Jurnal Data Mining.docx

by Wyatt Alcala

Submission date: 22-Jul-2025 08:46PM (UTC+0200)
Submission ID: 2719098103
File name: Jurnal_Data_Mining.docx (289.73K)
Word count: 3059
Character count: 20275

Analisis Kompleksitas Password Dengan Metode KNN, Naïve Bayes, Random Forest, SVM, Linear Regression

1st Melco Lauwento
Sistem Informasi
Universitas Komputer Indoensia
Bandung, Indonesia
melco05lauwento.1@gmail.com

2nd Daffa Yudis Pratama
Sistem Informasi
Universitas Komputer Indoensia
Bandung, Indonesia
melco05lauwento.1@gmail.com

3th Muhammad Naufal Sandu
Sistem Informasi
Universitas Komputer Indoensia
Bandung, Indonesia
melco05lauwento.1@gmail.com

***Abstrak*— Keamanan kata sandi telah menjadi askpek kritis dalam sistem proteksi Informasi digital, dengan penilaian kekuatan kata sandi yang bergantung pada kompleksitasnya. Penelitian ini bertujuan untuk menganalisis kompleksitas kata sandi Menggunakan berbagai algoritma machine learning, termasuk K-Nearest Neighbor (KNN), Naïve Bayes, Decision tree,Linear Regression, dan Support Vector Machine (SVM), Dengan menggunakan dataset dari Kaggle yang berisi berbagai kata sandi dengan tingkat kekuatan yang berbeda, model-model ini diuji untuk memprediksi tingkat keamanan kata sandi. Untuk mengilustrasikan data dengan dapa dimengerti, salah satu alat bantu yang dapa diunakan adalah Goggle Colabs menggunakan Bahasa Python. Dalam proses yang dilakukan adalah menganalisi bagaimana prediksi antara password dan tingkat kesulitannya menggunakan lima Machine Learning. Analisis data mining ini melakukan dengan penerapan Teknik klasifikasi dengan memanfaatkan lima metode algortima yang berbeda.**

Kata Kunci —Machine Learning, Password Strength, K-Nearest Neighbor, Naïve Bayess, Random Forest, Linear Regression, Support Vector Machine.

I. PENDAHULUAN

Keamanan Kata sandi menjadi aspek krusial dalam sistem perlindungan data di dunia digital saat ini. Di banyak sistem keamanan, kata sandi berfungsi sebagai lapisan pertahanan pertama yang melindungi Informasi sensitif dari akses yang tidak sah. Namun, meskipun keberadaanya sangat penting, bantak pengguna yang memilih kata sandi yang lemah dan mudah ditebak, seperti menggunakan kata yang ada dalam kamu atau pola keyboard yang sederhana. Keadaan ini menciptakan kerentanannya terhadap serangan seperti brute-force atau dictionary attack. Oleh karena itu, penting untuk memiliki sistem yang dapat scara efisien menilai kekuatan kata sandi, untuk memastikan bahwa pengguna membuat kata sandi yang cukup kuat guna melindungi data pribadi mereka dari potensi ancaman. Beberapa alat pengukur kekuatan kata sandi berbasis statistic dan model Meachine Learning dikembangkan, namun mereka masih terbatas dalam hal penerapan di berbagai Bahasa dan konteks, khususnya dalam menganalisis kata sandi yang lebih kompleks atau berbassis Bahasa tertentu.

Penelitian ini bertujuan untuk mengesskplorasi penerapan Machine Learning dalam menilai kekuatan kata sandi dengan menggunakan berbagai model, seperti K-Nearest Neighbors(KKN), Naïve Bayes, Random Forest, Linear Regression, dan Support Vector Machine(SVM). Motivasi utama dibalik penelitian ini adalah untuk meningkatkan akurasi dan efisien dalam mengklasifikasikan tingkat kekuatan kata sandi yang dapat digunakan dalam berbagai sistem keamanan. Melalui pendekatan ini, kami mencoba untuk menggali potensi teknik Machine Learning yang lebih canggih dalam menganalisis dan memberikan umpan balik yang lebih akurat terkait dengan kekuatan kata sandi. Dengan model-model ini, diharapkan dapat diperoleh prediksi yang lebih presisi dan dapat diadaptasi untuk berbagai dataset dan konteks, bauk untuk kata sandi berbahasa inggris maupun Bahasa lainnya.

Lingkup penelitian ini mencakup penggunaan lima algortima Machine Learning yang berbeda untuk menilai kekuatan kata sandi K-Nearest Neighbors(KKN), Naïve Bayes, Random Forest, Linear Regression, dan Support Vector Machine(SVM). Setiap model ini diuji menggunakan dataset yang berisi kata sandi dengan label kekuatan yang telah ditentukan, berdasarkan fitur-fitur seperti Panjang kata sandi, jumlah symbol, dan keberadaan huruf kapital serta angka. Kontribusi utama dari penelitian ini adalah perbandingan kinerja dari masing-masing model dalam hal akurasi prediksi dan waktu komputasi,serta penyesuaian model-model tersebut untuk meningkatkan kemampuan dalam mengklasifikasikan kata sandi yang lebih kompleks, seperti yang ditemukan dalam data dunia nyata, Inovasi lain yang dihasilkan adalah pemanfaatan berbagai ukuran kesamaan teks dalam meningkatkan akurasi model, serta penerapan ensemble methods seperti Random Forest yang terbukti memberikan hasil yang lebih baik dibandingkan dengan model yang lebih sederhana. Penelitian ini juga memberikan wawasan terkait dengan adaptasi model Machine Learning untuk analisis kekuatan kata sandi dalam berbagai konteks dan Bahasa.

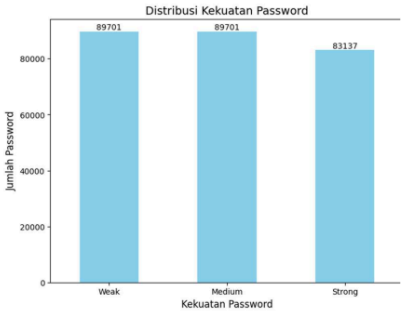
II. METODOLOGI

A. *Pengumpulan dan Preprocessing Data*

Dataset yang digunakan dalam penelitian ini diperoleh dari platfrom Kaggle, yang menyediakan data mengenai kekuatan password yang mencakup kategori-kategori seperti 0 = week, 1 = Medium, 2 = Strong. Data set ini lebih dari 600.000 pasword yang sudah diberi label kekuatannya, yang memungkinkan kami untuk melatih model Machine Learning dan memprediksi tingkat kekuatan password. Sumber dataset ini di ambil dari kumpulan data yang telah dipublikasikan oleh peneliti sebelumnya, dengan beberapa variasi kata sandi dari berbagai konteks, baik internasional maupun lokal, yang memastikan keberagaman data.

Langkah pertama dalam pengolahan data adalah membersihkan dataset untuk memastikan bahwa data yang

digunakan dapat dianfalkan. Proses pembersihan mencakup penghapusan baris-baris yang berisi nilai yang hilang atau duplikat, serta pemeriksaan lebih lanjut terhadap format data. Data yang memiliki nilai kosong pada kolom yang penting, seperti kolom password, dihapus untuk menghindari gangguna pada pelatihan model. Setelah data dibersihkan, langka selanjutnya adalah mengonversi semua password menjadi format yang dapat dianalisis, misalkan dengan menghitung Panjang password, jumlah symbol Khusus, huruf besar, dan angka dalam password.

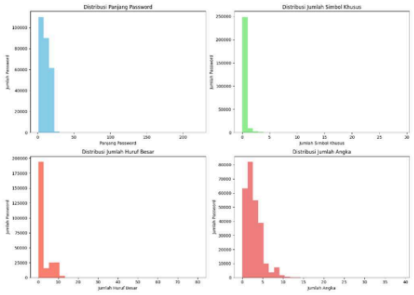


Gambar 1. Distribusi Kekuatan Password

B. Fitur yang Diekstrak

Dalam analisis ini beberapa fitur penting diekstrak dari setiap password untuk membangun represenetasi numerik yang akan digunakan dalam model Meaching Learning. Fitur yang diekstrak antara lain :

1. Panjang Password : Fitur ini dihitung dengan mengukur jumlah karakter dalam setiap password. Panjang password merupakan faktor penting menentukan kekuatannya, dimana password yang lebih Panjang umumnya lebih kuat dibandingkan yang lebih pendek.
2. Jumlah Simbol Khusus : Menghitung jumlah non-alfanumerik (seperti : @, #, $, dll) dalam password. Password dengan lebih banyak symbol Khusus biasanya dianggap lebih kuat.
3. Jumlah Huruf Besar : Fitur ini mengukur berapa banyak huruf kapital terdapata dalam password.keberadaan huruf besar seringkali meningkatkan komplesitas password.
4. Jumlah Anggka : Menghitung jumlah angka dalam password. Password yang mengandung angka cenderung lebih sulit ditebak daripada yang hanya terdiri huruf.



Gambar 2. Distribusi Fitur Password

C. Data yang hilang dan Ketidakseimbangan

Salah satu tantangan utama dalam pengolahan dataset ini adalah ketidakseimbangan data, dimana jumlah password yang di kategorikan sebagai 2 = Strong dan 0 = Week jauh lebih sedikit dibandingkan dengan password yang dikategorikan sebagai 1 = Medium.

Kategori	Persentasi
0	13%
1	74%
2	12%

Table 1. Persentasi Data Sebelum Under Sampling

Untuk mengatasi masalah ketidakseimbangan ini, metode undersampling diterapkan. Undersampling adalah Teknik dimana jumlah data pada kelas yang lebih dominan dikurangi sehingga jumlahnya lebih seimbang dengan kategori Medium. Dengan cara ini model tidak akan lebih focus pada kategori yang lebih sering muncul, tetapi akan lebih seimbang dalam mempelajari ketiga kategori. Proses ini membantu mencegah model terdistorsi oleh dominasi kategori yang lebih besar, yang pada akhirnya meningkatkan akurasi dan generalisasi model pada semua kategori.

Kategori	Persentasi
0	34%
1	34%
2	31%

Table 2. Persentasi Data Setelah Under Sampling

Selain itu, Langkah Langkah lain dalam pembersihan data juga diterapkan untuk menangani data yang hilang pada beberapa fitur penting. Baris dengan nilai yang hilang pada kolom yang relevan, seperti password dan tingkat kekuatannya, dihapus untuk menjaga integritas data yang digunakan dalam pelatihan model. Hal ini memastikan bahwa model hanya dilatih dengan data yang valid, meningkatkan efektivitas prediksi kekuatan password.

III. MODEL MACHINE LEARNING

A. *K-Nearest Neighbors (KNN)*

Model K-Nearest Neighbors (KNN) adalah salaah satu algoritma pembelajran mesin yang digunakan untuk klasifikasi berdasarkan kedekatan antara data yang ada. Dalam penelitian ini, KKN diterapkan untuk mengklafikasikan kekuatan password berdasarkan fitur-fitur yang diekstrak, seperti Panjang password,jumlah symbol Khusus, jumlah huruf kapital, dan angka.

Cara kerja KNN adalah dengan menghitung jarak antara password yang ingin diprediksi dengan data latih (training data) menggunakan metrik jarak seperti Euclidean atau Manhattan. Password yang berada paling dekat dengan password target akan mempengaruhi keputusan klasifikasi. Dalam konteks ini, model KKN menggunakan K(jumlah tetangga terdekat) untuk menentukan kategori kekuatan password. Model ini dipilih karena kesederhanaanya dan kemampuan untuk bekerja baik dengan data yang tidak memerlukan asumsi distribusi tertentu.

B. *NAÏVE BAYES*

Naïve Bayes adalah algoritma klasifikasi probalistrik yang beroperasi berdasarkan teorema Bayes, yang menghitung probabilitas suatu password berada pada kategori tertentu berdasarkan fitur fitur yang ada. Dalam penelitian ini, Naïve Bates digunakan untuk memperkirakan kekuatan password dengan mengasumsikan bahwa fitur-fitur yang ada dalam password bersifat independent satu sama lain.

Keunggulan dari Naïve Bayes adalahj kemampuannya dalam menangani data yang memiliki banyak fitur dan sangat efisien dalam hal komputasi. Meskipun sederhana, Model ini sangat efektif untuk maslaah klasifikasi seperti pengenalan teks dan analisis pola sederhana, yang cocok untuk mengklasifikasikan kata sandi berdasarkan fitur-fitur numerik dan kategorikal.

C. *RANDOM FOREST*

Random Forest adalah metode ensemble learning yang mengaabungkan banyak pohon keputusan (decision tree) untuk meningkatkan akurasi prediksi. Setiap pohon dalam random forest dilatih pada subset acak dari data pelatihan dan memprediksi kategori berdasarkan keputusan mayoritas dari seluruh pohon yang ada. Metode ini mengurangi masalah overfitting yang sering terjadi pada pohon keputusan tunggu dengan mengurangi variansi dari model.

Dalam penelitian ini, Random Forest diterapkan untuk mengklasifikasikan kekuatan password dengan memanfaatkan keputusan dari berbagai pohon keputusan untuk memberikan hasil yang lebih stabil dan akurat. Random Forest dipilih karena kemampuannya dalam menangani data yang kompleks dan adanya variabilitas antar pohon keputusan yang memungkinkan model mengahsilkan prediksi yang lebih kuat dan mengurangi kesalahan akibat outlier atau data yang tidak seimbang.

D. *LINEAR REGRESSION*

Meskipun Linear Regression umumnya digunakan untuk masalah predeksi numerik, dalam penelitian ini, Linear Regression diterapkan pada masalah klasifikasi dengan pendekatan yang sedikit berbeda. Biasanya, Linear Reggresion memodelkan hubungkan antara variable independent dan dependen dalam bentuk persamaan linier. Namun, untuk masalah klasifikasi seperti ini, pendekatannya melibatkan mengubah kategori kekuatan password menjadi nilai numerik dan memperlakukannya sebagai variable kontinu.

Dalam konteks ini, Linear Regression digunakna untuk memprediksi nilai yang berkaitan dengan kekuatan password berdasarkan fitur yang tersedia, seperti Panjang dan jumlah symbol pada password. Meskipun model ini lebih umum digunakan untuk regresi, penerapannya pada masalah klasifikasi dalam penelitian ini bertujuan untuk memberikan gambaran umum mengenai hubungan linear antara fitur dan kategori kekuatan password.

E. *SUPPORT VECTOR MACHINE (SVM)*

Support Vector Machine adalah algoritma klasifikasi yang efektif untuk menangani data yang memiliki banyak fitur, seperti pada maslaah pengklasifikasian kekuatan password. SVM beroperasi dengan mencari hyperplane yang memisahkan data dalam ruang fitur sehingga kelas yang berbeda terspisah dengan margin yang paling besar. Dalam penelitian ini, kernel liner digunakan pada model SVM karena efisiensinya dalam menangani dataset dengan dimensi yang relatinf tinggi, seperti pada password yang memiliki berbagai fitur numerik dan kategorikal.

Penggunaan kernel liner pada SVM dipilih karena cukup efektifd untuk data yang dapat dipisahkan secara linier (misalnya data password dengan berbagai karakteristik yang jelas terkalsifikasi). Dalam klasifikasi ini, SVM digunakan untuk menentukan kategori kekuatan password dengan memanfaatkan margin pemisahan yang optimal antar kelas, yaitu weak, medium, strong, berdasarkan fitur yang telah diekstraksi dari setiap password. Keunggulan SVM terletak pada kemampuannya untuk memaksimalkan margin antara kelas dan mengahsilkan prediksi yang akurat meskipun data yang tidak selalu linear.

IV. HASIL DAN PEMBAHASAN

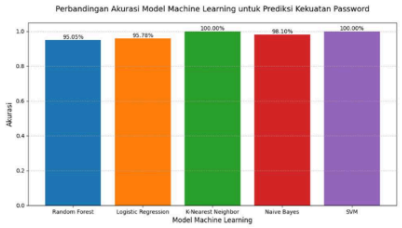
A. *Evaluasi Eksperimental*

Penelitian ini mengevaluasi lima algoritma Machine Learning untuk memprediksi kekuatan password, Logistic Regression, Support Vector Machine, K-Nearest Neighbors, Naïve Bayes, dan Random Forest. Metrik evaluasi yang di gunakan meliputu accuracy, precision, recall, F1 Score, dan analisis confusion matrix. Dataset dibagu menjadi tiga kelas kekuatan Week, Medium, dan Strong.

Hasil akurasi dari seluruh model dapat dilihat pada *Gambar 1,* menunjukan bahwa KKN dan SVM mencapai akurasi tertinggi (100%), disusul oleh Logistic Regression (95,78%), Naïve Bayes (98,10%), dan Random Forest (95,05%).

Performa masing-masing model dievaluasi lebih lanjut menggunakan confusion matrix untuk melihat distribusi klasifikasi benar dan salah pada tiap kelas untuk melihat distribusi klasifikasi benar dan salah pada tiap kelas. Model Support Vectro Machine dan K-Neares Neighbors menunjukan klasifikasi sempurna,tanpa adanya false positive maupun fals negatice di ketida kelas kekuatan password. Hal ini mencermikan kemampuan kedua model dalam memetakan fitur penting seperti Panjang password, jumlah

symbol, serta karakter kapital ke dalam kelas yang sesuai konsisten.



Gambar 3. Akurasi Machine Learning

B. Analisi Confusion Matrix dan Perfoma Model

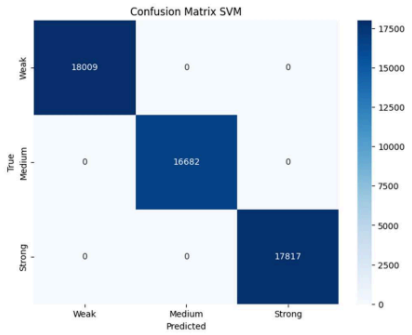
Evaluasi mendalam terhadap confusion matrix dari masing-masing model Machine Learning memberikan wawasan penting mengenai pola klasifikasi, jenis kesalahan yang terjadi (false positive dan false negative), serta kestabilan perfoma antar kelas prediksi.

1. Support Vector Machine (SVM)

Model SVM menunjukan perfoma tertinggi di antara seluruh algoritma yang di uji. Berdasarkan confusion matrix, SVM berhasil mengklasifikasikan seluruh data pada kelas weak, medium, dan strong secara sempurna, tanpa kesalahan klasifikasi. Hal ini mencerminkan nilai precision, recall, dan f1-score yang mencapai 1.00 pada keseluruhan,mengindikasikan ketepatan dan kepekaan model yang sangat optimal.

Keunggulan ini dapat dijelaskan oleh prinsip kerja SVM yang memaksimalkan margin antar kelas dalam fitur svm yang memaksimalkan margin antar kelas dalam ruang fitur berdimensi tinggi. Dengan memanfaatkan kernel trick, SVM mampu membentuk hypeplane optimal yang memisahkan data antar kelas dengan sangat efektif. Hal ini sangat berguna dalam domain seperti klasifikasi password, dimana fitur-fitur seperti Panjang, penggunaan huruf kapital, dan symbol menciptakan pemisahaan yang cukup jelas antar kelas kekuatan password.

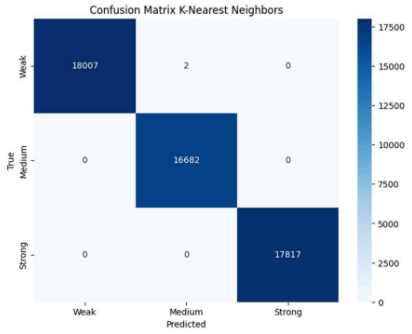
Hasil ini juga menguatkan studi sebelumnya yang menyatakan bahwa SVM sangat cocok digunakan dalam scenario klasifikasi dengan dimensi data terbatas namun memiliki distribusi yang dapat dipisahkan liner maupun-semi linier.



Gambar 4. Confusion Matrix SVM

2. K-Nearest Neighbors (KNN)

Kinerja KNN sangat sebanding dengan SVM. Confusion Matrix menunjukan bahwa seluruh sampel berhasil diklasifikasikan dengan benar ke kelas aslinya. Dengan nilai f1-score sebesar 1.00, model ini menunjukan bahwa ia mampu menangani perbedaan antar kelas tanpa overfitting atau underfitting. Keunggulan KKN dapa dikaitkan dengan sifat algoritma yang berbasis jarak (instance-based learning). Algoritma ini tidak membuat asumsu eksplisit terhdapap distribusi data, melainkan mengklasifikasikan data berdasarkan kesamaan fitur terdekat.ketika kelas dalam dataset terdistribusi secara baik maka Meachine Learning bekerja dengan sangat efektif.

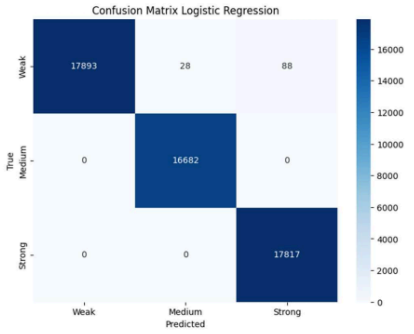


Gambar 4. Confusion Matrix K-Nearest Neighbors

3. Logistic Regression

Logistic Regression menunjukkan kinerja yang mendekati sempurna, namun masih terdapat sejumlah kecil kesalahan klasifikasi, khususnya pada kelas weak yang terkadang di prediksi sebagai medium, serta kesalahan monir antara kelas medium dan strong. Ini ditunjukan oleh nilai f1-score yang sedikit di bawah 1.00 untuk kelas tertentu.

Keterbatasa Logistic Regression muncul Ketika hubungan antar fitur dan label bersifat non-linier. Meskipun Logistic Regression adalah algoritma yang efesien dan mudah diinterpretasikan, kemampuannya untuk membedakan kelas dapat menurun saat data tidak sepenuhnya memenuhi asumsu linear separability. Password dengan pajang dan struktur yang tumpeng tindih antar kelas mungkin mempersulit dalam memetakan batas klasifikasi yangoptimal.



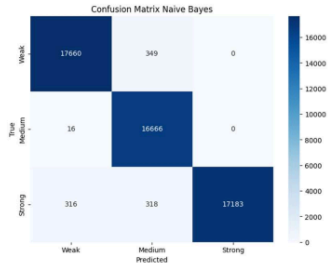
Gambar 5. Confusion Matrix Logistic Regression

Hasil dari Logistic Regression tetap kompetitif dan dapat digunakan sebagai baseline yang kuat, khususnya pada sistem yang membutuhkan interetabilitas dan kecepatan inferensi tinggi.

4. Naïve Bayes

Model Naïve Bayes mencapai hasil yang cukup tinggi namun konsisten menunjukan kelemahan pada kelas Strong. Kesalahan prediksi muncul Ketika password kompleks tidak sepenuhnya dikenali oleh model, mengakibatkan beberapa instance kelas strong salah diklasifikasikan sebagai Medium.

Hal ini terjadi karena asumsu utama Naïve bayes yaitu independensi antar fitur jarang terpenuhi dalam data password. Fitur-fitur seperti Panjang password,pengguna symbol dan kapitalisasi sering kali berkorelasi. Karena Naïve Bayes memperlakukan smua fitur sebgai independent secara probabilistic, model ini tidak dapat menangkap interaksi antar fitur dengan akurasi yang memadai



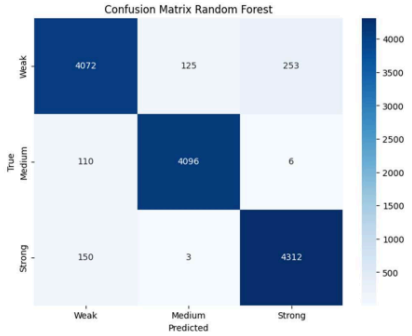
Gambar 5. Confusion Matrix Naïve Bayes

Namun demikian,karena sifatnya yang ringan dan cepat, Naïve Bayes tetap menjadi pilihan yang layak dalam situasu degan keterbatasan komputasi atau Ketika model digunakan sebagai sistem pendukung.

5. *Random Forest*

Random forest menghasilkan performa yang solid, namun tidak serta dengan SVM dan KKN. Confussion matrix memperlihatkan bahwa banyak instance dari kelas Weak di klasifikasikan secara keliru sebgai Medium atau Strong. Yang menurunkan nilai f1-score terutama pada kelas Weak.

Kesalahan ini terjadi karean ensemble tree dalam Random Forest mungkin terlalu dalam atau overfit pada kelas mayoritas, terutama jika distribusi fitur antara weak dan medium tidak terllau kontras. Meskipun Random Forest memiliki keunggulan dalam menangkap interaksi kompleks antar fitur dan menangani dataset berukuran besar, dalam kasus ini model tanpaknya kesulitan membedakan batas terfas antar kelas kekuatan password.



Gambar 5. Confusion Matrix Random Forest

V. KESIMPULAN

Penelitian ini membandingkan liam algoritma Machine Learning untuk mengklasifikasikan kekuran sandi, yakni K-Nearest Negihbors, Naïve Bayes, Logistic Regression, Random Forest, dan Support Vector Machine. Dari hasil pengujian, K-Nearest Neighbors dan Support Vector Machine mencatatkan performa tertinggi dengan akurasi dan F1-Score yang sempurna. Hal ini mengindikasikan bahwa kedua model ini sangat efektif dalam mengenali pola distribusi karakteristik kara sandi yang kompleks, bahkan tanpa memerlukan pelatihan intensif atau parameterisasi yang rumit.

Di sisi lain, Logistic Regression dan Naïve Bayes memberikan hasil yang cukup baik namun tidak tanpa kekurangan. Logistic Regression mengalami penurunan akurasi Ketika menghadapi data yang tidak sepenuhnya linear, sedangkan Naïve Bayes cenderumg keliru mengklasifikasikan kata sandi yang memiliki fitur saling bergantungan, seperti Panjang dan penggunaan symbol. Sementara Random Forest, Meskipun unggul dalam

mengatasi data tidak seimbang dan noise, menunjukan kelemahan dalam membedakan kelas Weak dan Medium secara akurat akibat overfitting pada distribusi fitur tertentu. Potensi aplikasi dari temuan ini cukup luas, khususnya dalam pewngembangan siste autentikasi yang adaptif di dunia maya. Model-model yang terbukti andal seperi support Vector Machine dan K-Nearest Neighbors dapat diintegrasikan ke dalam sistem keamanan platfrom daring, baik untuk, mendeteksi kekuatan password saat pembuatan akun maupun untuk menyarankan perbaikan secara real-time. Dengan demikikan, pendekatan ini tidak hanya meningkatkan keamanan pengguna secara signifikan, tetpai juga memberikan edukasi secara langsung tentang pentingnya membentuk kata sandi yang kuat.

DAFTAR PUSTAKA

[1] E. Mardiani, N. Rahmansyah, Y. F. Wijaya, A. A. Fitri, R. Mustafa, M. R. Rizki, and K. M. Pramesti, "Analisis Kompleksitas Password Dengan Metode KNN, Naïve Bayes, Decision Tree, Ensemble Methods Dan Linear Regression," *Digital Transformation Technology (Digitech)*, vol. 3, no. 2, pp. 955–966, Sep. 2023, doi: 10.47709/digitech.v3i2.3513.

[2] E. Darbutaite, P. Stefanovič, and S. Ramanauskaite, "Machine-Learning-Based Password-Strength-Estimation Approach for Passwords of Lithuanian Context," *Applied Sciences*, vol. 13, no. 13, p. 7811, Jul. 2023, doi: 10.3390/app13137811.

[3] S. Kuriakose, G. K. Teja, S. Duggi, A. H. Srivatsava, and V. Jonnalagadda, "Machine Learning Based Password Strength Analysis," *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, vol. 11, no. 8, pp. 5–8, Jul. 2022, doi: 10.35940/ijitee.H9119.0711822.

[4] D. Marutho, "Perbandingan Metode Naïve Bayes, KNN, Decision Tree Pada Laporan Water Level Jakarta," *Jurnal Ilmiah Infokam*, vol. 15, no. 2, 2019.

[5] E. Mardiani and F. A. Ramadhan, "Design Information System Sales of Nuts and Bolts at PT. Catur Naga Steelindo," *Jurnal SITEKIN: Jurnal Sains, Teknologi dan Industri*, vol. 20, no. 2, pp. 729–735, Jun. 2023.

[6] A. Lemantara, "Penerapan Algoritma Naïve Bayes dan ID3 untuk Memprediksi Segmentasi Pelanggan pada Penjualan Mobil," *Journal of Technology and Informatics (JoTI)*, vol. 4, no. 1, Oct. 2022.

[7] F. H. Pratama, A. Triayudi, and E. Mardiani, "Data Mining K-Medoids dan K-Means untuk Pengelompokan Potensi Produksi Kelapa Sawit di Indonesia," *Jurnal Ilmiah Penelitian dan Pembelajaran Informatika (JIPI)*, vol. 7, no. 4, 2022.

[8] H. Hozairi, A. Anwari, and S. Alim, "Implementasi Orange Data Mining Untuk Klasifikasi Kelulusan Mahasiswa dengan Model K-Nearest Neighbor, Decision Tree Serta Naive Bayes," *NERO (Networking Engineering Research Operation)*, vol. 6, no. 2, 2021.

[9] B. Bansal, "Password Strength Classifier Dataset," *Kaggle*, May 2022. [Online]. Available: https://www.kaggle.com/datasets/bhavikbb/password-strength-classifier-dataset

[10] M. Weir, S. Aggarwal, B. d. Medeiros, and B. Glodek, "Password Cracking Using Probabilistic Context-Free Grammars," in *Proc. 30th IEEE Symposium on Security and Privacy (SP)*, Berkeley, CA, 2009, pp. 391–405, doi: 10.1109/SP.2009.8.

Link Git HUB : https://github.com/melco-sv/Meachine-Learning-Password-Strength

Jurnal Data Mining.docx

ORIGINALITY REPORT

8%	7%	4%	1%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	jurnal.itscience.org Internet Source	1%
2	academic-accelerator.com Internet Source	1%
3	www.semanticscholar.org Internet Source	1%
4	Submitted to UPN Veteran Jawa Timur Student Paper	1%
5	digilib.batan.go.id Internet Source	<1%
6	ebin.pub Internet Source	<1%
7	publikasi.dinus.ac.id Internet Source	<1%
8	www.djournals.com Internet Source	<1%
9	repositorio.uss.edu.pe Internet Source	<1%
10	www.coursehero.com Internet Source	<1%
11	Dyan Prawita Sari, Zuhri Halim, Irlon Irlon, Bayu Waseso, Saromah Saromah. "Implementasi Machine Learning untuk Deteksi Intrusi pada Jaringan Komputer", Jurnal Minfo Polgan, 2024 Publication	<1%

12	journal.trunojoyo.ac.id Internet Source	<1%
13	marcelinawerda.blogspot.com Internet Source	<1%
14	123dok.com Internet Source	<1%
15	Arief Rahman Hakim, Alva Hendi Muhammad. "PERBANDINGAN MODEL TRANSFORMER, DEEP LEARNING, DAN MACHINE LEARNING UNTUK DETEKSI BERITA PALSU: STUDI KASUS PADA TEKS BERBAHASA INDONESIA", Jurnal Manajemen Informatika dan Sistem Informasi, 2025 Publication	<1%
16	journal.ilmudata.co.id Internet Source	<1%
17	repository.uin-malang.ac.id Internet Source	<1%

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	On		