Liver Diseases Prediction using KNN with Hyper Parameter Tuning Techniques

Sateesh Ambesange
Pragyan AI school
Bengaluru, India
sateesh.ambesange@gmail.com

Ranjana Nadagoudar
Department of CSE
Visvesvaraya Technological Univesity
Belegavi, India
sonalram303@gmail.com

Rashmi Uppin
Pragyan AI school
Bengaluru, India
rashmi.ambesange@gmail.com

Vilaskumar Patil
Member Student, IEEE
Department of E&CE
Sharnbasva University
Kalaburagi, India
shrutimnpatil@gmail.com

Shruti Patil
Department of CSE
Sharnbasva University
Kalaburagi,India
mpvilaskumar@gmail.com

Sushma Patil
Department of EEE
Sharnbasva University
Kalaburagi, India
sushgecw@gmail.com

Abstract: **The healthcare industry is producing massive volumes of data. The algorithms using ML can be used to discover hidden patterns for making diagnosis and critical decisions. In the past years, Liver disorders have increased persistently and it is the reason for a significant number of deaths in many countries like India. The number of patients with liver disease are rapidly increasing due to several reasons like over consumption of alcohol, breathing in injurious gases, eating unhygienic food, pickles and drugs. The aim of our work is to predict liver disease by Machine learning based prediction model trained with the dataset fetched from the northeast of Andhra Pradesh, India. Feature analysis is performed on data and checked for balanced/imbalanced, distribution of data and correlation between the features and between features and liver disease. Transformation techniques have been used to transform data into normal distribution. So before performing transformation, outliers are detected and removed using outlier removal and the best features are selected based on correlation matrix and feature selection approaches, which will transform the data set effectively. Grid Search techniques are used in Hyper parameter tuning. Performance is evaluated using various metrics such precision, recall, f1-score, precision, precision-recall curve (PRC) and Receiver operating curve (ROC). Eight models are proposed out of which the fine tuning parameters of K-NN using Grid search gives the better performance of 91% accuracy.**

Keywords: Liver disease prediction, Grid search, K-Nearest Neighbor, Machine Learning,

I. INTRODUCTION

The most significant organ in the human body is Liver as it is responsible for various functions of the body such as metabolism, digestion, and immunity and supplies nutrients in the body. Liver diseases are one of the most killer diseases by the most cause of Viral Hepatitis in the world. Liver disease is rising very rapidly, to diagnose liver disease effectively, it is important to detect disease early, which can increase the survival rate. The failure of liver functioning leads to serious health complications. Liver diseases are mainly caused by various factors such as stress, changed lifestyle, contaminated food, consumption of alcohol and drug intake, etc. The disease can be predicted based on health parameters, oral conditions - like alcohol, city pollution level, movement, body chemical compositions using advanced AI/ML techniques. ML is the branch of AI in which a machine learns from a dataset and its performance measures improve with real data over time. The different techniques of ML have been adopted for diagnosis and prophecy of various diseases in the field of medicine. Due to easy access to clinical data, ML algorithms play an important role in medical decision making. Therefore to identify the disease and make a real-time effective decision the design and develop of a ML model will play a major role. Several ML classification algorithms exist to predict the Liver disease. Each algorithm has different ways of learning from the data set and can be refined / performance tuned. The paper focuses on KNN algorithms, steps to be performance to optimise the model, step by step developing several models. The reason for picking a KNN algorithm is, which looks for several nearby values to classify diseases. Which helps to analyse various more effectively, as increased K value looks at several nearest values, before classifying disease.

Instead of building models using US/Europe based data sets, the paper works on building ML models effectively using Indian dataset and paper discussed how to analyze and predict with more accuracy step by step - preprocessing of data, Univariate and by variate analysis, feature selection, Feature engineering then ML model is trained using this data. As part of preprocessing data is analyzed and checked each feature distribution, most importantly is the data set is balanced/imbalanced and then appropriate methods used to transform the data to normal distribution and imbalance data set is balanced using various methods.. Feature engineering of the dataset is performed to get the important features for prediction and remove the less contributing features so that computation time of the model can be reduced. Hyper

978-1-7281-6828-9/20/$31.00 ©2020 IEEE

parameter tuning methods, tune the parameter's values of KNN ML model to get high accuracy. The paper builds several models, using all these techniques, to indicate performance importance and finally achieving a final KNN ML model to predict liver disease effectively for Indian Dataset. The several models built to predict liver disease and how performance improves is discussed in next sections.

In Section II the Related work is presented. Under section III, ML algorithm and preprocessing techniques which are used in our liver disease prediction ML models are discussed and In Section IV, results of all the models are shown and conclusion with future work is presented in section V.

II. RELATED WORK

Liver disease is considered as a crucial problem in the medicine domain. If the disease is not identified and treated, this can lead to death. Therefore it is necessary for the researchers to identify and investigate liver disease with different classification techniques.

The data classification using different datasets for liver disease [3] shows the result so that the Multilayer perceptron gives classification accuracy of 71.59% compared to other classifiers. To predict liver disease the liver datasets which are identified for developing classification models [4] have applied various classification algorithms to the datasets. With the experimental results we establish that the classification accuracy is superior using FT Tree algorithm compared to remaining algorithms which results in an Accuracy of 78.0%, precision of 77.5%, resulting in a sensitivity of 86.4% and 38.2% of Specificity. The various techniques of classification, such as KNN, SVM and decision trees are to the dataset [5], are gathered from 16,380 different analysis results per year. This may help further for minimizing the number of analyses. The evaluation performed on various types of liver dataset [6]. The performance of the Back Propagation algorithm was 71.59%. Further, the K-NN classification method has been applied above Indian liver dataset, which gives accuracy of 89.47%.

The accuracy for the Decision tree was 69.40%, with the other models being lower. With three models - Decision Tree, K-Nearest neighbor and Logistic [7] the Indian liver patient dataset from the UCI-ML repository.

An accuracy of 93.75% in analysis of liver disease DT [8] in comparison to other data mining classification methods used. The developed J48 classification model [9] resulted in a good performance that of other models after applying Particle Swarm Optimization (PSO) with characteristic engineering, which gives an accuracy of 95.04%.

The review on various classification and prediction techniques [10] for liver disease. Here they have basically compared FT growth and Naïve Bayes algorithms. We can observe that the Naïve Bayes (75.54%) gives better accuracy compared with the FT growth model (72.66%). Moreover this comparison was carried out on 29 datasets with 12 different features. By using different classification techniques [11] found various results namely The C4.5 algorithm gives 65.59%, The Naive Bayes algorithm gives 63.39%, and using BNND (Bayesian Network with Naïve Dependence)gives 61.83% and lastly using BNNF gives 61.42%.

III. METHODOLOGY

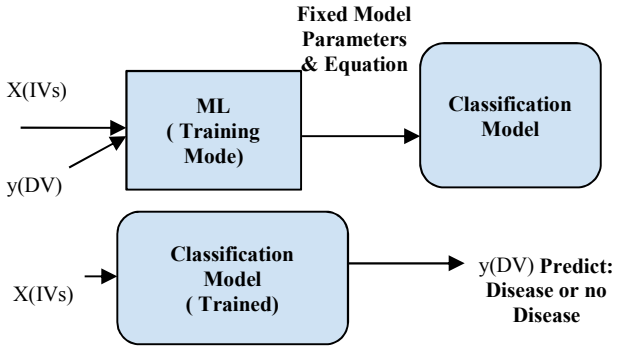


Fig. 1. Training and Testing Model Overview X: all features 1 to 10 and y: Target feature

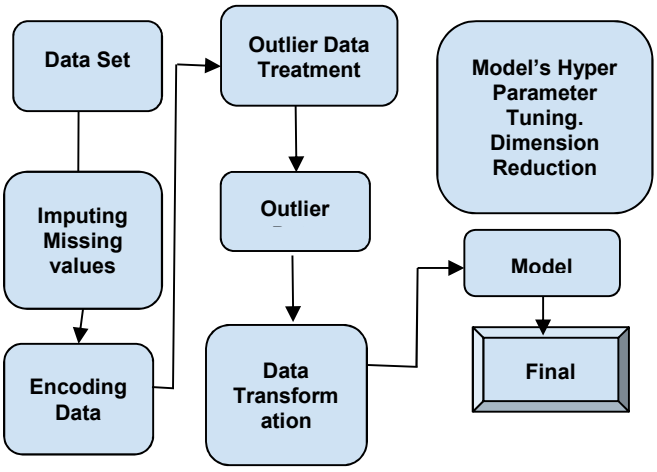


Fig. 2. Overall Model building Process

A. Liver Dataset

The Dataset for Indian Liver Patients obtained from Andhra Pradesh, India. It consists of records of 416 liver patients and records of 167 non-liver patients. The data set contains 441 male records and 142 female records. The patients were divided into 2 categories, one category represents healthy livers and other one represents without healthy livers. The attributes which are considered for our implementation were the following.

1) Data Pre-processing

The Pre-processing of the data is carried out before applying machine learning algorithms to the dataset. The datasets require purification and revision. Performance and accuracy of the predictive model are not only affected by the algorithms used but also by the quality of the dataset and pre-processing techniques. Fig.1(a) shows the converting categorical data to numerical data for gender and count plot is shown in Fig. 1(b).

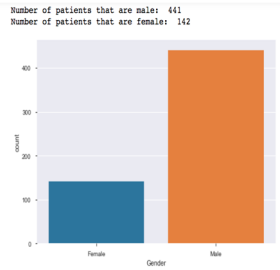


Fig. 1(a).

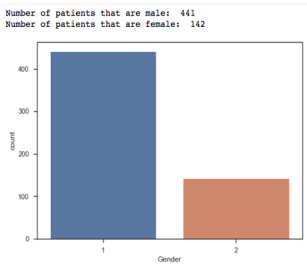


Fig. 1(b).

TABLE1: DATA SET ATTRIBUTES

SI.NO	*Attributes*	*Attribute Type*
1	Age	Numeric
2	Gender	Numeric
3	Total Bilirubin	Numeric
4	Direct Bilirubin	Numeric
5	Alkaline Phosphotase	Numeric
6	Alamine Aminotransferase	Numeric
7	Aspartate Aminotransferase	Numeric
8	Total Protiens	Numeric
9	Albumin	Numeric
10	Albumin and Globulin Ratio	Numeric

2) *Feature Selection*

To achieve good accuracy for a proposed model we need to recognize extremely significant features from the dataset by realizing a feature selection method such as Correlation Matrix by means of Heat map. The correlation between the features is shown through the Heat map in Fig. 2(a) & 2(b). We can observe that features have positive, negative and zero correlation with other features.

Alamine_Aminotransferase	Aspartate_Aminotransferase	Total_Protiens	Albumin	Albumin_and_Globulin_Ratio
-0.086883	-0.019910	-0.187461	-0.265924	-0.216089
-0.082332	-0.080336	0.089121	0.093799	0.003404
0.214065	0.237831	-0.008099	-0.222250	-0.206159
0.233894	0.257544	-0.000139	-0.228531	-0.200004
0.125680	0.167196	-0.028514	-0.165453	-0.233960
1.000000	0.791966	-0.042518	-0.029742	-0.002374
0.791966	1.000000	-0.025645	-0.085290	-0.070024
-0.042518	-0.025645	1.000000	0.784053	0.233904
-0.029742	-0.085290	0.784053	1.000000	0.686322
-0.002374	-0.070024	0.233904	0.686322	1.000000

Fig. 2(a).

	Age	Gender	Total_Bilirubin	Direct_Bilirubin	Alkaline_Phosphotase
Age	1.000000	-0.056560	0.011763	0.007529	0.080425
Gender	-0.056560	1.000000	-0.089291	-0.100436	0.027496
Total_Bilirubin	0.011763	-0.089291	1.000000	0.874618	0.206669
Direct_Bilirubin	0.007529	-0.100436	0.874618	1.000000	0.234939
Alkaline_Phosphotase	0.080425	0.027496	0.206669	0.234939	1.000000
Alamine_Aminotransferase	-0.086883	-0.082332	0.214065	0.233894	0.125680
Aspartate_Aminotransferase	-0.019910	-0.080336	0.237831	0.257544	0.167196
Total_Protiens	-0.187461	0.089121	-0.008099	-0.000139	-0.028514
Albumin	-0.265924	0.093799	-0.222250	-0.228531	-0.165453
Albumin_and_Globulin_Ratio	-0.216089	0.003404	-0.206159	-0.200004	-0.233960

Fig. 2(b).

3) *Hyperparameter Tuning Techniques*

ML model's hyper parameter's tuning essential to build an effective ML model for dataset. For KNN we have fine tuned key parameters like solver, random_state, penalty, C, class_weight, multi_class and mix_iter. Grid search finds out the combinations of parameter's values which gives better results in terms of accuracy performance measures.

The Total Bilirubin & Direct Bilirubin, Alamine Aminotransferase & Aspartate Aminotransferase and Total Protein & Albumin are the top-3 Correlations need to be looked into. But only Total Bilirubin & Direct Bilirubin has high (87%) correlation.

4) *Outlier Checking*

An outlier is an observation or a data point that lies outside the overall pattern of a distribution. If few data are very different or not in range of overall pattern then those are called outliers. This leads to skewness and affects the mean and standard deviation. So we need to detect and remove the outliers. The Total_Bilirubin, Direct_Bilirubin, Alkaline Phosphotase, Alamine Aminotransferase and Aspartate Aminotransferase are represented in Fig 3 as Outliers in Box Plots.

0

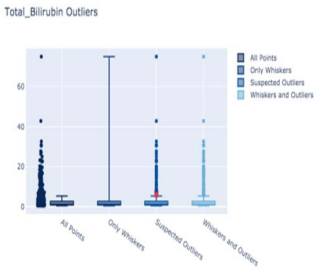


Fig. 3(a).

Fig. 3(b).

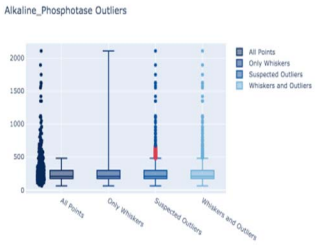


Fig. 3(c).

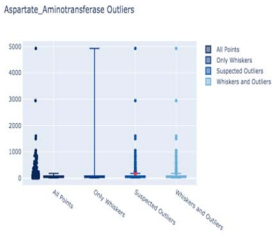


Fig. 3(d).

5) *K – Nearest Neighbor (KNN)*

The KNN method is a supervised ML method used for classification. KNN is used as our main machine learning model. For statistical estimation and pattern recognition this method is extensively used. KNN stores all of its available cases and it classifies latest cases supporting a similarity gauge. Since it is a non-parametric technique it won't formulate any prior assumption. KNN measures distance to predict the class by using Euclidean.

KNN algorithm uses Euclidean Distance to calculate the similarity between the data points. A majority vote of its neighbors will lead to a case classification, where case will be assigned to the class of most common along its K nearest neighbors calculated by a distance role. KNN algorithm is selected, as we can classify data using maximum number of nearby values, so that if a number of similar data looked more, the chance of generalizing based on few data will be avoided and hence mis-classifying non diseases person as diseases

person will be reduced. The divided dataset namely training and test sets. The training dataset is used for model building. Depending on the square root of the number of observations a k- value is decided. The test data is predicted on the model built, ML Model performance is tuned using sampling techniques and Hyper-parameter tuning.

III. IMPLEMENTATION AND RESULTS

Anaconda is a free and open-source distribution of Python. We have used Anaconda3 for our implementation work. It includes libraries of various algorithms written in python language. Anaconda3 uses python 3.7 and we have used Jupytor Notebook to run the codes.

A. Performance Measure

With the help of evaluation matrices we need to understand how good our ML model is going to perform on data. For balanced datasets the accuracy, precision and recall matrices are good ways to evaluate classification model, although if the data is imbalanced and there is class disparity, then other methods like ROC/AUC perform better in evaluating the model performance.
Recall Metric will measure from the total number of positive values how many true positives of the model has classified

$$Recall = \frac{T_P}{T_P + F_N} \quad (1)$$

Precision Metric indicates number of True Positives which are actually positive in contrast to the total number of positively predicted values

$$Precision = \frac{T_P}{T_P + F_P} \quad (2)$$

F1-score measures test accuracy, which considers together precision and recall to calculate the score, where F1 score attains its best value at 1 and worst at 0. The ROC is used when observation across classes of model predictions is balanced.
Confusion matrix, PRC and ROC for all ML models are shown in Fig. 4, 5, 6, 7, 8, 9, 10 and 11.

Model-1(KNN): Basic Data Processing, Categorical Data Transformation and NaN data replacement.

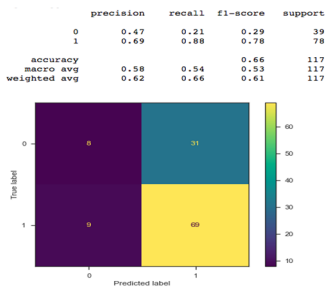


Fig. 4(a).

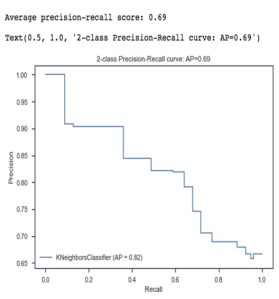
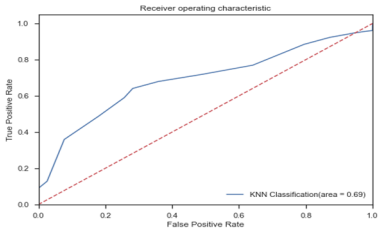


Fig. 4(b).



0.6581196581196581 0.5448717948717948 0.69 0.6873076923076923 0.7752808988764044

Fig. 4(c).

Fig. 4(a), 4(b) and 4(c) Represents Confusion matrix, PRC and ROC of Model-1 performance evaluation metrics resepectively.

Model-2 (KNN): After removing Outliers, performance of the model is given as below.

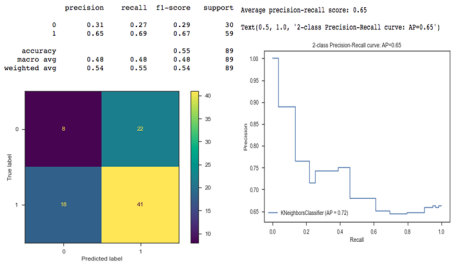
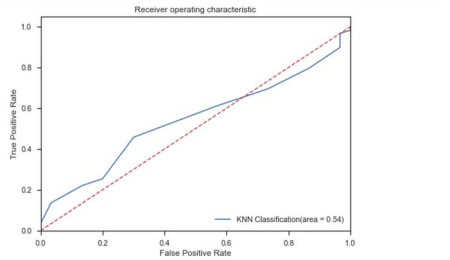


Fig. 5(a). Fig.5(b).



0.550561797752809 0.48079096045197744 0.6507936507936508 0.6544936263085187 398305084745763

Fig. 5(c).

Fig. 5(a), 5(b) and 5(c) Represents the Confusion matrix, PRC and ROC of Model-2 performance evaluation metrics respectively.

Model-3(KNN): Instead of Removing Outlier, we trimmed outliers with Extreme Values using Q1(15 Percentile), Q3(85 Percentile) and Min value trimmed at Q1 - 1.25 (Q3-Q1) and Max value trimmed at (Q3 + 1.25(Q3-Q1).

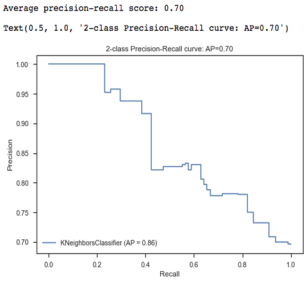
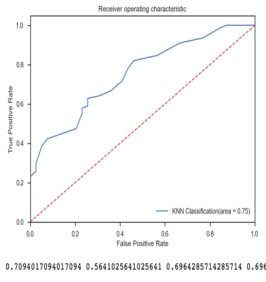


Fig. 6(a).



0.7094017094017094 0.5641025641025641 0.6964285714285714 0.6964285714285714

Fig. 6(b).

Fig.6(a) and 6(b) Represents the PRC and ROC of Model-3 performance evaluation metrics respectively.

Model-4: Trimming Extreme outliers - 3 IQR and removed one feature based on correlation Matrix.

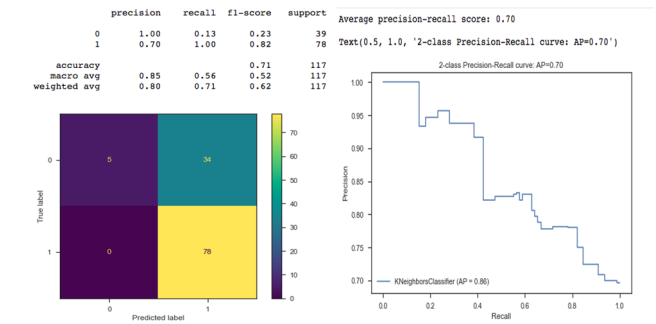


Fig. 7(a). Fig. 7(b).

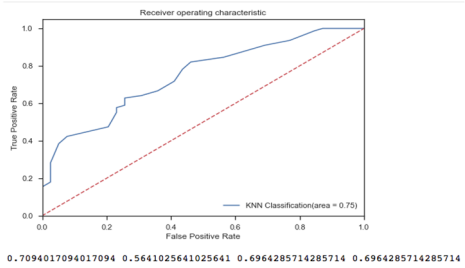


Fig. 7(c).

Fig.7(a), 7(b) and 7(c) Represents the Confusion matrix, PRC and ROC of Model-4 performance evaluation metrics respectively.

Model-5: The model extended model after Balancing Data Set using different re-sampling methods.

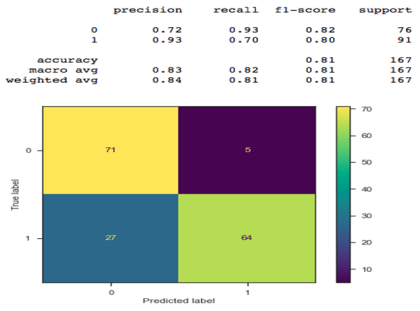


Fig. 8(a).

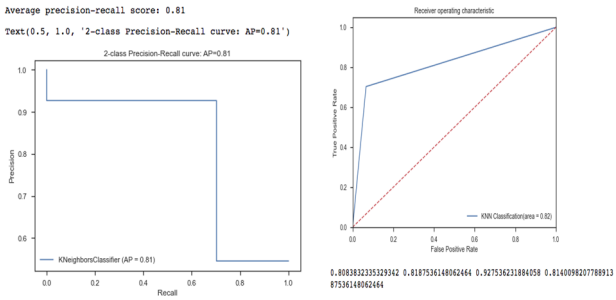


Fig. 8(b). Fig. 8(c).

Fig. 8(a), (b) PRC and (c) Represents the Confusion matrix, ROC of Model-5 performance evaluation metrics respectively.

Model-6: The model extends model 5, performing Robust Scalar Transformation of Data, and model trained using transformed data.

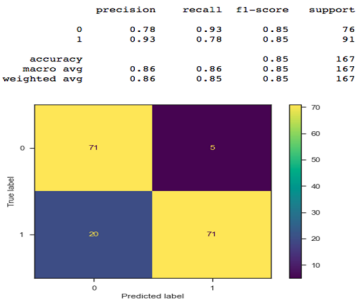


Fig. 9(a).

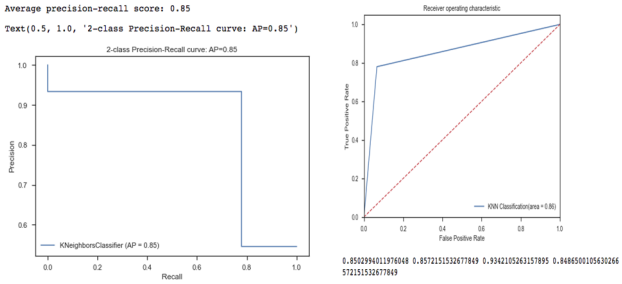


Fig. 9(b). Fig. 9(c).

Fig. 9(a), (b) and (c) Represents the Confusion matrix, PRC and ROC of Model-6 performance evaluation metrics respectively.

Model-7: Robust Scalar Transformation of Data and Dropping Gender Feature.

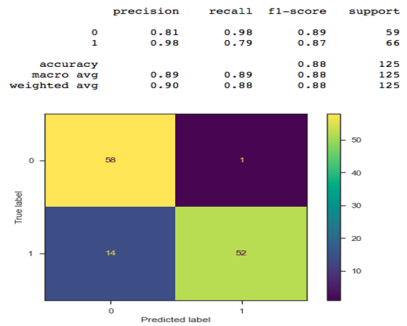


Fig. 10(a).

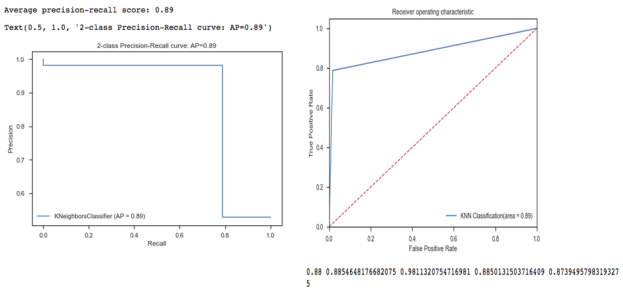


Fig. 10(b). Fig. 10(c).

Fig.10(a), 10(b) and 10(c) Represents the Confusion matrix, PRC and ROC of Model-7 pperformance evaluation metrics respectively.

Model-8: Robust Scalar Transformation of Data and Dropping Gender Feature.

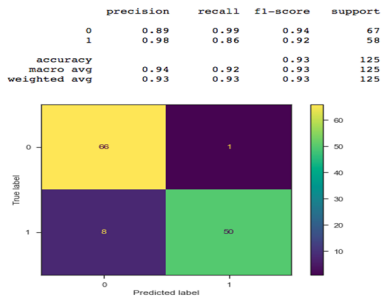


Fig. 11(a).

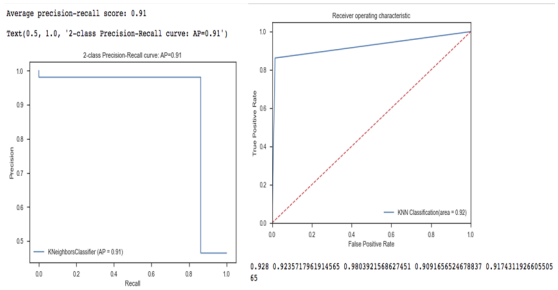


Fig. 11(b). Fig. 11(c).

Fig. 11(a), 11(b) and 11(c) Represents the Confusion matrix, PRC and ROC of Model-8 performance evaluation metrics respectively.

IV. CONCLUSION

In this work we have developed the K-Nearest Neighbor model to diagnose and predict liver disease. The data is transformed and further dimensionality reduction is performed to reduce the features to improve the model performance. The performance of classification and prediction techniques are evaluated on different performance measures some of them are precision , accuracy, recall and score of F-1. Grid Search is used for tuning the model's hyper parameters like solver, max-iterations, random-state etc. The model not only gives best accuracy, it also gives a perfect score in terms of AUC-ROC curve, precision, recall and other matrices of the model. The K-NN model performs better with an accuracy of 91%. In future this model can be utilized for larger and real time datasets with more attributes, so that the model can perform even more accurately.

REFERENCES

[1] Bendi Venkata Ramana, M. S. Prasad Babu and N. B. Venkateswarlu, A Critical Comparative Study of Liver Patients from USA and India. An Exploratory Analysis. International Journal of Computer Science Issues, ISSN: 1694-0784, May 2012.

[2] UCI Machine Learning Repository ILPD (Indian Liver Patient Dataset) https://archive.ics.uci.edu/ml /datasets/ILPD.

[3] Tapas RanjanBaitharua, Subhendu Kumar Panib, "Analysis of Data Mining Techniques For Healthcare Decision Support System Using Liver Disorder Dataset", International Conference on Computational Modeling and Security, 2016.

[4] Onwodi Gregory, Prediction of Liver Disease (Biliary Cirrhosis) Using Data Mining Technique, International Journal of Emerging Technology & Research, ISSN (E):2347-5900, ISSN (P):2347-6079, 2015.

[5] S. E. Sekar, Y. Unal,Z. Erdem, and H. Erdinc Kocer, Ensembled Correlation Between Liver Analysis Output, International Journal of Biology and Biomedical ngineering,ISSN:1998-4150

[6] B.V. Ramana, M. S. P. Babu, N. B. Venkateswarlu A critical study of selected classification algorithms for liver disease diagnosis Int J Database Manag Syst, 3 (2) (2011), pp. 101-114.

[7] H. Jin, S. Kim, J. KimDecision factors on effective liver patient data prediction. Int J Biosci Biotechnol, 6 (4) (2014), pp. 167-178.

[8] M. Abdar, M. Zomorodi Moghadam, R. Das, I.H. Ting Performance analysis of classification algorithms on early detection of liver disease Expert Syst Appl, 67 (2017), pp. 239-251.

[9] M. Priya, P. L. Juliet, P. R. Tamilselvi Performance analysis of liver disease prediction using machine learning algorithms Int Res J Eng Technol, 5 (1) (2018), pp. 206-211.

[10] S. Dhamodharan Liver disease prediction using bayesian classification 4th National Conference on Advanced Computing, Applications & Technologies (2014), pp. 1-3.

[11] A. S. Aneesh kumar, Dr. C. Jothi Venkateswaran, A novel approach for Liver disorder Classification using Data Mining Techniques , Engineering and Scientific International Journal ,ISSN 2394-7179,ISSN 2394-7187,2015.

[12] Karthik. S, Priyadarishini. A, Anuradha. J and Tripathi. B. K, Classification and Rule Extraction using Rough Set for Diagnosis of Liver Disease and its Types, Advances in Applied Science Research, 2011, 2 (3): page no 334-345

[13] S. Dhamodharan, "Liver Disease Prediction Using Bayesian Classification", 4th National Conference on Advanced Computing, Applications & Technologies, Special Issue, May 2014.

[14] V. A., S. S., S. N. and S. Ambesange, "Multi-Disease Prediction with Artificial Intelligence from Core Health Parameters Measured through Non-invasive Technique," 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2020, pp. 1252-1258, doi: 10.1109/ICICCS48265.2020.9121170.