PROSIDING SEMINAR NASIONAL & CALL FOR PAPER
ISSN Online: 2654-6590 | ISSN Cetak: 2654-5306
Website: https://prosiding.stie-aas.ac.id/index.php/prosenas

Pemanfaatan Data Mining Untuk Prediksi Resiko Penyakit *Acquired Immunodeficiency Syndrome* (AIDS) Dengan Metode Naïve Bayes

Muhammad Ausathul Fikri[1], Siti Rokhmah[2]
Institut Teknologi Bisnsi AAS Indonesia
lfikri882@gmail.com, Sitrokhmah.itbaas@gmail.com

Abstrak
HIV (Human Immunodeficiency Virus) adalah virus yang menyerang sistem kekebalan tubuh dan dapat berkembang menjadi AIDS (*Acquired Immunodeficiency Syndrome*), tahap lanjut dari infeksi HIV. Penyebaran HIV dapat terjadi melalui berbagai cara, termasuk kontak seksual. Pada penelitian ini, dilakukan analisis prediksi risiko HIV dengan menggunakan metode Naïve Bayes. Faktor-faktor yang dianalisis meliputi umur, status pernikahan, pendidikan, dan kebiasaan seksual. Data yang digunakan adalah data training sejumlah 100 data dan data testing sejumlah 50 data. Proses prediksi menggunakan Naïve Bayes dengan mencari probabilitas dan menghasilkan hasil prediksi. Pengujian hasil menggunakan Confusion Matrix untuk mengukur akurasi prediksi. Hasil penelitian ini memberikan gambaran tentang prediksi risiko HIV berdasarkan faktor-faktor yang dianalisis.

Abstract
HIV (Human Immunodeficiency Virus) is a virus that attacks the immune system, particularly CD4 cells, which are essential for fighting off infections and diseases. If left untreated, HIV can progress to AIDS (Acquired Immunodeficiency Syndrome), the advanced stage of HIV infection. HIV can be transmitted through various means, including sexual contact. In this study, the prediction of HIV risk is analyzed using the Naïve Bayes method. The factors analyzed include age, marital status, education, and sexual habits. The data used consists of 100 randomly selected training data and 50 testing data. The prediction process employs Naïve Bayes by calculating probabilities and generating prediction results. The accuracy of the predictions is evaluated using the Confusion Matrix. The results provide insights into the prediction of HIV risk based on the analyzed factors.

Pendahuluan

HIV (Human Immunodeficiency Virus) adalah virus yang menyerang sistem kekebalan tubuh, khususnya sel CD4, yang sangat penting untuk melawan banyak infeksi dan penyakit. HIV (Human Immunodeficiency Virus) merupakan salah satu tantangan kesehatan masyarakat yang terus menjadi perhatian global. Berdasarkan data dari kemeterian Republik Indonesia dilaporkan kasus HIV pada tahun 2022 mencapai 1.929 kasus, dengan penderita di kalangan remaja anatra umur 15-24 tahun (Hasibuan et al., 2024). Penyakit ini tidak hanya berdampak pada kesehatan individu tetapi juga memiliki konsekuensi sosial dan ekonomi yang signifikan(Magnolini et al., 2021). HIV menyerang sistem kekebalan pada tubuh dari waktu ke waktu, membuat hal itu lebih sulit bagi tubuh untuk untuk melawan banyak infeksi dan penyakit. Jika HIV tidak diobati, maka HIV dapat berkembang menjadi AIDS (Acquired Immunodeficiency Syndrome) yang mana hal itu adalah tahap paling lanjut dari infeksi HIV("1993 Revised Classification System for HIV Infection and Expanded Surveillance Case Definition for AIDS among Adolescents and Adults.," 1992). Ada beberapa cara untuk HIV bisa menular, dianataranya adalah melalui kontak seksual, jarum suntik, dan lainnya. Upaya untuk menekan angka penyebaran HIV

memerlukan pendekatan yang berbasis bukti dan pemanfaatan teknologi untuk mendukung keputusan yang lebih akurat dan efisien. Salah satu teknologi yang dapat dimanfaatkan adalah data mining, yang memungkinkan analisis data dalam skala besar untuk menemukan pola dan hubungan yang tidak terlihat secara langsung(Suhaimi et al., 2019),(Aresta & Jumaiyah, 2019).

Penyakit AIDS ditandai dengan menurunnya sistem kekebalan tubuh. Beberapa faktor yang mempengaruhi penyakit AIDS diantaranya adalah umur, status pernikahan, tingkat pendidikan, dan kebiasaan seksual diketahui memiliki pengaruh signifikan terhadap tingkat kerentanan seseorang terhadap HIV (Amelia M et al., 2016). Jumlah yang kasus AIDS yang semakin banyak memerlukan antisipasi sejak disni., salah satunya adalah dengan mengembangkan sistem prediksi. Dengan menganalisis data-data faktor resiko penyebab HIV, model prediktif dapat dibangun untuk mengidentifikasi individu atau kelompok yang berisiko tinggi, sehingga intervensi yang lebih tepat dapat dilakukan (Dodu et al., 2018). Data mining telah dikembangkan ke berbagai sekotor, termasuk kesehatan. Dalam konteks HIV, teknik data mining dapat digunakan untuk memprediksi risiko infeksi berdasarkan berbagai faktor risiko (Prakarsya & Prambayaun, 2020).

Algoritma *Naïve Bayes*, salah satu metode dalam data mining, menawarkan keunggulan dalam klasifikasi data dan prediksi. Metode ini sederhana namun sangat efektif dalam menangani dataset yang besar dan kompleks. Keunggulan *Naïve Bayes* terletak pada kemampuannya untuk memprediksi probabilitas suatu kejadian berdasarkan variabel yang saling independen. Dalam konteks penelitian ini, algoritma tersebut digunakan untuk memodelkan hubungan antara risiko HIV dengan faktor-faktor seperti umur, status pernikahan, pendidikan, dan kebiasaan seksual.algoritma naïve bayes merupakan salah satu metode yang efektif dalam mengklasifikasikan dan memprediksi peluang kejadian di masa depan berdasarkan data-data yang telah terjadi (Rokhmah & Rozaq Rais, 2022). Naïve bayes telah berhasil membantu dalam sistem klasifikasi data berbasis probabilitas bersayarat (Liliana et al., 2021).

Penelitian ini bertujuan untuk memanfaatkan data mining, khususnya algoritma *Naïve Bayes*, dalam memprediksi risiko HIV berdasarkan analisis data demografis dan perilaku. Dengan demikian, hasil penelitian ini diharapkan dapat memberikan kontribusi nyata dalam upaya pencegahan HIV, baik melalui peningkatan pemahaman terhadap faktor risiko maupun sebagai alat bantu dalam perencanaan program intervensi kesehatan yang lebih efektif dan efisien. Pendekatan ini tidak hanya berfokus pada aspek teknis pengembangan model prediktif tetapi juga mendukung implementasi solusi berbasis data yang dapat memperkuat strategi pencegahan dan penanganan HIV secara menyeluruh.

Kajian Teori

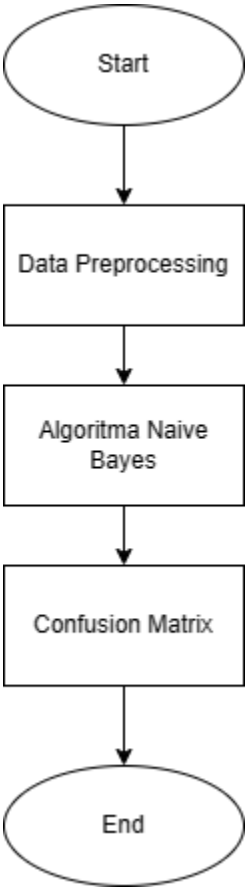
Beberapa penelitian yang melatarbelakangi penelitian ini diantaranya adalah penelitian yang berjudul "Penerapan Data Mining Untuk Memprediksi Pertumbuhan Jumlah Penderita Human Immunodeficiency Virus (HIV)Menggunakan Metode Multiple Linier Regression (Studi Kasus Dinas Kesehatan Provinsi Sumatera Utara)". Pada penelitian ini dikembangkan sistem prediksi dengan metode multiple linier. Penelitian ini dikembangkan ke dalam aplikasi berbasis desktop. Dari hasil penelitian tersebut diharapka dapat dijadikan acuan dalam menyediakan penanganan-penanganan terkait kasus HIV.

Penelitian lain adalah penelitian dengan judul

Metode Penelitian

Kerangka penelitian pada penelitian ini dapat dilihat pada gambar 1.

ISSN Online: 2654-6590 | ISSN Cetak: 2654-5306
Website: https://prosiding.stie-aas.ac.id/index.php/prosenas



Gambar 1. Kerangka penelitian

1. Studi pustaka

Pada tahapan ini dilakukan studi pustaka dari berbagai sumber, yaitu dari buku, jurnal dan media online lainnya yang membahas tentang penyakit AIDS, perkembangan virus HIV, data mining dan naïve bayes. Dari hasil studi pustaka tersebut dijadikan dasar dalam melaksanakan penelitian ini

2. Pengumpulan data

Pengumpulan data dari penelitian ini diperoleh dari data bersama dari kaggle.com. data yang dikumpulkan adalah data terkait gejala HIV dan data faktor resiko. data yang terkumpul kemudian dianalisis dan dilakukan pra pemrosesan dengan membuang data-data derau dan pencilan, sehingga menjadi data olah.

3. Pengolahan data

Data olah yang telah siap kemudian diolah dengan menggunakan algoritma naïve bayes. Algoritma *Naïve bayes* merupakan salah satu metode yang digunakan dalam klasifikasi data mining. Teknik untuk klasifikasi ini ditemukan oleh Thomas Bayes yang juga dikenal dengan Teorema Bayes(Nurmayanti, 2021). Pada Naïve bayes dilakukan penghitungan pada sekumpulan probabilitas dengan menjumlahkan beberapa frekuensi dan kombinasi nilai dari kumpulan data yang diberikan. Algoritma ini mengasumsikan

bahwa semua atribut dari setiap kelas tidak bergantung satu sama lain(Devita et al., 2018). *Naïve bayes* merupakan *supervised classification* yang mana dibutuhkan beberapa data training untuk melakukan proses klasifikasi yang berupa suatu nilai probabilitas.

$P(A \mid B) = P(B \mid A) \times P(A) / P(B)$

A = Hipotesis

B = Aktual

$P(A \mid B)$ = Probabilitas bahwa hipotesis(A) benar untuk aktual(B)

$P(B \mid A)$ = Probabilitas aktual(B) benar untuk hipotesis(A)

$P(A)$ = Probabilitas prior hipotesis(A)

$P(B)$ = Probabilitas prior aktual(B)

4. Pengujian

Hasil prediksi yang diperoleh dengan algoritma naïve bayes kemudian diuji dengan menggunakan model pengujian confussion matrix. Confussion matrix merupakan pengukuran kinerja untuk klasifikasi. Confussion matrix adalah alat evaluasi kinerja yang penting dalam klasifikasi data mining, memberikan gambaran menyeluruh tentang hasil prediksi model. Tabe confussion matrix dapat dilihat pada tabel 1.

Tabel 1. Confusion Matrix

Class		Aktual	
		Yes	No
Prediksi	Yes	True Positive	False Negative
	No	False Positive	True Negative

True Positive adalah kondisi dimana keadaan actual dan prediksi benar, *False Positive* adalah kondisi dimana keadaan actual salah dan prediksi benar, *False Negative* adalah kondisi dimana keadaan actual benar dan prediksi salah, *True Negative* adalah kondisi dimana keadaan actual dan prediksi salah. Pengujian ini dapat menghasilkan perhitungan dengan 5 output, yaitu:

$Precision = TP/\ TP+FP\ x\ 100\%$

$Recall = TP/\ TP+FN\ x\ 100\%$

$Akurasi = TP+TN/\ TP+TN+FP\_FN\ x\ 100\%$

$Error = FP+FN/\ TP+TN+FP+FN\ x\ 100\%$

$F\text{-}Measurement = 2\ x\ Precision\ x\ Recall\ /\ Precision+Recall$

Variable penelitian

Pada penelitian kali ini variable yang digunakan adalah umur, status pernikahan, latar belakang Pendidikan, tempat bertemu dengan patner seks, dan orientasi seksual.

Hasil dan Pembahasan

1.1. *Dataset*

Dataset yang digunakan pada penelitian ini memiliki 699 data dan dipilih sebanyak 100 data dengan acak sebagai data training dan juga mengambil senjumlah 50 data acak yang digunakan sebagai data testing.

3.2 *Algoritma Naïve bayes*

Mencari nilai probabilitas menggunakaan model *naïve baye*s dengan membagi jumlah value dalam suatu atribut dan membagi jumlah value tersebut dengan jumlah Positive/Negative.

Tabel 2. Mencari nilai probabilitas

	Aktual		Probabilitas	
Age	**positive**	**negative**	**positive**	**negative**
Adults	29	26	0,604167	0,5
Teenagers	13	13	0,270833	0,25
Children	1	1	0,020833	0,019231
Elderly	5	12	0,104167	0,230769
	100			
Marital Status	**positive**	**negative**	**positive**	**negative**
Married	12	15	0,25	0,288462
Unmarried	29	24	0,604167	0,461538
Widowed	1	6	0,020833	0,115385
Cohabiting	1	3	0,020833	0,057692
Divorced	5	4	0,104167	0,076923
	100			
Educational	**positive**	**negative**	**positive**	**negative**
College Degree	16	19	0,333333	0,365385
Junior High School	14	13	0,291667	0,25
Illiteracy	7	9	0,145833	0,173077
Primary School	2	5	0,041667	0,096154
Senior High School	9	6	0,1875	0,115385
	100			
Places of seeking sex partners	**positive**	**negative**	**positive**	**negative**
Bar	3	20	0,0625	0,384615
Internet	26	5	0,541667	0,096154
Park	4	9	0,083333	0,173077
Public Bath	11	7	0,229167	0,134615

Others	4	11	0,083333	0,211538
	100			
Sexual orietntation	**positive**	**negative**	**positive**	**negative**
Bisexual	11	14	0,229167	0,269231
Heterosexual	23	33	0,479167	0,634615
Homosexual	14	5	0,291667	0,096154
	100			
Result	0,48	0,52		
	1			

Setelah probabilitas ditemukan maka pemrosesan data testing dapat dilakukan dan hasil dari data testing sebagai berikut.

Tabel 3. Prediksi Menggunakan Naïve bayes

positive	negative	P Ya	P Tidak	Hasil Prediksi
0,0000240373	0,0003032982	0,073433226	0,926566774	NEGATIVE
0,0001682611	0,0002847633	0,371417312	0,628582688	NEGATIVE
0,0021476237	0,0000800097	0,964083106	0,035916894	POSITIVE
0,0038179977	0,0002533640	0,937769232	0,062230768	POSITIVE
0,0002261692	0,0002534706	0,471539602	0,528460398	NEGATIVE
0,0002944570	0,0001240765	0,703544541	0,296455459	POSITIVE
0,0000804157	0,0014715379	0,05181579	0,94818421	NEGATIVE
0,0000336522	0,0004271450	0,073030443	0,926969557	NEGATIVE
0,0003181665	0,0001013456	0,75842033	0,24157967	POSITIVE
0,0005468486	0,0002481531	0,687858435	0,312141565	POSITIVE
0,0002915069	0,0015489873	0,158385136	0,841614864	NEGATIVE
0,0066317824	0,0012673532	0,839557989	0,160442011	POSITIVE
0,0000447591	0,0001861148	0,193868213	0,806131787	NEGATIVE
0,0007652057	0,0050694128	0,13114922	0,86885078	NEGATIVE
0,0002261692	0,0008579006	0,208629706	0,791370294	NEGATIVE
0,0000292460	0,0002599699	0,10112173	0,89887827	NEGATIVE
0,0013418824	0,0007527310	0,640634879	0,359365121	POSITIVE
0,0000803064	0,0001280155	0,385492044	0,614507956	NEGATIVE
0,0006959892	0,0000320039	0,956038221	0,043961779	POSITIVE
0,0006219652	0,0001971438	0,75931918	0,24068082	POSITIVE
0,0000214151	0,0001056128	0,168585522	0,831414478	NEGATIVE
0,0011153308	0,0020431270	0,353125129	0,646874871	NEGATIVE
0,0000545926	0,0014623306	0,035989005	0,964010995	NEGATIVE
0,0003659679	0,0021506600	0,145419958	0,854580042	NEGATIVE
0,0000060658	0,0003217127	0,018505906	0,981494094	NEGATIVE
0,0001442238	0,0007567137	0,160081927	0,839918073	NEGATIVE
0,0000384597	0,0006242888	0,058030587	0,941969413	NEGATIVE
0,0014700996	0,0003899548	0,790352994	0,209647006	POSITIVE

0,0012499397	0,0001891784	0,868545591	0,131454409	POSITIVE
0,0003854709	0,0006934172	0,357285364	0,642714636	NEGATIVE
0,0000421971	0,0000561299	0,429151117	0,570848883	NEGATIVE
0,0000804157	0,0012039855	0,062609486	0,937390514	NEGATIVE
0,0012420730	0,0004992604	0,713288462	0,286711538	POSITIVE
0,0000804157	0,0012039855	0,062609486	0,937390514	NEGATIVE
0,0010814525	0,0007717856	0,583547529	0,416452471	POSITIVE
0,0014195120	0,0008918411	0,614147598	0,385852402	POSITIVE
0,0002944570	0,0000448054	0,867932843	0,132067157	POSITIVE
0,0000017331	0,0001787293	0,009603646	0,990396354	NEGATIVE
0,0000432671	0,0007467569	0,054766867	0,945233133	NEGATIVE
0,0023320554	0,0048159422	0,32625296	0,67374704	NEGATIVE
0,0001160421	0,0000436566	0,72663173	0,27336827	POSITIVE
0,0001057641	0,0003972747	0,210250432	0,789749568	NEGATIVE
0,0001442238	0,0007567137	0,160081927	0,839918073	NEGATIVE
0,0000214471	0,0001364842	0,135800061	0,864199939	NEGATIVE
0,0039036579	0,0005675353	0,873068495	0,126931505	POSITIVE
0,0039036579	0,0005675353	0,873068495	0,126931505	POSITIVE
0,0002596029	0,0004928596	0,345004413	0,654995587	NEGATIVE
0,0000099465	0,0001550957	0,060266243	0,939733757	NEGATIVE
0,0014195120	0,0007296882	0,66048383	0,33951617	POSITIVE
0,0007237413	0,0066888086	0,097637294	0,902362706	NEGATIVE

3.3 Confusion Matrix

Pengujian yang dilakukan menggunakan data sebanyak 50, dan didapatkan perhitungan *Accuracy, Error Rate, Precision, Recall, dan F-Measurement* yang dapat dilihat pada tabel 4.

Tabel 4. Hasil pengujian Confusion Matrix

Accuracy	0,84
error rate	0,16
precision	0,32
recall	0,838709677
F1	0,46325167

Tabel 4.

Class		Aktual	
		Yes	No
Prediksi	Yes	16	3
	No	5	26

Berdasarkan pengujian yang telah dilakukan, Akurasi yang didapat cukup baik, dengan nilai sebesar 84% dan Error Rate sebesar 16%. Pada tabel 4 ditampilkan hasil prediksi dengan hasil True Positif yaitu 16 dan True Negatif 26, dengan kesalahan 8 dari 50 data testing.

Kesimpulan

Penelitian ini menunjukkan bahwa pemanfaatan data mining, khususnya algoritma *Naïve Bayes*, dapat menjadi alat yang efektif dalam memprediksi risiko HIV berdasarkan variabel seperti umur, status pernikahan, tingkat pendidikan, dan kebiasaan seksual. Hasil analisis mengindikasikan bahwa faktor-faktor tersebut memiliki hubungan yang signifikan terhadap risiko infeksi HIV, dengan masing-masing variabel memberikan kontribusi berbeda terhadap tingkat kerentanan individu. Model prediksi berbasis *Naïve Bayes* yang dikembangkan dalam penelitian ini berhasil menunjukkan kinerja yang baik dalam mengklasifikasikan risiko HIV pada dataset yang dianalisis. Keunggulan algoritma ini, seperti kemampuannya untuk menangani dataset dengan banyak variabel independen, menjadikannya alat yang andal untuk digunakan dalam analisis risiko kesehatan masyarakat. Hasil dari penerapan metode klasifikasi *Naïve Bayes* dan melakukan pengujian menggunakan *Confusion Matrix* dalam dataset anemia, dari 100 data training dan 50 data testing didapat 84% akurasi sehingga dapat dikatakan pengklasifikasian berjalan dengan baik

Referensi

1993 revised classification system for HIV infection and expanded surveillance case definition for AIDS among adolescents and adults. (1992). *MMWR. Recommendations and Reports : Morbidity and Mortality Weekly Report. Recommendations and Reports*, *41*(RR-17), 1–19.

Amelia M, Hadisaputro S, Laksono B, Anies, & Sofro MA. (2016). Faktor Risiko yang Berpengaruh Terhadap Kejadian HIV/AIDSpada Laki-Laki Umur 25 - 44 Tahun di Kota Dili, Timor Leste. *Jurnal Epidemiologi Kesehatan Komunitas*, *I*(1), 39–46.

Aresta, A. S., & Jumaiyah, W. (2019). Pengetahuan Dan Dukungan Keluarga Dengan Kepatuhan Dalam Menjalankan Pengobatan Antiretroviral (ARV) Pada Pasien HIV/AIDS. *Indonesian Journal of Nursing Practices*, *2*(1), 51–61.

Devita, R. N., Herwanto, H. W., & Wibawa, A. P. (2018). Perbandingan Kinerja Metode Naive Bayes dan K-Nearest Neighbor untuk Klasifikasi Artikel Berbahasa indonesia. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, *5*(4), 427. https://doi.org/10.25126/jtiik.201854773

Dodu, A. Y. E., Nugraha, D. W., Informasi, J. T., Teknik, F., & Tadulako, U. (2018). *11900-37937-1-Pb*. *1*(1), 33–49.

Hasibuan, A., Maulana, M. F. Z., & Mauliah, S. (2024). Melonjaknya Kasus HIV Dikalangan Remaja Indonesia. *Amsir Community Service Journal*, *2*(1), 1–8. https://doi.org/10.62861/acsj.v2i1.392

Liliana, D. Y., Maulana, H., & Setiawan, A. (2021). Data Mining untuk Prediksi Status Pasien Covid-19 dengan Pengklasifikasi Naïve Bayes. *Multinetics*, *7*(1), 48–53. https://doi.org/10.32722/multinetics.v7i1.3786

Magnolini, R., Senkoro, E., Vanobberghen, F., & Weisser, M. (2021). "Linkage to care" among people living with HIV - definition in the era of "universal test and treat" in a sub-Sahara African setting. *Swiss Medical Weekly*, *151*(December 2020), w20535.

https://doi.org/10.4414/smw.2021.20535

Nurmayanti, W. P. (2021). Penerapan Naive Bayes dalam Mengklasifikasikan Masyarakat Miskin di Desa Lepak. *Geodika: Jurnal Kajian Ilmu Dan Pendidikan Geografi*, *5*(1), 123–132. https://doi.org/10.29408/geodika.v5i1.3430

Prakarsya, A., & Prambayaun, A. (2020). Implementasi Data Mining Untuk Prediksi Penyebaran Virus Hiv / Aids Di Bandar Lampung Dengan Tenkik Decision Tree. *Jurnal Sistem …*, *3*(2), 18–26. https://www.ejournal.lembahdempo.ac.id/index.php/SISKOMTI/article/view/120

Rokhmah, S., & Rozaq Rais, N. A. (2022). Application of Data Mining for Prediction of Long Covid on Covid-19 Survival With Feature Selection and Naïve Bayes Method. *Jurnal Teknik Informatika (Jutif)*, *3*(5), 1397–1405. https://doi.org/10.20884/1.jutif.2022.3.5.561

Suhaimi, D., Savira, M., & Krisnadi, S. R. (2019). Pencegahan dan penatalaksanaan infeksi hiv / aids pada kehamilan prevention and management of hiv infection (aids) in pregnancy. *Majalah Kedokteran Bandung*, *41*(2), 1–7.