IJIEMR Transactions, online available on 5th Jan 2023. Link :http://www.ijiemr.org/downloads.php?vol=Volume-12&issue=Issue 01

DOI: 10.48047/IJIEMR/V12/ISSUE 01/26

Title **Multi Disease Prediction Using Logistic Regression**

Volume 12, ISSUE 01, Pages: 267-275

Paper Authors

**P. Silpa Chaitanya, S. Hema Spandana, A. Maneesha, A. Saranya,
N. Chamundeswari**

Multi Disease Prediction Using Logistic Regression

P. Silpa Chaitanya*[1], S. Hema Spandana[2], A. Maneesha[3], A. Saranya[4], N. Chamundeswari[5]

[1]Assistant Professor, Department of Computer Science and Engineering, Vignan's Nirula Institute of Technologyand Science for Women
Email: Silpam86@gmail.com
[2,3,4,5] UG Student, Department of Computer Science and Engineering, Vignan's Nirula Institute of Technologyand Science for Women

Abstract

Tumours are the most common type of cancer in the human body, and they are the result of abnormal cell division. As a group, diabetes is a group of disorders in which the body either does not produce or does not correctly use insulin. In today's time, we have seen shortage of doctors in the world especially in India. Lot of people are suffering a lot without proper health care and they are not able to afford more amount of money in hospital. Most of the deaths are because of lack of timely medical check-up. In our paper, by taking some parameter value from the user and predict whether the person is suffering from cancer, diabetes, heart, liver, malaria, kidney and pneumonia. The parameters are age, sex (1 is taken if the person is male and 0 is taken if the person is female), chest pain type, Trestbps, serum cholesterol, Restecg, Thalach, Exang, Oldpeak, Slpe, Thal. With the help of the parameters, we can predict all the seven diseases. We utilise logistic regression to make illness predictions. A procedure called logistic regression is used to forecast the likelihood of a target variable. In the case of a goal or dependent variable, there are only two conceivable classes of outcomes. Binary data is used to represent the dependent variable, which is represented as either 1 (success/yes), or 0 (failure/no). Based on test results, our suggested model outperforms currently used categorization methods. We employ the metrics accuracy, precession, f-score, and recall to assess the suggested model's performance.

Introduction

Cancer is a dangerous disease that occurs when cells begin to divide improperly and uncontrollably. With a few significant exceptions, such as leukaemia, most cancer cells are found in the form of tumours. Tumours aren't always cancerous. A range of symptoms can be used to diagnose cancer. Symptoms may vary or not appear at all in other circumstances. There are a few common signs and symptoms: Irregular weight loss, recurring fever, and other symptoms. Under the umbrella of diabetes, there are a number of conditions in which the body either does not generate or does not use insulin properly. In the event of any of the following, sugar cannot be taken by cells. Blood sugar levels rise as a result of this. There are several disorders that may damage the heart together referred to as "heart disease." Coronary artery disease (CAD) and heart rhythm disturbances (arrhythmias) are both examples of cardiovascular disease (congenital heart defects). Heart disease is characterised by the build-up of fatty plaques in the arteries. In order to keep you healthy, the liver performs a variety of tasks. Your body uses it to convert nutrients into the chemicals it needs. Toxins are removed by a filtration process. Assists in the process of converting food into fuel. You may notice a decrease in your overall health when your liver isn't working as it should be. There are several ways in which infections may result in severe liver damage: Your liver may become inflamed because of an infection. For the most part, it's the result of viral hepatitis like Hepatitis A, B, or C as well as immune system problems. Kidney disease has rendered them unable to perform their vital function of removing waste from blood. Renal disease is more common in those with diabetes or high blood pressure. Swelling, dizziness, and faintness are all signs of a concussion. Using logistic regression, one may calculate the probability of a discrete result based on an input variable. Logistic regression is a statistical approach. Logistic regression models often use binary outcomes like "true" or "false," "yes" or "no." A logistic regression equation in statistical software accepts inputs from the equation and uses them to determine the relationship between a dependent variable and one or more predictor variables. Predicting the likelihood of an event or a choice happening may be made easier with the use of this kind of study. Logistic regression has emerged as a key tool in the area of machine learning. Using historical data, machine learning systems can categorise input messages.

Algorithms get better at anticipating categorization within data sets as more relevant elements are added. As part of the extract, transform, and load (ETL) process, logistic regression may aid with data preparation by enabling data sets to beput in certain buckets. This allows the information to be staged for analysis. A basic arithmetic problem is all that is required for complex probability calculations to be simplified via logistic regression. Despite the complexity of the calculation, modern statistical software has made most of the laborious work unnecessary. There are now many fewer affecting factors, making it easier to analyse the impacts of numerous variables. Thereby, mathematicians may quickly model and examine how several variables affect one particular result. Increasing the chance of an occurrence by expanding one of the predictor variables in a logistic model at a constant rate, with each independent variable having its own parameter, generalises the odds for a binary dependent variable. In addition to the logistic model, similar models with unique sigmoid activation functions such as the probit model may also be utilised. By modifying the raw data streams, logical models may be utilised to generate characteristics for other machine learning and artificial intelligence methods. When it comes to classification tasks, logistic regression is one of the most commonly used machine learning methods. This is due to the fact that logistic regression can be used to predict "yes or no" and "A or B" answers.

Literature Survey

This study article by Joseph A.Cruz and David S.Wishart on the use of machine learning in cancer prediction and prognosis was released in 2006. In this study, scientists looked at a wide range of machine learning approaches, the sorts of data they were using, and how well they predicted and predicted cancer outcomes. It performs better when evaluated on a dataset of chest radiographs depicting possible causes of pneumonia.

In 2021, Gresha Bhatia, Shravan Bhat, Vivek Choudhary, Aditya Deopurkar, and Sahil Talreja developed "Disease Prediction Using Deep Learning [2]".

Mosquito-borne infections may be predicted using time series analysis. Public health interventions may be more effective if they can be planned and implemented early on in the disease's transmission.

An author, Neha Sharma, released "Diabetes Detection and Prediction Using Machine Learning/IoT[3]" in 2018. Concern has grown about an alarming rise in the number of diabetes patients. A framework for storing and analysing diabetes data and spotting possible problems must be built using innovation.

Senthil Kumar Mohan, Chandrasekar Thirumalai, Gautam Srivastava wrote "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques[4]" in 2009. In an effort to increase the accuracy of cardiovascular disease prediction, he came up with a unique way to discover critical traits using machine learning. Using a hybrid random forest with a linear model, the prediction model for heart disease has an accuracy level of 88.7 percent.

C.Geetha, Dr.AR.Arunachalam will present "Evaluation based Approaches for Liver Disease Prediction using Machine Learning Algorithms [5]" in 2021. 96 percent of the time, the chance of liver illness is accurately predicted. Using liver records, this research aimed to create algorithms for identifying healthy individuals. It is the goal of this research to compare categorization algorithms and make predictions based on their success characteristics.

According to Prof.K. JayaMalini and Priyank Sonar in 2019, "Diabetes prediction using multiple machine learning algorithms" was published. Research in this area is focused on developing an algorithm that can better predict a patient's probability of developing diabetes. Decision Tree, Naive Bayes and SVM algorithms are used to categorise the data.

Using machine learning, Akash Maurya and Rahul Wable created "Chronic Kidney Disease Prediction and Recommendation of Suitable Diet plan [7]" in 2019. It was published in 2019.

For Individuals with chronic disease (CKD), a machine learning system may be used to recommend a healthy diet plan based on medical test results.

"Identifying pneumonia in chest X-rays[8]" was adopted in 2019 by Amit Kumar Jaiswal, Prayag Tiwari, Sachin Kumar, Deepak Gupta, Ashish Khanna and Joel J.P.C. Rodrigues. On the basis of this, we were able to locate pneumonia in chest X-ray pictures using a deep learning method.

When Octave Iradukunda and Haiying Che compared ELM to SVM and DENSENET, they found that it outperformed them all with 99 percent accuracy and 98 percent precision. This demonstrated the effectiveness of ELM in the application of malaria cell detection scenarios, which allowed it to be referred to by other researchers working in the field of machine learning in the future.

Detection Using Machine Learning [10]" in 2021. Feature extraction and machine learning methods are used in this study. This approach has a 98.30 percent success rate in eliminating cancer diagnoses.

Senthil Kumar K, Kavethanjali V, Preethi S, and Vasanthapriya V published a paper titled "Lung - Pleura Carcinoma Detection Using Machine Learning [10]" in 2021. Feature extraction and machine learning methods are used in this study. This approach has a 98.30 percent success rate in eliminating cancer diagnoses.

Proposed System

This mainly works on taking data sets from various sources ,processing and finally depicting the output that whether a person is healthy or not.

The first step is collecting data sets. Data set consists of many parameters these can be noisy data, missing values and unwanted features. The missing values can be removed by several methods like replacing with arbitrary values, replacing with mode value, median. These can also be done by replacing with previous or next value depending on the situation. Data processing methods such as data cleaning, data transformation, and data reduction may be used to reduce noise from data. It is capable of resolving discrepancies in data. Despite the fact that data sets include a wide range of characteristics, we must extract the most significant ones during training. Columns need to be extracted. Square brackets are used to extract the columns. Like the output feature, label encoding is categorical. Encoding is the first stage in changing data into a format that may be used for a range of tasks, including: compiling and running the programme, compressing and decompressing data, and transferring and storing data. An application data processing example is the conversion of a file. Separating training and testing information. Logistic regression may be fed pre- processed data after pre-processing is complete. Machine learning method that produces categorical results is the Logistic Regression.

We can forecast whether the output is 0 or 1 by using prediction. If the individual's score is 0 and the score is 1, then the individual is healthy. The sigmoid function produces an output of zero or one.

Algorithm

Input: Dataset

Output: Predicting whether data feeded is positive or not.

Step 1: To predict the disease we need to train the data. Dataset is taken as a input.

Step 2: Extracting the useful columns and scaling and splitting the data.

Step 3: Feeding the data to the logistic regression algorithm.

Z=mx+c

1. Y=sigmoid(x)
2. $\boldsymbol{y} = \frac{\mathbf{1}}{\mathbf{1}+e^{Nx+c}}$
3. But we have multiple feature so
4. $\boldsymbol{Z = m_1x_1 + m_2x_2 + \ldots + m_nx_n + C}$
5. $\boldsymbol{y} = \frac{\mathbf{1}}{\mathbf{1}+e^{(N_1s_1+N_2s_2+..+N_ns_n+C)}}$
6. $\boldsymbol{y} \quad = \begin{cases} \mathbf{0}\ \boldsymbol{represents\ don't\ have\ disease} \\ \mathbf{1}\ \boldsymbol{represents\ suffers\ from\ disease} \end{cases}$
7. Step 4**:** if y=0 it represents then the patient is not suffering from disease,if y=1 it represents then patient is suffering from disease.
8. Step 5: Calculating the accuracy
9. Step 6: Saving the model using joblib

Data sets are collected from various sources from kaggle. Datasets consists of noisy data and missing values and unwanted features. At first step of pre-

processing the data consists of Extraction of useful features and handling missing values. Datasets consists of various features but while training we need to give important feature through feature extraction. The output feature is categorical, so do label encoding. Splitting the data into training and testing.Logistic regression makes use of the pre-processed data. A supervised machine learning technique, Logistic Regression produces categorical output. The input feature is feed into algorithm.

Z= MX+c

Y=Sigmoid(Z)

Here X1, X2,...Xn are the independent features .

It is the dependent traits that rely on the independent ones.

A value of 0 or 1 is possible. The 0 indicates that the individual is healthy, whereas the 1 indicates that the person is ill. Either 0 or 1 is the result of applying the sigmoid function. After the model has built, we need calculate the accuracy. High accuracy of model will give greater results. Predicting the data using model. Saving the model using the joblib. This joblib saves the model and can be used predict the data without training lot many times.

Results and Discussions:

Dataset: From kaggle, data sets are

gathered from a variety of sources. The Datasets here are made up of noisy data, missing values, and undesirable features. Extraction of valuable characteristics and handling missing values are the first steps in preparing data. Each illness requires a distinct sort of data collection to be used in its training. For the gathering of liver patient data, Andhra Pradesh provided 416 liver patient files and 167 non-liver patient files. The "Dataset" field is used as a class label to separate patients with liver illness from those who do not (no disease). There are 441 records for men and 142 records for women in this data set, so we can determine who has liver disease andwho does not have it.

Accuracy:

Ratio between observed data and anticipated data is how accuracy is used.
Accuracy=(a+b)/(a+c+d+b)
Where,
a=Positive True, b=Negative True,
c=Positive False, d=Negative False

When predicting an illness using logistic regression and SVM Model, this graph indicates the accuracy of the prediction. A probabilistic approach of class predictions is a logical extension of logistic regression (multinomial regression). As a result, the logistic regression has a higher degree of accuracy. For huge datasets, the SVM method does not work well at all. Noise in the data set (i.e. the target classes overlap) affects SVM's ability to function.

Consequently, the SVM model is less accurate than the logistic regression model.

Precision:

The specific number of positive observations compared to the overall number of positive observations is what precisionis.

Precision =a/(a+c)

SVM and logistic regression models were used to predict an illness with high accuracy. Model coefficients may be interpreted as indications of feature

relevance using logistic regression. As a result, the illness prediction accuracy is improved. Having too many features per data point in an SVM model may lead to poor SVM performance since there is no statistical rationale for categorising data points above and below the classifying hyperplane, as is the case in most SVM models. As a result, as compared to logistic regression, the precision for illness prediction is lower.

Recall:

It's the proportion of precisely expected positive observations to all of the actual class observations.

Recall = a/(a+d)

Using logistic regression and an SVM model, the recall of a disease prediction is shown in the image. A probabilistic approach of class predictions is a logical extension of logistic regression (multinomial regression). As a result, the logistic regression has a higher recall. For huge datasets, the SVM method does not work well at all. Noise in the data set (i.e. the target classes overlap) affects SVM's ability to function. As a result, the SVM model's recall is lower than that of a logistic regression model.

F-Score:

Precision and Recall are the two weights that make up the average.

F-Score = 2*[e * f] / [e + f]

Where e=Recall, f=Precision

Using logistic regression and the SVM model, the F-Score is shown in the figure. Model coefficients may be interpreted as indications of feature relevance using logistic regression. Because of this, the illness prediction F-Score has increased. For an SVM, the number of features for each data point must be less than the number of training samples in order for it to perform well. There is no probabilistic explanation for classification since the support vector classifier operates by placing data points above and below the classifying hyper plane. A lower F-Score is a result of this, as opposed to a higher one in logistic regression.

Conclusion

Cancer is a dangerous disease that occurs when cells begin to divide improperly and uncontrollably and most cancer cells are found in the form of tumour. Under the umbrella of diabetes, there are a number of conditions in which the body either does not generate or does not use insulin properly. When we need to see a doctor these days, it costs more money. The impoverished are unable to afford this. Malaria, liver, pneumonia, heart disease, diabetes, cancer, and renal disease may all be detected with this method. The public will greatly benefit from this. We utilised the logistic regression approach for this task. You may utilise the provided independent variables to predict a categorical dependent variable. Logistic regression may be used to predict the outcome of a categorical dependent variable. There are many methods, which can identify the diseases. But they identify separately. Our method is very accurate and cost friendly. The accuracy is around 95%.

Our model is good in handling imbalanced data also.

References

Kourou, Konstantina, et al. "Machine learning applications in cancer prognosis and prediction." Computational and structural biotechnology journal 13 (2015): 8-17.

Imran, A., Posokhova, I., Qureshi, H. N., Masood, U., Riaz, M. S., Ali, K., ... & Nabeel, M. (2020). AI4COVID-19: AI enabled preliminary diagnosis for COVID-19 from cough samples via an app. Informatics in Medicine Unlocked, 20, 100378.

Chaki, Jyotismita, S. Thillai Ganesh, S. K. Cidham, and S. Ananda Theertan. "Machine learning and artificial intelligence based diabetes mellitus detection and self-management: a systematic review." Journal of King Saud University-Computer and Information Sciences (2020).

Reddy, G.T., Reddy, M., Lakshmanna, K., Rajput, D.S., Kaluri, R. and Srivastava, G., 2020. Hybrid genetic algorithm and a fuzzy logic classifier for heart disease diagnosis. Evolutionary Intelligence, 13(2), pp.185-196.

Geetha, C. and Arunachalam, A.R., 2021, January. Evaluation based Approaches for Liver Disease Prediction using Machine Learning Algorithms. In 2021 International Conference on Computer Communication and Informatics (ICCCI) (pp. 1-4). IEEE.

Sonar, Priyanka, and K. JayaMalini. "Diabetes prediction using different machine learning approaches." 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC). IEEE, 2019.

Maurya, A., Wable, R., Shinde, R., John, S., Jadhav, R. and Dakshayani, R., 2019, January. Chronic kidney disease prediction and recommendation of suitable diet plan by using machine learning. In 2019 International Conference on Nascent Technologies in Engineering (ICNTE) (pp. 1-4). IEEE

Maurya, A., Wable, R., Shinde, R., John, S., Jadhav, R. and Dakshayani, R., 2019, January. Chronic kidney disease prediction and recommendation of suitable diet plan by using machine learning. In 2019 International Conference on Nascent Technologies in Engineering (ICNTE) (pp. 1-4). IEEE

Madhu, Golla, and A. Govardhan. "Artificial Intelligence Based Diagnostic Model for the Detection of Malaria Parasites from Microscopic Blood Images." Intelligent Interactive Multimedia Systems for e-Healthcare Applications. Springer, Singapore, 2022.215-233.

Kavethanjali, V., S. Preethi, and V. Vasanthapriya. "Lung–Pleura Carcinoma Detection Using Machine Learning." 2021 3rd International Conference on Signal Processing and Communication (ICPSC). IEEE, 2021.