International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 04 | April - 2023 Impact Factor: 8.176 ISSN: 2582-3930

DISEASE PREDICTION AND DRUG RECOMMENDATION SYSTEM USING MACHINE LEARNING

Nikitha.V[1], L.Pallavi[2],K.Sharan Kumar[3],M.Akhay Raj[4] , G.Prasanna Kumar[5]

[1,2,3,4]*B.Tech. Student, Department of Computer Science and Engineering,*

nikkynikitha18@gmail.com, pallavireddylattupally@gmail.com,kattakittu333@gmail.com, akshayrajmargam@gmail.com, prassannakumar.cse@nmrec.edu.in

[5]*Assistant Professor, Department of Computer Science and Engineering,*

Nalla Malla Reddy Engineering College, Hyderabad, India

***Abstract*: Medical decisions could be extremely specialized and difficult jobs due to alternative factors or in case of rare diseases. This problem is complex due to the large amount of data required to make accurate predictions, the complexity of disease development, and the variability between patients is difficult. And it is a very time consuming process. Additionally, there is a need to ensure that patient data is protected and used ethically, while also providing actionable insights for healthcare providers. Our proposed system is developed using python. This website is developed to help the doctors, patients to identify the disease. To reduce the large number of variables and find the most probable diseases by using the Random Forest, decision trees, SVM algorithms. And this system recommends the drug based on the factors we provide such as age, gender, bp etc. It is not only useful for doctors but also for the people who want to predict their disease.**

Keywords: ***disease prediction, SVM, Random Forest, drug recomendation***

1. Introduction

The medical field has greatly benefited from advanced computing technology, particularly in areas such as surgical representation and x-ray photography. However, while technology can aid in accurate diagnoses, there are various factors that must be taken into consideration, including medical records, weather conditions, atmosphere, blood pressure, and more. Although these variables are crucial to understanding the complete working process, no model has been able to analyze them successfully. To address this issue, medical decision support systems have been developed to assist doctors in making the correct diagnosis. Medical decision support systems involve the process of attempting to identify possible diseases or disorders and arriving at a diagnostic opinion. This opinion can indicate either a degree of abnormality on a continuum or a classification of abnormality. It is important to note that non-medical factors such as power ethics and financial incentives for patients or doctors can also influence the diagnostic opinion. The diagnostic opinion can take the form of a brief summation or an extensive formulation, even taking the form of a story or metaphor. It can also serve as a means of communication, such as computer code through which it triggers payment, prescription, notification, information, or advice.In order to diagnose a medical condition accurately, it is important to have knowledge of what is normal and to measure the patient's current condition. Automated decision support systems are rule-based systems that can automatically provide solutions to repetitive management problems. These systems can

greatly assist doctors in making accurate diagnoses and providing better patient care.

2. Literature Survey

The existing research on disease prediction includes various methodologies and techniques that have been explored to improve the prediction systems' accuracy and efficiency. For instance, in the study [1] After evaluating four different machine learning algorithms, it was found that the Decision Tree (DT) classifier was the most effective. As a result, the DT classifier was chosen to create the diabetes prediction machinery for a mobile application. [11] The proposed paper suggests the utilization of data mining techniques that rely on classification, such as Rule-based, Decision Tree, Naïve Bayes, and Artificial Neural Network. These techniques can effectively discover previously unknown patterns and relationships in large healthcare datasets.[8] It is essential to create a decision support system capable of predicting heart disease in patients. To find this need, the authors of this paper proposed an efficient associative classification algorithm that utilizes a genetic approach for heart disease prediction.

3. Existing System

Making medical decisions can be a complex and specialized task, especially in cases involving rare diseases or multiple alternative factors. There are various factors that can contribute to misdiagnosis, including stress, fatigue, incomplete information, and ignorance on the part of doctors. Standard algorithms may consider numerous variables such as prevailing conditions, medical history, family records, and other factors related to patient records. However, the sheer number of available hidden factors can make accurate diagnosis challenging. To address this issue, differential diagnosis methods can be used to identify the presence of an entity where multiple alternatives are possible, and to include candidate alternatives. This involves a process of elimination or obtaining information that reduces the probability of candidate conditions to negligible levels. This method consists of three steps:1)Gathering all information about the patient and creating a list of symptoms.2)Creating a list of all possible causes of the symptoms.3)Prioritizing the list by placing the most dangerous possible cause of symptoms at the top, and ruling out or treating possible causes beginning with the most urgently dangerous conditions using scientific methods.The term "rule out" in this context refers to using test methods or other scientific approaches to eliminate possible causes of symptoms.

4. Proposed System

Our proposed system is providing accurate and personalized predictions about the likelihood of a patient developing a particular disease and recommending appropriate drugs or treatments. Our proposed system is developed using python. This website is developed to help the doctors, patients to identify the disease. A user interface will be developed to provide a simple and intuitive way for healthcare providers to input the body factors of patients and receive predictions of the likelihood of disease. Employing the Random Forest, Decision Tree, Voting Classifier (SVM + Random Forest + Decision Tree) algorithms to decrease the enormous number of variables and identify the diseases that are most likely to occur. These algorithms are more suited to grouping more disorders. One unsupervised learning approach used to address the clustering issue is SVM. Different tests carried out on the patients will be used as clustering attributes.

5. Methodology

Random Forest Algorithm

Random Forest is a popular machine learning algorithm for both classification and regression tasks. It is classified as an ensemble learning method as it combines multiple decision trees to increase accuracy and reduce overfitting.
The Random Forest algorithm involves the following steps:
The dataset is first prepared for analysis by cleaning, pre-processing, and transforming the data into a format that is suitable for machine learning algorithms. Random subsets of features and training data are selected from the original dataset, and a set of decision trees are constructed recursively by splitting the dataset into smaller subsets based on the best split that maximizes information gain or reduces impurity.Now the Predictions are made by combining the predictions of individual decision trees through majority voting. For classification tasks, the class with the highest number of votes is selected, while for regression tasks, the average of

the predicted values is taken. Out-of-bag (OOB) Error Estimation: The algorithm estimates its performance by calculating the prediction accuracy of each tree on the training dataset that was not used to build that tree. The performance of the Random Forest model can be optimized by adjusting hyperparameters such as the number of trees, maximum depth of each tree, and number of features considered for each split.

Random Forest has several advantages over traditional decision trees, including improved accuracy, reduced overfitting, and the ability to handle high-dimensional datasets. It is widely used in various applications, such as healthcare, finance, and marketing, for predicting outcomes, detecting anomalies, and recommending actions.

SVM

SVM, also known as Support Vector Machine, is a machine learning algorithm used for supervised regression and classification analysis. Its objective is to locate a hyperplane that can optimally separate the data points of distinct classes or groups to enable the accurate classification of new, unseen data points. The SVM algorithm comprises various stages. First, the dataset is processed for analysis by performing tasks like data cleaning, preprocessing , and data transformation to a format compatible with SVM. Subsequently, pertinent features are chosen or derived from the data to train the model. SVM locates a hyperplane that maximizes the margin between the classes or groups, and the margin denotes the distance between the hyperplane and the nearest data points from each class.If the data is not linearly separable, SVM can use a kernel function to transform the data into a higher-dimensional space, enabling it to capture more complex patterns. After training the model, its performance is evaluated using metrics such as accuracy, precision, recall, and F1 score.SVM offers many benefits, including its capability to deal with high-dimensional data and both linearly separable and non-linearly separable data. SVM has various applications in areas such as image classification, bioinformatics, and text classification.

Decision Trees

Decision trees are a popular type of supervised machine learning algorithm that can perform both classification and regression analysis. They use a graphical representation of a series of decisions and their possible outcomes, with each internal node representing a decision based on the features of the data, and each leaf node representing a final classification or prediction.

To build a decision tree, the dataset must first be prepared by cleaning, pre-processing, and transforming the data into a suitable format for analysis. The most important features are then selected from the dataset, and the decision tree is built by recursively dividing the data based on the selected features. The performance of the decision tree model is then evaluated using various methods such as accuracy, precision, recall, and F1 score.

One of the main advantages of decision trees is their interpretability, which allows users to understand how the model arrives at its predictions. They can handle both categorical and numerical data, and are used in many applications such as disease prediction, credit scoring, and customer segmentation.

However, overfitting can be a potential issue with decision trees, where the model is too closely tailored to the training data and performs poorly on new, unseen data. Techniques such as pruning and ensemble methods are commonly used to address this problem.

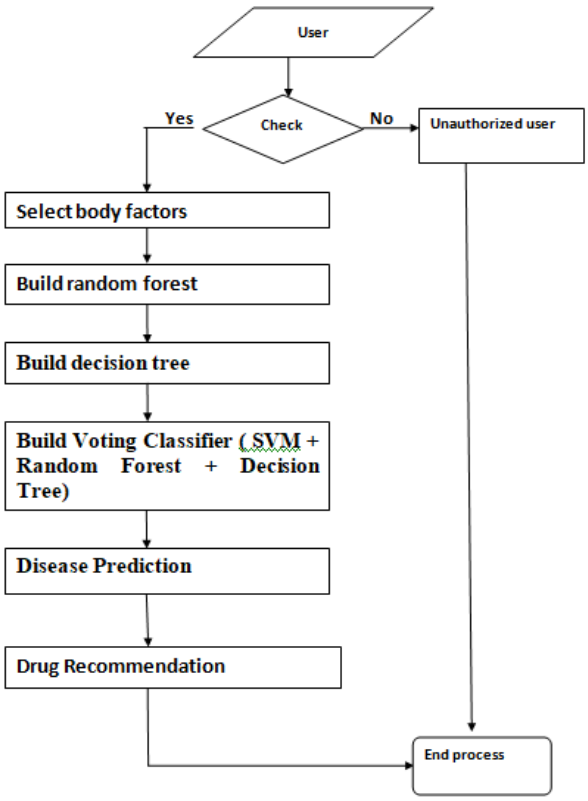


Fig 1 Architecture

7. Software and Hardware used

Software Used:

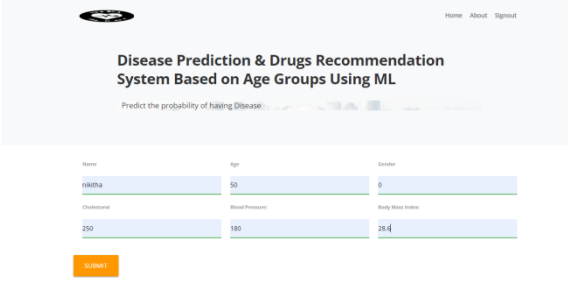
- Windows OS
- Python
- Jupyter Notebook
- Anaconda
- Flask

Hardware Used:

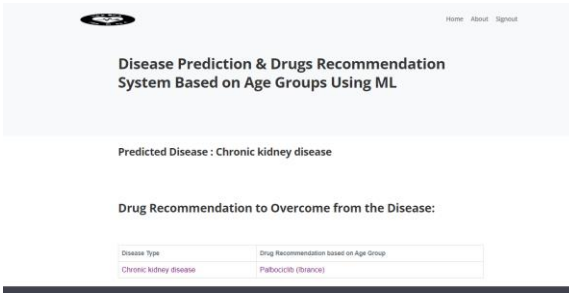
- Hard Disk – 1 TB
- Memory – 4 GB RAM

8. Results

The results for disease prediction and drug recommendation are obtained by taking the human body factors as inputs to the system. By using the machine learning Algorithms such as Random Forest ,Decision Trees, SVM to predict the disease.



The drug is recommended according to the predicted disease by using algorithms. The below image is the output for the above inputs.



9. Conclusion and Future Work

This paper gave a diagram of utilization of information machine learning procedures in regulatory, clinical, inquire about, furthermore, instructive parts of Clinical Predictions. This paper set up that while the current down to earth utilization of information machine learning in well being related issues is constrained, there exists an extraordinary potential for information mining systems to enhance different parts of Clinical Predictions. Besides, the inescapable ascent of clinical information will build the potential for information mining systems that enhances the quality and reduces cost of social insurance. This system has large scope as it has the following features which are: • Automation of Disease Diagnosis. • Paper free work helping the environment. • To increase the efficiency, accuracy for the patients to help them in future. • Managing the information related to diseases.

References

1] K Sowjanya, Ayush Singhal, Chaitali Choudhary, MobITest. A machine learning based system for predicting diabetes risk using mobile devices", Advance Computing Conference (IACC) 2015 IEEE International, pp. 397-402, 2015.

[2] Ameera M. Almasoud, Hend S. Al-Khalifa, Abdulmalik Al-Salman, "Recent developments in data mining applications and techniques", Digital Information Management (ICDIM) 2015 Tenth International Conference on, pp. 36-42, 2015

[4] Bharathan Venkatesh, Danasingh Asir Antony Gnana Singh, Epiphany Jebamalar Leavline, Advance in Intelligent Systems and Computing, vol. 517, pp. 633, 2017, ISSN 2194-5357, ISBN 978-981-10-3173-1

[5] Han, J. and Gao, J. (2009). Research Challenges for Data Mining in Science and Engineering. Next Generation of Data Mining, pages 1– 18.

[6] Text Data Mining of Care Life Log by the Level of Care Required Using KeyGraph Muneo Kushima, Kenji Araki, Tomoyoshi Yamazaki, Sanae Araki, Taisuke Ogawa, Noboru Sonehara Proceedings of the International MultiConference of Engineers and Computer Scientists 2017 Vol I, IMECS 2017, March 15 - 17, 2017, Hong Kong.

[7] Sun, Y., Han, J., Yan, X., and Yu, P. S. (2012). Mining Knowledge from Interconnected Data: A Heterogeneous Information Network Analysis Approach. Proceedings of the VLDB Endowment.

[8] M. Akhil jabbar & Dr. Priti Chandrab “Heart Disease Prediction System using Associative

Classification and Genetic Algorithm" International Conference on Emerging Trends in Electrical, Electronics and Communication Technologies-ICECIT, 2012.

[9] Nikhil N.Salvithal" Appraisal Management System using Data mining "International Journal of Computer Applications (0975 – 8887) Volume 135 – No.12, February 2016.

[10] DeFronzo Ralph. From the triumvirate to the ominous octet: a new paradigm for the treatment of type 2 diabetes mellitus. Diabetes. 2009;58:773-95.

[11] V.Krishnaiah & Dr.G.Narsimha, Dr.N.Subhash Chandra "Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques"(IJCSIT)International Journal of Computer Science and Information Technologies, Vol. 4 -No.1,2013, 39 - 45.

[12] Jaimini Majali, Rishikesh & Niranjan, Vinamra Phatak "Data Mining Techniques For Diagnosis And Prognosis Of Cancer" International Journal of Advanced Research in Computer and CommunicationEngineering Vol. 4, Issue 3, March 2015.

[13] Data Mining Techniques to Predict Diabetes Influenced Kidney Disease Swaroopa Shastri, Surekha, SaritaInternational Journal of Scientific Research in Computer Science, Engineering and Information Technology © 2017 IJSRCSEIT | Volume 2 | Issue 4 | ISSN : 2456-3307.

[14] Tanvi Sharma, Anand Sharma & Vibhakar Mansotra "Performance Analysis of Data Mining Classification Techniques on Public Health Care Data" International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization) Vol. 4, Issue 6, June 2016.