Machine Learning project
June 2024 Exam session

Forecasting Indian liver patient records to aid medical diagnosis

Team 19 : Alfio Leanza[1], Lorenzo Caliaro[1], Riccardo Tavecchio[1]
[1]Cdlm Data Science

Patients with Liver disease have been continuously increasing because of excessive consumption of alcohol, inhale of harmful gases, intake of contaminated food, pickles and drugs. This dataset was used to evaluate prediction algorithms in an effort to reduce burden on doctors.

This paper aims to find a model that can predict an individual's susceptibility to the occurrence of liver disease so that a targeted plan for prevention and early intervention can be implemented.

CONTENTS

1 INTRODUCTION

In recent years, the number of liver disease patients in north-eastern Andhra Pradesh, India, has been steadily increasing. This worrying trend is attributable to several risk factors prevalent in the region, including excessive alcohol consumption, inhalation of harmful gases, intake of contaminated food, pickles, and misuse of drugs. To address this health emergency, it is essential to have advanced tools that can assist physicians in the diagnosis and management of liver disease. In this context, a comprehensive dataset was collected and used to evaluate the effectiveness of various prediction algorithms. The main goal is to reduce physicians' workload by improving diagnostic accuracy and optimizing treatment pathways for patients.

The following analysis focuses on the detection of the main variables and prediction of the target variable "dataset" depending on the available attributes. After a preliminary cleaning of the dataset and exploration of

variables, the analysis is therefore focused on the search for an optimal classification method with the aim of improving clinical outcomes i.e., prediction in identifying patients with liver disease.

2 DATASET EXPLORATION

The dataset used to reach the main goal of the analysis is the Indian liver patient records dataset, available on the Kaggle platform [2].This data set contains 416 liver patient records and 167 non liver patient records collected from North East of Andhra Pradesh, India. The "Dataset" column is a class label used to divide groups into liver patient (liver disease) or not (no disease). This data set contains 441 male patient records and 142 female patient records.

The features are:

- *Age:* patient's age
- *Gender:* patient's gender
- *Total Bilirubin:* Total amount of bilirubin present in the blood measured in mg/dL (milligrams per decilitre)
- *Direct Bilirubin:* Conjugated Bilirubin present in the bloodstream measured in mg/dL (milligrams per decilitre)
- *Alkaline Phosphotase:* in IU/L(InternationalUnit/Litre), a measure of enzyme concentration in the blood
- *Alamine Aminotransferase:* in IU/L (InternationalUnit/Litre), a measure of enzyme concentration in the blood
- *Aspartate Aminotransferase:* in IU/L (InternationalUnit/Litre), a measure of enzyme concentration in the blood
- *Total Proteins:* Total Protein concentration in the blood measured in g/dL (grams per decilitre)
- *Albumin:* Albumin concentration in the blood measured in g/dL (grams per decilitre)
- *Albumin and Globulin Ratio:* measure of the balance between albumin and globulin proteins in the blood
- *Dataset:* field used to split the data into two sets (1: patient with liver disease ; 2: no disease)

Our target variable is *Dataset* of binary type. An initial descriptive statistic showed that 72% of the observations consisted of value 1 (patient with liver disease) while the remaining 28% consisted of value 2 (patient without liver disease) thus presenting a skewed.

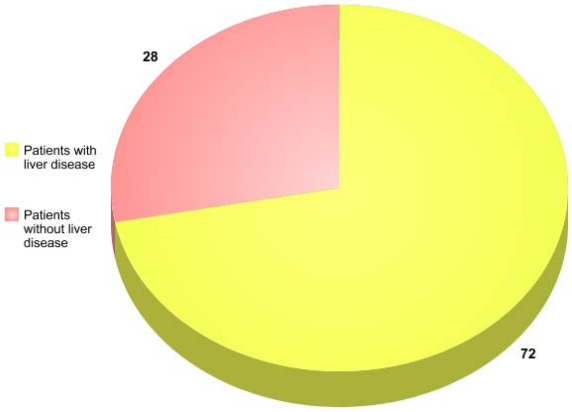

Figure 1: Pie chart of "Dataset" variable

The gender variable also presents a strong asymmetry in fact we note that the dataset is composed mainly of male subjects with a percentage of 76% while the female gender represents only 24%.

This marked asymmetry in the gender variable could be explained by the fact that the dataset was constructed to represent a specific reference population, which includes all individuals with potential liver disease, whether they actually have liver disease or not. Assuming that the sample was constituted following proper procedures and respecting the parameters of the reference population, it can be concluded that the gender proportions in the dataset reflect those of the reference census population.

3 PRE-PROCESSING

The dataset contained 4 null values, which were removed. Additionally, 13 duplicate

observations were detected and subsequently removed to ensure data integrity. The “gender” variable, originally in string format, was changed to a binary variable to avoid any errors during analysis, while the “dataset” variable was changed from numeric to binary.

After that, the correlation matrix was created to analyse the link between the variables.

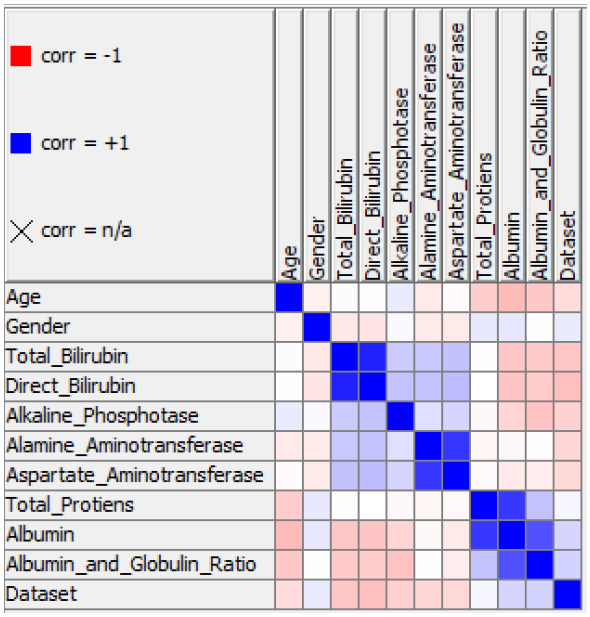

Figure 2: Linear correlation among variables

From the graph we can see a strong correlation of 87.4% between the Total_Bilirubin variable and the Direct_Bilirubin variable. Such a high correlation indicates a close connection between the two variables, suggesting that they essentially explain the same thing. Based on these considerations, it is possible to reduce the dimensionality of the data by eliminating one of the two variables, since it is redundant.

We also note that the correlations of the Total_Bilirubin and Direct_Bilirubin variables with each other are almost identical we then proceed with the removal of the variable Direct_Bilirubin.

By employing normalization techniques, we systematically transformed the attributes of the dataset to a standardized range, between 0 and 1. This process is particularly useful when dealing with features that exhibited disparate scales or ranges. Normalization ensures that each feature contributes proportionally to our analysis, preventing any unexpected influence from variables with larger numeric values.

4 MODELS AND PERFORMANCE MEASURES

4.1 CLASSIFICATION MODELS

Since the class attribute is a binary variable, it was decided to use regression models. In particular, the following models were selected:

Random Forest:

Is an ensemble learning method that constructs numerous decision trees. The "forest" created by the model comprises a collection of decision trees, typically trained using the bagging technique. This method generates multiple subsets of the training dataset by randomly sampling with replacement. Each subset is utilized to train a different decision tree, enabling the breakdown of information into multiple variables to reach the best decision for a classification problem. For each decision tree, only a random subset of features is considered at each split. This introduces diversity among the trees and helps mitigate overfitting to specific features.

Logistic regression:

Logistic regression is a binary classification algorithm that estimates the probability that an instance belongs to a specific class. Despite its name, logistic regression is used for classification tasks rather than regression. Specifically, the model output is interpreted

as the probability that the given instance belongs to the positive class: if the output is close to 1, the instance is predicted to be in the positive class; if the output is close to 0, it should be in the negative class. In our particular analysis, the output can be found in the prediction column, where we can observe the probability of each patient having liver disease or not

Decision tree:

The Decision Tree is a supervised learning model that organizes data into a tree structure. Each node represents an attribute and each branch a possible choice or outcome. Through a series of questions, the model classifies or predicts the output. It is simple to interpret and suitable for small to medium-sized datasets, but may be subject to overfitting on complex datasets.

Multi-Layer Perceptron (MLP):

Is a separation model, which partitions the attribute space via a mathematical function. During training, the model will optimize the weights and biases to produce accurate predictions about the target binary variable.

4.2 PERFORMANCE MEASURES

In order to evaluate the goodness of a model and compare it with others, different measures of performance. However, it is necessary to introduce the concept of a confusion matrix which provides a representation of the number of records classified correctly by model in question, i.e. the true positives (TP) and true negatives (TP), and those misclassified, false positives (FP) and false negatives (FN):

		Real value	
		-1	1
Prediction	-1	TN	FN
	1	FP	TP

considering the different ratios in which it is possible to combine these values, measures with different meanings are obtained:

- *Accuracy* : Accuracy measures the percentage of records correctly classified by the model compared to the totality of records in the Test set. Usually we tend to prefer the model that has the highest accuracy value but it is still essential to combine this measurement with other criteria.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

- *Recall* : Recall value indicates the percentage of records predicted as positive among all actually positive records. The relationship between true positives and actual positives is also called Sensitivity.

$$Recall = \frac{TP}{TP + FN}$$

- *Precision* : Precision of a model evaluates the percentage of records that are actually positive in proportion to those classified as such.

$$Precision = \frac{TP}{TP + FP}$$

- *F1 Measures* : F1-measure allows you to summarize Precision and Recall in a single measurement, in fact it represents the harmonic mean between the two

$$F1\ Measures = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

5 ANALYSIS AND RESULTS

5.1 RESEARCH OBJECTIVE

The research objective is to identify the best predictive model for the target variable “Dataset” which indicates whether a patient has liver disease or not. The aim is to develop a model that accurately predicts an individual's susceptibility to liver disease, facilitating early diagnosis and intervention, thus reducing the burden on healthcare providers.

5.2 HOLDOUT

Initially, the Holdout method was employed, which involves partitioning the dataset into two disjoint and exhaustive subsets using stratified sampling with respect to the target variable “dataset” . We divided the dataset into a training set (70%) and a test set (30%). The four classification models used were then trained on the Training set, validated with the Test set, and compared with each other. First, it was chosen to use all variables in the dataset as explanatory attributes since. We assume that they are all significant for correctly predicting the target variable.

Classification model	***Recall***	***Precision***	***F-Measure***	***Accuracy***
Decision Tree	0.73	0.75	0.74	0.63
Logistic Regression	0.96	0.72	0.82	0.70
Random Forest	0.85	0.77	0.80	0.71
Multi-layer perceptron	0.88	0.75	0.81	0.71

Figure 3: Performance measures with all attributes as an explanatory variable

Logistic Regression has the highest recall (0.96) and the best F-Measure (0.82), indicating that it has a good ability to correctly identify positive instances and balance between precision and recall. However, having a higher "accuracy" Random Forest and Multy-layer perceptron might be the better choice.

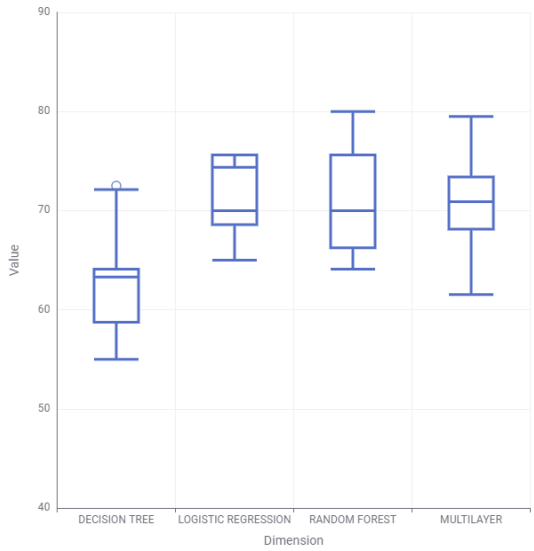

Figure 4: Box plot of accuracy with all variable selected

5.3 FEATURE SELECTION

The Feature Selection technique was applied to evaluate the significance of each attribute through both univariate and multivariate filters, using objective functions. In our case, we employed InfoGainAttribute for univariate filtering and CfsSubsetEval for multivariate analysis. This is because CfsSubsetEval evaluates feature subsets by taking into account not only the correlation between each feature and the target variable, but also the correlation between the features themselves. In the univariate approach, the function identified all attributes present in the dataset as significant attributes for the initial analysis. However, in the multivariate approach, the most relevant attributes highlighted were ***Total_Bilirubin, Alkaline_Phosphotase,Alamine_Aminotransferase,Aspartate_Aminotransferase,Albumin_and_Globulin_Ratio*.** After partitioning the dataset as in the previous steps, all the classifiers were re-run taking into account the attributes chosen by the CfsSubsetEval function.

Classification model	***Recall***	***Precision***	***F-Measure***	***Accuracy***
Decision Tree	0.697	0.722	0.709	0.6
Logistic Regression	0.605	0.828	0.699	0.635
Random Forest	0.882	0.761	0.817	0.724
Multi-layer perceptron	0.992	0.698	0.819	0.724

Figure 5: Performance measures with attributes chosen by feature selection

Considering the accuracy and overall balance of the metrics, Random Forest seems to be the best overall choice. It has high recall (0.882), good precision (0.761), excellent F-Measure (0.817), and best accuracy (0.724). Consequently the feature selection technique confirmed that the best classifier for our research objective is the *Random forest*

5.4 ROC CURVE

A further reliable measure for evaluating performance is the calculation of the area under the ROC (Receiver Operating Characteristic) curve, which graphically represents the relationship between the percentage of false positives on the x-axis and the percentage of true positives on the y-axis.

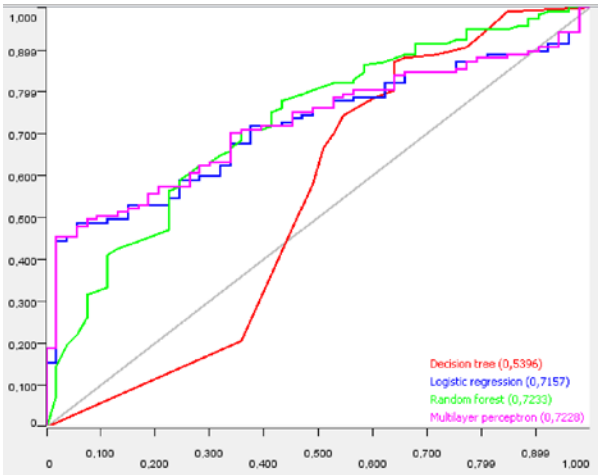

Figure 6: ROC curve with attributes chosen by feature selection

By the ROC Curve we can confirm that Random forest has the best performance because maximizes the area under the curve (0.723). The other models has also good performance except for Decision Tree.

6 CONCLUSION

The Random Forest model has been identified as the best classifier for predicting patients' susceptibility to the onset of liver diseases. This model not only achieved the best results in terms of recall, precision, F-Measure, and accuracy but also demonstrated excellent performance in ROC curve analysis. Implementing this model can significantly reduce the workload of physicians by improving diagnostic accuracy and optimizing treatment pathways for patients with liver diseases.

The topic lends itself to future developments and more in-depth analyses. it would be interesting to integrate the dataset with additional variables that can better explain the target variable.

7 REFERENCES

- Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, Vipin Kumar; Introduction to Data Mining (Second Edition); 2018
- [2]https://www.kaggle.com/datasets/uciml/indian-liver-patient-records
- Kevin P. Murphy, Machine Learning: A Probabilistic Perspective, 2012
- Fabio Antonio Stella: “Machine Learning – Anno 2023-2024”, E-learning unimib, 2024