iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** V **Month of publication:** May 2023

DOI: https://doi.org/10.22214/ijraset.2023.51712

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue V May 2023- Available at www.ijraset.com

Outbreak Prediction System using Machine Learning

Prof. Ashwini Yerlekar[1], Pranav Suralkar[2]
[1, 2]*Dept. Computer Science and Engineering Rajive Gandhi College of engineering and Research Nagpur, India*

Abstract: Outbreaks have significant social and political impacts such as clashes between nations, population displacement, and increased social tension and discrimination. The work proposes to help in the process of detecting the outbreak and finding the illness using data collected from verified medical professionals and using machine learning algorithm design tree to analysis data collected to identify communities which are in most danger of the outbreak. Thus the system helps to predict and monitor outbreak scenarios.
Keywords: Outbreak, diseases, machine learning

I. INTRODUCTION

A disease outbreak is the occurrence of cases of disease in excess of what would normally be expected in a defined community, geographical area or season. Outbreaks are maintained by infectious agents that spread directly from person to person, from exposure to an animal reservoir or other environmental source, or via an insect or animal vector. Human behaviour nearly always contribute to such spread. Early detection and reporting of such events is crucial in minimizing their negative social and economic impact.
The system applies machine learning and prediction Algorithm like Multiple Linear Regression to identify the pattern among data and then process as per input provided. This in turn will provide predictions about the outbreak. This system will requires the data of the patients their area of residence and it will identify and predict the outbreak progression.

II. LITERATURE SURVAY

Mohd Javaid, Abid Haleem and Ravi Pratap Singh Conveyed ML has been thrust into the spotlight since the outbreak of the COVID-19 pandemic. Organizations have turned to ML to stay competitive and gain an advantage, from streamlining operations to driving R&D in a sometimes volatile and uncertain work environment. ML has helped hospitals and health systems deal with unique challenges.
They have not only sought to establish a relationship between climatic factors and a possible malarial outbreak but they also tried to find out which algorithm is best suited for modeling the discovered relationship. For that purpose, historical meteorological data and records of malarial cases collected over six years have been combined and aggregated in order to be analyzed with various classification techniques such as KNN, Naive Bayes, and Extreme Gradient Boost among others. They were able to find out few algorithms which perform best in this particular use case after evaluating for each case, the accuracy, the recall score, the precision score, the Matthews correlation coefficient and the error rate.
[2] They took two datasets in consideration: the first is an Italian dataset obtained from the Istituto di Fisiologia Clinica of Consiglio Nazionale delle Ricerche of Reggio Calabria; the second is an American dataset provided by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) repository. From each one we obtained 5 datasets, according to the outcome of interest. They tested different types of algorithm (both linear and non-linear), but the final choice was to use Support Vector Machine. In particular, we obtained the best performances using the non-linear SVC with RBF kernel algorithm, optimizing it with GridSearch. The last is an algorithm useful to search the best combination of hyper-parameters (in our case, to find the best couple (C, γ)), in order to improve the accuracy of the algorithm.

III. PROBLEM STATEMENT

A disease outbreak is the occurrence of cases of disease in excess of what would normally be expected in a defined community. When outbreaks go undetected or are not monitored properly have significant social and political impacts and increased social tension and discrimination. To avoid such consequences we need to develop a system for outbreak detection and monitoring.

©IJRASET: All Rights are Reserved

IV. PROPOSED WORK

This system is designed to reduce damage done to society when outbreak of a disease occurs by detecting outbreak and its monitoring. This system will help CDC to handle outbreak of a disease by providing important data analysis. This system will require live data collected from the patients about their symptoms, treatment and followups.

A disease outbreak is the occurrence of cases of disease in excess of what would normally be expected in a defined community, geographical area or season. Outbreaks are maintained by infectious agents that spread directly from person to person, from exposure to an animal reservoir or other environmental source, or via an insect or animal vector. Human behaviors nearly always contribute to such spread. Early detection and reporting of such events is crucial in minimizing their negative social and economic impact.

A. Design And Development Of Outbreak Prediction System

There are three parts of this outbreak prediction system. The first part is data collection Second part is data storage and prepossessing and the Third part is to get conclusions and predictions from the collected data. This system will help in maintaining proper patients data, it will help in prediction of outbreaks.

B. Data Collection

Data for such applications is collected in following ways.There a a method of data scraping in this method the data is collected by tools which are automated to collect data which is available on the web. This is the method in which data is collected by specialized portals that are specifically designed to collect data for that system. This is the method where data is bought from a third party.

C. Prepossessing

Data prepossessing is an important step in the data mining process. It refers to the cleaning, transforming, and integrating of data in order to make it ready for analysis. The goal of data preprocessing is to improve the quality of the data and to make it more suitable for the specific data mining task.

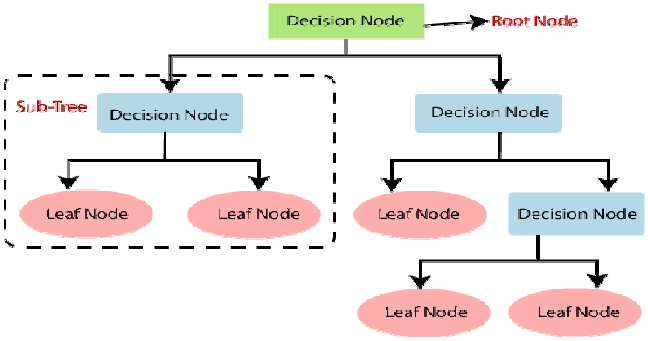
D. Prediction

“Prediction” refers to the output of an algorithm after it has been trained on a historical datasets and applied to new data when forecasting the likelihood of a particular outcome, such as whether or not a customer will churn in 30 days. The algorithm will generate probable values for an unknown variable for each record in the new data, allowing the model builder to identify what that value will most likely be. The word “prediction” can be misleading. In some cases, it really does mean that you are predicting a future outcome, such as when you’re using machine learning to determine the next best action in a marketing campaign. Other times, though, the “prediction” has to do with, for example, whether or not a transaction that already occurred was fraudulent. In that case, the transaction already happened, but you’re making an educated guess about whether or not it was legitimate, allowing you to take the appropriate action.

First the raw data about disease and its symptoms is taken from kaggle and the data is converted to SQL format from csv format using PHP algorithm. Then the raw data will be formatted accordingly. Using PHP algorithm in three steps these algorithms are non standard custom designed for the requirements. UI based data collection portal is created using PHP SQL HTML ZURB to collect data of patients city wise and other parameters as required by the portal by authorized personal.

The collected data will be in SQL format then it will be cleaned and prepossessed using PHP.

The ready use data is in SQL format it needs to be csv format to be operated in JUPITER notebook pandas library decision tree will be used to get the prediction from that data.



Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a data set, branches represent the decision rules and each leaf node represents the outcome.
For testing the process dummy data is created using the disease date gathered. The dummy date is randomized using PHP and SQL using custom made algorithm.

E. Tools Used

1) *HTML:* HTML (Hypertext Markup Language) is a markup language used to create the structure and content of web pages. It is a standard language used to create web pages and is used to define the various elements of the page and their layout.
2) *CSS:* CSS works by associating styles with HTML elements.Styles can be defined using selectors and attributes. Selectors target specific elements or groups of elements, and attributes define the style or layout of elements.
3) *PHP:* PHP (Hypertext Preprocessor) is known as a general-purpose scripting language that can be used to develop dynamic and interactive websites. It was among the first server-side languages that could be embedded into HTML, making it easier to add functionality to web pages without needing to call external files for data.
4) *SQL:* SQL is used to communicate with a database. According to ANSI (American National Standards Institute), it is the standard language for relational database management systems. SQL statements are used to perform tasks such as update data on a database, or retrieve data from a database.
5) *PYTHON PANDAS:* Pandas is a Python library used for working with data sets. It has functions for analyzing, cleaning, exploring, and manipulating data. The name "Pandas" has a reference to both "Panel Data", and "Python Data Analysis" and was created by Wes McKinney in 2008.

V. AIM & OBJECTIVE

"To implement a rule based system to predict and monitor outbreak scenarios from the collection of past data."

1) Detect Outbreak before it gets out of control.
2) Find illnesses which is causing the outbreak.
3) Generate Hypotheses.
4) Test Hypotheses.

VI. RESULT

Running the processed data in jupyter notebook we will be using DecisionTreeClassifier from sklearn. After importing data from the database into two parts x and y and then split them further 80% and 20% . 20% for traning the module.
We get accuracy of 46% , And are able to predict disease using age and city.

VII. CONCLUSION

Machine learning is a powerful tool for making predictions from data. However, it is important to remember that machine learning is only as good as the data that is used to train the algorithms. In order to make accurate predictions, it is important to use high-quality data that is representative of the real-world data that the algorithm will be used on.
This outbreak prediction system will collect data using specialized portal which will be operated by authorized individuals the data will be collected area wise. The data will be stored in SQL servers using PHP as a back-end . Then the data will be prepossessed and will be used to make prediction using Pandas API.

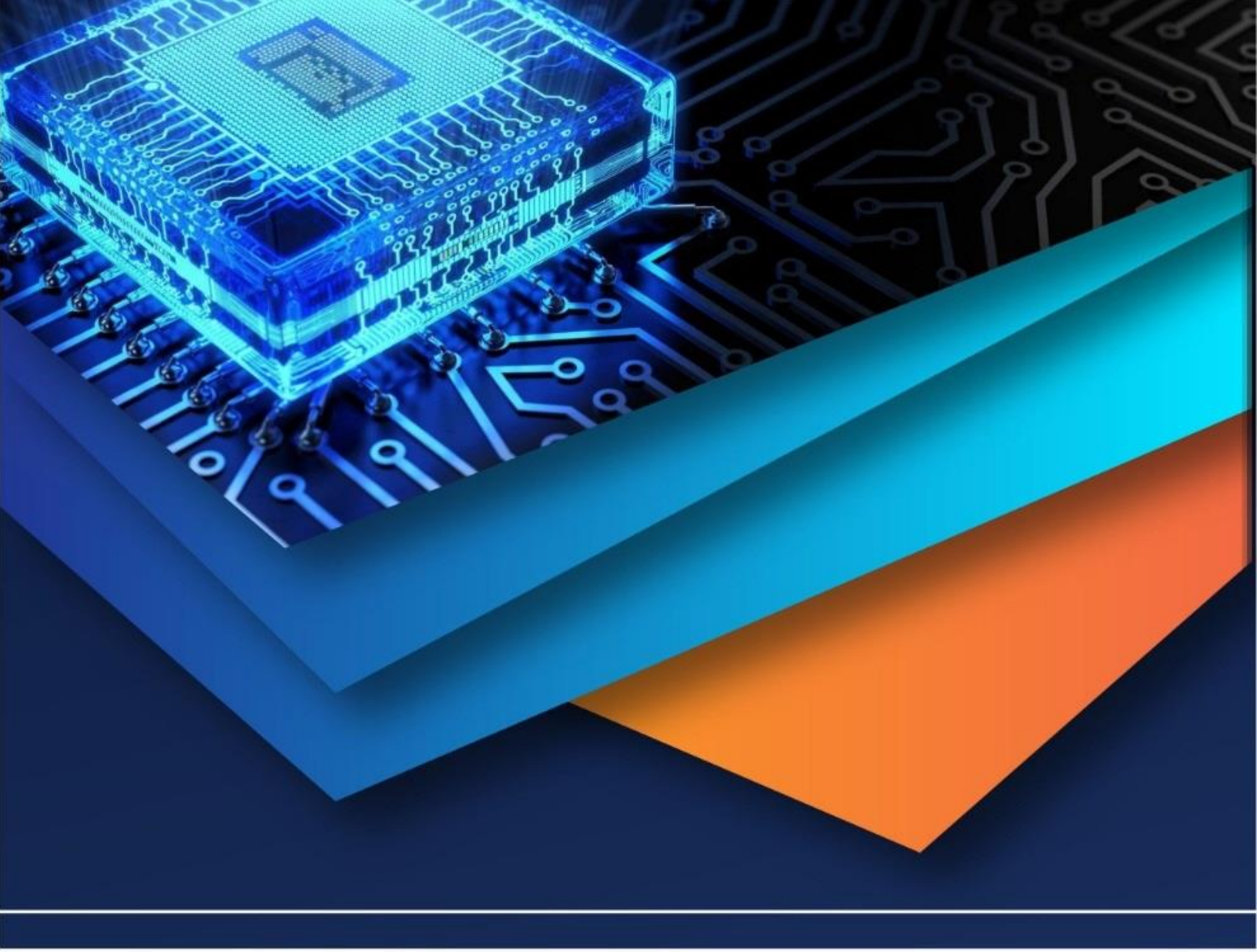
VIII. FUTURE SCOPE

1) Accuracy can be further increased by collecting more accurate data from reliable sources.
2) The application can be further developed to include more categories to improve accuracy.

REFERENCES

[1] Nurul Azam Mohd Salim, Yap Bee Wah, Caitlynn Reeves, Madison Smith, Wan Fairos Wan Yaacob, Rose Nani Mudin, Rahmat Dapari, Nik Nur Fatin Fatihah Sapri & Ubydul Haque, Prediction of dengue outbreak in Selangor Malaysia using machine learning techniques, Nature, 01-2021
[2] SabrinaMezzatesta, ClaudiaTorino, Pasquale DeMeo, GiacomoFiumara, AntonioVilasi, A machine learning-based approach for predicting the outbreak of cardiovascular diseases in patients on dialysis, Science Direct, 11-2018

[3] Godson Kalipe, Vikas Gautham, Rajat Kumar Behera, Predicting Malarial Outbreak using Machine Learning and Deep Learning Approach: A Review and Analysis, IEEE Xplore, 12-2018
[4] Smita Rath, Alakananda Tripathy, Alok Ranjan Tripathy, Prediction of new active cases of coronavirus disease (COVID-19) pandemic using multiple linear regression model, Science Direct,10-2020
[5] Ruxin Wang, Chaojie Ji, Zhiming Jiang, Yongsheng Wu, Ling Yin, Ye LI, A Short-Term Prediction Model at the Early Stage of the COVID-19 Pandemic Based on Multisource Urban Data, IEEE Xplore, 05-03-2021
[6] Haifeng Hu, Hong Du, Jing Li, Yage Wang,Xiaoqing Wu, Chunfu Wang, Ye Zhang, Gufen Zhang, Yanyan Zhao, Wen Kang†nd Jianqi Lian,Early prediction and identification for severe patients during the pandemic of COVID-19: A severe COVID-19 risk model constructed by multivariate logistic regression analysis,Journal of Global health,12-2020
[7] Junyi Gao, Rakshith Sharma, Cheng Qian, Lucas M Glass, Jeffrey Spaeder, Justin Romberg, Jimeng Sun, Cao Xiao,STAN: spatio-temporal attention network for pandemic prediction using real-world evidence,JAMIA,22-01-2021
[8] Saleh I Alzahrani ,Ibrahim A Aljamaan, Ebrahim A Alfakih,Forecasting the spread of the COVID-19 pandemic in Saudi Arabia using ARIMA prediction model under current public health interventions,Science Direct,07-2020
[9] Gergo Pinter, Imer Felde, Amir Mosavi, Pedram Ghamisi And Richard Gloaguen,COVID-19 Pandemic Prediction for Hungary; A Hybrid Machine Learning Approach,Mathematics,02-06-2020
[10] Sabrina Mezzatesta, Claudia Torino Pasquale, DeMeoc Giacomo,Fiumara Antonio Vilasi,A machine learning-based approach for predicting the outbreak of cardiovascular diseases in patients on dialysis,Science Direct,08-2019



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)