JINTEKS (Jurnal Informatika Teknologi dan Sains) ISSN 2686-3359 (Online)
Vol. 4 No. 4, November 2022, hlm. 386 – 390

PENERAPAN KNNIMPUTER DALAM MENGOLAH DATA MISSING VALUE UNTUK MEMBANTU MENINGKATKAN AKURASI SUPPORT VECTOR MACHINE KLASIFIKASI PENYAKIT TIROID

Supardianto[1*], Lalu Mutawali[2], Wafiah Murniati[3]
[1]Program Studi Teknik Informatika, Universitas Teknologi Mataram
[2,3]Program Studi Sistem Informasi, STMIK Lombok
email: supardianto88mkom@gmail.com[1*]

Abstrak: Tiroid adalah kondisi kelainan pada seorang akibat adanya gangguan tiroid. Berdasarkan data dari kementerian kesehatan di dunia prevalensi tiroid masih tergolong tinggi, jika kelahiran sebanyak lima juta bayi setiap tahunya maka terdapat seribu enam ratus bayi dengan hipertiroid. Algoritma yang digunakan untuk pengolahan data dan dimodelkan menjadi pengetahuan adalah support vector machine (SVM), SVM adalah digunakan untuk klasifikasi. Setelah melakukan melakukan explorasi pada datasets dari 23 atribut yang terdapat pada datasets terdapat 9 atribut yang memiliki missing value Antara lain age 4 baris, sex 307 baris, TSH 804 baris, T3 2604 baris, TT4 442 baris, T4U 809 baris, FTI 802 baris, TBG 8823 baris, dan target 1626 baris. Berdasarkan hasil evaluasi pada model yang telah dibuat pengujian presisi 94%, recall 100%, F1-score 97% dengan hasil akumulasi akurasi sebanyak 93%. Total keseluruhan evaluasi pada model adalah 93%.

Kata Kunci : klasifikasi, tyroid, svm, knnimputer

Abstract: *Thyroid is a condition of abnormalities in a person due to thyroid disorders. Based on data from the ministry of health in the world, the prevalence of thyroid is still relatively high, if five million babies are born each year, then there are one thousand six hundred babies with hyperthyroidism. The algorithm used for data processing and modeled into knowledge is a support vector machine (SVM), SVM is used for classification. After exploring the datasets of the 23 attributes contained in the datasets, there are 9 attributes that have missing values, including age 4 lines, sex 307 lines, TSH 804 lines, T3 2604 lines, TT4 442 lines, T4U 809 lines, FTI 802 lines, TBG 8823 lines, and target 1626 lines. Based on the evaluation results on the model that has been tested for precision 94%, recall 100%, F1-score 97% with an accumulated accuracy of 93%. The overall total evaluation on the model is 93%.*

Keywords **:** Classification, thyroid, svm, knnimputer

PENDAHULUAN

Tiroid merupakan endokrin murni terbesar dalam tubuh manusia, hormone tiroid memililiki peran penting dalam berbagai proses metabolisme (protein, karbohidrat, dan lemak), penyakit atau gangguan tiroid adalah kondisi kelainan pada seorang akibat adanya gangguan tiroid, ketika tiroid mengalami gangguan akan menakibatkan terseang gondok[1]. Berdasarkan data dari kementerian kesehatan di dunia prevalensi tiroid masih tergolong tinggi, jika kelahiran sebanyak 5 juta bayi setiap tahunya maka terdapat 1.600 bayi dengan hipertiroid, prevalensi di Indonesia mencapai 17 juta gangguan tiroid[2]. Data dari kementerian kesehatan menunjukkan penting untuk melakukan screening lebih dini, agar setiap orang dapat memeriksa kondisi. Proses skrining lebih cepat dilakukan apabila terdapat sebuah sistem yang dapat digunakan untuk membantu.

Pengembangan sistem untuk melakukan langkah awal pemeriksaan terhadap tiroid membutuhkan basis pengetahuan. Saat ini basis pengetahuan dapat dimodelkan dengan menggunakan pendekatan machine learning. Machine learning adalah cabang ilmu komputer yang berfokus pada penggunaan data dan algoritma untuk meniru yang dipelajari manusia dan dapat digunakan untuk melakukan prediksi[3]. Untuk memodelkan pengetahuan dibutuhkan juga dataset sebagai bahan yang dapat digunakan belajar oleh teknologi komputer. Datasets yang digunakan adalah datasets yang diambil dari situs UCI Datasets yang bersumber dari Garavan Institute in Sydnie, Australia[4].

Algoritma yang digunakan untuk pemodelan pengetahuan adalah support vector machine (SVM), SVM adalah algoritma machine learning yang tergolong serbaguna mampu melakukan klasifikasi secara linier dan non-linier[5]. Akan tetapi, saat melakukan pemodelan pada data yang terdapat di datasets tiroid SVM tidak dapat dijalankan, penyebabnya adalah terdapat banyak data yang bernilai nilai missing (missing value) pada datasets. Setelah melakukan uji coba explorasi pada datasets dari 23 atribut yang terdapat pada datasets terdapat 9 atribut yang memiliki missing value Antara lain age 4 baris, sex 307 baris, TSH 804 baris, T3 2604 baris, TT4 442 baris, T4U 809 baris, FTI 802 baris, TBG 8823 baris, dan target 1626 baris. Banyaknya atribut yang bernilai missing sangat berpengaruh terhadap pembuatan model. Upaya yang dilakukan dapat melakukan penghapusan data, namun menggunakan teknik penghapusan akan tidak relevan karena banyaknya data missing value pada datasets, sehingga pembuatan model dengan SVM tidak dapat dilakukan. Upaya yang dilakukan dalam menangani datasets yang mengandung data missing value adalah menerapkan algoritma K-Nearest Neighbours Imputation (KNNImputer) yang merupakan algoritma yang sangat baik dalam menangani data missing values. KNNImputer berfungsi untuk memberikan nilai yang hilang pada dataset diperhitungkan dengan nilai rata-rata dari neighbors yang memiliki kemiripan yang paling dekat[6].

 ISSN 2686-3359 (Online)

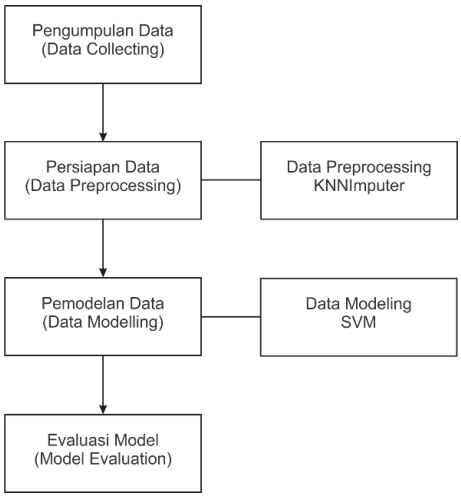
Dengan menerapkan KNNImputer untuk menangani data missing value akan dapat menambah performance SVM melakukan pemodelan data secara lebih baik, sehingga dapat menghasilkan basis pengetahuan yang akurat dalam mempresiksi penyakit tiroid.

TINJAUAN PUSTAKA

Pada bagian ini dilakukan tinjauan terhadap beberapa penelitian terdahulu, terkait dengan penerapan metododologi imputasi KNN dalam menangani kasus missing values atau data yang memiliki nilai kosong atau hilang. Penelitian menujukkan bahwa metode imputasi KNN dapat digunakan dan memiliki hasil yang baik dalam menagani data yang hilang[7]. Data dalam perawatan medis sering mengalami permasalahan kehilangan ini menyebabkan dataset akan menjadi missing, penelitian dilakukan untuk mengkoparasi imputasi dengan regresi dan KNN untuk mendapatkan hasil yang paling baik, diantara kedua metodologi tersebut KNNImputer lebih baik dalam menyelesaikan permasalahan data yang hilang[8]. Menerapkan metodologi imputasi dapat membantu dalam membentuk model yang lebih akurat dalam memprediksi penyakit kangker, penelitian menunjukkan penerapan KNNImputer lebih dibandingkan dengan melakukang penghapusan data secara manual[9]. Penerpaan metodologi imputasi juga diterapkan pada kasus klasifikasi cuaca, mengelola data dengan KNNImputer menjadi langkah yang baik untuk mendapatkan model dengan akurasi yang lebih baik[10]. Penerapan imputasi dapat membantu model-model yang dibagun dengan pendekatan machine learning[11]. Berdasarkan justifikasi penelitian terdahulu menyatakan bahwa pendekatan imputasi dapat meningkatkan akurasi terhadap model machine learning. Pada penelitian selanjutnya penerapan KNNImputer sebagai metodologi imputasi guna menyelesaikan permasalahan data missing pada data tiroid agar pemodelan dengan menggunakan SVM dapat lebih akurat dalam melakukan klasifikasi penyakit tiroid.

METODE

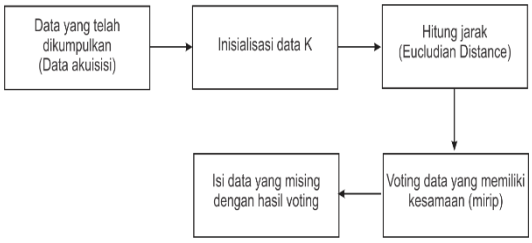
Penelitian ini bertujuan untuk membuat model yang lebih akurat dalam melakukan klasifikasi terhadap penyakit tiroid. Penelitian ini juga membutuhkan algoritma dalam melakukan data preposesing karena banyaknya jumlah data missing yang terkadung dalam dataset. Adapun tahapan pada penelitian meliputi empat tahapan yaitu pertama Pengumpulan data (data collecting) pengumpulan data melalui situs dapat diakases melalui link berikut https://archive.ics.uci.edu/ml/datasets/thyroid+disease[12]. Tahap kedua Persiapan data (data preprocessing) data yang telah dikumpuklan kemudian diekplorasi untuk mengetahui data yang memiliki missing value, apabila banyak terdapat data missing value maka tidak relvan apabila dilakukan reduksi. Upaya yang paling tepat adalah menerapkan algoritma KNNImputer untuk menyelesaikan permasalahan pada penelitian ini yang masih terdapat banyak data yang missing. Tahap ketiga pembuatan model, setelah data sudah siap missing value sudah teratasi selajutnya menerapkan support vector machine dengan kernel Radial Basic Function (RBF) untuk membuat model klasifikasi. Tahap keempat evaluasi pada model, model yang telah dibuat kemudian diuji untuk mengetahui seberapa baik model yang telah dibuat dalam melakukan klasifikasi terhadap penyakit tiroid. Gambar 1. menunjukkan alur penelitian.



Gambar 1. Tahapan Penelitian

Tahapan Data Preporcessing

Pada penelitian ini pencapaian model yang akurat dalam melakukan klasifikasi penyakit adalah prioritas. Akan tetapi capaian tersebut akan sulit dilakukan apabila sumber daya data yang digunakan (dimodelkan) masih terdapat banyak data yang missing. Tahapan ini menunjukan perbaikan pada datasets dengan menerapkan algotima KNNImputer. Gambar 2. menunjukkan tahapan-tahapan untuk menyelesaiakan permasalahan missing value.

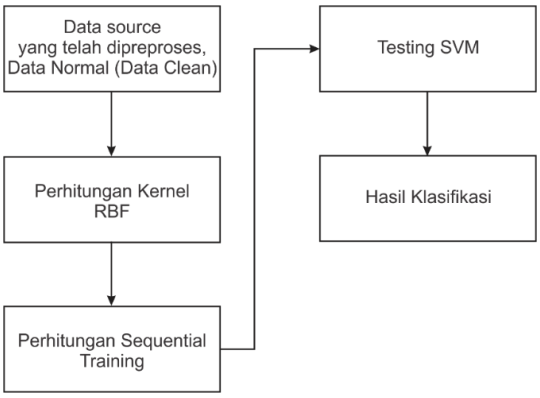


Gambar 2. Tahapan data preprocessing dengan KNNImputer

Tahapan pemodelan data

Data yang sudah dilakukan preprocess pada tahap sebelumnya dinyatakan sudah bersih dari missing value. Data yang sudah siap kemudian dimodelkan. Pada penelitian ini pembuatan model dengan menerapkan algoritma support vector machine (SVM). Gambar 3. menunjukkan alur pemodelan data menggunakan SVM.

 ISSN 2686-3359 (Online)



Gambar 3. Tahapan Pemodelan dengan SVM

Evaluasi Model

Model yang dibangun penting untuk dilakukasan evaluasi, tahapan ini bertujuan untuk menguji model telah dapat dengan baik dalam melakukan klasifikasi terhadap penyakit tiroid. Pada penelitian ini metode evaluasi yang digunakan adalah confusion matrix dan evaluasi model dengan menguji performance dengan mengukur akurasi, presisi, recall (sensitivity), F1-score atau dapat diartikan sebagai hasil dari hasil kategori dengan model dengan hasil yang sebenarnya.

HASIL DAN PEMBAHASAN

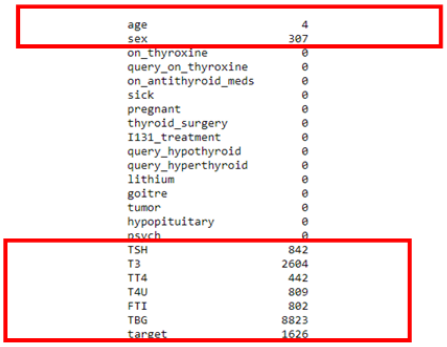
Data yang digunakan pada penelitian ini adalah data yang bersumber dari UCI Datasets, jumlah data yang terdapat di dalam datasets yaitu terdapat 31 kolom (fitur) dan masing-masing kolom memiliki nilai sebanyak 9172 baris data, dari 9172 terdapat 22 fitur yang digunakan untuk bahan pembuatan model, 22 fitur tersebut memiliki relevansi yang paling kuat terhadap penyakit tiroid. pada tebel 1. menunjukkan data keseluruhan fitur yang terdapat di dataset dan fitur yang digunakan sebagai bahan pembelajaran untuk pembuatan model.

Tabel 1. Data Fitur

Data Seluruh fitur	Data Yang digunakan untuk bahan ajar
atient_id, sex, on_thyroxine, query_on_thyroxine, on_antithyroid_meds,sick, pregnant, thyroid_surgery, I131_treatment, query_hypothyroid, query_hyperthyroid, lithium, goiter, tumor, hypopituitary, psych, TSH_measured, TSH, T3_measured, T3, TT4_measured, TT4, T4U_measured, T4U, FTI_measured, FTI, TBG_measured, TBG, target	Age, sex, on_thyroxine, query_on_thyroxine, on_antithyroid_meds, sick, Pregnant,thyroid_surgery, I131_treatment, query_hypothyroid, query_hypert, yroid, lithium, goiter, tumor, hypopituitary, psych, TSH, T3, TT4, T4U, FTI, TBG, target

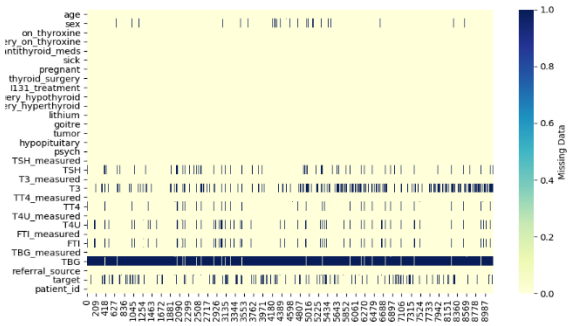
Eksplorasi Data dan Data Preprocessing

Sebelum melakukan pengolahan data penting untuk melakukan ekplorasi terhadap datasets, hal ini bertujuan untuk menentukan teknik-teknik yang paling tepat dalam preposesing, analisis, dan pemodelan. Melakukan kualtifikasi pada fitur yang memiliki data missing. Terdapat 9 fitur yang memiliki data missing value yaitu fitur age 4 baris, sex 307 baris, TSH 804 baris, T3 2604 baris, TT4 442 baris, T4U 809 baris, FTI 802 baris, TBG 8823 baris, dan target 1626 baris. . Pada gambar 4 merupakan hasil ekplorasi fitur dan jumlah datasets yang missing



Gambar 4. Data Missing Value

Data missing juga dapat direpresentasikan dalam bentuk visual, tujuanya adalah untuk mengetahui sebarannya. Pada gambar 5 ditunjukkan bahwa apabila terdapat warna biru yang padat maka menunjukkan semakin banyak terdapat data missing value pada fitur tersebut.



Gambar 5. Gambar visualisasi data missing

Data Preprocesing

Pada tahapan data preprocesing tahap awal melakukan pengecekan data redudansi, pada datasets terdapat 7 fitur yang redudansi. Data redudansi adalah merupakan beberapa salinan dari informasi yang sama disimpan di lebih dari satu tempat sekaligus[12]. Data redudansi akan menyebabkan data akan sulit diolah secara optimal. Upaya yang dilakukan adalah melakukan droping (menghapus) ke 7 fitur yang redudansi sehingga menyisakan 22 fitur yang akan diolah pada tahapan selanjutnya. Setelah melakukan proses droping tahapan selanjunya melakukan pembersihan data pada fitur age (usia) didalam dataset terdapat 4 baris yang memiliki nilai usia yang tinggi yaitu, pada baris 2976 berusia 455 tahun, baris 5710 berusia 65511 tahun, baris 6392

 ISSN 2686-3359 (Online)

berusia 65512 tahun, dan baris 8105 berusia 65526 tahun, gambar 7. menunjukkan data usia yang harus dibersihkan. Data tersebut diubah menjadi nilai null kemudian disi dengan nilai rata-rata usia yang terdapat pada dataset.

	age	sex	on_thyroxine	query_on_thyroxine	on_antithyroid_meds	sick	pregnant
2976	455	F	f	f	f	f	f
5710	65511	M	f	f	f	f	f
6392	65512	M	f	f	f	f	f
8105	65526	F	f	f	f	f	f

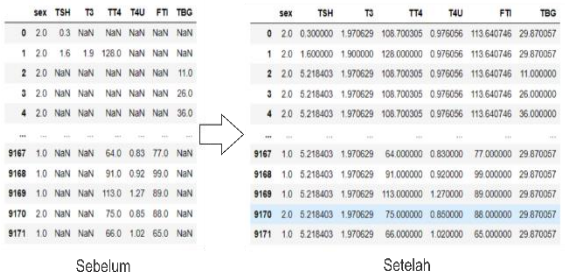
Gambar 7. Data Mising pada Fitur Age

Setelah membersihkan data missing pada fitur age dengan pendekatan menghitung nilai rata-rata, selannya pada fitur yang seperti pada fitur sex 307 baris, TSH 842 baris,T3 2604, TT4 442, T4U 809, FTI 802, TBG 8823 memiliki jumlah data missing yang sangat banyak. Hal ini akan tidak relevan jika melakukan droping atau menggunakan pedekatan nilai rata-rata. Tahapan yang paling cocok adalah dengan menerapkan algoritma KNNImputer. Gambar 8 menunjukkan penerapan algoritma KNNImputer.

```
from sklearn.impute import KNNImputer
imputer  = KNNImputer(n_neighbors=5)
df_col_miss = pd.DataFrame(
    imputer.fit_transform(df_col_miss), columns=df_col_miss.columns
)
```

Gambar 8. Penerapan KNNImputer

Setelah menerapkan KNNImputer fitur-fitur yang memiliki data missing kini telah teselesaikan. Pada gambar 9 merupakan data yang sebelum di preprocessing dan data yang olah menggunakan algoritma KNNImputer.



Gambar 9. Fitur sebelum di proses dan setelah diproses

Setelah melakukan preprocessing data missing menggunakan KNNImputer proses selanjunya adalah melakukan penskalaan. Penskalaan dataset bertujuan agar nilai pada masing-masing fitur tidak memiliki perbedaan nilai yang terlampau tinggi. Perbedaan nilai yang terlapau tinggi menyebabkan hasil prediksi dari model menjadi tidak akurat. Pada gambar 10 menunjukkan hasil penskalaan pada masing-masing nilai yang terdapat didalam fitur.

	sex	TSH	T3	TT4	T4U	FTI	TBG
0	1.0	0.000557	0.106999	0.178429	0.373174	0.127604	0.148925
1	1.0	0.003009	0.103064	0.210702	0.373174	0.127604	0.148925
2	1.0	0.009837	0.106999	0.178429	0.373174	0.127604	0.054527
3	1.0	0.009837	0.106999	0.178429	0.373174	0.127604	0.129565
4	1.0	0.009837	0.106999	0.178429	0.373174	0.127604	0.179590
...	...	...	...	...	...	...	...
9167	0.0	0.009837	0.106999	0.103679	0.305556	0.085948	0.148925
9168	0.0	0.009837	0.106999	0.148829	0.347222	0.110960	0.148925
9169	0.0	0.009837	0.106999	0.185619	0.509259	0.099591	0.148925
9170	1.0	0.009837	0.106999	0.122074	0.314815	0.098454	0.148925
9171	0.0	0.009837	0.106999	0.107023	0.393519	0.072306	0.148925

Gambar 10. Hasil Penskalaan Nilai Setiap Fitur

Data Modeling dan Evaluasi

Pada tahap sebelumnya telah dilakukan tahapan data preprocesing dengan menggunakan algoritma KNNImputer. Hasil pengolahan menggunkan KNNImputer menyelesaikan permasalahan data missing value, setelah itu tahap penskalaan telah dilakukan. Setelah melakukan pembersihan pada data kini data sudah siap untuk dimodelkan. Algoritma yang digunakan untuk pemodelan adalah Support Vector Machine (SVM). Pada tahap pemodelan membagi data training (data latih) dan data testing (data uji), pembagian diantaranya data training sebanyak 75% dari jumlah keseluruhan data sedang kan sisanya sebanyak 25% dijadikan sebagai data testing. Pada pembuatan model dengan SVM kernel yang digunakan adalah Radial Basis Function (RBF) yang dikenal akurat dalam memisahkan data non-linear. Gambar 11 menunjukkan proses penerapan algoritma SVM untuk pembuatan model.

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split (
    X, y, test_size=0.25, random_state=120183
)
```

```
from sklearn import svm
model = svm.SVC(kernel='rbf', C=1.0)
model.fit(X_train, y_train)
y_pred = model.predict(X_test)
```

Gambar 11. Penerapan Algoritma SVM

Evaluasi Model

Tahapan terakhir adalah melakukan evaluasi pada model yang telah dibuat, evaluasi model bertujuan untuk mengetahui tingkat akurasi dan balance model yang telah dibuat. Metode yang digunakan adalah confusion matrix dengan mempertimbangkan tingkat akurasi, presisi, recall, dan F1-score. Berdasarkan hasil evaluasi pada model yang telah dibuat pengujian presisi 94%, recall 100%, F1-score 97% dengan hasil akumulasi akurasi sebanyak 93%. Jadi total keseluruhan hasil evaluasi pada model adalah 93%, pada gambar 12 menujukkan proses evaluasi pada model yang telah dibuat.

	precision	recall	f1-score	support
0.0	0.94	1.00	0.97	2094
1.0	0.92	0.30	0.45	150
2.0	0.73	0.22	0.34	49
accuracy			0.93	2293
macro avg	0.86	0.51	0.59	2293
weighted avg	0.93	0.93	0.92	2293

Gambar 12. Evaluasi Model

ISSN 2686-3359 (Online)

KESIMPULAN DAN SARAN

Dalam pengembangan model pengetahuan menggunakan machine learning data yang baik dan terbebas dari nilai missing sangat berpengaruh terhadap performance model. Data yang digunakan pada penelitian ini memiliki data missing yang sangat tinggi sehingga menerapkan algoritma machine learning seperti support vector machine (SVM) tidak bisa digunakan membuat model yang baik, berdasarkan hasil explorasi pada datasets dari 23 atribut yang terdapat pada datasets terdapat 9 atribut yang memiliki missing value Antara lain age 4 baris, sex 307 baris, TSH 804 baris, T3 2604 baris, TT4 442 baris, T4U 809 baris, FTI 802 baris, TBG 8823 baris, dan target 1626 baris. Banyaknya atribut yang bernilai missing sangat berpengaruh terhadap pembuatan model. Upaya yang dilakukan dalam menangani datasets yang mengandung data missing value adalah menerapkan algoritma K-Nearest Neighbours Imputation (KNNImputer) yang merupakan algoritma yang sangat baik dalam menangani data missing values. Berdasarkan hasil evaluasi pada model yang telah dibuat pengujian presisi 94%, recall 100%, F1-score 97% dengan hasil akumulasi akurasi sebanyak 93%. Total keseluruhan evaluasi pada model adalah 93%, sehingga. Dengan menerapkan KNNImputer untuk menangani data missing value dapat menambah performance SVM klasifikasi lebih baik.

DAFTAR PUSTAKA

[1] Kemenkes, “Internasional Thyroid Award Bebaskan Dirimu dari Gangguan Tiroid,” Jakarta: Kementerian Kesehatan Republik Indonesia, 2015, p. 1. [Online]. Available: https://pusdatin.kemkes.go.id/resources/download/pusdatin/infodatin/infodatin-tiroid.pdf

[2] Kemenkes, “Prevalensi Gangguan Tiroid,” Bethesda Hospital, 2020. https://bethsaidahospitals.com/waspada-gangguan-tiroid/#:~:text=Prevalensi Gangguan Tiroid,gangguan tiroid umumnya tidak diketahui. (accessed Jul. 01, 2022).

[3] IBM, “Machine Learning,” IBM Cloude Educations, 2020. https://www.ibm.com/cloud/learn/machine-learning (accessed Jul. 01, 2022).

[4] R. Quinlan, “Thyroid Disease Data Set,” Thyroid Disease Data Set. https://archive.ics.uci.edu/ml/datasets/thyroid+disease (accessed Jul. 03, 2022).

[5] A. Geron, Hands-On Machine Learning With Scikit Learn and TensorFlow. California: O’Reilly, 2017.

[6] J. Brownlee, “KNN Imputation for Missing Values in Machine Learning,” Machine Learning Mastery, 2020. https://machinelearningmastery.com/knn-imputation-for-missing-values-in-machine-learning/ (accessed Jul. 05, 2022).

[7] C. Wongoutong, “Imputation methods for missing response values in the three parts of a central composite design with two factors,” J. Stat. Simul., vol. 92, no. 11, pp. 2273–2289, 2022.

[8] S. Batra, R. Khurana, M. Z. Khan, W. Baulilla, A. Kaubaa, and P. Srivastava, “A Pragmatic Ensemble Strategy for Missing Values Imputation in Health Records,” J. Entropy, vol. 24, no. 4, pp. 2–20, 2022.

[9] P. Keerin, “Improved KNN Imputation for Missing Values in Gene Expression Data,” Comput. aterials Contin., vol. 70, no. 22, pp. 4020–4023, 2022.

[10] W. Y. Lai, K. Kuok, Sh. Gato-Trinidad, and K. X. Ling, “A Study on Sequential K Nearest Neighbor (SKNN) Imputation for Treating Missing Rainfall Data,” Int. J. Adv. Trends Comput. Sci. Eng., vol. 8, no. 3, pp. 363–368, 2019.

[11] K. Hasan, A. Alam, S. Roy, and A. Dutta, “*Missing* value imputation affects the performance of machine learning: A review and analysis of the literature,” Informatics Med. Unlocked, vol. 27, pp. 1–23, 2021.

[12] M. Readdy, “What Is Data Redundancy,” Morgan Kaufmann, 2011.