

PROJECT DEVELOPMENT PHASE

DELIVERY OF SPRINT-4

DATE	08 NOVEMBER 2022
TEAM ID	PNT2022TMID36295
PROJECT NAME	STATISTICAL MACHINE LEARNING APPROACHES TO LIVER DISEASE PREDICTION

REVIEW OF LIVER DISEASE PREDICTION USING MACHINE LEARNING ALGORITHM

Abstract:

Liver Disease is the leading cause of global death that impacts the massive quantity of humans around the world. This disease is caused by an assortment of elements that harm the liver. For example, obesity, an undiagnosed hepatitis infection, alcohol misuse which is responsible for abnormal nerve function, coughing up or vomiting blood, kidney failure, liver failure, jaundice, liver encephalopathy and there are many more. Diagnosis of liver infection at preliminary stage is important for better treatment. In today's scenario devices like sensors are used for detection of infections. Accurate classification techniques are required for automatic identification of disease samples. This disease diagnosis is very costly and complicated. Therefore, the goal of this work is to evaluate the performance of different Machine Learning algorithms in order to reduce the high cost of chronic liver disease diagnosis by prediction. In this work, we used five algorithms Logistic Regression, Decision Tree, Support Vector Machine, Naïve Bayes, and Random Forest. The performance of different classification techniques was evaluated on different measurement techniques such as accuracy, precision, recall, and specificity. We found the accuracy 74%, 72%, 72%, 71%, and 57% for SVM,DT,RF,LR and NB. The analysis result shown the SVM achieved the highest accuracy. Moreover, our present study mainly focused on the use of clinical data for

liver disease prediction and explores different ways of representing such data through our analysis

Keywords: Classification, Logistic Regression, Support Vector Machine (SVM), Naïve Bayes, Decision Tree, Random Forest

1. INTRODUCTION

Liver is the largest internal organ in the human body, it is essential for digesting food and releasing the toxic element of the body and plays a major role in metabolism and serving several vital functions. The liver is the largest glandular organ of the body. It weighs about 3 lb (1.36 kg). The liver's main job is to strain the blood coming from the digestive tract, before passing it to the rest of the body. The liver also detoxifies chemicals and metabolizes drugs. As it does so, the liver hides bile that ends up back in the intestines. The liver also makes proteins important for blood clotting and other functions. The liver supports almost every organ in the body and is vital for our survival. Liver disease may not cause any symptoms at earlier stage or the symptoms may be vague, like weakness and loss of energy. Symptoms partly depend on the type and the extent of liver disease. Liver diseases are diagnosed based on the liver functional test[1]. Several diseases states can disturb the liver. Some of the diseases are Wilson's disease, hepatitis (an inflammation of the liver), liver cancer, and cirrhosis (a chronic inflammation that progresses ultimately to organ failure). Alcohol alters the metabolism of the liver, which can have on the whole detrimental effects if alcohol is taken over long periods of time. Hemochromatosis can cause liver problems.

Fatty liver is a reversible condition where large vacuoles of triglyceride fat acquire in liver cells via the process of lipid accumulation. It can occur in people with a high level of alcohol consumption as well as in people who never had alcohol. Hepatitis (usually caused by a virus spread by excess contamination or direct contact with infected body fluids). Cirrhosis of the liver is one of the most serious liver diseases. It is an action used to indicate all forms of diseases of the liver characterized by the significant loss of cells. The liver gradually contracts in size and becomes leathery and hard. The regenerative action continues under liver cirrhosis but the progressive loss of liver cells exceeds cell replacement. Liver cancer. The risk of liver cancer is higher in

those who have cirrhosis or who had valid types of viral hepatitis; but more often, the liver is the site of secondary (metastatic) cancers spread from other organs.

2. LITERATURE REVIEW

Health care and medicine handles huge data on daily basis. This data comprises of information about the patients, diagnosis reports and medical images. It is important to utilize this information to decipher a decision support system. To achieve this it is important to discover and extract the knowledge domain from the raw data. It is accomplished by knowledge discovery and data mining (KDD). The implementation of data mining techniques is widespread in biological domain. In recent years, liver disorders have excessively increased and liver diseases are becoming one of the most fatal diseases in several countries. In this study, liver patient datasets are investigate for building classification models in order to predict liver disease. Several feature model construction and comparative analysis are implemented for improving prediction accuracy of Indian liver patients. Different studies have been conducted for classification of liver disorders, they are discussed briefly. Classification algorithm is one of the greatest significant and applicable data mining techniques used to apply in disease prediction. Classification algorithm is the most common in several automatic medical health diagnoses. Many of them show good classification accuracy. In another study the UCI liver dataset was used for selection of sub features based on random forest classifier with multi-layer perceptron induced .Different approaches for artificial intelligence for the liver patient dataset, precise predictions of liver failure were applied [5, 6, 7 and 8]. Identification of liver infection at preliminary stage is important to combat the frequency and severity deaths of patients in India. The patients must be screened based on initial symptoms for development of personalized therapy. In this study, an attempt is made for prediction of liver disease in patients using data mining techniques. Based on the review of literature, it was depicted that the past research studies have implemented different data mining techniques for classification of liver dataset. A hybrid model can be adapted to further increase the prediction accuracy of liver disease. It is followed by development of a graphical user interface would further aid the scientific community

in early diagnosis of liver infection. It will provide a framework for end user application for generating promising treatment protocols.

3. METHODOLOGY

3.1. Data Collection

For this study, the Indian Liver Patient Dataset (ILPD) was selected from the UCI Machine Learning repository. It is a sample of the whole Indian population taken from the area of Andhra Pradesh. There were 583 instances based on ten different biological parameters in the dataset. Based on these criteria, the class value was stated as either yes (416 cases) or no (167 cases), reflecting the liver.

3.2. Pre-processing and Feature selection

To normalize the missing values, pre-processing techniques have been introduced. The missing values were replaced by null values along with their instances. Feature selection was followed to classify the appropriate attribute for classification. Using both filter and wrapper approaches, feature selection was carried out. The attributes with more than 70% correlation were initially excluded by correlation analysis from the dataset. The algorithm was implemented to estimate the value of different features in a dataset on the basis of random forests.

3.3. Randomization and splitting of dataset

To build classification models, the features selected in the preceding phase were accepted. The dataset was initially randomized to produce an arbitrary sample permutation. Splitting of the dataset into training (70 percent of the dataset) and test (30 percent) sets was followed. The training set consisted of 389 cases and the evaluation set consisted of the remaining 194 cases.

3.4. Classification algorithms

Classification algorithm is one of the greatest significant and applicable data mining techniques used to apply in disease prediction. Classification algorithm is the most common in several automatic medical health diagnoses. Many of them show good

classification accuracy. Different data mining algorithms like Naïve Bayes, Decision Tree, Logistic Regression, Random forest and Support vectormachine (SVM) were implemented. The algorithms are briefly discussed below:

3.4.1. Naïve Bayes :

It is based on the Bayes theorem of conditional probability. The algorithm assumes that each attribute contributes to the total outcome independent of other attributes [10]. In machine learning we are often interested in selecting the best hypothesis (h) given data (d). In a classification problem, our hypothesis (h) may be the class to assign for a new data instance (d). One of the easiest ways of selecting the most probable hypothesis given the data that we have that we can use as our prior knowledge about the problem. Bayes' Theorem provides a way that we can calculate the probability of a hypothesis given our prior knowledge. Bayes' Theorem is stated as: $P(h|d) = (P(d|h) * P(h)) / P(d)$ Where, $P(h|d)$ is the probability of hypothesis h given the data d. This is called the posterior probability. $P(d|h)$ is the probability of data d given that the hypothesis h was true. $P(h)$ is the probability of hypothesis h being true (regardless of the data). This is called the prior probability of h. $P(d)$ is the probability of the data (regardless of the hypothesis).

3.4.2. Logistic Regression:

Calculated Regression was for the most part utilized in natural research and applications in the mid-20th century [12]. Logistic regression can deal with any number of numerical as well as absolute factors. In addition, it introduces a discrete parallel item somewhere in the range of 0 and 1. Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes. In simple words, the dependent variable is binary in nature having data coded as either 1 (stands for success/yes) or 0 (stands for failure/no). Mathematically, a logistic regression model predicts $P(Y=1)$ as a function of X. It is one of the simplest ML algorithms that can be used for various classification problems such as spam detection, Diabetes prediction, cancer detection etc. Logistic regression can be divided into following types – Binary or Binomial In such a kind of classification, a dependent variable will have only two possible types

either 1 or 0. For example, these variables may represent success or failure, yes or no, win or loss etc. Multinomial In such a kind of classification, dependent variable can have 3 or more possible unordered types or the types having no quantitative significance. For example, these variables may represent “Type A” or “Type B” or “Type C”. Ordinal In such a kind of classification, dependent variable can have 3 or more possible ordered types or the types having a quantitative significance. For example, these variables may represent “poor” or “good”, “very good”, “Excellent” and each category can have the scores like 0,1,2,3

```

random_forest = RandomForestClassifier(n_estimators=100)
random_forest.fit(X_train, y_train)
# Predict Output
rf_predicted = random_forest.predict(X_test)

random_forest_score = round(random_forest.score(X_train, y_train) * 100, 2)
random_forest_score_test = round(random_forest.score(X_test, y_test) * 100, 2)
print('Random Forest Score: \n', random_forest_score)
print('Random Forest Test Score: \n', random_forest_score_test)
print('Accuracy: \n', accuracy_score(y_test, rf_predicted))
print(confusion_matrix(y_test, rf_predicted))
print(classification_report(y_test, rf_predicted))

Random Forest Score:
100.0
Random Forest Test Score:
72.02
Accuracy:
0.7202072538860184
[[123 10]
 [ 36 16]]

```

	precision	recall	f1-score	support
1	0.77	0.87	0.82	141
2	0.47	0.31	0.37	52

Fig 2. Implementation of Random Forest

RESULTS AND DISCUSSION

Dataset

It is a sample of the entire Indian population collected from Andhra Pradesh region. The dataset comprised of 583 instances based on ten different biological parameters. The class value was reported based on these parameters as either yes (416 cases) or no (167 cases) that represent the liver infection. The patients were described as either ‘1’ or ‘2’ on the basis of liver disease.

Application of classification algorithms

Five algorithms namely Naïve Bayes, Logistic Regression, Decision Tree, Random forest and SVM were implemented for classification of the Indian Liver Patient Dataset. The models were generated for the training set and evaluated on the test set.

Based on the accuracy of prediction it was observed that SVM achieved an accuracy of 74.09%.

	Model	Score	Test Score
4	SVM	99.74	74.09
3	Decision Tree	100.00	72.54
2	Random Forest	100.00	72.02
0	Logistic Regression	71.28	71.50
1	Gaussian Naive Bayes	53.59	57.51