**Disease Prediction Using Machine Learning**

Abstract

In recent years, the integration of machine learning techniques into healthcare systems has shown promising results in disease prediction, diagnosis, and prognosis. This paper presents a comprehensive review of the methodologies, challenges, and advancements in disease prediction using machine learning algorithms. The primary focus is on the application of various machine learning models, including but not limited to support vector machines, decision trees, logistic regression, K Nearest Neighbours and Naïve Bayes, in predicting diseases across different medical domains.

The paper begins by outlining the significance of disease prediction in early diagnosis and emphasizing the potential of machine learning to enhance predictive accuracy and clinical outcomes. It then discusses the process of data acquisition, preprocessing, and feature selection, highlighting the importance of high-quality data and feature representation in building robust predictive models. Furthermore, it reviews different evaluation metrics and validation techniques commonly employed to assess the performance of disease prediction models.

The main body of the paper delves into a comprehensive analysis of state-of-the-art machine learning techniques applied to various diseases, including cardiovascular diseases, cancer, diabetes, infectious diseases, neurological disorders, and others. The paper examines the specific challenges, datasets, features, and algorithms utilized, along with the reported accuracies and limitations of existing models.

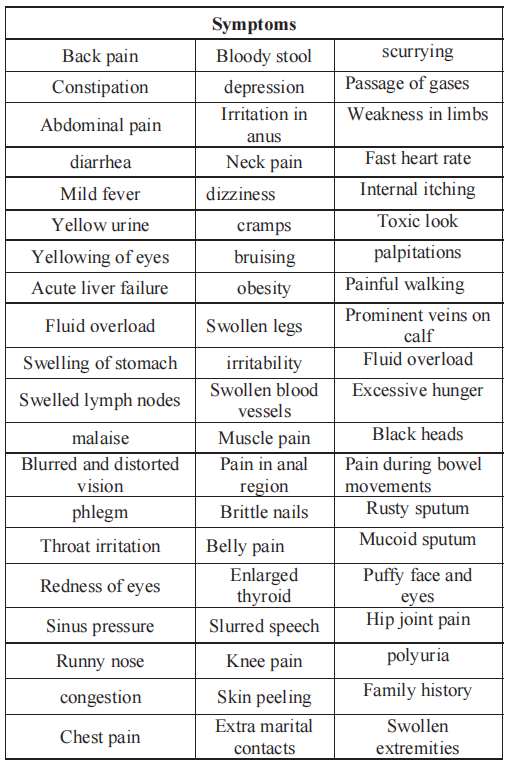
Finally, the paper concludes with a discussion on future directions and potential research avenues in disease prediction using machine learning, such as the integration of multimodal data sources, personalized medicine approaches, and real-time monitoring systems. It underscores the need for interdisciplinary collaboration between data scientists, clinicians, and policymakers to harness the full potential of machine learning in revolutionizing healthcare delivery and improving patient outcomes.

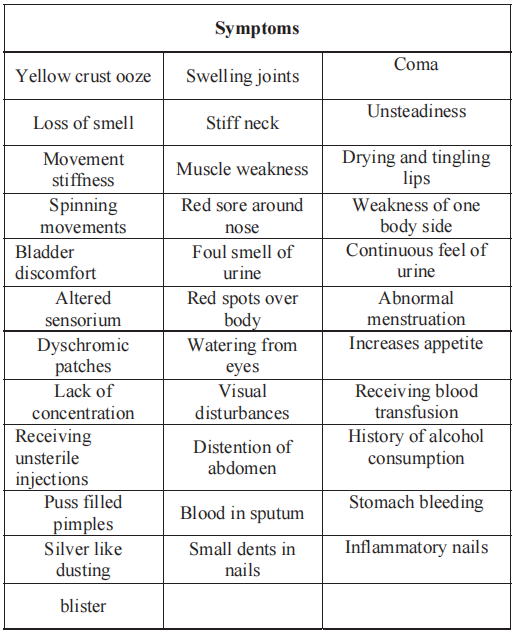
Overall, this review provides valuable insights into the current landscape of disease prediction using machine learning and offers guidance for researchers, clinicians, and policymakers aiming to leverage these technologies for advancing personalized medicine and public health initiatives.

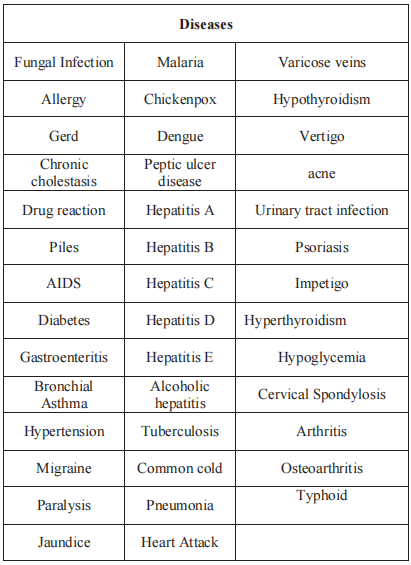
Introduction

Dataset

The dataset we have considered consists of 132 symptoms, the combination or permutations of which leads to 41 diseases. Based on the 4920 records of patients, we aim to develop a prediction model that takes in the symptoms from the user and predicts the disease he is more likely to have.

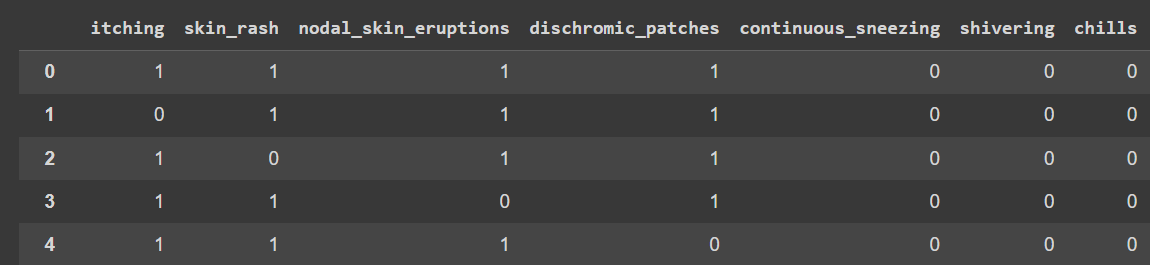






Data PreProcessing:

We convert categorical data to numerical data where symptoms are the columns and 1 represents symptom is present, 0 represent symptom absent.



Literature Review

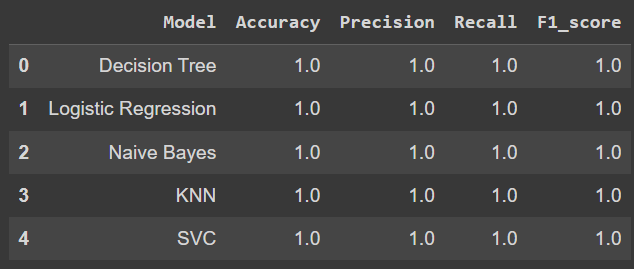
Proposed Work

Models Selected

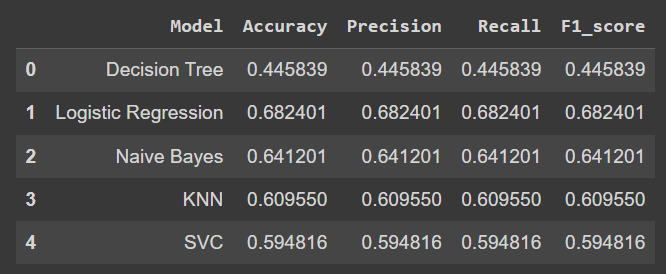
1. Decision Tree
2. Logistic Regression
3. Naïve Bayes
4. K Nearest Neighbours
5. Support Vector Machine

Analysis

Testing on Trained Data



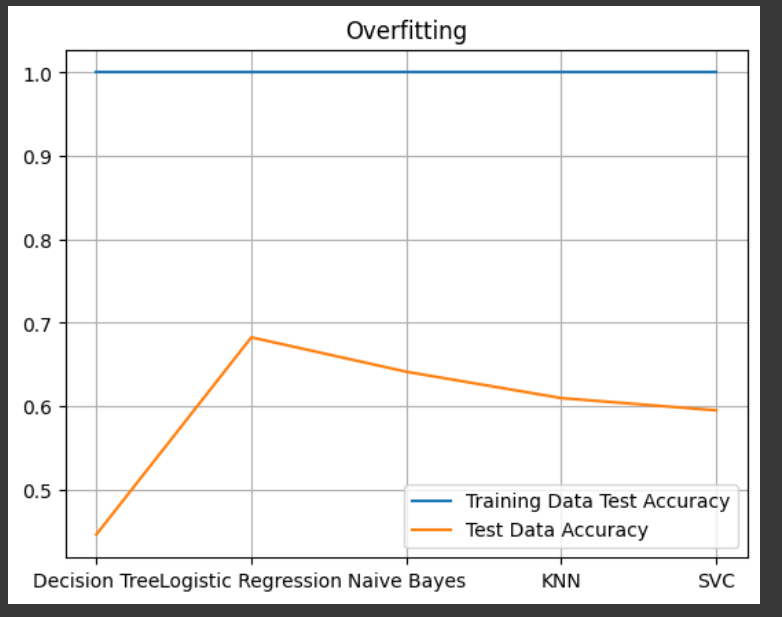
Testing on Test Data



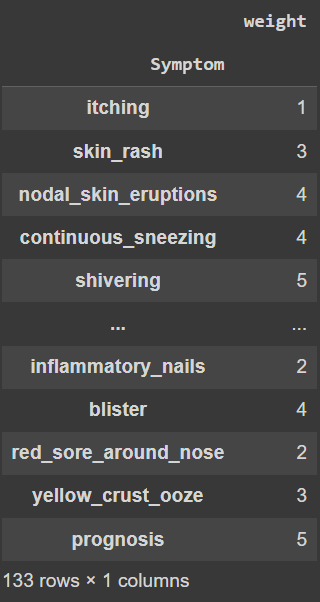
Overfitting Observed.

Limitation

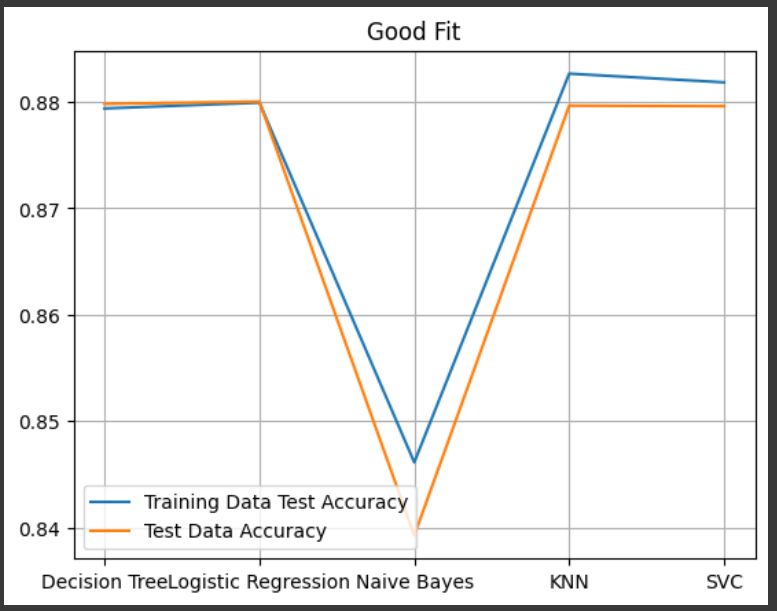
The models memorize the training data but fail on testing. There is overfitting.



Feature Weight Modification

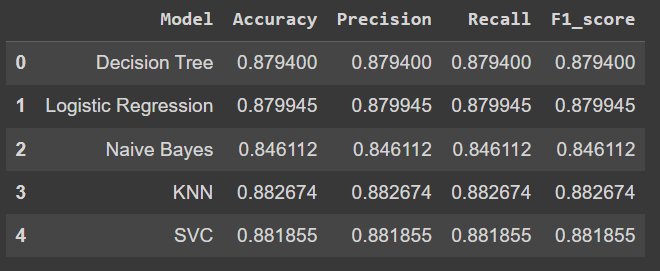


Model Trained on updated data provide good fit.



Result

Final Accuracy Achieved



Conclusion

Reference