Original Article

Early Disease Diagnosis Using Multivariate Linear Regression

Arul Natarajan[1],Panthagani vijayababu[2], M.Arsha [3],Sasibhushana RaoPappu[4],Vidya Rajasekaran[5]

St. Peter's Engineering College (SPEC), Hyderabad [1, 3], India
Vignan's Foundation for Science, Technology and Research, Andhra Pradesh[2], India
GITAM school of Technology, GITAM(Deemed to be university),Visakhapatnam[4], India
B.S.Abdur Rahman Crescent Institute of Science and Technology[5], India

arulthala82@gmail.com[1], panthagani.vijay@gmail.com [2], arsha.reddy@gmail.com[3],
sasipappu510@gmail.com[4], vidyarajesh23@gmail.com[5]
DOI: 10.47750/pnr.2023.14.S02.168

Abstract

The world population is rapidly increasing; people are prone to more diseases due to their food habits and lifestyle changes. The proper diagnosis of the disease at the initial stages will save the lives of millions. So this kind of disease prediction system will help in predicting the disease in the initial stages using their symptoms. We study the historical dataset of the patients with the symptoms and their diseases. We apply data analytics skills to extract hidden patterns from the data. We calculate the number of symptoms contributing to the disease and the weightage factor for every contributing symptom. The weightage indicates the score value contributing to the disease. The ultimate goal of this research is to develop a high-end prediction prototype for disease diagnosis with improved accuracy and efficiency. We implemented a Multivariate Linear Regression algorithm using python to predict the diseases. The model is evaluated using metrics like R2, MSE, and RMSE, and the Multivariate Linear Regression algorithm is found to be the best fit with an accuracy of 95%.

Keywords: Machine Learning; Predictive model; Data analytics; Data Preprocessing.

1. Introduction

Machine learning is continuously evolving and has been adopted in most fields. The applications of machine learning techniques in the medical and healthcare industry benefit society by making preventive measures for diseases at the initial stage based on the symptoms. Many machine learning algorithms can be applied to the medical dataset for making effective diagnoses and predictions [1]. In our research, we apply multivariate linear regression. In addition to making predictions, we find the weightage of the symptoms contributing to the disease. It is very difficult to accurately predict any disease at the initial stage and the prediction model should be developed with the very most attention to avoid misleads. Identifying the disease at a very early stage can lead to a complete cure and save lives in case of deadly diseases. Several systems predict generalized diseases or predict any particular disease [2] [4] [12]. Our work includes the weightage score calculation that is measured along with the symptoms to predict the disease. The weightage score denotes how strongly the symptom contributes to the disease. We work on a database for the prediction of generalized disease which may also turn into a deadly disease if left untreated at the initial stage. The proposed system is developed in python and multivariate linear regression is applied to make predictions. The research is illustrated in five sections. The preliminary section denotes a detailed outline of the research and its needs. The

second section explains the literature study of the relevant existing works. The third section illustrates the main effects model and its implementations. The fourth section explains the obtained results and their performance metrics. At last, we conclude the final part of the research with the conclusion and the future scope of the research.

2. Related Work

There are already several related works in disease prediction using machine learning [2] [11] for predicting diseases based on symptoms.

The research work in [3] has applied big data analytic techniques for medical data. Machine learning algorithms like Naïve Bayes, Decision Tree, and Random Forest are used in disease prediction using the symptoms. They have not taken steps to find the relationship among the variables or symptoms.

In [4] the machine learning techniques were applied to predict heart diseases in specific, based on the Indian population. The study stated that it worked efficiently in diagnosing heart disease at the early stage and can be used as a tool to screen the heart disease at the very early stage. In [9] [10] Similar heart-related research for developing predictions is carried out. Likewise, our model has been developed for predicting the generalized disease at the very early stage.

Here in [5], a classifier system has been developed using machine learning algorithms that can be used for predicting diseases at the early stage. Different classification algorithms were used for making predictions. In our work, we applied multivariate linear regression to find the association between the dependent and independent variables.

In our research, we use multiple over multivariate linear regression because two dependent variables are considered in our research. The disease type and the weightage score recommend the disease.

The work [6], has integrated several machine-learning algorithms for generic disease prediction. 97.32% of accuracy is obtained using the KNN algorithm which is the highest compared to other algorithms. This model has only one dependent variable and in our proposed model we have two dependent variables.

The work [7] is a similar disease prediction system based on the symptoms and other factors like age and gender. They have applied a weighted KNN algorithm and the resulting accuracy is 93.5% which is comparatively low compared to our prediction model with an accuracy of 95% using multivariate analysis.

The author [8] has designed an automated disease diagnosis model using a machine learning technique. The author has focused on specific diseases such as covid-19, diabetes, and heart-related diseases. The developed android app detected and resulted from the disease results. Logistic Regression was applied for prediction. So timely treatment can be carried out. This work is specific to a few types of disease and other early diagnoses are not possible.

3. Main Effects Model

We feed the input to our model to make predictions. The overall framework for the disease prediction system is shown in Figure 1. The input is the medical dataset used for disease diagnosis. Then we start with data preprocessing.

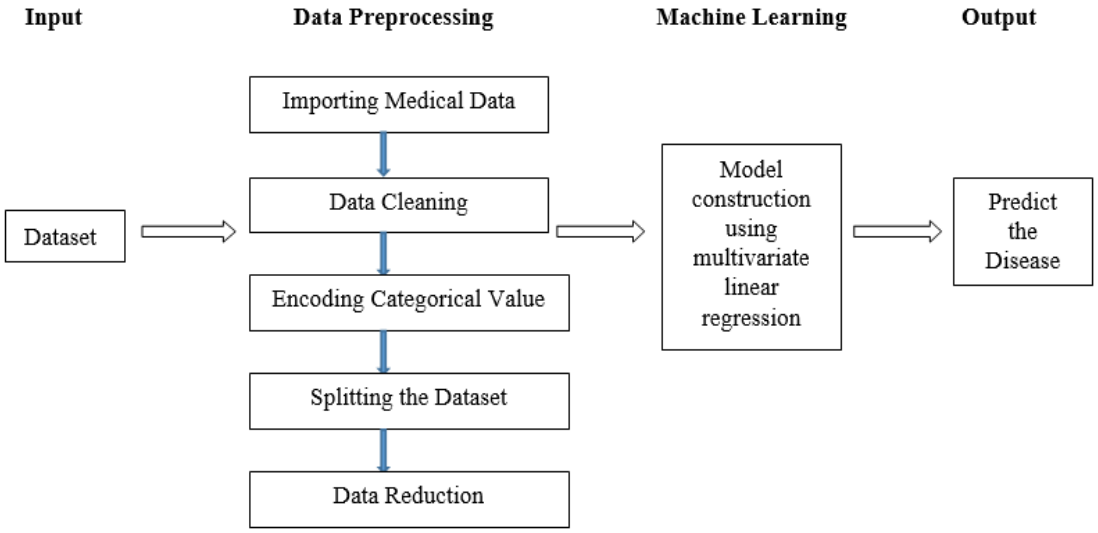


Figure 1. Block diagram of Disease Prediction System - This represents the input, processing, and machine learning algorithm to give the desired output.

We develop three different columns in the dataset representing the positive symptoms examined (α), the total count on the number of symptoms contributing to each disease (β), and finally, the weightage of the symptoms calculated for the particular disease (γ) based on the dataset as represented in Eq. (1),(2) and (3).

$$\alpha = \text{Total No. of symptoms examined} - \text{Total No. of negative symptoms} \quad (1)$$

$$\beta = \frac{\text{Total No. of symptoms contributing to the disease}}{\text{Total No. of symptoms examined}} \quad (2)$$

$$\gamma = \frac{\text{Total No. of symptoms contributing to the specific disease}}{\text{Total No. of symptoms examined for the particular disease}} \quad (3)$$

The below figure represents the dataset distribution of the various diseases along with their weightages. Figure (2) represents the average count on the number of positive symptoms, Figure (3) shows the average of the symptoms contributing to the disease, and figure (4) the weightage of the symptoms contributing to individual diseases. The final Figure (5) represents the chart containing the comparison of the average of the individual symptoms and their weightage contributing to a particular disease.

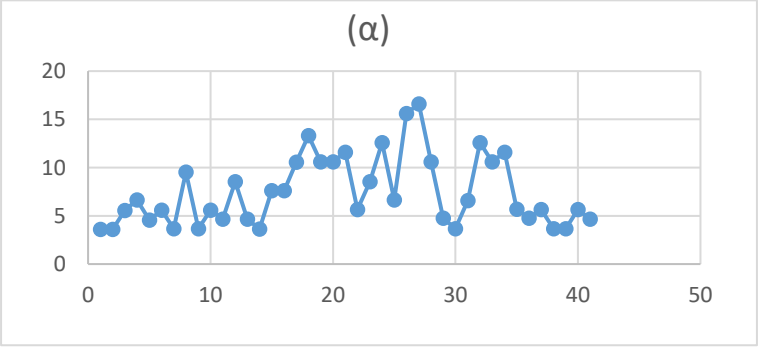


Figure 2. Average of positive symptoms – average on the total number of positive symptoms observed from the patients

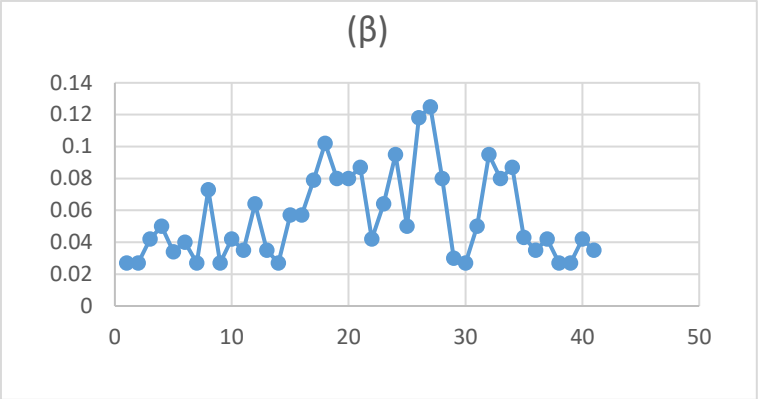


Figure 3. Average of the number of symptoms contributing to the disease – average on the total number of positive symptoms contributing a particular disease from the total number of symptoms

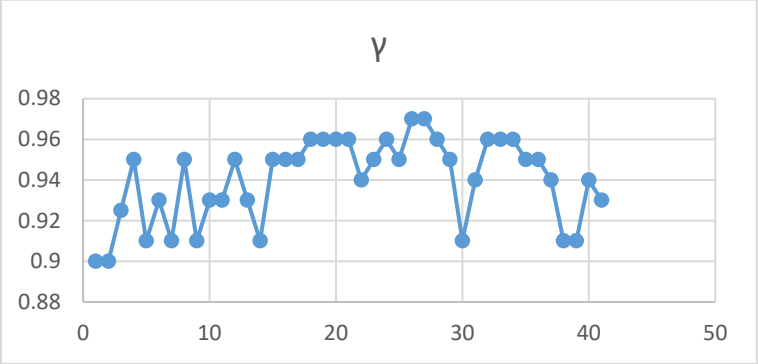


Figure 4. The weightage of symptoms contributing to individual disease (γ) – how much the particular symptom contributes to the disease

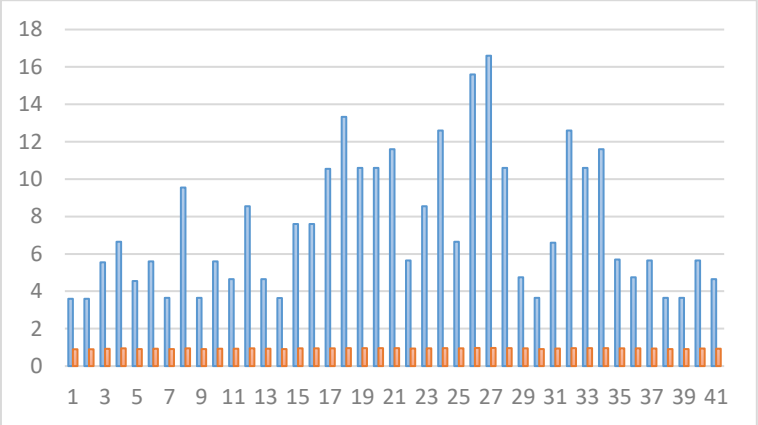


Figure 5. Comparison on individual disease and their weightage – diagnosed from the previously observed database the disease and their symptoms weightage making contribution

Then the dataset is allowed to be preprocessed. The very first step in data preprocessing is to import the medical data set [13]. This step involves extracting the dependent variable (y) and independent variables (x1, x2...xn). In our work disease prognosis column is the dependent variable and all other symptoms, the number of symptoms contributing to the disease, and the weightage factor for every contributing symptom are the independent variables. The “iloc [ ]”

function in python extracts the dependent and the independent variables. Next, we move on to data cleaning, which comprises the task of finding and removing or replacing inaccurate and incomplete data from the dataset. Once the data cleaning process is over we check the format of the data type. If the data are in a numerical format no issues occur, since machine learning models are primarily based on mathematical equations. In our work, the prognosis column is represented as textual data. So we convert it into numerical values.

After encoding the categorical variable we split the dataset into two different sets, training and test set. The training set represents the subset of the dataset that is used for training the model and the test set remaining part of the dataset that is used for testing the learning model and predicting the disease. We split our dataset into a 70:30 ratio, which means that 70% of the data is used in training the model and the left 30% of data is used for testing.

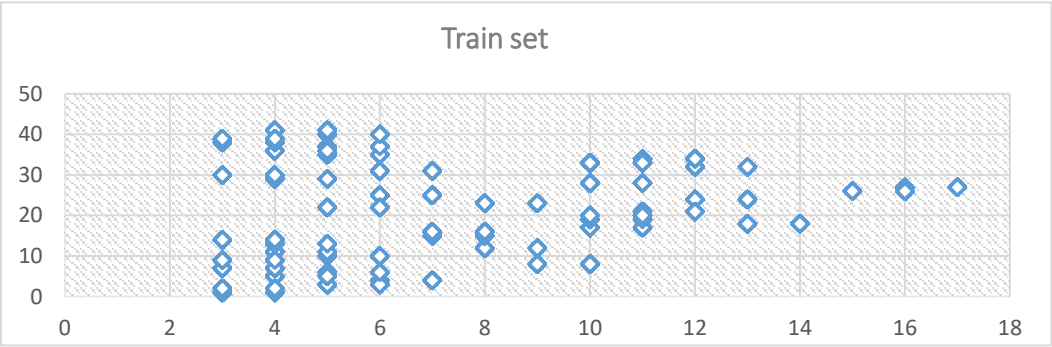
The next stage of preprocessing is data reduction. In our research, there are multiple independent variables representing the symptoms of the disease, and not all the symptoms contribute to the disease. So the Columns with no importance are removed from the dataset and a new dataset is formed. Data reduction will improve the overall efficiency of the predictive model. The data reduction process is carried out using the dimensionality reduction technique. Dimensionality reduction eliminates the attributes from the data set under consideration thereby reducing the volume of original data. There are three categories of dimensionality reduction techniques namely wavelet transformation, principal component analysis, and attribute subset selection. We apply an attribute subset selection method in our prototype. In the dataset used, there are several attributes, and the unwanted ones are eliminated and the data volume is reduced. The attribute subset selection method confirms that even after eliminating the unwanted attributes the resulting probability of data distribution is as close as possible to the original data distribution using all the attributes.

In machine learning, we develop the model using multivariate linear regression which is a supervised learning technique and is used to predict diseases. In the case of a single input variable, we can use Univariate linear regression. In our research, we have multiple input variables representing the symptoms of diseases and hence we choose multivariate linear regression to build the predictive model. The hypothesis function of our predictive model is defined using multiple variables. The multivariate linear regression algorithm produces the output value which is dependent on multiple input values. The process of creating and predicting a prototype using multivariate linear regression depends on the relationship between multiple input values, the different ranges of inputs, and their formats.

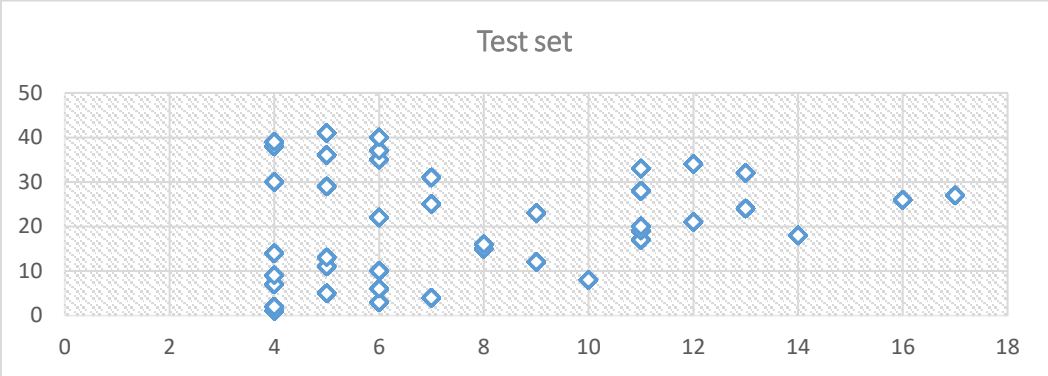
The hypothesis function of the multiple input values can be denoted as shown in Eq. (4).

$$h(x) = \theta_0 + \theta_1 * x_1 + \theta_2 * x_2 \ldots \theta_n * x_n \qquad (4)$$

Herex_1, $x_2 \ldots x_n$ are the multiple input values. In our work we consider the different symptoms, the count on the number of symptoms for disease prediction and, the weightage of the symptoms contributing to the disease are the input variables for the above-stated hypothesis Eq. (4). The distribution of the disease symptom and the disease type is represented in the below figure stated as Figure 6(a) and (b).



(a)



(b)

Figure 6. (a) Train set (b) Test set - the sample distribution of symptom and disease

4. Results and Discussion

The general evaluation of the prediction model can be carried out using the cost function which is represented as shown in (5).

$$R(\theta) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta x^i - y^i)^2 \qquad (5)$$

Here $R(\theta)$ represents the cost function. The best possible value can be chosen based on the fit of the line. The more sophisticated form of the cost function is called the mean squared error function which is used for evaluating our results. In Table 1 the prediction performance of the main effects model trained on 70% of the data and tested on 30% of the data is illustrated. The R^2 value represents the statistical measure calculating the proportion of variance for any dependent variable that is explained by an independent variable.

The normal range of R^2 lies in the range between 0 and 1. '0' indicates the worst fit and '1' indicates the best fit. We observe $R^2 = 0.95$ for the training set and $R^2 = 1$ for the test set. The test set denotes that the predictions are identical to the observed values. The higher the R^2 value, the better the model fits our data.

Table 1. Prediction performance

Variable	134
Train set	3444
R2	0.95
MAE	1.59
MSE	5.43
RMSE	2.33
Test set	1476
R2	1
MAE	3.75
MSE	2.12
RMSE	4.61

The MAE (Mean Absolute Error), MSE (Mean Squared Error), and RMSE (Root Mean Squared Error) denote the error metrics for the regression problem that measures the amount of error in statically models which are represented in Table 1 for the train set and the test set.

The final model will take the symptoms as input and the disease will be predicted as the output result based on the input fed. The initial input will be in textual format and the variable size is 134. The symptoms which are positive are marked as 1 and the negative symptoms are marked as 0. We create a dictionary to encode the symptoms into numerical form. The input data is then reshaped and converted to a suitable format for model prediction. We generate the output using the Multivariate Linear Regression model and the final disease prediction can be listed out. The estimated performance in Table 1 shows that the developed model results in high-rooted square metrics which makes the model the best fit.

5. Conclusion

There are several disease prediction systems using machine learning already available. We apply multivariate regression analysis to identify the relationship among the variables in the medical dataset and predict the disease based on several symptoms. So the correlation between dependent and independent variables can be analyzed and a deeper understanding of the disease and its leading symptoms and the weightage based on the most contributing symptom can be studied. Multivariate regression works better for this model and the drawback sometimes occurs where the loss and error outputs are not identical and most often the results are better for larger datasets.

6. References

1. Ramasamy, S., & Nirmala, K. (2017). Disease prediction in data mining using association rule mining and keyword based clustering algorithms. International Journal of Computers and Applications, 1–8. doi:10.1080/1206212x.2017.1396415

2. S. Grampurohit and C. Sagarnal, "Disease Prediction using Machine Learning Algorithms," 2020 International Conference for Emerging Technology (INCET), 2020, pp. 1-7, doi: 10.1109/INCET49848.2020.9154130.

3. Deepthi, Y., Kalyan, K.P., Vyas, M., Radhika, K., Babu, D.K., Krishna Rao, N.V. (2020). Disease Prediction Based on Symptoms Using Machine Learning. In: Sikander, A., Acharjee, D., Chanda, C., Mondal, P., Verma, P. (eds) Energy Systems, Drives and Automations. Lecture Notes in Electrical Engineering, vol 664. Springer, Singapore. https://doi.org/10.1007/978-981-15-5089-8_55

4. Ekta Maini, Bondu Venkateswarlu, Baljeet Maini, Dheeraj Marwaha, Machine learning–based heart disease prediction system for Indian population: An exploratory study done in South India,Medical Journal Armed Forces India, Volume 77, Issue 3, 2021, Pages 302-311, ISSN 0377-1237, https://doi.org/10.1016/j.mjafi.2020.10.013.

5. S. Grampurohit and C. Sagarnal, Disease Prediction using Machine Learning Algorithms, 2020 International Conference for Emerging Technology (INCET), 2020, pp. 1-7, doi: 10.1109/INCET49848.2020.9154130.

6. P. Jha, T. Biswas, U. Sagar and K. Ahuja, Prediction with ML paradigm in Healthcare System, 2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC), 2021, pp. 1334-1342, doi: 10.1109/ICESC51422.2021.9532752.

7. Keniya, Rinkal and Khakharia, Aman and Shah, Vruddhi and Gada, Vrushabh and Manjalkar, Ruchi and Thaker, Tirth and Warang, Mahesh and Mehendale, Ninad and Mehendale, Ninad, Disease Prediction From Various Symptoms Using Machine Learning (July 27, 2020). Available at SSRN: https://ssrn.com/abstract=3661426 or http://dx.doi.org/10.2139/ssrn.3661426

8. Naresh Kumar, Nripendra Narayan Das, Deepali Gupta, Kamali Gupta, Jatin Bindra, "Efficient Automated Disease Diagnosis Using Machine Learning Models", Journal of Healthcare Engineering, vol. 2021, Article ID 9983652, 13 pages, 2021. https://doi.org/10.1155/2021/9983652

9. Y. Khourdifi and M. Bahaj, “Heart disease prediction and classification using machine learning algorithms optimized by particle swarm optimization and ant colony optimization,” International Journal of Intelligent Engineering and Systems, vol. 12, no. 1, pp. 242–252, 2019.

10. S. Mohan, C. Thirumalai and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," in IEEE Access, vol. 7, pp. 81542-81554, 2019, doi: 10.1109/ACCESS.2019.2923707.

11. Chen, Min, Yixue Hao, Kai Hwang, Lu Wang, and Lin Wang. "Disease prediction by machine learning over big data from healthcare communities." Ieee Access 5 (2017): 8869-8879.

12. M. Dhilsath Fathima, S. Justin Samuel, R. Natchadalingam & V. Vijeya Kaveri (2022) Majority voting ensemble feature selection and customized deep neural network for the enhanced clinical decision support system, International Journal of Computers and Applications, DOI: 10.1080/1206212X.2022.2069643

13. Rajasekaran V., Priyadarshini R. (2021), An E-Commerce Prototype for Predicting the Product Return Phenomenon Using Optimization and Regression Techniques. In: Singh M., Tyagi V., Gupta P.K., Flusser J., Ören T., Sonawane V.R. (eds) Advances in Computing and Data Sciences. ICACDS 2021. Communications in Computer and Information Science, vol 1441. Springer, Cham. https://doi.org/10.1007/978-3-030-88244-0_22

14. Disease Prediction Using Machine Learning | Kaggle