Practical application 2: probabilistic supervised classification

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

November 21, 2023

Abstract

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

1 Introduction

The data provided are sourced from a Mayo Clinic [1] study on primary biliary cirrhosis (PBC) of the liver carried out from 1974 to 1984. The dataset is a multivariate dataset [2], which, after preprocessing, comprises 293 instances, each corresponding to an individual patient. It encompasses 25 distinct clinical features offering insights into demographics, comorbidities, treatment, and various other factors. The target class has two possible values: 0 (Ceased) and 1 (Death), with frequencies of 168 and 125, respectively. The study aims to achieve the following objectives:

- Evaluate the performance of probabilistic classification algorithms, for binary classification.
- 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.
- Analyze interpretability of probabilistic supervised classifiers.
- Finally, the above models are compared with those not based on probabilities.

2 Methodology

In this section, the problem of evaluating the models is described in subsection 2.1, the FSS methods are described in subsection 2.2, the explanation of the model selection subsection 2.3 and the software tools used are detailed in subsection 2.4. The proposed solution is introduced in subsection 2.5. Additionally, a comprehensive description of the experiments are presented in subsection 2.6. Finally, the importance of model interpretability is underscored in subsection 2.7.

Table 1: Data-set Analysis

Feature	**Mean**	**Std**	**Min**	**Max**	**Type**
Age	50.6271	10.5698	26.2958	78.4931	continuous
Albumin	3.5169	0.4229	1.96	4.64	continuous
Alk_Phos	2011.6709	2195.9565	289.0	13862.4	continuous
Ascites_N	0.918089	0.274699	0.0	1.0	binary
Ascites_Y	0.081911	0.274699	0.0	1.0	binary
Bilirubin	3.264164	4.648182	0.3	28.0	continuous
Cholesterol	365.2108	212.7557	120.0	1775.0	continuous
Copper	95.9395	84.2078	4.0	588.0	continuous
Drug_D-penicillamine	0.5051	0.5008	0.0	1.0	binary
Drug_Placebo	0.4948	0.5008	0.0	1.0	binary
Edema_N	0.8395	0.3676	0.0	1.0	binary
Edema_S	0.0921	0.2897	0.0	1.0	binary
Edema_Y	0.0682	0.2526	0.0	1.0	binary
Hepatomegaly_N	0.4948	0.5008	0.0	1.0	binary
Hepatomegaly_Y	0.5051	0.5008	0.0	1.0	binary
N_Days	2038.6655	1137.3298	41.0	4556.0	continuous
Platelets	259.5417	95.5279	62.0	563.0	continuous
Prothrombin	10.7488	1.0219	9.0	17.1	continuous
Sex_F	0.8873	0.3166	0.0	1.0	binary
Sex_M	0.1126	0.3166	0.0	1.0	binary
SGOT	122.0661	57.7574	26.35	457.25	continuous
Spiders_N	0.70989	0.4545	0.0	1.0	binary
Spiders_Y	0.2901	0.4545	0.0	1.0	binary
Stage	3.0170	0.8854	1.0	4.0	nominal
Tryglicerides	124.1343	61.5558	44.0	598.00	continuous

2.1 Evaluation problem

The evaluation of machine learning models stands as a pivotal aspect in Data Science, given the multiple configurations that can be applied. Comparing two models trained with different datasets, whether involving distinct features or preprocessing methods, presents a challenge in creating a pipeline that can robustly compare models employing various types of feature selection. The vast space of possible configurations introduces complexities. The questions within this space are:

- How many features will the Feature Subset Selection (FSS) select?
- What methods of univariate and multivariate FSS should be employed?
- How much data should be utilized for wrapper FSS?
- Which hyperparameters do we intend to select for the model? For Artificial Neural Networks (ANN), how many layers, what type of layers, and the number of neurons per layer?
- What constitutes the most equitable validation approach for comparing models with different subsets of the data-set?
- How should the data be partitioned to facilitate the selection of the best configuration in tuning parameters, evaluating models, and choosing the most effective features?
- All the models can be compared with the same preprocesing?

To address these challenges, I propose a simplified scenario to ensure fair evaluation, as depicted in subsection 2.5. This scenario is built on the following assumptions:

1. Models can only be compared when using the same dataset, with an identical number of features and instances except wrapper.

2. Given that FSS with wrapper methods involve training the model using a portion of the data and subsequently selecting features with the testing proportion, we can compare wrapper models even when utilizing different datasets beacuse all model have the oportunity to select their best features. However, it is essential to maintain consistency by employing the same split of the dataset for wrapper selection in every model. This ensures that each model starts from the same initial point.

2.2 Feature Subset Selection

The process of selecting an appropriate data subset is crucial for a model's learning capability. Models have diverse characteristics that may render certain methods more effective than others. In this paper, we have selectively identified one exemplary method for each type of feature selection taking into account their robustness and versatility for various modeling contexts:

- **Univariate Filtering**: We employ the Mutual Information method [3], which quantifies the amount of information gained about one random variable through observing another. This method assesses the dependency between each feature and the target variable independently, which is particularly useful for identifying individual features that have the strongest correlations with the outcome.
- **Multivariate Filtering**: We use the Relief algorithm [4]. This method extends beyond the scope of univariate methods by considering the interaction between features. It is adept at detecting subtle feature interactions, making it valuable for scenarios where the predictive power comes from the synergy between features rather than isolated effects.
- **Wrapper-Based Filtering**: The Sequential Feature Selector (SFS) [5] is implemented as our wrapper method. SFS is a search strategy that systematically assesses model performance across different feature combinations to identify the optimal subset in a greedy fashion. Our approach adopts a backward elimination strategy, beginning with the full set of features and iteratively removing variables until the best-performing set of k features is determined. It will select the best subset of features that maximizes the F-score, over the 5 stratified folds, in Algorithm 1 is depicted the process.

Algorithm 1 Wrapper Method for Feature Selection

```
Input: Dataset $(X, y)$, Model $(model)$, Number of Features $(n\_features)$
Output: Selected Data Set $X\_selected\_wrapper$
Initialize SequentialFeatureSelector($model$, $n\_features$, scoring="f1", direction="backward", CV=5)
Fit the selector with Dataset $(X, y)$
Transform $X$ using the selector to obtain $x\_selected\_wrapper$
Return $X\_selected\_wrapper$
```

2.3 Model Selection

In this section, its explain the base estimators of the meta-classifiers. The decision tree is chosen as the estimator in the Bagging ensemble model due to its inherent instability. In the stacking ensemble model, Support Vector Machine (SVM) and Decision Tree serve as base classifiers, leveraging their strengths in capturing complex patterns, while Logistic Regression is selected as the final classifier for its simplicity and effectiveness in combining outputs. The Voting Model combines SVM, Logistic Regression, and RandomForest to capitalize on the unique strengths of each classifier and mitigate their individual weaknesses. This ensemble strategy aims to improve overall performance by integrating diverse modeling capabilities.

2.4 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 libraries utilized are [6, 7, 8].

2.5 Evaluation

In this section, we explain the pipeline through which the models are evaluated. This pipeline represents a simplified version of the complex space, taking into account the following statements:

- The FSS only consider the best 15 features, due to high computational cost, this number is selected making a Principal Component Analysis. For more details see source code.
- The evaluation is configured with 5 folds, and the 7-fold option is excluded due to the dataset size, as it would result in a test set with only 15 samples, which appears to be insufficient.
- The models chosen for the experiments serve as representatives of the taxonomy elucidated in class, with the exception of those for which I did not find Python implementations, such as TAN or Naive Bayes Tree.
- The hyperparameters grid is estimated in a random homogeneous manner.

The automated pipeline (Algorithm 2) takes as input a binary classifier model, each with its parameter grid, and the dataset. It then outputs the mean and standard deviation of the F-score and Brier score for each fold.

Algorithm 2 Pipeline to evaluate models

```
Input: Dataset $(X, y)$, Model($model$), Parameter Grid ($params\_grid$)
Output: Model Evaluation Metrics ($FScore$ and $BrierScore$)

Initialize $f1\_score$ and $brie\_score$ lists
Prepare Stratified K-Fold Cross-Validator with $K$ splits

for each train-test split in K-Fold Cross-Validator do
    Split data into $X\_train, X\_test$ and $y\_train, y\_test$
    Perform Grid Search with cross-validation on $X\_train, y\_train$, over $model$ and $params\_grid$
    Select the best model based on the Grid Search
    Train the best model on $X\_train, y\_train$
    Make predictions on $X\_test$
    Evaluate predictions to compute $FScore$ and $BrierScore$
    Append scores to the respective lists ($f1\_score$ and $brie\_score$)
end for
Compute and return the arithmetic mean and the standard deviation of $FScore$ and $BrierScore$
```

2.6 Experiments

For each final dataset derived from the feature selection process outlined in Table 3, a 5 Cross-Validation validation procedure will be employed as explained in subsection 2.5. The chosen metric for evaluating the model is the F-score (see Equation 3). This choice is motivated by the presence of class imbalance, as the F-score enables a balanced assessment of both false positives and false negatives. Furthermore, in the realm of medicine, where precise probability estimation is crucial, model calibration becomes paramount. To assess the accuracy of these probability estimates, we utilize the Brier Score (refer to Equation 4). Lower Brier Scores indicate better-calibrated models, making this metric valuable in evaluating the reliability of probability estimates generated by the model.

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, with the training set of the upper validation and scoring F-score. The configuration of every model are displayed in Table 2.

$$\textbf{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{1}$$

$$\textbf{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{2}$$

$$\textbf{F-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

$$\textbf{Brier-score} = \frac{1}{N}\sum_{i=1}^{N}(p_i - o_i)^2 \tag{4}$$

Where:

- N is the number of samples.
- p_i is the predicted probability of the positive class for the i-th sample
- o_i is the actual binary outcome for the i-th sample (0 or 1).

2.7 Model Visualization and Interpretability

In the following section, we present visualizations depicting intrepretability of probabilistics of models. This models are trainned with all the data and with all the features. In a multivariate dataset with more than three features, visualizing boundary regions becomes impractical. The only viable approach for interpreting the model is to examine feature importance. Except for models like Decision Trees and RIPPER, whose decision-making processes can be traced and understood.

Logistic Regression, as illustrated in Figure 1, allows for the examination of feature importance by showcasing the coefficients of the model. A feature with a coefficient close to 0 indicates that the model does not consider the feature significant for decision-making. In this particular instance, it is evident that `Age` holds a positive importance, implying that as age increases, the likelihood of the model predicting death also increases. On the other hand, a high `Platelets` coefficient suggests a lower likelihood of being classified as deceased (Death).

Gaussian Naive Bayes also exhibits a form of interpretability. In this instance, we provide two examples wherein we take a single patient, alter one feature within the range from the minimum to the maximum, and demonstrate the resulting changes in probabilities, while fixing the other features. Figure 2 illustrates that from an `SGOT` level of 200, the models begin to incrementally raise the probability of Death (class 1) in a Gaussian-like pattern. Finally, in Figure 3 display the `Bilirubin` feature, which exhibits a behavior similar to the one observed previously but with a steeper curve.

3 Results

In this section, it is show the results of evaluating the models with 4 diferents FSS: All the features, univariate subset, multivariate subset and wrapper subset. A fair comparison can only be established for each subset due to differences in the data. This results are displayed in Table 4.

4 Discussion

In this section, we discuss the results in three ways: the subsets selected for the FSS (subsection 4.1), model performance (subsection 4.2), and model interpretability (subsection 4.3).

Table 2: Model configurations and Hyperparameter grid

Model	**Hyperparameter**	**Values**
sklearnMLP [6]	*hidden layer sizes*	[50], [100], [50, 50]
	activation	logistic, tanh, relu
	solver	adam
	alpha	0.0001, 0.001, 0.01
	learning rate	constant
	max iter	300, 400
kerasMLP [7]: 3 layers (input linear layer, 32 neurons ReLU layer, output linear layer)	*epochs*	500, 750, 1000
	batch size	64, 128
	learning rate	0.001
	lambda regu	0.01, 0.1
	hidden neurons	[50], [100], [50, 50]
RandomForest [6]	*n estimators*	50, 70, 90
	max features	auto, sqrt, log2
	max depth	4, 8, 12
	min samples split	2, 5
	min samples leaf	1, 2, 3
	bootstrap	True, False
AdaBoost [6]	*n estimators*	50, 70, 90
	learning rate	0.01, 0.1
	algorithm	SAMME, SAMME.R
Bagging [6], with DecisionTree as estimator	*n estimators*	10, 50, 100
	max samples	0.5, 1.0
	max features	0.5, 1.0
	base estimator max depth	3, 5, 10
DecisionTree [6]	*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
	criterion	gini, entropy, log_loss
KNearNeighbors [6]	*n neighbors*	10, 25, 50, 65, 80
	metric	minkowski
	p	2
LDA [6]	*solver*	svd, lsqr, eigen
	priors	None, [0.1, 0.9], [0.5, 0.5], [0.9, 0.1]
	tol	0.0001, 1e-05, 1e-06
LogisticRegression [6]	*C*	0.001, 0.01, 0.1, 1
	penalty	l1, l2
	max iter	1000, 2000, 3000, 4000, 5000
QDA [6]	*priors*	None, [0.1, 0.9], [0.5, 0.5], [0.9, 0.1]
	reg param	0.0, 0.1, 0.5
	store covariance	False, True
	tol	0.0001, 1e-05
RIPPER [8]	*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
SVM [6]	*C*	0.1, 1
	kernel	linear, rbf, sigmoid
	gamma	scale, auto
Stacking [6], two base classifiers: SVM and DT, final classifier: Logistic Regression	*svm C*	0.1, 1
	dt max depth	3, 5, 7
	final estimator C	0.1, 1
Voting [6], 3 classifiers: SVM, Logistic Regression and RandomForest	*lr C*	0.1, 1
	svc C	0.1, 1
	rf n estimators	10, 50, 100
	voting	soft, hard

Table 3: Selected features by each method

Feature	Univariate	Multivariate	Wrapper								
			LDA	QDA	LR	GNB	RF	AB	Bagging	Stacking	Voting
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	✓	✓	✓	✓		✓	✓	✓	✓	✓	
Sex_F		✓	✓	✓	✓	✓	✓	✓			✓
Sex_M		✓	✓	✓	✓	✓		✓		✓	
SGOT		✓	✓		✓	✓	✓	✓	✓	✓	✓
Spiders_N	✓	✓		✓	✓		✓		✓	✓	✓
Spiders_Y		✓	✓			✓	✓		✓	✓	
Stage	✓	✓		✓		✓					✓
Tryglicerides			✓		✓			✓			

* LDA: Linear Discriminant Analysis; QDA: Quadratic Discriminant Analysis; LR: Logistic Regression; GNB: Gaussian Naive Bayes; RF: Random Forest; AB: Ada Boost;

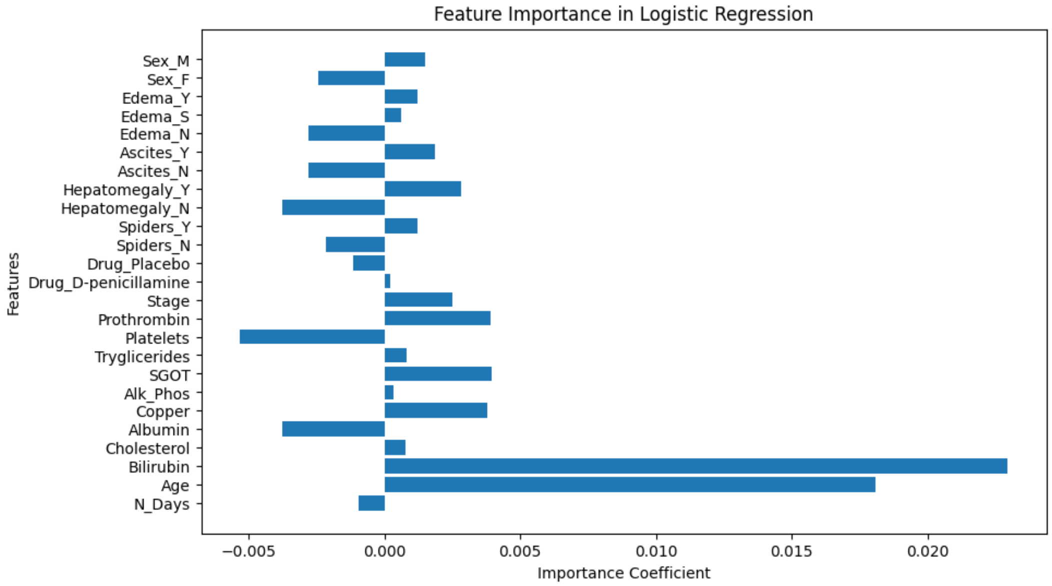

Figure 1: Interpretability Logistic Regression

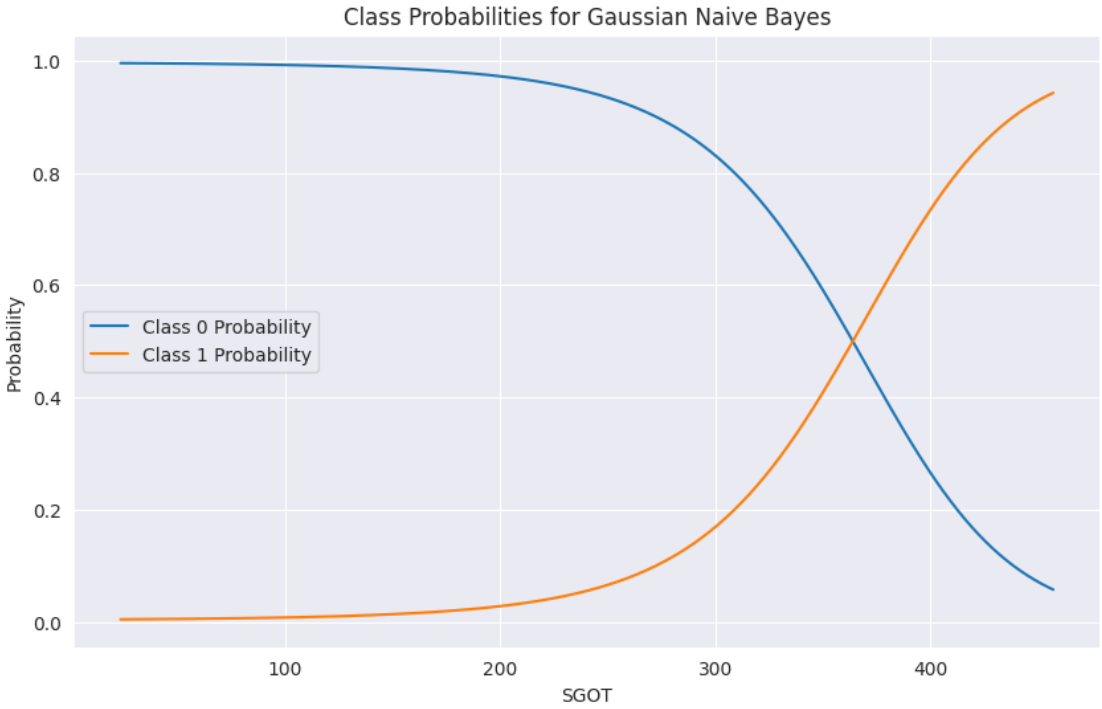

Figure 2: Interpretability GNB with Feature `SGOT`

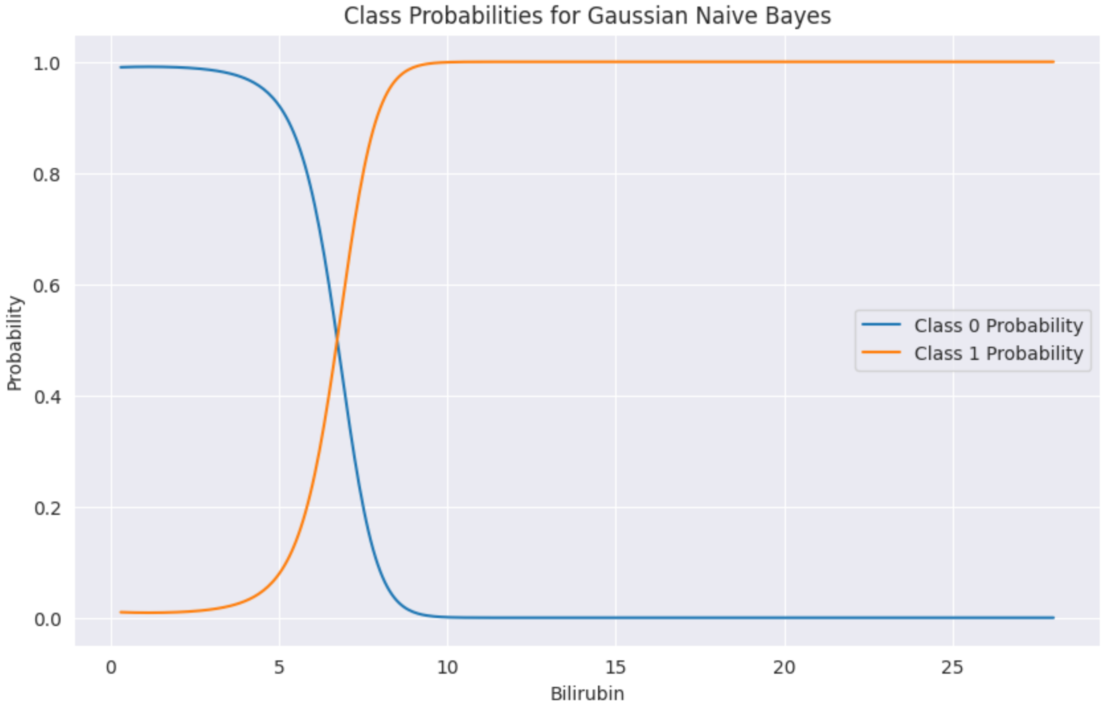

Figure 3: Interpretability GNB with Feature `Billirubin`

4.1 Feature Subset Selection

Some features, such as `Age`, `Albumin`, `Bilirubin`, and `Prothrombin`, consistently appear across multiple methods, indicating that these features are likely strong predictors and provide valuable information for the models. Additionally, there are features selected exclusively by specific methods; for instance,

Table 4: Models Comparative Results

Model	All		Univariate		Multivariate		Wrapper	
	F-score	Brier-score	F-score	Brier-score	F-score	Brier-score	F-score	Brier-score
DT	0.6881 ± 0.06	0.1908 ± 0.03	0.6877 ± 0.06	0.1951 ± 0.04	0.6988 ± 0.05	0.1895 ± 0.02	0.6644 ± 0.05	0.2162 ± 0.03
KNN	0.7190 ± 0.07	0.1747 ± 0.02	0.7134 ± 0.07	0.1736 ± 0.02	0.7134 ± 0.07	0.1744 ± 0.02	0.7190 ± 0.07	0.1752 ± 0.02
RIPPER	0.5595 ± 0.11	0.2251 ± 0.05	0.5461 ± 0.09	0.2078 ± 0.03	0.6264 ± 0.07	0.1970 ± 0.03	0.6067 ± 0.07	0.1905 ± 0.02
kerasMLP	0.6485 ± 0.03	–	0.6586 ± 0.04	–	0.5983 ± 0.10	–	0.6900 ± 0.05	–
sklearnMLP	0.7081 ± 0.03	0.1543 ± 0.02	0.6945 ± 0.05	0.1611 ± 0.03	0.7012 ± 0.06	0.1660 ± 0.01	0.7109 ± 0.05	0.1547 ± 0.02
SVM	**0.7715 ± 0.04**	–	**0.7604 ± 0.05**	–	0.7340 ± 0.06	–	0.7244 ± 0.05	–
LR	0.7593 ± 0.05	0.1484 ± 0.02	0.7596 ± 0.06	0.1484 ± 0.02	0.7507 ± 0.04	0.1591 ± 0.02	0.7728 ± 0.04	0.1563 ± 0.02
LDA	0.7516 ± 0.04	0.1467 ± 0.01	0.7452 ± 0.06	0.1491 ± 0.02	0.7577 ± 0.04	0.1483 ± 0.01	0.7714 ± 0.05	0.1464 ± 0.01
QDA	0.7351 ± 0.07	0.1800 ± 0.04	0.7563 ± 0.08	0.1765 ± 0.05	0.7422 ± 0.02	0.1885 ± 0.03	0.6583 ± 0.09	0.2700 ± 0.13
GNB	0.6885 ± 0.03	0.2115 ± 0.03	0.7098 ± 0.03	0.1969 ± 0.03	0.7355 ± 0.04	0.1874 ± 0.02	0.7659 ± 0.05	0.1756 ± 0.04
RF	0.7497 ± 0.04	0.1360 ± 0.02	0.7461 ± 0.04	0.1388 ± 0.02	**0.7589 ± 0.05**	**0.1355 ± 0.02**	**0.7881 ± 0.05**	**0.1389 ± 0.02**
AB	0.7247 ± 0.06	0.1972 ± 0.01	0.7380 ± 0.08	0.1932 ± 0.01	0.7467 ± 0.04	0.1890 ± 0.01	0.7215 ± 0.09	0.2026 ± 0.01
Bagging	0.7521 ± 0.03	0.1462 ± 0.01	0.7340 ± 0.04	0.1437 ± 0.02	0.7332 ± 0.04	0.1408 ± 0.02	0.7586 ± 0.04	0.1495 ± 0.02
Stacking	0.7261 ± 0.06	0.1704 ± 0.03	0.7133 ± 0.06	0.1674 ± 0.02	0.7086 ± 0.07	0.1694 ± 0.02	0.7520 ± 0.04	0.1630 ± 0.03
Voting	0.7351 ± 0.04	–	0.7601 ± 0.05	–	0.7471 ± 0.04	–	0.7524 ± 0.06	–

* DT: Decision Tree, MLP: Multi Layer Perceptron; SVM: Supper Vector Machine;
LDA: Linear Discriminant Analysis; QDA: Quadratic Discriminant Analysis; LR: Logistic Regression;
GNB: Gaussian Naive Bayes; RF: Random Forest; AB: Ada Boost;

`Drug_Placebo` is chosen by LDA, QDA, and some ensemble methods but not by others, suggesting that certain features may only demonstrate their predictive power under specific model conditions or when interacting with other features in complex ways. Furthermore, ensemble methods like Random Forest (RF), Ada Boost (AB), bagging, stacking, and voting display a broader range of feature selection compared to many individual models. This diversity is expected, as ensemble methods combine the predictions of several base estimators to enhance generalizability and robustness over a single estimator.

4.2 Models Performance

Marchine learning classifiers models were evaluated across different feature selection methods, with performance measured by F-score and Brier-score. The results displayed in Table 4, shows that the Random Forest (RF) model exhibited superior performance in the `Multivariate` and `Wrapper` feature selection categories, achieving the highest F-score and Brier-score, indicating both accuracy and reliability. The Support Vector Machine (SVM) demonstrated the highest F-scores in the `All` and `Univariate` categories, although Brier-scores were not reported because can no be estimate in a proper way. The consistency of model performance varied, with some models like KNN showing stability across different feature selection methods, while others like RIPPER were more variable. Furthermore, the Voting model with univariate filter exhibits an F-score of 0.7601, achieving a metric very close to that of SVM.

However, this project exhibits certain limitations demonstrating the statistical superiority of one model over another requires rigorous testing beyond what has been presented.

Ultimately, the results reveal biases in evaluation metrics across folds, suggesting the presence of outliers in the data and indicating a non-normally distributed dataset. Consequently, I assert that a personalized preprocessing approach should be applied for each model, depending on its inherent nature.

4.3 Models Interpretability

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 [9].

On the contrary, an elevated platelet count is associated with a diminished likelihood of death due to its crucial role in promoting blood clotting and preventing excessive bleeding. Platelets are essential for preserving vascular barrier function in the absence of injury or inflammation, emphasizing their importance in maintaining overall health and reducing the risk of fatal outcomes [10].

5 Conclusion

This project has successfully demonstrated the predictability of outcomes for patients with biliary cirrhosis using state-of-the-art classification models. Our findings reveal that the Support Vector Machine (SVM) achieves optimal performance in terms of F-score when employing all available features and when using the univariate feature subset. Conversely, the RandomForest model excels in scenarios involving multivariate and wrapper feature subsets, as evidenced by its lowest Brier scores, indicating effective calibration and reliability.

Furthermore, our experiments highlight the significant impact that different feature selection algorithms can have on a model's performance, either enhancing or diminishing its effectiveness. This observation underscores the importance of careful feature selection in model optimization.

A key takeaway from this study is the ability of certain models, particularly Logistic Regression and Gaussian Naive Bayes, to identify which features are most influential in the classification decision. This capability is especially valuable in the medical domain, where understanding the factors contributing to a model's predictions is crucial for clinical decision-making.

In summary, this research emphasizes the importance of model selection, feature selection, and the interpretability of models in medical applications. The nuanced differences in model performance across various feature selection techniques offer insights into the complexities of predictive modeling in healthcare.

References

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

[2] 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.

[3] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Physical review E*, vol. 69, no. 6, p. 066138, 2004.

[4] R. J. Urbanowicz, R. S. Olson, P. Schmitt, M. Meeker, and J. H. Moore, "Benchmarking relief-based feature selection methods for bioinformatics data mining," *Journal of biomedical informatics*, vol. 85, pp. 168–188, 2018.

[5] F. J. Ferri, P. Pudil, M. Hatef, and J. Kittler, "Comparative study of techniques for large-scale feature selection," in *Machine intelligence and pattern recognition*, vol. 16, pp. 403–413, Elsevier, 1994.

[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] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015. Software available from tensorflow.org.

[8] I. Moscovitz, "How to perform explainable machine learning classification — without any trees." https://github.com/imoscovitz/wittgenstein, 2019.

[9] 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.

[10] S. Gupta, C. Konradt, A. Corken, J. Ware, B. Nieswandt, J. Di Paola, M. Yu, D. Wang, M. T. Nieman, S. W. Whiteheart, *et al.*, "Hemostasis vs. homeostasis: Platelets are essential for preserving vascular barrier function in the absence of injury or inflammation," *Proceedings of the National Academy of Sciences*, vol. 117, no. 39, pp. 24316–24325, 2020.