INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor : 5.725

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 02, February 2024, pp : 444-447

PREDICTION OF LIVER DISEASE PATIENTS USING NAÏVEBAYES AND RANDOM FOREST

K. Jenifer[1], Mrs. M. Saranya[2]

[1]Department of Computer Science Sri Kaliswari College, Sivakasi, Tamil Nadu, India.

[2]M.Sc., M.Phil., Department of Computer Science Sri Kaliswari College, Sivakasi, Tamil Nadu, India.

DOI: https://www.doi.org/10.58257/IJPREMS32768

ABSTRACT

Globally liver disease is a major health concern. Improving patient outcomes depends heavily on early identification. Create a predictive model that can help identify those who are risk of liver disease by using these algorithms. The dataset includes a range of clinical and laboratory indicators, including alkaline phosphates levels, age, gender, total and direct bilirubin levels, and more.

The machine learning models are trained and tested using these features as inputs. Analysing the provided input parameters, the Naïve Bayes algorithm determines a probability that a patient has liver disease. It does this by assuming that features are independent of one another. In contrast, Random Forest builds a collection of decision trees and aggregates their forecasts to provide a single prediction. Through the process of fitting data to a Naive Bayes and Random forest calculates the possibility that liver disease patients prediction .Several criteria are used to assess each algorithm's performance, including accuracy, precision, recall. The outcomes of these assessments will be used to identify the best algorithm for hepatic disease prediction.

Keywords— Liver disease, prediction, Random Forest, Naïve Bayes.

1. INTRODUCTION

The early diagnosis of liver disease poses a challenging task for medical professionals due to inconspicuous symptoms complication arising from liver disease often unnoticed as the liver continues to function normally despite partial damage. While even experienced practioners may struggle to determine the presence of symptoms. It is possible to detect early warning sign.

The key to significantly prolonging a patients life span lies in early diagnosis, and in the modern era, machine learning techniques play a crucial role in preventive medicine by predicting disease from healthcare database. It plays a significant role in medical decision making and specialize in integrating multiple risk factors into a predictive tool with the gradual increase in health care data. Machine learning provides rapid access to analyse messive amount of data various industries are utilizing machine learning enhance medical diagnostics.

This paper aims to predict liver disease using the machine learning techniques of naive bayes classifier and random forest ,preprocessing techniques such us removing duplicate values, handling null values, encoding categorical data This model can serve as a valuable asset for clinical decision making in a business setting. Evoking a sense of mild positivity while maintaining a formal tone.

2. PROPOSED ALGORITHM

A. Random Forest– The Random Forest algorithm is a well-known machine learning method that belongs to the ensemble learning category. It operates by creating numerous decision trees in the training phase and provides the mode of the classes (for classification) or the average prediction (for regression) of the individual trees. Each decision tree in the Random Forest is developed using a subset of the training data and a random selection of features.

This randomness aids in reducing overfitting and enhancing the model's ability to generalize. During prediction, the algorithm combines the outcomes from all the trees to generate a final prediction. Random Forest is recognized for its high accuracy, resilience to noise, and capacity to handle large datasets with high dimensionality. It is extensively utilized in various machine learning applications such as classification, regression, and feature selection.

B. Naïve Bayes– Naive Bayes is a probabilistic classifier that is both simple and powerful. It is based on Bayes' theorem and operates under the assumption of independence between features. Despite this "naive" assumption, Naive Bayes has found widespread use in various applications such as text classification and spam filtering due to its efficiency and simplicity. The algorithm calculates the probability of a given instance belonging to a specific

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor : 5.725

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 02, February 2024, pp : 444-447

class by considering the probabilities of the features. It assumes that the presence of a particular feature in a class is independent of the presence of any other feature, hence the term "naive" assumption. Naive Bayes is known for its computational efficiency and ability to handle high-dimensional data effectively. It is particularly well-suited for text data, where the independence assumption can still yield accurate results. Although it may not capture complex relationships between features, Naive Bayes remains a popular choice for quick and reliable classification tasks.

C. EXPERIMENT AND RESULT

Data collection serves as the initial step in any data analysis process, where relevant data is acquired from various sources. Kaggle, a well-known data science platform, is a valuable resource for obtaining datasets from diverse domains. These datasets, including those from Kaggle, form the basis for subsequent analysis tasks. After data collection, data exploration is conducted to gain a comprehensive understanding of the dataset's characteristics and patterns.

This phase involves employing various techniques, such as descriptive statistics, to summarize the dataset's central tendencies and dispersion by calculating measures like mean, median, standard deviation, minimum, and maximum. Additionally, data visualization techniques are used to visually represent the distribution and relationships between variables. Scatterplots, lineplots, boxplots, barplots, countplots, and pointplots are generated, each providing unique insights into the dataset. Furthermore, missing value analysis is performed as part of data exploration to identify any missing or null values in the dataset.

This analysis is crucial for ensuring data integrity and reliability throughout the subsequent analysis stages, as appropriate strategies can be implemented to handle missing data effectively. Overall, data collection and exploration lay the foundation for informed decision-making and actionable insights in the data analysis process.

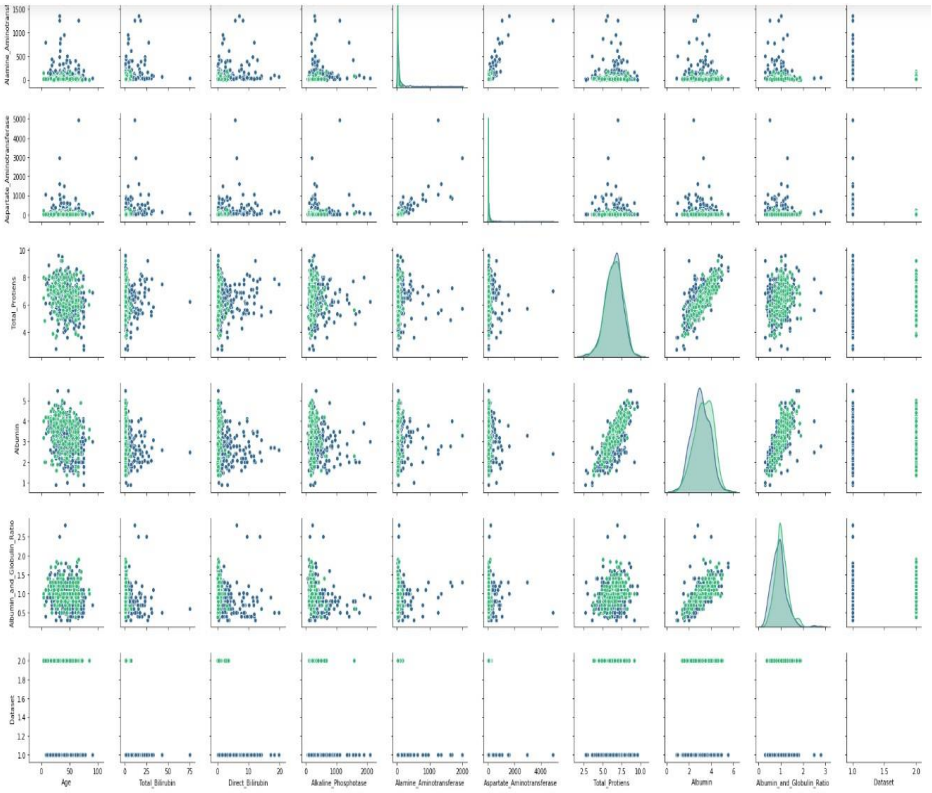

Fig. 1. Data Analysis

Data preprocessing plays a crucial role in the data analysis pipeline as it involves transforming and preparing raw data for further analysis.

A commonly used technique in preprocessing is binary encoding, which converts categorical variables into binary representations, making it easier for machine learning algorithms to process the data. Once the data is preprocessed, the next step is model evaluation.

This step involves assessing the performance of the trained model using various evaluation metrics such as the Confusion matrix, accuracy, precision, and recall.

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor : 5.725

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 02, February 2024, pp : 444-447

Fig. 2. Heatmap for an attributes

Accuracy provides an overall measure of how correct the model's predictions are, while precision quantifies the quality of positive predictions. Recall, on the other hand, evaluates the model's ability to correctly identify positive instances.By considering these evaluation metrics collectively, valuable insights can be gained into the effectiveness of the model. These insights can then guide further optimization efforts to improve the model's performance.

Table -1 Comparison between Random Forest and Naïve Bayes

	Random Forest	**Naïve Bayes**
Precision	**80**	**98**
Recall	**47**	**43**

Table -2 Experiment Result

Techniques	**Accuracy**
Random Forest	80%
Naïve Bayes	57%
Decision Tree classifier	65%
Logistic Regression	75%

D. CONCLUSION

To summarize, the research conducted a comparison between the Naïve Bayes and Random Forest algorithms in predicting liver disease patients using the ILPT dataset. Both algorithms exhibited promising outcomes in terms of accuracy and predictive capabilities. Naïve Bayes showcased efficient classification through its simplicity and assumption of feature independence, while Random Forest demonstrated robustness and high accuracy due to ensemble learning. In terms of predictive accuracy, precision, and recall, the Random Forest algorithm slightly outperformed Naïve Bayes.

However, Naïve Bayes still displayed competitive performance and could be a suitable choice for simpler classification tasks. Further optimization and fine-tuning of both algorithms have the potential to enhance their predictive power in diagnosing liver disease. Overall, this study emphasizes the effectiveness of machine learning algorithms in healthcare applications, particularly in predicting liver disease patients.

E. REFERENCE

[1] Jagdeep Sing, Sachin Bagga "Software Based prediction of Liver Disease with feature selection and classification technique"2019.

[2] PSM.Keerthana,Nimish phalinkar "Prediction model Detecting Liver Disease in Patients using machine learning", 2020.

[3] Rahul amin "Prediction of Chronic Liver Disease patients using Integrated Projection based statistical feature extraction with machine learning algorithm" 2023.

[4] Deepika ,Bhupathi Liver Disease Detection using machine learning , 2022.

[5] Easin Hasan, Fahad Mustafa "Statistical machine learning approaches to liver disease prediction, 2021.

[6] Shefi Tanvir fayas ,G.S Tejanmayi,yerra masetti kanaka Ruthin ,Prediction of machine learning , 16 November 2021.

[7] R.Kalaiselvi; K.Meena, V.Vanitha"Liver Disease Prediction Using Machine Learning Algorithms"2021Publisher: IEEE.2019.

[8] Alaa Ali Hameed ,Akhtar Jamil, Ayse Devim Deep Learning for Liver Disease Prediction First Online: 13 April 2022,

[9] R.H. LinAn intelligent model for liver disease diagnosisArtif Intell Med, 47 (1) (2009),

[10] Articles the global, regional, and national burden of cirrhosis by cause in 195 countries and territories, 1990 – 2017 : a systematic analysis for the Global Burden of Disease Study 2017,