International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Analysis of the Efficiency of Machine Learning Methods for Disease Prediction

[1]Japathi Sankeerthana, [2]Dr. G.Chauhan

Vaageswari College of Engineering, Karimnagar – 505 527, India

A B S T R A C T

Machine learning aids in the diagnosis and treatment of a wide variety of diseases. Improved disease prognosis and patient care can be achieved through the use of predictive analysis, which benefits from the application of powerful multiple machine learning algorithms. Disease prediction may be done quickly and accurately using machine learning techniques. These clinical characteristics are used in conjunction with machine learning algorithms for disease classification. In the end, we use a graph to compare the outcomes of various machine learning classification techniques.

KEYWORDS: Machine learning, classification, KNN, Data mining, Naviee bayes

1. INTRODUCTION

Keeping a patient healthy is a constant challenge due to the fact that each sickness has its own unique set of symptoms. Traditional medical practise assigns relative importance to each symptom in order to predict the presence of a disease and aid in the diagnosis process. The symptom with the greatest impact on the condition is given the most weight. Data mining supplements and replaces conventional medical expertise in aiding in the diagnosis and prognosis of disease. By using this strategy, we are able to extract previously unseen patterns, relationships, and knowledge that would have been inaccessible using more conventional statistical methods. Privilege and correct conclusion are ever-present contributors to the effective treatment. Data mining concepts are flexible enough to uncover hidden numbers, relationships within a database, and machine learning approaches to further assess a patient's case based on the documented clinical information. Machine learning algorithms are used to forecast a wide range of diseases and to develop effective regimens for improved health while minimizing the risks of excessive cost, slow recovery, and incorrect treatment. Thankfully, machine learning algorithms prove effective at disease prediction, and there are still a plethora of untapped methods to investigate. In this paper, we proposed a graphical user interface (GUI) application that makes use of machine learning methods for a disease prediction system, mining on the symptoms of the disease, and finally detecting the disease by comparing the performance or accuracy of different techniques in doing so. To get there, we'll use machine learning categorization techniques like Naive Bayes, Decision Tree, Support Vector Machines, and Random Forest, which can reliably predict the outcome of each new insertion.

2. RELATED WORK

One area where machine learning can be used is in the prediction of cardiovascular disease. The flexibility and adaptability of optimization algorithms make them well suited for handling difficult, non-linear issues. In order to enhance the quality of heart disease classification, we used the Fast Correlation-Based Feature Selection (FCBF) technique to eliminate redundant information. Then, we use a Multilayer Perception | Artificial Neural Network optimised with Particle Swarm Optimization (PSO) and Ant Colony Optimization (ACO) methods to perform a classification based on various classification algorithms like K-Nearest Neighbor, Support Vector Machine, Naive Bayes, Random Forest, and others. Applying the suggested hybrid method to the heart disease dataset, the results show the method's efficacy and robustness in processing different types of data for disease classification. Therefore, this research evaluates the outcomes of several machine learning algorithms based on a variety of metrics (such as accuracy, precision, recall, f1-score, etc.). Through the use of the proposed optimised model by FCBF, PSO, and ACO, we achieve a maximum accuracy of 99.65% in our classifications. The outcomes demonstrate that the proposed system outperforms the previously stated categorization method.

3. ARCHITECTURE

A system architecture or systems architecture is the conceptual model that architecture is the conceptual model that defines the structure, behavior and more views of a system. An architecture description is a formal description is a formal description and representation of a system, organized in a way that supports reasoning about the structures and behaviors of the system. System architecture can comprise system components, the externally visible properties of those components, the relationships between them.

The Implementation is Phase where we endeavor to give the practical output of the work done in designing stage and most of Coding in Business logic lay coms into action in this stage its main and crucial part of the project

- Dataset upload
- Pre-processing data
- Extracting dataset
- Splitting dataset traing and testing
- Applying models

4. METHODOLOGY

The proposed study includes the use of eight ML classification algorithms to predict and estimate the performance of six different medical diseases in humans. Two consecutive phases will be employed to diagnose the six different diseases.

1. The first Phase: consist of the following procedure Data Preprocessing: After collecting the datasets for the six various medical diseases, most of the datasets contain some attributes with nominal data and numerical data, and other datasets have attributes with null values. The role-play of data preprocessing is a data mining technique that can transform the raw data into an obvious and understandable format.

a. Data Transformation: All of the nominal data are transformed into numerical digital data. For example, the normal is converted into 1 and the abnormal is converted into 0, etc.

b. Missing Data Handling: The null values of the attributes are replaced using the KNN imputation technique which works on data that identifies the neighboring points through the measure of the distance between them then, the missing values can be estimated using completed values of neighboring observations.

2. Class Targeting: To classify the patient whether to be normal or abnormal according to the values of the available features.

3. Training & Testing Datasets Division: The whole datasets are divided into two ratios; 70% - 30 and 80% - 20% training and testing sets. Where the training sets are used to train the models and the testing sets are used to test their performance

The Second Phase: consist of the following procedure

1. Applying Machine Learning Algorithms: The main idea of the algorithms is to build ML models that could extradite input data. Then, statistically analyzing them to predict the output and to classify the disease as normal and abnormal with more accuracy. In this paper, eight different ML algorithms such as LR, KNN, SVM, NB, DT, RF, XGB, and ADB are applied to the medical datasets to predict and diagnose the presence of certain diseases.

2. Training & Testing Datasets: Each ML algorithm is trained using the training set, the learning algorithm finds patterns in the training data that map the input data attributes to the target which i

5. LIST OF MODULES

Tensor flow:

Open-source dataflow and differentiable programming framework TensorFlow is available for a wide range of activities.

Numpy :

A general-purpose array-processing toolkit, Numpy is available. High-performance multidimensional array objects and tools for interacting with these arrays are provided.

Pandas:

Pandas is an open-source Python library that provides strong data structures for data manipulation and analysis. Python was primarily utilised for data manipulation and prepping. It made a tiny dent in the data analysis process. This was a problem that Pandas solved. Data processing and analysis typically involves five steps: load, prepare, modify, model, and analyse, all of which may be accomplished with Pandas. Python and Pandas are widely utilised in a variety of sectors, including finance, economics, statistics, and analytic applications.

Matplotlib:

Mathematica's Matplotlib is a Python 2D plotting toolkit that produces publication-quality figures in a wide range of hardcopy formats and interactive settings across platforms. In Python scripts, the Python and IPython shells, the Jupyter Notebooks, web application servers, and four graphical user interface toolkits, Matplotlib can be utilised.

Scikit – learn

Scikit-learn provides a standard Python interface for a variety of supervised and unsupervised learning techniques. It is licenced under a BSD-style licence that allows for both academic and commercial use, and is available on a variety of Linux versions.

6. CONCLUSION

The fundamental aim of the paper is to anticipate all the more precisely the event of the ailment utilizing Machine Learning procedures. We used four algorithms like Decision Tree, Random Forest, SVM and Naïve Bayes. These algorithms are used to evaluate the parameters like Accuracy, Precision, Recall and F-score on a disease dataset by considering the symptoms as input variables. Among all the algorithms on the comparison, the highest accuracy for predicting the disease is given by the Naïve Bayes algorithm. Further enhancements can be added to this system because the new disease may arise in future. These new diseases can be predicted by adding new symptoms to the data set. By using more powerful machine learning supervised algorithms, the GUI is modified user friendly to provide more detail information about the disease to the patients and can be further extended by an additional component suggesting medicine to the patient if necessary on emergency.

REFERENCES:

[1] Thailand Motor Vehicle Registered. CEIC Data Global Database.

[2] Linda, S., Can public transport compete with the private car? Iatss Research, 2003. 27(2): p. 27-35.

[3] CO2 emissions (metric tons per capita) - Thailand. THE WORLD BANK.

[4] Cohen, A.J., et al., Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the Global Burden of Diseases Study 2015. The Lancet, 2017. 389(10082): p. 1907-1918.

[5] Litman, T. and D. Burwell, Issues in sustainable transportation. International Journal of Global Environmental Issues, 2006. 6(4): p. 331-347.

[6] Liu, M. and N. Choosri.. A technical solution to improve the red cab for touring in Chiang Mai: Chinese tourists' perspective. in 2016 Chinese Control and Decision Conference (CCDC). 2016. IEEE

[7] Farooq, M.U., A. Shakoor, and A.B. Siddique. GPS based Public Transport Arrival Time Prediction. in 2017 International Conference on Frontiers of Information Technology (FIT). 2017. IEEE.

[8] Bin, Y., Y. Zhongzhen, and Y. Baozhen, Bus arrival time prediction using support vector machines. Journal of Intelligent Transportation Systems, 2006. 10(4): p. 151-158.

[9] Maiti, S., et al. Historical data based real time prediction of vehicle arrival time. in 17th International IEEE Conference on Intelligent Transportation Systems (ITSC). 2014. IEEE.

[10] Fan, W. and Z. Gurmu, Dynamic travel time prediction models for buses using only GPS data. International Journal of Transportation Science and Technology, 2015. 4(4): p. 353-366.