International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 09 | Sep 2020 www.irjet.net p-ISSN: 2395-0072

Automated Prediction of Non Alcoholic Fatty Liver Disease using Machine Learning Algorithms

Adeep Kulkarni[1], Suprit Shinde[2], Professor Dipali Kadam[3]

[1]*Student, Dept. of Comp Engineering, P.I.C.T, Pune, Maharashtra, India*
[2]*Student, Dept. of Comp Engineering, P.I.C.T, Pune, Maharashtra, India*
[3]*Asst Professor, Dept. of Comp Engineering, P.I.C.T, Pune, Maharashtra, India*

Abstract - *The prevalence of obesity has lead the metabolic syndrome Non-Alcoholic Fatty Liver Disease (NAFLD) to be a serious health concern over the years. Early prediction on liver disease using classification algorithm is an efficacious task that can help the doctors to diagnose the disease within a short period of time. The main objective of this project is to identify the potential factors causing NAFLD by using Machine Learning algorithms like Naïve Bayes, Decision Tree classifier, SVM, Random Forest classifier, Logistic regression and K-means algorithm.*

Key Words: **Machine Learning, Data Science, Data Analysis, Electronic Health Records, Prediction, NAFLD**

1. INTRODUCTION

Non Alcoholic Fatty Liver Disease or NAFLD is a common and rising entity which leads to various liver disorders. Its prevalence is increasing at a rapid rate due to the increasing levels of obesity, diabetic patients and hypertension. These disorders are where the liver damage is unrelated to alcohol consumption. There are various methods to detect this early in order for quick treatment and recovery as in some cases this could be fatal. Hence, using Machine Learning classification algorithms, which use various biological features of the patient, useful in determining and predicting accurately the instance of NAFLD, is an efficient way to deal with these issues and in this project, we have used 6 features to determine the instance of this disease in a patient. Following risk factors were of NAFLD were observed like metabolic syndrome, Body Mass Index (BMI), Sex, Lipoprotein cholesterol (High density and Low density), Total and direct Bilirubin, onset of Fibrosis, Age etc. The performance of several methods mentioned above will be compared in this project. As NAFLD prediction is imperative for the prevention of its complications, we propose to evaluate whether a combination of blood-based bio markers and anthropocentric parameters can be used to predict NAFLD among overweight and obese adults.

2. METHODS

We collected the data of around 600 patients from a private hospital in Pune, India suffering from liver ailments from the years 2018-2020 from the Electronic Health Records (EHR) maintained by the hospital which gave us information about their individual health-based factors such as their sex, height, weight, and hence BMI, and other main factors such as lipoprotein, triglycerides, AST, ALT, total and direct bilirubin, HDL, LDL, instance of Diabetes, severity of Cystic Fibrosis in the patient, Thrombosis, WHR, Albumin and a few others. The health care data collected was completely local and was collected from a major hospital which had a history of multiple patients of various liver disorders, both alcoholic and nonalcoholic, which we had to manually segregate and reduce to obtain a clearer data-set to analyze the instance of specifically non-alcoholic fatty liver disorders.

The initial data gathered from the hospital had records of 605 patients, attaining to 27 different attributes which lead to the task of cleaning this data and performing subsequent data normalization and standardization techniques. The clear goal of this task was to improve the accuracy of classification of this data. Which we followed by splitting the data set into individual training and testing sets. We used various performance metrics -

1. Accuracy - Which simply refers to the correctness of predictions, both positive and negative, in the testing set.

2. Precision- Which refers to the number of positive predictions which were actually positive.

3. Recall - Refers to all the positive predictions out of all actually positive instances.

4. ROC curve- AUC-ROC curve, or Receiver Operating Characteristic curve is a measure which simply tells us how much the model is capable of actually differentiating between the different classes. Higher the AUC, or Area under Characteristic, the higher the chance that the model can actually differentiate the decision of disease or no disease in an instance. The curve is on the scale of True positive rate and False positive rate.

5. Confusion matrix - Confusion matrix is a sort of table layout which allows visualization of the performance of the algorithms.

2.1 FEATURE SELECTION AND DATA CLEANING TECHNIQUES

Since we had a vast data of 605 patients with 27 features, it was necessary for us to check the probable instance of null values, outliers, and other raw data in the data set. The next step, naturally, was data pre-processing which included treatment of missing values, and after checking the instance of null values, we found out that 154 entries had incomplete or null values, in some or another feature, or in all features and were subsequently removed from the data set, to get a higher accuracy score for prediction. The data was now reduced to 451 entries with the same 27 features. With the removal of these values, we needed to then standardize and normalize the data into a specific range (usually 0-1) to get faster classification. We used the python library Pandas, to assist us with the data cleaning techniques.

Feature engineering was the next step in managing our collected data, using domain knowledge, in this case, healthcare or knowledge about liver diseases, is to be taken, and used to extract features using raw data. Consulting health professionals is another way to reduce difficulties while feature selection. Hypothesis generation was also used, which involved removal of factors such as Cirrhosis, alcohol intake - which was obscure because the main purpose was to classify specifically Non-Alcoholic Fatty Liver Disease (NAFLD) in individuals.

For every algorithm we used, our sole aim was to increase the ROC and accuracy and for that purpose, we performed hyperparameter tuning. Hyperparameters are the parameters which include the model's properties such as the algorithm complexity, learning rate of an algorithm among many others.

Next, we made sure of these following things- measuring the right things in the data, using proper visualizations of data, and bringing out the meaning and context of the dataset, this step is called insight finding.

Another useful way we used was data visualization, for example, plotting the correlation matrix, of which we have a figure. The main features which had a high correlation to cystic fibrosis were age, weight, diabetes mellitus, and waist circumference which showed that NAFLD usually happens to individuals when these features are in a speculated risk factor. We also plotted scatterplots and histograms, which gave us the opportunity to have visual look at the dataset, before model building. We used, the Python library Seaborn, for assisting us with the visualization of our data.

We have also used 6-7 supervised machine learning models in this, SVM, Logistic regression, Random forest classifier, Decision tree classifier, Gradient boosting and Neural networks.

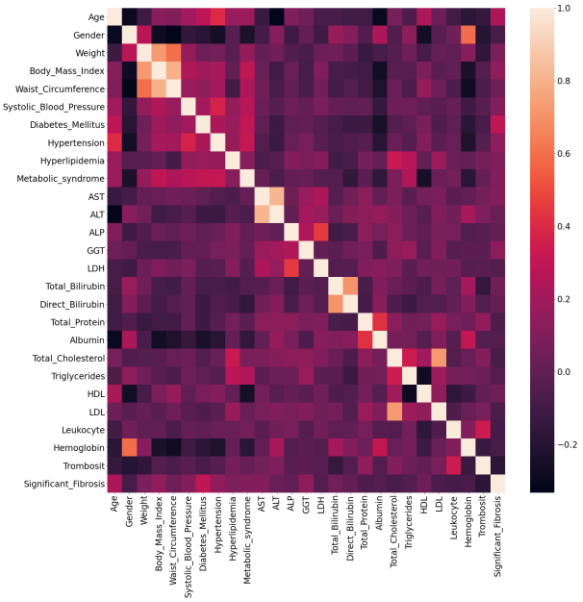

Chart -1: Correlation heat map

3. RESULTS

3.1 Analyzing the performance of various classifiers in the project

As discussed earlier, we used 6 classifiers in this project. Analyzing the results of every classifier, we determined the classifier with the highest accuracy to predict NAFLD.

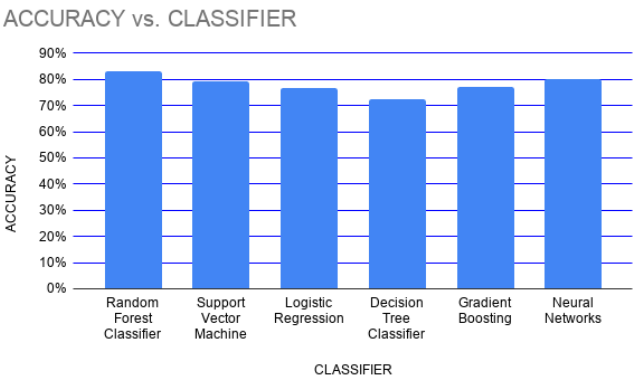

Fig -1: Accuracy comparison of all classifiers

1. Logistic Regression

Logistic Regression is a linear classifier. Hence, it misses out on the non-linearity required in the task of classification. It therefore has the lowest AUC-ROC rate at 73.81% and AUC of 60.71%.

The accuracy of the Logistic Regression model falls a bit short and is 77%.

Since its lack of linearity, in the AUC-ROC curve for Logistic Regression, the Area under the Curve is the smallest, meaning its ability to differentiate the classes and hence give an accurate prediction is relatively low.

However, the feature importance of this data in the ROC curve showed that Diabetes Miletus, followed by patient's Age and BMI are the most important features for predicting NAFLD using this model.

2. Decision Tree Classifier

As we know, a Decision Tree is the easiest interpreter in which the internal nodes of the trees are labeled as features. It is a non-parametric algorithm, which predicts the value of the target based on rules and then forms a tree, where the leaf nodes carry the predicted class.

The AUC-ROC score for Decision Tree was low, at 64.46%. The accuracy of this model lied at a mere 58.24%.

Again, Diabetes Miletus is the most significant feature in this model to predict NAFLD. Followed by Thrombosis, Age and number of Leukocytes.

3. Random Forest Classifier

Random Forest classifier is an improvement to the Decision Tree classifier. It works by creating a multitude of decision trees and outputs the class that is the mode or the mean prediction of the individual Decision Trees.

Hence, it gives the best possible output, and we can expect the highest accuracy in this. It's AUC-ROC score is an excellent 1.0 (100%).

It showed the second best accuracy of all the classifiers we have used at 85%

Thrombosis is the most significant feature in Random Forest classifier, followed by Age, Diabetes Miletus, AST and BMI.

4. Support Vector Machine classifier

Supervised learning models that analyze data for classification were also used in this project.

This classifier had the second highest AUC-ROC at 96.15%.

The accuracy was the best at 85.7%.

3.2 Results of the testing set

In the training data set, it was found out that 386 of the patient entries did not have significant fibrosis while 219 did. Meaning 386 of the original 605 patient entries were suffering from Non Alcoholic Fatty Liver Disease.

Analyzing the data further, we realize that the average age of the cohort is 46.30 and 323 patients are above the average age of 46. The data has 321 females and 284 males. The average weight of the patients is a whopping 86.40 meaning patient with larger weights are more at risk to the disease, which is also shown by the average BMI of 31.87 - which is considered obese.

Lastly, the highly significant factor Diabetes Miletus is studied, where we study that 225 of the entries have Diabetes Miletus. Since 219 of the entries, have fibrosis, it shows that Diabetes is a major factor in determining the instance of NAFLD in patients.

4. CONCLUSION

The findings of this project show that machine learning classification models especially the random forest model accurately predicts a non-alcoholic fatty liver disease patient. This project also helped us to find out some undiscovered factors majorly causing NAFLD. This method may lead to greater insights for doctors to effectively identify NAFLD for novel diagnosis, and for preventive and therapeutic purposes to mitigate the global burden of NAFLD.

REFERENCES

[1] Yu-Han Cheng, Cheng-Ying Chou, “Application of Machine Learning methods to predict NAFLD in Taiwanese high-tech industry workers”, International Conference of Data Mining, July 2017.

[2] Binish Khan, Piyush Kumar Shukla, Manish Kumar Ahriwar, “Strategic analysis in prediction of liver disease using different classification algorithms”, International Journal of Computer Science and Engineering, July 2019.

[3] Suruchi Fialoke, Anders Malarstig, Melissa R. Miller,Alexandra Duimitru, “Applications of Machine Learning methods to predict Non-Alcoholic Steatohepatitis (NASH) in NAFL patients”, December 2018.

[4] Tilman Kuhn, Tobias Nonenemacher, Johanna Nattenmuller, “Anthropometric and blood parameters for the prediction of NAFLD among overweight and obese adults”, July 2018.

[5] Younossi Zobair M., Koenig Aaron B., Abdelatif Dinan, Fazel Yousef, Henry Linda, Wymer Mark, “Global epidemiology of nonalcoholic fatty liver disease meta-analytic assessment of prevalence, incidence and outcomes” , Hepatology vol.2 , 1389-9,1389-90,19-22 February 2016.

[6] K. Sung, M.Y. Lee, Y. Kim et al, “Obesity and incidence of diabetes: except of absence of metabolic syndrome, insulin resistance, inflammation and fatty liver”, atherosclerosis, Vol. 275, PP. 50-57, 2018

[7] R. Kumar and S. Mohan "Non-Alcoholic Fatty Liver Disease In Lean Subjects: Characteristics and Implications" Journal of Clinical and Translational Hepatology, Vol.5, No.3, PP.216-223, 2017.

[8] Armstrong MJ, Houlihan DD, Bentham L, et al. "Presence and severity of nonalcoholic fatty liver disease in a large prospective primary care" cohort. J Hepatol 2012;56:234–40.

[9] Ekstedt M, Franzen LE, Mathiesen UL, et al. "Long-term follow-up of patients with NAFLD and elevated liver enzymes" . Hepatology 2006; 44:865–73.

[10] Adams LA, Sanderson S, Lindor KD, et al. "The histological course of nonalcoholic fatty liver disease: a longitudinal study of 103 patients with sequential liver biopsies" . J Hepatol 2005;42:132–8.

[11] Dixon JB, Bhathal PS, O'Brien PE. "Nonalcoholic fatty liver disease: predictors of nonalcoholic steatohepatitis and liver fibrosis in the severely obese". Gastroenterology 2001;121:91–100.

[12] McPherson S, Stewart SF, Henderson E, et al. "Simple non-invasive fibrosis scoring systems can reliably exclude advanced fibrosis in patients with non-alcoholic fatty liver disease" . Gut 2010;59:1265–9.

[13] Targher G, Bertolini L, Padovani R, et al. "Prevalence of nonalcoholic fatty liver disease and its association with cardiovascular disease among type 2 diabetic patients" . Diabetes Care 2007;30:1212–18.