Project Title: Using Machine Learning Algorithms for predicting the hepatitis C virus

Problem Statement :

Hepatitis C is a liver illness caused by the hepatitis C virus, according to the Centre for Disease Control and Prevention (CDC) (HCV). Hepatitis C is transmitted by blood contact with an infected individual. The majority of people nowadays get hepatitis C via sharing needles or other equipment used to manufacture and inject drugs. Hepatitis C is a short-term sickness for some people, but for more than half of those who contract the virus, it becomes a long-term, chronic condition. Cirrhosis and liver cancer are two major, potentially life-threatening side effects of chronic hepatitis C. Chronic hepatitis C patients frequently have no symptoms and do not feel ill. Symptoms of severe liver disease frequently show when they first develop. Hepatitis C does not have a vaccination. Avoiding actions that might transmit the disease, such as injecting drugs, is the greatest method to prevent hepatitis C. It's critical to get tested for hepatitis C since most people with the disease may be cured in 8 to 12 weeks with the right therapy.

Source - cdc.gov - Viral Hepatitis

Solution: Creating a predictive model that could perform early detection of Hepatits C and

other liver diseases would allow people to quickly and easily determine their risk/get

treatment.

Dataset Source - UCI Machine Learning Repository

All attributes except Category and Sex are numerical.

Attributes 1 to 4 refer to the data of the patient:

1) X (Patient ID/No.)

2) Category (diagnosis) (values: '0=Blood Donor', '0s=suspect Blood Donor', '1=Hepatitis',

'2=Fibrosis', '3=Cirrhosis')

3) Age (in years)

4) Sex (f,m)

Attributes 5 to 14 refer to laboratory data:

ALB: amount of albumin in patient's blood

ALP: amount of alkaline phosphatase in patient's blood

ALT: amount of alanine transaminase in patient's blood

AST: amount of aspartate aminotransferase in patient's blood

BIL: amount of bilirubin in patient's blood

CHE: amount of cholinesterase in patient's blood

CHOL: amount of cholesterol in patient's blood

CREA: amount of creatine in patient's blood

GGT: amount of gamma-glutamyl transferase in patient's blood

PROT: amount of protien in patient's blood

The target attribute for classification is Category (2): blood donors vs. Hepatitis C patients (including

its progress ('just' Hepatitis C, Fibrosis, Cirrhosis).

Significance of the topic

The prediction and identification of virus outbreaks helps to enhance surveillance systems. Most

of the challenge of predicting virus outbreaks or modeling illness frequencies using regression

approaches is based on the type of difficulties that result in. Most classification issues handled by

machine learning models, on the other hand, are related to detecting virus outbreaks.

Problem statement/description/background of the topic

Pandemics are huge infectious disease outbreaks that lead to considerable financial, societal, and

governmental disruption as well as an increase in the level of sickness and death across the

nation. Pandemics have become more frequent in the previous decades, and this may be because

of the inability to track or predict these viruses.

Why should be of interest to researchers, practitioners, and the public

Since the evolution of machine learning algorithms, the topic of medical health prediction has

been discussed. Recent research has revealed that machine learning algorithms can be used to

identify and confirm the various phases of HCV infection in the liver. Hoffman et al. used a

previously published study on hepatitis C patients to develop algorithms that build realistic

decision trees for detecting fibrosis and cirrhosis in the liver. The decision trees were validated

using the leave-one-out method, which yielded a model with an accuracy close to 80%, which

was greater than the baseline of pure guessing. The fact that they use the same published dataset

as we used to be an intriguing part of this research.

Datasets:

Table, Excel

Description automatically generated

[***https://archive.ics.uci.edu/ml/datasets/HCV+data***](https://archive.ics.uci.edu/ml/datasets/HCV+data)616 Columns and 14 rows .

Pre-Processing the Data

The purpose of data preprocessing is to prepare the data for analysis and model creation. Data could be cleansed

from outliers, missing values should be replaced with meaningful values, values could require conversion, some

input variables need to be placed in bins, continuous values could be normalized, features relevant to the output

should be selected, and new features could be constructed if needed .

• Outlier Detection: Data outliers are data points that are spread away from the rest of the data points.

Outliers do not always suggest that these values need to be eliminated. In some cases, outliers mean a high

spike in a level, in other times, they could be due to data entry errors. Understanding the data will help decide

if outliers are useful and should be kept or removed. In case of the HCV dataset, there are no existing outliers.

This was also asserted from the authors of the dataset through an email exchange.

• Missing Values: Having missing values in a real life science data is unavoidable. The HCV dataset was

generated from real patients who provided their results of blood tests for research. The missing values are due

to the fact that not all the patients had their blood examined for each feature. We discovered 31 missing values

in the dataset and had to make some choices:

– Ignore the missing values. The HCV dataset is relatively small. Deleting raws not only reduces the number

of instances and could be problematic in our study.

– Impute the missing values with the mean value of every feature respectively. Example: if values are

missing in the ALT column, using the unsupervised ReplaceMissingValues() filter in weka [10], we replace

these values with the mean value of all existing results of ALT.

– Predict the missing values by creating a classifier for each column and replace each missing value by the

predicted values.

Examining the data thoroughly made us realize that choosing only one approach might not satisfy the study.

Instead, we decided to apply the first two approaches and generate two datasets for each approach. Dataset D1

has no missing values, while dataset D2 contains imputed missing values. Due to time limitations we could not

explore option 3 and would like to do so in future work.

• Data Conversion: To make sense of the data, values need to be compared to some existing standard values.

The HCV dataset was created in a German laboratory that uses different units than the International System

of Units (SI Units). For the purpose of understanding the values of the data, a mapping tool needs to be

implemented that utilizes unit conversion as seen in the conversion table. Table 1 was constructed based on

information collected from Quest Diagnostics(QD), Lab Corp, and NIH websites [22] [7] [6]. We preserved the

existing values and created a new column for each attribute that requires conversion into SI Unites, using the

formula: newcolumns= oldcolumn \* conversion coefficient

***Feature Construction:*** Creating new features “may lead to a more concise and accurate classifiers”[16]. This

has been discussed an implemented in both Data Conversion and Discretization subsections. To summa-

rize, two columns have been generated: one to convert the data into SI units and the other to create bins for

each blood result values, as seen in Table 2.

• **Highly Imbalanced Classes:** As described in the previous section, the classes in the dataset are highly

imbalanced. The initial study did not guarantee good accuracy results using the 4 classes and we decided to

combine all HCV classes into one class. Such approach generated a more manageable datasets D3 and D4 with

ratio 6:1 (Healthy:HCV). D3 is created from D1 (no missing values) and D4 is created from D2 (mean imputed

missing values).

• ***Newly Generated Datasets:*** Since the main purpose of the study is to implement a 2-phase machine learning

prediction algorithm that will not only assist the medical professionals in detecting HCV, but will suggest the

type of treatment depending on the condition of the liver (medicine or liver transplant), we needed to isolate the

HCV patients in one dataset D5 (75 values) with initial class allocations for each patient (Hepatitis, Fibrosis,

Cirrhosis). Three new datasets have been generated based on the initial dataset:

– D3, used in Phase1: No missing Values, 486 instances, 2 classes (HCV or Donor)

– D4, used in Phase1: Mean imputed missing values: 516 instances, 2 classes (HCV or Donor)

– D5, used in Phase2: Mean imputed missing values: 75 instances, 3 classes (Hepatitis, Fibrosis, Cirrhosis)

Now that the data is been pre-processed it is ready for analysis.