Classification of Liver Patient Dataset Using Machine Learning Algorithm

INTRODUCTION :-

Overview

Prediction of the disease in the human being is the very long and difficult process in early days. Now a days , computer aided diagnosis is the important role in the medical industry for predicting , analyzing and storing medical information with the images. In this report I have discuss and classify the liver patients dataset with the help of machine learning algorithm. I have implemented some of the classification algorithms on the data selected

from the liver dataset and after the successful implementation of all the algorithms, the best algorithm is selected from the output of all the algorithms execution.

Purpose

Diagnosis of liver disease at a preliminary stage is important for better treatment. It is a very challenging task for medical researchers to predict the disease in the early stages owing to subtle symptoms. Often the symptoms become apparent when it is too late. To overcome this issue, this project aims to improve liver disease diagnosis using machine learning approaches. The main objective of this research is to use classification algorithms to identify the liver patients from healthy individuals. This project also aims to compare the classification algorithms based on their performance factors. To serve the medicinal community for the diagnosis of liver disease among patients, a graphical user interface will be developed using python.

LITERATURE SURVEY :-

EXISTING PROBLEM

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 patient's 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. Therefore, developing a machine that will enhance in the diagnosis of the disease will be of a great advantage in the medical field. These systems will help the physicians in making accurate decisions on patients and also with the help of Automatic classification tools for liver diseases (probably mobile enabled or web enabled), one can reduce the patient queue at the liver experts such

as endocrinologists.Classification techniques are much popular in medical diagnosis and predicting diseases.

PROPOSED SOLUTION

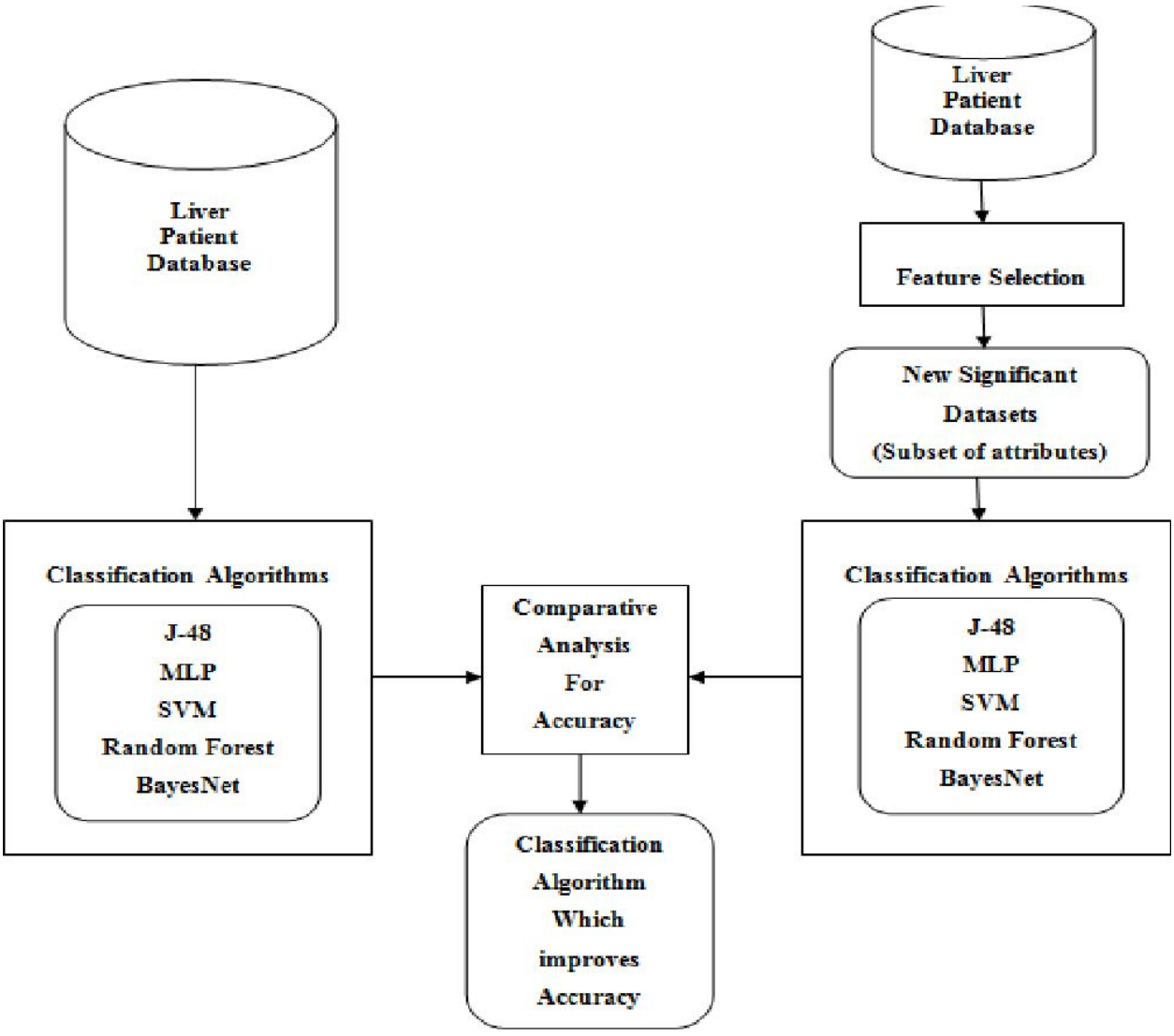
Michael J Sorich reported that SVM classifier produces best predictive performance for the chemical datasets. Lung-Cheng Huang reported that Naïve Bayesian classifier produces high performance than SVM and C 4.5 for the CDC Chronic fatigue syndrome dataset. Paul R Harper reported that there is not necessary a single best classification tool but instead the best performing algorithm will depend on the features of the dataset to be analyzed

The main objective of this research is to use classification algorithms to identify the liver patients from healthy individuals. In this study, FOUR classification algorithms Logistic Regression, Support Vector Machines (SVM), Decision tree classifier have been considered for comparing their performance based on the liver patient data. Further, the model with the highest accuracy is

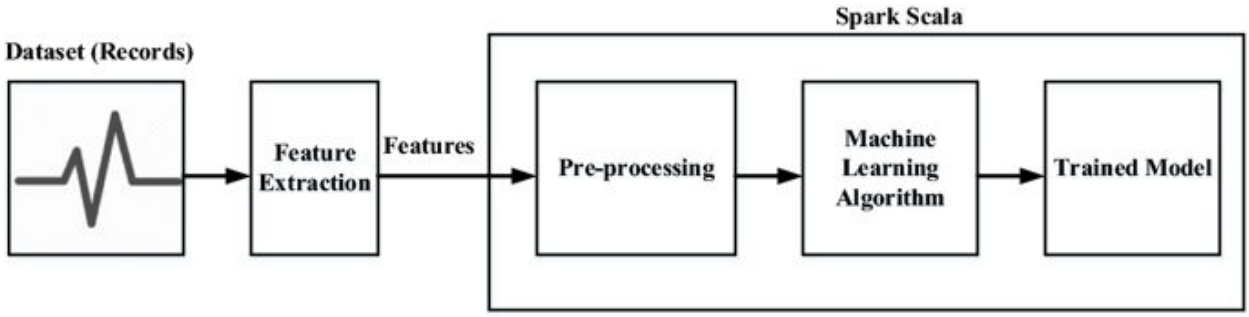
implemented as a user friendly Graphical User Interface (GUI) using package in python. The GUI can be readily utilized by doctors and medical practitioners as a screening tool for liver disease. The dataset used is The Indian Liver Patient Dataset (ILPD) which was selected from Kaggle repository for this study. It is a sample of the entire Indian population collected from Andhra Pradesh region and comprises of 585 patient data.

THEORITICAL ANALYSIS :-

- **BLOCK DIAGRAM**



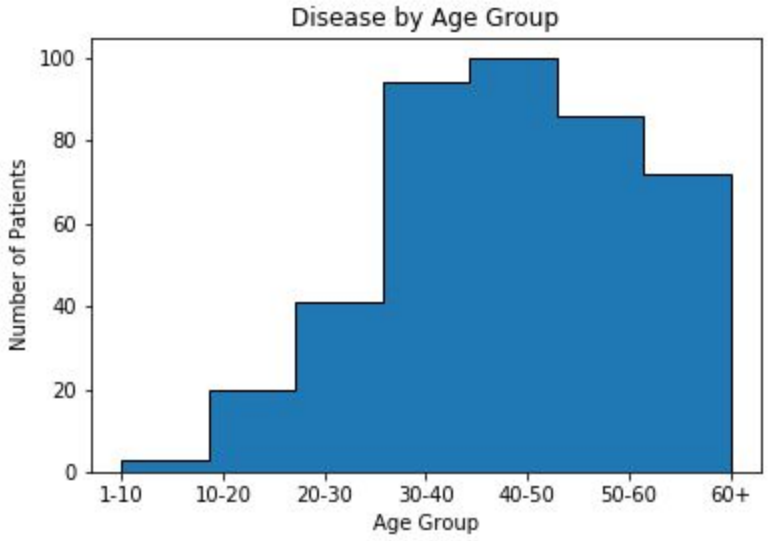
In this model I have downloaded the liver patient dataset and then did data preprosseing on it . After which I applied different types of Machine learning algorithm on the model and selected the algorith which gave the highest accuracy score.



According to the latest WHO data published in 2017 Liver Disease Deaths in India reached 259,749 or 2.95% of total deaths. These numbers point to a type of pandemic that needs investigation. In this analysis, I tried to analyse the factors responsible, the group of people predominantly affected and build a model that can accurately classify the sick from the healthy. So, let us have a look. The data set was collected from north east of Andhra Pradesh, India.During data preprocessing,I observed that Albumin-Globulin Ratio had four missing values. How did I solve this? The A-G ratio is the ratio of albumin to globulin. The feature albumin is already present in the data set and along with that we also have the total proteins readings. Now, with this data it is easy to calculate the globulin levels and thereby A-G ratio.

The first thing that I was concerned about was which age group was the disease mostly prevalent.

And



In a way, I breathed a sigh of relief as it appears that it strikes mostly people in the later phase of life.Next, it was time for the fun part; the model building or what I like to think "gazing at the crystal ball".Just like in life, you got to put your best foot forward. Here too, you need to do something similar;put your best "features or predictors" forward. We do that through a process called feature engineering.It said send "Direct_Bilirubin","Total_Bilirubin","A_G_Ratio","Alkaline_Phosphotase" into battle.Any one familiar with model building knows that you create a train set to train the model and a test set(usually 25% of the data) to test the model accuracy.A classification model that has strict implications on predictions expects a high degree of accuracy(i.e recall value in machine learning).

EXPERIMENTAL INVESTIGATION :-

DATASET

The Indian Liver Patient Dataset comprised of 10 different attributes of 583 patients. The patients were described as either 1 or 2 on the basis of liver disease. The detailed description of the dataset is shown in below in Table description. The table provide details about the attribute and attribute type. As clearly visible from the table, all the features except sex are real valued integers. The feature Sex is converted to numeric value (0 and 1) in the data pre-processing step.

ATTRIBUTES	ATTRIBUTE TYPES
AGE	NUMERIC
SEX	NOMINAL
TOTAL_BILLIRUBINM	NUMERIC
DIRECT_BILLIRUBINM	NUMERIC
ALKALINE	NUMERIC
ALAMINE_PHOSPHATASE	NUMERIC

TOTAL_PROTEINS	NUMERIC
ALBUMIN	NUMERIC
ALBUNIM AND GOBULIN RATION	NUMERIC
RESULT	NUMERIC(1,2)

DATA-PREPROCESSING

Data pre-processing is an important step of solving every machine learning problem. Most of the datasets used with Machine Learning problems need to be processed / cleaned / transformed so that a Machine Learning algorithm can be trained on it. Most commonly used preprocessing techniques are very few like missing value imputation, encoding categorical variables, scaling, etc. These techniques are easy to understand. But when we actually deal with the data, things often get clunky. Every dataset is different and poses unique challenges. All features, except Gender are real valued integers. The last column, Disease, is the label (with "1" representing presence of disease and "2" representing absence of disease). Total number of data points is 583, with 416 liver patient records and 167 non-liver patient records. In the description of this dataset, it is observed that some values

are Null for the Albumin and Globulin Ratio column. The columns which contain null values are replaced with mean values of the column.

CLASSIFICATION TECHNIQUES

a) SVM

SVM aims to find an optimal hyperplane that separates the data into different classes. The scikit-learn package in python is used for implementing SVM. The pre-processed data is split into test data and training set which is of 25% and 75% of the total dataset respectively. A support vector machine constructs a hyper plane or set of hyper planes in a high- or infinite-dimensional space. A good separation is achieved by the hyper plane that has the largest distance to the nearest training data point of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier.

b) LOGISTIC REGRESSION

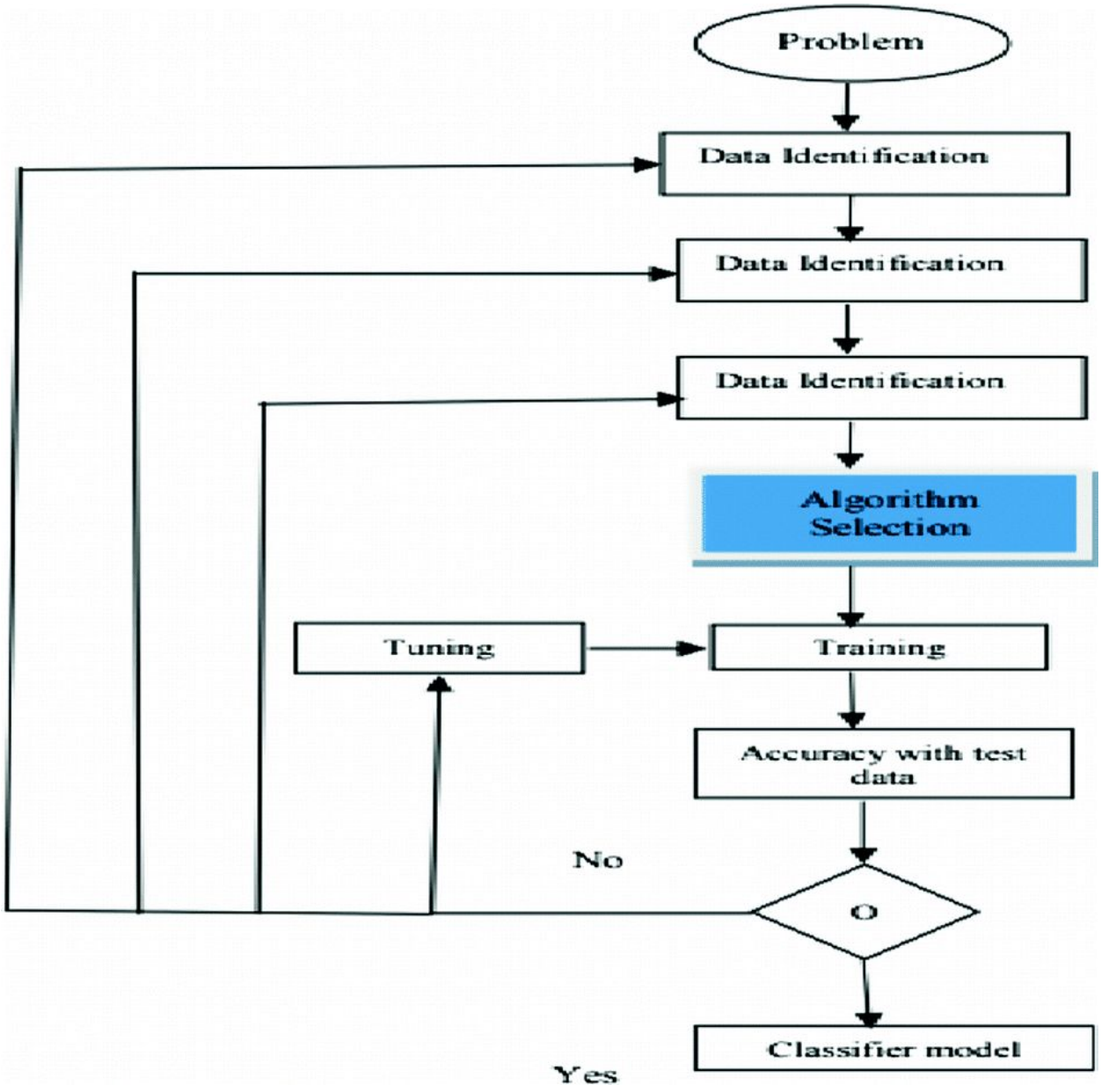
Logistic regression is one of the simpler classification models. Because of its parametric nature it can to some extent be interpreted by looking at the parameters making it useful when experimenters want to look at relationships between variables. A parametric model can be described entirely by a vector of parameters = (0, 1... p). An example of a parametric model would be a straight-line y = kx + m where the parameters are k and m. With known parameters the entire model can be recreated. Logistic regression is a parametric model where the parameters are coefficients to the predictor variables written as 0 +1+X1 + ...PXp Where 0 is called the intercept. For convenience we instead write the above sum of the parameterized predictor variables in vector form as X. The name logistic regression is a bit unfortunate since a regression model is usually used to find a continuous response variable, whereas in classification the response variable is discrete. The term can be motivated by the fact that we in logistic regression found the probability of the response variable belonging to a certain class, and this probability is continuous.

c) K-NN

This section describes the implementation details of KNN algorithm. The model for KNN is the entire training dataset. When a prediction is required for a unseen data instance, the KNN algorithm will search through the training dataset for the k-most similar instances. The prediction attribute of the most similar instances is summarized and returned as the prediction for the unseen instance. The similarity measure is dependent on the type of data. For real-valued data, the Euclidean distance can be used. Other types of data such as categorical or binary data, Hamming distance can be used. The KNN algorithm is belongs to the family of instance-based, competitive learning and lazy learning algorithms. Instance-based algorithms are those algorithms that model the problem using data in-stances (or rows) in order to make predictive decisions. The KNN algorithm is an extreme form of instance-based methods because all training observations are retained as part of the model. It is a competitive learning algorithm, because it internally uses competition between model elements (data instances) in order to make a predictive decision. The objective similarity

measure between data instances causes each data instance to compete to win or be most similar to a given unseen data instance and contribute to a prediction.

FLOWCHART



RESULT :-

Our main goal going into this project was to predict liverdisease using various machine learning techniques. Wepredicted using Support Vector Machine (SVM),

LogisticRegression, K-Nearest Neighbor (K-NN) and NeuralMNetwork. All of them predicted with better results. With Each algorithm, we have observed Accuracy, Precision, Sensitivity and Specificity which can be defined as follows:

Accuracy: The accuracy of a classifier is the percentage of the test set tuples that are correctly classified by the classifier.

$$\text{Acurracy} = \frac{\text{no of TP} + \text{no of TN}}{\text{no of TP+FP+FN+TN}}$$

The following is the result for the following algorithms :-

ALGORITHIMS	ACCURACY SCORE
NAIVE BAYES	60%
DECISION TREE	64%
SVM	66%

RANDOM FOREST	69%
LOGISTIC REGRESSION	70%

As clearly summarized in the table, logistic regression gave the best results.

Advantages of logistic algorithm

* Simplicity and transparency. Logistic Regression is just a bit more involved than Linear Regression, which is one of the simplest predictive algorithms out there. It is also transparent, meaning we can see through the process and understand what is going on at each step, contrasted to the more complex ones (e.g. SVM, Deep Neural Nets) that are much harder to track.

* Giving probabilistic output. Some other algorithms (e.g. Decision Tree) only produce the most seemingly matched label for each data sample, meanwhile, Logistic Regression gives a decimal number ranging from 0 to 1, which can be interpreted as the probability of the sample to be in the Positive Class. With that, we know how confident the prediction is, leading to a wider usage and deeper analysis.

▯ Disadvantages

* The assumption of linearity in the logit can rarely hold. It is usually impractical to hope that there are some relationships between the predictors and the logit of the response. However, empirical experiments showed that the model often works pretty well even without this assumption.

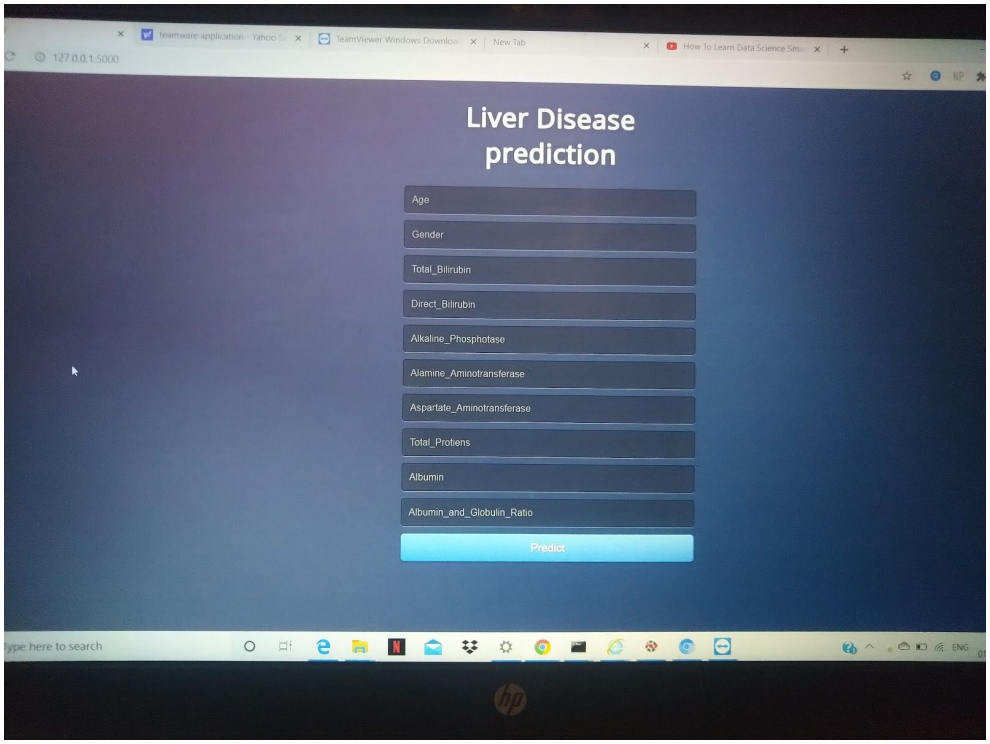
* Uncertainty in Feature importance. This trait is very similar to that of Linear regression. While the weight of each feature somehow represents how and how much the feature interacts with the response, we are not so sure

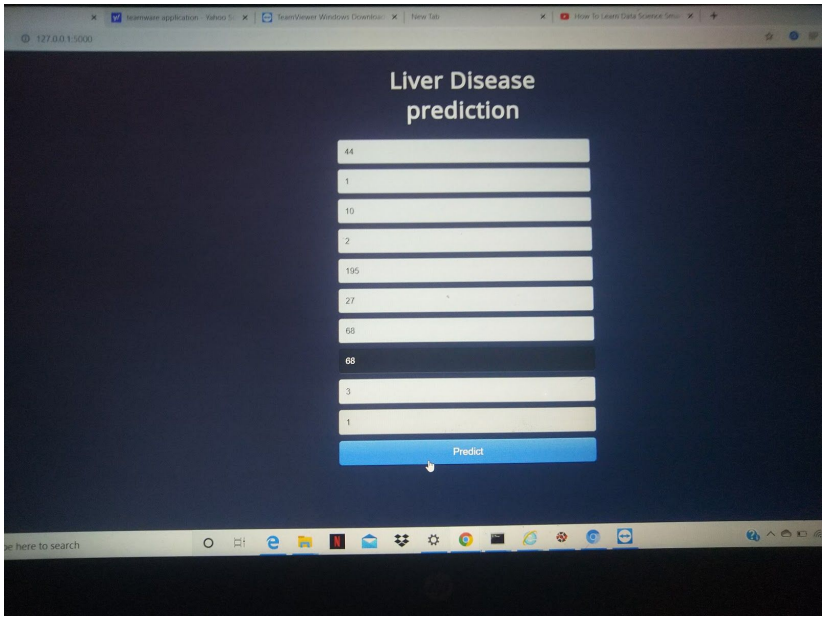
about that. The weight does not only depend on the association between an independent variable and the dependent variable, but also the connection with other independent variables.

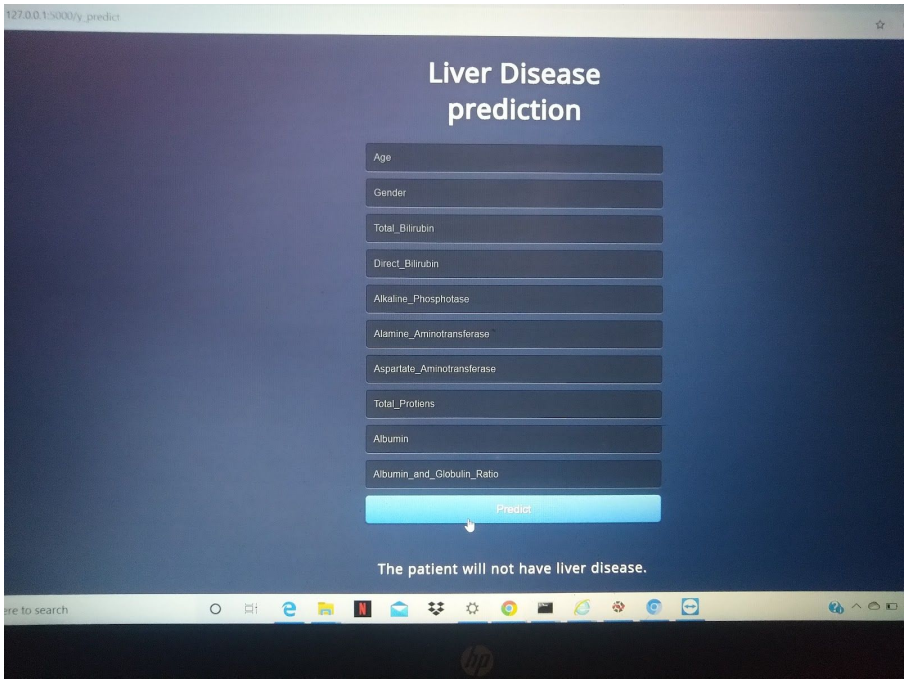
Conclusion

In this project, I have proposed methods for diagnosing liver disease in patients using machine learning techniques. The four machine learning techniques that were used include SVM, Logistic Regression, Random Forest and Decision Tree. The system was implemented using all the models and their performance was evaluated. Performance evaluation was based on certain performance metrics. Logistc Regression was the model that resulted in the highest accuracy with an accuracy of 71%. Comparing this work with the previous research works, it was discovered that Logistic Regression was proved to be much more efficient than the other algorithms.

- **Application**







Bibliography

[1] Michael J Sorich. An intelligent model for liver disease diagnosis. Artificial Intelligence in Medicine 2009;47:53—62.

[2] Paul R. Harper, A review and comparison of classification algorithms for medical decision making.

[3] BUPA Liver Disorder Dataset. UCI repository machine learning databases.