PROJECT DEVELOPMENT PHASE

PROJECT DEVELOPMENT - DELIVERY OF SPRINT-1

Date	26 October 2022
Team members	Vinisha S, Susma T, Shreedhar T, Vijay M, Syed Asif Z
Team ID	PNT2022TMID18834
Project Name	Statistical Machine Learning Approaches To Liver Disease Prediction

ABSTRACT

In this paper we are going discuss how to predict risk of liver disease for a person, based on the blood test report results of the user. In this paper, the risk of liver disease was predicted using various machine learning algorithms.The final output was predicted based on the most accurate machine learning algorithm.Based on the accurate model we designed a system which asks a person to enter the details of his/her blood test report. Then the system uses the most accurate model which is trained to predict, whether a person has risk of liver disease or not.

Keywords Machine learning, Liver disease, Confusion matrix ,Use case diagram ,backpropagation algorithm

1. INTRODUCTION

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. Liver diseases cause millions of deaths every year. Viral hepatitis alone causes 1.34 million deaths every year. 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. Liver failures are at high rate of risk among Indians. It is expected that by 2025 India may become the World Capital for Liver Diseases. The widespread occurrence of liver infection in India is contributed due to deskbound lifestyle, increased alcohol consumption and smoking. There are about 100 types of liver infections.

With such alarming figures, it is necessary to have a concern towards tackling these diseases. Afterall, we cannot expect a developed and prosperous nation, with unhealthy youths.

In this project we have taken UCI ILPD Dataset which contains 10 variables that are age, gender, total Bilirubin, direct Bilirubin, total proteins, albumin, A/G ratio, SGPT, SGOT and Alkphos and contains 415 as liver disease patients and 167 as non liver disease patients.As we got through the next parts of this paper we will explain what proccess as taken place for the selection of best model and building necccessary sytem for the prediction of liver disease.

The major outcomes that can be expected through this project are:

- Increased convenience for predicting a liver disease

2. LITERATURE SURVEY

1. Naive Bayes

Naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features. They are among the simplest Bayesian network models, but coupled with kernel density estimation, they can achieve higher accuracy levels.

Naive Bayes classifiers are highly scalable, requiring a number of parameters linear in the number of variables (features/predictors) in a learning problem. Maximum- likelihood training can be done by evaluating a closed- form expression, which takes linear time, rather than by expensive iterative approximation as used for many other types of classifiers.

2. ANN

Artificial Neural networks (ANN) or neural networks are computational algorithms .They intend to simulate the behavior of biological systems composed of neurons. ANNs are computational models inspired by an animals central nervous systems. They are capable of machine learning as well as pattern recognition. These are present as systems of interconnected neurons which can compute values from inputs.

A neural network is an oriented graph. It consists of nodes which in the biological analogy represent neurons, connected by arcs. It corresponds to dendrites and synapses. Each arc is associated with a weight while at each node. To do the prediction, we need to apply the values received as input by the node and define activation function along the incoming arcs, adjusted by the weights of the arcs.An ANN is trained based on backpropagation algorithm.

Algorithm:

1. XTraining Data set of size m*n
2. yLabels for records in X
3. wThe weights for respective layers
4. lThe number of layers in the neural network,1.L

   ij

   ij

5. Dij(l) The error for all l,i,j
   - Reduction in number of deaths due to liver diseases
6. t(l)

   0.For all l,i,j

   - More accurate diagnosis of liver disease by the doctors
7. For i=1 to m

al feedforward(x(I),w) al a(L)-y(i)

ij+aj .t i

ij+aj .t i

tij(l) t(l) (l) l+1

8. if j0 then

   Dij(l) 1 tij(l) +w(l)

   enables Python to be used as an alternative application development language to C++ on all supported platforms

9. else

ij

Dij(l) 1 tij(l)

including iOS and Android.

F. Spyder Notebook

3. **KNN**

where

()

()

J(w)=Dij(l)

Spyder is an open-source cross-platform integrated development environment (IDE) for scientific programming in the Python language. Spyder is extensible with first-party and third-party

The k-nearest neighbors (KNN) algorithm is a simple, easy-to-implement supervised machine learning algorithm that can be used to solve both classification and regression problems.

The KNN algorithm assumes that similar things exist in close proximity. In other words, similar things are near to each other. The KNN algorithm hinges on this assumption being true enough for the algorithm to be useful. KNN captures the idea of similarity (sometimes called distance, proximity, or closeness) with some mathematics we might have learned in our childhood calculating the distance between points on a graph.

4. SVM

Support Vector Machine or SVM algorithm is a simple yet powerful Supervised Machine Learning algorithm that can be used for building both regression and classification models. SVM algorithm can perform really well with both linearly separable and non-linearly separable datasets. Even with a limited amount of data, the support vector machine algorithm does not fail to show its magic.

The objective of SVM algorithm is to find a hyperplane in an N-dimensional space that distinctly classifies the data points. The dimension of the hyperplane depends upon the number of features. If the number of input features is two, then the hyperplane is just a line. If the number of input features is three, then the hyperplane becomes a 2-D plane. In short, the hyperplane is (n-1)- D plane for n features.

5. PyQt Library

PyQt is a GUI widgets toolkit. It is a Python interface for Qt, one of the most powerful, and popular cross- platform GUI library. PyQt is a blend of Python programming language and the Qt library. PyQt API is a set of modules containing a large number of classes and functions. While QtCore module contains non-GUI functionality for working with file and directory etc., QtGui module contains all the graphical controls. In addition, there are modules for working with XML (QtXml), SVG (QtSvg), and SQL (QtSql), etc.

For this paper, we have used the PyQt version 5, which is implemented as more than 35 extension modules and

plugins, includes support for interactive tools fr data inspection and embeds Python-specific code quality assurance and introspection instruments, such as Pyflakes, Pylint and Rope.

It is available cross-platform through Anaconda, on Windows, on macOS through MacPorts, and on major Linux distributions. Spyder uses Qt for its GUI and is designed to use either of the PyQt or PySide Python bindings. QtPy, a thin abstraction layer developed by the Spyder project and later adopted by multiple other packages, provides the flexibility to use either backend.

6. SYSTEM DESIGN

Proposed system

The system being proposed here uses concept of machine learning, and the models are first trained, then tested. Finally the most accurate model will predict the final result.

At first, the system asks you to enter your details including age, gender, total Bilirubin, direct Bilirubin, total proteins, albumin, A/G ratio, SGPT, SGOT and Alkphos. Values of last eight parameters mentioned here, can be known by blood test report of the user.

After taking these inputs from the user, the system compares the data input with the training dataset of most accurate model and then predicts the result accordingly as risk or no risk.

The system has following advantages:

1. No medical expertise required: You dont need to have any knowledge of medical

   science and liver diseases to predict the liver disease using this application. All you need to do is enter the details being asked, which are already present in the blood test report(some like age, gender are already known) and then you will get the results of prediction.

2. High accuracy: The system predicts the results with 100 % accuracy for the dataset that we have used while creating this application. While the accuracy might be different in some cases, it will still be high enough to be trustworthy at a large scale.
3. Immediate results: The results here are predicted within seconds of entering the details. You dont need to wait for a doctor to come, unlike in traditional method.

7. General workflow of the system related to creation and working

The application mainly consists of the following tasks:

- Building and training the system: The phase is totally worked upon by developer of the system, and end user has nothing to do with it. In this phase, we split the dataset into training dataset and test dataset, and then trained the models using training dataset.
- Testing the models: In this phase we tested the accuracy of the models with the test dataset that was formed in previous phase and the most accurate model is figured out.
- Entering details and prediction: In this phase, the end user comes into picture. He/she enters the details of blood test report using GUI of the application. The application then matches the details with the training dataset of the most accurate model, and then predicts final result displaying, Risk or No Risk on the screen.

8.Diagrammatic representations:

The diagrammatic representations of working of the system are as follows:

Use case diagram: The use case diagram of the system is as follows:

As we can see from the use case diagram first the user enters the blood test details and desktop app takes it as an input and predicts the output based on trained accuracy model and displays the result to the user whether the person is at the risk of liver disease or not.

Work flow diagram:

It represents flow of process which we have implemented to develop the prediction system.

Fig 2: Work flow diagram

3. IMPLEMENTATION

The data preprocessing was done using Jupyter Notebook and Desktop Application was Implemented using Sypder IDE.The programming langauge which was used is python and machine learning Sklearn was used to build the model using classification algorithm like KNN,SVM,Naive Bayes and ANN and we Found that SVM was giving most accurate result.

4. RESULT

The results for all the ML models and of final completed project are shown in the following figures and tables:

USE CASE DIAGRAM

SL.NO	ML MODEL	ACCURACY
1	SVM	100%
2	ANN	99.9%
3	KNN	70%
4	Naive Bayes	55.56%