Analisis Kekuatan Kata Sandi Berbasis Konteks Bahasa Indonesia Menggunakan Machine Learning

Putu Dena Satwika Sandi[a1], I Wayan Supriana[a2]

[a]Program Studi Informatika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Udayana
Jalan Raya Kampus Udayana, Bukit Jimbaran, Kuta Selatan, Badung, Bali, Indonesia
[1]denasatwika28@gmail.com
[2]wayan.supriana@unud.ac.id

Abstract

The widespread reliance on password authentication is persistently undermined by users creating contextually weak passwords, a vulnerability often overlooked by standard, English-centric password strength meters. This research addresses this security gap by developing and evaluating a machine learning model specifically tailored for password strength analysis within the Indonesian linguistic context. We trained a Decision Tree classifier and benchmarked it against a robust XGBoost model using a dataset enriched with local passwords and contextual features, including a custom heuristic score and Levenshtein similarity to a comprehensive Indonesian dictionary. To overcome severe class imbalance, the Synthetic Minority Oversampling Technique (SMOTE) was applied to the training data. While the XGBoost model achieved superior predictive performance, the most significant finding emerged from the feature importance analysis, which revealed that our custom heuristic score and the password's length were the two most dominant predictors. This study successfully validates that a context-aware machine learning approach can effectively analyze password strength, underscoring the critical need to integrate local linguistic patterns into security mechanisms and providing a robust foundation for developing more secure authentication systems for Indonesian users.

Keywords: *Password Strength Analysis, Machine Learning, Decision Tree, XGBoost, Cybersecurity, Indonesian Language Context*

1. Pendahuluan

Keamanan informasi tetap menjadi tantangan utama dalam pengelolaan sistem digital. Meskipun telah diperkenalkan berbagai metode autentikasi alternatif seperti biometrik dan autentikasi dua faktor, kata sandi tetap menjadi metode autentikasi yang paling banyak digunakan karena kemudahan implementasi dan biaya yang rendah [1]. Namun, penggunaan kata sandi memiliki kelemahan yang signifikan. Penelitian menunjukkan bahwa sekitar 30% pelanggaran data yang terjadi disebabkan oleh faktor manusia, seperti penggunaan kata sandi yang lemah atau mudah ditebak, yang menyoroti pentingnya peningkatan dalam pengelolaan kata sandi [1].

Berbagai alat bantu telah dikembangkan untuk membantu pengguna membuat kata sandi yang lebih kuat. Alat-alat ini umumnya mengevaluasi kekuatan kata sandi berdasarkan panjang karakter, keragaman simbol, dan kesamaan dengan kata sandi yang telah bocor dalam data pelanggaran sebelumnya. Namun, sebagian besar alat ini mengabaikan konteks semantik, seperti informasi pribadi yang tersebar di media sosial, yang dapat menyebabkan penilaian yang tidak akurat terhadap kekuatan kata sandi [2]. Keterbatasan pendekatan berbasis aturan (rule-based) dalam menangani konteks semantik inilah yang menjadi celah penelitian utama yang mendorong perlunya adopsi metode yang lebih canggih.

Kegagalan pendekatan konvensional dalam menangani konteks semantik mendorong adopsi metode pembelajaran mesin (machine learning) yang mampu mengidentifikasi pola dalam data historis dan mengklasifikasikan kata sandi dengan lebih akurat. Berbagai penelitian telah berhasil menerapkan algoritma supervised learning untuk klasifikasi kekuatan kata sandi. Model seperti

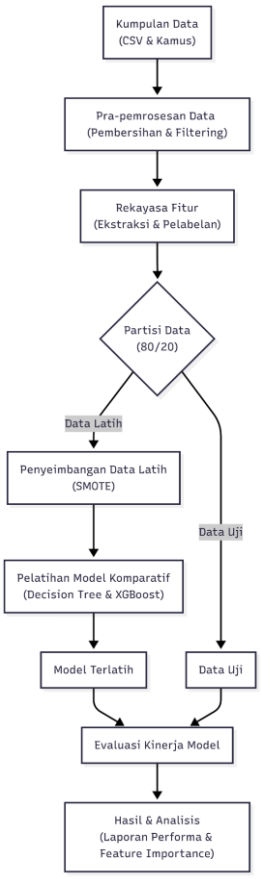
p-ISSN: 2986-3929
e-ISSN: 3032-1948

Decision Tree dan XGBoost telah menunjukkan kinerja yang sangat baik, dengan akurasi mencapai 99% dalam mengklasifikasikan kata sandi ke dalam kategori lemah, sedang, dan kuat [3][4].

Penulis disarankan untuk menggunakan struktur makalah sebagai berikut: **Pendahuluan**, **Metode Penelitian**, **Hasil dan Pembahasan** dan **Kesimpulan**.

2. Metode Penelitian

Penelitian ini menerapkan metodologi kuantitatif eksperimental untuk mengembangkan dan mengevaluasi model *machine learning* dalam mengestimasi kekuatan kata sandi. Alur kerja penelitian yang sistematis diilustrasikan pada Gambar 1.



Gambar 1. Alur Kerja Penelitian

2.1. Akuisisi dan Pra-pemrosesan Data

Dataset penelitian dikonstruksi dengan menggabungkan dua sumber utama: **kumpulan 485 kata sandi yang relevan dengan konteks Indonesia** dan **9,969 kata sandi dari *dataset* publik internasional**. Untuk analisis kontekstual, sebuah kamus Bahasa Indonesia kustom dikompilasi dari berbagai sumber terbuka. Setelah digabungkan, *dataset* final yang digunakan dalam penelitian ini terdiri dari 10.454 entri kata sandi. Seluruh data kemudian melalui tahap pra-pemrosesan yang meliputi eliminasi entri duplikat dan nilai yang hilang (*null values*) untuk menjamin integritas data.

p-ISSN: 2986-3929
e-ISSN: 3032-1948

Tabel 1. Daftar Kamus yang Digunakan Dalam Pengukuran

Kamus	Ukuran	Sumber
Kamus bahasa Indonesia	14556 kata	github.com
Kata sandi yang sering digunakan	47,023 kata	Kamus zxcvbn bawaan
Kata-kata bahasa Inggris di Wikipedia	100.000 kata	Kamus zxcvbn bawaan
Nama-nama pria dalam bahasa Inggris yang umum	1219 kata	Kamus zxcvbn bawaan
Nama-nama wanita dala bahasa Inggris yang umum	4275 kata	Kamus zxcvbn bawaan
Nama keluarga umum dalam bahasa Inggris	88.799 kata	Kamus zxcvbn bawaan

2.2. Ekstraksi Fitur dan Pelabelan

Untuk analisis *machine learning*, setiap kata sandi mentah dikonversi menjadi vektor fitur numerik. Fitur yang diekstraksi dikategorikan menjadi tiga jenis: standar, struktural, dan kontekstual, yang dirinci pada Tabel 2. Sebagai variabel target (*ground truth*), setiap kata sandi diberi label kekuatan (kelas 0-4) secara objektif menggunakan skor mentah yang dihasilkan oleh *library* zxcvbn.

Tabel 2. Fitur yang Digunakan dalam Penelitian

Kategori Fitur	Nama Fitur	Deskripsi
Standar	length, upper, lower, digit, symbol	Kuantitas total karakter, huruf besar, huruf kecil, angka, dan simbol.
Struktural	entropy, consecutive_chars, repeated_chars, keyboard_pattern	Nilai keacakan (*entropi*), jumlah pola berurutan, karakter berulang, dan indikator pola *keyboard*.
Kontekstual	levenshtein_similarity, custom_score	Tingkat kemiripan dengan kamus Indonesia dan skor kekuatan berdasarkan aturan heuristik kustom.

2.3. Pengembangan dan Evaluasi Model

Dataset yang telah melalui rekayasa fitur dipartisi menjadi data latih (80%) dan data uji (20%) menggunakan metode *stratified split*. Untuk mengatasi ketidakseimbangan kelas yang signifikan, teknik *oversampling* SMOTE (*Synthetic Minority Oversampling Technique*) diterapkan secara eksklusif pada data latih. Penelitian ini mengembangkan dua model untuk analisis komparatif:

- *Decision Tree Classifier*
- *XGBoost Classifier* (sebagai *baseline* performa tinggi)

Kedua model dilatih menggunakan data latih yang telah diseimbangkan. Kinerja model dievaluasi pada data uji menggunakan metrik standar, meliputi *Accuracy*, *Precision*, *Recall*, *F1-Score*, dan *Confusion Matrix*. Terakhir, analisis *Feature Importance* dilakukan pada model XGBoost untuk mengidentifikasi prediktor kekuatan kata sandi yang paling signifikan.

3. Hasil dan Diskusi

Bagian ini menyajikan dan membahas hasil kuantitatif dari penelitian yang dilakukan.

3.1. Lingkungan Penelitian

Eksperimen diimplementasikan menggunakan Python 3.10 dengan *library* utama meliputi pandas untuk manipulasi data, scikit-learn dan xgboost untuk pemodelan, imblearn untuk *oversampling*, serta zxcvbn dan fuzzywuzzy untuk rekayasa fitur.

3.2. Kinerja Model Klasifikasi

Kinerja kedua model dievaluasi pada data uji untuk mengukur kemampuan generalisasinya terhadap distribusi data riil. Model XGBoost menunjukkan performa yang lebih unggul dibandingkan *Decision Tree* dengan akurasi keseluruhan 64% dan F1-Score makro 0.68. Sebaliknya, *Decision Tree* mencatat akurasi 61% dan F1-Score makro 0.54. Performa detail dari model XGBoost yang superior disajikan pada Tabel 3.

Tabel 3. Laporan Klasifikasi Detail Model XGBoost

Kelas	***Precision***	***Recal***	***F-1 Score***	***Support***
0 (Sangat Lemah)	0.49	0.76	0.59	696
1 (Lemah)	0.81	0.56	0.66	1329
2 (Sedang)	0.53	0.71	0.61	14
3 (Kuat)	0.62	0.81	0.70	31
4 (Sangat Kuat)	0.76	0.90	0.83	21
Rata-rata Makro	0.64	0.75	0.68	2091

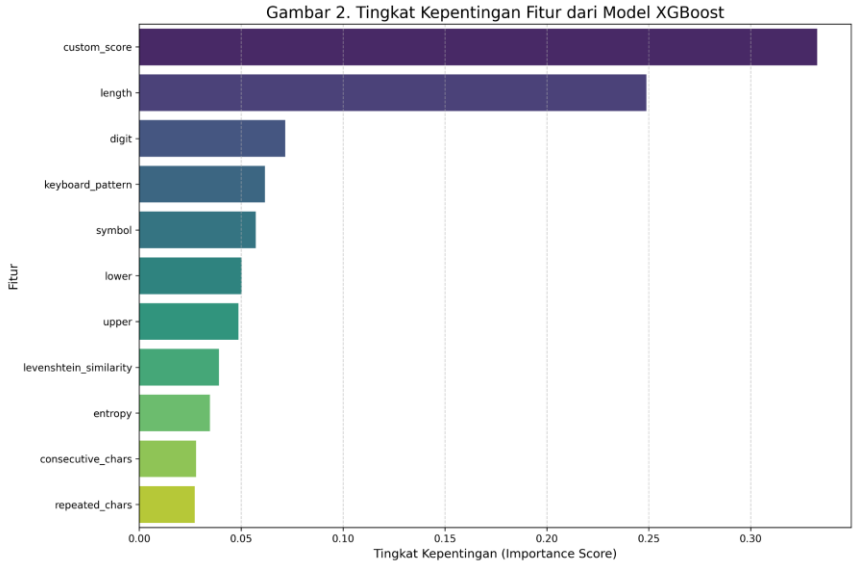
Tabel 4. Laporan Klasifikasi Detail Model Decision Tree

Kelas	***Precision***	***Recal***	***F-1 Score***	***Support***
0 (Sangat Lemah)	0.47	0.80	0.59	696
1 (Lemah)	0.82	0.51	0.63	1329
2 (Sedang)	0.35	0.50	0.41	14
3 (Kuat)	0.44	0.68	0.53	31
4 (Sangat Kuat)	0.58	0.52	0.55	21
Rata-rata Makro	0.53	0.60	0.54	2091

3.3. Analisis Kepentingan Fitur (*Feature Importance*)

Analisis *feature importance* dilakukan pada model XGBoost untuk mengidentifikasi prediktor paling signifikan. Hasilnya, diilustrasikan pada Gambar 2, menyoroti dua fitur yang dominan.

p-ISSN: 2986-3929
e-ISSN: 3032-1948



Gambar 2. Tingkat Kepentingan Fitur dari Model XGBoost

Fitur **custom_score (0.333)** dan **length (0.249)** secara kolektif menjadi dua prediktor paling berpengaruh. Temuan ini mengindikasikan bahwa kombinasi antara aturan heuristik sederhana dan panjang karakter merupakan sinyal terkuat untuk membedakan kekuatan kata sandi dalam model ini.

3.4. Diskusi

Hasil penelitian menunjukkan bahwa pendekatan *machine learning* dapat secara efektif menganalisis kekuatan kata sandi dalam konteks Indonesia. Keunggulan performa XGBoost atas *Decision Tree* menegaskan efektivitas algoritma *ensemble* dalam menangani pola data yang kompleks.

Temuan paling signifikan adalah dominasi fitur custom_score dan length. Hal ini mengimplikasikan bahwa sebuah pendekatan hibrida yang menggabungkan aturan-aturan kontekstual sederhana (tercermin dalam custom_score) dengan metrik fundamental seperti panjang karakter sangat efektif. Model ini tidak hanya belajar dari data, tetapi juga berhasil memanfaatkan pengetahuan domain yang telah dikodekan dalam fitur heuristik. Ini menyarankan bahwa sistem keamanan di masa depan dapat memperoleh manfaat besar dengan mengintegrasikan aturan bisnis atau konteks lokal sebagai fitur eksplisit dalam model prediktif.

Meskipun demikian, penelitian ini memiliki beberapa keterbatasan. Performa model, meskipun menjanjikan, menunjukkan masih ada ruang untuk peningkatan, yang mungkin dapat dicapai melalui pengayaan *dataset* dengan lebih banyak data asli Indonesia atau melalui *tuning hyperparameter* yang lebih ekstensif. Selain itu, ketergantungan pada zxcvbn sebagai sumber label *ground truth* berarti model ini pada dasarnya belajar untuk meniru dan memperkaya logika zxcvbn, bukan mendefinisikan kekuatan dari prinsip pertama.

4. Kesimpulan

Penelitian ini berhasil menunjukkan bahwa pendekatan *machine learning* yang diperkaya dengan konteks lokal efektif untuk menganalisis kekuatan kata sandi berbahasa Indonesia. Hasil analisis komparatif membuktikan keunggulan model XGBoost atas *Decision Tree*, dan yang lebih signifikan, analisis *feature importance* mengidentifikasi bahwa panjang karakter dan skor heuristik kustom merupakan dua prediktor paling dominan. Temuan ini menggarisbawahi pentingnya pendekatan hibrida yang mengintegrasikan aturan kontekstual ke dalam model prediktif untuk meningkatkan akurasi sistem keamanan. Meskipun demikian, terdapat ruang untuk

p-ISSN: 2986-3929
e-ISSN: 3032-1948

pengembangan di masa depan, di mana penelitian selanjutnya dapat berfokus pada pengayaan *dataset* dengan lebih banyak data otentik dari Indonesia, eksplorasi model *deep learning* untuk menangkap pola yang lebih kompleks, serta melakukan optimalisasi *hyperparameter* untuk lebih meningkatkan kinerja model.

Daftar Pustaka

[1] W. P. K. Fernando, D. A. N. P. Dissanayake, S. G. V. D. Dushmantha, D. L. C. P. Liyanage, and C. Karunatilake, "Challenges and opportunities in password management: A review of current solutions," *Sri Lanka Journal of Social Sciences and Humanities*, vol. 3, no. 2, pp. 9-20, Aug. 2023. [**ResearchGate**] [**SLJSSH**]

[2] M. Atzori, E. Calò, L. Caruccio, S. Cirillo, G. Polese, and G. Solimando, "Evaluating password strength based on information spread on social networks: A combined approach relying on data reconstruction and generative models," *Online Social Networks and Media*, vol. 42, p. 100278, 2024. [**ScienceDirect**]

[3] S. Sarkar and M. Nandan, "Password Strength Analysis and its Classification by Applying Machine Learning Based Techniques," in *2022 Second International Conference on Computer Science, Engineering and Applications (ICCSEA)*, 2022, pp. 1–6. [**ResearchGate**] [**IEEEXplore**]

[4] U. Farooq, "Real Time Password Strength Analysis on a Web Application Using Multiple Machine Learning Approaches," *International Journal of Engineering Research & Technology (IJERT)*, vol. 9, no. 12, pp. 359–364, Dec. 2020. [**IJERT**]

[5] E. Darbutaitė, P. Stefanovič, and S. Ramanauskaitė, "Machine-learning-based password-strength-estimation approach for passwords of Lithuanian context," *Appl. Sci.*, vol. 13, no. 13, p. 7811, Jul. 2023. [**MDPI**]

[6] Z. Damzaky, "Kumpulan kata bahasa Indonesia KBBI (legacy list)," *GitHub*. [Online]. Available: https://github.com/damzaky/kumpulan-kata-bahasa-indonesia-KBBI/tree/master/legacy

[7] A. Ramadhan, "Daftar password yang sering dipakai orang Indonesia di 2022," *Kumparan Tech*, Nov. 15, 2022. [Online]. Available: https://kumparan.com/kumparantech/daftar-password-yang-sering-dipakai-orang-indonesia-di-2022-1zJEhTf3s0o

[8] T. Vohra, "Bruteforce database - password dictionaries," *Kaggle*, 2022. [Online]. Available: https://www.kaggle.com/datasets/taranvee/bruteforce-database-password-dictionaries