**HEPATITIS DISEASE PREDECTION**

**1.Abstract:** Hepatitis is a chronic disease with the potential to cause a worldwide health care crisis. Hepatitis means injury to the liver with inflammation of liver cells. It is caused by alcohol and some other toxins. It is an excessively increased disease and becoming one of the most fatal diseases in several countries. It involves machine learning algorithms and techniques to analyse medical data. However early prediction of hepatitis is a quite challenging task. Machine learning is an emerging scientific field in data science with the ways in which machines learn from experience. The aim of this project is to develop a system which can perform early prediction of hepatitis based on the condition of the patient with high accuracy by combining the results of different machine learning techniques

**2.Introduction:**

Machine learning is the scientific field dealing with the ways in which machines learn from experience. For many scientists, the term “machine learning” is identical to the term “artificial intelligence”, given that the possibility of learning is the main characteristic of an entity called intelligent in the broadest sense of the word. The purpose of machine learning is the construction of computer systems that can adapt and learn from their experience. With the rise of Machine Learning approach we have the ability to find a solution to this issue, we have developed a system using data mining which has the ability to predict whether the patient has hepatitis or not. The main aim is to allow computers to learn automatically without human intervention and also adjust actions accordingly. Hepatitis disease has a great deal of attention in medical research. The diagnosis of hepatitis disease is a challenging task, which can offer automated prediction about the liver condition of patient so that further treatment can be made effective. The early identification is the only remedy to stay away from the complications. Data mining has the ability to extract hidden knowledge from a huge amount of hepatitis-related data. Because of that, it has a significant role in hepatitis research, now more than ever. . Classification strategies are broadly used in the medical field for classifying data into different classes according to some constrains comparatively an individual classifier. The aim of this research is to develop a system which can predict the diabetic risk level of a patient with a higher accuracy. This research has focused on developing a system based on three classification methods namely, Decision Tree Classifier, Logistic regression algorithms.

3.**Literary survey:**

Data mining is widely used in the medical field such as prediction of hepatitis disease since it is a multidisciplinary field. Using data mining researchers are developing various techniques in-order to predict the hepatitis diseases with high accuracy. Large no. of research work is carried out for medical diagnosis for various diseases.

Life Prognosis of hepatitis patients can be predicted by using classifier such as Support Vector Machine. In support vector machine, dataset is classified into training and testing data. Support vector machine analyze the training data and makes prediction on testing data. Predictive accuracy of classifiers can be enhanced by applying the techniques of feature selection. In this paper, Wrapper methods were incorporated to remove noise features before classification. After removal of noisy attributes, accuracy of the algorithm was further increased. SVM algorithm provides more and improved accuracy with the 10 attributes identified using wrapper method using a data mining tool called weka. Data mining concepts and techniques [2] provide us the how to preprocess data and handle with missing values.

A.H.Roslina [2] revealed that the hepatitis patients are the individuals who require continuous specific clinical treatment to diminish the demise rate. By logical examination discoveries information and AI innovation, for example, Support Vector Machines(SVM),the order and prediction of their life finding should be possible. Regardless of whether, it can‟t guarantee with the point of the considerable number of highlights esteems in the information are identified with each other. Their investigation shows the upgrade of forecast among information associating highlight determination technique prior to arrangement strategy.

Fadl Mutaher et al. [3] introduced the relative investigation in the prognostic of hepatitis information utilizing rough set method. It was discovered that the presentation and time taken to run the hepatitis information is quick in Naïve Bayes algorithms. The outcomes acquired were contrasted and different calculations like Decision Tree algorithm, Random Forest algorithm, Logistic Regression and SVM algorithm.

Duygu et al. [4] proposed a keen hepatitis conclusion framework utilizing Principle Component Analysis(PCA) which is underlined on highlight extraction and decrease of information for productive order.

Using inconsistency rate estimation and discretization, Huang et al. applied a filter-feature selection approach to a medical claims database to determine the adequacy of antidepressant usage length. They used algorithms similar to logistic regression and decision tree. Their findings suggest that using the filter-based feature selection approach to lower the dimensionality of healthcare databases can be feasible and efficient[5].

Prediction of risk score for hepatitis disease using associative classification and hybrid feature subset selection was obtained by Jabbar Akhil[5] used Feature selection as a pre- processing step in used to reduce dimensionality, removing irrelevant data and increasing accuracy and improves comprehensibility. Associative classification is a recent and rewarding technique that applies the methodology of association into classification and achieves high classification accuracy. Most associative classification algorithms adopt exhaustive search algorithms like in Apriori, and generate huge no. of rules from which a set of high quality of rules are chosen to construct efficient classifier.Effective attributes are selected through statistical methods such as Correlation coefficient

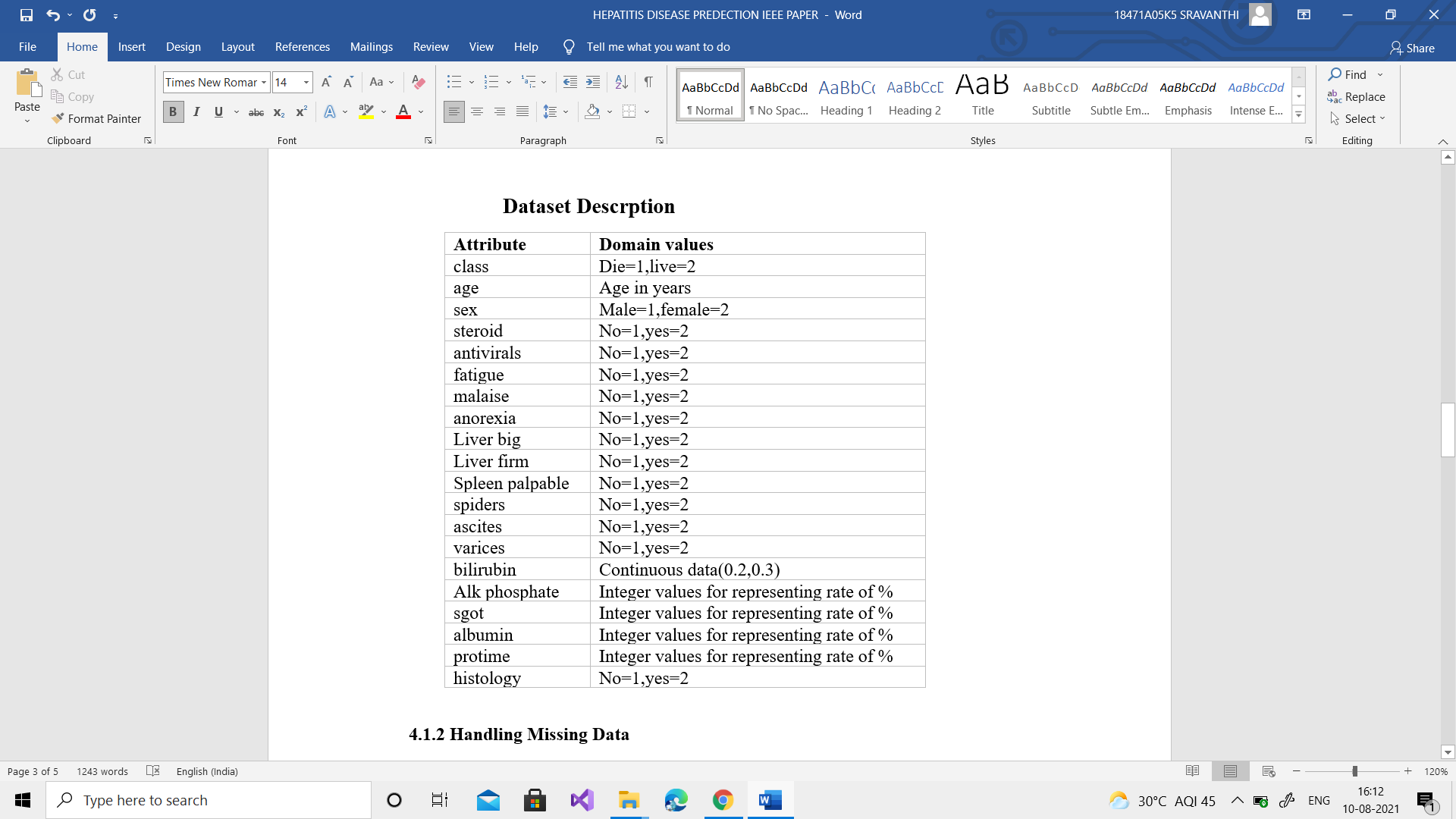
**4. Proposed System**

**4.1 Data Pre-processing:**

Before feeding data to an algorithm we have to apply transformations to our data which is referred as pre-processing. By performing pre-processing the raw data which is not feasible for analysis is converted into clean data. In-order to achieve better results using a model in Machine Learning, data format has to be in a proper manner. The data should be in a particular format for different algorithms.

**4.1.1 Dataset Description:**

Dataset collected from UC Irvine machine learning repository which consists of 155 instances and 19 attributes with the class stating the life prognosis yes (or) no. The dataset consist of 14 nominal attribute and 6 multi-valued attributes. The attributes which are identified are



**4.1.2 Handling Missing Data**

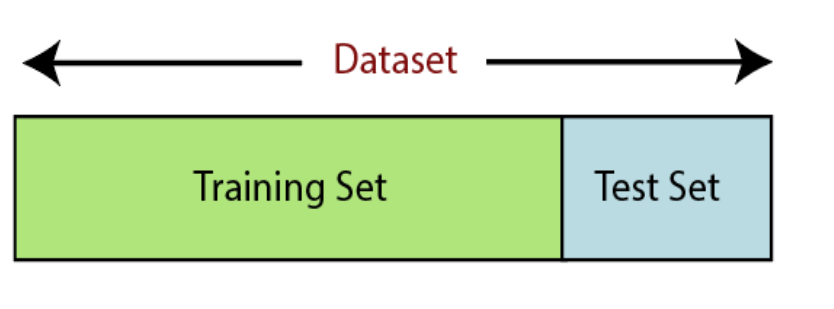
After importing dataset into the model,missing data can be handled in two ways:

a) By deleting those particular rows

b) By replacing those values with mean or median after pre processing data looks like



**4.1.3 Splitting the Set of data**

We split our 80 percent of dataset into a training system data after treating missing data and remaining data in the test collection ****

**4.1.4 Feature Scaling**

Feature scaling is the final step of data preprocessing in machine learning. Using Standard scalar we have standardized the independent variables of our dataset in a specific range.

**4.2 Correlation**

We have used correlation matrix and constructed correlation heatmap for identify highly correlated features based on it and remove them.



Fig 3: Correlation Heat Map

## Algorithms Used

* + 1. **Logistic Regression**

Logistic regression is a supervised classification algorithm in which we analyze the dataset using logistic function and predict the output.

**4.1.2 Decision Tree**

A decision tree is also a supervised classification algorithm that automatically constructs a decision tree from a given (training) dataset and predict the output.

**4.1.3 Random Forest:**

Random forest algorithm generates decision trees on information tests and afterward gets the forecast from every one of them lastly chooses the best yield via voting

**4.1.4 Gaussian Naïve Bayes**

Naive Bayes classifiers are a collection of classification algorithms based on **Bayes’ Theorem**.It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other.

## Data Visualization

**4.2.1Histogram**

## 

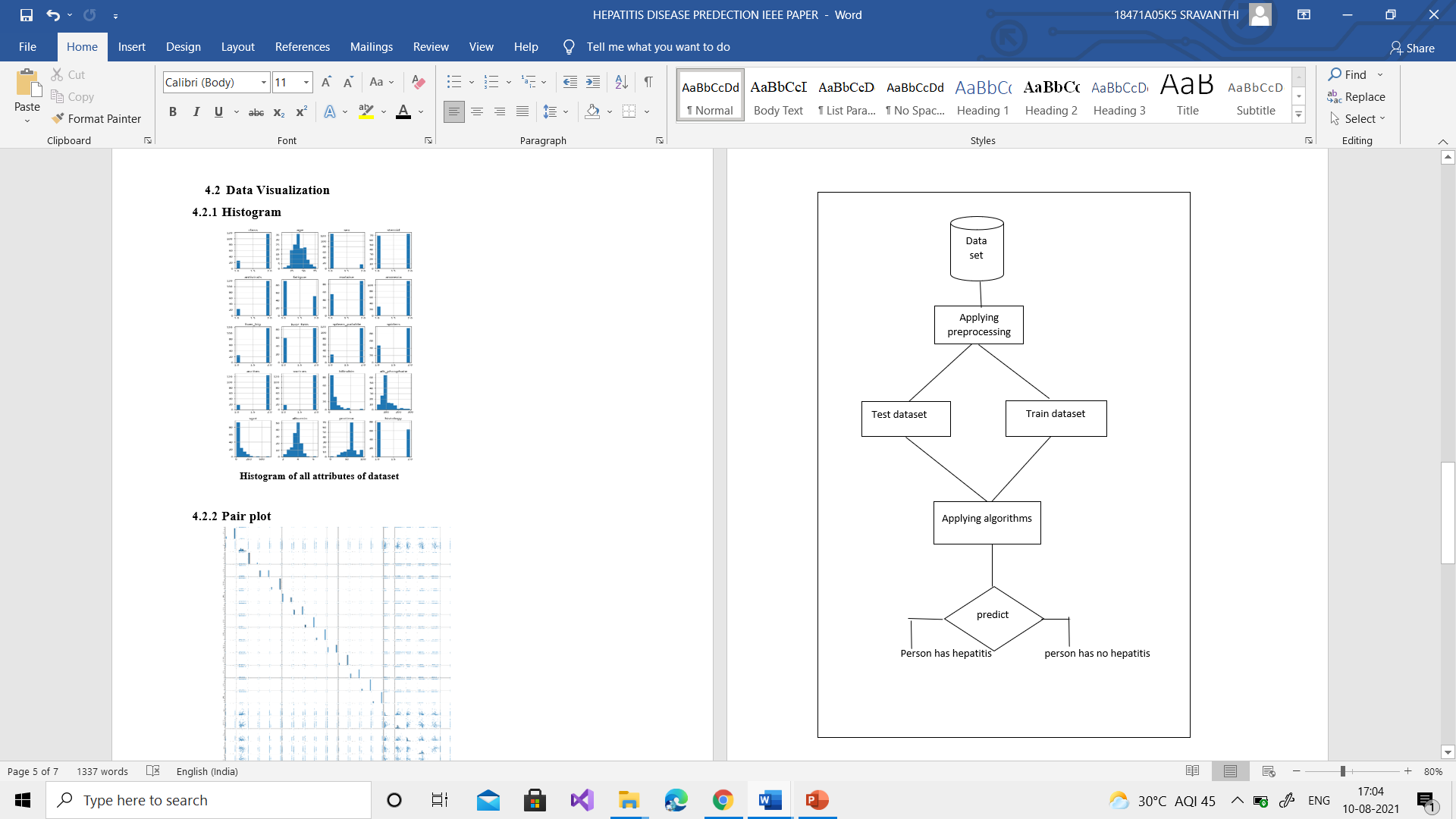
## Histogram of all attributes of dataset

## Pair plot

## 

Pair-plot of all attributes of dataset

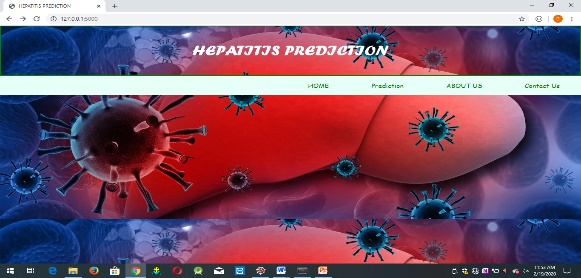
## Proposed Model

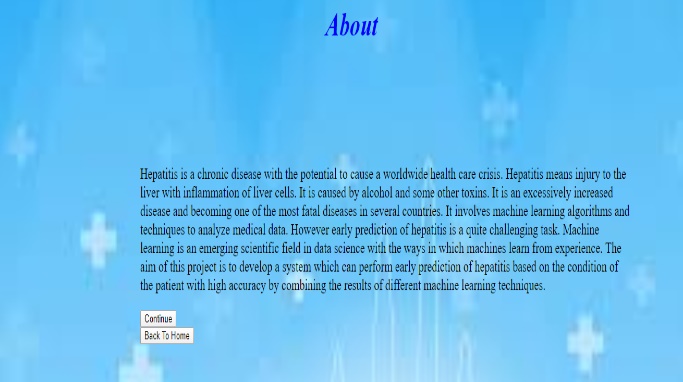


**5.Result Analysis**

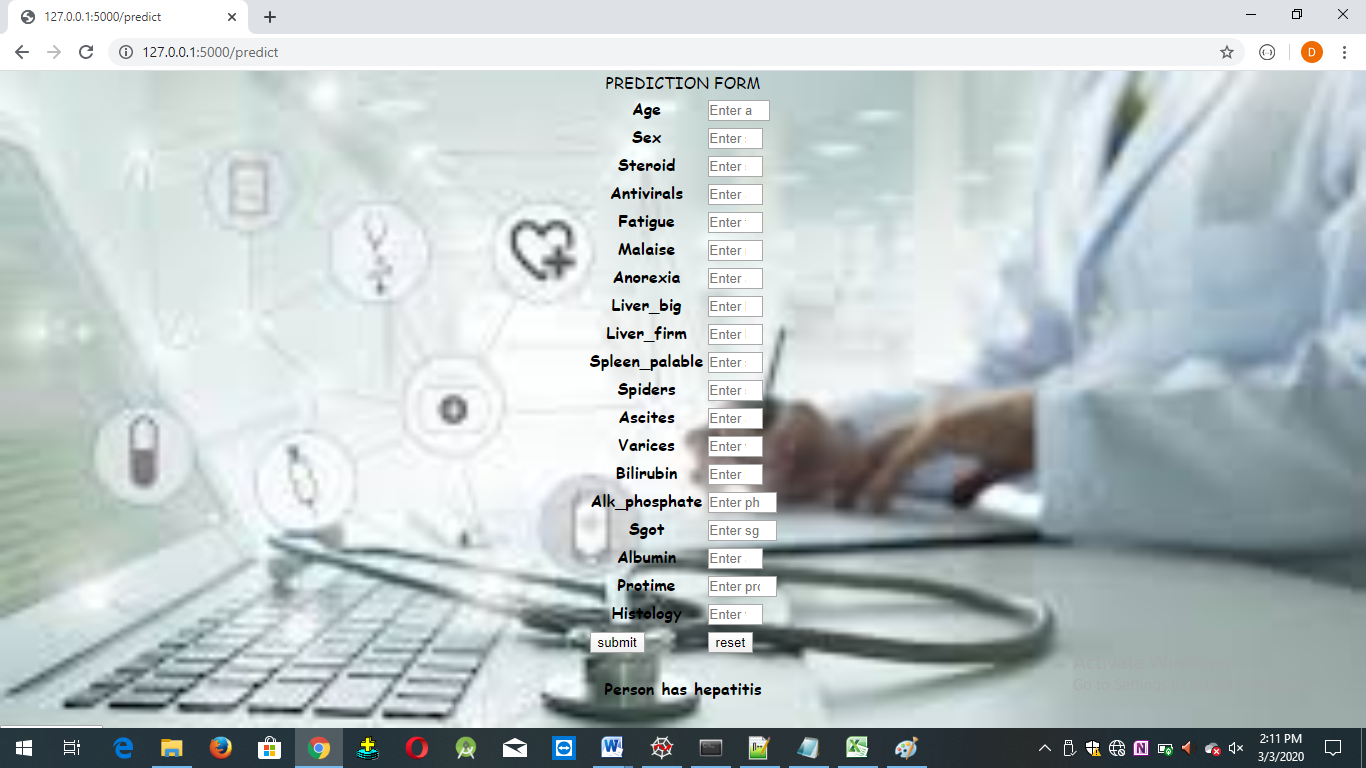


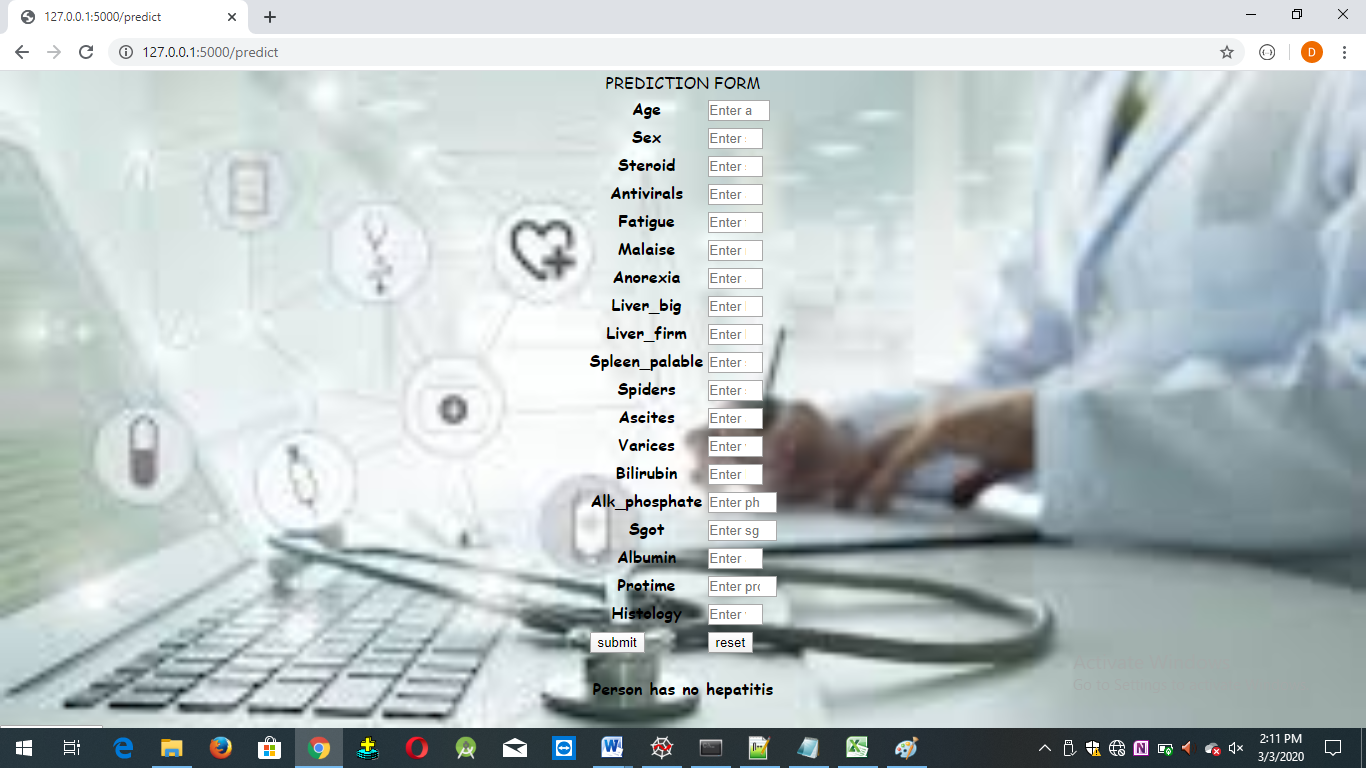
**6.output screens**





**7.Testcase**





**8.conclusion**

We have used 5 algorithms like Decision Trees, Random Forests, Naive Bayes, SVM and Logistic Regression in-order to predict presence or absence of hepatitis disease . The accuracy varies for different algorithms. The accuracy for Decision tree algorithm is 68.42. The accuracy for Naïve Bayes algorithm is 84.21 The accuracy for Logistic Regression algorithm is 84.21. The accuracy for SVM algorithm is 84.21.

In most of the cases highest accuracy is given when we have used Random Forest algorithm using which is nearly 85%.

**9.References**

1. [https://www.kaggle.com/harinir/hepatitis](https://www.kaggle.com/harinir/hepatitis%20)
2. https://ieeexplore.ieee.org/abstract/document/9032585