Generalized Informative Discrimination Measure and Its Properties with Applications

D.K. Sharma¹* and Sonali Saxena²

¹Jaypee University of Engineering and Technology, A.B. Road, Raghogarh, Dist. Guna, M.P., India.
²Government Shrimant Maharaja Madhav Rao Sindhiya College, Kolaras, Distt. Shivpuri -473770 (M.P.) India.
¹dilipsharmajiet@gmail.com,

*Corresponding Author

Abstract. The Shannon interval uncertainty was suggested as a convenient functional uncertainty metric for dual-sided shortened latent variables in the composite reliability. A new measure of disparity has been recently suggested between two doubly truncated distributions of life. The generalised informative discrimination measure for lifetime distribution in the given time interval is described in the present paper. Few features of the new generalised measure are also being investigated. An analysis of lung-cancer data is done using survival function and Kaplan Meier estimator using Python libraries.

Keywords: Discrimination Measure; Interval Entropy; Survival Function; Mathplotlib; Pandas; Numpy; Kaplan Meier Estimator; Machine Learning

AMS Classification: 62N05, 62B10

1. Introduction

[5] and [3] in 1996 concluded that the data measure of Shannon is an effective method for calculating the volatility and reliability of spontaneous lifetime distributions. In particular, the mainstream narrative on information measures is essential for the form of divergence and the kind of entropy measure of data. Entropy/divergence acts as a fundamental feature in all data measurements. The most common entropy and divergence measures [13] are also the definition of information-theoretical entropy by [12] as well as [10] and the relative entropy or estimate of divergence by [6]. Shannon information was used to calculate diffusion, risk, and uncertainty, while the estimate of divergence was applied for assessing the extent among probability distributions and has a major role in assumption and discrimination issues. Some stochastic processes were also studied by [11]. The possible exertion of information and divergence method can be found in econometric measurements in the literature.

It is to be assumed that X and Y be positive valued random variables which are totally continuous that define that lifetime of two objects. These can be the intermediate stages of malfunction, including the left and right organs, for instance, or the failure periods of several computer components. Let the pdf, cdf as well as the survival function of X be \( f(t), F(t) \) and \( \bar{F}(t) = 1 - F(t) \) respectively. Let the functions \( g(t), G(t) \) or \( \bar{G}(t) \) be the same in relation to Y.

Kullback and Leibler [6] suggested the measure of discrimination which is given below and also known as relative information of X and Y, as an information separation between two random variables F and G:

\[
I_{XY} = \int_{0}^{\infty} f(x) \log \frac{f(x)}{g(x)} \, dx. \tag{1.1}
\]

(1.1) is the generalization of Shannon’s differential entropy of X given below:

\[
H(f) = \int_{0}^{\infty} f(x) \log f(x) \, dx. \tag{1.2}
\]

Here natural logarithm is denoted by ’log.’ Distance (1.1) is symmetric with shift and size. Ullah (1996) observed that if \( g(x) \) is a constant density, it’ll become (1.1) (1.2). In testing hypotheses and model assessment, Kullback [7] heavily generated its use. It has been extensively used during statistics and economics ever since.

In general, a time-dependent (dynamic) type of Kullback-Leibler discrimination information was described by [4] as non-negative life time experience data of systems X and Y at time t.

\[
I_{XY}(t) = \int_{t}^{\infty} f(x) \log \frac{f(x)}{G(t)} \, dx. \tag{1.3}
\]

© IJAICT India Publications 2021

Suman Rajest S. et al.(eds.). ICT based Framework for Data Science and Machine Learning Applications, https://doi.org/10.46532/978-81-950008-7-6_002
Roughly equivalent to (1.1), \( I_{X,Y}(t) \) thus distinguishes \([X - t | X > t] \) and \([Y - t | Y > t] \) with relative information. The knowledge measure (1.3) is useful for collecting the residual lifetimes of two remaining variables prior to period \( t \). A further measure of knowledge was described by ([1], [2]). \( I_{X,Y}(t) \) Dual to (1.3) in the context that it is indeed a measure of knowledge inequality among two previous lives \([X | X \leq t] \) and \([Y | Y \leq t] \). It is defined as

\[
I_{X,Y}(t) = \int_{t}^{1} \log \frac{f(x)}{g(x)/|g(t_1)|} \, dx, \quad t > 0
\]  

(1.4)

Since both structures were observed depressed at time \( t \), \( I_{X,Y}(t) \) tests the insightful distance among their previous lives. Currently, on the comparable lines identified by the latest discrepancy measure what concludes the analysis of the knowledge gap between random lifetimes \( X \) and \( Y \), ([8],[9]).

**Definition 1.1** The relative information measure between the compressed lives \((X|t_1 < X < t_2)\) and \((Y|t_1 < Y < t_2)\) is the interval distance among random lifetimes \( X \) and \( Y \) at \((t_1, t_2)\) interval:

\[
I_{X,Y}(t_1, t_2) = \int_{t_1}^{t_2} \log \frac{f(x)/|f(t_2)|}{g(x)/|g(t_1)|} \, dx
\]  

(1.5)

Clearly, \( I_{X,Y}(0, t) = I_{X,Y}(t) \), \( I_{X,Y}(t, \infty) = I_{X,Y}(t) \) and \( I_{X,Y}(0, \infty) = I_{X,Y}(t) \).

Since both processes \( X \) and \( Y \) also continued until time \( t_1 \) and been failing at time \( t_2 \), \( I_{X,Y}(t_1, t_2) \) in the interval \((t_1, t_2)\), calculate the discrepancy among their unspecified random variables. \( I_{X,Y}(t_1, t_2) \) fulfils all of the relative information measure attributes. Out of (1.5), we get

\[
I_{X,Y}(t_1, t_2) = \int_{t_1}^{t_2} \log \frac{f(x)}{g(x)/|g(t_1)|} \, dx - \bar{H}_X(t, t_2),
\]  

(1.6)

where

\[
\bar{H}_X(t, t_2) = -\int_{t_1}^{t_2} \log \frac{f(x)}{F(t_2) - F(t_1)} \, dx.
\]  

Alternatively (1.5) can be written as

\[
I_{X,Y}(t_1, t_2) = \log \frac{g(t_2)}{g(t_1)} + \int_{t_1}^{t_2} f(x) \log \frac{g(x)}{g(x)} \, dx.
\]  

(1.7)

We implement a multivariate generalisation of the measure throughout the present study (1.5). Owing to the \( \beta \) variable described and discussed in section 2, this new generalised measure has far more versatility in implementation. Some characteristics of this measure were analysed in section 3 and even some useful details are summarized. In section 4 survival analysis of lung-cancer data is performed using survival function and Kaplan Meier estimator with the help of libraries of python. Conclusion of this study is given in section 5.

## 2. A New Generalized Informative Discrimination Measure

The generalized \( X \) and \( Y \) discrimination measure for the interval of time \( t_1, t_2 \) is provided by

\[
I_{X,Y}^\beta(t_1, t_2) = \frac{1}{\beta - 1} \left[ \int_{t_1}^{t_2} \frac{f(x)}{F(t_2) - F(t_1)} \right] \left( \frac{g(x)}{g(t_1)} \right)^{1-\beta} \, dx, \quad \beta \neq 1, \beta > 0
\]  

(2.1)

It can be mentioned that (2.1) get reduced to (1.5) when \( \beta \to 1 \). But we may call (2.1), the degree \( \beta \) simplified informative discrimination measure.

Likewise, it is possible to generalize the measure (1.6) as

\[
\bar{I}_{X,Y}^\beta(t_1, t_2) = \int_{t_1}^{t_2} \frac{f(x)}{F(t_2) - F(t_1)} \left( \frac{g(x)}{g(t_1)} \right)^{1-\beta} \, dx - \bar{H}_X^\beta(t_1, t_2),
\]  

(2.2)

where

\[
\bar{H}_X^\beta(t_1, t_2) = -\int_{t_1}^{t_2} \frac{f(x)}{F(t_2) - F(t_1)} \left( \frac{g(x)}{g(t_1)} \right)^{1-\beta} \, dx.
\]  

Measure (2.2) can also be written as

\[
\bar{I}_{X,Y}^\beta(t_1, t_2) = \frac{1}{F(t_2) - F(t_1)} \log \frac{g(t_2)}{g(t_1)} \left( \frac{F(t_2) - F(t_1)}{F(t_2) - F(t_1)} \right) \left( \frac{g(t_2)}{g(t_1)} \right)^{1-\beta} \int_{t_1}^{t_2} f(x) \left( \frac{g(x)}{f(x)} \right)^{1-\beta} \, dx +
\]
\[
\frac{1}{F(t_2) - F(t_1)} \left( \frac{f(t_2) - F(t_1)}{G(t_2) - G(t_1)} \right)^{1-\beta} \int_{t_1}^{t_2} f(x) \cdot \left( \frac{g(x)}{f(x)} \right)^{1-\beta} \log \left( \frac{f(x)}{g(x)} \right) dx.
\] (2.3)

It may be noted that as \(\beta \to 1\) (2.2) and (2.3) reduces to (1.6) and (1.7) respectively.

**Example 2.1** Consider \(X, Y\) as random life time of two systems with marginal entropies \(f(x) = \frac{2-x}{2}, 0 < x < 2, g(y) = \frac{4-y}{4}, 0 < y < 4\). Since \(X\) and \(Y\) are belonging in the different domain therefore using relative entropy is not interpretable. In (2.3) for \(\beta = 2\), in the time interval (0,1) and (1, 1.5) discrimination measure is (0.0562) and (0.2798). Hence discrimination measure among \(X\) and \(Y\) in time interval (1, 1.5) is greater than of it in time interval (0,1). It is clear from the example that as time interval for failure of two systems is less discrimination measure became large.

Furthermore, using the words “increasing” and “decreasing” in the non-strict context, we acquire a few other bounds of \(I_{X;Y}(t_1, t_2)\). We also start making the use stochastic appropriate referral that refer to Shaked and Shanty Kumar (1994).

### 3. Some properties of Generalised Informative Discrimination measure

In this section lower and upper bounds for generalized discrimination measure are studies, for this we first give the concept of probability ratio ordering.

**Definition 3.1** A random variable \(X\) is said to be larger than another random variable \(Y\) in the likelihood ratio \(X \geq_{lr} Y\) (or \(Y \leq_{lr} X\)) if \(\frac{f(x)}{g(x)}\) is increasing (decreasing) in \(x\) over the union of the support of \(X\) and \(Y\).

**Proposition 3.1** (i) If \(\frac{f(x)}{g(x)}\) is increasing (decreasing) in \(x > 0\) i.e. \(X \geq_{lr} Y\) (or \(Y \leq_{lr} X\)), then

\[
\left( \frac{r_{1}(t_1, t_2)}{r_{2}(t_1, t_2)} \right)^{-1} \log \left( \frac{r_{1}(t_1, t_2)}{r_{2}(t_1, t_2)} \right) \leq \left( \frac{r_{1}(t_1, t_2)}{t_{2}(t_1, t_2)} \right)^{\beta - 1} \log \left( \frac{r_{1}(t_1, t_2)}{t_{2}(t_1, t_2)} \right).
\] (3.1)

(ii) If \(g(x)\) is decreasing (increasing) in \(x > 0\), then

\[
\left( \frac{r_{1}(t_1, t_2)}{r_{2}(t_1, t_2)} \right)^{-1} \log \left( \frac{1}{r_{1}(t_1, t_2)} \right) \leq \left( \frac{r_{1}(t_1, t_2)}{t_{2}(t_1, t_2)} \right)^{\beta - 1} \log \left( \frac{1}{r_{2}(t_1, t_2)} \right).
\] (3.2)

**Proof:** (i) It is clear from (2.1), that

\[
I_{X;Y}^\beta(t_1, t_2) = \frac{1}{\beta - 1} \left[ \int_{t_1}^{t_2} \left( \frac{f(x)}{g(x)} \right)^{\beta - 1} \left( \frac{g(t_2) - g(t_1)}{f(t_2) - F(t_1)} \right)^{\beta - 1} dx - 1 \right]
\] (3.3)

as \(\frac{f(x)}{g(x)}\) is increasing (decreasing) in \(x > 0\), (3.3) reduces to

\[
I_{X;Y}^\beta(t_1, t_2) \leq \left( \frac{f(t_2)}{F(t_2)} - F(t_1) \right) \log \left( \frac{f(t_2)}{F(t_2)} - F(t_1) \right)
\]

and

\[
I_{X;Y}^\beta(t_1, t_2) \geq \left( \frac{f(t_2)}{F(t_2)} - F(t_1) \right) \log \left( \frac{f(t_2)}{F(t_2)} - F(t_1) \right).
\]
(ii) Additionally, for all $t_1 < x < t_2$ decreasing (increasing) $g(x)$ and for $x > 0$ $g(t_1) < g(x) < g(t_2)$ holds and consequently from (2.2), we have
\[
I_{XY}^\beta(t_1, t_2) \leq (\geq) - \frac{r_2^X(t_1, t_2)}{r_2^Y(t_1, t_2)} \beta^{-1} \log t_2^Y(t_1, t_2) - H^\beta_X(t_1, t_2)
\]
and
\[
I_{XY}^\beta(t_1, t_2) \geq (\leq) - \frac{r_1^X(t_1, t_2)}{r_1^Y(t_1, t_2)} \beta^{-1} \log r_1^Y(t_1, t_2) - H^\beta_X(t_1, t_2)
\]
which is (3.2).

**Example 3.1** Interpret X and Y as two distinct random variables relating to the modified exponential distribution only with rates $\lambda$ and $\mu$ and the weight function $\phi(.)$. Their densities are determined by
\[
f(x) = \frac{\phi(x)e^{-\lambda x}}{h(x)}, \quad g(x) = \frac{\phi(x)e^{-\mu x}}{h(x)},
\]
respectively, and $h(.)$ denotes the Laplace transform of $\phi(.)$.

$h(\xi) = L\xi(\phi(.)) = f_0^\infty e^{-\xi x} \phi(x) dx, \quad \xi > 0,$

therefore, for $\lambda \neq \mu$ the generalized informative discrimination measure among X and Y at the interval $(t_1, t_2)$ is as follows:
\[
\tilde{I}_{XY}^\beta(t_1, t_2) = \int_{F(t_2) - F(t_1)} \left( \frac{G(t_2) - G(t_1)}{F(t_2) - F(t_1)} \right)^{\beta-1} \log \left( \frac{G(t_2) - G(t_1)}{F(t_2) - F(t_1)} \right) f_1^t \left( e^{(\mu-\lambda) x} \right) \frac{h(\mu) \cdot h(\lambda)}{h(\xi)} dx
\]
\[
+ \left( \frac{G(t_2) - G(t_1)}{F(t_2) - F(t_1)} \right)^{\beta-1} (\mu - \lambda) \int_{F(t_2) - F(t_1)} f_1^t \left( e^{(\mu-\lambda) x} \right) \frac{h(\mu) \cdot h(\lambda)}{h(\xi)} dx
\]
\[
+ \left( \frac{G(t_2) - G(t_1)}{F(t_2) - F(t_1)} \right)^{\beta-1} \log \frac{h(\mu) \cdot h(\lambda)}{h(\xi)} \int_{F(t_2) - F(t_1)} f_1^t \left( e^{(\mu-\lambda) x} \right) \frac{h(\mu) \cdot h(\lambda)}{h(\xi)} dx
\]
(3.6)

It may be noted that when $\beta \rightarrow 1$, (3.6) reduces to the measure given below:
\[
\tilde{I}_{XY}^\beta(t_1, t_2) = \log \left( \frac{G(t_2) - G(t_1)}{F(t_2) - F(t_1)} \right) + \log \frac{h(\mu)}{h(\lambda)} - (\lambda - \mu)E(X|t_1 < X < t_2),
\]
(3.7)

**Example 3.2.** For $\phi(x) = x^n$ and $h(\xi) = (n! \xi^n)$, the random variable distributions in example 3.1 above are termed as Erlang distributions with both the scale parameters $\lambda$ and $\mu$ as well as the common shape parameters.

**Solution.** Substituting Erlang Distributions in (3.6) we get
\[
\tilde{I}_{XY}^\beta(t_1, t_2) = \int_{F(t_2) - F(t_1)} \left( \frac{G(t_2) - G(t_1)}{F(t_2) - F(t_1)} \right)^{\beta-1} \log \left( \frac{G(t_2) - G(t_1)}{F(t_2) - F(t_1)} \right) f_1^t \left( e^{(\mu-\lambda) x} \right) \frac{h(\mu) \cdot h(\lambda)}{h(\xi)} dx
\]
\[
+ \left( \frac{G(t_2) - G(t_1)}{F(t_2) - F(t_1)} \right)^{\beta-1} (\mu - \lambda) \int_{F(t_2) - F(t_1)} f_1^t \left( e^{(\mu-\lambda) x} \right) \frac{h(\mu) \cdot h(\lambda)}{h(\xi)} dx
\]
\[
+ \left( \frac{G(t_2) - G(t_1)}{F(t_2) - F(t_1)} \right)^{\beta-1} \log \frac{h(\mu) \cdot h(\lambda)}{h(\xi)} \int_{F(t_2) - F(t_1)} f_1^t \left( e^{(\mu-\lambda) x} \right) \frac{h(\mu) \cdot h(\lambda)}{h(\xi)} dx
\]
(3.8)

In case $\beta \rightarrow 1$, (3.8) will be reduces to the following measure:
\[
\tilde{I}_{XY}^\beta(t_1, t_2) = \log \frac{G(t_2) - G(t_1)}{F(t_2) - F(t_1)} - n \log \frac{\lambda}{\mu} + (\lambda - \mu) \frac{y^{(n+1,2\lambda t_2)} - y^{(n,2\lambda t_1)}}{\lambda^{(n+1)}(F(t_2) - F(t_1))}
\]
(3.9)

which is the result due to [9].

In the above examples (3.1) and (3.2) since $\lambda \neq \mu$, then either $\mu > \lambda$ or $\mu < \lambda$, considering these two cases we have plotted following two graphs(fig.3.1 and fig.3.2) by taking some particular values for $\beta$ and for time interval (5.15) and (5.45) respectively.

It is clear from these plots that as $\beta \rightarrow 1$ values of $\tilde{I}_{XY}^\beta(t_1, t_2)$ comes close to $\tilde{I}_{XY}(t_1, t_2)$.

Fig.3.1: plot for example (3.2), $\tilde{I}_{XY}^\beta(t_1, t_2)$ for $\beta = 0.3, 0.6, 1$ and $\mu > \lambda$ (from bottom to top) against $t \in (5, 15)$
ICT based Framework for Data Science and Machine Learning Applications

**Proposition 3.2** Consider \(X, Y\) as random variables with usual assist \((0, \infty)\) and \(\varphi\) be a continuous and monotonic increasing function, then

\[
I_{X|Y}^\beta(t_1, t_2) = I_{X,Y}^\beta(\varphi^{-1}(t_1, t_2)).
\]

Following remarks clear the invariance of generalized informative discrimination measure under location and scale transformation.

**Remark 3.1** For all \(0 \leq a < t_1\), we get

\[
I_{X+a,Y+a}^\beta(t_1, t_2) = I_{X,Y}^\beta(t_1 - a, t_2 - a).
\]

**Remark 3.2** Let \(Y = bX\), then

\[
I_{X,Y}^\beta(t_1, t_2) = I_{X,Y}^\beta \left( \frac{t_1}{b}, \frac{t_2}{b} \right).
\]

4. **Survival Analysis**

All of us still have a thought in mind about how long an occurrence will take to happen. The human species is impacted by some kind of illness, like the breakdown of a mechanical machine, how often time it would require to cure the diseases. Then, since making a medical diagnosis, which one will survive a particular, to which rate one will die or fail? Is it necessary to take a wider view of the various causes of mortality or failure into account? We’re learning Survival Analysis to address so many questions. An significant branch of statistics that’s still factored to address all these questions is Survival Analysis.

The study of Survival Analysis must establish the time period within which this study is conducted. As in certain examples, the specified time-period for the occurrence to occur can be alike, Analysis of survival requires the modelling of time to event results. So in the research, we need to describe the scope of Survival Analysis as the "Event" in the scope of Survival Analysis as period. There are numerous ways in which we conduct survival research. It’s done in many forms, such as when we identify a group. Kaplan Meier Curves, Cox Regression Models, Hazard Rate, Survival Function are some of them, etc. The survival analysis of two separate categories is compared when the survival analysis is completed. We are performing the Log-Rank test there. We like to do Cox proportional hazard regression, Parametric Survival Systems, etc. where the Survival Analysis explains the categorical but quantitative survival variables.

We have to define such words in the Survival Analysis before one continues, such as the case, time, censorship, survival mechanism, etc. Case, once we speak about, is the activity that happens or is going to be happening in the study of survival research, such as the death of a person from a specific illness, time to be cured by a medical diagnosis, time to be cured by vaccines, time of malfunction of machines in the manufacturing store floor, time of incidence of diseases, etc.

**The Period**

The time from the start of the survival analysis evaluation mostly on subject matter to the time when the incident will occur is the case study of survival analysis. As in the case of a malfunction of a mechanical system, we have to know the

(a) the time of the case in which the computer begins

(b) if the system fails.
4.1 Censoring/ Censored Observation

This term is interpreted like if the topic in which we're doing the survival analysis research also isn't influenced by study's described case, then that is represented as censored. Just after end of a survival analysis evaluation, the censored topic may still not have a case. The issue is referred to as censored in the sense and after the censoring era, nothing has been observed outside the issue.

Censoring Observation are also of 3 types-

1. Censored correct

In several difficulties, proper censoring has been used. It occurs when, after a certain moment in time, they are not sure what happened with individuals.

It occurs when the true event time is greater than the censored time when \( c < t \). This occurs if, since they expired but were missed to follow-up or withdrew from of the research, any persons can not be followed the whole time.

2. Censored left

Left censoring is that we're not sure what's happening until a certain moment in time to people. In comparison, left censoring happens when the actual event time is much less than censored time while \( c > t \).

3. Censored Interval

Duration censoring would be when we know that something occurred in an interval (not before the study's kickoff date not after the current study end time), so we don't know exactly when it happened in the period.

Duration censoring is a mixture of left side censoring where the time among two points is considered to have happened.

**Survival function \( S(t) \):** This is a function of chance that relies mostly on study time. The subject survives more than time \( t \). The Survivor function indicates the probability that the specified time \( t \) is exceeded by the random variable \( T \). Here, we will discuss the Kaplan Meier Estimator.

4.2 Kaplan Meier Estimator

With lifetime results, the Kaplan Meier Estimator is being used to calculate the survival function. It is a methodology for non-parametric statistics. It is also referred to as the product limit estimator, as well as the idea is to measure the life time over a certain period of time as a major medical test occurrence, the certain death time, system malfunction, or other major huge moment.

There are several examples, such as

1. Inability of machine parts after many hours of operation.
2. How often time will be required for COVID 19 vaccine to heal the patient.
3. How often time is taken to get a remedy from either a medical diagnosis etc.
4. To predict how often workers will sell the organization in a given period of time.
5. How often people will get healed by lung cancer

To Calculate the Kaplan Meier Survival we first need to estimate that Survival Function \( S(t) \) is the likelihood of incident time \( t \)

Here \( (d) \) are the number of death incidents only at time \( (t) \), and \( (n) \) is the number of subjects at risk of injury or death just
prior to the time $(t)$.

### 4.3 Assumptions of Kaplan Meier Survival

In real-life situations, we will not have an idea of the true survival rate function. However in Kaplan Meier Estimator we approximate and calculate the true survival function from of the study data. There's many 3 hypotheses of Kaplan Meier Survival

1) Survival Odds are the same as for both the samples that entered late in the analysis and those who have joined early. The Survival analysis that can affect is not presumed to change.

2) Incidence of Event are achieved at a given time.

3) The study's censoring is not contingent on the result. The Kaplan Meier technique does not rely on the result of interest.

The Y-axis interpretation of Survival Analysis shows the likelihood of a subject not included in the case study. Since surviving until time, the X-axis shows the reflection of the interest of the subject. Each decrease in the survival function (approximated by the Kaplan-Meier estimator) is caused by at least one occurrence of an event of interest.

To explain the uncertainty about confidence intervals, the graph is always followed by confidence intervals—wider confidence intervals indicate high uncertainty, this occurs while we have few other participants—occurs in both dying and also being censored findings.

![Theoretical S(t) and Practical S(t)](image)

### 4.4 Important things to consider for Kaplan Meier Estimator Analysis

1) To make some kind of assumptions, we have to conduct the Log Rank Test.

2) The outcomes of Kaplan Meier can be conveniently skewed. A univariate approach to fixing the issues is the Kaplan Meier

3) The elimination of censored data would cause that shape to change. This will build prejudices in fit-up modelling

4) When Continuous Variable Dichotomization is done, statistical tests or observations become misleading.

5) We use statistical measures including the median to construct classes by dichotomizing methods, but this may lead to data set issues.

**Let us take the example in Python**

Let us import the important library required to work in python [14]
Next, we're importing various python libraries. There, the lung-cancer data collection is taken. After libraries are loaded, we will be using the panda library [16] to read the data. The dataset includes various kinds of knowledge.

Treatment 1= Norm, 2= Examination, 1= Squamous form of cell, 2= Tiny Cell, 3= adeno, 4= big, Day survival, Status 1= dead, 0= censored, Karnofsky score (general success measurement, 100= best), Diagnostic months, Age in years Prior therapy 0= no, 10= yes, etc.

Here we see the Head and tail.

Now, Here we import the python code for performing the Kaplan Meier Estimator
There, on the Karnofsky score, they analyse, which displays the timeline by x-axis as well as the score by y-axis using matplotlib [15]. It means that subject is fit, a score of 0 indicates the worst score. The best score is 1.

Then we submit the Survival Code, Previous Therapy, and here they will do the Kaplan Meier Estimator Analysis for the procedure.

To fit the Kaplan Meier function, they then fit kmf1 = KaplanMeierFitter() and execute the following code for various data related to lung cancer issues.

```python
kmf1 = KaplanMeierFitter()
T = df['Survival']
E = df['Prior_therapy']
groups = df['Treatment']
i1 = (groups == 1)
i2 = (groups == 2)
```

After executing the algorithm, Kaplan Meier estimation method shows the plot among Treatment Test Norm & Treatment Test.
Our main objective was to clarify the Kaplan Meier Estimator Survival Study. The stuff connected to it and the definition of an issue in real life.

4.5 Advantages and Dis-Advantages of Kaplan Meier Estimator

**Ventajas**
1) Doesn’t really involve many characteristics: only time is needed for the survival analysis case.
2) Offers an event-related average summary.

**Inconveniences**
1) It is not possible to compare and simultaneously track several variables.
2) The model would be biased at the time of fit unless the censoring information is removed.
3) It is not possible to estimate the correct calculation of the degree of change in the event.

5. Conclusion

In the challenge of device failure among time interval $(t_1,t_2)$, Miasg and Yari [8] introduced the concept of interval entropy and informative distance. In the present study, we expect a stochastic sweeping generalisation of informational differentiation and call it a generalised degree β informative discrimination measure. A standardized measure is flexible to better fit findings and to take into account subjective variables that can be agreed in the absence of parameters. In reality, Renyi [10] was the first person to offer the definition of parameter generalisation. One essential argument will be that sweeping generalization is always done at some cost. There are some characteristics of a particular measure, and all of these are not even present in the generalised measure. There we took advantage of the insightful discrimination measure by taking an example. We have also examined certain generalised measurement characteristics. We made some comments in support of both outcomes. The results obtained here are by no way final, but will preferably emerge in the study of expressed confidence in an overall situation, but will be useful in the philosophy of reliability and engineering, whereby information plays a significant role. At last survival analysis of lung-cancer data is performed using survival function and Kaplan Meier estimator with the help of libraries of python like numpy, matplotlib and pandas and after executing the algorithm, Kaplan Meier estimation method shows the comparison among treatment test norm and treatment test.

**References**

[1]. Di Crescenzo, A. and Longobardi, M.: A Measure of Discrimination between past lifetime distributions. *Probab. Letters.* 67, 173-182 (2004).

[2]. Di Crescenzo, A. and Longobardi, M.. Entropy-based measure of uncertainty in past lifetime distributions. *J. Appl. Prob.* 39, 434-440 (2002).

[3]. Ebrahimi, N.: How to measure uncertainty in the residual life time distribution. *Sankhya Ser. A.* 58 48-56 (1996).

[4]. Ebrahimi N. and Kirmani, S.N.U.A.: A measure of discrimination between two residual life time distributions and its applications. *Ann. Inst. Statist. Math.* 48, 257-265 (1996a).

[5]. Ebrahimi, N. and Pellerey, F.: New partial ordering of survival functions based on the notion of uncertainty. *J. Appl. Prob.* 32, 202-211 (1995).
[6]. Kullback, S. and Leibler, R.A.: On information and sufficiency. *Ann. Math. Statist.* 22, 79-86 (1951).
[7]. Kullback, S.: Information Theory and Statistics. New York, *John Wiley*, (1959).
[8]. Misagh, F. and Yari, G.H. A novel entropy-based measure of uncertainty to lifetime distributions characterizations. *In Proceedings of the First International Conference on Mathematics and Statistics, AUS-ICMS’10, American University of Sharjah, Sharjah, UAE; ISBN 978-9948-427-17-9, ID: 100196* (2010).
[9]. Misagh, F. and Yari, G.H.: Interval Entropy and Informative Distance Entropy, *MDPI*, 13,1(2011).
[10]. Rényi, A.: On measures of entropy and information. *In: Proceedings of the fourth Berkeley symposium*, 1, 547–561 (1961).
[11]. Shaked, M. and Shantikumar, J.G.: Stochastic Orders and Their applications. *Academic Press, San Diego* (1994).
[12]. Shannon, CE.: A mathematical theory of communications. *Bell Syst Tech J* 27,623–656 (1948).
[13]. Ullah, A.: Entropy, divergence and distance measures with econometric applications. *J Stat Plan,Inference* 49:137–162 (1996).
[14]. https://plot.ly/python/
[15]. https://matplotlib.org/
[16]. https://realpython.com/python-data-cleaning-numpy-pandas/