International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 01 | January - 2024 SJIF Rating: 8.176 ISSN: 2582-3930

Health Guard: Machine Learning Powered Preliminary Diseases Predicto

Nisha Wilvicta J [1], Jitendra Prajapati [2], M.A. Qhizar [3], MD Shadab Arshad[4], Mohd. Fardeen Shaiekh[5]

[1]Nisha Wilvicta J, Computer Science Engineering Department, T John Institute of Technology, Bengaluru

[2]Jitendra Prajapati, Computer Science Engineering Department, T John Institute of Technology, Bengaluru

[3]M.A. Qhizar, Computer Science Engineering Department, T John Institute of Technology, Bengaluru

[4]MD Shadab Arshad, Computer Science Engineering Department, T John Institute of Technology, Bengaluru

[5]Mohd. Fardeen Shaiekh, Computer Science Engineering Department, T John Institute of Technology, Bengaluru

---

Abstract: The evolving landscape of health information-seeking behaviour poses challenges for users navigating online platforms to comprehend diseases, diagnoses, and treatments. Implementing recommendation systems for doctors and medicines through review mining could streamline this process, saving considerable time. However, a significant hurdle lies in users, often laymen, grappling with complex medical vocabulary amidst the vast array of available information. Integrating advanced data mining techniques, particularly in healthcare and biomedicine, contributes to the extraction of valuable insights. Machine learning applications, like disease prediction systems, play a pivotal role in early diagnosis and patient care. In a recent study, machine learning algorithms, including Decision Tree, Random Forest, and Naïve Bayes classifiers, were employed to predict and diagnose diseases, showcasing promising results in enhancing healthcare outcomes.

Keywords: Machine Learning, Data mining, Decision Tree classifier, Random Forest classifier, Naive Bayes classifier, Disease Prediction.

I. INTRODUCTION

Healthcare is a critical domain where technological advancements can significantly impact the well-being of individuals. Timely diagnosis and treatment are fundamental to managing various diseases effectively. The "Health Guard: Machine Learning Powered Preliminary Diseases Predicto" project is a novel application that harnesses the power of machine learning to assist in disease prediction, making healthcare more accessible and convenient. The primary objective of this project is to provide users with a simple and efficient means of predicting diseases based on their reported symptoms. Developed by our team, this application features a user-friendly GUI that allows individuals to input their symptoms effortlessly. The system then employs machine learning models to make predictions, and the predicted disease is displayed to the user. If a match is found, the disease name is displayed; otherwise, it indicates "Not Found," encouraging users to consult a healthcare professional for further evaluation.

The project incorporates three machine learning models: Decision Tree, Random Forest, and Naive Bayes. These models are trained on a dataset that contains a list of symptoms and their associated diseases. By utilizing this dataset, the models learn to recognize patterns and associations between symptoms and diseases, enabling them to make informed predictions.

Furthermore, the application calculates and displays the accuracy of the machine learning models, providing users with insights into the reliability of the predictions. This transparency is essential as it instils trust in the system's ability to provide accurate disease predictions.

The "Health Guard: Machine Learning Powered Preliminary Diseases Predicto" project serves as a demonstration of the practical application of machine learning in the healthcare sector. It emphasizes the potential of technology to enhance and expedite the medical diagnosis process, potentially saving lives by promoting early intervention. The user-friendly interface ensures that individuals, regardless of their technical expertise, can benefit from this innovative tool.

II. LITERATURE REVIEW

2.1 Algorithms Precision

Machine learning techniques are used by Gupta et al. [6] to develop a prediction system that evaluates symptoms and predicts the best treatment for each newly identified condition. The illness prediction system is made using the three data mining algorithms: Random Forest Classifier, Decision Tree Classifier, and Naive Bayes Classifier. There is a tentative list of diseases that are known to exist along with their symptoms. The drugs and their ingredients are then discussed in connection with the mentioned conditions. The system was assessed using the dataset that was obtained from New York-Presbyterian Hospital. This study shows that the Nave Bays Classifier has a greater accuracy (approximately 98%) than the Random Forest and Decision Tree algorithms (both of which have an accuracy of roughly 97%).

2.2 Heart Disease

The results show that RF and LM are the best. The RF error rate for a dataset is high (20.9%) [2] compared to the other datasets. The LM method for the dataset is the best (9.1%) compared to DT and RF methods. We combine the RF method with LM and propose HRFLM method to improve the results.

2.3 Liver Diseases

Moreover, C Geeta [3] found a technique from Prof Christopher N. New Automatic Diagnosis of Liver Status Using Bayesian Classification that considers six benchmarks in the diagnosis of medical diseases: liver, hepatitis, heart, diabetes, breast, and lymph illnesses. With a precision of 64.60 percent with 19 liver problem dataset rules and 62.89 percent with 43 WSO and C4.5 rules, respectively, the researchers constructed WSO- and C4.5-based systems.

2.4 Dengue Disease

Tan [4] presented a method in which the wrapper approach is used to effectively connect two machine learning algorithms, Genetics Algorithm (GA) and SVM hybrid approach. LIBSVM and the WEKA data mining tool are employed in this work. Two data sets—diabetes and heart disease—will be taken from the UC Irvine machine learning repository for this investigation. The GA and SVM hybrid method yields an 84.07 percent precision for heart disease. Accuracy rates for the collection of diabetes data are 78.26 percent. Additionally, it is a binary classifier, which makes it less prone to overfitting, robust against noise, and has certain disadvantages. For the classification of multiple classes, pairwise identification may be used. The cost of computation is high, so it works slowly.

2.5 Cancer

John Adeoye's [5] study revealed that in terms of sensitivity and precision, machine learning models with 26 variables outperformed those with 15 variables. This refers to the 26-feature models' ability to better distinguish between high- and low-risk patients with oral leukaemia and oral lichenoid mucositis when clinical factors like risk factor category, clinical history of comorbidities, viral hepatitis status, number of lesions, and presence of induration are available. The machine learning models' performance was observed to differ depending on the class imbalance correction method used. Specifically, the models that used ADASYN yielded higher F1 scores than SMOTE, regardless of the number of variables or machine learning methodology used.

2.6 Multiple Symptoms

Suvendu Kumar Nayak [1] used four models for disease prediction Multinomial Naive Bayes, Extra Tree Classifier, Decision Tree Classifier, and Support Vector Machine. Multinomial Naive Bayes has an accuracy of 86.90%.

The accuracy rate of the Additional Tree Classifier is 88.18%. 86.95% accuracy is provided via a decision tree classifier with a maximum depth of 120. The Support Vector Machine classification yielded 86.96%. The first and second symptom projections for the condition were incorrect. For three symptoms, multinomial Naive Bayes generated the correction prediction.

III. METHODOLOGY

3.1 Dataset

To begin with, we need some datasets in order to train our model and gain some insights. To that end, we have conducted a number of surveys within the medical sector, looked through some online data, and combined all of it into a raw dataset.

International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 01 | January - 2024 SJIF Rating: 8.176 ISSN: 2582-3930

Thus, a dataset is now available.

skin_peelir	silver_like_	small_den	inflammat	blister	red_sore_	yellow_cru	prognosis
0	0	0	0	0	0	0	Fungal infection
0	0	0	0	0	0	0	Fungal infection
0	0	0	0	0	0	0	Fungal infection
0	0	0	0	0	0	0	Fungal infection
0	0	0	0	0	0	0	Fungal infection
0	0	0	0	0	0	0	Fungal infection
0	0	0	0	0	0	0	Fungal infection
0	0	0	0	0	0	0	Fungal infection
0	0	0	0	0	0	0	Fungal infection
0	0	0	0	0	0	0	Fungal infection
0	0	0	0	0	0	0	Allergy
0	0	0	0	0	0	0	Allergy
0	0	0	0	0	0	0	Allergy
0	0	0	0	0	0	0	Allergy
0	0	0	0	0	0	0	Allergy
0	0	0	0	0	0	0	Allergy
0	0	0	0	0	0	0	Allergy
0	0	0	0	0	0	0	Allergy
0	0	0	0	0	0	0	Allergy
0	0	0	0	0	0	0	Allergy

Fig 1: Symptoms Dataset

3.2 Data Preparation/ Preprocessing

Signs and Diseases Lists: Symptoms and the diseases, they correspond with are then recorded in lists. Training and testing datasets are needed to load from CSV files, and lists and data frames will be initialized. After gathering the data, we must prepare it for use in training our machine learning model because it is raw data. The utilization of Python tools such as NumPy, pandas, and scikit-learn has enabled us to prepare the data for machine learning models.

3.3 Machine Learning Models

3.3.1 Random Forest

In supervised machine learning, where a target variable has been labelled, random forests are used. Classification and regression problems with a numeric target variable can be resolved using random forests. As an ensemble approach, random forests integrate predictions from various models. In the random forest ensemble, every smaller model is a decision tree. It's an ensemble classifier that uses multiple decision tree models and may be applied to both regression and classification. The Random Forest classifier, as its name implies, is capable of including many decision trees on different dataset subsets and using the average to improve the dataset's predictive accuracy. Rather than depending solely on a single decision tree, the random forest utilizes the forecast from each tree to predict the final result, which is backed by the majority of predictions.

3.3.2 Decision Tree

This is done by using a decision tree classifier (Decision Tree Classifier) from the `scikit-learn` library. The training data ({X} features, {y} goal) is used to train the model. Decision trees require less pre-processing work to prepare the data than other algorithms. Normalizing data is not necessary for a decision tree. Scaling data is not necessary when using a decision tree. Additionally, the process of creating a decision tree is NOT significantly impacted by missing values in the data. Technical teams and stakeholders may both easily understand and comprehend a decision tree model.

3.3.3 Naive Bayes

The `scikit-learn` library's Gaussian Naive Bayes classifier uses Bayes' theorem to determine the likelihood that an object with a given set of features will belong to a particular class. Because of its "naive" assumption of feature independence given the class, computations are made simpler. The approach classifies objects based on their probabilities and is best suited for datasets with continuous attributes that follow a Gaussian distribution. For example, the algorithm combines these probabilities to forecast the class when recognizing a disease by symptoms; for example, it can identify a fungal infection based on its itching, skin rash, and skin peeling conditions.

$$P (A/ B) = (P (B/ A) * P (A))/ P (B)$$

3.3.4 GUI Setup

To develop a graphical user interface (GUI), this makes use of the `tkinter` package. For displaying results and facilitating user interaction, the GUI has buttons, labels, entry fields, and text fields.

IV. GENERALIZED PREDICTION MODEL

Fig 2: Prediction Model

A. Input

In constructing the model, it is assumed that users possess a clear understanding of the symptoms they are experiencing. The developed prediction system accommodates symptoms, allowing users to input their specific symptoms for analysis.

B. Data Preprocessing

Data preprocessing involves transforming raw data into a format easily interpretable by algorithms. This work utilizes data cleaning and data reduction techniques to refine the data for effective algorithmic interpretation.

C. Selected Models

The system employs three algorithms for disease prediction:

1. Decision Tree Classifier

2. Random Forest Classifier

3. Naïve Bayes Classifier

The conclusion of the study involves a comparative analysis, evaluating the performance of each algorithm on the considered database.

D. Output (Diseases)

Following training with the mentioned algorithms, the system generates a rule set. When users input their symptoms, the model processes them based on the established rule set, facilitating classification and predicting the most probable disease. If a match is found, the disease name is displayed; otherwise, it indicates "Not Found".

V. USE CASE DIAGRAM

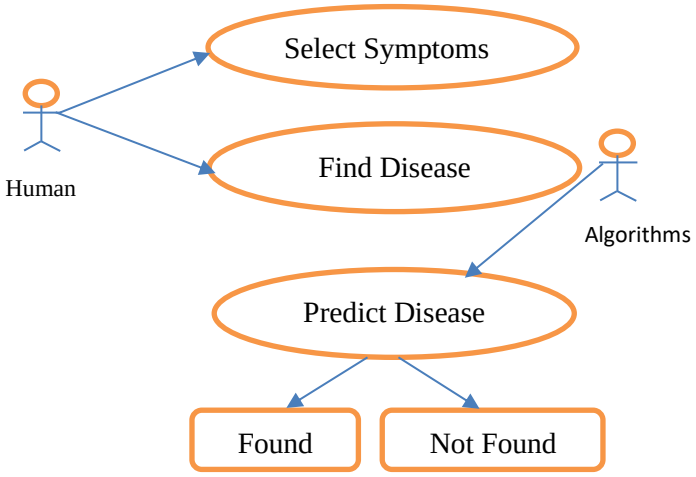

Figure 3: Use Case Diagram

Users can enter symptoms into the AI-Based Disease Prediction System for analysis. The system makes illness predictions using naive Bayes, random forest, and decision tree algorithms. The message "Found" indicates if a match has been found; if not, it says "Not Found." By inputting symptoms, users start the process, which prompts the machine learning system to analyze and forecast illnesses. Users are then informed of the outcome. By utilizing machine learning to forecast diseases effectively and accurately based on symptoms entered by the user, this streamlined method improves medical diagnostics.

VI. CONCLUSION

Over the course of machine learning's historical development and its applications in the medical field, a number of methods and approaches have evolved that make it possible to employ machine learning algorithms to simply analyze complex data. In-depth comparison research is conducted in this work to assess the efficacy of three algorithms on medical records; all algorithms achieve remarkable accuracy of up to 95%. Analyzing performance indicators including confusion matrices and accuracy ratings is part of the appraisal process. Artificial Intelligence (AI) is expected to play a more significant role in data analysis in the future due to the exponential rise of data made available by modern technologies.

It follows that machine learning can be used to track health in an efficient manner. By enabling people to

periodically check on their health at no expense, this strategy promotes wellbeing. It will be freely available to all users after a machine learning model has been developed. Users can take proactive steps by using the procedure, which gives them insights into their current health situation. Users may consult their healthcare providers for advice in cases of seriousness. People all throughout the world are empowered to live healthier lives thanks to this democratized approach to health tracking.

REFERENCES

[1] Suvendu Kumar Nayak, Mamata Garanayak, Sangram Keshari Swain, Sandeep Kumar Panda, And Deepthi Godavarthi, "An Intelligent Disease Prediction And Drug Recommendation Prototype By Using Multiple Approaches Of Machine Learning Algorithms", IEEE Access (Volume: 11), Electronic ISSN: 2169-3536, 11 September 2023, DOI: 10.1109/Access.2023.3314332

[2] Senthilkumar Mohan, Chandrasegar Thirumalai, And Gautam Srivastava, Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques, DOI: 10.1109/ACCESS.2019.2923707

[3] C.Geetha, Dr.AR. Arunachalam, "Evaluation based Approaches for Liver Disease Prediction using Machine Learning Algorithms", 2021 International Conference on Computer Communication and Informatics (ICCCI) | 978-1-7281-5875-4/21/$31.00 ©2021 IEEE |DOI: 10.1109/ICCCI50826.2021.9402463

[4] Sarma, D., Hossain, S., Mittra, T., Bhuiya, M. A. M., Saha, I., & Chakma, R. (2020). Dengue Prediction using Machine Learning Algorithms. 2020 IEEE 8th R10 Humanitarian Technology Conference (R10-HTC). DOI: 10.1109/R10-HTC49770.2020.9357035

[5] John Adeoye, J., Koohi-Moghadam, M., Choi, SW. *et al.* Predicting oral cancer risk in patients with oral leukoplakia and oral lichenoid mucositis using machine learning. *J Big Data* **10**, 39 (2023). https://doi.org/10.1186/s40537-023-00714-7

[6] J. P. Gupta, A. Singh and R. K. Kumar, "A computer-based disease prediction and medicine recommendation system using machine learning approach", *Int. J. Adv. Res. Eng. Technol. (IJARET)*, vol. 12, no. 3, pp. 673-683, 2021.
https://iaeme.com/Home/article_id/IJARET_12_03_062