**Research Article**

International Journal of Data Science and Applications (JOINDATA) 8(1), 11-27, 2025

Received: 20-Sep-24 Accepted: 13-Jan-25

homepage: https://dergipark.org.tr/tr/pub/joindata

# Using the Fleming-Harrington Estimator Method to Process Censored Data in Machine Learning: A Methodological Study

1st Pelin AKIN[1*], 2nd Yüksel TERZİ[2]

[1] Department of Statistics, Faculty of Science, Çankırı Karatekin University, pelinakin@karatekin.edu.tr

[2] Department of Statistics, University Faculty of Science, Ondokuz Mayıs University, yukselt@omu.edu.tr

## ABSTRACT

The Cox regression method is generally used to model censored data. Recently, with the increase in data, new methods have been sought. This study aims to reclassify the censored data using the Fleming-Harrington method to apply machine learning techniques, thereby conducting survival analysis through machine learning classification methods. In practice, the censored data of acute leukemia patients were used, with four distinct sample sizes simulated using a correlation matrix obtained from this acute leukemia dataset. The data were adapted to the machine learning algorithm using the Fleming-Harrington method. Naive Bayes, Decision Tree, Random Forest, and Support Vector Machines methods were applied to the datasets from among the classification algorithms. Performance metrics, including accuracy, the area under the ROC Curve (AUC), and the F score, were used to compare these algorithms. Results showed that the Random Forest algorithm performed best for the actual dataset, while the Naive Bayes algorithm produced the best outcomes for the simulated dataset. When examining the machine learning algorithm results, close values were found, with Naive Bayes outperforming other algorithms in all situations. Comparisons between these datasets using the Cox regression method and Naive Bayes algorithm AUC values revealed similar outcomes. However, as the sample size increased, the performance of the Cox regression method decreased, while the machine learning algorithms' performance increased. Therefore, machine learning algorithms can provide valuable insights into cancer patients' mortality status or the likelihood of disease recurrence in studies incorporating survival analyses, especially when the sample size is large.

**Keywords:** Survival analysis; Naive Bayes; Censored data; Fleming-Harrington Estimator.

* Corresponding Author's email: pelinakin@karatekin.edu.tr

# Fleming-Harrington Tahmin Yöntemini Kullanarak Makine Öğrenmesinde Sansürlü Verileri İşlemek: Metodolojik Bir Çalışma

## ÖZ

Cox regresyon yöntemi genellikle sansürlü verileri modellemek için kullanılır. Son zamanlarda, verilerin artmasıyla birlikte yeni yöntemler aranmıştır. Bu çalışma, makine öğrenmesi tekniklerini uygulamak için Fleming-Harrington yöntemini kullanarak sansürlü verileri yeniden sınıflandırmayı ve böylece makine öğrenmesi sınıflandırma yöntemleri aracılığıyla sağkalım analizi yapmayı amaçlamaktadır. Uygulamada, akut lösemi hastalarının sansürlü verileri kullanıldı ve bu akut lösemi veri setinden elde edilen bir korelasyon matrisi kullanılarak dört ayrı örneklem boyutu simüle edildi. Veriler, Fleming-Harrington yöntemi kullanılarak makine öğrenmesi algoritmasına uyarlandı. Sınıflandırma algoritmaları arasından veri setlerine Naive Bayes, Karar Ağacı, Rastgele Orman ve Destek Vektör Makineleri yöntemleri uygulandı. Bu algoritmaları karşılaştırmak için doğruluk, ROC Eğrisi Altındaki Alan (AUC) ve F puanı gibi performans ölçütleri kullanıldı. Sonuçlar, Rastgele Orman algoritmasının gerçek veri kümesi için en iyi performansı gösterdiğini, Naive Bayes algoritmasının ise simüle edilmiş veri kümesi için en iyi sonuçları ürettiğini gösterdi. Makine öğrenimi algoritması sonuçları incelendiğinde, Naive Bayes'in tüm durumlarda diğer algoritmalardan daha iyi performans gösterdiği yakın değerler bulundu. Cox regresyon yöntemi ve Naive Bayes algoritması AUC değerleri kullanılarak bu veri kümeleri arasında yapılan karşılaştırmalar benzer sonuçlar ortaya koydu. Ancak, örneklem boyutu arttıkça Cox regresyon yönteminin performansı azalırken, makine öğrenimi algoritmalarının performansı arttı. Bu nedenle, makine öğrenimi algoritmaları, özellikle örneklem boyutu büyük olduğunda, sağkalım analizlerini içeren çalışmalarda kanser hastalarının ölüm durumu veya hastalığın tekrarlama olasılığı hakkında değerli bilgiler sağlayabilir.

**Anahtar Kelimeler:** Sağkalım Analizi, Naive Bayes; Sansürlü Veri, Fleming-Harrington Tahmin

## 1 Introduction

Censored data in health research often fails to capture an individual's life expectancy fully, presenting significant challenges for survival analysis. Traditionally, Cox regression and Kaplan-Meier methods have been used to model such data. However, advancements in storage, processing power, and network connectivity have enabled the application of machine learning techniques to increasingly complex datasets that were previously unmanageable with traditional methods.

Machine learning is still relatively young in its lifecycle, mainly when applied to censored data, which remains challenging. Early studies often ignored censored data, leading to biased models. For instance, Snow, et al. [1] developed a neural network to predict recurrence after radical prostatectomy but treated censored data as non-recurrent, resulting in a model biased towards non-recurrent cases. Similarly, other studies excluded patients with short follow-up periods, leading to biased outcomes by only considering recurrent cases among those with limited follow-up. One of the first studies to adequately address censored data was by Faraggi and Simon [2], who used an Artificial Neural Network (ANN) instead of the traditional log-linear relationship between independent variables and the hazard function. Kattan and colleagues suggested that neural networks outperform traditional statistical models. Zupan [3] and team used machine learning techniques to analyze prostate cancer recurrence and proposed new methods to handle censored data by determining the distribution of outcomes rather than assuming non-occurrence. They applied Naive Bayes and Decision Trees, finding that Bayesian models and Cox regression outperformed Decision Trees, although the difference was insignificant. Fard [4]and team presented a

model to predict future events using early-stage data, finding that Bayesian algorithms performed better than Cox regression. This study involved real data from acute leukemia patients and simulated data with varying sample sizes, applying classification algorithms (Naive Bayes, Decision Trees, Random Forest, and Support Vector Machines) and evaluating performance based on criteria like accuracy and AUC. Billichová [5] et al. compared the Cox Proportional Hazards (CPH) model with the Random Survival Forest (RSF) algorithm for tumor progression prediction, showing that while CPH was superior in some datasets, RSF had advantages in complex, high-dimensional data. Tizi [6] and Berrado demonstrated that machine learning algorithms generally provide higher prediction accuracy than traditional cancer research methods. Stefan Leger [7] and colleagues evaluated various machine learning methods for radiomics risk modeling and found them generally outperform traditional methods. Annette Spooner and team assessed machine learning methods for dementia prediction in high-dimensional clinical data, revealing significant advantages of some algorithms.Özbay Karakuş and Er [8] compared machine learning algorithms for heart failure survival prediction, finding more accurate methods.

This study distinguishes itself by using the Fleming-Harrington estimator to predict right-censored data and integrating these predictions with machine learning methods for more accurate and reliable survival analysis. It also compares machine learning algorithms across different sample sizes and performance criteria using real and simulated data. The article is structured as follows: Section 2 defines the machine learning algorithms and the proposed method. Section 3 discusses the application of these algorithms with data. Section 4 provides a brief discussion, and the final section presents the conclusions.

## 2 Research Methodology

### 2.1 Suggested Handling Censored Data

Kaplan – Meier estimator is a nonparametric method incorporated in estimating the survival function from the survival time data [9]. This method is not limited to medicine but is used in different fields. The Kaplan – Meier estimator is calculated using the Equation (1) formula.

$$\widehat{S(t)} = \prod\prod(1 - \frac{d_i}{n_i}) \qquad (1)$$

$d_i$: indicating the number of deaths during ti,

$n_i$: indicating the number of individuals at risk and is obtained with $n_i = n_{i-1} - d_{i-1} - c_{i-1}$

$c_i$: indicating the number of censored observations during ti,

Nelson proposed an alternative method to the Kaplan – Meier estimator (1972), and Aalen expanded on this estimator in 1978 to create the Nelson – Aalen estimator (NAE) [10]. Although Kaplan – Meier and Nelson – Aalen outcomes are nearly identical, Nelson – Aalen is more effective in cumulative hazard function estimates and small samples [11]. The survival function estimation is calculated using the formula in Equation (2).

$$\hat{S}_{NA}(t_i) = \exp\left(-\sum \frac{d_i}{n_i}\right) \quad (2)$$

The Fleming – Harrington method emerged in 1984 with some changes to the Nelson – Aalen method. The survival function is given in Equation (3).

$$\hat{S}_{FH}(t_i) = \exp\left(-\sum_{k=1}^{i} \sum_{j=0}^{d_k-1} \frac{1}{d_k - j}\right) \quad (3)$$

The probability of an event occurring before time t, using the Kaplan-Meier survival function equation, is given in Equation (4).

$$\hat{F}_e(t) = 1 - \hat{S}_{FH}(t_i) \quad (4)$$

The probability of it being censored before time t is given in Equation (5), and the probability of censored data occurring before time t is given in Equation (6)[12].

$$\widehat{G(t)} = \prod_{i:t_i} \left(1 - \frac{c_i}{n_i}\right) \quad (5)$$

$$\hat{F}_c(t) = 1 - \widehat{G(t)} \quad (6)$$

If $\hat{F}_e(t) > \hat{F}_c(t)$ , then the data is labeled as the event, and if $\hat{F}_e(t) < \hat{F}_c(t)$, then it is labeled as event-free [4].

By comparing these two probability values, a new state variable is created. Upon follow-up on the censored data patient group, we determine whether an event will occur or non-occur. In Table 1, calculations such as event estimates, survival rates, and censored event probabilities for the first 15 observations are given in detail.

Table 1 shows the event estimates and survival calculations for the first 15 observations of the acute leukemia dataset used in the study. The values of $d_i$ (number of deaths), $c_i$ (number of censored observations), $n_i$ (number of individuals at risk), $\widehat{\mathbf{G}(\mathbf{t})}$(probability of non-censoring), $\hat{\mathbf{F}}_\mathbf{c}(\mathbf{t})$ (probability of censoring), ST* (survival probability), $\widehat{\boldsymbol{F}}_{\boldsymbol{e}}(\boldsymbol{t})$ (event probability) and "eventPRE" (event prediction) in this table were calculated using the study's Kaplan-Meier and Fleming-Harrington methods described in detail. During the calculations, the value of ni was iteratively updated with the formula $n_i = n_{i-1} - d_{i-1} - c_{i-1}$, and other probability values were obtained using the relevant formulae. The table includes the first 15 observations to provide an example of the analysis applied to the full dataset; results for the full dataset are included in the scope of the study. Figure 1 provides the probabilities of the occurrence and non-occurrence of an event through the incorporation of the Fleming-Harringhton estimators.

**Table 1.** *Event Estimates and Survival Calculations for the First 15 Observations*

| DAYS | DI | CI | NI | $\widehat{G(t)}$ | $\hat{F}_c(t)$ | ST* | $\hat{F}_e(t)$ | EVENTPRE |
|---|---|---|---|---|---|---|---|---|
| **0.000** | 0.000 | 0.000 | 164.000 | 1.000 | 0.000 | 1.000 | 0.000 | 0.000 |
| **5.000** | 1.000 | 0.000 | 164.000 | 0.994 | 0.006 | 1.000 | 0.000 | 1.000 |
| **11.000** | 1.000 | 0.000 | 163.000 | 0.988 | 0.012 | 1.000 | 0.000 | 1.000 |
| **17.000** | 2.000 | 0.000 | 162.000 | 0.976 | 0.024 | 1.000 | 0.000 | 1.000 |
| **19.000** | 0.000 | 1.000 | 160.000 | 0.976 | 0.024 | 0.994 | 0.006 | 1.000 |
| **21.000** | 0.000 | 2.000 | 159.000 | 0.976 | 0.024 | 0.981 | 0.019 | 1.000 |
| **22.000** | 0.000 | 1.000 | 157.000 | 0.976 | 0.024 | 0.975 | 0.025 | 0.000 |
| **25.000** | 1.000 | 1.000 | 156.000 | 0.969 | 0.031 | 0.969 | 0.031 | 0.000 |
| **34.000** | 1.000 | 0.000 | 154.000 | 0.963 | 0.037 | 0.969 | 0.031 | 1.000 |
| **40.000** | 1.000 | 0.000 | 153.000 | 0.957 | 0.043 | 0.969 | 0.031 | 1.000 |
| **48.000** | 0.000 | 1.000 | 152.000 | 0.957 | 0.043 | 0.962 | 0.038 | 1.000 |
| **57.000** | 1.000 | 0.000 | 151.000 | 0.951 | 0.049 | 0.962 | 0.038 | 1.000 |
| **65.000** | 0.000 | 1.000 | 150.000 | 0.951 | 0.049 | 0.956 | 0.044 | 1.000 |
| **70.000** | 0.000 | 1.000 | 149.000 | 0.951 | 0.049 | 0.950 | 0.050 | 0.000 |
| **71.000** | 0.000 | 1.000 | 148.000 | 0.951 | 0.049 | 0.943 | 0.057 | 0.000 |
| **76.000** | 1.000 | 0.000 | 147.000 | 0.944 | 0.056 | 0.943 | 0.057 | 0.000 |

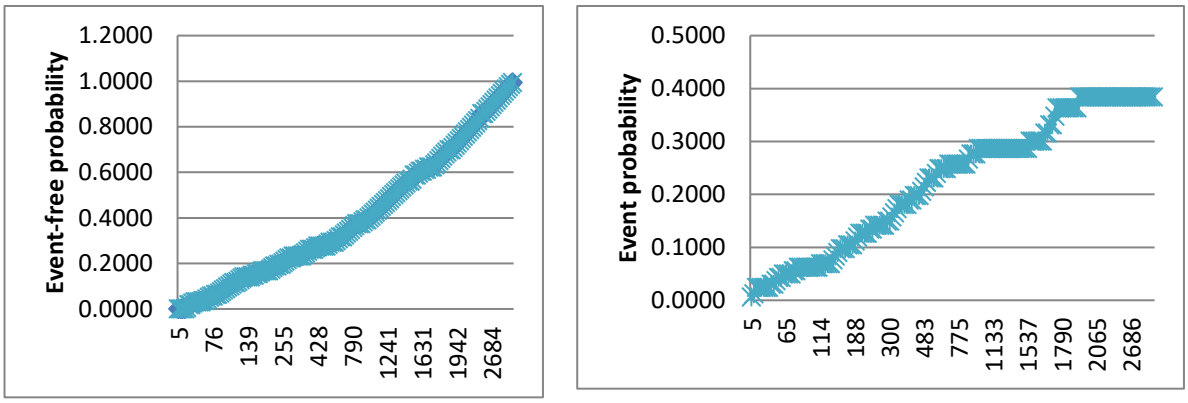


**Figure1.** Graph of the Probabilities of Events Occurring and Not Occurring for Each Predictor

Figure 1 provides the probabilities of the occurrence and non-occurrence of an event through the incorporation of the Fleming – Harrington estimators. The probability of occurrence of the event variable, that is, the probability of death, showed stabilization after approximately 1800 days. The probability of the event not occurring at all increased as the number of days increased.

## 2.2 Cox Regression Method

The Cox regression model is a semi-parametric model examining independent variables with survival time [13]. The basic assumption in the Cox regression model is that the hazard function is constant. Thus, the Cox regression model is also called the proportional hazard model. Equation (7) shows the mathematical representation of the Cox regression model [14].

$$h(t,X) = h_0(t).\exp\left(\sum_{i=1}^{p} \beta_i X_i\right) \quad (7)$$

The vector of X=( $X_1, X_2, \ldots, X_P$ ) represents independent variables, while β represents regression coefficients vector, and $h_0(t)$ is an essential, unspecified hazard function of t.

Akaike information criterion (AIC) and Bayesian Information Criteria (BIC) are often used to compare Cox regression models. AIC and BIC are calculated using the formulas in Equations (8) and (9).

$$AIC = -2logL + 2p \quad (8)$$

$$BIC = -2logL + plogn \quad (9)$$

In statistical modeling, the maximized log-likelihood of the data given the model parameter estimates is represented by (L), with p denoting the number of parameters in the model and n representing the sample size.

### 2.3 Decision Trees

Decision tree algorithms are among the most favored machine learning techniques due to their interpretability, error detection capabilities, and applicability [15].

One of the classification tree algorithms is the C4.5 algorithm. The C4.5 algorithm employs the entropy technique. Entropy is defined as a measure of uncertainty. This measure is based on the work of Claude Shannon, who developed an information theory that examines the value or "information content" of messages [16]. Information gain is defined as the difference between the original information need, which is Entropy (D) based solely on the ratio of classification, and the new information need, $Entropi_A(D)$. The calculation determines the extent to which information gain will be achieved by utilizing attribute A.

$$\text{Information Gain(A)} = Entropy(\text{D}) - Entropy_{\text{A}}(\text{D}) \quad (10)$$

(Eq.11) shows the split information formula.

$$Split\ Information_{\text{A}}(\text{D}) = -\sum_{\text{J}=1}^{\text{v}} \frac{|\text{D}_\text{j}|}{|\text{D}|} * \log_2 \frac{|\text{D}_\text{j}|}{|\text{D}|} \quad (11)$$

Split information interpretation is generated by splitting the training data set, D, into v partitions, corresponding to the v outcomes of a test on attribute A. The gain ratio is defined as the quotient of the

mean of the outcomes in each partition and the mean across all partitions, as shown in Equation 12. While this ratio is obtained with Equations 10 and 11, it is represented here as Equation 12 for simplicity.

$$\text{Gain Ratio} = \frac{\text{Information Gain (A)}}{Split\ Information_{\text{A}}(\text{D})} \tag{12}$$

The variable exhibiting the highest gain rate is selected for splitting.

### 2.4 Naive Bayes Classifier

The Naive Bayes classifier is a statistically supervised learning method [17]. Although this classifier is theoretically simple, it is often effective [18]. Naive Bayes classifiers assume that the effect of a property value in a given class is independent of the values of other attributes. This assumption is called class conditional independence [19]. The Naive Bayes classifier is constructed using Bayes' theorem.

In the Naive Bayes classifier, suppose the X class label is an unknown data sample and is comprised of values X=($x_1$, $x_2$, . . .,$x_m$). Moreover, suppose the dataset has a m number of classes named ($C_1$, $C_2$, …, $C_m$). The class of the sample is determined by calculating the probabilities, as shown in Equation (13).

$$\text{P}(C_{\text{i}} \mid X) = \frac{\text{P}(\text{X} \mid \text{C}_{\text{i}})\text{P}(\text{C}_{\text{i}})}{\text{P}(\text{X})} \tag{13}$$

Given that $P$(X) is constant for all classes and that only the numerator needs to be maximized, it can be assumed that the Naive Bayes classes are conditionally independent. Consequently, we can express $\text{P}(\text{X} \mid \text{C}_{\text{i}})$ as follows:

$$\text{P}(\text{X} \mid \text{C}_{\text{i}}) = \prod_{\text{j}=1}^{\text{n}} \text{P}(x_{\text{aj}} \mid \text{C}_{\text{i}}) \tag{14}$$

In Equation (14), only the numerator part of the fraction is compared since the denominators will be equal. The biggest number among the numbers compared is chosen and designated as belonging to this class. This Equation is called the maximum posterior classification method.

### 2.5 The Random Forest

The Random Forests algorithm represents an ensemble learning method whereby a prediction model is created by combining the strengths of several simpler, more basic models [20]. Breiman's Random Forest classification represents an advanced iteration of the bagging method, achieved by incorporating randomness. The Random Forest algorithm comprises the following steps:

i) The original dataset is drawn into $n$ bootstrap samples.

ii) An unpruned classification or regression tree (CART) is created for each bootstrap sample.

iii) In random classification, two parameters are utilized: the number of variables used in each node ($m$) and the number of trees to be developed ($N$). The optimal split is identified through the utilization of these parameters. A new estimate is derived by aggregating the estimates from the N trees. In the case of classification trees, the class with the majority of votes is selected as the final estimate. In contrast, the estimate is calculated by averaging the votes for regression trees [21].

### 2.6 Support Vector Machines

Support Vector Machines (SVMs) have been proposed to solve classification and regression problems. They are supervised learning techniques based on statistical learning theory and the principle of minimizing structural risk[22]. The inaugural study on support vector machines was presented in 1992 by Vladimir Vapnik and his colleagues Bernhard Boser and Isabelle Guyon. However, the origins of this study can be traced back to 1960, as evidenced by its inclusion in Vapnik and Alexei Chervonenkis's seminal work on the theory of statistical learning [23].

Let each $x_i$ be defined as an input with D number of attributes, and let yi be defined as an output representing the class with samples, each of which can take on one of two values, +1 or –1. A linear hyperplane that optimally separates the training set S, comprising n pairs of ($x_i$, $y_i$), into distinct classes can be identified.

$$\begin{aligned} \mathrm{wx} + \mathrm{b} \geq +1 \quad & y_i{=}{+}1 \\ \mathrm{wx} + \mathrm{b} \leq -1 \quad & y_i{=}{-}1 \end{aligned} \tag{15}$$

In Equation (15), W is defined as a weight vector, and b is a constant. A multitude of linear classifiers are capable of data partitioning; however, only one class can achieve the optimal margin, which entails maximizing the distance between the nearest data points of each class. This linear classifier is the optimal hyperplane for splitting [24]. The boundary between the two classes' support vectors is maximized to identify the optimal splitting plane, referred to as the support vector. The mathematical function of the SVM algorithm is as follows:

$$\mathrm{f(x)} = \mathrm{sgn}\big(\mathrm{(\ w}x_i) + \mathrm{b}\big) = \mathrm{sgn}(\sum_{i=1}^{l} y_i \alpha_i (x_i x_j)) \tag{16}$$

"In Equation (16), j refers to the index of the other data points (or support vectors) used in the summation, while i represents the specific data point being evaluated.

$$Subject\ to \begin{cases} min\ \dfrac{\|w\|^2}{2} \\ \mathrm{y_i(wx + b)} \geq 1 \end{cases} \tag{17}$$

### 2.7 Performance Evaluation Criteria for Classification Algorithms

The data set is partitioned into a training set and a test set, the objective being to assess the model's efficacy. The data set is employed to train the model. The model's performance is evaluated using data from the test set [25]. The most common ratios employed are (60:40),(70:30), and (80:20). It has been

demonstrated that dropout rates can impact performance. A confusion matrix is typically employed to ascertain the model's performance in two-class classification models. The confusion matrix for a two-class dataset is presented in Table 2.

**Table 2.** *Confusion matrix*

| | | Estimate | |
|---|---|---|---|
| | | Positive | Negative |
| True | Positive | True Positive (TP) | False Negative (FN) |
| | Negative | False Positive (FP) | True Negative (TN) |

The following section presents the formulas used to determine the classification performance.

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP}$$

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$\mathrm{F_{score}} = \frac{2 * \mathrm{Sensitivity} * \mathrm{Precision}}{\mathrm{Sensitivity} + \mathrm{Precision}}$$

### 2.8 Simulation

The multivariate normal distribution assumption data for leukemia data was produced with the "Binnor" package in the R program. However, since the day and age variables should be positive integers, absolute and rounding codes were added to this package. A correlation matrix was created from actual data before running the simulation. This correlation matrix was created using biserial correlation for the correlation between the continuous and categorical variables, the Phi correlation coefficient for the relationship between the two categorical variables, and the Pearson-Spearman correlation coefficient between the quantitative variables. In addition, the correlation matrix created should be positively defined, and if it is not, the *compute.sigma.star* function from the binnor package gives the closest positively defined matrix.

The data were derived by taking four different sample sizes (500, 1000, 1500, and 2000) and repeating this scenario 100 times, with the results averaged.

## 3 Results

This study aimed to utilize machine learning techniques to classify mortality status in patients diagnosed with acute leukemia. The data set comprised 165 patients diagnosed with acute leukemia between 1992 and 2002 at the Ondokuz Mayis University Faculty of Medicine Research Hospital Chest Diseases Department [26]. The work complies with the principles of the Helsinki Declaration. The data on leukemia patients consists of two groups: acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL). The dataset includes 14 independent, one target, and time variables. Among the independent variables, only age and days are quantitative, and other variables are qualitative (Table 3).

**Table 3.** *The variables used in the analysis*

| **Independent Variables** | | **Encoding Format** |
|---|---|---|
| Group | | AML, ALL |
| Age | | |
| Swallowing Difficulty | | No, Yes |
| Weakness | | No, Yes |
| Cough | | No, Yes |
| Fire | | No, Yes |
| Mass | | No, Yes |
| Leukocyte | | ≤100000, >100000 |
| Uric acid | | <7, ≥7 |
| kidney function | | All normal, Impaired |
| bone marrow blas | | ≤80, >80 |
| Treatment | | All, Other |
| 15th-day bone marrow | | ≤%5, >%5 |
| 30 days bone marrow | | No, Yes |
| **Dependent Value** | | |
| Time | Day | Numeric |
| Status | Categorical | Sansürlü, Ölüm |

To ensure the generalizability of the results and to prevent overfitting, 10-fold cross-validation was applied during the evaluation of the classification algorithms. This approach divides the dataset into ten equal parts; 9 parts are used for training, and one is used for testing each iteration. This process is repeated 10 times, with each subset used as the test set exactly once. The mean performance metrics, such as AUC and F-score, were calculated across these iterations to provide a robust estimate of the algorithms' effectiveness. The test set results were also carefully analyzed to evaluate the model's predictive capability on unseen data. This ensures that the reported performance metrics reflect the training data and the model's ability to generalize to new, unseen data.

In this study, the Multiple Imputation with Chained Equations (MICE) method, one of the multiple imputation methods, was employed to address the issue of missing data. This multiple imputation method yields a statistical distribution based on the dataset. Subsequently, the distribution above is utilized to fill in the missing data. This process is repeated on more than one occasion, with each data set stored for subsequent utilization. Furthermore, the error rate of each dataset is calculated [27]. Furthermore, the data were derived from four sample sizes: 500, 1,000, 1,500, and 2,000. Furthermore, this procedure was repeated 100 times, and the resulting data were averaged.

After preprocessing the leukemia data through the Fleming-Harrington method, it was determined whether the patient's death or life. By comparing these two probability values, a new state variable is created. Upon follow-up on the censored data patient group, we determine whether an event will occur or non-occur. Table 4 presents a comparison of the performance of the classification algorithms for varying training-test data ratios (60%-40%, 70%-30%, 80%-20%) and sample sizes (N=500, 1000, 1500, 2000) utilizing the AUC (Area Under the Curve) criterion.

**Table 4.** *AUC Results for Classifying Algorithms*

| | | (60%-40%) (Training-Test) | (70%-30%) (Training-Test) | (80%-20%) (Training-Test) |
|---|---|---|---|---|
| **Actual data** | **RF** | 0.7324 | 0.563 | 0.5722 |
| | **SVM** | 0.6439 | 0.6 | 0.6245 |
| | **NB** | 0.6633 | 0.6294 | 0.5727 |
| | **C4.5** | 0.4683 | 0.5088 | 0.7295 |
| **N=500** | **RF** | 0.6509 | 0.6334 | 0.6782 |
| | **SVM** | 0.6358 | 0.5786 | 0.6403 |
| | **NB** | 0.6793 | 0.6785 | 0.6949 |
| | **C4.5** | 0.4787 | 0.5 | 0.4786 |
| **N=1000** | **RF** | 0.6751 | 0.6808 | 0.7027 |
| | **SVM** | 0.6542 | 0.6638 | 0.6809 |
| | **NB** | 0.7048 | 0.7118 | 0.7231 |
| | **C4.5** | 0.4675 | 0.4754 | 0.4966 |
| **N=1500** | **RF** | 0.6925 | 0.6851 | 0.6838 |
| | **SVM** | 0.6748 | 0.6678 | 0.6585 |
| | **NB** | 0.7181 | 0.7117 | 0.7122 |
| | **C4.5** | 0.4487 | 0.4644 | 0.4593 |
| **N=2000** | **RF** | 0.6858 | 0.6873 | 0.6878 |
| | **SVM** | 0.6655 | 0.6678 | 0.6637 |
| | **NB** | 0.7117 | 0.7158 | 0.7043 |
| | **C4.5** | 0.4677 | 0.4773 | 0.4608 |

The Random Forest (RF) algorithm showed the highest performance in the real dataset with an AUC of 0.7324 at 60% training - 40% test rate. In the N=500 sample size, the Naive Bayes (NB) algorithm achieved the highest success with an AUC value of 0.6949 at an 80% training - 20% testing ratio. Likewise, in the N=1000 sample size, the Naive Bayes (NB) algorithm stood out with an AUC value of 0.7048 at an 80% training - 20% testing ratio. In N=1500 and N=2000 sample sizes, the Naive Bayes (NB algorithm exhibited the highest AUC performance with AUC values of 0.7181 and 0.7158 at 70% training - 30% testing ratio, respectively. Among the other algorithms in the table, Support Vector Machines (SVM) variable performance at different rates, for example, with an AUC value of 0.6809 at 80% training - 20% testing ratio in the N=1000 dataset. The C4.5 algorithm, the other algorithm, presented generally lower AUC values, with the highest AUC value measured as 0.7295 at 80% training - 20% testing ratio on N=1000 sample size. In conclusion, the Naive Bayes (NB algorithm performed best with generally high AUC values, while the Random Forest algorithm was somewhat effective. Table 5 shows the performance of the classification techniques on the F-score criterion using three different training-test ratios (60%-40%, 70%-30%, 80%-20%) for the real dataset and different sample sizes (N=500, 1000, 1500, 2000).

On the actual dataset, the Random Forest (RF) algorithm showed the highest performance in the F-score criterion at 80% training - 20% test ratio. In particular, a significant improvement in F-score values was observed as the sample size increased, reaching approximately 0.90. While the other algorithms, Support Vector Machines (SVM) and Naive Bayes (NB), also achieved high F-score values at these rates, the C4.5 algorithm generally presented lower F-scores. At N=500 sample size, Random Forest (RF) 0.8441, Support Vector Machines (SVM) 0.8419, Naive Bayes (NB) 0.8447 and C4.5 0.8347 F-score values at 80% training - 20% testing ratio. At N=1000 sample size, Random Forest (RF) showed F-score values of 0.8418, Support Vector Machines (SVM) 0.8425, Naive Bayes (NB) 0.8411 and C4.5 0.8369. At

N=1500 sample size, Random Forest (RF) 0.8414, Support Vector Machines (SVM) 0.8398, Naive Bayes (NB) 0.8421 and C4.5 0.8382 F-score values were obtained. At N=2000 sample size, Random Forest (RF) achieved F-score values of 0.8375, Support Vector Machines (SVM) 0.8366, Naive Bayes (NB) 0.8376 and C4.5 0.8321. As a result, the Random Forest (RF) algorithm exhibited the highest F-score performance at 80% training - 20% testing ratio, and a significant improvement in F-score was achieved as the sample sizes increased. Table 6 used three different split ratios (60% - 40%, 70% - 30%, 80% - 20%) and accuracy comparison criteria to compare the classification techniques. In Table 5, the accuracy of the classification techniques is compared, and the effects of different split ratios and their performances are evaluated.

**Table 5.** *F-score Results for Classifying Algorithms*

| | | **60%-40% (Training-Test)** | **(70%-30%) (Training-Test)** | **(80%-20%) (Training-Test)** |
|---|---|---|---|---|
| **Actual data** | **RF** | 0.7186 | 0.6763 | 0.6197 |
| | **SVM** | 0.6925 | 0.6737 | 0.5968 |
| | **NB** | 0.6769 | 0.6738 | 0.6563 |
| | **C4.5** | 0.7231 | 0.6326 | 0.5 |
| **N=500** | **RF** | 0.8446 | 0.8504 | 0.8441 |
| | **SVM** | 0.8458 | 0.8526 | 0.8419 |
| | **NB** | 0.8457 | 0.8527 | 0.8447 |
| | **C4.5** | 0.8409 | 0. 8527 | 0.8347 |
| **N=1000** | **RF** | 0.8462 | 0.8483 | 0.8418 |
| | **SVM** | 0.8456 | 0.8484 | 0.8425 |
| | **NB** | 0.8465 | 0.8488 | 0.8411 |
| | **C4.5** | 0.8385 | 0. 8432 | 0.8369 |
| **N=1500** | **RF** | 0.8305 | 0.8419 | 0.8414 |
| | **SVM** | 0.8302 | 0.8420 | 0.8398 |
| | **NB** | 0.8315 | 0.8417 | 0.8421 |
| | **C4.5** | 0.8222 | 0. 8344 | 0.8382 |
| **N=2000** | **RF** | 0.8374 | 0.8415 | 0.8375 |
| | **SVM** | 0.8375 | 0.8404 | 0.8366 |
| | **NB** | 0.8375 | 0.8411 | 0.8376 |
| | **C4.5** | 0.8314 | 0. 8367 | 0.8321 |

Table 6 shows the accuracy performance of various machine learning algorithms using three different training-test ratios (60%-40%, 70%-30%, 80%-20%) for the real dataset and different sample sizes (N=500, 1000, 1500, 2000). On the actual dataset, the C4.5 algorithm showed the highest performance in terms of accuracy at 60% training - 40% test ratio. At sample sizes N=500 and N=1000, the Naive Bayes (NB) algorithm achieved the highest AUC value at 70% training - 30% testing ratio. At N=1500 and N=2000 sample sizes, the Naive Bayes (NB) algorithm performed the best in the accuracy criterion at 70% training - 30% testing ratio. In machine learning algorithms, the results are close to each other. Naive Bayes was found to be the best overall performer. The Naive Bayes algorithm was chosen to compare Cox regression. Figure 2 shows the AIC and BIC values obtained from Cox regression for leukemia and simulated data sets.

**Table 6.** *Accuracy Results for Classifying Algorithms*

| | | 60%-40% (Training-Test) | (70%-30%) (Training-Test) | (80%-20%) (Training-Test) |
|---|---|---|---|---|
| **Actual data** | **RF** | 0.196 | 0.2667 | 0.381 |
| | **SVM** | 0.006 | 0.014 | 0.184 |
| | **NB** | 0.15 | 0.1333 | 0 |
| | **C4.5** | 0.15 | 0.3333 | 0.1 |
| **N=500** | **RF** | 0.9865 | 0.9973 | 0.9917 |
| | **SVM** | 0.9916 | 1 | 0.9848 |
| | **NB** | 0.9973 | 1 | 0.9945 |
| | **C4.5** | 0.9749 | 1 | 0.9591 |
| **N=1000** | **RO** | 0.9959 | 0.9971 | 0.9966 |
| | **SVM** | 0.9936 | 0.9944 | 0.9899 |
| | **NB** | 1 | 1 | 1 |
| | **C4.5** | 0.9755 | 0.9789 | 0.9756 |
| **N=1500** | **RF** | 0.9953 | 0.9981 | 0.9976 |
| | **SVM** | 0.9888 | 0.9940 | 0.9896 |
| | **NB** | 1 | 1 | 1 |
| | **C4.5** | 0.9666 | 0.9736 | 0.9780 |
| **N=2000** | **RF** | 0.9982 | 0.9988 | 0.9989 |
| | **SVM** | 0.9925 | 0.9919 | 0.9949 |
| | **NB** | 1 | 1 | 1 |
| | **C4.5** | 0.9755 | 0.9784 | 0.9818 |

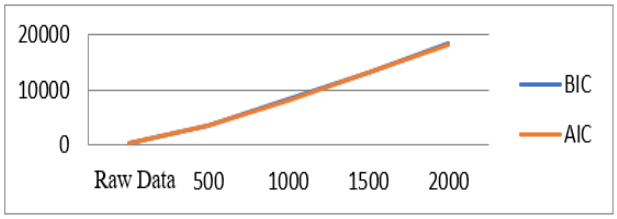


**Figure 2.** AIC and BIC values for leukemia and derive data

Lower values of AIC and BIC indicate a better model fit. The higher the sample size, the higher the AIC and BIC values. Table 7, Naive Bayes and Cox regression were compared for the actual (leukemia data) and the simulation data.

Table 7. AUC values of Naive Bayes and Cox Regression analysis for leukemia and simulation data

| | NaiveBayes | Cox Regression |
|---|---|---|
| Actual data | 0.650 | 0.705 |
| N=500 | 0.695 | 0.708 |
| N=1000 | 0.723 | 0.727 |
| N=1500 | 0.718 | 0.715 |
| N=2000 | 0.716 | 0.704 |

According to Table 6, the Cox regression model provided better results than the Naive Bayes algorithm (0.695 and 0.723, respectively) with AUC values of 0.708 for the n=500 dataset and 0.727 for the

n=1000 dataset. This indicates that Cox regression provides higher discrimination power in these two data sets. However, in the n=1500 data set, Naive Bayes is ahead of Cox regression (0.715) with an AUC value of 0.718. Similarly, in the n=2000 data set, Naive Bayes outperformed Cox regression (0.704) with an AUC value of 0.716. As a result, while Cox regression shows superiority in small data sets (n=500 and n=1000), Naive Bayes achieves better results in larger data sets (n=1500 and n=2000). This reveals that both methods show performance differences depending on the data set size and provide advantages for specific data set sizes.

## 4 Discussion

Studies in the field of health have been among the most researched in recent years. Censored data are seen in these data. In cancer studies with censored data, prognostic factors affecting survival times are determined by Cox regression analysis. Machine learning is a technique that has grown in popularity in recent years. Therefore, machine learning studies gain importance in cancer studies with censored data.

This study aims to show the use of machine learning for censored data. Simulation data were generated with a sample size of 500, 1000, 1500, and 2000 using the leukemia dataset. This process was repeated 100 times, and the results were averaged. After, Fleming–Harrington estimators were proposed to process censored data. The probability of survival and the probability of being censored were calculated using the Fleming-Harrington method. A new event variable is obtained by comparing these two probabilities by taking the event to occur and not to occur. Classification algorithms and Cox regression analysis were applied to these samples, and their performances were compared. When the actual dataset was split at an 80% training - 20% test data ratio, the C4.5 algorithm performed best with a 72% accuracy value. When the dataset for n=500, n=1000, and n=2000 was split at 70 % training, and 30% test data, Naive Bayes was the best algorithm with an approximately %85 accuracy ratio. When the actual data and n= 500 are compared, it is seen that the performance criteria have increased. When the amount of data increased from 500 to 2000, it was seen that the results were approximately the same. In machine learning algorithms, no one algorithm gives the best results. Performance varies according to the data set and separation rate. Therefore, different algorithms and separation rates should be tried, and the best algorithm of that set should be found. For the data set we have, there is not much difference between the separation rates and the performance of the algorithms. In most cases, Naive Bayes was chosen because it was better than other algorithms. For this data set, the best performance Naive Bayes algorithm was obtained. The success of this algorithm was compared with the classical method, Cox regression.

The superior performance of the Naive Bayes algorithm in this study can be attributed to several factors. Firstly, the dataset primarily consists of categorical variables, and Naive Bayes is particularly well-suited for such data types due to its ability to efficiently calculate class probabilities. Additionally, the algorithm's simplicity and assumption of independence between predictors allow it to generalize well, particularly in datasets with low multicollinearity. In contrast, more complex models like Random Forest or SVM may suffer from overfitting or computational inefficiency, especially with smaller sample sizes. Lastly, as the sample size increases, the Naive Bayes algorithm's performance remains robust, demonstrating its scalability and effectiveness in handling larger datasets.

As a result, the success of Naive Bayes increases as the number of samples increases. It has been seen that Naive Bayes gives better results as the number of samples increases. It has been seen that Naive Bayes gives better results as the number of samples increases.

# 5 Conclusion

Machine learning techniques have grown in popularity in recent years. Likewise, machine learning studies became increasingly important for cancer studies with censored data. The biggest problem is the processing of censored data. In this study, a new method of adaptation is proposed. This method is the Fleming-Harrington classifier. Censored data were reclassified using this estimator. Then, machine learning classification algorithms are compared. Different separation rates and classification algorithms compare results to achieve the best in machine learning. In this study, the results of the machine learning algorithm were found to be close to each other. In general, the Naive Bayes algorithm gave better results. This algorithm was compared with the classical method, Cox regression. As a result, Naive Bayes performs better as the sample size increases. This result also checks in with other studies in the literature [3-5]. It is possible to obtain better outcomes using machine learning approaches in survival analysis studies containing censored data. In recent years, there has been a significant rise in the application of machine learning techniques, particularly in cancer studies involving censored data. Addressing the challenge of censored data, this study introduces the Fleming-Harrington classifier as a novel method for reclassifying such data. By employing this estimator, the study compares various machine learning classification algorithms to determine their effectiveness. The results indicate that, despite the generally close performance across different machine learning algorithms, the Naive Bayes algorithm consistently provided superior results, especially as the sample size increased. This observation aligns with findings in the literature [5, 20], further validating the effectiveness of Naive Bayes in large datasets. The comparison was based on [specific metrics, e.g., accuracy, F-score, AUC], which showed that Naive Bayes outperformed the classical Cox regression method as the sample size grew. The study highlights that machine learning approaches, particularly Naive Bayes, can enhance predictive accuracy in the survival analysis of censored data. This advancement underscores the potential of machine learning to improve outcomes in cancer research and similar fields. Future research could explore additional machine learning methods or hybrid approaches to refine survival analysis techniques further. Additionally, addressing any limitations of the current study, such as [mention any specific limitations, e.g., dataset size, types of cancer studied], will be crucial for advancing the field.

## 5.1 Study Limitations

The fact that the study is based on a specific data set may limit the applicability of the findings in a broader context. However, these limitations may contribute to a more in-depth examination of the topic by providing an essential basis for future research.

## 5.2 Acknowledgments

This paper is derived from the first author's doctoral thesis.

## 5.3 Funding source

There is no funding source.

## 5.4 Competing Interests

There is no conflict of interest in this study.

## 5.5 Authors' Contributions

Study conception and design: Akın, Terzi. Data acquisition: Terzi. Data analysis and interpretation: Akın. Article drafting and revising: Akın, Terzi. All authors approved the final manuscript.2

### References

[1] P. B. Snow, D. S. Smith, and W. J. Catalona, "Artificial neural networks in the diagnosis and prognosis of prostate cancer: a pilot study," *The Journal of urology*, vol. 152, no. 5, pp. 1923-1926, 1994.

[2] D. Faraggi and R. Simon, "A neural network model for survival data," *Statistics in medicine*, vol. 14, no. 1, pp. 73-82, 1995.

[3] B. Zupan, J. DemšAr, M. W. Kattan, J. R. Beck, and I. Bratko, "Machine learning for survival analysis: a case study on recurrence of prostate cancer," *Artificial intelligence in medicine*, vol. 20, no. 1, pp. 59-75, 2000.

[4] M. J. Fard, P. Wang, S. Chawla, and C. K. Reddy, "A Bayesian Perspective on Early Stage Event Prediction in Longitudinal Data," (in English), *Ieee Transactions on Knowledge and Data Engineering*, vol. 28, no. 12, pp. 3126-3139, Dec 1 2016, doi: 10.1109/Tkde.2016.2608347.

[5] M. Billichová, L. J. Coan, S. Czanner, M. Kováčová, F. Sharifian, and G. Czanner, "Comparing the performance of statistical, machine learning, and deep learning algorithms to predict time-to-event: A simulation study for conversion to mild cognitive impairment," *Plos one*, vol. 19, no. 1, p. e0297190, 2024.

[6] W. Tizi and A. Berrado, "Machine learning for survival analysis in cancer research: A comparative study," *Scientific African*, vol. 21, p. e01880, 2023.

[7] S. Leger *et al.*, "A comparative study of machine learning methods for time-to-event survival data for radiomics risk modelling," *Scientific reports*, vol. 7, no. 1, p. 13206, 2017.

[8] M. Özbay Karakuş and O. Er, "A comparative study on prediction of survival event of heart failure patients using machine learning algorithms," *Neural Computing and Applications*, vol. 34, no. 16, pp. 13895-13908, 2022.

[9] E. L. Kaplan and P. Meier, "Nonparametric estimation from incomplete observations," *Journal of the American statistical association*, vol. 53, no. 282, pp. 457-481, 1958.

[10] J. P. Klein, "Small Sample-Moments of Some Estimators of the Variance of the Kaplan-Meier and Nelson-Aalen Estimators," (in English), *Scandinavian Journal of Statistics*, vol. 18, no. 4, pp. 333-340, 1991. [Online]. Available: <Go to ISI>://WOS:A1991HF92400006.

[11] E. A. Colosimo, F. F. Ferreira, M. D. Oliveira, and C. B. Sousa, "Empirical comparisons between Kaplan-Meier and Nelson-Aalen survival function estimators," (in English), *Journal of Statistical Computation and Simulation*, vol. 72, no. 4, pp. 299-308, Apr 2002, doi: 10.1080/00949650212847.

[12] G. A. Satten and S. Datta, "The Kaplan–Meier estimator as an inverse-probability-of-censoring weighted average," *The American Statistician*, vol. 55, no. 3, pp. 207-210, 2001. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5568678/pdf/nihms810169.pdf.

[13] A. Ihwah, "The Use of Cox Regression Model to Analyze the Factors that Influence Consumer Purchase Decision on a Product," (in English), *International Conference on Agro-Industry (Icoa): Sustainable and Competitive Agro-Industry for Human Welfare Yogyakarta-Indonesia 2014*, vol. 3, pp. 78-83, 2015, doi: 10.1016/j.aaspro.2015.01.017.

[14] D. R. Cox, "Regression models and life-tables," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 34, no. 2, pp. 187-202, 1972.

[15] S. B. Kotsiantis, "Decision trees: a recent overview," (in English), *Artificial Intelligence Review*, vol. 39, no. 4, pp. 261-283, Apr 2011, doi: 10.1007/s10462-011-9272-4.

[16] X. Jinguo and X. Chen, "Application of decision tree method in economic statistical data processing," in *E-Business and E-Government (ICEE), 2011 International Conference on*, 2011: IEEE, pp. 1-4.

[17] Vikramkumar, B. Vijaykumar, and Trilochan, "Bayes and Naive Bayes Classifier," *Computer Science & Engineering. Rajiv Gandhi University of Knowledge Technologies Andhra Pradesh, India,* 2014.

[18] R. R. Yager, "An extension of the naive Bayesian classifier," (in English), *Information Sciences,* vol. 176, no. 5, pp. 577-588, Mar 6 2006, doi: 10.1016/j.ins.2004.12.006.

[19] O. T. Bişkin, M. Kuntalp, and D. G. Kuntalp, "Classification of arrhythmias according to the energy spectral density features by using Kernel density estimation," in *Biomedical Engineering Meeting (BIYOMUT), 2010 15th National*, 2010: IEEE, pp. 1-4.

[20] C. Friedman and S. Sandow, *Utility-based learning from data* (Machine learning & pattern recognition series.). Boca Raton: Chapman & Hall/CRC, 2011, p. 397 p.

[21] A. Liaw and M. Wiener, "Classification and regression by randomForest," *R news,* vol. 2, no. 3, pp. 18-22, 2002.

[22] Ş. Haciefendioğlu, "Makine öğrenmesi yöntemleri ile glokom hastalığının teşhisi," Selçuk Üniversitesi Fen Bilimleri Enstitüsü, 2012.

[23] V. Vapnik, "Principles of risk minimization for learning theory," *Advances in neural information processing systems,* vol. 4, 1991.

[24] E. Alpaydin, *Machine learning : the new AI* (MIT Press essential knowledge series.). pp. xv, 206 pages.

[25] S. Uğuz, "Makine öğrenmesi teorik yönleri ve Python uygulamaları ile bir yapay zeka ekolü," *Nobel Yayıncılık. Ankara,* 2019.

[26] A. Dirican, "Kliniğimizde akciğer kanseri tanısı alan hastaların prospektif olarak değerlendirilmesi ve sağkalıma etki eden faktörlerin belirlenmesi " Tıpta Uzmanlık, Ondokuz Mayıs University, 2004.

[27] S. v. Buuren and K. Groothuis-Oudshoorn, "mice: Multivariate imputation by chained equations in R," *Journal of statistical software,* pp. 1-68, 2010.