IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Multi-disease prediction model using machine learning.

Ridham Sharma
Dept. Of Computer Science, Lovely Professional University, Punjab.

Mohit Sharma
Dept. Of Computer Science, Lovely Professional University, Punjab.

Arjun Chaudhary
Dept. Of Computer Science, Lovely Professional University, Punjab.

Surbhi Sharma
Dept. Of Computer Science, Lovely Professional University, Punjab.

Abstract

In this advanced world, data is a resource, and tremendous data was creating altogether fields. Data in the medical care industry comprises of patient data and infection related data. This clinical data and AI method will assist us with examining a lot of data to discover the secret examples in the infection, to give customized treatment to the patient, and used to foresee the illness. In this work, an overall design has proposed for foreseeing the sickness in the medical services industry. This framework was tested utilizing with decreased set highlights of Constant Heart and cancer disease dataset utilizing improved AI prediction strategy consist of AI methods, for example, Knn neighbors classifier, Random Forrest Classifier in PhyCharm

From the output results, knn neighbors classifier and random Forrest produces accuracy as 95% and 89% in Breast Cancer and Heart Disease dataset respectively.

Keywords Random forest ,Knn neighbors classifier , Data analytic , Breast Cancer , Heart disease Clinical data analytics , Healthcare analytics .

Introduction

Lately, because of digitization, data was developing dramatically, taking all things together fields. Huge data is a term, which are monstrous, and it can't be prepared with standard PCs. Huge data investigation is a movement of inspecting huge datasets to reveal new bits of knowledge, esteem, and covered up designs. Enormous data examination has utilized in a few applications like climate expectation, extortion and hazard location, Calculated Conveyance and Medical care. AI calculations will assist us with contemplating the calculations that utilization enormous informational index to learn, sum up and foresee. AI is firmly identified with computational measurements and in deciding. Machine learning calculations are utilized in different applications like foreseeing the deals of the item, finding the likelihood of the event of precipitation in a specific area and so on Deliberate investigation of existing clinical information.

Saystematic examination with the AI calculations will assist us with building the prescient models for customized treatment, screen the mischief indications of the patients during their preliminary attempt and furthermore assists the specialists with recognizing the prescription for the patients. In this paper, we proposed a mix structure of a choice emotionally supportive network for anticipating the sickness utilizing AI strategies. The choice emotionally supportive network was carried out with Knn neighbors classifier and Random Forrest method for anticipating the infection and its exhibition were contrasted and different AI procedures.

In the fast arising nations, the value and standard of the healthcare programs and services are still needed to be enhanced to a level where it is good for the people. Example :- In Indonesia, we lack medical doctors and good medical study where the ratio is 0.34 doctor per 900 residents [5]. With concentrative development , the machine to machine technology and the applied science will be enormously used in many fields and practices , including various healthcares. Data and Correspondence Innovation can give a job in diseases. Indeed, as well as propelling information, huge data additionally make esteem through this. For an instance, data mining is the approach that worked in clinical study is expanding quickly. “The reason is high prescient outcomes, diminishing medication costs, improving patient wellbeing, worth, and nature of well being, and settling on choices to save individuals' lives” [8]. “Grouping is the issue of distinguishing which set of Categories, a novel perception has a place with” [9]. The nonstop estimation of prescient credits can be one of the normal order issues. “AI typically used to analyze in the event that somebody is experiencing a specific illness or not. For model, Malignancy Characterization utilizing Fluffy C-Means with Highlight Choice” [10], “Application Piece Adjusted Fluffy C-Means for Gliomatosis Cerebri” [11], “Utilization of AI on cerebrum malignant growth multiclass characterization” [12], “Normed bit work based fluffy possibilistic C-implies (NKFPCM) calculation for highly dimensional bosom disease database grouping with the highlight determination depends on Laplacian Score” [13]

Literature Review

As the world develops the new technologies evolves it is same for machine learning. This is one of the best method in AI world. The method of Machine learning is used in wast and verity of fields and it is becoming the best modern helping hand. It is consists of many algorithms which are used in different ways to achive best and effective outcomes. Some of the technology which use machine learning are:

1. Pintrest
2. Facebook
3. Google
4. Twitter
5. Baidu
6. Hubspot

We have used two machine learning algorithm in our model, one of them is K nearest neighbour algorithm which is used to solve classification as well as regression problems. We have used and experiment data set , we need to realize how significant every factor is, to know which factors are a highly prioritize. This data is a reference when boundary weighting is done in K Nearest Neighbor. We have achieved good accuracy using this algorithm. The second algorithm that we have used is “Random forest”. This method gave us a accuracy of 95% which is very good for prediction. This method made some random groups of data from data set in form of branchesand then use the algorithem for prediction. This method for arrangement of disease dataset is propitious to deliver great accuracy. Utilizing a particularly productive method, specialists can give precise choices.

Proposed Method

A. Dataset

For Heart Disease Predictin model

The data set portrays the substance of the coronary illness index. This index contains 4 data sets concerning coronary illness conclusion. All credits are numericaly-esteemed. The information was gathered from:

A. V.A. Clinical Center, Long Sea shore, CA
B. Cleveland Center Establishment
C. College Medical clinic, Zurich, Switzerland
D. Hungarian Organization of Cardiology, Budapest

The used dataset has 76 attributes, but we use 14 attributes in the experiments. Machine learning researchers and analysts only use this database upto this date.
Out of 76 attributes 8 are given below in the Table A.

Table A

No	Parameter	Description
1	Age	Age of the patient, in year
2	Sex	0 = Female, 1 = Male
3	CP	Chest Pain type: 1 = Typical angina 2 = Atypical angina 3 = Non-angina pain 4 = Asymptomatic
4	Trestbps	Resting blood pressure systolic
5	Trestbpd	Resting blood pressure diastolic
6	Restecg	Resting ECG: 0 = Normal 1 = Having ST-T wave abnormality 2 = Showing probable or definite left ventricular hypertrophy by Estes' criteria
7	Thalrest	Resting heart rate
8	Exang	Exercise induced angina: 0 = No; 1 = Yes

The last boundary is the determination result which is the expectation the outcome, regardless of whether a patient is healthy (0) or has coronary illness (1). significant motivation to pick those boundaries is that the genuine data in emergency clinic oftentimes deficient. It very well may be perceived since patient with coronary failure some of the time need brisk assistance from paraprofessional doctor and then they disregard to complete or write in the datastructure totally. This is the reason, we directed study in a clinic known as (HARKIT) in Jakarta and we have also gathered 386 diagnostic clinical or medical records and information or data with configuration depicted in Table B.

Table B

No	Field	Description	Complete / Not
1	Medical record ID	Patient's medical record ID	Complete (387 records)
2	Sex	Sex	Complete (387 records)
3	Age	Age in years	Complete (387 records)
4	Symptom	Patient's complaint description (pain, illness, etc)	Complete (387 records)
5	Additional Symptom	Additional patient's complaints	Almost complete (351 records)
6	Blood pressure	Blood pressure sys & dia	Almost complete (381 records)
7	Heart rate	Patient's heart rate	Almost complete (381 records)
8	Cholesterol level	Cholesterol level	Not Complete (15 records)
9	Trop T	Troponin T rate	Not Complete (100 records)
10	CKMB	Creatine Kinase MB level	Not Complete (74 records)
11	GDP	GDP level	Not Complete (149 records)
12	Echo	Echocardiogram test result	Not Complete (84 records)

After we checked the data-design from the HARKIT(Jakarta), we found out that a few boundaries were indistinguishable with the dataset fields . The fields that we found indistinguishable were: Sex (male or female), Age, Pulse (Trestbps and Trestbpd), Pulse (Thalrest), and Electrocardiography (Restecg). While, Cerebral palsy and Exang can be surmised from the side effects or extra Side effect in HARKIT information. We overlook all the collected data and information again and again.

For Breast Cancer Prediction model

This data was made by Dr. William H. Wolberg from University of Wisconsin Hospitals, Madison and it is consists of ten attributes.

Name of the samples	**Range**
i. Clump Thickness :-	1-10
ii. Uniformity of Cell Size :-	1-10
iii. Uniformity of Cell Shape :-	1-10
iv. Marginal Adhesion :-	1-10
v. Single Epithelial Cell Size :-	1-10
vi. Bare Nuclei :-	1-10
vii. Bland Chromatin :-	1-10
viii. Normal Nucleoli :-	1-10
ix. Mitoses :-	1 -10
x. Class:-	(2 for benign, 4 for malignant)

B. Method

Heart diseases prediction

Before we play out the KNN examination and contrast and other method, we may investigate the boundaries/factors. We need to realize how significant every factor is, to know which factors are a higher priority than the other factors. All this information and data could also be a testimonial reference when we will be needed to do the boundary weight in K Nearest Neighbour Algorithm. Also there are numerous and various approaches to process the significance of variable , out of all of them one is with the Chi-Square characteristic assessment. With Weka (Waikato Environment for Knowledge Analysis) instrument, we decide characteristic importance and significance with the Chi-Square property assessment also with the outcomes educated or declared. Due to this test we also realise that the (three) most of the significant and important factors are: A.Exang, B.Cerebral palsy, and C.Sex(male and female), and all the other different factors finished up as very less significant or important. afterwards in the KNN(k-nearest neighbors algorithm) weighting tests we will be using these 3 factors with the weight of 2 and the other factors with the weight of 1. After this knn weighting ,based from the result we will later test the KNN figuring with the extra knn weight on the Boundaries (Exang and Cerebral palsy) and Boundaries (Exang, Cerebral palsy, and Sex(male and female)

We do similar strides with 8 boundaries, presently utilizing 13 boundaries. For all these 13 boundaries, the weight grouping is A .Age, B. Sex(male and female), C. Cerebral palsy, D. Trestbps , E. Chol, F. FBS,G. Restecg, H. Thalach,I. Exang, J. Oldpeak, K. Incline,L. CA, and M. Thal. First of all we direct Chi Squared quality assessment utilizing WEKA (Waikato Environment for Knowledge Analysis)apparatuses to figure out that which factors are significant than others, that referenced in Table 3.

Table 3

Result of these 13 Tested parameters

Rank	Score	Attribute
1	110.334	11 slope
2	100.456	9 exang
3	90.583	3 cp
4	90.227	10 oldpeak
5	29.239	8 thalach
6	21.876	2 sex
7	0	4 trestbps
8	0	13 thal
9	0	7 restecg
10	0	5 chol
11	0	6 fbs
12	0	12 ca
13	0	1 age

From the test, we tracked down that the best six factors are A. Incline, B. Test, C. Cerebral palsy, D. Oldpeak, E. Thalach, and F.Sex. We could express that all these six factors are of a higher priority than other 7 factors , so in the KNN(k-nearest neighbors algorithm) weight tests we give all the best six factors as a weight of 2, while the other 7 factors as a weight of 1. At that point we will do KNN(k-nearest neighbors algorithm) weighting investigations to restrain and check the accuracy. After this We get a accuracy of 89% form Chi-Squared with Top 4 Variables and value of k=9.

Breast Cancer prediction

Assume the training-test $\mathscr{L}p = \{(A1, B1), ..., (Ap, Bp)\}$ of free indistinguishably appropriated [0,1] 1 * R-esteemed factors (l greater than and equal to 2) with a similar circulation, which is similar to a free non exclusive duo (A,B) fulfilling $B2 < \infty$, ie. B2 should not be equal to infinity. Utilizing the data $\mathscr{L}p$, objective is going to appraise the relapse work r(a) = [B [A = a]] for $a \in [0,1]l$. Thusly, we can say that the assuming $E[rp(A) – r(A)]2 \rightarrow 0$ as $p \rightarrow \infty$ the assessment of relapse work r is mostly reliable. All in all, a random forest classifier is a classifier that is comprising with bunch of randomly based relapse trees { $rp(a, \gamma q, \mathscr{L}p)$, $q \geq 1$}, where $\gamma 1, \gamma 2, ...$ are autonomous indistinguishably dispersed yields with randomization variable γ. After that all the available random trees are combined and will frame total relapse gauges

$$rp(A, \mathscr{L}p) = E[rp(A, \gamma, \mathscr{L}p]$$

where $E\gamma$ is exception in relation to some randomly used parameters of A and dataset $\mathscr{L}p$. Dependent is denote by $\bar{r}p$ (A). Now we exclude some another strategies that could relies on some of the information and data to make trees. This way every tree is formed. Every single hubs are associated with cells same as assortment of cells (i.e., outer hubs at progression of development tree shapes a segment itself. After that this procedure work again [log2jp]times, where

a) log2 = base-2 algorithm

b) ceiling function ≥ 2

c) $jp \geq 2$

Every random tree $rp(A, \gamma)$ yields a normal over all Bm and vector Am, of all random segment as comparing A comes in a similar cell. It lets $Bp(A, \gamma)$ to be the cell of the random parcel.

In particular, each tree should think about an alternate random subset of boundaries. While picking the most educational boundary, we ascertain the mistake if the boundary esteem is shown on the whole out-of-sack (OOB) perceptions. An increment in forecast blunder happens if the boundary esteem shown altogether OOB perceptions is determined for each tree, at that point the whole outfit is found the middle value of and partitioned by the standard deviation of the whole outfit

Experimental Data And Result

Before we play out the KNN examination and contrast and other method, we may investigate the boundaries/factors. We need to realize how significant every factor is, to know which factors are a highly prioritize. This data is a reference when boundary weighting is done in K Nearest Neighbor. We have numerous approaches for processing variable significance, e.g. Chi-Square property assessment. With apparatus, the property significance is decided having Chi-Square characteristic assessment. From the output results, knn neighbors classifier and random Forrest produces accuracy as 95% and 89% in Breast Cancer and Heart Disease dataset respectively.We utilize "a" % of data set as training set and other used if form of testing set on trial for approving calculations, where "a" = 10, 20, 30,40…… , 90. Examination then rehashed again and again. Order accuracy was estimated by:

Accuracy = ((True Positive+True Negative) / (True Positive+True Negative+Fasle Positive+False Negative)) X 100

From the outcomes , This method for arrangement of disease dataset is propitious to deliver great accuracy. Utilizing a particularly productive method, specialists can give precise choices.

CONCLUSION

The method of mining is utilized in numerous fields for example medical care. This paper's goal is to check coronary failure expectation could be founded in less boundaries and than suggested in past investigations. Boundaries are utilized, which are: Age, Sex, Chesttorment, Resting pulse systolic, Tests utilizing boundaries with KNN shows great precision in the event that we contrasted and 14 boundaries, with mining calculations as Bayes and Choice Tree. In future exploration, exceptions tends to utilized same to boundaries in distant checking utilizing machine to machine innovation, particularly for homely treated patients. Random forest is troupe learning calculations which is an entirely adaptable classifier. Random forest calculation is a calculation which frames some group of characterization techniques that rely upon a blend of a few choice trees. The arbitrary forest likewise runs effectively in enormous data sets. Notwithstanding, there is a shortcoming in the arbitrary woods; it is acceptable at order however not as great concerning relapse. The outcome is 95 % in the paper. In addition , when 80 % to 90 % of the data is utilized for the preparation information, the precision is 95 %, now this implies that it is generally precision for forecasting breast cancer.

REFERENCES

[1] Cleveland Clinic Foundation (cleveland.data)
[2] Hungarian Institute of Cardiology, Budapest (hungarian.data)
[3] V.A. Medical Center, Long Beach, CA (long-beach-va.data)
[4] University Hospital, Zurich, Switzerland (switzerland.data)
[5] Kementerian Kesehatan Republik Indonesia. Profil Kesehatan Indonesia
[6] O. L. Mangasarian and W. H. Wolberg: "Cancer diagnosis via linear programming", SIAM News, Volume 23, Number 5, September 1990, pp 1 & 18.
[7] William H. Wolberg and O.L. Mangasarian: "Multisurface method of pattern separation for medical diagnosis applied to breast cytology", Proceedings of the National Academy of Sciences, U.S.A., Volume 87, December 1990, pp 9193-9196.
[8] H. Asri, H. Mousannif, H. Al Moatassime and T. Noel, in Procedia Comput. Sci. (2016), pp. 1064-1069.
[9] J. Tang, S. Alelyani and H. Liu, *Feature Selection for Classification: A Review* (CRC Press, Boca Raton, 2014).
[10] A. Wulan, M.V. Jannati, Z. Rustam and A.A. Fauzan, in Proc. - 2016 12th Int.Conf. Math. Stat. Their Appl. ICMSA 2016 Conjunction with 6th Annu. Int. Conf. Syiah Kuala Univ. (2017), pp. 35-38.
[11] V. Panca and Z. Rustam, AIP Conf. Proc. 1862, 030133 (2016).
[12] A. W. Lestari and Z. Rustam, AIP Conf. Proc. 1862, 030143 (2016).
[13] L Breiman, Machine Learning 45, 5-32 (2001).
[14] T. L. Octaviani, Z. Rustam. "Random forest for breast cancer prediction".