applied sciences

MDPI

Article

Predicting the Survival of Primary Biliary Cholangitis Patients

Diana Ferreira [1], Cristiana Neto [1], José Lopes [2], Júlio Duarte [1], António Abelha [1] and José Machado [1,*]

[1] Algoritmi Research Center, University of Minho, 4710 Braga, Portugal
[2] Department of Informatics, University of Minho, 4710 Braga, Portugal
* Correspondence: jmac@di.uminho.pt; Tel.: +351-253604430; Fax: +351-253604471

Abstract: Primary Biliary Cholangitis, which is thought to be caused by a combination of genetic and environmental factors, is a slow-growing chronic autoimmune disease in which the human body's immune system attacks healthy cells and tissues and gradually destroys the bile ducts in the liver. A reliable diagnosis of this clinical condition, followed by appropriate intervention measures, can slow the damage to the liver and prevent further complications, especially in the early stages. Hence, the focus of this study is to compare different classification Data Mining techniques, using clinical and demographic data, in an attempt to predict whether or not a Primary Biliary Cholangitis patient will survive. Data from 418 patients with Primary Biliary Cholangitis, following the Mayo Clinic's research between 1974 and 1984, were used to predict patient survival or non-survival using the Cross Industry Standard Process for Data Mining methodology. Different classification techniques were applied during this process, more specifically, Decision Tree, Random Tree, Random Forest, and Naïve Bayes. The model with the best performance used the Random Forest classifier and Split Validation with a ratio of 0.8, yielding values greater than 93% in all evaluation metrics. With further testing, this model may provide benefits in terms of medical decision support.

Keywords: classification; data mining; predictive models; primary biliary cholangitis

Citation: Ferreira, D.; Neto, C.; Lopes, J.; Duarte, J.; Abelha, A.; Machado, J. Predicting the Survival of Primary Biliary Cholangitis Patients. *Appl. Sci.* **2022**, *12*, 8043. https://doi.org/10.3390/app12168043

Academic Editor: Keun Ho Ryu

Received: 1 July 2022
Accepted: 8 August 2022
Published: 11 August 2022

1. Introduction

In terms of the disease's definition, Primary Biliary Cholangitis (PBC) is described as being a chronic, progressive cholestatic liver disease of unknown cause [1], but because of autoantibodies, it is generally thought to be an autoimmune disease [2], that eventually leads to liver failure and the need of a liver transplantation [1]. PBC was formerly known as Primary Biliary Cirrhosis, but the designation was around 2015, because cirrhosis was not a necessary condition for the diagnosis of this disease [3].

PBC is hypothesised to be related to environmental exposure in genetically vulnerable individuals. Prominent clinical features include fatigue, pruritis, jaundice, xanthomas, osteoporosis, and dyslipidemia [4]. Additionally, it is known that about 90% of patients with PBC are women and that the disease is normally diagnosed in patients between the ages of 40 and 60 years [2], meaning that genetic factors, sex, and age are commonly associated as risk factors of PBC. Moreover, previous research has shown that the risk of PBC can be increased by alterations in sex hormones in women over time [5].

Nowadays, its diagnosis is becoming more common as a result of increased physician recognition and the widespread use of automated blood testing and the antimitochondrial antibody test, which is somewhat specific for this disease [1]. Given current screening methods, PBC is now often diagnosed when the patient is still asymptomatic, with abnormal liver biochemistry and/or Antimitochondrial Antibodies (AMA) identified in blood during a routine check-up or as part of the work-up for an associated disorder [6]. The Mayo Risk score is the most widely used and the best prognostic system [4].

In this study, a public dataset of 418 patients diagnosed with PBC at various stages during the disease's progression, containing 19 important features to the identification and monitoring of the disease, is examined. Since the chosen dataset contained information

about the patient's status, i.e., if the patient was alive, in need of a liver transplant, or deceased, this study focused on the prediction of the survival of patients diagnosed with PBC using Data Mining (DM) techniques.

DM is useful for analysing and exploring large datasets in order to discover meaningful patterns and rules [7]. The implementation of DM techniques can also facilitate the reduction in medical errors, enhance patient safety, standardise clinical practice, and improve patient outcomes [8]. Since the dataset is labelled, this study will use supervised techniques, which can include regression and classification [9]. As expected, and as previously stated, the focus of this paper will be on classification techniques, with different classifiers from the RapidMiner software being used to predict PBC patients' status, i.e., the label attribute, which corresponds to the attribute that predicts the patients' survival or non-survival. With this study, it will be possible to identify which patients have a greater chance of survival. As a result, hospitals and healthcare professionals will be able to concentrate their efforts, time, resources, and treatment options on patients who are more likely to survive. On the other hand, in terms of research, the cases of patients classified as patients who will not survive can be used for the search and development of innovative treatments.

The paper is structured in five sections. After introducing the problem, it will be presented the Related Work Section 2, regarding works related to the PBC topic. This section is followed by a detailed description of the DM process carried out using the CRossIndustry Standard Process for Data Mining (CRISP-DM) methodology in the Materials and Methods Section 3. Next, an analysis of the obtained results and its interpretation is performed in the Discussion Section 4. Finally, the Conclusions Section 5, summarises the whole process, draws the final conclusions based on the results achieved, and presents a brief proposal for future work.

2. Related Work

The subject of this paper has been mentioned and studied in several articles, given the fact that it is not only a chronic liver disease but also a rare condition that is often only treated through liver transplantation. As a result, the majority of the studies in the literature are devoted to improving treatment management and evaluating the influence of certain substances in disease spread.

Reference [10] studied the progression of PBC in 312 patients treated at the Mayo Clinic between January 1974 and May 1984. This study was based on repeated patient visits (a total of 1945 patient visits) that were important to evaluate the changes in the prognostic variables of PBC, such as age, albumin value, prothrombin time, bilirubin value, and edema. This dataset is similar to the one used in this study but not exactly the same. Using these data and the Cox proportional-hazards regression model, it was built an updated model to be used in the short-term survival prediction at any stage of the disease. It was concluded that the new updated model makes more accurate predictions in terms of short-term survival than the original Mayo model, meaning that the Mayo model is still the best choice for predicting the patient's survival for more than 3 years, but for periods up to 2 years, this new model offers a more accurate estimation of PBC patients' survival.

The biochemical response to ursodeoxycholic acid (UDCA) in PBC was analysed by [11] and is related to the long-term prognosis of the disease, thus allowing the identification of the patients' need of new therapeutic approaches. The aim of this study was to determine, in a population of patients in an early-stage of the disease, the most efficient biochemical response to UDCA, allowing the prediction the absence of poor outcome, as described by liver-related death, liver transplantation, complications of cirrhosis, or histological evidence of cirrhosis development.

A similar study was developed by [12], in which it was explored the effect of azathioprine on the survival of PBC patients. Using a randomised clinical trial containing 248 patients, where the immunosuppressive medication azathioprine was given to 127 of those patients, with the remaining 121 patients receiving placebo, which has no therapeutic

value. Similarly to the approach taken by [10], it was applied a Cox multiple regression analysis and an adjustment relative to the imbalance between the treatment groups. The therapeutic effect of azathioprine has been shown to reduce the risk of non-survival to 59% of those observed with placebo, as well as to improve the survival time by up to 20 months for the average patient, revealing a statistical significance of azathioprine therapy.

In a global manner, as it can be seen from the previous articles, there is more emphasis in the literature on the analysis of treatments for PBC and not specifically on its prognosis or evolution. Thus, more studies are needed to broaden the knowledge and understanding on the diagnosis of this condition as early as possible [13]. Machine learning algorithms can be used to perform early diagnosis and risk stratification, as well. Hence, developing diagnostic algorithms for PBC patients based on demographic variables, symptomatology, and laboratory results is a resourceful tool for improving the quality of clinical practice. We present a novel DM approach for PBC where the emphasis is the use of classified data to predict the survival of patients diagnosed with PBC. Several experiments were carried out, in which different DM techniques and feature selection setups were taken into account. Such knowledge is particularly useful to perform early diagnosis and risk stratification, thus improving the quality of clinical practice and lowering the mortality rate of patients diagnosed with PBC.

3. Materials and Methods

The DM process followed the CRISP-DM methodology, which is one of the most used for increasing the success of a DM project [14]. One of the main advantages of its utilisation is that it allows the construction and implementation of a DM model that can be used in a real environment, helping to support business decisions [15]. CRISP-DM is defined as a cyclic process, in which six phases are defined [16,17]:

- Business Understanding, where it is defined the project objectives and requirements, as well as the Data Mining problem definition;
- Data Understanding, in which the initial data are collected, proceeded by their familiarisation and analysis, where initial conclusions are taken in terms of data quality problems and obvious results;
- Data Preparation, characterised by the selection of data (decision of the important attributes to analyse), data cleansing (remove duplicates, decision of the best approach in terms of missing values and outliers), data transformation (includes production of derived attributes or even entire new records), and, in case of unbalanced datasets, data sampling, by applying oversampling or undersampling techniques to reduce the imbalance in the classes' distribution;
- Modeling, where the modeling techniques that will be used are selected followed by their execution on the given dataset to compare its results;
- Evaluation, determines if the results meet business objectives and identifies business problems that should have been addressed earlier;
- Deployment, the final phase of the methodology, which will not be taken into consideration in this article, refers to the practical implementation of the resulting models, where we take the evaluation results and determine a strategy for their development [18].

Figure 1 illustrates the lifecycle of the CRISP-DM methodology.

The RapidMiner software was the tool selected to conduct this study, which is a prominent data science platform that binds data preparation, Machine Learning, and predictive model deployment. Its user-friendly interface, efficiency, and wide range of algorithms and techniques available were additional factors for this choice.

All methods were performed in accordance with the relevant guidelines and regulations.

In the following subsections, each phase of the CRISP-DM will be discussed in relation to the data of PBC patients.

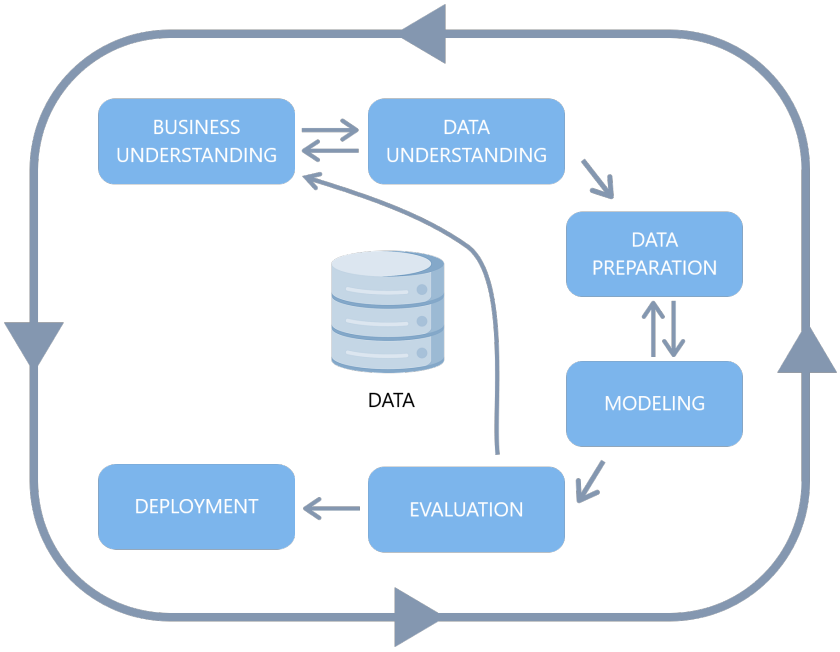


Figure 1. Stages of the CRISP-DM lifecycle.

3.1. Business Understanding

The current DM process focuses its business objective on the study of the health status of patients with PBC. With this business objective in mind, the study could pursue two different approaches, predict the survival/non-survival of patients with PBC, or determine the need of surviving patients for liver transplantation. However, in this article, and in the processing of the given data, it was only taken into consideration a binary classification, namely, the survival/non-survival of patients with PBC.

Additionally, there are some aspects that can influence the business goal. This is the case, for example, of the type of drug administered to patients, where two different types of drugs—D-penicillamine or placebo—can be administered. Based on the number of deceased patients vs. the number of non-deceased patients, the effect of these drugs may be helpful, because if a drug type has a lower number of deceased patients associated in comparison to the other drug, it can be assumed that this type of drug may have more success in the survival of patients with PBC, which is very helpful to physicians who are medicating patients with this condition.

It is also essential to check the influence of the patient's histologic stage. To do this, it is important to understand the different stages of the condition:

- Stage 1—known as portal hepatitis;
- Stage 2—associated with periportal hepatitis;
- Stage 3—septal fibrosis, bridging necrosis, or both;
- Stage 4—final stage, commonly referred to as cirrhosis.

Obviously, the death rate can be expected to be higher in patients with stage 4 PBC, which means that it is probably more difficult to save them than to save a patient with stage 1 or stage 2 PBC.

The study schema for this project is shown in Figure 2. This schema shows the structuring main steps of the research presented in this paper. Each one of these steps will be detailed in the next sections.

After defining the business goal, the need to analyse the data arose. Hence, the next subsection focuses on the data understanding stage of the CRISP-DM methodology.

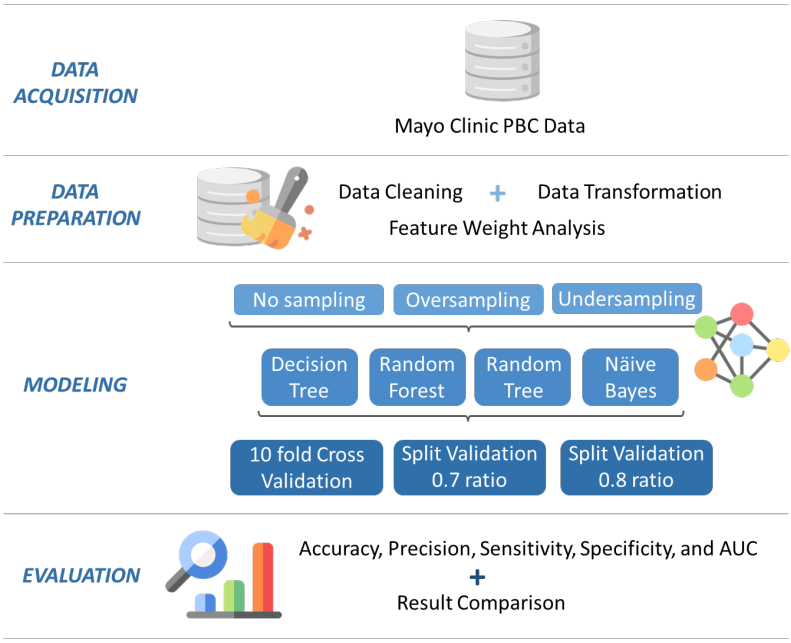


Figure 2. Project plan of steps taken in the present study.

3.2. Data Understanding

The data used in the present study are from a Mayo Clinic trial in PBC of the liver conducted between 1974 and 1984 [19]. The dataset contains information about 418 patients with PBC. The information of each patient, defined in the dataset as a data instance, consists of a set of 19 variables plus the patient's id, which was later disregarded from the study. The description of the attributes is presented in Table 1.

Table 1. Description of the attributes in the given data before the modifications made in the Data Preparation phase.

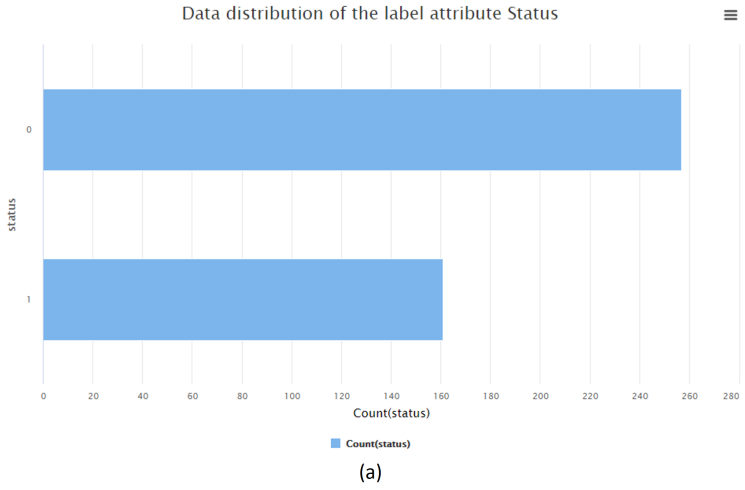
Variable	Type	Missing Values	Range	Average	Description
futime	Integer	0	[41–4795]	1917.78	number of days between registration and the earlier of death or study analysis time
age	Integer	0	[9598–28,650]	18,533.35	age of each patient, in days
sex	Integer	0	[0–1]	0.90	0 = male; 1 = female
status	Integer	0	[0–2]	0.83	0 = alive; 1 = needs a liver transplant; 2 = deceased
drug	Integer	106	[1–2]	1.49	1 = D-penicillamine; 2 = placebo
stage	Integer	6	[1–4]	3.02	1 = portal hepatitis; 2 = periportal hepatitis; 3 = septal fibrosis, bridging necrosis, or both; 4 = cirrhosis
ascites	Integer	106	[0–1]	0.08	existence of abnormal accumulation of fluid in the abdomen: 0 = no; 1 = yes
hepato	Integer	106	[0–1]	0.51	existence of hepatomegaly (enlarged liver condition): 0 = no; 1 = yes
spiders	Integer	106	[0–1]	0.29	existence of blood vessel malformations in the skin (spiders): 0 = no; 1 = yes
edema	Real	0	[0–1]	0.10	accumulation of fluids in body tissues: 0 = no edema and no diuretic therapy for edema; 0.5 = presence of edema but without diuretics or edema resolved by diuretics; 1 = presence of edema despite diuretic therapy
bili	Real	0	[0.30–28]	3.22	amount of serum bilirubin (mg/dL)
chol	Integer	134	[120–1775]	369.51	amount of serum cholesterol (mg/dL)
albumin	Real	0	[1.96–4.64]	3.50	amount of albumin (gm/dL)
copper	Integer	108	[4–588]	97.65	amount of copper in urine (µg/day)

Table 1. *Cont.*

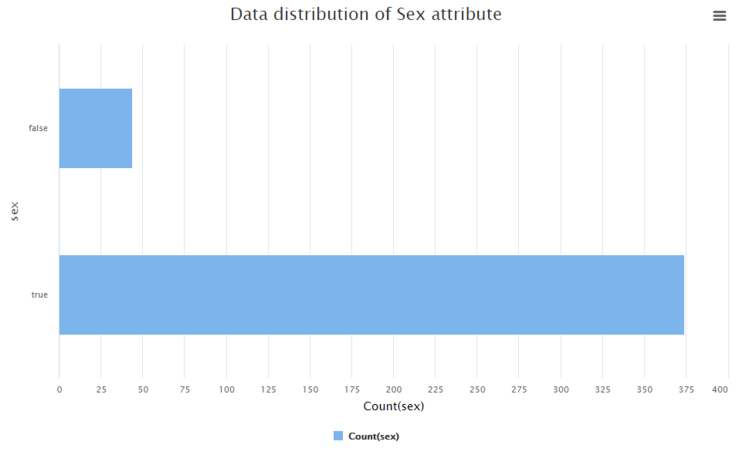
Variable	Type	Missing Values	Range	Average	Description
alk_phos	Real	106	[289–13862.40]	1982.66	amount of alkaline phosphatase (U/L)
sgot	Real	106	[26.35–457.25]	122.56	amount of Serum Glutamic Oxaloacetic Transaminase (U/mL)
trig	Integer	136	[33–598]	124.70	number of triglycerides (mg/dL)
platelet	Integer	11	[62–721]	257.03	number of platelets per cubic mL/1000
protime	Real	2	[9–18]	10.73	prothrombin time (in seconds), to evaluate the extrinsic pathway of coagulation

As already mentioned, this study is based on a binary classification, namely, the survival/non-survival of patients with PBC, so the *status* attribute was transformed to the binary type, where patients that were in need of liver transplantation were considered surviving patients.

Lastly, it was also important to underline some interesting aspects of the given data. Therefore, Figure 3 presents the distribution of the label attribute—*status*—and the *sex* attribute, where it can be seen that the data are a little unbalanced in the label attribute and that there are much more women than men associated with this disease (which goes as expected in a real life scenario, as described in Section 1).



(a)



(b)

Figure 3. (**a**) Visualisation of the value discrepancy in the sex attribute. (**b**) Data distribution by attribute.

When comparing the label attribute and the *stage* attribute (Figure 4), it can also be concluded that, as expected, there are more deceased patients in the last stage of the disease—stage 4—followed by stage 3, which makes sense, given the fact that a patient is more likely to pass away when the disease is more advanced.

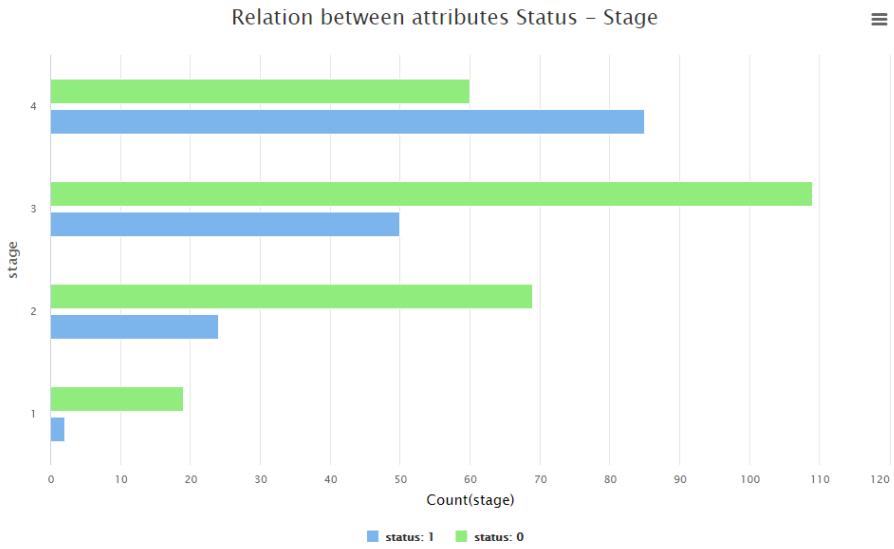


Figure 4. Association of the *status* attribute (label) and the *stage* attribute, where the alive patients are represented by the green bars and the deceased patients by the blue bars.

3.3. Data Preparation

First, as mentioned in the Business Understanding phase, the *status* attribute only considered the values referring to patient's survival/non-survival. In order to do so, patients in need of a liver transplant were referred to as surviving patients, which means that patients with a status value equal to 1 were replaced by a status value equal to 0. Additionally, patients with status value equal to 2 were replaced by a status value equal to 1, meaning that this attribute can now only take the value 0 or the value 1, where 0 stands for the patient's survival and 1 stands for the patient's death, making it a binomial attribute.

In addition, there was also the need to change the data type of some values, since most of them had the integer type, which was not the most accurate choice. Thus, the attributes *status*, *sex*, *ascites*, *hepato*, *spiders*, and *drug* were transformed to the binomial type, since they can only take two values, and the attributes *edema* and *stage* were transformed to the polynomial type, given the fact that they can take more than two specific values.

After adjusting the data types, it was important, as seen in Table 1, to deal with missing values in some attributes. Since the dataset has a reduced number of instances (patients), the best way to deal with these missing values is to replace them. To do so, these missing values were estimated by learning models for each attribute, with the exception of the label attribute. As a result, the missing values were treated by imputing the values resulting from the Neural Network (NN) model for missing numerical values and from the Decision Tree (DT) model for missing nominal values.

In terms of outlier removal, it was established that there were no outliers present in the dataset. More precisely, the possible outliers present in the given data were probably related to unusually high/low natural values, meaning that these values were not related to errors in the dataset itself, and consequently, there was no need to remove them.

Lastly, since there were numerous attributes related to substances found in the human body in different scales, where the higher values associated with higher scales would lead to an unrealistic influence, it became crucial to normalise each of these numerical attributes to the same scale. A minmax normalisation was therefore performed, so that all numeric attributes were at the same scale, with values ranging between 0 and 1, in order to correctly

determine which ones are actually more significant, leading to a more accurate use of the dataset.

3.4. Modelling

In this phase, the data resulting from the data preparation phase are used to feed the different Data Mining Models (DMMs) through the usage of the RapidMiner software. Given the fact that this is a classification approach, the following predictive models were used: DT, Random Forest (RF), Random Tree (RT), and Naïve Bayes (NB), using a fine-tuning process through the optimisation of the parameters of each model.

DT is an algorithm widely used for classification and regression tasks. The process of creating a DT works by greedily selecting the best split point in order to make predictions and repeat the process until the tree has a fixed depth. After the tree is constructed, it is pruned to improve the ability of the model to generalise to new data [20].

A common drawback of DTs is that they tend to overfit the training data. RF is one way to address this problem. RF is essentially a collection of DTs where, initially, a bootstrap sample is selected from the training data (random sample obtained with replacement) with the aim of inducing DT. This step is repeated until a set of DTs was created, each of which has its own predictive value. Hence, the final prediction is achieved by combining the output of all trees, which corresponds to the most frequent output of the ensemble [21].

RT works exactly like DT, with one exception: for each split, only a random subset of attributes is available. RT works similar to Quinlan's C4.5 or CART, but it selects a random subset of attributes before being applied [22]. Although the first three algorithms are tree-based, they do not perform the same. Because decision trees are an effective method of decision-making, it was decided to test different tree variations.

In addition, the NB classifier was also used in this study. It is based on the Bayes Theorem, which is a probabilistic theorem that is used to find out the probability of something happening (A) by knowing that (B) has occurred, as mathematically expressed in (1). Therefore, given the problem at hand, the A variable can be considered the label/target, the attribute *status*. The B variable can be extended and seen as the rest of the features of the dataset [20].

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)} \quad (1)$$

Following the selection of the models, it was necessary to define the different sampling methods for testing. Full training was not considered in this study; instead, three sampling methods were considered, namely, Cross Validation (CV), using 10 folds, and Split Validation (SV), with a split ratio of 0.7 and a split ratio of 0.8. CV uses all data for training, whereas SV uses a certain percentage for training and the remaining for testing, as can be seen in Figure 5.

Two missing values approaches were tested: the application of data replacement and the removal of instances with missing values.

In addition, three data approaches were taken into consideration, the original dataset, the dataset with oversampling, and the dataset with undersampling. Oversampling consisted of the replication of cases of the minority class until a balanced distribution was achieved between the two classes. On the other hand, undersampling consisted of the removal of instances from the majority class until a balanced distribution was achieved. Both oversampling and undersampling techniques were achieved using the sampling operator, which creates a sample from a dataset by selecting examples randomly. The number of examples in the sample can be specified on absolute, relative, or probability basis depending on the setting of the sample parameter. The class distribution of the sample can be controlled by the balance data parameter.

Additionally, to evaluate which attributes were the most relevant in the label's prediction, several scenarios were defined using different sets of attributes. This selection was made based on the feature weight analysis using two different operators:

- The *Weight by Correlation* operator, which calculates the weight based on the correlation of the feature with the *status* attribute (label). Highly correlated attributes can erroneously influence the label's prediction, which could be misleading or disguise the other features. Consequently, this operator was applied in the definition of the scenarios, where the attributes with high weights were eliminated, as they could be deceptive to the label's prediction.
- The *Weight by Information Gain Ratio* operator, which calculates the weight of each attribute related to the *status* attribute (label) through the usage of the information gain ratio, which means that the higher the weight of an attribute, the more relevant it is to the forecast. Consequently, this operator was applied in the definition of the scenarios, where the attributes with lower weights were eliminated, since they do not have much impact in the label's prediction.

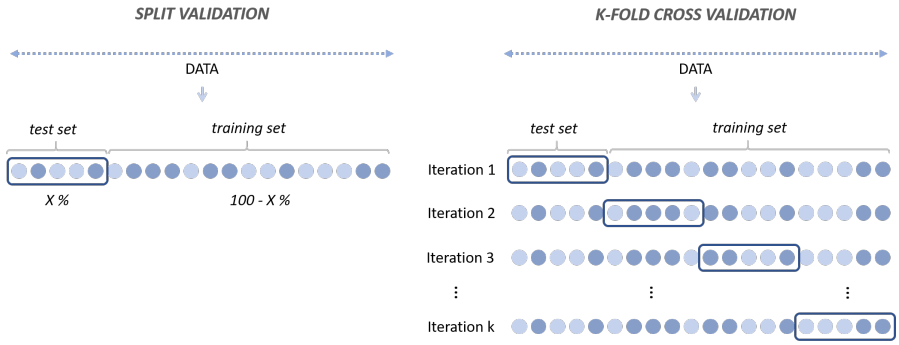


Figure 5. Comparison between Split and Cross Validation functioning.

Such results are displayed in Figure 6, where the Weight by Correlation operator is shown in Figure 6a and the Weight by Information Gain Ratio operator is displayed in Figure 6b.

Thereby, the scenarios defined were:

- S1: {All attributes};
- S2: {All attributes, except *drug*, *sex*, and *stage*}, where the attributes with weight inferior to 0.05 were eliminated, consistent with the information viewed in the Weight by Information Gain Ratio operator;
- S3: {All attributes, except *age*, *drug*, *edema*, *hepato*, *sex*, *sgot*, *spiders*, and *stage*}, where, similar to the S2 scenario, a threshold above 0.1 was applied, meaning that the attributes with weight below 0.1 were eliminated, in accordance with the Weight by Information Gain Ratio operator;
- S4: {All attributes, except *bili*}, since it is the attribute with higher weight correlation, according to the Weight by Correlation operator.

Conclusively, each attempt can be described as belonging to an Approach (A)= {Classification}, being associated to a Data Mining Technique (DMT) = {Decision Tree, Random Forest, Random Tree, Naïve Bayes}; a Scenario (S) = {S1, S2, S3, S4}; a Sampling Method (SM) = {Cross Validation, Split Validation 0.7, Split Validation 0.8}; a Missing Values Approach (MVA) = {Replacement (MVA1), Deletion (MVA2)}; a Data Approach (DA) = {None (DA1), Oversampling (DA2), Undersampling (DA3)}; and a Target (T) = {status}, as expressed in (2):

$$DMM = \{A, S, DMT, SM, MVA, DA, T\} \quad (2)$$

In total, 288 models were induced according to (3):

$$DMM = \{1(A) \times 4(S) \times 4(DMT) \times 3(SM) \times 2(MVA) \times 3(DA) \times 1(T)\} \quad (3)$$

The next phase shows the results achieved for each of these attempts.

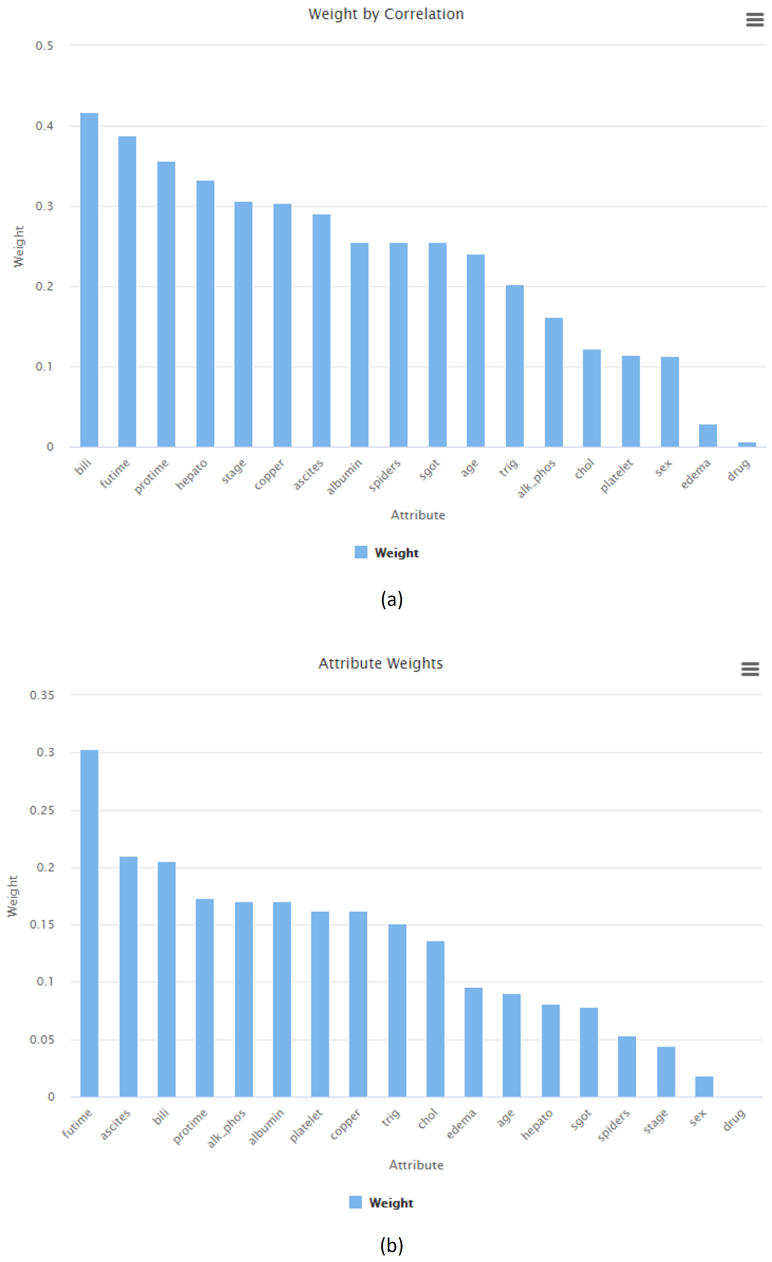


Figure 6. (**a**) Weight by Information Gain Ratio. (**b**) Feature weight analysis.

3.5. Evaluation

During this phase, several metrics were used to evaluate the performance of the different DMTs developed at the previous stage, as well as to determine the quality and reliability of the results. Since this study fits into a binary classification scope, the applied metrics were derived from the Confusion Matrix, which is a predictive classification table that contains the number of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). The evaluation metrics extracted from the Confusion Matrix were Accuracy, Precision, Sensitivity, and Specificity. In addition, the Area Under the ROC (Receiver Operating Characteristic) Curve (AUC) was also used to assess the performance of the models. The metrics chosen are the most commonly used when addressing this type of problems; specifically, the sensitivity measure is the most important in domains, where FNs have a high cost, such as medical diagnosis.

Accuracy, mathematically defined according to Equation (4), represents the ratio between the instances that the model was able to correctly classify and all the classified

instances. In this study, this metric refers to the number of PBC patients' status that were correctly classified either as surviving or dying out of all the patients [23]:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (4)$$

Precision, expressed by Equation (5), measures a classifier's exactness and is the ratio of the positive instances that were correctly classified by the model among all the positive instances classified by the model, i.e., it provides information on how many positive values were actually correctly predicted [23]. This metric evaluates the ability of the model to identify PBC patients who will not survive, i.e., the fraction of PBC patients who are classified as non-surviving and who have actually died.

$$Precision = \frac{TP}{TP + FP} \quad (5)$$

Sensitivity, also known as Recall, is calculated through Equation (6) and measures a classifier's completeness. It is associated with the ratio between the positive instances correctly predicted by the model and all the actual positive instances [23].

In the context of this study, this metric evaluates the ability of the model to correctly identify non-surviving PBC patients, i.e., the fraction of PBC patients classified as dying, labelled as 1, who actually died among all PBC patients who did not survive, described as 1 in the dataset. Sensitivity is therefore a good measure to evaluate models in domains where there is a high cost associated with FN, as it is with medical diagnosis, where it is harmful to predict that a PBC patient will not die when, in fact, he/she will die (FN). Thus, the closer the sensitivity is to 100% the better because a higher FN value means that the survival of some patients was incorrectly predicted, leading physicians to neglect intensive treatments or therapeutics that could save the patient's life. Therefore, and because of the critical nature of this problem, FN must be avoided at all costs.

$$Sensitivity = \frac{TP}{TP + FN} \quad (6)$$

Specificity, mathematically described by Equation (7), measures the ratio of negative instances correctly predicted by the model and all the negative instances, i.e, it provides information on how many negative values were actually correctly predicted [23]. In this study, the metric informs the proportion of surviving PBC patients that the model was able to classify correctly, evaluating the ability of the model to identify the PBC patients who will survive, i.e., the patients labelled as 0. A higher FP value means that the non-survival of patients was incorrectly predicted.

$$Specificity = \frac{TN}{TN + FP} \quad (7)$$

The AUC value is associated with the ROC probabilistic curve and is defined as a measure that informs the ability of the model to distinguish classes, where a higher AUC value indicates that the model predicts more correctly 0 s as 0 s and 1 s as 1 s.

Based on the previously defined DMTs, sampling methods, scenarios, as well as missing values and data approaches, all the attempts were evaluated. For each attempt, it was calculated the Accuracy, Precision, Sensitivity, Specificity, and AUC value. The best results are shown in Table 2, where the first column is an identifier of each attempt, the second is the machine learning model, the third is the scenario, the fourth is the sampling method, the fifth is the missing values approach, the sixth is the data-balancing approach, and the remaining columns are the evaluation metrics used. It is important to mention that a threshold of Accuracy values inferior to 85% was established to select the best models.

Table 2. Best results obtained regarding Accuracy, Precision, Sensitivity, Specificity, and AUC value.

#	DMT	S	SM	MVA	DA	Accuracy	Precision	Sensitivity	Specificity	AUC
1	DT	S1	SV (0.7)	MVA1	DA1	85.60%	84.71%	93.51%	72.92%	0.855
2	DT	S1	SV (0.8)	MVA1	DA1	89.16%	88.89%	94.12%	81.25%	0.500
3	RF	S1	SV (0.8)	MVA1	DA1	92.77%	95.92%	92.16%	93.75%	0.849
4	RT	S1	SV (0.8)	MVA1	DA1	85.54%	85.45%	92.16%	75.00%	0.688
5	DT	S1	SV (0.8)	MVA2	DA1	89.09%	88.57%	93.94%	81.82%	0.826
6	RF	S1	SV (0.8)	MVA2	DA1	94.55%	91.67%	100%	86.36%	0.884
7	RF	S1	SV (0.8)	MVA1	DA2	93.14%	90.74%	96.08%	90.20%	0.975
8	RF	S1	SV (0.8)	MVA1	DA3	93.75%	91.18%	96.88%	90.62%	0.967
9	RF	S2	SV (0.8)	MVA1	DA1	93.98%	91.07%	100%	84.38%	0.946
10	RF	S3	SV (0.8)	MVA1	DA1	92.77%	90.91%	98.04%	84.38%	0.968
11	RF	S4	SV (0.8)	MVA1	DA1	90.36%	92.16%	92.16%	87.50%	0.825
12	RF	S2	SV (0.8)	MVA1	DA3	89.16%	87.50%	96.08%	78.12%	0.871
13	RF	S3	SV (0.8)	MVA1	DA3	93.75%	88.89%	100%	87.50%	0.954
14	RF	S4	SV (0.8)	MVA1	DA3	95.31%	96.77%	93.75%	96.88%	0.906

4. Discussion

As seen in Table 2, the highest metric values are achieved through the usage of the RF predictive model. Additionally, the lowest values (which didn't make it to Table 2) were associated to the usage of the NB method, followed by the RT and the DT methods. In what concerns the sampling approach, the higher values were accomplished using split validation with a 0.8/0.2 ratio, meaning that 80% of the dataset was used for the training set (335 instances) and 20% was used for the test set (83 instances).

Despite the fact that attempts 5 and 6 have interesting metric values, they are not the finest predictive models, since they eliminate almost half of the patients' data. Initially, the dataset had 418 instances, and in the end, after removing all instances with missing values (MVA2), there were just 276 patients in the dataset, making these models somewhat inadequate, since its results are derived from few patients' information. Consequently, the attempts with the best results are also associated with the replacement of the missing values (MVA1), which allows us to use all the patients' data without losing substantial information.

Since it was more difficult to achieve good quality results for the Precision and Specificity metrics, they were the most relevant criteria when choosing the most fitting model. Furthermore, it was even more difficult to achieve high results in terms of Specificity; thus, it can be concluded that these models generate a substantial amount of FP, suggesting that it was incorrectly predicted the survival of some patients, when in reality, these patients have died, which in this particular case is not the ideal scenario, since the number of FPs needs to be as low as possible.

As for the best overall values for the calculated measures, the attempts 12 and 14 have the highest and coincidentally, the same values in terms of Accuracy, Sensitivity, Precision, and Specificity. However, the attempt 12 has a higher AUC value, much closer to 1, meaning that the quality of this model's prediction is superior. Consequently, the best predictive method is the one displayed in attempt 12, using the S2 scenario (all attributes, except *drug*, *sex*, and *stage*) and undersampling, in addition to the utilisation of the Random Forest method, data split validation with a 0.8/0.2 ratio, and the replacement of the missing values. This DMM achieved values superior to 93% in all the calculated metrics, namely, an Accuracy of 95.31%, a Precision of 96.77%, a Sensitivity of 93.75%, a Specificity of 96.88%, and an AUC of 0.990.

Nonetheless, using the different scenarios with split validation with a 0.8/0.2 ratio (with or without undersampling/oversampling)—attempts 3, 7, 8, 9, 10, 11, and 13—also

results in values superior to 90%, which are also promising results, but not as high as the ones obtained in attempts 12 and 14.

Additionally, comparing attempts 12 and 13, increasing the threshold value, from 0.05 (S2) to 0.1 (S3), was not beneficial to the results, and the removal of five more attributes in the S3 scenario, although resulting in a Sensitivity value of 100%, slightly reduced the percentage values of all the other metrics, subsequently implying a slight loss in predictive power in attempt 13.

Looking now specifically to attempt 12, it is still possible to interpret each one of the obtained metrics. Therefore: its 95.31% accuracy value demonstrates that this model correctly predicted almost every patient's status (survival or non-survival); its 96.77% precision value is related to the correctly predicted positive values, where this elevated value allows us to infer that there were only a few FP values in the calculated model, thus revealing that only a small fraction of deceased patients (status = 1) were predicted to have survived (status = 0); its 93.75% sensitivity value, as already mentioned, gives information about how many patients were correctly labelled as 1 (non-surviving) of those who were initially described as 1 in the dataset, meaning that, since this value is superior to 90%, the survival of many non-survival patients was not incorrectly predicted (low FN); its 96.88% specificity value, as said before, provides how many patients were correctly labelled as 1 (not survived) of the initial deceased patients, where this high value allows us to infer that the survival of many patients that were actually deceased was not incorrectly predicted. Lastly, its 0.990 AUC value is also vital to show that this model is close to the "perfect" classifier (that has an AUC value equal to 1), which means that this model has the ability to correctly distinguish the classes of the status attribute, fittingly predicting almost all 0 s as 0 s and 1 s as 1 s. Although promising results have been achieved, a direct comparison cannot be performed because the studies mentioned in the Related Work focus on the analysis of treatments for PBC rather than specifically on its prognosis or evolution.

5. Conclusions

PBC is a life-threatening disease, especially in patients with severe symptoms that are representative of advanced stages of the disease. In this sense, the prediction of the risk of mortality associated with a patient with this clinical condition is a key aspect in the medical decision-making process. Hence, this project involved the implementation of DMTs with the aim of predicting the survival or non-survival of patients with PBC, which requires a process that ensures that the results are reliable and statistically significant.

Consequently, performance metrics were applied to ensure an appropriate evaluation of the quality and characteristics of the models, consequently assuring the reliability of the results. Accuracy, Precision, Sensitivity, Specificity, and AUC were the metrics used in this project to quantify the classifiers' performance. Based on these metrics, some DMMs attained high results, with some of these results being higher than 90% in all evaluation metrics, making them almost "ideal" classifiers. The model that achieved the best results used the Random Forest classifier, the Split Validation method with a 0.8 ratio, the S2 scenario, the replacement of missing values, and the undersampling technique, resulting in values superior to 93% in all evaluation metrics, more specifically, an Accuracy of 95.31%, a Precision of 96.77%, a Sensitivity of 93.75%, a Specificity of 96.88%, and an AUC of 0.990.

Although promising results have been achieved, the amount of data used in this study is not sufficient to assume that the best predicted model is credible and could provide benefits regarding the support of medical decisions. In future work, therefore, more data should be collected from healthcare institutions in order to not only have a richer and more varied dataset but also to achieve a more balanced distribution of classes, avoiding the need to use data sampling techniques, making the models more reliable and realistic. In addition, it would also be advantageous to use other DM approaches and techniques to compare its results, such as Logistic Regression, Multi-Layer Perceptron, K-Neural Networks, Neural Networks, and Extreme Gradient Boosting. Additionally, it would be interesting to consider

other feature selection techniques, such as Lasso Regression, as well as other attributes, particularly those related with medication and therapeutics, in order to study the influence that these have in the label attribute—*status*—the patient's survival or non-survival and thus also provide insights on the most suitable drugs for treating patients with this clinical condition. Finally, it will be necessary to conduct a more in-depth study regarding the patients that were initially classified as in need of a transplant, since a patient who requires a transplant may survive but may also die.

Author Contributions: Conceptualisation, D.F., C.N., J.D. and J.L.; methodology, J.L., D.F., C.N., J.D. and A.A.; software, C.N., J.L. and D.F.; validation, J.D., D.F. and C.N.; formal analysis, J.D., A.A. and J.M.; investigation, J.D., D.F., C.N., A.A. and J.M.; resources, A.A. and J.M.; data curation, D.F., C.N., J.L., J.D., A.A. and J.M.; writing—original draft preparation, J.L., D.F., C.N. and J.M.; writing—review and editing, D.F. and C.N.; visualisation, J.L., D.F. and C.N.; supervision, A.A. and J.M.; project administration, J.M. and A.A.; funding acquisition, J.M. and A.A. All authors have read and agreed to the published version of the manuscript.

Funding: This work is funded by "Fundação para a Ciência e Tecnologia (FCT)" within the R&D Units Project Scope: UIDB/00319/2020.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are available in a publicly accessible repository that does not issue DOIs. Publicly available datasets were analysed in this study. These data can be found here: https://www.kaggle.com/jixing475/mayo-clinic-primary-biliary-cirrhosis-data (accessed on 1 July 2022).

Acknowledgments: D.F. and C.N. thank the Fundação para a Ciência e Tecnologia (FCT) Portugal for the grants 2021.06308.BD and 2021.06507.BD, respectively.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AMA	Antimitochondrial Antibodies
AUC	Area Under the Curve
CRISP-DM	CRossIndustry Standard Process for Data Mining
CV	Cross Validation
DA	Data Approach
DM	Data Mining
DMM	Data Mining Model
DMT	Data Mining Technique
DT	Decision Tree
FN	False Negatives
FP	False Positives
MVA	Missing Values Approach
NB	Naïve Bayes
NN	Neural Network
PBC	Primary Biliary Cholangitis
RF	Random Forest
ROC	Receiver Operating Characteristic
RT	Random Tree
SM	Sampling Method
S	Scenario
SV	Split Validation
T	Target
TN	True Negatives
TP	True Positives
UDCA	Ursodeoxycholic Acid

References

1. Kaplan, M.M. Primary Biliary Cirrhosis. *N. Engl. J. Med.* **1996**, *335*, 1570–1580. [CrossRef] [PubMed]
2. Revett, K.; Gorunescu, F.; Gorunescu, M.; Ene, M. Mining A Primary Biliary Cirrhosis Dataset Using Rough Sets and a Probabilistic Neural Network. In Proceedings of the 3rd International IEEE Conference Intelligent Systems, London, UK, 4–6 September 2006; pp. 284–289.
3. Morgan, M.A.; Sundaram, K.M. Primary biliary cholangitis: Review for radiologists. *Abdom. Radiol.* **2021**, 1–9. [CrossRef] [PubMed]
4. Purohit, T.; Cappell, M.S. Primary biliary cirrhosis: Pathophysiology, clinical presentation and therapy. *World J. Hepatol.* **2015**, *7*, 926. [CrossRef] [PubMed]
5. Zhang, L.; Ding, D.; Yu, L.; Qi, H.; Han, C.; Jiang, J.; Jiang, J. Primary biliary cirrhosis associated with myasthenia gravis after postpartum: A case report. *J. Med. Case Rep.* **2021**, *15*, 1–3. [CrossRef]
6. Heathcote, E. Management of Primary Biliary Cirrhosis. *Hepatology* **2000**, *31*, 1005–1013. [CrossRef] [PubMed]
7. Pujari, A.K. *Data Mining Techniques*; Universities Press: Madison, CT, USA, 2001.
8. Srinivas, K.; Rani, B.K.; Govrdhan, A. Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks. *Int. J. Comput. Sci. Eng. (IJCSE)* **2010**, *2*, 250–255.
9. Olson, D.L.; Delen, D. *Advanced Data Mining Techniques*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2008.
10. Murtaugh, P.A.; Dickson, E.R.; Van Dam, G.M.; Malinchoc, M.; Grambsch, P.M.; Langworthy, A.L.; Gips, C.H. Primary Biliary Cirrhosis: Prediction of Short-term Survival Based on Repeated Patient Visits. *Hepatology* **1994**, *20*, 126–134. [CrossRef] [PubMed]
11. Corpechot, C.; Chazouillères, O.; Poupon, R. Early primary biliary cirrhosis: Biochemical response to treatment and prediction of long-term outcome. *J. Hepatol.* **2011**, *55*, 1361–1367. [CrossRef]
12. Christensen, E.; Neuberger, J.; Crowe, J.; Altman, D.G.; Popper, H.; Portmann, B.; Doniach, D.; Ranek, L.; Tygstrup, N.; Williams, R. Beneficial effect of azathioprine and prediction of prognosis in primary biliary cirrhosis: Final results of an international trial. *Gastroenterology* **1985**, *89*, 1084–1091. [CrossRef]
13. Parés, A.; Leading PBC Group. Practical management of primary biliary cholangitis. *Rev. Esp. Enferm. Dig. Organo Of. Soc. Esp. Patol. Dig.* **2021**, *114*, 410–417. [CrossRef] [PubMed]
14. Ferreira, D.; Silva, S.; Abelha, A.; Machado, J. Recommendation system using autoencoders. *Appl. Sci.* **2020**, *10*, 5510. [CrossRef]
15. Moro, S.; Laureano, R.; Cortez, P. Using data mining for bank direct marketing: An application of the CRISP-DM methodology. In Proceedings of the 25th European Simulation and Modelling Conference—ESM'2011, Guimaraes, Portugal, 24–26 October 2011.
16. Martins, B.; Ferreira, D.; Neto, C.; Abelha, A.; Machado, J. Data Mining for Cardiovascular Disease Prediction. *J. Med. Syst.* **2021**, *45*, 1–8. [CrossRef] [PubMed]
17. Nogueira, M.; Ferreira, D.; Neto, C.; Abelha, A.; Machado, J. Data Mining for the Prediction of Fetal Malformation Through Cardiotocography Data. In *Proceedings of the International Conference on Information Technology & Systems*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 60–69.
18. Azevedo, A.I.R.L.; Santos, M.F. KDD, SEMMA and CRISP-DM: A parallel overview. In Proceedings of the IADIS European Conference on Data Mining 2008, Amsterdam, The Netherlands, 24–26 July 2008.
19. Kaggle. *Mayo Clinic Primary Biliary Cirrhosis Data*; Kaggle: San Francisco, CA, USA, 2019.
20. Neto, C.; Peixoto, H.; Abelha, V.; Abelha, A.; Machado, J. Knowledge discovery from surgical waiting lists. *Procedia Comput. Sci.* **2017**, *121*, 1104–1111. [CrossRef]
21. Neto, C.; Brito, M.; Lopes, V.; Peixoto, H.; Abelha, A.; Machado, J. Application of data mining for the prediction of mortality and occurrence of complications for gastric cancer patients. *Entropy* **2019**, *21*, 1163. [CrossRef]
22. Mohamed, M.H.; Waguih, H.M. A proposed academic advisor model based on data mining classification techniques. *Int. J. Adv. Comput. Res.* **2018**, *8*, 129–136. [CrossRef]
23. Ghoneim, S. Accuracy, Recall, Precision, F-Score & Specificity, Which to Optimize On? 2019. Availble online: https://towardsdatascience.com/accuracy-recall-precision-f-score-specificity-which-to-optimize-on-867d3f11124 (accessed on 1 July 2022).