CS5710:Machine learning

Project + increment

PREDICTION OF HEPATITIES DISEASE USING MACHINE LEARNING ALGORITHMS.

CRN - 22921

Team members:

Kasoju sushma – 700747358

Mahesh kumar Uppu – 700741747

Maheswari pulagam – 700744329

Varna Nemulla – 700744920

GOALS AND OBJECTIVES:

Motivation:

Many Hepatitis B (human) carriers are utterly ignorant about their diseases and treatment options. A chronic stage of hepatitis, which is nearly untreatable and so expensive that a poor person could not afford such expenses, is brought on by a lack of adequate medical facilities, poor economic standing, incompetent medical staff, and ignorance about the disease and its prevention. Although there are immunizations, there is still no proven treatment for hepatitis. Hepatitis also places a significant financial strain on the healthcare system due to the expense of treating liver failure.

Many afflicted persons can be saved by early disease prediction and proper diagnosis. The main goal of the research is to analyze data from a hepatitis dataset using various classification approaches in order to precisely predict the outcome in each example of data. The following are the paper's main contributions:

- Measuring useful classification accuracy for predicting hepatitis illnesses.
- Evaluation of different machine learning techniques using the hepatitis dataset .
- Find the algorithm that performs the best for predicting hepatitis illnesses.

Significance:

The prediction of hepatitis disease using machine learning has significant important in healthcare. Here are some of the key significance of prediction of hepatitis disease.

- Early detection of hepatitis using machine learning can allow healthcare professionals to intervene earlier, leading to better treatment outcomes.

- Personalized treatment leads to improved accuracy in predicting hepatitis, which is critical for effective diagnosis and treatment .

Implementing the algorithms by understanding the decision tree algorithm and support vector machine and other algorithms.

Objective:

In the field of medicine, it is difficult to identify hepatitis disease in a patient's body at an early stage. Currently, if we look at the medical sector, we can see that the volume of health-related data is growing daily. The findings of patients' diagnostic tests and various clinical reports are key sources of information for the healthcare industry. By observing the hidden pattern and the linked features that are present in the dataset, it is used to determine the class name from the dataset. The patient's hepatitis status can be determined using both the concealed pattern and the linked features. Its method of operation is comparable to an expert system.

Nonetheless, a lot of machine learning methods are employed in prediction. But finding the ideal method is a difficult task. The goal of this research is to apply several machine learning approaches to detect hepatitis. In order to determine the best tool for diagnosing hepatitis disease, multiple ML algorithms were used to compare the accuracy for a specific data set. Support Vector Machine (SVM) and K Nearest Neighbor (KNN) techniques are used in this study to accurately forecast the disease.

Features:

The features are essential in ensuring accurate predictions and effective treatment plans . here are some of the key features are age, sex, steroid, antivirals, fatigue, malaise, Anorexia, Anorexia, Liver_big, Liver_firm, Spleen_palpable, Spiders, Ascites, Varices, Bilirubin, alk_phosphate, sgot, albumin, protime were being used in the dataset to get better results accordingly with other existing problem.

Increment :

Dataset:

Based upon data of a person we calculate whether a person has hepatitis disease or not .

This csv file has 17 datasets.

Age- age of a person

Sex- M/F

Steroid

Antivirals

Fatigue

Malaise

Anorexia

Liver_big

Liver_firm

Spleen_palpable

Spiders

Ascites

Varices

Bilirubin

alk_phosphate-

sgot- serum glutamic oxaloacetic transaminase.

albumin

protime

Detail design of features:

There are 17 attributes included in our dataset. Age, sex, steroids, antivirals, tiredness, malaise, anorexia, big, firm, palpable liver, spleen, spiders, ascites, varices, bilirubin, alk phosphate, sgot, albumin, protime, histology, and class are some of the other factors.

Datasets to achieve the aim and objective of this work online hepatitis datasets were used.

The Hepatitis patients' results based on the datasets are classified into two classes, either "Live" or "Die" based on predefined attributes. The hepatitis dataset contains 19 attributes and 156 data samples. For this work, due to the amount of time it takes to train the perceptron, 156 data samples are selected from the original hepatitis dataset with 14 attribute values. The classes are renamed as 1 and -1 (for Live and Die respectively) for use with the perceptron.

To predict hepatitis for unclassified patients who will survive or die if they possess certain attributes of symptoms as indicated in the dataset. The K-NN and the perceptron algorithm were also implemented to predict this disease. As mentioned above, the confusion matrix and Rand Index were used to evaluate the performance of these algorithms and the results show that the Decision Tree and K-NN classifiers predict hepatitis better than the perceptron classifier. We could speed up the classification algorithms using dimensionality reduction techniques.

Several values in our dataset are missing. There are numerous methods available to deal with missing values. In our

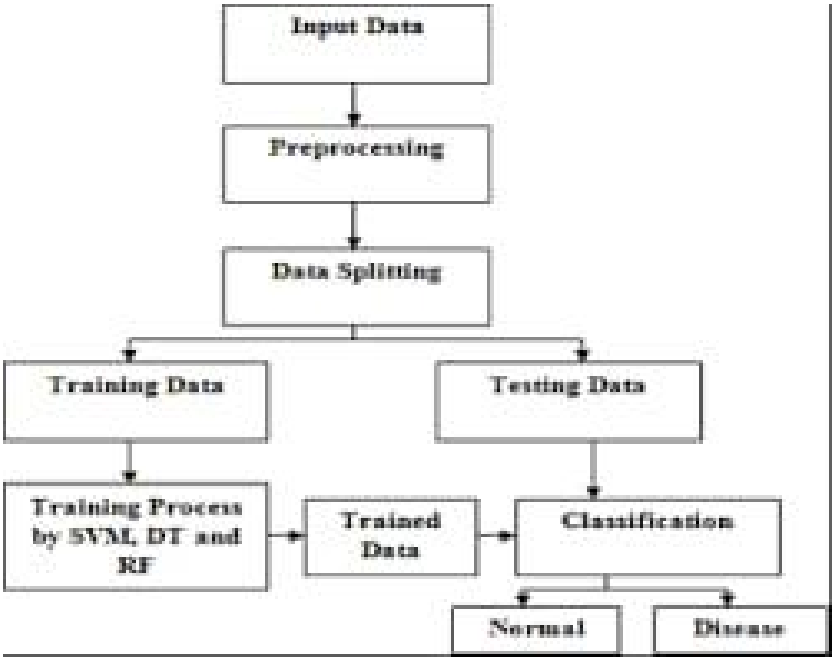
study, we first examined to see which records had missing data, then we eliminated such instances as in a prior study on hepatitis, and we used a total of 156 records to train the classifier. Because records with missing data occasionally may have detrimental consequences on classification accuracy, we took this action. Missing values could make classification less accurate.

Analysis:

It's the first and foremost stage of the any design as our is a an academic leave for essentials amassing we followed of IEEE Journals and Amassed so numerous IEEE Relegated papers and final tagged a Paper designated " Individual web revisitation by setting and substance significance input and for analysis stage we took arbiters from the paper and did literature check of some papers and amassed all the essentials of the design in this stage

Implementation:

Classification Techniques Classification is one of the major techniques used in data mining for prediction of diseases. It is used to classify each disease stored as one row of a dataset into a predefined class such as "die" or "live". Classes are sometimes called "labels" or "categories". For this research, the following classification techniques will be used due to their flexibility and uniqueness.



Preliminary Results:

As we have worked on three different algorithms, with the insights we have concluded by comparing the accuracies. We decided to add few more algorithms for better accuracy**.**

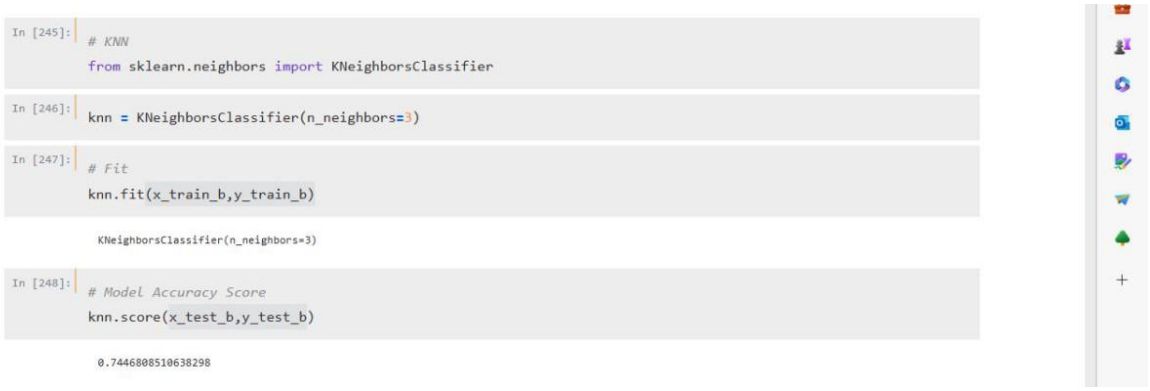
Project management:

Work completed:

Description:

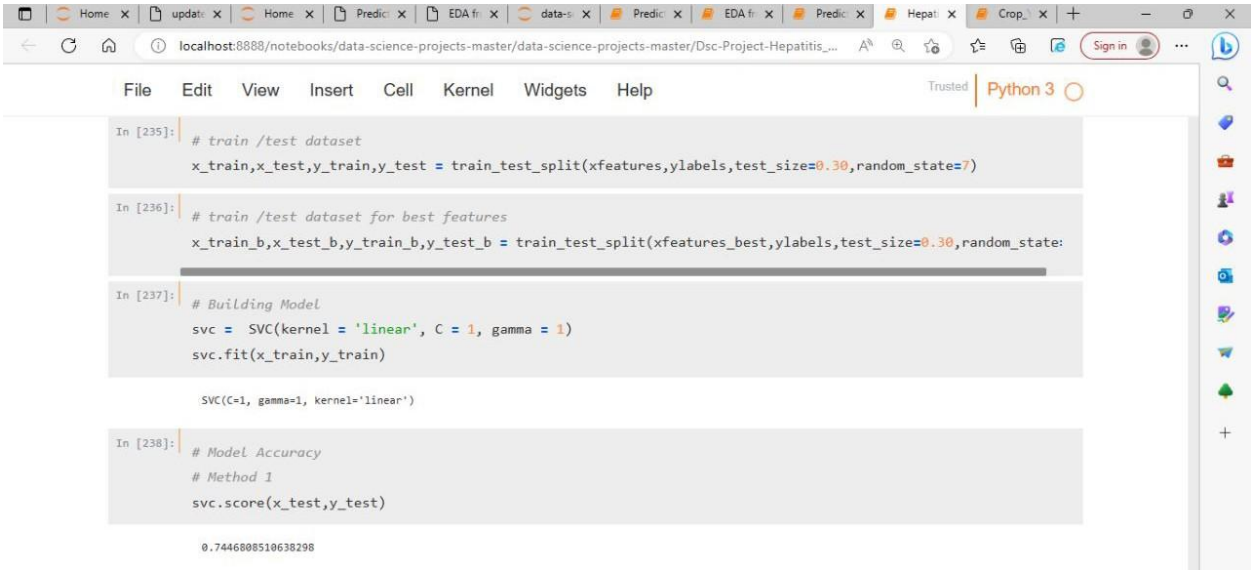
Decision tree algorithm : the datasets is split in to smaller subsets based on a set of conditions or features. The conditions are selected based on their ability to separate the data into homogeneous groups. The algorithm continues to split the datasets until a stooping criterion is met, such as the maximum depth of the tree or the minimum number of samples required to split a node .

K-nearest neighbor: knn can be used to predict the likelihood of a patient having hepatitis based on certain input features such as age, gender, symptoms . knn algorithm find the k nearest neighbor to give patient record based on the similarities of their features value and the use the labels of those neighbor to predict the label of the given patient record.



```
In [245]: # KNN
          from sklearn.neighbors import KNeighborsClassifier

In [246]: knn = KNeighborsClassifier(n_neighbors=3)

In [247]: # Fit
          knn.fit(x_train_b,y_train_b)

          KNeighborsClassifier(n_neighbors=3)

In [248]: # Model Accuracy Score
          knn.score(x_test_b,y_test_b)

          0.7446808510638298
```

Svm: the basic idea behind the svm algorithm is to find the two classes of data with the largest margin , while minimizing the misclassification error. In case of hepatitis disease prediction , the two classes would be patients with and without hepatitis.



```
In [235]: # train /test dataset
          x_train,x_test,y_train,y_test = train_test_split(xfeatures,ylabels,test_size=0.30,random_state=7)

In [236]: # train /test dataset for best features
          x_train_b,x_test_b,y_train_b,y_test_b = train_test_split(xfeatures_best,ylabels,test_size=0.30,random_state=

In [237]: # Building Model
          svc =  SVC(kernel = 'linear', C = 1, gamma = 1)
          svc.fit(x_train,y_train)

          SVC(C=1, gamma=1, kernel='linear')

In [238]: # Model Accuracy
          # Method 1
          svc.score(x_test,y_test)

          0.7446808510638298
```

Responsibility (Task, Person):

Most of the development work is already completed using KNN and SVM algorithms and results are compared with different datasets and analyzed the outputs to draw conclusions with the results. Both person and the tasks are mentioned below.

Sushma Kasoju: Backend logic for algorithms

Mahesh Kumar Uppu: Research and idea

Maheshwari Pulagam: front end design and working on dataset

Varna Nemulla: documentation and testing

Contributions (members/percentage):

Sushma Kasoju: 25%

Mahesh Kumar Uppu: 25%

Maheswari Pulagam: 25%

Varna Nemulla: 25%

Work to be completed:

Description:

Rigorous testing must be completed with all invalid and negative scenarios to ensure everything is working well. And, in the coming part, we would like to include xgboost and random

forest algorithms to improve the accuracy. Both person and the tasks are mentioned below.

Responsibility (Task, Person):

Sushma Kasoju: testing and code changes after testing if required.

Mahesh Kumar Uppu: Analyzing results.

Maheshwari Pulagam: Documenting results and issues

Varna Nemulla: Testing

Contributions (members/percentage):

Sushma Kasoju: 25%

Mahesh Kumar Uppu: 25%

Maheshwari Pulagam: 25%

Varna Nemulla: 25%

Issues/Concerns:

1. Comparing the accuracies

References/Bibliography:

1. M. J. Nayeem, S. Rana, F. Alam and M. A. Rahman, “Prediction of Hepatitis Disease Using K-Nearest Neighbors, Naïve Bayes, Support Vector Machine, Multi-Layer Perceptron and Random Forest,” 2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD), Dhaka, Bangladesh, 2021, pp. 280-284
2. V. K. Yarasuri, G. K. Indukuri and A. K. Nair, “Prediction of Hepatitis Disease Using Machine Learning Technique,” 2019 Third International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), Palladam, India, 2019, pp. 265-26
3. G. V. Nivaan and A. W. R. Emanuel, “Analytic Predictive of Hepatitis using The Regression Logic Algorithm,” 2020 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), Yogyakarta, Indonesia, 2020, pp. 106-110
4. P. Idrovo-Berrezueta, D. Dutan-Sanchez, R. Hurtado-Ortiz and V. Robles-Bykbaev, “Data Analysis Architecture using Techniques of Machine Learning for the Prediction of the Quality of Blood Fonations against the Hepatitis C Virus,” 2022 IEEE International Autumn Meeting on Power, Electronics and Computing (ROPEC), Ixtapa, Mexico, 2022, pp. 1-7
5. T. I. Trishna, S. U. Emon, R. R. Ema, G. I. H. Sajal, S. Kundu and T. Islam, “Detection of Hepatitis (A, B, C and E) Viruses Based on Random Forest, K-nearest and Naïve Bayes Classifier,” 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kanpur, India, 2019, pp. 1-7\

6. M. K. Lee, J. H. Paik and I. S. Na, “Outbreak Prediction of Hepatitis A in Korea based on Statistical Analysis and LSTM Network,” 2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), Fukuoka, Japan, 2020, pp. 379- 381
7. T. M. Ghazal, S. Abbas, M. Ahmad and S. Aftab, “An IoMT based Ensemble Classification Framework to Predict Treatment Response in Hepatitis C Patients,” 2022 International Conference on Business Analytics for Technology and Security (ICBATS), Dubai, United Arab Emirates, 2022, pp. 1-4