Penerapan Algoritma J48 Decision Tree Untuk Analisis Tingkat Kemiskinan di Indonesia

Application of J48 Decision Tree Algorithm For Analyzing Poverty Level in Indonesia

Fergie Joanda Kaunang
Sistem Informasi, Fakultas Ilmu Komputer, Universitas Klabat
e-mail: fergie@unklab.ac.id

Abstrak

Kemiskinan telah menjadi masalah sosial dan tantangan bagi masyarakat di seluruh dunia yang terus dicari penyelesaiannya. Berdasarkan identifikasi dari Badan Program Pembangunan PBB (UNDP) yang bekerjasama dengan Oxford Poverty and Human Development Initiative (OPHI), 1.3 miliar penduduk dunia teridentifikasi sebagai penduduk miskin pada bulan September tahun 2018. Di tingkat nasional, Indonesia, tingkat kemiskinan tertinggi terjadi pada tahun 1999 dengan persentase sebesar 23.43%. Berdasarkan data dari Badan Pusat Statistik Indonesia (BPS), penduduk miskin di Indonesia mencapai 25.95 juta orang dengan persentase 9.82% pada tahun Maret 2018. Oleh karena itu penelitian ini bertujuan untuk menganalisis tingkat kemiskinan menggunakan dimensi dasar dari indeks pembangunan manusia (IPM) menggunakan metode data mining dan machine learning yakni algoritma J48 Decision Tree. Akurasi dari model prediksi yang telah dibuat menunjukan hasil yang baik yakni sebesar 88.6% dimana dengan kata lain model prediksi yang dikembangkan dapat digunakan untuk membantu para pembuat kebijakan maupun para pemangku kepentingan untuk mengambil keputusan.

Kata kunci—Angka Kemiskinan, Indeks Pembangunan Manusia, Algoritma J48 Decision Tree, Data Mining, Machine Learning

Abstract

Poverty has been a social issues and a challenge to the societies all over the world that need to be solved. Based on the finding of United Nation of Development Programme (UNDP) and Oxford Poverty and Human Development Initiative (OPHI) on September 2018, 1.3 billion of world population identified as poor people. In Indonesia itself, the poverty level hits the highest point on 1999 with 23.43%. Based on the Central Bureau of Statistics of Indonesia, on March 2018 the poverty level reached 9.82% with the amount of 25.95 million of poor people. Therefore this study aims to analyze the poverty level based on the dimensions of Human Development Index using J48 Decision Tree Algorithm. The accuracy of the developed prediction model is 88.6%. Meaning that this prediction model can help the government as well as the decision maker to decide and make the right policy to reduce the poverty level in Indonesia.

Keywords—Poverty Level, Human Development Index, J48 Decision Tree Algorithm, Data Mining, Machine Learning

1. PENDAHULUAN

Kemiskinan telah menjadi masalah sosial dan tantangan bagi masyarakat di seluruh dunia yang terus dicari penyelesaiannya. Kemiskinan sendiri dapat didefinisikan ke dalam berbagai macam konsep. Bank Dunia mendefinisikan kemiskinan sebagai ketidakmampuan untuk mencapai kehidupan yang layak dengan pendapatan $1.9 perhari. Sementara itu, Badan Pusat Statistik Indonesia (BPS) mendefinisikan kemiskinan berdasarkan kemampuan dalam memenuhi kebutuhan dasar dari sisi ekonomi [1]. Dengan kata lain, penduduk yang tergolong miskin adalah penduduk yang rata-rata pengeluaran perkapita perbulan berada dibawah garis kemiskinan.

Berdasarkan identifikasi dari Badan Program Pembangunan PBB (UNDP) yang bekerjasama dengan Oxford Poverty and Human Development Initiative (OPHI), 1.3 miliar penduduk dunia teridentifikasi sebagai penduduk miskin pada bulan September tahun 2018 [2]. Dari jumlah tersebut 83% penduduk miskin adalah mereka yang tinggal di Sub-Saharan Afrika dan Asia Selatan. Data lainnya menyebutkan bahwa masalah kemiskinan paling banyak menyerang negara-negara berkembang [3]. Di tingkat nasional, Indonesia, tingkat kemiskinan tertinggi terjadi pada tahun 1999 dengan persentase sebesar 23.43%. Berdasarkan data dari Badan Pusat Statistik Indonesia (BPS), penduduk miskin di Indonesia mencapai 25.95 juta orang dengan persentase 9.82% [4]. Berdasarkan angka-angka tersebut dapat dilihat bahwa terjadi penurunan angka kemiskinan. Namun, kemiskinan masih menjadi masalah yang harus dituntaskan. *Data mining* merupakan proses pencarian pola yang sebelumnya belum diketahui. Informasi yang diperoleh kemudian digunakan untuk membangun sebuah model prediksi. Hasil prediksi yang diperoleh berupa pengetahuan baru yang dapat digunakan dalam hal lainnya seperti pengambilan keputusan dan sebagainya.

Data mining dan *machine learning* sudah digunakan secara luas di berbagai bidang kehidupan seperti bidang ekonomi pemasaran, telekomunikasi, kesehatan dan pengobatan, pendidikan dan bidang lainnya. Weng et al menggunakan teknik data mining dan machine learning untuk membuat suatu model prediksi dalam mengidentifikasi karbonilasi protein (*protein carbonylation*) [5]. Penerapan lainnya diberikan oleh Sano and Nindito yang mengelompokkan daerah miskin di Indonesia menggunakan algoritma *K-Means Clustering* [6]. Hasil penelitian tersebut menunjukkan kelompok provinsi yang harus dijadikan prioritas pengentasan kemiskinan oleh para pembuat kebijakan. Studi lainnya yakni studi oleh Jean et al yang menggabungkan citra satelit dan teknik *machine learning* untuk memprediksi kemiskinan [7]. Studi ini memberikan hasil berupa sebuah model yang dapat digunakan dalam melacak dan menentukan daerah-daerah yang seharusnya dijadikan target untuk mengurangi kemiskinan. Oleh karena itu penelitian ini bertujuan untuk menganalisis tingkat kemiskinan menggunakan dimensi dasar dari indeks pembangunan manusia (IPM) menggunakan metode *data mining* dan *machine learning* yakni algoritma J48 Decision Tree.

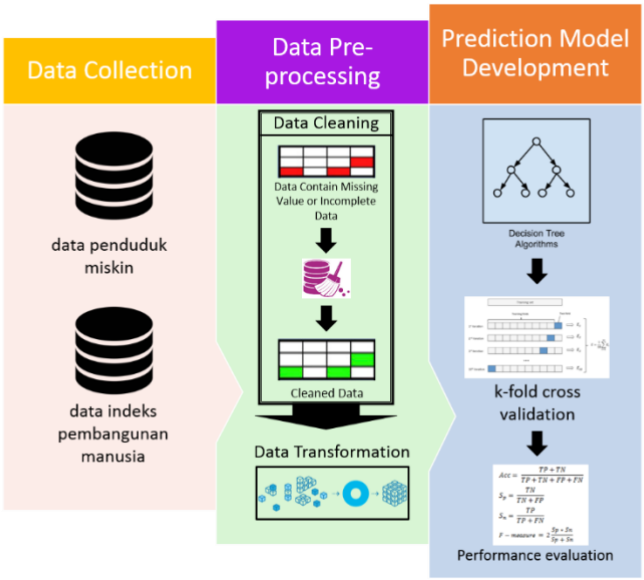
2. METODE PENELITIAN

Metode yang digunakan dalam penelitian ini terbagi ke dalam 3 tahapan yakni pengumpulan data (*data collection*), pra-proses data (*data pre-processing*), dan pengembangan model prediksi (*prediction model development*) seperti yang terlihat pada Gambar 1.

2.1 Data Collection

Sumber data yang digunakan dalam penelitian ini yakni situs resmi Badan Pusat Statistik (BPS) Indonesia (www.bps.go.id). Data yang digunakan dalam penelitian ini adalah data kemiskinan dan ketimpangan serta indeks pembangunan manusia (IPM). Data tersebut adalah data jumlah penduduk miskin, data rata-rata lama sekolah, data angka harapan hidup, serta data pengeluaran perkapita berdasarkan provinsi dari tahun 2010 sampai dengan tahun 2017. Format data yang digunakan adalah dalam bentuk *comma separated values* (.csv). Jumlah penduduk

miskin tersedia baik pada daerah perkotaan juga pada daerah pedesaan. Dalam penelitian jumlah penduduk miskin pada daerah perkotaan dan pedesaan dijumlahkan ke dalam satu nilai keseluruhan untuk jumlah penduduk miskin pada satu provinsi.



Gambar 1 Alur pembuatan model prediksi

Indeks pembangunan manusia sendiri menjelaskan mengenai bagaimana masyarakat memakai dan menikmati hasil pembangunan untuk mendapatkan pendapatan, kesehatan, pendidikan, dan sebagainya [8]. IPM ini digunakan sebagai indikator pengukur tingkat keberhasilan pembangunan kualitas hidup masyarakat juga sebagai tolak ukur kinerja Pemerintah dalam hal ini Pemerintah Indonesia. Pembentukan IPM dilakukan dengan memperhatikan 3 (tiga) dimensi dasar yakni (1) umur panjang dan hidup sehat, (2) pengetahuan, dan (3) standar hidup layak. Adapun metodologi dalam penghitungan IPM digunakan rumus-rumus sebagai berikut:

$$I_{kesehatan} = \frac{AHH - AHH_{min}}{AHH_{maks} - AHH_{min}} \quad \text{.....(3)}$$

$$I_{HLS} = \frac{HLS - HLS_{min}}{HLS_{maks} - HLS_{min}} \quad \text{.....(4)}$$

$$I_{RLS} = \frac{RLS - RLS_{min}}{RLS_{maks} - RLS_{min}} \quad \text{.....(5)}$$

$$I_{pendidikan} = \frac{I_{HLS} + I_{RLS}}{2} \quad \text{.....(6)}$$

$$I_{pengeluaran} = \frac{\ln(pengeluaran) - \ln(pengeluaran_{min})}{\ln(pengeluaran_{maks}) - \ln(pengeluaran_{min})} \quad \text{.....(7)}$$

Rumus 3 sampai dengan rumus 7 adalah rumus penghitungan nilai dari dimensi kesehatan, dimensi pendidikan dan dimensi pengeluaran dimana $I_{kesehatan}$ adalah nilai dimensi

kesehatan, $I_{pendidikan}$ adalah nilai dimensi pendidikan, dan $I_{pengeluaran}$ adalah nilai dimensi pengeluaran. AHH adalah angka harapan hidup, HLS adalah harapan lama sekolah, dan RLS adalah rata-rata lama sekolah. Sementara angka jumlah penduduk miskin yang digunakan dalam penelitian ini adalah data yang bersumber dari Survei Sosial Ekonomi Nasional (Susenas) Modul Konsumsi dan Pengeluaran periode Maret dan September seperti yang disediakan pada situs dari BPS.

2.2 Data Preprocessing

Setelah data dikumpulkan, langkah selanjutnya adalah melakukan pra-proses data (*data preprocessing*). Teknik praproses data yang digunakan dalam penelitian ini dibagi ke dalam empat bagian yakni pembersihan data (*data cleaning*). Pada tahap ini data yang tidak lengkap, data yang tidak konsisten maupun data yang tidak memiliki nilai (*missing values*) akan dibersihkan. Selanjutnya yakni tahap penyatuan data (*data integration*) dimana data dari tabel-tabel ataupun basis data yang berkaitan akan digabungkan menjadi satu. Tahap yang ketiga yaitu pemilihan data (*data selection*). Pada tahap ini dilakukan pemilihan data yang berkaitan dengan penelitian. Tahap yang keempat adalah tahap transformasi data (*data transformation*) dimana data yang sudah bersih dan diseleksi akan diubah. Data jumlah penduduk miskin diubah menjadi data nominal menggunakan metode *Equal Width Binning*. Mekanisme metode ini yakni dengan membagi data ke dalam *k* interval dimana masing-masing interval yang sudah dibagi memiliki ukuran yang sama. Dalam menentukan nilai batas *w* untuk setiap interval digunakan rumus yakni:

$$w = (max - min)/k \quad \text{.....(8)}$$

Batas interval yang dihasilkan yakni $min + w, min + 2w, \dots, min + (k-1)w$ [9]. Setelah menerapkan metode *binning* pada data jumlah penduduk miskin maka diperoleh 3 kategori yakni kategori rendah (*low*), cukup (*fair*), dan tinggi (*high*). Kategori rendah berarti jumlah penduduk miskin berada dibawah nilai interval *w*, kategori cukup berarti jumlah penduduk miskin berada pada nilai tengah interval, sementara kategori tinggi berarti jumlah penduduk miskin berada diatas nilai interval *w*. Parameter jumlah penduduk miskin ini kemudian digunakan sebagai parameter untuk menentukan klasifikasi jumlah penduduk miskin di setiap provinsi. Pada akhir dari tahapan ini akan dihasilkan *training dataset* dengan jumlah 272 data dan *testing dataset* berjumlah 82 data yang akan digunakan pada tahapan selanjutnya yakni tahap pengembangan model prediksi. Rasio antara *training dataset* dan *testing dataset* adalah 70:30 dimana 70% dari keseluruhan data digunakan sebagai *training dataset* dan 30% digunakan sebagai *testing dataset*.

2.3 Model Development

Tahapan selanjutnya dari penelitian ini yakni tahap pengembangan model prediksi. Pada tahapan ini teknik *data mining* dan *machine learning* diaplikasikan. Algoritma yang digunakan pada penelitian ini adalah algoritma J48 *Decision Tree*. Algoritma *Decision Tree* telah banyak digunakan dalam berbagai bidang. Pada studi yang dilakukan oleh Raditya, algoritma *decision tree* digunakan untuk mencari pola prediksi hujan [10]. Studi lainnya menggunakan algoritma *decision tree* untuk pemetaan daerah rawan longsor [11]. Selanjutnya, algoritma ini juga digunakan untuk mendeteksi diabetes pada seseorang [12, 13].

Algoritma J48 sendiri adalah sebuah algoritma turunan dari C4.5. Algoritma ini menghasilkan pohon biner dimana dalam proses klasifikasi pohon akan dibangun dan setiap tupel dari pohon tersebut akan diterapkan pada basis data dan hasil klasifikasi dari tupel tersebut [10, 14, 15]. Algoritma J48 akan mengabaikkan nilai yang tidak lengkap dalam proses pembuatan pohon. Dasar dari algoritma ini adalah untuk membagi data ke dalam beberapa bagian berdasarkan nilai atribut dari item yang ada pada *training dataset*. Algoritma J48 dapat melakukan klasifikasi baik melalui decision tree ataupun rules yang diperoleh dari pohon tersebut[16]. Adapun langkah-langkah dalam algoritma J48 adalah [17]:

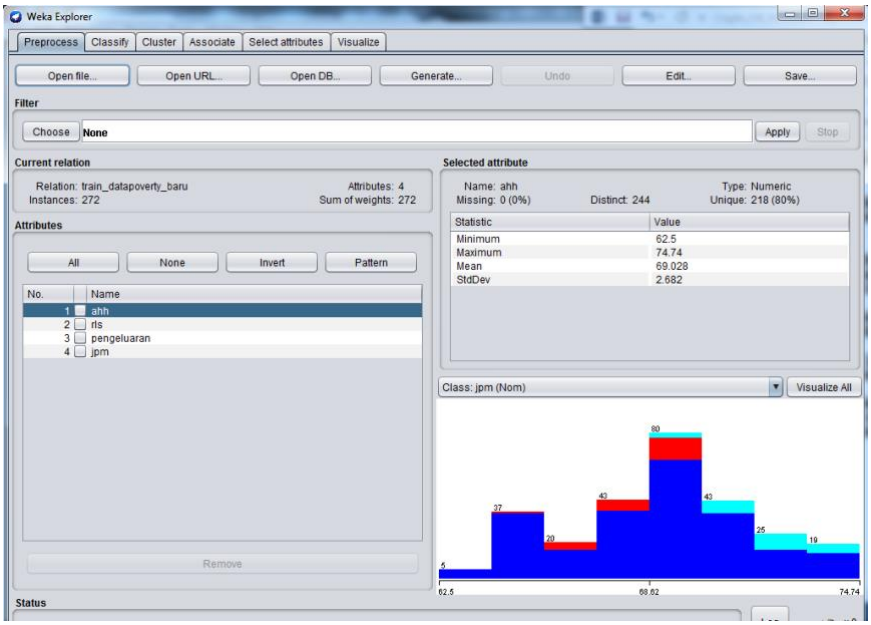
1. Menetapkan *training dataset*.
2. Penentuan akar dari pohon keputusan.

3. Penghitungan nilai Gain menggunakan persamaan
   $Entropy(S) = \sum_{i=1}^{n} - pi * log_2\ pi$.....(1)
4. Mulai kembali langkah ke-2 sampai semua tupel terbagi dengan menggunakan persamaan $Gain(S, A) = S - \sum_{i=1}^{n} \frac{|S_i|}{|S|} * S_i$.....(2)
5. Proses pembagian akan berhenti ketika semua tupel dalam titik N telah memperoleh kelas yang sama dana tau sudah tidak ada atribut dalam tupel yang dibagi lagi atau tidak ada tupel dalam cabang yang kosong.

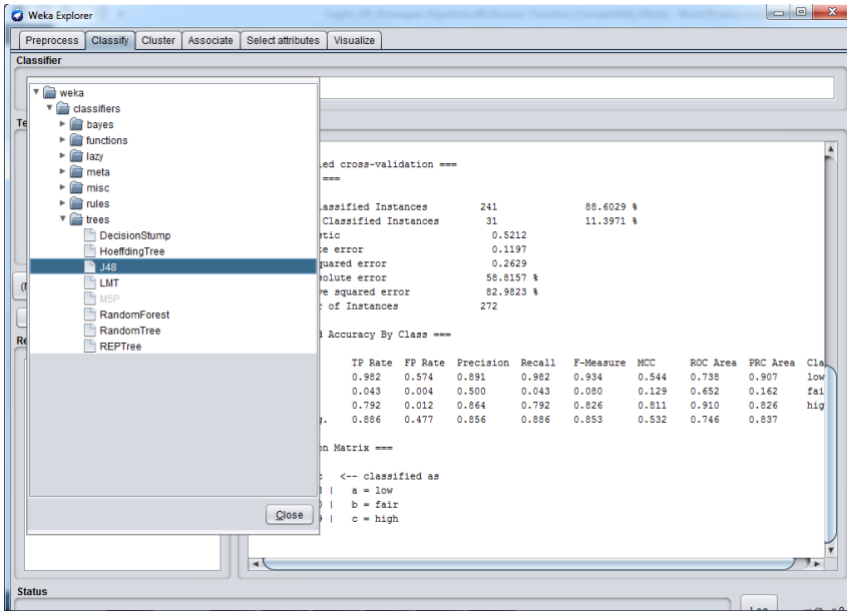
Untuk mengoptimasi parameter klasifikasi digunakan *10-fold* validasi silang (*cross validation*). Kemudian *testing dataset* digunakan untuk mengukur dan menguji validitas model prediksi yang dikembangkan. Hasil akhir dari tahapan-tahapan di atas adalah berupa model prediksi akhir yang dapat memberikan pengetahuan baru.

3. HASIL DAN PEMBAHASAN

Pada penelitian ini terdapat empat atribut yang digunakan yakni angka harapan hidup, rata-rata lama sekolah, pengeluaran perkapita, serta atribut jumlah penduduk miskin yang dijadikan sebagai atribut kelas. Proses klasifikasi pada penelitian ini digunakan WEKA [13, 18] untuk mengevaluasi model prediksi yang dibuat. WEKA adalah perangkat lunak yang memuat algoritma *machine learning* yang digunakan dalam menyelesaikan tugas-tugas atau analisa pada *data mining*[19, 20].



Gambar 2 Tampilan WEKA ketika training dataset diupload



Gambar 3 Tampilan WEKA ketika dengan algoritma machine learning

Adapun metrik-metrik yang dievaluasi pada penelitian ini yakni *accuracy*, *precision (specificity)*, *recall (sensitivity)*, dan *F-measure* menggunakan persamaan-perasamaan berikut:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}$$

$$S_p = \frac{TN}{TN + FP}$$

$$S_n = \frac{TP}{TP + FN}$$

$$F - measure = 2\frac{Sp * Sn}{Sp + Sn}$$

Accuracy adalah sebuah pengujian untuk mencari tahu kemampuan model yang dikembangkan untuk membedakan penduduk miskin pada tiga klasifikasi yakni rendah (*low*), wajar (*fair*), ataupun high (*tinggi*) secara keseluruhan. *Specificity* adalah pengujian untuk mencari tahu kemampuan model yang dikembangkan dalam menentukan instans yang diklasifikasikan benar dengan jumlah total instans dalam data. *Sensitivity* menyatakan rasio instans positif yang diklasifikasikan benar terhadap semua instans dalam kelas aktual. Sementara metrik F-measure adalah untuk menghitung nilai mean yang merupakan gabungan dari *Sensitivity* dan *Specificity* [21, 22].

Tabel 1 memberikan hasil evaluasi dari keempat metric tersebut menggunakan 10-fold validasi silang (*cross validation*). Seperti yang terlihat pada Tabel 1 hasil akurasi yang bagus yakni sebesar 85.66% dengan 233 instans yang diklasifikasikan dengan benar (*correctly classified instances*) dan 39 instans diklasifikasikan sebagai tidak benar (*incorrectly classified instansces*) seperti yang terlihat pada gambar 2.

```
  a   b   c   <-- classified as
217   3   5 |   a = low
 22   0   1 |   b = fair
  8   0  16 |   c = high
```

Gambar 4 Gambar confusion matrix yang menyatakan instans yang diklasifikasi menggunakan validasi silang

Tabel 1 Hasil evaluasi model prediksi dengan 10-fold validasi silang

Evaluasi Metrik	Nilai hasil
Accuracy	85.66%
Precision	0.791
Recall	0.857
F-measure	0.822

Hasil evaluasi model prediksi yang dikembangkan menggunakan *testing dataset* dapat dilihat pada Tabel 2. Hasil ini menyatakan validitas dari model prediksi penelitian ini menggunakan data independen. Dengan kata lain *testing dataset* yang digunakan tidak dimodifikasi. Pengujian independen ini menghasilkan 241 instans yang diklasifikasikan dengan benar dan 31 instans yang diklasifikasikan tidak benar oleh model yang dikembangkan seperti yang terlihat pada gambar 3.

```
  a   b   c   <-- classified as
221   1   3 |   a = low
 22   1   0 |   b = fair
  5   0  19 |   c = high
```

Gambar 5 Gambar confusion matrix yang menyatakan instans yang diklasifikasi

Dari hasil yang ada dapat dilihat bahwa model prediksi yang dikembangkan dapat dijadikan model sebagai bahan pertimbangan para pembuat kebijakan sebelum mengambil keputusan.

Tabel 2 Hasil evaluasi model prediksi menggunakan *testing dataset*

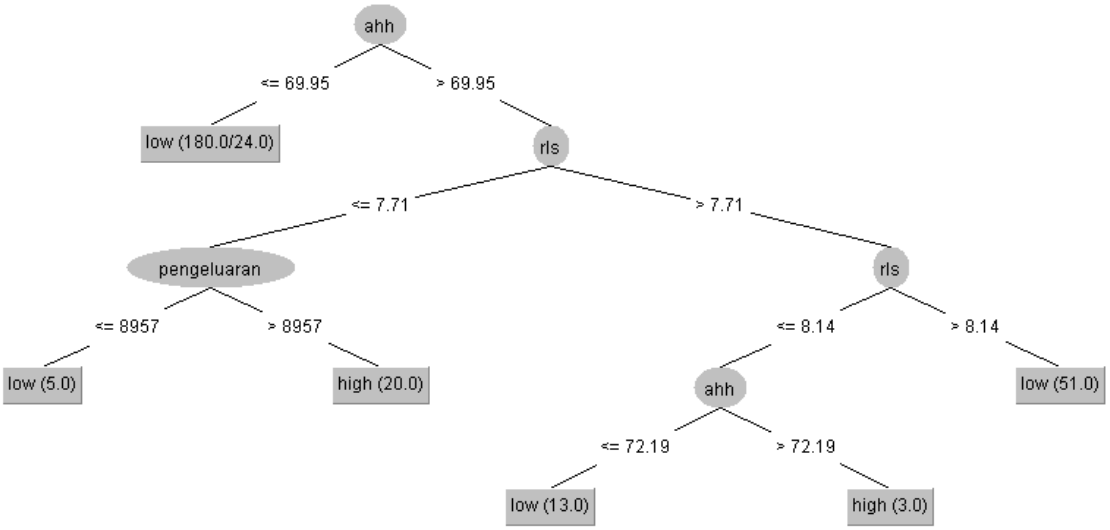
Evaluasi Metrik	Nilai hasil
Accuracy	88.60%
Precision	0.856
Recall	0.886
F-measure	0.853

Hasil lainnya dari penelitian ini yakni hasil visualisasi dari algoritma J48 Decision Tree yang dapat dilihat pada Gambar 2. Hasil tersebut menunjukkan atribut yang paling mempengaruhi prediksi jumlah penduduk miskin yakni atribut angka harapan hidup yang menyatakan dimensi kesehatan pada penghitungan indeks pembangunan manusia. Berdasarkan hasil visualisasi jika angka harapan hidup berada lebih kecil atau sama dengan 69.95 maka jumlah penduduk miskin tergolong rendah. Sementara jika angka harapan hidup berada di atas 69.95 maka atribut selanjutnya yang ditinjau oleh algoritma yakni rata-rata lama sekolah yang merepresentasikan dimensi pendidikan pada penghitungan IPM. Jika rata-rata lama sekolah berada di atas 7.71 dan pengeluaran perkapita berada di atas Rp. 8957 maka jumlah penduduk miskin tergolong tinggi.

Dari hasil yang ada diharapkan dapat membantu para pengambil keputusan dalam hal pemerataan indeks pembangunan manusia di Indonesia.

4. KESIMPULAN

Penelitian yang telah dilakukan menyediakan informasi mengenai keterkaitan antara dimensi pembentuk indeks pembangunan manusia dan jumlah penduduk miskin di Indonesia. Hasil yang ada menunjukkan bahwa tingginya jumlah penduduk miskin di Indonesia diprediksi terjadi pada provinsi dengan angka harapan hidup lebih besar dari 69.95, rata-rata lama sekolah lebih kecil dari 7.71, dan pengeluaran perkapita lebih besar dari Rp. 8.957,00.- Hasil ini diharapkan dapat membantu para pembuat kebijakan untuk memperhatikan indeks pembangunan manusia di daerah-daerah yang memiliki angka-angka di atas. Akurasi dari model prediksi yang telah dibuat menunjukan hasil yang baik yakni sebesar 88.6% dimana dengan kata lain model prediksi yang dikembangkan dapat digunakan untuk membantu para pembuat kebijakan maupun para pemangku kepentingan untuk mengambil keputusan.



Gambar 6 Hasil visualisasi decision tree

5. SARAN

Pengembangan dari penelitian ini yakni dapat disertakan atribut indeks pembangunan manusia agar dapat menyediakan hasil yang lebih lengkap dan menyeluruh. Di samping itu, dapat juga digunakan algoritma data mining lainnya seperti Random Forest, SVM, Naïve Bayes, dan sebagainya untuk membandingkan hasil evaluasi dan memungkinkan mengembangkan model prediksi yang lebih baik lagi.

DAFTAR PUSTAKA

[1] BPS. (2018, December 5, 2018). *Kemiskinan dan Ketimpangan*. Available: https://www.bps.go.id/subject/23/kemiskinan-dan-ketimpangan.html

[2] OPHI. (2018, December 6, 2018). *Global MPI 2018*. Available: https://ophi.org.uk/multidimensional-poverty-index/global-mpi-2018/

[3] A. Shah. (2011, December 6, 2018). *Poverty Around The World*. Available: http://www.globalissues.org/article/4/poverty-around-the-world

[4] BPS. (2018, December 5, 2018). *Persentase penduduk miskin Maret 2018 turun menjadi 9,82 persen*. Available: https://www.bps.go.id/pressrelease/2018/07/16/1483/persentase-penduduk-miskin-maret-2018-turun-menjadi-9-82-persen.html

[5] H.-J. Kao, S.-L. Weng, K.-Y. Huang, F. J. Kaunang, J. B.-K. Hsu, C.-H. Huang, *et al.*, "MDD-carb: a combinatorial model for the identification of protein carbonylation sites with substrate motifs," *BMC systems biology*, vol. 11, p. 137, 2017.

[6] A. V. D. Sano and H. Nindito, "Application of K-Means Algorithm for Cluster Analysis on Poverty of Provinces in Indonesia," *ComTech: Computer, Mathematics and Engineering Applications*, vol. 7, pp. 141-150, 2016.

[7] N. Jean, M. Burke, M. Xie, W. M. Davis, D. B. Lobell, and S. Ermon, "Combining satellite imagery and machine learning to predict poverty," *Science*, vol. 353, pp. 790-794, 2016.

[8] BPS. (2018, December 4, 2018). *Indeks Pembangunan Manusia*. Available: https://www.bps.go.id/subject/26/indeks-pembangunan-manusia.html#subjekViewTab1

[9] J. Dougherty, R. Kohavi, and M. Sahami, "Supervised and unsupervised discretization of continuous features," in *Machine Learning Proceedings 1995*, ed: Elsevier, 1995, pp. 194-202.

[10] A. Raditya, "Implementasi data mining classification untuk mencari pola prediksi hujan dengan menggunakan algoritma C4. 5," 2012.

[11] D. T. Bui, T. C. Ho, I. Revhaug, B. Pradhan, and D. B. Nguyen, "Landslide susceptibility mapping along the national road 32 of Vietnam using GIS-based J48 decision tree classifier and its ensembles," in *Cartography from pole to pole*, ed: Springer, 2014, pp. 303-317.

[12] G. Kaur and A. Chhabra, "Improved J48 classification algorithm for the prediction of diabetes," *International Journal of Computer Applications*, vol. 98, 2014.

[13] S. Drazin and M. Montag, "Decision tree analysis using weka," *Machine Learning-Project II, University of Miami*, pp. 1-3, 2012.

[14] J. Han, J. Pei, and M. Kamber, *Data mining: concepts and techniques*: Elsevier, 2011.

[15] A. K. Sharma and S. Sahni, "A comparative study of classification algorithms for spam email data analysis," *International Journal on Computer Science and Engineering*, vol. 3, pp. 1890-1895, 2011.

[16] T. R. Patil and S. Sherekar, "Performance analysis of Naive Bayes and J48 classification algorithm for data classification," *International journal of computer science and applications*, vol. 6, pp. 256-261, 2013.

[17] N. S. Diwandari and N. A. Setiawan, "Perbandingan Algoritma J48 dan NBTREE Untuk Klasiifikasi Diagnosa Penyakit Pada Soybean," in *Seminar Nasional Teknologi Informasi dan Komunikasi 2015 (SENTIKA 2015)*, 2015, p. 208.

[18] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, pp. 10-18, 2009.

[19] E. Frank, M. Hall, G. Holmes, R. Kirkby, B. Pfahringer, I. H. Witten, *et al.*, "Weka-a machine learning workbench for data mining," in *Data mining and knowledge discovery handbook*, ed: Springer, 2009, pp. 1269-1277.

[20] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical machine learning tools and techniques*: Morgan Kaufmann, 2016.

[21] D. M. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," 2011.

[22] E. A. Amrieh, T. Hamtini, and I. Aljarah, "Preprocessing and analyzing educational data set using X-API for improving student's performance," in *Applied Electrical Engineering and Computing Technologies (AEECT), 2015 IEEE Jordan Conference on*, 2015, pp. 1-5.