Indian Liver Patient

Paolo Guerini Rocco
Giovanni Cornacchia
Dariia Haryfullina

Abstract:

The main goal of our study is to elaborate and evaluate the efficiency of different Machine Learning algorithms at predicting whether a patient has liver disease. The source data provides information regarding the levels of enzymes in the blood and the biological features of patients from the North-East of Sandhra Pradesh in India [13].
To solve this classification problem with a binary response variable, heuristic, regression-based and probabilistic models were implemented in KNIME, using 4 different variations of the dataset (original scale with outliers and without outliers, log-scale with and without outliers). After computing the metrics, such as F1-score, Balanced Accuracy, Mean Accuracy and its standard deviation, we were able to compare the performance of the models carried out on versions of the data stated above.

Introduction:

Liver diseases have progressively become more and more prevalent in developing societies due to unhealthy lifestyle and disruptive behaviours, such as smoking, drinking and consuming drugs. For example, severe acute alcoholic hepatitis is associated with a mortality rate of up to 50%, and that treatment of this form is extremely difficult [14]. In addition, problems with liver patients are not easily discovered in an early stage as it will be functioning normally even when it is partially damaged, so an early diagnosis of liver problems by analysing the levels of enzymes in the blood is essential to increase patient's survival rate [12].
The aim of our research is then to determine whether a patient has liver disease based on his medical data. Machine Learning algorithms will be used to tackle the issue of providing accurate diagnoses, which can help the doctors to provide their patients with the appropriate treatments in a timely manner.
The classification models that were implemented in this classification study were Random Forest, Gradient Booster Tree, Logistic regression, Bayesnet, J48 and NBtree. All the models are implemented via Weka predictor 3.7 in KNIME.

Dataset:

The original source dataset for the analysis contains disease patients' records from the North-East of Sandhra Pradesh in India [13]. It is hosted on kaggle.com, an online community of data scientists and machine learning practitioners. It allows users to find and publish datasets, explore and build models in a web-based data-science environment.
The dataset is composed of 583 records and 11 columns, 10 of which are explanatory variables which provide information regarding the levels of enzymes in the blood and the biological features of patients. The last one is the response variable used to perform and evaluate the performance of the classification task.

The type, meaning and ranges for the variables are the following:

- Age of the patient (numerical - ordinal), the age is normally-distributed among the sample population (see chart A).

- Gender of the patient (categorical - binary), the sample we are working on is composed of 75.6% of men and 24.4% of women.

- Total Bilirubin in mg/dL (numerical - ordinal), bilirubin is a red-yellow compound which human body uses to eliminate the waste. The high level of this substance in the blood can indicate several diseases, among which the liver dysfunction. This attribute indicates the total quantity of the bilirubin (direct and indirect) in the patient's blood sample [15].

- Direct Bilirubin in mg/dL (numerical - ordinal), or Conjugated Bilirubin is a type of Bilirubin, which underwent a chemical change In the liver. High levels of this substance may indicate that the liver is not working well. The normal rage of Direct Bilirubin in adult's body is 1.2 mg/dL [2].

- Alkaline Phosphatase in IU/L (numerical - ordinal) is an enzyme present in most of the body tissues. High level of this substance may indicate the presence of liver or bone problems [1].

- Alanine Aminotransferase in IU/L (numerical - ordinal), or commonly ALT, is an enzyme used as a marker of liver disease, if found in high concentration [16].

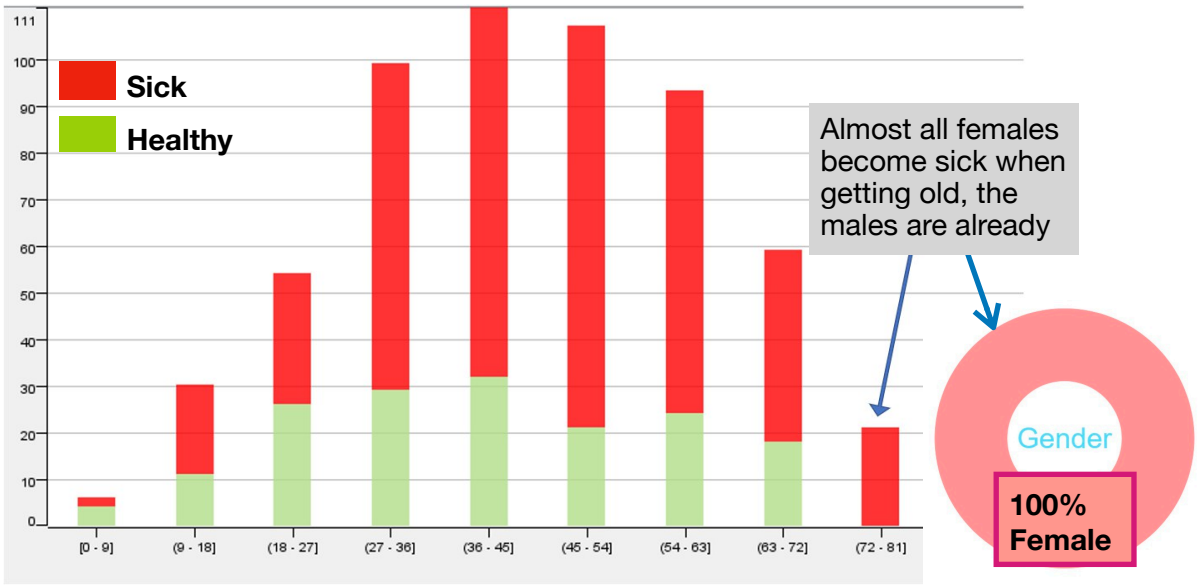
- Aspartate Aminotransferase in IU/L (numerical - ordinal), or AST, is an enzyme generally located in the liver or in the muscle tissue. If the liver is damaged, it is releasing a high quantity of this substance in the blood [3].

- Total Proteins in g/dL (numerical - ordinal), refers to the total amount of proteins (Globulin and Albumin) in the blood [5]. The normal range is from 6 to 8.3 g/L, higher or lower quantities point out liver disease or other sicknesses [6].

- Albumin in g/dL (numerical - ordinal), is an enzyme which has various functions, among them carrying hormones and vitamins through our blood system. Lower levels of albumin point out the presence of different kinds of diseases, among which liver dysfunctions [4].

- Albumin and Globulin Ratio (numerical - ratio), if found in low quantity, can be used as an indicator of a liver disease the normal range is between 1.1 and 2.5 [5].

- The response variable is called "dataset" (categorical - binary), it is equal to 1 in case of disease and 2 otherwise. The proportions show that 71.4% of the patients are suffering from liver disease, while 28.6% are not. This means the classes are a bit imbalanced, but not extremely.



a) Binned density distribution of the age, divided by health status.

Data pre-processing and EDA:

Before starting to apply the algorithms on our dataset, it is first necessary to pre-process the data in order to make sure that it is correct and in an appropriate format to allow effective training of the models. The procedure of cleaning the dataset includes several steps:

1. Deletion of wrong records: 3 rows were deleted due to an impossible value for the direct bilirubin. Since direct bilirubin is a component of the sum to calculate total bilirubin (direct + indirect), it is then impossible that its value alone were to exceed the total. Afterwards, the variable total bilirubin is substituted with the newly calculated “Indirect bilirubin” to reduce collinearity.

2. Deletion of records with missing values: 4 rows were deleted due to missing values in the Albumin and Globulin ratio column.

3. Handling multicollinearity: collinearity among explanatory variables may lead the model to suffer from overfitting [17]. Overfitting means that the model focuses too much on the noise present in the data, thereby increasing considerably the variance of its performance. The most correlated explanatory variables have been identified the help of the correlation matrix and their multicollinearity measured through the variance inflation factor (VIF). The higher the VIF, the more correlated and inflated the explanatory variable. Usually a VIF value greater than 5 (or even more strictly 3) is considered to be worth treating [18]. One of the most problematic variables is Albumin, due to having high correlation values to both Total Proteins (78.20%) and Albumin/Globulin ratio (69.04%), plus its VIF value is the highest (10.38). For these reasons, eliminating this attribute from the dataset appears as a necessary decision to create a well-fitted model. In addition, another major correlated pair of variables is Alamine-Aspartate (79.20%). The VIF does not provide any indication to which variable to eliminate, so the criterion chosen is to discard the variable less correlated with the response variable. Since the correlation values are -48.89% for Aspartate and -70.19% for Alamine, Aspartate is removed from the dataset.

VIF Values

The table shows Variance Inflation Factors (VIF) across all numeric variables.

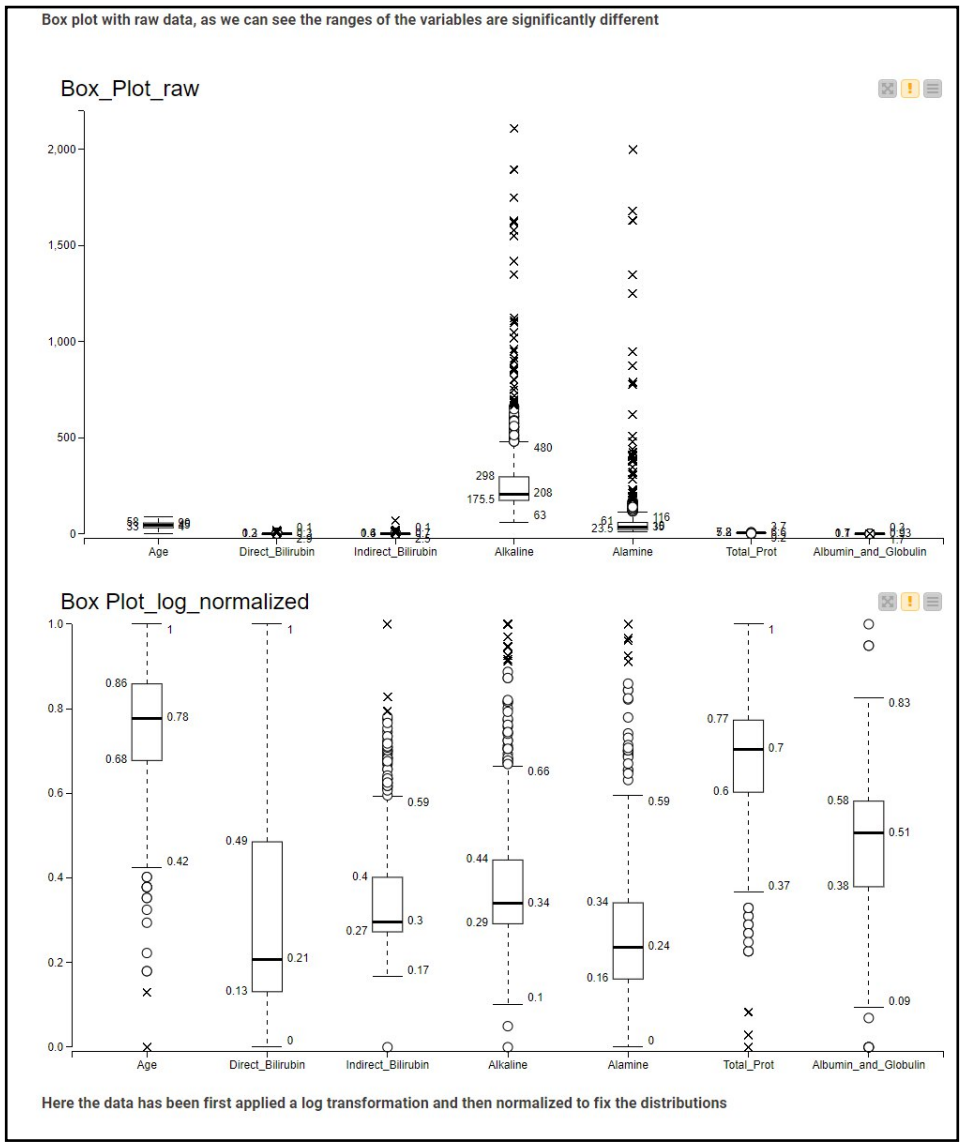
Age	1,11
Is_Female	1,04
Direct_Bilirubin	2,24
Alkaline_Phosphotase	1,13
Alamine_Aminotransferase	2,85
Aspartate_Aminotransferase	2,82
Total_Protiens	*5,62 (consider eliminating)*
Albumin	*10,38 (consider eliminating)*
Albumin_and_Globulin_Ratio	3,79
Liver_Status	1,13
Indirect_Bilirubin	1,89

b) VIF values computed over all the variables

4. Outlier detection: an observation was flagged as an outlier when the its value exceeded the range $R = [Q_1 - k \cdot (Q_3 - Q_1); Q_3 + k \cdot (Q_3 - Q_1)]$, where Q_n represents the nth quartile of the distribution [19] and the k value chosen was 3 due to 1.5 causing too many observations to be classified as outliers. Outliers are not necessarily a negative trait of the dataset which causes overfitting *per se*, since they might contain useful information to teach the model. For this reason, the outliers were not discarded *a priori*, but were instead kept in the datasets and flagged as such.

5. Feature transformation: since the Alkaline and Alamine have extremely long-tailed distributions with many outliers, a logarithmic transformation is applied to the dataset to reduce the effect of extreme values over the learning process. In addition, since the distributions have extremely different ranges of values, the variables are also constrained in a $[0; 1]$ interval via normalisation. The normalisation procedure aims to prevent wider-ranged variables from prevailing over the others during the learning process via the following formula:

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}}$$



c) Distribution boxplots pre and post transformations

Data Partitioning:

10-fold cross-validation was chosen to partition the dataset in order to better monitor the variance of the classification performance. In a nutshell, the dataset was partitioned in 10 training-test set pairs via stratified sampling over the response variable so that each model could be run ten times instead of one as it would be in a standard holdout procedure. Given the relatively small size of the dataset (583 records), the computational burden to run the cross-validation is trivial and it aids to make better use of the limited amount of records by re-using them multiple times.

Implementation of the models:

The classification models that were implemented in this study were Random Forest, Gradient Booster Tree, Logistic regression, Bayesnet, J48 and NBtree. All the models are implemented via Weka predictor 3.7 in KNIME.
Random Forest and J48 are heuristic classification techniques, which use decision trees to predict the results based on the majority rule in the decisions outcomes, to find a solution of the classification problem.
Logistic regression is a binary regression-based classifier which predicts the class of a binary response variable given the input of the regressors.
Bayesnet or Bayes Network is a probabilistic model which allows to predict the class of the response variable based on conditional dependencies between the explanatory variables [20].
NBtree, where NB stands for “Naïve Bayes” is a probabilistic classification technique which predicts the class of an observation based on the naïve independence assumptions and using Bayes’ Theorem, is a mixed model which has both probabilistic and heuristic elements, as it implements the Naïve Bayes classifiers at every leaf node.
Gradient Boosted Tree uses gradient boosting (creating a loss function, which is applied to a weak learner) on the decision tree model. This is a strong learning algorithm which however, may overfit the data.

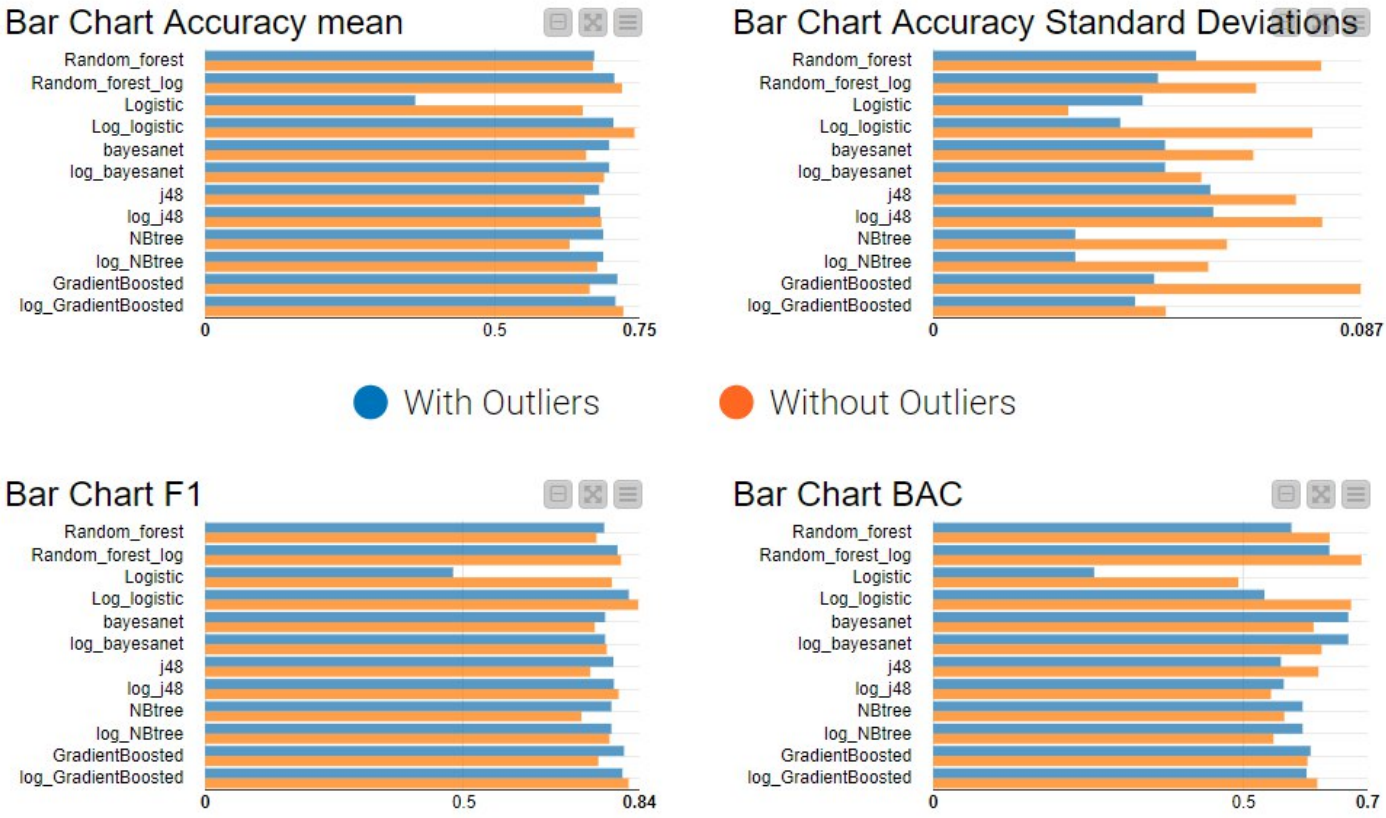
Performance evaluation and results:

The main evaluation metrics chosen are accuracy and F1 score. The accuracy is calculated with the ratio of correct predictions out of total predictions, and the F1 score is defined as the harmonic mean between Precision and Recall. The F1 score is chosen in addition to the accuracy because it better explains performance over unbalanced datasets.

The charts below (see chart D) show the results of the various models that were implemented (with or without transformations applied), using all the features available after the preprocessing. In addition, each one is shown when trained on a version of the dataset first including outliers (blue color) and then excluding them (orange color).
From the top-right chart it is evident that the models running on the dataset without outliers have double the standard deviation in accuracy as compared to their counterpart. For this reason, we can deduce that keeping the outliers tends to make the performance of the classification more consistent.
From the F1 chart, it is evident that the transformation increases performance in most of them. A particular callout is regarding the logistic regression, which jumps by almost 200% (from 0.48 to 0.84).

The performance of the top models tends to be similar, but the top scoring in term of accuracy and F1 score are the following:

- Log Logistic (F1=0.84, accuracy=0.65)
- Log Gradient Boost (F1=0.82, accuracy=0.72)



d) Model performance with and without outliers and transformations

References:

1. https://labs.selfdecode.com/blog/alkaline-phosphatase/
2. https://medlineplus.gov/lab-tests/bilirubin-blood-test/
3. https://medlineplus.gov/lab-tests/ast-test/
4. https://medlineplus.gov/lab-tests/albumin-blood-test/
5. https://medlineplus.gov/lab-tests/total-protein-and-albumin-globulin-a-g-ratio/
6. https://www.healthline.com/health/total-protein#results
7. https://en.wikipedia.org/wiki/Naive_Bayes_classifier
8. https://weka.sourceforge.io/doc.stable/weka/classifiers/trees/NBTree.html
9. https://en.wikipedia.org/wiki/Random_forest
10. https://www.ibm.com/cloud/learn/random-forest#toc-what-is-ra-DaEaNVdG
11. https://statisticaloddsandends.wordpress.com/2020/01/23/what-is-balanced-accuracy/
12. Ramana, Bendi & Babu, M.S.P.. (2012). "Liver Classification Using Modified Rotation Forest", *International Journal of Engineering Research and Development*. 1. 17-24.
13. https://www.kaggle.com/uciml/indian-liver-patient-records
14. Cainelli, Francesca. “Liver diseases in developing countries.” *World journal of hepatology* vol. 4,3 (2012): 66-7. doi:10.4254/wjh.v4.i3.66
15. https://www.urmc.rochester.edu/encyclopedia/content.aspx?contenttypeid=167&contentid=bilirubin_direct
16. https://medlineplus.gov/lab-tests/alt-blood-test/
17. https://towardsdatascience.com/multi-collinearity-in-regression-fe7a2c1467ea
18. https://hub.knime.com/adm/spaces/Public/latest/Components/Variance%20Inflation%20Factor%20(VIF)~dTcw3ZzkAiitPDRf
19. https://hub.knime.com/knime/extensions/org.knime.features.stats/latest/org.knime.base.node.stats.outlier.handler.NumericOutliersNodeFactory
20. https://www.bayesfusion.com/bayesian-networks/?gclid=Cj0KCQiAuP-OBhDqARIsAD4XHpduH0vrhsn9kn3tuLJPifqbBl6Zok_f0P3fIz7ZhyB2BWnPQ0oPn_kaApaqEALw_wcB