Penggunaan Metode *K Nearest Neighborhood* untuk Imputasi Data Tersensor Kanan pada Pasien Kanker Paru-Paru Sel Kecil

Caecilia A Rahman*, Abdul kudus

Prodi Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Islam Bandung, Indonesia.

*caeciliaar24@gmail.com, abdul.kudus@unisba.ac.id

Abstract. In a study, it is usually necessary to have complete data for the accuracy of parameter estimation, but in *survival* analysis incomplete data is often found called censored data, this can happen due to limited research time and others. To complete the censored data, imputation is needed, one of method to imputating the censored data is *K-Nearest Neighborhood* (KNN) method. KNN imputation is designed to find K nearest neighbors from censored data to all complete data and then fill in the censored data with events that are most similar to its neighbors. If the target variable (or attribute) is categorical then imputation refers to the majority of neighbors but if the variable is numeric, then the imputation uses the average of the nearest neighbors. This study used data from 121 small cell lung cancer patients from the *North Central Cancer Treatment Group* in the United States. KNN imputation was used to impute the right-censored *survival* time of patients based on the average of the K nearest neighbors' complete data of *survival* time. The cens variable is used as an indicator of censorship, while the age and arm variables measure the distance between the complete data and the censored data. The smaller the distance data becomes the closest neighbor because it has similar characteristics. The average of the K complete data will be the imputed value for the censored data. This study succeeded in imputing 23 censored data based on 46 closest neighbors (K = 46).

Keywords: *Survival Analysis.Right-censored Data, K-Nearest Neighborhood.*

Abstrak. Dalam suatu penelitian biasanya diperlukan kelengkapan data untuk ketepatan pendugaan parameter, namun pada analisis *survival* kerap ditemukannya data yang tidak lengkap yang disebut data tersensor, hal ini bisa terjadi karena terbatasnya waktu penelitian dan lain-lain. Untuk melengkapi data yang tidak lengkap tersebut diperlukannya imputasi, salah satunya yaitu metode *K-Nearest Neighborhood* (KNN). Imputasi KNN dirancang untuk mencari K tetangga terdekat dari data yang tidak lengkap ke seluruh kejadian suatu data, kemudian mengisi data yang hilang dengan kejadian yang paling mirip dengan tetangganya, jika target variabel (atau atribut) berupa kategorik maka imputasi merujuk kepada mayoritas tetangga namun apabila variabel berupa numerik maka imputasi menggunakan rata-rata dari tetangga terdekat. Penelitian ini menggunakan data dari 121 pasien kanker paru-paru sel kecil dari *North Central Cancer Treatment Group* di Amerika Serikat. Imputasi KNN digunakan untuk mengimputasi waktu *survival* pasien yang tersensor kanan berdasarkan rata-rata dari sebanyak K tetangga terdekat data lengkap waktu *survival*. Variabel cens digunakan sebagai indikator penyensoran sedangkan variabel usia dan Arm (jenis perawatan) digunakan untuk mengukur jarak antara data lengkap dengan data tersensor, semakin kecil jarak maka data tersebut menjadi tetangga terdekat karena memiliki karakteristik yang mirip. Rata-rata dari sebanyak K data lengkap akan menjadi nilai imputasi bagi data tersensor. Pada penelitian ini berhasil mengimputasi 23 data tersensor berdasarkan 46 tetangga terdekatnya (K = 46).

Kata Kunci: *Analisis Survival, Data Tersensor Kanan, K-Nearest Neighborhood.*

Corresponding Author
Email: abdul.kudus@unisba.ac.id

A. Pendahuluan

Dalam suatu penelitian biasanya diperlukan kelengkapan data untuk ketepatan pendugaan parameter, namun pada analisis *survival* kerap ditemukannya data yang tidak lengkap, hal ini bisa terjadi karena terbatasnya waktu penelitian dan lain-lain. Data yang tidak lengkap pada analisis *survival* disebut dengan data tersensor, dikatakan tersensor karena informasi waktu *survival* tidak diketahui secara lengkap, dalam kata lain informasi mengenai waktu hingga terjadinya suatu event pada individu hanya diketahui sebagian.

Untuk melengkapi data perlu dilakukan imputasi, imputasi adalah metode untuk melengkapkan data dengan memperkirakan nilai yang cukup layak untuk mengisi data yang hilang dan dilanjutkan dengan analisis menggunakan metode baku [1]. Terdapat berbagai macam metode untuk mengimputasi seperti Nearest Neighbors, Self-Organizing Maps, Decision Tree, dan Jaringan Bayesian.

Salah satu penelitian yang dilakukan oleh Ahmed dkk (2020)[2] mengimputasi data menggunakan *K-Nearest Neighborhood* (KNN), penelitian tersebut mengimputasi waktu *survival* pasien kanker otak yang tersensor kanan dengan total 30 orang pasien dan tingkat penyensoran sebesar 27%, hasil menunjukkan bahwa KNN dapat digunakan sebagai alternatif untuk imputasi karena memiliki hasil yang lebih baik daripada model prediksi.

Penelitian oleh Jerez dkk (2010)[3] yang membandingkan metode-metode untuk imputasi, diperoleh hasil bahwa KNN mengungguli metode lain yang berbasis machine learning Kemudian penelitian oleh Malarvizhi & Thanamani (2012)[4] bandingkan KNN dengan K-Means untuk imputasi, diperoleh bahwa KNN memberikan hasil yang lebih baik dibandingkan K-Means. Meskipun memiliki banyak kelebihan, KNN juga memiliki kekurangan yaitu memerlukan waktu pengerjaan yang lama, sangat sensitif kepada variabel yang tidak berhubungan atau berlebihan, ketidakjelasan tipe pengukuran jarak yang tepat untuk mendapatkan hasil terbaik serta biaya komputasi yang tinggi karena harus mengukur jarak antar seluruh data[5].

Metode *K-Nearest Neighborhood* merupakan salah satu metode yang sering digunakan untuk masalah imputasi. Penelitian Batista & Monard (2003)[6] menggunakan KNN karena memiliki keunggulan yaitu bisa memprediksi atribut kualitatif dan kuantitatif serta tidak harus membuat model prediktif untuk setiap atribut data yang hilang. Penelitian oleh Mawarsari & Irhamah (2016)[7] menggunakan KKN untuk imputasi karena sederhana dan fleksibel serta dapat digunakan untuk variabel diskrit dan kontinu.

Berdasarkan eksplanasi di atas maka penelitian ini akan melakukan imputasi data tersensor kanan menggunakan *K-Nearest Neighborhood*. Penelitian ini menggunakan data pasien kanker paru-paru sel kecil dari *North Central Cancer Treatment Group* di Amerika Serikat. Data waktu *survival* pasien yang tersensor kanan berdasarkan variabel cens akan diimputasi menggunakan rata-rata dari sebanyak *K* tetangga terdekat data lengkap variabel waktu *survival*. Penentuan banyaknya *K* tetangga terdekat menggunakan jarak terdekat berdasarkan variabel usia dan Arm.

Berdasarkan latar belakang yang telah diuraikan, maka perumusan masalah dalam penelitian ini sebagai berikut: "Bagaimana penerapan metode *K-Nearest Neighborhood* pada data pasien kanker paru-paru sel kecil *North Central Cancer Treatment Group*?". Selanjutnya, tujuan dalam penelitian ini diuraikan dalam pokok-pokok sbb.

1. Mengetahui deskripsi statistik dari data pasien kanker paru-paru sel kecil *North Central Cancer Treatment Group*
2. Mengetahui penerapan metode *K-Nearest Neighborhood* pada data pasien kanker paru-paru sel kecil *North Central Cancer Treatment Group*

B. Metodologi Penelitian

Data yang digunakan untuk penelitian ini adalah data sekunder yang berasal dari jurnal berjudul "Sequencing and schedule effects of cisplatin plus etoposide in small-cell lung cancer: results of a *North Central Cancer Treatment Group* randomized clinical trial" oleh (Maksymiuk, et al., 1994) [8], dijelaskan dalam (Ying, Jung, & Wei, 1995)[9] dan diunggah pada situs web figshare oleh Valerie Poynor. Data dari 121 pasien kanker paru-paru sel kecil *North Central*

Cancer Treatment Group (NCCTG) yang memiliki 4 variabel yaitu usia (X_1), Arm (X_2), waktu *survival* (Y) dan cens sebagai indikator penyensoran.

Pasien secara acak mengikuti satu dari dua perawatan yang disebut sebagai Arm A dan Arm B. Pasien Arm A menerima cisplatin (P) diikuti oleh etoposide (E), sedangkan pasien Arm B menerima (E) diikuti oleh (P). Ada total 62 pasien di Arm A dengan 15 data tersensor kanan, sedangkan Arm B terdiri dari 59 pasien dengan 8 data tersensor kanan. Sehingga dari total data sebanyak 121 terdapat 98 data lengkap dan 23 data tersensor kanan.

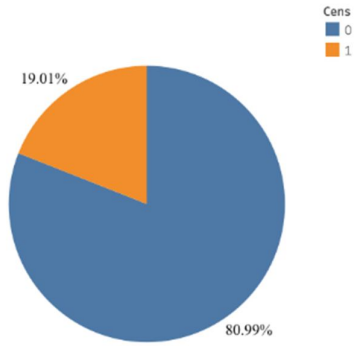
Adapun input yang diperlukan dalam tahapan analisis yaitu data tersensor kanan (z_i), indikator penyensoran (δ_i), nilai K dan nilai dari variabel prediktor (x_i) yang akan menghasilkan output nilai hasil imputasi (y_i^{knn}). Adapun langkah-langkah penelitian ini adalah:

1. Mulai
2. Jika $\delta = 0$ (data tersensor), lakukan
3. Dari $(i = 1 \text{ sampai } p)$ buat kolom
4. Jika $\delta = 1$ (data lengkap), lakukan
5. Dari $(j = 1 \text{ sampai } p)$ buat baris
6. Menghitung jarak campuran berdasarkan [10]:
   Untuk variabel usia karena bertipe numerik maka menggunakan rumus:
   $$d_{ij}^{(f)} = \frac{|x_{if} - x_{jf}|}{max_h x_{hf} - min_h x_{hf}}$$
   Nilai h diperoleh dari semua objek yang lengkap untuk atribut f. Atribut numerik harus diskalakan ke rentang [0,0;1,0].
   Untuk variabel Arm karena bertipe nominal maka menggunakan rumus:
   $$d_{ij}^{(f)} \begin{cases} 0 \; jika \; x_{if} = x_{jf} \\ 1 \; lainnya \end{cases}$$
   Kemudian menggabungkan kedua nilai jarak antara x_i dan x_j untuk tiap titik data menggunakan rumus:
   $$d(i,j) = \frac{\sum_{f=1}^{p} \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^{p} \delta_{ij}^{(f)}}$$
   Di mana $\delta_{ij}^{(f)}$=0 jika x_{if} atau x_{jf} kosong (missing) atau $x_{if} = x_{jf} = 0$ dan f adalah atribut biner asimetris dan $\delta_{ij}^{(f)}$=1 untuk kondisi yang lain. Sedangkan $d_{ij}^{(f)}$ adalah kontribusi atribut f terhadap ketidakmiripan antara objek data i dan objek data j, yang dihitung berdasarkan jenis atribut
7. Mencari sebanyak *K* tetangga terdekat yang sama untuk setiap data tersensor
8. Memperoleh sebanyak *K* data lengkap terdekat (z_i)
9. Menghitung rata-rata dari sebanyak *K* data lengkap terdekat dari z_i untuk memperoleh (y_i^{knn})
10. Menggantikan data tersensor (z_i, δ_i=0) dengan nilai rata-rata sebanyak *K* tetangga terdekat (y_i^{knn})
11. Selesai

C. Hasil Penelitian dan Pembahasan

Statistika Deskriptif

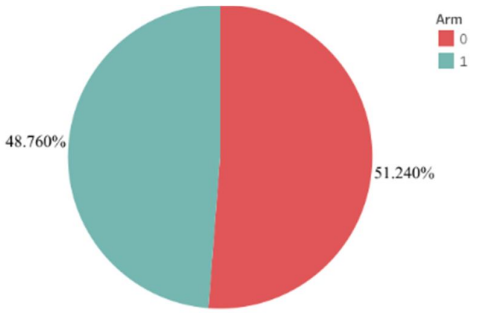
Deskripsi data bertujuan untuk mengetahui karakteristik data pasien kanker paru-paru sel kecil. Data terdiri dari 121 pasien yang memiliki 4 variabel, variabel cens dan Arm bertipe nominal sedangkan variabel waktu *survival* (*Time*) dan usia (*Age*) merupakan variabel bertipe rasio. Berikut diagram lingkaran dari variabel cens data pasien kanker paru-paru sel kecil NCCTG, disajikan pada Gambar 1.



Gambar 1. Persentase Data Tersensor Kanan pada Pasien Kanker Paru-paru Sel Kecil NCCTG

Data tersensor kanan ditandai dengan bagian diagram lingkaran berwarna jingga dan notasi yang digunakan yaitu satu (1) sedangkan data lengkap ditandai dengan bagian diagram lingkaran berwarna biru dengan notasi nol (0). Terlihat bahwa persentase data tersensor kanan pada pasien kanker paru-paru sel kecil sebesar 19,01% atau sebanyak 23 data yang tersensor kanan, sedangkan data lengkap memiliki persentase sebesar 80,99% atau sebanyak 98 data yang lengkap.

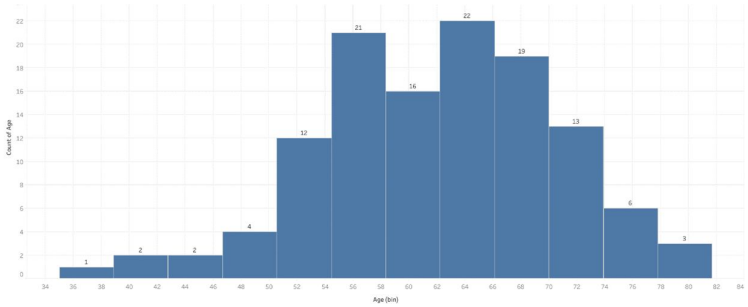
Berikut diagram lingkaran dari variabel Arm data pasien kanker paru-paru sel kecil NCCTG, disajikan pada Gambar 2.



Gambar 2. Persentase Arm Pasien Kanker Paru-paru Sel Kecil NCCTG

Pada Gambar 2 bagian diagram lingkaran yang berwarna merah adalah pasien yang menerima Arm A dengan notasi nol (0), sedangkan bagian diagram lingkaran yang berwarna hijau adalah pasien yang menerima Arm B dengan notasi satu (1). Terlihat bahwa pasien yang menerima Arm A terdapat sebanyak 51,24% atau sebanyak 62 pasien, sedangkan pasien yang menerima Arm B terdapat sebanyak 48,76% atau sebanyak 59 pasien.

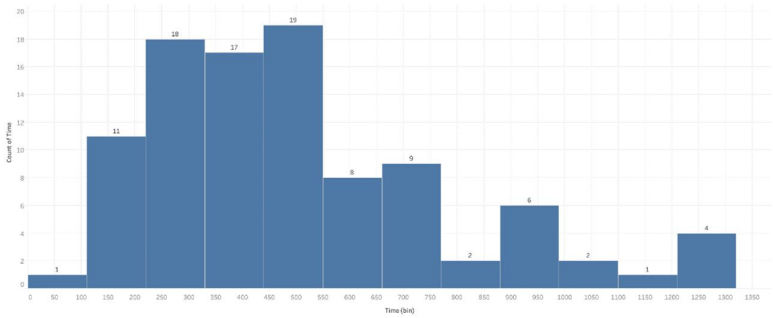
Berikut diagram histogram dari variabel usia data pasien kanker paru-paru sel kecil NCCTG, disajikan pada Gambar 3.



Gambar 3. Usia Pasien Kanker Paru-paru Sel Kecil

Gambar 3 merupakan histogram dari usia pasien kanker paru-paru sel kecil, pasien terbanyak berada pada rentang usia 62-66 tahun sedangkan pasien tersedikit berada pada rentang usia 35-39 tahun.

Berikut diagram histogram dari variabel waktu *survival* data lengkap pasien kanker paru-paru sel kecil NCCTG, disajikan pada Gambar 4.



Gambar 4. Waktu *Survival* Pasien Kanker Paru-paru Sel Kecil

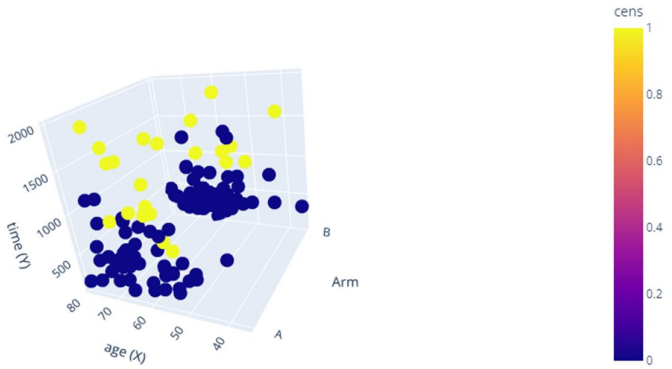
Gambar 4. adalah histogram dari data lengkap waktu *survival* pasien kanker paru-paru sel kecil, rentang waktu *survival* paling lama berada pada rentang 440-550 hari rentang yang paling sedikit berada pada rentang 0-110 hari dan 1100-1210 hari. Statistik deskriptif untuk variabel usia dan waktu dideskripsikan berdasarkan rata-rata, maksimum, minimum, modus dan median (nilai tengah) ditampilkan pada Tabel 1.

Tabel 1. Statistik Deskriptif Variabel Usia dan Waktu

Variabel	Usia (*Age*)	Waktu (*Time*)
Rata-rata	62,12	505,01
Minimum	36	83
Maksimum	79	1315
Modus	68	490
Median	63	441,5

Gambar 5 disajikan diagram tebar untuk data pasien kanker paru-paru sel kecil, dengan variabel usia (Age) sebagai sumbu X, variabel waktu *survival* (Time) sebagai sumbu Y dan variabel Arm sebagai sumbu Z variabel cens diwakili oleh perbedaan warna di mana data lengkap dengan notasi nol (0) berwarna biru sedangkan data tersensor berwarna kuning dengan notasi satu (1).

Berikut diagram tebar dari data pasien kanker paru-paru sel kecil NCCTG, disajikan pada Gambar 5.



Gambar 5. Diagram Tebar Pasien Kanker Paru-paru Sel Kecil

Imputasi KNN

Imputasi KNN memperkirakan nilai yang cukup layak untuk mengganti nilai yang tersensor (z_i) berdasarkan K tetangga terdekat. Penentuan tetangga terdekat dihitung berdasarkan jarak, variabel yang digunakan untuk mengukur jarak yaitu usia (X_1) dan Arm (X_2). Data yang tersensor (z_i) merupakan variabel waktu *survival* yang berupa variabel numerik. Indikator penyensoran (δ) adalah variabel cens, data dikatakan lengkap apabila pasien meninggal selama penelitian (cens=0) sedangkan data dikatakan tersensor kanan (cens=1) apabila pasien masih hidup saat penelitian berakhir. Karena data yang ingin diimputasi bertipe numerik maka imputasi menggunakan nilai rata-rata dari K tetangga terdekat.

Setelah mendapatkan jarak dari seluruh nilai $d(i,j)$ antara data lengkap dan data tersensor, maka langkah selanjutnya adalah menentukan banyaknya K yang digunakan. Karena ketika K=1 memiliki banyak tetangga yang berbeda-beda perlu dilakukan iterasi untuk mencari banyaknya K yang sesuai atau memiliki nilai yang setara agar banyaknya tetangga terdekat sama untuk setiap data tersensor (z_i). Setelah dilakukan iterasi diperoleh banyaknya K tetangga terdekat yang setara yaitu *K=46*

Kemudian rata-rata dari sebanyak 46 tetangga terdekat data tersensor (z_i) akan menjadi nilai imputasi (y_i^{knn}) untuk data tersensor kanan. Tabel 2 menampilkan hasil imputasi untuk data tersensor kanan Pasien Kanker Paru-paru Sel Kecil NCCTG.

Tabel 2. Hasil Imputasi untuk Data Tersensor Kanan (y^{knn}) Pasien Kanker Paru-paru Sel Kecil

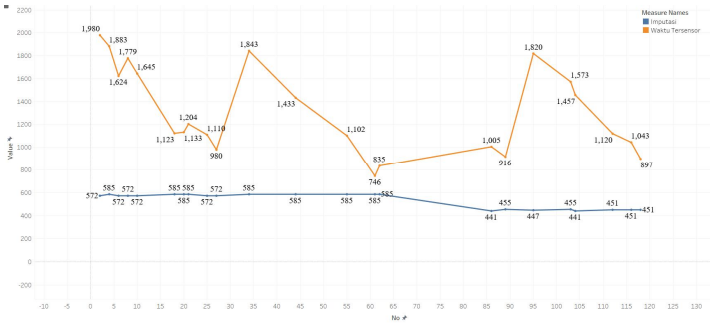
No	**2**	**4**	**6**	**8**	**10**	**18**	**20**	**21**
y^{knn}	572.4	585.4	572.4	572.4	572.4	585.4	585.4	585.4

No.	**25**	**27**	**34**	**44**	**55**	**61**	**62**
y^{knn}	572.4	572.4	585.4	585.4	585.4	585.4	585.4

No.	**86**	**89**	**95**	**103**	**104**	**112**	**116**	**118**
y^{knn}	440.9	454.9	447.5	454.9	440.9	451.0	451.0	451.0

Gambar 6 membandingkan nilai saat tersensor dengan nilai hasil imputasi, terlihat bahwa nilai hasil imputasi memiliki kemiripan sebab K yang dimiliki besar yaitu sebanyak 46, sehingga rata-rata satu nilai dengan nilai yang lain mirip.

Berikut diagram garis hasil imputasi data tersensor kanan (z_i) pasien kanker paru-paru sel kecil NCCTG, disajikan pada Gambar 6.



Gambar 6. Grafik Perbandingan Nilai saat Tersensor dengan Nilai Hasil Imputasi

D. Kesimpulan

Berdasarkan pembahasan dalam penelitian ini, peneliti menyimpulkan beberapa hasil penelitian sebagai berikut:

1. Data pasien kanker paru-paru sel kecil *North Central Cancer Treatment Group* (NCCTG) memiliki tingkat penyensoran sebesar 19,01%, pasien yang menerima Arm A terdapat 51,24% dan Arm B sebesar 48,76%, variabel usia memiliki rata-rata 62,12 tahun, minimum 36 tahun, maksimum 79 tahun, modus 68 tahun dan median 63 tahun, sedangkan variabel waktu memiliki rata-rata 505,01 hari, minimum 83 hari, maksimum 1315 hari, modus 490 hari dan median 441,5 hari.
2. Imputasi *K Nearest Neighborhood* (KNN) pada data pasien kanker paru-paru sel kecil NCCTG efektif diterapkan. KNN digunakan untuk mengimputasi waktu *survival* yang tersensor kanan dengan rata-rata sebanyak *K* data lengkap waktu *survival* berdasarkan karakteristik variabel usia dan Arm (jenis perawatan) yang paling mirip. Data pasien kanker paru-paru sel kecil berhasil diimputasi dengan *K* sebesar 46.

Acknowledge

Peneliti ingin mengucapkan terima kasih kepada Allah swt yang telah memberikan rezeki dan berkah serta selalu ada kapan pun dan dimana pun untuk peneliti. Peneliti juga ingin mengucapkan terima kasih kepada Pak Abdul Kudus, M.Si., Ph.D. yang telah memberikan bimbingan hingga penelitian ini dapat terselesaikan, dosen-dosen statistika Unisba yang telah memberikan ilmu, bimbingan dan wawasan kepada peneliti, orangtua, keluarga, rekan-rekan seperjuang statistika Unisba atas bantuan, semangat, doa serta bimbingannya.

Daftar Pustaka

[1] Evriyanto, Y. (2004). Perbandingan Metode Imputasi untuk Mengestimasi Data Hilang Pada Data Kesehatan Ibu dan Anak di Jawa Timur. Skripsi. Diambil kembali dari http://repository.unair.ac.id/id/eprint/35857

[2] Ahmed, S. E., Aydin, D., & Yilmaz, E. (2020). Nonparametric Regression Estimates Based on Imputation Techniques for Right-Censored Data. Proceedings of the Thirteenth International Conference on Management Science and Engineering Management, 109-120. doi:10.1007/978-3-030-21248-3_8

[3] Jerez, J. M., Molina, I., Garcia-Laecina, P. J., Alba, E., Ribelles, N., Martin, M., & Franco, L. (2010). Missing data imputation using statistical and machine learning methods in a real. Artificial Intelligence in Medicine, 50(2), 105-115. doi:10.1016/j.artmed.2010.05.002

[4] Malarvizhi, M. R., & Thanamani, A. S. (2012, November). Framework for Missing Value Imputation. International Journal of Engineering Research and Development, 4(7), 14-16.

[5] Imandoust, S. B., & Bolandraftar, M. (2013, Sept-Oct). Application of K-Nearest Neighbor (KNN) Approach for Predicting Economin Events: Theoretical Background. Int. Jurnal of ENgineering Research and Applications, 3(5), 605-610

[6] Batista, G. E., & Monard, M. (2003). An Analysis of Four Missing Data Treatment Methods for Supervised Learning. Applied Artificial Intelligence: An International Journal, 17, 519-533. doi:10.1080/08839510390219309

[7] Mawarsari, U., & Irhamah. (2016). Imputasi Missing Data dengan *K-Nearest Neighborhood* dan Algoritma Genetika. AdMathEdu, 6(1), 1-13.

[8] Maksymiuk, A. W., Jett, J. R., Earle, J. D., Su, J. Q., Diegert, F. A., Mailliard, J. A., . . . Levitt, R. (1994, Januari). Sequencing and Schedule Effects of Cisplatin Plus Etoposide in Small-Cell Lung Cancer: Results of a *North Central Cancer Treatment Group* Randomized Clinical Trial. Journal of Clinical Oncology, 12(1), 70-76.

[9] Ying, S., Jung, S. H., & Wei, L. J. (1995). *Survival* Analysis with Median Regression Models. Journal of the American Statistical Association, 90, 178-184. Diambil kembali dari http://www.jstor.org/stable/2291141 .

[10] Suyanto. (2017). Data Mining untuk Klasifikasi dan Klasterisasi Data. Bandung: Informatika Bandung.

[11] Yulianto, Anggi Priliani, Darwis, Sutawanir. (2021). *Penerapan Metode K-Nearest Neighbors (kNN) pada Bearing*, Jurnal Riset Statistika, 1(1), 10-18.