**Digital Transformation Technology (Digitech) | e-ISSN: 2807-9000**
**Volume** 3, **Number** 2, September 2023
https://doi.org/10.47709/digitech.v3i2.3513

# Analisis Kompleksitas Password Dengan Metode KNN, Naïve Bayes, Decision Tree, Ensemble Methods Dan Linear Regression

**Eri Mardiani[1*], Nur Rahmansyah[2], Yunan Fauzi Wijaya[3], Annisa Amalia Fitri[4], Rayhan Mustafa[5],Muhammad Romadhoni Rizki[6], Komang Mustika Pramesti[7]**
[1,3]Universitas Nasional, [,2]Politeknik Negeri Media Kreatif, Indonesia, [4.5,6,7]FEB, UPN Veteran Jakarta, Indonesia
[1]erimardiani1@gmail.com, [2]nur_rahmansyah@polimedia.ac.id, [3]yunanfw@gmail.com, [4]annisamalia07@gmail.com, [5]rayhanmustafa36@gmail.com, [6]2010111172@mahasiswa.upnvj.ac.id, [7]mustikapramesti55@gmail.com

open access
**Histori Artikel:**
Diajukan: 26 Januari 2024
Disetujui: 30 Januari 2024
Dipublikasi: 31 Januari 2024
**Kata Kunci:**
Password, data mining, keamanan, orange, algoritma

***Digital Transformation Technology (Digitech)*** *is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0).*

**Abstrak**
Dalam era digital yang semakin kompleks dan penuh tantangan, keamanan kata sandi menjadi krusial untuk melindungi informasi sensitif dan mencegah potensi ancaman keamanan siber. Kata sandi merupakan lapisan pertama pertahanan dalam banyak sistem keamanan digital, oleh karena itu, pemahaman mendalam tentang metode yang efisien dalam menilai dan memprediksi kompleksitas password sangatlah penting. Ketika dihadapkan dengan data yang sangat kompleks, diperlukan analisis dan representasi visual data agar informasi dapat lebih mudah dipahami. Untuk mengilustrasikan data dengan cara yang interaktif dan dapat dimengerti oleh berbagai kalangan, salah satu software atau alat bantu yang dapat digunakan adalah Orange. Dalam pengolahan data ini menggunakan aplikasi orange, kami menganalisis bagaimana prediksi hubungan antara password dan tingkat kompleksitasnya menggunakan fitur-fitur yang telah dikonstruksi. melakukan analisis data mining melalui penerapan teknik klasifikasi dengan memanfaatkan lima metode algoritma yang berbeda. Dataset yang akan dijadikan dasar proyek berasal dari publikasi data pada situs Kaggle.com.

## PENDAHULUAN

Dalam era digital yang semakin kompleks dan penuh tantangan, keamanan kata sandi menjadi krusial untuk melindungi informasi sensitif dan mencegah potensi ancaman keamanan siber. Kata sandi merupakan lapisan pertama pertahanan dalam banyak sistem keamanan digital, oleh karena itu, pemahaman mendalam tentang metode yang efisien dalam menilai dan memprediksi kompleksitas password sangatlah penting. Ketika dihadapkan dengan data yang sangat kompleks, diperlukan analisis dan representasi visual data agar informasi dapat lebih mudah dipahami. Untuk mengilustrasikan data dengan cara yang interaktif dan dapat dimengerti oleh berbagai kalangan, salah satu software atau alat bantu yang dapat digunakan adalah Orange. (Mardiani, et al., 2023)

Studi Literatur Tentang Naive Bayes Algorithm for Sentiment analysis Windows PhoneStore Application Reviews dapat disimpulkan bahwa, menganalisa pengguna aplikasi Windows Phone Store dengan mengklasifikasikan review ke dalam kategori opini positif atau negative. (Mardiani & Ramadhan, Design Information System Sales of Nuts and Bolts at PT. Catur Naga Steelindo, 2023). Dan studi literatur tentang Penerapan Algoritma Naïve Bayes dan ID3 untuk Memprediksi Segmentasi Pelanggan pada Penjualan Mobil. (Lemantara, 2022)

Perkembangan zaman memberikan peningkatan data yang semakin banyak data yang dibutuhkan oleh setiap manusia. (Rahmansyah, Mulyani, Mardiani, & Rahman, 2022). Dengan banyaknya data yang ada, maka semakin sulit juga untuk menganalisis data tersebut dengan cara manual. Oleh karena itu, manusia membutuhkan bantuan data mining untuk mengumpulkan informasi menarik dari data yang tersedia. Dengan menggunakan data mining maka akan mendapatkan penemuan pengetahuan yang menarik dari sejumlah besar data yang disimpan dalam basis data, gudang data. (Laksono, Achmadi, & Sasmito, 2023)

Big Data Analytics dapat menawarkan keuntungan besar bagi bisnis yang menggunakannya secara efektif, termasuk pengambilan keputusan yang lebih baik dan lebih cepat, lebih banyak kepuasan pelanggan, penemuan peluang baru, pengoptimalan proses, dan pembuatan barang dan layanan yang lebih baik. (Indrapras, Alfaris, Falentina, & Triana, 2022). Dataset digunakan untuk melihat dampak dari data categorical (Gender, Race/ethnicity, Parental level of education, Lunch, and Test preparation course) terhadap data numeric (nilai matematika, membaca, dan menulis). (https://www.kaggle.com/datasets/morph1max/password-security-sber-dataset/data) Dengan menggunakan aplikasi Orange Data Mining , maka pengolahan data supaya proses analisis udapat dilakukan untuk memprediksi hubungan antara password dan tingkat kompleksitasnya menggunakan fitur-

fitur yang telah dikonstruksi, menggunakan dataset password security sber untuk melakukan analisis data mining melalui penerapan teknik klasifikasi dengan memanfaatkan lima metode algoritma yang berbeda. (Marutho, 2019)

Dengan prediksi menggunakan metode-metode K-Nearest Neighbour (KNN), Naive Bayes, Decision Tree, Ensemble Method, dan Linear Regression dapat dilakukan analisa diagnosa keganasan kanker berdasarkan ciri-ciri yang ditunjukkan oleh sel kanker yang terbentuk menggunakan dataset Benign and Malignant Cancer Data.

## STUDI LITERATUR

Data Mining menggunakan Teknik Klasifikasi dengan 5 Model Algoritma (Priyanti, 2019)

1.K-Nearest Neighbor (k-NN) (Hozairi, Anwari, & Alim, (2021),Vol 6, No 2 )

Algoritma K-Nearest Neighbor (k-NN) merupakan sebuah model algoritma yang digunakan untuk melakukan klasifikasi terhadap objek yang berdasarkan dari data pembelajaran yang jaraknya paling dekat dengan objek tersebut. Ataupun dapat dipahami juga bahwa k-nearest neighbor adalah salah satu algoritma yang paling sederhana dan banyak digunakan. Titik data akan diklasifikasikan berdasarkan kesamaan kelompok tertentu dari titik data lain yang berdekatan. Sehingga, algoritma ini akan memberikan hasil yang kompetitif. (Indriyanti, Ichsan, Fatah, Wahyuni, & Ermawati, Aug 30, 2022)

Teorema K-NN

$$d_i = \sqrt{\sum_{i=1}^{n} (x_{ij} - p_j)^2}$$

Keterangan
$d_i$ = Jarak sampel
$x_{ij}$ = Data sampel pengetahuan
$p_j$= Data input var ke-j
n = Jumlah sampel

2.Naive Bayes (Marutho, 2019)

Salah satu metode data mining ialah klasifikasi Naive Bayes. Naive Bayes Classifier adalah metode klasifikasi yang berakar pada teorema Bayes. Naive bayes merupakan metode pengklasifikasian berdasarkan probabilitas sederhana dan dirancang agar dapat dipergunakan dengan asumsi antar variabel penjelas saling bebas (independen). Pada algoritma ini pembelajaran lebih ditekankan pada pengestimasian probabilitas. tujuan dari metode Naïve Bayes adalah untuk menemukan probabilitas ketika kita mengetahui probabilitas tertentu lainnya. Hasil dari perhitungan data mining menggunakan metode klasifikasi Naïve Bayes akan makin berguna jika penyajiannya menarik dan dapat dipahami dengan baik oleh penerima data. (Mardiani, et al., 2023)

Teorema Naive Bayes (Giri, 2018)

$$P(H|X)=\frac{P(X|H).P(H)}{P(X)}$$

Keterangan
X : Data dengan class yang belum diketahui
H : Hipotesis data merupakan suatu class spesifik
P(H|X) : Probabilitas hiotesis H berdasar kondisi X (Posteriori Probabilitas)
P(H) : Probabilitas Hipotesis H (Prior Probabilitas)
P(X|H) : Probabilitas X berdasarkan kondisi hipotesis H
P(X) : Probabilitas X

3.Decision Tree

Decision tree adalah algoritma machine learning yang menggunakan seperangkat aturan untuk membuat keputusan dengan struktur seperti pohon yang memodelkan kemungkinan hasil, biaya sumber daya, utilitas dan kemungkinan konsekuensi atau resiko. Konsepnya adalah dengan cara menyajikan algoritma dengan pernyataan bersyarat, yang meliputi cabang untuk mewakili langkah-langkah pengambilan keputusan yang dapat mengarah pada hasil yang menguntungkan. Klasifikasi ini menggunakan observasi pada node untuk menemukan target pada leaves. Decision Tree merupakan salah satu metode klasifikasi yang paling populer karena mudah untuk diinterpretasi oleh manusia dengan kemampuannya untuk mem-break down proses pengambilan keputusan yang kompleks menjadi lebih simple. (Hadi, Agustina, & Subhani, Vol 3, No 1 (2019) )

Teorema Decision Tree (Indriyawati & Khoirudin, 2019-12-05)

$$Entropy(S) = \sum_{i=1}^{n} pi * \log_2 pi$$

Keterangan:
S = Himpunan kasus
A = Atribut
n = Jumlah patisi S
$P_i$ = Proporsi dari $S_i$ terhadap S

4.Ensemble Method
Ensemble Method adalah algoritma dalam pembelajaran mesin (machine learning) dimana algoritma ini sebagai pencarian solusi prediksi terbaik dibandingkan dengan algoritma yang lain karena metode ensemble ini menggunakan beberapa algoritma pembelajaran untuk pencapaian solusi prediksi yang lebih baik daripada algoritma yang bisa diperoleh dari salah satu pembelajaran algoritma konstituen saja. Tidak seperti ensemble statistika dalam mekanika statistika biasanya selalu tak terbatas. Ensemble Pembelajaran hanya terdiri dari seperangkat model alternatif yang bersifat terbatas, namun biasanya memungkinkan untuk menjadi lebih banyak lagi struktur fleksibel yang ada diantara alternatif model itu sendiri. Evaluasi prediksi dari ensemble biasanya memerlukan banyak komputasi daripada evaluasi prediksi model tunggal (single model). (Karo, Erwansyah, & Suharsil, 2020)

$$F(x) = \sum_{t=1}^{T} \propto_t h_t(x)$$

Di mana:
$h(x)$ : Pengklasifikasi dasar atau lemah
$\propto$ : Tingkat pembelajaran (learning rate)
F(x) : Hasil, berupa pengklasifikasi kuat atau akhir

5.Linear Regression
Linear regression merupakan salah satu algoritma yang memodelkan suatu persamaan untuk menghitung estimasi. Pada metode ini bertujuan untuk mencari pola pada nilai numerik, sehingga data yang dibutuhkan berupa data numerik agar dapat diolah dengan model algoritma ini. Linear regression memprediksi nilai data yang tidak diketahui dengan menggunakan nilai data lain yang terkait dan diketahui. Secara matematis memodelkan variabel yang tidak diketahui atau tergantung dan variabel yang dikenal atau independen sebagai persamaan linier.
Teorema Linear Refression (Pratama, Triayudi, & Mardiani, 2022, Vol 7, No 4)

$$Y = a + bX$$

Mengidentifikasi Atribut dan Label
Keterangan:
Y : Variabel terikat (dependen)
a : konstanta
b : koefisien regresi (kemiringan); besaran Respon yang ditimbulkan oleh variabel
X : Variabel tidak terikat (Independen)
Data Mining menggunakan Teknik Klasifikasi dengan 5 Model Algoritma(Mardiani, Rahmansyah, Ningsih, Lantana, dkk., 2023)

Model algoritma yang akan digunakan adalah sebagai berikut:
1. K-Nearest Neighbor (kNN)

Algoritma K-Nearest Neighbor (kNN) merupakan sebuah model algoritma yang digunakan untuk melakukan klasifikasi terhadap objek yang berdasarkan dari data pembelajaran yang jaraknya paling dekat dengan objek tersebut. Ataupun dapat dipahami juga bahwa K-Nearest Neighbor adalah salah satu algoritma yang paling sederhana dan banyak titik data akan diklasifikasikan berdasarkan kesamaan kelompok tertentu dari titik data lain yang berdekatan. Sehingga, algoritma ini akan memberikan hasil yang kompetitif (Mardiani et al., 2023).
2.Naïve Bayes

Klasifikasi Naïve Bayes merupakan suatu pendekatan yang berakar pada teorema Bayes. Naive Bayes Classifier digunakan untuk mengklasifikasikan data dengan mengasumsikan bahwa variabel penjelas saling bebas atau independen satu sama lain. Algoritma ini fokus pada pembelajaran probabilitas, dan tujuan utamanya adalah untuk menemukan probabilitas ketika probabilitas lainnya sudah diketahui. Keberhasilan metode Naïve Bayes

dalam data mining sangat tergantung pada presentasi hasil yang menarik dan dapat dipahami dengan baik oleh penerima data (Mardiani et al., 2023).

3.Decision Tree

Decision tree merupakan suatu algoritma machine learning yang menggunakan serangkaian aturan untuk mengambil keputusan dengan struktur mirip pohon, yang memodelkan kemungkinan hasil, biaya sumber daya, utilitas, serta risiko atau konsekuensi yang mungkin terjadi. Ide dasarnya adalah menyajikan algoritma dengan pernyataan bersyarat, yang terdiri dari cabang-cabang untuk merepresentasikan langkah-langkah pengambilan keputusan yang dapat membawa pada hasil yang menguntungkan. Decision tree menjadi salah satu metode klasifikasi yang populer karena kemudahan interpretasinya oleh manusia, mampu menyederhanakan proses pengambilan keputusan yang kompleks menjadi lebih mudah dimengerti (Mardiani et al., 2023).

4.Ensemble Method

Ensemble Method merupakan salah satu algoritma sebagai pencarian solusi prediksi terbaik dibandingkan dengan algoritma lainnya karena metode ini menggunakan beberapa algoritma untuk pencapaian solusi prediksi yang lebih baik. Algoritma yang umum dipakai pada ensemble method ini diantaranya ada algortima Decision Tree, Random Forest, dan AdaBoost (Mardiani et al., 2023).

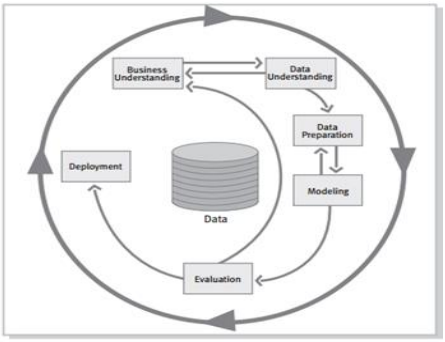
5.Linear Regression

Algoritma Linear Regression merupakan salah satu algoritma yang memodelkan suatu persamaan untuk menghitung estimasi. Pada metode ini bertujuan untuk mencari pola pada nilai numerik, sehingga data yang dibutuhkan berupa data numerik agar dapat diolah dengan model algoritma ini. Linear regression memprediksi nilai data yang tidak diketahui dengan menggunakan nilai data lain yang terkait dan diketahui. Secara matematis memodelkan variabel yang tidak diketahui atau tergantung dan variabel yang dikenal atau independen sebagai persamaan linier (Mardiani et al., 2023).

## METODE

Metode penelitian yang digunakan adalah deskriptif , observasi, wawancara dan dokumentasi di lapangan memberikan informasi tentang fakta yang ada. Setelah mengumpulkan data, disajikan dalam bentuk yang memungkinkan untuk ditarik kesimpulan.

Penelitian ini untuk melakukan analisa perbandingan metode K-NN, *Naive Bayes , Decision tree, Ensemble Method, Linear regression* menggunakan dataset Password Security: Sber Dataset menggunakan Orange Tools untuk simulasinya. (Mardiani, Eri, Nur Rahmansyah, Endah Tri Esti Handayani, Sari Ningsih, Deny Hidayatullah, Dhieka Avrilia Lantana, Yuni Latifah, Alica Dwi Fahira, Keysha Belynda Tyva Panggabean, Imelta Natalia Ginting.2023). Data yang dikumpulkan untuk penelitian ini yaitu primer dan sekunder. Teknik pengumpulan data primer berasal dari studi literatur seperti buku maupun jurnal yang memiliki keterkaitan terhadap sistem terdistribusi pada sekolah dasar. Adapun Teknik pengumpulan data sekunder pada penelitian ini diperoleh dari media online maupun sumber – sumber lainnya. Studi pustaka akan digunakan oleh peneliti guna menjabarkan dan menganalisis data yang sudah dikumpulkan yang memiliki kaitan dengan penelitian ini. (Mardiani, Rahmansyah dkk., 2021).




Gambar 1. Tahapan Penelitian

Business understanding

Tahap awal berfokus pada pemahaman tujuan dan persyaratan proyek dari perspektif bisnis, kemudian mengubah pengetahuan ini menjadi definisi masalah data mining dan rencana awal yang dirancang untuk mencapai tujuan.

Data understanding

Tahap pemahaman data dimulai dengan pengumpulan data awal dan dilanjutkan dengan aktivitas yang memungkinkan untuk mengenal data, mengidentifikasi masalah kualitas data, menemukan informasi tentang data, dan / atau mendeteksi subset yang menarik untuk membentuk hipotesis mengenai informasi tersembunyi.

Data Preparation / Data Preprocessing
Tahap persiapan data mencakup semua kegiatan yang diperlukan untuk membangun data akhir (data yang akan dimasukkan ke dalam alat pemodelan) dari data mentah awal. Tugas persiapan data kemungkinan dilakukan beberapa kali dan tidak dalam urutan yang ditentukan. Tugas meliputi pemilihan tabel, catatan, dan atribut, serta transformasi dan pembersihan data untuk alat pemodelan.
Modelling
Pada tahap ini, berbagai teknik pemodelan dipilih dan diterapkan, dan parameternya dikalibrasi dengan nilai optimal. Biasanya, ada beberapa teknik untuk jenis masalah data mining yang sama. Beberapa teknik memiliki persyaratan khusus berupa data. Karena itu, tahap persiapan data sangat diperlukan.
Evaluation
Sebelum dilanjutkan ke penerapan model akhir, penting untuk mengevaluasi secara menyeluruh dan meninjau langkah-langkah yang dijalankan untuk membuatnya, untuk memastikan model tersebut benar mencapai tujuan bisnis. Tujuan utamanya adalah untuk menentukan apakah ada beberapa masalah bisnis penting yang belum cukup dipertimbangkan. Pada akhir tahap ini, keputusan penggunaan hasil data mining harus dicapai.
Deployment
Pembuatan model umumnya bukan akhir dari proyek. Sekalipun tujuan dari model ini adalah untuk meningkatkan pengetahuan tentang data, pengetahuan yang didapat perlu diorganisir dan disajikan sedemikian rupa sehingga pelanggan dapat menggunakannya. Bergantung pada persyaratan, fase deployment dapat sesederhana menghasilkan laporan atau serumit penerapan proses pengumpulan data berulang diseluruh perusahaan. Jika analis akan melakukan upaya deployment, penting bagi pelanggan untuk memahami terlebih dahulu tindakan apa yang perlu dilakukan agar benar-benar memanfaatkan model yang dibuat.

## HASIL

Dataset dalam analisis ini diperoleh dari *kaggle.com* yang mencakup informasi terkait 100.000 kata sandi dan kategori kelas dari kata sandi tersebut (0 – kata sandi tidak dapat diandalkan, 2 – kata sandi sangat dapat diandalkan). Adapun perolehan data tersebut dapat dilihat pada tabel berikut.

Tabel 1. Dataset Password Security: Sber Dataset

| | **Password** | **Strength** |
|---|---|---|
| **0** | yrtzuab476 | 1 |
| **1** | yEdnN9jc1NgzkkBP | 2 |
| **2** | sarita99 | 1 |
| **…** | … | … |
| **99998** | toji25 | 0 |
| **99999** | midnight000 | 1 |

Atribut dataset untuk penelitian ini diperoleh dari Kaggle.com yang kemudian diolah menggunakan aplikasi Orange. Atribut tersebut mencakup informasi mengenai kata sandi dan tingkat kompleksitas yang kelompok kami kategorikan ulang menggunakan Edit Domain untuk menggambarkan kekuatan suatu password. Sebagai contoh, angka "0" mengindikasikan "Weak," angka "1" mengindikasikan "Moderate," dan angka "2" mengindikasikan "Strong". Kemudian, kami juga menambahkan feature pada dataset ini menggunakan Feature Constructor atau Formula agar dataset ini dapat digunakan secara optimal dalam algoritma Orange Data Mining. Atribut yang akan kami gunakan pada analisis ini dapat dilihat dalam tabel berikut.

Tabel 2. Keterangan Atribut Dataset

| **Name** | **Type** | **Role** | **Explanation** |
|---|---|---|---|
| **Password** | Text | Meta | |
| **Strength** | Categorical | Target | Tingkat kompleksitas *password* |
| **Length** | Numerical | Feature | Jumlah huruf dan angka pada *password* |
| **Digit Count** | Numerical | Feature | jumlah digit angka pada *password* |
| **Uppercase Count** | Numerical | Feature | jumlah huruf kapital pada *password* |
| **Lowercase Count** | Numerical | Feature | Jumlah huruf kecil pada |

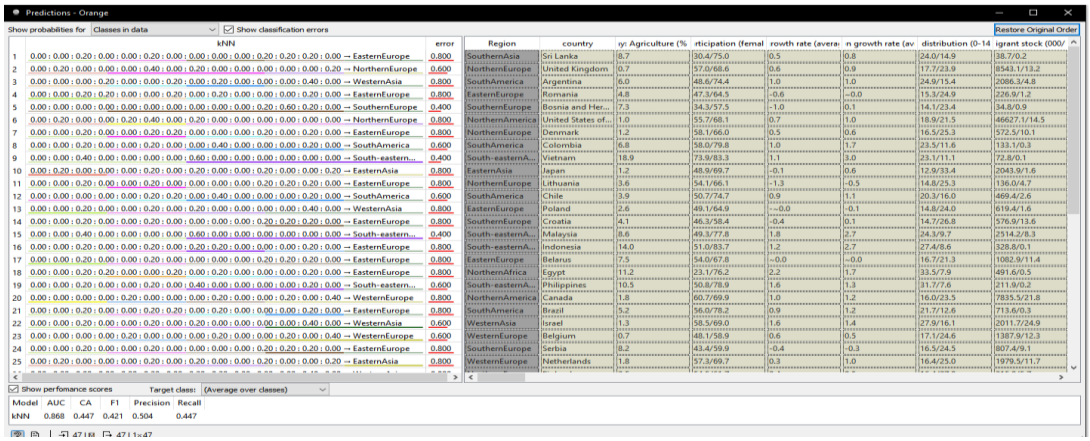
| | | | |
|---|---|---|---|
| | | | *password* |
| **Special Char Count** | Numerical | Feature | Jumlah symbol pada *password* |

Kemudian, dalam analisis ini kami menggunakan 15% dari total dataset, mencakup sekitar 15.000 data dari jumlah keseluruhan 100.000 data yang dapat diakses. Pendekatan ini dipilih dengan maksud untuk melakukan evaluasi dan pengujian pada subset yang representatif, sehingga hasil analisis dapat memberikan gambaran yang lebih akurat dan relevan terhadap keseluruhan dataset.

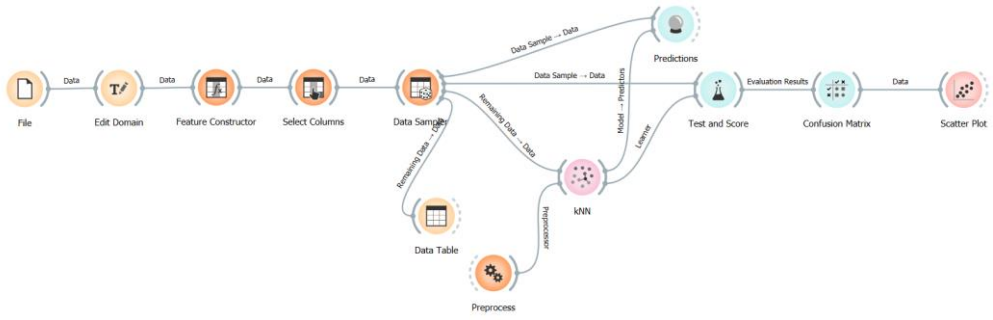
## PEMBAHASAN

Berikut merupakan workflow teknik klasifikasi dari dataset Password Security: Sber Dataset dengan menggunakan lima jenis model algoritma. Datasets ini diperoleh dari website kaggle dan digunakan sebagai penyaji informasi terkait prediksi nilai dari sekelompok atribut.

**Metode KNN Algoritma K-Nearest Neighbor (k-NN)**

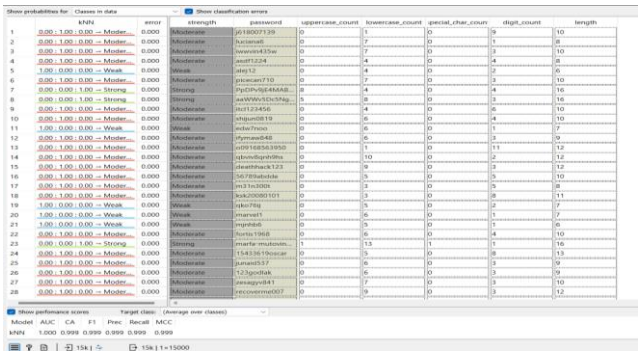


Gambar 2 Penggunaan *Data Sampler* Pada *Orange*



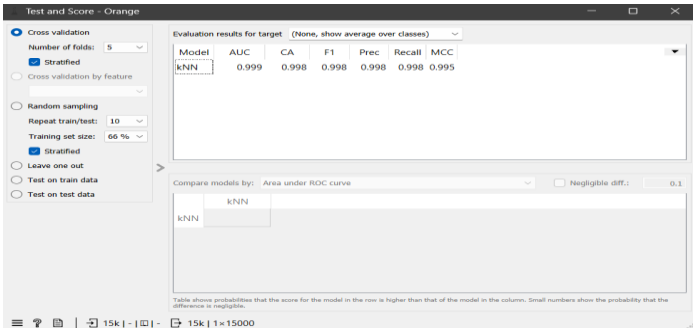

Gambar 3. *Work Flow* Analisa k-NN

Gambar 3 di atas merupakan *work flow* yang digunakan untuk melakukan analisis prediksi menggunakan algotritma *K-Nearest Neighbor.* Selanjutnya hasil *predictions* dapat dilihat pada gambar di bawah ini.



Gambar 4. Analisis *Prediction* k-NN

Berdasarkan hasil penelitian menggunakan *Predictions* dapat dilihat tabel nilai prediksi dan tabel nilai aktualnya. Apabila dilihat secara manual oleh tabel pada *Predictions*, maka nilai prediksi menggunakan metode K-NN hampir secara keseluruhan dapat mendeskripsikan dengan benar nilai aktualnya. Hal ini didukung dengan nilai K-NN yang berada pada posisi kiri bawah gambar. Nilai K-NN memperoleh nilai AUC sebesar 100% yang berarti nilai tersebut dianggap sangat baik, nilai CA sebesar 99,9%, nilai F1 sebesar 99,9%, nilai *Precisions* sebesar 99,9%, nilai *Recall* sebesar 99,9%, dan nilai MCC sebesar 99,9%. Secara keseluruhan melalui penelitian *Predictions* nilai K-NN dapat dikatakan sangat baik.




Gambar 5. Analisis *Test and Score* k-NN

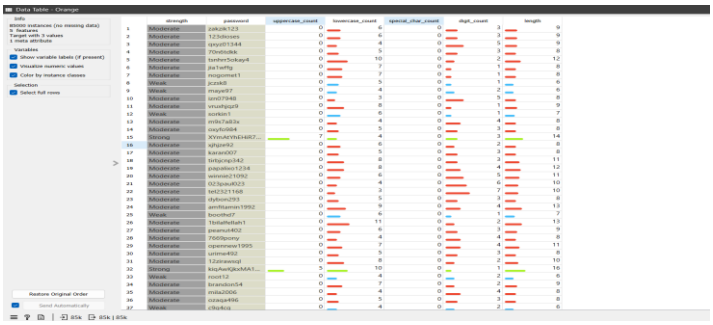
Melalui penelitian nilai *test and score* secara aktual terdapat perbedaan dengan hasil penelitian menggunakan *Predictions*. Pada hasil penelitian nilai *test and score*, dapat diketahui bahwa nilai AUC sebesar 99,9% sehingga dapat dikatakan sangat baik karena mendekati angka 100%, nilai CA yaitu sebesar 99,8%, nilai F1 sebesar 99,8%, nilai *Precisions* sebesar 99,8%, nilai *Recall* sebesar 99,8%, dan nilai MCC sebesar 99,5%. Secara keseluruhan Secara keseluruhan melalui penelitian *Predictions* nilai K-NN dapat dikatakan sangat baik.

Predicted

| Actual | Weak | Moderate | Strong | Σ |
|---|---|---|---|---|
| **Weak** | 2001 | 10 | 0 | **2011** |
| **Moderate** | 6 | 11106 | 12 | **11124** |
| **Strong** | 0 | 6 | 1859 | **1865** |
| **Σ** | **2007** | **11122** | **1871** | **15000** |

Gambar 6. Analisis *Confusion Matrix* k-NN

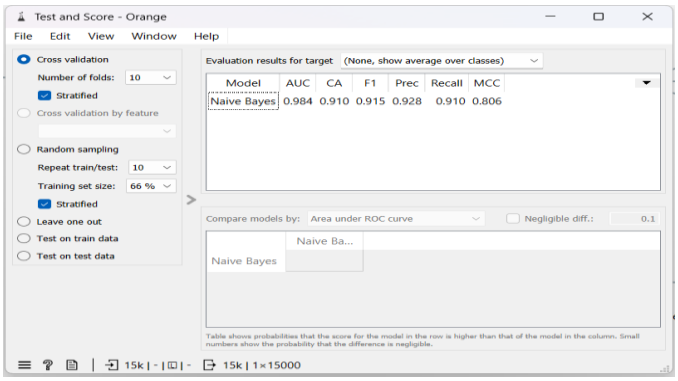
Selanjutnya, penilaian metode K-NN dapat dilihat melalui evaluasi menggunakan *Confusion Matrix*. Pada gambar di atas dijelaskan bahwa pada nilai aktual *Weak* dengan *Predicted Weak* sebanyak 2.001, nilai aktual *Weak* dengan *Predicted Moderate* sebanyak 10, dan nilai aktual *Weak* dengan *Predicted Strong* sebanyak 0. Selanjutnya nilai aktual *Moderate* dengan *Predicted Weak* sebanyak 6, nilai aktual *Moderate* dengan *Predicted Moderate* sebanyak 11.106, dan nilai aktual *Moderate* dengan *Predicted Strong* sebanyak 12. Selanjutnya nilai aktual *Strong* dengan *Predicted Weak* sebanyak 0, nilai aktual *Strong* dengan *Predicted Moderate* sebanyak 6, dan nilai aktual *Strong* dengan *Predicted Strong* sebanyak 1.859. Hasil penelitian menggunakan *Confusion Matrix* juga dapat dillihat menggunakan ikon *Data table*, yang bertujuan untuk melihat secara manual dan lebih detail manakah nilai prediksi yang sesuai dari metode K-NN dengan Variabel *Weak, Moderate*, dan *Strong*. Untuk hasil *data table* ditampilkan seperti gambar 6 yang tertera di bawah.




Gambar 7. *Data Table* k-NN

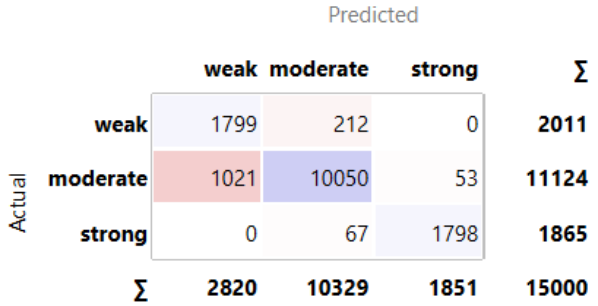
### *A. Analisis Model Algoritma Naïve Bayes*

hasil *test and score* dapat dilihat pada gambar di bawah ini.



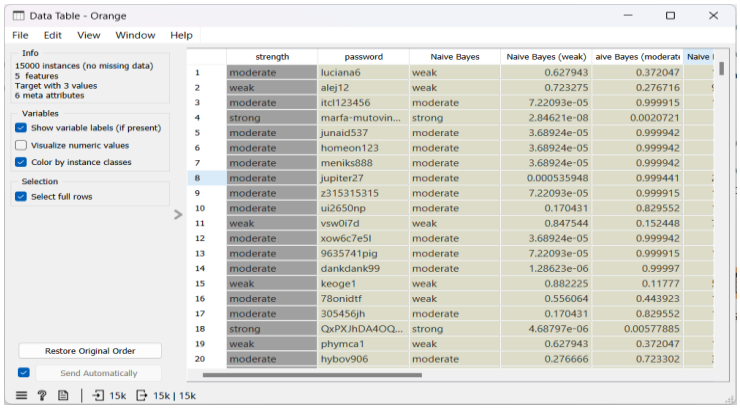

Gambar 8. *Test and Score Naïve Bayes*

Hasil penelitian menggunakan metode Naïve Bayes, seperti yang terlihat pada gambar 10 di atas, menunjukkan bahwa nilai AUC sebesar 98.4%. Hasil ini dapat dikategorikan sangat baik, karena mendekati 100%. Selain itu, metode Naive Bayes juga menghasilkan nilai CA sebesar 91%, nilai F1 sebesar 91.5%, nilai Presisi sebesar 92.8%, dan nilai Recall sebesar 91%.




Gambar 9. Confusion Matrix Naive Bayes

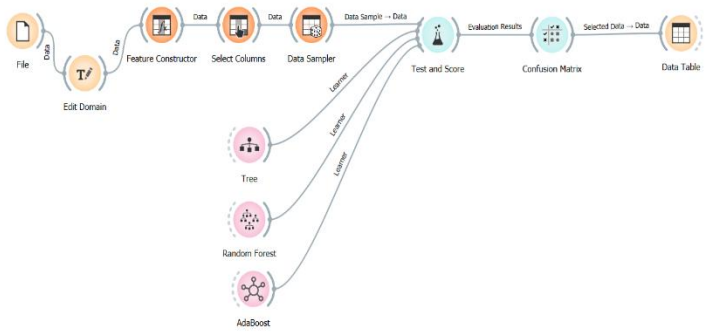
Selanjutnya, penilaian kinerja metode Naive Bayes dapat dievaluasi menggunakan Confusion Matrix. Pada ilustrasi 11 di atas dijelaskan bahwa untuk kategori weak, hasil prediksi menggunakan metode Naive Bayes memiliki 1799 prediksi yang benar dan 1021 prediksi yang salah. Sementara itu, pada kategori moderate, hasil prediksi dengan metode Naive Bayes menunjukkan 10050 prediksi yang benar dan 279 prediksi yang salah. Pada kategori strong, hasil prediksi menggunakan metode Naive Bayes mencapai 1798 prediksi yang benar dan 53 prediksi yang salah.




Gambar 10. Data Table Naïve Bayes

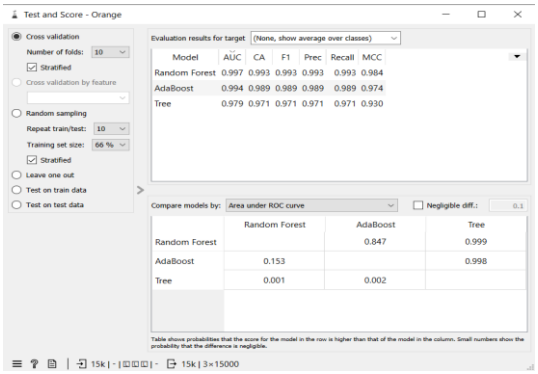
Hasil penelitian dengan Test and Score juga dapat dilihat menggunakan Data Table yang tertera pada gambar 12 di atas. Hal ini memungkinkan untuk secara manual dan dengan lebih jelas melihat nilai prediksi yang benar dan yang salah dari metode Naive Bayes.

### *B. Analisis Model Algoritma Decision Tree dan Ensemble Method*



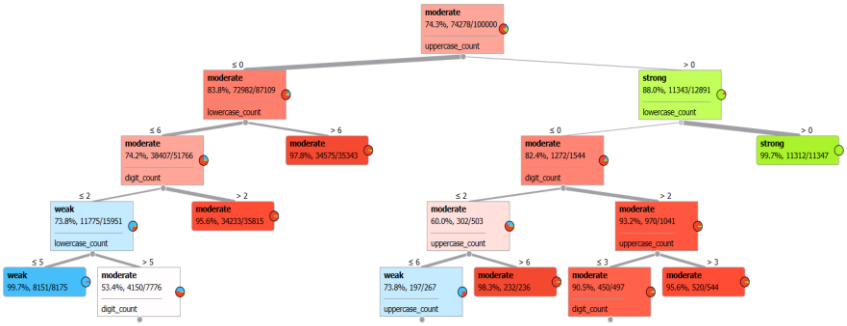

Gambar 11. Work Flow Analisa Decision Tree dan Ensemble Method

Gambar 10 di atas merupakan *work flow* yang digunakan untuk melakukan analisis prediksi menggunakan algotritma *Decision Tree* dan *Ensemble Method.* Untuk *feature* yang dipakai pada algoritma *Ensemble Method* yakni huruf kapital pada *password* dan huruf kecil pada *password*. Selanjutnya hasil *test and score* dapat dilihat pada gambar di bawah ini.




Gambar 12. Test and Score Decision Tree dan Ensemble Method

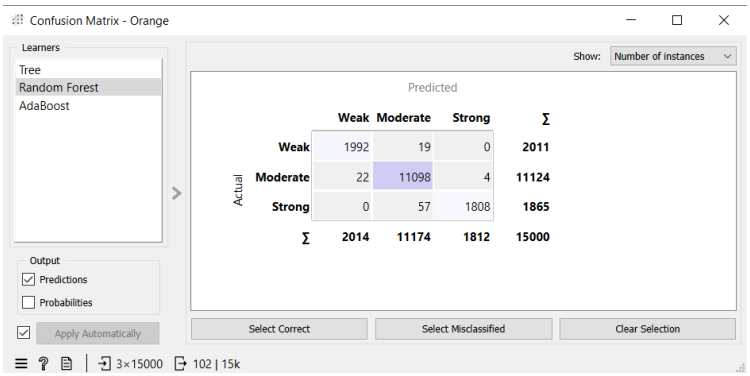
Berdasarkan gambar diatas, dari data sample yang diambil sebesar 15%, diketahui bahwa nilai AUC tertinggi berada pada model algoritma *Random Forest* dengan nilai sebesar 0,997 dan disusul oleh *AdaBoost* dengan nilai 0,994 serta *Tree* dengan nilai 0,979. Begitupun dengan nilai CA, F1, *Precision*, *Recall*, dan MCC dari model *Random Forest* yang menunjukkan bahwa model Random Forest memiliki nilai tertinggi diantara model *AdaBoost* dan *Tree*. Hal ini sekaligus membuktikan bahwa model algoritma *Random Forest* merupakan model algoritma paling baik dalam memprediksi dataset yang digunakan kelompok ini.




Gambar 23. Hasil Tree Viewer

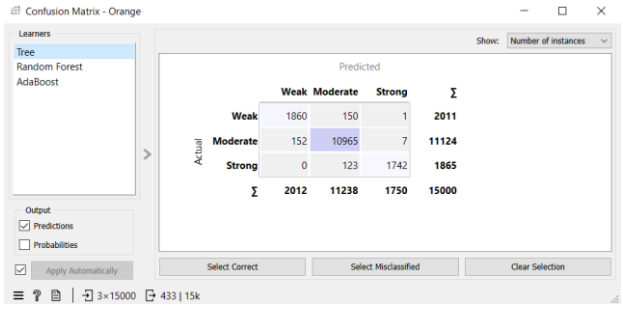
Untuk mengevaluasi hasil dari Tree, dapat digunakan Tree Viewer. Pada ilustrasi, dijelaskan bahwa variabel-variabel terbagi ke dalam beberapa klasifikasi, dengan penjelasan yang terdiri dari berbagai cabang dalam gambar tersebut. Cabang utama menjelaskan mengenai jumlah huruf kapital yang digunakan pada suatu password yang kemudian dibagi ke dalam dua cabang yaitu password yang memiliki huruf kapital lebih dari 0 dan password yang memiliki huruf kapital lebih sedikit atau sama dengan 0. Setelah itu, setiap cabang kembali dibagi berdasarkan

jumlah digit angka yang terdapat dalam kata sandi, dan proses pembagian ini berlanjut hingga mencapai cabang paling akhir yang berfungsi sebagai simpulan atau kesimpulan dari kompleksitas kata sandi tersebut.

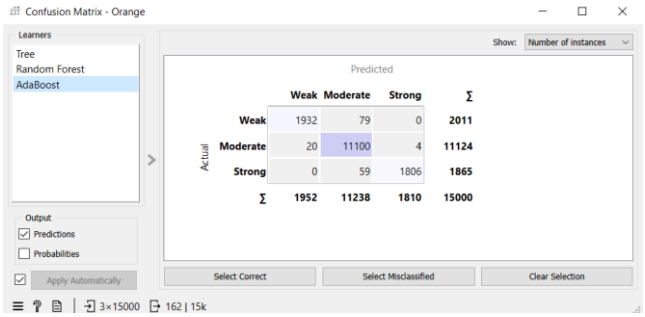



Gambar 14. Confusion Matrix Random Forest

Setelah diketahui *Random Forest* merupaka model algoritma terbaik untuk memprediksi, langkah selanjutnya yakni melihat *Confussion Matrix* dari *Random Forest* untuk mengetahui jumlah pasti dari prediksi model algoritma *Random Forest* mengenai dataset yang digunakan kelompok ini. Diketahui bahwa dari masing-masing total *target* dalam dataset ini, hampir keluruhan prediksi *Random Forest* dinyatakan benar. Sebagai contoh dari total aktual *target* "Weak" sejumlah 2.011, diketahui bahwa tingkat prediksi model *Random Forest* yang benar mencapai 1.992, sedangkan sisanya kesalahan prediksi sebesar 19.
Selain itu, untuk mendukung pernyataan *Random Forest* merupakan model algoritma terbaik dalam memprediksi dataset ini, diperlihatkan juga *Confussion Matrix* dari *AdaBoost* dan juga *Tree*.



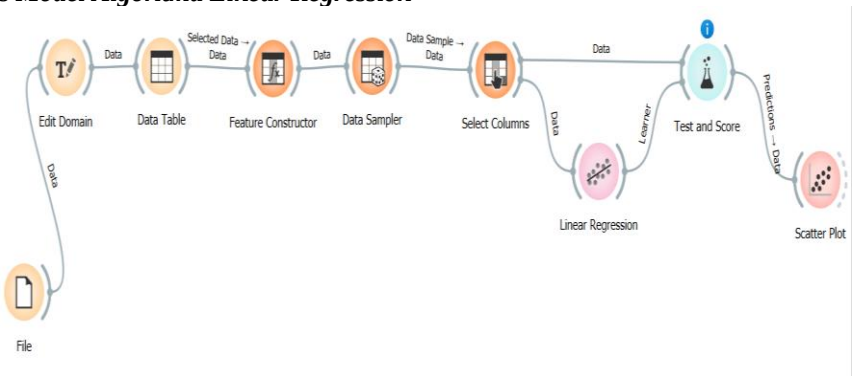

Gambar 15. Confusion Matrix Tree




Gambar 36. Confusion Matrix AdaBoost

Dapat dilihat bahwa dari kedua *Confussion Matrix*, kedua model tersebut menunjukkan bahwa tingkat kesalahan prediksi dari model *Tree* dan *AdaBoost* lebih besar dari model *Random Forest*. Sebagai contoh untuk total aktual *target* "Weak" sejumlah 2.011, untuk *Tree* memiliki prediksi yang salah sebesar 150, dan untuk *AdaBoost* memiliki prediksi yang salah sebesar 79. Dengan demikian, hasil dari *Confussion Matrix* dari masing-masing model mendukung pernyataan bahwa model algoritma *Random Forest* merupakan model dengan pengolahan data dan prediksi terbaik dalam *Ensemble Method* dengan nilai *Test Score* tertinggi dibanding kedua model lainnya dan hasil *Confussion Matrix* yang menunjukkan tingkat kebenaran dari prediksi yang paling besar dibanding kedua model lainnya.

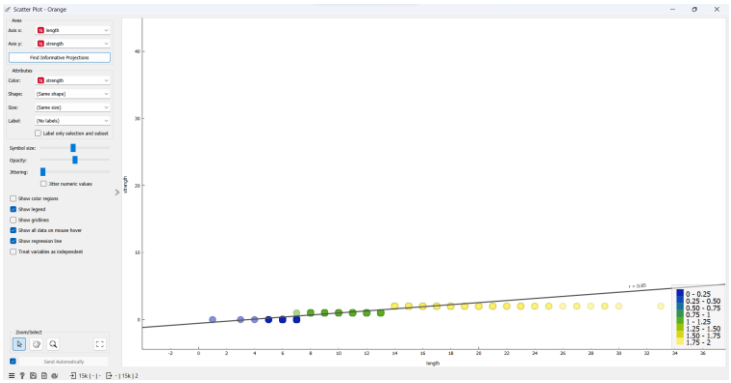
### *D. Analisis Model Algoritma Linear Regression*




Gambar 17. Work Flow Analisa Linear Regression

Gambar 16 di atas merupakan *work flow* yang digunakan untuk melakukan analisis prediksi menggunakan algotritma *Linear Regression.* Untuk *feature* yang dipakai pada algoritma *Linear Regression* yakni huruf kapital pada *password* dan huruf kecil pada *password.* Selanjutnya hasil *test and score* dapat dilihat pada gambar di bawah ini.

Berdasarkan hasil analisis menggunakan metode linear regression, didapatkan hasil prediksi model yang dapat dilihat melalui test and score seperti pada gambar di atas. Hasil tersebut menunjukkan bahwa nilai MSE sebesar 0,073. Hal ini berarti bahwa model yang diprediksi seperti variabel x yaitu length (jumlah password) terhadap variabel y yaitu strength (kekuatan password) akan terdapat kesalahan sebesar 7,3%. Adapun, untuk nilai R-Square ($R^2$) yang didapat adalah sebesar 0,716. Hal ini menandakan bahwa model yang digunakan dalam memprediksi variabel y yaitu kekuatan password (strength) melalui variabel x yaitu panjang password (length) dapat dijelaskan sebesar 71,6%. Sedangkan, sebesar 28,4% sisanya dipengaruhi oleh faktor lain yang tidak digunakan dalam model.



Gambar 18. Scatter Plot Linear Regression

Selain itu, berdasarkan analisis menggunakan scatter plot dapat diketahui bahwa variabel panjang password (length) berpengaruh positif terhadap kekuatan password (strength). Berdasarkan gambar 18 di atas, dapat dilihat bagaimana semakin panjang password yang dibuat, maka semakin kuat juga keamanan passwordnya. Sebagai contoh, panjang password paling rendah adalah satu, maka kekuatan password berada di kategori 0 (lemah). Sedangkan, untuk panjang password terpanjang berjumlah 33 karakter dan masuk ke dalam kategori 2 (kuat). Evaluasi melalui scatter plot ini juga dapat memberikan informasi bahwa titik yang berada semakin jauh dari garis linear regression, maka nilai prediksi terdapat kesalahan.

## KESIMPULAN

Berdasarkan analisis prediksi password dan kompleksitasnya menggunakan K-Nearest Neighbor, Naïve Bayes, Decision Tree, Ensemble Methods, dan Linear Regression, dapat disimpulkan bahwa model-model tersebut mampu menghasilkan nilai Test and Score atau prediksi dengan tingkat akurasi dan ketepatan yang tinggi. Algoritma model yang mendapatkan nilai paling tinggi yaitu Random Forest, yang kemudian diikuti Naive Bayes, AdaBoost, Tree, dan k-NN. Dengan begitu, Random Forest dianggap dapat sebagai model algoritma yang paling akurat dalam penelitian ini. Sedangkan untuk algoritma atau model Linear Regression tidak bisa digunakan dalam dataset ini

## REFERENSI

https://www.kaggle.com/datasets/morph1max/password-security-sber-dataset/data

Giri, G. A. (2018). Klasifikasi Musik Berdasarkan Genre dengan Metode K-Nearest Neighbor. *Jurnal Ilmu Komputer VOL. XI No. 2*.

Hadi, H., Agustina, S., & Subhani, A. ( Vol 3, No 1 (2019) ). Penguatan Kesiapsiagaan Stakeholder dalam Pengurangan Risiko Bencana Alam Gempabumi. *Jurnal Kajian Ilmu dan Pendidikan Geografi*.

Hozairi, H., Anwari, A., & Alim, S. ((2021),Vol 6, No 2 ). Implementasi Orange Data Mining Untuk Klasifikasi Kelulusan Mahasiswa Dengan Model K-Nearest Neighbor, Decision Tree Serta Naive Bayes. *Nero (Networking Engineering Research Operation)*.

(n.d.). *https://www.kaggle.com/datasets/morph1max/password-security-sber-dataset/data.* www.kaggle.com.

Indrapras, K., Alfaris, M., Falentina, & Triana, A. (2022). Analisis Big Data dan Official Statistics dalam Melakukan Nowcasting Pertumbuhan Ekonomi Indonesia Sebelum dan Selama Pandemi COVID-19. *Seminar Nasional Official Statistics 2022, https://prosiding.stis.ac.id/index.php/semnasoffstat*.

Indriyanti, I., Ichsan, N., Fatah, H., Wahyuni, T., & Ermawati, E. (Aug 30, 2022). Implementasi Orange Data Mining Untuk Prediksi Harga Bitcoin. *Jurnal Responsif : Riset Sains dan Informatika*, 118~125.

Indriyawati, H., & Khoirudin. (2019-12-05). Penerapan Metode Regresi Linier Dalam Koherensi Pengolahan Data Bahan Baku Tiandra Store Guna Meningkatkan Mutu Produksi. *Sintak PORSIDING,https://www.unisbank.ac.id/ojs/index.php/sintak/article/view/7603*.

Karo, G. E., Erwansyah, K., & Suharsil. (2020). Implementasi Data Mining Dalam Mengestimasi Pendapatan Pada Pt Citosarana Jasa. *Jurnal Cyber Tech*.

Laksono, R. A., Achmadi, S., & Sasmito, A. P. (2023). Implementasi Data Mining Menggunakan Metode Least Square Untuk Memprediksi Jumlah Pendapatan. *Jati (Jurnal Mahasiswa Teknik Informatika). Vol. 7 No. 5, Oktober 2023*.

Lemantara, L. (2022). Penerapan Algoritma Naïve Bayes dan ID3 untuk Memprediksi Segmentasi Pelanggan pada Penjualan Mobil. *Journal of Technology and Informatics (JoTI)*, Vol. 4, No. 1, Bulan Oktober Tahun 2022.

Mardiani, E., & Ramadhan, F. A. (2023). Design Information System Sales of Nuts and Bolts at PT. Catur Naga Steelindo. *Jurnal SITEKIN: Jurnal Sains, Teknologi dan Industri , Vol. 20, No. 2, June 2023*, .729 – 735.

Mardiani, E., Rahmansyah, N., Ningsih, S., Lantana, D. A., Wirawan, A. S., Wijaya, S. A., & Putri, D. N. (2023). Komparasi Metode KNN, Naive Bayes. Decision Tree, Ensemble, Linear Regression Terhadap Analisis Performa Pelajar Sma. *Jurnal INNOVATIVE: Journal Of Social Science Research*, 13880-13892.

Marutho, D. (2019). Perbandingan Metode Naïve Bayes, Knn, Decision Tree Pada Laporan Water Level Jakarta. *Jurnal Ilmiah Infokam*, Vol 15, No 2.

Pratama, F. H., Triayudi, A., & Mardiani, E. (2022, Vol 7, No 4). Data Mining K-Medoids Dan K-Means Untuk Pengelompokan Potensi Produksi Kelapa Sawit Di Indonesia. *Jurnal Ilmiah Penelitian dan Pembelajaran Informatika(JIPI)*.

Priyanti, E. (2019). Priyanti, Evy(2019). Komparasi Klasifikasi Pada Prediksi Pendapatan Rumah Tangga. Jurnal Swabumi Vol.7 No.2 September 2019, pp.114~121. *Jurnal Swabumi Vol.7 No.2 September 2019, pp.114~121*, 114-121.

Rahmansyah, N., Mulyani, D., Mardiani, E., & Rahman, A. (2022). Perancangan Sistem Transaksi Berbasis Web Pada UKM Pangkas Rambut Tasik. *Jurnal JUNSIBI: Jurnal Sistem informasi Bisnis April 2022, 22-31*, 22-31.