Using Big Data Analytics for Advanced Disease Prediction Module

Saksham Sharma
Kalinga Institute of Industrial Technology
Bhubaneswar ,Odisha
Saksham2810200@gmail.com

Om Ashok Rade
Kalinga Institute of Industrial Technology
Bhubaneswar ,Odisha
om.rade28@gmail.com

Kaustubh Srivastava
Kalinga Institute of Industrial Technology
Bhubaneswar ,Odisha

***Abstract*—The emergence of big data analytics and machine learning techniques has revolutionized the field of disease prediction and diagnosis. This paper explores the application of various machine learning algorithms, including Naive Bayes, Decision Trees, Random Forests, and Gradient Boosting, in predicting diseases based on patient-reported symptoms. The vast amount of patient data, encompassing symptoms, medical histories, and laboratory results, can be leveraged through advanced analytical techniques to uncover valuable insights and patterns. By training these algorithms on large datasets containing information about symptoms and corresponding diagnoses, they can learn to recognize the underlying relationships between symptoms and diseases. When presented with a new set of symptoms, the trained models can predict the most likely disease, assisting healthcare professionals in making informed decisions, recommending further tests, and suggesting appropriate treatments. The paper highlights the significance of big data analysis in disease prediction, explains the methodological approach, provides a practical example, and discusses the challenges and future directions of integrating machine learning and big data analytics in healthcare. The integration of these technologies holds immense potential for improving patient outcomes, enabling personalized medicine, and enhancing the efficiency of healthcare systems.**

***Index Terms*—Disease prediction, machine learning, big data analytics, healthcare informatics, Naive Bayes, Decision Trees, Random Forests, Gradient Boosting, symptoms,diagnosis.**

I. Introduction

In the era of big data, the health-care industry has witnessed a significant transformation in disease prediction and diagnosis. The vast amount of patient data, including symptoms, medical histories, and laboratory results, can be leveraged through advanced analytical techniques to uncover valuable insights and patterns [1]. Big data analysis plays a crucial role in disease prediction from symptoms, enabling health-care professionals to make informed decisions and provide timely interventions [2].

Disease prediction from symptoms involves the utilization of machine learning algorithms to analyze patient data and identify patterns that correlate with specific diseases. By training these algorithms on large datasets containing information about symptoms and corresponding diagnoses, they can learn to recognize the underlying relationships between symptoms and diseases [3]. The process typically involves data pre-processing, feature engineering, algorithm selection, model training, and evaluation. During data pre-processing, the raw patient data is cleaned, formatted, and transformed into a suitable representation for analysis. Feature engineering involves selecting or creating relevant features from the data that are informative for the prediction task. Subsequently, appropriate machine learning algorithms, such as Naive Bayes, Decision Trees, Random Forests, or Gradient Boosting, are chosen based on the characteristics of the data and the desired performance metrics [4]. The selected algorithm is then trained on the preprocessed data, allowing it to learn the patterns and relationships between symptoms and diseases. During the training process, the algorithm adjusts its internal parameters to minimize the prediction error on a validation dataset. Once trained, the model can be evaluated on a separate test dataset to assess its performance and generalization capabilities [5]. When presented with a new set of symptoms, the trained model can predict the most likely disease based on the patterns it has learned from the training data. This prediction can then be used by healthcare professionals to support decision-making, recommend further tests, or suggest appropriate treatments.

Consider a scenario where a patient presents with symptoms such as fever, cough, and fatigue. By inputting these symptoms into a disease prediction algorithm, the model can analyze the patterns and relationships it has learned from past data [6]. Based on its analysis, the algorithm may predict that the patient is likely suffering from influenza or a similar respiratory illness. This prediction can then assist healthcare professionals in making an informed diagnosis and prescribing appropriate treatment or further tests [7]. The process begins with gathering the patient's symptoms and relevant medical history.

This data is then preprocessed and transformed into a format suitable for the chosen machine learning algorithm. For instance, the algorithm may require the symptoms to be encoded as numerical values or binary indicators. Once the data is prepared, it is fed into the trained disease prediction model. The model, which could be based on techniques like Naive Bayes, Decision Trees, Random Forests, or Gradient Boosting, evaluates the input data against the patterns it has learned from the training dataset [8].

By identifying the correlations between the symptoms and potential diseases, the model can provide a ranked list of probable diagnoses. The healthcare professional can then review the model's predictions, considering the patient's overall medical history and any additional tests or examinations. The predicted diagnosis can guide further investigations, aid in ruling out alternative conditions, and inform the selection of appropriate treatments or medications [9].

Disease prediction from symptoms using machine learning algorithms holds immense potential in revolutionizing healthcare delivery [10]. By leveraging the power of big data analysis and advanced computational techniques, healthcare professionals can gain valuable insights, make accurate predictions, and provide timely interventions. As technology continues to evolve, the integration of machine learning and big data analytics in disease prediction will become increasingly crucial, ultimately leading to improved patient outcomes and more efficient healthcare systems [11].

The aim of this paper is to explore the potential of machine learning algorithms in disease prediction from symptoms, leveraging the power of big data analysis in the healthcare domain.

II. Methods

"Disease Prediction from Symptoms" is a medical diagnostic system that leverages machine learning algorithms to predict potential diseases based on reported patient symptoms. In this system, patients input their symptoms, and the application employs multiple algorithmic techniques, including Naive Bayes, Decision Trees, Random Forests, and Gradient Boosting, to analyze the data and generate disease predictions.

- Naive Bayes

- Decision Tree
- Random Forest
- Gradient Boosting

Each function of each algorithm:-

The Naive Bayes algorithm is a probabilistic classifier that calculates the probability of each disease given the set of symptoms reported by the patient, based on Bayes' theorem and the assumption of feature independence.

Decision Trees are hierarchical models that recursively partition the input space based on the most discriminative features, in this case, symptoms, to construct a tree-like structure for predicting the target disease.

Random Forests are ensemble learning methods that construct multiple decision trees from randomly sampled subsets of the data and features, with the final prediction being the mode of the predictions from all individual trees, effectively reducing overfitting and improving overall accuracy.

Gradient Boosting is another ensemble technique that iteratively trains weak learners, such as decision trees, on the residuals of the previous model, gradually improving the overall model's performance by combining the weak learners in an additive manner.

For a patient reporting fever, headache, body aches, and fatigue, the Naive Bayes algorithm calculates disease probabilities based on symptom independence. The Decision Tree recursively splits data on discriminative symptoms like fever presence. Random Forests create ensembles of decision trees on random subsets. Gradient Boosting iteratively trains weak learners on residuals from previous models. Combining outputs from these algorithms, leveraging Naive Bayes' probabilistic approach, Decision Trees' hierarchical splitting, Random Forests' ensemble diversity, and Gradient Boosting's iterative residual minimization, the system could robustly predict a viral illness like influenza for the given symptom set.

By employing these diverse algorithmic approaches, the system aims to provide accurate and reliable disease predictions based on the reported symptomatology, potentially assisting medical professionals in the diagnostic process and enabling more informed decision-making for patient care..

III. Results and Discussion

Within the realm of medical data training, a dataset comprising 149 diseases and 405 symptoms is utilized. Each row within this dataset pertains to distinct diseases or conditions, whereas the columns correspond to a range of symptoms or health indicators.

The binary values (0 or 1) within the cells signify the presence (1) or absence (0) of specific symptoms for each disease or condition. This structured dataset holds potential for constructing decision trees or other machine learning models aimed at aiding in disease diagnosis, symptom recognition, and risk evaluation.

The dataset offers comprehensive coverage of a wide range of medical conditions, including Alzheimer's, cancers, infections, and other disorders. Its symptom diversity allows for a multi-dimensional analysis of disease relationships. The binary encoding design focuses on categorical or Boolean classification tasks, allowing for a focus on symptom presence or absence. This structured data could provide insights into patterns, correlations, and decision rules, aiding healthcare professionals in understanding disease relationships and improving diagnostic methods or treatment strategies.

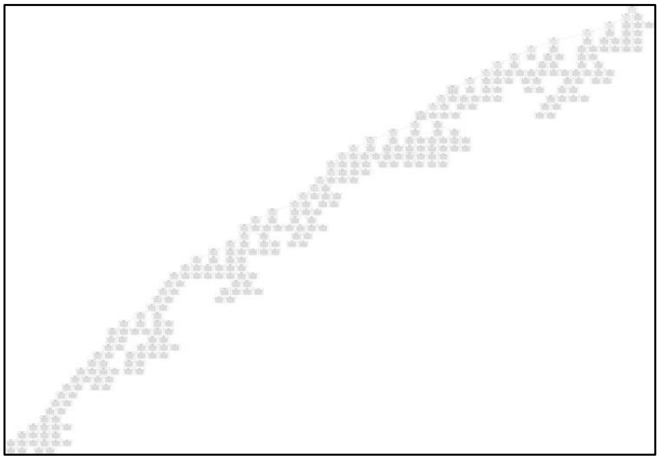

Fig 1.1. A complex Decision Tree of diseases with represent to Symptoms.

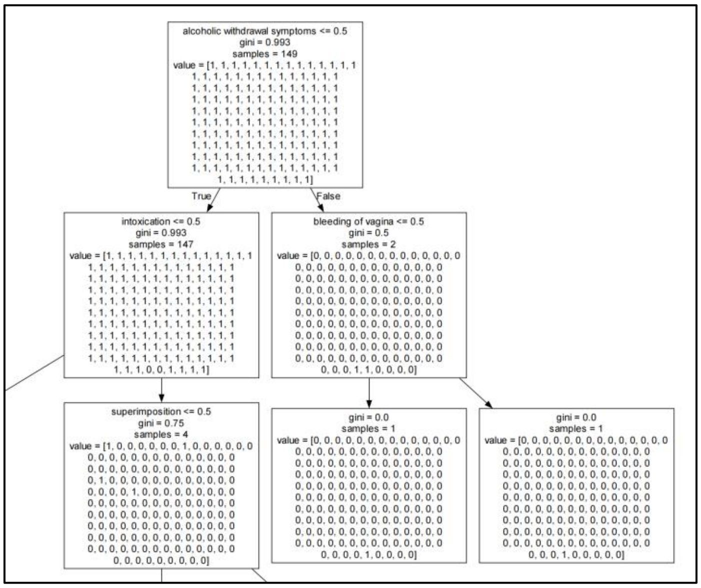

Fig 2 Top nodes of the Decisions Tree

Using the trained data, a decision tree is created. The tree has interconnected nodes that represent specific features or conditions being evaluated. Each node contains information such as the feature name, the threshold or gini value used for evaluation, the lower the gini value, the more accurate the disease prediction, and the number of samples associated with that node. The branches from each node represent possible outcomes or decisions based on the feature evaluation. The leaf nodes at the bottom of the tree represent the final classifications or decisions made by the model. The values assigned to the leaf nodes are binary (0 or 1), indicating the model's output or prediction.

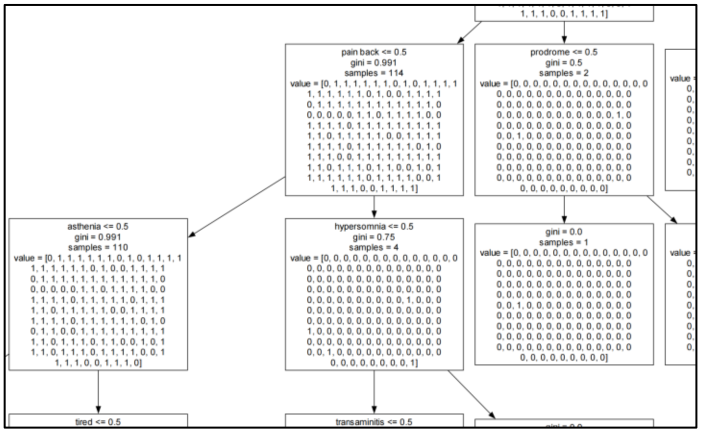

Fig.3 Middle part of the Decision Tree

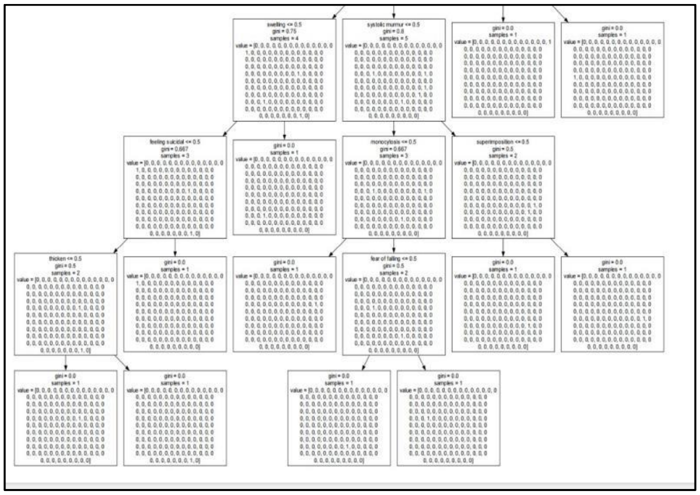

Fig.4 End nodes of Decision Tree

In terms of the results, the decision tree appears to be classifying or evaluating some kind of input data based on various features or conditions. The final output values at the leaf nodes suggest that the system is producing numerical results or predictions based on the input data and the decision-making process represented by the tree.

By utilizing Decision Tree classification, we can effectively forecast diseases based on their symptoms. This enables us to present both the actual and predicted diseases. Furthermore, by inputting a dataset of prescribed drugs for various diseases, we can associate drugs with specific ailments. However, for diseases where we are unable to provide corresponding medications, we advise consulting a doctor. This innovative system can be considered as a compact virtual doctor.

IV. Conclusions

Based on the study findings, extensive datasets have been gathered, which establish connections between symptoms and particular diseases. These datasets are subjected to prepressing and then utilized to train machine learning algorithms. It is worth noting that the medical field has limited resources. However, the use of decision trees can enhance disease prediction based on symptoms, enabling the display of both potential and confirmed illnesses. By incorporating a medication database, users can further investigate potential treatments for specific conditions. Nevertheless, it is crucial to emphasize the significance of consulting a healthcare professional, particularly in cases where no medication is recommended. This system holds the potential to serve as a convenient self-assessment tool.

Acknowledgment

The authors would like to express their sincere gratitude to KIIT University for providing the opportunity to work on this project on disease prediction using symptoms and drug prescription. We are thankful for the support and resources that KIIT University has provided, which have been instrumental in the development of this research.

References

[1] Qiu, J., Wu, Q., Ding, G., Xu, Y., & Feng, S. (2016). A survey of machine learning for big data processing. EURASIP Journal on Advances in Signal Processing, 2016(1), 67.

[2] Raghupathi, W., & Raghupathi, V. (2014). Big data analytics in healthcare: promise and potential. Health Information Science and Systems, 2(1), 3

[3] Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., ... & Wang, Y. (2017). Artificial intelligence in healthcare: past, present and future. Stroke and Vascular Neurology, 2(4), 230-243.

[4] Deo, R. C. (2015). Machine learning in medicine. Circulation, 132(20), 1920-1930.

[5] Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the future—big data, machine learning, and clinical medicine. The New England Journal of Medicine, 375(13), 1216-1219.

[6] Kononenko, I. (2001). Machine learning for medical diagnosis: history, state of the art and perspective. Artificial Intelligence in Medicine, 23(1), 89-109.

[7] Shirabad, J. S., Sadreddini, M. H., & Das, S. (2020). Machine learning applications in medical diagnosis. In Machine Learning for Predictive Analysis (pp. 61-85). Academic Press.

[8] Deo, R. C. (2015). Machine learning in medicine. Circulation, 132(20), 1920-1930.

[9] Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the future—big data, machine learning, and clinical medicine. The New England Journal of Medicine, 375(13), 1216-1219.

[10] Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. Nature Medicine, 25(1), 44-56.

[11] Beam, A. L., & Kohane, I. S. (2018). Big data and machine learning in health care. Jama, 319(13), 1317-1318.