Practical application 1: non-probabilistic supervised classification

Álvaro García Barragán
alvaro.gbarragan@alumnos.upm.es
Universidad Politécnica de Madrid

October 15, 2023

Abstract

This paper proposes a comprehensive study aimed at predicting the survival states of patients diagnosed with liver cirrhosis with non-probabilistic supervised classification models. Data is sourced from a Mayo Clinic study on primary biliary cirrhosis. Objectives include evaluating algorithm performance with original variables, employing feature selection methods, and assessing their impact on classification results.

1 Introduction

Liver cirrhosis is a late stage of scarring (fibrosis) of the liver caused by many forms of liver diseases and conditions, such as hepatitis and chronic alcoholismis [1]. It is currently the 14th most common cause of death among adults worldwide [2].

A comprehensive study is conducted to predict the survival states of liver cirrhosis patients using non-probabilistic supervised classification models. The data provided is sourced from a Mayo Clinic [3] study on primary biliary cirrhosis (PBC) of the liver carried out from 1974 to 1984.

The objectives of this study are the following:

- Evaluate the performance of non-probabilistic classification algorithms, which have been covered in our class curriculum.
- Conduct a comprehensive analysis of the classification algorithms using all original variables as input features.
- Perform a second analysis using a univariate filter feature subset selection method to determine its impact on classification performance.
- Conduct a third analysis using a multivariate filter feature subset selection technique to assess its influence on the algorithms' performance.
- Perform a fourth analysis employing a wrapper feature subset selection method to evaluate its effect on the classification results.

This approach assumes that the factors influencing patient outcomes can be effectively captured and discriminated. In light of this perspective, it seeks to address the following inquiries:

- Which non-probabilistic supervised classification model, among the chosen candidates, demonstrates superior predictive performance when classifying patient outcomes?
- How do the results of the classification approach change when various feature selection methods are applied?
- What are the implications for model accuracy, interpret-ability, and the identification of key predictive factors?

2 Problem description

Liver cirrhosis is a critical medical condition characterized by the scarring of liver tissue, often resulting from chronic liver diseases [1]. Accurate prognosis of patients with liver cirrhosis is of utmost importance for healthcare providers and patients alike [4]. It allows for informed clinical decisions, timely interventions, and improved patient outcomes.

This study aims to address the following research problem: predicting the survival states of patients diagnosed with liver cirrhosis based on a dataset from Mayo Clinic. The survival states include three categories: 0 (D) representing death, 1 (C) indicating censoring, and 2 (CL) denoting censoring due to liver transplantation.

The scope of this study encompasses the development and evaluation of predictive models for survival state prediction in patients with liver cirrhosis. The primary focus will be on exploring the predictive power of these features in determining the survival state of patients.

Limitations and Constraints:

- **Data Availability:** This study relies on the data available in the Mayo Clinic dataset, and any limitations or biases in this data could impact the model's accuracy.
- **Generalizability:** The findings of this study will primarily apply to patients with liver cirrhosis within the context of the Mayo Clinic dataset. Generalizing the results to broader populations may require further validation.
- **Model Complexity:** The complexity of predictive models and their interpretability will be considered. Highly complex models may provide better predictive performance but could be challenging to interpret in a clinical context.

Addressing the prediction of survival states in liver cirrhosis patients is highly significant and relevant for several reasons:

- **Clinical Decision-Making:** Accurate predictions can assist clinicians in tailoring treatment plans and interventions based on a patient's predicted survival state. This personalized approach can improve patient care.
- **Resource Allocation:** Healthcare resources, including liver transplantation, are limited. Predicting which patients are more likely to require such interventions can aid in efficient resource allocation.
- **Research Advancement:**: This study contributes to the broader understanding of predictive modeling in healthcare, demonstrating the applicability of machine learning techniques to clinical problems.

3 Methodology

In this section, we outline the steps to evaluate the non-probabilistic models, encompassing various model configurations and the applied feature selection methods. To begin, subsection 3.1 and subsection 3.2 provides a description of the source data-set, while subsection 3.3 discusses the software tools utilized.

As is customary in the field of Data Science, the data is typically not ready for training every model. Therefore, a preprocessing step is carried out on the dataset, which is described in detail in subsection 3.4.

Finally, the methods employed for feature selection are detailed in subsection 3.5 and the evaluation of the models and a comprehensive description of the experiments presented in subsection 3.6.

3.1 Dataset

The dataset, as documented in [5], contains 418 instances, each representing an individual patient, and includes 17 distinct clinical features. The contents of Table 1 provide information on each variable, including its role, data type, and a brief description.

Table 1: Dataset Variables

Variable	**Role**	**Type**	**Description**
ID	Identifier	Integer	Unique identifier for a patient
N Days	Other	Integer	Number of days between registration and the earlier of death, transplantation, or study analysis time in July 1986
Drug	Feature	Categorical	Type of drug (D-penicillamine or placebo)
Age	Feature	Integer	Age in [days]
Sex	Feature	Categorical	Gender (M: male, F: female)
Ascites	Feature	Categorical	Presence of ascites (N: No, Y: Yes)
Hepatomegaly	Feature	Categorical	Presence of hepatomegaly (N: No, Y: Yes)
Spiders	Feature	Categorical	Presence of spiders (N: No, Y: Yes)
Edema	Feature	Categorical	Presence of edema (N: no edema and no diuretic therapy for edema, S: edema present without diuretics, or edema resolved by diuretics, Y: edema despite diuretic therapy)
Bilirubin	Feature	Continuous	Serum bilirubin in [mg/dl]
Cholesterol	Feature	Continuous	Serum cholesterol in [mg/dl]
Albumin	Feature	Continuous	Albumin in [gm/dl]
Copper	Feature	Continuous	Urine copper in [ug/day]
Alk Phos	Feature	Continuous	Alkaline phosphatase in [U/liter]
SGOT	Feature	Continuous	SGOT in [U/ml]
Triglycerides	Feature	Continuous	Triglycerides in [mg/dl]
Platelets	Feature	Integer	Platelets per cubic [ml/1000]
Prothrombin	Feature	Continuous	Prothrombin time in seconds [s]
Stage	Feature	Integer	Histologic stage of disease (1, 2, 3, or 4)
Status	Target	Categorical	Status of the patient (C: censored, CL: censored due to liver tx, D: death)

3.2 Data Analysis

In this section, an overview of feature distributions is presented after the step of subsection 3.4. As illustrated in Figure 1, a notable gender imbalance is evident, with approximately four times as many females as males in the study cohort. This gender disproportion has the potential to introduce bias into the study's results.

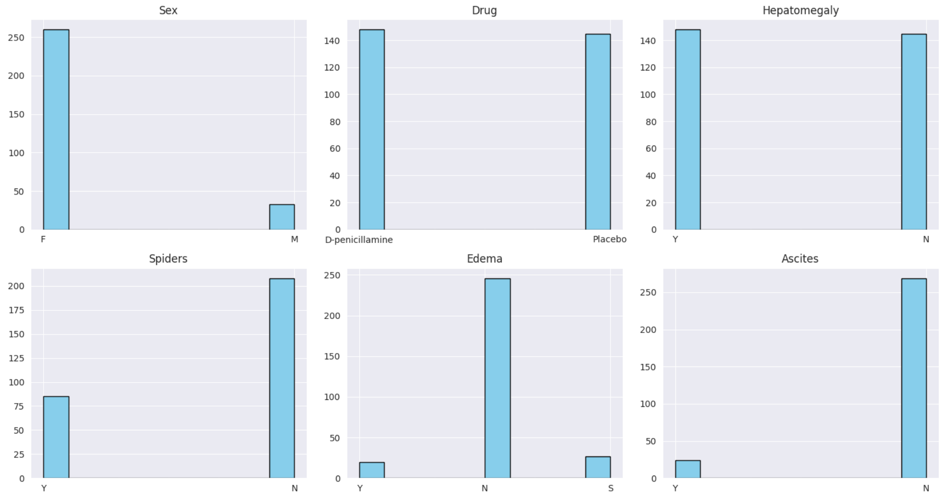

Figure 1: Categorical Features

Additionally, as depicted in Figure 2, the age distribution of study participants predominantly falls within the 40-60 years old range.

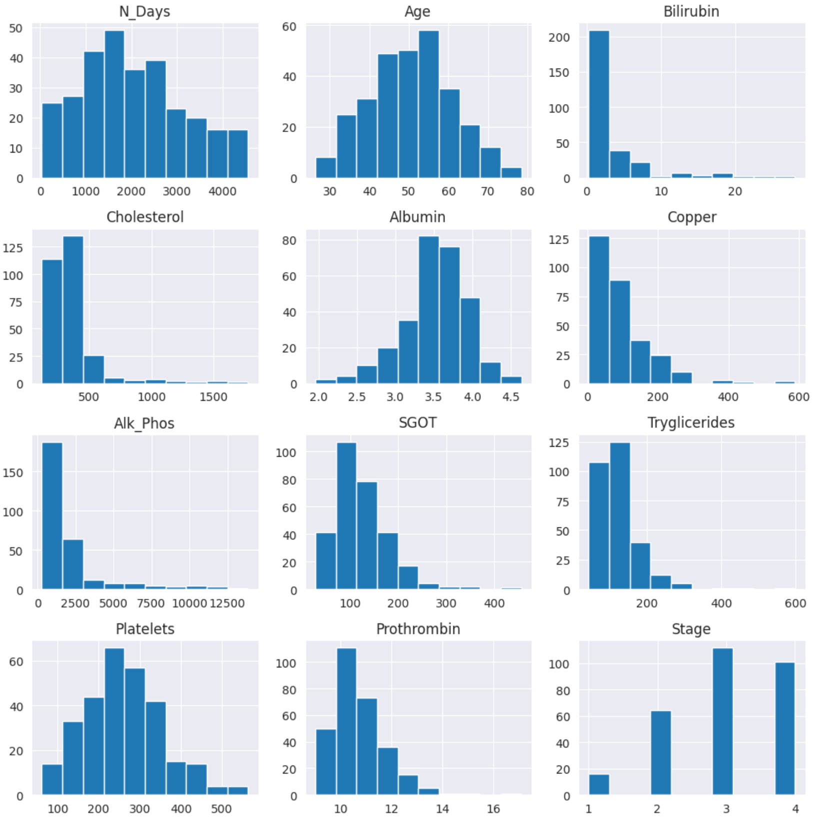

Figure 2: Numerical Features

3.3 Software Tools

The software utilized for the execution of this project was implemented within the Python Jupyter Notebooks environment to systematically document all experimental procedures. The codebase and corresponding experiment records are accessible via the provided repository link: `https://github.com/Alvaro8gb/PracticalApplicationsML`. The primary library utilized is scikit-learn [6].

3.4 Preprocessing

In this section, we outline the steps for obtaining a suitable dataset for model input. This step reduces the number of instances from 418 to 293. The purpose of this preprocessing step is to encode the information in a format that is compatible with all the models and enhances the interpretability of the model. This process involves the following stages:

1. **Remove Metadata Information:** Remove metadata information, such as the ID column, which is not relevant for modeling purposes.

2. **Convert Age to Years:** Convert the age values from days to years for improved model interpretability.

3. **Remove Null Values in Drug Feature:** Eliminate instances with missing values in the ”Drug” feature to maintain data integrity.

4. **Replace Missing Numerical Values:** For numerical features with missing values, apply the imputation method by replacing these missing values with the arithmetic mean of the respective feature.

5. **Class Imbalance Mitigation:** Subsequent to the preceding preprocessing steps, an observation surfaced regarding the class distribution: 168 instances classified as ”CL”, 19 as ”CL” and 125

as "D". Notably, the class labeled as "CL" comprises only 19 instances, a notably insufficient number for robust analysis. Consequently, to address this imbalance and enhance model efficacy, the problem was re-framed as a binary classification task, assigning a value of 0 to patients who are censored and 1 to those who have death.

6. **One-Hot Encoding for Categorical Features:** Perform one-hot encoding on categorical features to convert them into a numerical format suitable for modeling purposes.

3.5 Feature selection

In this section, we expound upon the process of generating experimental datasets designated for evaluation. After conducting the one-hot encoding procedure, as elucidated in subsection 3.4, a total of 25 features are acquired. Subsequently, within this comprehensive feature set, we have introduced three distinct methods for the extraction of the ten most relevant features, as depicted in Table 2.

Table 2: Feature Selection Methods

Method	Description	Selected Features
No Filter	All original variables	Age, Albumin, Alk-Phos, Ascites-N, Ascites-Y, Bilirubin, Cholesterol, Copper, Drug-D-penicillamine, Drug-Placebo, Edema-N, Edema-S, Edema-Y, Hepatomegaly-N, Hepatomegaly-Y, N-Days, Platelets, Prothrombin, SGOT, Sex-F, Sex-M, Spiders-N, Spiders-Y, Stage, Tryglicerides
Univariate Filter	Analysis of Variance (ANOVA) [7]	Albumin, Ascites-N, Ascites-Y, Bilirubin, Copper, Hepatomegaly-N, Hepatomegaly-Y, N-Days, Prothrombin, Stage
Multivariant Filter	Relief Algorithm [8, 9]	Bilirubin, Copper, Hepatomegaly-N, Hepatomegaly-Y, N-Days, Prothrombin, SGOT, Spiders-N, Spiders-Y, Stage
Wrapper	Sequential Feature Selector [10], forward direction, with a Random Forest Classifier [11]	Ascites-N, Ascites-Y, Bilirubin, Edema-N, Edema-S, Edema-Y, Hepatomegaly-N, Hepatomegaly-Y, Sex-F, Sex-M

3.6 Model Evaluation and Experiments

For each final dataset derived from the feature selection process outlined in subsection 3.5, a 70-30 hold-out validation procedure will be employed. In this procedure, 70% of the dataset will be utilized for training, while the remaining 30% will be reserved for evaluation. This division will remain consistent across all experiments.

Each type of model exhibits distinct characteristics that necessitate different parameter configurations. While the primary objective of this project is not to obtain the optimal configuration, in order to ensure a fair comparison, model parameters and hyperparameters have been selected through a 5-fold cross-validation, hereafter referred to as *GridSearchCV*, using the dataset with all variables.

The following shows the configurations considered for each model:

- **Multilayer Perceptron (MLP):**
  - 3-Fold Cross Validation, measure of f1-score, comparing the number of neurons in the intermediate layer: 2, 10, 18, 26, 34, 42 (keeping hyper-parameters fixed). The experiments show the more number of neurons the more f1-score.
- **Support Vector Machine (SVM):**
  - Default configuration (due to limited hyper-parameters).

- **Repeated Incremental Pruning to Produce Error Reduction (RIPPER):**
  - GridSearchCV with the following parameter grid:
    - prune-size: 0.1, 0.3, 0.5,
    - k: 2, 3, 4,
    - alpha: 0.1, 1.0, 2.0,
    - n-discretize-bins: 10, 15, 20
- **k-Nearest Neighbors (k-NN):**
  - GridSearchCV with parameters: k ranging from 1 to 100 with a step size of 5.
- **Decision Tree (DT):**
  - GridSearchCV with the following parameter grid:
    - max-depth: None, 4, 8, 10,
    - min-samples-split: 2, 8, 16, 32,
    - min-samples-leaf: 2, 8, 16, 32, 48,
    - max-features: None, 'sqrt', 'log2'

The definitive configuration chosen for conducting the experiments is presented in Table 3.

Table 3: Models and Configurations

Model	**Configurations**	**Python Package**
MLP	3 layers (input linear layer, 42 neurons ReLU layer, output linear layer), learning-rate=1e-3, lambda=0.01, epochs=1000, batch-size=128	tensorflow, keras.layers.Dense
SVM	kernel=linear, C=1.0	sklearn.svm.SVM
RIPPER	prune-size=0.33, k=2, alpha=1.0, n-discretize-bins=10	wittgenstein.RIPPER
DT	max-depth=4, max-features=sqrt, min-samples-leaf=2, min-samples-split=8	sklearn.tree.DecisionTreeClassifier
k-NN	k=80, metric=minkowski, p=2	sklearn.neighbors.KNeighborsClassifier

4 Results

In this section, we present the results of our experiments using a 70-30 holdout method explained in subsection 3.6. This approach led to the creation of 16 different models, comprising 4 types of datasets and 4 types of models. To make it easier to digest, we have divided this section into 4 subsections, with each subsection corresponding to one of the dataset types.

Within each experiment, we provide statistics that include accuracy, recall, and F1-score for each class as well as overall statistics for the model. These overall statistics include the corresponding macro average and weighted average, along with the support values. We emphasize the selection of the F1-score as a metric in medicine due to its ability to strike a balance between precision and recall, as well as its significance in reducing false negatives.

4.1 All features data-set

In this subsection, we present the outcomes of the initial experiment, which involves using a dataset containing all variables. The results are displayed in the following tables:

Table 4: MLP all-features report

	precision	recall	f1-score	support
Censored	0.77	0.72	0.74	50
Death	0.66	0.71	0.68	38
accuracy			0.72	88
macro avg	0.71	0.72	0.71	88
weighted avg	0.72	0.72	**0.72**	88

Table 5: SVM all-features report

	precision	recall	f1-score	support
Censored	0.86	0.84	0.85	50
Death	0.79	0.82	0.81	38
accuracy			0.83	88
macro avg	0.83	0.83	0.83	88
weighted avg	0.83	0.83	**0.83**	88

Table 6: RIPPER all-features report

	precision	recall	f1-score	support
Censored	0.71	0.94	0.81	50
Death	0.86	0.50	0.63	38
accuracy			0.75	88
macro avg	0.79	0.72	0.72	88
weighted avg	0.78	0.75	**0.73**	88

Table 7: DT all-features report

	precision	recall	f1-score	support
Censored	0.77	0.88	0.82	50
Death	0.81	0.66	0.72	38
accuracy			0.78	88
macro avg	0.79	0.77	0.77	88
weighted avg	0.79	0.78	**0.78**	88

Table 8: k-NN all-features report

	precision	recall	f1-score	support
Censored	0.75	0.94	0.83	50
Death	0.88	0.58	0.70	38
accuracy			0.78	88
macro avg	0.81	0.76	0.77	88
weighted avg	0.80	0.78	**0.77**	88

4.2 Uni-variant filter data-set

In this subsection, we showcase the results of the second experiment, which employs a uni-variant filter method resulting in a dataset consisting of 10 variables. You can find the results in the tables below:

Table 9: MLP uni-variant report

	precision	recall	f1-score	support
Censored	0.78	0.92	0.84	50
Death	0.86	0.66	0.75	38
accuracy			0.81	88
macro avg	0.82	0.79	0.80	88
weighted avg	0.82	0.81	**0.80**	88

Table 10: SVM uni-variant report

	precision	recall	f1-score	support
Censored	0.78	0.92	0.84	50
Death	0.86	0.66	0.75	38
accuracy			0.81	88
macro avg	0.82	0.79	0.80	88
weighted avg	0.82	0.81	**0.80**	88

Table 11: RIPPER uni-variant report

	precision	recall	f1-score	support
Censored	0.76	0.94	0.84	50
Death	0.88	0.61	0.72	38
accuracy			0.80	88
macro avg	0.82	0.77	0.78	88
weighted avg	0.81	0.80	**0.79**	88

Table 12: DT uni-variant report

	precision	recall	f1-score	support
Censored	0.81	0.76	0.78	50
Death	0.71	0.76	0.73	38
accuracy			0.76	88
macro avg	0.76	0.76	0.76	88
weighted avg	0.76	0.76	**0.76**	88

Table 13: k-NN uni-variant report

	precision	recall	f1-score	support
Censored	0.71	0.98	0.82	50
Death	0.95	0.47	0.63	38
accuracy			0.76	88
macro avg	0.83	0.73	0.73	88
weighted avg	0.81	0.76	**0.74**	88

4.3 Multi-variant filter data-set

In this subsection, we showcase the results of the third experiment, which employs a multi-variant filter method resulting in a dataset consisting of 10 variables. You can find the results in the tables below:

Table 14: MLP multi-variant report

	precision	recall	f1-score	support
Censored	0.72	0.86	0.78	50
Death	0.75	0.55	0.64	38
accuracy			0.73	88
macro avg	0.73	0.71	0.71	88
weighted avg	0.73	0.73	**0.72**	88

Table 15: SVM multi-variant report

	precision	recall	f1-score	support
Censored	0.69	0.82	0.75	50
Death	0.69	0.53	0.60	38
accuracy			0.69	88
macro avg	0.69	0.67	0.67	88
weighted avg	0.69	0.69	**0.69**	88

Table 16: RIPPER multi-variant report

	precision	recall	f1-score	support
Censored	0.73	0.90	0.80	50
Death	0.81	0.55	0.66	38
accuracy			0.75	88
macro avg	0.77	0.73	0.73	88
weighted avg	0.76	0.75	**0.74**	88

Table 17: DT multi-variant report

	precision	recall	f1-score	support
Censored	0.72	0.76	0.74	50
Death	0.66	0.61	0.63	38
accuracy			0.69	88
macro avg	0.69	0.68	0.68	88
weighted avg	0.69	0.69	**0.69**	88

Table 18: k-NN multi-variant report

	precision	recall	f1-score	support
Censored	0.71	0.98	0.82	50
Death	0.95	0.47	0.63	38
accuracy			0.76	88
macro avg	0.83	0.73	0.73	88
weighted avg	0.81	0.76	**0.74**	88

4.4 Wrapper subset filter data-set

In this subsection, we present the outcomes of the ultimate experiment, where we utilized a wrapper feature selection method to create a dataset containing 10 variables. The results are available in the tables below:

Table 19: MLP wrapper report

	precision	recall	f1-score	support
Censored	0.69	0.90	0.78	50
Death	0.78	0.47	0.59	38
accuracy			0.72	88
macro avg	0.74	0.69	0.69	88
weighted avg	0.73	0.72	**0.70**	88

Table 20: RIPPER wrapper report

	precision	recall	f1-score	support
Censored	0.74	0.90	0.81	50
Death	0.81	0.58	0.68	38
accuracy			0.76	88
macro avg	0.78	0.74	0.74	88
weighted avg	0.77	0.76	**0.75**	88

Table 21: SVM wrapper report

	precision	recall	f1-score	support
Censored	0.69	0.82	0.75	50
Death	0.69	0.53	0.60	38
accuracy			0.69	88
macro avg	0.69	0.67	0.67	88
weighted avg	0.69	0.69	**0.69**	88

Table 22: DT wrapper report

	precision	recall	f1-score	support
Censored	0.72	0.76	0.74	50
Death	0.66	0.61	0.63	38
accuracy			0.69	88
macro avg	0.69	0.68	0.68	88
weighted avg	0.69	0.69	**0.69**	88

Table 23: k-NN wrapper report

	precision	recall	f1-score	support
Censored	0.71	0.98	0.82	50
Death	0.95	0.47	0.63	38
accuracy			0.76	88
macro avg	0.83	0.73	0.73	88
weighted avg	0.81	0.76	**0.74**	88

4.5 Model Visualization

In the following section, we present visualizations depicting the training process of models that utilize all available variables, as discussed in subsection 4.1. Specifically, we focus on the Decision Tree and RIPPER model, as these are the only models with interpretability. Moreover, the MNN can provide a form of interpretation by revealing how much of each feature passes to the next layer when examining the input layer, as show in Figure 5.

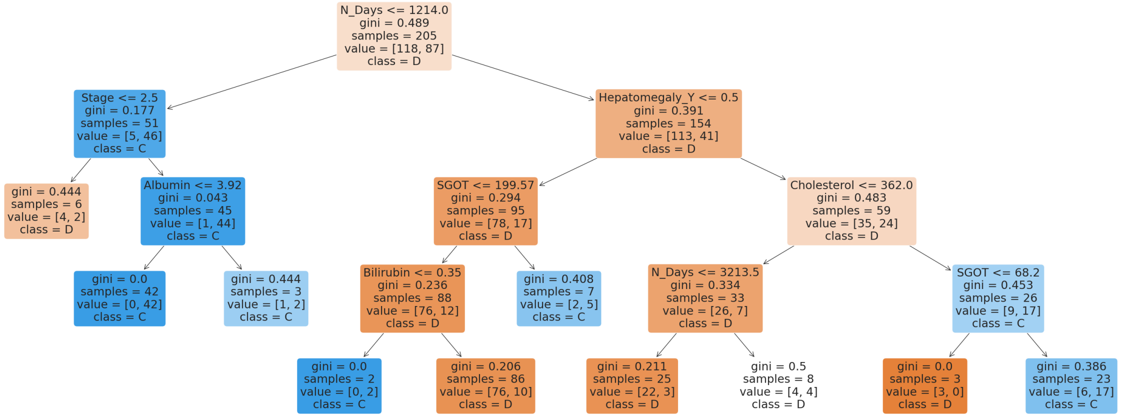

Figure 3: Decision Tree

The primary feature of significance for the tree, as depicted in Figure 3, is N_Days, followed by SGOT. The blue nodes correspond to class C (Censored), while the orange nodes denote class D (Death). Furthermore, in the rules of the RIPPER model Figure 4, it is demonstrated which conditions determine the classification of a patient as class D.

```
[[Hepatomegaly_N=False ^ Bilirubin=>7.16] V
[Hepatomegaly_N=False ^ N_Days=602.6-999.8] V
[N_Days=<602.6] V
[SGOT=>198.4 ^ Copper=104.6-141.8] V
[Prothrombin=11.5-12.0]]
```

Figure 4: RIPPER model, when positive == Death (D)

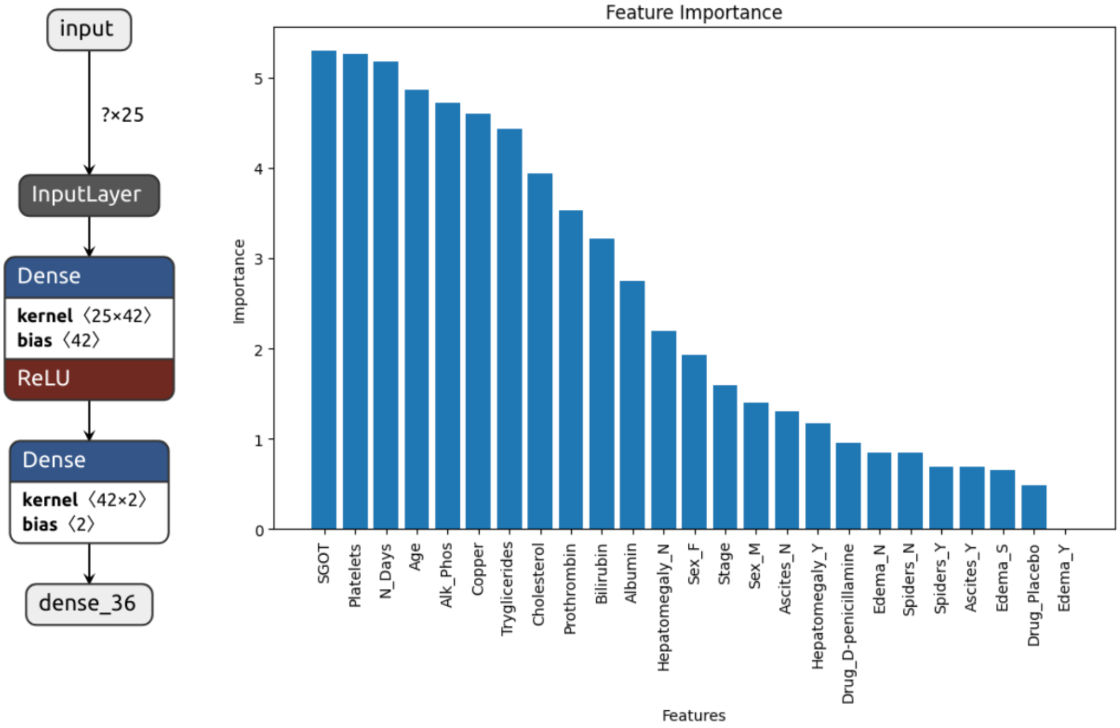

Figure 5: MNN first layer

5 Discussion

The model visualization indicates that the SGOT feature plays a critical role in the classification decision. Elevated SGOT levels can serve as an indicator of potential liver damage because the liver cells release SGOT into the bloodstream, leading to its association with liver-related issues [12].

The best results are achieved when employing all 25 features, with the SVM model delivering the highest score, 0.83 of weighed avg f1-score.

The k-NN model consistently maintains the F1-score value across all feature selection methods.

It's crucial to highlight that using only 10 features leads to reduced classification performance but for the MLP model, the univariate filter improves its F1-score by 0.08 and the RIPPER model sees an enhancement of 0.06.

We illustrate that Decision Trees and RIPPER models offer a high level of comprehensibility, especially within the medical domain, where model interpretability is a crucial factor to consider.

However, this project exhibits certain limitations. The determination of the optimal model configuration is highly dependent on the specific characteristics of the dataset. Employing a uniform configuration across all types of datasets may not yield optimal results. Furthermore, demonstrating the statistical superiority of one model over another requires rigorous testing beyond what has been presented.

6 Conclusion

This project demonstrates the predictability of biliary cirrhosis using non-probabilistic classifier models. It reveals that SVM yields the highest F1-score, when using all the features during a 70-30 hold-out. Additionally, the experiments illustrate that altering the feature selection algorithm can either enhance or diminish the model's performance.

Finally, it underscores the significance of the SGOT feature in patient outcome classification and emphasizes the crucial role of model interpretation in medical domain context.

References

[1] M. Clinic, “Cirrhosis - symptoms and causes,” 2023. https://www.mayoclinic.org/es/diseases-conditions/cirrhosis/symptoms-causes/syc-20351487.

[2] Y. Kim, K. Kim, and I. Jang, “Analysis of mortality prognostic factors using model for end-stage liver disease with incorporation of serum-sodium classification for liver cirrhosis complications: A retrospective cohort study,” *Medicine*, vol. 98, no. 45, 2019.

[3] M. Clinic, 2023. https://www.mayoclinic.org.

[4] Mayo Clinic, “Cirrhosis clinical trials,” 2023. https://www.mayo.edu/research/clinical-trials/diseases-conditions/cirrhosis/.

[5] D. E., G. P., F. T., F. L., and L. A., “Cirrhosis patient survival prediction.” UCI Machine Learning Repository, 2023. https://doi.org/10.24432/C5R02G.

[6] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, *et al.*, “Scikit-learn: Machine learning in python,” *Journal of machine learning research*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[7] L. St, S. Wold, *et al.*, “Analysis of variance (anova),” *Chemometrics and intelligent laboratory systems*, vol. 6, no. 4, pp. 259–272, 1989.

[8] K. Kira and L. A. Rendell, “The feature selection problem: Traditional methods and a new algorithm,” in *Proceedings of the tenth national conference on Artificial intelligence*, pp. 129–134, 1992.

[9] R. J. Urbanowicz, R. S. Olson, P. Schmitt, M. Meeker, and J. H. Moore, “Benchmarking relief-based feature selection methods.” arXiv e-print. https://arxiv.org/abs/1711.08477, 2017.

[10] I. Guyon and A. Elisseeff, “An introduction to variable and feature selection,” *Journal of machine learning research*, vol. 3, no. Mar, pp. 1157–1182, 2003.

[11] L. Breiman, “Random forests,” *Machine learning*, vol. 45, pp. 5–32, 2001.

[12] J. A. Cohen and M. M. Kaplan, “The sgot/sgpt ratio—an indicator of alcoholic liver disease,” *Digestive diseases and sciences*, vol. 24, pp. 835–838, 1979.