International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 07 | July 2020 www.irjet.net p-ISSN: 2395-0072

IMPLEMENTATION OF MACHINE LEARNING TECHNIQUES FOR PREDICTING NON-ALCOHOLIC FATTY LIVER DISEASE

Srikanth Viswanadhuni[1], Kalyan Shankar Ragam[2], Anjali Mathur[3]

[1]Srikanth Viswanadhuni: Student, Dept. of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Andhra Pradesh, India
[2]Kalyan Shankar Ragam: Student, Dept. of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Andhra Pradesh, India
[3]Anjali Mathur: Professor, Dept. of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Andhra Pradesh, India

---

Abstract - *Non alcoholic fatty liver disease (NAFLD) is one of the most common chronic liver diseases worldwide and has become a significant public health concern. Various Machine Learning techniques can be introduced to predict the clinical model of NAFLD optimally. Machine Learning techniques can be implemented using various open software tools. The machine learning techniques involved in this research are Support Vector Machine, K-Nearest Neighbor and Decision Trees. Based on a set of statistical testing techniques, uric acid, triglycerides, BMI, the serum alanine aminotransferase (ALT), and gamma-glutamyl transpeptidase (γGT) are the most prominent features contributing to NAFLD. The proposed researched work takes the dataset of patients suffering from NAFLD, their symptoms and other factors, and use them in calculating accuracy of the NAFLD perceptivity using various Machine Learning techniques.*

Key Words: **Support Vector Machine(SVM), Decision Tree Classifier, K-Nearest Neighbor(K-NN), Triglycerides, BMI, Gamma-glutamyl transpeptidase.**

1. INTRODUCTION

1.1 Non-alcoholic Fatty Liver Diseases

Non alcoholic Fatty liver diseases are the bunch of diseases affecting people who drink little to no alcohol. It occurs due to large amount of fat stored in the liver cells. It is observed commonly all around the world. Some people with NAFLD may develop non-alcoholic steatohepatitis (NASH), an aggressive form of fatty liver disease, marked my inflammation and may progress to cirrhosis and liver failure. There are not many symptoms which can be observed to find if a person is suffering from NAFLD apart from fatigue and pain in the upper right abdomen. But symptoms can be found with people suffering from NASH and cirrhosis like ascites, Enlarged blood cells beneath the skin, Enlarged spleen, Red palms and Jaundice.

Doctors across the world are still trying to find a reason why people are affected with NAFLD and why people accumulate fat in the liver even if they don't drink alcohol. There is little to no understanding on why patients develop inflammation which leads to cirrhosis. The main complication of NAFLD is cirrhosis which is the late-stage scarring in the liver. Liver tries to stop the inflammation in turn producing areas of fibrosis. This continues on and on, leading to fibrosis taking up more liver tissue which ultimately leads to ascites, esophageal varices, hepatic encephalopathy, liver cancer and Liver failure.

1.2 Machine Learning techniques

Machine learning techniques are mathematical expressions that represent data in the context of the problem helping to go from data to insight. Machine learning models predicts or makes decisions from the datasets by building a mathematical model. There are few machine learning algorithms which are used widely in our studies and deduction. They are Naive Bayes, Logistic Regression, K-Means, Linear Regression, Dimensionality Reduction Algorithms, Support Vector Machine, AdaBoost and Gradient Boosting.

Linear Regression helps in establishing a relationship by fitting independent and dependent variables in a single line. Logistic Regression is used to estimate discreet values from the set of independent variables. It helps in finding probability of an event by fitting data to a logic function. Decision tree is a supervised algorithm which works by classifying both categorical and continuous independent variables. A collection of decision trees is called Random Forests. SVM is used to plot data in a graph where raw data is depicted in points and lines called classifiers can be used to split data and plot them on a graph. Naïve Bayes assumes the presence of particular feature in class is unrelated to any other feature. KNN stores all available cases and classifies any new cases by taking majority vote of its k neighbors. In K-Means, datasets are classified into clusters in such a way that all data points are within a cluster. Gradient Boosting and Ad boost are used when massive loads of data have to be handled to make predictions with high accuracy

2. METHODOLOGY

We are using three different machine learning algorithms to classify the disease on the basis of symptoms. The fatty liver disease has some symptoms that are very much closer to diabetes and hepatitis. Most of the symptoms like fatigue, swollen eyes and headache are symptoms of these two diseases.

2.1 Decision Tree

A decision tree is a structure where each internal node is a test of a feature, branches represent conjunctions of features which lead the class labels and each leaf node is a class label. The path from leaf to nodes represent classification rules. Decision tree is a supervised learning algorithm that identifies ways to split a dataset based on different conditions which are mainly used for classification and regression tasks. It generally works on a top down approach. It can be both numerical and categorical data.

We are using all the symptoms are the featured set and the type of disease as the target set. Our target is to classify the disease in any one of the four classes. They are Hepatitis B, fatty liver, Jaundice and diabetes. After implementing decision tree, we got the following results

1. For the age over 40, the entropy value is 1.728 for hepatitis disease while it is 0.28 for fatty liver disease and 0.291 for diabetes.
2. The major focus was on Red palms and the age factor to find the entropy
3. The Red Palms are further divided into age factor and fatigue symptom and that further divides to age factor
4. A total of 1500 values are used for experiment in which 70% were set as training data and 30% is set as the testing data.
5. We bound the maximum depth limit of decision tree classifier by 4.

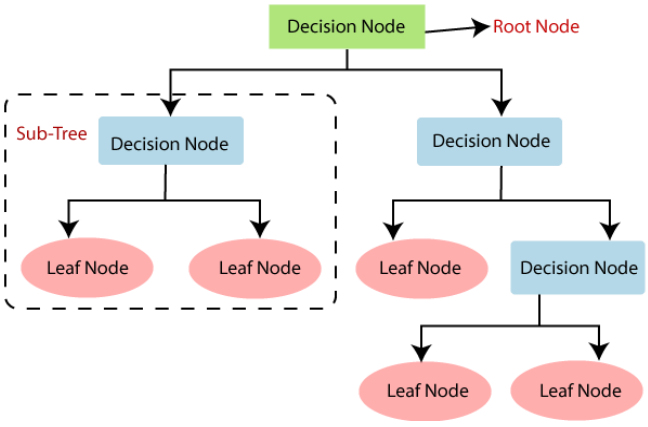

Figure -1: Architecture of Decision Tree Classifier

2.2 K-Nearest Neighbor

K-Nearest Neighbor is supervised machine learning algorithm that can solve both classification and regression problems. KNN uses the idea of proximity with operations which is used to calculate the distance between cluster of data points on a graph. We

run the algorithm several times to choose the right K value for our data which is ideal in making accurate predictions without much errors. KNN is versatile and can be used for regression, search and classification. KNN gets a bit slower as the volume of the data keeps increasing which is not practical in some environments where the predictions need to be made quickly rather than accurate.

In this algorithm, we consider K=4. After predicting the KNN values on testing and training dataset, we approximately got the accuracy of training dataset as 0.87 and testing dataset as 0.84

For K=11, the confusion matrix shows the mean accuracy in the range of 0.72 to 0.87.

The best accuracy was 0.8789 with K=6

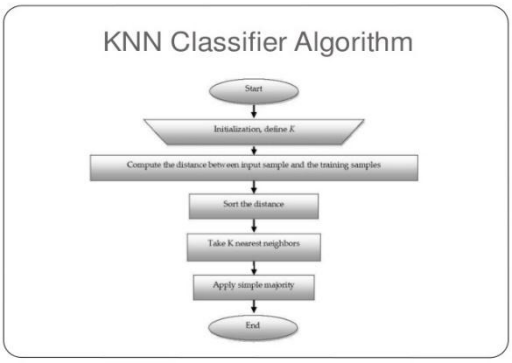

Figure -2: Architecture of K-Nearest Neighbor

2.3 Support Vector Machine

In Support Vector Machine, there is a hyperplane in a N- dimensional space that divides the data points. Hyperplanes are boundaries that divide the data points. Support Vectors are the data points that influence the position and orientation of the hyperplane. In python, SVM is used by taking any two features and plotting them to visualize. We extract the required features and split them into training and testing data. Most of it used in training data while a few used in testing data. SVM algorithm can also be implemented from scikit learn library and call the related functions.

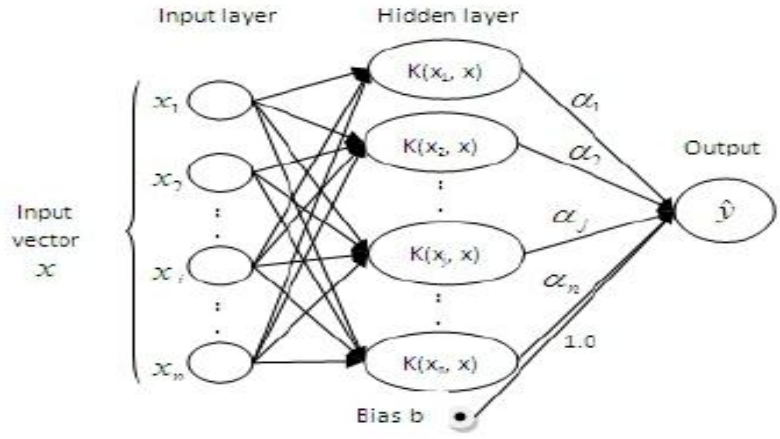

Figure -3: Architecture of Support Vector Machine

3. RESULTS AND ANALYSIS

On the training and testing of our dataset with the above mentioned machine learning techniques the accuracy of decision tree classifier ,K-Nearest Neighbor , Support Vector Machine are like 0.84288,0.87898,0.87704 respectively.

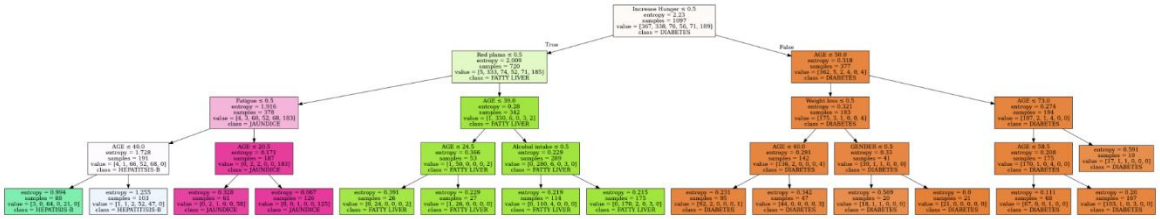

Figure -4: shows the visualization decision tree classifier

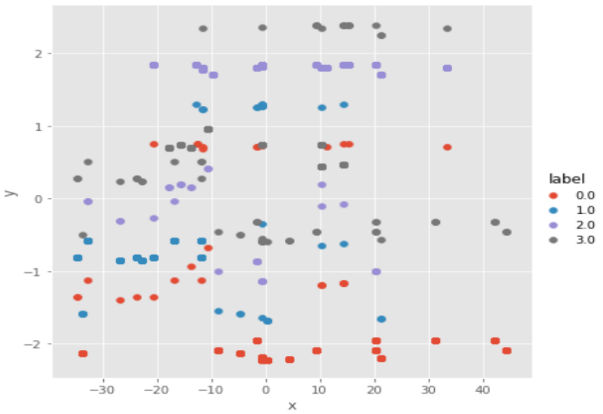

Figure -5: shows the visualization of K-nearest neighbor

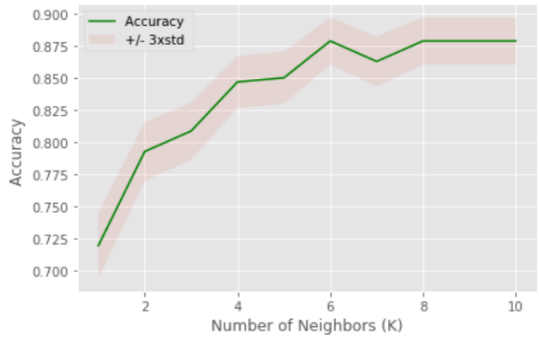

Figure -6: shows the accuracy graph of the best model among the three techniques

Table -1: The accuracy of our machine learning models

S.NO	MODEL	ACCURACY
1	Decision tree classifier	0.842
2	K-Nearest Neighbor	0.878
3	Support Vector Machine	0.877

After implementing the three algorithms we found that decision tree classifier yield less accuracy compared to Support vector machine followed by K-nearest neighbor algorithm has highest accuracy of 0.878.

4. CONCLUSION

The accuracy for decision tree classifier, K-Nearest Neighbor and Support Vector Machine are 0.84, 0.878 and 0.877 respectively. From the given accuracy we conclude that K-nearest neighbor has more accuracy with our dataset when compared with other two machine learning models. Hence we conclude that from our work on non-alcoholic fatty liver disease has completed which is used to reduce the time and cost for doctors to diagnosis the disease, besides which will also save the patients life with early diagnosis .The limitation is that there are no works containing these machine learning techniques to compare them for their accuracy.

5. REFERENCES

[1] Aleksandra Mojsilovi´c , Miodrag Popovi´ c, Member, IEEE, Srdjan Markovi´c, Miodrag Krsti´ c, Characterization of Visually Similar Diffuse Diseases from B-Scan Liver Images Using Non-separable Wavelet Transform, IEEE TRANSACTIONS ON MEDICAL IMAGING, VOL. 17, Aug 1998.

[2] Andreas Mueller, Scikit-Learn Tutorial: Statistical-Learning for Scientific Data Processing,(NYU Center for Data Science)

[3] Binish Khan, Piyush Kumar Shukla, Manish Kumar Ahirwar, Strategic Analysis in Prediction of Liver Disease Using Different Classification Algorithms, IJCSE vol-7 issue-7

[4] Han Ma, Cheng-fu Xu, Zhe Shen, Chao-hui Yu, You-ming Li, Application of Machine Learning Techniques for Clinical Predictive Modelling: A Cross-Sectional Study on Non-alcoholic Fatty Liver Disease in China, Hindawi, BioMed Research International, Volume 2018

[5] Johan M. Thijssen and his team, Computer-aided B-mode ultrasound diagnosis of hepatic steatosis: a feasibility study, IEEE transactions on ultrasonics, ferroelectrics, and frequency control, vol. 55

[6] Nazmun Nahar1 and Ferdous Ara2, LIVER DISEASE PREDICTION BY USING DIFFERENT DECISION TREE TECHNIQUES, International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.8, No.2, March 2018

[7] Ricardo and his team, An Ultrasound-Based Computer-Aided Diagnosis Tool for Steatosis Detection, IEEE Journal of Biomedical and Health Informatics (Volume: 18 , Issue: 4 , July 2014)

[8] S. Pavlopoulos', E. Kyriatou', D. Koutsouris', K. Blekasl, Ai Stafylopotis2, P. Zoumpoulis3, FUZZY Neural Network-Based Test Analysis of Ultrasonic Images, IEEE Engineering in Medicine and biology vol:19.

[9] Samuel burns, Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn and Tensorflow (Step-By-Step Tutorial for Beginners), Published Independently.

[10] Shai Shalev-Shwartz and Shai Ben-David, Understanding Machine Learning from Theory to Algorithms. Cambridge University Press; 1 edition (19 May 2014)

[11] Stephen Marsland, Machine learning an algorithmic perspective, Chapman and Hall/CRC, 2 edition(17 November 2014)

[12] Willi Richert and Luis Pedro Coelho, Building machine learning systems with python, Packt Publishing Limited(26 July 2013)