UNIVERSITA' POLITECNICA DELLE MARCHE
FACOLTA' DI INGEGNERIA

Corso di Laurea triennale in Ingegneria Gestionale

TECNICHE DI SURVIVAL ANALYSIS PER LA MANUTENZIONE PREDITTIVA DI IMPIANTI INDUSTRY 4.0

Survival Analysis techinques for the predictive maintenance of Industry 4.0 plants

Relatore:

Chiar.mo Prof. Maurizio Bevilacqua

Correlatore:

Chiar.mo Prof. Giovanni Mazzuto

Tesi di Laurea di :

Mattia Vergari

A.A. 2020/2021

Indice

Capitolo Primo

ANALISI DELLA SOPRAVVIVENZA

1.1 Introduzione

Con il concetto di Analisi della sopravvivenza si intende una serie di metodi e modelli matematici che hanno l'obiettivo di prevedere il verificarsi di un determinato evento. L'analisi della sopravvivenza può essere applicata a diversi campi, come medicina, sanità pubblica, scienze sociali, ingegneria (Tableman and Kim, 2004). In ogni ambito, si studia il periodo di tempo che intercorre tra uno specifico evento di inizio (starting point) e il verificarsi di uno specifico risultato (endpoint). Nel campo medico nelle prime applicazioni del metodo, l'evento in studio era quasi sempre riferito alla morte del paziente; da ciò deriva il nome *dati di sopravvivenza*. Nelle più recenti applicazioni in ambito ingegneristico,

quindi produttivo, con l'evento "morte del paziente" ci si riferisce ad un'eventuale rottura o malfunzionamento di una macchina.

1.2 La funzione di sopravvivenza

Questo tipo di analisi si adatta bene alle situazioni in cui il problema generale è valutare la probabilità di sopravvivenza in funzione del tempo. L'analisi della sopravvivenza è costituita infatti dalla presenza di una variabile aleatoria non negativa, con distribuzione tipicamente asimmetrica, legata al tempo di accadimento di un particolare evento. Per definizione, la funzione di sopravvivenza $s_T(t)$ esprime la probabilità che il tempo di sopravvivenza T dell'unità sperimentale sia maggiore di t, ossia

$$s_T(T) = 1 - F_T(t) = P(T > t)$$

Con $F_T(t)$ funzione di ripartizione di T.

1.3 La Censura

Durante il periodo di osservazione, solo alcuni individui sperimentano l'evento finale, mentre per gli altri che non lo hanno sperimentato non si potrà conoscere il loro tempo di sopravvivenza. Da ciò nasce l'esigenza dei dati censurati (Selvin, 2008, 73). Generalmente ci sono tre tipi di censura : a destra, a sinistra, a intervallo di tempo. La censura a destra è quella più utilizzata e si verifica quando:

1. Un soggetto/macchina abbandona lo studio prima che l'evento accada
2. Un soggetto/macchina non sviluppa l'evento atteso
3. Vengono persi dati relativi al soggetto/macchina

La censura a sinistra, sebbene raramente applicata, si utilizza quando l'evento non è riconducibile ad un tempo esatto ma si inizia lo studio nel momento in cui si presenta l'evento in questione (Kleinbaum and Klein, 2011, 6-8).

Capitolo Secondo

MODELLI DELLA SURVIVAL ANALYSIS

2.1 Modelli parametrici e non parametrici

I modelli dell'analisi di sopravvivenza possono essere classificati in : modelli non parametrici e modelli parametrici. I modelli non parametrici vengono utilizzati quando non vi è la necessità di fare assunzioni sulla distribuzione *T*. Al contrario, i modelli parametrici richiedono che la distribuzione del tempo di sopravvivenza sia nota e la funzione di azzardo $\lambda(t)$ completamente specificata :

$$\lambda(t) = \lim_{\Delta t \to 0} + \frac{P_r\{t \leq T < t + \Delta t | T \geq t\}}{\Delta t} = \frac{f_T(t)}{s_T(t)}$$

Tra i modelli parametrici più noti vale la pena citare quello di Cox. Il modello di Cox è una particolare tecnica di regressione multipla che

permette di analizzare il rapporto tra un fattore di rischio (per esempio il fumo) e l'incidenza di un determinato esito clinico (per esempio l'infarto del miocardio), correggendo per uno o più fattori di confondimento (quali l'obesità e l'ipertensione) (Provenzano, D'Arrigo, Zoccali, Tripepi, 2011). Nella regressione di Cox, che si utilizza negli studi di coorte (analisi fattori di rischio), la variabile dipendente è il tasso di incidenza di un determinato evento, cioè il numero di eventi per persona-tempo. Pertanto, a un determinato tempo *t*, per ogni individuo della coorte è indispensabile conoscere la condizione (vivo/morto, evento/non evento, affetto/non affetto) e il tempo intercorso tra l'ingresso nello studio e la data dell'ultima osservazione (Provenzano, D'Arrigo, Zoccali, Tripepi, 2011).

Il metodo non parametrico più diffuso per la stima di probabilità di sopravvivenza è il metodo del prodotto limite, noto anche come lo stimatore di

Kaplan-Meier (Kaplan and Meier, 1958). Lo stimatore è così definito :

$$\hat{s} = \prod_{y(i) \leq t} \hat{p}_i = \prod_{y(i) \leq t} \frac{n_i - d_i}{n_i}$$

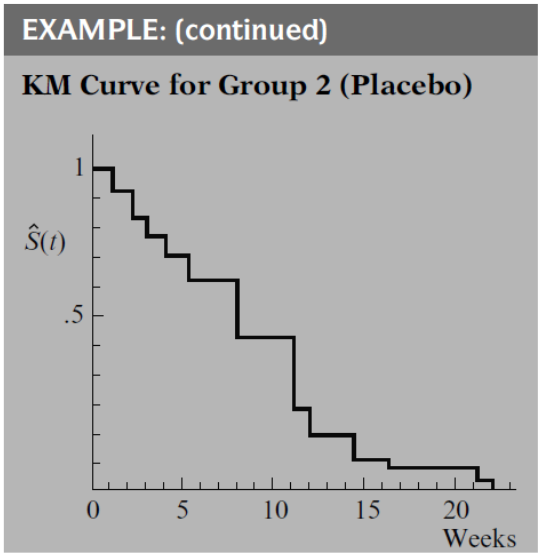
dove $y_{(k)} \leq t < y_{(k+1)}$

essendo :

- n_i il numero dei soggetti a rischio prima di $y_{(i)}$
- d_i numero di soggetti che sperimentano l'evento al tempo $y_{(i)}$
- $p_i = P\left(T > y_{(i)} \middle| T > y_{(i-1)}\right)$

L'obiettivo dello stimatore di Kaplan-Meier è quello di andare a generare una curva di sopravvivenza (output) attraverso l'utilizzo di dati definiti all'inizio di uno studio (input). La curva generata dallo stimatore è una curva a gradini, in cui ogni step rappresenta il verificarsi dell'outcome. Il numero degli step rappresenta il numero di eventi che si sono verificati durante il

corso dello studio. Sull'asse delle ascisse viene riportato il tempo, mentre sull'asse delle ordinate la probabilità di sopravvivenza.



(Kleinbaum and Klein, 2011, 63).

Capitolo Terzo

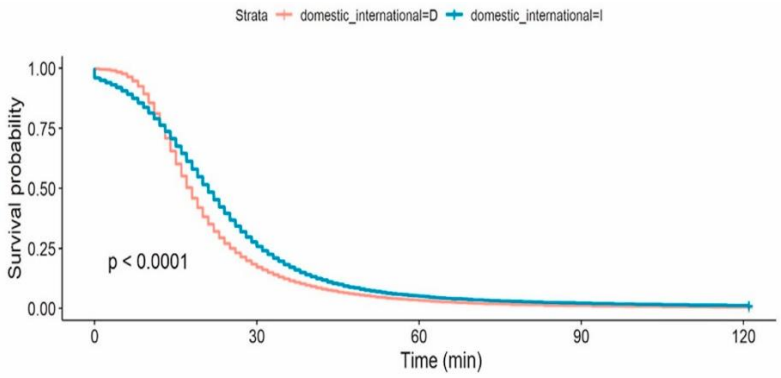
APPLICAZIONI DELL'ANALISI DI SOPRAVVIVENZA

3.1 Il ritardo nei voli arei della Corea del Sud

Studiando la letteratura applicata all'analisi di sopravvivenza ci rendiamo conto che i campi in cui questo metodo viene applicato sono diversi. Come anticipato precedentemente quest'analisi viene spesso applicata nel campo medico ma può essere modellata con l'obiettivo di applicarla a ciò che si vuole studiare.

Per esempio, nello studio che verrà riportato qui di seguito, l'analisi della sopravvivenza è stata utilizzata per analizzare i ritardi dei voli arei nella Corea del Sud. Secondo quanto riportato dallo studio, dagli anni 2000, la Corea del Sud ha vissuto un incremento della domanda dei voli arei che

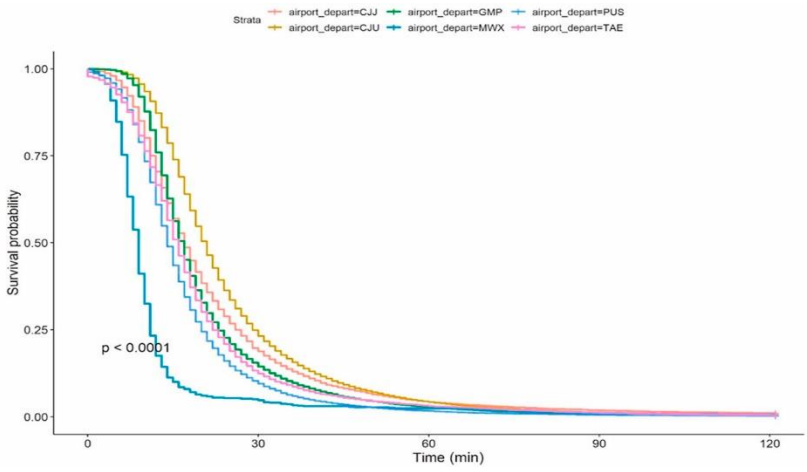
hanno portato all'ampliamento del settore facendo sorgere nuove aziende. L'incremento della domanda e di conseguenza del traffico aereo ha portato a dei ritardi nei voli aerei. Lo studio si pone l'obiettivo di analizzare i ritardi, capire quale nesso c'è tra l'incremento della domanda e i ritardi e a quale causa attribuire quest'ultimi. Nello studio viene specificato che il ritardo può esser causato da diversi fattori (voli cancellati, aeroporti diversi), per questo motivo i dati sono stati suddivisi in varie categorie. Una volta definiti i dati, è stata calcolata la curva di Kaplan-Meier così riportata :



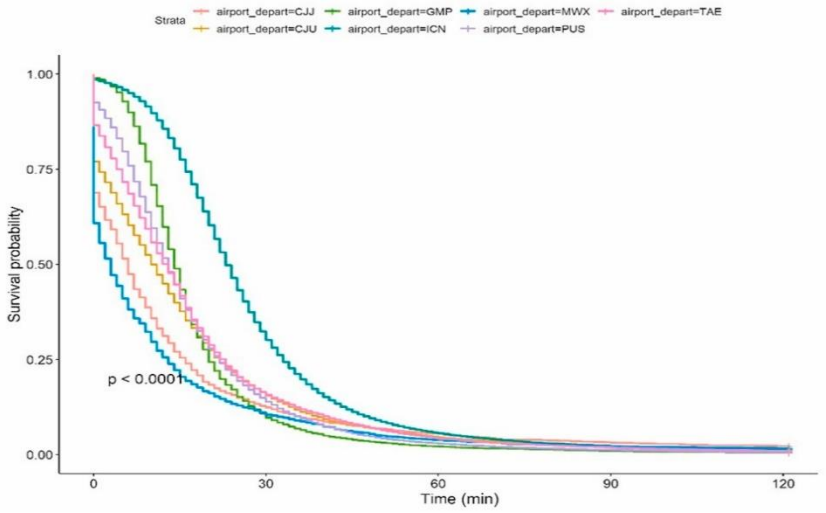
Per generare la curva sono stati analizzati 441,312 voli; 195,161 voli domestici e 246,151 voli internazionali. Nel caso dei voli internazionali la

percentuale di aerei che partono senza ritardo è maggiore dei voli domestici. Con il passare del tempo la curva per i voli domestici si inverte, ed i ritardi dei voli associati diminuiscono rispetto ai voli internazionali. In sintesi, i voli domestici subiscono spesso un ritardo di 10-30minuti, mentre i voli internazionali subiscono ritardi ben più ampi.

Per avere dei risultati più precisi è stato fatto il LogRankTest considerando i voli di ogni singolo aeroporto per tipologia di volo (domestici e internazionali). Le curve dei voli domestici per singolo aeroporto è la seguente :



Mentre la curva dei voli internazionali per i singoli aeroporti è :



Dal LogRankTest emerge che le curve generate dai ritardi dei voli domestici e internazionali sono diverse anche considerando le partenze dai diversi aeroporti. Con maggiore attenzione si può notare che i risultati tendono ad essere simili alla curva generata di Kaplan-Meier. Per alcuni aeroporti non è semplice comparare soltanto la curva di sopravvivenza, motivo per cui viene utilizzato Cox per ottenere dei risultati più precisi.

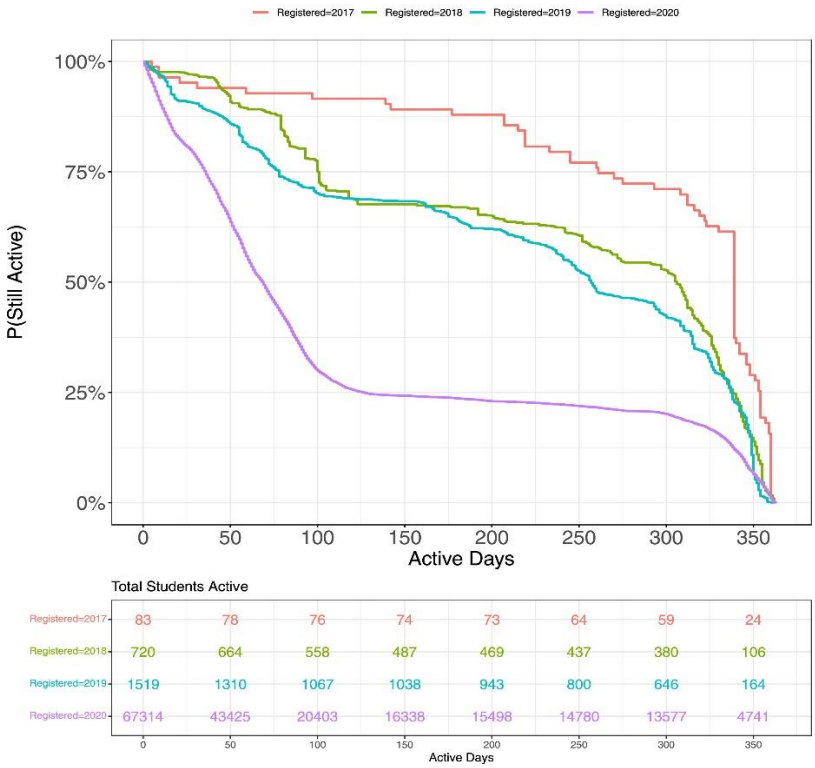
In conclusione, abbiamo visto che sono stati messi a confronto i ritardi dei voli di vari aeroporti della Corea del Sud. I risultati ottenuti ci hanno chiarito che sebbene i voli domestici abbiano più ritardi

rispetto ai voli internazionali, l'entità del ritardo è inferiore (Myeonghyeon and Jiheon, 2020).

3.2 Gli effetti del COVID-19 su Bettermark

L'analisi della sopravvivenza è stata modellata anche per scoprire gli effetti che il COVID-19 ha causato agli studenti. In particolare, nello studio proposto è stata studiata la perseverenza nel coinvolgimento degli studenti su Bettermark.

Bettermark è un sistema di apprendimento adattivo per la matematica che sostituisce i libri di testo. È una piattaforma online dove gli iscritti hanno a disposizione degli strumenti per imparare step by step la matematica. Nello studio proposto sono state analizzate le iscrizioni degli studenti alla piattaforma per lo stesso periodo di tempo dal 2017 al 2020. I dati analizzati hanno generato queste curve :



Analizzando i dati si scopre che, sebbene gli iscritti alla piattaforma aumentino ogni anno, probabilmente grazie alla maggior visibilità della piattaforma, il numero degli studenti attivi diminuisce di molto nel 2020 rispetto agli altri periodi considerati. I risultati di questo studio evidenziano l'applicabilità dell'analisi di sopravvivenza con ambienti di lavoro di apprendimento online. L'analisi della sopravvivenza può essere uno strumento utile per studiare il coinvolgimento degli studenti o il tasso

di abbandono scolastico (Hermann Spitzer, Gutsfeld, Wirzerberger & Moeller, 2021).

3.3 Il caso Airbnb

Vediamo un altro esempio dell'applicazione dell'analisi della sopravvivenza. In questo studio sono stati analizzati gli annunci della piattaforma Airbnb ad Ibiza; precisamente 9000 proprietà delle Isole Baleari sono state inserite nello studio che è durato da Luglio 2015 a Settembre 2016, per verificare quali caratteristiche degli annunci influiscono maggiormente sulla probabilità di lasciare la piattaforma. Le caratteristiche prese in considerazione sono : la particolarità e i dettagli degli annunci , la posizione della struttura (distanza dai punti di interesse) e l'esperienza del proprietario.

Il dataset studiato è il seguente :

Period	Failures	Entries	Balance
0	–	–	4240
1	0	543	4783
2	620	378	4541
3	49	193	4685
4	207	147	4625
5	184	166	4607
6	290	210	4527
7	260	439	4706
8	303	445	4848
9	372	436	4912
10	450	469	4931
11	502	527	4956
12	410	604	5150
13	521	554	5183
14	769	370	4784
15	214	23	4593

a

Each period corresponds to a month where period 1 is July 2015.

Come riportato dalla tabella, all'inizio dello studio (Luglio 2015) sono stati considerati 4240 annunci. Durante il periodo di osservazione, 5151 annunci su 9744, circa il 53%, hanno sperimentato l'evento, ovvero hanno lasciato la piattaforma Airbnb. Per ogni periodo sono stati resi noti i numeri degli

annunci che hanno abbandonato la piattaforma e quelli che sono stati inseriti successivamente. Inoltre, per semplificare la lettura della tabella viene mostrato il bilancio attivo tra il numero degli annunci che restano a far parte dello studio e quello che lo hanno abbandonato. Successivamente sulla base di quanto specificato sono state calcolate le curve di Kaplan-Meier.

Come è stato anticipato le caratteristiche che lo studio prende in considerazione sono : la qualità descrittiva dell'annuncio, la posizione della struttura e l'esperienza dell'host. I risultati hanno confermato che la posizione gioca un ruolo importante. In primo luogo l'ambiente competitivo è una forte determinante della sopravvivenza. Infatti, trovarsi in una zona con una forte concorrenza aumenta il tasso di mortalità dell'annuncio, in secondo luogo sulla posizione influisce anche la distanza della struttura dai punti di interesse.

Un altro aspetto fondamentale che permette all'annuncio una maggiore probabilità di sopravvivenza oltre all'ubicazione è l'esperienza dell'host. Host più esperti, sia in termini di annunci gestiti che di longevità sulla piattaforma, hanno meno probabilità di lasciarla. Dallo studio emerge anche che gli host che implementano strategie di mercato dinamico hanno minori probabilità di abbandonare Airbnb. Infine, un altro aspetto fondamentale per la longevità degli annunci riguarda le caratteristiche intrinseche dello stesso, ovvero, gli annunci più dettagliati (Leoni, 2020).

Capitolo Quarto

KAPLAN MEIER

4.1 Definizione del dataset

In questa tesi verranno simulate più curve di Kaplan-Meier attraverso l'utilizzo di alcuni dati selezionati, con l'obiettivo di confrontare i risultati ottenuti dalle curve.

I dati dello studio proposto, sono stati estrapolati da un dataset della NASA, caratterizzato da diverse variabili e sensori. Il dataset grezzo è formato da 26 colonne numerate, ognuna delle quali si riferisce ad una precisa caratteristica.

Esso è definito :

1. Numero delle macchine
2. Tempo misurato delle macchine
3. Condizioni di lavoro n°1

4. Condizioni di lavoro n°2
5. Condizioni di lavoro n°3
6. Sensori di rilevamento n°1
7. Sensori di rilevamento n°2

…

26. Sensori di rilevamento n°26

Dato che verranno proposte delle curve di Kaplan-Meier, il dataset è stato ridimensionato includendo soltanto i dati necessari per lo sviluppo di tali curve. Sono state selezionate le colonne numero 1 e 2 adeguandole allo standard applicativo dello stimatore. Infatti, per la colonna numero 2 (tempo delle macchine) sono stati presi in considerazione e messi in ordine crescente soltanto gli istanti di tempo finali.

4.2 Sviluppo di Kaplan-Meier

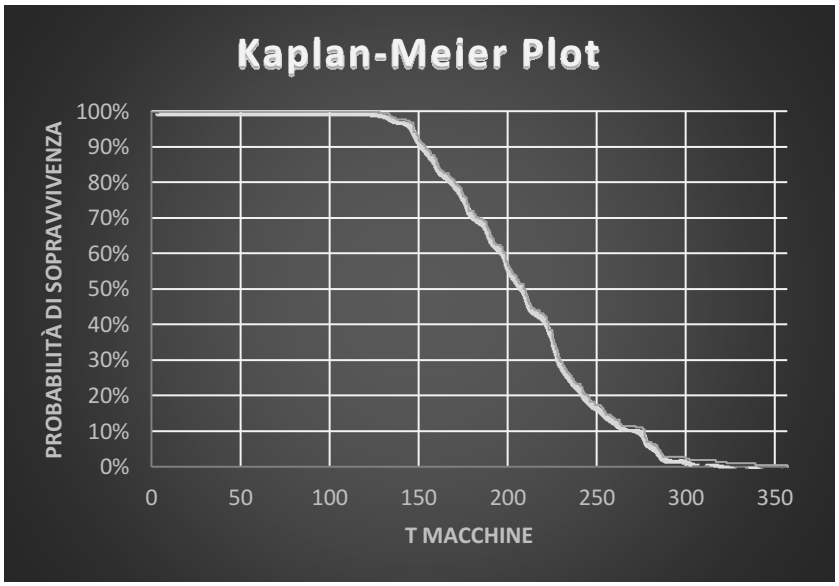
La matrice del dataset è così rappresentata :

A	B	C	D
1	128	0,9954	217
2	133	0,9908	216
3	134	0,9862	215
4	135	0,9816	214
5	137	0,9771	213
6	143	0,9725	212
7	145	0,9679	211
8	147	0,9404	205
9	147	0,9404	205
10	147	0,9404	205
11	147	0,9404	205
12	147	0,9404	205
13	147	0,9404	205
14	149	0,9312	203
15	149	0,9312	203
16	150	0,9174	200
…	…	…	…
204	281	0,0596	14
205	283	0,0504	13
206	284	0,0504	11
207	284	0,0458	11
208	285	0,0458	10
209	286	0,0367	8
210	286	0,0367	8
211	287	0,0275	6
212	287	0,0275	6
213	300	0,0229	5
214	302	0,0183	4
215	317	0,0138	3
216	323	0,0092	2
217	339	0,0046	1
218	357	0	0

(Saxena and Goebel, 2008)

con A *numero delle macchine*, B *istante di tempo finale in ordine crescente*, C $s(t)$, D *numero restante delle macchine*.

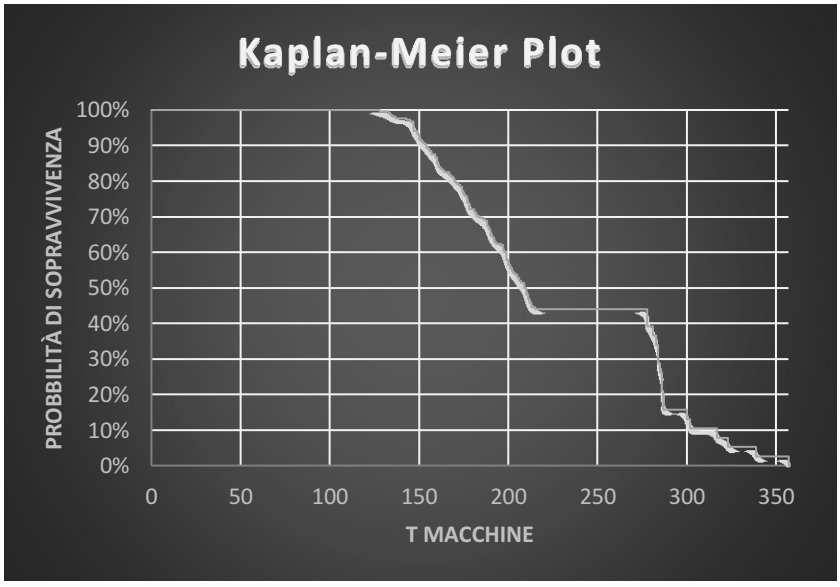
La curva di Kaplan-Meier che si genera utilizzando la matrice proposta è la seguente :



Questa curva è caratterizzata dal fatto che sono state prese in esame tutte le macchine, senza considerare eventuali censure. Sull'asse delle ascisse viene riportato l'intervallo di tempo dello studio relativo ai dati macchina, mentre sull'asse delle ordinate troviamo la probabilità di sopravvivenza $s(t) = \frac{n_i - d_i}{n_i}$, con n_i il numero delle

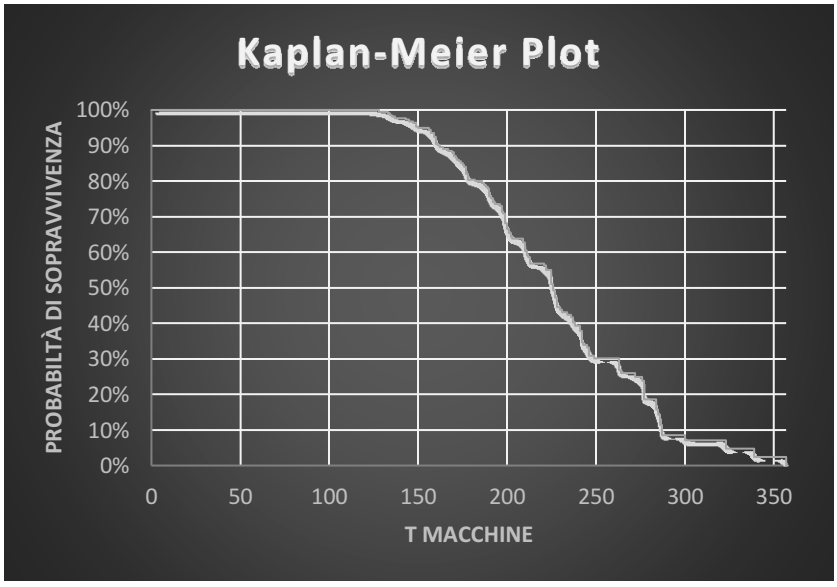
macchine a rischio (colonna B), d_i macchine che sperimentano l'evento (rottura).

Prendiamo in considerazione l'intervallo di tempo *t*, con $200 \leq t \leq 250$, ed analizziamo cosa succede alla curva di sopravvivenza in alcuni casi. Nella curva inizialmente proposta, la probabilità di sopravvivenza delle macchine, nel periodo considerato, oscilla circa tra il 56% ed il 17%. Nello specifico, in $t = 225$ la probabilità di sopravvivenza delle macchine è del 38,5%. Ipotizzando che si debba censurare, per colpa di una perdita di dati, l'intervallo di tempo *t,* $218 \leq t \leq 278$, la curva sarà la seguente :



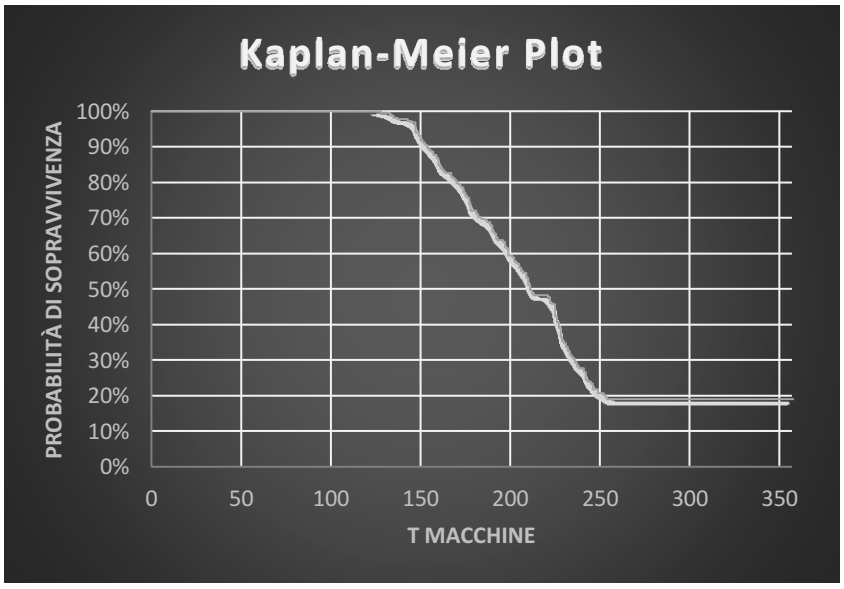
Possiamo notare che, per lo stesso intervallo di tempo considerato nella curva precedente, la probabilità di sopravvivenza delle macchine è cambiata, infatti ora oscilla tra il 56% ed il 44%. Notiamo un incremento della probabilità di sopravvivenza rispetto al primo caso per via della censura, con la probabilità al tempo $t = 225$ del 44%.

Nel terzo caso, ipotizziamo di aver avuto dei problemi con le registrazioni dei dati, motivo per cui abbiamo dovuto censurare un gran quantità di sensori in più istanti temporali. La curva che otteniamo applicando Kaplan-Meier è :



Nel solito intervallo di tempo considerato, la probabilità di sopravvivenza è cambiata di nuovo, ora oscilla tra il 66,5% ed il 30%. Mentre per $t = 225$, la percentuale è del 55% con una variazione al tempo $t = 225$ del 16.5% rispetto al caso iniziale.

Nella curva finale vedremo che sono stati censurati diversi dati in più istanti di tempo. La maggior concentrazione di dati censurati si è verificata a partire da $t = 250$. La curva che ne consegue è la seguente:



La probabilità di sopravvivenza delle macchine rispetto alla curva iniziale anche in questa simulazione è diversa. Nell'istante di tempo $t = 225$, la percentuale di sopravvivenza corrisponde al 45%, mentre dall'istante di tempo $t = 250$ fino alla fine dello studio è del 20% circa.

Supponiamo di dover studiare le seguenti curve con l'obiettivo di fare una previsione delle rotture delle macchine in modo da poter agire tempestivamente con la manutenzione. Riassumendo brevemente i risultati di ogni curva abbiamo visto che per $t = 225$, nella prima curva, la probabilità di sopravvivenza è del 38,5%. Nella seconda è del 44%, nella terza del 55% mentre nell'ultima del 45%. Osservando i seguenti dati, notiamo delle evidenti discrepanze delle percentuali di sopravvivenza per lo stesso intervallo di tempo considerato. L'abuso della censura nei tre esempi riportati, rende il metodo di Kaplan-Meier meno affidabile restituendo dei

risultati in grado di compromettere lo studio predittivo.

CONCLUSIONI

Le curve generate mostrano la probabilità di sopravvivenza dello studio considerato. Nonostante vengono simulati degli eventi avversi che portano a rotture improvvise e perdite di dati, le curve che ne conseguono non differiscono molto dalla prima curva generata. Questo accade perché la popolazione considerata dallo studio non è suddivisa in gruppi/categorie, se così fosse stato avremmo ottenuto diverse curve per ogni sottocategoria della popolazione. In quel preciso caso, avremmo potuto comparare i risultati delle diverse curve con l'obiettivo di scovare la sottocategoria più a rischio in modo da poter agire di conseguenza.

Affinché lo studio delle curve di Kaplan-Meier sia quanto più affidabile possibile, necessita anche di altri strumenti come per esempio il LogRankTest definito come :

$$LogR.Test = \frac{(Mortalità\ attesa - Mortalità\ Osservata)^2}{Somma\ dei\ prod.delle\ mortalità\ attese\ nei\ due\ gruppi}$$

Inoltre, come descritto precedentemente l'analisi della curva di sopravvivenza porterebbe a dei risultati ancora più affidabili utilizzando la regressione di Cox, attraverso l'utilizzo di variabili qualitative e quantitative da associare alla popolazione dello studio.

RIFERIMENTI

Hermann Spitzer , M.W., Gutsfeld, R., Wirzberger, M., Moeller, K. (2021). *Evaluating students'engagement with an online learning enviroment during and after COVID-19 related school closures : A survival analysis approach* in "Trends in Neuroscience and Education", Volume 25, December 2021, 100168.

Kaplan, E. L. and Meier, P. (1958). Nonparamentric estimation from incomplete observations. *Journal of the American Statistical Association,* 53, 457-481.

Kleinbaum, David G. and Klein, M. (2011). *Survival Analysis, third edition*. Springer.

Leoni, V. (2020). *Stars vs lemons. Survival analysis of peer-to peer marketplaces: the case of Airbnb* in "Tourism Management", Volume 79, August 2020, 104091.

Myeonghyeon, K. and Jiheon, B. (2020). *Modeling the flight departure delay using survival analysis in South Korea in* "Journal of Air Trainsport Management" , Volume 91, March 2021, 101996.

Provenzano, F. ; D'Arrigo, G. ; Zoccali, C. & Tripepi, G. (2011). *La regressione di Cox*. G Ital Nefrol 2011; 28(3) : 319-322.

Saxena, A. and Goebel, K. (2008). "PHM08 Challenge Data Set" , NASA Ames Prognostics Data Repository (http://ti.arc.nasa.gov/project/progrnostic-data-repository), NASA Ames Research Center, Moffett Field, CA.

Selvin, S. (2008). *Survival Analysis for Epidemiologic and Medical Research.* Cambridge University Press.

Tableman, M . and Kim, J. (2004). *Survival Analysis Using S. Chapman & Hall*.