BULLETIN OF INFORMATION TECHNOLOGY (BIT)
Vol 6, No 2, Juni 2025, Hal. 23 - 35
ISSN 2722-0524 (media online)
DOI 10.47065/bit.v5i2.1783
https://journal.fkpt.org/index.php/BIT

Penerapan Data Mining Untuk Klasifikasi Penduduk Miskin Di Kabupaten Labuhanbatu Menggunakan Random Forest Dan K-Nearest Neighbors

Andi Ernawati*, Khairul, Zulham Sitorus, Muhammad Iqbal, Darmeli Nasution

Magister Teknologi Informasi, Universitas Pembangunan Panca Budi Medan , Medan, Indonesia
Email: aernawati296@gmail.com, Khairul@dosen.pancabudi.ac.id, zulhamsitorus@dosen.pancabudi.ac.id, muhammadiqbal@dosenpancabudi.ac.id, darmelinasution@gmail.com
Email Penulis Korespondensi: aernawati296@gmail.com

Abstrak- Penelitian ini bertujuan untuk menerapkan dan membandingkan performa dua algoritma data mining, yaitu Random Forest (RF) dan K-Nearest Neighbors (KNN), dalam klasifikasi status kemiskinan penduduk di Kabupaten Labuhanbatu. Data yang digunakan mencakup informasi pekerjaan, pendapatan, tempat tinggal, dan pendidikan dari 21.137 individu. Setelah melalui tahapan praproses, pelatihan, optimasi, dan evaluasi model, dilakukan pengujian dengan lima metrik utama: akurasi, presisi, recall, F1-score, dan AUC. Hasil evaluasi menunjukkan bahwa algoritma Random Forest memberikan performa yang sedikit lebih baik dibandingkan KNN dengan nilai akurasi 0,6023, presisi 0,4827, recall 0,4177, F1-score 0,4479, dan AUC 0,5681. Sedangkan KNN memperoleh nilai akurasi 0,5990, presisi 0,4771, recall 0,4006, F1-score 0,4355, dan AUC 0,5622. Berdasarkan hasil ini, dapat disimpulkan bahwa Random Forest lebih efektif dalam tugas klasifikasi kemiskinan pada dataset ini, meskipun perbedaannya tidak signifikan.

Kata Kunci: Data Mining, Random Forest, K-Nearest Neighbors, Kemiskinan, Klasifikasi

Abstract-This study aims to apply and compare the performance of two data mining algorithms—Random Forest (RF) and K-Nearest Neighbors (KNN)—in classifying poverty status among residents of Labuhanbatu Regency. The dataset includes information on occupation, income, housing, and education from 21,137 individuals. After undergoing preprocessing, model training, hyperparameter optimization, and evaluation, both models were assessed using five key metrics: accuracy, precision, recall, F1-score, and AUC. The results show that Random Forest performed slightly better than KNN, achieving an accuracy of 0.6023, precision of 0.4827, recall of 0.4177, F1-score of 0.4479, and an AUC of 0.5681. In comparison, KNN obtained an accuracy of 0.5990, precision of 0.4771, recall of 0.4006, F1-score of 0.4355, and an AUC of 0.5622. Based on these findings, it can be concluded that Random Forest is more effective for poverty classification on this dataset, although the performance difference is relatively small.

Keywords: Data Mining, Random Forest, K-Nearest Neighbors, Poverty, Classification

1. PENDAHULUAN

Kemiskinan adalah salah satu masalah sosial yang paling mendesak di dunia, mencerminkan kondisi di mana individu atau kelompok tidak memiliki akses yang memadai terhadap sumber daya dasar untuk memenuhi kebutuhan hidup seperti makanan, tempat tinggal, pendidikan, dan kesehatan. Kemiskinan tidak hanya menjadi tantangan ekonomi, tetapi juga berdampak luas pada aspek sosial, budaya, dan politik. Kemiskinan dapat terjadi akibat berbagai faktor, termasuk ketidakmerataan distribusi sumber daya, rendahnya tingkat pendidikan, kurangnya kesempatan kerja, hingga kebijakan yang tidak mendukung pembangunan yang inklusif. Selain itu, faktor-faktor eksternal seperti bencana alam, konflik, dan perubahan iklim juga sering memperburuk tingkat kemiskinan di banyak wilayah.

Dampak kemiskinan meluas dari individu hingga komunitas. Pada tingkat individu, kemiskinan sering kali menghambat akses ke pendidikan yang berkualitas, kesehatan yang layak, dan peluang ekonomi. Pada tingkat komunitas, kemiskinan dapat memicu ketimpangan sosial, meningkatkan angka kriminalitas, dan menghambat pertumbuhan ekonomi secara keseluruhan. Oleh karena itu, pemahaman dan penanganan terhadap kemiskinan membutuhkan pendekatan yang holistik dan berkelanjutan, melibatkan kerja sama antara pemerintah, sektor swasta, dan masyarakat. Pendekatan ini mencakup kebijakan yang mendukung redistribusi kekayaan, peningkatan akses pendidikan dan kesehatan, penciptaan lapangan kerja, serta pemberdayaan masyarakat miskin agar mereka dapat meningkatkan taraf hidup mereka secara mandiri. Dengan usaha kolektif, kemiskinan dapat ditekan, memberikan harapan akan masa depan yang lebih adil dan sejahtera bagi semua.

Kabupaten Labuhanbatu salah satu kabupaten yang ada di Provinsi Sumatera Utara yang memiliki wilayah ± 2.561,38 km² dengan jumlah penduduk ±511.704 jiwa. Kabupaten ini memilih 9 Kecamatan dengan 23 Keluarahan dan 75 Desa. (Wikipedia, 2024). Angka kemiskinan Kabupaten Labuhanbatu mengalami penurunan sebesar 0,15 persen poin yaitu dari 7,99 persen pada Maret 2023 menjadi 7,84 persen pada Maret 2024. Angka kemiskinan ini setara dengan 42,45 ribu jiwa pada Maret 2024,atau berkurang sekitar 0,13 ribu jiwa. (BPS, 2024) Angka ini menunjukkan adanya tantangan besar dalam upaya pengentasan kemiskinan di daerah tersebut. Dalam konteks ini, penerapan teknik data mining, khususnya Random Forest (RF) dan K-Nearest Neighbors (KNN), menjadi sangat relevan untuk menganalisis faktor-faktor yang mempengaruhi kemiskinan dan memprediksi tren kedepannya.

Salah satu permasalahan utama yang dihadapi adalah kurangnya pemahaman tentang faktor-faktor yang menyebabkan kemiskinan di Labuhanbatu. Berbagai penelitian sebelumnya menunjukkan bahwa faktor ekonomi, pendidikan, dan tempat tinggal memiliki pengaruh signifikan terhadap tingkat kemiskinan. Namun, analisis yang mendalam dan sistematis mengenai hubungan antara variabel-variabel ini dan kemiskinan masih terbatas. Selain itu, adanya fluktuasi ekonomi

BULLETIN OF INFORMATION TECHNOLOGY (BIT)
Vol 6, No 2, Juni 2025, Hal. 23 - 35
ISSN 2722-0524 (media online)
DOI 10.47065/bit.v5i2.1783
https://journal.fkpt.org/index.php/BIT

yang dipengaruhi oleh faktor eksternal, seperti Bencana Alam, menambah kompleksitas dalam memahami dinamika kemiskinan di daerah ini.

Untuk mengatasi permasalahan tersebut, beberapa alternatif solusi dapat dipertimbangkan. Pertama, melakukan survei dan pengumpulan data yang lebih komprehensif mengenai kondisi sosial ekonomi masyarakat di Labuhanbatu. Kedua, menerapkan teknik analisis data yang lebih canggih, seperti data mining, untuk menggali informasi yang lebih mendalam dari data yang ada. Ketiga, kolaborasi antara pemerintah, akademisi, dan organisasi non-pemerintah dalam merumuskan kebijakan yang berbasis data dapat membantu dalam pengentasan kemiskinan secara lebih efektif.

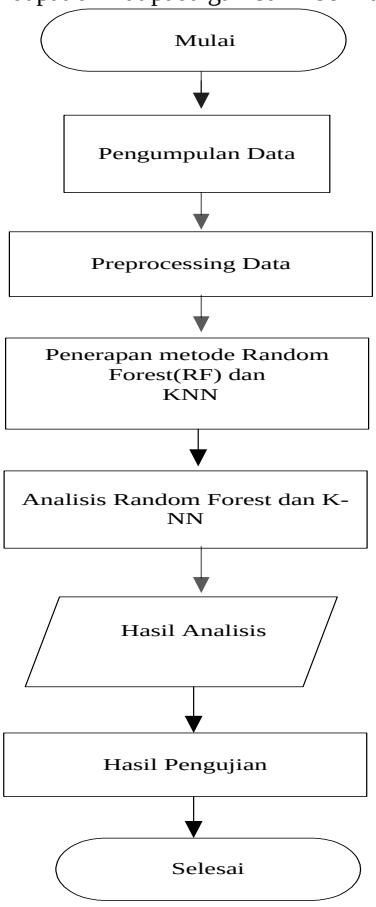
Random Forest (RF) pertama kali diperkenalkan oleh Leo Breiman pada tahun 2001, Random Forest (RF) adalah salah satu metode pembelajaran mesin berbasis ensemble yang dirancang untuk meningkatkan akurasi prediksi. Algoritma ini bekerja dengan membangun sekumpulan decision tree secara acak dan menggabungkan hasilnya untuk menghasilkan prediksi yang lebih andal. (Suci Amaliah et al., 2022) Random forest adalah algoritma yang menggabungkan prediksi dari banyak decision tree untuk menghasilkan hasil yang lebih akurat dan stabil. Algoritma ini bekerja dengan membuat banyak pohon keputusan (decision trees) pada subset acak dari data pelatihan, dan kemudian mengambil rata-rata atau suara mayoritas (tergantung pada tugasnya) untuk memberikan prediksi akhir.

K-Nearest Neighbors (KNN) adalah salah satu algoritma machine learning yang digunakan untuk melakukan klasifikasi dan regresi. Algoritma ini merupakan metode lazy learning, yang berarti tidak memiliki tahap pelatihan eksplisit, melainkan membuat prediksi berdasarkan data yang ada secara langsung saat ada permintaan klasifikasi atau prediksi baru. K-Nearest Neighbors (KNN) memiliki pengaruh yang signifikan terhadap metode nonparametrik dalam bentuk klasifikasi, namun tingkat kinerjanya secara umum bergantung pada titik ekuilibrium variabel yang berkorelasi dengan titik jauh. Jarak antara pembacaan dari batas yang ditentukan dari nilai standar deviasi (W et al., 2022)

Dalam penelitian ini, solusi yang dipilih adalah penerapan data mining, khususnya menggunakan Random Forest untuk mengklasifikasikan data yang mempengaruhi kemiskinan dan K-Nearest Neighbors (KNN). Metode ini dipilih sebagai perbandingan metode data mining. Dengan menggunakan teknik ini, diharapkan dapat diperoleh pemahaman yang lebih baik mengenai dinamika kemiskinan serta rekomendasi kebijakan yang lebih tepat sasaran.

2. METODOLOGI PENELITIAN

Adapun diagram penelitian yang diusulkan dapat dilihat pada gambar 1 berikut :



Gambar 1: Diagram Penelitian

Berdasarkan gambar diagram tersebut, dapat dijelaskan prosedur Penelitian Sebagai berikut:

1. Penelitian di mulai dengan merencanakan tujuan dan langkah langkah penelitian yang akan dilakukan.
2. Pengumpulan Data: Data yang berkaitan dikumpulkan yang bersumber dari instansi resmi dan kredibel. Mengambil

BULLETIN OF INFORMATION TECHNOLOGY (BIT)
Vol 6, No 2, Juni 2025, Hal. 23 - 35
ISSN 2722-0524 (media online)
DOI 10.47065/bit.v5i2.1783
https://journal.fkpt.org/index.php/BIT

data yang sudah tersedia dari sumber resmi yaitu Dinas Sosial Kabupaten Labuhanbatu untuk data resmi terkait data kemiskinan, yang meliputi, Pekerjaan, pendidikan, Pendapatan dan Tempat tinggal.

3. *Preprocessing Data*: Data yang sudah dikumpulkan diolah terlebih dahulu untuk memperbaiki kualitasnya. Langkah-langkah ini mencakup:
   a. Pembersihan data (menghapus data yang tidak relevan atau kesalahan).
   b. Transformasi data ke format yang tepat untuk analisis.
   c. Normalisasi atau standarisasi data jika diperlukan.
4. Penerapan *Metode Random Forest*(RF) dan *K-Nearest Neighbors* (KNN): Setelah pengolahan data selesai, metode *Random Forest(*RF) dan *K-Nearest Neighbors* (KNN) digunakan untuk menciptakan model prediktif atau klasifikasi.
5. Analisis *Random Forest* (RF) dan *K-Nearest Neighbors* (KNN): Output dari penerapan kedua metode tersebut dianalisis untuk menilai performa model. Analisis ini mungkin melibatkan pengukuran akurasi, presisi, recall, dan metrik evaluasi yang lain.
6. Hasil Analisis: Hasil dari analisis dibandingkan untuk menentukan metode mana yang memberikan performa terbaik dalam konteks data dan tujuan riset.
7. Hasil Pengujian: Model yang telah dianalisis selanjutnya diuji menggunakan data uji (*testing data*) untuk menilai kemampuan generalisasi model pada data yang baru.
8. Selesai: Penelitian diakhiri setelah semua langkah selesai, dan hasilnya didokumentasikan dalam bentuk laporan atau publikasi.

2.1 Random Forest (RF)

Random Forest merupakan salah satu algoritma pembelajaran mesin berbasis ensemble yang digunakan untuk tugas klasifikasi dan regresi. Algoritma ini diperkenalkan oleh Leo Breiman pada tahun 2001 sebagai pengembangan dari metode pohon keputusan (*decision tree*) dengan tujuan meningkatkan akurasi prediksi dan mengurangi risiko overfitting (Sis, 2024)

Metode *Random Forest* bekerja dengan membangun sejumlah besar pohon keputusan secara acak, di mana setiap pohon dilatih menggunakan subset acak dari data pelatihan. Prediksi akhir dihasilkan melalui agregasi mayoritas *(voting*) untuk klasifikasi atau perataan *(averaging*) untuk regresi. Dengan pendekatan ini, Random Forest mampu menangani data yang kompleks serta memberikan hasil yang lebih stabil dibandingkan dengan model pohon keputusan tunggal.

Keunggulan utama dari *Random Forest* terletak pada kemampuannya dalam menangani data berdimensi tinggi serta ketahanannya terhadap outlier dan data yang hilang. Selain itu, algoritma ini juga menyediakan estimasi pentingnya fitur *(feature importance*), yang berguna dalam proses seleksi variabel untuk meningkatkan efisiensi model. Berkat karakteristik tersebut, Random Forest telah banyak digunakan dalam berbagai bidang, termasuk analisis keuangan, diagnosis medis, deteksi intrusi jaringan, dan penelitian genomik.

Algoritma ini membangun sekumpulan pohon keputusan independen, yang masing-masing dilatih menggunakan subset data pelatihan yang dipilih secara acak dengan pengembalian (*bootstrap sampling*), serta subset fitur yang juga dipilih secara acak pada setiap percabangan. Melalui agregasi prediksi individu dari setiap pohon menggunakan metode suara mayoritas (*majority voting*) dalam klasifikasi atau rata-rata dalam regresi, *Random Forest* mengurangi varians model tanpa meningkatkan bias secara signifikan. Secara matematis, prediksi akhir untuk tugas klasifikasi dapat dirumuskan sebagai:

Dapat dirumuskan sebagai :

$\hat{y}=mode\ (\{\hat{y}_t: t = 1,2, \ldots\ldots, T\})$

Sedangkan untuk tugas regresi:

$\hat{y}=\frac{1}{T}\sum_{t=1}^{T}\hat{y}_t$

di mana $\hat{y}_t$ adalah prediksi dari pohon ke-*t* dan *T* adalah jumlah total pohon. Random Forest juga menggunakan metrik seperti Gini Impurity untuk klasifikasi:

$Gini = 1 - \sum_{K=1}^{K} P_k^2$

dan pengurangan variansi untuk regresi:

$$Var = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2$$

Kemampuan algoritma ini untuk menangani data besar, fitur yang multivariat, dan interaksi non-linear membuatnya menjadi alat yang sangat relevan dalam berbagai aplikasi ilmiah dan industri.

2.2 *K-Nearest Neighbors* (KNN)

K-Nearest Neighbors (KNN) adalah algoritma pembelajaran mesin yang digunakan untuk klasifikasi dan regresi. Algoritma ini bekerja dengan mencari sejumlah k tetangga terdekat dari data yang ingin diprediksi, berdasarkan metrik jarak tertentu seperti jarak Euclidean. Prediksi dilakukan dengan menganalisis label atau nilai dari tetangga-tetangga terdekat tersebut. *K-Nearest Neighbors* (KNN) dikenal karena kesederhanaannya dan efektivitasnya dalam berbagai aplikasi, seperti pengenalan pola dan pengambilan keputusan.(Sakti & Daulay, 2024)

BULLETIN OF INFORMATION TECHNOLOGY (BIT)
Vol 6, No 2, Juni 2025, Hal. 23 - 35
ISSN 2722-0524 (media online)
DOI 10.47065/bit.v5i2.1783
https://journal.fkpt.org/index.php/BIT

K-Nearest Neighbors (K-NN) merupakan salah satu algoritma pembelajaran mesin yang bersifat sederhana namun efektif dalam tugas klasifikasi dan regresi. Prinsip dasar dari algoritma ini adalah asumsi bahwa objek-objek yang berada dalam jarak yang dekat satu sama lain cenderung memiliki karakteristik yang serupa. Dengan kata lain, jika karakteristik suatu objek telah diketahui, maka status objek lain dapat diprediksi berdasarkan kemiripan dengan tetangga terdekatnya.
Sebagai pengembangan dari teknik *K-Nearest Neighbors*, algoritma K-NN mengklasifikasikan data baru berdasarkan suara mayoritas dari K tetangga terdekat. Nilai K merupakan bilangan bulat positif yang menentukan jumlah tetangga yang akan dipertimbangkan dalam proses klasifikasi. Umumnya, nilai K dipilih dalam jumlah ganjil untuk menghindari kemungkinan hasil seri dalam proses pemungutan suara (voting) antar tetangga.
Sebagai metode non-parametrik, K-NN tidak bergantung pada asumsi distribusi data tertentu. Hal ini menjadikannya fleksibel dalam menangani berbagai jenis dataset. Dalam tugas klasifikasi, K-NN menentukan kategori suatu objek berdasarkan mayoritas kelas dari K tetangga terdekat. Sementara itu, dalam regresi, nilai yang diprediksi dihitung sebagai rata-rata atau median dari nilai-nilai tetangga terdekat.
Dalam berbagai bidang, K-NN telah banyak diterapkan untuk berbagai kebutuhan analisis data. Salah satu penerapannya dalam studi sosial ekonomi adalah dalam analisis kemiskinan, di mana algoritma ini dapat digunakan untuk mengelompokkan individu atau rumah tangga berdasarkan karakteristik ekonomi, pendidikan, dan sosial. Dengan membandingkan atribut individu dengan data yang telah diklasifikasikan sebelumnya, algoritma K-NN dapat membantu dalam menentukan status kemiskinan suatu kelompok masyarakat.
Keunggulan utama dari K-NN terletak pada kemudahannya dalam implementasi serta kemampuannya dalam menangani data yang tidak memiliki pola distribusi yang jelas. Namun, kelemahan utama algoritma ini adalah sensitivitasnya terhadap jumlah K yang dipilih serta kompleksitas komputasi yang meningkat seiring bertambahnya jumlah data. Oleh karena itu, pemilihan nilai K yang optimal dan penggunaan teknik optimasi, seperti feature scaling dan dimensionality reduction, menjadi faktor penting dalam meningkatkan performa algoritma K-NN.$d_{(a,b)} = \sqrt{\sum_{g=1}^{p}(x_{ag} - x_{bg})^{\ 2}}$

Keterangan:
$d_{(a,b)}$ = jarak antara objek a dengan b
$x_{(a,g)}$ = nilai objek $training\ a$ pada variabel ke-g
$x_{(a,g)}$ =nilai objek data $testing\ b$ pada variabel ke-g
P = banyaknya variabel bebas
Algoritma pengerjaan metode K-Nearest Neighbor adalah sebagai berikut:
1. Tetapkan parameter K (jumlah tetangga terdekat).
2. Hitung jarak data baru/data testing menggunakan jarak Euclidean dengan semua data yang ada pada data training (persamaan 2.1)
3. Tentukan tetangga mana yang paling dekat berdasarkan jarak ke-K yang paling minimal.
4. Menguraikan kategori dari tetangga terdekat
5. Menggunakan kategori yang paling dapat diandalkan dari tetangga terdekat sebagai prediksi data baru.

Hasil penelitian menunjukkan bahwa kedua metode, Naive Bayes dan KNN, memiliki kelebihan dan kekurangan masing-masing. Naive Bayes lebih efisien dalam hal waktu komputasi, sementara KNN memberikan akurasi yang lebih tinggi. Pemilihan metode yang tepat harus mempertimbangkan ukuran dataset, kompleksitas data, dan tujuan analisis.
Integrasi data dari berbagai sumber juga menjadi faktor penting dalam analisis kemiskinan. Data yang lebih lengkap dan beragam dapat meningkatkan akurasi model prediksi. Selain itu, kolaborasi antara berbagai disiplin ilmu dapat memperkaya analisis dan memberikan wawasan yang lebih mendalam tentang faktor-faktor yang mempengaruhi kemiskinan.

3. HASIL DAN PEMBAHASAN

3.1 Pengujian Random Forest (RF)

Random Forest (RF) merupakan algoritma pembelajaran ensemble berbasis pohon keputusan (decision tree) yang bekerja dengan cara membentuk sejumlah pohon keputusan pada saat pelatihan dan menghasilkan prediksi berdasarkan hasil voting (untuk klasifikasi) atau rata-rata (untuk regresi) dari seluruh pohon. RF dikenal tangguh terhadap overfitting karena menggabungkan hasil dari banyak pohon keputusan yang dibentuk secara acak melalui teknik bootstrap sampling dan pemilihan fitur acak pada setiap pemisahan simpul (node)
Dalalam pengujian random Forest yang kita lakukana adalah menentukan

1. Bootstrap Sampling
   Yaitu mengambil sampel acak dari data yang akan di uji dalam hal ini kita menggukan sampel data pada tabel 4.8. dalam hal ini kami menguji dengan 10 data dikarenakan jika terlalu banyak data dengan pengujian manual terlalu sulit.

BULLETIN OF INFORMATION TECHNOLOGY (BIT)
Vol 6, No 2, Juni 2025, Hal. 23 - 35
ISSN 2722-0524 (media online)
DOI 10.47065/bit.v5i2.1783
https://journal.fkpt.org/index.php/BIT

Tabel 1. Bootstrap Sampling

No	Nama	Jenis_Kelamin	Sts_Kawin	Pekerjaan	Penghasilan	Tempat Tinggal	Pendidikan	Hasil
1	Solehuddin	0	0	14	0	0	2	Miskin
2	Samsuddin	0	0	27	3,2	1	2	Tidak Miskin
3	Yusnani Batubara	1	1	10	0	1	0	Miskin
4	Sugiono	0	1	12	3,5	1	0	Miskin
5	Risma	1	1	10	0	0	1	Tidak Miskin
6	Syahran Hrp	0	1	1	2	0	0	Tidak Miskin
7	Parlindungan Sipahutar	0	1	21	4	1	1	Tidak Miskin
8	Azhar Hsb	0	1	1	1,5	0	0	Miskin
9	Nasruddin Hasibuan	0	1	21	4	1	0	Miskin
…	…	…	…	…	…	…	…	…
24000	Limah	1	0	10	0	1	2	Miskin

Data diatas

Miskin = 6 Orang

Tidak Miskin = 4 Orang

2. Gini Awal (Sebelum Pemisahan)

Pada tahap awal pembangunan model Random Forest, langkah krusial adalah mengidentifikasi titik pemisahan (split) terbaik pada setiap node dalam setiap pohon keputusan. Proses ini diawali dengan penghitungan Gini Impurity pada node awal (node root) dari setiap pohon. Gini Impurity mengukur tingkat 'ketidakmurnian' atau heterogenitas kelas dalam suatu subset data. Perhitungan Gini dilakukan menggunakan rumus :

$$Gini = 1 - \left(\frac{6}{10}\right)^2 - \left(\frac{4}{10}\right)^2 = 1 - 0.36 - 0.04 = 1 - 0.52 = 0.48$$

Nilai Gini Impurity sebesar **0.48** ini menunjukkan bahwa node root memiliki tingkat campuran kelas yang cukup tinggi, mengindikasikan perlunya pemisahan lebih lanjut untuk mencapai node yang lebih murni. Selanjutnya, untuk mengidentifikasi fitur dan nilai ambang terbaik untuk pemisahan pertama, algoritma Random Forest menerapkan prinsip keacakan yang unik. Dari seluruh fitur prediktor yang tersedia (yaitu, Pekerjaan, Penghasilan, Tempat Tinggal, dan Pendidikan), secara acak dipilih sebagian kecil fitur untuk dipertimbangkan pada langkah pemisahan ini.

3. Split Terbaik – Uji Semua Fitur.
Setelah menghitung Gini Impurity awal sebesar 0.48, proses selanjutnya dalam membangun model Random Forest adalah mengidentifikasi atribut terbaik yang dapat digunakan untuk melakukan pemisahan (split) pada node root. Langkah ini dilakukan dengan menghitung Gini Impurity Split dan Gini Gain untuk setiap atribut prediktor yang tersedia Fitur Tempat Tinggal, Jenis Kelamin, Status Kawin, Pekerjaan, Pendidikan danPenghasilan.
   a. Fitur Tempat Tinggal
      - Tempat Tinggal (sewa) : 0= 4 data → Miskin = 2, Tidak Miskin = 2
        Gini(Cabang Tempat Tinggal)= $1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2$
        Gini(Cabang Tempat Tinggal)=$1 - (0.5)^2$-$(0.5)^2$
        Gini(Cabang Tempat Tinggal)=1-0.25−0.25
        Gini(Cabang Tempat Tinggal)=**0.5**
      - Tempat Tinggal (Milik Sendiri) = 1: 6 data → Miskin = 4,Tidak Miskin =2
        Gini(Cabang Tempat Tinggal) =$1 - \left(\frac{4}{6}\right)^2 - \left(\frac{2}{6}\right)^2$
        Gini(Cabang Tempat Tinggal)=1− 0.6667− 0.3333
        Gini(Cabang Tempat Tinggal)=**0.4445**
        $GiniSplit = \frac{4}{10}$ x 0,5 + $\frac{6}{10}$x 0.4445= 0.4667 => $Gain = 0.48 - 0.4667 =$ **0.0133**
   b. Fitur Jenis Kelamin
      - Jenis Kelamin (Lk) : 0 = 7 data → Miskin = 4, Tidak Miskin = 3
        Gini(Cabang Jenis Kelamin)= $1 - \left(\frac{4}{7}\right)^2 - \left(\frac{3}{7}\right)^2$
        Gini(Cabang Jenis Kelamin)=1−(0.3265)−(0.1837)
        Gini(Cabang Jenis Kelamin)=**0.4898**

BULLETIN OF INFORMATION TECHNOLOGY (BIT)
Vol 6, No 2, Juni 2025, Hal. 23 - 35
ISSN 2722-0524 (media online)
DOI 10.47065/bit.v5i2.1783
https://journal.fkpt.org/index.php/BIT

- Jenis Kelamin (Pr)=1 : 3 data → Miskin = 2, Tidak Miskin = 1

  Gini(Cabang Jenis Kelamin)= $1-\left(\frac{2}{3}\right)^2-\left(\frac{1}{3}\right)^2$

  Gini(Cabang Jenis Kelamin)=1− 0.4444- 0.1111

  Gini(Cabang Jenis Kelamin)=**0.4445**

$GiniSplit = \frac{7}{10}$ x 0.4898 + $\frac{3}{10}$x 0.4445 = 0.4762=> $Gain = 0.48 - 0.4762 =$ **0.0038**

c. Fitur Status Kawin

- Jenis Status Kawin (Blm) : 0 = 3 data → Miskin = 2, Tidak Miskin = 1

  Gini(Cabang Status Kawin)= $1-\left(\frac{2}{3}\right)^2-\left(\frac{1}{3}\right)^2$

  Gini(Cabang Status Kawin)= 1− 0.4444- 0.1111

  Gini(Cabang Status Kawin)= **0.4445**
- Jenis Status Kawin (Sdh)= 1:7 data → Miskin =4 , Tidak Miskin = 3

  Gini(Cabang Status Kawin)= $1-\left(\frac{2}{6}\right)^2-\left(\frac{0}{6}\right)^2$

  Gini(Cabang Status Kawin)= 1−(0.3265)−(0.1837)

  Gini(Cabang Status Kawin)=**0.4898**

$GiniSplit = \frac{3}{10}$ x 0.4445 + $\frac{7}{10}$x 0.4898 = 0.4762=> $Gain = 0.48 - 0.4762 =$ **0.0038**

d. Fitur Pekerjaan

- Jenis Pekerjaan (bhl) :1= 2 data → Miskin =1 ,Tidak Miskin=1

  Gini(Cabang Pekerjaan)= $1-\left(\frac{1}{2}\right)^2-\left(\frac{1}{2}\right)^2$

  Gini(Cabang Pekerjaan)=1−0.25-0.25

  Gini(Cabang Pekerjaan)=**0.5**
- Jenis Pekerjan (irt)= 10:3 data → Miskin=2,Tidak Miskin=1 Gini(Cabang Pekerjaan)= $1-\left(\frac{2}{3}\right)^2-\left(\frac{1}{3}\right)^2$

  Gini(Cabang Pekerjaan)= 1− 0.4444- 0.1111

  Gini(Cabang Pekerjaan)= **0.4445**
- Jenis Pekerjan (pedagang)= 12:1 data → Miskin = 1, Tidak Miskin= 0

  Gini(Cabang Pekerjaan)= $1-\left(\frac{1}{1}\right)^2-\left(\frac{0}{1}\right)^2$

  Gini(Cabang Pekerjaan)=1−1−0

  Gini(Cabang Pekerjaan)=**0**
- Jenis Pekerjan (Pelajar)= 14:1 data → Miskin = 1, Tidak Miskin= 0

  Gini(Cabang Pekerjaan)=$1-\left(\frac{1}{1}\right)^2-\left(\frac{0}{1}\right)^2$

  Gini(Cabang Pekerjaan)= 1−1-0

  Gini(Cabang Pekerjaan)=**0**
- Jenis Pekerjan Petani)= 21:2 data → Miskin =1, Tidak Miskin= 1

  Gini(Cabang Pekerjaan)= $1-\left(\frac{1}{2}\right)^2-\left(\frac{1}{2}\right)^2$

  Gini(Cabang Pekerjaan)=1−0.25-0.25

  Gini(Cabang Pekerjaan)=**0.5**
- Jenis Pekerjan (wiraswasta)= 27:1 data → Miskin = 1, Tidak Miskin=0

  Gini(Cabang Pekerjaan)= $1-\left(\frac{1}{1}\right)^2-\left(\frac{0}{1}\right)^2$

  Gini(Cabang Pekerjaan)= 1−1-0

  Gini(Cabang Pekerjaan)= **0**

  $GiniSplit = \frac{2}{10}$ x 0.5 + $\frac{3}{10}$x 0.4445 x $\frac{1}{10}$ x 0 + $\frac{1}{10}$x 0 $\frac{2}{10}$ x 0.5 + $\frac{1}{10}$x 0 = 0.3334=> $Gain = 0.48 - 0.3334 =$ **0.1466**

e. Fitur Pendidikan

- Jenis Pendidikan (SD) : 0 = 5 data → Miskin = 4, Tidak Miskin = 1

  Gini(Cabang Pendidikan)= $1-\left(\frac{4}{5}\right)^2-\left(\frac{1}{5}\right)^2$

  Gini(Cabang Pendidikan)=1−0.64-0.04

  Gini(Cabang Pendidikan)=**0.32**
- Jenis Pendidikan (SMP) : 1 = 2 data → Miskin = 0, Tidak Miskin = 2

  Gini(Cabang Pendidikan)= $1-\left(\frac{0}{2}\right)^2-\left(\frac{2}{2}\right)^2$

  Gini(Cabang Pendidikan)=1−0−1

  Gini(Cabang Pendidikan)=**0**
- Jenis Pendidikan (SMA) : 2 = 3 data → Miskin = 2, Tidak Miskin = 1

BULLETIN OF INFORMATION TECHNOLOGY (BIT)
Vol 6, No 2, Juni 2025, Hal. 23 - 35
ISSN 2722-0524 (media online)
DOI 10.47065/bit.v5i2.1783
https://journal.fkpt.org/index.php/BIT

Gini(Cabang Pendidikan)= $1-\left(\frac{2}{3}\right)^2-\left(\frac{1}{3}\right)^2$
Gini(Cabang Pendidikan)=1−0.4444-0.1111
Gini(Cabang Pendidikan)=1-0.5555=**0.4445**
$GiniSplit = \frac{5}{10}$ x 0.32 + $\frac{2}{10}$x 0 x $\frac{3}{10}$ x 0.4445= 0.2934=> $Gain = 0.48 - 0.2934 =$ **0.1866**

f. Fitur Penghasilan

- Jenis Penghasilan 0 = 4 data → Miskin = 3, Tidak Miskin = 1
  Gini(Cabang Penghasilan)= $1-\left(\frac{3}{4}\right)^2-\left(\frac{1}{4}\right)^2$
  Gini(Cabang Penghasilan)=1−0.5625-0.0625
  Gini(Cabang Penghasilan)=**0.375**
- Jenis Penghasilan 1.5 = 1 data → Miskin = 1, Tidak Miskin = 0
  Gini(Cabang Penghasilan)= $1-\left(\frac{1}{1}\right)^2-\left(\frac{0}{1}\right)^2$
  Gini(Cabang Penghasilan)=1−1-0
  Gini(Cabang Penghasilan)=**0**
- Jenis Penghasilan 2 = 1 data → Miskin = 0, Tidak Miskin = 1
  Gini(Cabang Penghasilan)= $1-\left(\frac{0}{1}\right)^2-\left(\frac{1}{1}\right)^2$
  Gini(Cabang Penghasilan)=1−0-1
  Gini(Cabang Penghasilan)=**0**
- Jenis Penghasilan 3.2 = 1 data → Miskin = 0, Tidak Miskin = 1
  Gini(Cabang Penghasilan)= $1-\left(\frac{0}{1}\right)^2-\left(\frac{1}{1}\right)^2$
  Gini(Cabang Penghasilan)=1−0-1
  Gini(Cabang Penghasilan)=**0**
- Jenis Penghasilan 3.5 = 1 data → Miskin = 0, Tidak Miskin = 1
  Gini(Cabang Penghasilan)= $1-\left(\frac{0}{1}\right)^2-\left(\frac{1}{1}\right)^2$
  Gini(Cabang Penghasilan)=1−0-1
  Gini(Cabang Penghasilan)=**0**
- Jenis Penghasilan 4 = 2 data → Miskin = 1, Tidak Miskin = 1
  Gini(Cabang Penghasilan)= $1-\left(\frac{1}{2}\right)^2-\left(\frac{1}{2}\right)^2$
  Gini(Cabang Penghasilan)=1−0.25-0.25
  Gini(Cabang Penghasilan)=**0.5**
  $GiniSplit = \frac{4}{10}$ x 0.375 + $\frac{1}{10}$x 0 + $\frac{1}{10}$ x 0 + $\frac{1}{10}$x 0 +$\frac{1}{10}$x 0+ $\frac{1}{10}$ x 0.5= 0.20=> $Gain = 0.48 - 0.20 =$ **0.28**

Hasil pengujian terhadap seluruh fitur prediktor diatas menunjukkan bahwa fitur "Penghasilan" memberikan informasi pemisahan terbaik, ditandai dengan nilai Gini Gain tertinggi yaitu sebesar **0,28.** Artinya, penghasilan merupakan variabel paling signifikan dalam membedakan status kemiskinan dalam dataset ini.

4. Pemilihan Fitur Acak

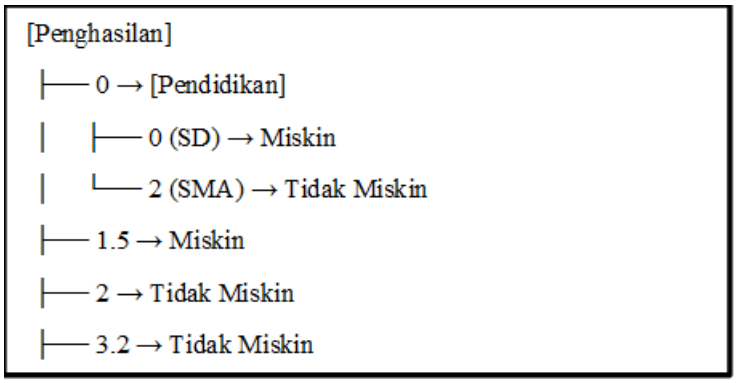
a. Pohon 1: Fitur Acak → Penghasilan, Pendidikan, Jenis_Kelamin
Root Node: Penghasilan (Gini Gain = 0.28 — tertinggi)
Penghasilan = 0 (4 data): Gini ≠ 0

- Split terbaik selanjutnya → Pendidikan (Gain = 0.125 > Jenis Kelamin)
  Pendidikan = 0 (SD): 3 data → 3 Miskin → Miskin (pure)
  Pendidikan = 2 (SMA): 1 data → 1 Tidak Miskin → Tidak Miskin (pure)

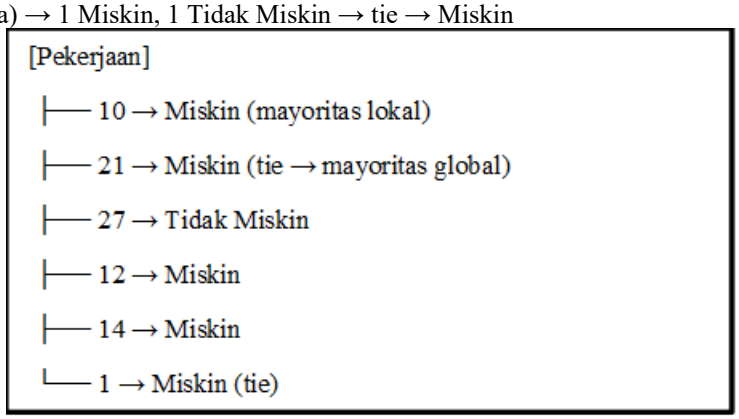
Penghasilan = 1.5, 2, 3.2, 3.5: semua pure → tidak perlu split
Penghasilan = 4 (2 data): 1 Miskin, 1 Tidak Miskin

- Split → Pendidikan atau Jenis_Kelamin
  Karena datanya cuma 2, dan tidak bisa memisah lebih lanjut, → tie, voting ke mayoritas global → Miskin

BULLETIN OF INFORMATION TECHNOLOGY (BIT)
Vol 6, No 2, Juni 2025, Hal. 23 - 35
ISSN 2722-0524 (media online)
DOI 10.47065/bit.v5i2.1783
https://journal.fkpt.org/index.php/BIT



b. Pohon 2: Fitur Acak → Tempat Tinggal, Pekerjaan, Status Kawin
Root Node: Pekerjaan (Gain = 0.1466 — tertinggi di subset)
Pekerjaan = 10 (3 data) → 2 Miskin, 1 Tidak Miskin
Split terbaik: Tempat Tinggal atau Status Kawin
Misalnya split Tempat Tinggal → 1 data = Tidak Miskin, 2 data = Miskin → Bisa klasifikasi berdasarkan mayoritas: Miskin
Pekerjaan = 21 (2 data) → 1 Miskin, 1 Tidak Miskin → tie → Miskin (mayoritas global)
Pekerjaan = 27, 12, 14 → pure → langsung klasifikasi
Pekerjaan = 1 (2 data) → 1 Miskin, 1 Tidak Miskin → tie → Miskin



c. Pohon 3: Fitur Acak → Pekerjaan, Penghasilan
Root Node: Penghasilan (Gain = 0.28 — tertinggi)
Seperti Tree 1:
- Penghasilan = 0 → split oleh Pekerjaan
  Pekerjaan = 10 → Miskin
  Pekerjaan = 1 → tie → Miskin
- Penghasilan = 1.5 → Miskin
- Penghasilan = 2 → Tidak Miskin
- Penghasilan = 3.2 → Tidak Miskin
- Penghasilan = 3.5 → Tidak Miskin
- Penghasilan = 4 → split Pekerjaan
- Pekerjaan = 21 → tie → Miskin

BULLETIN OF INFORMATION TECHNOLOGY (BIT)
Vol 6, No 2, Juni 2025, Hal. 23 - 35
ISSN 2722-0524 (media online)
DOI 10.47065/bit.v5i2.1783
https://journal.fkpt.org/index.php/BIT

5. Prediksi dengan semua pohon
   Pohon 1 (Penghasilan, Pendidikan, Jenis Kelamin):
   Penghasilan = 0 → split ke Pendidikan
   Pendidikan = 0 → klasifikasi **Miskin**
   Pohon 2 (Tempat Tinggal, Pekerjaan, Status Kawin):
   Root: Pekerjaan = 10
   Pekerjaan = 10 → split ke Tempat Tinggal
   Tempat Tinggal = 1 → mayoritas miskin → klasifikasi **Miskin**
   Pohon 3 (Penghasilan, Pekerjaan):
   Penghasilan = 0 → split ke Pekerjaan
   Pekerjaan = 10 → klasifikasi **Miskin**

6. Voting Mayoritas
   Voting mayoritas adalah proses di mana kelas hasil prediksi yang paling sering muncul di antara seluruh pohon dalam hutan Random Forest dipilih sebagai hasil akhir klasifikasi. Karena sampel data diatas semua pohon memprediksi "Miskin", maka hasil akhir berdasarkan voting mayoritas juga adalah Miskin.

3.2 Pengujian K-Nearest Neighbor (KNN)

Didalam pengujian K-Nearest Neighbor (KNN) dari tabel 4.3 kita menguji Muhammad Ikhsan apakah statusnya miskin atau tidak miskin, yaitu dengan cara menghitung jarak Euclidien antara data uji dengan tiap baris data latih yaitu baris ke 1 sampai dengan baris ke 300 dan mencari kedekatannya dengan K3 serta Voting berdasarkan kedekatan hasil ujinya. Dengan perhitungan sebagai berikut:

$$\text{Jarak} = \sqrt{\sum_{i-1}^{n} (x_i - y_i)^2}$$

Langkah Hitung Jarak (Euclidean Distance) sebagai berikut Berdasarkan data Fitur

Tabel 2 Hasil Normalisasi

No	Nama	Fitur
1	Solehuddin	[0, 0, 14, 0, 0, 2]
2	Samsuddin	[0, 0, 27, 3.2, 1, 2]
3	Yusnani Batubara	[1, 1, 10, 0, 1, 0]
4	Sugiono	[0, 1, 12, 3.5, 1, 4]
5	Risma	[1, 1, 10, 0, 0, 1]
6	Syahran Hrp	[0, 1, 1, 2, 0, 0]
7	Parlindungan Sipahutar	[0, 1, 21, 4, 1,1]
8	Azhar Hsb	[0, 1, 1, 1.5, 0, 0]
9	Nasruddin Hasibuan	[0, 1, 21, 4, 1, 0]
10	Limah	[1, 0, 10, 0, 1, 2]
11	Ahmad Efendi	[0, 1, 3, 3, 1, 2]
12	Susanti	[1, 1, 10, 0, 0, 4]
13	M. Najri Hasibuan	[0, 1,27, 3.2, 1, 1]
14	Supartik	[1, 1, 10, 0, 0, 2]
15	Sarijan	[0, 1, 21, 4, 1, 2]
16	Sukarni	[1, 1, 10, 0, 0, 2]
17	Kamiso	[0, 1, 21, 4, 1, 2]
18	Waijo	[0, 1, 21, 4, 1, 2]
19	Surep	[0, 1, 14, 0, 1, 2]
20	Purnama Raharja	[0, 1, 21, 4, 1, 2]
21	Isma Baswara	[0, 0, 14, 0, 1, 2]
22	Selasi Prima	[1, 1, 10, 0, 0, 0]
23	Hendi Penerangan	[0, 1, 21, 4, 1, 2]
24	Suroyo	[0, 1, 21, 4, 1, 2]
25	Miswadi	[0, 1, 21, 4, 1, 2]
26	Arya Pramadi	[0, 0, 14, 0, 1, 2]
27	Widodo	[0, 0, 0, 0, 0,2]
28	Fitriya	[1, 0, 0, 0, 0, 2]
29	Suwanto	[0, 1, 21, 4, 1, 2]
30	Muaiman	[0, 0, 14, 0, 1, 2]
	…	[…….]
24000	Tuginem	[1,1,10,0,1,0]

BULLETIN OF INFORMATION TECHNOLOGY (BIT)
Vol 6, No 2, Juni 2025, Hal. 23 - 35
ISSN 2722-0524 (media online)
DOI 10.47065/bit.v5i2.1783
https://journal.fkpt.org/index.php/BIT

$$\text{Jarak} = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2} + \cdots \ldots . . + (x_n - y_n)^2$$

0 3 27 3,2 0 5 ?

Baris 1 – Solehuddin (Miskin)

Fitur = [0, 0, 14, 0, 0,2]

$$=\sqrt{(0-0)^2 + (0-3)^2 + (14-27)^2 + (0-3{,}2)^2 + (0-0)^2} + (2-5)^2$$

$$=\sqrt{0 + 9 + 169 + 10{,}24 + 0 + 9}$$

$$=\sqrt{197{,}24}$$

=14,04

Baris 2 – Samsuddin (Tidak Miskin)

Fitur = $[0, 0, 27, 3.2, 1, 2]$

$$=\sqrt{(0-0)^2 + (0-3)^2 + (27-27)^2 + (3{,}2-3{,}2)^2 + (1-0)^2} + (2-5)^2$$

$$=\sqrt{0 + 9 + 0 + 0 + 1 + 9}$$

$$=\sqrt{19}$$

=4,36

Baris 3 – Yusnani Batubara (Miskin)

Fitur = $[1, 1, 10, 0, 1, 0]$

$$=\sqrt{(1-0)^2 + (1-3)^2 + (10-27)^2 + (0-3{,}2)^2 + (1-0)^2} + (0-5)^2$$

$$=\sqrt{1 + 4 + 289 + 10{,}24 + 1 + 25}$$

$$=\sqrt{330.24}$$

=18,17

Penelitian ini bertujuan untuk menerapkan metode data mining dalam klasifikasi penduduk miskin di Kabupaten Labuhanbatu dengan menggunakan algoritma Random Forest (RF) dan K-Nearest Neighbors (KNN). Pengujian dilakukan untuk mengevaluasi performa kedua metode dalam mengklasifikasikan penduduk miskin berdasarkan variabel pekerjaan, penghasilan, tempat tinggal dan pendidikan yang tersedia. Pengujian ini menggunakan Google Colab berbasis Python. Data set yang digunakan berjumlah 21137 Jiwa yang akan di uji. Dengan menggunakan Algoritam Algoritma Random Forest (RF) dan K-Nearest Neighbors (KNN). Proses yang akan dilaksanakan adalah Eksplorasi data, Pembersihan Data, Persiapan Data, Pemisahan Data, Pelatihan Model, Optimasi Model, Evaluasi Model.
Pemuatan data atau data loading adalah tahapan penting dalam proses pengolahan data, yaitu proses mengimpor dan memuat data dari berbagai sumber dalam hal ini data yang di input dalam bentuk xlsx. Pemuatan data ini bertujuan untuk membaca Fram data baik jumlah data serta kategori dan jumlah kolom yang ada di data yang akan di uji. Berikut code di python
Berikut hasil dari pemuatan data

Tabel 3. Hasil Pemuatan data.

NO	NAMA	JENIS\nKELAMIN	ALAMAT	STS_\nKAWIN	\nPEKERJAAN	PENGHASILAN	TEMPAT TINGGAL	PENDIDIKAN	Target
1	Yuyun Sarwendha	Perempuan	Bagan Bilah	Kawin	Mengurus Rumah Tangga	0	Milik Pribadi	SD	0
2	Solehuddin	Laki-laki	Bagan Bilah	Belum Kawin	Pelajar/Mahasiswa	0	Sewa	SMA	0
3	Samsuddin	Laki-laki	Dusun I Bagan Bilah	Belum Kawin	Wiraswasta	3200000	Milik Pribadi	SMA	1
4	Yusnani Batubara	Perempuan	Dusun I Bagan Bilah	Kawin	Mengurus Rumah Tangga	0	Milik Pribadi	SD	0
5	Sugiono	Laki-laki	Dusun I	Kawin	Pedagang	3500000	Milik Pribadi	S1	1
…	…	…	…	…	…	…	…	…	…
x	xxxx	xxxx	xxxx	xxxx	xxxx	xxxx	xxxx	xxxx	xxxx

BULLETIN OF INFORMATION TECHNOLOGY (BIT)
Vol 6, No 2, Juni 2025, Hal. 23 - 35
ISSN 2722-0524 (media online)
DOI 10.47065/bit.v5i2.1783
https://journal.fkpt.org/index.php/BIT

3.3 Pemisahan Data

Pemisahan data merupakan tahap lanjutan dalam proses persiapan data yang bertujuan untuk membagi dataset menjadi dua bagian utama, yaitu data latih (training set) dan data uji (testing set). Proses ini dilakukan agar model yang dibangun dapat diuji keandalannya secara objektif, tanpa dipengaruhi oleh data yang sudah digunakan saat proses pelatihan. Secara umum, data latih digunakan untuk membangun dan menyesuaikan parameter model, sedangkan data uji digunakan untuk mengevaluasi kinerja model terhadap data yang belum pernah dilihat sebelumnya. Dengan demikian, pemisahan data membantu menghindari masalah overfitting, yaitu kondisi ketika model terlalu cocok dengan data latih tetapi tidak mampu melakukan generalisasi terhadap data baru.

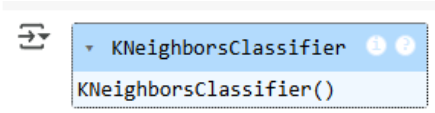
Dalam praktiknya, pemisahan data dilakukan dengan proporsi tertentu, misalnya 80% untuk pelatihan dan 20% untuk pengujian, meskipun proporsi ini dapat disesuaikan dengan ukuran dan karakteristik dataset. Dalam penelitian ini, pemisahan data menjadi langkah krusial untuk memastikan bahwa model prediksi status kemiskinan yang dibangun menggunakan algoritma seperti Random Forest dan K-Nearest Neighbors dapat diuji efektivitasnya secara adil dan terukur. Hasil dari tahap ini akan menjadi dasar untuk menilai sejauh mana model mampu mengklasifikasikan individu ke dalam kategori miskin atau tidak miskin secara akurat pada data baru.

```
X_train shape: (19394, 58)
y_train shape: (19394,)
X_test shape: (2424, 58)
y_test shape: (2424,)
X_val shape: (2425, 58)
y_val shape: (2425,)
```

3.4 Pelatihan Model

Pelatihan model merupakan tahapan inti dalam proses pengembangan sistem prediktif berbasis data mining, di mana algoritma pembelajaran mesin (machine learning) digunakan untuk mengenali pola dari data latih yang telah dipersiapkan sebelumnya. Pada tahap ini, model dilatih untuk memahami hubungan antara variabel-variabel independen (fitur) dan variabel dependen (target), sehingga mampu membentuk fungsi prediksi yang optimal. Dalam konteks penelitian kemiskinan, proses pelatihan dilakukan menggunakan algoritma seperti Random Forest dan K-Nearest Neighbors (KNN), yang masing-masing memiliki pendekatan berbeda dalam membangun model prediktif.

Random Forest bekerja dengan membangun sejumlah pohon keputusan (decision trees) secara acak, kemudian menggabungkan hasil dari masing-masing pohon untuk menghasilkan prediksi yang lebih stabil dan akurat. Sementara itu, KNN menentukan klasifikasi berdasarkan kedekatan data baru terhadap sejumlah tetangga terdekat dalam ruang fitur. Selama pelatihan, algoritma akan mengevaluasi performanya menggunakan metrik tertentu, seperti akurasi, presisi, dan recall, terhadap data latih. Tahap ini sangat penting karena kualitas model yang dihasilkan sangat bergantung pada bagaimana data dipelajari oleh algoritma. Oleh karena itu, pelatihan model harus dilakukan dengan pendekatan yang cermat, disertai dengan validasi agar model tidak hanya baik dalam mengenali data yang telah dipelajari, tetapi juga mampu melakukan generalisasi pada data baru secara efektif.



3.5 Optimasi Model

Optimasi model merupakan tahapan penting dalam pengembangan sistem pembelajaran mesin yang bertujuan untuk meningkatkan kinerja model melalui penyesuaian parameter-parameter yang memengaruhi proses pembelajaran. Dalam konteks ini, optimasi dilakukan dengan cara mengatur hyperparameter, yaitu parameter yang nilainya ditentukan sebelum proses pelatihan model dimulai dan tidak dipelajari langsung dari data. Contohnya pada algoritma Random Forest, jumlah pohon keputusan (n_estimators) dan kedalaman maksimum pohon (max_depth) dapat diatur untuk memperoleh hasil klasifikasi yang lebih akurat. Sementara pada K-Nearest Neighbors (KNN), pemilihan nilai k yang tepat (jumlah tetangga terdekat) menjadi kunci dalam menentukan keandalan prediksi.

Proses optimasi umumnya dilakukan melalui teknik seperti Grid Search atau Random Search, di mana sistem secara sistematis menguji berbagai kombinasi parameter untuk menemukan konfigurasi terbaik berdasarkan metrik evaluasi tertentu, seperti akurasi, precision, recall, atau nilai F1-score. Dalam penelitian ini, optimasi menjadi langkah strategis agar model yang dibangun tidak hanya mampu mengenali pola dari data latih, tetapi juga menunjukkan performa yang konsisten dan dapat diandalkan ketika diuji dengan data baru. Dengan optimasi yang tepat, model prediksi kemiskinan dapat bekerja lebih efisien, mengurangi kesalahan klasifikasi, serta memberikan hasil yang lebih valid secara ilmiah.

```
from sklearn.model_selection import GridSearchCV
Best hyperparameters for Random Forest: {'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 5,
'n_estimators': 200}
Best hyperparameters for K-NN: {'metric': 'euclidean', 'n_neighbors': 3, 'weights': 'uniform'}
```

BULLETIN OF INFORMATION TECHNOLOGY (BIT)
Vol 6, No 2, Juni 2025, Hal. 23 - 35
ISSN 2722-0524 (media online)
DOI 10.47065/bit.v5i2.1783
https://journal.fkpt.org/index.php/BIT

3.6 Evaluasi Model

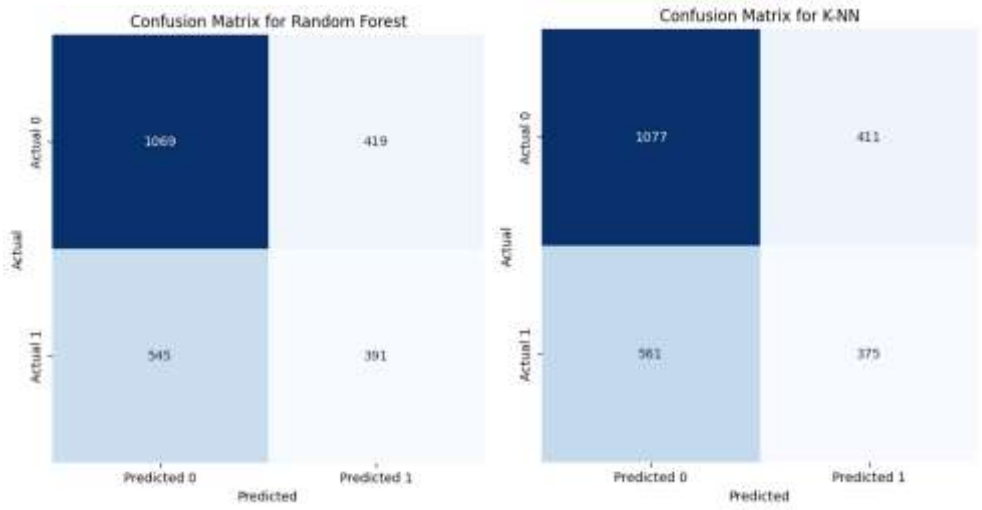
Evaluasi model merupakan tahapan akhir dalam proses pembangunan sistem pembelajaran mesin yang bertujuan untuk menilai kinerja model berdasarkan data uji yang tidak digunakan selama pelatihan. Langkah ini sangat penting untuk mengukur sejauh mana model mampu melakukan generalisasi terhadap data baru dan memberikan prediksi yang akurat. Evaluasi dilakukan dengan menggunakan sejumlah metrik yang relevan, seperti akurasi, presisi, recall, dan F1-score, yang masing-masing memberikan perspektif berbeda terkait performa model. Akurasi mengukur proporsi prediksi yang benar secara keseluruhan, sementara presisi dan recall lebih menekankan pada kualitas prediksi positif, terutama penting dalam kasus klasifikasi ketidakseimbangan kelas seperti pada data kemiskinan.

Dalam penelitian ini, evaluasi dilakukan terhadap model yang dibangun menggunakan algoritma Random Forest dan K-Nearest Neighbors (KNN). Hasil evaluasi tidak hanya memberikan gambaran tentang kekuatan model dalam mengklasifikasikan individu ke dalam kategori miskin atau tidak miskin, tetapi juga menjadi dasar dalam menentukan model mana yang paling sesuai untuk digunakan dalam implementasi nyata. Evaluasi juga membuka ruang untuk interpretasi yang lebih luas mengenai variabel-variabel yang paling berpengaruh terhadap status kemiskinan. Dengan demikian, proses evaluasi tidak hanya bersifat teknis, melainkan juga mendukung pengambilan keputusan berbasis data secara ilmiah dan bertanggung jawab.

```
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score, roc_auc_score, confusion_matrix
import matplotlib.pyplot as plt
import seaborn as sns
```

Evaluation Metrics for Random Forest:
- Accuracy: 0.6023
- Precision: 0.4827
- Recall: 0.4177
- F1-score: 0.4479
- AUC: 0.5681

Evaluation Metrics for K-NN:
- Accuracy: 0.5990
- Precision: 0.4771
- Recall: 0.4006
- F1-score: 0.4355
- AUC: 0.5622



Model Comparison:

	Metric	Random Forest	K-NN
0	Accuracy	0.602310	0.599010
1	Precision	0.482716	0.477099
2	Recall	0.417735	0.400641
3	F1-score	0.447881	0.435540
4	AUC	0.568075	0.562216

BULLETIN OF INFORMATION TECHNOLOGY (BIT)
Vol 6, No 2, Juni 2025, Hal. 23 - 35
ISSN 2722-0524 (media online)
DOI 10.47065/bit.v5i2.1783
https://journal.fkpt.org/index.php/BIT

Evaluasi Model: Random Forest mencapai akurasi 0.6023, presisi 0.4827, recall 0.4177, F1-score 0.4479, dan AUC 0.5681 pada dataset uji. Performa KNN sedikit lebih rendah di semua metrik ini.

4. KESIMPULAN

Berdasarkan analisis:Model Random Forest berkinerja sedikit lebih baik daripada model KNN dalam hal akurasi, presisi, recall, F1-score, dan AUC.Namun, perbedaan kinerja tersebut tidak signifikan. Pembersihan Data: Kolom 'PENGHASILAN', yang awalnya bertipe objek, berhasil diubah menjadi numerik.Nilai yang hilang dalam dataset diperhitungkan menggunakan median untuk fitur numerik dan modus untuk fitur kategorikal. Pencilan dalam fitur numerik diurutkan pada persentil ke-1 dan ke-99. Baris duplikat dihapus. Optimasi Model: Hiperparameter terbaik yang ditemukan untuk model Random Forest adalah {'max_depth': Tidak ada, 'min_samples_leaf': 1, 'min_samples_split': 5, 'n_estimator': 200}. Hiperparameter terbaik untuk model KNN adalah {'metric': 'euclidean', 'n_tetangga': 3, 'bobot': 'uniform'}. Evaluasi Model: Random Forest mencapai akurasi 0.6023, presisi 0.4827, recall 0.4177, F1-score 0.4479, dan AUC 0.5681 pada dataset uji. Performa KNN sedikit lebih rendah di semua metrik ini..

REFERENCES

[1] Ernawati, A., & Wahyuni, S. (2024). Analisis Data Mining Pola Penggunaan Seluler dan Klasifikasi Perilaku Pengguna di Berbagai Perangkat Menggunakan Metode C4 . 5. 5(4), 162–168. https://doi.org/10.47065/bit.v5i2.1689
[2] Kamaliah, A. (2024). AI Bisa Bantu Atasi Masalah Keluarga Miskin di Indonesia. Detikinet. https://inet.detik.com/cyberlife/d-7368981/ai-bisa-bantu-atasi-masalah-keluarga-miskin-di-indonesia
[3] Khaerunisa, S., Nur Padilah, T., & Haerul Jaman, J. (2024). Implementasi Data Mining Menggunakan Metode Regresi Data Panel Untuk Memprediksi Capaian Indeks Pembangunan Manusia. JATI (Jurnal Mahasiswa Teknik Informatika), 7(5), 3399–3406. https://doi.org/10.36040/jati.v7i5.7260
[4] Mukhlis, I. R., Hayam, U., Perbanas, W., Pipin, S. J., & Mikroskil, U. (2024). BIG DATA (Mengenal Big Data & Implementasinya di Berbagai Bidang) (Issue February).
[5] Sakti, R., & Daulay, A. (2024). Analisis Kritis dan Pengembangan Algoritma K-Nearest Neighbor (KNN): Sebuah Tinjauan Literatur. 4(2), 131–141.
[6] Saputra, F. A., & Iskandar, A. (2023). Data Mining Penerapan Asosiasi Apriori Dalam Penentuan Pola Penjualan. Journal of Computer System and Informatics (JoSYC), 4(4), 778–788. https://doi.org/10.47065/josyc.v4i4.4043
[7] Sari, R. P. (2024). Apa itu Data Mining? Pengertian, Metode dan Penerapannya. Cloud Computing. https://www.cloudcomputing.id/pengetahuan-dasar/apa-itu-data-mining
[8] Sis. (2024). Random Forest vs Decision Tree. https://sis.binus.ac.id/2024/04/02/random-forest-vs-decision-tree/
[9] sumut. (2019). Sejarah Sumatera Utara. https://sumutprov.go.id/artikel/halaman/sejarah
[10] Wahyuni, S. (2018). Implementation of Data Mining to Analyze Drug Cases Using C4.5 Decision Tree. Journal of Physics: Conference Series, 970(1). https://doi.org/10.1088/1742-6596/970/1/012030
[11] Ali, M. M., Hariyati, T., Pratiwi, M. Y., & Afifah, S. (2022). Metodologi Penelitian Kuantitatif dan Penerapannya dalam Penelitian. Education Journal.2022, 2(2), 1–6.
[12] BPS. (2024). Profil Kemiskinan Kabupaten Labuhanbatu Maret 2024. BPS. https://labuhanbatukab.bps.go.id/id/pressrelease/2024/07/29/266/profil-kemiskinan-kabupaten-labuhanbatu-maret-2024.html
[13] BREIMAN, L. (2001). Random Forests LEO. Kluwer Academic Publishers. Manufactured in The Netherlands, 12343 LNCS, 503–515. https://doi.org/10.1007/978-3-030-62008-0_35
[14] Sitorus, Z., Hariyanto, E., & Kurniawan, F. (2023). Analysis of Artificial Intelligence Machine Learning Technology for Mapping and Predicting Flood Locations in Pahlawan Batu Bara Village. 2(2). CV Hawari
[15] Suci Amaliah, Nusrang, M., & Aswi, A. (2022). Penerapan Metode Random Forest Untuk Klasifikasi Varian Minuman Kopi di Kedai Kopi Konijiwa Bantaeng. VARIANSI: Journal of Statistics and Its Application on Teaching and Research, 4(3), 121–127. https://doi.org/10.35580/variansiunm31
[16] sumut. (2019). Sejarah Sumatera Utara. https://sumutprov.go.id/artikel/halaman/sejarah
[17] W, R. S. A., Hariyanto, E., & Sitorus, Z. (2022). COMPARISONAL ANALYSIS OF EUCLIDEAN , CANBERRA , AND CHEBECHEV DISTANCE MODELS ON KNN METHOD ON STUDENTS ' VALUE. 10(3), 315–318.
[18] Wikipedia. (2024). Kabupaten Labuhanbatu. https://id.wikipedia.org/wiki/Kabupaten_Labuhanbatu
[19] B. Bangun and A. K. Karim, "Pengembalian Data Yang Hilang Pada Dataset Dengan Menggunakan Algoritma K-Nearest Neighbor Imputation Data Mining," Jurnal Media Informatika Budidarma, vol. 8, no. 3, p. 1706, 2024, doi: 10.30865/mib.v8i3.8014.
[20] A. Karim, "Penerapan Algoritma Entropy dan Aras Menentukan Desa Terbaik Di Pemerintah Kabupaten Labuhanbatu," vol. 3, no. 1, pp. 33–43, 2022.
[21] A. Karim, "Implementation of the Multi-Objective Optimization Method on the Basic of Ratio Analysis (MOORA) and Entropy Weighting in New Employee Recruitment," vol. 5, no. 2, pp. 704–712, 2024, doi: 10.47065/josh.v5i2.4859.
[22] A. Karim, "Clusterisasi Tingkat Pengangguran Terbuka Menurut Provinsi di Indonesia Menggunakan Algoritma K-Medoids," 2024, doi: 10.47065/bits.v6i3.6198.
[23] A. Karim, "Sistem Pendukung Keputusan Penerimaan Analis Di Pusat Penelitian Kelapa Sawit Menggunakan Metode Complex Proportional Assessment (Copras)," Buletin Ilmiah Informatika Teknologi, vol. 2, no. 1, pp. 32–42, [Online]. Available: https://ejurnal.amikstiekomsu.ac.id/index.php/BIIT
[24] Abdul Karim, "Implementasi Metode Multi-Objective Optimization On The Basis Of Ratio Analysis dalam Seleksi Mahasiswa Program Indonesia Pintar," Bulletin of Computer Science Research, vol. 3, no. 5, pp. 351–356, 2023, doi: 10.47065/bulletincsr.v3i5.283.

BULLETIN OF INFORMATION TECHNOLOGY (BIT)
Vol 6, No 2, Juni 2025, Hal. 23 - 35
ISSN 2722-0524 (media online)
DOI 10.47065/bit.v5i2.1783
https://journal.fkpt.org/index.php/BIT

[25] Z. Budiarso, H. Listiyono, and A. Karim, "Optimizing LSTM with Grid Search and Regularization Techniques to Enhance Accuracy in Human Activity Recognition," Journal of Applied Data Sciences, vol. 5, no. 4, pp. 2002–2014, Dec. 2024, doi: 10.47738/jads.v5i4.433.