M.Kumarasamy
College of Engineering
NAAC Accredited Autonomous Institution
Approved by AICTE & Affiliated to Anna University
ISO 9001:2015 Certified Institution
Thalavapalayam, Karur, Tamilnadu.

STATISTICAL MACHINE LEARNING APPROACHES TO LIVER DISEASE PREDICTION

TEAM ID:

PNT2022TMID15966

TEAM MEMBERS:

1. MTHEESH C (927619BEC4123)
2. PARITHI MALAVAN G (927619BEC4140)
3. RAGUNATH SD (927619BEC4158)
4. SANJAY R (927619BEC4172)

INDUSTRY MENTORS NAME:

NIDHI

FACULTY MENTOR NAME:

Dr. K KARTHIKEYAN

ABSTRACT

The improvement of patient care, research, and policy is significantly impacted by medical diagnoses. Medical practitioners employ a variety of pathological techniques to make diagnoses based on medical records and the conditions of the patients. Disease identification has been significantly enhanced by the application of artificial intelligence and machine learning in conjunction with clinical data. Data driven, machine learning (ML) techniques can be used to test current approaches and support researchers in potentially innovative judgments. The goal of this work was to use ML algorithms to derive meaningful predictors of liver disease from the medical data.

1.INTRODUCTION

The number of patients with liver disease has been steadily rising as a result of excessive alcohol use, exposure to hazardous gases, ingestion of tainted foods such pickles and cucumbers, and drug usage. In an effort to lighten the load on doctors, this dataset was used to assess prediction systems. The data set consists of the patient's age, gender, and total bilirubin. Direct bilirubin, alkaline phosphatase, alamine aminotransferase, aspartate aminotransferase, total proteins, albumin, and the ratio of albumin to globulin are other examples. Set: the field that was utilized to divide the data into two sets (patient with liver disease, or no disease). This study attempts to find an appropriate machine learning algorithm that can determine whether a person has liver disease or not given a dataset containing biological and diagnostic data of Indian patients.

A. Relevance of the Project

Using certain characteristics such as total bilirubin, direct bilirubin, alkaline phosphatase, total protein, albumin, and globulin, this software can determine whether a patient has liver disease or not.

B. Scope of the Project

It is necessary to use supervised learning to resolve this binary classification issue. Each data point has ten attributes, and there is a label that indicates if the patient has liver disease or not.In order to find the answer, our goal should be to train a variety of supervised learning models on this dataset in order to create a high – performing model that can accurately identify any new data point as positive or negative and outperform the benchmarks.

2. EXISTING SYSTEM

Only two systems exist in the same domain, according to a thorough investigation into the subject. First, the system is entirely manual. It has the capacity to store patient information and medical records. The initial system's key characteristics are as follows. The second system is more effective than the first. It was discovered from a related research study that the system is constructed utilizing the KNN method.

LIMITATIONS:

- The entire system was manual
- It fails to accurately predict a value using the KNN algorithm.
- This system takes a long time to provide the user with an output.

3.PROPOSED SYSTEM

Using the various machine learning algorithm and finds out the best algorithm having high accuracy is considered for building the model.This system forecasts liver illness. Compare the capacity to forecast binary classifications of liver disease among several statistical learning techniques. Obtain confusion matrices for contrasting predictive classes with actual classes, then evaluate several ML techniques to gauge how well they work at diagnosing liver illness. Analyze receiver operating characteristic (ROC) curves to assess the diagnostic value of the binary disease classification.

The following module make up the bulk of the proposed system:

Liver Disease Prediction:

Use the patient's age, total bilirubin, direct bilirubin, alkaline phosphatase total proteins, albumin, albumin, an globulin ratio to determine whether the patient has liver disease or not.

Web-page module:

By providing the required information, a person can view about his/her liver condition on the website using this module.

4.METHODOLOGY

In this project, we gather data from a data set, and the health specialist can enter the data for testing using our web application. In this application, we perform data cleaning and pre-processing, extensive data analysis, data visualization, and machine learning using supervised learning algorithms, decision trees, K nearest neighbor's, logistic regression, and support vector machines. This approach makes predictions about a person's liver condition based on variables including total bilirubin, direct bilirubin, albumin, total protein, etc.

5. LITERATURE REVIEW

AUTHOR: Muktevi Srivenkatesh

DESCRIPTION: The use of information digging systems for prescient examination is significant in the wellbeing field since it enables us to confront ailments prior and accordingly spare individuals' lives through the expectation of fix. In this work, we utilized a few learning calculations K-Nearest Neighbors, Support Vector Machines, Logistic Regression, Navi Bayes, Random Forest to foresee patients with constant liver disappointment infection, and patients who are not experiencing this illness.

AUTHOR: MH Zweig, G Campbell

DESCRIPTION: The capacity to appropriately divide people into clinically significant subgroups, or diagnostic accuracy, is a clinical performance metric for laboratory tests. Diagnostic accuracy, which should be distinguished from usefulness or real practical worth of the information, relates to the standard of the information produced by the categorization device.

Receiver-operating characteristic (ROC) plots show the boundaries of a test's capacity to distinguish between various states of health over the full range of operating conditions, providing a pure index of accuracy. Additionally, ROC charts have a crucial or unifying role in the evaluation and application of diagnostic technologies. A user can easily continue with many additional tasks after the plot has been created, such as doing a quantitative ROC analysis.

AUTHOR: Keerthana Jaganathan

DESCRIPTION: One of the major issues to patient safety and the pharmaceutical business is drug-induced liver damage. It results in the withdrawal of licensed medications from the market as well as the discontinuation of drug candidates in clinical trials. Therefore, it is crucial to find hepatotoxic substances early in the medication development process. The aim of this study is to build quantitative structure activity connection models for molecular descriptor sets utilizing machine learning algorithms and rigorous feature selection approaches. The models were created using a substantial and varied collection of 1253 pharmacological molecules, and they underwent internal 10-fold cross-validation. To improve prediction in this work, we used a range of feature selection strategies to obtain the ideal subset of descriptors as modelling features.

AUTHOR: Jagdeep Singh

DESCRIPTION: Today, everyone's health is a very essential concern, so it is necessary to offer medical services that are freely accessible to everyone. The primary goal of this study is to forecast liver illness using a software engineering methodology that makes use of feature selection and classification techniques.

The Indian Liver Patient Dataset (ILPD) from the University of California, Irvine database is used to carry out the proposed research. The many variables of the liver patient dataset, including age, direct bilirubin, gender, total bilirubin, Alphas, spot, albumin, globulin ratio, and shot, among others, are used to forecast the risk level of liver illnesses. The Liver platform is used to implement many classification methods, including Logistic Regression, SMO, Random Forest algorithm, Naive Bayes, J48, and k-nearest neighbor (IBK).

AUTHOR: James Brink & Daniel Rosenthal

DESCRIPTION: For healthcare facilities to offer patients high-quality, cost-effective care, contemporary technology must be powerful. The development of effective and reliable management tools has not yet occurred, despite the significant advancements in the computerization and digitalization of medicine. The enormous complexity and variety of healthcare operations, whose needs have exceeded conventional management, is a significant factor in this. Scalable and adaptable to complicated patterns, machine learning algorithms may be especially well adapted to overcoming these issues. The ability to create robust models from many weakly predictive characteristics.

AUTHOR: Federico Divina

DESCRIPTION: The ratio of the two variables, radius and margin, determines the generalization error of the traditional support vector machine (SVM). The typical SVM priorities margin maximization but disregards radius minimization, which lowers the SVM classifier's overall performance. To achieve a trade-off between the margin and radius, numerous strategies are devised. Due to the demands of matrix transformation, all these techniques still have a high computational cost.

An SVM is employed in many non-linear and complex issues since it attempts to set the optimal hyperplane across classes and because of several reliable kernel methods. Because the optimum hyperplane design across classes is ineffective, binding a class to its specific area is necessary to improve.

AUTHOR: Michael Friendly & Leland Wilkinson

DESCRIPTION: The cluster heat map is a clever visualization that shows a data matrix's row and column hierarchical cluster structure at the same time. It comprises of a rectangular tiling where each tile is darkened according to the value of the appropriate data matrix element. The tiling's rows (columns) are arranged so that comparable rows (columns) are close to one another. Hierarchical cluster trees are located on the tiling's edges, both vertically and horizontally. This cluster heat map is a combination of many visual representations created by statisticians over the course of more than a century. We identify the earliest roots for this display in late 19th-century works and trace a wide range of statistical literature from the 20th century that served as a foundation for this most recent.

AUTHOR: Fatah Chatoyance & Christine Duguay

DESCRIPTION: The emergency room of Dr. Georges-L. Dumont Hospital in Moncton was the subject of a discrete event simulation research, which is described in this publication (Canada). The study's goals were to decrease patient wait times while also enhancing system throughput and service delivery. Numerous options were developed using resource scenario addition since patient wait times are related to resource availability. Software from Arena was used to create the models.

6.REFERENCE

[1] Asrani, S.K.; Devarbhavi, H.; Eaton, J.; Kamath, P.S. “Burden of liver diseases in the world”. J. Hepatol. 2019.

[2] Chalasani, N.; Younossi, Z.; Lavine, J.E.; Charlton, M.; Cusi, K.; Rinella, M.; Harrison, S.A.; Brunt, E.M.; Sanyal, A.J. “The diagnosis and management of nonalcoholic fatty liver disease: Practice guidance from the American Association for the Study of Liver Diseases”. Hepatology 2018.

[3] Wang, Y.; Li, Y.; Wang, X.; Gacesa, R.; Zhang, J.; Zhou, L.; Wang, B. “Predicting Liver Disease Risk Using a Combination of Common Clinical Markers: A Screening Model from Routine Health Check-Up”. Dis. Markers 2020.