Classifiers compared for prediction of liver diseases

The purpose of the project is to identify a functional classification model for predictive purposes to determine whether, given certain clinical values, a subject has liver disease. For this purpose, several various classification models were tested to identify the best one.

UNIVERSITA' DEGLI STUDI DI MILANO BICOCCA

Open for Innovation KNIME

Introduction

Two million deaths each year globally are caused by liver disease, one half is attributed to complications from cirrhosis while the other from viral hepatitis and hepatocellular carcinoma.

Cirrhosis is the 11th leading cause of death while liver liver cancer the 16th: they constitute together 3.5 percent of deaths. Cirrhosis is also among the the top 20 causes of disability in the world.

About two billion people consume alcohol and, more than 75 million of them have been diagnosed with disorders related to its consumption with

a good chance of contracting diseases of the liver. Approximately two billion people suffer from obesity or are overweight and more than four hundred million have diabetes. Such conditions could lead to hepatic steatosis nonalcoholic liver disease and hepatocellular carcinoma.

The global percentage of those with viral hepatitis remains high, and liver damage induced by drugs continues to increase.

Liver transplantation is the second most most common solid organ (only 10% of the transplant requirements are met). Although this information may be considered disheartening, they highlight an important opportunity to improve public health, by supporting physicians with tools that can ease the workload they have to

cope with. In particular, some techniques of machine learning could be an important aid in the recognition of patterns that characterize an individual with liver disease

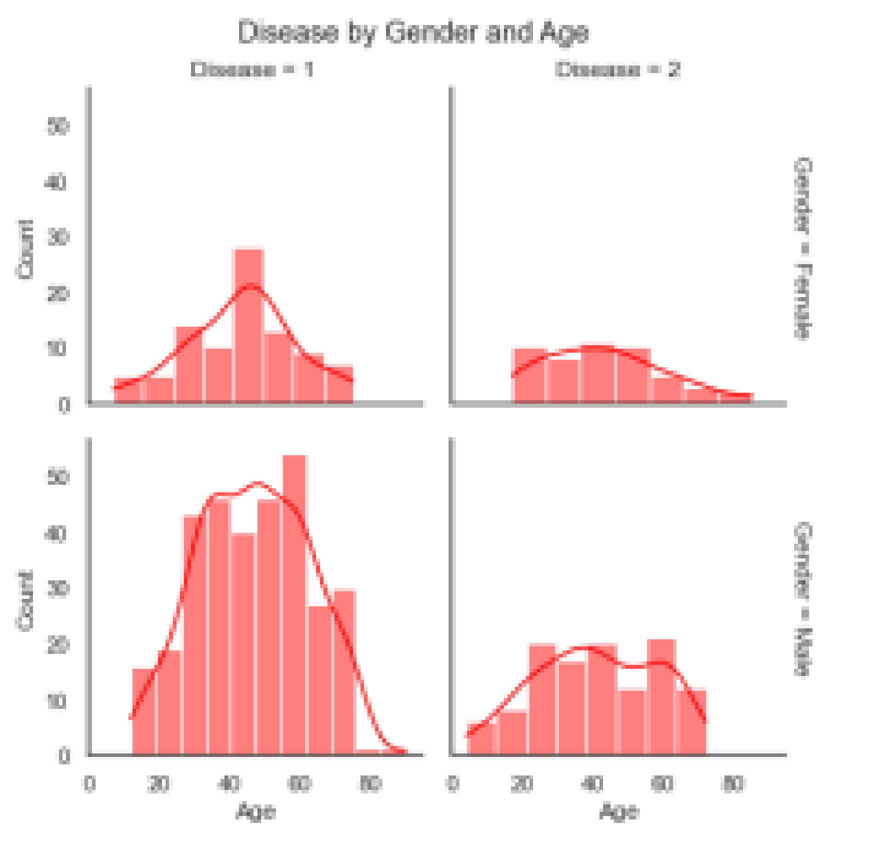
Dataset

In this study, data on liver disease were were collected from the University of California Irvine Machine Learning repository. The dataset contains 583 records consisting of 416 records medical records of liver patients and 167 medical records non-hepatic patients collected in northeastern Andhra Pradesh, India.

For more details : https://www.kaggle.com/datasets/uciml/indian-liver-patient-records

Pre-processing

To optimize usability, it was found useful to Binarize the " Gender " feature. The attributes categorical " Male " and " Female " weretransformed into binary attributes (0,1)



In the dataset used, the missing data (four) were replaced with the results obtained from the conditional mean, as this is one of the most effective methods despite critical issues. It is calculated the average of the values involved considering a given condition. In this specific, the reference conditions are:

- Sick/non-sick
- Male / female
- Age (divided into 10-year intervals)

It was decided to normalize all the data to improve the performance of the classifiers.

Classifiers

- SVM (Support Vector Machine)
- MLP (Multi-layer Perceptron)
- LogReg (Logistic Regression)
- Naive Bayes & Tree augmented naive bayes
- J48 (Decision Tree)

Performance evaluation

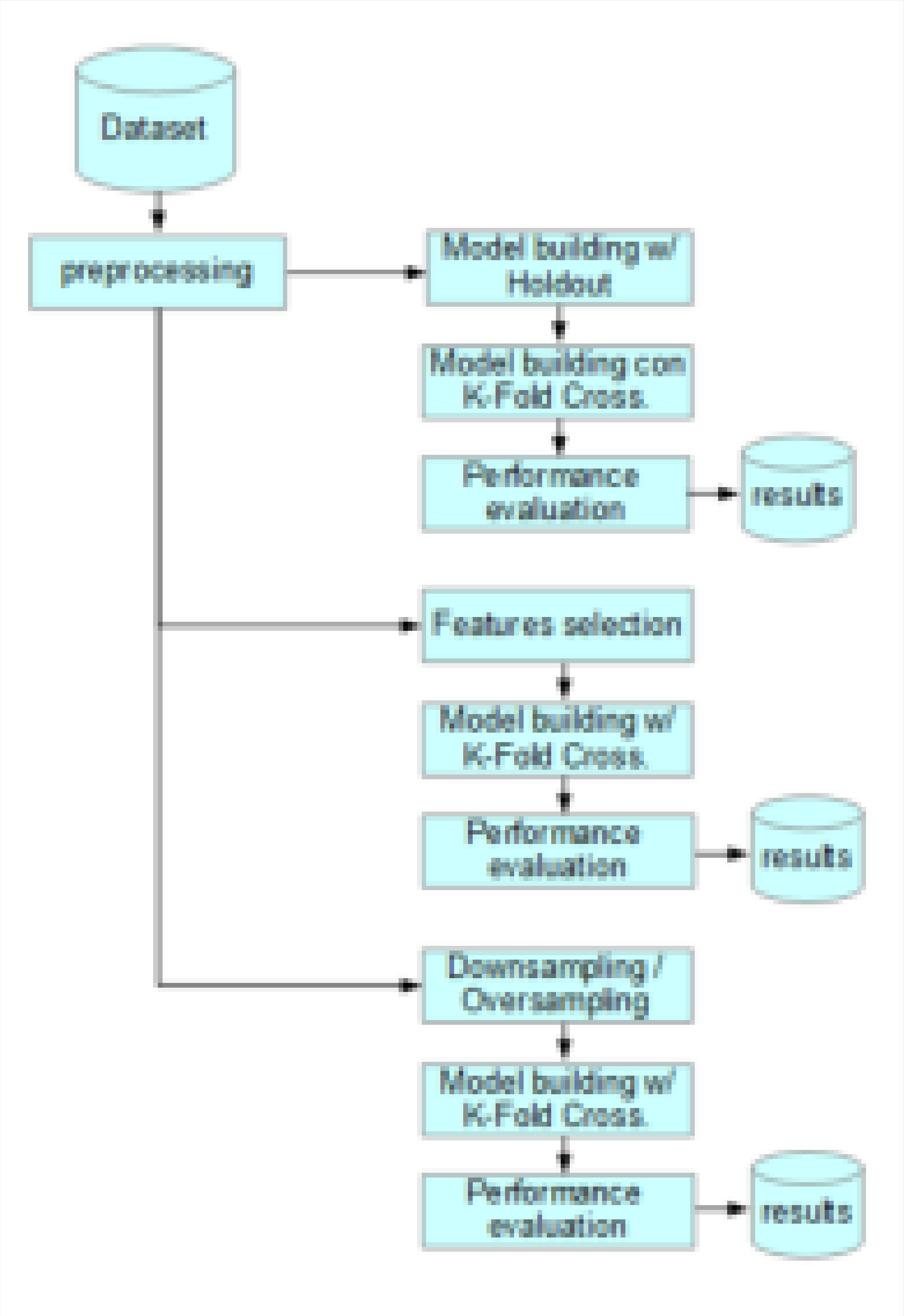
The dataset contains an unbalanced distribution of the Disease class variable, in fact there are 147 patients with liver disease, while there are 416 healthy patients, thus 71.4% of the entire dataset.

In cases like this measuring the performance of classification models through accuracy is not the most suitable method: the classifier, in fact, tends to neglect the least represented class (the positive class), focusing on the most present class.

Consequently, in addition to accuracy, the metrics used will be F-measure (along with its Precision and Recall components) and the ROC curve study.

Methods of approaching classification

- Holdout
- K-Fold cross validation
- Classification with feature selection
- SMOTE
- Equal size sample



Dataset

preprocessing

Model building w/ Holdout

Model building con K-Fold Cross.

Performance evaluation

results

Features selection

Model building w/ K-Fold Cross.

Performance evaluation

results

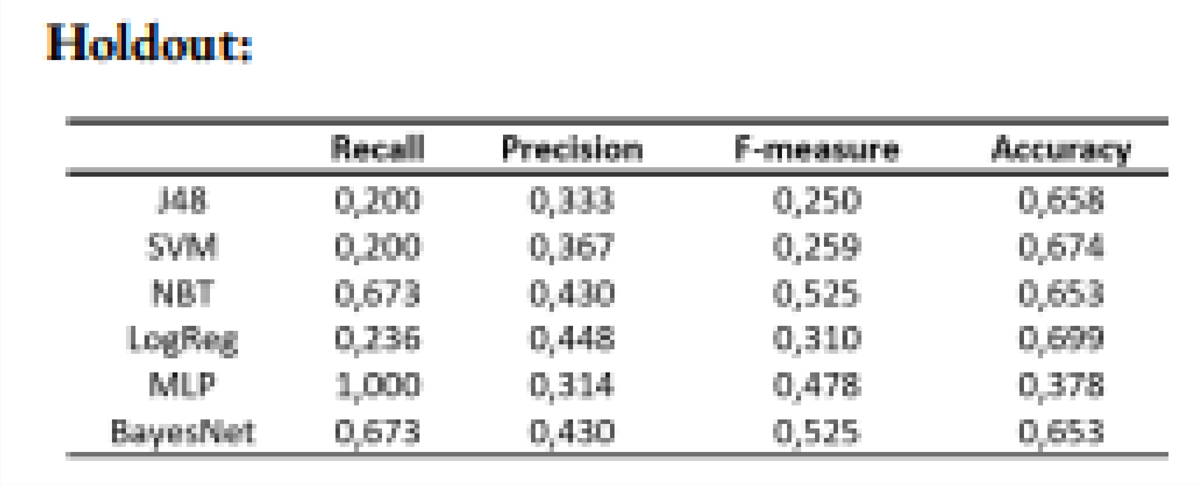
Downsampling / Oversampling

Model building w/ K-Fold Cross.

Performance evaluation

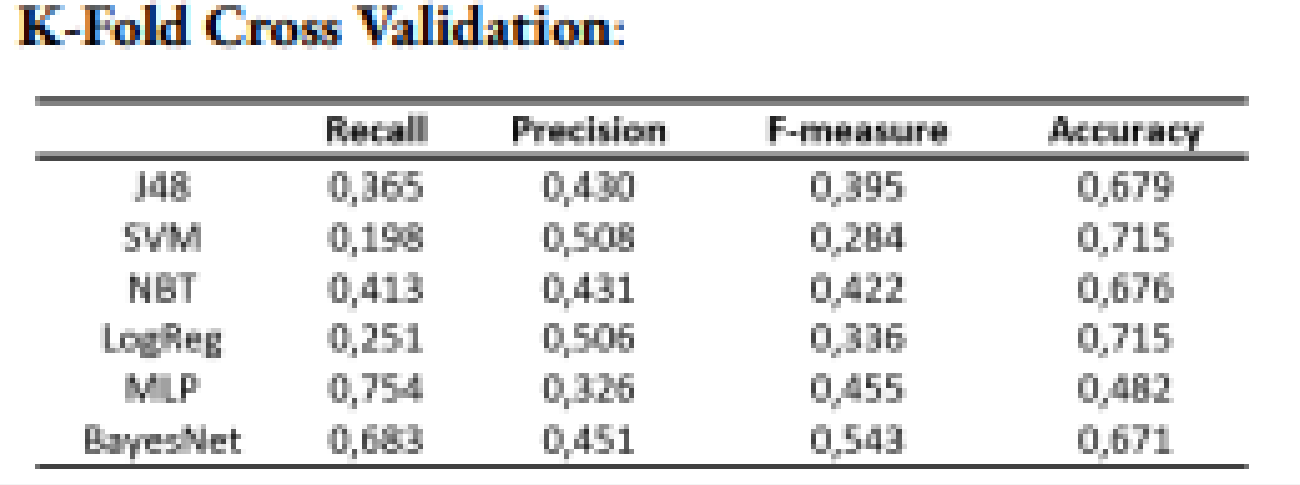
results

Result analysis



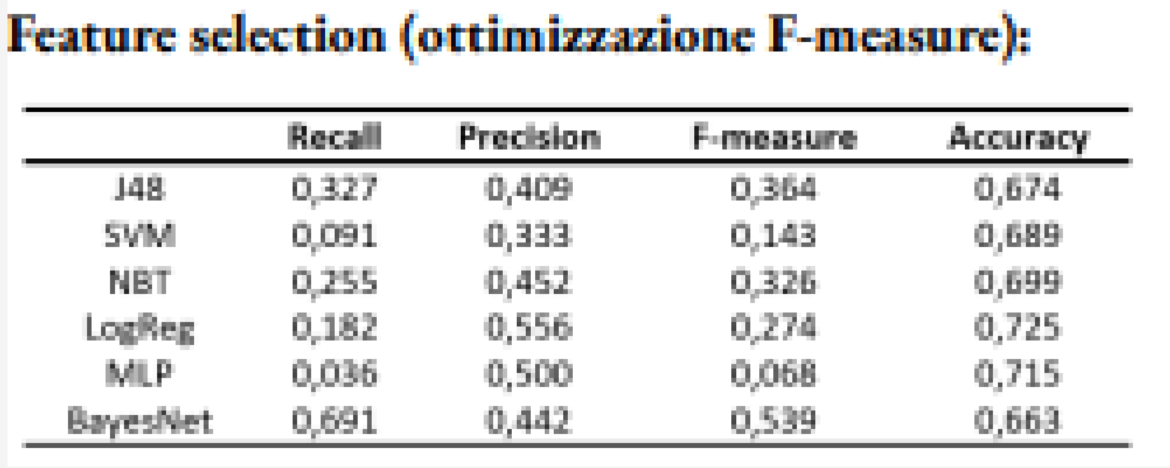
Holdout:

	Recall	Precision	F-measure	Accuracy
J48	0,200	0,333	0,250	0,658
SVM	0,200	0,367	0,259	0,674
NBT	0,673	0,430	0,525	0,653
LogReg	0,236	0,448	0,310	0,699
MLP	1,000	0,314	0,478	0,378
BayesNet	0,673	0,430	0,525	0,653



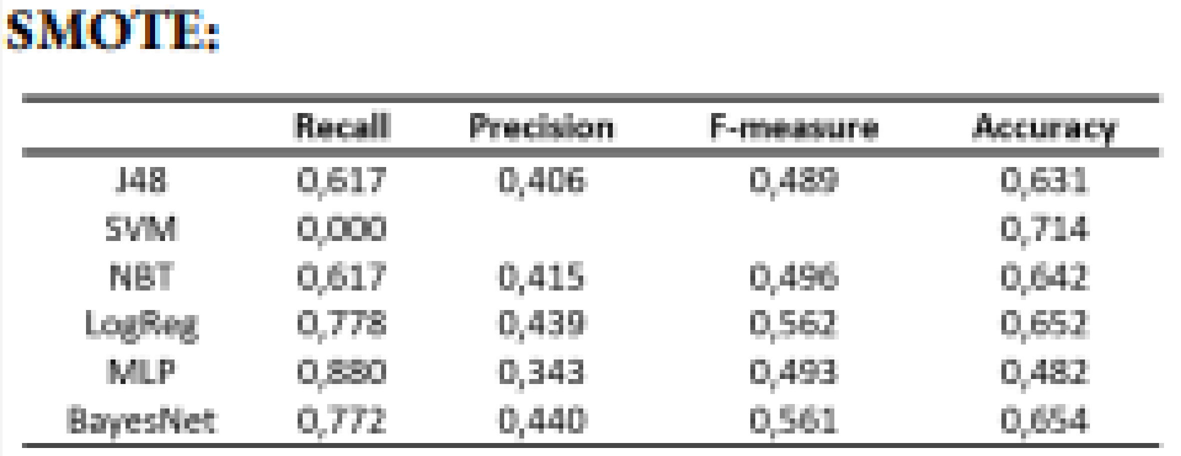
K-Fold Cross Validation:

	Recall	Precision	F-measure	Accuracy
J48	0,365	0,430	0,395	0,679
SVM	0,198	0,508	0,284	0,715
NBT	0,413	0,431	0,422	0,676
LogReg	0,251	0,506	0,336	0,715
MLP	0,754	0,326	0,455	0,482
BayesNet	0,683	0,451	0,543	0,671



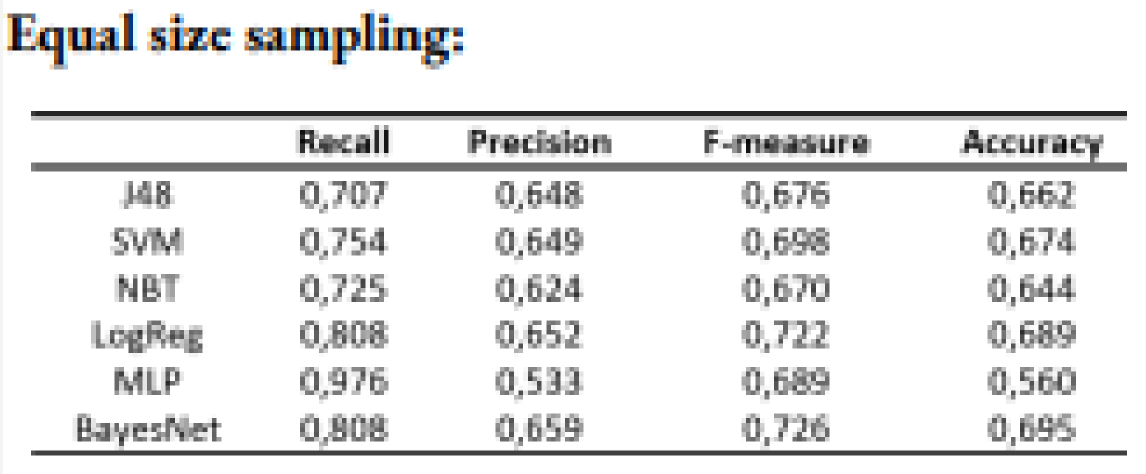
Feature selection (ottimizzazione F-measure):

	Recall	Precision	F-measure	Accuracy
J48	0,327	0,409	0,364	0,674
SVM	0,091	0,333	0,143	0,689
NBT	0,255	0,452	0,326	0,699
LogReg	0,182	0,556	0,274	0,725
MLP	0,036	0,500	0,068	0,715
BayesNet	0,691	0,442	0,539	0,663



SMOTE:

	Recall	Precision	F-measure	Accuracy
J48	0,617	0,406	0,489	0,631
SVM	0,000			0,714
NBT	0,617	0,415	0,496	0,642
LogReg	0,778	0,439	0,562	0,652
MLP	0,880	0,343	0,493	0,482
BayesNet	0,772	0,440	0,561	0,654



Equal size sampling:

	Recall	Precision	F-measure	Accuracy
J48	0,707	0,648	0,676	0,662
SVM	0,754	0,649	0,698	0,674
NBT	0,725	0,624	0,670	0,644
LogReg	0,808	0,652	0,722	0,689
MLP	0,976	0,533	0,689	0,560
BayesNet	0,808	0,659	0,726	0,695

Conclusion

The first phase of the process was devoted to the analysis of the dataset and possible approaches solving with regard to the problem of class unbalanced classes. As recommended in the literature the techniques of F-measure and the ROC curve - in addition to the value accuracy.

Initial results showed that K-Fold Cross Validation was the better method of breakdown (and validation) better than holdout, because the the latter does not guarantee that all instances are included in the training or test set. For this reason, it will then be used as the final breakdown.

In order to optimize the F-measure value the feature selection proved to be somewhat inefficient, the cause could be attributable to the limited size of the database. In addition, the elimination of the -already scarce- features at available could lead to the reduction of information relevant to the algorithm. When considering F-measure as the main parameter, however, the most effective method for handling the unbalanced class in this dataset turns out to be Equal Size Sampling.

In contrast, with the SMOTE method, an overall improvement in accuracy can be seen. Specifically, the behavior of the SVM can be attributable to the fact that the
dataset being contaminated by artificial observations that are not a support for the algorithm.
Finally, the comparison showed that Logistic Regression and Bayesian Network appear to be the best performing classifiers.

Bibliography

[1] - Sumeet K Asrani, Harshad Devarbhavi, John Eaton, Patrick S Kamath. Burden of liver diseases in the world.
[2] - Fahad Mostafa 1, Easin Hasan, Morgan Williamson and Hafiz Khan. Statistical Machine Learning Approaches to Liver Disease Prediction.
[3] - Javad Hassannataj Joloudari, Hamid Saadatfar, Abdollah Dehzangi, Shahaboddin Shamshirband. Computer-aided decision-making for predicting liver disease using PSO-based optimized SVM with feature selection.
[4] - Dorian Pyle. Data Preparation for Data Mining.
[5] - Muzamil Basha, Reva University, Dharmendra Singh Rajput, Ravi Kumar, Bharath Bhushan. Evaluating the Performance of Supervised Classification Models: Decision Tree and Naïve Bayes Using KNIME. International Journal of Engineering and Technology, 7(4):248-253
[6] - Alberto Fernández, Salvador García, Mikel Galar, Ronaldo C. Prati, Bartosz Krawczyk, Francisco Herrera. Learning from Imbalanced Data Sets.
[7] - Massimiliano Morrelli. Addestramento con Dataset Sbilanciati.
[8] - Max Kuhn, Kjell Johnson. Applied predictive modeling.