CrossRef DOI of original article: 10.34257/GJCSTHVOL21IS1PG23

AROGYA Intelligent Health Care Application

Anupama Kannangara[1] and Rathnayaka W.G.P.N.[2]

[1] Sri Lanka Institute of Information Technology

Received: 15 June 2021 Accepted: 1 July 2021 Published: 15 July 2021

Abstract

People of today pay less attention to their daily diet due to their busy lifestyles. Therefore, there is a great tendency to contract chronic non-communicable diseases. Furthermore, the lack of nutritious meals and daily exercise causes chronic non-communicable diseases in people with no age difference. In our first part, we developed to predict specialization in cardiology using symptoms. However, when we refer to a doctor, we must at least know what specialist should know based on their symptoms. In addition, there is a problem with the recipe. If the pharmacist has misread the prescription given by the doctor, patients can receive bad medications, leading to terrible side effects and even death due to careless writing by the doctor. As a solution, an application function can be proposed that will be developed in the project and the function should be able to improve the readability and intelligibility of the patient with prescription drugs. Therefore, the patient always knows the prescribed medications through the application to avoid the problem mentioned above. According to the 2016 pharmaceutical magazine, there are cholesterol and diabetic patients suffering mainly from chronic non-communicable diseases in Sri Lanka.

Index terms—

1 Introduction

uman health is main part of the life. The meaning of health has evolved over time. From the biomedical perspective, early definitions of health focused on the issue of the body's ability to function; health was viewed as a normal functioning state which could be interrupted from time to time by illness. Also, in current era, people too much busy with their works. Because of that, people are suffering with noncommunicable chronic diseases [7]. According to the Pharmaceutical Journal of Sri Lanka 2016 [7], there are Hypertension 48.5%, Diabetes mellitus 45.3% and Ischemic Heart disease 29.4%. Proposed application based on non-communicable chronic diseases. First part mainly based on heart diseases. Because the heart diseases area is wide in the medical world. There are many categories based on heart diseases named blood vessel diseases, coronary artery disease; heart rhythm problems (arrhythmias); and heart defects born with (congenital heart defects), among others [8]. Also, Cardiovascular disease usually refers to conditions that involve narrowed or blocked blood vessels that can lead to a heart attack, chest pain (angina), or a stroke. Other heart conditions, such as those that affect muscle, valves, or heart rate, are also considered to be forms of heart disease. Other part based on cholesterol, diabetic and blood pressure. Cholesterol is a chemical compound that the body needs as a building block for cell membranes and hormones like estragon and testosterone. The liver produces about 80% of the body's cholesterol and the rest comes from food sources such as meat, chicken, eggs, fish, and dairy products. Plant-based foods do not contain cholesterol [9]. Cholesterol divided in to three parts known as Highdensity lipoprotein (HDL), Low-density lipoprotein (LDL) and very low-density lipoprotein(VLDL). There are levels known as VLDL $<$ 40 mg/dL, LDL $<$ 160 mg/dL, HDL $>=$ 45mg/dL, Triglycerides$<$ 150 mg/dL.

2 II.

3 Method a) Data sets i. Heart disease prediction model

The data set that we used in this research, is used various researchers for their research purpose. We get it the web site called kaggle.com [10].This dataset was used in this research designing for heart disease diagnosis for machine-learning-based system. This heart disease related dataset has a sample size of 4240 patients, 16 features.

ii

4 . Prediction for future cholesterol level model

The data set that we used to predict cholesterol level is a created data set by me. This data set used in this research for predict cholesterol level for future six months suing time series analysis. This cholesterol level related dataset has a sample size of 20 months of one patient.

iii. Model for diet plan I created dataset to similar medicine using SPC guidelines. Because there is no dummy, data set to this part. In that case, I met a doctor and created a sample dataset to create the model.

To predict the diabetic level and cholesterol levels part we got the dummy data using https://www.kaggle.com/ website.

5 iv. Give an idea of the prescription

Collected cholesterol prescription as an image data set and created a stranded dataset.

6 b) Data processing

There is feature called education. That one is not related to the heart disease. Therefore, we drop that feature. During the cleaning, remove null values. Some null values fill with the mean value of the feature and get a value, which will increase the efficiency. In the prediction of the diabetic and cholesterol level part, started to create the model using jupyter notebook. First, imported necessary libraries and added the dataset. Also dropped the unnecessary data column in the data set.

7 c) Methodology of the Proposed System i. Heart disease prediction model

The proposed system developing with the aim to classify weather people should channel cardiology or not. One of the popular machine learning classifiers logistic regression used for classification of this system. Logistic regression is the one of best classifier to get binary value output. The methodology of the proposed system structured into four stages including (1) preprocessing of dataset, (2) feature selection, (3) machine learning classifiers, and (4) classifiers' performance evaluation methods. Figure 1 shows the framework of the proposed system. In order to classify two classes 0 and 1, a hypothesis will be designed and threshold classifier output is at 0.5. If the value of hypothesis, it will predict y = 1 which mean that the person has heart disease and if value of, then predict y = 0 which shows that the person is healthy.

Hence, the prediction of logistic regression under the condition is done.

My ratio is 80-20. 80% data will train and 20% will be test. Import confusion matrix to represent the false positive, false negative, true positive and true negative.

8 ii. Prediction for future cholesterol level model

This proposed system developing to predict a cholesterol level for about 6 months of future and store patient past data records of cholesterol level. The time series is one of the popular machine learning prediction algorithms. In time series analysis have one variable at that time. There have an independent variable and a dependent variable. Time series prediction is a form of data mining that predicts future behaviors by analyzing historical data.

( )H Year 2021

The objectives of a time series prediction ,t is estimated value of x and ,X[t+s]=f(x[t],x[t-1],?..,x[t-N]), s>0 is called the horizon of prediction. Figure **??** shows the prediction of a time series using auto regression integrated moving average (ARIMA-model) **??**11].

9 Figure 2

The way a Simply Moving Average is calculated is that it takes the subset of the data mentioned in the moving average model description, adds together the data points, and then takes the average over the subset of data. It can help identify the direction of trends in your data and identify levels of resistance wherein business or trading data. **??**12] Forecasting is one of the most relevant tasks when working with time-series data. You can forecast with a simple moving average, another moving average model called 'Autoregressive Integrated Moving Average' is popular for fairly accurate and quick forecasting of time series. The Autoregressive Integrated Moving Average, or ARIMA model, is a linear function that is used for predicting future data points based on past data.

ARIMA combines the models the past data points to determine future points to the linear regression model on an independent variable to predict the dependent variable. Because of ARIMA's using past data, a longer series is preferable to get results that are more accurate. [13] iii. Model for diet plan

In decision tree classification data model have two main types known as classification tree and regression tree. This is a non-parametric supervised learning method [1]. In this data model, predict the value of the target variable in the data set by learning simple decision rules. In classification, tree outcome was yes/no type. Those decision variables are categorical or discrete. Also, it known as binary recursive partitioning. However, regression tree is taking continuous values or real numbers [2]. There are many different algorithms but in here, mainly used ID3 (Iterative Dichotomies 3) algorithm [3] invented by Ross Quinlan. Simple meaning of this is greedy search via the space of possible branches without no reverse. It is built top-down from a root node and create subsets using similar values. This is known as homogenous [4]. ID3 algorithm used entropy (figure **??**) to appraise the homogeneity of the data set.

10 Figure 3

Entropy using the frequency table to calculate the decision tree within two types. First, one is for one attribute (Figure 4). Second, one is for two attributes (Figure 5). There are two formulas for above mention types.

11 iv. Give an idea of the prescription

First recognized of handwritten medical forms. For that, I used Lexicon Driven Word Recognizer Algorithm [5]. All lexicon entries are treated as detached words and matched the input word image as containing handwriting to recognize in word model-based recognition. Lexicon entry is the best top choice of this. To develop this model, we created word recognition methodology (Figure **??**). Segments are matched against individual characters without using any contextual information in character model-based recognition. In addition, we used Latent Semantic Analysis [6] to compute the relationship between the context of words and terms to a semantic category. For the cholesterol prediction section, we used the performance of time series analysis to predict future cholesterol levels. The result of selecting time series analysis, inputs of cholesterol levels are used to convert to log scale and giving a plot graph and showing prediction line for six months. Other than that in that graph shows the confidence level of prediction In the cholesterol level and diabetic level, get the 90% as accuracy score. Also, predict the result of the data set. As we expected, predict data was the same as the actual data (Figure7).

12 Figure 10 d) Give an idea of the prescription

In the handwriting recognition of the prescription part, get an example to give the result.

Selected a random word as 'word'. Matched the word between a sample image and lexicon entry 'word' (Figure 8).

13 Figure 11

In the first part, it shows segment point of the image. There are 9 points. Second part is the confidence of the match word. Final part is matching the paths and confidences. For this result will be as the expected one. classification. We will perform more experiments to increase the performance of these predictive classifiers for heart disease prediction by using others feature selection algorithms and optimization techniques. If someone follow the heart disease prediction, you all can use different data set and can be use other algorithms for classification. If researchers can implement hybrid model using many algorithms. We think it is also new era of this heart disease prediction model.

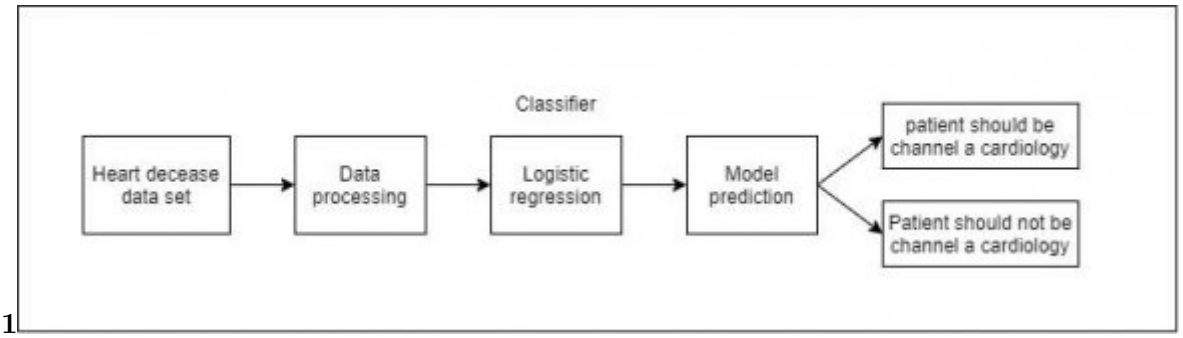
In cholesterol level prediction section, we try to implement a model of predict cholesterol level for about six months for future. We train and test model for predict cholesterol level using given dataset and get prediction line and confidence area.

Researchers can develop this model for other diseases and they can try to develop this system using other algorithms and techniques. If someone trying to follow cholesterol level prediction, you can try to get a

14 Conclusion

In this research, we try to implement a model for predict heart disease; predict cholesterol and diabetic levels for best meal plan using machine learning algorithm. We train and test model for using given data set. In predict heart disease, part, its accuracy score is 87% up to now. For that, we used logistic regression for classification. In predict cholesterol and diabetic levels for best meal plan part, its accuracy score is 90%. For that, we used Decision tree for classification.

Researchers can increase accuracy level of the model. However, there are numbers of algorithms to very smooth line using another way for stationarity. If researchers implement this using more algorithm, that also a new thing for cholesterol prediction model.



1

Figure 1: Figure 1 Logistic

Logistic regression sigmoid function can be written as follows:

$$h\theta\,(x) = g\left(\theta^T X\right),$$

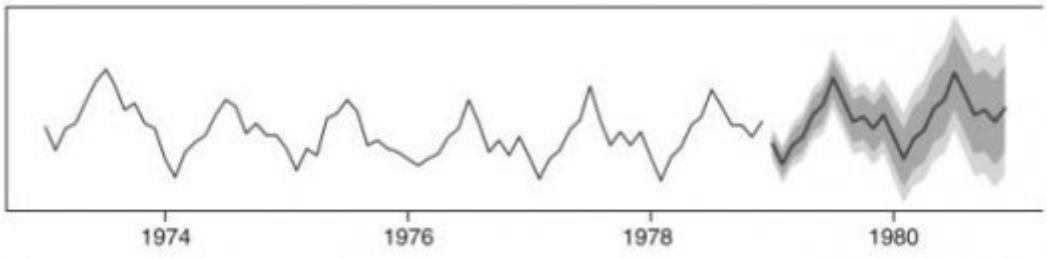
where $g(z) = 1/(1 + x^{-z})$ and $h\theta(x) = 1/(1 + x^{-z})$.

Similarly, the logistic regression cost function can be written as follows:

$$J\,(\theta) = \frac{1}{m}\sum_{i=1}^{m}\text{cost}\left(h\theta\left(x^{(i)}\right), y^{(i)}\right).$$

4

Figure 2: Figure 4 :



5

Figure 3: Figure 5 :

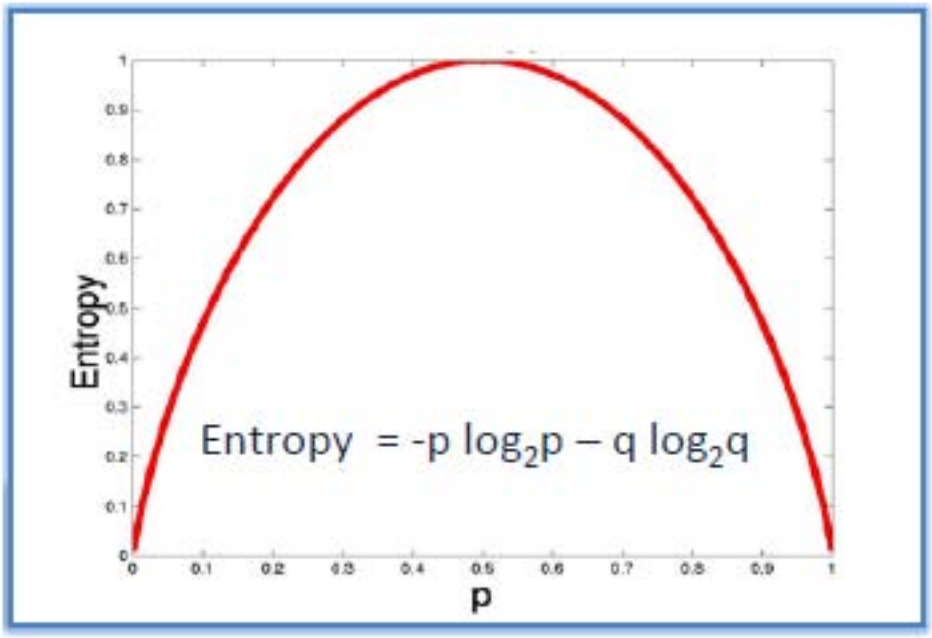


Figure 4:

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

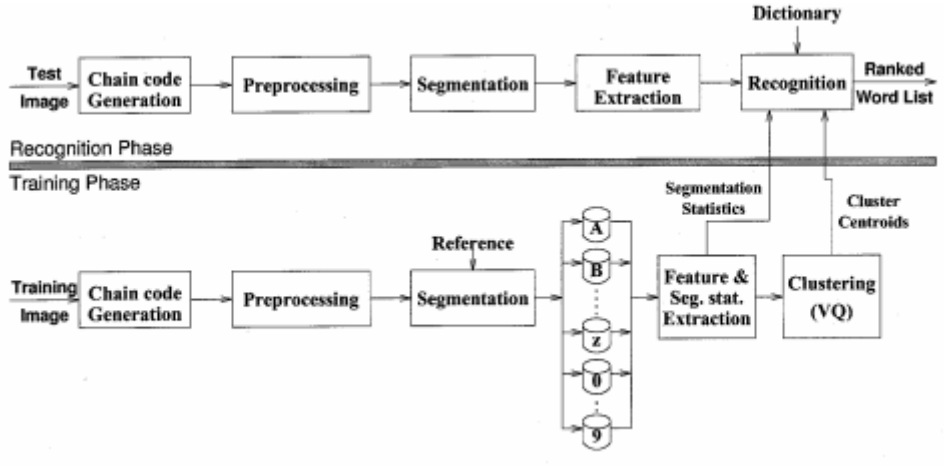
7

Figure 5: Figure 7 I

$$E(T, X) = \sum_{c \in X} P(c) E(c)$$

8

Figure 6: Figure 8 b



9

Figure 7: Figure 9 c

```
In [24]: print('Accuracy Score : ' + str(accuracy_score(y_test,y_pred)))
         print('Precision Score : ' + str(precision_score(y_test,y_pred)))
         print('Recall Score : ' + str(recall_score(y_test,y_pred)))
         print('F1 Score : ' + str(f1_score(y_test,y_pred)))
         from sklearn.metrics import confusion_matrix
         print('Confusion Matrix : \n' + str(confusion_matrix(y_test,y_pred)))

Accuracy Score : 0.8702830188679245
Precision Score : 0.8
Recall Score : 0.10084033613445378
F1 Score : 0.17910447761194032
Confusion Matrix :
[[726   3]
 [107  12]]
```

Figure 8:

 10.34257/GJCSTHVOL21IS1PG23

In prescription reading via image, processing is a big challenge for us. However, using Lexicon Driven Word Recognizer Algorithm, it simplifies the model work. Use of variable duration in word recognition process improved performance. [1]

[1]( ) H © 2021 Global Journals Year 2021

.1 Acknowledgement

The Sri Lanka Institute of Information Technology supported for this work.

[Pdfs and Org ()] *A Lexicon Driven Approach to Handwritten Word Recognition for Real-Time Applications*, Semanticscholar Pdfs , Org . https://pdfs.semanticscholar.org/9dde/54b4c73b866cde5bdc013ba0de8cf72ee003.pdf 1997.

[Shanika ()] *Adverse drug reactions and associated factors in a cohort of Sri Lankan patient with non-communicable chronic diseases*, N , Wijekoonand L Shanika . https://www.researchgate.net/publication/307545534_Adverse_Drug_reactions_and_associated_factors_in_a_cohort_of_Sri_Lankan_patients_with_non-communicable_chronic_diseases 2016. (ResearchGate.net)

[Brownlee ()] 'Available at: https://machinele arningmastery.com/parametric-and-nonparametr icmachine-learningalgorithms/#:~:text=under lying% 20mapping%20function'. J Brownlee . *Machine Learning Mastery*, 2016. Nonparametric%20Mach ine%20Learning%20Algorithms. 20 p. . (Parametric and Nonparametric Machine Learning Algorithms)

[Saedsayad] 'Decision Tree'. Saedsayad . https://www.saedsayad.com/decision_tree.htm *#:~:text=Decision%20Tree%20%2D%20Cl assification*, (decision%20nodes%20and%20leaf%20 nodes)

[Chakure ()] *Decision Tree Classification*, A Chakure . https://towardsdatascience.com/decision-tree-classification-de64fc4d5aac 2019.

[Kaggle and Com (2020)] *Framingham Heart Study Dataset*, Kaggle , Com . <https://www.kaggle.com/amanajmera1/framingham-heart-study-dataset> 2020. July 2020.

[Silva and Leong ()] *Grammar-Based Feature Generation for Time-Series Prediction*, A M Silva , P H W Leong . 2015. Berlin, Germany: Springer.

[Handbook Of Latent Semantic Analysis ()] *Handbook Of Latent Semantic Analysis*, https://books.google.lk/books?hl=en&lr=&id=JbzCzPvzpmQC&oi=fnd&pg=PP1&dq=Latent+Semantic+Analysis&ots=aNO3G2P0HE&sig=aQIXfzinenum39EpoSmmEwc5R-s&redir_esc=y#v=onepage&q=Latent%20Semantic%20Analysis&f=false 2011. (Google Books)

[Heart disease -Symptoms and causes Mayo Clinic staff ()] 'Heart disease -Symptoms and causes'. https://www.mayoclinic.org/diseases-conditions/heart-disease/symptoms-causes/syc-20353118 *Mayo Clinic staff* 2018. (Mayo Clinic)

[Davis and Wedro ()] 'What Is Cholesterol? HDL and LDL Ranges and Diet'. R , Charles Patrick Davis , Benjamin Wedro . https://www.medicinenet.com/cholesterol_management/article.htm#why_is_high_cholesterol_dangerous *Medicine Net* 2016.