SKRIPSI

PREDIKSI POTENSI DESA CERDAS MENGGUNAKAN ALGORITMA KLASIFIKASI *RANDOM FOREST* DAN *LOGISTIC REGRESSION*

Oleh:

Salma Amanda

065119196

PROGRAM STUDI ILMU KOMPUTER

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS PAKUAN

2024

SKRIPSI

PREDIKSI POTENSI DESA CERDAS MENGGUNAKAN ALGORITMA KLASIFIKASI *RANDOM FOREST* DAN *LOGISTIC REGRESSION*

Diajukan sebagai salah satu syarat untuk memperoleh Gelar Sarjana Komputer
Jurusan Ilmu Komputer Fakultas Matematika dan Ilmu Pengetahuan Alam

Oleh:

Salma Amanda

065119196

PROGRAM STUDI ILMU KOMPUTER

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS PAKUAN

2024

HALAMAN KREASI / PERSEMBAHAN SKRIPSI

"*Because love's such an old-fashioned word*
And love dares you to care for the people on the (people on streets) edge of the night
And love (People on streets) dares you to change our way of caring about ourselves
This is our last dance"

(Queen & David Bowie – *Under Pressure*)

Saya persembahkan skripsi ini untuk kedua orang tua, keluarga besar, ibu dan bapak dosen program studi Ilmu Komputer Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Pakuan Bogor,

teman-teman angkatan 19 khususnya kelas G&H dan Himpunan Mahasiswa Ilmu Komputer (Himakom)

yang telah memberikan banyak dukungan dan semangat kepada penulis sehingga bisa menyelesaikan skripsi ini

HALAMAN PENGESAHAN

Judul : Prediksi Potensi Desa Cerdas Menggunakan Algoritma Klasifikasi *Random Forest* dan *Logistic Regression*
Nama : Salma Amanda
NPM : 065119196

Mengesahkan,

Pembimbing Pendamping	Pembimbing Utama
Program Studi Ilmu Komputer	Program Studi Ilmu Komputer
FMIPA - UNPAK	FMIPA - UNPAK
Erniyati, M.Kom.	**Dr. Eneng Tita Tosida, S.T.P., M.Si., M.Kom.**

Mengetahui,

Ketua Program Studi Ilmu Komputer	Dekan
FMIPA - UNPAK	FMIPA - UNPAK
Arie Qur'ania, M.Kom.	**Asep Denih, S.Kom., M.Sc., Ph.D.**

PERNYATAAN KEASLIAN KARYA TULIS SKRIPSI

Dengan ini saya menyatakan bahwa:

Sejauh yang saya ketahui, karya tulis ini bukan merupakan karya tulis yang pernah dipublikasikan atau sudah pernah dipakai untuk mendapatkan gelar sarjana di Universitas lain, kecuali pada bagian-bagian di mana sumber informasinya dicantumkan dengan cara referensi yang semestinya.

Demikian penyataan ini saya buat dengan sebenar-benarnya. Apabila kelak dikemudian hari terdapat gugatan, penulis bersedia dikenakan sanksi sesuai dengan peraturan yang berlaku.

Bogor, Juli 2024

Salma Amanda

PERNYATAAN PELIMPAHAN SKRIPSI DAN SUMBER INFORMASI SERTA PELIMPAHAN HAK CIPTA

==========

Saya yang bertandatangan di bawah ini :

Nama	: Salma Amanda
NPM	: 065119196
Judul Skripsi	: Prediksi Potensi Desa Cerdas Menggunakan Algoritma Klasifikasi *Random Forest* dan *Logistic Regression*

Dengan ini saya menyatakan bahwa Paten dan Hak Cipta dari produk Skripsi dan Tugas Akhir di atas adalah benar karya saya dengan arahan dari komisi pembimbing dan belum diajukan dalam bentuk apapun kepada perguruan tinggi manapun.

Sumber informasi yang berasal atau dikutip dari karya yang diterbitkan maupun tidak diterbitkan dari penulis lain telah disebutkan dalam teks dan dicantumkan dalam Daftar Pustaka di bagian akhir skripsi ini.

Dengan ini saya melimpahkan Paten, hak cipta dari karya tulis saya kepada Universitas Pakuan.

Bogor, Juli 2024

Salma Amanda
NPM.065119196

RIWAYAT HIDUP

Penulis dilahirkan di Bogor pada tanggal 5 September 2001 dari pasangan Bapak Ade Hermawan dan Ibu Suci Pujiarti sebagai anak pertama dari dua bersaudara.

Penulis memulai pendidikan di Sekolah Dasar yang bertempat di SDN Depok 2, kemudian tahun 2013 menempuh pendidikan di SMPN 1 Depok. Kemudian tahun 2016 melanjutkan pendidikan di SMAN 12 Depok dan lulus pada tahun 2019.

Pada tahun 2019 penulis meneruskan pendidikan ke Universitas Pakuan Bogor, Program Studi Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan Alam.

Selama di Universitas Pakuan, penulis pernah mengikuti organisasi Himpunan Mahasiswa Ilmu Komputer (HIMAKOM) periode 2020-2021 menjabat sebagai anggota Departemen Pendidikan dan Keilmuan dan periode 2021-2022 menjabat sebagai anggota Departemen Kesekreriatan dan Wirausaha. Pada bulan Juni tahun 2024 penulis menyelesaikan penelitian dengan judul "Prediksi Potensi Desa Cerdas Menggunakan Algoritma Klasifikasi *Random Forest* dan *Logistic Regression*"

RINGKASAN

Salma Amanda. Prediksi Potensi Desa Cerdas Menggunakan Algoritma Klasifikasi Random Forest dan Logistic Regression. Dibawah bimbingan Dr. Eneng Tita Tosida, M.Si., M.Kom dan Erniyati, M.Kom.

Pertumbuhan penduduk yang semakin lama meningkat menimbulkan terjadinya migrasi dari desa ke kota, karena selama ini pembangunan cenderung berorientasi dan bias sehingga menyebabkan pembangunan di desa menjadi terhambat. Hal tersebut lah yang menciptakan ketimpangan antara desa dengan kota. Salah satu cara untuk mengurangi ketimpangan antara desa dan kota yaitu dengan adanya pembangunan desa cerdas atau *smart village*. Oleh karena itu, dilakukan prediksi potensi desa cerdas dengan membandingkan dua algoritma klasifikasi, yaitu *random forest* dan *logistic regression* untuk mencari tahu mana algoritma yang kinerjanya lebih baik. Data penelitian ini bersumber dari Badan Pusat Statistik tahun 2021 yang kemudian diambil beberapa jumlah variasi data, yaitu 1500, 1550, 1600, dan 1650 yang diambil secara acak dan proposional. Evaluasi pada penelitian ini menggunakan *confusion matrix* dengan mengambil nilai *precision* yang paling tinggi sebagai pengambilan kesimpulan kinerja algoritma. Dari beberapa percobaan pembagian dataset 90:10, 85:15, 80:20, dan 70:30, pembagian dataset 85:15 dengan dataset 1500 acak menghasilkan nilai precision yang paling tinggi baik dari algoritma *random forest* maupun *logistic regression*, yaitu berturut-turut sebesar 92.37% dan 92.30%. Karena algoritma *random forest* lebih unggul daripada algoritma *logistic regression*, maka hasil prediksi yang digunakan adalah prediksi *random forest*.

Kata Kunci : Desa Cerdas, Potensi Desa, Random Forest, Logistic Regression

KATA PENGANTAR

Puji syukur kehadirat Allah SWT, karena rahmat dan hidayah- Nya penulis dapat menyelesaikan skripsi ini yang berjudul: **"Prediksi Potensi Desa Cerdas Menggunakan Algoritma Klasifikasi *Random Forest* dan *Logistic Regression*"**. Penulisan tugas akhir ini merupakan salah satu syarat memperoleh gelar Sarjana Komputer di Program Studi Ilmu Komputer FMIPA UNPAK Bogor.

Dalam penulisan tugas akhir ini, penulis dengan senang hati ingin mengucapkan terima kasih yang sebesar-besarnya kepada:

1. Dr. Eneng Tita Tosida, M.Si., M.Kom., selaku Pembimbing Utama yang telah memberikan dorongan moril dan motivasi kepada penulis.
2. Erniyati, M.Kom, selaku pembimbing Pendamping yang telah memberikan bimbingan, semangat dan motivasi.
3. Arie Qurania, M.Kom, selaku Ketua Program Studi Ilmu Komputer FMIPA Universitas Pakuan Bogor.
4. Kedua orang tua, yang tidak pernah lelah untuk mendoakan dan mendukung saya selama ini.
5. Seluruh dosen Program Studi Ilmu Komputer, yang telah memberikan banyak ilmu selama perkuliahan.
6. Angel, Nabila, Dita, Raja, Wianda, Virnanda, dan Sony yang selalu memberikan saya semangat, hiburan, dan bantuan.
7. Seluruh pihak yang tidak dapat penulis sebutkan satu persatu yang telah memberikan segala dukungan, semangat, bantuan secara langsung maupun tidak langsung.

Saran dan kritik yang membangun dalam penulisan tugas akhir ini akan diterima dengan senang hati. Mudah-mudahan Allah SWT akan membalas semua kebaikan kepada semua pihak yang membantu. Akhir kata, semoga laporan ini dapat bermanfaat bagi kita semua. Aamiin.

Bogor, 25 Juli 2024

Salma Amanda

065119196

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

DAFTAR LAMPIRAN

BAB I
PENDAHULUAN

1.1 Latar Belakang

Indonesia adalah negara kepulauan dengan jumlah desa lebih banyak daripada kota. Menurut Kepmendagri No. 050-145 tahun 2022, terdapat 34 provinsi, 416 kabupaten, 98 kota, 7.266 kecamatan, 8.506 kelurahan, 74.961 desa, dan 16.772 pulau. Namun, pertumbuhan penduduk yang semakin lama meningkat malah menimbulkan terjadinya migrasi dari desa ke kota, karena kebijakan pembangunan sering kali memberikan prioritas pada daerah perkotaan, sementara perhatian yang lebih sedikit diberikan pada daerah pedesaan (Santoso *et al*, 2019) sehingga menyebabkan pembangunan di desa menjadi terhambat. Hal tersebut lah yang menciptakan ketimpangan antara desa dengan kota. Adapapun hal lainnya yaitu masyarakat pedesaan terus dirundung masalah seperti kemiskinan, keterbelakangan, dan ketidakamanan sosial-ekonomi lainnya (Santoso *et al*, 2019). Pada BPS (2022) menunjukkan bahwa persentase penduduk miskin di daerah perdesaan cukup tinggi yaitu sebesar 12,36% dari total jumlah penduduk Indonesia, sementara di daerah perkotaan sebesar 7,53%. Kemiskinan juga termasuk faktor pemicu yang mendorong masyarakat pedesaan untuk mengadu nasib di kota karena desa dianggap tidak memberikan mata pencaharian yang memadai.

Kementerian Desa, Pembangunan Daerah Tertinggal dan Transmigrasi mempunyai tugas untuk mengurasi ketimpangan tersebut dengan mengacu pada Peraturan Menteri Desa, Pembangunan Daerah Tertinggal, dan Transmigrasi Republik Indonesia Nomor 6 Tahun 2015 Pasal 2, yaitu “Kementerian Desa, Pembangunan Daerah Tertinggal, dan Transmigrasi mempunyai tugas menyelenggarakan urusan pemerintahan di bidang pembangunan desa dan kawasan perdesaan, pemberdayaan masyarakat desa, percepatan pembangunan daerah tertinggal, dan transmigrasi untuk membantu Presiden dalam menyelenggarakan pemerintahan negara.” Selain kebijakan tersebut, ada salah satu cara untuk mengurangi ketimpangan antara desa dan kota, yaitu dengan adanya pembangunan desa cerdas atau *smart village*.

Konsep desa cerdas sendiri mengambil gagasan dari *smart city*, tetapi dirancang agar sesuai dengan kondisi dan masalah di daerah pedesaan (Santoso *et al*, 2019). Program desa cerdas diharapkan bisa meningkatkan kesejahteraan masyarakat desa dengan meningkatkan kualitas infrastruktur, layanan kesehatan, sektor pendidikan, dan pertumbuhan ekonomi yang berkelanjutan. Akan tetapi, pembangunan desa cerdas juga tetap harus disesuaikan dengan keadaan, potensi, keperluan, dan masalah yang dihadapi oleh setiap desa di Indonesia.

Penelitian mengenai desa cerdas pernah dilakukan oleh (Tosida *et al*., 2021) yang berjudul "Klasifikasi Potensi Desa Cerdas menggunakan *Deep Learning*". Dalam penelitian ini, data potensi desa diklasifikasikan menjadi lima kelompok desa untuk mengidentifikasi potensi desa cerdas, yakni: tidak berpotensi untuk desa cerdas berjumlah 11.032 desa, kurang berpotensi untuk desa cerdas berjumlah 433 desa, cukup berpotensi untuk desa cerdas berjumlah 217 desa, berpotensi untuk desa cerdas berjumlah 153 desa, dan sangat berpotensi untuk desa cerdas berjumlah 204 desa, dengan nilai *Accuration* sebesar 0,9996.

Penelitian yang berkaitan dengan desa cerdas juga dilakukan oleh (Tosida *et al*., 2020) dengan judul "*Clustering of Citizen Science Prospect to Construct Big Data-based Smart Village in Indonesia*" dengan menggunakan kombinasi algoritma k-means, expected maximum k-means, expected maximum, dan algoritma berbasis densitas. Hasil penelitian ini terbagi menjadi menunjukkan bahwa terdapat 3 cluster, yaitu cluster sangat potensial (11%) yang meliputi Provinsi Kepulauan Bangka Belitung, Jawa Barat, Jawa Tengah, Jawa Timur, Yogyakarta, Banten dan Bali, lalu cluster potensial (60%) dan cluster cukup potensial (29%).

Random Forest merupakan algoritma *machine learning* yang telah membuktikan keefektifannya dalam masalah regresi dan klasifikasi dalam beberapa tahun terakhir dan merupakan salah satu algoritma *machine learning* terbaik yang digunakan di berbagai bidang (Apriliah *et al*., 2021), dapat menghasilkan *Accuration* yang lebih tinggi dan mengatasi data dalam jumlah yang besar secara efisien (Nilwanda *et al*., 2024), serta algoritma ini juga efektif baik pada data diskrit maupun kontinu (Huda, 2023). Terdapat salah satu algoritma klasifikasi dari metode statistika yang mampu menangani kondisi serupa *Random Forest*, yaitu *Logistic Regression* yang merupakan model yang sederhana untuk diterapkan dan dapat memberikan prediksi yang baik (Handayani *et al*, 2021), menggunakan respon biner dan prediktor yang dapat terdiri dari data kontinu, diskrit, atau kategori dan dapat digunakan untuk berbagai skala data (Cahyani *et al*., 2022) (Phylosta & Febryansyah, 2022).

Perbandingan antara algoritma *Random Forest* dan *Logistic Regression* pernah dilakukan oleh (Cahyana & Nurlayli, 2023) menghasilkan *accuration* sebesar 75% untuk *Random Forest*, 75% *decision gree*, dan 80% *Logistic Regression* untuk mendeteksi kanker payudara. (Peerbasha *et al*, 2023) menyimpulkan bahwa *Random Forest* lebih unggul dengan *accuration* sebesar 99% dan *Logistic Regression* sebesar 80.66% untuk memprediksi diabetes. Gripsy & Divya, (2023) menyimpulkan *Logistic Regression* menghasilkan *accuration* lebih tinggi (91.90%) dan *Random Forest* (90.93%) dalam memprediksi penyakit paru-paru. Lalu, pada penelitian (Prasetyo *et al*, 2021) didapatkan hasil *Random Forest* mempunyai *accuration* tertinggi sebesar 88,75% sedangkan *Logistic Regression* sebesar 88,35% dalam memprediksi cacat *software*.

Berdasarkan permasalahan di atas, penulis mengajukan judul "Prediksi Potensi Desa Cerdas Menggunakan Algoritma Klasifikasi *Random Forest* dan *Logistic Regression*" untuk menentukan algoritma mana yang lebih ungul dan sesuai dengan kebutuhan penelitian.

1.2 Tujuan Penelitian

Penelitian ini bertujuan untuk membandingkan algoritma *Random Forest* dan *Logistic Regression* untuk mengetahui algoritma mana yang menghasilkan kinerja lebih baik untuk memprediksi potensi desa cerdas.

1.3 Ruang Lingkup Penelitian

Ruang lingkup penelitian sebagai dasar batasan penelitian diuraikan seperti berikut ini:

1. Data yang digunakan bersumber dari data potensi desa tahun 2021.
2. Data potensi desa tersebut akan diolah dalam model prediksi menggunakan algoritma *Random Forest* dan *Logistic Regression* dengan dua kelas target, yaitu "Tidak Berpotensi" dan "Berpotensi".

3. Penelitian ini akan menghasilkan perbandingan kinerja antara algoritma *Random Forest* dan *Logistic Regression* dengan membandingkan nilai *Precision*.
4. Bahasa pemrograman yang digunakan dalam penelitian ini yaitu *Python* dengan tools *Google Colab*.

1.4 Manfaat Penelitian

Manfaat yang diharapkan dari penelitian ini adalah:

1. Mengetahui cara menerapkan algoritma *Random Forest* dan *Logistic Regression* untuk memprediksi potensi desa.
2. Mengetahui algoritma klasifikasi mana yang kinerjanya lebih baik.
3. Menghasilkan informasi kepada pemerintah khususnya Kementerian Desa, Pembangunan Daerah Tertinggal, dan Transmigrasi agar dapat mengarahkan sumber daya dan kebijakan secara lebih tepat guna untuk mendukung perkembangan pedesaan yang lebih merata.
4. Membantu para peneliti lainnya dalam mengembangkan topik terkait prediksi potensi desa menggunakan algoritma klasifikasi.

BAB II
TINJAUAN PUSTAKA

2.1 Landasan Teori

2.1.1 Desa

Menurut (Martadala *et al*, 2021), desa adalah kesatuan masyarakat hukum yang berdasarkan atas kesatuan budaya dan adat istiadat, yang berada dalam suatu wilayah tertentu, memiliki ikatan lahir batin yang sangat kuat, memiliki kepentingan politik, ekonomi, sosial, dan keamanan yang sama, serta memiliki susunan pemerintahan yang dipilih secara bersama-sama.

Adapun definisi desa berdasarkan Peraturan Menteri Desa Pembangunan Daerah Tertinggal dan Transmigrasi Republik Indonesia Nomor 2 Tahun 2016 yaitu, "Desa adalah desa dan desa adat atau yang disebut dengan nama lain yang selanjutnya disebut desa adalah kesatuan masyarakat hukum yang memiliki batas wilayah yang berwenang untuk mengatur dan mengurus urusan pemerintahan, kepentingan masyarakat setempat berdasarkan prakarsa masyarakat, hak asal usul, dan/atau hak tradisional yang diakui dan dihormati dalam sistem pemerintahan Negara Kesatuan Republik Indonesia."

2.1.2 Potensi Desa

Sesuai dengan Peraturan Menteri Desa Pembangunan Daerah Tertinggal dan Transmigrasi Republik Indonesia Nomor 2 Tahun 2016 tentang potensi desa, "Potensi desa atau disingkat podes, adalah sumber daya sosial, ekonomi dan ekologi yang terdapat di Desa, yang dapat dikembangkan untuk meningkatkan kesejahteraan masyarakat desa."

2.1.3 Desa Cerdas

Runanto *et al*, (2021) menjelaskan bahwa desa cerdas adalah komunitas dan daerah pedesaan yang mengembangkan kekuatan dan aset mereka sambil berusaha mengembangkan peluang baru di mana penggunaan pengetahuan yang lebih baik, teknologi digital, telekomunikasi, dan inovasi meningkatkan jaringan dan layanan tradisional dan baru.

Desa cerdas pun dapat diartikan sebagai desa yang dapat memajukan kesejahteran masyarakat dan kualitas hidup masyarakat dengan penggunaan teknologi informasi (Prasetya *et al*, 2022). Dalam penerapan konsep desa cerdas terdapat 6 pilar yaitu, Masyarakat Cerdas, Ekonomi Cerdas, Tata Kelola Cerdas, Lingkungan Cerdas, Kehidupan Cerdas, dan Mobilitas Cerdas.

2.1.4 *Proportionate Stratified Random Sampling*

Proportionate stratified random sampling adalah metode yang digunakan pada situasi di mana populasi mengandung elemen atau anggota yang tidak homogen dan berstrata secara proporsional. Berikut adalah persamaannya (Machali, 2021):

$$Strata = \frac{\text{Jumlah Populasi Strata x Sampel}}{\text{Jumlah Populasi}} \quad (1)$$

2.1.5 *Data mining*

Proses menemukan pola dan informasi yang berguna dari kumpulan data yang besar dikenal sebagai data mining. Tujuan utama dari data mining adalah untuk mengenali pola, hubungan, dan pengetahuan tersembunyi yang dapat digunakan untuk pengambilan keputusan. (Sundari *et al*, 2023).

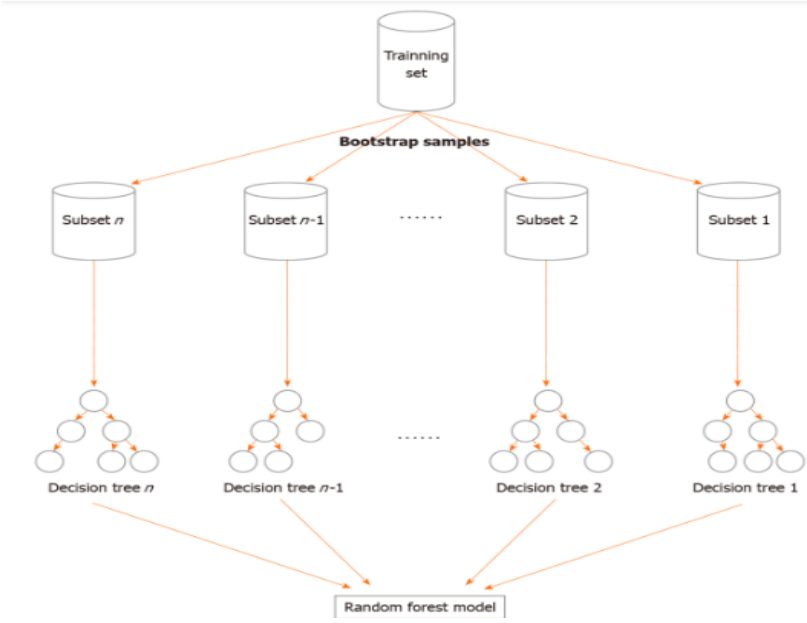
Ada juga banyak istilah lain yang memiliki arti yang sama dengan data mining, misalnya, *knowledge mining from data*, *knowledge extraction*, *data/pattern analysis*, *data archaeology*, dan *data dredging*. Istilah lain yang digunakan oleh banyak orang adalah *Knowledge Discovery in Database* atau KDD. Kedua istilah tersebut memiliki konsep yang berbeda, namun masih ada keterkaitan di antara keduanya. (Syahranita, 2022).

2.1.6 *Hyperparameter Tuning*

Hyperparameter Tuning merupakan proses untuk menemukan hyperparameter yang tepat, karena kinerja model dipengaruhi oleh nilai parameter. Salah satu metode yang paling umum digunakan dalam *Hyperparameter Tuning* adalah *GridSearch*. (Rizky Mubarok *et al*., 2022).

2.1.7 *Random Forest*

Random Forest dicetuskan oleh Breiman pada tahun 2001. *Random Forest* menggabungkan hasil dari beberapa pohon keputusan untuk memperoleh pengetahuan yang lebih kuat dengan memanfaatkan dua konsep utama, yaitu bagging dan random feature *selection* (Yan & Shen, 2022) yang merupakan pengembangan dari metode Classification and Regression Tree (CART) (Lumbanraja *et al*, 2019).



Gambar 1. Tahapan *Random Forest*

Tahap - tahap pembangunan *Random Forest* menurut (Breiman, 2001) adalah sebagai berikut:

1. Ambil dataset secara acak sebanyak n (jumlah total data) dari *training dataset* secara bergantian, sehingga sehingga masing-masing dataset memiliki data yang sama.
2. Buatlah pohon keputusan dari dataset acak dengan menggunakan algoritma CART (*Classification and Regression Trees*) dan menggunakan aturan *splitting Gini Index* atau *Entropy*. Pohon keputusan ini nantinya akan menjadi sebuah *decision tree*.

$$Gini\ (A) = 1 - \sum_{i=1}^{n} P_i^2 \tag{2}$$

$$Entropy\ (A) = - \sum_{i=1}^{n} P_i log_2(Pi) \tag{3}$$

Keterangan:

n : Jumlah kelas target

Pi : Proporsi jumlah sampel kelas i terhadap jumlah total sampel

3. Ulangi tahap 1 dan 2 sebanyak M kali untuk membuat M *decision tree*. Setiap *decision tree* dibangun dengan menggunakan dataset yang diacak, yang di mana setiap model yang dibangun akan menggunakan variasi dataset yang berbeda.
4. Gunakan setiap pohon keputusan untuk mengklasifikasikan setiap observasi dalam dataset. Proses klasifikasi ini akan menghasilkan M prediksi untuk setiap observasi.
5. Untuk menghasilkan satu prediksi akhir, kombinasikan hasil prediksi dari pohon keputusan M. Jika menggunakan regresi, gunakan prediksi rata-rata dari pohon keputusan M, tetapi jika menggunakan klasifikasi, gunakan metode *voting atau weighted voting* untuk menentukan hasil akhir.

2.1.7.1 *Variable Importance*

Cara sederhana untuk menilai pentingnya *variable importance* adalah dengan menghitung seberapa sering variabel tersebut muncul dalam kumpulan pohon keputusan. (Christy & Suryowati, 2021).

2.1.8 *Logistic Regression*

Logistic Regression dapat diartikan sebagai sebuah metode atau algoritma yang digunakan untuk klasifikasi, di mana ia menggabungkan variabel respons dan variabel prediksi, dan menghasilkan tingkat probabilitas sebagai hasilnya (Putra *et al*, 2023), probabilitas antara 0 dan 1 mengindikasikan apakah suatu peristiwa akan terjadi atau tidak. (Handayani *et al*, 2021).

Persamaan dari *Logistic Regression* dengan *k* variabel independen adalah sebagai berikut (Hosmer & Lemeshow, 2000):

$$\pi(x_i) = \frac{e^{(\beta_0+\beta_1 x_1+\cdots+\beta_i x_i)}}{1+e^{(\beta_0+\beta_1 x_1+\cdots+\beta_i x_i)}} \quad (4)$$

Dimana:

$\pi(x_i)$: Peluang kejadian dengan nilai probabilitas $0 \leq \pi(x_i) \leq 1$
β_0: Konstanta
X_i: Variabel bebas ke-i
β_1: Koefisien dari variabel bebas ke-i

2.1.8.1 Estimasi Parameter

Maximum likelihood estimation (MLE) merupakan metode untuk menentukan nilai parameter dalam model *logistic regression*. Metode MLE dipilih karena pada kasus regresi linier, penerapan metode kuadrat terkecil menghasilkan estimasi parameter yang tidak memiliki varians minimum dan kualitas statistik yang tidak bias. (Umaroh, 2020).

Adapun menurut (Hosmer & Lemeshow, 2000: 8-9), fungsi *likelihood* yang diperoleh adalah fungsi probabilitas observasi untuk mendapatkan estimasi parameter yang tidak diketahui. Hasil estimasi parameter yang tidak diketahui merupakan hasil dari memaksimumkan nilai *likelihood* dari data observasi. Metode MLE mengestimasi koefisien atau parameter β dengan memaksimumkan fungsi *likelihood*.

2.1.8.2 Uji Parsial

Pengujian parsial adalah tahap di mana setiap variabel independen diuji secara terpisah untuk menilai apakah setiap variabel tersebut memiliki nilai yang signifikan untuk dimasukkan ke dalam model. Adapun hipotesis yang diuji (Zaen, 2019):

1. H_0: $\beta_j = 0$ tidak terdapat pengaruh antara variabel independen ke-p dengan variabel dependen.
2. H_0: $\beta_j \neq 0$ terdapat pengaruh antara variabel independen ke-p terhadap variabel dependen.

Uji Parsial yang akan digunakan yaitu uji Wald, berikut persamaannya (Harlan, 2018):

$$W = \left[\frac{\beta_j}{SE(\beta_j)}\right] \quad (5)$$

Keterangan:
W = Uji Wald
β_j = Estimasi Parameter
SE = Standar Error
Dengan kriteria penolakan (tolak H_0) $|W_i| > Z_{\alpha/2}$ atau sig. < α.

2.1.8.3 *Odds Ratio*

Kecenderungan antara satu kategori dengan kategori lainnya untuk variabel prediktor kualitatif dikenal sebagai rasio odds. Rasio odds menunjukkan seberapa mungkin suatu peristiwa sukses terjadi pada suatu kelompok dibandingkan dengan kelompok lainnya. Nilai rasio odds untuk setiap variabel prediktor dalam model regresi logistik dengan p prediktor dapat dihitung sebagai berikut (Harlan, 2018):

(6)

$$OR = e^{\beta j}$$

2.1.9 *Confusion Matrix*

Matriks konfusi atau *Confusion Matrix* merupakan alat ukur yang digunakan untuk mengevaluasi hasil akhir dari algoritma klasifikasi. *Confusion Matrix* ini menggambarkan klasifikasi "*confusion*" atau biasa disebut *classfication error*, yang dimana menampilkan kelas aktual pada setiap barisnya dan menampilkan kelas prediksi di setiap kolomnya (Christy & Suryowati, 2021).

Tabel 1. *Confusion Matrix*

Actual Class	***Predicted Class***	
	Positive	***Negative***
Positive	*True Positive* (TP)	*False Negative* (FN)
Negative	*False Positive* (FP)	*True Negative* (TN)

1. *True Positive* (TP): kelas yang diprediksi positif dan pada kenyataannya positif.
2. *True Negative* (TN): kelas yang diprediksi negatif dan pada kenyataannya negatif.
3. *False Positive* (FP): kelas yang diprediksi positif namun pada kenyataannya negatif.
4. *False Negative* (FN): kelas yang diprediksi negatif namun pada kenyataannya positif.

Terdapat berbagai macam ukuran dalam mengevaluasi kinerja model berdasarkan *Confusion Matrix*, diantaranya yaitu:

1. *Accuration*, merupakan nilai yang menunjukkan tingkat kedekatan antara nilai prediksi dengan nilai aktual (Pradana *et al*., 2022).
$$Akurasi = \frac{TP + TN}{TP + FP + TN + FN} \quad (7)$$
2. *Recall*, merupakan nilai yang menunjukkan hasil dari jumlah data yang diprediksi sebagai positif kenyataannya positif dibagi dengan keseluruhan prediksi yang kenyataannya positif (Marlina & Bakri, 2021).
$$Recall = \frac{TP}{TP + FN} \quad (8)$$
3. *Specificity*, merupakan nilai yang menunjukkan proporsi data yang kenyataannya negatif benar-benar diindentifikasi sebagai negatif (Pradana *et al*., 2022)
$$Specificity = \frac{TN}{TN + FP} \quad (9)$$
4. *Precision*, merupakan nilai yang menunjukkan hasil dari jumlah data yang diprediksi sebagai positif kenyataannya positif dibagi dengan keseluruhan data yang diprediksi positif (Marlina & Bakri, 2021).
$$Precision = \frac{TP}{FP + TP} \quad (10)$$
5. *F-Measure*, merupakan nilai perbandingan *mean* dari presisi dan *Recall* yang dibobotkan (Maulidah *et al*., 2020).

(11)

$$F - Measure = 2\,x\,\frac{Presisi\ x\ Recall}{Presisi + Recall}$$

2.2 Penelitian Terdahulu

1. Nama : Tosida *et al*, (2021)
   Judul : Klasifikasi Potensi Desa Cerdas Menggunakan *Deep Learning*
   Isi : Desa cerdas didefinisikan sebagai daerah dan masyarakat pedesaan yang membangun kekuatan dan aset mereka sendiri sambil berusaha mengembangkan peluang baru di mana jaringan dan layanan tradisional dan baru ditingkatkan melalui teknologi digital, telekomunikasi, inovasi, dan penggunaan pengetahuan yang lebih baik. tradisional maupun baru dan pelayananan ditingkatkan melalui teknologi digital, telekomunikasi, inovasi dan penggunaan pengetahuan yang lebih baik. Tujuan dari penelitian ini adalah untuk membuat model klasifikasi dan menerapkan teknik deep learning pada data potensi desa. Sebanyak 10 kali pengujian dilakukan dengan arsitektur jaringan syaraf tiruan yang berbeda dan didapatkan hasil klasifikasi dengan arsitektur terbaik dengan 5-4-5 hidden layer dengan nilai RMSE (Root Mean Squared Error) sebesar 0.04783, Rsquared sebesar 0, 9039 dan MAE (Mean Squared Error) 0.01324 serta memiliki nilai SSE (Sum of Squared Error) paling rendah dibandingkan dengan arsitektur lainnya yaitu 0.042, dengan nilai threshold error 0.092 dan jumlah step/epoch yang dibutuhkan 9391 serta nilai akurasi 0.9996.
2. Nama : Tosida *et al*, (2020)
   Judul : *Clustering of Citizen Science Prospect to Construct Big Data-based Smart Village in Indonesia*
   Isi : Salah satu solusi untuk mengurangi kemiskinan di pedesaan adalah dengan mengembangkan *citizen science* sebagai fondasi dari desa cerdas. Memetakan klaster prospek citizen science dalam membangun desa cerdas berbasis big data di Indonesia adalah tujuan utama dari penelitian ini. Kontribusi dari penelitian ini adalah menghasilkan peta klaster prospek *citizen science* untuk desa cerdas di 33 provinsi di Indonesia. Hasil cluster menunjukkan bahwa terdapat 3 klaster potensi citizen science untuk mengembangkan desa cerdas Indonesia, yaitu klaster sangat potensial (11%), potensial (60%) dan cukup potensial (29%). Peta klaster prospek citizen science ini divisualisasikan dalam bentuk data spasial berdasarkan provinsi di Indonesia. Provinsi Kepulauan Bangka Belitung, Jawa Barat, Jawa Tengah, Jawa Timur, Yogyakarta, Banten, dan Bali merupakan provinsi yang potensial untuk mengembangkan citizen science dalam rangka membangun desa cerdas berbasis big data. Hasil klaster ini divalidasi dengan struktur dendogram dari *Systematic Literature Review* (SLR).
3. Nama : Cahyana & Nurlayli (2023)
   Judul : Analisis Performa *Logistic Regression*, *Naïve Bayes*, dan *Random Forest* Sebagai Algoritma Pendeteksi Kanker Payudara
   Isi : Kanker payudara adalah jenis penyakit kronis yang dapat menyebabkan kematian. Tujuan dari penelitian ini adalah untuk

menentukan metode prediksi kanker payudara yang paling akurat di Coimbra, menggunakan metode *Logistic Regression, Naïve Bayes*, atau *Random Forest*. Penelitian ini diharapkan mampu membantu masyarakat dan tenaga medis dalam deteksi dini penyakit kanker payudara. Berdasarkan pengujian yang dilakukan menggunakan algoritma Logistics Regression didapatkan nilai akurasi sebesar 80%, pada algoritma Naïve Bayes mendapatkan nilai sebesar 75%, dan terakhir dengan algoritma Random Forest didapatkan nilai sebesar 75%. Dari pengujian tersebut dapat disimpulkan bahwa algoritma Logistics Regression terbukti memiliki tingkat akurasi yang paling baik dalam hal prediksi penyakit kanker payudara dibandingkan dengan kedua algoritma lainnya.

4. Nama : Peerbasha *et al*, (2023)
   Judul : *Diabetes Prediction using Decision tree, Random Forest, Support Vector Machine, KNearest Neighbors, Logistic Regression Classifier*
   Isi : Diabetes termasuk ke dalam salah satu penyakit yang paling mematikan di dunia. Contoh: Kekecewaan koroner, gangguan penglihatan, penyakit organ kemih, dan lain sebagainya. Pasien harus menyisihkan uang dan waktu mereka untuk berkonsultasi dengan dokter di rumah sakit. Namun, dengan adanya perkembangan AI, peneliti membangun sebuah model untuk mengetahui apakah pasien tersebut memiliki penyakit poligenik. Hasil yang paling penting dari penelitian ini adalah pembentukan kerangka kerja teoretis yang dapat memprediksi tingkat risiko pasien untuk terkena diabetes dengan menggunakan algoritma *Decision tree, Random Forest, Support Vector Machine, K-Nearest Neighbors*, dan *Logistic Regression*. Dari kelima algoritma tesebut, *Random Forest* mendapatkan hasil *Accuration* yang paling tinggi, yaitu sebesar 99%, *decision tree* 98,41%, *Logistic Regression* 80,66%, SVM 79,45%, dan KNN 78,85%.
5. Nama : Gripsy & Divya (2023)
   Judul : *Lung Cancer Disease Prediction and Classification based on Feature Selection method using Bayesian Network, Logistic Regression, J48, Random Forest, and Naïve Bayes Algorithms*
   Isi : Orang yang tidak pernah merokok dapat terjangkit kanker paru-paru, namun yang merokok memiliki risiko terjangkit yang lebih tinggi. Kanker paru-paru dapat mempengaruhi semua bagian sistem pernapasan, mulai di mana saja di dalam paru-paru. Tujuan dari penelitian ini yaitu untuk memprediksi kanker paru-paru menggunakan lima algoritma klasifikasi, yaitu *Bayesian Network*, *Logistic Regression*, J48, *Random Forest*, dan Naive Bayes. Data penelitian ini didapatkan dari *Kaggle* kemudian diolah menggunakan WEKA. Dari hasil penelitian didapatkan bahwa *Logistic Regression* memiliki kinerja yang paling baik dengan *Accuration* sebesar 91,90%, diikuti oleh *naïve bayes* 90.29 %, *bayesian network* 88.34 %, J48 86.08 %, dan *Random Forest* 90.93 %.

6. Nama : Prasetyo *et al*, (2022)
   Judul : Komparasi Algoritma *Logistic Regression* dan *Random Forest* pada Prediksi Cacat Software
   Isi : Dalam proses membuat perangkat lunak yang berkualitas tinggi, pengujian menjadi bagian penting dari proses. Pengujian dapat dinilai dengan menggunakan ukuran dan teknik tertentu; salah satu tolak ukur kualitas perangkat lunak adalah ISO. Pada prediksi cacat perangkat lunak, kesalahan prediksi cacat perangkat lunak adalah kesalahan yang sangat mengerikan, dan hasil prediksi dapat berdampak pada perangkat lunak itu sendiri. Peneliti membandingkan kinerja Algoritma *Logistic Regression* dan *Random Forest* untuk memprediksi cacat *software*. Hasil yang didapatkan dari penelitian ini adalah algoritma *Random Forest* menghasilkan *Accuration* yang lebih tinggi dibandingkan *Logistic Regression*, yaitu berturut-turut sebesar 88,75% dan 88,35% untuk sebelum resampling, lalu 95,6% dan 91,4% untuk sesudah resampling. Dapat disimpulkan bahwa *Random Forest* dengan metode *resampling* lebih efektif pada prediksi cacat software

2.3 Tabel Perbandingan Penelitian

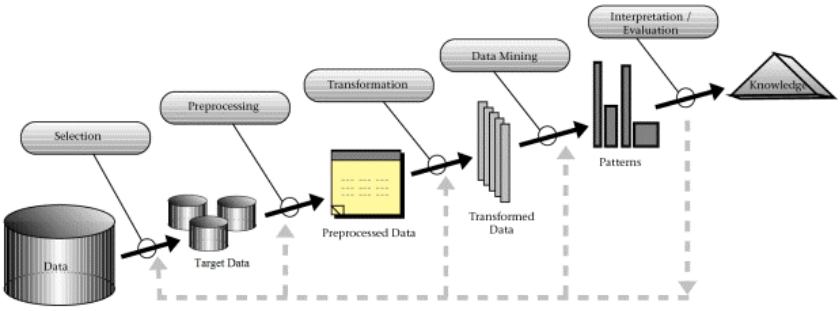
Tabel 2. Perbandingan Penelitian

No.	Nama	Algoritma			Bahasa Pemrograman			Jumlah Data
		Random Forest	***Logistic Regression***	**Lainnya**	**R**	**Python**	**Lainnya**	
1.	Tosida *et al*, (2021)			✓	✓			62.487 data
2.	Tosida *et al*, (2020)			✓		✓		62.847 data
3.	Cahyana & Nurlayli (2023)	✓	✓	✓		✓		116 Data
4.	Peerbasha *et al*, (2023)	✓	✓	✓		✓		2000 Data
5.	Gripsy & Divya (2023)	✓	✓	✓			✓	309 Data
6.	Prasetyo *et al*, (2022)	✓	✓				✓	498 Data
7.	Salma Amanda (2023)	✓	✓			✓		1500, 1550, 1600, dan 1650 Data

BAB III
METODE PENELITIAN

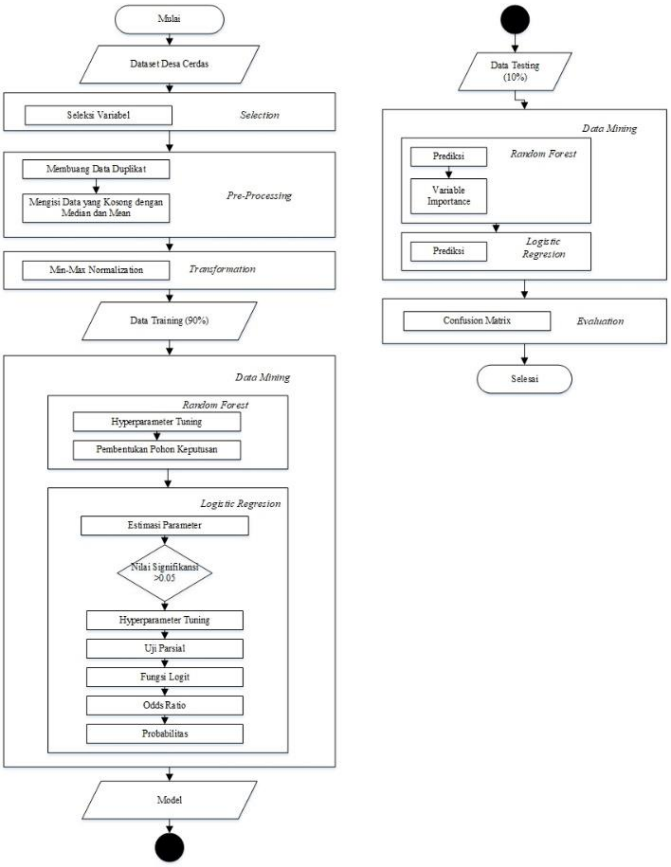
3.1 Metodologi Penelitian

Metodologi penelitian yang digunakan pada penelitian ini yaitu *Knowledge Discovery in Database* (KDD), yang merupakan sebuah proses untuk menemukan informasi atau pengetahuan yang berguna dari data. Adapun beberapa tahapan KDD menurut (Nurzanah *et al*, 2022), yaitu *selection*, *prepocessing*, *transformation*, *data mining*, dan *evaluation*.



Sumber: (Winarta & Kurniawan, 2021)

Gambar 2. Tahapan KDD



Gambar 3. Alur Penelitian

3.1.1 *Selection*

Tahap *selection* yaitu proses untuk menyeleksi data dan variabel yang akan digunakan pada proses *data mining* nanti. Data yang digunakan pada penelitian ini adalah data potensi desa tahun 2021. Data tersebut terdiri dari variabel-variabel yang termasuk ke dalam dimensi untuk memprediksi potensi desa cerdas, seperti *ICT*, *mobility*, *economy*, *sosial* and *cultural*, *community and government*, dan *environment.*

3.1.2 *Preprocessing / Cleaning*

Tahap *preprocessing* yaitu proses melakukan pembersihan terhadap data-data, dengan cara membuang data duplikat dan mengisi data yang kosong.

3.1.3 *Transformation*

Tahap *Transformation* yaitu proses mengubah data agar dapat diolah atau diproses pada saat tahap data mining nanti.

3.1.4 *Data mining*

Data mining merupakan proses untuk menemukan pola atau informasi yang menarik dalam data yang telah dipilih menggunakan algoritma tertentu (Bahtiar, 2023).

3.1.5 *Interpretation / Evaluation*

Informasi yang diperoleh dari proses *data mining* akan dievaluasi untuk mengetahui kinerja algoritmanya. Kemudian, informasi yang dihasilkan akan ditampilkan ke dalam bentuk yang mudah dimengerti.

3.2 Alat dan Bahan

3.2.1 Alat

Alat yang digunakan yaitu berupa perangkat lunak dan perangkat keras yang meliputi:

a. Perangkat Lunak
   1. Sistem Operasi Windows 10 Pro 64bit
   2. Google Chrome 111.0.5563.65
   3. Microsoft Office 2019

b. Perangkat Keras
   1. Laptop Dell Inspiron 14
   2. Processor Intel(R) Core (TM) i3-4005U
   3. RAM 4.00 GB
   4. SSD 512 GB

3.2.2 Bahan

Bahan yang digunakan penelitian ini yaitu:

1. Data potensi desa tahun 2021 yang diperoleh dari Badan Pusat Statistik (BPS).
2. Buku, jurnal, dan skripsi sebagai bahan referensi pembuatan laporan penelitian.
3. Panduan skripsi Program Studi Ilmu Komputer, Fakultas Matematika Dan Ilmu Pengetahuan Alam, Universitas Pakuan.

BAB IV
PERANCANGAN DAN IMPLEMENTASI

4.1 Tahap Perancangan

4.1.1 *Selection*

Data yang digunakan pada penelitian ini adalah data potensi desa tahun 2021 yang berjumlah 84.096 data. Namun penulis menggunakan beberapa dataset sebagai percobaan, yang dimana setiap dataset memiliki 1350 data desa yang sudah desa cerdas sedangkan jumlah untuk desa yang belum desa cerdas akan diambil sebanyak 150, 200, 250, dan 300 data dengan cara acak dan proposional. Cara proposional dilakukan menggunakan Persamaan 1.

Sebagai contoh, penulis akan mengambil sampel dengan jumlah sampel sebanyak 200 pada Provinsi Jawa Tengah yang jumlah desa yang belum berlabel desa cerdas sebanyak 8.427 dan jumlah anggota populasi seluruhnya sebanyak 82.473. Maka jumlah anggota sampel:

Provinsi Jawa Tengah: 8,427 Desa

$$Strata = \frac{8427\ x\ 200}{82473} = 20.4357\ \approx 20$$

Adapun jumlah variabel/atribut sebanyak 857, namun pada penelitian ini hanya menggunakan 29 variabel independen yang mengacu pada dimensi desa cerdas oleh BPS dan 1 variabel dependen. Sedangkan untuk kelas target mengacu pada Surat Keputusan (SK) penetapan desa cerdas fase I tahun 2021 sebanyak 350 desa dan fase II tahun 2022 sebanyak 1000 desa yang dikeluarkan oleh Kementerian Desa, Pembangunan Daerah Tertinggal dan Transmigrasi Republik Indonesia.

Tabel 3. Variabel Penelitian

No.	Atribut (Tipe)	Keterangan	Rentang Nilai
1.	R809A (Numerik)	Jumlah jenis-jenis lembaga kemasyarakatan desa: PKK	Min = 0, Max = n
2.	R809B (Numerik)	Jumlah jenis-jenis lembaga kemasyarakatan desa: Karang taruna	Min = 0, Max = n
3.	R809C (Numerik)	Jumlah jenis-jenis lembaga kemasyarakatan desa: Lembaga adat	Min = 0, Max = n
4.	R809D (Numerik)	Jumlah jenis-jenis lembaga kemasyarakatan desa: Kelompok tani	Min = 0, Max = n
5.	R809E (Numerik)	Jumlah jenis-jenis lembaga kemasyarakatan desa: Lembaga pengelolaan air	Min = 0, Max = n
6.	R809F (Numerik)	Jumlah jenis-jenis lembaga kemasyarakatan desa: Kelompok masyarakat (pokmas)	Min = 0, Max = n
7.	R1205A1 (Numerik)	Jumlah Bank Umum Pemerintah (misalnya: BRI, BNI, MANDIRI, BPD, dll.) yang beroperasi di wilayah desa kelurahan :	Min = 0, Max = n
8.	R1205A2 (Numerik)	Jumlah Bank Umum Swasta (misalnya: Danamon, BCA, Niaga, dll.) yang beroperasi di wilayah desa kelurahan :	Min = 0, Max = n

No.	Atribut (Tipe)	Keterangan	Rentang Nilai
9.	R1205A3 (Numerik)	Jumlah Bank Perkreditan Rakyat (BPR) yang beroperasi di wilayah desa kelurahan :	Min = 0, Max = n
10.	R1206A1 (Numerik)	Jumlah KUD	Min = 0, Max = n
11.	R1206A2 (Numerik)	Jumlah koperasi Industri Kecil dan Kerajinan Rakyat (Kopinkra) yang masih aktif beroperasi	Min = 0, Max = n
12.	R1206A3 (Numerik)	Jumlah koperasi Simpan Pinjam (Kospin) yang masih aktif beroperasi : ..... unit	Min = 0, Max = n
13.	R1206A4 (Numerik)	Jumlah koperasi lainnya ..... (misalnya koperasi: pegawai, karyawan, pensiunan, sekolah, dll) yang masih aktif beroperasi : ..... Unit	Min = 0, Max = n
14.	R1007C (Kategorik)	Perusahaan agen jasa ekspedisi (pengiriman barang dokumen) swasta:	1=Beroperasi 2=Jarang beroperasi 3=Tidak beroperasi 4=Tidak ada
15.	R1001C1 (Kategorik)	Keberadaan angkutan umum:	1=Ada, dengan trayek tetap 2=Ada, tanpa trayek tetap 3=Tidak ada angkutan umum
16.	R1001A (Kategorik)	Lalu lintas dari ke desa kelurahan melalui :	1=Darat 2=Air 3=Darat dan air 4=Udara
17.	R1007A (Kategorik)	Keberadaan kantor pos pos pembantu rumah pos :	1=Beroperasi 2=Jarang beroperasi 3=Tidak beroperasi 4=Tidak ada
18.	R1003B (Kategorik)	Keberadaan warga yang menggunakan telepon seluler handphone HP	1=Sebagian besar warga 2=Sebagian kecil warga 3=Tidak ada
19.	R1004 (Kategorik)	Keberadaan internet untuk warnet, game online, dan fasilitas lainnya di desa kelurahan	1=Ada 2=Tidak ada
20.	R1005C (Kategorik)	Sinyal telepon seluler handphone di sebagian besar wilayah desa kelurahan	1=Sinyal sangat kuat 2=Sinyal kuat 3=Sinyal lemah 4=Tidak ada sinyal
21.	R1006B (Kategorik)	Fasilitas internet di kantor kepala desa lurah:	1=Berfungsi 2=Jarang berfungsi 3=Tidak befungsi 4=Tidak ada
22.	R1401A (Kategorik)	Keberadaan sistem informasi desa:	1=Ada, diperbaharui 2=Ada, tidak diperbaharui 3=Tidak ada

No.	Atribut (Tipe)	Keterangan	Rentang Nilai
23.	R1401C (Kategorik)	Pembangunan sistem keuangan desa:	1=Ada, diperbaharui 2=Ada, tidak diperbaharui 3=Tidak ada
24.	R501C (Kategorik)	Keluarga yang menggunakan lampu tenaga surya	1=Ada, sebagian besar 2=Ada, sebagian kecil 3=Tidak ada
25.	R502A (Kategorik)	Penerangan di jalan utama desa kelurahan yang menggunakan lampu tenaga surya	1=Ada 2=Tidak ada
26.	R604A (Kategorik)	Sistem peringatan dini bencana alam	1=Ada 2=Tidak ada
27.	R604B (Kategorik)	Sistem peringatan dini khusus tsunami	0=Bukan wilayah potensi tsunami 3=Ada 4=Tidak ada
28.	R808A (Kategorik)	Kebiasaan gotong royong warga di desa kelurahan untuk kepentingan umum komunitas :	1=Ada, sebagian besar warga terlibat 2=Ada, sebagian kecil warga terlibat 3=Tidak ada kebiasaan
29.	R808B (Kategorik)	Kegiatan gotong royong warga untuk membantu warga yang sedang mengalami musibah :	1=Ada, sebagian besar warga terlibat 2=Ada, sebagian kecil warga terlibat 3=Tidak ada kebiasaan

Adapun perbandingan variabel yang digunakan antara penelitian terdahulu (Tosida *et al*, 2021) dan (Tosida *et al*, 2020) dengan penelitian saat ini terlampir pada Lampiran 1.

4.1.2 *Preprocessing / Cleaning*

Pada penelitian ini, data yang kosong akan diisi dengan nilai mean atau rata-rata untuk data bertipe numerik dan nilai modus untuk data bertipe kategorik yang mengacu pada penelitian (Tosida *et al*., 2021).

```
# Menghitung jumlah total data yang duplikat berdasarkan kolom tertentu
total_duplikat = data.duplicated(subset=kolom_duplikat).sum()

if total_duplikat > 0:
    print(f"Total data yang duplikat berdasarkan kolom {kolom_duplikat}: {total_duplikat}")
    print("Data yang duplikat:")
    print(data[data.duplicated(subset=kolom_duplikat)])
else:
    print(f"Tidak ada data yang duplikat berdasarkan kolom {kolom_duplikat}.")
```

Gambar 4. Mengecek Data Duplikat

```
[ ]  # Mendapatkan nilai modus dari kolom 'R1401A'
     modus_kolom_R1401A = int(data['R1401A'].mode()[0])

     # Mendapatkan nilai modus dari kolom 'R1401C'
     modus_kolom_R1401C = int(data['R1401C'].mode()[0])
```

Gambar 5. Pengisian *Missing Value* menggunakan Nilai Modus

4.1.3 *Transformation*

Proses tranformasi data pada peneltiain ini menggunakan metode *Min-max Normalization*, dengan cara selisih data yang akan dinormalisasi dengan data minimum lalu dibagi dengan selisih data maksimum dengan data minimum dari satu kolom, yang di mana mengacu pada penelitian (Suryanegara *et al*, 2021).

```
# Menentukan kolom yang akan dinormalisasi
kolom_normalize = data.columns[5:34]

# Melakukan Min-Max Normalization pada kolom tertentu
data[kolom_normalize] = (data[kolom_normalize] - data[kolom_normalize].min()) / (data[kolom_normalize].max() - data[kolom_normalize].min())
```

Gambar 6. Transformasi Data

4.1.4 *Data mining*

Proses *data mining* pada penelitian ini menggunakan dua algoritma klasifikasi, yaitu *Random Forest* dan *Logistic Regression*.

4.1.4.1 *Random Forest*

1. **Hyperparameter Tunning**

   Proses *Hyperparameter Tuning* pada penelitian ini menggunakan metode *Grid Search* dan menggunakan metode *k-fold cross validation* sebanyak 10 *fold* untuk mengevaluasi kinerja model (Aini, 2023). Berikut adalah parameter yang digunakan pada algoritma *Random Forest*, merujuk pada penelitian (Aini, 2023).

   Tabel 4. Parameter *Random Forest*

Parameter	Keterangan
n_estimators	Jumlah pohon pada tree
max_depth	Kedalaman maksimum pada tree
min_samples_split	Jumlah minimum sampel yang diperlukan untuk memisahkan node internal
min_samples_leaf	Jumlah sampel minimum yang dibutuhkan leaf node
max_features	Jumlah fitur yang dipertimbangkan saat mencari split terbaik
criterion	Pengukuran untuk kualitas split

2. **Perhitungan Manual**

   Berikut adalah contoh penerapan algoritma *Random Forest* menggunakan data 1500 acak yang sudah ditransformasi dengan pembagian dataset 90:10 yaitu 1350:150. Untuk mempermudah proses perhitungan, peneliti membuat 5 *bootsrapped dataset*, dengan menggunakan 5 data *train* dan 5 fitur atau variabel.

   Tabel 5. Dataset Perhitungan Manual

ID	R809D	R1206A4	R502A	R604B	R809F	Target
1	0.15217	0	1	0	0.09091	1
2	0.02174	0	1	1	0.01136	0
3	0.34783	0.6	0	0	0.22727	1
4	0.32609	0.6	1	0	0.20455	1
5	0.06522	0	0	0	0.05682	0

Tabel 6. *Bootstrapped* Dataset Pertama

R809D	R1206A4	Target
0.02174	0	0
0.15217	0	1
0.34783	0.6	1
0.34783	0.6	1
0.32609	0.6	1

Pada *Bootstrapped* dataset pertama, ID yang digunakan yaitu (2 1 3 3 4) dan ID yang tidak digunakan yaitu 1. Lalu Fitur yang digunakan adalah R809D dan R1206A4. Setelah itu menghitung *Gini Impurity* (GI) menggunakan persamaan 2.

Hasil dari GI(R809D) adalah GI(R809D) = $1 - ((1/5)^2 + (4/5)^2) = 0.38$ dan GI GI(R1206A4) = $1 - ((1/5)^2 + (4/5)^2) = 0.38$. Karena GI(R809D) dan GI(R1206A4) memiliki nilai yang sama, maka dapat memilih salah satu variabel sebagai *root node*, misalnya R1206A4. Selanjutnya, perlu untuk membuat split pada R1206A4 dengan *threshold* tertentu. Pada penelitian ini akan menggunakan nilai mean sebagai *threshold*, didapatkan nilai mean pada fitur R1206A4 adalah $\frac{0+0+0.6+0.6+0.6}{5} = 0.36$. Maka, data akan terbagi menjadi 2 subset:

Subset 1 (R1206A4 ≤ 0.36)

R809D	R1206A4	Target
0.02174	0	0
0.15217	0	1

Subset 2 (R1206A4 > 0.36)

R809D	R1206A4	Target
0.34783	0.6	1
0.34783	0.6	1
0.32609	0.6	1

Leaf node pada subset 2 sudah terbentuk yaitu 1 dikarenakan homogen. Selanjutnya yaitu mencari mean dari R809D dari subset 1, didapatkan mean sebesar $\frac{0.02174+0.15217}{2} = 0.086955$. Sehingga subset yang dihasilkan sebagai berikut

Subset 3 (R809D ≤ 0.086955)

R809D	R1206A4	Target
0.02174	0	0

Subset 4 (R809D > 0.086955)

R809D	R1206A4	Target
0.15217	0	1

Jadi, apabila nilai R809D ≤ 0.086955, maka hasil leaf node nya 0 dan jika nilai R809D > 0.086955, maka hasil leaf nodenya 1.

Tahap *Bootstrapped* kedua sampai *Bootstrapped* kelima terlampir pada Lampiran 3.

3. *Variable Importance*

Pada Penelitian ini, untuk mengukur pentingnya sebuah variabel menggunakan *Features Importance* yang didapat dari *RandomForestClassifier*.

4.1.4.2 *Logistic Regression*

1. Estimasi Parameter

Sebelum membuat analisis regresi logistik dengan 29 variabel, harus terlebih dahulu melakukan uji variabel dan mencari variabel yang berpengaruh dengan mengeliminasi variabel yang memiliki nilai signifikansi > α. Eliminasi berakhir ketika nilai signifikansi semua variabel ≤ α. Merujuk pada penelitian (Zaen, 2019)) nilai α (alpha) yang digunakan adalah sebesar 0.05 atau 5% dan penelitian ini menggunakan *Forward Selection* untuk proses eliminasi variabelnya, yang dimana untuk proses pemodelannya akan dimulai dengan tidak menggunakan fitur apa pun, dan kemudian secara iteratif fitur satu persatu dimasukkan sampai tidak ada lagi fitur yang memenuhi kriteria (Cahyaningtyas *et al*., 2022). Berikut adalah contoh estimasi parameter menggunakan 1500 dataset acak 90:10.

Iterasi Pertama

Hasil dari iterasi pertama didapatkan variabel R1006B memiliki p-value yang paling kecil yaitu sebesar 2.58E-13 ≤ alpha (0.05). Maka variabel tersebut berpengaruh terhadap prediksi potensi desa cerdas dan akan dihapus dari iterasi.

Tabel 7. Iterasi Pertama Estimasi Parameter

Variabel	P-Value
R1006B	**2.58E-13**
R1206A4	0.298586407
R809F	0.015090203
R809B	0.828441312
R1001A	0.005983184
R1206A3	0.089043071
R1206A2	0.480672516
R1001C1	0.300585222
R604A	0.03567857
R1007A	0.471494805
R501C	0.02258427
R1003B	1.17E-08
R502A	0.925561518
R604B	0.1196872
R1206A1	0.167187128
R1205A1	0.951693374
R808A	0.359030284
R809D	0.005290403
R1004	0.003210353
R1401A	0.016795054
R1205A3	0.103013988

Variabel	P-Value
R809A	0.269753426
R1205A2	0.052711794
R1005C	6.78E-12
R809C	0.136178274
R808B	0.78571736
R1401C	0.129074079
R809E	0.017551739
R1007C	0.152153751

Tahap iterasi selengkapnya terlampir pada Lampiran 4.
Dari ketujuh iterasi, didapatkan variabel R1006B, R1005C, R1003B, R1205A2, dan R809D merupakan variabel yang berpengaruh secara signifikan terhadap prediksi potensi desa cerdas, sehingga variabel-variabel tersebut akan dimasukkan ke dalam model regresi logistik.

2. *Hyperparameter Tuning*

Proses *Hyperparameter Tuning* pada penelitian ini menggunakan metode *Grid Search* dan menggunakan metode *k-fold cross validation* sebanyak 10 *fold* untuk mengevaluasi kinerja model (Aini, 2023). Berikut adalah parameter yang digunakan pada algoritma *Logistic Regression*, merujuk pada penelitian (Oberoi *et al*., 2023).

Tabel 8. Parameter *Logistic Regression*

Parameter	Keterangan
penalty	Tipe regularisasi
C	Kekuatan regularisasi

3. Uji Parsial

Uji parsial pada penelitian ini menggunakan uji wald yaitu pada Persamaan 5. Dengan kriteria penolakan H_0 apabila $|W_i| > Z_{\alpha/2}$ atau sig. < α.

Tabel 9. Variabel Uji Wald 85:15

Variabel	β	Std Error
(Konstanta)	2.8866	
R1006B	-1.0555	0.1936
R1005C	-1.5543	0.3909
R1003B	-1.2868	0.3806
R1205A2	0.7845	0.46393
R809D	0.8884	0.3795

Berdasarkan Tabel 9, variabel yang tersisa sebanyak 5 variabel, yaitu variabel R1006B, R1005C, R1205A2, R809D, R1003B dan R1004, oleh karena itu uji wald hanya dilakukan pada keenam variabel tersebut. Berikut contoh perhitungan uji wald menggunakan Persamaan 5:

Variabel R1006B

$$\left[\frac{\beta_j}{SE(\beta_j)}\right] = \left[\frac{-1.0555}{0.1936}\right] = -5.4519$$

Berdasarkan perhitungan di atas didapatkan nilai W = $|-5.4519| = 5.4519 > Z_{\alpha/2}$ (1.96), maka bisa disimpulkan bahwa variabel R1006B mempunyai pengaruh terhadap variabel dependen.

Perhitungan Uji Wald selengkapnya terlampir pada Lampiran 5.

4. Model Terbaik

Variabel yang berdampak pada variabel dependen atau memiliki nilai signifikansi ≤ alpha (0,05) digunakan dalam model *logistic regression*, yaitu variabel R1006B, R1005C, R1003B, R1205A2, dan R809D. Berdasarkan Tabel menggunakan Persamaan 4, maka diperoleh persamaan *Logistic Regression* sebagai berikut:

$$= \frac{e^{(2.8866-1.0555_{\mathrm{R1006B}}-1.5543_{\mathrm{R1005C}}-1.2868_{\mathrm{R1003B}}+0.7845_{\mathrm{R1205A2}}+0.8884_{\mathrm{R809D}})}}{1+e^{(2.8866-1.0555_{\mathrm{R1006B}}-1.5543_{\mathrm{R1005C}}-1.2868_{\mathrm{R1003B}}+0.7845_{\mathrm{R1205A2}}+0.8884_{\mathrm{R809D}})}}$$

$$g(x) = 2.8866-1.0555_{\mathrm{R1006B}} - 1.5543_{\mathrm{R1005C}} - 1.2868_{\mathrm{R1003B}} + 0.7845_{\mathrm{R1205A2}}+0.8884_{\mathrm{R809D}}$$

5. *Odds Ratio*

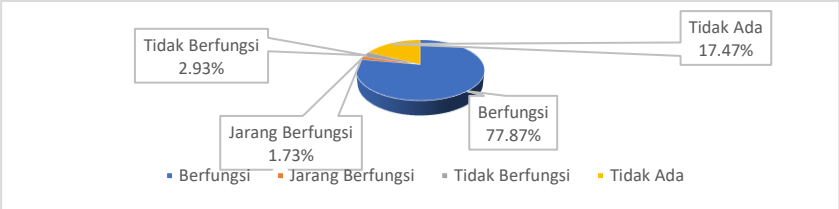
Berikut pada Tabel 10 adalah perhitungan *Odds Ratio* dari masing-masing variabel yang berpengaruh atau signifikan dengan menggunakan Persamaan 6.

Tabel 10. *Odds Ratio*

Variabel	β	*Odds Ratio*
R1006B	-1.0555	0.3480
R1005C	-1.5543	0.2113
R1003B	-1.2868	0.2761
R1205A2	0.7845	0.4563
R809D	0.8884	2.4314

Contoh interpretasi salah satu variabel adalah sebagai berikut:

Variabel R1006B (Fasilitas internet di kantor kepala desa lurah)



Gambar 7. *Odds Ratio* Variabel R1006

Variabel R1006B menghasilkan nilai *Odds Ratio* sebesar 0.3480, yang menunjukkan bahwa fasilitas internet di kantor kepala desa yang berfungsi mempunyai peluang untuk berpotensi menjadi desa cerdas 0.3480 kali lebih besar daripada status fasilitas internet yang lainnya.

6. Probabilitas

Menghitung besarnya probabilitas setiap desa untuk berpotensi menjadi desa cerdas dapat dilakukan menggunakan persamaan yang sudah didapatkan pada tahap pembentukan model terbaik, berikut contoh perhitungan probabilitas menggunakan Persamaan 4 :

$$= \frac{e^{(2.8866 - 1.0555_{\mathrm{R1006B}} - 1.5543_{\mathrm{R1005C}} - 1.2868_{\mathrm{R1003B}} + 0.7845_{\mathrm{R1205A2}} + 0.8884_{\mathrm{R809D}})}}{1 + e^{(2.8866 - 1.0555_{\mathrm{R1006B}} - 1.5543_{\mathrm{R1005C}} - 1.2868_{\mathrm{R1003B}} + 0.7845_{\mathrm{R1205A2}} + 0.8884_{\mathrm{R809D}})}}$$

Misal ingin menghitung besar probabilitas desa Srikuncoro di Bengkulu:

$$= \frac{e^{(2.8866 - 1.0555_{(0)} - 1.5543_{(0.3333)} - 1.2868_{(0)} + 0.7845_{(0)} + 0.8884_{(0.0217)})}}{1 + e^{(2.8866 - 1.0555_{(0)} - 1.5543_{(0.3333)} - 1.2868_{(0)} + 0.7845_{(0)} + 0.8884_{(0.0217)})}}$$

$$= \frac{10.8898}{1 + 10.8898} = 0.9159$$

Kesimpulannya, Desa Srikuncoro akan berpotensi desa cerdas sebesar 91.59%.

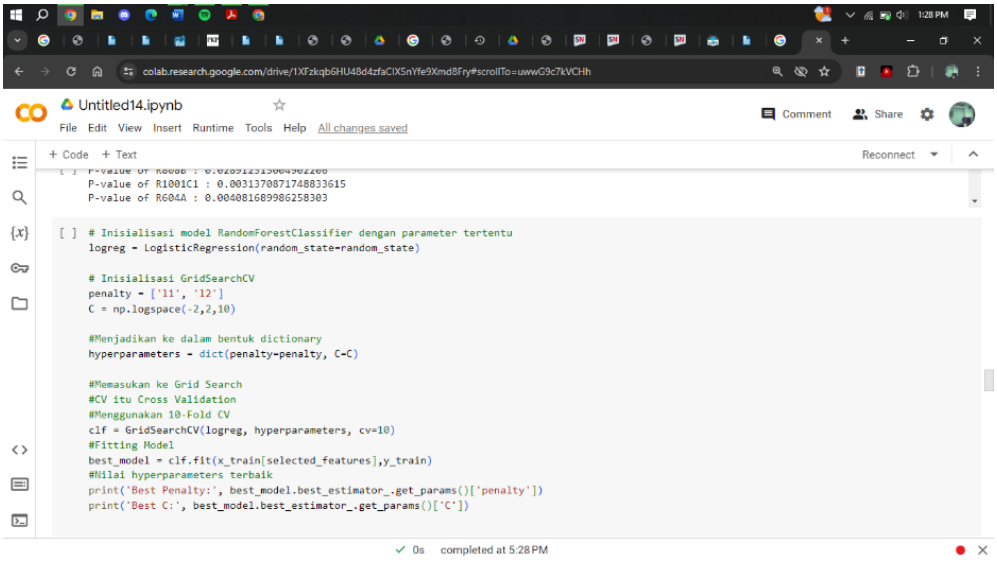
Nilai probabilitas yang didapat digunakan untuk proses prediksi potensi desa cerdas, yang dimana jika nilai probabilitas kurang dari 0.5 akan diprediksi sebagai kelas 0 dan jika lebih atau sama dengan 0.5 akan diprediksi sebagai ke kelas 1 (Oberoi *et al*., 2023)

4.1.5 *Interpretation / Evaluation*

Evaluasi dalam penelitian ini dilakukan menggunakan *Confusion Matrix* lalu hasilnya ditampilkan berupa tabel perbandingan.

4.2 Implementasi

Tahap implementasi merupakan tahap pembangunan sistem yang sudah dirancang. Dalam penelitian ini, sistem dibangun menggunakan *Google Colaboratory* dengan bahasa pemrograman *Python*.



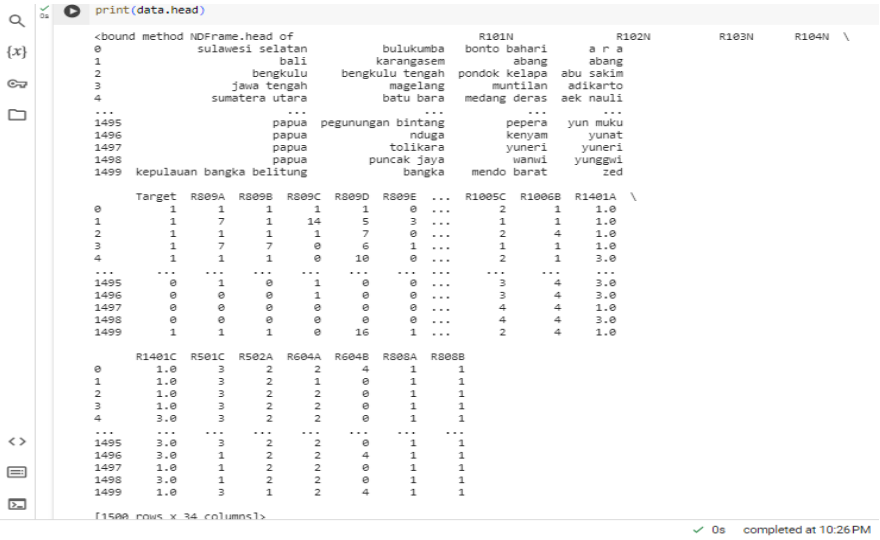
Gambar 8. Implementasi Sistem

BAB V
HASIL DAN PEMBAHASAN

5.1 Hasil

5.1.1 *Selection*

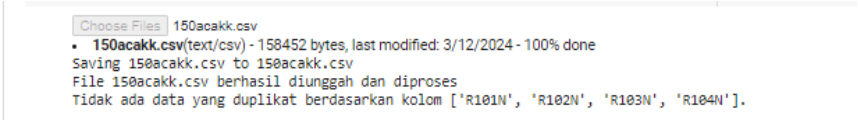
Data yang digunakan pada penelitian ini adalah data potensi desa tahun 2021 yang berjumlah 84.096 data. Namun penulis menggunakan beberapa dataset sebagai eksperimen, yang dimana setiap dataset memiliki 1350 data desa yang sudah desa cerdas sedangkan jumlah untuk desa yang belum desa cerdas akan diambil sebanyak 150, 200, 250, dan 300 data dengan cara acak dan proposional. Adapun jumlah variabel/atribut sebanyak 871, namun pada penelitian ini hanya menggunakan 29 variabel independen dan 1 variabel dependen. Berikut adalah 5 data teratas dan 5 data terbawah pada dataset 1500 dengan data acak. Untuk dataset selengkapnya terlampir pada Lampiran 2.



Gambar 9. Lima Dataset Teratas dan Terbawah

5.1.2 *Preprocessing / Cleaning*

Tahap *preprocessing* pada penelitian ini adalah memeriksa dan menghapus data duplikat, serta mengisi data yang kosong dengan nilai mean atau rata-rata untuk data bertipe numerik dan nilai modus untuk data bertipe kategorik. Dari proses pengecekkan data duplikat, didapatkan semua dataset tidak ada data duplikat. Berikut pada Gambar 10 contoh hasil pengecekkan data duplikat pada dataset 1500 dengan data acak.



Gambar 10. Hasil Pengecekkan Data Duplikat

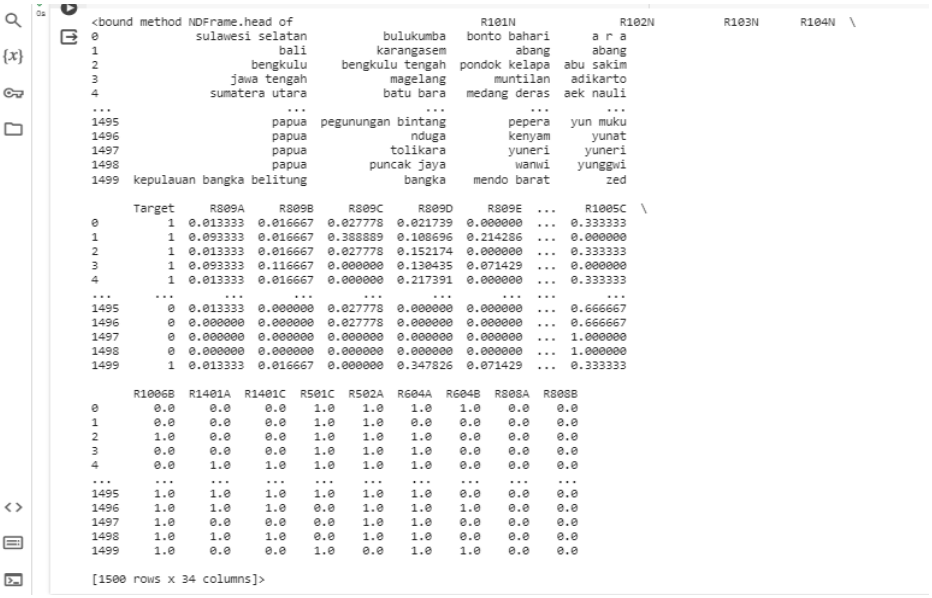
Lalu pada Gambar 11 adalah pengecekkan *Missing value* atau data yang kosong. Terdapat *missing value* pada variabel R1401A dan R1401C yang bertipe kategorik, maka *missing value* akan diisi dengan nilai modus.

```
Mengecek Missing Value:
R101N       0
R102N       0
R103N       0
R104N       0
Target      0
R809A       0
R809B       0
R809C       0
R809D       0
R809E       0
R809F       0
R1205A1     0
R1205A2     0
R1205A3     0
R1206A1     0
R1206A2     0
R1206A3     0
R1206A4     0
R1007C      0
R1001C1     0
R1001A      0
R1007A      0
R1003B      0
R1004       0
R1005C      0
R1006B      0
R1401A     21
R1401C     21
R501C       0
R502A       0
R604A       0
R604B       0
R808A       0
R808B       0
dtype: int64
```

Gambar 11. Hasil Pengecekkan *Missing Value*

5.1.3 *Transformation*

Tahap *Transformation* merupakan proses dimana data diubah agar dapat diolah atau diproses dalam tahap *data mining*. Pada penelitian ini, data ditransformasi menggunakan metode *Min-max Normalization*. Berikut pada Gambar 12 adalah hasil data yang sudah ditransformasi.



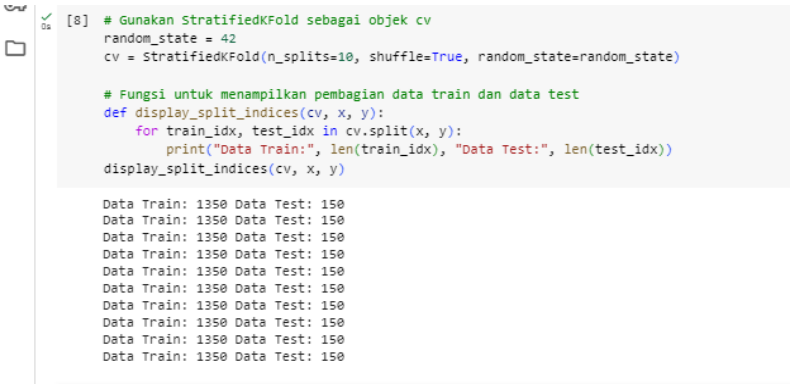
```
<bound method NDFrame.head of                          R101N              R102N         R103N      R104N  \
0              sulawesi selatan          bulukumba   bonto bahari      a r a
1                         bali         karangasem          abang      abang
2                    bengkulu     bengkulu tengah  pondok kelapa  abu sakim
3                  jawa tengah           magelang       muntilan   adikarto
4                sumatera utara          batu bara   medang deras  aek nauli
...                         ...                ...            ...        ...
1495                     papua  pegunungan bintang         pepera   yun muku
1496                     papua              nduga          kenyam      yunat
1497                     papua           tolikara          yuneri     yuneri
1498                     papua        puncak jaya           wanwi    yunggwi
1499  kepulauan bangka belitung             bangka    mendo barat        zed

      Target     R809A     R809B     R809C     R809D     R809E  ...     R1005C  \
0          1  0.013333  0.016667  0.027778  0.021739  0.000000  ...  0.333333
1          1  0.093333  0.016667  0.388889  0.108696  0.214286  ...  0.000000
2          1  0.013333  0.016667  0.027778  0.152174  0.000000  ...  0.333333
3          1  0.093333  0.116667  0.000000  0.130435  0.071429  ...  0.000000
4          1  0.013333  0.016667  0.000000  0.217391  0.000000  ...  0.333333
...      ...       ...       ...       ...       ...       ...  ...       ...
1495       0  0.013333  0.000000  0.027778  0.000000  0.000000  ...  0.666667
1496       0  0.000000  0.000000  0.027778  0.000000  0.000000  ...  0.666667
1497       0  0.000000  0.000000  0.000000  0.000000  0.000000  ...  1.000000
1498       0  0.000000  0.000000  0.000000  0.000000  0.000000  ...  1.000000
1499       1  0.013333  0.016667  0.000000  0.347826  0.071429  ...  0.333333

      R1006B  R1401A  R1401C  R501C  R502A  R604A  R604B  R808A  R808B
0        0.0     0.0     0.0    1.0    1.0    1.0    1.0    0.0    0.0
1        0.0     0.0     0.0    1.0    1.0    0.0    0.0    0.0    0.0
2        1.0     0.0     0.0    1.0    1.0    1.0    0.0    0.0    0.0
3        0.0     0.0     0.0    1.0    1.0    1.0    0.0    0.0    0.0
4        0.0     1.0     1.0    1.0    1.0    1.0    0.0    0.0    0.0
...      ...     ...     ...    ...    ...    ...    ...    ...    ...
1495     1.0     1.0     1.0    1.0    1.0    1.0    0.0    0.0    0.0
1496     1.0     1.0     1.0    0.0    1.0    1.0    1.0    0.0    0.0
1497     1.0     0.0     0.0    0.0    1.0    1.0    0.0    0.0    0.0
1498     1.0     1.0     1.0    0.0    1.0    1.0    0.0    0.0    0.0
1499     1.0     0.0     0.0    1.0    0.0    1.0    1.0    0.0    0.0

[1500 rows x 34 columns]>
```

Gambar 12. Data yang Sudah Ditransformasi

5.1.4 *Data mining*

Data mining merupakan proses untuk menemukan pola atau informasi yang menarik dalam data yang telah dipilih. Proses *data mining* pada penelitian ini menggunakan dua algoritma klasifikasi, yaitu *Random Forest* dan *Logistic Regression*.

Sebelum masuk ke dalam proses pemodelan menggunakan kedua algoritma tersebut, dataset dibagi menjadi data *training* dan data *testing* yang didapat dari hasil metode k-fold cross validation dengan jumlah k = 10. Berdasarkan Gambar 13, didapatkan hasil pembagian dataset adalah 90:10. Kemudian menggunakan pembagian 85:15, 80:20, dan 70:30 juga sebagai perbandingan.



```
[8]  # Gunakan StratifiedKFold sebagai objek cv
     random_state = 42
     cv = StratifiedKFold(n_splits=10, shuffle=True, random_state=random_state)

     # Fungsi untuk menampilkan pembagian data train dan data test
     def display_split_indices(cv, x, y):
         for train_idx, test_idx in cv.split(x, y):
             print("Data Train:", len(train_idx), "Data Test:", len(test_idx))
     display_split_indices(cv, x, y)

     Data Train: 1350 Data Test: 150
     Data Train: 1350 Data Test: 150
     Data Train: 1350 Data Test: 150
     Data Train: 1350 Data Test: 150
     Data Train: 1350 Data Test: 150
     Data Train: 1350 Data Test: 150
     Data Train: 1350 Data Test: 150
     Data Train: 1350 Data Test: 150
     Data Train: 1350 Data Test: 150
     Data Train: 1350 Data Test: 150
```

Gambar 13. Pembagian Dataset

5.1.4.1 *Random Forest*

1. *Hyperparameter Tuning*

Berikut pada Tabel 11 dan Tabel 12 adalah hasil parameter terbaik dari *Random Forest* yang didapatkan dari proses *grid searchCV* dengan melakukan pencarian secara menyeluruh terhadap parameter menggunakan 10 *fold cross validation* untuk mengevaluasi kinerja model. Hasil parameter terbaik pada dataset yang lain terlampir pada Lampiran 6.

Tabel 11. Parameter Terbaik *Random Forest* Dataset Acak 85:15

Parameter	**Grid Search Values**	**Best Parameter Dataset**			
		1500	**1550**	**1600**	**1650**
n_estimators	100, 200, 500	200	100	100	100
Max_depth	2, 5, 6, 7, 8	8	7	8	7
Criterion	Entropy, Gini	entropy	entropy	gini	Entropy
Min_samples_leaf	1, 2, 4	2	4	1	2
Max_features	Auto, log2	Log2	Auto	Auto	auto
Min_samples split	2, 5, 10	2	2	2	10

Tabel 12. Parameter Terbaik *Random Forest* Dataset Proposional 85:15

Parameter	**Grid Search Values**	**Best Parameter Dataset**			
		1500	**1550**	**1600**	**1650**
n_estimators	100, 200, 500	200	500	200	100
Max_depth	2, 5, 6, 7, 8	8	7	8	5
Criterion	Entropy , Gini	Gini	Gini	Entropy	Gini
Min_samples_leaf	1, 2, 4	1	1	1	1
Max_features	Auto, log2	Auto	Auto	Log2	Auto
Min_samples split	2, 5, 10	2	2	5	10

Keterangan:

- *Max_features* auto, maka *Max_features* = √(Jumlah fitur)
- *Max_features* log2, maka *Max_features* = $\log_2$(Jumlah fitur)

2. Prediksi

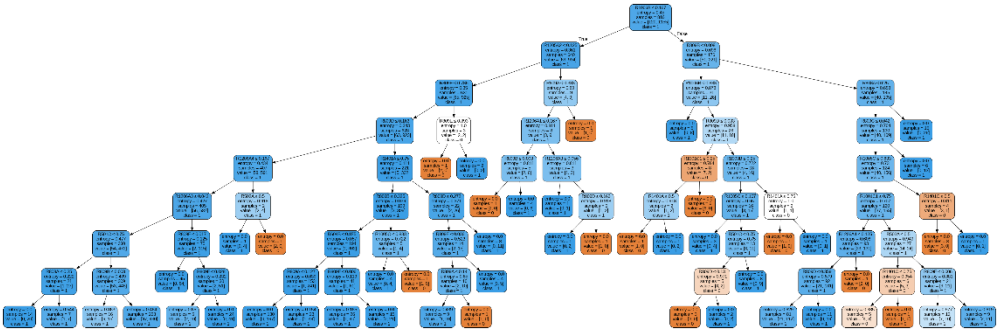
Setelah melakukan proses *Hyperparameter Tuning*, dilakukan proses pembentukan model yang di mana model tersebut digunakan untuk memprediksi potensi desa cerdas. Dari Tabel 11, diketahui dataset 1500 acak 85:15 menggunakan 200 pohon, kemudian setiap pohon membuat maksimal 8 percabangan. Jumlah fitur yang dipertimbangkan saat mencari split terbaik adalah 5. Lalu, untuk mengukur kualitas split pada pohon menggunakan nilai Entropy. Adapun jumlah sampel minimum yang digunakan pada setiap leaf node sebanyak 2 dan untuk menghentikan proses split diperlukan minimal 2 sampel pada node tersebut.

Berikut pada pada Tabel 13 adalah hasil prediksi 5 desa teratas dan dan Gambar 14 adalah pohon keputusan pada dataset 1500 acak 85:15

Tabel 13. Hasil Prediksi *Random Forest* 5 Teratas

Nama Provinsi	Nama Desa	Aktual	Prediksi
Bengkulu	Srikuncoro	1	1
Di Yogyakarta	Demangan	0	1
Kalimantan Barat	Jongkong Kiri Tengah	1	1
Sumatera Selatan	Jejawi	1	1
Jawa Tengah	Kaligentong	1	1

Hasil prediksi dan pohon keputusan selengkapnya terlampir pada Lampiran 7.

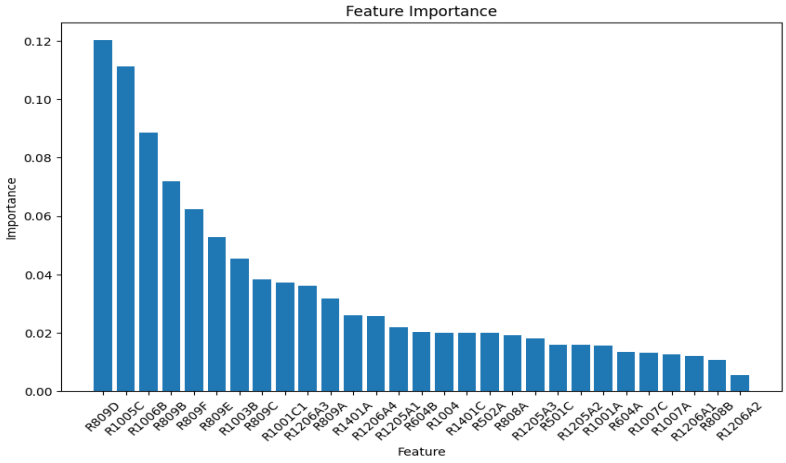


Gambar 14. Salah Satu Pohon Keputusan pada Dataset 1500 Acak 85:15

Pada gambar pohon keputusan di atas, variabel utama klasifikasi desa cerdas adalah fasilitas internet di kantor kepala lurah. Jika fasilitas internet tersebut berfungsi, maka dicek apakah jumlah bank umum di desa kurang atau sama dengan 1. Jika benar, maka dicek kembali apakah jumlah lembaga pengelolaan air di desa kurang atau sama dengan 11, dan begitu seterusnya. Untuk rule lebih lengkapnya terlampir pada Lampiran 8 dan untuk gambar pohon keputusan

3. ***Variable Importance***

Berikut adalah hasil *Variable Importance* dataset 1500 acak 85:15



Gambar 15. *Variable Importance* 1500 Acak 85:15

Gambar 15 di atas menyajikan nilai *Variable Importance*. Variabel R809D (Jumlah kelompok tani, R1005C (Sinyal telepon seluler handphone di sebagian besar wilayah desa kelurahan), R1006B (Fasilitas internet di kantor kepala desa lurah), R809B (Jumlah karang taruna), dan R809F (Jumlah Kelompok Masyarakat) menjadi 5 variabel teratas yang mempengaruhi prediksi potensi desa cerdas.

Variable Importance dataset yang lain terlampir pada Lampiran 9.

5.1.4.2 *Logistic Regression*

1. **Estimasi Parameter**

Sebelum membuat analisis regresi logistik 29 variabel, terlebih dahulu melakukan uji variabel dan mencari variabel yang berpengaruh dengan mengeliminasi variabel yang memiliki nilai signifikansi > α. Eliminasi berakhir ketika nilai signifikansi semua variabel ≤ α. Nilai α atau alpha yang digunakan pada penelitian ini adalah 0.05. Berikut pada Tabel 14 adalah hasil estimasi parameter. Estimasi Parameter pada dataset selengkapnya terlampir pada Lampiran 10.

Tabel 14. Hasil Estimasi Parameter 85:15

Dataset		Estimasi Parameter
Acak	**1500**	R1006B, R1005C, R1003B, R1205A2, R809D
	1550	R1005C, R1006B, R809E, R1401A, R809F, R1206A4, R809A, R809D
	1600	R1006B, R1005C, R501C
	1650	R1006B, R1005C, R809D, R604A, R1401A
Proposional	**1500**	R1005C, R1006B, R1401C, R809A, R809C, R1001C1, R809D, R1004
	1550	R1005C, R1006B, R809D, R1003B, R1206A3, R809A
	1600	R1005C, R1003B, R1006B, R809D, R1401C, R1206A3, R809A, R1206A1, R1401A, R1205A1
	1650	R1006B, R1005C, R809D, R604A, R1003B, R1401C, R1401A

2. Hyperparameter Tuning

Berikut pada Tabel 15 adalah hasil parameter terbaik dari *Logistic Regression* yang didapatkan dari proses *grid searchCV* dengan melakukan pencarian secara menyeluruh terhadap parameter menggunakan 10 *fold cross validation* untuk mengevaluasi kinerja model. Hasil parameter terbaik pada dataset yang lain terlampir pada Lampiran 11.

Tabel 15. Parameter Terbaik *Logistic Regression* Dataset Acak 85:15

Parameter	Grid Search Values	Best Patameter Dataset			
		1500	1550	1600	1650
Penalty	L1 (Lasso), L2 (Ridge)	L2	L2	L2	L2
C	Logspace (-2,2)	0.599484250 3189409	1.668100537 2000592	0.599484250 3189409	12.91549665 0148826

Tabel 16. Parameter Terbaik *Logistic Regression* Dataset Proposional 85:15

Parameter	Grid Search Values	Best Patameter Dataset			
		1500	1550	1600	1650
Penalty	L1 (Lasso), L2 (Ridge)	L2	L2	L2	L2
C	Logspace (-2,2)	12.91549665 0148826	4.64158883361 27775	0.599484250 3189409	35.93813663 804626

3. Uji Parsial

Berikut pada Tabel adalah hasil uji wald. Hasil uji wald selengkapnya terlampir pada Lampiran 12.

1500 acak

Tabel 17. Hasil Uji Wald Dataset 1500 Acak 85:15

Variabel	β	Std Error	Uji Wald
(Konstanta)	2.8866		
R1006B	-1.0555	0.1936	-5.4519
R1005C	-1.5543	0.3909	-3.9762
R1003B	-1.2868	0.3806	-3.3809
R1205A2	0.7845	0.46393	1.6909
R809D	0.8884	0.3795	2.3409

Dari hasil uji wald, didapatkan semua variabel pada masing – masing dataset menghasilkan nilai wald $|W_i| > Z_{\alpha/2}$ (1.96) atau sig. < α (0.05), maka variabel – variabel tersebut mempunyai pengaruh terhadap variabel dependen.

4. Model Terbaik

Berdasarkan Tabel 14, Tabel 17, dan Lampiran 12 maka diperoleh persamaan *Logistic Regression* dari masing – masing dataset menggunakan Persamaan 4 adalah sebagai berikut:

Hasil model terbaik dataset 1500 Acak 85:15

$$= \frac{e^{(2.8866-1.0555_{\mathrm{R1006B}}-1.5543_{\mathrm{R1005C}}-1.2868_{\mathrm{R1003B}}+0.7845_{\mathrm{R1205A2}}+0.8884_{\mathrm{R809D}})}}{1+e^{(2.8866-1.0555_{\mathrm{R1006B}}-1.5543_{\mathrm{R1005C}}-1.2868_{\mathrm{R1003B}}+0.7845_{\mathrm{R1205A2}}+0.8884_{\mathrm{R809D}})}}$$

$$g(x) = 2.8866-1.0555_{\mathrm{R1006B}} - 1.5543_{\mathrm{R1005C}} - 1.2868_{\mathrm{R1003B}} +0.7845_{\mathrm{R1205A2}}+0.8884_{\mathrm{R809D}}$$

Model terbaik pada dataset yang lain terlampir pada Lampiran 13.

5. *Odds Ratio*

Berikut pada Tabel 18 adalah hasil *Odds Ratio* dari masing-masing variabel yang berpengaruh atau signifikan.

Tabel 18. Hasil *Odds Ratio* Dataset 1500 Acak 85:15

Variabel	β	*Odds Ratio*
(Konstanta)	2.8866	
R1006B	-1.0555	0.3480
R1005C	-1.5543	0.2113
R1003B	-1.2868	0.2761
R1205A2	0.7845	0.4563
R809D	0.8884	2.4314

Hasil *Odds Ratio* pada dataset yang lain terlampir pada Lampiran 14.

6. Probabilitas

Berikut pada Gambar adalah hasil prediksi 5 teratas dataset 1500 acak menggunakan nilai probabilitas. Hasil prediksi selengkapnya terlampir pada Lampiran 15.

Tabel 19. Hasil Prediksi *Logistic Regression* Dataset 1500 Acak 85:15

Nama Provinsi	Nama Desa	Aktual	Prediksi	Probabilitas Prediksi
Bengkulu	Srikuncoro	1	1	0.9158978214
Di Yogyakarta	Demangan	0	1	0.9402468487
Kalimantan Barat	Jongkong Kiri Tengah	1	1	0.9188260315
Sumatera Selatan	Jejawi	1	1	0.910647641
Jawa Tengah	Kaligentong	1	1	0.951811668

5.1.5 *Interpretation / Evaluation*

Evaluasi dalam penelitian ini dilakukan menggunakan *Confusion Matrix* lalu hasilnya ditampilkan berupa tabel perbandingan. Berikut pada Tabel adalah hasil *Confusion Matrix*

Tabel 20. *Confusion Matrix Random Forest*

Dataset			*Accuration*	*Recall*	*Specificity*	*Precision*	*F-Measure*
90:10	**Acak**	1500	0.92	1	0.0769	0.9194	0.0769
		1550	0.8774	0.9926	0.05	0.8881	0.9374
		1600	0.8438	0.9925	0.076	0.8471	0.914
		1650	0.8182	1	0.09	0.814	0.8975
	Proposional	1500	0.8933	1	0	0.8933	0.9436
		1550	0.8645	1	0	0.8645	0.9273
		1600	0.8125	1	0	0.8125	0.8966
		1650	0.8606	1	0.148	0.8571	0.923
85:15	**Acak**	**1500**	**0.9244**	**1**	**0.1052**	**0.9237**	**0.9603**
		1550	0.8712	1	0.0909	0.8695	0.9302
		1600	0.8667	0.9951	0.1142	0.8680	0.9272
		1650	0.8306	1	0.1063	0.8271	0.9054
	Proposional	1500	0.9022	1	0	0.9022	0.9485
		1550	0.8712	0.9950	0	0.8750	0.9311
		1600	0.8458	1	0.0512	0.8445	0.9157
		1650	0.8589	1	0.1250	0.8559	0.9223
80:20	**Acak**	1500	0.9133	0.9927	0.0769	0.9189	0.9543
		1550	0.8806	1	0.0975	0.8790	0.9356
		1600	0.8594	0.9963	0.0638	0.8607	0.9235
		1650	0.8112	1	0.0781	0.8184	0.9001
	Proposional	1500	0.9033	0.9963	0	0.9063	0.9492
		1550	0.8839	1	0	0.8839	0.9383
		1600	0.85	1	0.0769	0.8481	0.9178
		1650	0.8364	0.9926	0.1186	0.8380	0.9087
70:30	**Acak**	1500	0.9111	0.9975	0.0714	0.9125	0.9531
		1550	0.8753	0.9975	0.0806	0.8758	0.9327
		1600	0.8542	0.9975	0.0281	0.8553	0.9209
		1650	0.8222	0.9925	0.0659	0.8251	0.9011
	Proposional	1500	0.9178	1	0.0263	0.9175	0.9570
		1550	0.8903	1	0.0377	0.8898	0.9417
		1600	0.8562	1	0.0675	0.8547	0.9216
		1650	0.8525	0.9951	0.1012	0.8536	0.9189

Berdasarkan hasil *Confusion Matrix* untuk *Random Forest* pada Tabel di atas, dataset 1500 acak dengan pembagian dataset 85:15 menghasilkan kinerja yang paling unggul dari nilai *Accuration*, *Precision*, dan *F-Measure*.

Tabel 21. *Confusion Matrix Logistic Regression*

Dataset			*Accuration*	*Recall*	*Specificity*	*Precision*	*F-Measure*
90:10	**Acak**	1500	0.9133	0.9927	0.0769	0.9189	0.9543
		1550	0.8903	1	0.055	0.8896	0.9416
		1600	0.8438	1	0.038	0.8428	0.9147
		1650	0.8121	0.98485	0.1212	0.8176	0.8935
	Proposional	1500	0.8667	0.9701	0	0.8904	0.9285
		1550	0.871	1	0.0476	0.8701	0.9305
		1600	0.8125	1	0	0.8125	0.8966
		1650	0.8606	0.9927	0.1852	0.8616	0.9225
85:15	**Acak**	**1500**	**0.9156**	**0.9902**	**0.1052**	**0.9230**	**0.9555**
		1550	0.8627	0.995	0.0606	0.8652	0.9255
		1600	0.8583	1	0.0285	0.8577	0.9234
		1650	0.8185	0.9850	0.1063	0.8250	0.8979
	Proposional	1500	0.8933	0.9901	0	0.9013	0.9436
		1550	0.8841	1	0.0689	0.8831	0.9379
		1600	0.8375	1	0	0.8375	0.91156
		1650	0.8508	0.9807	0.1750	0.8607	0.9168
80:20	**Acak**	1500	0.9133	0.9927	0.0769	0.9189	0.9543
		1550	0.8677	0.9962	0.0243	0.8701	0.9289
		1600	0.8562	0.9963	0.0425	0.8580	0.9220
		1650	0.8121	0.9887	0.0781	0.8167	0.8945
	Proposional	1500	0.90	0.9926	0	0.9060	0.9473
		1550	0.8871	0.9963	0.0555	0.8892	0.9397
		1600	0.8438	1	0.0384	0.8427	0.9146
		1650	0.8333	0.9926	0.1016	0.8354	0.9072
70:30	**Acak**	1500	0.9067	0.9950	0.0476	0.9103	0.9508
		1550	0.8667	1	0	0.8667	0.9285
		1600	0.8562	1	0.0281	0.8556	0.9222
		1650	0.8121	0.9801	0.0659	0.8232	0.8949
	Proposional	1500	0.9044	0.9854	0.0263	0.9164	0.9497
		1550	0.8882	0.9975	0.0377	0.8896	0.9405
		1600	0.8542	0.9950	0.0810	0.8559	0.9202
		1650	0.8465	0.9807	0.1392	0.8571	0.9147

Berdasarkan hasil *Confusion Matrix* untuk *Logistic Regression* pada Tabel di atas, dataset 1500 acak dengan pembagian dataset 85:15 menghasilkan kinerja yang paling unggul dari nilai *Accuration*, *Precision*, dan *F-Measure*.

5.2 Pembahasan

Topik yang diangkat pada penelitian ini adalah prediksi potensi desa cerdas di Indonesia, yang dimana data yang digunakan bersumber dari BPS tahun 2021. Penelitian dengan topik potensi desa cerdas sebelumnya pernah dilakukan oleh Tosida *et al*, (2021) dan Tosida *et al*, (2020). Berdasarkan penjelasan sebelumnya, penelitian ini bertujuan untuk memprediksi data potensi desa yang berpotensi menjadi desa cerdas.

Algoritma yang diimplementasikan pada penelitian ini menggunakan *Random Forest* dan *Logistic Regression*. Kedua algoritma bisa digunakan pada data yang besar dan juga efektif baik pada data diskrit maupun kontinu. Oleh karena itu, penulis menggunakan kedua algoritma tersebut untuk proses prediksi dengan melakukan percobaan menggunakan 8 dataset pada masing – masing algoritma. Untuk perbandingan kinerja kedua algoritma menggunakan *confusion matrix* terdapat pada Tabel 20 dan Tabel 21.

Pengambil keputusan pada penelitian ini lebih menginginkan terjadinya *True Positive* dan sangat tidak menginginkan terjadinya *False Positive*. Oleh karena itu, penulis memilih algoritma yang *precision*-nya tinggi. Kondisi yang sangat diminimalisir adalah memprediksi bahwa desa tersebut berpotensi desa cerdas padahal sebenarnya desa tersebut tidak berpotensi desa cerdas. Karena kondisi tersebut dapat menyebabkan sumber daya tidak dialokasikan secara efisien dan dapat menyebabkan pemborosan sumber daya. Penelitian dengan menggunakan nilai *Precision* sebagai parameter untuk pengambilan kesimpulan peforma algoritma juga pernah dilakukan oleh (Effendi *et al*., 2023) dan (Kristiawan & Widjaja, 2021).

Berdasarkan Tabel 20, nilai *Precision* paling tinggi untuk *Random Forest* dihasilkan oleh dataset 1500 acak dengan pembagian dataset 85:15 yaitu sebesar **0.9237 atau 92.37%** pada saat n_estimator = 200, *max_depth* = 8, *criterion* = Entropy, *min_samples_leaf* = 2, *max_features* = Log2, dan *min_samples_split* = 2. Dari hasil penelitian Fai *et al*, (2023), nilai n_estimator lebih baik rendah karena jika ditingkatkan dapat menyebabkan *overfitting*, kondisi tersebut juga berlaku pada nilai *max_features* jika ditingkatkan akan menyebabkan *overfitting* dan jika dikurangi dapat membuat model kurang sensitive terhadap *noise*. Nilai *max_depth* yang lebih tinggi menghasilkan model yang lebih kompleks, namun pada saat bersamaan juga dapat menyebabkan *overfitting* (Minastireanu & Mesnita, 2019). Menurut Hashemi *et al*, (2023), nilai *min_samples_split* tinggi menghasilkan model yang lebih sederhana, sedangkan nilai yang rendah dapat mengurangi *overfitting* namun model tidak dapat menangkap pola cukup baik yang dapat menyebabkan bias dan untuk nilai *min_samples_leaf* lebih baik tinggi karena dapat mencegah *overfitting* dan memperhalus model. Lalu, nilai *precision* paling tinggi pada *Logistic Regression* juga dihasilkan oleh dataset 1500 acak dengan pembagian dataset 85:15 dengan nilai sebesar **0.9230 atau 92.30%** pada saat nilai C = 0.59948 yang di mana menurut (Tanone & Emmanuel, 2020) regularisasi yang lebih kuat ditunjukkan oleh nilai C yang lebih rendah.

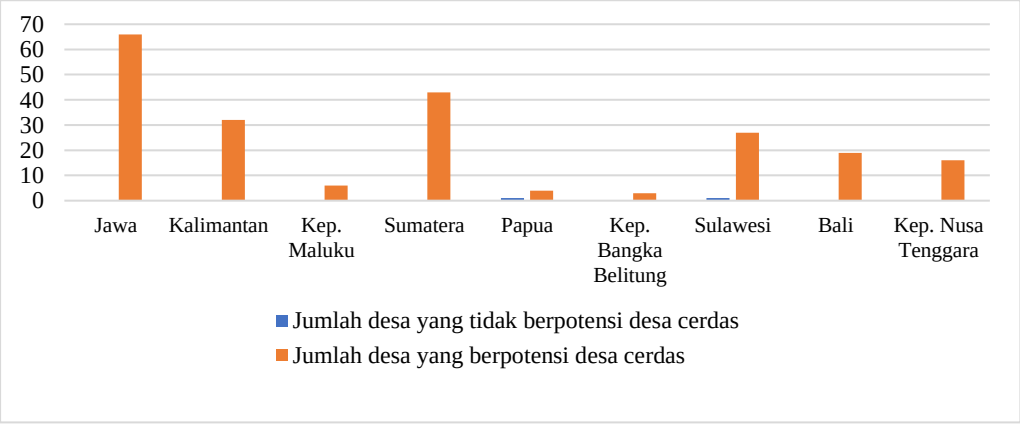
Variabel yang mempengaruhi prediksi dari algoritma *Random Forest* dan *Logistic Regression* yang nilai *Precision*-nya paling tinggi adalah sebagai berikut:

Tabel 22. Variabel yang Mempengaruhi Prediksi

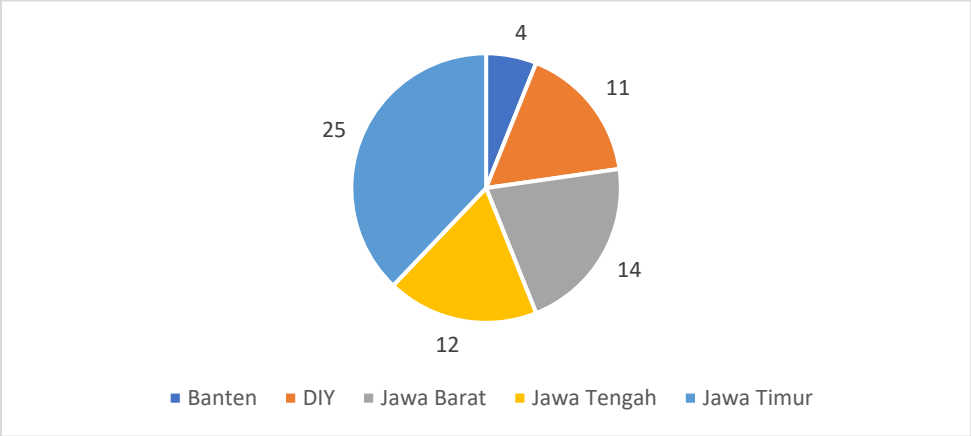
Algoritma	Variabel yang Mempengaruhi Prediksi
Random Forest	R809D (Jumlah kelompok tani)
	R1005C (Sinyal telepon seluler handphone di sebagian besar wilayah desa kelurahan)
	R1006B (Fasilitas internet di kantor kepala desa lurah)
	R809B (Jumlah karang taruna)
	R809F (Jumlah kelompok Masyarakat)
Logistic Regression	R1006B (Fasilitas internet di kantor kepala desa lurah)
	R1005C (Sinyal telepon seluler handphone di sebagian besar wilayah desa kelurahan)
	R1003B (Keberadaan warga yang menggunakan telepon seluler handphone HP)
	R1205A2 (Jumlah Bank Umum Swasta)
	R809D (Jumlah kelompok tani)

Dari Tabel 22 di atas, didapatkan variabel R809D (Jumlah kelompok tani), R1005C (Sinyal telepon seluler handphone di sebagian besar wilayah desa/kelurahan), dan R1006B (Fasilitas internet di kantor kepala desa lurah) muncul di kedua algoritma sebagai variabel yang mempengaruhi prediksi potensi desa cerdas. Menurut Fatimah *et al* (2020), dengan memanfaatkan Teknologi Informasi dan Komunikasi (TIK), kelompok tani dapat bekerja sama, mencari informasi, dan memasarkan produk mereka, yang di mana dapat meningkatkan produktivitas pertanian dan perekonomian desa. Hal tersebut sesuai dengan pengertian desa cerdas menurut (Prasetya *et al*, 2022), yaitu desa yang dapat memajukan kesejahteran masyarakat dan kualitas hidup masyarakat dengan penggunaan teknologi informasi. Maka variabel R809D (Jumlah kelompok tani), R1006B (Fasilitas internet di kantor kepala desa lurah), dan R1005C (Sinyal telepon seluler handphone di sebagian besar wilayah desa kelurahan) berpengaruh terhadap prediksi potensi desa cerdas.

Hasil perbandingan kinerja kedua algoritma menunjukkan bahwa algoritma *Random Forest* lebih unggul dari pada algoritma *Logistic Regression* dalam memprediksi potensi desa cerdas. Maka, hasil prediksi yang digunakan adalah prediksi *Random Forest* dataset 1500 acak dengan pembagian data 85:15. Hasil prediksi terlampir pada Lampiran 7.



Gambar 16. Kelompok Desa Berdasarkan Pulau



Gambar 17. Jumlah Desa Berpotensi Desa Cerdas Per Provinsi di Pulau Jawa

Berdasarkan Gambar 16 dan Gambar 17, Pulau Jawa menjadi pulau yang jumlah desa berpotensi desa cerdasnya paling banyak yang berjumlah 66, yang di mana Provinsi Jawa Timur menjadi provinsi yang paling banyak yaitu sebanyak 25. Sedangkan kelompok desa yang tidak berpotensi desa cerdas hanya terdapat di Pulau Papua dan Pulau Sulawesi dengan jumlahnya masing-masing yaitu 1. Menurut survey Asosiasi Penyelenggara Jasa Internet Indonesia (APJII) tahun 2023, 3 dari 5 provinsi yang penetrasi pengguna internetnya tinggi adalah provinsi yang terletak di Pulau Jawa, yaitu Provinsi Banten, Provinsi Jawa Barat, dan Provinsi Jawa Timur. Adapun menurut Indeks Pembangunan Teknologi Informasi dan Komunikasi dari BPS, Provinsi DIY menjadi provinsi kategori subindeks akses dan infrastruktur TIK tinggi selama 2020–2021, sedangkan provinsi lain yang terletak di Pulau Jawa menempati kategori subindeks akses dan infrastruktur TIK sedang. Kemudian berdasarkan Buku Statistik Penyuluhan Pertanian, 3 dari 5 provinsi yang presentase jumlah kelompok tani paling tinggi adalah provinsi yang terletak di pulau jawa, yaitu Provinsi Jawa Tengah, Provinsi Jawa Barat, dan Provinsi Jawa Timur. Hal tersebut berkesinambungan antara variabel yang yang mempengaruhi prediksi dengan Pulau Jawa yang unggul pada hasil prediksi potensi desa cerdas.

Adapun dari hasil prediksi potensi keseluruhan desa cerdas, desa yang diidentifikasi memiliki potensi besar untuk menjadi desa cerdas umumnya memiliki jumlah kelompok tani sebanyak 4, sinyal telepon seluler/HP yang kuat, fasilitas internet di kantor kepala desanya berfungsi, jumlah karang taruna per desanya sebanyak 1, dan tidak memiliki kelompok masyarakat per desanya.

BAB VI
KESIMPULAN DAN SARAN

6.1 Kesimpulan

Berdasarkan hasil penelitian yang telah dilakukan, dapat disimpulkan bahwa algoritma *Random Forest* dan *Logistic Regression* dapat digunakan untuk memprediksi potensi desa cerdas. Algoritma *Random Forest* menghasilkan prediksi melalui majority voting, sedangkan *Logistic Regression* menghasilkan prediksi menggunakan nilai probabilitas yang dimana jika probabilitas < 0.5 akan masuk kelas 0 sedangkan ≥ 0.5 masuk ke kelas 1.

Data yang digunakan pada penelitian ini bersumber dari Badan Pusat Statistik 2021 menggunakan 29 variabel dan 8 percobaan dataset yang diambil secara acak dan proposional dengan pembagian data *training*:data *testing* sebanyak 90:10, 85:15, 80:20, dan 70:30. Berdasarkan hasil kinerja *Confusion Matrix*, algoritma *Random Forest* menghasilkan nilai *Precision* paling tinggi sebesar **0.9237 atau 92.37%** yang di mana lebih unggul dari pada algoritma *Logistic Regression* dengan nilai *Precision* sebesar **0.9230 atau 92.30%**, dan keduanya dihasilkan dari dataset 1500 acak dengan pembagian dataset 85:15. Didapatkan parameter optimal *Random Forest* adalah pada saat n_estimator = 200, *max_depth* = 8, *criterion* = Entropy, *min_samples_leaf* = 2, *max_features* = Log2, dan *min_samples_split* = 2, sedangkan parameter dari *Logistic Regression* pada saat nilai C = 0.59948 dan penalty = L2. Adapun tiga variabel yang mempengaruhi prediksi yang muncul di kedua algoritma, yaitu variabel R809D (Jumlah kelompok tani), R1005C (Sinyal telepon seluler handphone di sebagian besar wilayah desa/kelurahan), dan R1006B (Fasilitas internet di kantor kepala desa lurah). Dari hasil prediksi algoritma yang lebih unggul yaitu *Random Forest*, desa yang jumlah berpotensi desa cerdasnya paling banyak terletak di Pulau Jawa, tepatnya di Provinsi Jawa Timur. Desa yang diidentifikasi memiliki potensi besar untuk menjadi desa cerdas umumnya memiliki jumlah kelompok tani sebanyak 4, sinyal telepon seluler/HP yang kuat, fasilitas internet di kantor kepala desanya berfungsi, jumlah karang taruna per desanya sebanyak 1, dan tidak memiliki kelompok masyarakat per desanya. Kesimpulan dalam berbentuk tabel terlampir pada Lampiran 16.

Kementerian Desa, Pembangunan Daerah Tertinggal, dan Transmigrasi serta pemangku kepentingan lainnya dapat mengetahui pulau mana dan provinsi mana yang paling banyak berpotensi menjadi desa cerdas, agar dapat mengarahkan sumber daya dan kebijakan secara lebih tepat guna untuk mendukung perkembangan pedesaan yang lebih merata, namun hasil tersebut perlu dikaji ulang kembali.

6.2 Saran

Diharapkan untuk penelitian selanjutnya dapat menggunakan metode klasifikasi yang lain untuk memprediksi potensi desa cerdas dan menggunakan dataset yang lebih besar.

DAFTAR PUSTAKA

Aini, Z. 2023. Implementasi *Random Forest* Dan Gradient Boosting Pada Klasifikasi Indeks Pembangunan Manusia (IPM). Skripsi. Jurusan Matematika FST UIN Syarif Hidayatullah, Jakarta.

Apriliah, W., Kurniawan, I., Baydhowi, M., & Haryati, T. (2021). SISTEMASI: Jurnal Sistem Informasi Prediksi Kemungkinan Diabetes pada Tahap Awal Menggunakan Algoritma Klasifikasi *Random Forest*. *Jurnal Sistem Informasi*, *10*, 163–171. http://sistemasi.ftik.unisi.ac.id

Bahtiar, R. (2023). Implementasi Data Mining Untuk Prediksi Penjualan Kusen Terlaris Menggunakan Metode K-Nearest Neighbor. *Jurnal Informatika MULTI*, *1*(3), 203–214. https://jurnal.publikasitecno.id/index.php/jim203

Breiman, L. (2001). *Random Forest*s. In Machine Learning (Vol. 45, Issue 1). Cambridge University Press; 1st edition. https://doi.org/doi.org/10.1023/A:1010933404324

Briandy, B. T. P., Yulianingsih, E., Fatmasari, & Ferdiansyah. (2023). Analisis Tingkat *Accuration* Prediksi Gejala COVID - 19 Dengan Menggunakan Metode *Logistic Regression* dan Support Vector Machine. *Jurnal Fasilkom*, *13*(02), 269–278. https://doi.org/10.37859/jf.v13i02.5629

Cahyana, C. W., & Nurlayli, A. (2023). Analisis Performa *Logistic Regression*, Naïve Bayes, dan *Random Forest* sebagai Algoritma Pendeteksi Kanker Payudara. *INSERT: Information System and …*, *4*(1), 51–64. https://ejournal.undiksha.ac.id/index.php/insert/article/view/62362%0Ahttps://ejournal.undiksha.ac.id/index.php/insert/article/download/62362/26415

Cahyani, Q. R., Finandi, M. J., Rianti, J., Arianti, D. L., Dwi, A., Putra, P., & Artikel, G. (2022). Prediksi Risiko Penyakit Diabetes menggunakan Algoritma Regresi Logistik Diabetes Risk Prediction using *Logistic Regression* Algorithm Article Info ABSTRAK. *JOMLAI: Journal of Machine Learning and Artificial Intelligence*, *1*(2), 2828–9099. https://doi.org/10.55123/jomlai.v1i2.598

Cahyaningtyas, C., Manongga, D., & Sembiring, I. (2022). Algorithm Comparison and Feature Selection for Classification of Broiler Chicken Harvest. *Jurnal Teknik Informatika (Jutif)*, *3*(6), 1717–1727. https://doi.org/10.20884/1.jutif.2022.3.6.493

Christy, E., Suryowati, K., Statistika, J., Sains Terapan, F., & AKPRIND Yogyakarta, I. (2021). Analisis Klasifikasi Status Bekerja Penduduk Daerah Istimewa Yogyakarta Menggunakan Metode *Random Forest*. *Jurnal Statistika Industri Dan Komputasi*, *6*(1), 69–76.

Dani, M., & Setiawati, D. (2022). *Sistem Informasi Desa Kiringan Berbasis Website Menuju Desa Cerdas Menggunakan Metode Prototype*. *6*(2), 52–59.

Data Statistik Penyuluhan Pertanian. (2020). Badan Penyuluhan dan Pengembangan Sumber Daya Manusia Pertanian Kementrian Pertanian, Jakarta.

E.-A. MINASTIREANU and G. MESNITA, "Light GBM Machine Learning Algorithm to Online Click Fraud Detection," *J. Inf. Assur. Cybersecurity*, no. April, pp. 1–12, 2019, doi: 10.5171/2019.263928.

Edris Effendi, M., Yuadi, I., & Puspitasari, I. (2023). Prediksi Guru Kemungkinan Tetap Bekerja di Sekolah Al Uswah Surabaya Menggunakan Machine Learning. *Jurnal Informasi Dan Teknologi*, *5*(1), 129–137.

https://doi.org/10.37034/jidt.v5i2.361

Fatimah, S., Judawinata, M. G., Barkah, M. N., Trimo, L., & Deliana, Y. (2020). Towards Smart Village: A Case Study of Genteng Village Development in Sumedang, West Java, Indonesia. *Society*, *8*(2), 663–676. https://doi.org/10.33019/society.v8i2.264

Gde Agung Brahmana Suryanegara, Adiwijaya, & Mahendra Dwifebri Purbolaksono. (2021). Peningkatan Hasil Klasifikasi pada Algoritma *Random Forest* untuk Deteksi Pasien Penderita Diabetes Menggunakan Metode Normalisasi. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *5*(1), 114–122. https://doi.org/10.29207/resti.v5i1.2880

Greessheilla Phylosta P.B, & Rido Febryansyah. (2022). Permohonan Pinjaman Pada Koperasi Simpan Pinjam. *Ilmudata.Org*, *2*(12), 1–12.

Handayani, F. (2021). Komparasi Support Vector Machine, *Logistic Regression* Dan Artificial Neural Network Dalam Prediksi Penyakit Jantung. *Jurnal Edukasi Dan Penelitian Informatika (JEPIN)*, *7*(3), 329. https://doi.org/10.26418/jp.v7i3.48053

Harlan, J. (2018). Analisis Regresi Logistik. Gunadarma, Depok.

Hosmer, D. W., & Lemeshow, S. (2000). Applied Logistic Regresion. John Wiley & Sons Inc, New York

Kristiawan, K., & Widjaja, A. (2021). Perbandingan Algoritma Machine Learning dalam Menilai Sebuah Lokasi Toko Ritel. *Jurnal Teknik Informatika Dan Sistem Informasi*, *7*(1), 35–46. https://doi.org/10.28932/jutisi.v7i1.3182

Lumbanraja, F. R., Mudyaningsih, W., Hermanto, B., Syarif, A., & Komputer, J. I. (2019). Implementasi Metode *Random Forest* Untuk Prediksi Posisi Metilasi Pada Sekuens Protein. *Seminar Nasional Sains, Matematika, Informatika, Dan Aplikasinya*, 105–112.

Machali, I. (2021). Metode Penelitian Kuantitatif Panduan Praktis Merencanakan, Melaksanakan Dan Analisis Dalam Penelitian Kuantitatif. Fakultas Ilmu Tarbiyah dan Keguruan Universitas Islam Negeri (UIN) Sunan Kalijaga Yogyakarta, Yogyakarta.

Marlina, D., & Bakri, M. (2021). Penerapan Data Mining Untuk Memprediksi Transaksi Nasabah Dengan Algoritma C4.5. *Jurnal Teknologi Dan Sistem Informasi (JTSI)*, *2*(1), 23–28.

Martadala, D. A., Neneng, Susanto, E. R., & Ahmad, I. (2021). Model Desa Cerdas Dalam Pelayanan Administrasi (Studi Kasus: Desa Kotabaru Barat Kecamatan Martapura Kabupaten Oku Timur). *Jurnal Teknologi Dan Sistem Informasi (JTSI)*, *2*(2), 40–51. http://jim.teknokrat.ac.id/index.php/JTSI

Maulidah, M., Windu Gata, Rizki Aulianita, & Cucu Ika Agustyaningrum. (2020). Algoritma Klasifikasi *Decision tree* Untuk Rekomendasi Buku Berdasarkan Kategori Buku. *E-Bisnis : Jurnal Ilmiah Ekonomi Dan Bisnis*, *13*(2), 89–96. https://doi.org/10.51903/e-bisnis.v13i2.251

Nilwanda, L. E., Arifiyanti, A. A., & Hadiwiyanti, R. (2024). 2024 Madani : Jurnal Ilmiah Multidisipline Penerapan Algoritma *Decision tree* Pada Penentuan Penerima Program Keluarga Harapan di Desa Turirejo , Kedamean Gresik 2024

Madani : Jurnal Ilmiah Multidisipline. *Jurnal Ilmiah Multidisiplin*, *2*(1), 221–230. https://jurnal.penerbitdaarulhuda.my.id/index.php/MAJIM/article/view/1673

N. J. Fai, W. K. Wey, and G. J. Xian, "Digits Classification Using Random Forest Classifier," vol. 7, no. 3, 2023.

Nurzanah, S. C., Alam, S., & Hermanto, T. I. (2022). *ANALISIS ASSOCIATION RULE UNTUK IDENTIFIKASI POLA GEJALA PENYAKIT HIPERTENSI MENGGUNAKAN ALGORITMA APRIORI (STUDI KASUS : KLINIK RAFINA MEDICAL CENTER) ASSOCIATION RULE ANALYSIS FOR IDENTIFICATION OF HYPERTENSION SYMPTOMS PATTERNS USING APRIORI ALGORITH*. *5*(2), 132–141. https://doi.org/10.33387/jiko

Oberoi, A., Sehgal, O., & Malik, C. (2023). Optimizing a Binary *Logistic Regression* model by Hyperparameter tuning. *International Journal of Scientific and Research Publications*, *13*(7), 120–127. https://doi.org/10.29322/ijsrp.13.07.2023.p13913

Peerbasha, S., Raja, A. S., Praveen, K. P., Iqbal, M., & Surputheen, M. (2023). Diabetes Prediction using *Decision tree, Random Forest*, Support Vector 38 Machine, K- Nearest Neighbors, *Logistic Regression* Classifiers. Journal of Advanced Applied Scientific Research, 5-4, 42-54

Pradana, D., Luthfi Alghifari, M., Farhan Juna, M., & Palaguna, D. (2022). Klasifikasi Penyakit Jantung Menggunakan Metode Artificial Neural Network. *Indonesian Journal of Data and Science*, *3*(2), 55–60. https://doi.org/10.56705/ijodas.v3i2.35

Prasetyo, R., Nawawi, I., Fauzi, A., & Ginabila, G. (2021). Komparasi Algoritma *Logistic Regression* dan *Random Forest* pada Prediksi Cacat Software. *Jurnal Teknik Informatika UNIKA Santo Thomas*, *06*(Siringoringo 2017), 275–281. https://doi.org/10.54367/jtiust.v6i2.1522

R. Tanone and A. B. Emmanuel, "Prediction of Not Operational Transaction Using Logistic Regression at XYZ Bank in Kupang City (Prediksi Not Operational Transaction Menggunakan Logistic Regression pada Bank XYZ di Kota Kupang)," Aiti, vol. 17, no. 1, pp. 42–55, 2020, doi: 10.24246/aiti.v17i1.42-55.

Rizky Mubarok, M., Herteno, R., Komputer Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lambung Mangkurat Jalan Ahmad Yani Km, I., & Selatan, K. (2022). Hyper-Parameter Tuning Pada Xgboost Untuk Prediksi Keberlangsungan Hidup Pasien Gagal Jantung. *Klik - Kumpulan Jurnal Ilmu Komputer*, *9*(2), 391–401. http://klik.ulm.ac.id/index.php/klik/article/view/484

Rovidatul, Yunus, Y., & Nurcahyo, G. W. (2023). Perbandingan algoritma c4.5 dan naive bayes dalam prediksi kelulusan mahasiswa. *Jurnal CoSciTech (Computer Science and Information Technology)*, *4*(1), 193–199. https://doi.org/10.37859/coscitech.v4i1.4755

Runanto, R., Mislahudin, M. F., Alfiansyah, F. A., Maisun Taqiyyah, M. K., & Tosida, E. T. (2021). Potential classification of Smart Village – Smart Economy with Deep Learning methods. *International Journal of Quantitative Research and Modeling*, *2*(3), 147–162. https://doi.org/10.46336/ijqrm.v2i3.147

Santoso, A. D., Fathin, C. A., Effendi, K. C., Novianto, A., Sumiar, H. R., Angendari, D. A. D., & Putri, B. P. (2019). Desa Cerdas: Transformasi Kebijakan dan Pembangunan Desa Merespon Era Revolusi Industri 4.0. Center for Digital Society, Yogyakarta.

Sundari, M. A., & Pane, R. (2023). *Data Mining Clustering Korban Kejahatan Pelecehan Seksual dengan Kekerasan Berdasarkan Provinsi Menggunakan Metode AHC*. *5*(1). https://doi.org/10.47065/bits.v5i1.3499

Syahranita, R. (2022). Prediksi Kategori Kelulusan Mahasiswa Menggunakan Metode Regresi Logistik Multinomial. Skripsi. Jurusan Teknik Informatika FST Universitas Islam Negeri Maulana Malik Ibrahim, Malang.

Tanone, R., & Emmanuel, A. B. (2020). Prediksi Not Operational Transaction Menggunakan *Logistic Regression* pada Bank XYZ di Kota Kupang. *Aiti*, *17*(1), 42–55. https://doi.org/10.24246/aiti.v17i1.42-55

Tosida, E. T., Suprehatin, S., Herdiyeni, Y., Marimin, & Solihin, I. P. (2020). Clustering of Citizen Science Prospect to Construct Big Data-based Smart Village in Indonesia. Proceedings - 2nd International Conference on Informatics, Multimedia, Cyber, and Information System, ICIMCIS 2020, July 2023, 58–63. https://doi.org/10.1109/ICIMCIS51567.2020.9354323

Tosida, E. T., Wihartiko, F. D., Lingga, A., & Putra, E. (2021). *KLASIFIKASI POTENSI DESA CERDAS*. 1–14.

Umaroh, A. K. (2020). Perbandingan Metode Regresi Logistik Biner Classification And Regression Tree Pada Klasifikasi Status Kesejahteraan Rumah Tangga Di Kota Batu. Skripsi. Jurusan Matematika FST Universitas Islam Negeri Maulana Malik Ibrahim, Malang.

Viji Cripsy, J., & Divya, T. (2023). Lung Cancer Disease Prediction and Classification based on Feature Selection method using Bayesian Network, *Logistic Regression*, J48, *Random Forest*, and Naïve Bayes Algorithms. *Proceedings - 2023 3rd International Conference on Smart Data Intelligence, ICSMDI 2023*, *January 2019*, 335–342. https://doi.org/10.1109/ICSMDI57622.2023.00066

Winarta, A., Kurniawan, W. J., & Komputer, F. I. (2021). *OPTIMASI CLUSTER K-MEANS MENGGUNAKAN METODE ELBOW PADA DATA PENGGUNA NARKOBA DENGAN PEMROGRAMAN*. *5*(1).

Yan, M., & Shen, Y. (2022). Traffic Accident Severity Prediction Based on *Random Forest*. *Sustainability (Switzerland)*, *14*(3), 1–13. https://doi.org/10.3390/su14031729

Yati, R. 2023. Survei APJII: Pengguna Internet di Indonesia Tembus 215 Juta Orang. https://teknologi.bisnis.com/read/20230308/101/1635219/survei-apjii-pengguna-internet-di-indonesia-tembus-215-juta-orang. 29 Maret 2024.

Zaen, N. J. 2019. Diagnosis Penyakit Stroke Dengan Metode Regresi Logistik Biner. Skripsi. Jurusan Matematika FMIPA Universitas Islam Negeri Sunan Ampel, Surabaya.

LAMPIRAN

Lampiran 1. Perbandingan Variabel

No.	Atribut/Variabel Penelitian Saat Ini	Atribut/Variabel Penelitian Terdahulu
1.	R809A (Jumlah jenis-jenis lembaga kemasyarakatan desa: PKK)	R1003A (Jumlah keluarga yang berlangganan telepon kabel)
2.	R809B (Jumlah jenis-jenis lembaga kemasyarakatan desa: Karang taruna)	R1003B (Keberadaan warga yang menggunakan telepon seluler/handphone/HP)
3.	R809C (Jumlah jenis-jenis lembaga kemasyarakatan desa: Lembaga adat)	R1004 (Keberadaan internet untuk warnet, game online, dan fasilitas lainnya di desa kelurahan)
4.	R809D (Jumlah jenis-jenis lembaga kemasyarakatan desa: Kelompok tani)	R1005A (Jumlah menara base transceiver station (BTS))
5.	R809E (Jumlah jenis-jenis lembaga kemasyarakatan desa: Lembaga pengelolaan air)	R1005B (Jumlah operator layanan komunikasi telepon seluler/handphone yang menjangkau)
6.	R809F (Jumlah jenis-jenis lembaga kemasyarakatan desa: Lembaga pengelolaan air)	R1005C (Sinyal telepon seluler/handphone di sebagian besar wilayah desa/kelurahan)
7.	R1205A1 (Jumlah Bank Umum Pemerintah (misalnya: BRI, BNI, MANDIRI, BPD, dll.) yang beroperasi di wilayah desa kelurahan)	R1005D (Sinyal internet GSM atau CDMA telepon seluler/handphone di sebagian besar wilayah di desa/kelurahan)
8.	R1205A2 (Jumlah Bank Umum Swasta (misalnya: Danamon, BCA, Niaga, dll.) yang beroperasi di wilayah desa kelurahan)	R1006A (Komputer/PC/laptop yang masih berfungsi di kantor kepala desa/lurah)
9.	R1205A3 (Jumlah Bank Umum Swasta (misalnya: Danamon, BCA, Niaga, dll.) yang beroperasi di wilayah desa/kelurahan)	R1006B (Fasilitas internet di kantor kepala desa/lurah)
10.	R1206A1 (Jumlah KUD)	R1601B3K2 (Kegiatan : Pengelolaan transportasi Desa)

11.	R1206A2 (Jumlah koperasi Industri Kecil dan Kerajinan Rakyat (Kopinkra) yang masih aktif beroperasi)	R1601B3K3 (Sumber dana: Pengelolaan transportasi desa)
12.	R1206A3 (Jumlah koperasi Simpan Pinjam (Kospin) yang masih aktif beroperasi)	R1601B3K4 (Pelaksana: Pengelolaan transportasi desa)
13.	R1206A4 (Jumlah koperasi lainnya ..... (misalnya koperasi: pegawai, karyawan, pensiunan, sekolah, dll) yang masih aktif beroperasi)	R1601B3K5 (Penerima manfaat langsung : Pengelolaan transportasi desa)
14.	R1007C (Perusahaan agen jasa ekspedisi (pengiriman barang dokumen) swasta)	R1601B4K2 (Kegiatan: Pengembangan energi terbarukan)
15.	R1001C1 (Keberadaan angkutan umum)	R1601B4K3 (Sumber dana: Pengembangan energi terbarukan)
16.	R1001A (Lalu lintas dari ke desa kelurahan melalui)	R1601B4K4 (Pelaksana: Pengembangan energi terbarukan)
17.	R1007A (Keberadaan kantor pos pembantu rumah pos)	R1601B4K5 (Penerima manfaat langsung : Pengembangan energi terbarukan)
18.	R1003B (Keberadaan warga yang menggunakan telepon seluler /handphone/HP)	R1601B5K2 (Kegiatan: Pengelolaan informasi dan komunikasi)
19.	R1004 (Keberadaan internet untuk warnet, game online, dan fasilitas lainnya di desa kelurahan)	R1601B5K3 (Sumber dana: Pengelolaan informasi dan komunikasi)
20.	R1005C (Sinyal telepon seluler/handphone di sebagian besar wilayah desa kelurahan)	R1601B5K4 (Pelaksana: Pengelolaan informasi dan komunikasi)
21.	R1006B (Fasilitas internet di kantor kepala desa lurah)	R1601B5K5 (Penerima manfaat langsung : Pengelolaan informasi dan komunikasi)
22.	R1401A (Keberadaan sistem informasi desa)	R1601B6K2 (Kegiatan: Pengelolaan usaha produktif berbasis pertanian dan industri kecil desa non pertanian)
23.	R1401C (Pembangunan sistem keuangan desa)	R1601B6K3 (Sumber dana: Pengelolaan usaha produktif berbasis pertanian dan

		industri kecil desa non pertanian)
24.	R501C (Keluarga yang menggunakan lampu tenaga surya)	R1601B6K4 (Pelaksana: Pengelolaan usaha produktif berbasis pertanian dan industri kecil desa non pertanian)
25.	R502A (Penerangan di jalan utama desa kelurahan yang menggunakan lampu tenaga surya)	R1601B6K5 (Penerima manfaat langsung Pengelolaan usaha produktif berbasis pertanian dan industri kecil desa non pertanian)
26.	R604A (Sistem peringatan dini bencana alam)	
27.	R604B (Sistem peringatan dini khusus tsunami)	
28.	R808A (Kebiasaan gotong royong warga di desa kelurahan untuk kepentingan umum komunitas)	
29.	R808B (Kegiatan gotong royong warga untuk membantu warga yang sedang mengalami musibah)	

Keterangan:

: Variabel pada penelitian terdahulu Tosida *et al*, (2020) & Tosida *et al*, (2021) yang digunakan pada penelitian saat ini (Salma, 2024)

Lampiran 2. Dataset Penelitian

Dataset penelitian keseluruhan dapat diakses melalui link berikut:

https://drive.google.com/drive/folders/1OpAzqg3bWSzu_eDw6-z7QQEAjMoicM-i?usp=sharing

Lampiran 3. Perhitungan Manual *Random Forest*

ID	R809D	R1206A4	R502A	R604B	R809F	Target
1	0.15217	0	1	0	0.09091	1
2	0.02174	0	1	1	0.01136	0
3	0.34783	0.6	0	0	0.22727	1
4	0.32609	0.6	1	0	0.20455	1
5	0.06522	0	0	0	0.05682	0

***Bootstrapped* Kedua**

R1206A4	R502A	Target
0.6	1	1
0.6	0	1
0.6	0	1
0.6	1	1
0	0	0

Pada *Bootstrapped* dataset kedua, ID yang digunakan yaitu (3 3 5 4 4) dan ID yang tidak digunakan yaitu 1 dan 2. Hasil dari GI(R502A) adalah GI(R502A) = $1 - ((4/5)^2 + (1/5)^2) = 0{,}38$. Karena GI(R502A) dan GI(R1206A4) memiliki nilai yang sama, maka dapat memilih salah satu variabel sebagai *root node*, misalnya R502A. Nilai mean dari R502A adalah 0,4. Maka, data akan terbagi menjadi 2 subset:
Subset 1 (R502A ≤ 0.4)

R1206A4	R502A	Target
0.6	0	1
0.6	0	1
0	0	0

Subset 2 (R502A > 0.4)

R1206A4	R502A	Target
0.6	1	1
0.6	1	1

Leaf node pada subset 2 sudah terbentuk yaitu 1 dikarenakan homogen. Selanjutnya yaitu mencari mean dari R1206A4 dari subset 1, didapatkan mean sebesar 0,4. Sehingga subset yang dihasilkan sebagai berikut
Subset 3 (R1206A4 ≤ 0.4)

R1206A4	R502A	Target
0	0	0

Subset 4 (R1206A4 > 0.4)

R1206A4	R502A	Target
0.6	0	1
0.6	0	1

Jadi, apabila nilai R1206A4 ≤ 0.4, maka hasil leaf node nya 0 dan jika nilai R1206A4 > 0.4, maka hasil leaf nodenya 1.

***Bootstrapped* Ketiga**

R502A	R604B	Target
1	0	1
1	1	0
1	1	0
1	0	1
0	0	0

Pada *Bootstrapped* dataset ketiga, ID yang digunakan yaitu (5 2 2 1 1) dan ID yang tidak digunakan yaitu 3 dan 4. Karena GI(R502A) dan GI(R604B) memiliki nilai yang sama yaitu 0.48, maka dapat memilih salah satu variabel sebagai root node, misalnya R604B. Nilai mean dari R604B adalah 0,4. Maka, data akan terbagi menjadi 2 subset:
Subset 1 (R604B ≤ 0.4)

R502A	R604B	Target
1	0	1
1	0	1
0	0	0

Subset 2 (R604B > 0.4)

R502A	R604B	Target
1	1	0
1	1	0

Leaf node pada subset 2 sudah terbentuk yaitu 0 dikarenakan homogen. Selanjutnya yaitu mencari mean dari R502A dari subset 1, didapatkan mean sebesar 0.67. Sehingga subset yang dihasilkan sebagai berikut
Subset 3 (R502A ≤ 0.67)

R502A	R604B	Target
0	0	0

Subset 4 (R502A > 0.67)

R502A	R604B	Target
1	0	1
1	0	1

Jadi, apabila nilai R502A ≤ 0.67, maka hasil leaf node nya 0 dan jika nilai R502A > 0.67, maka hasil leaf nodenya 1.

***Bootstrapped* Keempat**

R604B	R809F	Target
0	0.09091	1
1	0.01136	0
0	0.22727	1
0	0.05682	0
0	0.05682	0

Pada *Bootstrapped* dataset keempat, ID yang digunakan yaitu (5 5 3 1 2) dan ID yang tidak digunakan yaitu 4. Karena GI(R809F) dan GI(R604B) memiliki nilai yang sama yaitu 0.48, maka dapat memilih salah satu variabel sebagai root node, misalnya R809F. Nilai mean dari R809F adalah 0.088636. Maka, data akan terbagi menjadi 2 subset:
Subset 1 (R809F ≤ 0.088636)

R604B	R809F	Target
1	0.01136	0
0	0.05682	0
0	0.05682	0

Subset 2 (R809F > 0.088636)

R604B	R809F	Target
0	0.09091	1
0	0.22727	1

Dikarenakan kedua subset hanya memiliki 1 kelas, maka tidak perlu melakukan split lebih lanjut dan dapat diambil kesimpulan bahwa subset 1 (R809F ≤ 0.088636) mempunyai target 0, sedangkan subset 2 (R809F > 0.088636) mempunyai target 1.

***Bootstrapped* Kelima**

R809D	R604B	Target
0.02174	1	0
0.02174	1	0
0.34783	0	1
0.32609	0	1

R809D	R604B	Target
0.34783	0	1

Pada *Bootstrapped* dataset kelima ID yang digunakan yaitu (2 2 3 3 4) dan ID yang tidak digunakan yaitu 1 dan 5. Karena GI(R809D) dan GI(R604B) memiliki nilai yang sama yaitu 0.48, maka dapat memilih salah satu variabel sebagai root node, misalnya R604. Nilai mean dari R604B adalah 0,213046. Maka, data akan terbagi menjadi 2 subset:

Subset 1 (R809D $\leq$ 0.213046)

R809D	R604B	Target
0.02174	1	0
0.02174	1	0

Subset 2 (R809D > 0.213046)

R809D	R604B	Target
0.34783	0	1
0.32609	0	1
0.34783	0	1

Dikarenakan kedua subset hanya memiliki 1 kelas, maka tidak perlu melakukan split lebih lanjut dan dapat diambil kesimpulan bahwa subset 1 (R809D $\leq$ 0.213046) mempunyai target 0, sedangkan subset 2 (R809D > 0.213046) mempunyai target 1.

Lampiran 4. Iterasi Estimasi Parameter

Iterasi Kedua

Hasil dari iterasi kedua didapatkan variabel R1005C memiliki p-value yang paling kecil yaitu sebesar 3.80E-07 ≤ alpha (0.05). Maka variabel tersebut berpengaruh terhadap prediksi potensi desa cerdas dan akan dihapus dari iterasi.

Variabel	P-Value
R1206A4	0.58221
R809F	0.09136
R809B	0.44549
R1001A	0.22165
R1206A3	0.33447
R1206A2	0.43521
R1001C1	0.42353
R604A	0.23979
R1007A	0.81524
R501C	0.14756
R1003B	1.61E-05
R502A	0.66959
R604B	0.95446
R1206A1	0.37334
R1205A1	0.34009
R808A	0.28548
R809D	0.01085
R1004	0.38099
R1401A	0.42013
R1205A3	0.48348
R809A	0.4502
R1205A2	0.01364
R1005C	**3.80E-07**
R809C	0.1434
R808B	0.86162
R1401C	0.47803
R809E	0.12017
R1007C	0.63105

Iterasi Ketiga

Hasil dari iterasi ketiga didapatkan variabel R1003B memiliki p-value yang paling kecil yaitu sebesar 0.00424 ≤ alpha (0.05). Maka variabel tersebut berpengaruh terhadap prediksi potensi desa cerdas dan akan dihapus dari iterasi.

Variabel	P-Value
R1206A4	0.672286
R809F	0.148548
R809B	0.518949
R1001A	0.484541

Variabel	P-Value
R1206A3	0.454073
R1206A2	0.423496
R1001C1	0.956233
R604A	0.302234
R1007A	0.406657
R501C	0.391927
R1003B	**0.00424**
R502A	0.953811
R604B	0.911569
R1206A1	0.600353
R1205A1	0.046628
R808A	0.24773
R809D	0.012993
R1004	0.858411
R1401A	0.924011
R1205A3	0.671116
R809A	0.480904
R1205A2	0.004864
R809C	0.150653
R808B	0.73717
R1401C	0.986077
R809E	0.137618
R1007C	0.604474

Iterasi Keempat

Hasil dari iterasi keempat didapatkan variabel R1205A2 memiliki p-value yang paling kecil yaitu sebesar 0.00536 ≤ alpha (0.05). Maka variabel tersebut berpengaruh terhadap prediksi potensi desa cerdas dan akan dihapus dari iterasi.

Variabel	P-Value
R1206A4	0.69375
R809F	0.153028
R809B	0.451991
R1001A	0.520724
R1206A3	0.445864
R1206A2	0.379211
R1001C1	0.936778
R604A	0.326703
R1007A	0.397037
R501C	0.410266
R502A	0.821544
R604B	0.919441
R1206A1	0.644184
R1205A1	0.060434
R808A	0.277103
R809D	0.021933

R1004	0.915066
R1401A	0.951927
R1205A3	0.666795
R809A	0.510924
R1205A2	**0.00502**
R809C	0.154845
R808B	0.789759
R1401C	0.725668
R809E	0.14902
R1007C	0.599489

Iterasi Kelima

Hasil dari iterasi kelima didapatkan variabel R809D memiliki p-value yang paling kecil yaitu sebesar 0.03235 ≤ alpha (0.05). Maka variabel tersebut berpengaruh terhadap prediksi potensi desa cerdas dan akan dihapus dari iterasi.

Variabel	**P-Value**
R1206A4	0.697779
R809F	0.146078
R809B	0.740264
R1001A	0.501331
R1206A3	0.40162
R1206A2	0.385353
R1001C1	0.902056
R604A	0.366702
R1007A	0.71471
R501C	0.385964
R502A	0.920464
R604B	0.988741
R1206A1	0.668156
R1205A1	0.764912
R808A	0.270388
R809D	**0.03235**
R1004	0.691733
R1401A	0.992305
R1205A3	0.315755
R809A	0.344299
R809C	0.166585
R808B	0.945675
R1401C	0.731217
R809E	0.177705
R1007C	0.988016

Iterasi Keenam

Hasil dari iterasi keenam didapatkan tidak ada variabel yang memenuhi kriteria yaitu p-value ≤ alpha (0.05), maka proses iterasi dihentikan.

Variabel	P-Value
R1206A4	0.735756
R809F	0.161287
R809B	0.476746
R1001A	0.648199
R1206A3	0.415725
R1206A2	0.354679
R1001C1	0.877863
R604A	0.35082
R1007A	0.64297
R501C	0.328461
R502A	0.747902
R604B	0.84638
R1206A1	0.724125
R1205A1	0.879318
R808A	0.284191
R1004	0.661192
R1401A	0.928837
R1205A3	0.294252
R809A	0.433711
R809C	0.181375
R808B	0.996383
R1401C	0.54034
R809E	0.290205
R1007C	0.884007

Lampiran 5. Perhitungan Manual Uji Wald

Variabel	β	Std Error
(Konstanta)	2.8866	
R1006B	-1.0555	0.1936
R1005C	-1.5543	0.3909
R1003B	-1.2868	0.3806
R1205A2	0.7845	0.46393
R809D	0.8884	0.3795

Variabel R1005C

$$\left[\frac{\beta_j}{SE(\beta_j)}\right] = \left[\frac{-1.5543}{0.3909}\right] = 2.1751$$

Berdasarkan perhitungan di atas didapatkan nilai $|W|$ = 2.1751> $Z_{\alpha/2}$ (1.96), maka bisa disimpulkan bahwa variabel R1005C mempunyai pengaruh terhadap variabel dependen.

Variabel R1003B

$$\left[\frac{\beta_j}{SE(\beta_j)}\right] = \left[\frac{-1.2868}{0.3806}\right] = -3.3809$$

Berdasarkan perhitungan di atas didapatkan nilai $|W|$ = 3.3809> $Z_{\alpha/2}$ (1.96), maka bisa disimpulkan bahwa variabel R1003B mempunyai pengaruh terhadap variabel dependen.

Variabel R1205A2

$$\left[\frac{\beta_j}{SE(\beta_j)}\right] = \left[\frac{0.7845}{0.4639}\right] = 1.6909$$

Berdasarkan perhitungan di atas didapatkan nilai $|W|$ = 1.6315 atau Sig. (0.00502) < α (0.05) maka bisa disimpulkan bahwa variabel R1205A2 mempunyai pengaruh terhadap variabel dependen.

Variabel R809D

$$\left[\frac{\beta_j}{SE(\beta_j)}\right] = \left[\frac{0.8884}{0.3795}\right] = 2.3409$$

Berdasarkan perhitungan di atas didapatkan nilai $|W|$ = 2.3409> $Z_{\alpha/2}$ (1.96), maka bisa disimpulkan bahwa variabel R1003B mempunyai pengaruh terhadap variabel dependen.

Lampiran 6. Hasil Parameter Terbaik *Random Forest*

Berikut adalah hasil parameter terbaik dari dataset acak 85:15, 80:20, dan 70:30. Untuk hasil lengkapnya pada dataset lain, penulis lampirkan pada link di bawah ini:

https://drive.google.com/file/d/1GBGFgyBmTb0J6OypdMr038yFJz1eg8tb/view?usp=sharing

Parameter Terbaik Dataset Acak 90:10

Parameter	Grid Search Values	Best Parameter Dataset			
		1500	**1550**	**1600**	**1650**
n_estimators	100, 200, 500	100	100	100	200
Max_depth	2, 5, 6, 7, 8	7	7	8	6
Criterion	Entropy, Gini	Gini	Gini	Entropy	Gini
Min_samples_leaf	1, 2, 4	1	1	2	2
Max_features	Auto, log2	Auto	Auto	Auto	Auto
Min_samples split	2, 5, 10	2	2	5	2

Parameter Terbaik Dataset Acak 80:20

Parameter	Grid Search Values	Best Parameter Dataset			
		1500	**1550**	**1600**	**1650**
n_estimators	100, 200, 500	200	100	500	100
Max_depth	2, 5, 6, 7, 8	8	8	7	7
Criterion	Entropy, Gini	Gini	Entropy	Gini	Entropy
Min_samples_leaf	1, 2, 4	1	1	1	1
Max_features	Auto, log2	Auto	Log2	Log2	Auto
Min_samples split	2, 5, 10	5	2	2	2

Parameter Terbaik Dataset Acak 70:30

Parameter	Grid Search Values	Best Parameter Dataset			
		1500	**1550**	**1600**	**1650**
n_estimators	100, 200, 500	500	100	100	100
Max_depth	2, 5, 6, 7, 8	6	8	7	6
Criterion	Entropy, Gini	Gini	Gini	Entropy	Gini
Min_samples_leaf	1, 2, 4	1	2	4	4
Max_features	Auto, log2	Auto	log2	Auto	Auto
Min_samples split	2, 5, 10	10	2	2	2

Lampiran 7. Hasil Prediksi dan Pohon Keputusan *Random Forest*

Prediksi

Berikut adalah sebagian data hasil prediksi *Random Forest* pada data 1500 acak dari dataset 90:10, 85:15, 80:20, dan 70:30. Untuk hasil prediksi lengkapnya pada dataset lain dan pohon keputusan, penulis lampirkan pada link di bawah ini:

https://drive.google.com/drive/folders/1e6vKBCNVzwHOeOyPfiSo-UzgTz48_813?usp=sharing

Prediksi Dataset 1500 Acak 90:10

Nama Provinsi	Nama Desa	Aktual	Prediksi
Bengkulu	Srikuncoro	1	1
Di Yogyakarta	Demangan	0	1
Kalimantan Barat	Jongkong Kiri Tengah	1	1
Sumatera Selatan	Jejawi	1	1
Jawa Tengah	Kaligentong	1	1
Sulawesi Utara	Paslaten	1	1
Bali	Serongga	1	1
Sulawesi Utara	Motongkad Utara	1	1
Bali	Tojan	1	1
Papua	Yuneri	0	0

Prediksi Dataset 1500 Acak 85:15

Nama Provinsi	Nama Desa	Aktual	Prediksi
Bengkulu	Srikuncoro	1	1
Di Yogyakarta	Demangan	0	1
Kalimantan Barat	Jongkong Kiri Tengah	1	1
Sumatera Selatan	Jejawi	1	1
Jawa Tengah	Kaligentong	1	1
Sulawesi Utara	Paslaten	1	1
Bali	Serongga	1	1
Sulawesi Utara	Motongkad Utara	1	1
Bali	Tojan	1	1
Sulawesi Barat	Ralleanak	0	0

Prediksi Dataset 1500 Acak 80:20

Nama Provinsi	Nama Desa	Aktual	Prediksi
Bengkulu	Srikuncoro	1	1
Di Yogyakarta	Demangan	0	1
Kalimantan Barat	Jongkong Kiri Tengah	1	1
Sumatera Selatan	Jejawi	1	1
Jawa Tengah	Kaligentong	1	1
Sulawesi Utara	Paslaten	1	1
Bali	Serongga	1	1
Sulawesi Utara	Motongkad Utara	1	1

Nama Provinsi	Nama Desa	Aktual	Prediksi
Bali	Tojan	1	1
Nusa Tenggara Timur	Gaura	1	0

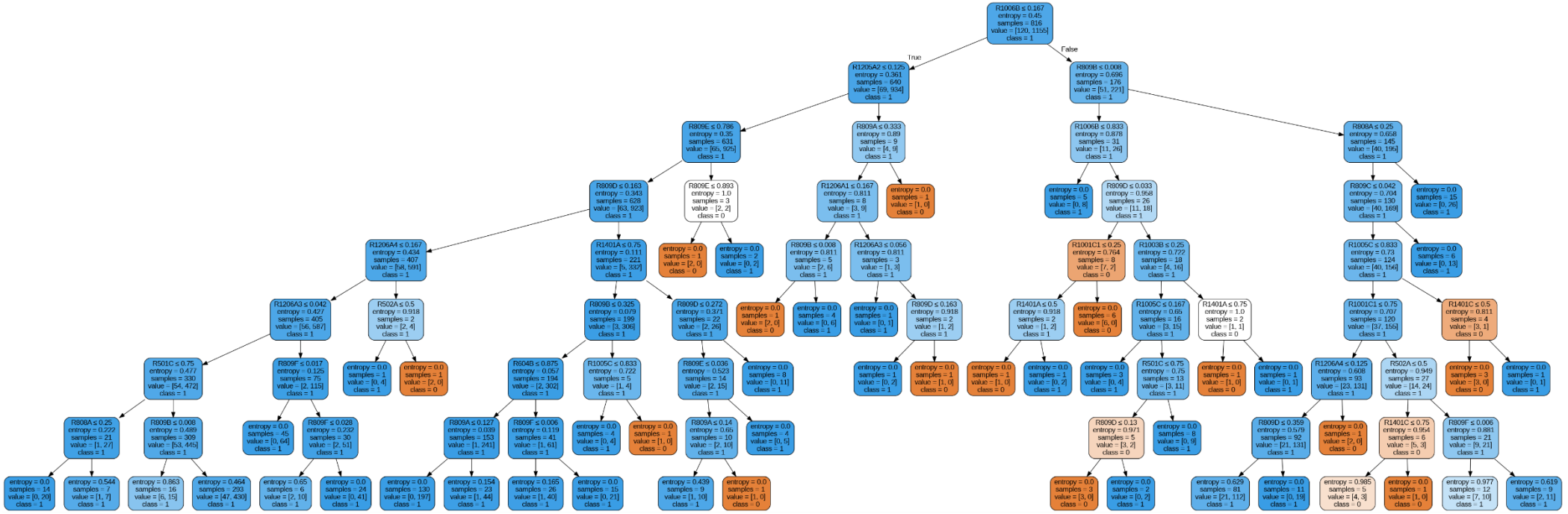
Prediksi Dataset 1500 Acak 70:30

Nama Provinsi	Nama Desa	Aktual	Prediksi
Bengkulu	Srikuncoro	1	1
Di Yogyakarta	Demangan	0	1
Kalimantan Barat	Jongkong Kiri Tengah	1	1
Sumatera Selatan	Jejawi	1	1
Jawa Tengah	Kaligentong	1	1
Sulawesi Utara	Paslaten	1	1
Bali	Serongga	1	1
Sulawesi Utara	Motongkad Utara	1	1
Bali	Tojan	1	1
Papua	Yunat	0	0

Pohon Keputusan

Berikut adalah gambar salah satu pohon keputusan pada data 1500 acak dari dataset 85:15. Untuk pohon keputusan lengkapnya pada dataset lain penulis lampirkan pada link di bawah ini:

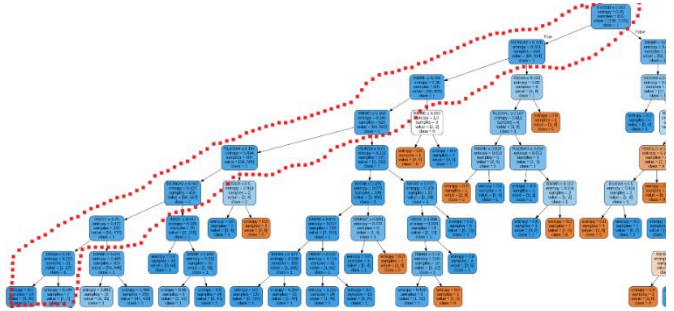
https://drive.google.com/drive/folders/1dsOfyBEAPrDmNEHtRQoryXQWGIsMqzKQ?usp=sharing

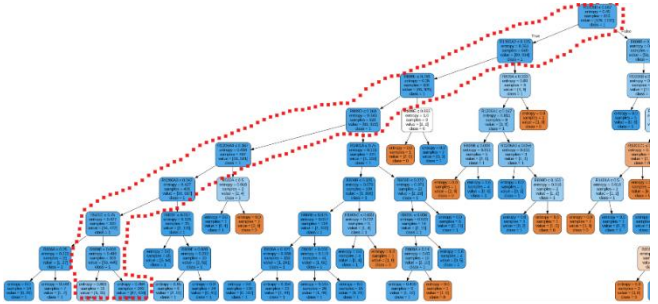
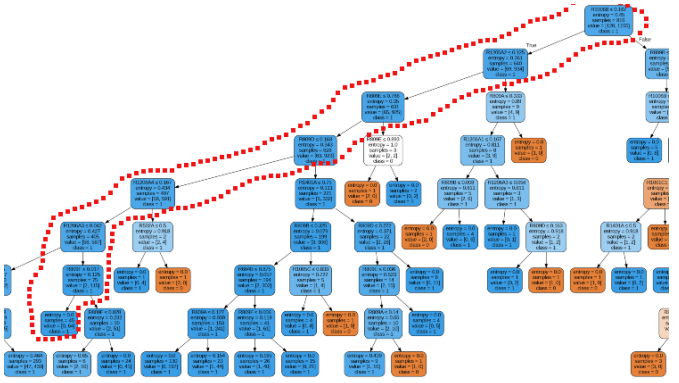


Lampiran 8. Tabel *Rules Random Forest 85:15*

Berikut adalah tabel dari sebagian dari *rules random forest* dari dataset 85:15. Untuk *rules* selengkapnya, penulis lampirkan pada link di bawah ini:

https://drive.google.com/file/d/1Y4mygFgYULUTjZCjty9btIL3NHVe0gM6/view?usp=sharing

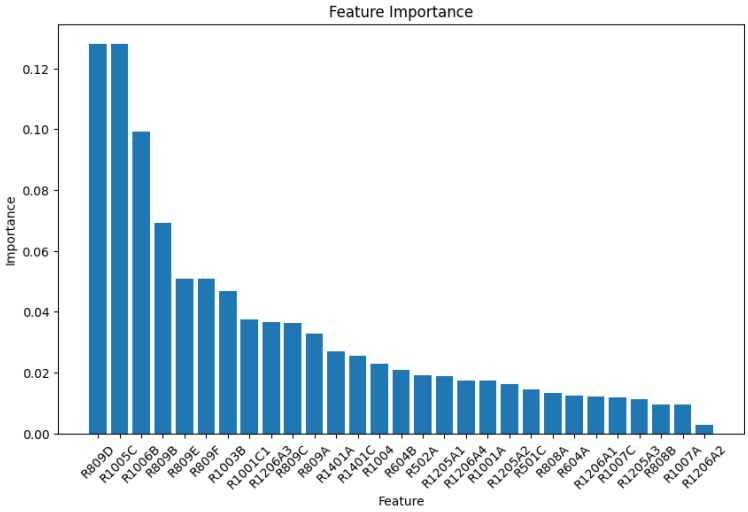
Rule No.	Kondisi	Keputusan	Pohon
1	Fasilitas Internet di Kantor Kepala Desa Berfungsi **AND** Jumlah Bank Umum Swasta <= 1 **AND** Jumlah Lembaga Pengelolaan Air <= 11 **AND** Jumlah Kelompok Tani <= 7 **AND** Jumlah Koperasi Lainnya yang Masih Aktif <= 4 **AND** Jumlah Koperasi Simpan Pinjam yang Masih Aktif <= 1 **AND** Sebagian Besar atau Sebagian Kecil Ada Keluarga yang Menggunakan Lampu Tenaga Surya **AND** (Sebagian Besar Keluarga Terlibat Kebiasaan Gotong Royong Warga di Desa untuk Kepentingan Umum/Komunitas **OR** Sebagian Kecil atau Tidak Ada Kebiasaan Keluarga Terlibat Kebiasaan Gotong Royong Warga di Desa untuk Kepentingan Umum/Komunitas)	Berpotensi Desa Cerdas	

Rule No.	Kondisi	Keputusan	Pohon
2	Fasilitas Internet di Kantor Kepala Desa Berfungsi **AND** Jumlah Bank Umum Swasta <= 1 **AND** Jumlah Lembaga Pengelolaan Air <= 11 **AND** Jumlah Kelompok Tani <= 7 **AND** Jumlah Koperasi Lainnya yang Masih Aktif <= 4 **AND** Jumlah Koperasi Simpan Pinjam yang Masih Aktif <= 1 **AND** Tidak Ada Keluarga yang Menggunakan Lampu Tenaga Surya **AND** (Tidak Ada Karang Taruna OR Jumlah Karang Taruna > 0)	Berpotensi Desa Cerdas	
3	Fasilitas Internet di Kantor Kepala Desa Berfungsi **AND** Jumlah Bank Umum Swasta <= 1 **AND** Jumlah Lembaga Pengelolaan Air <= 11 **AND** Jumlah Kelompok Tani <= 7 **AND** Jumlah Koperasi Lainnya yang Masih Aktif <= 4 **AND** Jumlah Koperasi Simpan Pinjam yang Masih Aktif > 1 **AND** Jumlah Kelompok Masyarakat <= 1	Berpotensi Desa Cerdas	

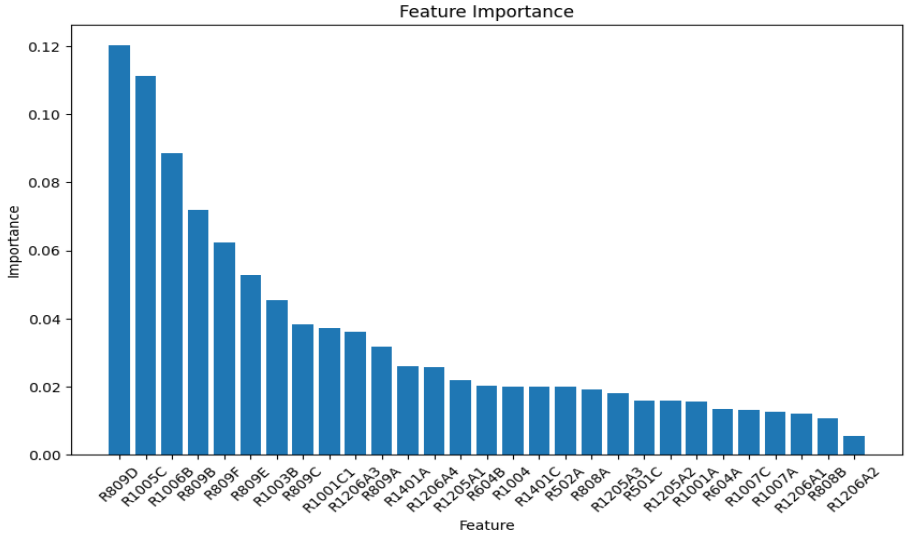
Lampiran 9. *Variable Importance*

Berikut adalah sebagian hasil *variable importance random forest* pada data 1500 acak dari dataset 85:15 dan 80:20. Untuk hasil lengkapnya, peneliti lampirkan pada link di bawah ini.
https://drive.google.com/drive/folders/1BqL3cceitxQfYkxkTViZQYRXOtvfJmZ2?usp=sharing

Dataset 1500 Acak 90:10



Dataset 1500 Acak 80:20



Lampiran 10. Estimasi Parameter

Estimasi Parameter Dataset 90:10

Dataset		Estimasi Parameter
Acak	**1500**	R1006B, R1005C, R1003B, R1205A2, R809D
	1550	R1005C, R1006B, R809D, R809F, R1206A4, R1401A, R1004, R809A
	1600	R1006B, R1005C, R501C, R1206A1, R502A, R809D
	1650	R1006B, R1005C, R604A, R809D, R1401A, R1001A
Proposional	**1500**	R1005C, R1006B, R809D, R809A, R809C, R1401C, R1206A1, R604B, R1001C1
	1550	R1005C, R1006B, R809D, R1003B, R1206A3, R809A
	1600	R1005C, R1003B, R809D, R1006B, R1401C, R604A, R809A, R1206A3, R1206A1
	1650	R1006B, R1005C, R1003B, R809D, R604A, R1001C1, R1401C, R1401A, R809E

Estimasi Parameter Dataset 80:20

Dataset		Estimasi Parameter
Acak	**1500**	R1005C, R1006B, R1003B, R1205A2, R809D
	1550	R1005C, R1006B, R809E, R809F, R1206A4, R809A, R1401A, R1001A
	1600	R1006B, R1005C, R501C
	1650	R1006B, R1005C, R809D, R604A, R1401A, R1206A3
Proposional	**1500**	R1005C, R1006B, R1401C, R809A, R1001C1, R809C, R1004
	1550	R1005C, R809D, R1006B, R1003B, R809A, R1206A3
	1600	R1005C, R1006B, R1003B, R809D, R1206A3, R1205A1, R1206A1, R1401C, R1401A
	1650	R1006B, R1005C, R809D, R604A, R1003B

Estimasi Parameter Dataset 70:30

Dataset		Estimasi Parameter
Acak	**1500**	R1005C, R1006B, R1205A1, R1003B, R809D
	1550	R1006B, R1005C, R809E, R809F, R809A, R1401A
	1600	R1006B, R1005C, R501C
	1650	R1006B, R1005C, R809D, R1401A, R604A
Proposional	**1500**	R1005C, R1006B, R1401C, R809A, R809C, R1004, R1001C1, R1007A, R1206A3, R1206A1
	1550	R1005C, R1006B, R809D, R1206A3, R809A, R1003B, R604B
	1600	R1005C, R1006B, R1003B, R809E, R809D, R1206A3, R1206A1, R1205A1, R809C
	1650	R1006B, R1005C, R809D, R604A, R1003B, R1205A2

Lampiran 11. Parameter Terbaik *Logistic Regression*

Parameter Terbaik Dataset 90:10 Acak

Parameter	Grid Search Values	Best Patameter Dataset			
		1500	1550	1600	1650
Penalty	L1 (Lasso), L2 (Ridge)	L2	L2	L2	L2
C	Logspace (-2,2)	0.59948425 03189409	35.93813663 804626	0.59948425 03189409	4.64158883 36127775

Parameter Terbaik Dataset 90:10 Proposional

Parameter	Grid Search Values	Best Patameter Dataset			
		1500	1550	1600	1650
Penalty	L1 (Lasso), L2 (Ridge)	L2	L2	L2	L2
C	Logspace (-2,2)	35.9381366 3804626	4.64158883 36127775	35.938136 63804626	0.59948425 03189409

Parameter Terbaik Dataset 80:20 Acak

Parameter	Grid Search Values	Best Patameter Dataset			
		1500	1550	1600	1650
Penalty	L1 (Lasso), L2 (Ridge)	L2	L2	L2	L2
C	Logspace (-2,2)	0.599484250 3189409	12.91549665 0148826	4.6415888336 127775	12.91549665 0148826

Parameter Terbaik Dataset 80:20 Proposional

Parameter	Grid Search Values	Best Patameter Dataset			
		1500	1550	1600	1650
Penalty	L1 (Lasso), L2 (Ridge)	L2	L2	L2	L2
C	Logspace (-2,2)	4.6415888336 127775	1.668100537 2000592	0.599484250 3189409	0.599484250 3189409

Parameter Terbaik Dataset 70:30 Acak

Parameter	Grid Search Values	Best Patameter Dataset			
		1500	1550	1600	1650
Penalty	L1 (Lasso), L2 (Ridge)	L2	L2	L2	L2
C	Logspace (-2,2)	0.599484250 3189409	0.01	1.668100537 2000592	4.6415888336 127775

Parameter Terbaik Dataset 70:30 Proposional

Parameter	Grid Search Values	Best Patameter Dataset			
		1500	1550	1600	1650
Penalty	L1 (Lasso), L2 (Ridge)	L2	L2	L2	L2
C	Logspace (-2,2)	12.91549665 0148826	1.668100537 2000592	100.0	100.0

Lampiran 12. Hasil Uji Wald

Berikut adalah sebagian hasil uji wald pada data 1500 acak dari dataset 90:10, 85:15, 80:20, dan 70:30. Untuk hasil lengkapnya pada dataset lain, penulis lampirkan pada link di bawah ini:

https://drive.google.com/drive/folders/1rUEkDS9Xosb3QOLYeyh6XKnW5OLLP3jK?usp=drive_link

Dataset 1500 Acak 85:15

Variabel	β	Std Error	Uji Wald
(Konstanta)	2.9043		
R1006B	-1.0710	0.1986	-5.3927
R1005C	-1.5717	0.3859	-4.0728
R1003B	-1.2352	0.4192	-2.9465
R1205A2	-0.7706	0.4723	-1.6315
R809D	0.9450	0.3854	2.4519

Dataset 1500 Acak 80:20

Variabel	β	Std Error	Uji Wald
(Konstanta)	2.9018		
R1005C	-1.7027	0.4133	-4.1197
R1006B	-0.9967	0.1939	-5.1402
R1003B	-1.3168	0.3757	-3.5049
R1205A2	-0.7852	0.4559	-1.7223
R809D	0.9691	0.3944	2.4571

Dataset 1500 Acak 70:30

Variabel	β	Std Error	Uji Wald
(Konstanta)	2.9306		
R1005C	-1.7135	0.4082	-4.1976
R1006B	-1.0016	0.2080	-4.8153
R1205A1	-0.7005	0.4838	-1.4479
R1003B	-1.2211	0.3527	-3.4621
R809D	0.9182	0.3758	2.4433

Lampiran 13. Model Terbaik *Logistic Regression*

Berikut adalah sebagian hasil model terbaik *Logistic Regression* pada data 1500 acak dari dataset 90:10. 85:15. 80:20. dan 70:30. Untuk hasil lengkapnya pada dataset lain. penulis lampirkan pada link di bawah ini.

https://drive.google.com/drive/folders/1Xpg6FYraxzlowQvgTwy3kkohdn0UpvRy?usp=sharing

Dataset 1500 Acak 90:10

$$= \frac{e^{(2.9043-1.0710_{\mathrm{R1006B}}-1.5717_{\mathrm{R1005c}}-1.2352_{\mathrm{R1003B}}-0.7706_{\mathrm{R1205A2}}+0.9450_{\mathrm{R809D}})}}{1+e^{(2.9043-1.0710_{\mathrm{R1006B}}-1.5717_{\mathrm{R1005c}}-1.2352_{\mathrm{R1003B}}-0.7706_{\mathrm{R1205A2}}+0.9450_{\mathrm{R809D}})}}$$

$$g(x) = 2.9043-1.0710_{\mathrm{R1006B}} - 1.5717_{\mathrm{R1005c}} -1.2352_{\mathrm{R1003B}} -0.7706_{\mathrm{R1205A2}} + 0.9450_{\mathrm{R809D}}$$

Dataset 1500 Acak 80:20

$$= \frac{e^{(2.9018-1.7027_{\mathrm{R1005C}}-0.9967_{\mathrm{R1006B}}-1.3168_{\mathrm{R1003B}}-0.7852_{\mathrm{R1205A2}}+0.9691_{\mathrm{R809D}})}}{1+e^{(2.9018-1.7027_{\mathrm{R1005C}}-0.9967_{\mathrm{R1006B}}-1.3168_{\mathrm{R1003B}}-0.7852_{\mathrm{R1205A2}}+0.9691_{\mathrm{R809D}})}}$$

$$g(x) = 2.9018-1.7027_{\mathrm{R1005C}} - 0.9967_{\mathrm{R1006B}} - 1.3168_{\mathrm{R1003B}} -0.7852_{\mathrm{R1205A2}}+0.9691_{\mathrm{R809D}}$$

Dataset 1500 Acak 70:30

$$= \frac{e^{(2.9306-1.7135_{\mathrm{R1005C}}-1.0016_{\mathrm{R1006B}}-0.7005_{\mathrm{R1205A1}}-1.2211_{\mathrm{R1003B}}+0.9182_{\mathrm{R809D}})}}{1+e^{(2.9306-1.7135_{\mathrm{R1005C}}-1.0016_{\mathrm{R1006B}}-0.7005_{\mathrm{R1205A1}}-1.2211_{\mathrm{R1003B}}+0.9182_{\mathrm{R809D}})}}$$

$$g(x) = 2.9306-1.7135_{\mathrm{R1005C}} - 1.0016_{\mathrm{R1006B}} - 0.7005_{\mathrm{R1205A1}} -1.2211_{\mathrm{R1003B}}+0.9182_{\mathrm{R809D}}$$

Lampiran 14. *Odds Ratio*

Berikut adalah sebagian hasil *Odds Ratio* pada data 1500 acak dari dataset 90:10. 85:15. 80:20. dan 70:30. Untuk hasil lengkapnya pada dataset lain. penulis lampirkan pada link di bawah ini:
https://drive.google.com/drive/folders/1ngi6afHMTQ08K2tLZ1pIDybpWeewwRM6?usp=sharing

Dataset 1500 Acak 90:10

Variabel	**β**	***Odds Ratio***
(Konstanta)	2.987	
R1006B	-1.0710	0.3427
R1005C	-1.5717	0.2076
R1003B	-1.2352	0.29078
R1205A2	-0.7706	0.4627
R809D	0.9450	2.5728

Dataset 1500 Acak 80:20

Variabel	**β**	***Odds Ratio***
(Konstanta)	2.9018	
R1005C	-1.7027	0.1822
R1006B	-0.9967	0.3691
R1003B	-1.3168	0.2680
R1205A2	-0.7852	0.4560
R809D	0.9691	2.6357

Dataset 1500 Acak 70:30

Variabel	**β**	***Odds Ratio***
(Konstanta)	2.9306	
R1005C	-1.7135	0.1802
R1006B	-1.0016	0.3673
R1205A1	-0.7005	0.4963
R1003B	-1.2211	0.2949
R809D	0.9182	2.5049

Lampiran 15. Hasil Prediksi *Logistic Regression*

Berikut adalah sebagian data hasil prediksi *Logistic Regression* pada data 1500 acak dari dataset 90:10. 85:15. 80:20. dan 70:30. Untuk hasil lengkapnya pada dataset lain. penulis lampirkan pada link di bawah ini:

https://drive.google.com/drive/folders/1e6vKBCNVzwHOeOyPfiSo-UzgTz48_813?usp=sharing

Prediksi Dataset 1500 Acak 90:10

Nama Provinsi	Nama Desa	Aktual	Prediksi	Probabilitas Prediksi
Bengkulu	Srikuncoro	1	1	0.916903375
Di Yogyakarta	Demangan	0	1	0.941361414
Kalimantan Barat	Jongkong Kiri Tengah	1	1	0.919980797
Sumatera Selatan	Jejawi	1	1	0.914115115
Jawa Tengah	Kaligentong	1	1	0.952893491
Sulawesi Tenggara	Wapia Pia	1	1	0.871934
Jawa Tengah	Masaran	1	1	0.880833
Sumatera Barat	Sambungo	1	1	0.919981
Nusa Tenggara Timur	Gaura	1	0	0.498316
Kalimantan Selatan	Panaitan	0	1	0.916903375

Prediksi 1500 Dataset Acak 85:15

Nama Provinsi	Nama Desa	Aktual	Prediksi	Probabilitas Prediksi
Bengkulu	Srikuncoro	1	1	0.915897821
Di Yogyakarta	Demangan	0	1	0.940246848
Kalimantan Barat	Jongkong Kiri Tengah	1	1	0.918826031
Sumatera Selatan	Jejawi	1	1	0.910647641
Jawa Tengah	Kaligentong	1	1	0.951811668
Sulawesi Utara	Paslaten	1	1	0.947352896
Bali	Serongga	1	1	0.921661009
Sulawesi Utara	Motongkad Utara	1	1	0.713160554
Bali	Tojan	1	1	0.950008617
Papua	Yongsu Dosoyo	1	0	0.266959797

Prediksi 1500 Dataset Acak 80:20

Nama Provinsi	Nama Desa	Aktual	Prediksi	Probabilitas Prediksi
Bengkulu	Srikuncoro	1	1	0.913357434
Di Yogyakarta	Demangan	0	1	0.941085828
Kalimantan Barat	Jongkong Kiri Tengah	1	1	0.916634364
Sumatera Selatan	Jejawi	1	1	0.915760530

Nama Provinsi	Nama Desa	Aktual	Prediksi	Probabilitas Prediksi
Jawa Tengah	Kaligentong	1	1	0.952896799
Sulawesi Utara	Paslaten	1	1	0.947999927
Bali	Serongga	1	1	0.919798240
Sulawesi Utara	Motongkad Utara	1	1	0.710204456
Bali	Tojan	1	1	0.950969025
Papua	Yuneri	0	0	0.387935457

Prediksi 1500 Dataset Acak 70:30

Nama Provinsi	Nama Desa	Aktual	Prediksi	Probabilitas Prediksi
Bengkulu	Srikuncoro	1	1	0.915244470
Di Yogyakarta	Demangan	0	1	0.933314870
Kalimantan Barat	Jongkong Kiri Tengah	1	1	0.918290536
Sumatera Selatan	Jejawi	1	1	0.914543400
Jawa Tengah	Kaligentong	1	1	0.953930626
Sulawesi Utara	Paslaten	1	1	0.947768741
Bali	Serongga	1	1	0.921236549
Sulawesi Utara	Motongkad Utara	1	1	0.712270178
Bali	Tojan	1	1	0.952143954
Nusa Tenggara Timur	Gaura	1	0	0.486081276

Lampiran 16. Tabel Kesimpulan Penelitian

Tabel Perbandingan Kinerja Kedua Algoritma yang Unggul

Algoritma	Dataset	*Precision*	*Accuration*	*Recall*	*Specificity*	*F-Measure*
Random Forest	1500 Acak 85:15	92.37%	92.44%	1	10.52%	96.03%
Logistic Regression	1500 Acak 85:15	92.30%	91.56%	99.02%	10.52%	92.30%

Variabel yang Paling Mempengaruhi Prediksi pada Algoritma yang Paling Unggul (Random Forest)

Variabel Yang Mempengaruhi Prediksi
R809D (Jumlah kelompok tani)
R1005C (Sinyal telepon seluler handphone di sebagian besar wilayah desa kelurahan)
R1006B (Fasilitas internet di kantor kepala desa lurah)
R809B (Jumlah karang taruna)
R809F (Jumlah kelompok Masyarakat)

Lampiran 17. Tabel Z

z	0	1	2	3	4	5	6	7	8	9
-0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.5	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002
-3.6	0.0002	0.0002	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
-3.7	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
-3.8	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
-3.9	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Lampiran 18. *Source Code*

Random Forest

Berikut di bawah ini adalah potongan *Source Code* untuk algoritma Random Forest yang digunakan pada penelitian ini. Untuk *Source Code* lengkapnya, penulis lampirkan pada link: https://colab.research.google.com/drive/14mcv9VLRnxxEepCaOIVlw8rigvu56OE0?usp=sharing

```
# Inisialisasi model RandomForestClassifier dengan parameter
tertentu
rf = RandomForestClassifier(random_state=random_state)

# Inisialisasi GridSearchCV
grid_space = {
    'max_features': ['auto', 'sqrt', 'log2'],
    'n_estimators': [100, 200, 500],
    'max_depth': [4,5,6,7,8],
    'min_samples_leaf': [1,2,4],
    'criterion': ['entropy','gini'],
    'min_samples_split': [2,5,10]
}

grid = GridSearchCV(rf, param_grid=grid_space, cv=10,
scoring='accuracy')
model_grid = grid.fit(x_train,y_train)
# grid search results
print('Best grid search hyperparameters are:
'+str(model_grid.best_params_))
print('Best grid search score is: '+str(model_grid.best_score_))
# Inisialisasi model RandomForestClassifier dengan parameter
terbaik
rf_best = model_grid.best_estimator_
```

Logistic Regession

Berikut di bawah ini adalah potongan *Source Code* untuk algoritma Logistic Regression yang digunakan pada penelitian ini. Untuk *Source Code* lengkapnya, penulis lampirkan pada link: https://colab.research.google.com/drive/1XFzkqb6HU48d4zfaCIX5nYfe9Xmd8Fry?usp=sharing

```
# Inisialisasi model RandomForestClassifier dengan parameter
tertentu
logreg = LogisticRegression(random_state=random_state)
```

```
# Inisialisasi GridSearchCV
penalty = ['l1', 'l2']
C = np.logspace(-2,2,10)

#Menjadikan ke dalam bentuk dictionary
hyperparameters = dict(penalty=penalty, C=C)

#Memasukan ke Grid Search
#CV itu Cross Validation
#Menggunakan 10-Fold CV
clf = GridSearchCV(logreg, hyperparameters, cv=10)
#Fitting Model
best_model = clf.fit(x_train[selected_features],y_train)
#Nilai hyperparameters terbaik
print('Best Penalty:',
best_model.best_estimator_.get_params()['penalty'])
print('Best C:', best_model.best_estimator_.get_params()['C'])
```