Assignment 2: *Nested CV on Hepatitis C dataset*

Konstantinos Mexis

Department of Process Engineering, NTUA, Iroon Politechniou 6 Zografou, Athens, Greece

Abstract

Hepatitis C is a significant public health concern, resulting in substantial morbidity and mortality worldwide. Early diagnosis and effective treatment are essential to prevent the disease's progression to chronic liver disease. Machine learning algorithms have been increasingly used to develop predictive models for various diseases, including hepatitis C. In this study, a complete machine learning pipeline was developed to classify individuals as either Hepatitis C patients or healthy blood donors. To achieve this, a *nested cross-validation* pipeline was utilized to systematically compare the performance of multiple classification algorithms on new, unseen data. This pipeline offers a promising alternative method for the diagnosis of Hepatitis C-related diseases.

Keywords: Hepatitis C; machine learning; nested cross validation;

1. Introduction

Hepatitis C is a viral infection that causes the inflammation of the liver. It is caused by the hepatitis C virus (HCV), which is one of the most important global health problems in the world. Worldwide, 350,000 people out of 185 million patients infected with HCV die from diseases caused by HCV. HCV infections pose serious problems on a global scale. Since there is no vaccine yet to prevent HCV infection, it is essential to prevent infection. HCV is a disease that can be difficult to diagnose in its early stages, which can have serious consequences. Early diagnosis and treatment have an important place for the disease. The diagnosis of the disease can be made with the use of machine learning methods. Early detection of patients and people at risk is critical to prevent the spread of HCV infection. The contribution of this study is to create a complete ML pipeline to process the collected data and classify successfully future patients.

2. Methods

2.1. Dataset description

The data set used in this study is a part of the HCV data set in the UCI Machine Learning Repository. The given dataset consists of 204 rows and 12 features. The features are described in the following table:

Feature	Description	Data type
Age	Numerically it is the age value in years	int64
Sex	Female=1, Male=0	int64
ALB	Numerical value of laboratory test data	float64
ALP	Numerical value of laboratory test data	float64
ALT	Numerical value of laboratory test data	float64
AST	Numerical value of laboratory test data	float64
BIL	Numerical value of laboratory test data	float64

CHE	Numerical value of laboratory test data	float64
CHOL	Numerical value of laboratory test data	float64
CREA	Numerical value of laboratory test data	float64
GGT	Numerical value of laboratory test data	float64
PROT	Numerical value of laboratory test data	float64

The target variable is the *label* column. *'label=1'* corresponds to a Hepatitis C patient (positive class) and *'label=0'* corresponds to a healthy blood donor (negative class).

2.2. Data pre-processing

Data pre-processing is very important for the correct operation and high performance of AI algorithms. In the given dataset, there was no missing data. As a result, no data imputation techniques were used in this study.

The given dataset was observed to be imbalanced, with a significant disparity in the number of instances belonging to each class or category. In **Figure 1** *(Supplementary Material*), we can see the plot of the class distribution, which illustrates the imbalanced nature of the dataset with the majority of instances belonging to one class and relatively few instances belonging to the other class.

As a preprocessing step, we performed data normalization on all the numerical features, meaning all the dataset features except the categorical feature '*Sex'*. Data normalization ensures that each feature is on the same scale, reducing the impact of outliers and extreme values. Normalization helps improve the performance and reliability of a machine learning model. In this study, we used the *Standardization Scaling*, subtracting the mean of each observation and diving by the standard deviation.

2.3. Pipeline

In this study, the expected performance of the following classification algorithms was examined in the problem of diagnosis of Hepatitis C patients:

Classification algorithm
Logistic Regression (LR)
Gaussian Naïve Bayes (GNB)
K-Nearest Neighbors (kNN)
Linear Discriminant Analysis (LDA)
Support Vector Machines (SVM)

In order to select the best machine learning model, we built a nested Cross Validation pipeline to systematically compare the performance of the mentioned algorithms.

We performed a **nested Cross Validation**, using a K=5 fold for the outer loop and a L=3 folds for the inner loop. We used a *Stratified k-fold cross-validation* approach to evaluate the performance of the classification algorithms on the dataset. This was necessary because the dataset was observed to be imbalanced. By using a stratified k-fold approach, we ensured that each fold of the cross-validation retained the same class distribution as the original dataset, which helped to prevent bias in our evaluation results.

Nested cross-validation is an approach to model hyperparameter optimization and model selection that attempts to overcome the problem of overfitting the training dataset. The procedure involves treating model hyperparameter optimization as part of the model itself and evaluating it within the broader k-fold cross-validation procedure for evaluating models for comparison and selection. For the evaluation of the algorithms (outer loop), we used the Mathews Correlation Coefficient (MCC). For training and model selection (inner loop) we used the F1 score.

After finding the winner algorithm using nested CV, we used the whole dataset and a simple 5-fold Cross Validation to determine the *final model* that will be deployed in the field. To find the optimal set of hyperparameters for that model, we defined a grid and performed a Randomized Search. For training and model selection we used the Mathews Correlation Coefficient (MCC).

3. Results and Discussion

After performing the nested Cross Validation pipeline, it was found that the *Support Vector Machines* (SVM) algorithm achieved the best performance among the considered methods. The mean score for each model, in terms of the Matthews Correlation Coefficient (MCC), was computed for the 50 nCV folds and reported in **Table 1** of the *Supplementary Material*. Additionally, we generated boxplots of the evaluation metrics over the 50 nCV outer loop folds for all algorithms, which are shown in **Figure 2** of the *Supplementary Material*.

To select the final model for deployment, a 5-fold cross validation was performed on the entire dataset. The resulting model achieved a cross validation score of 0.8674 in terms of MCC. Details of the hyperparameters used for the model can be found in Table 2 of the Supplementary Material. The trained model was saved using the pickle library in Python for future use.

4. Conclusion

The identification of patients at risk for Hepatitis C viral infection is a challenge for the clinicians and public health specialists. The aim of this study was to evaluate and compare the predictive performances of machine learning-based models for the prediction of HCV status using a nested Cross Validation pipeline.

Our results showed that the Support Vector Machines achieved the highest predictive performance over the 50 nCV outer loop folds. SVM increases class separation and reduces expected prediction error and is applicable for the analysis of high-dimensionality data with small sample size [1].

To further improve the prediction accuracy, data augmentation techniques such as the Synthetic Minority Oversampling Technique (SMOTE) could be utilized to handle the imbalance data and increase the size of the dataset [2].

Feature Selection techniques should also be implemented in order to select the most informative features from the dataset. Feature selection techniques, such as Recursive Feature Elimination and SelectKBest, can play a critical role in improving the performance of the classification models. By selecting the most relevant features, one can reduce the noise in the dataset and focus on the most informative signals, leading to better generalization and more accurate predictions.

Finally, explainable AI techniques could also be used to better understand the problem of Hepatitis C diagnosis. he complexity of the hepatitis C data and the presence of numerous confounding variables make it difficult to understand how a model arrives at its decision. Therefore, by using explainable machine learning techniques such as SHAP values, we can identify which features are most critical in making predictions, understand the direction and magnitude of their influence, and potentially reveal previously unknown relationships between the predictors and the response variable. These techniques can provide insights into the underlying biological mechanisms driving the disease and help clinicians make informed decisions based on the model's predictions.

The machine learning-based models could be useful tools for HCV infection prediction and for the risk stratification process of adult patients who undergo a viral hepatitis screening program. These findings are important for clinicians and public health specialists because they can be further validated and incorporated into national screening programs in order to optimize them and to reduce their costs.

References

[1] Knights D., Costello E.K., Knight R. Supervised classification of human microbiota. FEMS Microbiol. Rev. 2011;35:343–359. doi: 10.1111/j.1574-6976.2010.00251.x

[2] Ali, Ali Mohd, Mohammad R. Hassan, Faisal Aburub, Mohammad Alauthman, Amjad Aldweesh, Ahmad Al-Qerem, Issam Jebreen, and Ahmad Nabot. 2023. "Explainable Machine Learning Approach for Hepatitis C Diagnosis Using SFS Feature Selection" Machines 11, no. 3: 391. https://doi.org/10.3390/machines11030391

Supplementary Material

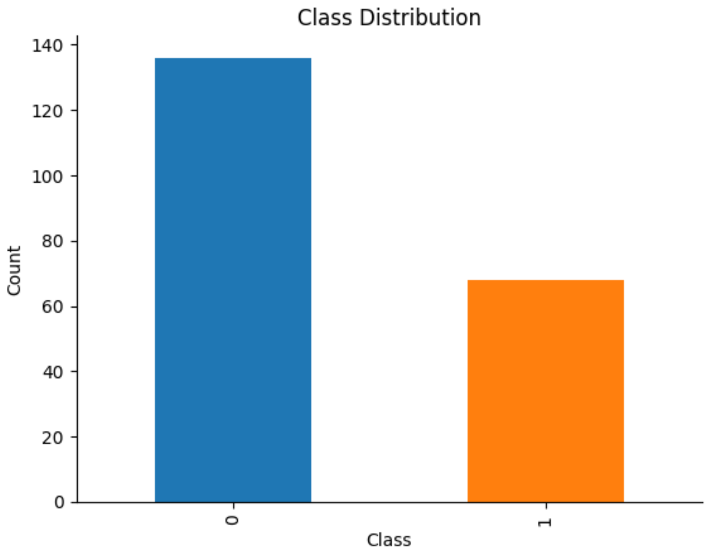

Figure 1: Class distribution for the dataset. Negative Class (label 0 - *healthy blood doners*) accounting for 67% of the instances and Positive Class (label 1 - *Hepatitis C patient*) accounting for the remaining 33%. This suggests that the dataset may pose a challenge for classification tasks

Classification algorithm	**Mean MCC over 50 nCV outer loop folds**
Logistic Regression (LR)	0.82
Gaussian Naïve Bayes (GNB)	0.72
K-Nearest Neighbors (kNN)	0.71
Linear Discriminant Analysis (LDA)	0.72
Support Vector Machines (SVM)	0.84

Table 1: Average MCC score over the 50 nCV outer loop folds for all algorithms

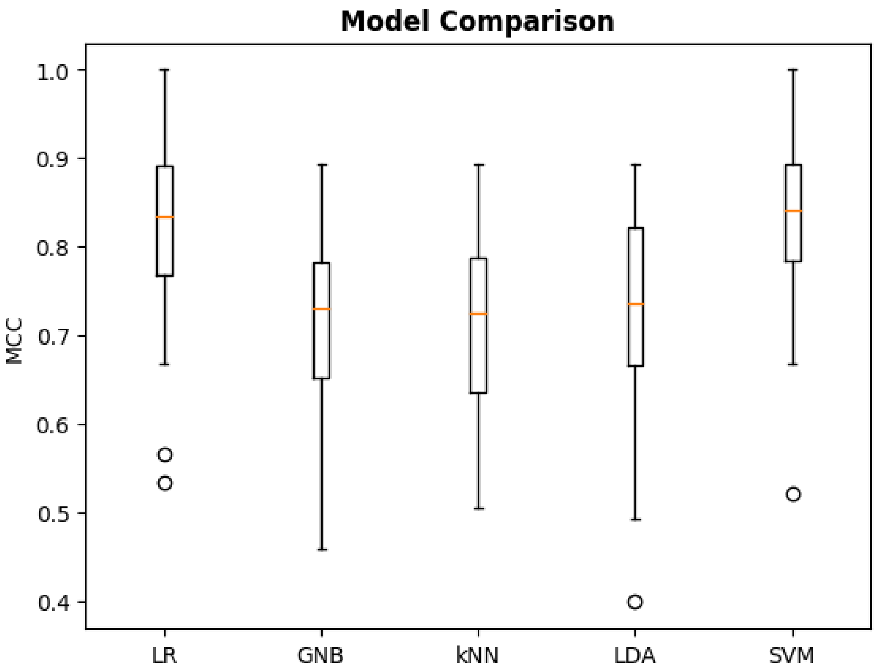

Figure 2: Boxplots of the MCC score over the 50 nCV outer loop folds for all algorithms

Support Vector Machines (SVM)	
Hyperparameter	**Value**
kernel	linear
C	0.4013603603603604
gamma	6.650289333776228

Table 2: Hyperparameter values of the final model to be deployed in the field

Appendix

The analysis presented in this technical report was performed using Python programming language. The code was executed in Jupyter notebooks, which allows for easy documentation of the code and results.

In order to successfully run the notebooks, the following Python libraries are required: pandas, numpy, matplotlib, seaborn, scikit-learn, and tqdm. These can be installed via pip or conda.

To install the required libraries via pip, simply run the following command in the command line: `pip install pandas numpy matplotlib seaborn scikit-learn tqdm`

To install the required libraries via conda, simply run the following command in the command line: `conda install pandas numpy matplotlib seaborn scikit-learn tqdm`

It is recommended to use Python version 3.7 or higher for running the notebooks.