EasyChair Preprint

№ 5451

Classification of Liver Diseases Using Intelligent Techniques

Shreyansh Jain, Ritik Sharma and R. Rajkamal

May 4, 2021

Classification of Liver Diseases Using Intelligent Techniques

Shreyansh Jain
"SRM Institute of Science and Technology Chennai, Tamil Nadu"
ss1774@srmist.edu.in

Ritik Sharma
"SRM Institute of Science and Technology Chennai, Tamil Nadu"
rr4513@srmist.edu.in

Dr. R. Rajkamal
"SRM Institute of Science and Technology Chennai, Tamil Nadu"

Abstract

Liver disease is one of the leading causes of death in several countries. Patients with liver disease have been steadily increasing as a result of alcohol abuse, inhaling harmful gases, eating contaminated food, heavy drinking of local beverage and drugs. Prolonged drinking habits are directly linked to an increased risk of developing various liver diseases that can lead to preventable deaths if diagnosed early. "Identifying a liver patient in the early stages of the disease (i.e., even minor liver damage) is difficult. Early detection of liver problems will increase a patient's survival rate." [4] In the database, we have about 600 patient details thus creating the learning filters for Good or Bad machine. "Performance is measured in terms of accuracy, precision and recall measures. The results of the various dividers are obtained using the proposed algorithm" [11] for selecting the proposed feature. From the analysis and comparative analysis, we will be able to increase the accuracy of the categories and lead to a reduction in the Classification time and ultimately help to predict the disease accurately or accurately.

Introduction

"Exact identification or classification of diseases is always being a very important but a very complex task for doctors. Many diagnostic tests have been developed over a long period time to perform human disease identification. In many cases a single test or even several tests cannot exactly identify the correct disease. Identification of correct disease is the first and the foremost task for correct treatment and for that we need specialist doctors with accurate diagnostic reports. However, expert specialist doctors are not always available in many parts of the world and more specifically in third world countries. For liver disease identification, analysis is performed with the help of several types of enzymes in the patient blood. In this regard, medical field is now gradually incorporating computational techniques in biology to automate various tasks including the disease classification problem." [4]

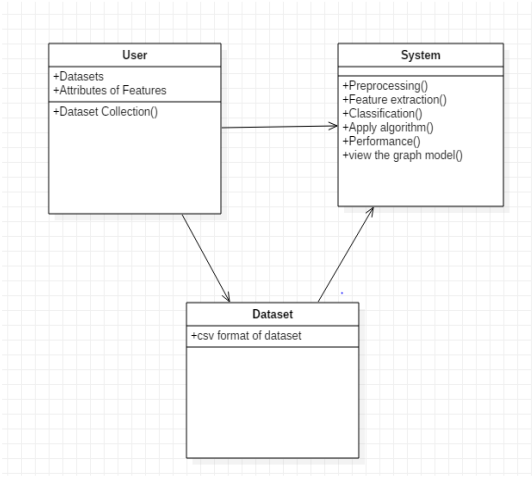
In this paper, we propose checking of Liver disease patients using intelligent techniques by classifying them according to the percentage of patients getting the disease as a positive or else negative information measure.

"The World Health Organization reported in 2002 (World Health Report, 2002) that alcohol use was responsible for 4% of the total global disease burden compared with 4.1% (tobacco) and 4.4% (hypertension). In developed countries" [10] there was also an obligation of 9.2% of all years of life lost to disability. These include conditions related to neuropsychiatric, suicide, homicide, and physical injuries (road accidents, falls, burns, drownings, etc.). Machine learning or Artificial intelligence in a whole is a powerful tool in the automated diagnosis of various harmful diseases. "Data mining techniques have widely been used for the prediction of various diseases and use of classification algorithms like Decision tree

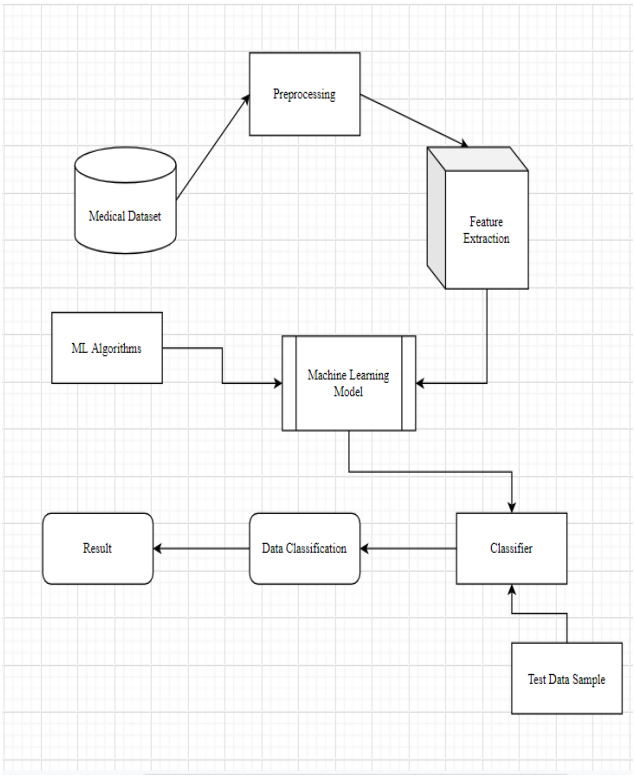
algorithm, Bayes Algorithm and Rule based Algorithm for diabetes disease prediction and they are considered popular classification algorithms at the time" [3]

In this paper, we aim to find Liver disease patients using intelligent techniques by classifying them according to the percentage of patients getting the disease as a positive or else negative information measure. Thus giving a representation of accuracy, precision and recall measure for each algorithm.

Diagrams



Class Diagram



Architecture Diagram

Modules

1. Dataset (Medical)

The Indian Liver Patient Dataset has a variety of symptoms for around 600 patients. Patients were defined as 1 or 2 on the basis of whether they are having liver disease or not. The gender is calculated on the basis of 0's and 1's meaning conversion of Male tags and Female tags as 0 and 1 respectively. We also have an age range providing us insights on the Age Gap and Audience insights.

2. Pre-processing:

Pre-processing of Data or particularly datasets is a major and rudimentary step in generation of ML models. Most of the data used in these machine learning problems requires us to perform multiple processing / refinement / modification in order for the our Algorithms to be trained in and perfected. Widely used processing techniques are generally very small such as inputation or removing of missing

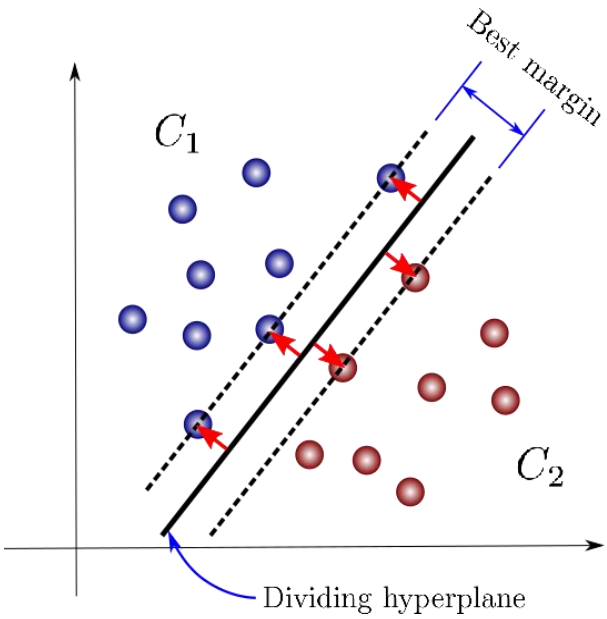
values, variable code coding, measurement, Splitting and feature scaling.

3. Classification Technique:

SVM: -

Support Vector Machine "aims to find an optimal hyper plane that separates the data into different classes. The scikit package in python is used for the SVM model. The pre-processed data is split into test data and training set which is of 25% and 75% of the total dataset respectively" [10]

A SVM or simply support vector machine creates a simple hyper-plane or a set of hyper planes in an infinite or higher domain position. The best distinction is obtained by the hyper plane with its highest distance in the training-data acquisition area of category which is known as margin, because the limit is usually greater when it reduces the standard distortion error.

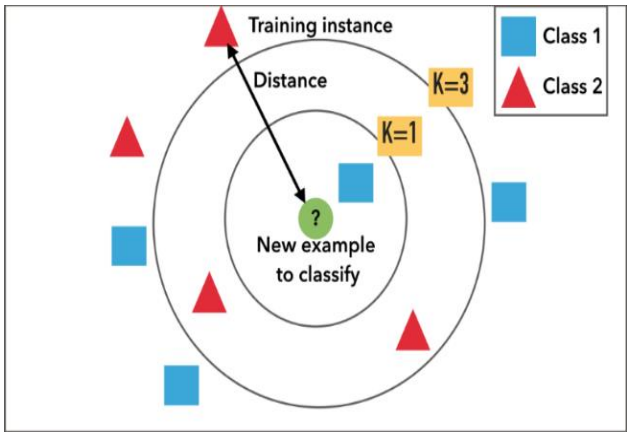


Support Vector Machine

K-NN: -

The KNN model is as the name suggest is trying to find the nearest datasets or values to each other to try to club them together. When forecasting is required for an invisible data event, the KNN algorithm will search for training databases of very similar conditions.

The predictive prediction of very similar conditions is summarized and restored as an unpredictable prediction.

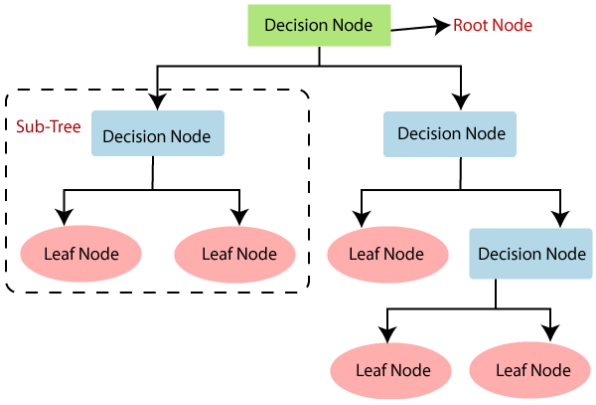


K-Nearest Neighbor

Decision Tree: -

Decision tree is considered to be one of the most loved and used prediction/classification type algorithm.

The solution tree is what you call as a tree-like shaped diagram of nodes in it each and every internal node defines a test in a particular attribute and each of it's branches represents a test result, and each leaf node or simply terminal node holds a section label.



Requirements Specification

Since the involvement of the dataset preprocessing and other manipulation Python

3.0 or higher with IDE like Jupyter Notebook for ease or simply Anaconda Navigator is required. Also, basic things like Python's standard library is needed So it can be installed using PIP install.

For the Hardware requirements, it basically involve Windows 7,8,10 64 bit/32 bit or any other Operating System with an RAM of 4GB and above

Existing Model

Logistic Regression, Support Vector Machine, k-nearest neighbors and one ensemble method. Different combinations of hyperparameters for individual algorithms, like kernel, degree and for SVM and weights, n-neighbors and algorithms for k-Nearest Neighbors will be tried across the training sets.

Feature selection used in here as well as the classifiers (Bayes theorem, Parson, SVM-Support vector machine, and K-Nearest Neighbour-KNN) are optimally selected for each stage.

"A meta-analysis study stated that there are promising results for the detection of cirrhosis but it simultaneously reveals that there is this great variability within the detection of severe fibrosis and other liver diseases.

These methods, however, are not the most reliable alternatives to biopsy, as they present several limitations and do not meet the criteria in medical literature" [1]

Experimental Results

"In this section, results are analysed which are given by different classification algorithms involving K-nearest neighbour (KNN), Decision Tree, Support Vector Machine (SVM). In the experiment, the dataset is divided into training set and testing set. The ratio of the training set to testing set is 80% and 20% respectively. In this work, 10- fold cross-validation is used to train and test the machine learning model. The experiment is conducted in Python programming language and the library used" [5] are pandas, NumPy, Sci-kit learn,matplotlib and XG Boost.

A. Performance Measure

We will measure the performance measure of the model by using confusion matrix

- Confusion Matrix is the tabular "representation of actual or predicted values. Accuracy is calculated by (true positive (TP) + true negative (TN))/ (true positive (TP) + true negative(TN) + false positive (FP) + false negative (FN))" [5]
- Precision tells us how many of the correctly predicted cases actually turned out to be positive and it is calculated as Precision = TP true positive /(TP true positive +FP false positive)
- Recall tells us how many of the actual positive cases we were able to predict correctly with our model and its calculation is as follows Recall = TP meaning true positive / (TP true positive + FN meaning false negative)

B. Results.

All the above-mentioned classification algorithms are been tested as follows.

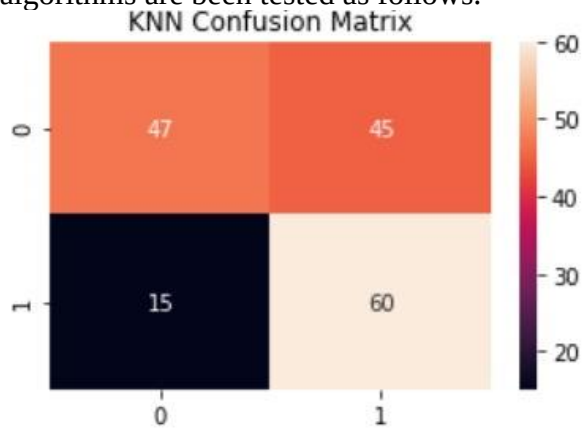


Fig. 1: K-nearest Neighbour Matrix

From this above confusion matrix of K-nearest neighbour (KNN) model,we can

clearly see the accuracy to be 64% and the Precision value of it will be approx. of 0.57 But Recall values giving us a 0.80 upon calculation.

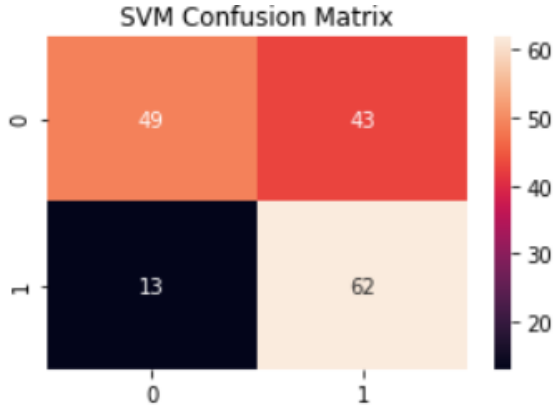


Fig. 2: Support Vector Machine SVM Matrix.

Secondly, from this above oversampled confusion matrix of Support Vector Machine(SVM) model, we found that accuracy is a above average 66% whereas Precision value showing value of 0.59 and Recall values coming at 0.83.

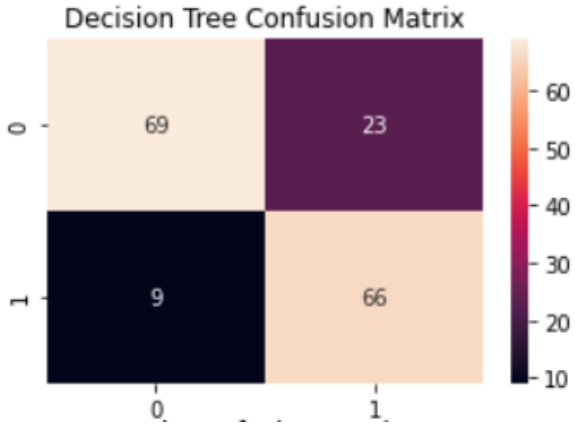


Fig. 3: Decision Tree Matrix.

Thirdly, From this above oversampled confusion matrix of Decision Tree model, we can see the accuracy to be coming at an approximation of 81% and the Precision value giving us a whopping 0.74 and Recall values are equal to 0.88.

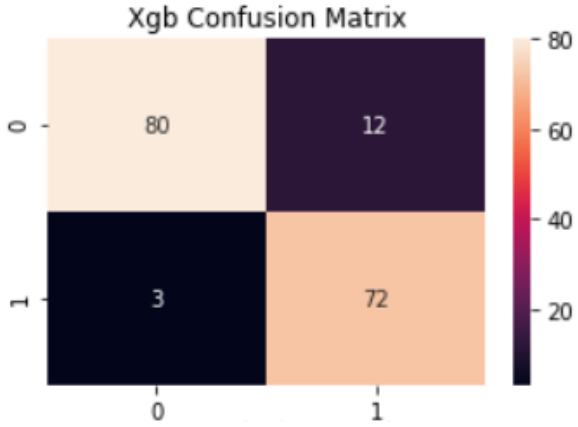


Fig.4: XG Boost Matrix.

And at last we will see for xgb, from the above oversampled confusion matrix of XG Boost model we can clearly see our accuracy is at an all-time high and equating upwards of 91% While simultaneously the value of Precision is equal to 0.86 and Recall values are through the roof and making it to over 0.96.

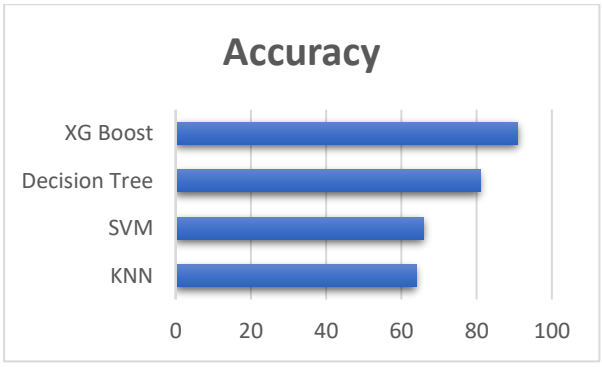


Fig. 5: Comparison in accuracy of all the algorithm in percentage (%)

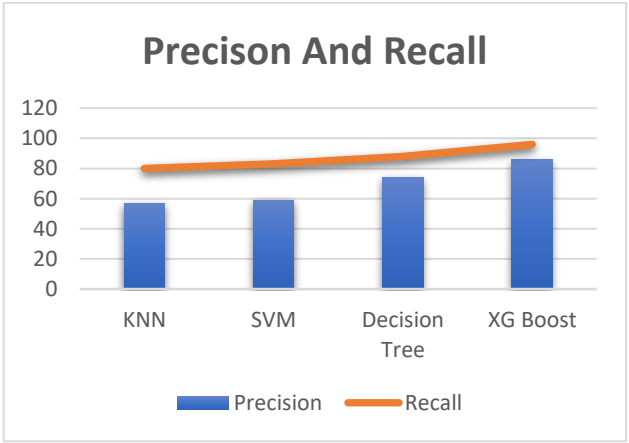


Fig. 6: Comparison according to Precision and Recall of our Machine learning models.

According to our oversampling results, we can say that Support Vector Machine(SVM) and K-Nearest Neighbour(KNN) has shown little to no difference in accuracy as well as precision and recall values. According to the medical-science terminology, by simply testing Precision we will be able to unlock the ability of the testing to correctly identify those with the disease thereby giving us accurate results . XG Boost and decision tree are the ones which are giving high accuracy in predicting liver disease.

Conclusion

The proposed model will therefore provide the right level of accuracy compared to other previous works. Analyze the performance of the proposed task during the execution, similarity calculations etc.

In this paper, the prognosis for liver disease is studied and appropriate analysis is done. Machine Learning Datasets are cleared by introduction of various pre-processing techniques such as median missing values, coding labels to convert a category into numerical easy to read data for its analysis, dual termination and retail locations are removed using the Isolate forest to improve performance. XGBoost is used to download the best qualities needed in predicting liver disease. “Various differentiated algorithms are used to predict the presence or absence of liver disease. Performance metrics such as accuracy, precision and recall are effectively used to analyze the performance of various classification algorithms” [3]

References

[1] Ricardo T. Ribeiro, Rui Tato Marinho, and J. Miguel Sanches, Senior Member, IEEE, “Classification and Staging of Chronic Liver Disease from Multimodal Data” in IEEE Transactions on Biomedical Engineering (Volume: 60, Issue: 5, May 2013)

[2] Li Dan dan, IEEE member, Miao Huanhuan, Li Xiang, Jiang Yu, Jin Jing, Shen Yi, ”Classification of diffuse liver diseases based on ultrasound images with multimodal features” 2019 IEEE International Instrumentation and Measurement Technology (I2MTC) (09 September 2019)

[3] Maria Alex Kuzhippallil Carolyn Joseph, Kannan A, “Comparative Analysis of Machine Learning Techniques for Indian Liver Disease Patients” in 2020 6th IEEE on Advanced Computing and Communication Systems (ICACCS)(23 April 2020)

[4] Syed Hasan Adil, Mansoor Ebrahim, Kamran Raza, Syed Saad Azhar Ali, Manzoor Ahmed Hashmani, “Liver Patient Classification using Logistic Regression” 2018 IEEE Computer and Information Sciences (ICCOINS) 29 October 2018

[5] Thirunavukkarasu, k. Singh, A. S. Irfan, M., & Chowdhury, A., “Prediction of Liver Disease using Classification Algorithms” 2018 IEEE Computing Communication and Automation (ICCCA) (29 July 2019)

[6] Sontakke, S., Lohokare, J., & Dani, R, “Diagnosis of liver diseases using machine learning proposes Back propagation and Micro Array Analysis” 2017 IEEE Emerging Trends & Innovation in ICT (Feb 3, 2017)

[7] Jagdeep Singh, Sachin Bagga, Ranjodh Kaur, “Software-based Prediction of Liver Disease with Feature Selection and Classification Techniques” Journal of Computational Intelligence and Data Science (June 2018)

[8] Damodar Reddy Edla, Kunal Mangalorekar, Gauri Dhavalikar, Shubham Dodia,”Classification of EEG data for human mental state

analysis using Random Forest Classifier" International Conference on Computational Intelligence and Data Science (ICCIDS 2018)
[9] www.irjet.net
[10] academic.oup.com
[11] ieeexplore.ieee.org