BAB II
LANDASAN TEORI

2.1 Penelitian terkait

2.1.1 Kemiskinan dan *Machine Learning*

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 2. 1 Penelitian Terdahulu Tentang Prediksi Kemiskinan dengan Machine Learning

Peneliti	**Himpunan data**	**Teknik Pengambilan Sampel**	**Seleksi Fitur**	**Klasifikasi**	**Hasil**
(1)	(2)	(3)	(4)	(5)	(6)
Fitzpatrick dkk. (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. Klasifikasi adalah dilakukan dengan menggunakan full fitur dataset dan satu set	10 open-source ML classification algoritma klasifikasi adalah diterapkan untuk memprediksi kemiskinan dan hasil dibandingkan. Variabel prediksi target	MALAWI: Akurasi = 78 - 87% untuk rangkaian fitur lengkap; dan = 73 - 77% untuk set fitur sederhana INDONESIA: Akurasi = 88% - 91% untuk fitur lengkap mengatur; dan = 90 - 91% lebih sedikit set fitur

Peneliti	Himpunan data	Teknik Pengambilan Sampel	Seleksi Fitur	Klasifikasi	Hasil
			dari fitur terpilih Variance inflation factor (VIF) pengujian untuk menghilangkan fitur yang berlebihan menghasilkan set terpilih	dari semua metode adalah label biner "miskin" atau "tidak miskin", untuk mengklasifikasikan rumah tangga	
Thoplan (2014)	data survei, 296 294	Random Forest Bagging	Gini Score	Bootstrap pada pohon diterapkan pada a sampel acak dari observasi Prediksi out-of-bag (OOB). diperoleh dengan menggunakan suara terbanyak melintasi pepohonan	Keluar dari tas (OOB) Kesalahan Mean = 0,175
McBride dan Nichols (2018)	data survei, 1800- 11280	Regression forest Quantile regressionforest bootstrap aggregation	Random forest for feature selection	Metode ansambel (Bagging dan kemudian menerapkan random forest untuk klasifikasi) Pemilihan model berdasarkan validasi silang	Akurasi Total Rata-Rata Malawi = 80% Rata-rata Timor Timur = 75% Rata-rata Bolivia = 64%
Kshirsagar,	LCMS Zambia 2015 data survei	Elastic net logistic regression	Variabel bootstrap pilihan	Elastic net logistic regression	Probabilitas = 0,85

Peneliti	Himpunan data	Teknik Pengambilan Sampel	Seleksi Fitur	Klasifikasi	Hasil
Wieczorek ,Ramathan dan Wells (2017)					
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) LASSO r^2= 56,4% Random Forest r^2 = 55,6%
Sohnesen dan Stender (2017)	data survei, 1800 – 18000 Di 6 negara	Random Forest	Entropy loss function Gini loss function	Random Forest	National Mean square error (MSE) Gini = 1,71 Entropi = 1,94 UMK Perkotaan/Pedesaan Gini = 2,58 Entropi = 2,58
Gravemeyer , Gries dan Xue (2010)	data survei, 1056 rumah tangga dan 3256 orang	Regresi logit Regresi Tobit Regresi probit	Empirical Truncated Censored	Metode statistik regresi untuk mengukur kemiskinan Regresi diterapkan, dan variables are truncated; others are censored. Ini memungkinkan kita untuk memiliki koefisien	Probit r^2 = 74% Menggigit r^2 = 75% OLS r^2 = 53,6%

Studi oleh Fitzpatrick et al. [13] membandingkan hasil dari regresi linier yang dioptimalkan dengan 10 metode klasifikasi modern, termasuk decision tree, algoritma genetik, dan metode deep learning. Dalam penelitian ini, mereka mengevaluasi kompromi antara kompleksitas komputasi metode dan kinerja model. Algoritme klasifikasi yang digunakan semuanya bersifat open source dan diimplementasikan pada data survei di Malawi dan Indonesia.

Respons survei diubah menjadi fitur-fitur yang digunakan sebagai prediktor untuk mengidentifikasi apakah rumah tangga tersebut "miskin" atau "tidak miskin". Hasil penelitian menunjukkan bahwa tidak ada perbedaan signifikan dalam kemampuan prediksi antara regresi linier dan metode ML yang lebih canggih. Bahkan, dalam semua kasus, regresi linier memberikan performa yang lebih baik daripada metode yang lebih canggih, dan muncul sebagai salah satu dari tiga prediktor terbaik. Tingkat akurasi prediksi berkisar antara 73% - 87% untuk Malawi dan 88% - 91% untuk Indonesia sebelum dilakukan penyempurnaan model, serta 86% - 87% untuk Malawi dan 81% - 85% untuk Indonesia setelah penyempurnaan model. Pada dataset sederhana dengan 10 fitur, akurasi model berkisar antara 73% - 77%. Namun, perlu dicatat bahwa 10 fitur teratas diperoleh menggunakan pendekatan bertahap yang serupa dengan yang digunakan dalam teknik regresi untuk mengidentifikasi prediktor terbaik.

Thoplan [14] melakukan penelitian menggunakan metode hutan acak dengan data survei sensus untuk memprediksi kemiskinan. Pendekatan ini dipilih karena memiliki kebutuhan komputasi yang rendah, tingkat akurasi yang baik, dan mampu menghindari overfitting pada data pelatihan. Selain itu, hutan acak juga efisien

dalam memilih fitur-fitur penting yang berpengaruh terhadap variabel target. Hasil prediksi dengan menggunakan hutan acak menunjukkan rata-rata kesalahan "out of the bag" (OOB) dalam mengklasifikasikan rumah tangga miskin dan tidak miskin sebesar 0,175.

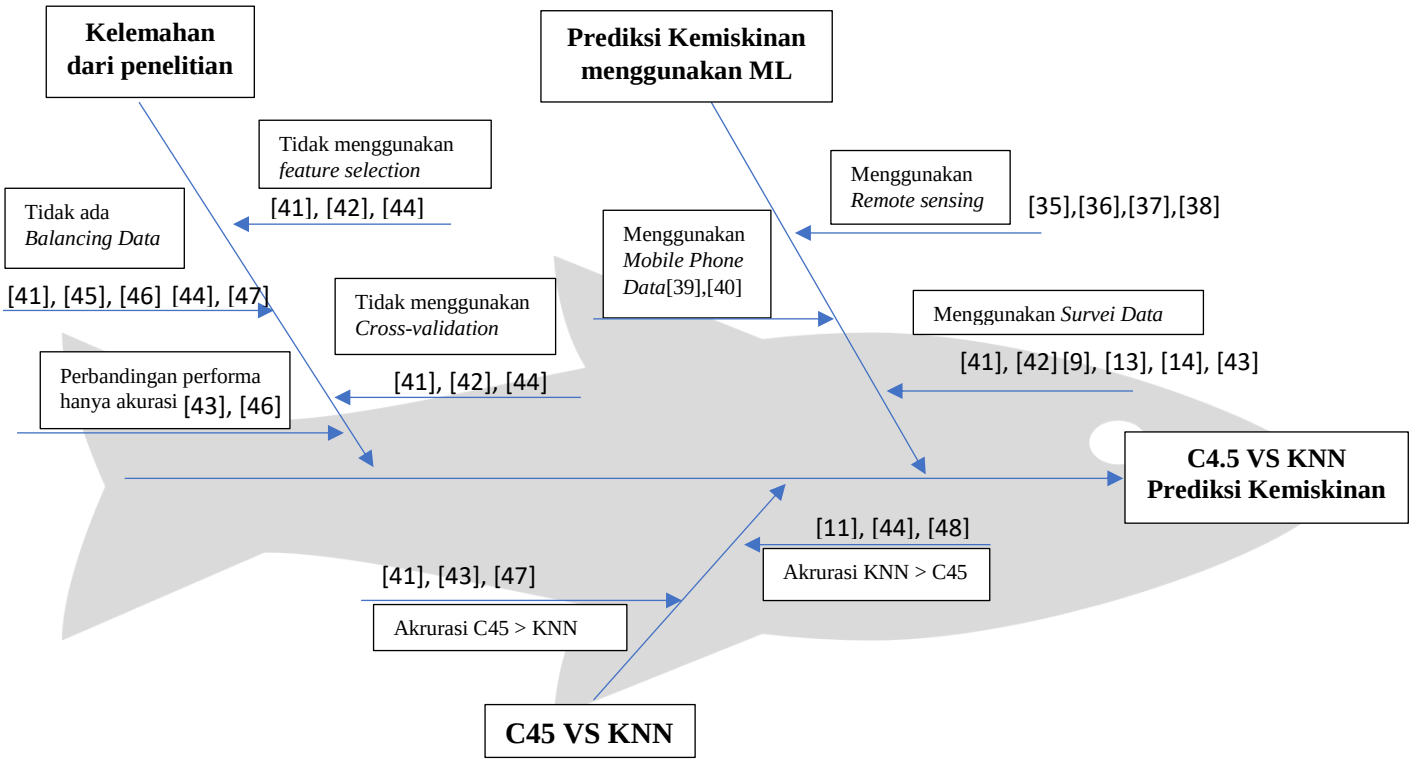
McBride dan Nichols [15] menggunakan data survei untuk memprediksi kemiskinan dengan menggunakan metode validasi silang dan pendekatan ensemble stokastik untuk meningkatkan kemampuan prediksi pengujian alat proxy means test (PMT) yang digunakan oleh Badan Pembangunan Internasional Amerika Serikat (USAID). Studi tersebut menyimpulkan bahwa pendekatan ensemble stokastik dapat mengurangi waktu yang diperlukan untuk pemilihan fitur dan meningkatkan proses perbandingan kinerja beberapa model pembelajaran mesin.

Secara keseluruhan, hasil prediksi dari pendekatan ensemble menunjukkan peningkatan yang konservatif dalam kinerja dibandingkan metode lainnya. Tingkat akurasi prediksi berkisar antara 55% untuk Bolivia dan 87% untuk Malawi. Penggunaan data survei membantu mengatasi tantangan dalam memperoleh data kepemilikan dan persyaratan pemrosesan komputer dengan resolusi tinggi yang biasa terjadi pada jenis data lainnya.

2.1.2 Perbandingan C4.5 dan KNN

Beberapa penelitian sebelumnya telah membandingkan algoritma C4.5 dan KNN dalam klasifikasi. Penelitian-penelitian ini memeriksa akurasi klasifikasi dari kedua algoritma tersebut. Dengan mempelajari penelitian-penelitian sebelumnya, kita dapat memperoleh pemahaman yang lebih baik tentang bagaimana kedua algoritma

ini bekerja dan karakteristiknya dalam klasifikasi. Untuk mempermudah pemahaman mengenai penelitian-penelitian tersebut, akan disajikan dalam bentuk diagram fishbone seperti yang terlihat pada gambar 2.1.



Gambar 2. 1 Diagram Fishbone

Dalam analisis fishbone diagram ini, terdapat beberapa cabang tulang ikan yang mewakili penelitian-penelitan terdahulu yang perlu diperhatikan dalam penelitian perbandingan antara algoritma C45 dan KNN dalam memprediksi kemiskinan. Berikut adalah penjelasan untuk setiap cabang tulang ikan:

1. Prediksi Kemiskinan menggunakan ML : Dalam beberapa tahun terakhir, penelitian tentang prediksi kemiskinan dengan menggunakan machine learning telah menjadi topik yang semakin diminati. Pendekatan ini memanfaatkan

kemajuan teknologi dalam analisis data untuk mengidentifikasi pola dan hubungan yang terkait dengan kemiskinan, dengan harapan dapat memberikan wawasan yang lebih baik kepada para pengambil keputusan dalam merancang kebijakan dan intervensi yang tepat. Dalam analisis prediksi kemiskinan menggunakan machine learning, sumber data yang digunakan menjadi faktor penting dalam mempengaruhi akurasi dan ketepatan prediksi. Diagram Ishikawa dapat digunakan untuk membagi sumber data menjadi tiga kategori utama: remote sensing [35],[36],[37],[38], mobile survey [39],[40], dan data survey [9], [13], [14], [43], [41], [42]

2. Perbandingan Algoritma C4.5 dan K-Nearest Neighbors (KNN) dalam Penelitian Sebelumnya : Dalam penelitian sebelumnya, telah dilakukan perbandingan antara algoritma C4.5 dan K-Nearest Neighbors (KNN) dalam berbagai konteks dan dataset. Tujuan dari perbandingan ini adalah untuk memahami kinerja relatif kedua algoritma tersebut dalam memecahkan masalah klasifikasi. Beberapa penelitian menunjukkan bahwa algoritma C4.5 memiliki kinerja yang lebih baik daripada KNN dalam beberapa kasus [41], [43], [47]. C4.5 adalah algoritma pohon keputusan yang populer dan efektif dalam melakukan klasifikasi. Keunggulan C4.5 terletak pada kemampuannya untuk menghasilkan aturan keputusan yang dapat diinterpretasikan dengan mudah oleh manusia. Dalam beberapa situasi, C4.5 mampu memberikan akurasi klasifikasi yang tinggi dengan tingkat kesalahan yang rendah. Namun, penelitian lain menunjukkan bahwa ada situasi di mana KNN dapat memiliki kinerja yang lebih baik dibandingkan dengan C4.5 [11], [44], [48]. KNN

adalah algoritma pembelajaran berbasis instance yang mengklasifikasikan objek berdasarkan kesamaan dengan tetangga terdekatnya. Algoritma ini memiliki fleksibilitas yang tinggi dan tidak mengasumsikan distribusi data tertentu. KNN sering kali efektif dalam kasus-kasus di mana batasan keputusan yang kompleks sulit didefinisikan secara formal.

3. Kelemahan Preprocessing Data dalam Penelitian Terdahulu : Dalam penelitian-penelitian terdahulu, telah diidentifikasi beberapa kelemahan yang sering muncul dalam proses preprocessing data. Preprocessing data merupakan tahap yang penting dalam analisis data, yang melibatkan transformasi dan pengaturan data mentah agar siap untuk analisis lebih lanjut. Meskipun kelemahan ini dapat bervariasi tergantung pada konteks dan penelitian yang dilakukan, beberapa kelemahan yang umum ditemui adalah kurangnya penggunaan feature selection [41], [42], [44], ketidakseimbangan data [41], [45], [46], [44], [47] , perbandingan performa yang terbatas pada akurasi [43], [46], dan kurangnya penggunaan cross-validation [41], [42], [44]

2.2 Teori Dasar

2.2.1 Kemiskinan

Kemiskinan dianggap sebagai ketidakmampuan dalam hal keuangan untuk memenuhi kebutuhan penting seperti makanan dan barang kebutuhan lainnya, yang diukur berdasarkan pengeluaran [16]. Penduduk yang dianggap miskin adalah orang-orang yang memiliki rata-rata pengeluaran per kapita (per bulan) di bawah garis kemiskinan. Sedangkan menurut World Bank, kemiskinan dapat diartikan sebagai ketidakmampuan seseorang untuk memenuhi kebutuhan hidupnya dan

kesulitan dalam memanfaatkan sumber daya yang tersedia untuk mencukupi kebutuhan tersebut [17].

Dalam konsepnya, kemiskinan dapat dibagi menjadi dua kategori, yaitu kemiskinan relatif dan kemiskinan absolut, yang dibedakan berdasarkan standar penilaiannya. standar penilaian kemiskinan absolut merujuk pada kebutuhan minimum yang dibutuhkan untuk memenuhi kebutuhan dasar seperti makanan dan non-makanan, dan disebut sebagai garis kemiskinan. Sementara itu, Standar penilaian kemiskinan relatif ditetapkan oleh masyarakat setempat dan bersifat lokal, di mana mereka yang berada di bawah standar tersebut dianggap miskin secara relatif. [18]. Pernyataan tersebut menjelaskan bahwa kemiskinan tidak hanya dapat diukur dari pendapatan yang rendah, tetapi juga melibatkan faktor-faktor lain yang mempengaruhi kemampuan seseorang untuk mencapai standar kehidupan yang memadai. Aspek-aspek seperti sandang (pakaian), pangan (makanan), dan papan (tempat tinggal) sangat penting dalam mempengaruhi kapabilitas atau kemampuan seseorang untuk memenuhi kebutuhan dasar mereka.

Dalam konteks ini, jika seseorang tidak memiliki akses yang memadai terhadap sandang, pangan, dan papan, hal ini dapat menyebabkan penurunan kapabilitas mereka dalam mencukupi kebutuhan kesehatan dan pendidikan. Misalnya, ketidakmampuan untuk membeli makanan yang bergizi atau tidak memiliki tempat tinggal yang layak dapat berdampak negatif pada kesehatan seseorang. Begitu pula, jika seseorang tidak memiliki pakaian yang cukup atau akses yang memadai terhadap pendidikan, hal itu dapat menghambat kemampuan mereka untuk

meningkatkan kualitas hidup dan memperoleh pengetahuan serta keterampilan yang diperlukan.

Dengan demikian, pandangan ini mengakui bahwa kemiskinan adalah masalah yang kompleks dan melibatkan lebih dari sekadar pendapatan. Faktor-faktor seperti sandang, pangan, serta papan juga harus dipertimbangkan dalam upaya untuk mengatasi kemiskinan secara menyeluruh.

Menurut Alkire dan Seth [19] ada tiga dimensi dalam pengukuran kemiskinan multidimensi yaitu, dimensi Kesehatan, Dimensi Pendidikan dan Dimensi Standar Hidup.

1. Dimensi Kesehatan

Menurut Undang-Undang Republik Indonesia Nomor 36 Tahun 2009 tentang Kesehatan, kesehatan merupakan hak asasi manusia dan merupakan salah satu unsur kesejahteraan yang harus diwujudkan sesuai dengan cita-cita bangsa Indonesia sebagaimana dimaksud dalam Pancasila dan Undang-Undang Dasar Negara Republik Indonesia Tahun 1945. Kesehatan yang dimaksud adalah keadaan sehat, baik secara fisik, mental, spiritual, maupun sosial yang memungkinkan setiap orang hidup produktif secara sosial dan ekonomi.

Kesehatan adalah aset penting dalam kehidupan yang dapat menunjang kelancaran berbagai aktivitas manusia. United Nations Fund for Population Activities atau UNFPA [20] menyatakan bahwa kesehatan yang buruk merupakan penyebab sekaligus dampak dari kemiskinan. Efek dari kesehatan yang buruk akan lebih besar dirasakan oleh orang miskin terhadap produktivitas dan pendapatannya, karena

cenderung dituntut untuk melakukan pekerjaan fisik yang berat. Intervensi untuk memperbaiki kesehatan dari pemerintah juga merupakan suatu alat kebijakan penting untuk mengurangi kemiskinan [21]. Salah satu faktor yang mendasari kebijakan ini adalah perbaikan kesehatan akan meningkatkan produktivitas penduduk miskin.

2. Dimensi Pendidikan

Menurut Undang-Undang No.23 Tahun 2003 tentang Sistem Pendidikan Nasional, pendidikan didefinisikan sebagai usaha sadar dan terencana untuk mewujudkan suasana belajar dan proses pembelajaran agar peserta didik secara aktif mengembangkan potensi dirinya untuk memiliki kekuatan spiritual keagamaan, pengendalian diri kepribadian, kecerdasan, akhlak mulia, serta keterampilan yang diperlukan dirinya, masyarakat, bangsa, dan negara.

Pendidikan merupakan investasi sumber daya manusia karena pendidikan dapat mengubah pola pikir dan meningkatkan martabak seseorang. Pendidikan memainkan peranan kunci dalam meningkatkan pertumbuhan ekonomi dan mengurangi kemiskinan. Pendidikan diharapkan mampu meningkatkan potensi, pekerjaan, dan mobilitas tenaga kerja. Todaro [22] menyatakan bahwa terdapat korelasi positif antara pendidikan seseorang dengan penghasilan yang akan diperolehnya. Sementara itu menurut World Bank [23], menaikkan tingkat pendidikan dapat meningkatkan standar hidup dan mengurangi kemiskinan.

3. Dimensi Perumahan

Menurut WHO, perumahan yang layak harus memberikan perlindungan terhadap paparan agen dan vektor penyakit menular, melalui: pasokan air yang aman, sanitasi pembuangan tinja, pembuangan limbah padat rumah tangga, drainase air permukaan, kebersihan pribadi dan rumah tangga, penyiapan makanan yang aman, dan perlindungan struktural terhadap penularan penyakit.

Hal tersebut sejalan dengan pengertian rumah dalam Pasal 1 ayat 7 Undang-Undang No.1 Tahun 2011, yaitu bangunan gedung yang berfungsi sebagai tempat tinggal yang layak huni, sarana pembinaan keluarga, cerminan harkat dan martabat penghuninya, serta aset bagi pemiliknya. Rumah yang ditempati diharapkan adalah rumah yang layak huni dalam lingkungan yang sehat, aman, serasi, dan teratur. Dapat disimpulkan bahwa kemiskinan dapat dilihat dari kondisi perumahan suatu rumah tangga. Semakin baik kondisi perumahan suatu rumah tangga maka akan semakin baik standar hidupnya yang akan menjauhkan dari kemiskinan.

4. Dimensi Kualitas Hidup

Aset adalah sumber daya ekonomi yang dikuasai dan/atau dimiliki oleh masyarakat dan mempunyai manfaat ekonomi sosial serta dapat diukur dalam satuan uang. Aset termasuk sumber daya yang non keuangan baik yang digunakan dalam menghasilkan barang atau jasa atau digunakan untuk tujuan lainnya, perlengkapan yang dibeli dan disimpan untuk digunakan, maupun yang mempunyai masa manfaat atau dimanfaatkan sebagai investasi jangka panjang. Rendahnya tingkat kepemilikan aset merupakan salah satu faktor yang menyebabkan Kemiskinan [24].

5. Dimensi Perlindungan

Faktor ketenagakerjaan dan perlindungan sosial juga memiliki kontribusi yang cukup besar terhadap angka kemiskinan. Pekerjaan merupakan sumber utama penghasilan bagi keluarga. Meskipun tidak dapat dikatakan sebagai sebuah dimensi yang baru dalam kesejahteraan, ketenagakerjaan sering kali dilupakan dalam studi pembangunan dan pemberantasan kemiskinan. Mendapatkan pekerjaan yang layak secara umum dapat diasosiasikan dengan keluar dari kemiskinan. Lugo [25] menyebutkan beberapa indikator yang menggambarkan kelayakan pekerjaan antara lain proteksi pekerjaan informal dan kekurangan atau kelebihan jam kerja.

Perlindungan sosial juga berimplikasi positif pada pembangunan ekonomi dan sosial suatu negara. Pemberian perlindungan sosial memberikan kontribusi yang positif dalam pengentasan kemiskinan, khususnya dalam bidang pendidikan dan kesehatan.

Dari kelima dimensi tersebut dibentuklah atribut-atribut sebagai berikut :

Tabel 2. 2 Atribut-Atribut Penelitian Ini

No.	Dimensi	Fitur
1	Dimensi Kesehatan	Keluhan Kesehatan Disabilitas Kepemilkan Asuransi Kesehatan
2	Dimensi Pendidikan	Ijazah Tertinggi Kepala Rumah Tangga Ijazah Tertinggi Anggota Rumah Tangga Mendapatkan Bantuan PIP

No.	Dimensi	Fitur
3	Dimensi Perumahan	Status Kepemilikan Rumah Kondisi Rumah (Jenis Atap, Jenis Lantai, Luas Lantai) Sumber Air Minum Kepemilikan Sanitasi Bahan Bakar yang digunakan untuk Memasak
4	Dimensi Kualitas Hidup	Kepemilikan Aset (Properti, Motor, Mobil, Lemari Es) Akses terhadap Telekomunikasi dan Informasi
5	Dimensi Perlindungan	Status Pekerjaan Kepala Rumah Tangga Status Pekerjaan Anggota Rumah Tangga Jumlah Jam Kerja Menerima Bantuan Sosial (Raskin, KPS/KKS, dan PKH) Pengeluaran Rumah Tangga
6	Demografi	Usia Status Anggota Rumah Tangga Status Perkawinan Jumlah Anggota Rumah tangga
7	Klasifikasi Daerah	Perdesaan atau Perkotaan

No.	Dimensi	Fitur
8	Kelas	Miskin atau Tidak Miskin

2.2.2 Klasifikasi

Klasifikasi adalah suatu proses pembentukan model atau fungsi yang dapat mengkategorikan data atau objek pengamatan ke dalam kelas-kelas tertentu.[7] Dalam melakukan klasifikasi, terdapat dua tahapan, yaitu tahap pembelajaran (training) dan tahap klasifikasi (testing). Pada tahap pertama, model klasifikasi dibangun dengan menggunakan data latih (training data) yang telah memiliki label atau kelas yang telah diketahui. Pada tahap kedua, model yang telah dibangun digunakan untuk memprediksi kelas label pada data yang belum memiliki kelas label (testing data), dan akurasi dari model dihitung dengan membandingkan hasil prediksi dengan kelas yang sebenarnya. Jika akurasi model dianggap bisa diterima, maka model tersebut dapat digunakan untuk memprediksi kelas pada data yang belum memiliki kelas label.[26]

2.2.3 Decision Tree

Decision tree adalah salah satu metode klasifikasi yang digunakan untuk memprediksi nilai kelas berdasarkan sekumpulan fitur atau atribut pada data.[27] Secara sederhana, decision tree dapat dianggap sebagai sebuah pohon yang terdiri dari node dan cabang-cabangnya, dimana setiap node pada pohon mewakili suatu keputusan atau pengujian yang dilakukan terhadap nilai atribut tertentu, sedangkan

cabang-cabangnya mewakili kemungkinan hasil atau nilai yang diperoleh dari pengujian tersebut.

Pada decision tree, proses pembentukan pohon dimulai dari root node dan dilanjutkan dengan memilih atribut yang paling signifikan untuk dipilih sebagai node selanjutnya. Atribut yang dipilih akan membagi data menjadi beberapa bagian yang lebih kecil, dimana setiap bagian tersebut merupakan cabang dari node yang dipilih. Proses ini dilakukan secara rekursif hingga tidak ada lagi atribut yang dapat dipilih atau seluruh data telah terklasifikasi dengan benar.

Pada saat pengujian dilakukan, suatu instance data akan diberikan ke root node, dan kemudian diperiksa pada atribut yang terkait dengan node tersebut. Kemudian instance data tersebut akan dikirim ke cabang yang sesuai dengan nilai dari atribut tersebut, dan proses pengujian akan berlanjut pada node cabang tersebut. Proses ini akan terus berlanjut hingga mencapai node daun, dimana nilai kelas atau kategori dari instance data akan ditentukan berdasarkan mayoritas nilai kelas pada data latih yang ada pada node tersebut.

Keuntungan dari penggunaan decision tree adalah kemampuannya untuk memberikan interpretasi yang mudah dan intuitif, serta kemudahan dalam mengambil keputusan. Decision tree juga cocok digunakan untuk data dengan atribut kategorikal, dan dapat digunakan untuk memprediksi kelas yang bernilai diskrit maupun kontinu.[28] Namun, kelemahan dari decision tree adalah kemungkinan terjadinya overfitting atau underfitting pada data, serta sensitivitasnya terhadap data yang tidak seimbang (imbalanced data).

Secara matematika, D, yang merupakan partisi data, harus dikonfigurasikan sebagai himpunan data pelatihan yang terdiri dari tupel-tupel dengan label kelas. Asumsikan bahwa atribut label kelas memiliki m nilai yang berbeda untuk menentukan m kelas yang berbeda, yaitu $Ci(i = 1 \ldots . m)$. Selanjutnya, tetapkan CiD sebagai himpunan tupel kelas Ci dalam D. Modulus dari D merupakan jumlah tupel dalam D, sementara modulus dari CiD adalah jumlah tupel dalam CiD. Simpul N digunakan untuk memisahkan tupel-tupel dalam D. Sebagai atribut pemisah di simpul N, dipilih atribut dengan nilai tertinggi yang memberikan informasi terbesar.

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.

2.2.4 Algoritma C4.5

C4.5 adalah algoritma decision tree yang dikembangkan oleh Ross Quinlan pada tahun 1993. C4.5 adalah pengembangan dari algoritma sebelumnya, yaitu ID3 (Iterative Dichotomiser 3), dan kemudian menjadi dasar dari algoritma decision tree

yang lebih modern seperti CART (Classification and Regression Tree) dan CHAID (Chi-squared Automatic Interaction Detection).[8]

C4.5 menggunakan pendekatan top-down divide and conquer, dimana pada awalnya seluruh data digunakan untuk membuat root node, kemudian data dibagi-bagi ke dalam subset yang lebih kecil berdasarkan aturan tertentu, dan proses ini diulang hingga terbentuk sebuah tree yang dapat memprediksi kelas target dari suatu data input.

C4.5 memiliki beberapa kelebihan, antara lain kemampuan untuk menangani data yang kompleks, robust terhadap nilai missing data, dapat bekerja dengan variabel target berupa nominal maupun numerik, serta dapat memilih atribut yang paling penting dalam memisahkan data.

C4.5 menggunakan metode gain ratio untuk memilih atribut yang akan digunakan pada setiap node. Gain ratio mengukur seberapa banyak informasi yang diberikan oleh sebuah atribut dalam memisahkan data, dan memperhitungkan jumlah subset yang dihasilkan dari atribut tersebut. Hal ini memungkinkan C4.5 untuk memilih atribut yang lebih banyak memberikan informasi dan menghindari atribut yang cenderung menghasilkan banyak subset dengan jumlah data yang sedikit.

Setelah tree terbentuk, C4.5 melakukan pruning atau pemotongan cabang yang tidak penting untuk mengurangi overfitting pada data latih. C4.5 juga memperhitungkan nilai probabilitas pada setiap leaf node, sehingga dapat menghasilkan prediksi yang lebih akurat. C4.5 memiliki banyak aplikasi, seperti

pada bidang klasifikasi data, pengenalan pola, pengambilan keputusan, dan lain sebagainya.

Algoritma C4.5 dan C5.0 telah ditingkatkan dari ID3, di mana rasio perolehan (gain ratio) digunakan untuk menyelesaikan karakteristik yang cenderung mempengaruhi hasil uji dengan beberapa pengukuran hasil, sehingga dapat mengatasi fitur-fitur kontinu. Rasio perolehan didefinisikan sebagai berikut:

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

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

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.

2.2.5 Algoritma KNN

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

Konsep dasar dari algoritma KNN adalah mencari tetangga terdekat dari suatu data berdasarkan jarak Euclidean atau jarak Minkowski dengan k. Kemudian, data yang akan diklasifikasikan akan diberikan label yang sama dengan mayoritas label tetangga terdekatnya.

Pada algoritma KNN, k merupakan 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.

Kelebihan dari algoritma KNN adalah kemudahan dalam implementasi dan hasil klasifikasinya yang cukup akurat. Namun, algoritma ini memiliki kelemahan dalam mengklasifikasikan data yang memiliki fitur atau atribut yang kompleks dan banyak, serta memiliki nilai-nilai yang tidak terstandarisasi.

Berikut adalah beberapa rumus yang digunakan dalam algoritma K-NN:

1. Jarak Euclidean: 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. K-NN Klasifikasi:

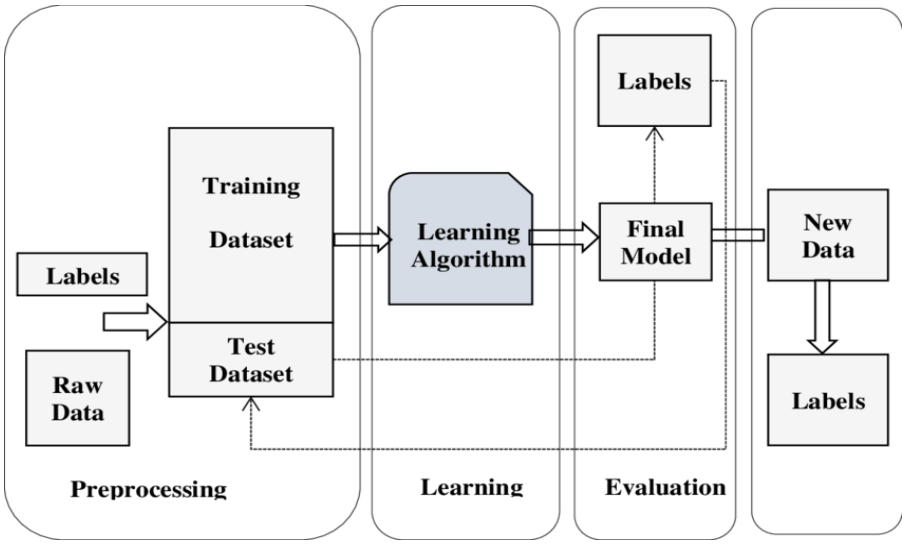
Dalam K-NN untuk klasifikasi, langkah-langkah umum adalah:

- Hitung jarak antara data yang akan diklasifikasikan dengan setiap data pelatihan.
- Pilih k tetangga terdekat berdasarkan jarak.
- Tentukan kelas mayoritas dari tetangga terdekat sebagai label atau kelas prediksi untuk data yang akan diklasifikasikan.

Pada dasarnya, algoritma K-NN menghitung jarak antara data yang akan diprediksi dengan data pelatihan, dan menggunakan k tetangga terdekat untuk melakukan prediksi atau regresi. Jarak Euclidean adalah salah satu metode yang umum digunakan untuk mengukur kedekatan antara data.

2.2.6 Roadmap Membangun Sistem *Mechine Learning*

Raschka [29] menjelaskan langkah-langkah umum dalam membangun sistem pembelajaran mesin. Pertama, ada pra-pemrosesan data yang bertujuan untuk mempersiapkan data agar dapat digunakan oleh berbagai algoritme ***Mechine Learning*** (ML) guna mendapatkan hasil yang lebih baik. Ada banyak algoritma ML yang dikembangkan untuk menyelesaikan berbagai masalah. Langkah kedua dalam alur kerja ML adalah pembelajaran, di mana algoritma ML dilatih menggunakan data input dari tugas masalah yang ada untuk mengembangkan model prediktif. Setelah itu, dilakukan evaluasi model untuk mengukur kinerja model prediktif berdasarkan metrik kinerja yang telah ditentukan. Jika kinerjanya memuaskan, model tersebut bisa digunakan untuk melakukan prediksi pada data baru atau data yang belum pernah ada sebelumnya. Diagram alur kerja Raschka [29] dapat ditemukan pada Gambar 2.2 di bawah ini.



Gambar 2. 2 Diagram Alur Kerja Machine Learning