

Prediksi Hipertensi Menggunakan *Decision Tree*, *Naïve Bayes* dan *Artificial Neural Network* pada Software KNIME

Prediction of Hypertension using Decision Tree, Naïve Bayes and Artificial Neural Networks in KNIME Analytics Platform

Mayanda Mega Santoni¹, Nurul Chamidah², Nurhafifah Matondang³

^{1,2}Program Studi Informatika, Fakultas Ilmu Komputer, UPN Veteran Jakarta

³Program Studi Sistem Informasi, Fakultas Ilmu Komputer, UPN Veteran Jakarta

E-mail: ¹megasantoni@upnvj.ac.id, ²nurulchamidah@upnvj.ac.id,

³nurhafifahmatondang@upnvj.ac.id

Abstrak

Hipertensi merupakan salah satu penyakit tidak menular yang dapat menyebabkan kematian karena meningkatkan resiko munculnya berbagai penyakit seperti gagal ginjal, gagal jantung, bahkan stroke. Resiko hipertensi disebabkan oleh beberapa faktor penyebab seperti usia, keturunan, pola makan dan olahraga, dan merokok. Teknologi artificial intelligence yakni machine learning dimanfaatkan di bidang kesehatan khususnya prediksi penyakit hipertensi. Pada penelitian ini diimplementasi tiga algoritma machine learning yakni *decision tree*, *naïve bayes* dan *artificial neural networks*. Data yang digunakan pada penelitian ini sebanyak 274 data yang diperoleh dari hasil kuesioner dengan 26 pertanyaan, dimana 25 pertanyaan adalah variabel faktor resiko dan satu pertanyaan merupakan kelas yang menyatakan responden memiliki riwayat hipertensi atau tidak. Data diolah menggunakan platform analisis data yakni KNIME. Sebelum data diolah untuk membangun model klasifikasi menggunakan *decision tree*, *naïve bayes* dan *artificial neural network*, data dipraproses terlebih dahulu dengan melakukan imputasi missing value, oversampling dan normalisasi data. Selanjutnya pembagian data menggunakan 5-fold cross validation. Model klasifikasi yang diperoleh dievaluasi menggunakan nilai akurasi, recall dan precision. Hasil evaluasi dari eksperimen yang dilakukan diperoleh bahwa algoritma *artificial neural network* memiliki tingkat performa lebih baik dibandingkan *decision tree* dan *naïve bayes* dengan nilai akurasi sebesar 94.7%, *recall* sebesar 91.5% dan *precision* sebesar 97.7%.

Kata kunci: Hipertensi, *Decision Tree*, *Naïve Bayes*, *Artificial Neural Networks*, KNIME

Abstract

Hypertension is a non-communicable disease that can cause death because it increases the risk of various diseases such as kidney failure, heart failure, and even stroke. The risk of hypertension is caused by several. Artificial intelligence technology is utilized in the health sector, especially hypertension prediction. In this research, three machine learning algorithms are implemented such as decision tree, naïve bayes and artificial neural network (ANN). The data used in this study were 274 data obtained from questionnaire results with 26 questions, consisting of 25 questions were risk factor variables and one question was a class that stated the respondent had a history of hypertension or not. The data is processed using a data analysis platform KNIME. Before the data is processed to build classification models using decision trees, naïve bayes and neural networks, the data is preprocessed by imputing missing values, oversampling and data normalization. Furthermore, we split data for training and testing using 5-fold cross validation. Evaluation classification model using accuracy, recall and precision. The evaluation results of the experiments carried out showed that the ANN algorithm has a better level of performance than the Decision Tree and Naïve Bayes with 94.7% accuracy, 91.5% recall, and 97.7% precision.

Keywords: Hypertension, *Decision Tree*, *Naïve Bayes*, *Artificial Neural Network*, KNIME

1. PENDAHULUAN

Hipertensi atau biasa juga disebut dengan tekanan darah tinggi, merupakan suatu kondisi dimana seseorang memiliki tekanan darah yang melebihi dari standar normalnya. Kondisi seperti ini sangat berbahaya, karena jantung akan bekerja lebih keras untuk memompa darah ke seluruh tubuh sehingga dapat menyebabkan munculnya berbagai penyakit, seperti gagal ginjal, gagal jantung, bahkan stroke. Banyak faktor penyebab yang dapat meningkatkan resiko hipertensi pada seseorang, seperti usia, keturunan, pola makan dan olahraga, dan merokok.

Kemajuan teknologi saat ini dapat dimanfaatkan di berbagai bidang, termasuk bidang kesehatan. Di era Industri 4.0, salah satu teknologi yang sering dimanfaatkan yakni teknologi *artificial intelligence* (AI) [1]. Teknologi AI merupakan sebuah teknologi dimana mesin memiliki kecerdasan seperti layaknya manusia berpikir. Pada bidang kesehatan, AI dapat digunakan untuk meminimalisir kesalahan yang disebabkan oleh manusia (*human error*) dan membuat segala sesuatu menjadi lebih efisien. Salah satu contohnya adalah teknologi kecerdasan buatan digunakan untuk memprediksi penyakit sebagai deteksi dini terhadap suatu penyakit.

Salah satu sub bidang dalam *artificial intelligence* adalah *machine learning*, yakni bidang ilmu yang memungkinkan komputer untuk belajar dari sekumpulan data sehingga menghasilkan sebuah model pembelajaran yang dapat digunakan oleh mesin. Proses belajar pada *machine learning* menggunakan algoritma khusus yang biasa disebut dengan *machine learning algorithms*. Beberapa algoritma pembelajaran mesin yang populer seperti *artificial neural networks*, *decision tree* dan *naïve bayes*.

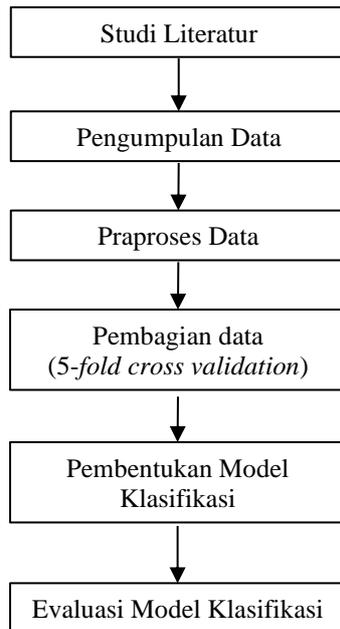
Penelitian di bidang kesehatan yang menggunakan teknologi AI yakni *machine learning* sudah banyak dilakukan, khususnya terkait deteksi hipertensi. Penelitian [2] melakukan penelitian tentang prediksi hipertensi berdasarkan data klinis pasien menggunakan algoritma *machine learning* yaitu *artificial neural networks*. Pada penelitian ini diperoleh tingkat akurasi model sebesar 82%. Penelitian [3] juga melakukan penelitian tentang deteksi hipertensi seseorang berdasarkan data klinis menggunakan algoritma *naïve bayes*. Diperoleh tingkat akurasi model sebesar 83.69%. Penelitian [4] juga melakukan penelitian tentang prediksi hipertensi berdasarkan data kuesioner menggunakan algoritma *logistic regression* dan *artificial neural network*. Tingkat akurasi dari algoritma yang digunakan pada penelitian ini adalah 72%. Selain itu, penelitian [5] juga mengembangkan model prediksi untuk mengklasifikasikan tingkat resiko hipertensi menggunakan algoritma *decision tree*. *Decision tree* dapat digunakan untuk mengetahui variabel yang berpengaruh dalam prediksi tingkat resiko hipertensi pada seseorang.

Untuk melakukan proses pembelajaran, banyak *tools* yang dapat digunakan. Pada penelitian [6] membandingkan lima *data mining tools* yakni WEKA, Rapidminer, Tanagra, Orange and KNIME dalam memprediksi penyakit *liver*. Diperoleh dari kelima tools tersebut, KNIME memiliki nilai akurasi yang paling tinggi pada tiga algoritma *machine learning* yang diujikan. Selain itu, penelitian [7] juga merekomendasikan KNIME sebagai sebuah platform analisis data bagi para pemula karena *tool*-nya yang sederhana namun memiliki hasil yang *robust*.

Oleh karena itu, berdasarkan latar belakang masalah dan penelitian-penelitian sebelumnya yang telah dibahas di atas, maka penelitian ini akan menggunakan KNIME sebagai platform analisis data untuk memprediksi hipertensi yang akan diterapkan pada tiga algoritma *machine learning*, yakni *decision tree*, *naïve Bayes* dan *artificial neural network*.

2. METODE PENELITIAN

Metodologi yang digunakan pada penelitian ini dapat dilihat pada Gambar 1. Dimulai dari pengumpulan data penelitian, kemudian dilakuka praproses data agar data siap untuk diolah. Sebelum masuk ke proses pembentukan model klasifikasi, data akan dibagi menjadi data latih dan data uji. Data latih digunakan untuk pembentukan model, dan data uji digunakan untuk mengevaluasi model hasil klasifikasi.



Gambar 1 Metodologi Penelitian

2.1 Studi Literatur

Penelitian terkait prediksi hipertensi telah banyak dilakukan dengan menggunakan beberapa algoritma *machine learning* seperti *artificial neural network*, *naïve bayes*, dan *decision tree*.

Pada penelitian [2] melakukan prediksi hipertensi berdasarkan data klinis pasien menggunakan algoritma *artificial neural networks*. Data yang digunakan sebanyak 185,371 data pasien dengan 11 parameter yakni umur, jenis kelamin, *body mass index*, tekanan darah *systolic*, tekanan darah *diastolic*, *high density lipoprotein*, *low density lipoprotein*, *Triglycerides*, kolesterol, *micro-albumin*, dan *Urine Albumin-Creatinine Ratio*. Dengan menggunakan algoritma *artificial neural network*, diperoleh nilai akurasi sebesar 82%. Akurasi yang dihasilkan baik untuk dataset yang besar.

Penelitian [4] melakukan prediksi hipertensi tanpa melakukan pengukuran klinis. Data yang digunakan diperoleh dari hasil kuesioner yang diambil dari data Sistem Surveilans Faktor Risiko Perilaku (*Behavior Risk Factor Surveillance System*). Jumlah data yang digunakan sebanyak 308,711 data responden dengan 13 pertanyaan, yakni umur, jenis kelamin, tinggi badan, berat badan, status pernikahan, tingkat pendapatan, tingkat pendidikan, frekuensi olahraga, diabetes, kolesterol, frekuensi merokok, frekuensi dan konsumsi alkohol. Data tersebut dilatih dengan menggunakan algoritma *artificial neural network* dengan tingkat akurasi yang dihasilkan sebesar 72%.

Penelitian [8] melakukan evaluasi pada algoritma *artificial neural network* dalam memprediksi hipertensi. Data yang digunakan 981 data klinis pasien dengan 13 parameter, yakni umur, tekanan darah *systolic*, tekanan darah *diastolic*, serum protein, serum albumin, *hematocrit*, *Erythrocyte Sedimentation Rate*, *Serum Cholesterol*, *Serum Triglycerides*, *Whole Blood Viscosity*, *Plasma Viscosity*, dan *Red Cell Aggregation*. Tingkat performa maksimum yang dihasilkan oleh *artificial neural network* pada data ini sebesar 92.85%.

Penelitian [3] melakukan prediksi hipertensi dengan menggunakan algoritma *naïve bayes*. Data yang digunakan sebanyak 52 data pasien dengan 10 parameter yakni umur, jenis kelamin, *chest pain*, tekanan darah *systolic*, tekanan darah *diastolic*, kolesterol, gula darah puasa, *thalach*, *exang* dan *old peak*. Performa dari algoritma *naïve Bayes* pada data ini sebesar 83.67%.

Penelitian [5] mengembangkan model prediksi untuk mengklasifikasikan tingkat resiko hipertensi menggunakan algoritma *decision tree*. Data yang digunakan pada penelitian ini

merupakan data kuesioner dari 32 data pasien yang dirawat di rumah sakit. Kuesioner berisikan informasi pribadi pasien seperti jenis kelamin, umur, tingkat pendidikan, pekerjaan, status pernikahan, suku, dan juga informasi faktor resiko hipertensi, seperti penyakit turunan, body mass index, frekuensi olahraga, tingkat stres pekerjaan, jam kerja, merokok, konsumsi alkohol, dan lain-lain. Algoritma *decision tree* dapat digunakan untuk mengetahui variabel yang berpengaruh dalam memprediksi tingkat resiko hipertensi.

Penelitian [9] melakukan analisa terhadap pola penyakit hipertensi menggunakan algoritma *decision tree* yakni C4.5. Parameter yang digunakan pada penelitian ini sebanyak 8 parameter, yakni usia, jenis kelamin, lama dirawat, memiliki diabetes, gagal ginjal, gagal jantung, stroke, *Hipogilkemia*. Dari 8 parameter tersebut, penyakit penyerta seperti gagal ginjal, gagal jantung, stroke, diabetes dan hipogilkemia memiliki pengaruh terhadap prediksi hipertensi. Dari parameter tersebut, gagal ginjal merupakan parameter paling berpengaruh dalam prediksi hipertensi.

2.2 Pengumpulan Data

Data penelitian diperoleh melalui kuesioner dengan total responden sebanyak 274. Jumlah pertanyaan sebanyak 26 pertanyaan, dimana 25 pertanyaan adalah variabel faktor resiko dan 1 pertanyaan merupakan kelas yang menyatakan responden memiliki riwayat hipertensi atau tidak. Daftar pertanyaan dikelompokkan menjadi 6 bagian, yakni informasi data diri, gaya hidup, pola makan, kebiasaan istirahat, kebiasaan merokok, dan riwayat kesehatan. data untuk masing-masing kelas, yakni 40 responden yang memiliki riwayat hipertensi dan 234 responden yang tidak memiliki riwayat hipertensi. Deskripsi data penelitian dapat dilihat pada Tabel 1.

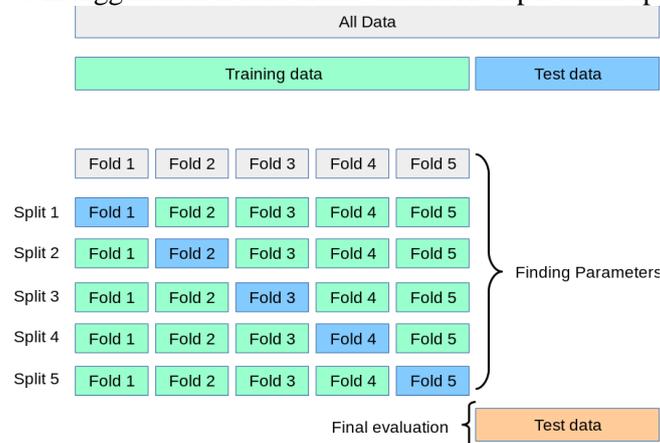
Tabel 1 Deskripsi Data Penelitian

No	Nama variabel	Deskripsi variable	Tipe Data
1	Usia	Usia	Number (Integer)
2	JK	Jenis Kelamin	Pria, Wanita
3	SP	Status Pernikahan	Single, Married, Divorced
4	TB	Tinggi Badan	Number (Integer)
5	BB	Berat Badan	Number (Integer)
6	OR	Olahraga 30 menit perhari	Ya, Tidak
7	ORH	Olahraga minimal 1 kali perminggu	Ya, Tidak
8	Daging	Konsumsi daging <3 kali seminggu	Ya, Tidak
9	Lemak	Konsumsi makanan berlemak tinggi < 3 kali seminggu	Ya, Tidak
10	Goreng	Konsumsi gorengan < 3 kali seminggu	Ya, Tidak
11	FF	Konsumsi makanan cepat saji < 3 kali seminggu	Ya, Tidak
12	Kafein	Konsumsi kafein < 3 kali seminggu	Ya, Tidak
13	Asin	Konsumsi makanan yang diasinkan < 3 kali seminggu	Ya, Tidak
14	Sayur	konsumsi sayuran >= 3 kali seminggu	Ya, Tidak
15	Buah	konsumsi buah-buahan >= 3 kali seminggu	Ya, Tidak
16	Mie	konsumsi mie instan > 2 bungkus seminggu.	Ya, Tidak
17	Bangun	frekuensi terbangun <2 kali saat tidur malam	Ya, Tidak
18	Insom	frekuensi susah tidur <2 kali dalam seminggu	Ya, Tidak
19	Nap	tidur siang 1-2 jam >= 3 kali dalam seminggu	Ya, Tidak
20	Sleep	tidur teratur 6-8 jam pada malam hari	Ya, Tidak

No	Nama variabel	Deskripsi variable	Tipe Data
21	Smoke	Apakah merokok?	Ya, Tidak
22	Highsmoke	merokok > 20 batang dalam sehari	Ya, Tidak
23	Wine	kebiasaan minum minuman beralkohol	Ya, Tidak
24	Diabet	Memiliki riwayat diabetes	Ya, Tidak
25	Kolesterol	Memiliki riwayat kolesterol	Ya, Tidak
26	Hipertensi	Memiliki riwayat hipertensi	Ya, Tidak

2.3. Pembagian Data

Setelah data selesai dipraproses, maka selanjutnya akan dilakukan pembagian data. Data akan dibagi menjadi data latih dan data uji menggunakan metode k-fold cross validation dengan nilai k yang digunakan sebanyak lima. 5-fold cross validation artinya akan membagi data menjadi lima bagian dan setiap bagiannya pernah digunakan sebagai data latih dan juga data uji. Ilustrasi proses pembagian data menggunakan k-fold cross validation dapat dilihat pada Gambar 2.



Gambar 2 K-fold cross validation [10]

2.4. Pembentukan Model Klasifikasi

Pada penelitian ini, terdapat tiga model klasifikasi yang akan diimplementasikan, yakni *decision tree*, *naïve bayes*, dan *artificial neural network*. Data latih pada pembagian data akan digunakan untuk menghasilkan model klasifikasi untuk masing-masing algoritma.

1. Decision tree

Algoritma *decision tree* atau biasa disebut dengan pohon keputusan merupakan algoritma klasifikasi yang menggunakan struktur pohon yang merepresentasikan aturan dari sebuah pengetahuan. Model klasifikasi pada algoritma ini mudah diinterpretasikan oleh manusia. Untuk melakukan pemilihan node pada tree menggunakan nilai dari Gini Index. Nilai gini index bervariasi dari rentang 0 sampai dengan 1. Gini index bernilai 0 menunjukkan bahwa semua data memiliki kelas tertentu atau hanya ada 1 kelas. Sementara itu, gini index bernilai 1 artinya bahwa data terdistribusi secara acak di semua kelas. Gini index bernilai 0.5 menunjukkan bahwa data tersebar secara merata di semua kelas. Gini index dapat dihitung menggunakan Persamaan 1.

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

dengan nilai n merupakan jumlah kelas pada data, dan P_i merupakan probabilitas objek yang diklasifikasikan ke sebuah kelas.

2. Naïve Bayes

Algoritma *naïve Bayes* merupakan algoritma klasifikasi yang didasarkan pada probabilitas dan teorema bayesian. Dikatakan “naïve” karena algoritma ini menerapkan asumsi bahwa semua variabel bersifat independen, yaitu tidak adanya kaitan sebuah variabel dengan variabel yang lain. Padahal kenyataannya, kita jarang sekali menemukan kasus dimana variabel - variabelnya tidak berkaitan. Namun kinerja dari algoritma *naïve Bayes* ini baik dan proses pembelajarannya cepat di berbagai masalah klasifikasi. Probabilitas pada *naïve Bayes* dapat dihitung menggunakan Persamaan 2.

$$P(Y | X) = P(X | Y).P(Y) / P(X) \quad (2)$$

Dimana:

X : data sampel dengan kelas label yang belum diketahui,

Y : hipotesis bahwa X adalah data dengan kelas label C ,

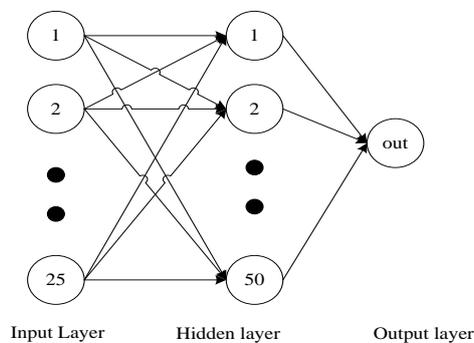
$P(Y)$: adalah peluang dari hipotesis Y (prior), $P(X)$: peluang data sampel yang diamati (evidence),

$P(X/Y)$: peluang data sampel X , bila diasumsikan bahwa hipotesis Y benar (likelihood).

Untuk masalah klasifikasi, yang dihitung adalah $P(Y|X)$, yaitu peluang bahwa hipotesis Y benar untuk data sampel X yang diamati.

3. Artificial Neural Network

Artificial Neural Network (ANN) merupakan *multilayer neural network* yang digunakan untuk membangun model ANN dalam penelitian ini. Arsitektur ANN yang digunakan meliputi, jumlah input neuron pada input layer sejumlah 25 neuron, satu hidden layer dengan 50 hidden neuron, dan output layer berisi satu neuron untuk menentukan kelas hipertensi (1) atau tidak hipertensi (0) dengan epoch atau iterasi 1000. Arsitektur ANN ini dapat dilihat pada Gambar 3.



Gambar 3 Arsitektur ANN

2.5. Evaluasi Model Klasifikasi

Model klasifikasi yang sudah dibentuk dari data latih, akan diujikan dengan menggunakan data uji. Hasil pengujian akan dipetakan ke dalam sebuah matrik yang disebut dengan *confusion matrix*. Ilustrasi *confusion matrix* dapat dilihat pada Tabel 2.

Tabel 2 *Confussion Matrix*

		Kelas Prediksi	
		hipertensi	nonhipertensi
Kelas Sebenarnya	hipertensi	TP	FN
	nonhipertensi	FP	TN

Dimana :

- TP (*True Positive*) : jumlah kelas hipertensi diprediksi benar sebagai kelas hipertensi,
 FP (*False Positive*) : jumlah kelas nonhipertensi yang salah diprediksi sebagai kelas hipertensi,
 TN (*True Negative*) : jumlah kelas nonhipertensi yang diprediksi benar sebagai kelas nonhipertensi,
 FN (*False Negative*) : jumlah kelas hipertensi yang salah diprediksi sebagai kelas nonhipertensi.

Nilai evaluasi yang akan digunakan pada penelitian ini adalah nilai akurasi, recall, dan precision. Nilai akurasi merupakan nilai yang menunjukkan perbandingan jumlah data yang teridentifikasi benar dengan seluruh jumlah data yang diujikan. Evaluasi ini diperoleh didasarkan pada banyaknya data uji yang diprediksi secara benar dan tidak benar oleh model. Nilai akurasi dapat dihitung menggunakan Persamaan 3.

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \quad (3)$$

Nilai recall merupakan nilai perbandingan prediksi benar pada kelas hipertensi dibandingkan dengan keseluruhan data yang sebenarnya kelas hipertensi. Nilai recall dapat dihitung menggunakan Persamaan 4.

$$Recall = \frac{TP}{TP + FN} \quad (4)$$

Nilai precision merupakan perbandingan prediksi benar pada kelas hipertensi dengan keseluruhan data yang diprediksi sebagai kelas hipertensi. Nilai precision dapat dihitung menggunakan Persamaan 5.

$$Recall = \frac{TP}{TP + FP} \quad (5)$$

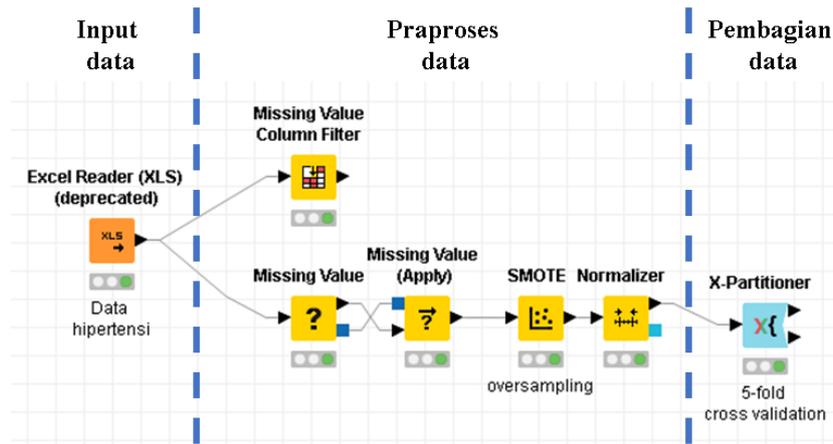
3. HASIL DAN PEMBAHASAN

Data pada penelitian ini diolah menggunakan platform KNIME sebagai alat analisis data. Data penelitian akan disimpan pada file excel dan akan dibaca oleh KNIME menggunakan node excel reader. Jumlah total data awal yakni 274 data dengan distribusi 234 data merupakan data kelas nonhipertensi dan 40 data kelas hipertensi. Semua data akan dilakukan praproses terlebih dahulu, yakni melakukan imputasi pada missing value yakni mengisi nilai yang kosong dengan nilai maksimum pada kolom variabel yang memiliki missing value. Pada penelitian ini terdapat 12 variabel yang memiliki missing value.

Selanjutnya dilakukan penyeimbangan data menggunakan teknik oversampling. Pada KNIME, kita dapat menggunakan node SMOTE. Oversampling dilakukan dengan cara menambahkan jumlah kelas minoritas (hipertensi) sama dengan kelas mayoritas (nonhipertensi). Data kelas hipertensi setelah dilakukan oversampling menjadi 234 data sama dengan data kelas nonhipertensi, sehingga total data setelah dilakukan oversampling adalah 468 data.

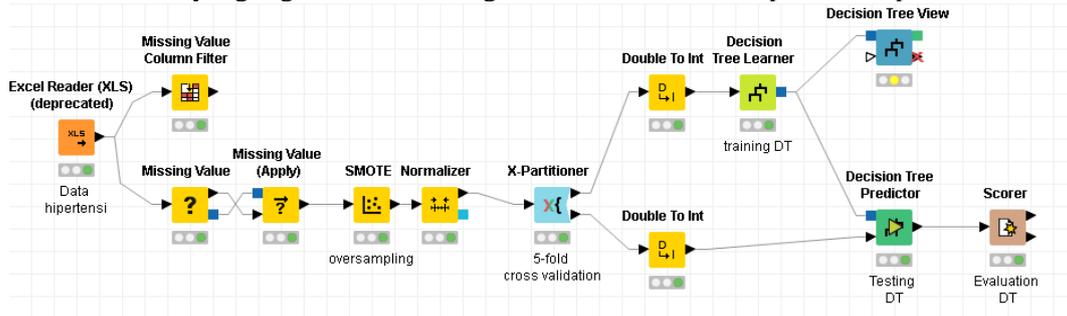
Kemudian data yang telah dioversampling dilakukan normalisasi yakni menyamakan rentang data untuk semua variabel dari 0 – 1. Teknik normalisasi yang digunakan adalah min-max normalization. Pada KNIME, kita dapat menggunakan node Normalizer.

Setelah tahapan praproses selesai, tahapan selanjutnya adalah pembagian data. Pada KNIME, kita dapat menggunakan node X-partitioner dengan menggunakan 5-fold cross validation. Hasil dari pembagian data ini, data akan terbagi menjadi data latih dan data uji. Data latih sebanyak 374 data dan data uji sebanyak 94 data dengan perbandingan data setiap kelasnya sama. Workflow pada KNIME untuk tahapan input, praproses, pembagian data dapat dilihat pada Gambar 4.



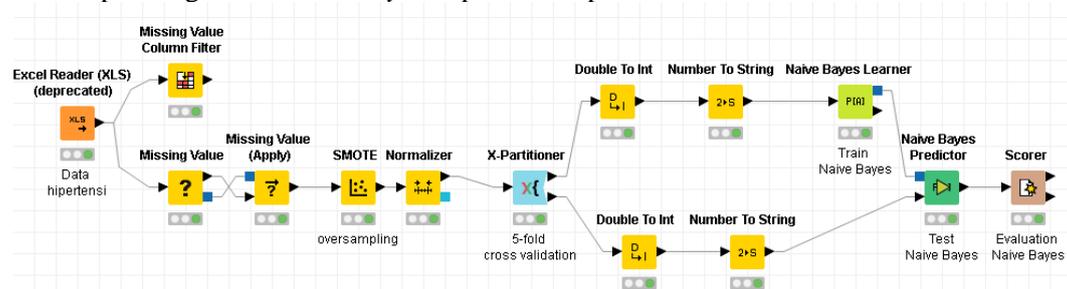
Gambar 4 Workflow KNIME untuk tahapan input, praproses dan pembagian data

Tahap selanjutnya adalah pembentukan model klasifikasi untuk tiga algoritma machine learning yakni *decision tree*, *naïve Bayes* dan *artificial neural network*. Pada KNIME, model *decision tree* dapat diperoleh dengan menggunakan node *Decision tree Learner*. Model yang dihasilkan dapat ditampilkan menggunakan node *Decision tree View*. Kemudian, model yang dihasilkan akan diuji menggunakan data uji. Node *Decision tree Detector* pada KNIME digunakan untuk melakukan proses pengujian. Hasil pengujian dapat dilihat menggunakan node *Scorer*. Workflow yang digunakan untuk algoritma *decision tree* dapat dilihat pada Gambar 5.



Gambar 5 Workflow KNIME untuk *decision tree*

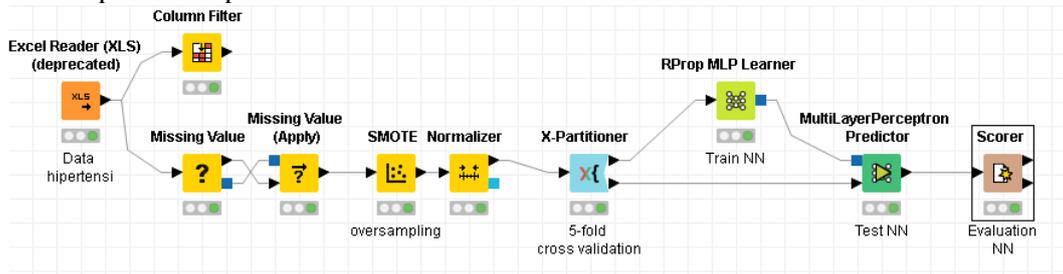
Pembentukan model klasifikasi pada algoritma *naïve Bayes* pada KNIME dapat menggunakan node *Naïve Bayes Learner*. Model yang dihasilkan akan diujikan menggunakan node *Naïve Bayes Predictor* dan hasilnya pengujianya akan ditampilkan pada node *Scorer*. Workflow pada algoritma *naïve Bayes* dapat dilihat pada Gambar 6.



Gambar 6 Workflow KNIME untuk *naïve bayes*

Node RProp MLP Learner pada KNIME merupakan node yang digunakan untuk melakukan proses pelatihan pada data latih dengan menggunakan algoritma *artificial neural network*. Model pelatihan atau model klasifikasi yang dihasilkan, akan diujikan pada data uji

menggunakan node Multilayer Perceptron Predictor. Node Scorer digunakan untuk menampilkan hasil pengujian model klasifikasinya. Workflow yang digunakan untuk algoritma *artificial neural network* dapat dilihat pada Gambar 7.



Gambar 7 Workflow KNIME untuk *artificial neural network*

Hasil evaluasi model klasifikasi dari ketiga algoritma machine learning yakni *decision tree*, *naïve Bayes* dan neural network dapat dilihat pada Tabel 3.

Tabel 3 Hasil Evaluasi Model Klasifikasi

Nilai evaluasi	<i>Decision tree</i>	<i>Naïve Bayes</i>	<i>Artificial neural network</i>
Akurasi	0.830	0.745	0.947
Recall	0.766	0.681	0.915
Precision	0.878	0.780	0.977

Nilai akurasi pada algoritma *decision tree* yakni 0.83 yang artinya algoritma *decision tree* dapat memprediksi masing-masing kelas dengan benar sebesar 83%. Kelas hipertensi benar terklasifikasi sebanyak 36 data dan kelas nonhipertensi benar terklasifikasi sebanyak 42 data dengan total keseluruhan data uji adalah 94 data. Nilai recall pada algoritma *decision tree* yakni 0.766 yang artinya *decision tree* mampu memprediksi benar kelas hipertensi sebesar 76% dari total keseluruhan data yang benar-benar kelas hipertensi. Sementara itu, untuk tingkat precision pada algoritma *decision tree* untuk memprediksi benar kelas hipertensi dari keseluruhan data yang diprediksi kelas hipertensi sebesar 0.878. Pada algoritma ini nilai recall memiliki nilai yang rendah artinya masih terdapat banyak kesalahan pada kelas hipertensi yang diprediksi sebagai kelas tidak hipertensi (false negatif). Confussion matrix pada algoritma *decision tree* dapat dilihat pada Tabel 4.

Tabel 4 *Confussion Matrix* Algoritma *Decision tree*

		Kelas Prediksi	
		hipertensi	nonhipertensi
Kelas Sebenarnya	hipertensi	36	11
	nonhipertensi	5	42

Untuk algoritma *naïve bayes*, nilai akurasi yang diperoleh sebesar 0.745. Kelas hipertensi yang terprediksi benar sebanyak 32 data dan kelas nonhipertensi yang terprediksi benar sebanyak 38 data. Nilai recall pada algoritma *naïve Bayes* sebesar 0.681 dan precision sebesar 0.78. Nilai evaluasi dari algoritma *naïve Bayes* ini memberikan nilai yang paling rendah dibandingkan algoritma *decision tree* dan *artificial neural network*. Pada *naïve Bayes* masih banyak terjadi kesalahan prediksi baik itu pada kelas hipertensi maupun kelas nonhipertensi. Confussion matrix pada algoritma *naïve Bayes* dapat dilihat pada Tabel 5.

Tabel 5 *Confussion Matrix* Algoritma *Naïve Bayes*

		Kelas Prediksi	
		hipertensi	Nonhipertensi
Kelas Sebenarnya	hipertensi	32	15
	nonhipertensi	9	38

Tabel 6 Confussion Matrix Algoritma *Artificial Neural Network*

		Kelas Prediksi	
		hipertensi	Nonhipertensi
Kelas Sebenarnya	hipertensi	43	4
	nonhipertensi	1	46

Tabel 6 menunjukkan *confussion matrix* pada algoritma *artificial neural network*. Algoritma *artificial neural network* memiliki nilai evaluasi model klasifikasi paling tinggi dibandingkan algoritma *decision tree* dan *naïve bayes*. Nilai akurasi pada algoritma *artificial neural network* sebesar 0.947 yang artinya 94.7% dari data uji memprediksi benar sesuai dengan kelas sebenarnya. Kelas hipertensi diprediksi benar sebanyak 43 data dan kelas nonhipertensi diprediksi benar sebanyak 46 data. Nilai recall pada algoritma *artificial neural network* adalah 0.915 dan precision sebesar 0.977. Nilai presisi pada *artificial neural network* memiliki nilai yang lebih tinggi dari nilai recall. Hal ini menunjukkan bahwa ketepatan/presisi *artificial neural network* dalam memprediksi kelas hipertensi baik, karena dari 43 yang diprediksi sebagai kelas hipertensi hanya 1 kelas yang salah prediksi yang seharusnya adalah kelas nonhipertensi (*false positif*). Nilai *recall* pada algoritma *artificial neural network* memiliki nilai lebih rendah dari nilai *precision*. Hal ini menunjukkan bahwa masih terdapat data yang sebenarnya kelas hipertensi namun diprediski sebagai kelas nonhipertensi (*false negatif*). Namun demikian, hasil yang diberikan algoritme *artificial neural network* sudah sangat baik dalam memprediksi penyakit hipertensi pada seseorang.

Tabel 7 Perbandingan Hasil Evaluasi Model Klasifikasi dengan Penelitian Sebelumnya

Nilai evaluasi	Penelitian saat ini (KNIME)			Penelitian Sebelumnya (Python) [11]		
	<i>Decision tree</i>	<i>Naïve Bayes</i>	<i>Artificial neural network</i>	<i>Decision tree</i>	<i>Naïve Bayes</i>	<i>Artificial neural network</i>
Akurasi	0.830	0.745	0.947	0.860	0.710	0.910
Recall	0.766	0.681	0.915	0.879	0.700	0.990
Precision	0.878	0.780	0.977	0.860	0.720	0.860

Pengolahan data menggunakan KNIME menghasilkan performa lebih baik jika dibandingkan dengan penelitian sebelumnya yang diolah menggunakan python. Perbandingan hasil evaluasi dapat dilihat pada Tabel 7. Pada penelitian sebelumnya, algoritme *artificial neural network* memberikan akurasi paling tinggi yakni sebesar 0.91, namun hasil evaluasi ini masih lebih kecil jika dibandingkan dengan hasil evaluasi penelitian ini yakni sebesar 0.947. Oleh karena itu, KNIME dapat digunakan untuk pembelajaran data dengan penggunaan lebih sederhana dan tidak membutuhkan kemampuan *programming*, namun memiliki hasil evaluasi yang *robust*.

4. KESIMPULAN DAN SARAN

Pada penelitian ini, telah diimplementasikan tiga algoritma *machine learning* yakni *decision tree*, *naïve Bayes* dan *artificial neural network* untuk memprediksi penyakit hipertensi. Data dipraproses melalui beberapa tahapan yakni melakuakn imputasi *missing value*, *oversampling* dan normalisasi data. Pembagian data dilakukan menggunakan *k-fold cross validation* menggunakan nilai k=5. Eksperimen dilakukan dengan menggunakan alat analisis data yakni software KNIME.

Hasil yang diperoleh pada penelitian ini, tiga algoritma *machine learning* yang diimplementasikan dapat memprediksi penyakit hipertensi. Algoritma *artificial neural network* memiliki performa yang lebih baik dibandingkan algoritma *decision tree* dan *naïve bayes* dengan nilai akurasi sebesar 94.7%, recall sebesar 91.5% dan precision sebesar 97.7%. Tingkat ketepatan algoritma *artificial neural network* dalam memprediksi kelas hipertensi sangat baik, walaupun tingkat kebenaran algoritma *artificial neural network* dalam memprediksi semua data kelas

hipertensi masih lebih rendah karena masih terdapat kelas hipertensi yang diprediksi sebagai kelas nonhipertensi (*false negatif*).

Penggunaan software KNIME untuk prediksi penyakit hipertensi memberikan hasil yang robust, namun sederhana dalam penggunaannya serta tidak membutuhkan kemampuan *programming*. Oleh karena itu, KNIME sangat direkomendasikan bagi para pemula dalam melakukan analisis atau pembelajaran dalam penambangan data.

UCAPAN TERIMA KASIH

Terima kasih kepada Universitas Pembangunan Nasional Veteran Jakarta yang telah mendanai penelitian ini.

DAFTAR PUSTAKA

- [1] J. Lee, H. Davari, J. Singh, and V. Pandhare, "Industrial Artificial Intelligence for industry 4.0-based manufacturing systems," *Manuf. Lett.*, vol. 18, pp. 20–23, 2018.
- [2] D. Lafreniere, F. Zulkernine, D. Barber, and K. Martin, "Using machine learning to predict hypertension from a clinical dataset," in *2016 IEEE Symposium Series on Computational Intelligence, SSCI 2016*, 2017.
- [3] B. O. Afeni, T. I. Aruleba, and I. A. Oloyede, "Hypertension Prediction System Using Naive Bayes Classifier," *J. Adv. Math. Comput. Sci.*, pp. 1–11, Sep. 2017.
- [4] A. Wang, N. An, G. Chen, L. Li, and G. Alterovitz, "Predicting hypertension without measurement: A non-invasive, questionnaire-based approach," *Expert Syst. Appl.*, vol. 42, no. 21, pp. 7601–7609, 2015.
- [5] I. P. Adebayo, "Idowu Peter Adebayo. Predictive Model for the Classification of Hypertension Risk Using Decision Trees Algorithm," *Am. J. Math. Comput. Model.*, vol. 2, no. 2, pp. 48–59, 2017.
- [6] A. Naik and L. Samant, "Correlation Review of Classification Algorithm Using Data Mining Tool: WEKA, Rapidminer, Tanagra, Orange and Knime," *Procedia Comput. Sci.*, vol. 85, no. Cms, pp. 662–668, 2016.
- [7] S. Dwivedi, P. Kasliwal, and S. Soni, "Comprehensive study of data analytics tools (RapidMiner, Weka, R tool, Knime)," *2016 Symp. Colossal Data Anal. Networking, CDAN 2016*, 2016.
- [8] R. Samant and S. Rao, "Evaluation of Artificial Neural Networks in Prediction of Essential Hypertension," *Int. J. Comput. Appl.*, vol. 81, no. 12, pp. 34–38, 2013.
- [9] N. Azwanti, E. Elisa, U. P. Batam, and J. R. S. Kuning, "InfoTekJar : Jurnal Nasional Informatika dan Teknologi Jaringan Analisis Pola Penyakit Hipertensi Menggunakan Algoritma C4 . 5," vol. 2, 2019.
- [10] Sci-kit-Learn, "Cross-validation: evaluating estimator performance," 2015. [Online]. Available: http://scikit-learn.org/stable/modules/cross_validation.html.
- [11] N. Chamidah, M. M. Santoni, and N. Matondang, "Pengaruh Oversampling pada Klasifikasi Hipertensi dengan Algoritma," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 4, no. 4, pp. 635–641, 2021.