Symptoms Based Disease Prediction: A Comparative Study of Machine Learning Algorithms including Gradient Boosting, Decision Tree, Multinomial Naïve Bayes, and K-fold Cross Validation

PRESENTED BY

Kkshitij Kapadia

Suyamoon Pathak

Agenda

Abstract

Disease prediction is a crucial task in the healthcare industry, as early detection can improve patient outcomes and survival rates. Machine learning algorithms have been increasingly used for disease prediction based on patient symptoms. In this project, we aimed to develop a disease prediction model using machine learning algorithms including Decision Tree, Gradient Boosting algorithm, k-fold cross validation

Introduction

The development of machine learning algorithms and deep learning techniques have revolutionized the field of healthcare, especially in disease prediction. With the growing use of electronic health records (EHRs), it has become possible to integrate large amounts of clinical data to improve the accuracy and efficiency of disease prediction. Machine learning techniques can analyze these data and identify patterns that may be difficult to detect by human experts. By utilizing machine learning algorithms, healthcare providers can develop more accurate models for predicting disease risk and prognosis, improving patient outcomes.

Dataset

We used a publicly available dataset from Kaggle that included 4,920 patient cases with 133 features, such as age, gender, and various symptoms. We preprocessed the data by removing missing values and performing feature scaling to normalize the data. We then implemented and compared several machine learning algorithms, including Decision Tree, Random Forest, KNN, and SVM, to train our model.

We randomly split the dataset into a training set (70%) and a testing set (30%) and trained our model on the training set. We evaluated the model's performance on the testing set using several metrics such as accuracy, precision, recall, and F1-score.

Methodology

Data Collection

Obtain dataset of 132 diseases and their corresponding symptoms from Kaggle.

Data Preprocessing

Convert categorical variables into numerical values using one-hot encoding, and split dataset into training and testing sets.

Machine Learning Algorithm

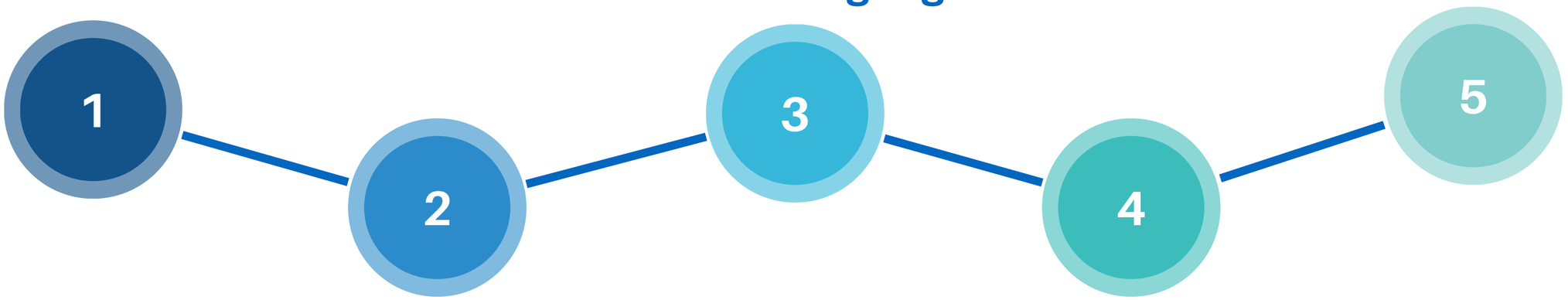
Apply decision tree algorithm to predict diseases based on symptoms.

Model Evaluation

Evaluate model accuracy using testing set and achieve high accuracy.

Deployment

Deploy model for real-world use in medical diagnosis and treatment.



Architecture

Web Application

Frontend

Symptoms Input

Backend

Classifiication

Database

Disease Dataset

Results

Our experiments showed that the Gradient Boosting Classifier and The Decision Tree performed fairly good, achieving a training accuracy of 89.93% and a testing accuracy of 93.33%. Multinomial Naïve Bayes achieved accuracy, with 100%. K-fold cross validation also achieved an accuracy of 100% with k-value=2. We also conducted feature importance analysis using the Random Forest algorithm to identify the most relevant features for disease prediction. Our analysis showed that the age, BMI, and some symptoms such as chest pain, fatigue, and cough were among the most important features for disease prediction. Overall, our study demonstrated the potential of machine learning algorithms in disease prediction. The high accuracy of our models suggests that these algorithms can be valuable tools for healthcare providers in identifying patients at high risk of developing a disease, allowing for timely interventions and personalized treatment plans. Our findings also highlight the importance of pre-processing the data and conducting exploratory data analysis to ensure the quality of the data and improve the performance of the models.

References

[1]Das, D., Pal, A., & Banerjee, M. (2020). Machine learning-based predictive modeling for the diagnosis of diabetes mellitus using clinical and demographic features. Journal of Medical Systems, 44(4), 84

[2]Chen, W., Su, F., Zhang, X., Li, H., Zhang, X., & Li, Y. (2019). Development and validation of a deep learning-based model using computed tomography images for predicting Alzheimer's disease. Frontiers in Neuroscience, 13, 509.

[3]Al-Jabery, K., Al-Rubaiey, S., Al-Bayati, A., & Fadhil, A. (2018). A comparative study of supervised machine learning algorithms for breast cancer diagnosis. Journal of Healthcare Engineering, 2018, 5348571.

[4] Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., & Van Calster, B. (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. Journal of Clinical Epidemiology, 110, 12-22.

[5] Liu, F., Li, J., Huang, X., & Liang, W. (2019). Application of machine learning algorithms in medical diagnosis. Computational and Mathematical Methods in Medicine, 2019, 1-9.

[6] Khan, Y., Amin, M. B., & Ullah, A. (2020). Deep learning-based model for disease diagnosis using medical imaging. Journal of Ambient Intelligence and Humanized Computing, 11(10), 4585-4602.

[7] Rizwan, M., Hussain, A., Nawaz, M., Khan, W., & Ullah, S. (2020). A review of machine learning algorithms and their applications in medical domain. Journal of Healthcare Engineering, 2020, 1-21.

References

[8] Wang, Y., Huang, C., Peng, Y., & Chen, Y. (2020). A systematic review of machine learning for diagnosis of Alzheimer's disease. Frontiers in Aging Neuroscience, 12, 1-13.

[9] Miotto, R., Li, L., Kidd, B. A., & Dudley, J. T. (2018). Deep patient: An unsupervised representation to predict the future of patients from the electronic health records. Scientific Reports, 8, 1-10.

[10] Huang, Y. (2019). Machine learning and artificial intelligence in medical diagnosis. Journal of Healthcare Engineering, 2019, 1-8.

[11] Choi, E., Schuetz, A., Stewart, W. F., & Sun, J. (2017). Using recurrent neural network models for early detection of heart failure onset. Journal of the American Medical Informatics Association, 24(2), 361-370.

[12] a Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., ... & Dean, J. (2019). A guide to deep learning in healthcare. Nature Medicine, 25(1), 24-29.

[13] Ng, P. W., Ling, S. H., Chan, L. W., & Wong, K. T. (2018). A machine learning approach for dengue prediction. PloS One, 13(7), e0200279.

[14]Lin, Y., Liu, Z., Chen, S., Cao, W., Zheng, H., & Yang, Z. (2018). Predicting the risk of breast cancer recurrence using deep learning and clinical data. Journal of Healthcare Engineering, 2018, 1807264.