ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 8, Special Issue 1, August 2021

PREDICTING THE EFFECTIVENESS OF COVID-19 VACCINES USING MACHINE LEARNING TECHNIQUES

Neha Seirah Biju
Department of Computer Applications
Saintgits College of Applied Sciences
Pathamuttom, Kottayam- 686532
nehasb.bca1922@saintgits.org

Abstract**—***Covid-19 pandemic had unfavorable effects on mankind and it has changed the view of one's life. People had to go through tough times during this period. This pandemic has taken millions of lives, governments all around the world was forced to take grating restrictions on humans to minimize the spread of the virus. This pandemic has affected people in all walks of life, the whole world was fighting over a deadly virus. To rescue came covid-19 vaccines, covid-19 vaccinations is providing a passage to move out of this pandemic, as natural immunity was not sufficient to fight covid-19.*
By the beginning of 2021, many vaccines were given approval and began to roll out worldwide. Immunization reduces the risk of getting diseases. Efficiencies of vaccine differ from one another and from person to person. Therefore, through this paper we try to predict the effectiveness of covid-19 vaccines available in India through data mining and machine learning. This focuses on prediction algorithms that can sort out an optimal result. The optimal classifier must be able to deliver the near results to that of the real-world clinical outcomes. Data Mining classification techniques such as Naive Bayes, Random Forest, Logistic Regression are used to conduct the analysis.

Keywords**—**Logistic Regression, Data mining, Naïve Bayes, Random Forest, Machine Learning.

I. INTRODUCTION

Covid-19 pandemic, that originated from Wuhan (China) had spread like wildfire across various countries which effected the mankind in a disastrous manner. Many countries had faced substantial difficulty in controlling the transmission of the disease. One such country was India. The outbreak of this pandemic had adverse effects on the citizens of India. Multiple strategies were used to bring the situation under control like implementing lockdowns, spreading awareness etc. Covid-19 had affected the day to day life of citizen declining the economy slowly. This pandemic has taken lives of many which caused immense pressure on different countries to take prominent decision to tackle the situation. One such decision was the production of vaccine. To provide vaccination was the only way to save lives. Vaccination is an easy, safe, and successful way for shielding people against injurious diseases. Vaccines reduce the possibility of getting diseases by functioning with body's natural defenses to build protection. When vaccines are injected, body's immune system responds to it by producing antibodies to fight against pathogens. Extensive testing and monitoring have shown that covid-19 vaccines are safe and effective and all citizens are advised to take vaccine and immunize themselves.

Data mining is a very promising area for making different types of decisions in the Medical field. During this period, it is crucial for us to understand the importance and efficiency of vaccinations. We could use different classification techniques and algorithms such as Naïve Bayes,

Random Forest, Logistic Regression, J48, MLP, Bagging e.tc to find the efficiency in different dimensions of medical fields. The utilization of machine learning and its applications deliver efficient results by extracting useful information from a dataset of group of citizens who have received the covid-19 vaccine. Selected information from the dataset could be used to analyze and predict the effectiveness of covid-19 vaccines.

II. PROPOSED METHODOLOGY

Machine learning can be used to predict the efficiency of vaccines. The key objective of this paper is to find the optimal algorithm that can predict the efficiency of vaccines. A survey was taken from a group of people belonging to different age groups to predict the efficiency of covid-19 vaccines. Questionnaire was prepared based on criteria on whether they have received covid-19 vaccinations dose 1, dose 2, whether they have been infected after dose 1 or dose 2. By using Google form, responses were collected from a group of people. These responses were collected and evaluated using the Weka tool. The results are envisioned using classification techniques like, Naive Bayes, Random Forest, and Logistic Regression.

III. IMPLEMENTATION METHODS

The application of data mining would be different for different sectors and is a vast topic. In the medical field, scope of Data mining (KDD) is very large and huge number of data sets are generated and computing a predication is out of the league for humans. So, data mining and machine learning algorithms come to play. Without data mining and machine learning, we could only draw this conclusion by plotting graphs manually, questionnaires etc. This would make the analysis difficult and prone to error (human error). Here using different implementation methods data were analyzed and results were evaluated.

A. Dataset Description

Attribute	Possible values
Email address	Email address of a person.
Age category	18+, 40-44, 45+, 60+
Existing diseases	Yes /No
Name of vaccine	Covaxin, Covishield, others
Received dose 1	Yes /No
Received dose2	Yes /No
Covid-19 infected after dose 1	Yes /No
Covid-19 infected after dose 2	Yes /No
Post vaccine side effects	Yes /No

Implementation was done through different stages. First stages involved the data collection from various group of citizens, second stage mainly consisted removing the duplicated data and extracting the relevant data from the data set and the last step involved the evaluation classification techniques to form a decision to obtain. Firstly, described how the data set should be formed and the different types of attributes to be involved. Table 1 shows the possible attributes and their descriptions.

Attribute	**Description**
Email address	Email address of a person.
Age category	Age of a person.
Existing diseases	Suffering from any other diseases.
Name of vaccine	Name of the vaccine a person has received.
Received dose 1	If a person has received dose 1 of vaccine.
Received dose2	If a person has received dose 2 of vaccine.
Covid-19 infected after dose 1	A person has been infected after dose1.
Covid-19 infected after dose 2	A person has been infected after dose 2.
Post vaccine side effects	Any post vaccine side effects.

Table. 1. Dataset Description.

B. Data collection

A questionnaire was created through google form and the collected data was stored in Microsoft Excel. Around 70 – 100 responses were collected and evaluated. Machine learning used in medical sector is gaining popularity due to the effectiveness in gaining accurate and flexible output. This dataset was collected from a group of citizens of different age group. This dataset was used to find the efficiency covid-19 vaccines in India. This paper mainly has two phases: First they were classified into their attributes as yes or no and second this paper deals with the use of classification techniques such as Random Forest, Naïve Bayes and logistic regression. Table 2 describes all the possible ranges of values of an attribute.

Table. 2. Possible values for attribute

C. Data preprocessing and feature selection

a. Data selection is a process where only essential information that could produce accurate results are selected. It is basically a filtration process to get useful information and discard wanted data. To get accurate result only useful data are required unwanted and unnecessary features would reduce the accuracy in the outcomes [**3**].

Data pre-processing is one of the most important process of data mining, that converts feasible data to useful information. Pre-processing includes steps like Data Cleaning, Data Transformation, Data Integration and Data Reduction.

D. Data mining

Data mining is commonly used technique to extract beneficial information from raw data to useful information. Data mining is also called knowledge discovery from data (KDD) which is used in different application like science, genetics, sales, marketing etc. Data mining include many steps such as extraction, data cleaning, transformation, management of data etc. These results are then manipulated into different visual representation such as decision trees, visualization graph etc. to make the data user understandable. All the useful data is retrieved through data mining algorithm and the main algorithm used are K Mean Algorithm, Naive Bayes Algorithm, J48, Random Forest, etc.

E. Applying Machine learning techniques

Data mining is an interdisciplinary area, that includes various sets of data. Classification helps to understand and match requirements. Classification algorithms are used to identify new observations on the basis of data sets. Classification algorithms in machine learning accepts data to predict a result that a particular data will fall into one of the pre-existing sets. Classification algorithms could be used to categorize data into different categories. It could be done on data that are structured or unstructured. There are different types of classification: binary classification, multiclass classification and multilabel classification. Classification is an important aspect as it used to predict instances of various data sets to receive a valid, classified, accurate, discrete outcome [**4**].

1. Random Forest

Random Forest is a type of supervised learning algorithm used for regression analysis and classification. As the name suggests, like a forest is made up of trees, a random forest algorithm is made up of decision trees and each decision tree in random forest gives a predicted outcome. By the process of voting an optimal solution or result is found. It can run efficiently on large datasets. This method is used for classification problems. It is said to be better than single decision tree because it reduces each method by averaging the results.

2. Naive Bayes

Naive Bayes algorithm [**19**] is a supervised algorithm that roots from the Bayes theorem used for classification. It is one of the most simple and effective classification algorithms which helps to make quicker predictions in machine learning. It is called Naive Bayes because it uses Bayes law for prediction, and it assumes that all features are independent and not related (naive).

3. Logistic Regression

Logistic regression [**20**] is a machine learning algorithm that comes under supervised learning, Logistic regression operates on categorical dependent variable and predicts feasible outcome. Hence the outcome could only be a categorical or discrete value. It can be either (Yes or No), (0 or 1), (true or False), etc. and it can't give exact value as 0 or 1 but gives only the probabilistic values that lie between 0 and 1.

IV. EXPERIMENTAL METHODOLOGY AND RESULTS.

A survey was conducted using google form and the data received was stored in Microsoft excel and converted to csv file. An .arff file, (Weka) file is generated from csv file through online converters. The Weka is a software that gives tools for data preprocessing, implement machine learning techniques etc. The accuracy, precision, recall, F-measure of each classification algorithm could be found using the equations given below.

$$\text{Accuracy} = \frac{\text{No.of correct prediction}}{\textit{Total No of predictions}}$$

$$\text{Precision} = \frac{\textit{True Positive}}{\textit{Actual Results}}$$

$$\text{Recall} = \frac{\textit{True Positive}}{\textit{Predicted Results}}$$

$$\text{F- measure} = \frac{2 * (precision * recall)}{precision + recall}$$

Table 3 is the description of results obtained from the classification techniques used:

Algorithm	Accuracy	Precision	Recall	F-measure
Random forest	94.50%	0.94	1.00	0.97
Naïve Bayes	71.57%	0.97	0.70	0.81
Logistic Regression	88.99%	0.89	1.00	0.94

Table. 3. Performance Comparison of Classifiers

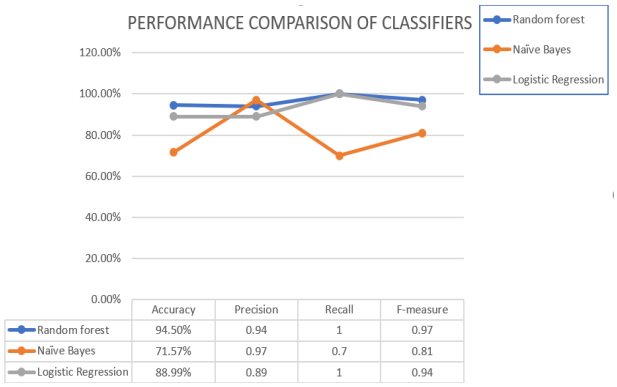

Fig. 1. Graphical representation of Performance comparison of Classifiers

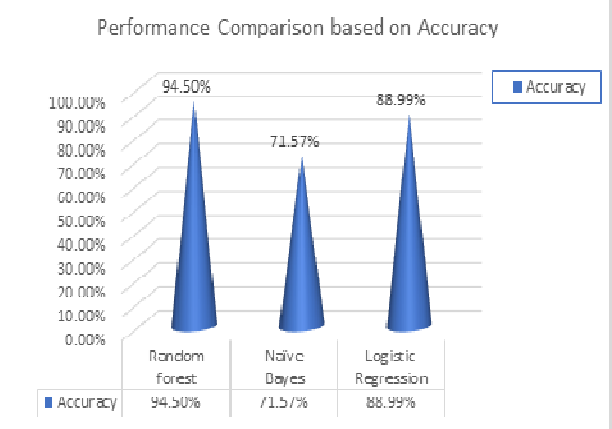

Fig. 2. Performance comparison on accuracy.

It is very important to plot data in graphical form, as graphical visualization plays a crucial role in understanding the data and its characteristics. Figures 1 and 2 are created based on results obtained, for better understanding.

Through these observations we say that the accuracy of these classification techniques is more in Random forest than other techniques. By classification techniques we found that random forest has 94.5 % accuracy, Naïve Bayes has 71.57% accuracy and logistic regression has 88.99% accuracy. And from the above observations, Random forest is the most appropriate technique for predicting efficiency of Covid-19 vaccines accurately.

V. CONCLUSION

In this paper we have tried to predict the efficiency of covid-19 vaccines available in India using classification techniques. The study shows that majority of the citizen have been vaccinated efficiently and more than 90% of citizens have not been infected after dose 1 or dose 2 of the vaccination. This shows that being immunized will help people fight against covid-19. The effectiveness of vaccines can differ from one person to another, and the type of vaccine taken. Not all vaccines have the same efficiency, to predict the efficiency of vaccines machine learning techniques could be used with the help of machine learning attributes like Data mining. Results obtained from the google form was not sufficient to conclude a decision, so with the help of Weka tool and classification techniques like Random Forest, Naïve Bayes, Logistic regression we could predict the efficiency of covid-19 vaccines. In terms of accuracy random forest stands at position one. It has the highest accuracy among all classification techniques. From this paper, we can conclude that, the best classification technique used to predict the efficiency of covid-19 vaccines is Random Forest.

ACKNOWLEDGMENT

First and foremost, praises to God almighty for showering his blessing throughout my research work. I would like to express my hearty gratitude to my mentor, Assistant Professor Miss Ambily Merlin Kuruvilla (HOD, Department of computer Application, Saintgits College of Applied Sciences) for giving me this opportunity and providing me the guidance to complete the research. Her motivation is what really inspired me. It was a great honor to have been working under her guidance and support.

I am extremely thankful to my parents who have supported me and prayed for my success at all times. I also extend my gratitude to my teachers for their support and valuable instructions. Last but not the least, I also would like to thank my friends for their endless support and motivations.

REFERENCES

[1] Aishwarya, R. Gayathri Pand Jaisankar N, “A Method for Classification Using Machine Learning Technique for Diabetes.” Int. J. Eng. Technol. (IJET) 2013, 5, pp. 2903–2908.

[2] Al-Karawi, Dhurgham, et al. "Machine learning analysis of chest CT scan images as a complementary digital test of coronavirus (COVID-19) patients." MedRxiv (2020).

[3] Ambily Merlin Kuruvilla, Dr. N V. Balaji," Predicting diabetes mellitus using feature selection and classification techniques in machine learning algorithms", Karpagam jcs vol.13 Sep. - oct. 2019.

[4] Amjad Abu Saa Information Technology Department Ajman University of Science and Technology Ajman, United Arab Emirates," Educational data mining & students’ performance prediction", (ijacsa) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 5, 2016

[5] An, C., Lim, H., Kim, D.W., Chang, J.H Y.J. and Kim, S.W., 2020. “Machine learning prediction for mortality of patients diagnosed with COVID-19: a nationwide Korean cohort study”. Scientific reports, 10(1), pp.1-11.

[6] Arora and R. Suman, “Comparative Analysis of Classification Algorithms on different datasets using WEKA”, International Journal of Computer Applications, 54 (2012), pp. 21- 25 i:10.5120/8626-2492.

[7] Estiri, H., Strasser, Z. H., Klann, J. G., Naseri, P., Wagholikar, K. B., & Murphy, S. N. (2021). “Predicting COVID-19 mortality with electronic medical records”. NPJ digital medicine, 4(1), pp. 1-10.

[8] Forsati, R., Moayedikia, A., Keikha, A., & Shamsfard, M. (2012). A novel approach for feature selection based on the bee colony optimization. International Journal of Computer Applications, 43(8), 30-34.

[9] Jamshidi, M., Lalbakhsh, A., Talla, J., Peroutka, Z., Hadjilooei, F., Lalbakhsh, P., & Mohyuddin, W. (2020). “Artificial intelligence and COVID-19: deep learning approaches for diagnosis and treatment.” IEEE Access, 8, 109581-109595.

[10] Kannan, S., Subbaram, K., Ali, S., & Kannan, H. (2020). “The role of artificial intelligence and machine learning techniques:

Race for covid-19 vaccine". Archives of Clinical Infectious Diseases, 15(2).
[11] Kwekha-Rashid, A.S., Abduljabbar, H.N. and Alhayani, B., 2021. "Coronavirus disease (COVID-19) cases analysis using machine-learning applications". Applied Nanoscience, pp.1-13.
[12] Lazarus, J. V., Ratzan, S. C., Palayew, A., Gostin, L. O., Larson, H. J., Rabin, K., ... & El-Mohandes, "A. (2021). A global survey of potential acceptance of a COVID-19 vaccine. Nature medicine", 27(2), pp. 225-228.
[13] Leia wedlund & Joseph "Machine learning model predicts who may benefit most from covid-19 vaccination" kvedar npj digital medicine by volume 4, article number: 59 (2021).
[14] Mannan, D. K. A., & Farhana, K. M. (2020). Knowledge, attitude and acceptance of a COVID-19 vaccine: A global cross-sectional study. International Research Journal of Business and Social Science, 6(4) doi: December 7, 2020.
[15] Priyam A, Gupta R, Rathee A and Srivastava S "Comparative Analysis of Decision Tree Classification Algorithms" International Journal of Current Engineering and Technology" Vol., 3 (2013), pp. 334-337 doi: June 2013, aXiv: ISSN 2277-4106.
[16] Sallam M. (2021). COVID-19 vaccine hesitancy worldwide: a concise systematic review of vaccine acceptance rates. Vaccines, 9(2), 160.
[17] Snider, B., McBean, E. A., Yawney, J, Gadsden, S. A., & Patel, B. (2021). Identification of Variable Importance for Predictions of Mortality from COVID-19 Using AI Models for Ontario, Canada. Frontiers in Public Health, 9.
[18] Webb, G. (2021). "A COVID-19 epidemic model predicting the effectiveness of vaccination". Mathematics in Applied Sciences and Engineering, pp. 1-15. doi:10.5206/mase/13889.
[19] Ambily Merlin Kuruvilla, Dr. N V. Balaji (2021). "Heart disease prediction system using Correlation Based Feature Selection with Multilayer Perceptron approach", IOP Conference Series: Materials Science and Engineering.
[20] Ambily Merlin Kuruvilla, Dr. N V. Balaji (2020). "A Review and Analysis on Data Mining Algorithms to Predict Diabetes."