P(0)) + p(1) * log(P(1)))

2) LANGUAGE MODEL

For given input features *xi*; *yi* with input vector *xi* of data *D* the linear form of solution $f(x)= mx+b$ is solved by subsequent parameters:

$$P_{(w_1,w_2,\ldots,w_n)} = p(w_1)p(w_2|w_1)p(w_3|w_1,w_2)\ldots p(w_n|w_1,w_2,\ldots,w_{n-1}) = \prod_{i=1}^{n} p(w_i|w_1,\ldots,w_{i-1}) \quad (1)$$

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)

Volume 6- Issue 1, Paper 27,January 2023

3) SUPPORT VECTOR MACHINE

Let the training samples having dataset *Data* D f*yi*; *xi*g I *i* D 1; 2; : : : ; *n* where *xi* 2 *Rn* represent the *ith* vector and *yi* 2 *Rn* represent the target item. The linear SVM ends the optimal hyperplane of the form *f* (*x*) D *wT x* C *b* where *w* is a dimensional coefficient vector and *b* is a offset. This is done by solving the subsequent optimization problem:

α is called the Lagrange multiplier. In terms of the SVM optimization problem,

$$f(w)=12\|w\|2,\ g(w,b)=yi(w\cdot x+b)-1,i=1...\ m.$$

The Lagrangian function is then

$$L(w,b,\alpha)=12\|w\|2-\sum mi=1\alpha i[yi(w\cdot x+b)-1].$$

4) RANDOM FOREST

This ensemble classifier builds several decision trees and incorporates them to get the best result. For tree learning, it mainly applies bootstrap aggregating or bagging. For a given data, *X*= f{*x*1; *x*2; *x*3; :::; *xn*} with responses *Y*= f{*x*1; *x*2; *x*3; :::; *xn}* which repeats the bagging from *b*= 1 to *B*.

The uncertainty of prediction on these tree is made through its standard deviation,

$$\sigma = \sqrt{\frac{\sum_{b=1}^{B}\left(fb\,(x') - \hat{f}\right)^2}{B-1}}$$

5) NAIVE BAYES

This learning model applies Bayes rules through independent features. Every instance of data *D* is allotted to the class of highest subsequent probability. The model is trained through the Gaussian function with prior probability *P (Xf)= priority* € (0,1)

$$P\left(X_{f1}, X_{f2}, \ldots, X_{f_n}|c\right) = \prod_{i=1}^{n} P\left(X_{fi}|c\right)$$

$$P\left(X_f|c_i\right) = \frac{P\left(c_i|X_f\right)P\left(X_f\right)}{P\left(c_i\right)} \quad c \in \{\, benign,\ malignant \,\}$$

At last, the testing data is categorized based on the probability of association:

$$c_{nb} = \arg\max P\left(c_k\right) \prod_{i=1}^{n} P\left(X_{fi}|c_k\right), \quad for\ k = 1, 2$$

6) NEURAL NETWORKS

The neuron components includes inputs *xi*; hidden layers and output *yi*. The result is produced through the activation function like sigmoid and a bias constant *b*.

$$f\left(b + \sum_{i=1}^{n} x_i u_i\right)$$

Algorithm 1 Decision Tree-Based Partition

Require: Input: D dataset _ features with a target class
for features **do**
for Each sample **do**
Execute the Decision Tree algorithm
end for
Identify the feature space f1, f2, . . . , f*x* of dataset UCI.

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)

Volume 6- Issue 1, Paper 27,January 2023

(9)
end for
Obtain the total number of leaf nodes l1, l2, l3,. . . , l*n* with its constraints (10)
Split the dataset D into d1, d2, d3,. . . , d*n* based on the leaf nodes constraints. (11)
Output: Partition datasets d1, d2, d3,. . . , d*n*

Algorithm 2 Apply ML to Find Less Error Rate

Require: Input: Datasets with partition _ d1, d2, d3,. . . , d*n*
for 8apply the rules **do**
On the dataset R(d1,d2,d3,. . . . . dn)
end for
Classify the dataset based on the rules C(R(d1),R(d2)
. . . . . R(d*n*)) (12)
Output: Classi_ed datasets with rules C(R(d1),R(d2)
. . . . . R(d*n*))

7) K-NEAREST NEIGHBOUR
It extract the knowledge based on the samples Euclidean distance function *d xi*; *xj*_ and the majority of k-nearest neighbors.

$$d\ (xi;xi) = \sqrt{(xi,1 - xj,1)^2 + \ldots\ldots + (xi,m - xj,m)^2}$$

IV.PERFORMANCE MEASURES

Several standard performance metrics such as accuracy, precision and error in classification have been considered for the computation of performance efficiency of this model. Accuracy in the current context would mean the percentage of instances correctly predicting from among all the available instances. Precision is defined as the percentage of corrective prediction in the positive class of the instances. Classification error is defined as the percentage of accuracy missing or error available in the instances. To identify the significant features of heart disease, three performance metrics are used which will help in better understanding the behavior of the various combinations of the feature-selection. ML technique focuses on the best performing model compared to the existing models.
We introduce HRFLM, which produces high accuracy and less classification error in the prediction of heart disease. The performance of every classifier is evaluated individually and all results are adequately recorded for further investigation.

V. CONCLUSION

Identifying the processing of raw healthcare data of heart information will help in the long term saving of human lives and early detection of abnormalities in heart conditions. Machine learning techniques were used in this work to process raw data and provide a new and novel discernment towards heart disease. Heart disease prediction is challenging and very important in the medical field. However, the mortality rate can be drastically controlled if the disease is detected at the early stages and preventative measures are adopted as soon as possible. Further extension of this study is highly desirable to direct the investigations to real-world datasets instead of just theoretical approaches and simulations. The proposed hybrid HRFLM approach is used combining the characteristics of Random Forest (RF) and Linear Method (LM). HRFLM proved to be quite accurate in the prediction of heart disease. The future course of this research can be performed with diverse mixtures of machine learning techniques to better prediction techniques. Furthermore, new feature selection methods can be developed to get a broader perception of the significant features to increase the performance of heart disease prediction.

REFERENCES

[1] A. S. Abdullah and R. R. Rajalaxmi, ``A data mining model for predicting the coronary heart disease using random forest classi_er," in *Proc. Int.Conf. Recent Trends Comput. Methods, Commun. Controls*, Apr. 2012,pp. 22_25.
[2] A. H. Alkeshuosh, M. Z. Moghadam, I. Al Mansoori, and M. Abdar,``Using PSO algorithm for producing best rules in diagnosis of heart disease," in *Proc. Int. Conf. Comput. Appl. (ICCA)*, Sep. 2017, pp. 306_311.
[3] N. Al-milli, ``Backpropogation neural network for prediction of heart disease," *J. Theor. Appl.Inf. Technol.*, vol. 56, no. 1, pp. 131_135, 2013.
[4] C. A. Devi, S. P. Rajamhoana, K. Umamaheswari, R. Kiruba, K. Karunya, and R. Deepika, ``Analysis of neural networks based heart disease prediction system," in *Proc. 11th Int. Conf. Hum. Syst. Interact. (HSI)*, Gdansk,Poland, Jul. 2018, pp. 233_239.

[5] P. K. Anooj, ``Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules," *J. King Saud Univ.-Comput. Inf.Sci.*, vol. 24, no. 1, pp. 27_40, Jan. 2012. doi:
[6] L. Baccour, ``Amended fused TOPSIS-VIKOR for classi_cation (ATOVIC) applied to some UCI data sets," *Expert Syst. Appl.*, vol. 99, pp. 115_125, Jun. 2018. doi: 10.1016/j.eswa.2018.01.025.
[7] C.-A. Cheng and H.-W. Chiu, ``An arti_cial neural network model for the evaluation of carotid artery stenting prognosis using a national-wide database," in *Proc. 39th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2017, pp. 2566_2569.
[8] H. A. Esfahani and M. Ghazanfari, ``Cardiovascular disease detection using a new ensemble classi_er," in *Proc. IEEE 4th Int. Conf. Knowl.-Based Eng. Innov. (KBEI)*, Dec. 2017, pp. 1011_1014.
[9] F. Dammak, L. Baccour, and A. M. Alimi, ``The impact of criterion weights techniques in TOPSIS method of multi-criteria decision making in crisp and intuitionistic fuzzy domains," in *Proc. IEEE Int. Conf. Fuzzy Syst. (FUZZ-IEEE)*, vol. 9, Aug. 2015, pp. 1_8.
[10] R. Das, I. Turkoglu, and A. Sengur, ``Effective diagnosis of heart disease through neural networks ensembles," *Expert Syst. Appl.*, vol. 36, no. 4, pp. 7675_7680, May 2009. doi: 10.1016/j.eswa.2008.09.013.
[11] M. Durairaj and V. Revathi, ``Prediction of heart disease using back propagationMLPalgorithm," *Int. J. Sci. Technol. Res.*, vol. 4, no. 8, pp. 235_239, 2015.

[12] M. Gandhi and S. N. Singh, ``Predictions in heart disease using techniques of data mining," in *Proc. Int. Conf. Futuristic Trends Comput. Anal. Knowl.Manage. (ABLAZE)*, Feb. 2015, pp. 520_525.