v. 4, n. 12, 2024

International Journal of Health Science

SURVIVAL ANALYSIS APPLYING COX MODEL AND MACHINE LEARNING TO COVID-19 DATA IN THE CITY OF BUCARAMANGA BETWEEN MARCH 2020 TO MARCH 2023

Giann Axel Leguízamo Jordán
Universidad El Bosque
Bogotá D.C
https://orcid.org/0009-0009-0503-342X

Emiliano Rodríguez Arango

Abstract: This study performs a comparative analysis of the performance of Machine Learning, neural networks and traditional survival analysis techniques. The techniques compared are the traditional Cox proportional hazards (CPH) model, Machine Learning Random Survival Forest (RSF) technique and neural networks such as DeepSurv. These techniques are applied to the study of Covid-19 cases of patients diagnosed in the city of Bucaramanga between March 2020 and March 2023. This study demonstrates better performance was obtained with the Random Survival Fores technique in predicting the survival function measured through the C – index, Brier score and AUC.
Keywords: Survival analysis; Covid-19; Random Survival Forest; proportional hazards; Deepsurv.

INTRODUCTION

The health emergency declared by Covid-19 worldwide by the WHO has perhaps been that milestone in the history of all governments, which allows us to determine the socioeconomic implications of it, in terms of international public health; that beyond the damages and losses that its nature reveals, it brought with it effects after the social confinement to which the world population was subjected and its implications that are recorded in its high mortality rates as well as the survival rates compared. by the different entities allied to this problem.

Talking about survival in a contingency like this means talking about all those socioeconomic efforts formulated and implemented for periods of time by government entities in conjunction with related health authorities, in order to favor the implementation of measures that safeguard lives, regardless of the situation. the robustness of the effectiveness of the treatments carried out or their secondary implications.

This article provides an overview of survival analysis in Covid-19 research, which is intended to provide a practical tool for understanding the variety of statistical approaches to address outcomes observed during the time to event.

In this study, prediction models are developed using a Cox proportional hazards model, Random Survival Forest, and DeepSurv. Their performance is compared in terms of time-dependent C-index, Brier score and area under the AUC curve.

DATA

The data analyzed corresponds to 141,394 patients with Covid-19 who attended the medical service in the city of Bucaramanga between March 2020 and March 2023, data that was taken from the digital platform "Colombian Open Data", where a follow-up of patients from the beginning of treatment until their death or event of interest, or until the end of the study, so some observations are censored.

STATISTIC ANALYSIS

During the development of the process, the survival function was estimated by applying the estimator (Kaplan & Meier, 1958). Comparing the variables of sex and age. There, several Cox models (1972) were adjusted in order to obtain significant covariates, eliminating non-significant variables. Then, a residual analysis was applied to the final models to verify the assumptions of the model of (Cox, 1972).

ELEMENTS OF SURVIVAL ANALYSIS THEORY

The purpose of survival analysis is to verify the follow-up time until the occurrence of the event of interest, highlighting its application at the time when censored observations are evident. According to Rebasa, (2005), there are different types of censorship: right, left

and intervals. In this case, censorship with common application or right-wing censorship will be addressed, which occurs until the last moment in which the individual has been followed and at that moment the event of interest has not yet occurred.

BASIC DEFINITIONS

SURVIVAL FEATURES

Considered as any probability that a person will survive or that an event of interest will not occur, at least until time t. Likewise, the survival function is given under the condition that T is a positive (or non-negative) random variable with distribution function F(t) and probability density function f(t). The survival function S(t) Arribalzaga, E. B. (2007).

$$S(t) = 1 - F(t) = P[T > t]$$

HAZARD RATIO FUNCTION (HAZARD RATE)

Denoted as the instantaneous failure rate $\lambda(t)$ and specifies the quotient between the density function and the survival function, expressed as follows:

$$\lambda(t) = \frac{f(t)}{S(t)}$$

It is explained as the probability that the event of interest occurs to an individual in the next unit of time.: $\Delta(t)$, and that this probability has occurred until the time:

This function is derived from the average failure rate, which gives us the conditional probability of failures in the period (t; t + Δt) since the person survives in the period (0;t), the average failure rate (AMR), said by Estévez, G. and Quintela, A. (2001), is defined as:

$$TMF = \frac{F(t+\Delta t) - F(t)}{\Delta t}\frac{1}{S(t)}$$

KAPLAN AND MEIER ESTIMATOR

The most widely used estimator of the survival function is that of Kaplan & Meier (1958), expressed as follows:

$$\hat{S}_{K\,M}(t) = \prod_{t_i \leq t} \frac{r(t_i) - d(t_i)}{r(t_i)}$$

Where, $r(t_i)$ y $d(t_i)$ are the number of individuals at risk and the number of deaths (or occurrence of the event of interest at the time: t_i). The confidence interval, calculated by default by statistical programs, affects the identity interval or flat scale, given for a confidence level of 90% by the following expression:

$$\hat{S}^2_{K\,M}(t) \pm 1.645\, ee\left(\hat{S}_{K\,M}(t)\right)$$

Where, $ee(\hat{S}_{KM}(T))$, is the standard error of estimation of the Kaplan & Meier estimator.

LOG RANK TEST (LOG RANK)

Based on the same assumptions as the Kaplan Meier survival curve, according to Bland JM, Altman DG (1998), so censoring is not related to prognosis; Survival probabilities are the same for subjects recruited at the beginning and end of the study, and for events occurring at the specified times. Deviations from these assumptions are more important if they hold differently in the groups being compared, for example, if censoring is more likely in one group than another. Bouliotis, G., & Billingham, L. (2011).

COX REGRESSION MODEL

The Cox regression model (1972) is the most used for survival data in the medical area. In this, the risk for the ith individual is expressed as follows:

$$\lambda\left(t; Z_i(t)\right) = \lambda_0(t) e^{B' Zi(t)}$$

Where, Zi(t) is the vector of covariates for the ith individual at time t. Also called a semiparametric model, since it includes a

parametric part and a non-parametric part:

1. The parametric part is expressed as,*ri(t)* = $e^{B`Zi(t)}$, so-called risk score (risk score) yβ is the regression parameter vector.
2. The non-parametric part is expressed as, $\lambda_0(t)$, called basis risk function and is an arbitrary unspecified function.

The Cox regression model, also called the proportional hazards model, since the ratio between the risk for two subjects with the same vector of covariates is constant over time, and is expressed as follows:

$$\frac{\lambda(t;Z_i(t))}{\lambda(t;Z_j(t))} = \frac{\lambda_0(t)e^{B\prime Z_i(t)}}{\lambda_0(t)e^{B\prime Z_j(t)}} = \frac{e^{B\prime Z_i(t)}}{e^{B\prime Z_j(t)}}$$

Where, if a death has occurred at time t^*, it is related to the likelihood that the death will occur to the ith individual and not to another, expressed as follows:

$$L_i(\beta) = \frac{\lambda_0(t^*)r_i(t^*)}{\sum_j Y_j(t^*)\lambda_0(t^*)r_j(t^*)} = \frac{r_i(t^*)}{\sum_j Y_j(t^*)r_j(t^*)}$$

As it is, the product: L(B) = ∏ L_i(B) or partial likelihood and the maximization of Log(L(B)) gives an estimate for the *B* without needing to estimate the noise parameter the λ_0(t)

PROPORTIONAL HAZARDS ASSUMPTIONS

STATISTICAL METHOD THROUGH HYPOTHESIS TESTING

This method offers greater reliability, since graphic inspection can often be misleading. It is done through a linear correlation between the scaled Schoenfield residuals and the study time. Under the null hypothesis that the absence of correlation (Ho = p = 0) is synonymous with proportional risks. Values p<0.05 mean rejecting the null hypothesis of proportional hazards (Domenech, Navarro 2008).

HYPOTHESIS CONTRAST FOR COX MODEL

After adjusting the Cox model, there are three hypothesis tests that aim to verify the significance of the model. These tests are characterized by being asymptotically equivalent. However, this characteristic is not always externalized in the practical exercise of its implementation.

- Wald test
- Likelihood ratio test
- Score test

RANDOM SURVIVAL FORESTS

Random forests are one of the most interesting machine learning techniques for classification and regression. This technique was applied to survival analysis. (Ishwaran, Kogalur, Blackstone, Lauer. 2008)

An advantage of this model is that it is completely nonparametric and therefore does not assume a distribution for the relationship between the predictors and the response variable. Furthermore, it captures the linear and non-linear relationships between the explained variable and the predictor variables. Perhaps, another important characteristic is that it finds interactions between variables because the learning comes from the set of decision trees (Ishwaran, Kogalur, Blackstone, Lauer. 2008)

The RSF method is an extension of Breiman's random forest method for right-censored survival data by using a forest of survival trees for prediction. Similar to regression and classification configurations, RSF is an ensemble learner formed by averaging a tree base learner. In survival environments, a binary survival tree is the base learner, and the ensemble learner is formed by averaging the Nelson - Aalen cumulative hazard function of each tree (Ishwaran, Kogalur, Chen, Minn. 2011).

This model does not require the

CPH model's assumption that there are proportional risks between individuals, but rather allows survival functions to be constructed with different shapes for each insured. Furthermore, the assumption of the same basic hazard rate for all insured is avoided because it is inconsistent with reality (Hothorn, Lausen, Benner. 2004).

Survival trees are constructed by splitting each parent node into two child nodes from the root, which comprises the entire data set. A division is performed according to a survival criterion that maximizes the difference between the child nodes; this division is repeated in each subsequent node in a binary manner.

HYPERPARAMETERS

There are several tunable hyperparameters to consider when training a model. The main ones include:

1. The number of trees in the forest.
2. The number of functions to be considered in any given division.
3. The division rule to use during tree construction.

DEEPSURV

DeepSurv is a deep feedback neural network that predicts the effects of a patient's covariates on their hazard rate parameterized by network weights θ. The input to the network is the reference data of patient x. The hidden layers of the network consist of a fully connected node layer, followed by a dropout layer. (Katzman, Shaham, Cloninger, Bates, Jiang, Kluger. 2018). Network output: $\widehat{h_\theta}(x)$. It is a single node with a linear activation that estimates the log-hazard function in the Cox model. We train the network by setting the objective function as the average negative log partial likelihood with regularization:

$$l(\theta) \coloneqq \frac{1}{N_{E=1}} \sum_{i:E_{i=1}} \left(\widehat{h_\theta}(X_i) log \sum_{j \in \mathbb{R}(T_i)} e^{\widehat{h_\theta}(X_j)} \right) + \lambda * ||\theta||_2^2$$

Where, $N_{E=1}$ is the number of patients with an observable event and is the regularization parameter: ℓ_2. We then use gradient descent optimization to find the network weights that minimize the equation.

METHODOLOGY

A quantitative analysis that allows obtaining Covid-19 data, where the measurement system is through the screening technique and antigen and PCR tests.

Obtaining the source information lies in choosing those data that were processed and analyzed, in order to estimate the survival function using the Kaplan & Meier (1958) estimator and the techniques used to compare the C - index of each model.

1. Understanding of the data and formulation of the problem. Identify the set of data with which you are going to work, understand the limitations at the level of data availability (NA, etc.), be able to understand the different variables that are included in the base, and the possibility of developing new attributes that serve to improve the understanding and disposition of data for subsequent analysis.
2. Development of survival analysis through each model. Once the database is analyzed, prepared, and understood, the survival analysis will be developed through different models, which take into account the information from different perspectives. To do this we will use non-parametric, semi-parametric and Machine Learning models. We will analyze the behavior/performance of each of the models in relative terms.
3. Comparative analysis of the models. Once the analyzes defined in the previous stage are available, the Machine Learning Random Survival Forest model is used.

International Journal of Health Science ISSN 2764-0159 DOI https://doi.org/10.22533/at.ed.15941224230110

Under the assumptions that performance will have relevant differences, a comparison will be made based on the main performance indicators of the statistical models.

4. Conclusion and recommendation. Depending on the results, different conclusions and recommendations will be drawn up for future theoretical and empirical work. As a summary, an outline of the methodological process is presented:

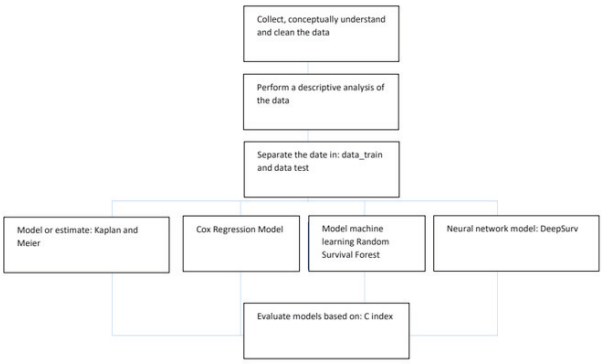


Figure 1. Flowchart methodology

Source: Own elaboration

RESULTS

KAPLAN AND MEIER ESTIMATOR

The values obtained with the Kaplan & Meier estimator with the interval, the standard error and the 95% confidence interval can be seen in Table 1. Performed with the RStudio statistical software.

Table 1 includes the following information:

- Time: Number of months of follow-up
- Risk number: Number of individuals at risk before time.
- Number of events: Number of deaths between the time and the following week in which a death occurs.
- Survival: Probability that an individual will survive for a greater number of weeks than the time.
- Err.st: Standard error of survival.
- LCI 95%: Lower 95% confidence limit for survival.
- LCS 95%: Upper 95% confidence limit for survival.

The probability of survival of patients with covid-19 until week 15 is 75%, 47% manage to survive up to 24 weeks, 25% survive until week 31 and 0.0493% of patients survive more than 53 weeks. The mean survival time was 24 weeks, with a standard error of 1.33e-03 weeks.

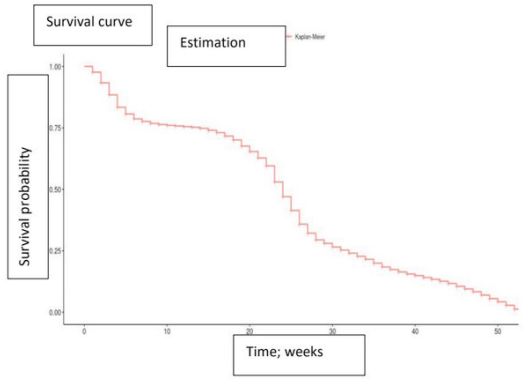


Figure 2. Covid-19 survival curve with the Kaplan & Meier estimator

Source: Own elaboration

MODELS TO BE COMPARED

We compare the performance of survival models for a Covid-19 dataset consisting of survivors and deceased patients. The study is applied to an open data set called "Covid-19" to compare the performance of the models:

- Cox model
- Random Survival Forest
- DeepSurv

COX MODEL

The best Cox model adjusted for the Covid-19 data and death as an event of interest is shown in Table 2. It can be stated that the variable age and sex are significant at the 5% level, because the p-value is less than 0.05.

Time	Number of risk	Number of events	Survival	Err.est	Lower 95% CI	Upper 95% CI
1	141393	3279	0.976809	4.00e-04	0.97601	0.977581
2	138034	6144	0.933331	6.64e-04	0.93202	0.934619
3	131798	6911	0.884390	8.51e-04	0.88271	0.886047
4	124748	7140	0.833772	9.91e-04	0.83182	0.835704
5	117456	3855	0.806407	1.05e-03	0.80434	0.808459
6	113472	2795	0.786544	1.09e-03	0.78440	0.788673
7	110607	1418	0.776460	1.11e-03	0.77428	0.778625
8	109139	1140	0.768350	1.12e-03	0.76614	0.770543
9	107962	653	0.763702	1.13e-03	0.76148	0.765911
10	107284	469	0.760364	1.14e-03	0.75813	0.762583
11	106804	347	0.757893	1.14e-03	0.75565	0.760121
12	106445	383	0.755166	1.15e-03	0.75291	0.757402
13	106049	387	0.752411	1.15e-03	0.75015	0.754655
14	105645	585	0.748244	1.16e-03	0.74597	0.750501
15	105033	1090	0.740479	1.17e-03	0.73818	0.742759
16	103915	1349	0.730866	1.18e-03	0.72854	0.733174
17	102514	1987	0.716700	1.20e-03	0.71434	0.719046
18	100450	2202	0.700989	1.22e-03	0.69859	0.703373
19	98153	3522	0.675836	1.25e-03	0.67338	0.678274
20	94491	3074	0.653849	1.27e-03	0.65136	0.656329
21	91327	3680	0.627503	1.29e-03	0.62497	0.630024
22	87505	4452	0.595577	1.31e-03	0.59300	0.598139
23	82927	9122	0.530063	1.33e-03	0.52745	0.532673
24	73626	8303	0.470287	1.33e-03	0.46767	0.472900
25	65171	7811	0.413921	1.32e-03	0.41134	0.416504
26	57190	7723	0.358025	1.28e-03	0.35551	0.360543
27	49294	5017	0.321586	1.25e-03	0.31913	0.324042
28	44163	3743	0.294330	1.22e-03	0.29193	0.296729
29	40329	1924	0.280288	1.21e-03	0.27793	0.282654
30	38336	2053	0.265278	1.19e-03	0.26296	0.267605

Table 1. Kaplan and Meier estimate

Source: Own elaboration

Covariable	Coefficient	p-value
Age	0.069523	<2e-16
Gender	-0.003668	<2e-16

Table 2. Significant covariates

Source: Own elaboration

This model is significant by any of the two criteria for a 5% significance level, because the p-values are all less than 0.05. For the Wald test it was <2e-16.

PROPORTIONAL HAZARDS ASSUMPTIONS

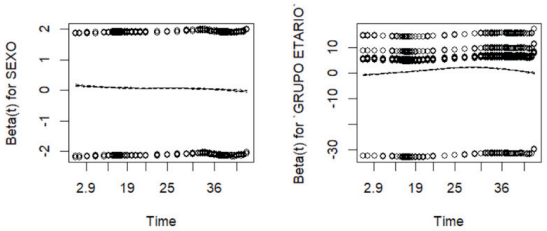


Figure 3. Verification of the assumptions of the Cox model

Source: Own elaboration

It must be noted that systemic deviations from a horizontal line indicate that the proportional risk assumption is not met, since proportional risks assume that the estimates of the β's coefficients do not vary much over time.

Verification of the proportional hazards assumption can be carried out through a hypothesis test, where the null hypothesis is associated with compliance with the proportional hazards assumption. The results of this contrast indicate that the risk assumption is violated. Therefore, we can say that Ho is rejected and there is no proportionality.

The main component of the Cox proportional model is the proportionality assumption, in this case we observe that the proportionality assumptions are not met, it is possible that the large sample size is responsible for the apparently strong evidence against the PH assumptions, the p-values are a function of the sample size and their usefulness decreases when the sample size grows a lot (Talavera, Rivas, Bernal 2011). Some alternatives are classic methods such as stratify and Aalen model, on the other hand, more recent models such as random forests and neural networks.

The calculation of the stratified cox model was carried out, where it was also possible to identify that it does not comply because the proportionality assumption is violated.

Since the assumption of proportional hazards is not met, the idea that some covariates depend on time can be considered. As an alternative, we calculate the additive Aalen model, in this case (AGE GROUP and SEX) using the timereg library and the Aalen function in the statistical software. RSTUDIO, this gives us an output of a Cramer Vos Mises and Kolmogorov-Smirnov hypothesis test to validate the effects with respect to time, where the null hypothesis is rejected and has an effect that depends on time, which is evident in Table xx where the age group category that depends on time is early childhood, adulthood and older people according to the p-value.

Furthermore, in this model the covariate has separate effects on the response variable. For this reason it is called the additive risk model or Aalen additive model (Alayo, 2016).

On the other hand, rejecting the null hypothesis for a covariate means that there is sufficient evidence to say that this covariate is significant, therefore, it will remain as part of the model.

	Kolmogoorv-Smirnov test	**p-valor Ho: Constant effect**
Intercept	2.820	0.000
Gender: women	0.117	0.730
Age group - Adulthood	0.680	0.019
Age group – Childhood	0.467	0.679
Age group – Young people	0.508	0.184
Age group – Older	0.681	0.050
Age Group – Early Childhood	1.980	0.004

Table 3. Test Kolmogorov Smirnov

Source: Own elaboration

In Table 3, it is statistically evident that the age group covariate depends on time in some categories, because as time passes the patients possibly changed categories.

RANDOM SURVIVAL FOREST

A survival random forest ensures that individual trees are uncorrelated by building each tree on a different bootstrap sample of the original training data and at each node, evaluating only the splitting criterion for a randomly selected subset of features and thresholds.

To demonstrate Random Survival Forest, we use Covid-19 data in the city of Bucaramanga in a period from March 2020 to March 2023, on the treatment of diagnosed

patients. It contains data from 75,959 women, 65,434 men and 4 prognostic factors: 1. age, 2. year, 3. sex, 4. age group.

The goal is to predict recurrence-free time in the following steps:

1. We load the data
2. We divide into 75% for training and 25% for testing, so that we can define how well our model generalizes.
3. The most widespread criterion is based on the log-rank test, which you are probably familiar with when comparing survival curves between two or more groups. Using the training data, we fit a random survival forest comprising several tests with 50, 100, 200, and 500 trees.
4. Check how well the evaluated model works with the test data.
5. Estimate C index, AUC and Brier score.

Hyperparameter	Value
Number of trees	50, 100, 200, 500
Variables used in division	4
Division rule	Log-rank

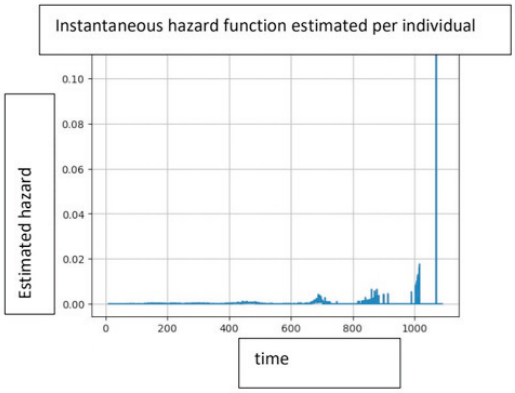
Table 4. Hyperparameters

Source: Own elaboration

MODEL TRAINING

Random SurvivalForest (min_samples_leaf=10, min_samples_split=7, n_jobs=-1, random_state=1234, verbose=1)

RESULTS OF THE MODEL



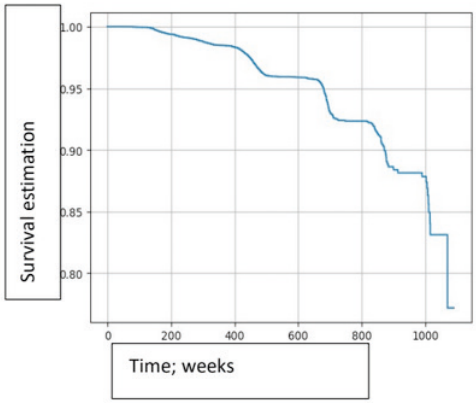


Figure 4. Survival estimation and Hazard function

Source: Own elaboration

INDEXES

In survival analysis, a common way to evaluate a model, can be done by calculating the probability of agreement or the concordance index C index, Brier score and AUC (Gönen, Heller 2005).

Concordance	Brier score	AUC
0.863	0.918	0.607

Table 6. C-index, AUC and Brier Score Random Survival Forest model

Source: Own elaboration

We also obtained a concordance index of 0.863, Brier score of 0.918 and AUC of 0.607, considering that it is a strong model indicating the capacity of the model to correctly provide a reliable classification of survival times based on individual risk scores.

DEEPSURV

Personalized treatment recommendation system using a Cox proportional hazards deep neural network.

We performed a series of experiments training DeepSurv with real and simulated survival data. We demonstrate that DeepSurv performs as well as or better than other state-of-the-art survival models and validate that

DeepSurv successfully models increasingly complex relationships between a patient's covariates and her risk of failure.

The network has multiple hidden layers, and the number of nodes in each layer is determined by the node list. In this case, all hidden layers have 256 nodes. Therefore, the network architecture would be a fully connected network with four hidden layers, each with 256 nodes.

Input and output:

The number of input features is determined by the way of training. The network expects each training example to have this number of features.

The output layer has a single node, since out features is equal to 1. This suggests that the network is designed for regression, where a continuous value is predicted (in this case, possibly related to survival time).

Batch Normalization:

This implies that after each hidden layer, the output values are normalized before moving to the next stage, this helps to stabilize and speed up the training of the network.

Dropout:

It is applied at a rate of 0.4 after each hidden layer. Dropout is a regularization technique that helps prevent overfitting by randomly "turning off" some nodes during training.

Output Bias:

The output layer has a bias (output_bias). If true, a bias will be added to the output layer.

Optimizer:

The Adam optimizer is used for training. Adam is a popular optimization algorithm in deep learning.

We defined the neural network

n_nodes = 256

in_features = x_train.shape[1]

num_nodes = [n_nodes, n_nodes, n_nodes, n_nodes]

out_features = 1

batch_norm = True

dropout = 0.4

output_bias = False

net_ds = tt.practical.MLPVanilla(in_features, num_nodes, out_features, batch_norm, dropout, output_bias = output_bias

model_ = CoxPH(net_ds, tt.optim.Adam

CURVE: KAPLAN & MEIER

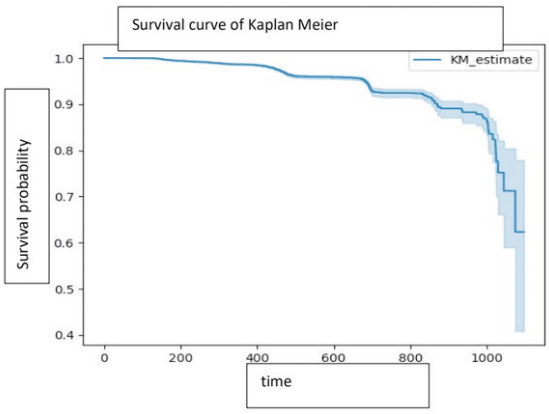


Figure 5. Kaplan and Meier survival curve

Source: Own elaboration

INDEXES

In survival analysis, a common way to evaluate a model, can be done by calculating the probability of agreement or the concordance index C index, Brier score and AUC (Gönen, Heller 2005).

Concordance	Brier score	AUC
0.428	0.918	0.586

Table 7. Concordance index DeepSurv Model

Source: Own elaboration

We obtain a concordance index of 0.428, Brier score of 0.918 and AUC of 0.586, considering that it is a poor model since its C-index is below 0.5 and the Brier score value is close to one, with one being the worst model. and the AUC reflects the rate of true positives as a percentage of 58.6% of events that occur.

MODEL EVALUATION

We compare the performance of the above-mentioned models through the concordance index.

	C - Index
Cox Proportional Hazard Model	0,517
Random Survival Forest	**0.863**
DeepSurv	0.428

Table 8. Concordance index evaluation
Source: Own elaboration

Table 8 shows the concordance indices of the models trained on the Covid-19 data set. The reference models, CoxPH and DeepSurv, start with decent performance, but are outperformed by the Random Survival Forest model that achieved the higher agreement index indicating that it is a strong model.

DISCUSSION

Survival analysis is a very powerful tool for modeling event-time data and the most suitable for censored data, it is poorly linked to statistical study programs.

We must highlight that the field of action of survival analysis is not only linked to the medical area, but to any area where we want to determine the functions of the time elapsed from the moment at which the follow-up of a group of individuals begins until the an event of interest, and if the event of interest does not occur, the observations are censored.

This document presents the estimation of survival functions, obtaining predictive covariates of the survival function, verification of the assumptions of the Cox model and the application of techniques such as Cox Propotional, Random Survival Forest and DeepSurv.

The analysis presented was performed in R-STUDIO and PYTHON statistical software.

CONCLUSIONS

Classic survival analysis is suitable for estimating survival functions and adjusting regression models to obtain significant covariates, which is evidenced in this work.

DeepSurv's prediction and modeling capabilities will enable medical researchers to use deep neural networks as a tool in their exploration, understanding and prediction of the effects of a patient's characteristics on their risk of failure.

In this study, four models are applied to estimate the survival function of patients diagnosed with Covid-19 and the predictive power of each of them is evaluated using the C-index measure, AUC and Brier score, obtaining better performance with the technique. by Random Survival Forest.

The cox model was applied, which does not comply with the proportional hazards assumption due to the amount of data and some categories of the age group variable that had a time-dependent effect; on the other hand, when comparing the cox model with some models of machine learning Random Survival Forest and the DeepSurv neural network model, it was found that the model with the best performance was RSF with a performance of 0.863.

REFERENCES

Alayo Bueno, I. (2016). *El modelo aditivo de Aalen. Una alternativa al modelo de riesgos proporcionales* (Master›s thesis, Universitat Politècnica de Catalunya).

Arribalzaga, E. B. (2007). Interpretación de las curvas de supervivencia. *Revista chilena de cirugía*, *59*(1), 75-83.

Bland JM, Altman DG. Probabilidades de supervivencia. El método Kaplan-Meier. *BMJ* 1998; 317 : 1572.

Bouliotis, G., & Billingham, L. (2011). Crossing survival curves: alternatives to the log-rank test. *Trials*, *12*(1), 1-1.

Cox, D. (Arribal), 'Regression models and life tables (with discussion)', Journal of the Royal Statistical Society: Series B (34), 187–220.

Domenech Massons JM, Navarro Pastor JB. Análisis de la supervivencia y modelo de riesgos proporcionales de Cox. Editorial Signo. Barcelona. 2008.

Estévez, G. y Quintela, A. (2001). Estimación no paramétrica de la función de riesgo: Fleming, D. P. (1991), Counting Processes and Survival Analysis, John Wiley & Sons, Inc., N.Y.

Gönen, M. y Heller, G. (2005). Probabilidad de concordancia y poder discriminatorio en regresión de riesgos proporcionales. *Biometrika* , *92* (4), 965-970.

Hothorn T, Lausen B, Benner A, Radespiel-Troger M. Embolsado de árboles de supervivencia. *Estadísticas médicas.* 2004; 23 : 77–91.

Hougaard, P. (1995), 'Frailty models for survival data', Lifetime Data Analysis (1), 255–273.

Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Bosques de supervivencia aleatorios. *Estadísticas de aplicaciones de Ann.* 2008; 2 : 841–860.

Ishwaran H, Kogalur UB, Chen X, Minn AJ. Bosques de supervivencia aleatorios para datos de alta dimensión. *Stat Anal Data Mining ASA Data Sci J.* 2011; 4 :115–132.

Kaplan, E. & Meier, P. (1958), 'Nonparametric estimation from incomplete observations', Journal of the American Statistical Association (53), 457–481.

Katzman, J. L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., & Kluger, Y. (2018). DeepSurv: Sistema personalizado de recomendacion de tratamiento que utiliza una red neuronal profunda de riesgos proporcionales de Cox. Metodologia de investigacion medica BMC, 18(1), 24.

Rebasa, P. (2005). Conceptos básicos del análisis de supervivencia. *Cirugía española*, *78*(4), 222-230.

Talavera, J. O., Rivas-Ruiz, R., & Bernal-Rosales, L. P. (2011). Investigación clínica V. Tamaño de muestra. *Revista Médica del Instituto Mexicano del Seguro Social*, *49*(5), 517-522.

Therneau, T. & Grambsch, P. (2000), Modeling Survival Data: Extending the Cox Model, Springer-Verlag, N.Y.

Therneau, T., Grambsch, P. & Fleming, T. (1990), 'Martingale-based residuals for survival models', Biometrika (77), 147–160.