Journal of Applied Informatics and Computing (JAIC)
Vol.6, No.2, Desember 2022, pp. 194~204
e-ISSN: 2548-6861

Optimasi Algoritma K-Nearest Neighbors dengan Teknik Cross Validation Dengan Streamlit (Studi Data: Penyakit Diabetes)

Aditya Budi Prasetyo [1]*, Tri Ginanjar Laksana [2]*
* Teknik Informatika, Fakultas Informatika, Institut Teknologi Telkom Purwokerto
18102002@ittelkom-pwt.ac.id [1], anjarlaksana@ittelkom-pwt.ac.id [2]

Article Info

Article history:

Received 2022-06-23
Revised 2022-10-07
Accepted 2022-10-11

Keyword:

Cross Validation,
Diabetes,
K-Nearest Neighbors,
Optimization,
Streamlit.

ABSTRACT

The problem that occurs in the application of K-Nearest Neighbors as a classification algorithm is the frequent occurrence of overfitting in data processing. This can be overcome by using cross-validation techniques in evaluating the algorithm model and minimizing overfitting. Then the performance of diabetes prediction accuracy is unknown using the K-Nearest Neighbors algorithm with cross-validation technique. The data used comes from the National Institute of Digestive and Kidney Diabetes in 2021. The case study in this study is to find out the initial screening for diabetes is supported by the results of algorithm accuracy and real time application of streamlit-based users. The purpose of this study was to optimize the accuracy results with a cross validation technique supported by the k-nearest neighbors algorithm in the study of diabetes data. The method used is the k-nearest neighbors algorithm which is supported by cross validation technique for optimal accuracy results. Then the application of a streamlit-based interactive web application for testing the accuracy results used by the user to see the probability that the user has diabetes. The results showed that the optimization of the Cross Validation technique supported by the KNearest Neighbors algorithm model worked well. The results of the confusion matrix using the cross validation technique are more accurate in terms of the advantages of using the cross-validation technique itself. So that the classification report which has a value of 95% is more accurate than the accuracy which is worth 92% because of the use of cross-validation techniques that can minimize overfitting in addition to considerations of the accuracy value and the implementation of streamlit-based interactive web applications for user testing is going well.

I. PENDAHULUAN

Dalam perkembangan teknologi dengan semakin banyak sumber informasi yang sudah tersedia dari berbagai kalangan, hal ini memudahkan orang mencari informasi sehingga orang melakukan pengulasan tentang sebuah karya ilmiah maupun penelitian dengan baik. Tujuannya untuk lebih singkat dalam pembacaan artikel dan juga menambah pengetahuan kepada kita sehingga karya ilmiah banyak diketahui dan disukai orang[1]. Ketika mereka menghadapi permasalahan dan bagaimana solusi yang mereka gunakan dalam menyelesaikan permasalahan tersebut, sehingga solusi tersebut menjadi dasar pemikiran terciptanya ide-ide baru[2]. Walaupun demikian, perkembangan teknologi memiliki aspek positif maupun negatif. Salah satunya adalah aktifitas manusia itu sendiri dalam memanfaatkan teknologi dan informasi yang ada.

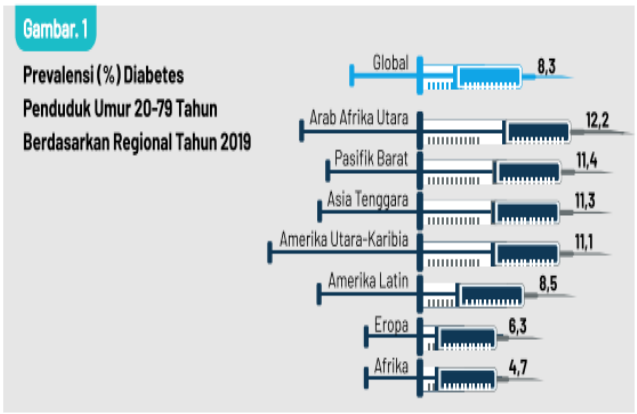
Informasi yang sudah tersedia semakin meningkat tiap tahunnya. Salah satunya adalah informasi mengenai penyakit seseorang. Setiap tahun bahkan setiap hari, ada pasien yang ingin mengecek dirinya mengalami atau periksa rutin penyakit yang dialami. Istilah dalam Kesehatan dalam mengecek seseorang dinamakan diagnosis penyakit. Diagnosis penyakit adalah fungsi perawatan kesehatan yang paling kritis. Jika suatu penyakit didiagnosis sebelum periode normal atau yang direncanakan, itu dapat menyelamatkan nyawa orang[3]. Industri *Healthcare* berisi data yang sangat besar dan sensitif dan perlu ditangani dengan sangat hati-hati. Ini adalah akar penyebab banyak penyakit kesehatan

terkait[4]. Misalnya efek samping dari suatu penyakit yang diderita pasien seperti nyeri saraf pada penderita diabetes.

Masalah kesehatan masyarakat yang utama, diabetes, adalah peningkatan kadar gula dalam darah yang berlebihan dan terjadi ketika pankreas tidak memproduksi insulin, atau meskipun pankreas memproduksi insulin, tubuh tidak dapat menggunakannya secara efektif[5]. Gula darah sangatlah vital bagi kesehatan karena merupakan sumber energi yang penting bagi sel-sel dan jaringan[6]. Gula darah yang tinggi ini berdampak pada berbagai organ tubuh manusia dan terkadang menimbulkan komplikasi pada banyak fungsi tubuh, khususnya pembuluh darah dan saraf[7].

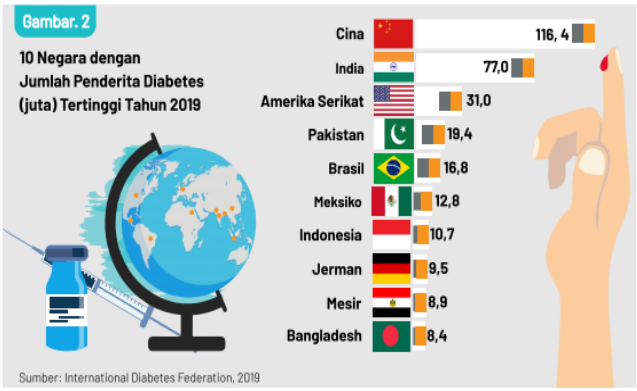
Diabetes menjadi penyakit mematikan jika tidak ditangani dengan baik. Berbagai penelitian mengenai Kesehatan atau healthcare ada banyak sekali dengan berbagai studi kasus yang berbeda sehingga data penelitian sangat diperlukan. Healthcare adalah industri penting yang menawarkan pertimbangan berbasis penghargaan kepada banyak orang[8]. Organisasi medis, di seluruh dunia, mengumpulkan data tentang berbagai masalah terkait kesehatan[9]. Salah satunya adalah data para penderita penyakit diabetes di dunia.

Diabetes melitus (DM) merupakan salah satu penyakit mematikan di dunia[10]. Statistik meningkatkan bahwa, penyakit diabetes adalah salah satu penyakit paling berbahaya yang menyebabkan penyakit berbahaya lainnya dan dapat menyebabkan kematian dalam beberapa kasus[11]. Masalah diabetes melitus di Indonesia sudah terjadi sejak tahun 1980 sampai sekarang. Jumlah penderita diabetes mellitus di Indonesia diperkirakan sekitar 10 juta orang dengan angka prevalensi 6,2% dan penyakit ini merupakan salah satu penyebab kematian di Indonesia[12]. Jumlah penderita diabetes mellitus di kalangan orang dewasa diperkirakan mencapai 463 juta (9,3% dari total populasi) di seluruh dunia pada 2019[13].



Gambar 1 Prevalensi (%) Diabetes Penduduk Umur 20-79 tahun di Tahun 2019[14].

Berdasarkan referensi dari International Diabetes Federation pada tahun 2019, terlihat Gambar 1 yang menunjukkan prevalensi diabetes dalam persentase penduduk umur 20 sampai 79 tahun[14]. Daerah yang memiliki prevalensi paling tinggi adalah Arab Afrika Utara dengan tingkat pravelensi 12,2 %.



Gambar 2 Negara dengan Jumlah Penderita Diabetes(juta) Tertinggi Tahun 2019[14].

Kemudian sepuluh negara dengan jumlah penderita diabetes tertinggi dalam tingkatan juta yang terlihat pada Gambar 2. Negara yang memiliki jumlah penderita diabetes tertinggi adalah Cina dan berjumlah 116,4 juta penderita.

Dengan melihat data penderita diabetes yang semakin meningkat tiap tahunnya di seluruh dunia, prevalensi diabetes diperkirakan akan terus meningkat secara eksplosif, dari 578,4 juta (10,2%) pada tahun 2030 menjadi 700,2 juta (10,9%) pada tahun 2045[13]. Sehingga penyakit diabetes ini menjadi rahasia umum masyarakat dan tidak sedikit yang mengalaminya, khususnya masyarakat dari negara Indonesia.

Dalam hal ini, ada aspek positif bahwa manusia mencoba mengatasi masalah Kesehatan diabetes agar tidak terkena ataupun kondisi penderita penyakit diabetes semakin parah. Diabetes adalah penyakit jangka panjang yang tidak dapat disembuhkan. Namun, gejala dan komplikasi selanjutnya dapat dikendalikan dengan perawatan yang diperlukan dan gaya hidup sehat. Dengan demikian, penelitian ini memiliki arti penting karena memungkinkan mendeteksi diabetes pada tahap awal[5].

Machine Learning (ML) kini menjadi semakin populer dan telah dilaporkan sebagai salah satu metode paling efektif dalam berbagai aplikasi dalam perawatan kesehatan preventif[15]. Ini memiliki keuntungan terkait seperti komputasi yang relatif murah, ketahanan, kemampuan generalisasi dan kinerja tinggi[5].

Dalam berkembangnya alat, perlengkapan dan peralatan medis, pengetahuan lanjutan dapat diperoleh di bidang diagnosis penyakit. Pengambilan keputusan dengan bantuan komputer, yaitu Machine Learning membantu manusia dengan memproses kumpulan data medis yang kompleks dan menganalisisnya untuk memberikan wawasan klinis[16], [17]. Ekstraksi pengetahuan dari data merupakan faktor penting untuk prediksi dan diagnosis penyakit dalam industri medis[18], [19].

Dalam hal ini, penelitian menggunakan model algoritma machine learning K-Nearest Neighbors untuk proses

perhitungan akurasi dari dataset yang sudah dipersiapkan sebelumnya. Dataset yang akan diuji, penulis ambil dari dokumen asli National Institute of Diabetes and Digestive and Kidney Diseases[20]. Model Machine Learning Supervised Learning yang digunakan adalah K nearest neighbors[21].

Sebelumnya, ada tambahan Teknik cross validation untuk optimasi algoritma K-nearest neighbors dengan mengevaluasi model algoritma secara optimal. Teknik ini akan mengoptimalkan hasil akurasi dan konsep yang penting dalam Ilmu data dan analisis data. Itupun digunakan untuk mencegah atau setidaknya meminimalkan overfitting. Overfitting berarti bahwa model yang disesuaikan terlalu banyak dengan data pelatihan.

Banyak penelitian medis modern didukung oleh statistik dan teknik pembelajaran mesin. Untuk keandalan studi ini, penting untuk mengukur efektivitas model yang dikembangkan dalam membuat prediksi pada data medis nyata. Validasi silang adalah alat, yang menyediakan metode berprinsip untuk mengukur efektivitas model dan membandingkan model satu sama lain. Hasilnya menunjukkan metode berulang (seperti pengambilan sampel dan bootstrap) tampaknya menghasilkan hasil yang lebih stabil[22].

Model yang dikembangkan pada akhirnya di-deploy agar layak digunakan secara real time. K-Nearest Neighbors (KNN) dan grid search cross validation (CV) telah digunakan untuk melatih dan mengoptimalkan model untuk memberikan hasil terbaik. Keuntungannya adalah akurasi dalam prediksi yang telah terlihat sebesar 80%[23]. Kemudian membuat interactive web application menggunakan streamlit. Streamlit adalah library python untuk membuat model yang dibangun menjadi berbasis web dan mudah digunakan[24].

Selain itu, ada penelitian terdahulu lainnya yang relevan terhadap penelitian ini. Penelitian dengan judul "Perbandingan Metode Klasifikasi *Supervised Learning* pada Data Bank Customers Menggunakan Python" dilakukan oleh F. Sodik, B. Dwi, and I. Kharisudin pada tahun 2020. Pada penelitian ini menggunakan dan membandingkan algoritma Regresi logistik, K-nearest neighboor, naive bayes, super vector machine, dan random forest. Prediksi dilakukan menggunakan data churn modelling dari Kaggle. Perbandingan 9:1 pada data training dan testing. Kemudian dilakukannya confusion matrix tiap model algoritma. Model Algoritma klasifikasi random forest lebih baik dari model algoritma lainnya yaitu nilai akurasi 86,2%, nilai precision 0,740, nilai recall 0,482, dan nilai f1 adalah 0,584[25]. Penelitian dengan judul "Predicting Diabetes Mellitus and Analysing Risk-Factors Correlation" dilakukan oleh M. F. Faruque, Asaduzzaman, S. M. M. Hossain, M. H. Furhad, and I. H. Sarker pada tahun 2020. Pada penelitian ini menggunakan algoritma Support Vector Machine Naive Bayes, K-Nearest Neighbors (KNN) dan C4.5 Decision Tree. Prediksi menggunakan dataset dari Medical Centre Chittagong, Bangladesh dengan 200 pasien diabetes mengggunakan 16 attributes data. Kemudian mencari model terbaik dengan korelasi tertinggi dengan beberapa variable atau attributes data yang ada. Model Decision Tree lebih baik dari model lainnya dengan hasil akurasi 73.5%, F-measure 72%, dan AUC 0.69[7]. Penelitian dengan judul "Predictive Supervised Machine Learning Models for Diabetes Mellitus" dilakukan oleh L. J. Muhammad, E. A. Algehyne, and S. S. Usman pada tahun 2020. Pada penelitian ini menggunakan dan membandingkan algoritma supervised learning yang terdiri dari Logistic regression, support vector machine, K-nearest neighbor, random forest, naive Bayes dan gradient booting. Prediksi menggunakan dataset diabetes type 2 dari rumah sakit Murtala Mohammed Specialist, kano. Kemudian penggunaan confusion matrix, dsb. Selanjutnya ditemukan sebagai model terbaik di antara model adalah random forest yang dikembangkan dengan akurasi 88,76%, sedangkan dalam hal kurva karakteristik pengoperasian penerima, random forest dan gradient boosting tampaknya menjadi model klasifikasi prediktif terbaik dengan 86,28%[10].

Berdasarkan permasalahan yang telah disebutkan sebelumnya, bahwa Algoritma K-Nearest Neighbors adalah salah satu algoritma terbaik dalam klasifikasi data dan banyak digunakan di berbagai penelitian. Penerapan K-Nearest Neighbors sebagai algoritma klasifikasi harus memperhatikan Teknik validasi silang dalam evaluasi model algoritma. Proses evaluasi data tanpa validasi silang bisa mengalami kesulitan dikarenakan memakan waktu komputasi yang lama pada jumlah attributes dataset yang banyak.

Selain itu, Teknik validasi silang digunakan untuk meminimalkan overfitting. Kemudian belum diketahuinya performasi akurasi mengenai prediksi penyakit diabetes menggunakan algoritma K-Nearest Neighbors dengan Teknik cross validation. Oleh karena itu, penulis akan optimasi algoritma K Nearest Neighbors dengan Teknik cross validation untuk prediksi penyakit. Selain itu, ada tambahan library streamlit dan basisnya adalah interactive website application.

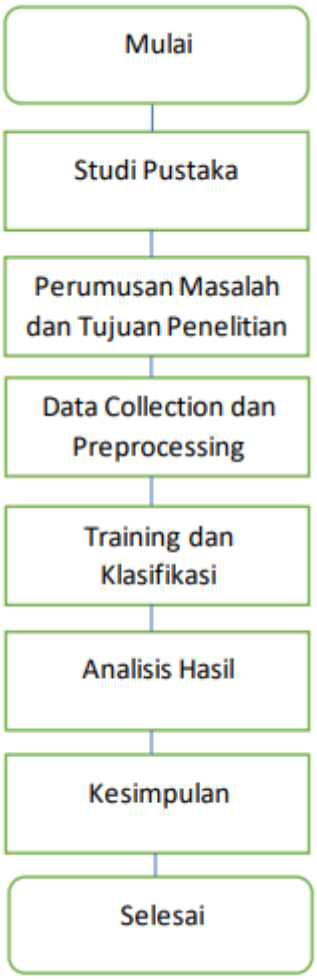
Penelitian ini akan melakukan optimalisasi proses komputasi yang lama dan meminimalkan overfitting dalam memberikan estimasi akurasi pada algoritma K-Nearest Neighbors dengan menggunakan Teknik cross validation pada prediksi penyakit diabetes. Hasil Confusion Matrix bisa berpengaruh dalam proses perhitungan akurasi model algoritma dengan dan tanpa menggunakan Teknik cross validation.

II. METODE PENELITIAN

A. Subyek dan Obyek Penelitian

Subyek penelitian merupakan data yang akan diamati. Subyek penelitian ini adalah dataset penyakit diabetes. Obyek penelitian ini merupakan permasalahan yang akan diamati. Obyek penelitian ini merupakan penerapan algoritma K-Nearest Neighbors dengan Teknik cross validation.

B. Diagram Alir Penelitian



Gambar 3. Tahapan-tahapan penelitian

1) Studi Pustaka

Penulis melakukan studi Pustaka. Pada tahap ini, penulis membaca dan memahami konsep dan permasalahan machine learning yang ada pada jurnal, buku maupun penelitian sebelumnya. Kemudian, hasil yang didapatkan menjadi landasan penulisan dan penelitian yang akan dilakukan. Pada penelitian ini berfokus pada data yang berkaitan dengan penyakit diabetes dalam berbagai sumber penelitian terkait. Dengan melakukan studi pustaka ini berharap penulis akan lebih menguasai topik-topik yang berada didalamnya.

2) Perumusan Masalah dan Tujuan

Perumusan masalah dilakukan untuk mengetahui permasalahan yang ada sehingga diperlukan penelitian ini dan penyusunan tujuan penelitian dilakukan untuk mengetahui tujuan dari penelitian ini. Dalam hal ini, penulis berfokus kepada permasalahan di bidang Kesehatan. Salah satunya penyakit diabetes dan jenis model algoritma machine learning yang digunakan untuk tujuan penelitian ini terhadap data penyakit tersebut.

3) Data Collection dan Preprocessing

Pada tahap ini, penulis melakukan data collection dan preprocessing. Langkah pertama dari tugas ilmu data adalah untuk mendapatkan, mengumpulkan, dan mengukur data yang diperlukan dan ditargetkan dari sumber data internal atau eksternal yang tersedia, dan kemudian dikompilasi ke dalam sistem yang mapan. Dalam hal ini, penulis mendapatkan dataset dari National Institute of Diabetes and Digestive and Kidney Diseases pada tahun 2021.

Kemudian melakukan prepocessing. Preprocessing adalah teknik penambangan data yang mengubah data mentah menjadi format yang dapat dipahami. Proses ini memiliki empat tahap utama yaitu data cleaning, data integration, data transformation, and data reduction. Pada tahap ini, akan menyaring, mendeteksi, dan menangani data kotor untuk memastikan kualitas data dan hasil analisis yang berkualitas. Dalam hal ini, mungkin ada noise dari nilai dan outlier yang tidak mungkin dan ekstrim, dan nilai yang hilang. Kesalahan mungkin termasuk data yang tidak konsisten dan atribut dan data yang berlebihan. Dan empat tahap utama dalam preprocessing ikut terlibat

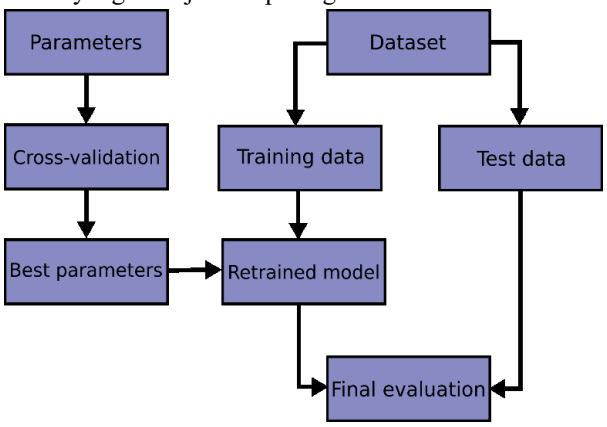
4) Training dan Klasifikasi

Pada tahap ini, dimulai dari eksplorasi analisis data, data modelling, dan evaluasi model algoritma yang diuji. Kemudian akan tampil hasil akurasi dari model algoritma yang digunakan untuk proses implementasi akurasi model algoritma ke input manual dari pengguna berbasis streamlit. Eksploraai analisis data bertujuan untuk melakukan penyelidikan awal pada data sebelum pemodelan formal dan representasi grafis dan visualisasi, untuk menemukan pola, melihat asumsi, dan menguji hipotesis.

Data modelling dimana dataset dibagi menjadi dua set terpisah yaitu set pelatihan dan set tes. Keduanya terdiri dari atribut yang sama, tapi nilai atributnya berbeda. Training set digunakan untuk melatih dan membangun model klasifikasi. Test set digunakan untuk memprediksi klasifikasi data baru yang tidak bias yang tidak digunakan untuk melatih model, sebelum mengevaluasi kinerja model berdasarkan metrik kinerja akurasi, presisi, recall, dan skor F1 dari klasifikasi tersebut.

Kemudian ada Teknik cross validation yang digunakan untuk mengevaluasi dan dan membandingkan algoritma pembelajaran dengan membagi data menjadi dua segmen: satu digunakan untuk mempelajari atau melatih model dan yang lainnya digunakan untuk memvalidasi model. Dalam tipikal cross validation, set pelatihan dan validasi harus saling silang dalam putaran yang berurutan sehingga setiap titik data

memiliki peluang untuk divalidasi. Biasanya Cross Validation K-fold digunakan karena dapat mengurangi waktu komputasi dengan tetap menjaga keakuratan estimasi. Berikut alur cross validation yang ditunjukkan pada gambar 2.2 ini:



Gambar 4 Alur cross validation

Merujuk pada Gambar 4, terlihat alur cross validation yang bermula dataset awal yang menggunakan train-test split untuk membagi data menjadi dua yaitu data pelatihan dan data uji. K-Fold Cross Validation ini digunakan untuk menyelesaikan permasalahan train-test split. Karena K-Fold Cross Validation akan memilih nilai k sebagai pembagian data. Misalnya nilai k=5 maka jumlah data dibagi 5 dan melakukan putaran training dan testing sebanyak 5 kali. Kemudian hasilnya pembagian sebagai data uji dan sisanya sebagai data pelatihan serta dilakukan sebanyak 5 kali. Tiap putaran akan menghasilkan akurasi dan banyaknya akurasi tersebut dilakukan mean atau rata-rata. Hasil mean tersebut adalah akurasi sebenarnya dari model algoritma yang digunakan menggunakan teknik cross validation.

Selanjutnya model evaluasi dilihat dari hasil confusion matrixnya yang digunakan sebagai pengevalusi kinerja metode klasifikasi. Sehingga hasil tiap model algoritma dengan dan tanpa menggunakan Teknik cross validation bisa dilakukan perbandingan model yang lebih baik. Confusion matrix juga bisa menghasilkan sebuah perhitungan nilai untuk hasil accuracy, precision, dan recall. Dalam hal ini, perhitungan model algoritma dengan dan tanpa menggunakan Teknik cross validation. Berikut contoh confusion matrix dan dilanjut contoh perhitungan akurasi, presisi, akurasi, dan f1 score dari data yang terlihat pada tabel I:

Tabel I
CONTOH CONFUSION MATRIX

		Nilai Sebenarnya	
		TRUE	FALSE
Nilai Prediksi	TRUE	90	20
	FALSE	10	880

$$\text{Presisi} = \frac{90}{90+20} = \frac{90}{110} = 0.82 = 82\%$$

$$\text{Recall} = \frac{90}{90+10} = \frac{90}{100} = 0.9 = 90\%$$

$$\text{Akurasi} = \frac{90+880}{90+880+20+10} = \frac{970}{1000} = 0.97 = 97\%$$

$$\text{F1 Score} = 2 \text{ x } \frac{0.82\ X\ 0.9}{0.82 + 0.9} = 2\ x\ \frac{0.738}{1.72} = 2\ X\ 0.43 = 0.86 = 86\%$$

5) Analisis Hasil

Pada tahap analisis hasil dilihat dari optimasi model algoritma k-nearest neighbors dengan Teknik cross validation. Percobaan menguji model algoritma k-nearest neighbors tanpa Teknik cross validation memiliki perbedaan hasil akurasi. Sehingga ada perbandingan hasil akurasi model algoritma k-nearest neighbors dengan dan tanpa menggunakan K-Fold Cross Validation. Hasil optimasi model algoritma k-nearest neighbors seperti penggunaan dan tanpa Teknik cross validation yang kemudian dianalisis melalui beberapa parameter nilai seperti akurasi, presisi, recall dan f1 score serta merujuk pada manfaat penggunaan Teknik cross validation itu sendiri pada suatu data.

6) Kesimpulan

Setelah melakukan proses pengolahan data sampai tahap evaluasi hasil, dilakukan tahap pengambilan kesimpulan. Kesimpulan didapatkan berdasarkan hasil dari proses pengolahan data, hasil, evaluasi data sampai analisis akurasi yang akurat yang melibatkan Teknik cross validation dalam menguji model algoritma k-nearest neighbors.

III. HASIL DAN PEMBAHASAN

A. Hasil Pengujian

Langkah pertama dari tugas ilmu data adalah untuk mendapatkan, mengumpulkan, dan mengukur data yang diperlukan dan ditargetkan dari sumber data internal atau eksternal yang tersedia, dan kemudian dikompilasi ke dalam sistem yang mapan. Dataset ini berasal dari Kaggle dan diperoleh dari Institut Nasional Diabetes dan Penyakit Pencernaan dan Ginjal. Dataset ini akan digunakan untuk memprediksi secara diagnostik apakah pasien menderita diabetes atau tidak, berdasarkan pengukuran diagnostik

tertentu yang disertakan dalam kumpulan data. Nantinya ada beberapa kolom yang dihapus karena variable tersebut tidak diperlukan dalam proses perhitungan akurasi.

patient_number	cholesterol	glucose	hdl_chol	chol_hdl_ratio	age	gender	height
1	193	77	49	3,9	19	female	61
2	146	79	41	3,6	19	female	60
3	217	75	54	4	20	female	67
4	226	97	70	3,2	20	female	64
5	164	91	67	2,4	20	female	70

weight	bmi	systolic_bp	diastolic_bp	waist	hip	waist_hip_ratio	diabetes
119	22,5	118	70	32	38	0,84	No diabetes
135	26,4	108	58	33	40	0,83	No diabetes
187	29,3	110	72	40	45	0,89	No diabetes
114	19,6	122	64	31	39	0,79	No diabetes
141	20,2	122	86	32	39	0,82	No diabetes

Gambar 5 Dataset Pelatihan

Data tersebut, akan mengalami proses *preprocessing*. Proses ini adalah teknik penambangan data yang mengubah data mentah menjadi format yang dapat dipahami. Proses ini memiliki empat tahap utama yaitu pembersihan data, integrasi data, transformasi data, dan reduksi data. Pembersihan data atau *data cleaning* akan menyaring, mendeteksi, dan menangani data kotor untuk memastikan kualitas data dan hasil analisis yang berkualitas. Kemudian integrasi data yang bertujuan agar data menjadi lebih halus dari hasil pembersihan data sebelumnya. Selanjutnya data transformasi adalah data yang akan dinormalisasi dan digeneralisasikan. Dan reduksi data atau pengurangan data yang dapat meningkatkan efisiensi penyimpanan dan mengurangi representasi data di gudang data. Selain itu, reduksi data adalah transformasi informasi digital numerik atau abjad yang diperoleh secara empiris atau eksperimental ke dalam bentuk yang dikoreksi, diurutkan, dan disederhanakan. Hasil dari tahap preprocessing ini dilakukan pengecekan untuk mencegah duplikasi data sebelum ke tahap selanjutnya.

```
#Checking for missing data
print('No missing data') if sum(df.isna().sum()) == 0 else df.isna().sum()
```

```
No missing data
```

Gambar 6. Hasil pengecekan terdapat missing data atau tidak.

height	weight
61	119
60	135
67	187
64	114
70	141

Gambar 7 Data awal dari dataset yang memiliki kesalahan penempatan kolom dengan value yang ada.

weight	height
61	119
60	135
67	187
64	114
70	141

Gambar 8 Hasil pengubahan kolom dengan value yang sesuai.

	cholesterol	glucose	hdl_chol	chol_hdl_ratio	age	weight	height
0	193	77	49	3,9	19	61	119
1	146	79	41	3,6	19	60	135
2	217	75	54	4	20	67	187
3	226	97	70	3,2	20	64	114
4	164	91	67	2,4	20	70	141

bmi	systolic_bp	diastolic_bp	waist	hip	waist_hip_ratio	diabetes
22,5	118	70	32	38	0,84	0
26,4	108	58	33	40	0,83	0
29,3	110	72	40	45	0,89	0
19,6	122	64	31	39	0,79	0
20,2	122	86	32	39	0,82	0

Gambar 9. Dataset yang telah dihapus beberapa kolom yang tidak diperlukan dalam proses perhitungan akurasi.

Selain itu, ada pengecekan value data seperti koma pemisah desimal yang tidak diketahui oleh parser csv di panda yang mengharuskan mengubah koma menjadi titik pemisah desimal agat diketahui oleh program saat dieksekusi dalam

proses perhitungan akurasi. Kemudian ada pengecekan duplikasi data pada dataset dan lanjut ke tahap selanjutnya.

	cholesterol	glucose	hdl_chol	chol_hdl_ratio	age	weight
0	193	77	49	3.9	19	61
1	146	79	41	3.6	19	60
2	217	75	54	4.0	20	67
3	226	97	70	3.2	20	64
4	164	91	67	2.4	20	70

height	bmi	systolic_bp	diastolic_bp	waist	hip	waist_hip_ratio	diabetes
119	22.5	118	70	32	38	0.84	0
135	26.4	108	58	33	40	0.83	0
187	29.3	110	72	40	45	0.89	0
114	19.6	122	64	31	39	0.79	0
141	20.2	122	86	32	39	0.82	0

Gambar 10. Dataset yang diubah dari koma ke titik pemisah decimal

```
# detect duplicated records
df[df.duplicated(subset = None, keep = False)]
```

cholesterol	glucose	hdl_chol	chol_hdl_ratio	age	weight

There are no duplications in the dataset.

Gambar 11. Hasil pengecekan duplikasi data

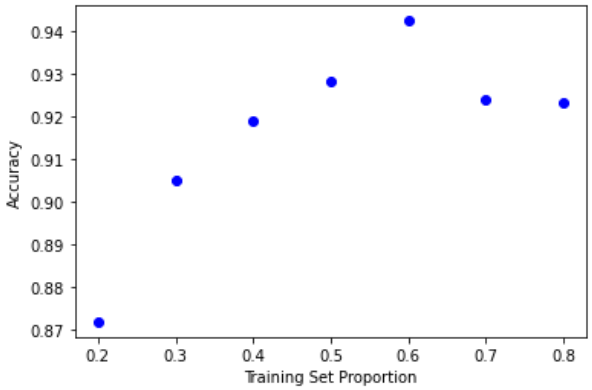
Selanjutnya melakukan eksplorasi analisis data yang bertujuan untuk melakukan penyelidikan awal pada data sebelum pemodelan formal dan representasi grafis dan visualisasi, untuk menemukan pola, melihat asumsi, dan menguji hipotesis. Ringkasan informasi tentang karakteristik utama dan tren tersembunyi dalam data dapat membantu dokter mengidentifikasi area dan masalah yang menjadi perhatian, dan penyelesaiannya dapat meningkatkan akurasi dalam mendiagnosis diabetes.

Dalam dataset terdapat 330 orang tidak terkena diabetes dan 60 orang terkena diabetes dengan value data tiap kolom yang beragam. Ada histogram tiap kolom dan korelasi heatmap menggunakan korelasi matrix. Dan penggunaan code describe() untuk mendapatkan ringkasan statistic.

	cholesterol	glucose	hdl_chol	chol_hdl_ratio	age	weight	height
count	390	390	390	390	390	390	390
mean	207.2308	107.3385	50.2667	4.5246	46.7744	65.9513	177.4077
std	44.6660	53.7982	17.2791	1.7366	16.4359	3.9189	40.4078
min	78	48	12	1.5000	19	52	99
25%	179	81	38	3.2000	34	63	150.2500
50%	203	90	46	4.2000	44.5000	66	173
75%	229	107.7500	59	5.4000	60	69	200
max	443	385	120	19.3000	92	76	325

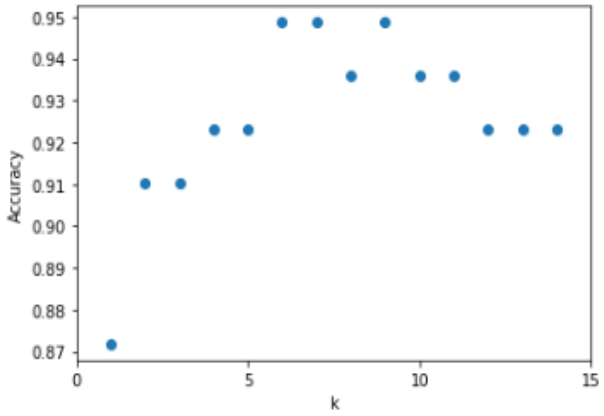
Gambar 12. Ringkasan statistic dari dataset beberapa kolom

Kemudian tahap data modelling dimana dataset dibagi menjadi dua set terpisah yaitu set pelatihan dan set tes. Keduanya terdiri dari atribut yang sama, tetapi nilai atributnya tidak sama. Training set digunakan untuk melatih dan membangun model klasifikasi. Test set digunakan untuk memprediksi klasifikasi data baru yang tidak bias yang tidak digunakan untuk melatih model algoritma, sebelum mengevaluasi kinerja model algoritma berdasarkan metrik kinerja akurasi, presisi, recall, dan skor F1 dari klasifikasi tersebut.



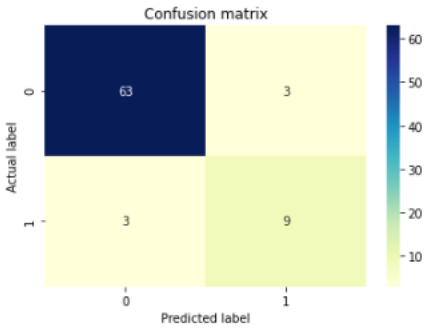
Gambar 13. Training Set Proportion

Terlihat hasil dari Training Set Proportion memiliki akurasi yang beragam dari 0.87 sampai 0.94. Training Set Proportion dapat berpengaruh pada pemilihan proporsi training set yang optimal untuk hasil akurasi yang akurat. Pada train test split, penulis menggunakan 20% untuk test data dan 80% untuk train data untuk perolehan tingkat akurasi yang optimal.



Gambar 14. Penentuan nilai K dari angka 0 sampai 15.

Subset pelatihan membutuhkan 312 instances yang terdiri dari non diabetes yang berjumlah 264 orang dan diabetes berjumlah 48 orang. Pemilihan k paling optimal menghasilkan akurasi 94 % dan tampilan confusion matrix sebelum penggunaan Teknik K-Fold Cross Validation



Gambar 15. Confusion Matrix sebelum penggunaan K-Fold Cross Validation

Kemudian model evaluasi yang memiliki subset test yang berjumlah 78 instances terdiri dari non diabetes yang berjumlah 66 orang dan diabetes berjumlah 12 orang. Terlihat classification report dari model algoritma KNN tanpa K-Fold Cross Validation pada gambar 16.

	precision	recall	f1-score	support
0	0.95	0.95	0.95	66
1	0.75	0.75	0.75	12
accuracy			0.92	78
macro avg	0.85	0.85	0.85	78
weighted avg	0.92	0.92	0.92	78

Gambar 16. Classification Report KNN sebelum ada K-Fold Cross Validation

Selanjutnya penggunaan K-Fold Cross Validation pada model algoritma KNN. Terlihat Confusion matrix dan hasil Classification Report dari model algoritma KNN dengan K-Fold Cross Validation pada gambar 17.

```
[[65  1]
 [ 3  9]]
```

Confusion Matrix setelah penggunaan K-Fold Cross Validation

-	precision	recall	f1-score	support
positive	0.96	0.98	0.97	66
negative	0.90	0.75	0.82	12
accuracy			0.95	78
macro avg	0.93	0.87	0.89	78
weighted avg	0.95	0.95	0.95	78

Gambar 17. Classification Report KNN setelah ada K-Fold Cross Validation

Setelah itu, export file dump dari model algoritma K-Nearest Neighbors dengan Teknik Cross Validation yang telah diproses sebelumnya sampai ada hasil akurasinya. Dan implementasi rangkaian tahapan yang telah dilakukan ke dalam streamlit beserta penambahan fitur prediksi diabetes.

B. Analisis

Pada bagian analisis, dilakukan untuk menganalisis hasil optimasi model algoritma K-Nearest Neighbors dengan Teknik Cross Validation Prediksi penyakit diabetes dengan streamlit berdasarkan hasil berbagai tahapan yang telah dilakukan dengan Teknik cross validation dan hasil akurasi melalui perhitungan confusion matrix dari proses optimasi. Analisis ini juga mencakup pembahasan terkait hasil pengujian sebelum penggunaan k-fold cross validation yang telah dilakukan.

Hasil dari penerapan Teknik cross validation untuk kebutuhan optimasi model algoritma K-Nearest Neighbors prediksi penyakit diabetes berjalan dengan baik. Dimulai dari normalisasi dataset dengan melakukan pemilihan kolom yang perlu di hapus untuk keperluan proses perhitunagan akurasi. Selain itu, ada pengubahan koma pemisah desimal menjadi titik dan nama kolom agar sesuai value pada data yang ada seperti pada gambar 7 sampai gambar 10

Merujuk pada gambar 11, terlihat tidak ada missing data dan duplikasi data pada dataset. Kolom hasil diabetes yang semula memiliki value text diubah menjadi value integer yang bernilai 1 dan 0 saja untuk tiap hasil pada dataset. Angka 1 merujuk pada value diabetes dan angka 0 merujuk pada value non diabetes. Ada ringkasan statistic yang mecakup ukuran tendensi sentral seperti mean, median dan ukuran dispersi seperti standar deviasi, yang berguna dalam memberikan deskripsi dataset dan karakteristiknya dengan cepat dan sederhana.

Hasil dari pengujian ini akan menghasilkan sebuah nilai akurasi, presisi, recall dan f1-score dari model algoritma K-Nearest Neighbors tanpa dan menggunakan Teknik cross validation. Hasil pengujian ini ditampilkan dengan confusion matrix yang berfungsi untuk mengevaluasi hasil kinerja model algoritma yang dibangun berdasarkan hasil actual dan prediksi serta classification report untuk mengukur kualitas hasil prediksi yang dibangun dari algoritma klasifikasi.

Merujuk pada gambar 14 dan 15, ada penentuan nilai k untuk mengoptimalkan hasil akurasi. Kemudian ada confusion matrix yang menggunakan dan tanpa ada Teknik cross validation. Bisa dilihat confusion matrix memiliki hasil yang beragam dan stabil atau sama pada prediksi label. Hal tersebut juga terjadi pada classification report yang memiliki value data yang berbeda. Sehingga model algoritma K-Nearest Neighbors yang menggunakan Teknik cross validation memiliki perbedaan akurasi dari pengujian sebelumnya.

1) Perbandingan pengujian penggunaan dan tanpa Teknik cross validation

Sebagai sebuah perbandingan, terlihat dari gambar dan penjelasan yang sudah dipaparkan pada poin sebelumnya. Bahwa hasil akurasi dari proses perhitungan yang dilakukan menghasilkan hasil yang berbeda. Confusion matrix 2 pengujian tersebut menghasilkan data yang berbeda dan classification report yang berbeda pula seperti pada gambar 16 dan 17. Itupun membuktikan bahwa penggunaan Teknik

cross validation menghasilkan tingkat akurasi yang berbeda dan berusaha meminimalkan overfitting yang ada.

Confusion Matrix
Before K-Fold
[[63 3]
[3 9]]
After K-Fold
[[65 1]
[3 9]]

CLASSIFICATION REPORT
BEFORE K-FOLD

	precision	recall	f1-score	support
0	0.95	0.95	0.95	66
1	0.75	0.75	0.75	12
accuracy			0.92	78
macro avg	0.85	0.85	0.85	78
weighted avg	0.92	0.92	0.92	78

AFTER K-FOLD

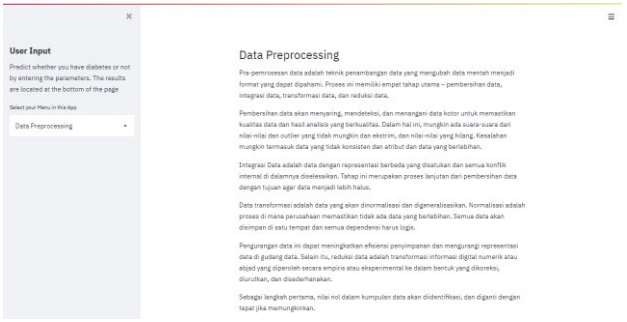
-	precision	recall	f1-score	support
positive	0.96	0.98	0.97	66
negative	0.90	0.75	0.82	12
accuracy			0.95	78
macro avg	0.93	0.87	0.89	78
weighted avg	0.95	0.95	0.95	78

Gambar 18. Perbandingan Confusion Matrix dan Perbandingan Classification Report

Dari hasil pengujian yang didapatkan, bisa dilihat bahwa hasil akurasi terbaik didapatkan oleh dataset dengan Teknik cross validation. Hasil akurasi yang didapatkan sebesar 92 %, dimana dataset tanpa menggunakan Teknik cross validation mendapatkan hasil sebesar 94%. Bisa disimpulkan, meskipun hasil akurasi lebih sedikit dari pengujian tanpa penggunaan Teknik cross validation, tapi salah satu manfaat Teknik cross validation adalah bisa meminimalkan overfitting. Sehingga hasil akurasi lebih baik daripada akurasi sebelum penggunaan Teknik cross validation.

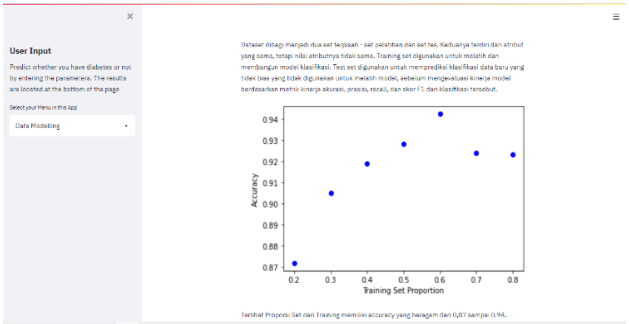
Setelah itu, kita implementasikan hasil akurasi ke dalam streamlit berbasis interactive website application. Dalam implementasinya, kita buat beberapa menu yang menggambarkan proses pengolahan dataset sampai tahap proses perhitungan akurasi. Kemudian implementasi pengguna mengisi data dari beberapa attribute yang menentukan prosentase pengguna mengalami diabetes berdasarkan data yang di inputkan pengguna dari attribute yang berguna untuk proses perhitungan yaitu 14 attributes dataset yang digunakan untuk proses pengolahan data sampai hasil akurasi optimal menggunakan Teknik cross validation. Berikut tampilan menu-menu dari streamlit berbasis interactive website application.

Gambar 19. Tampilan Home

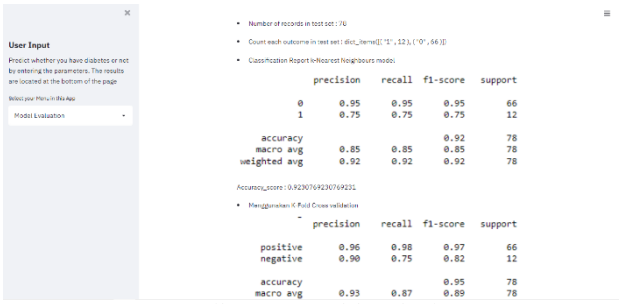


Tampilan Menu Data Preprocessing

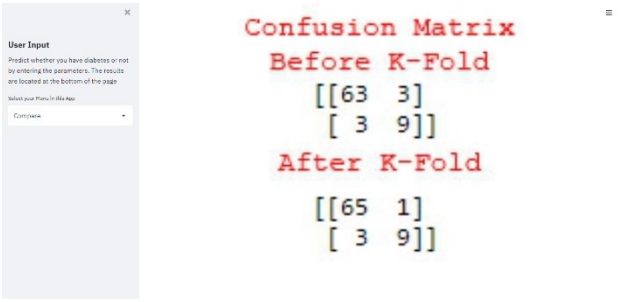
Tampilan Menu Exploratory Data Analysis (EDA)



Gambar 20. Tampilan Menu Data Modelling



Tampilan Menu Model Evaluation



Tampilan Menu Compare

Gambar 21. Tampilan K Nearest Neighbors

Merujuk pada gambar 19 sampai gambar 21, terlihat implementasi berbagai menu interactive web application basis streamlit. Pada gambar 19 terlihat menu home yang berisi informasi umum data yang digunakan dan kegunaan programnya. Pada gambar 20 sampai gambar 21 terlihat menu-menu untuk informasi pengolahan datanya dari preprocessing sampai evaluasi model algoritma yang digunakan. Selanjutnya analisis hasil akurasi model algoritma yang terlihat pada gambar 21 yang berisi perbandingan akurasi dengan dan tanpa Teknik cross validation. Pada gambar 21 terlihat menu K-Nearest Neighbors yang berisi pengujian input user dari attribute data yang ada dengan hasil akurasi optimal menggunakan Teknik cross validation dan terlihat porsentase user terkena diabetes dari data yang telah diinputkan.

V. KESIMPULAN

Berdasarkan hasil dan analisis pada penjelasan diatas, terdapat beberapa kesimpulan yang diperoleh dari penelitian tugas akhir ini, yaitu sebagai berikut:

Pertama, penelitian yang telah dilakukan untuk optimasi hasil akurasi yang optimal menggunakan model algoritma K-Nearest Neighbors dengan Teknik cross validation prediksi penyakit diabetes dengan streamlit berjalan dengan baik. Proses optimasi ini dilakukan secara berurutan dari tahap preprocessing sampai evaluasi model algoritma membutuhkan normalisasi data untuk kebutuhan proses perhitungan akurasi model algoritma K-Nearest Neighbors prediksi penyakit diabetes ini. Sehingga tingkat akurasi bisa lebih akurat tanpa adanya reduksi data dan penggunaan Teknik cross validation untuk meminimalkan overfitting.

Kedua, hasil confusion matrix yang menggunakan Teknik cross validation lebih akurat dilihat dari manfaat penggunaan Teknik cross validation itu sendiri. Sehingga classification report yang memiliki nilai sebesar 95% lebih akurat daripada akurasi yang bernilai 92% karena ada penggunaan Teknik cross validation yang bisa meminimalkan overfitting selain pertimbangan dari value akurasinya.

DAFTAR PUSTAKA

[1] F. D. Telaumbanua, P. Hulu, T. Z. Nadeak, R. R. Lumbantong, and A. Dharma, "Penggunaan Machine Learning Di Bidang Kesehatan," *J. Teknol. dan Ilmu Komput. Prima*, vol. 2, no. 2, pp. 57–64, 2020, doi: 10.34012/jutikomp.v2i2.657.

[2] B. Triandi, "Keamanan Informasi secara Aksiologi Dalam Menghadapi Era Revolusi Industri 4.0," *Jurikom*, vol. 6, no. 5, pp. 477–483, 2019, [Online]. Available: http://ejurnal.stmik-budidarma.ac.id/index.php/jurikom%7CPage477.

[3] M. Diwakar, A. Tripathi, K. Joshi, M. Memoria, P. Singh, and N. Kumar, "Latest trends on heart disease prediction using machine learning and image fusion," *Mater. Today Proc.*, vol. 37, no. Part 2, pp. 3213–3218, 2020, doi: 10.1016/j.matpr.2020.09.078.

[4] Y. Jeevan Nagendra Kumar, N. Kameswari Shalini, P. K. Abhilash, K. Sandeep, and D. Indira, "Prediction of diabetes using machine learning," *Int. J. Innov. Technol. Explor. Eng.*, vol. 8, no. 7, pp. 2547–2551, 2019, doi: 10.35940/ijrte.e6290.018520.

[5] B. Pranto, S. M. Mehnaz, E. B. Mahid, I. M. Sadman, A. Rahman, and S. Momen, "Evaluating machine learning methods for predicting diabetes among female patients in Bangladesh," *Inf.*, vol. 11, no. 8, 2020, doi: 10.3390/INFO11080374.

[6] A. Maulida, "Penerapan Metode Klasifikasi K-Nearest Neigbor pada Dataset Penderita Penyakit Diabetes," *Indones. J. Data Sci.*, vol. 1, no. 2, pp. 29–33, 2020.

[7] M. F. Faruque, Asaduzzaman, S. M. M. Hossain, M. H. Furhad, and I. H. Sarker, "Predicting diabetes mellitus and analysing risk-factors correlation," *EAI Endorsed Trans. Pervasive Heal. Technol.*, vol. 5, no. 20, pp. 1–15, 2020, doi: 10.4108/eai.13-7-2018.164173.

[8] M. Atif, J. Siddiqui, F. Talib, S. S. Sohail, J. Hamdard, and N. Delhi, "Applications of Machine Learning Techniques for Disease Diagnosis : a review," vol. 7, no. 17, pp. 2652–2661, 2020.

[9] V. V. Ramalingam, A. Dandapath, and M. Karthik Raja, "Heart disease prediction using machine learning techniques: A survey," *Int. J. Eng. Technol.*, vol. 7, no. 2.8 Special Issue 8, pp. 684–687, 2018, doi: 10.14419/ijet.v7i2.8.10557.

[10] L. J. Muhammad, E. A. Algehyne, and S. S. Usman, "Predictive Supervised Machine Learning Models for Diabetes Mellitus," *SN Comput. Sci.*, vol. 1, no. 5, pp. 1–10, 2020, doi: 10.1007/s42979-020-00250-8.

[11] H. Torkey, E. Ibrahim, E. E.-D. Hemdan, A. El-Sayed, and M. A. Shouman, "Diabetes classification application with efficient missing and outliers data handling algorithms," *Complex Intell. Syst.*, no. 0123456789, 2021, doi: 10.1007/s40747-021-00349-2.

[12] P. S. Nugroho, N. A. Tianingrum, S. Sunarti, A. Rachman, D. S. Fahrurodzi, and R. Amiruddin, "Predictor risk of diabetes mellitus in Indonesia, based on national health survey," *Malaysian J. Med. Heal. Sci.*, vol. 16, no. 1, pp. 126–130, 2020.

[13] S. Kohsaka, N. Morita, S. Okami, Y. Kidani, and T. Yajima,

"Current trends in diabetes mellitus database research in Japan," *Diabetes, Obes. Metab.*, vol. 23, no. S2, pp. 3–18, 2021, doi: 10.1111/dom.14325.

[14] Kemenkes, "Infodatin tetap produktif, cegah, dan atasi Diabetes Melitus 2020," *Pusat Data dan Informasi Kementrian Kesehatan RI*. pp. 1–10, 2020, [Online]. Available: https://pusdatin.kemkes.go.id/resources/download/pusdatin/infodatin/Infodatin-2020-Diabetes-Melitus.pdf.

[15] M. S. Amin, Y. K. Chiam, and K. D. Varathan, "Identification of significant features and data mining techniques in predicting heart disease," *Telemat. Informatics*, vol. 36, pp. 82–93, 2019, doi: 10.1016/j.tele.2018.11.007.

[16] S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, "Comparing different supervised machine learning algorithms for disease prediction," *BMC Med. Inform. Decis. Mak.*, vol. 19, no. 1, pp. 1–16, 2019, doi: 10.1186/s12911-019-1004-8.

[17] D. Dahiwade, G. Patle, and E. Meshram, "Designing disease prediction model using machine learning approach," *Proc. 3rd Int. Conf. Comput. Methodol. Commun. ICCMC 2019*, no. Iccmc, pp. 1211–1215, 2019, doi: 10.1109/ICCMC.2019.8819782.

[18] K. M. F. Fuhad, J. F. Tuba, M. R. A. Sarker, S. Momen, N. Mohammed, and T. Rahman, "Detection from Blood Smear and Its Smartphone Based Application," *Diagnostics*, vol. 10, no. 329, 2020.

[19] J. H. Joloudari *et al.*, "Coronary artery disease diagnosis; ranking the significant features using a random trees model," *Int. J. Environ. Res. Public Health*, vol. 17, no. 3, 2020, doi: 10.3390/ijerph17030731.

[20] U. S. Department of Health and Human Services, "National Institute of Diabetes and Digestive and Kidney Diseases." https://www.niddk.nih.gov/ (accessed Jul. 13, 2021).

[21] A. D. Kumari, J. P. Kumar, V. S. Prakash, and K. S. Divya, "Supervised Learning Algorithms : A Comparison," vol. 1, no. 1, pp. 1–12, 2020.

[22] M. Rafało, "Cross validation methods: Analysis based on diagnostics of thyroid cancer metastasis," *ICT Express*, no. xxxx, 2021, doi: 10.1016/j.icte.2021.05.001.

[23] G. S. K. Ranjan, A. Kumar Verma, and S. Radhika, "K-Nearest Neighbors and Grid Search CV Based Real Time Fault Monitoring System for Industries," *2019 IEEE 5th Int. Conf. Converg. Technol. I2CT 2019*, pp. 9–13, 2019, doi: 10.1109/I2CT45611.2019.9033691.

[24] A. Saxena, M. Dhadwal, and M. Kowsigan, "Indian Crop Production : Prediction And Model Deployment Using Ml And Streamlit," vol. 32, no. 3, pp. 1874–1886, 2020.

[25] S. Raschka, "Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning," 2018, [Online]. Available: http://arxiv.org/abs/1811.12808.