Forecasting Chronic Disease using Gradient Boosting Algorithm

K.Sindhanaiselvan, Shridevi

Abstract---A Chronic disease is exists inside human body long term. The continual ailment typically final for three months or more as per defined with the aid of the USA National center of Health Statistics. The leading persistent sicknesses in advanced international locations encompass arthritis, cardiovascular ailment which includes coronary heart attacks, stroke and oral fitness issues. These illnesses are resistant to vaccination and can't be prevented by means of way of remedy. Eating conduct, lack of workout and awful meals conduct are the essential contribution to the persistent sickness prevalence. In propose system, used superior device gaining knowledge of algorithms like assist vector machines, Gradient boosting, to expect the superiority of chronic sickness. Also, as part of t observe have provided the comparative observe of all the fashions with go validation techniques.

Keywords: Chronic Disease, Gradient Descent Boosting,Supervised Machine Learning, K Fold Validation Confusion matrix

I. INTRODUCTION

Machine learning is area where we don't had to explicitlyprogramfortheoutcomeofthevariable.Wewill be using the historical pattern to predict the occurrence of chronic disease. There are three types of machine learning algorithms. Supervised, unsupervised and re enforcement algorithms. Supervised algorithms where you will have labeled data and unsupervised algorithms will have unlabeleddata.The outcome of the variable is class we call these algorithms as classification algorithms and the outcome is continues variable, we call these algorithms as regression algorithms the domain has revolution across the areas that .It will be using machine learning algorithms to predict the occurrence of chronic disease researchers have used this advanced machine learning algorithms for making the prediction in health care field. One of the use cases of this behind prominent in chronic diseasepredictions

Chronic disease refers to a disorder that persists for lengthy period. Eighty percent of Americans over sixty five years of age have as a minimum one of the chronic ailments. The persistent sicknesses aren't suppressible by way of vaccines or cured with the aid of medicine no longer do they disappear within the human body. They just persist long time till it damages the human frame. Heart attacks, breast cancer, diabetes, epilepsy and obesity are a number of the frequently passed off ailment

As per American association for retired professional's people's who has more than sixty five years of age can have as a minimum one chronic sickness. Lack or physical activity, bad consuming behavior and utilization of tobacco and alcohol are fundamental motives for occurrence of the persistent disease. A chronic disease is persists for lengthy duration in human being's health condition that may not have a treatment. Examples of persistent ailments are: Alzheimer disease and dementia, Arthritis, Asthma, Cancer, COPD, Cystic fibrosis, Diabetes, Epilepsy, Heart sickness, HIV/AIDS, Mood disorders (bipolar, cyclothymiacs, and melancholy), Parkinson sickness. Symptoms of persistent ailment: Loss of memory and awareness and throat contamination. Headaches and Joint muscle Problems. Machine studying for predicting continual ailment to construct an ensemble system mastering model that are expecting the incidence of the persistent sickness in the sufferers of various demographics (Age, Gender, Habits, Race...And so on). As those are long time illnesses proposed structures required to devise good enough stock of the medicine in advance. Hence, this model can assist in planning the drugs supply higher within the hospitals

II. RELATED WORK

Supervised machine leaning algorithms (Classification) and Clustering algorithms (Unsupervised Machine learning was used to predict the occurrence of the chronic disease. These algorithms have the following advantages when compared to other models. A combination supervised and unsupervised machine learning algorithms has been used to predict the final outcome with highest accuracy. However, this research work doesn't consider the Dimensionality and variable reductions methods could have yielded better results. The model is not generalized; it works for each cluster differently. In early days researches used to predict the number of chronic patients visitthe hospitals which will help in planning for the medication between. A series of linear, non linear, time series for casting models were used to predict the outcome of the occurrence of the chronic patient counts. As the rise of the technology people started capturing the data in different formats hence it leads to capturing more details about the patient. This will give us scope to perform netter machine learning algorithms

Theodora[7] offered a aggregate of categorized and unlabelled strategies like supervised and unsupervised studying algorithms has been used to offer cease result of the fashions. Initially the chronic patient is classified into more than one categories and very last classifier has been run on every cluster to predict the outcome of the models. Each

Revised Manuscript Received on 14 August, 2019.

Dr.K.Sindhanaiselvan, Dept. of CSE, MVJ College of Engineering, Bangalore, Karnataka, India 56007

Shridevi, Dept. of CSE, MVJ College of Engineering, Bangalore, Karnataka, India 56007

cluster could be dealt with as a separate dataset and for every cluster there might be separate classifier and are expecting the whether a patient can have chronic ailment or not disease. Krishan L. Khatri[1] proposed that predicting persistent ailment is one of the key aspects to control the supply of the medicine in the course of the height seasons. This will assist us to are expecting the demand upfront and plan for the treatment of the patients upfront. An superior neural community changed into used to predict the affected person's.

Joshua[2] suggested to track efficaciously the patient fitness records and device is hooked up to display the health of the patient and constantly we will be track in the development of all the important key function of the factors within the body. Hence, there are more accurate statistics will be used to are expecting the persistent sickness prevalence. By using distinctive algorithms there may be prediction of persistent sickness within the affected person. Lambda Jena [4] offers persistent disorder is one of the key illnesses throughout the season inside the present day studies, this technique have mentioned approximately the few category algorithms with a purpose to assist us in predicting continual disorder occurrences accurately. Out of many machines leaning algorithms available out of the ensemble fashions were select as a part of the studies paintings. To degree the effect of the model accuracy precision, recollect and accuracy rating have been used.

III. METHODOLOGY

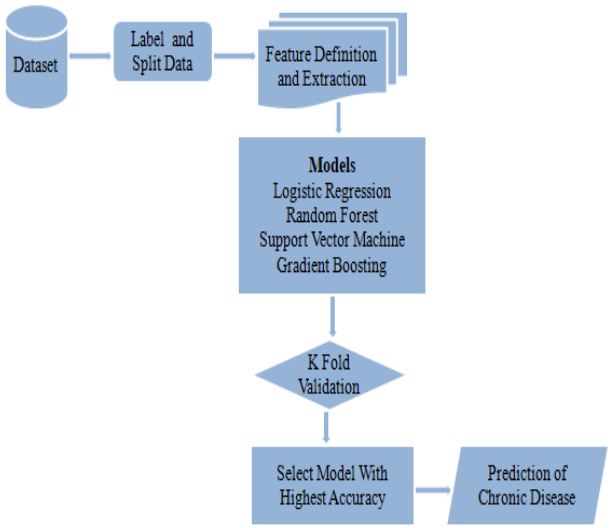

Fig1: System architecture

Solving the problem starts off evolved with identifying the right factors required for chronic disease. The prioritized factors are taken into consideration for growing analytical data set. Different models are carried out on analytical dataset. All algorithms are verified thru k fold validation methodology to find the proper accuracy. Any supervised system reading algorithms will require a scientific flow of the below steps. These are generalized frameworks that can assist in defining the hassle better and executing the all of the degrees indexed out underneath in hooked up way.

Defining the problem: Identifying the right components required for flaw forecast

Dataset Labels creation: Converting the base dataset to logical informational series

Model: Creating the baseline version for predicting the chronic sickness occurrence.

Model Validation: Validate the version normal overall performance on confusion matrix, ROC-AUC Curve.

K Fold Validation: Test the version usual overall performance across the extraordinary datasets.

A. Problem solving framework

One of the essential tiers of the any device getting to know trouble is defining the trouble higher. The key elements required for the evaluation are diagnosed in the system. It will help in listing down the factors without bias. The issue map and Hypothesis are generated. Identify and prioritize the essential elements based totally on movements and feasibility matrix.

B. Data pre-processing (exploratory statistics analysis)

Majority of the time the information accumulated for walking system learning algorithms isn't to be had readymade. Hence, the information calls for pre-processing an amazing way to assist to get the proper predictions. The information may be handled as one of the maximum important steps of system getting to know set of rules predictions. Data Collection, Data Merging, Null Value Treatment, Outlier Treatment, Garbage Value Removal and growing the analytical geared up information set are the essential steps in EDA. In our works, the above preprocessing steps are included to create the analytical data set.

Data Sets

The dataset used in the analysis changed into referenced from Kaggle public licensed datasets.

Table1. DATA SETS

Patient Id	Ethnicity
RB Count	Kidney Dysfunction
Gender	Symptoms
Education	CCIS
Marriage	Disturbance of Equilibrium
BMI Index	Shortness of Breath

C. Models and Comparison

The data model was built using Logistic regression Random forest, SVM and ensemble classifier. This project have created different model with following: Logistic regression is one of the base models utilized in our work. As the malicious program prediction is a binary response variable, the malicious program conduct can be predicted by more than one algorithms. Logistic Regression is taken into consideration as base version and other algorithms may be taken into consideration as benchmarking models. Data set is divided in to train and take a look at set. Validating the performance of the models is completed by discover the Accuracy score, Confusion Matrix, ROC- AUC curve designing, Probability curve, Improving model overall performance

Logistic Regression Steps

ALGORITHM

Logistic Regression

Step 1: Data Prepared
Step 2: Data Preprocessing
Step 3: Unsupervised Cluster
Use K-means Algorithm with different value of data, remove incorrectly classified data
Step 4: Calculate: rate = remaining (data / sum)
Step 5: If the rate is lower accuracy, then execute again with another value of data
Step 5: Use logistic supervised classification regression algorithm and model validation

Random Forest Steps

ALGORITHM

Random Forest

To create training model, many decision trees are needed, for each decision tree performs following steps:
Step 1: M features Training samples
Step 2: Extract m from M
Step 3: Randomly choose a training set for this tree by estimated the prediction error of the tree

- Randomly choose M
- Calculate the best split based on m in the training set

Step 4: Each tree is fully grown and not pruned
Step 5: Random forest model (Many decision trees)
A test sample and class label produces decision trees

Support Vector Machine Steps

ALGORITHM

Gradient Descent Boosting

Step 1: Given a data sample distribution D and determine the total number of base models as M
Step 2: Define the initial training sample distribution as D1= D for m= 1 to M do
Step 3: Train a base model BM(x) from the training sample distribution Dm
Step 4: Compute the error of the model, adjust the distribution Dm to Dm+1 to make the mistake of the model more evident
Step 5: Output the constructed base model Bm(x)
Step 6: Output the prediction of the ensemble trees for a given new input $x = \frac{1}{y}\sum_{m=1}^{m} Bm(x)$

Gradient Boosting Steps

Bagging is a method used to reduce the variance of our predictions by means of combining the end result of a couple of classifiers modeled on distinct sub-samples of the equal facts set. The steps observed in bagging are:

- *Create Multiple Datasets:*

Sampling is performed with alternative at the unique information and new datasets are formed. The new statistics sets will have a fragment of the columns in addition to rows, which might be typically hyper-parameters in a bagging version. Taking row and column fractions less than 1 enables in make strong models, much less liable to over becoming.

- *Build Multiple Classifiers and Combine Classifiers:*

Classifiers are constructed on each statistics set. Generally, the equal classifier is modeled on every records set and predictions are made. The predictions of all the classifiers are combined the usage of a median, median or mode price relying on the hassle handy. The combined values are typically extra strong than a unmarried version.

ALGORITHM

SVM algorithm (k SVM)

Step 1: Input training dataset D, number of local models k
Hyper-parameter of kernel function γ
C for tuning margin and errors of SVMs
Step 2: Output k local support vector machines models
Step 3: Applied k-means performs the data clustering on D
Creating k clusters denoted by C1, C2, C3 and their corresponding centers Ck
for i ← 1 to k do
Learning a local SVM model from Ci
lSVMi = SVM (Di, γ, C)
Return kSVM Model = {(c1, lSVM1), (c2, lSVM2), (ck, lSVMk)}
Step 4: End

IV. MODEL VALIDATION

K-Fold Cross Validation: K-Fold Cross Validation is a commonplace kind of cross validation this is broadly used in system getting to know. For every particular group:

•Take the organization as a preserve out or test information set
•Take the last corporations as a schooling facts set
•Fit a model at the schooling set and examines it on the take a look at set
•Retain the evaluation score and discard the model
•Summarize the skill of the model the usage of the pattern of version assessment ratings

A assignment is gone thru 10-Fold go validation at the algorithms along with Random Forest, SVM and Logistic Regression and gradient descent boosting and the consequences obtained are consolidated in the Table 2 and the corresponding graphs (x axis: folds, y axis: Accuracy) are shown in Fig 4.The Following observations can be made from Table 2.Gradient boosting produces the very best accuracy. Our prediction model produces Less Variance and Biasness. Better choice of variables thru feature engineering and prioritization matrix has been obtained.

Proposed system implemented for 10-fold move validation on the algorithms which include logistic regression, random woodland, linear hyper plane algorithm and gradient descent boosting algorithm to get a better accuracy and the effects received are consolidated and the corresponding in to graphs.

Confusion matrix condenses the aftereffects of the testing calculation and offers a file of the quantity of True Positive (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN).

Predicted	Actual	
	Class X	*Class Y*
Class X	TP	FP
Class Y	FN	TN

Accuracy, Precision, F- degree

Accuracy = (TP + TN) / (TP + TN+ FP + FN)
Precision = TP / (TP + FP)
F-measure = (2* Recall * Precision) / (Recall+Precision)

Gradient Descent Boosting (10 Fold: Accuracy)				
88	87	88	87	87
87	86	85	87	89

Random Forest (10 Fold: Accuracy)				
85	86	86	86	87
87	86	86	86	88

Logistic Regression (10 Fold: Accuracy)				
86	86	86	86	86
86	86	86	86	86

Support Vector Machines (10 Fold: Accuracy)				
86.17	86	86	86.17	86
86	86.13	86.13	86	86.13

Table 2. Accuracy scores of all models

V RESULTS

because of its additional specific drivability and warmth efficiencies. As respects to oil based good crisis and growing car populace the quest for plausibility gas has have turned out to be out to be basic for diesel engines as a result of ventured forward environmental issues, and money related edges. Marvelous

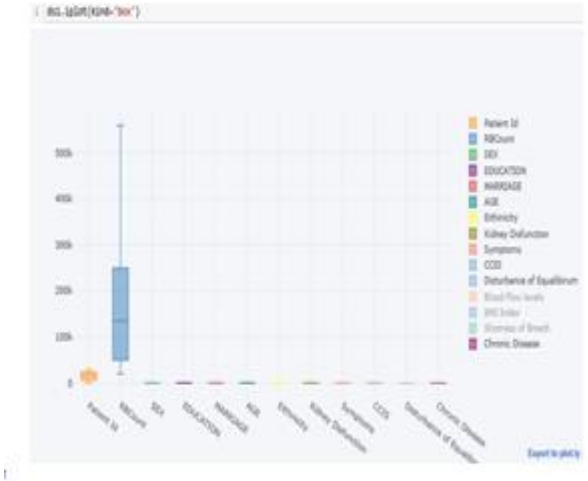

Fig 2. Box plot

```
kelbow(ds1,8)
```

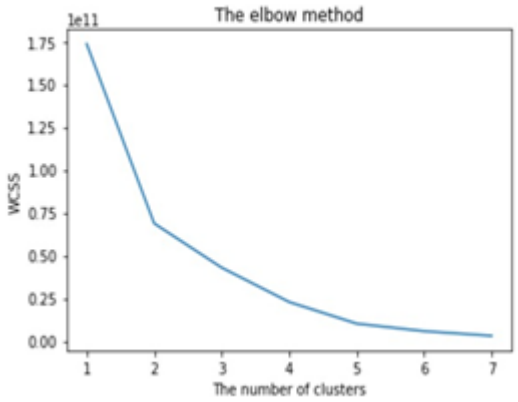

Fig 3.Clusters of datasets.

Table 3: Final accuracy comparison table

	LogisticRegression	RandomForest	SVM	Gradientboosting
0	0.861702	0.856383	0.861702	0.880319
1	0.861702	0.869681	0.861702	0.872340
2	0.861702	0.869681	0.861702	0.880319
3	0.861702	0.867021	0.861702	0.877660
4	0.861702	0.867021	0.861702	0.877660
5	0.861702	0.869681	0.861702	0.867021
6	0.861333	0.864000	0.861333	0.861333
7	0.861333	0.861333	0.861333	0.858667
8	0.861333	0.861333	0.861333	0.877333
9	0.861333	0.880000	0.861333	0.888000

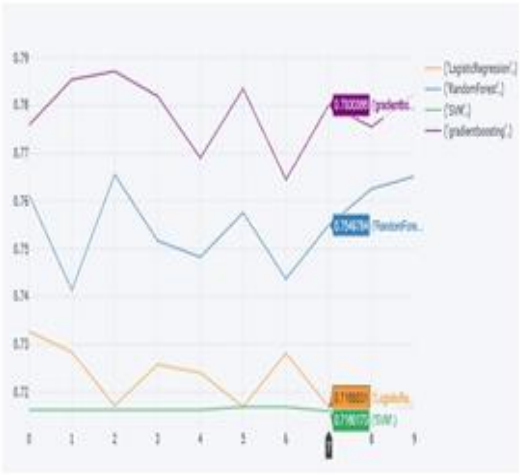

Fig 4. Final accuracy comparisongraph

Table 3. Accuracy score

Gradient Descent Boosting					
Accuracy Score		Precision	Recall	F1-score	Support
88.83 %	0	0.90	0.98	0.94	816
	1	0.71	0.26	0.38	124

VI. CONCLUSION

As in line with American association for retired experts, People who're above50 age organization can have at least one continual sickness in United States alone. Hence, as that is the out of the want, the prediction systems for chronic ailment detection are critical. This will assist in planning the medication in advance and also plan for the medication stock out efficiently. Proposed machine has examined multiple device mastering models to expect the outcome of the and finally selected gradient descent boosting fashions for the very last choice. The gradient boosting model has been examined

VII. FUTURE ENHANCEMENT

To further enhancements and superior neural community models like (LSTM and Resnet can be used to predict the incidence of the chronic sickness). A stop to end product which may be utilized by the physicians to be expecting continual disorder also may be dealt with as enhancement for future work.

REFERENCES

1. Krishna L. Khatri, and Lakshman S. Tamil, Senior Meeberrie "Early detection of peak demand days of chronic respiratory diseases emergency department visits using artificial neural networks" Vol.22,No.1, JAN, 2018.
2. Joshua June "A natural walking monitor for pulmonary patients using mobile phones" in: IEEE Journal of Biomedical and Health Informatics Vol.19, No.4, JUL, 2015.
3. Lambda Jena and Ramakrishna Swain "Chronic disease risk prediction using distributed machine learning classifiers" International Conference on Information Technology IEEE,2017.
4. Benjamin Letham, Cynthia Rudin, Tyler H. Mccormick, and David Madigan, "Interpretable classifiers using rules and Bayesian analysis: building a better stroke prediction model" The Annals of Applied Statistics Vol. 9, No. 3, JAN 2015.
5. Jinghe Zhang, Kamran Kowsari, James H. Harrisonyz, Jennifer M. Lobo and Laura E. Barnes. "Patient2Vec: A Personalized Interpretable Deep Representation of the Longitudinal Electronic Health Record", IEEE Access JUL 2018.
6. Fei Wang, Ping Zhan, Buyup Qian, Xiang Wang, Ian Davidson IBM T. J. "Clinical risk prediction with multi linear sparse logistic regression" ACM SIGKDD international conference on Knowledge discovery and data mining ,AUG 2014.
7. Tinting Xu, Theodora S. Brahma, Tamiya Wang, Ouyang Dai, and Iohannis Ch. Paschalis's, Fellow, "A joint sparse clustering and classification approach with applications to hospitalization prediction" IEEE 55th Conference on Decision and Control (CDC) DEC 16.
8. Joerg Habetha "The heart Project Fighting Cardiovascular Diseases by Prevention and Early Diagnosis" Proceedings of the 28th IEEE EMBS Annual International Conference New York City, USA, AUG 2006.
9. www.quora.com,www.google.com,https://en.wikipedia.org/wiki/Chronic_condition
10. https://www.who.int/chp/about/integrated_cd/en/
11. Li Wang, Brian Porter, Charles Maynard, Ginger Evans et.al, "Predicting Risk of Hospitalization or Death Among Patients Receiving Primary Care in the Veterans Health Administration" Medical Care, Vol.51, No. 4, APR 2013.
12. Dr. S. Vijayarani1, Mr.S.Dhayanand2 "Data Mining Classification Algorithms for Kidney Disease Prediction" Vol. 4, No. 4, Aug 2015.
13. Liu, Zhang and Razavian "Deep EHR: Chronic Disease Prediction Using Medical Notes" Proceedings of the 3rd Machine Learning for Healthcare Conference AUG 2018.
14. Lauren Crain Patrick J. O'Connor, MD, MPH,William A. Rush, JoAnn M.Jay J. Gutenkauf, Jane E. Duncan, MPH "Impact of an Electronic Medical Record on Diabetes Quality of Care" Annals of family medicine Vol. 3, No. 4 JULY 2005.
15. Asmaa S. Hussein, Wail M. Omar, Xue Li "Efficient Chronic Disease Diagnosis Prediction and Recommendation System" International Conference on Biomedical Engineering and Sciences DEC 2012

AUTHORS PROILE

K.Sindhanaiselvan, received the Bachelor of Technology in Information Technology from Anna University, Chennai in 2005 and his Master of Engineering Degree in Computer Science from the Anna University, Chennai in 2007 and his Doctor of Philosophy in Information and Communication Engineering from the Anna University, Chennai in 2017. His current area of research interest is Energy Efficient in Mobile Adhoc networks, Wireless Sensor Networks and Software Defined Networks

Shridevireceived the Bachelor of Engineering in Computer Science and Engineering from VTU University, Belagavi in 2017 and her Master of Technology in Computer Science and Engineering Bangalore from the VTU University,Belagavi in 2019. Her current area of research interest is Machine Learning,CloudComputing,Chronicdisease,IoT.