

Aplikasi *Web* Prediksi Dampak Gempa di Indonesia Menggunakan Metode *Decision Tree* dengan Algoritma C4.5

Diory Pribadi Sinaga^{1, a)}, Rini Marwati^{1, b)}, Bambang Avip Priatna Martadiputra^{1, c)}

¹Departemen Pendidikan Matematika, Universitas Pendidikan Indonesia, Jl. Dr. Setia Budhi No. 229, Bandung 40154, Indonesia

Email: ^adiorypribadisinaga@upi.edu, ^brini-marwati@upi.edu, ^cbambangavip@upi.edu

Abstract

An event or problem sometimes needs to be predicted to determine the impact caused. One of the events that need to be predicted is the impact of the earthquake. The Meteorology, Climatology and Geophysics Agency (BMKG) classifies earthquake impacts based on the BMKG Earthquake Intensity Scale (SIG-BMKG) which consists of 5 scales. In making predictions on a problem, you can use data mining that extracts data into useful information. Grouping the impact of an earthquake is one of the tasks of data mining, namely classification. Prediction can be viewed as a classification that groups data into predefined classes. One classification method is the Decision Tree. This method can handle both categorical and numerical data on large data. Some of the algorithm of the Decision Tree method are ID3, CART, and C4.5. The C4.5 algorithm is an improved ID3 algorithm so that it can handle missing values and continuous data. This study aims to construct a model and analyze the performance of the model obtained using the Decision Tree method with the C4.5 algorithm. In determining the best model, you can utilize Split Validation and k-fold Cross Validation. The best model was obtained in the first iteration of 10-fold Cross Validation. The best model is then used in a web application that can be used by the community to predict the impact of earthquakes that occur in Indonesia.

Keywords: Data Mining, Classification, Decision Tree, C4.5 Algorithm, Split Validation, k-fold Cross Validation

Abstrak

Suatu peristiwa atau masalah terkadang perlu dilakukan prediksi guna mengetahui dampak yang ditimbulkan. Salah satu peristiwa yang perlu dilakukan prediksi salah satunya adalah dampak gempa. Badan Meteorologi Klimatologi dan Geofisika (BMKG) mengelompokkan dampak gempa berdasarkan Skala Intensitas Gempabumi BMKG (SIG-BMKG) yang terdiri dari 5 skala. Dalam melakukan prediksi pada suatu masalah dapat memanfaatkan data *mining* yang mengekstraksi data menjadi sebuah informasi berguna. Pengelompokan dampak gempa merupakan salah satu tugas data *mining* yaitu klasifikasi. Prediksi dapat dipandang sebuah klasifikasi yang mengelompokkan data ke dalam kelas yang sudah ditentukan. Salah satu metode klasifikasi adalah *Decision Tree*. Metode ini dapat menangani data kategorik maupun numerik pada data yang berukuran besar. Beberapa algoritma metode *Decision Tree* adalah ID3, CART, dan C4.5. Algoritma C4.5 merupakan algoritma ID3 yang diperbaiki sehingga mampu menangani *missing value* dan data kontinu. Penelitian ini bertujuan mengonstruksi model dan menganalisis performa model yang diperoleh menggunakan metode *Decision Tree* dengan algoritma C4.5. Dalam penentuan model terbaik dapat memanfaatkan *Split*

Validation dan *k-fold Cross Validation*. Model terbaik diperoleh pada iterasi pertama *10-fold Cross Validation*. Model terbaik tersebut kemudian digunakan pada aplikasi *web* yang dapat dimanfaatkan oleh masyarakat untuk melakukan prediksi dampak gempa yang terjadi di Indonesia.

Kata Kunci: *Data Mining*, Klasifikasi, *Decision Tree*, Algoritma C4.5, *Split Validation*, *k-fold Cross Validation*

PENDAHULUAN

Pada tanggal 21 November 2022, Indonesia dikejutkan adanya peristiwa gempa yang terjadi di Kabupaten Cianjur, Provinsi Jawa Barat. Kekuatan gempa saat itu sebesar 5,6 Magnitudo yang berdampak banyaknya korban jiwa dan kerusakan baik ringan, sedang, maupun berat. BMKG menyebutkan melalui media Kompas gempa susulan yang terjadi pasca gempa tersebut adalah sebanyak 259 kali. Berdasarkan peristiwa tersebut, maka penting melakukan prediksi dampak gempa yang terjadi guna mengurangi risiko. BMKG mengelompokkan dampak gempa menjadi 5 skala yang disebut Skala Intensitas Gempa bumi BMKG (SIG-BMKG). SIG-BMKG merupakan skala *Modified Mercalli Intensity* (MMI) yang disederhanakan untuk menyesuaikan dengan tipikal bangunan di Indonesia. Dalam melakukan prediksi dampak gempa dapat memanfaatkan data *mining*.

Data *mining* merupakan proses menemukan pola, korelasi pada sejumlah besar data menggunakan teknik matematika dan statistika (Larose & Larose, 2014). Dengan menggunakan data *mining* sehingga dapat diperoleh informasi yang bisa dimanfaatkan pada data yang baru. Beberapa tugas dari data *mining* diantaranya adalah prediksi, estimasi, klasifikasi, dan pengklasteran. Prediksi dapat dipandang sebagai klasifikasi yang memetakan ke dalam kelas yang sudah ditentukan.

Klasifikasi merupakan pengelompokan data ke dalam kelas yang sudah ditentukan sedemikian sehingga setiap data memiliki kelasnya masing-masing. Beberapa metode klasifikasi diantaranya *Decision Tree*, *Naïve Bayes*, *Neural-Network*, *Rule-based Classifier*, dan *Support Machine*. Keunggulan metode *Decision Tree* mampu menangani data kategorik dan data numerik (Khadafy & Wahono, 2015). Metode ini juga termasuk metode klasifikasi yang mudah dipahami tanpa memiliki pemahaman matematika yang mendalam (Singh & Gupta, 2014). Beberapa algoritma metode *Decision Tree* adalah ID3, CART, dan C4.5. Penelitian ini menggunakan algoritma C4.5 dalam membangun pohon keputusan.

Algoritma C4.5 merupakan perbaikan algoritma ID3 yang sudah mampu menangani *missing value* dan data kontinu (Hssina et al., 2014). Berdasarkan penelitian-penelitian terdahulu seperti Mukminin & Riana (2017), Anam & Santoso (2018), dan Haekal et al. (2021) diperoleh bahwa dengan menggunakan algoritma C4.5 memiliki akurasi yang baik daripada algoritma lain atau metode klasifikasi lainnya.

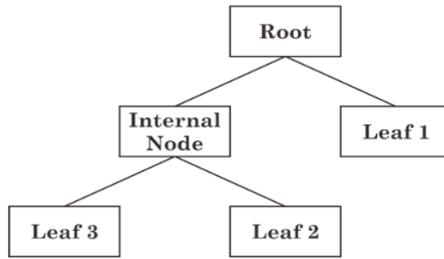
Dalam melakukan prediksi menggunakan metode *Decision Tree* dengan algoritma C4.5 perlu memilih model terbaik yang digunakan untuk melakukan prediksi pada kasus baru. Penelitian ini menggunakan *Split Validation* dan *k-fold Cross Validation* untuk menentukan model terbaik berdasarkan performa model. *Confusion Matrix* dapat digunakan untuk mengukur performa model. Model terbaik yang diperoleh digunakan dalam aplikasi *web* untuk melakukan prediksi dampak gempa.

METODE

Model *Decision Tree* C4.5

Metode *Decision Tree* adalah salah satu metode klasifikasi yang prinsipnya membuat struktur pohon seperti diagram alur di mana setiap *node* (*node non leaf*) mewakili atribut dan setiap cabang pada pohon mewakili hasil uji, serta *node leaf* menunjukkan kelas (Han et al., 2012). Contoh struktur pohon yang dibentuk dari metode *Decision Tree* seperti gambar di bawah. Penentuan atribut sebagai *node* (*node non*

leaf) didasarkan pada pemilihan atribut terbaik, di mana pemilihan atribut tergantung pada algoritma yang dipilih.



Gambar 1. Struktur Pohon Keputusan (Setio et al., 2020)

Metode *Decision Tree* memiliki beberapa algoritma salah satunya adalah algoritma C4.5. Algoritma C4.5 adalah algoritma ID3 yang diperbaiki sehingga dapat menangani *missing value* dan variabel kontinu dalam membuat pohon keputusan. Algoritma C4.5 menggunakan *Gain Ratio* sebagai pemilihan atribut terbaik. Misalkan data $D = \{t_1, t_2, t_3, \dots, t_n\}$, himpunan kelas $C = \{C_1, C_2, \dots, C_m\}$, himpunan atribut $A = \{A_1, A_2, \dots, A_h\}$. Berikut langkah-langkah membangun pohon keputusan menggunakan metode *Decision Tree* dengan algoritma C4.5.

- 1) Menghitung nilai *Entropy D*

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

di mana,

$$p_i = \frac{freq(C_i, D)}{|D|} \quad (2)$$

- 2) Menghitung nilai *Entropy* pada masing-masing atribut

$$info_A(D) = \sum_{j=1}^v \frac{|D_j|}{|D|} \times info(D_j) \quad (3)$$

- 3) Menghitung *Information Gain* setiap atribut

$$Gain(A) = info(D) - info_A(D) \quad (4)$$

- 4) Menghitung *Gain Ratio* setiap atribut

$$GainRatio(A) = \frac{Gain(A)}{SplitInfo(A)} \quad (5)$$

di mana,

$$SplitInfo(A) = - \sum_{j=1}^v \frac{|D_j|}{|D|} \times \log_2 \left(\frac{|D_j|}{|D|} \right) \quad (6)$$

Berdasarkan perhitungan *Gain Ratio* inilah kemudian memilih atribut terbaik menjadi *node* (*node non leaf*).

K-means Clustering

Penelitian ini mengategorikan data numerik menjadi tiga kategori untuk masing-masing atribut. Dalam mengategorikan dapat memanfaatkan *k-means Clustering*. *K-means Clustering* merupakan algoritma yang digunakan untuk mempartisi data menjadi *k* bagian di mana setiap objek dalam partisi (kelompok) memiliki kemiripan yang tinggi. Secara matematis algoritma *k-means Clustering* sebagai berikut.

Diberikan $x_1, x_2, x_3, \dots, x_n$ dengan masing-masing x_i memiliki dimensi d dan k yang menyatakan banyaknya partisi (*clustering*).

- 1) Pilih sembarang k objek sembarang sebagai *centroid*.
- 2) Setiap objek yang tersisa kemudian dikelompokkan ke dalam *cluster* (kelompok) yang paling dekat berdasarkan perhitungan jarak *Euclidean distance*.
- 3) Kemudian perbarui *centroid* setiap *cluster* yang diperoleh dari rata-rata *cluster*. Lakukan langkah sebanyak 2 – 3 kali, sampai tidak adanya perubahan pada setiap *cluster*.

Pada langkah pertama di atas adalah memilih sembarang k objek sebagai pusat *cluster* (*centroid*). Pemilihan k objek tersebut dapat memanfaatkan teknik *sampling*.

Sampling

Penelitian ini menggunakan *sampling* dalam mengategorikan data pada kasus prediksi dampak gempa di Indonesia. *Sampling* adalah suatu proses pengambilan sampel dari suatu populasi sehingga dapat mewakili populasi dalam menarik kesimpulan. Teknik *sampling* dibagi menjadi dua, yaitu *probability sampling* dan *nonprobability sampling*. Penelitian ini menggunakan *probability sampling* yang memberikan peluang yang sama pada setiap unsur populasi. *Probability sampling* diantaranya adalah *Simple Random Sampling*, *Systematic Random Sampling*, *Stratified Random Sampling*, dan *Cluster Sampling*. Penelitian ini menggunakan *Simple Random Sampling*, *Systematic Random Sampling* dalam pengambilan sampel.

Validasi model

Dalam memilih model yang digunakan pada prediksi perlu dilakukan validasi model. Validasi adalah mengetahui apakah model yang diperoleh representasi akurat pada keadaan nyata (Law & Kelton, 1999). Validasi memiliki dua teknik yaitu *Split Validation* dan *k-fold Cross Validation*. Berikut penjelasan kedua teknik tersebut.

a) *Split Validation*

Teknik validasi ini membagi data menjadi dua bagian yaitu data *training* dan data *testing*. Data *training* digunakan untuk memodelkan dan data *testing* digunakan untuk mengukur performa

model yang diperoleh. Pada penelitian-penelitian terdahulu kebanyakan menggunakan *split* 8:2 pada *Split Validation*.

b) *K-fold Cross Validation*

Berbeda dengan *Split Validation*, *k-fold Cross Validation* melakukan validasi berulang. Teknik ini membagi data menjadi *k* partisi sama besar. Setiap partisi akan mempunyai kesempatan menjadi data *testing* yang digunakan untuk mengukur performa model. Banyak model dan performa model yang diperoleh adalah sebanyak *k*. Dalam melakukan validasi *k-fold Cross Validation* direkomendasikan *k* = 10 (Han et al., 2012).

Confusion Matrix

Ketika melakukan validasi maka akan mengukur performa model pada data *testing*. Pengukur performa model dapat menggunakan bantuan *Confusion Matrix*. Tabel *Confusion matrix* dapat dilihat pada Tabel 1.

Tabel 1. *Confusion Matrix*

		Predicted class					
		Class	C_1	C_2	...	$C_{(m-1)}$	C_m
Actual Class	C_1	$e_{1,1}$	$e_{1,2}$...	$e_{(1,m-1)}$	$e_{(1,m)}$	
	C_2	$e_{2,1}$	$e_{2,2}$...	$e_{(2,m-1)}$	$e_{(2,m)}$	
	
	C_m	$e_{(m,1)}$	$e_{(m,2)}$...	$e_{(m,m-1)}$	$e_{(m,m)}$	

Tabel di atas menyatakan sebuah matriks berukuran $m \times m$, dengan m menyatakan banyak kelas. Setiap elemen $e_{i,j}$ menyatakan banyak $t_i \in D$ yang kelas sebenarnya C_i tetapi kelas prediksi adalah C_j . Terdapat istilah pada *Confusion Matrix* yaitu TP (*True Positive*), FP (*False Positive*), TN (*True Negative*), dan FN (*False Negative*) yang digunakan untuk mempermudah dalam mengukur performa model. Misalkan C_j menyatakan kelas makna istilah TP, FP, TN, dan FN sebagai berikut:

- 1) TP (*True Positive*) menyatakan banyaknya $t_i \in D$ yang kelas prediksi dan kelas sebenarnya adalah C_j .
- 2) FP (*False Positive*) menyatakan banyaknya $t_i \in D$ yang kelas prediksinya sama dengan C_j tetapi kelas sebenarnya tidak sama dengan kelas prediksinya.
- 3) TN (*True Negative*) menyatakan banyaknya $t_i \in D$ yang kelas prediksinya sama dengan kelas sebenarnya tetapi kelas tersebut bukan C_j .
- 4) FN (*False Negative*) menyatakan banyaknya $t_i \in D$ yang kelas prediksinya tidak sama dengan C_j dan kelas prediksi tidak sama dengan kelas sebenarnya.

Dalam mengukur performa model maka dibutuhkan metrik evaluasi yang dihitung menggunakan *Confusion Matrix*. Terdapat beberapa metrik evaluasi yang digunakan dalam penelitian ini yaitu *accuracy*, *precision*, *recall*. Berikut rumus dalam menghitung masing-masing metrik evaluasi.

1. *Accuracy*

Metrik evaluasi ini digunakan untuk mengukur ketepatan model dalam mengklasifikasikan data dengan benar.

$$accuracy = \frac{\sum_{i=1}^m e_{i,i}}{\sum_{i=1}^m \sum_{j=1}^m e_{i,j}} \quad (7)$$

2. *Precision*

Metrik evaluasi ini digunakan untuk mengukur seberapa baik model dalam menghindari FP (*False Positive*).

$$precision = \frac{\sum_{i=1}^m precision_i}{m} \quad (8)$$

di mana, $precision_i$ menyatakan $precision_i$ pada kelas C_i

$$precision_i = \frac{e_{i,i}}{\sum_{j=1}^m e_{j,i}} = \frac{TP_i}{TP_i + FP_i} \quad (9)$$

3. *Recall*

Metrik evaluasi ini digunakan untuk mengukur seberapa baik model dalam menghindari FN (*False Negative*).

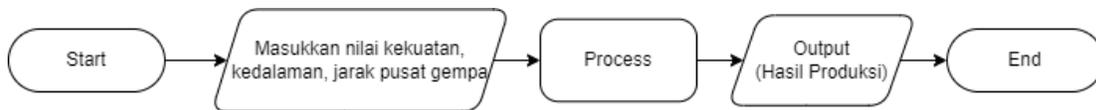
$$recall = \frac{\sum_{i=1}^m recall_i}{m} \quad (10)$$

di mana, $recall_i$ menyatakan $recall_i$ pada kelas C_i

$$recall_i = \frac{e_{i,i}}{\sum_{j=1}^m e_{i,j}} = \frac{TP_i}{TP_i + FN_i} \quad (11)$$

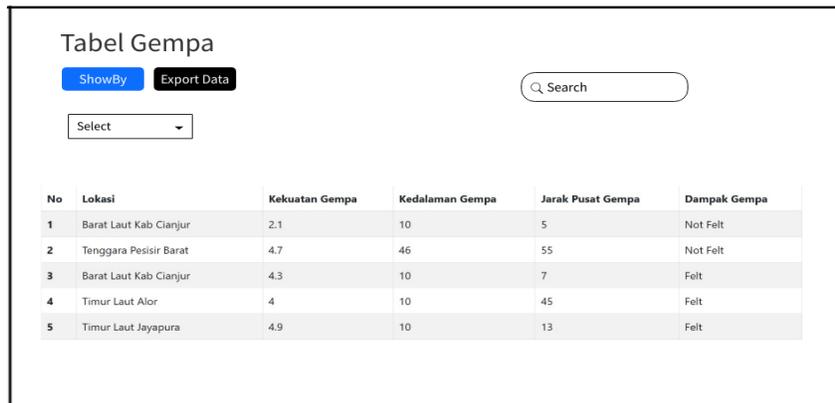
Konstruksi Aplikasi

Model terbaik yang diperoleh digunakan untuk memprediksi dampak gempa di Indonesia. Model terbaik dibungkus menggunakan *web* yang dapat digunakan oleh pengguna. *Web* ini dapat melakukan prediksi, melihat statistik data, dan perbandingan performa model. Pada penelitian ini menggunakan bahasa pemrograman *Javascript* dengan *NodeJs* sebagai *runtime* dalam menjalankan *Javascript* di sisi server. Dalam proses mempermudah membangun *web* penelitian ini menggunakan *ExpressJs* dalam membangun *RESTful API*. Berikut alur penggunaan aplikasi *web* prediksi dampak gempa.



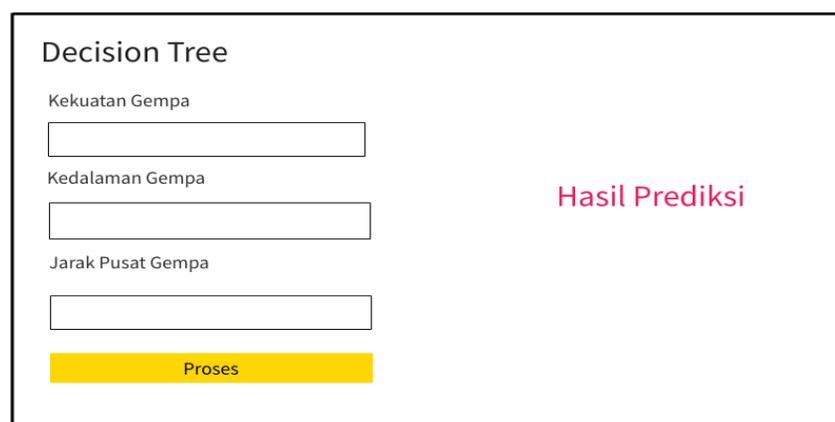
Gambar 2. Alur Aplikasi Prediksi Dampak Gempa

Adapun rancangan tampilan *web* sebagai berikut.



No	Lokasi	Kekuatan Gempa	Kedalaman Gempa	Jarak Pusat Gempa	Dampak Gempa
1	Barat Laut Kab Cianjur	2.1	10	5	Not Felt
2	Tenggara Pesisir Barat	4.7	46	55	Not Felt
3	Barat Laut Kab Cianjur	4.3	10	7	Felt
4	Timur Laut Alor	4	10	45	Felt
5	Timur Laut Jayapura	4.9	10	13	Felt

Gambar 3. Rancangan Tampilan Halaman Utama



Decision Tree

Kekuatan Gempa

Kedalaman Gempa

Jarak Pusat Gempa

Proses

Hasil Prediksi

Gambar 4. Rancangan Tampilan Halaman Prediksi

HASIL DAN PEMBAHASAN

Penelitian ini menggunakan proses pengembangan data *mining* yaitu CRISP-DM (*Cross-Industry Standard Process for Data Mining*) dalam membangun model hingga tahap penyebaran model. Adapun langkah-langkah penelitian adalah sebagai berikut:

1. Pemahaman Bisnis

Pada tahapan ini merupakan berhubungan masalah dan tujuan yang diselesaikan. Seperti yang dibahas pada pendahuluan, penelitian ini bertujuan mengonstruksi model klasifikasi untuk menyelesaikan permasalahan prediksi dampak gempa yang terjadi di Indonesia.

2. Pemahaman Data

Pada tahapan ini berkaitan dengan pengumpulan data yang digunakan dalam penelitian. Penelitian ini menggunakan data yang diperoleh dari BMKG diambil dari tanggal 24 Januari 2023 sampai dengan 30 Maret 2023. Sumber data yang digunakan dapat diakses melalui tautan berikut <https://www.bmkg.go.id/gempabumi-dirasakan.html>. Berdasarkan data yang diambil diperoleh 152

gempa dengan 91 data gempa menunjukkan kelas *Felt* (dirasakan) dan 61 data gempa menunjukkan kelas *Not Felt*.

3. Persiapan Data

Tahapan ini berhubungan dengan mempersiapkan data sehingga siap untuk dimodelkan dan memilih variabel yang digunakan. Variabel yang digunakan sebagai atribut adalah kekuatan gempa (Mag), kedalaman gempa (km), dan jarak pusat gempa (km) serta variabel target adalah dampak gempa berdasarkan SIG-BMKG. Membersihkan data dari *missing value* atau memformat data sesuai kebutuhan termasuk pada tahapan ini. Setiap atribut dikategorikan menjadi 3 diantaranya kekuatan gempa terdiri dari lemah, sedang, dan kuat. Kedalaman gempa terdiri dari dangkal, sedang, dan dalam. Jarak pusat gempa terdiri dari dekat, sedang, dan jauh. Pengkategorian data penelitian ini memanfaatkan *k-means Clustering*. Berdasarkan algoritma *k-means Clustering* diperoleh untuk kekuatan gempa terdiri dari 59 kategori lemah, 63 kategori sedang, dan 30 kategori kuat. Kedalaman gempa terdiri dari 133 kategori dangkal, 11 kategori sedang, dan 8 kategori dalam. Jarak pusat gempa terdiri dari 113 kategori dekat, 26 kategori sedang, dan 13 kategori jauh.

4. Pemodelan

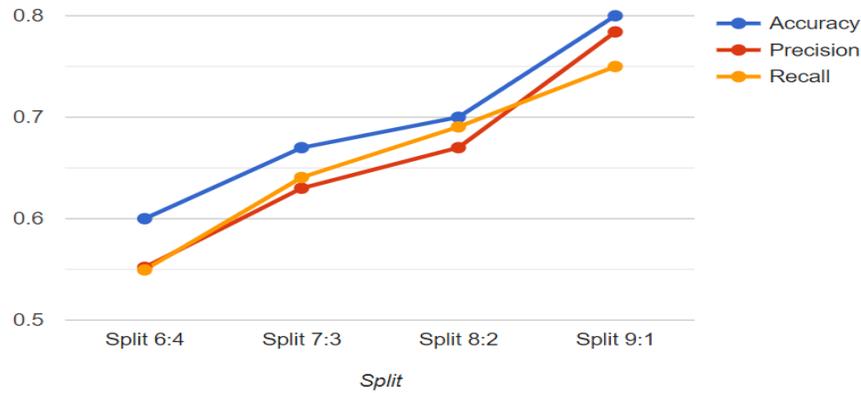
Tahapan ini berkaitan dengan memodelkan data yang sudah dipersiapkan sebelumnya. Berdasarkan langkah-langkah menggunakan metode *Decision Tree* dengan algoritma C4.5 diperoleh model untuk masing-masing teknik validasi. Selanjutnya performa masing-masing model diukur menggunakan metrik evaluasi pada tahap evaluasi.

5. Evaluasi

Pada tahapan ini bertujuan untuk mengetahui performa model yang diperoleh untuk kemudian dipakai dalam tahap penyebaran model. Berikut perbandingan performa masing-masing teknik validasi.

Tabel 2. Performa Model *Split Validation*

<i>Split</i>	<i>Accuracy</i>	<i>Precision</i>	<i>Recall</i>
6:4	60 %	55.20 %	54.95 %
7:3	66.67 %	62.82 %	64.06 %
8:2	70 %	66.67 %	69.05 %
9:1	80 %	78.41 %	75 %



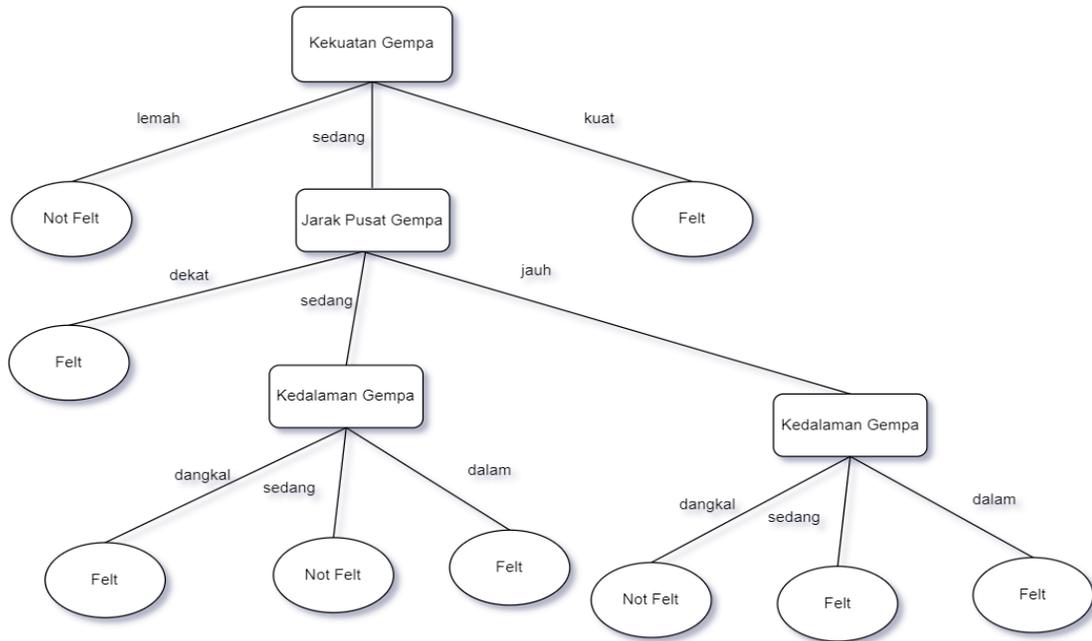
Gambar 5. Perbandingan Performa *Split Validation*

Berdasarkan Tabel 2 diperoleh model terbaik pada *Split* 9:1 dengan perolehan metrik evaluasi *accuracy* = 80%, *precision* = 78.41% dan *recall* = 75%. Adapun untuk *10-fold Cross Validation* sebagai berikut

Tabel 3. Performa Model *10-fold Cross Validation*

Iterasi	Accuracy	Precision	Recall
1	80 %	80.36 %	80.36 %
2	73.33 %	74.11 %	75 %
3	73.33 %	75 %	72.32 %
4	80 %	75 %	78.41 %
5	40 %	40 %	41.07 %
6	73.33 %	77.78 %	80 %
7	66.67 %	69.44 %	69.44 %
8	66.67 %	60 %	61.36 %
9	60 %	61.61 %	64.77 %
10	80 %	77.78 %	80 %

Berdasarkan Tabel 3 diperoleh model terbaik pada iterasi pertama dengan *accuracy* = 80 %, *precision* = 80.36 % dan *recall* = 80.36 %. Model ini ternyata lebih baik dibandingkan dengan model terbaik pada *Split Validation*, sehingga model pada iterasi pertama ini yang dipakai pada tahap penyebaran. Gambar 6 merupakan pohon keputusan dari model terbaik yang diperoleh.



Gambar 6. Pohon Keputusan Model Terbaik

Berdasarkan Gambar 6 diperoleh bahwa atribut kekuatan gempa menjadi faktor yang mempengaruhi dampak gempa yang terjadi di Indonesia.

6. Penyebaran

Tahap terakhir pada proses pengembangan data *mining* adalah penyebaran. Penyebaran dalam hal ini adalah model yang siap digunakan dalam mengatasi permasalahan yang sudah dijelaskan sebelumnya. Model terbaik yang diperoleh dibungkus pada sebuah *website* sehingga pengguna dengan mudah menggunakannya. Berikut hasil tampilan *website* prediksi dampak gempa (sebagian).

Tabel Gempa

Show By [Export data](#)

Show

No	Lokasi	Kekuatan Gempa	Kedalaman Gempa	Jarak Pusat Gempa	Dampak Gempa
1	Barat Laut Kab Cianjur	2.1	10	5	Not Felt
2	Tenggara Pesisir Barat	4.7	46	55	Not Felt
3	Barat Laut Kab Cianjur	4.3	10	7	Felt
4	Timur Laut Alor	4	10	45	Felt

Gambar 7. Tampilan Halaman Utama

Prediksi

KNN(K-Nearest Neighbors) Naive Bayes

Decision Tree

Kekuatan Gempa

Kedalaman Gempa

Jarak Pusat Gempa

Algoritma

Hasil Prediksi

Felt

Gambar 8. Tampilan Halaman Prediksi

Pada Gambar 7 menunjukkan tampilan halaman utama yang menampilkan data gempa yang diperoleh. Pada Gambar 8 merupakan tampilan halaman prediksi yang dapat digunakan pengguna dalam memprediksi dampak gempa yang terjadi. Sebagai contoh seperti pada Gambar 8 ketika pengguna menginputkan kekuatan gempa sebesar 4.3 Mag, dengan kedalaman gempa 2 km dan jarak pusat gempa 5 km diperoleh dampak gempa *Felt* (dirasakan). Hasil *website* selengkapnya dapat dilihat dengan mengunjungi tautan <https://decisiontreec45-production.up.railway.app>.

KESIMPULAN

Metode *Decision Tree* dengan algoritma C4.5 dapat digunakan untuk melakukan prediksi dampak gempa di Indonesia berdasarkan SIG-BMKG. Algoritma *k-means Clustering* dapat dimanfaatkan untuk mengategorikan masing-masing atribut. Pengambilan acak *k* objek pada langkah pertama *k-means Clustering* dapat menggunakan *Simple Random Sampling* dan *Systematic Random Sampling*. Berdasarkan hasil performa model pada masing-masing teknik validasi, diperoleh model terbaik yaitu pada iterasi pertama *10-fold Cross Validation*. Perolehan metrik evaluasi model terbaik diantaranya *accuracy* = 80 %, *precision* = 80.36 % dan *recall* = 80.36 % dengan kekuatan gempa sebagai atribut yang memiliki faktor paling berpengaruh dalam memprediksi dampak gempa. Model terbaik disebarkan dengan dibungkus *website* sehingga pengguna dapat menggunakan dengan mudah dan dapat diakses kapan saja.

REFERENSI

- D. T. Larose and C. D. Larose, *Discovering knowledge in data: an introduction to data mining*, vol. 4. John Wiley & Sons, 2014.
- A. R. Khadafy and R. S. Wahono, 'Penerapan Naive Bayes Untuk Mengurangi Data Noise Pada Klasifikasi Multi Kelas Dengan Decision Tree', *Journal of Intelligent Systems*, vol. 1, no. 2, 2015.
- P. Singh, S. and Gupta, 'Comparative study ID3, cart and C4. 5 Decision tree algorithm: a survey', *International Journal of Advanced Information Science and Technology (IJAIST)*, vol. 27, no. 27, 2014.
- B. HSSINA, A. MERBOUHA, H. EZZIKOURI, and M. ERRITALI, 'A comparative study of decision tree ID3 and C4.5', *International Journal of Advanced Computer Science and Applications*, vol. 4, no. 2, 2014, doi: 10.14569/specialissue.2014.040203.

- A. Mukminin and D. Riana, 'Komparasi Algoritma C4 . 5 , Naïve Bayes Dan Neural Network Untuk Klasifikasi Tanah', *Jurnal Informatika Universitas Bina Sarana Informatika*, vol. 4, no. 1, 2017.
- C. Anam and H. B. Santoso, 'Perbandingan Kinerja Algoritma C4.5 dan Naive Bayes untuk Klasifikasi Penerima Beasiswa', *Energy - Jurnal Ilmiah Ilmu-Ilmu Teknik*, vol. 8, no. 1, 2018.
- B. V. Haekal, I. Ernawati, and N. Chamidah, 'Klasifikasi Kepuasan Pengguna Layanan Aplikasi Shopee Menggunakan Metode Decision Tree C4.5', *Informatik : Jurnal Ilmu Komputer*, vol. 17, no. 3, 2021, doi: 10.52958/iftk.v17i3.3648.
- J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. 2012. doi: 10.1016/C2009-0-61819-5.
- P. B. N. Setio, D. R. S. Saputro, and Bowo Winarno, 'Klasifikasi Dengan Pohon Keputusan Berbasis Algoritme C4.5', *PRISMA, Prosiding Seminar Nasional Matematika*, vol. 3, 2020.
- A. Law and D. Kelton, *Simulation Modeling and Analysis*. McGraw Hill, 1999.