PREDICTION OF HEPATITIS DISEASE USING MACHINE LEARNING ALGORITHMS

Kasoju Sushma- 700747358

Email:sxk73580@ucmo.edu

University Of Central Missouri, MO, USA.

Department Of Computer Science.

Mahesh Kumar Uppu- 700741747

Email: mxu17470@ucmo.edu

University Of Central Missouri, MO, USA.

Department Of Computer Science.

Maheswari Pulagam-700744329

Email:mxp43290@ucmo.edu

University Of Central Missouri, MO, USA.

Department Of Computer Science.

Varna Nemula- 700744920

Email: vxn49200@ucmo.edu

University Of Central Missouri, MO, USA.

Department Of Computer Science.

ABSTRACT

The life-threatening illness hepatitis is brought on by the enlargement of the liver. Hepatitis viruses (hepatitis B, hepatitis C, and viral hepatitis) may infect the liver and cause illness. For extensive diagnoses, manual inspection may be taxing, which can have a serious financial effect on the individual health program. Hepatitis prediction software that uses machine learning methods like decision trees, kNN, Random Forest, and other algorithms has promise as a useful diagnostic tool. A decision tree is built in this research to predict hepatitis, and various algorithms like kNN, Random Forest, support vector, and xgboost are used to predict the same illness. The degree to which each of the five learning algorithms predicts hepatitis is compared to the others. Graphs are used to make the comparisons. Biomedical data analysis often makes use of data mining methods. These methods have produced effective outcomes for the diagnosis and categorization of illnesses as well as the prediction of disease severity. All ages are susceptible to hepatitis, an inflammation of the liver. Hepatitis is estimated to impact millions of individuals worldwide. Numerous people may be saved by a timely and accurate hepatitis diagnosis. Due to the difficulty in clinically diagnosing the illness in its early stages, hepatitis poses a significant problem for public health care providers. The findings of several strategies in terms of accuracy and training time are presented in this research along with distinct data mining techniques utilized for hepatitis diagnosis. we have predicted hepatitis disease by using different data mining techniques. Besides this, we have proposed a decent way by which we can improve the performance of our prediction models. We have handled missing values present in ourdataset by removing the observations having missing values. We have measured accuracy, precision, recall, F1-score and ROC whose help us to compare the performance of the classification models. Removing the observations having missing values as well as the info-gain feature selection technique has helped us to improve the accuracy of our prediction models. We have got best performance from Random Forest

KEYWORDS

knn , svm , xgboost , random forest , decision tree.

INTRODUCTION

Hepatitis is defined as an inflammation of the liver and most commonly it is occured by a viral infection causing overall 1.5 million deaths all over the world every year. This viruses tent to target the cells in the liver.Hepatitis can be occurred due to viruses, bacteria, drugs, etc. This disease can be classified as Acute or Chronic. There are five main Hepatitis viruses, referred to as types A, B, C, D and E.Hepatitis B virus is the most familiar problem and a serious health issue which affects all most 2 billion people all over the world including 350 million chronic carriers from which 75% are in the Asia-Pacific zone . Chronic hepatitis C is a worldwide public health issue estimated to have over 180 million patients that are expected to increase three times by 2020 and worldwide up to 250 million people are thought to be attacked by hepatitis C .Hepatitis E and A are caused by oral infection, contaminated water and unhygienic food . It is specifically occurs in developing countries especially rural communities due to the poor hygienic conditions. Hepatitis B, C and D are considered as global health problem and can easily transfer to individuals through intercourse, blood and birth from infected mothers . Hepatitis C more principally spread through blood transfusions and use of un sterilized needles and equipment. For the prevention of Hepatitis B effective vaccine is available globally and locally although

very expensive for poor people to afford [keyani1]. For Hepatitis C, globally no vaccines are available and therefore preventive measures and awareness information's disseminated is mandatory.HEV also occurs through transmission by the fecal–oral same as HAV

In the human body, the liver is an important organism that performs an important role in many body functions. Hepatitis disease is responsible for liver damage. Due to this reason, a patient can die. In medical science, the detection of hepatitis disease within a patient's body at an early stage is a challenging task. At present, if we observe the medical industry, then we can see that day-by-day amount of data related with health is increasing. Data mining is a field related to machine learning has the ability to manage huge data as well as solve the complex problem very efficiently so that researchers can take correct decision from a huge database. It is applied to identify unknown patterns and find out valuable information from a huge dataset. Due to this reason, it is the first choice for the researcher in order to solve many problems in the real world. However, health care industries gather significant information from different clinical reports and patient's diagnostic test results. It is applied to know the class name from the dataset by noticing the unseen pattern along with correlated features present in the dataset. Both the hidden pattern and the correlated features helps to distinguish either the patient is affected by hepatitis disease or not. Its working approach has the similarity with an expert system. Besides this, it will also save cost and diagnosis time. However, there are many machine learning algorithms that are used for prediction purposes. It is a difficult work for us to find out the best technique There are three contributions in our study. The first contribution is that we have collected real world diagnostic datasets of hepatitis disease having different types of features of 155 patients from the UCI machine learning repository. Our second contribution is that we have found out the unnecessary features by using info-gain feature selection procedure with ranker search which helped us to increase our classification model performance. Our third contribution is that we have made a performance comparison of our five techniques as well as compare the performance result with the previous research result and also evaluate the prediction outcome based on different risk factors.

Machine learning is one of the effective ways that can be used in the biomedical world. This is seen from how machine learning approaches an approach to making good and automatic algorithms that can be used in the process of diagnosis or disease prediction for the decision making process looking at the amount of data generated in the health field, and also the difficult data management process, various approaches are offered using existing machine learning methods. Some research was carried out by looking at some parameters such as Age, Sex, Steroids, Antivirals, Fatigue, Malaise, Anorexia, Liver Big, Liver Firm, Spleen Palpable, Ascites Spiders, Arices, Bilirubin, Alk 'Phosphate', Sgot, Albumin, and Protime and Histology, to see the advantages of ensemble learning.

Medical diagnosis is a vital and difficult process that requires precise identification. It is crucial to identify the illness at the right moment and to begin treatment as soon as possible. The important organ of a human body is the liver. Hepatitis, which results in liver inflammation, is one of the serious disorders that impairs the liver's ability to operate. The main factor for Hepatitis disease is the presence of virus in liver . Hepatitis is a worldwide disease with high mortality rate. If accurate measures are not taken in proper time, it may affect the vital functions of the body and may cause to cirrhosis, severe scarring and increase the risk of liver cancer . Early detection through proper diagnosis and proper medication can cure the disease. For diagnosis of any disease, the two important things are:
(I) The selection of right parameters of diagnosis and
(II) proper analysis of the data with an experienced expertise. Machine Learning (ML) is the tool which could make a system to learn by itself by detecting different patterns and different relationships for the given data using different algorithms . This would enable automatic diagnosis of any diseases, where the two important things considered with utmost care are: selection of parameters and the tool used for analyzing these parameters. In this work, a study of three different tools that are used for Hepatitis prediction namely: KNN, SVM , xgboost , random forest , decision tree are carried out. Different researches have undergone for the diagnosis and the prediction of diseases using machine learning techniques . Somaya et al. evaluated different machine learning techniques in the prediction of advanced fibrosis that incorporates serum biomarkers. Haydon et al. used artificial neural networks for the prediction of cirrhosis in patients using routine clinical host and viral parameters . An automatic diagnosis system was proposed by Jiaxin et al. using extreme learning machine on serum indices data of patients to predict the fibrosis stage and inflammatory activity grade of chronic hepatitis C . Sushrutha et al. proposed a hybrid model for the prediction of hepatitis . They have developed a combination of genetic search algorithm and multilayer perceptron technique. The main objective of this work is to perform a comparative study for a specific dataset by training the same dataset using different ML tool and choosing those best tool for diagnosis of Hepatitis disease. Comparative study of various technique is performed based on the accuracy. In this work, we used SVM , xgboost , decision tree , random forest, KNN ML algorithms.

MOTIVATION

Many Hepatitis 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.

Different researches have undergone for the diagnosis and the prediction of diseases using machine learning techniques. Medical diagnosis is an important and a quite complex task which requires accurate identification. It is important to diagnose the disease at proper time and to be cured at the earliest. Liver is the vital part of a human body. One of the severe diseases that affect the functionality of liver is hepatitis, which causes inflammation of the liver. The main factor for Hepatitis disease is the presence of virus in liver . Hepatitis is a worldwide disease with high mortality rate. If accurate measures are not taken in proper time, it may affect the vital functions of the body and may cause to cirrhosis, severe scarring and increase the risk of liver cancer . Early detection through proper diagnosis and proper medication can cure the disease

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.

MAIN CONTRIBUTIONS & OBJECTIVES

The primary objective of the paper is analysis of data from a hepatitis dataset using different classification techniques to predict the result accurately in each case of data. Major contributions of the paper are:

- To measure useful classification accuracy for the prediction of hepatitis diseases.
- Comparision of various data mining algorithms on the hepatitis dataset.
- Identify the best algorithm performance for the prediction of hepatitis diseases.
- To develop an accurate MACHINE LEARNING algorithm that can predict the likelihood of a patient having hepatitis disease and able to classify patients as either having hepatitis or not having hepatitis with a high degree of accuracy.
- To evaluate the performance of different machine learning algorithms and identify the best performing algorithm for predicting hepatitis disease.
- Simple to improvise - The software is easy to create and accuracy will be shown instantly.

RELATED WORK

In this section, we examine similar publications, notably those that use data mining methods to predict the presence of hepatitis.

Examined and used a number of machine learning techniques to analyze data from the NS3 serine protease cleavage of the hepatitis C virus. According to the research, hepatitis C virus detection may be done using machine learning techniques. ROC-AUC curves were used to examine the effectiveness of these methods and assess the model performance.

Used a data mining technique called a decision tree to predict the locations where the hepatitis C virus would cleave. The GINI index served as the algorithm's spitting criterion, and it was implemented using the Matlab toolkit for classification and regression trees (CART). The decision tree classifier model produced findings with a prediction accuracy of 96%, which was not the best. However, the rules established by the decision tree prediction model made the results more illuminating.

Presented a decision tree algorithm-based prediction method for hepatitis diagnosis. The decision tree was built using the C4.5 algorithm, which focuses on 19 attributes from the dataset, including age, sex, steroids, antivirals, spleen, fatigue, malaise, anorexia, big, firm, and spidery livers, ascites, varices, bilirubin, albumin, time, and histology for the diagnosis of hepatitis. The algorithm's accuracy is shown using the confusion matrix, and the calculated accuracy is 85.81%.The hepatitis illness was investigated using several neural network algorithms, including Quick, Multiple, Dynamic, and RBFN, along with various parameters, including data size, learning cycle, and processing time, to accomplish the diagnostic accuracy and estimated error. Utilizing the confusion matrix, the neural network's accuracy was determined.

Investigated computer learning models to forecast the course of hepatitis C in veterans. A large dataset of people with chronic hepatitis C was used to create two machine learning models and assess their performance in predicting the development of cirrhosis.

Nireekshith.Yarraguntla implemented a system using biosensor that can predict Hepatitis viruses(HAV, HBV, HCV) by immobilizing specific antibodies on the sensing element of the sensor when the sample of the patient bearing hepatitis viruses surface antibody (anti-HbsAg) is settled on the sensing part, Biochemical interactions occur between Hepatitis viruses antigens and the antibodies standing on the sensing element. For this reason, beam structure is changed so that diseases can be detected.

S.L. Znoyko presented real time optical methods for finding the HBV surface antigen. An optical technique has been elaborated for screening of functionalized nanoparticles and reagents for the design of immunochromatographic and capacitive biosensors. The developed immunosensors were tested for rapid quantitative detection of HBV surface antigen, which is the marker for diagnostic of hepatitis B.

E. Alipour designed a capacitive immunosensor to detect the HBV surface antigen. Their result show the presence of HBsAg which affects most of the time to the surface area of the capacitors and for this there is an enhancement in the capacitance. They found the results by developing an immunoassay capacitor using gold nanoparticles as amplifiers.

Chunyan Yang, JinJise Song gave a model on Hepatitis diagnosis which is based on GA-BP neural network algorithm and fuzzy integral. In order to develop the accuracy of hepatitis diagnosis in computer-aided diagnosis system, the medical data from the network public database were optimized by GA-BP neural network algorithm. From the results of calculation, using fuzzy integral can greatly improve the accuracy of hepatitis auxiliary diagnosis.

Huina Wang, Yihui Liu, Wei Huang did their work to detect the risk factors for hepatitis B virus reactivation after the precise radiotherapy in patients with primary liver cancer (PLC). We use sequential forward selection and sequential backward selection to extract features which would be combined into an optimal feature subset, and then establish Bayesian and support vector machine (SVM) classification model. The experimental results showed that the key feature subset has a better classification performance than the initial
feature set clearly.

Sara Omer Hussien esented an overview of the recent state-of-the-art data mining techniques used for diagnosing hepatitis and gives,the performance of various techniques in term of training time and accuracy. Such review helps in the implementation, development and evaluation of
efficient clinical decision support systems; where
accurate diagnosis is the most important factor.

Jiaxin Cai, Tingting Chen, Xuan Qiu proposed an automatic diagnosis system of chronic hepatitis C using serum indices data of patients to predict the fibrosis stage and inflammatory activity grade of chronic hepatitis C by training the extreme learning machine. Due to the superiority of extreme learning machine such as simple structure and fast calculation speed, the presented automatic diagnosis system can achieve good diagnosis performance. The proposed automatic diagnosis system is test on real clinical cases of chronic hepatitis C based on serum indices. Experimental results explain that the performance of the given method overcomes that of the state-of-the-art baselines concerning the diagnosis of fibrosis stage and inflammatory activity grade of chronic hepatitis C.

Tasneem A. Gameel predicted the hepatitis C infection progression into cirrhosis or liver cancer. For the prediction of the disease progression, a knowledge discovery framework is proposed consisting of three phases: preprocessing, data mining and prediction. While the preprocessing phase focuses on the discretization of the training data, the data mining phase focuses on mining patients' records using a rule based classifier built by the proposed algorithm to generate a set of unique rules. Eventually, the predictor uses the rules to predict patients' disease progression.

We out a research on the effectiveness of machine learning techniques for predicting esophageal variations in chronic hepatitis C patients in Egypt. The goal of the research was to identify methods for diagnosing the condition by examining the information obtained from data via classification analysis and utilizing machine learning techniques for early prediction in cirrhotic patients based on their clinical examination. To accomplish this goal, the model employs six machine learning techniques, including Bayesian networks, Naive Bayes, Decision Trees, Support Vector Machines, and Naive Bayes.

Presented a research on hepatitis B surface antigen seroclearance prediction using machine learning techniques. Four algorithms—the decision tree (DCT), logistic regression (LR), extreme gradient boosting (XGBoost), and random forest—were used to build the models. Area under the receiver operating characteristic curve (AUC), a measurement tool, was used to choose the optimum model. XGBoost has the greatest predictive performance, with AUCs of 89%, Random Forest at 82%, Decision Tree at 61%, and Logistics Regression at 68%, respectively.

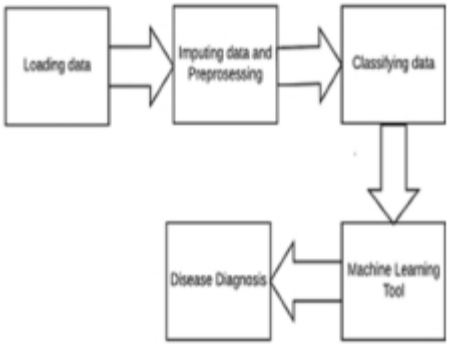
PROPOSED FRAMEWORK

In this work, the required data set is chosen from UCI repository, considering different clinical cases. This dataset consists of 156 instances with 19 attributes, one among the same attributes is the class to decide the life expectancy of a hepatitis patient. The 19 attributes are used. Machine learning algorithm such as SVM and KNN and other algorithms were applied to the dataset for training and testing. Restrictions apply. performed on the same dataset for performance analysis. Comparison was evaluated based on the prediction accuracy of the tool .
Most hepatitis A virus infections are the mostly unseen symptom and are detected only by the presence of IgM antibody. It is important to detect IgM antibody for HAV test. The increment of 4- fold in anti-HAV IgG antibody can also be used as a current infection. Isolation of the virus in cell culture is possible but not available in the clinical laboratory.

Attributes	Value
Age	Numerical value
Sex	male(1), female(2)
Steroid	no(1), yes(2)
Liver Big	no(1), yes(2)
Liver Firm	no(1), yes(2)
Spiders	no(1), yes(2)
Antivirals	no(1), yes(2)
Fatigue	no(1), yes(2)
Malaise	no(1), yes(2)
Spleen Palpable	no(1), yes(2)
Ascites	no(1), yes(2)
Varices	no(1), yes(2)
Varices	no(1), yes(2)
Bilirubin	0.39, 0.80, 1.20, 2.00, 3.00, 4.00
Alkaline Phosphate	33, 80, 120, 160, 200, 250
Aspartate transaminase	13, 100, 200, 300, 400, 500
Albumin	2.1, 3.0, 3.8, 4.5, 5.0, 6.0
Pro-time	10, 20, 30, 40, 50, 60, 70, 80, 90
Histology	no(1), yes(2)

The two most important serological tests for the diagnosis of early hepatitis B are the tests for HBsAg and for IgM antibody. Both occur in the serum earlier on the disease. The availability of HBsAg, Anti-HBV IgM, and HBeAg is indicating the early stage (acute) of hepatitis B. If only Anti-HBc IgG is presented it indicates Window period. The persistent HBsAg, HBeAg, and attendance of Anti-HBc IgG refer to the chronic stage with viral activity and infectivity. The emergence of Anti-HCV indicates the current infection by hepatitis C virus. There is another test namely RIBA (Recombinant immunoblot assay) for confirmation of HCV. HEV antibody test is not readily obtainable, the diagnosis is therefore typically made by detecting IgM antibody to HEV. High probability of infected by HEV in pregnancy period of a woman. Mode of transmission for HBV and HCV are almost the same but the fact which matters is the percentage. According to the percentage we categorize mode of transmission into Horizontal and Vertical.

The main processes involved in this model can be described as: loading the data, imputing the data and data preprocessing, classifying the data, applying ML technique and diagnose the disease. Steps involved in the process are described in detail in the following sections:



- **Loading the data**: The data set is extracted from UCI repository which consists of 155 instances with 20 attributes. Since ML learns from examples, sufficient and smoothened data have to be given to the network model. To get sufficient data, data imputation was performed on the available dataset.
- **Imputing the data and data preprocessing:** The presence of missing data in any field, if not handled wisely may result in incorrect prediction and affects the quality of the result. In this hepatitis database, out of 155instances 75 of them have missing values. To get sufficient data for training, validation and testing, the data augmentation technique was used. Data augmentation was performed thrice to increase the precision of the results. Out of all the instances of the dataset, the missing values corresponding to the instances are removed and the imputation process was performed on the basis of remaining data. The performance of the imputation method was evaluated using the metric error rate, in which the imputed data is compared with the data with non-missing values in the attribute.
- **Classifying the data**: After loading the data and preprocessing phase, classification of the data is carried out. Data classification is carried

out in two phases, namely training phase and classification phase. In the training phase, the data is classified into training set and validation set. After that, the classifier algorithm builds the classifier with the training dataset. In the second classification phase, the trained model is used for disease classification and the life expectancy of the Hepatitis person. The data is divided into training (60%), testing (20%) and validation (20%) in a stratified manner.

Applying ML tool and diagnose the disease: There are mainly three stages in implementing the machine learning code as training, validating and testing. In our research, with the given hepatitis dataset, initially the data was split in to these three categories using stratified splitting. After this, we train the data using suitable machine learning tool and then the data for validation is given to the network. Using the trained network, the testing data is validated, and this is the phase which gives the accuracy of prediction of life expectancy of patients with Hepatitis. SVM, KNN , xgboost , decision tree, Random Forest

Random Forest A group of decision tree classifiers make up the Random Forest classifier. To increase the variety of the decision trees, each decision tree in the forest is trained using a bootstrap sample or a subsample of the original training data. In each tree, samples are pushed from the root node to the leaf node by performing a binary test at each internal node. In In a binary test, a feature's value is compared to a threshold. In the training step, for each node a binary test on a particular feature is identified by optimizing the information gain in the training dataset. Depending on the outcome of the binary test, each training sample is sent to the corresponding child node and this process is recursively repeated until the number of samples falls below a certain threshold in a node.

K Nearest-Neighbour (KNN) is one of the easiest classification techniques. KNN is one of the initial options for these identification issues if there is no previous knowledge of the distribution of nation data. A benchmark classification algorithm is the KNN classifier. KNN Kanpur, India classifier performance determined by choice of K as well as the distance metric applied. In the
predetermining of the K value difficult when the points are uniformly distributed.

The classification itself is performed in two steps: finding K nearest neighbors in the training dataset, and: assigning a label based on majority voting amongst the K neighbors . The neighbors are determined by calculating the Euclidian distance between the new observation and the instances in the training dataset.

NCA-XG Boosting: The decision-tree-based ensemble Machine Learning approach known as XG Boost makes use of a gradient-boosting framework to increase both the accuracy and the speed of its predictions. XG Boost was developed by Microsoft Research and was given its current moniker in honour of the founder of the corporation. When it comes to solving prediction issues that include unstructured data, artificial neural networks often outperform all other algorithms and frameworks (images, text, etc.). Decision tree-based algorithms are widely acknowledged as the most effective method for the management of structured or tabular data sets that range in size from relatively small to relatively medium.

Two common types of ensemble tree algorithms are XG Boost and Gradient Boosting Machines (GBMs). Both of these ensemble tree approaches improve the performance of weak learners (in general, CARTs) by using the gradient descent architecture.

Since it has helped people and teams win almost every structured data competition on Kaggle, people and teams have developed a unique love for the tool known as XG Boost. Participants in these competitions are asked to submit data, after which statisticians and data miners compete to see who can construct the most reliable models for analysing and interpreting the data. Python was first used in the development of XB Boost, and then R took over. Because of the overwhelming demand for its services, XG Boost has begun providing package implementations for a variety of languages, including Java, Scala, Julia, and Perl, amongst others. The popularity of XB Boost has grown among the Kaggle community as a direct result of the new implementations that have been made possible as a result of these.

Decision tree The root of a decision tree is shown at the top and is depicted upside down. A condition or internal node, upon which the tree divides into branches or edges, is shown by the bold text in black on the left-hand figure. The choice or leaf, in this example, whether the passenger lived or died, is the end of the branch that no longer splits and is represented as red and green text, respectively.Even though a real dataset will have many more attributes and this will just be one branch of a much larger tree, you can't discount how straightforward this technique is. It is easy to see the relationships between the features and their value. The above tree is referred to as a "Classification tree" since its goal is to categorize passengers as having survived or having passed away. This process is more frequently known as "learning decision tree from data."

Regression trees are modeled similarly, except instead of predicting discrete variables like home prices, they forecast continuous values. The abbreviation CART, or Classification and Regression Trees, is used to describe Decision Tree algorithms in general.

Support Vector Machine: The goal of the support vector machine technique is to locate a hyperplane that clearly categorizes the data points in an N-dimensional space (N is the number of characteristics). There are a variety of different hyperplanes that might be used to split the two classes of data points. Finding a plane with the greatest margin—that is, the greatest separation between data points from both classes—is our goal. Maximizing the margin distance adds some support, increasing the confidence with which future data points may be categorised.

Decision boundaries known as hyperplanes assist in categorizing the data points. Different classifications may be given to the data points that lie on each side of the hyperplane. Additionally, the amount of features affects how big the hyperplane is. The hyperplane is essentially a line if there are just two input characteristics. The hyperplane turns into a two-dimensional plane if there are three input characteristics. When there are more than three characteristics, it gets harder to visualize. Support vectors are data points that are closer to the hyperplane and have an impact on the hyperplane's location and orientation. By using these support vectors, we increase the classifier's margin. The hyperplane's location will vary if the support vectors are deleted. These are the ideas that aid in the development of our SVM.

DATA DESCRIPTION

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.

Preprocessing stroke prediction datasets is important for machine learning. The dataset should be cleaned and organized in a way that is easy to use for the machine learning algorithm. The first step is to remove any invalid data points. Invalid data points can be caused by errors in the data collection process or by incorrect data entry. Invalid data points can also be caused by outliers in the data set. Outliers are data points that are far from the rest of the data points in the set. They can distort the results of the machine learning algorithm if they are not removed.
Because of missing values and/or noisy data, the quality of the raw data may be worse than the quality of the final forecast.

Standardizing the data means that all of the data points are converted to the same unit of measurement. This is important because it ensures that the machine learning algorithm is comparing apples to apples. The data sets are combined in the third stage. This is necessary if the data set is divided into multiple files. The fourth step is to label the data. This is necessary if the data set is not already labeled. Labeling the data means assigning a name to each data point. The fifth step is to remove any duplicate data points. Duplicate data points can distort the results of the machine learning algorithm.

The sixth step is to split the data into training and testing sets. The machine learning algorithm is trained using the training set. The testing set is used to test the accuracy of the machine-learning algorithm. The seventh step is to format the data. This is necessary if the data is not in a format that the machine learning algorithm can use. The eighth step is to filter the data. This is necessary if the data set is too large to use for the machine learning algorithm. The ninth step is to normalize the data. Normalizing the data means adjusting the data so that the mean is zero and the standard deviation is one. This is important because it ensures that the machine learning algorithm is comparing apples to apples.

The tenth step is to choose the machine learning algorithm. The machine learning algorithm is the algorithm that will be used to learn from the data set. The eleventh step is to choose the parameters for the machine learning algorithm. The parameters are the settings that the machine learning algorithm will use to learn from the data set. The twelfth step is to run the machine learning algorithm. This is the step where the machine learning algorithm is actually run on the data set. The thirteenth step is to evaluate the results of the machine learning algorithm. This is the step where the accuracy of the machine-learning algorithm is determined. The fourteenth step is to modify the machine learning algorithm if necessary. This is the step where the machine learning algorithm is modified based on the results of the evaluation. The fifteenth step is to repeat the steps from six to fourteen until the machine learning algorithm reaches the desired accuracy.

Data visualization is a powerful tool for understanding complex data sets. In machine learning, data visualization can be used to help identify patterns in data, understand the performance of a machine learning algorithm, and diagnose problems with a machine learning model. A correlation plot is a graphical representation of the correlation between two variables. In machine learning, it is often used to help identify relationships between input and output variables. The plot displays the strength and direction of the correlation and can help to identify relationships that may be useful for predictive modeling.

Online hepatitis databases were utilized to collect the data needed to accomplish the goal and purpose of this inquiry. 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. For usage with the perceptron, the classes are changed to 1 and -1 (for Live and Die, respectively).

to forecast whether unclassified individuals with hepatitis would live or die based on the presence of certain symptom characteristics as shown 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.

RESULTS

```
In [100]: # Building Model
          svc =  SVC(kernel = 'linear', C = 1, gamma = 1)
          svc.fit(x_train,y_train)

Out[100]: SVC(C=1, gamma=1, kernel='linear')

In [101]: # Model Accuracy
          # Method 1
          svc.score(x_test,y_test)
          results.update({"SVC":svc.score(x_test,y_test)})

In [102]: # Method 2
          y_pred=svc.predict(x_test)

In [103]: print(list(y_test))

          [2, 1, 2, 2, 2, 1, 1, 2, 2, 2, 1, 1, 2, 2, 2, 2, 1, 2, 2, 1, 2,
          1, 1, 2, 1, 2]

In [104]: accuracy_score(y_test,svc.predict(x_test))

Out[104]: 0.7446808510638298

In [105]: accuracy=metrics.accuracy_score(y_test_b, y_pred)
          print("Accuracy:",accuracy)
          print(confusion_matrix(y_test_b, y_pred))
          print(classification_report(y_test_b, y_pred))

          Accuracy: 0.7446808510638298
          [[ 4  9]
           [ 3 31]]
                        precision    recall  f1-score   support

                     1       0.57      0.31      0.40        13
                     2       0.78      0.91      0.84        34

              accuracy                           0.74        47
             macro avg       0.67      0.61      0.62        47
          weighted avg       0.72      0.74      0.72        47
```

```
Out[107]: DecisionTreeClassifier()

In [108]: # Model Accuracy Score
          clf.score(x_test_b,y_test_b)
          results.update({"DT":clf.score(x_test_b,y_test_b)})

In [109]: y_pred=clf.predict(x_test_b)

In [110]: accuracy=metrics.accuracy_score(y_test_b, y_pred)
          print("Accuracy:",accuracy)
          print(confusion_matrix(y_test_b, y_pred))
          print(classification_report(y_test_b, y_pred))

          Accuracy: 0.7872340425531915
          [[ 6  7]
           [ 3 31]]
                        precision    recall  f1-score   support

                     1       0.67      0.46      0.55        13
                     2       0.82      0.91      0.86        34

              accuracy                           0.79        47
             macro avg       0.74      0.69      0.70        47
          weighted avg       0.77      0.79      0.77        47
```

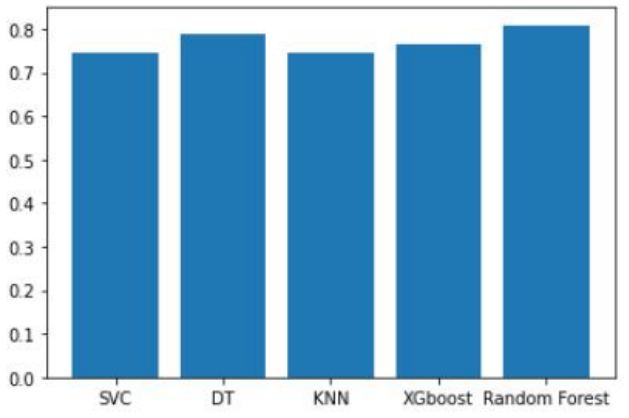
```
Out[113]: KNeighborsClassifier(n_neighbors=3)

In [114]: # Model Accuracy Score
          knn.score(x_test_b,y_test_b)
          results.update({"KNN":knn.score(x_test_b,y_test_b)})

In [115]: y_pred=knn.predict(x_test_b)

In [116]: accuracy=metrics.accuracy_score(y_test_b, y_pred)
          print("Accuracy:",accuracy)
          print(confusion_matrix(y_test_b, y_pred))
          print(classification_report(y_test_b, y_pred))

          Accuracy: 0.7446808510638298
          [[ 2 11]
           [ 1 33]]
                        precision    recall  f1-score   support

                     1       0.67      0.15      0.25        13
                     2       0.75      0.97      0.85        34

              accuracy                           0.74        47
             macro avg       0.71      0.56      0.55        47
          weighted avg       0.73      0.74      0.68        47
```

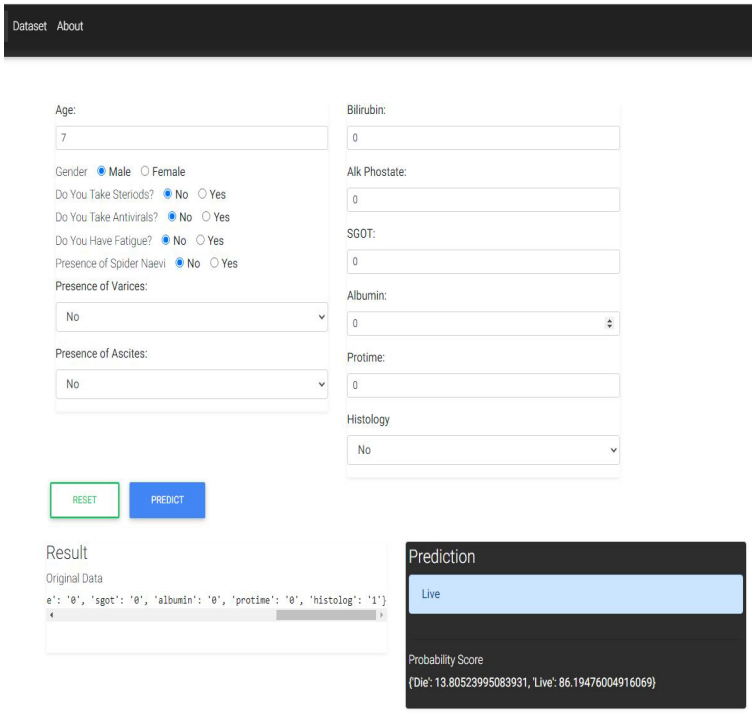
```
In [127]: # Save the plot
          graph.write_png("hep_decisition_tree_plot.png")

Out[127]: True

In [128]: from sklearn import metrics
          from sklearn.metrics import confusion_matrix
          from sklearn.metrics import classification_report

In [129]: from xgboost import XGBClassifier
          xgb = XGBClassifier()
          xgb.fit(x_train_b,y_train_b)
          y_pred = xgb.predict(x_test_b)
          accuracy=metrics.accuracy_score(y_test_b, y_pred)
          print("Accuracy:",accuracy)
          print(confusion_matrix(y_test_b, y_pred))
          print(classification_report(y_test_b, y_pred))
          results.update({"XGboost":accuracy})

C:\Users\Mahi\anaconda3\envs\project\lib\site-packages\xgboost
ssifier is deprecated and will be removed in a future release.
el_encoder=False when constructing XGBClassifier object; and 2
2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)

[12:52:39] WARNING: C:/Users/Administrator/workspace/xgboost-w
0, the default evaluation metric used with the objective 'bina
t eval_metric if you'd like to restore the old behavior.
Accuracy: 0.7659574468085106
[[ 6  7]
 [ 4 30]]
              precision    recall  f1-score   support

           1       0.60      0.46      0.52        13
           2       0.81      0.88      0.85        34

    accuracy                           0.77        47
   macro avg       0.71      0.67      0.68        47
weighted avg       0.75      0.77      0.76        47
```

```
In [130]: from sklearn.ensemble import RandomForestClassifier
          rclf = RandomForestClassifier(n_estimators=100,n_jobs=-1)
          rclf.fit(x_train_b,y_train_b)
          y_pred = rclf.predict(x_test_b)
          accuracy=metrics.accuracy_score(y_test_b, y_pred)
          print("Accuracy:",accuracy)
          print(confusion_matrix(y_test_b, y_pred))
          print(classification_report(y_test_b, y_pred))
          results.update({"Random Forest":accuracy})

Accuracy: 0.8085106382978723
[[ 4  9]
 [ 0 34]]
              precision    recall  f1-score   support

           1       1.00      0.31      0.47        13
           2       0.79      1.00      0.88        34

    accuracy                           0.81        47
   macro avg       0.90      0.65      0.68        47
weighted avg       0.85      0.81      0.77        47
```

```
In [131]: import matplotlib.pyplot as plt
```

```
plt.xticks(range(len(results)), list(results.keys
plt.show()
```





ANALYSIS & CONCLUSION

In this study, hepatitis was diagnosed using a variety of machine learning methods including neural networks.A comparison on the accuracy for a particular data set was performed by using various techniques, for identifying the best tool for Hepatitis disease diagnosis.

With this study, it is inferred that out of all models considered and its performance, Random Forest is most accurate that gives a good prediction accuracy

we have shown the importance of handling missing values and feature selection in order to progress classification model accuracy. Due to get the best classifier we have made a comparison among our classification models. By removing observations from the dataset having missing values as well as applying info gain feature selection technique with ranker search on our dataset, each of our classification algorithms gives a remarkable performance. The less contribution features present in the dataset as well as missing values may be the reason for poor classification accuracy.

In the future, larger data set will be considered to build the model which will produce more unique rules so giving higher accuracy. To enhance the accuracy more, different rule weighting techniques are suggested. It is possible to extend the research by using different classification techniques.

REFERENCES

[1] M. J. Nayeem, S. Rana, F. Alam and M. A. Rahman, "Prediction of Hepatitis Disease Using K-Nearest Neighbors, Naive 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

[8] M. Ramasamy, S. Selvaraj, and Dr. M. Mayilvaganan. "An empirical
analysis of decision tree algorithms: Modeling hepatitis data." IEEE
International Conference on Engineering and Technology (ICETECH), India, pp. 1-4, 20 March. 2015.

[9] A. H. Roslina, and A. Noraziah. "Prediction of hepatitis prognosis
using Support Vector Machines and Wrapper Method." IEEE Seventh
International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), China, Vol. 5, pp. 2209-2211, 10-12 August. 2010.

[10] G. Sathyadevi"Application of CART algorithm in hepatitis disease
diagnosis." IEEE International Conference on Recent Trends in Information Technology (ICRTIT), India, pp. 1283-1287, 3-5 June.
2011.

[11] K. S. Bhargav, T. D. Kumari, D. S. S. B. Thota, and V. B "Application of Machine Learning Classification Algorithms on Hepatitis Dataset." International Journal of Applied Engineering Research, vol. 13, no. 16, pp. 12732-12737, 2018.

[12] S. Hashem, G. Esmat, W. Elakel, S. Habashy, S. A. Raouf, M. Elhefnawi, M. El-Adawy, and M. Elhefnawi "Comparison of Machine Learning Approaches for Prediction of Advanced Liver Fibrosis in Chronic Hepatitis C Patients." IEEE/ACM transactions on computational biology and bioinformatics, vol. 15 no. 3, pp. 861-868, 2018.

[13] S. M. M. Hasan, M. A. Mamun, M. P. Uddin, and M. A. Hossain, "Comparative Analysis of Classification Approaches for Heart Disease Prediction," IEEE International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2), Rajshahi, Bangladesh, pp. 1-4, 8-9 February. 2018.

[14] V. Kumar. M, V. Sharathi. V, and G. D. BR. "Hepatitis prediction model based on data mining algorithm and optimal feature selection to improve predictive accuracy." International Journal of Computer Applications, vol. 51, no. 19, pp. 13-16, 2012.

[15] N. Nahar, and F. Ara. "Liver Disease Prediction by using Different Decision Tree Techniques." International Journal of Data Mining & Knowledge Management Process (IJDKP), vol. 8, no. 2, pp. 1-9, 2018.

[16] M. F. Faruque, Asaduzzaman, and I. H. Sarkar. “Performance analysis of Machine Learning Techniques to Predict diabetis Mellitus” IEEE International Conference on Electrical, Computer and Communication Engineering (ECCE), Cox’s Bazar, Bangladesh, pp. 1-4, 7-9 February. 2019.

[17] 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” IEEE International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kanpur, India, pp. 1-7, 6-8 July. 2019.

[18] V. K. Yarasuri, G. K. Indukuri and A. K. Nair. “Prediction of Hepatitis Disease Using Machine Learning Technique” IEEE International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), Palladam, India, pp. 1-5, 12-14 December. 2019.

[19] A. K. M. S. Rahman, F. M. J. M. Shamrat, Z, Tasnim, J. Roy and S. A. Hossain. “A Comparative Study On Liver Disease Prediction Using Supervised Machine Learning Algorithms.” International Journal of Scientific & Technology Research (IJSTR), vol. 8, no. 11, pp. 1-4, 2019.

[20] J. Cabrera, A. Dionisio and G. Solano. “Lung cancer classification tool using microarray data and support vector machines.” IEEE International Conference on Information, Intelligence, Systems and Applications (IISA), Corfu, Greece, pp. 1-6, 6-8 July. 2015.

[21] H. Yan, Y. Jiang, J. Zheng, C. Peng and Q. Li. “A multilayer perceptron-based medical decision support system for heart disease diagnosis.” Expert Systems with Applications, vol. 30, no. 2, pp. 272-281, 2006.

[22] UCI Machine Learning Repository. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/hepat

GitHubLink:
https://github.com/VarnaNemulla/ML-Final-Project