ISSN: 2598-0256, E-ISSN: 2598-0238

# Kajian Algoritma C4.5 dan K-NN Untuk Memprediksi Penduduk Miskin

Muhamad Septa Utama SP[1a*], Handoyo Widi Nugroho[2b]

[ab] Institut Informatika dan Bisnis Darmajaya Lampung
[a] *septautama.2221210047@mail.darmajaya.ac.id*
[b]*handoyo.wn@darmajaya.ac.id*

**Abstract**

Poverty is a serious issue that affects people all over the world. The United Nations (UN) has recognised poverty as a top priority in its Sustainable Development Goals. Although there has been a decline in poverty rates in recent years, there are still many individuals who struggle to fulfil their basic needs. Therefore, more effective efforts are needed in identifying the poor so that aid programmes can be well-targeted. This research aims to compare two classification methods, namely C4.5 and K-Nearest Neighbors (KNN), in predicting poverty rates. The C4.5 method uses a decision tree to classify the data, while KNN uses the closest distance to perform the classification. The data used in this research is poverty data in Indonesia. This research methodology involves data pre-processing stages, including data cleaning, feature selection, data exploration, and data balancing. Next, model training and testing using the C4.5 and KNN algorithms were conducted. The model performance is evaluated using metrics such as accuracy, recall, precision, F1 measure, and Area Under Curve (AUC). This research is still at the model design stage, and the follow-up will be to continue the research until the algorithm evaluation results. Using the confusion matrix, the best algorithm that can detect the poor with high accuracy will be selected. The results of this research are expected to provide useful insights in developing effective assistance programmes for poverty alleviation in Indonesia.

***Keywords*** : *Poverty; C.45; K-Nearest Neighbours (KNN), Prediction*

**Abstrak**

Kemiskinan adalah isu serius yang mempengaruhi masyarakat di seluruh dunia. Perserikatan Bangsa-Bangsa (PBB) telah mengakui kemiskinan sebagai prioritas utama dalam Tujuan Pembangunan Berkelanjutan. Meskipun ada penurunan tingkat kemiskinan dalam beberapa tahun terakhir, masih banyak individu yang berjuang untuk memenuhi kebutuhan dasar mereka. Oleh karena itu, diperlukan upaya yang lebih efektif dalam mengidentifikasi penduduk miskin agar program-program bantuan dapat tepat sasaran. Penelitian ini bertujuan untuk membandingkan dua metode klasifikasi, yaitu C4.5 dan K-Nearest Neighbors (KNN), dalam memprediksi tingkat kemiskinan. Metode C4.5 menggunakan decision tree untuk mengklasifikasikan data, sementara KNN menggunakan jarak terdekat untuk melakukan klasifikasi. Data yang digunakan dalam penelitian ini adalah data kemiskinan di Indonesia. Metodologi penelitian ini melibatkan tahapan pra-pemrosesan data, termasuk pembersihan data, seleksi fitur, eksplorasi data, dan balancing data. Selanjutnya, dilakukan pelatihan dan pengujian model menggunakan algoritma C4.5 dan KNN. Hasil evaluasi kinerja model menggunakan metrik seperti akurasi, recall, presisi, F1 measure, dan Area Under Curve (AUC). Penelitian ini masih berada pada tahap desain model, dan tindak lanjut yang akan dilakukan adalah melanjutkan penelitian hingga hasil evaluasi algoritme. Dengan menggunakan confusion matrix, akan dipilih algoritme terbaik yang dapat mendeteksi penduduk miskin dengan akurasi yang tinggi. Hasil penelitian ini diharapkan dapat memberikan wawasan yang berguna dalam mengembangkan program-program bantuan yang efektif untuk pengentasan kemiskinan di Indonesia.

**Kata Kunci** : Kemiskinan; C.45; K-Nearest Neighbors (KNN), Prediksi.

## 1. PENDAHULUAN

Kemiskinan adalah isu serius yang mempengaruhi masyarakat di seluruh dunia. Perserikatan Bangsa-Bangsa (PBB) telah mengakui kemiskinan sebagai prioritas utama dalam Tujuan Pembangunan Berkelanjutan (SDG). Pada tahun 2015, sekitar 736 juta orang, atau sekitar 10% dari populasi dunia, hidup dengan pendapatan kurang dari $1,90 per hari. Angka ini mengalami penurunan yang signifikan sejak tahun 1990, ketika sekitar 36% populasi dunia hidup dalam kemiskinan. Meskipun demikian, penurunan kemiskinan telah melambat dalam beberapa tahun terakhir, dan masih banyak individu yang berjuang untuk memenuhi kebutuhan dasar mereka (World Bank, 2018).

Menurut World Bank, kemiskinan didefinisikan sebagai ketidakcukupan atau kekurangan kesejahteraan(World Bank, 2020). World Bank mengukur kemiskinan secara absolut dengan menggunakan pendapatan sebagai parameter, dengan menetapkan batas kemiskinan pada pendapatan di bawah $2 per hari. Namun, Badan Pusat Statistik memiliki

ISSN: 2598-0256, E-ISSN: 2598-0238

pendekatan yang sedikit berbeda, menganggap kemiskinan sebagai ketidakmampuan memenuhi kebutuhan dasar, terutama dalam hal pangan, berdasarkan kemampuan ekonomi daripada pengeluaran.

Selama dekade terakhir, Indonesia telah mengalami penurunan yang signifikan dalam tingkat kemiskinan, seperti yang tercatat oleh Badan Pusat Statistik. Upaya terus dilakukan untuk mengatasi masalah kemiskinan di negara ini dengan hasil yang positif terlihat dari penurunan tingkat kemiskinan dari tahun ke tahun. Pada tahun 2018, persentase penduduk miskin di Indonesia mencapai 9,66%, yang menandai pertama kalinya tingkat kemiskinan berada di bawah angka 10%. Meskipun pandemi Covid-19 telah mempengaruhi Indonesia pada awal tahun 2022 dan menyebabkan peningkatan angka kemiskinan, dengan tingkat kemiskinan mencapai 10,19% pada tahun 2021, tren secara keseluruhan menunjukkan penurunan tingkat kemiskinan dalam periode tahun 2011 hingga 2022.

Pemerintah telah meluncurkan banyak program untuk mengurangi kemiskinan, tetapi ironisnya beberapa program bantuan belum tepat sasaran. Menurut penelitian yang dilakukan oleh Arif Sofianto, sekitar 21,54 persen penerima Program Keluarga Harapan (PKH) bukanlah orang miskin (Sofianto, 2020). Hal ini menunjukkan pentingnya penggunaan teknologi mesin learning klasifikasi dalam mengidentifikasi penduduk miskin. Dengan adanya teknologi ini, diharapkan dapat lebih akurat dalam mengklasifikasikan penerima manfaat program-program bantuan yang ditujukan untuk pengentasan kemiskinan.

Proses klasifikasi melibatkan pembuatan model atau kerangka kerja yang dapat mengategorikan dan menjelaskan berbagai kelas atau konsep data yang berbeda. Tujuan utamanya adalah untuk memprediksi kelas yang tidak diketahui dari objek yang diamati (Han et al., 2011). Para peneliti telah mengembangkan berbagai metode klasifikasi, di antaranya adalah pendekatan decision tree C4.5 dan K-Nearest Neighbour (KNN). Metode C4.5 menawarkan beberapa keuntungan, seperti hasil yang mudah diinterpretasikan, penghapusan fitur yang tidak relevan, dan efisiensi yang lebih tinggi dibandingkan dengan model yang lebih kompleks (Lantz, 2015). Dengan menggunakan metode ini, indikator-indikator yang paling relevan yang secara efektif menjelaskan kemiskinan dapat diidentifikasi dari berbagai variabel yang tersedia.

Decision tree mampu mengintegrasikan model yang sederhana ke dalam sistem basis data dan memiliki tingkat akurasi yang baik dalam mengklasifikasikan kemiskinan, seperti yang terlihat dalam penelitian Kaunang, di mana decision tree memberikan performa hingga 80 persen (F.J. Kaunang, 2018). Di sisi lain, algoritma KNN memiliki kelebihan seperti ketangguhan terhadap data latih yang memiliki banyak noise dan efektivitasnya ketika digunakan pada data latih yang besar (W. Yustanti, 2012). Pardomuan melakukan klasifikasi kemiskinan menggunakan metode KNN dan mencapai tingkat akurasi hingga 73 persen (Sihombing & Arsani, 2021). Hasil penelitian Dita Noviana menunjukkan bahwa algoritma KNN memiliki akurasi 90,7% dan tingkat kesalahan 9,3%, sedangkan algoritma C4.5 memiliki akurasi 88,3% dan tingkat kesalahan 11,7% (Noviana et al., 2019). Berdasarkan penelitian-penelitian sebelumnya, peneliti merasa perlu membandingkan metode C4.5 dan KNN dalam memprediksi tingkat kemiskinan.

## 2. KERANGKA TEORI

### *2.1. Kemiskinan*

Kemiskinan didefinisikan sebagai ketidakmampuan finansial untuk memenuhi kebutuhan penting seperti makanan dan barang kebutuhan lainnya, yang diukur berdasarkan pengeluaran (BPS, 2022a). Orang yang dianggap miskin adalah mereka yang memiliki pengeluaran per kapita (per bulan) di bawah garis kemiskinan. World Bank, di sisi lain, mengartikan kemiskinan sebagai ketidakmampuan seseorang untuk memenuhi kebutuhan hidupnya dan kesulitan dalam memanfaatkan sumber daya yang tersedia untuk memenuhi kebutuhan tersebut (World Bank., 2001).

Konsep kemiskinan dapat dibagi menjadi dua kategori, yaitu kemiskinan relatif dan kemiskinan absolut, dengan standar penilaian yang berbeda. Standar penilaian kemiskinan absolut mengacu pada kebutuhan minimum untuk memenuhi kebutuhan dasar seperti makanan dan non-makanan, yang dikenal sebagai garis kemiskinan. Di sisi lain, standar penilaian kemiskinan relatif ditetapkan oleh masyarakat setempat dan bersifat lokal, di mana mereka yang berada di bawah standar tersebut dianggap miskin secara relatif (BPS, 2022b).

Hal ini menunjukkan bahwa kemiskinan tidak hanya melibatkan pendapatan yang rendah, tetapi juga melibatkan faktor-faktor lain yang memengaruhi kemampuan seseorang untuk mencapai standar kehidupan yang memadai. Aspek-aspek seperti sandang (pakaian), pangan (makanan), dan papan (tempat tinggal) memainkan peran penting dalam mempengaruhi kemampuan seseorang untuk memenuhi kebutuhan dasar mereka.

Ketika seseorang tidak memiliki akses yang memadai terhadap sandang, pangan, dan papan, ini dapat mengurangi kapabilitas mereka dalam memenuhi kebutuhan kesehatan dan pendidikan. Misalnya, ketidakmampuan untuk membeli makanan bergizi atau tidak memiliki tempat tinggal yang layak dapat berdampak negatif pada kesehatan seseorang. Demikian pula, ketika seseorang tidak memiliki pakaian yang cukup atau akses yang memadai terhadap pendidikan, hal ini dapat menghambat kemampuan mereka untuk meningkatkan kualitas hidup dan memperoleh pengetahuan serta keterampilan yang diperlukan.

Oleh karena itu, kemiskinan harus dipahami sebagai masalah yang kompleks yang melibatkan lebih dari sekadar pendapatan. Aspek-aspek seperti sandang, pangan, dan papan juga harus diperhatikan dalam upaya untuk mengatasi kemiskinan secara holistik.

ISSN: 2598-0256, E-ISSN: 2598-0238

*2.2. Decision Tree C4.5*

C4.5 adalah algoritma decision tree yang dikembangkan oleh Ross Quinlan pada tahun 1993. Algoritma ini merupakan pengembangan dari ID3 (Iterative Dichotomiser 3) dan menjadi dasar bagi algoritma decision tree yang lebih modern seperti CART dan CHAID. C4.5 menggunakan pendekatan top-down divide and conquer untuk membangun pohon keputusan. Data awal digunakan untuk membuat root node, kemudian data dibagi menjadi subset yang lebih kecil berdasarkan aturan tertentu. Proses ini diulang hingga terbentuk sebuah tree yang dapat memprediksi kelas target dari data input.

C4.5 memiliki beberapa kelebihan. Algoritma ini dapat menangani data yang kompleks, seperti data dengan atribut yang beragam. Selain itu, C4.5 juga robust terhadap nilai yang hilang pada data (missing data) dan dapat bekerja dengan variabel target yang bersifat nominal maupun numerik. Kelebihan lainnya adalah kemampuannya dalam memilih atribut yang paling penting dalam memisahkan data. C4.5 menggunakan metode gain ratio untuk memilih atribut pada setiap node. Metode ini mengukur informasi yang diberikan oleh sebuah atribut dan mempertimbangkan jumlah subset yang dihasilkan dari atribut tersebut.

Setelah pohon keputusan terbentuk, C4.5 melakukan pruning atau pemotongan cabang yang tidak penting untuk mengurangi overfitting pada data latih. Pruning dilakukan dengan memperhitungkan nilai probabilitas pada setiap leaf node, sehingga menghasilkan prediksi yang lebih akurat. Algoritma C4.5 memiliki berbagai aplikasi, seperti dalam klasifikasi data, pengenalan pola, dan pengambilan keputusan. Dengan keunggulan-keunggulannya, C4.5 menjadi salah satu algoritma decision tree yang banyak digunakan dalam analisis dan pemrosesan data. Pertama-tama, entropi data (informasi yang diharapkan) yang diperlukan untuk mengklasifikasikan tupel dalam D didefinisikan sebagai berikut:

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i) \quad (1)$$

di mana:

Info(D) adalah entropi dari set data D.
$p_i$ adalah proporsi jumlah sampel yang termasuk dalam kelas i terhadap total jumlah sampel.
log2 adalah logaritma basis 2.

Kita anggap bahwa telah dilakukan penerapan beberapa atribut A pada tupel yang terbagi dalam D. Atribut A memiliki v nilai unik sesuai dengan data pelatihan $\{a\ 1\ ,\ a\ 2\ ,\ \ldots\ .\ a\ v\ \}$. Dengan menggunakan atribut A tersebut, tupel-tupel dalam D dapat dibagi menjadi v subset $\{D\ 1\ ,\ D\ 2\ ,\ \ldots\ .\ D\ v\ \}$.

Untuk menentukan klasifikasi yang benar, kita perlu mengevaluasi informasi dengan menggunakan langkah-langkah berikut:

$$Info\ A(D) = \sum_{i=1}^{m} \frac{|D_v|}{|D|} X_i Info(D_i) \quad (2)$$

Dimana $\frac{|D_v|}{|D|}$ adalah penimbang dari split jth

Perolehan Information Gain adalah hasil dari perbedaan antara informasi awal dan informasi yang diperoleh setelah menggunakan atribut tertentu.

$Gain$ (D, $A$ ) = $Info$ ( $D$ ) - $Info\ A$ ($D$) (3)

di mana:

Gain(D, A) adalah Information Gain dari set data D setelah pemisahan menggunakan atribut A.
Info(D) adalah entropi dari set data D.
$D_V$ adalah subset dari set data D yang hasil pemisahannya menggunakan atribut A.
$|D_V|$ adalah jumlah sampel dalam subset $D_V$.
|D| adalah jumlah sampel dalam set data D.

Rasio perolehan didefinisikan sebagai berikut:

$$Gain \text{ Ratio (D,}A) = \frac{Gain(D,A)}{\text{Splitinfo (D,}A)} \quad (4)$$

$$\text{SplitInfo(D, A)} = -\sum \frac{|\text{Dv}|}{|\text{D}|} \log_2 \frac{|\text{Dv}|}{|\text{D}|} \quad (5)$$

ISSN: 2598-0256, E-ISSN: 2598-0238

di mana:

SplitInfo(D, A) adalah Split Info dari set data D setelah pemisahan menggunakan atribut A.
Dv adalah subset dari set data D yang hasil pemisahannya menggunakan atribut A.
|Dv| adalah jumlah sampel dalam subset Dv.
|D| adalah jumlah sampel dalam set data D.
log2 adalah logaritma basis 2.
GainRatio(D, A) adalah Gain Ratio dari set data D setelah pemisahan menggunakan atribut A.
Gain(D, A) adalah Information Gain dari set data D setelah pemisahan menggunakan atribut A.

*2.3. K-Nears Neighboard (KNN)*

K-Nearest Neighbor (KNN) adalah salah satu algoritma klasifikasi yang sederhana dan populer dalam Machine Learning. Algoritma KNN digunakan untuk mengklasifikasikan data berdasarkan kategori atau label yang dimiliki oleh tetangga terdekatnya.

Ide dasar dari algoritma KNN adalah mencari tetangga terdekat dari suatu data dengan menggunakan jarak Euclidean atau jarak Minkowski dengan parameter k. Setelah itu, data yang akan diklasifikasikan akan diberi label yang sama dengan mayoritas label tetangga terdekatnya.

Dalam algoritma KNN, nilai k menunjukkan jumlah tetangga terdekat yang akan digunakan untuk menentukan label klasifikasi data. Proses pencarian tetangga terdekat dilakukan dengan menghitung jarak antara data yang akan diklasifikasikan dengan data latih yang sudah diberi label kelasnya.

Keunggulan dari algoritma KNN adalah kemudahan implementasinya dan tingkat akurasi klasifikasinya yang cukup baik. Namun, algoritma ini memiliki kelemahan dalam mengklasifikasikan data yang memiliki fitur atau atribut yang kompleks dan banyak, serta memiliki nilai-nilai yang tidak terstandarisasi. Jarak Euclidean digunakan untuk mengukur jarak antara dua titik dalam ruang dengan menggunakan koordinat Euclidean. Rumus jarak Euclidean antara dua titik $(x_1, y_1)$ dan $(x_2, y_2)$ adalah:

$$D(x_1, y_1) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \quad (6)$$

di mana:

D(x1, x2) adalah jarak Euclidean antara dua titik.
x1, y1 adalah koordinat titik pertama.
x2, y2 adalah koordinat titik kedua.

*2.4. Penelitian Terkait*

Metode pembelajaran mesin atau Machine Learning (ML) juga telah diterapkan pada data survei sebagai alat untuk mengklasifikasikan kemiskinan, baik secara mandiri maupun dalam kombinasi dengan data lain untuk validasi atau melengkapi informasi. Beberapa studi yang menggunakan metode klasifikasi ML pada data survei untuk memprediksi kemiskinan dijelaskan pada tabel 2.1 di bawah ini.

Tabel 1. Penelitian Terdahulu Tentang Prediksi Kemiskinan dengan Machine Learning

| Belajar | Himpunan data | Teknik Pengambilan Sampel | Seleksi Fitur | Klasifikasi | Pertunjukan |
|---|---|---|---|---|---|
| (1) | (2) | (3) | (4) | (5) | (6) |
| (Fitzpatrick, 2018) | Data survei rumah tangga dari sebuah nasional sampel perwakilan Malawi, Indonesia | Synthetic Minority Oversampling Technique (SMOTE),, untuk Dataset Indonesia | Variabel kategori (menanggapi survei pertanyaan) adalah sebagai kemiskinan prediktor.<br><br>Klasifikasi adalah dilakukan dengan menggunakan full fitur dataset dan satu set dari fitur terpilih | 10 open-source ML classification algoritma klasifikasi adalah diterapkan untuk memprediksi kemiskinan dan hasil dibandingkan.<br><br>Variabel prediksi target dari semua metode adalah label biner "miskin" atau "tidak miskin", untuk mengklasifikasikan rumah tangga | MALAWI: Akurasi = 78 - 87% untuk rangkaian fitur lengkap; dan = 73 - 77% untuk set fitur sederhana<br><br>INDONESIA: Akurasi = 88% - 91% untuk fitur lengkap mengatur; dan = 90 - 91% lebih sedikit set fitur |

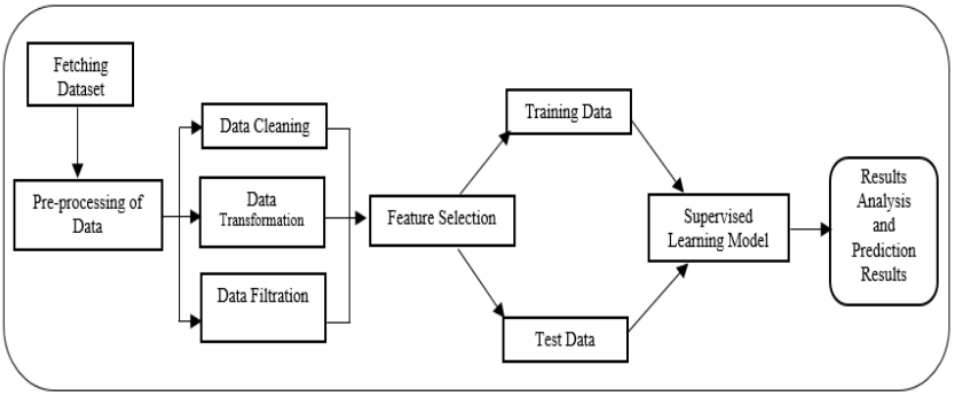
ISSN: 2598-0256, E-ISSN: 2598-0238

| | | | | | |
|---|---|---|---|---|---|
| | | | Variance inflation factor (VIF) pengujian untuk menghilangkan fitur yang berlebihan menghasilkan set terpilih | | |
| (Thoplan, 2014) | data survei, 296 294 | Random Forest Bagging | Gini Score | Bootstrap pada pohon diterapkan pada a sampel acak dari observasi<br><br>Prediksi out-of-bag (OOB). diperoleh dengan menggunakan suara terbanyak melintasi pepohonan | Keluar dari tas (OOB) Kesalahan Mean = 0,175 |
| (McBride, 2018) | data survei, 1800- 11280 | Regression forest<br><br>Quantile regressionforest bootstrap aggregation | Random forest for feature selection | Metode ansambel ( Bagging dan kemudian menerapkan random forest untuk klasifikasi)<br><br>Pemilihan model berdasarkan validasi silang | Akurasi Total<br>Rata-Rata Malawi = 80%<br><br>Rata-rata Timor Timur = 75%<br><br>Rata-rata Bolivia = 64% |
| Kshirsagar , Wieczorek ,Ramathan dan Wells (2017) | LCMS Zambia 2015 data survei | Elastic net logistic regression | Variabel bootstrap Pilihan | Elastic net logistic regression | Probabilitas = 0,85 |
| Knippenberg, Jensenand Constas (2019) | data survei, 576 rumah tangga | N/A | Least Absolute Shrinkageand Selection Operator (LASSO) dan Random Forest untuk mengidentifikasi Prediktor yang terbaik. | LASSO dan Random Forest untuk mengidentifikasi prediktor terbaik dibutuhkan. | Keluar dari sampel (April, Mei)<br>LASSO $r^2$= 56,4%<br>Random Forest $r^2$ = 55,6% |
| Sohnesen dan Stender (2017) | data survei, 1800 – 18000 Di 6 negara | Random Forest | Entropy loss function<br>Gini loss function | Random Forest | National Mean square error (MSE)<br>Gini = 1,71<br>Entropi = 1,94<br>UMK Perkotaan/Pedesaan<br>Gini = 2,58<br>Entropi = 2,58 |
| Gravemeyer , Gries dan Xue (2010) | data survei, 1056 rumah tangga dan 3256 orang | Regresi logit<br>Regresi Tobit<br>Regresi probit | Empirical Truncated Censored | Metode statistik regresi untuk mengukur kemiskinan<br>Regresi diterapkan, dan variables are truncated; others are censored. Ini memungkinkan kita untuk memiliki koefisien | Probit<br>$r^2$ = 74%<br>Menggigit<br>$r^2$ = 75%<br>OLS<br>$r^2$ = 53,6% |

## 3. METODOLOGI

Metodologi penelitian ini akan dilaksanakan berdasarkan tahapan penelitian yang diilustrasikan pada gambar 1 sebagai berikut.

Seminar Nasional Hasil Penelitian dan Pengabdian Masyarakat 2023
Institut Informatika dan Bisnis Darmajaya, 3 Agustus 2023
ISSN: 2598-0256, E-ISSN: 2598-0238




Gambar 1. Kerangka Pikir Penelitian

Gambar 1 mengilustrasikan langkah-langkah penelitian, dimulai dengan pengumpulan data dan berlanjut ke *preprocessing*. Setelah itu, *algoritme C4.5* dan KNN di *training* dan *testing* menggunakan data yang telah disediakan. Dengan menggunakan *confusion matrix, algoritme* ini dievaluasi sebagai langkah terakhir penelitian.

### *3.1. Pre-Prosessing Data*

Tahapan awal dalam pemodelan pada penelitian ini adalah pra-pemrosesan data. Pra-pemrosesan data dilakukan sebelum data dapat digunakan untuk melatih dan menguji model prediktif. Tujuan dari tahap ini adalah membersihkan, mengubah, dan mempersiapkan data agar sesuai dengan kebutuhan analisis. Tahap pra-pemrosesan data memiliki peranan penting dalam memastikan bahwa data yang digunakan dalam pemodelan prediktif memiliki kualitas yang baik, terstandarisasi, dan siap untuk dianalisis. Dengan melakukan pra-pemrosesan data yang baik, penelitian ini dapat mengurangi kemungkinan terjadinya kesalahan dan bias, serta meningkatkan kemampuan model prediktif untuk menghasilkan hasil yang akurat dan memiliki makna yang signifikan.

#### *3.1.1. Data Cleaning*

Pentingnya pembersihan data (data cleaning) dalam penelitian ini sangatlah signifikan. Data yang bersih dan terstruktur merupakan aspek yang krusial dalam memastikan validitas dan reliabilitas hasil penelitian. Dalam rangka membersihkan data, penelitian ini akan menggunakan metode replace dengan parameter rata-rata untuk menggantikan data yang hilang (missing data) dan data yang mengandung noise. Metode ini akan mengisi nilai yang hilang dengan nilai rata-rata dari data yang tersedia, sehingga mempertahankan konsistensi dan mencegah terjadinya bias dalam analisis data. Dengan melakukan pembersihan data yang efektif dan menggunakan pendekatan yang tepat, penelitian ini akan meningkatkan kualitas data yang digunakan dalam analisis dan memastikan hasil penelitian yang lebih akurat dan dapat diandalkan.

#### *3.1.2. Seleksi Fitur*

Pada penelitian ini, dilakukan seleksi fitur dengan menggunakan metode filter. Metode filter adalah pendekatan yang digunakan untuk mengurutkan fitur-fitur dengan melakukan uji statistik guna menghitung koefisien korelasi antara fitur-fitur input. Pendekatan ini telah diterapkan dalam metode klasifikasi yang dikembangkan oleh Fitzpatrick (Fitzpatrick, 2018). Dalam metode filter, fitur-fitur tersebut diberi peringkat berdasarkan perkalian antara koefisien korelasi dan deviasi standar dari parameter yang terkait dalam data. Dengan menggunakan metode ini, kita dapat mengidentifikasi fitur-fitur yang memiliki hubungan yang kuat dengan variabel target, serta mengurutkannya berdasarkan tingkat relevansi dan kepentingannya dalam analisis yang lebih lanjut.

#### *3.1.3. Eksplorasi Data*

Setelah tahap pra-pemrosesan selesai, dataset digunakan untuk melakukan eksplorasi guna memperoleh pemahaman yang lebih mendalam mengenai hubungan antara fitur-fitur yang ada. Analisis eksplorasi data memiliki peran yang sangat penting dalam menentukan model Machine Learning (ML) yang akan digunakan dan fitur-fitur mana yang perlu diberikan perhatian lebih. Sebuah artikel menjelaskan bahwa analisis data tidak hanya berguna untuk menguji hipotesis yang sudah ada, tetapi juga untuk menemukan hipotesis baru dengan memanfaatkan data yang tersedia (Paper, 2018).

ISSN: 2598-0256, E-ISSN: 2598-0238

Dalam penelitian ini, hubungan antara fitur-fitur dan variabel target "kemiskinan" dieksplorasi. Data dianalisis dengan memvisualisasikan distribusi fitur-fitur terhadap variabel target "kemiskinan" atau dengan mengukur korelasinya. Langkah awal yang dilakukan dalam penelitian ini adalah memetakan fitur numerik yang terdapat dalam dataset Kabupaten Pesawaran terhadap variabel target "kemiskinan". Melalui langkah ini, penelitian ini dapat memperoleh wawasan yang lebih baik mengenai bagaimana fitur-fitur tertentu berkaitan dengan variabel target dan memberikan dasar untuk analisis yang lebih mendalam.

### *3.1.4. Balancing Data*

Dalam penelitian ini, dilakukan upaya untuk menjaga keseimbangan distribusi kelas dalam dataset dengan melakukan balancing data. Ketidakseimbangan kelas dalam data dapat menyebabkan bias pada model yang dilatih, di mana kelas mayoritas memiliki dominasi yang kuat dan kelas minoritas sulit diprediksi dengan akurasi yang tinggi.

Untuk mengatasi masalah ini, teknik yang digunakan dalam penelitian ini adalah SMOTE. Metode SMOTE dapat menghasilkan sampel sintetis yang meningkatkan variasi dalam kelas minoritas, sehingga mengurangi risiko overfitting pada data tersebut.

Dengan melakukan balancing data ini, penelitian ini bertujuan untuk menghasilkan model yang lebih baik dalam memprediksi kelas minoritas, mengurangi bias, dan meningkatkan keseimbangan performa model di semua kelas yang ada.

### *3.2. Pembelajaran: Melatih dan Memilih Model Prediksi*

Ada dua tahap utama dalam *supervised learning*. Pertama, dilakukan pelatihan model prediksi menggunakan data pelatihan yang telah diberi label. Tujuannya adalah menghasilkan model klasifikasi yang dapat dengan akurat memprediksi apakah seseorang termasuk dalam kategori miskin atau tidak.

Pada tahap kedua, model yang telah dilatih diuji menggunakan data uji. Model tersebut akan mencoba memprediksi variabel target dengan benar, yaitu menentukan apakah seseorang termasuk dalam kategori miskin atau tidak. Kinerja model dinilai menggunakan berbagai metrik kinerja, dan model terbaik dipilih berdasarkan hasil perbandingan.

Secara keseluruhan, pelatihan model dalam *supervised learning* melibatkan langkah-langkah ini untuk menghasilkan model yang akurat dalam memprediksi status kemiskinan seseorang.

### *3.2.1 Pemilihan Model*

Karena tujuan dari penelitian ini adalah untuk memprediksi variabel target "miskin" atau "tidak miskin", maka masalah yang dihadapi adalah klasifikasi biner. Setelah data dimasukkan ke dalam Rapidminer, algoritma di dalamnya menganalisis dan mempelajari data pelatihan untuk mencari parameter prediksi terbaik. Parameter ini akan digunakan untuk membuat model klasifikasi. Model klasifikasi tersebut kemudian diuji menggunakan dataset pengujian untuk mengevaluasi kinerjanya dalam memprediksi dengan akurat apakah seseorang termasuk dalam kategori "miskin" atau "tidak miskin".

Dalam penelitian ini, dua model klasifikasi yang dianggap paling sesuai berdasarkan tinjauan literatur di Bab 2, yang membahas metode pembelajaran mesin yang digunakan dalam klasifikasi kemiskinan menggunakan data survei, telah dipilih. Model-model tersebut adalah: i) C4.5 dan; ii) K-Nearest Neighbors. Kedua model ini digunakan untuk melatih, menguji, mengevaluasi, dan memprediksi label dalam masalah klasifikasi biner ini.

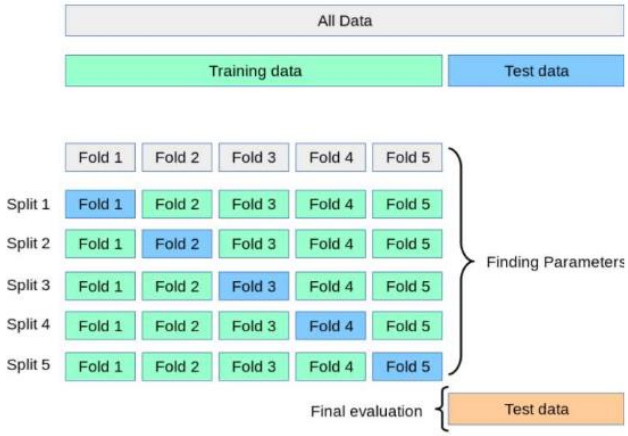
### *3.2.2 Cross-Validation*

Dalam pelatihan dan pengujian model, tidak cukup hanya menggunakan dua subset data tersebut, karena dapat menyebabkan model terlalu terikat pada data pelatihan dan kinerjanya akan menurun saat diterapkan pada data baru. Untuk mengatasi masalah ini, digunakan subset data terpisah yang digunakan sebagai validasi model. Oleh karena itu, dataset dibagi menjadi tiga bagian, yaitu pelatihan, pengujian, dan validasi.

Namun, membagi data menjadi tiga bagian dapat mengurangi jumlah sampel untuk melatih algoritme, yang berpotensi mempengaruhi kinerja model. Untuk mengatasi hal ini, digunakan metode validasi silang. Dalam validasi silang, dataset pelatihan dibagi menjadi k subset yang lebih kecil atau fold. Model kemudian dilatih pada k-1 fold dan diuji pada fold yang tersisa. Proses ini diulang sebanyak k kali, dan kinerja model yang dilaporkan adalah rata-rata hasil pengujian dari setiap iterasi. Proses ini dikenal sebagai k-fold cross-validation.

Dengan menggunakan validasi silang, model dapat dilatih dan dievaluasi dengan lebih banyak data, sambil tetap mempertahankan subset yang cukup untuk validasi. Proses ini membantu mencegah overfitting dan memberikan estimasi yang lebih akurat tentang kinerja model pada data yang belum pernah dilihat sebelumnya. Ilustrasi dari proses k-fold cross-validation dapat dilihat pada Gambar 3.2.

Seminar Nasional Hasil Penelitian dan Pengabdian Masyarakat 2023
Institut Informatika dan Bisnis Darmajaya, 3 Agustus 2023
ISSN: 2598-0256, E-ISSN: 2598-0238




Gambar 2. Proses *Cross Validation*

### *3.3. Matriks Performa*

Dalam klasifikasi, terdapat beberapa metrik yang dapat digunakan untuk mengevaluasi kinerja suatu model. Metrik-metrik tersebut meliputi: (i) akurasi; (ii) Area Under Curve (AUC); (iii) Recall; (iv) Presisi; (v) F1 measure; dan (vi) Kappa. Pada penelitian ini, kinerja model yang telah dilatih dievaluasi menggunakan semua metrik tersebut. Berikut ini adalah penjelasan singkat mengenai metrik-metrik tersebut.

#### *3.3.1.* Confusion Matrix

*Confusion Matrix* merupakan sebuah alat yang merangkum kinerja model klasifikasi dalam memprediksi dengan benar kelas variabel target, yaitu kelas positif "miskin" dan kelas negatif "tidak miskin". Dalam setiap model klasifikasi, matriks kebingungan menunjukkan prediksi yang benar (positif dan negatif) dan prediksi yang salah atau salah klasifikasi (positif palsu dan negatif palsu). Semua metrik kinerja yang digunakan untuk mengevaluasi kinerja dalam penelitian ini, seperti yang dijelaskan di bawah, dapat diperoleh dari confusion matrix (Tiziana Rancati, 2019).

Tabel 2. *Confussion Matrix*

| Predicted → / Actual ↓ | Poor | Non-Poor |
|---|---|---|
| Poor | TP | FN |
| Non-Poor | FP | TN |

Dimana : TP = True Positif ; FP = False Positif; FN = False Negatif; dan TN = True Negatif

#### *3.3.2 Akurasi*

Akurasi adalah metrik yang paling sering digunakan dalam klasifikasi (Raschka, 2015). Secara sederhana, akurasi melaporkan porsi atau persentase prediksi yang benar, yang secara matematis dinyatakan sebagai berikut.

$$\text{Akurasi} = \frac{TP + TN}{TP + FP + TN + FN} \quad (7)$$

Di mana TP = true positive (yaitu rumah tangga miskin); TN = true negative (yaitu rumah tangga tidak miskin); FP = positif False dan FN = negatif False. Akurasi dapat diandalkan sebagai metrik kinerja tunggal ketika dataset memiliki keseimbangan yang baik.

#### *3.3.3 Recall*

Recall berfokus pada bagian dataset yang positif. Oleh karena itu, ini adalah tingkat di mana pengklasifikasi dapat memprediksi dengan benar positif yang sebenarnya. Ini adalah rasio dari true positive (TP) terhadap semua positif, yang merupakan TP dan FN. Representasi persamaan disediakan berikut ini.

ISSN: 2598-0256, E-ISSN: 2598-0238

$$\text{TPR} = \frac{TP}{TP+FN} \quad (8)$$

Pengklasifikasi yang baik adalah pengklasifikasi yang memiliki sedikit FN, dan oleh karena itu, memiliki tingkat positif yang tinggi (TPR). Recall sangat membantu untuk mengukur kinerja model ketika dataset tidak seimbang (Fitzpatrick, 2018). Dalam studi ini, recall yang rendah berarti terlalu banyak rumah tangga miskin yang tidak menerima bantuan. bantuan.

### *3.3.4 Presisi*

Presisi didefinisikan sebagai nilai prediksi positif (PPV) (Fitzpatrick, 2018); (Zheng, 2018). Hal ini melihat totalitas dari apa yang diidentifikasi oleh model sebagai positif, dan berapa banyak dari mereka yang benar-benar positif. yang benar-benar positif. Oleh karena itu, persamaan untuk presisi adalah sebagai berikut, di mana FP = False positif.

$$\text{Presisi} = \frac{TP}{\text{TP} + \text{FP}} \quad (9)$$

Semakin tinggi jumlah positif palsu, semakin rendah presisi model. Oleh karena itu, ketepatan adalah penyeimbang dari recall, dan akan sangat membantu untuk melihat kedua metrik tersebut secara bersamaan. Dalam penelitian ini, presisi yang rendah berarti bantuan ditargetkan pada terlalu banyak rumah tangga yang tidak membutuhkannya.

### *3.3.5 F1 measure*

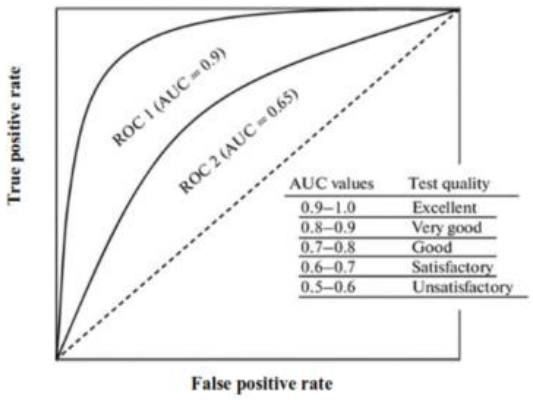
Nilai F1 menggabungkan presisi dan recall yang memberikan "rata-rata harmonis". Hal ini memberikan cara yang mudah untuk melihat ketepatan dan recall secara bersamaan (Fitzpatrick, 2018) merepresentasikan F1 sebagai berikut.

$$\text{F1} = \frac{2TP}{2\text{TP} + \text{FN} + \text{FP}} \quad (10)$$

Sebuah pengklasifikasi mendapatkan nilai F1 yang tinggi jika recall dan presisi tinggi.

### *3.3.6 Area Under The Curve (AUC)*

Kurva Karakteristik Operasi Penerima (ROC) adalah metrik kinerja yang dapat direpresentasikan dalam bentuk grafik, yang menampilkan hubungan antara kelas True Positive Rate (TPR) dan False Kelas prediksi False Positive Rate (FPR). Ini adalah salah satu alat yang paling sering digunakan untuk mengukur kinerja pengklasifikasi ML, terutama secara visual (Davis, 2006) dan (Fitzpatrick, 2018). ROC dibangun dengan memplot tingkat positif yang benar (atau Recall) terhadap tingkat positif palsu. Satu titik pada kurva ROC, TPR dan FPR terkait, adalah diperoleh dari confussion matrix pengklasifikasi. Nilai AUC yang tinggi mewakili TPR yang tinggi dan FPR yang rendah. Gambar 3.2 memberikan panduan untuk menginterpretasikan kurva ROC dan nilai AUC.




Gambar 3. Kurva AUC

ISSN: 2598-0256, E-ISSN: 2598-0238

## 4. HASIL DAN PEMBAHASAN

Penelitian ini masih di tahap desain model, dengan menggunakan dua metode klasifikasi, yaitu C4.5 dan K-Nearest Neighbors (KNN), penelitian ini berupaya untuk memprediksi tingkat kemiskinan dengan lebih akurat. Metode C4.5 akan menggunakan decision tree untuk mengklasifikasikan data, sehingga hasilnya dapat memberikan wawasan yang lebih jelas tentang faktor-faktor yang mempengaruhi kemiskinan. Sementara itu, KNN akan memanfaatkan jarak terdekat untuk mengelompokkan wilayah-wilayah dengan karakteristik serupa, yang dapat membantu dalam perencanaan program bantuan yang lebih tepat sasaran. Evaluasi kinerja model menggunakan berbagai metrik akan membantu menentukan algoritme terbaik yang dapat mendeteksi penduduk miskin dengan akurasi yang tinggi. Diharapkan hasil dari penelitian ini dapat memberikan wawasan berharga bagi pembuat kebijakan dan organisasi bantuan dalam mengembangkan program-program yang lebih efektif dan efisien dalam mengentaskan kemiskinan di Indonesia. Dengan kontribusi penerapan kedua metode tersebut, diharapkan upaya penanggulangan kemiskinan dapat menjadi lebih berdaya guna dan membawa dampak positif bagi masyarakat yang membutuhkan.

## 5. KESIMPULAN

Dalam penelitian ini, dilakukan perbandingan antara metode klasifikasi C4.5 dan K-Nearest Neighbors (KNN) dalam memprediksi tingkat kemiskinan di Indonesia. Metodologi penelitian meliputi tahapan pra-pemrosesan data, pelatihan dan pengujian model, serta evaluasi kinerja menggunakan berbagai metrik. Dalam tahap pra-pemrosesan data, dilakukan pembersihan data, seleksi fitur, eksplorasi data, dan balancing data untuk memastikan kualitas data yang digunakan dalam analisis. Selanjutnya, dilakukan pelatihan dan pengujian model menggunakan algoritma C4.5 dan KNN. Meskipun hasil evaluasi belum dilakukan dalam penelitian ini, penelitian sebelumnya telah menunjukkan tingkat akurasi yang signifikan dari kedua metode. Metode C4.5 memiliki keuntungan dalam interpretabilitas hasil, penghapusan fitur yang tidak relevan, dan efisiensi yang tinggi. Di sisi lain, metode KNN memiliki kelebihan dalam menangani data latih yang memiliki banyak noise dan efektivitas pada data latih yang besar.

Penelitian ini memberikan gambaran tentang pentingnya penggunaan metode klasifikasi dalam mengidentifikasi penduduk miskin secara akurat. Dengan mengoptimalkan penggunaan teknologi machine learning seperti C4.5 dan KNN, diharapkan program-program bantuan yang ditujukan untuk mengatasi kemiskinan dapat lebih tepat sasaran. Namun, penelitian ini masih berada pada tahap desain model, dan tindak lanjut yang diperlukan adalah melanjutkan penelitian hingga tahap evaluasi algoritme untuk memilih algoritme terbaik yang dapat mendeteksi penduduk miskin dengan akurasi tinggi. Kesimpulannya, penelitian ini memberikan kontribusi penting dalam upaya pengentasan kemiskinan di Indonesia dengan membandingkan metode klasifikasi C4.5 dan KNN. Hasil penelitian ini diharapkan dapat digunakan sebagai landasan untuk mengembangkan program bantuan yang lebih efektif dalam mengatasi masalah kemiskinan di negara ini.

## UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada semua pihak yang telah membantu penelitian ini, khususnya sumber data penelitian.

## DAFTAR PUSTAKA

BPS. (2022a). *Data dan Informasi Kemiskinan Kabupaten_Kota Tahun 2022*. https://www.bps.go.id/publication/2022/11/30/3b084878f782dfa44e0025e0/data-dan-informasi-kemiskinan-kabupaten-kota-tahun-2022.html

BPS. (2022b). *Penghitungan dan Analisis Kemiskinan Makro Indonesia Tahun 2022*.

Davis, J. , & G. M. (2006). The Relationship Between Precision-Recall and ROC Curves. *Procedings of the 23rd International Conference on Machine Learning*.

Fitzpatrick, C. A. , B. P. , & D. O. (2018). *Machine learning for poverty prediction:A comparative assessment of classification algorithms*.

ISSN: 2598-0256, E-ISSN: 2598-0238

F.J. Kaunang. (2018). Penerapan Algoritma J48 Decision Tree Untuk Analisis Tingkat Kemiskinan di Indonesia. *Cogito Smart Journal*, *4*, 348–357.

Han, J., Kamber, M., & Pei, J. (2011). *Data Mining. Concepts and Techniques, 3rd Edition (The Morgan Kaufmann Series in Data Management Systems)*.

Lantz, B. (2015). *Machine learning with R : discover how to build machine learning algorithms, prepare data, and dig deep into data prediction techniques with R*. Packt Publishing.

McBride, L. , & N. A. (2018). Retooling Poverty Targeting Using Out-of-Sample Validation and Machine Learning. *The World Bank Economic Review*, 531–550.

Noviana, D., Susanti, Y., & Susanto, I. (2019). *Analisis Rekomendasi Penerima Beasiswa Menggunakan Algoritma K-Nearest Neighbor (K-Nn) Dan Algoritma C4.5*.

Paper, D. (2018). *Exploring Data: Data Science Fundamentals for Python and MongoDB*. 167–209.

Raschka, S. (2015). *Python Machine Learning*. Packt Publishing.

Sihombing, P. R., & Arsani, A. M. (2021). Comparison Of Machine Learning Methods In Classifying Poverty In Indonesia In 2018. *Jurnal Teknik Informatika (Jutif)*, *2*(1), 51–56. https://doi.org/10.20884/1.jutif.2021.2.1.52

Sofianto, A. (2020). Implementasi Program Keluarga Harapan (PKH) di Provinsi Jawa Tengah. *Sosio Konsepsia*, *10*(1). https://doi.org/10.33007/ska.v10i1.2091

Thoplan, R. (2014). Random Forests for Poverty Classification. *International Journal of Sciences:Basic and Applied Research (IJSBAR)*, 252–259.

Tiziana Rancati, C. F. (2019). Modelling Radiotherapy Side Effects: Practical Applications for Planning Optimisation. *CRC Press*.

World Bank. (2001). *Attacking poverty : overview.* World Bank.

World Bank. (2018). *Poverty and Shared Prosperity 2018: Piecing Together the Poverty Puzzle*.

World Bank. (2020). *SDG 9.1.1. Rural Access Index Metadata*.

W. Yustanti. (2012). Algoritma K-Nearest Neighbour untuk Memprediksi Harga Jual Tanah. *Jurnal Matematika, Statistika, & Komputasi*, *9*, 57–68.

Zheng, A. , & C. A. (2018). Feature Engineering for Machine Learning. *O'Reilly Media Inc*.