A REVIEW OF LIVER PATIENT ANALYSIS USING

MACHINE LEARNING

SUBMITTED BY

TEAM LEADER

P.DEEPANA

TEAM MEMBERS

S.DHARANI

P.DURGA

G.DHARSHANA

A Review Of Liver Patient Analysis Methods Using Machine Learning

**INTRODUCTION:**

Liver diseases averts the normal function of the liver. This disease is caused by an assortment of elements that harm the liver. Diagnosis of liver infection at the 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 liver disease diagnosis. Early prediction of liver disease using classification algorithms is an efficacious task that can help the doctors to diagnose the disease within a short duration of time. In this project we will analyse the parameters of various classification algorithms and compare their predictive accuracies so as to find out the best classifier for determining the liver disease. This project compares various classification algorithms such as Random Forest, Logistic Regression, KNN and ANN Algorithm with an aim to identify the best technique. Based on this study, Random Forest with the highest accuracy outperformed the other algorithms and can be further utilised in the prediction of liver disease and can be recommended to the user.

**Technical Architecture:**

**Project Flow:**

User interacts with the UI to enter the input.

Entered input is analysed by the model which is integrated.

Once model analyses the input the prediction is showcased on the UI

To accomplish this, we have to complete all the activities listed below,

**Define Problem / Problem Understanding**

Specify the business problem

Business requirements

Literature Survey

Social or Business Impact.

**Data Collection & Preparation**

Collect the dataset

Data Preparation

**Exploratory Data Analysis**

Descriptive statistical

Visual Analysis

**Model Building**

Training the model in multiple algorithms

Testing the model

**Performance Testing & Hyperparameter Tuning**

Testing model with multiple evaluation metrics

Comparing model accuracy before & after applying hyperparameter tuning

**Model Deployment**

Save the best model

Integrate with Web Framework

**Project Demonstration & Documentation**

Record explanation Video for project end to end solution

Project Documentation-Step by step project development procedure

**Project Structure:**

Create the Project folder which contains files as shown below

We are building a flask application which needs HTML pages stored in the templates folder and a python script app.py for scripting.

ETC.pkl is our saved model. Further we will use this model for flask integration.

Training folder contains a model training file.

Define Problem / Problem Understanding

In this milestone, we will go through the problem understanding.

**Specify The Business Problem**

Liver diseases averts the normal function of the liver. This disease is caused by an assortment of elements that harm the liver. Diagnosis of liver infection at the 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 liver disease diagnosis. Early prediction of liver disease using classification algorithms is an efficacious task that can help the doctors to diagnose the disease within a short duration of time. In this project we will analyse the parameters of various classification algorithms and compare their predictive accuracies so as to find out the best classifier for determining the liver disease. This project compares various classification algorithms such as Random Forest, Logistic Regression, KNN and ANN Algorithm with an aim to identify the best technique. Based on this study, Random Forest with the highest accuracy outperformed the other algorithms and can be further utilised in the prediction of liver disease and can be recommended to the user.

**Business Requirements**

Patients with Liver disease have been continuously increasing because of excessive consumption of alcohol, inhale of harmful gases, intake of contaminated food, pickles and drugs and other factors. This dataset was used to evaluate prediction algorithms in an effort to reduce burden on doctors. Use these patient records to build a prediction model that will predict which patients have liver disease and which ones do not.

**Literature Survey**

With a growing trend of sedentary and lack of physical activities, diseases related to liver have become a common encounter nowadays. In rural areas the intensity is still manageable, but in urban areas, and especially metropolitan areas the liver disease is a very common sighting nowadays. Problems with liver patients are not easily discovered in an early stage as it will be functioning normally even when it is partially damaged. An early diagnosis of liver problems will increase patients survival rate. There are various algorithms that have been used with varying levels of success. Logistic regression, decision tree, random forest, and neural networks have all been used and have been able to accurately predict liver disease.

**Social Or Business Impact**

Social Impact:- Today almost everybody above the age of 12 years has smartphones with them, and so we can incorporate these solutions into an android app or ios app. Also it can be incorporated into a website and these app and website will be highly beneficial for a large section of society.

Business Model/Impact:- Its now more feasible Blood test centers to give the result. As for this model user don’t need to have any deep knowledge of medical science and liver diseases. User need to do pass the details being asked, which are already present in the blood test report( some like age, gender are already known) and then user will get the results of prediction

**Data Collection & Preparation**

ML depends heavily on data. It is the most crucial aspect that makes algorithm training possible. So this section allows you to download the required dataset.

**Collect The Dataset**

There are many popular open sources for collecting the data. Eg: kaggle.com, UCI repository, etc.

In this project we have used .csv data. This data is downloaded from kaggle.com. Please refer to the link given below to download the dataset

Link: <https://www.kaggle.com/datasets/uciml/indian-liver-patient-records>

As the dataset is downloaded. Let us read and understand the data properly with the help of some visualisation techniques and some analysing techniques.

**Importing The Libraries**

Import the necessary libraries as shown in the image. (optional) Here we have used visualisation style as fivethirtyeight.

**Read The Dataset**

Our dataset format might be in .csv, excel files, .txt, .json, etc. We can read the dataset with the help of pandas.

In pandas we have a function called read\_csv() to read the dataset. As a parameter we have to give the directory of the csv file.

**Data Preparation**

As we have understood how the data is, let's pre-process the collected data. The download data set is not suitable for training the machine learning model as it might have so much randomness so we need to clean the dataset properly in order to fetch good results. This activity includes the following steps.Handling missing values

**Handling categorical data**

These are the general steps of pre-processing the data before using it for machine learning. Depending on the condition of your dataset, you may or may not have to go through all these steps

**Handling Missing Values**

Let’s find the shape of our dataset first. To find the shape of our data, the df.shape method is used. To find the data type, df.info() function is used.

C:\Users\Acer\Desktop\SmartBridge\Liver Analysis\Output\data.info().PNG

Handling Categorical Values As we can see our dataset has categorical data we must convert the categorical data to integer encoding or binary encoding.

To convert the categorical features into numerical features we use encoding techniques. There are several techniques but in our project we are using manual encoding with the help of list comprehension. In our project, for Gender, encoding is done.

**Exploratory Data Analysis**

In this milestone, we will see the exploratory data analysis.

**Descriptive Statistical**

Descriptive analysis is to study the basic features of data with the statistical process. Here pandas has a worthy function called describe. With this describe function we can understand the unique, top and frequent values of categorical features. And we can find mean, std, min, max and percentile values of continuous features.

**Visual Analysis**

Visual analysis is the process of using visual representations, such as charts, plots, and graphs, to explore and understand data. It is a way to quickly identify patterns, trends, and outliers in the data, which can help to gain insights and make informed decisions

**Univariate Analysis**

In simple words, univariate analysis is understanding the data with a single feature. Here we have displayed two different graphs such as distplot and countplot. The Seaborn package provides a wonderful function distplot. With the help of distplot, we can find the distribution of the feature. To make multiple graphs in a single plot, we use subplot.

**Countplot:-**

A count plot can be thought of as a histogram across a categorical, instead of quantitative, variable. The basic API an d options are identical to those for barplot() , so you can compare counts across nested variables.

**Multivariate Analysis**

In simple words, multivariate analysis is to find the relation between multiple features. Here we have used a heat plot from the seaborn package.

**Handling Imbalance Data**

Data Balancing is one of the most important step, which need to be performed for classification models, because when we train our model on imbalanced dataset, we will get biased results, which means our model is able to predict only one class element

For balancing the data we are using the SMOTE Method.

SMOTE: Synthetic minority over sampling technique, which will create new synthetic data points for under class as per the requirements given by us using KNN method.

**Model Building**

In this milestone, we will see model building.

Training The Model In Multiple Algorithms

Now our data is cleaned and it’s time to build the model. We can train our data on different algorithms. For this p roject we are applying four classification algorithms. The best model is saved based on its performance.

**Random Forest Model**

A function named RandomForestClassifier is imported and train and test data are passed as the parameters. Inside the function, RandomForestClassifier algorithm is initialised and training data is passed to the model with .fit() function. Test data is predicted with .predict() function and saved in a new variable. For evaluating the model, a confusion matrix and classification report is done.

**Performance Testing & Hyperparameter Tuning**

In this milestone, we will see performance testing and hyperparameter turning.

**Testing Model With Multiple Evaluation Metrics**

Multiple evaluation metrics means evaluating the model's performance on a test set using different performa nce measures. This can provide a more comprehensive understanding of the model's strengths and weaknesses. We are using accuracy, score to compare between models

Model Deployment

In this milestone, we will see the model deployement.

Save The Best Model

Saving the best model after comparing its performance using different evaluation metrics means selecting the model with the highest performance and saving its weights and configuration. This can be useful in avoiding the need to retrain the model every time it is needed and also to be able to use it in the future.

Integrate With Web Framework

## In this section, we will be building a web application that is integrated to the model we built. A UI is provided for the uses where he has to enter the values for predictions. The enter values are given to the saved model and prediction is showcased on the UI.A

#### ADVANTAGGES&DISADVANTAGES:

###### ADVANTAGES

\*Liver transplant

**\***Diaganosis,grades and stages

1.Hepatititis C

2.Hepaititis B

\*Identifies hiepatotoxity

\*Evoluates abnormal liver function tests

DISADVANTAGES

\*Invasive

\*Accessibility to the procedure

\*Need for training

\*Repeated testing caused

**This section has the following tasks**

Building HTML Pages

Building server side script

Run the web application

Project Demonstration & Documentation

Project deliverables to be submitted along with other deliverables

Record Explanation Video For Project End To End Solution

Record explanation Video for project end to end solution

Project Documentation-Step By Step Project Development Procedure

Create document as per the template provided

###### CONCLUSION:

In the proposed work different classifiers were implmented on liver patient diseases dataset to predict liver diesases based on develpoed software. Dataset was processed and implemented on WEKA tool using features selection techinques with ten fold cross validation testing option. The result of the proposed work were compared using features selection and without using features selection techinques after the implemention of different classifiers in terms of execution time and accuracy. The best result was achieved using logistics regration on classifer with feature selection techinques and execution time of different classifiers was decrease after the implemention of features selection techinques.