PENERAPAN ALGORITMA DECISION TREE C.45 UNTUK KLASIFIKASI DATA STATUS HUNI RUMAH REHABILITASI PASCA ERUPSI MERAPI

APPLICATION OF C.45 DECISION TREE ALGORITHM FOR REHABILITATION HOUSEHOLD DATA CLASSIFICATION POST ERUPTION OF MERAPI

Nurhadi Wijaya[1*], Marselina Endah[2], Mujatia Feliati[3]

[1, 2, 3] Program Studi Informatika Program Sarjana, Fakultas Sains dan Teknologi, Universitas
[1]nurhadi@respati.ac.id,[2]marselina.endah@respati.ac.id,[3]mujastia@gmail.com
***Penulis Korespondensi**

Abstrak

Erupsi gunung Merapi berikut lahar hujan di Tahun 2010 berdampak pada kerusakan infrastruktur berikut ribuan hunian rumah di Kabupaten Sleman D.I.Yogyakarta dan Kabupaten Magelang Jawa Tengah. Melalui Peraturan Kepala BNPB No.5 Tahun 2011, rehabilitasi dan rekonstruksi perumahan yang terdampak erupsi Merapi, dilakukan dengan skema program Rehabilitasi dan Rekonstruksi Masyarakat dan Permukiman Berbasis Komunitas. Skema tersebut telah membangun rumah hunian sebanyak 2.516-unit. Berdasarkan Key Performance Indikator (KPI) oleh The World Bank, status huni rumah merupakan indikator keberhasilan kinerja skema program ini. Pelaksanaan progam rehabilitasi rumah pasca erupsi Merapi didokumentasikan dan terekam ke dalam basis data. Dibidang data mining, basis data merupakan aset yang dapat digunakan sebagai bahan pengenalan dan penemuan pola-pola data yang dapat dipelajari dan diteliti guna menyelesaikan permasalahan baik pengelompokan data maupun klasifikasi data. Pada penelitian ini dilakukan penerapan algoritma decision tree C.45 untuk mengklasifikasi data status huni rumah rehabilitasi pasca erupsi gunung Merapi. Hasil klasifikasi penelitian diperoleh angka nilai tingkat akurasi klasifikasi mencapai 91.34%, dengan demikian terjawab bahwa algoritma decision tree C.45 dapat diterapkan untuk mengklasifikasi data status huni rumah rehabilitasi pasca erupsi gunung Merapi.

Kata kunci : Data mining, Decision Tree C.45, Klasifikasi, Rehabilitasi, Status huni, Merapi

Abstract

The eruption of Mount Merapi and lava rain in 2010 had an impact on infrastructure damage as well as thousands of residential houses in Sleman D.I.Yogyakarta and Magelang, Central Java. Through the Regulation of the Head of BNPB No.5 of 2011, the rehabilitation and reconstruction of housing affected by the eruption of Merapi is carried out with the Community-Based Settlement and Community Rehabilitation and Reconstruction program scheme. The scheme has built 2,516 residential houses. Based on the Key Performance Indicator (KPI) by The World Bank, the occupancy status of the house is an indicator of the successful performance of this scheme. The implementation of the post-Merapi home rehabilitation program was documented and recorded in a database. In the field of data mining, databases are assets that can be used as a material for recognition and discovery of data patterns that can be studied and researched in order to solve problems both data grouping and data classification. In this study, the application of the C.45 decision tree algorithm was carried out to classify the occupancy status data of post-Merapi volcano rehabilitation houses. The results of the research classification showed that the value of the classification accuracy rate reached 91.34%, thus it is answered that the C.45 decision tree algorithm can be applied to classify the occupancy status data of post-Merapi volcano rehabilitation houses.

Keywords: Data mining, Decision Tree C.45, Classification, Rehabilitation, Habitat Status, Merapi

1. PENDAHULUAN

Letusan erupsi gunung Merapi ditahun 2010 mengakibatkan kerusakan Rumah dan infrastruktur di wilayah Propinsi Jawa Tengah dan Propinsi D.I.Yogyakarta. Tercatat Sebanyak 2.682 rumah di wilayah Kabupaten Sleman mengalami rusak berat, dan di Provinsi Jawa Tengah 174 rumah rusak berat. Sedangkan susulan bencana aliran air hujan atau lahar hujan gunung Merapi telah mengakibatkan sebanyak 2.856 unit rumah terdampak primer dan 1.087 unit terdampak sekunder[1]. Melalui Peraturan Kepala BNPB No 5, tahun 2011 ditetapkan kegiatan rehabilitasi sektor perumahan pasca erupsi Merapi menggunakan skema Rehabilitasi dan Rekonstruksi Masyarakat dan Permukiman berbasis Masyarakat [1]. Melalui skema ini telah terehabilitasi rumah sebanyak 2.516-unit hunian rumah [9].

Berdasarkan *Key Performance Indicators* yang dikeluarkan oleh The World Bank, status huni hunian rumah teatau disebut terehabilitasi atau disebut dengan hunian tetap (huntap) merupakan salah satu indikator keberhasilan skema program rehabilitasi pasca erupsi merapi. Semakin banyak rumah dihuni, maka indikator keberhasilan kinerja semakin baik [2].

Satuan kerja rehabilitasi dan rekonstruksi (Satker Rehabrekon) sebagai pelaksana kegiatan rehabilitasi dan rekonstruksi rumah pasca bencana erupsi gunung Merapi, telah mendokumentasikan data rehabilitasi rumah kedalam database/basis data. Satker telah merecord status huni, namun masih terdapat status huni yang belum jelas apakah sudah dihuni atau belum dihuni.

Data Mining adalah sebuah proses yang bertujuan untuk menemukan pola dari –data-data yang terdapat di dalam basis data. Penemuan pola ini dapat dimanfaatkan untuk pengelompokan objek atau data berdasarkan kemiripan data. Teknik klasifikasi dalam data mining merupakan teknik pembelajaran yang digunakan untuk dapat memprediksi nilai dari atribut kategori target. Klasifikasi bertujuan untuk membagi objek yang ditugaskan hanya ke salah satu nomor kategori yang disebut *class*/kelas[3].

Penelitian terkait klasifikasi status data huni rumah rehabilitasi pasca bencana erupsi gunung Merapi Tahun 2010 telah dilakukan. Klasifikasi data status huni pasca erupsi Merapi dicapai akurasi sebesar 89,59% dengan menggunakan Algoritma Naive Bayes [12]. Berdasarkan hasil penelitian tersebut, memunculkan pertanyaan apakah metode atau algoritma yang digunakan dari penelitian sebelumnya dapat diganti atau diubah dengan dengan algoritma lain seperti algoritma decision tree C.45?. Apakah perubahan model berikut metode algoritma yang digunakan akan berpengaruh dalam meningkatkan akurasi klasifikasi?

2. DASAR TEORI /MATERIAL DAN METODOLOGI/PERANCANGAN

2.1 Data Mining

Data mining merupakan suatu proses yang memiliki tujuan untuk menemukan pola dari data yang sudah ada di dalam basis data agar dapat dimanfaatkan untuk menyelesaikan masalah, mengelompokkan objek atau data berdasarkan kemiripan data, sehingga anggota dalam kelompok memiliki banyak kemiripan dibandingkan dengan kelompok lain. Teknik klasifikasi dalam data mining merupakan teknik pembelajaran yang digunakan untuk dapat memprediksi nilai dari atribut kategori target. Klasifikasi bertujuan untuk membagi objek yang ditugaskan hanya ke salah satu nomor kategori yang disebut kelas[3]. Dijelaskan oleh [7] data mining adalah sebuah aplikasi untuk memberikan keunggulan kompetitif untuk mencapai keputusan yang tepat.

2.2 Klasifikasi

Menurut [4], klasifikasi adalah proses penemuan model (atau fungsi) yang menggambarkan dan membedakan kelas data atau konsep yang bertujuan agar bisa digunakan untuk memprediksi kelas dari objek yang label kelasnya tidak diketahui. Terdapat dua langkah tahapan proses klasifikasi. Tahap pertama adalah learning (*fase training*), dimana algoritma klasifikasi dibuat agar dapat menganalisa data training lalu direpresentasikan ke dalam bentuk rule klasifikasi. Tahap kedua adalah klasifikasi, dimana data tes digunakan untuk memperkirakan akurasi dari *rule* klasifikasi. Metode yang dapat digunakan untuk mengukur akurasi algoritma klasifikasi, antara lain adalah: *cross validation, confusion matrix*, dan kurva ROC (*Receiver Operating Characteristic*).

2.3 Algoritma Klasifikasi Decision Tree C.45

Decision Tree adalah flow-chart seperti struktur tree, dimana tiap internal node menunjukkan sebuah test pada sebuah atribut, tiap cabang menunjukkan hasil dari test dan leaf node menunjukkan class-class atau class distribution [10]. Algoritma C4.5 merupakan kelompok algoritma Decision Tree. Algoritma ini mempunyai input berupa training samples dan samples. Training samples berupa data contoh yang akan digunakan untuk membangun sebuah tree yang telah diuji kebenarannya. Sedangkan samples merupakan field-field data yang nantinya akan digunakan sebagai parameter dalam melakukan klasifikasi data [10]. Ada tiga prinsip kerja algoritma C4.5 pada tahap belajar dari data, yaitu : Pembuatan Pohon Keputusan; pemangkasan Pohon Keputusan dan Evaluasi (Opsional) dan Pembuatan Aturan-Aturan dari Pohon Keputusan (Opsional). Algoritma C4.5 dapat menangani data numerik dan diskret. Algoritma C4.5 menggunakan rasio perolehan (gain ratio). Sebelum menghitung rasio perolehan, perlu dilakukan perhitungan nilai informasi dalam satuan bits dari suatu kumpulan objek, yaitu dengan menggunakan konsep entropi. Langkah – langkah dalam membuat sebuah decision tree dengan algoritma C4.5 adalah sebagai berikut [11]:

1) Menyiapkan data training. Data training ini diambil dari data yang pernah terjadi sebelumnya atau disebut dengan data masa lalu yang telah dikelompokan ke dalam kelas-kelas tertentu.
2) Menghitung akar dari pohon. Akar akan diambil dari atribut yang akan dipilih, dengan cara menghitung nilai gain dari masing-masing atribut, nilai gain yang paling tinggi yang akan menjadi akar pertama. Sebelum menghitung nilai gain dari atribut, hitung dahulu nilai entropy. Perhitungan Nilai entropy diperoleh dengan persamaan berikut :

$$Entropy(S) = \sum_{i=1}^{n} - P_i * \log_2 P_i \quad (1)$$

Keterangan :
S = Himpunan kasus
N =Jumlah Partisi
P_i= Proporsi S_i terhadap S

3) Menghitung nilai *Gain* dengan menggunakan persamaan berikut :

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{|S_i|}{S} * Entropy(S_i) \quad (2)$$

Keterangan :
S = Himpunan kasus
A = fitur
N = Jumlah Partisi atribut A
$|S_i|$ = Himpunan kasus
$|S|$ = Jumlah kasus dalam S

4) Ulangi langkah kedua dan langkah ketiga sampai dengan semua *record* terpatisi.
5) Proses partisi *decision tree* akan berhenti apabila :
   a) Semua *record* dalam simpul *N* mendapat kelas yang sama.
   b) Tidak ada atribut di dalam *record* yang dipartisi lagi.
   c) Tidak ada *record* di dalam cabang yang kosong.

2.4 *Confusion Matrix*

Confusion matrix merupakan salah satu alat ukur berbentuk matrik 2 x 2 yang digunakan untuk mendapatkan jumlah ketepatan klasifikasi dataset terhadap kelas aktif dan tidak aktif pada algoritma yang dipakai. Evaluasi model klasifikasi dilandaskan pada pengujian untuk memperkirakan obyek yang benar dan salah. Urutan pengujian ditabulasikan dalam *confusion matrix* dimana *class* yang diprediksi ditampilkan dibagian atas matriks dan *class* yang diamati disisi kiri. Setiap sel memiliki isi angka yang menunjukkan berapa banyak kasus yang sebenarnya dari kelas yang diamati untuk diprediksi [8]. Berikut ini merupakan contoh tabel *confusion matrix*

Tabel 1 Confusion Matrix

Classification	***Predicted Class***		
Oberved Class		***Class = Yes***	***Class = No***
	Class = Yes	*a* *(True Positive-TP)*	*b* *(False Negative-FN)*
	Class = No	*c* *(False Positive-FP)*	*d* *(True Negative-TN)*

Keterangan:
TP = Prediksi positif yang positif FN = Prediksi positif yang negatif
FP = Prediksi negatif yang positif TN = Prediksi negatif yang negatif

Untuk pengujian nilai dari hasil akurasi klasifikasi pada penelitian ini akan dihitung nilai akrasi hasil klasifikasi menggunakan persamaan sebagai berikut:

$$Accuracy = \left(\frac{a+d}{a+b+c+d}\right) = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

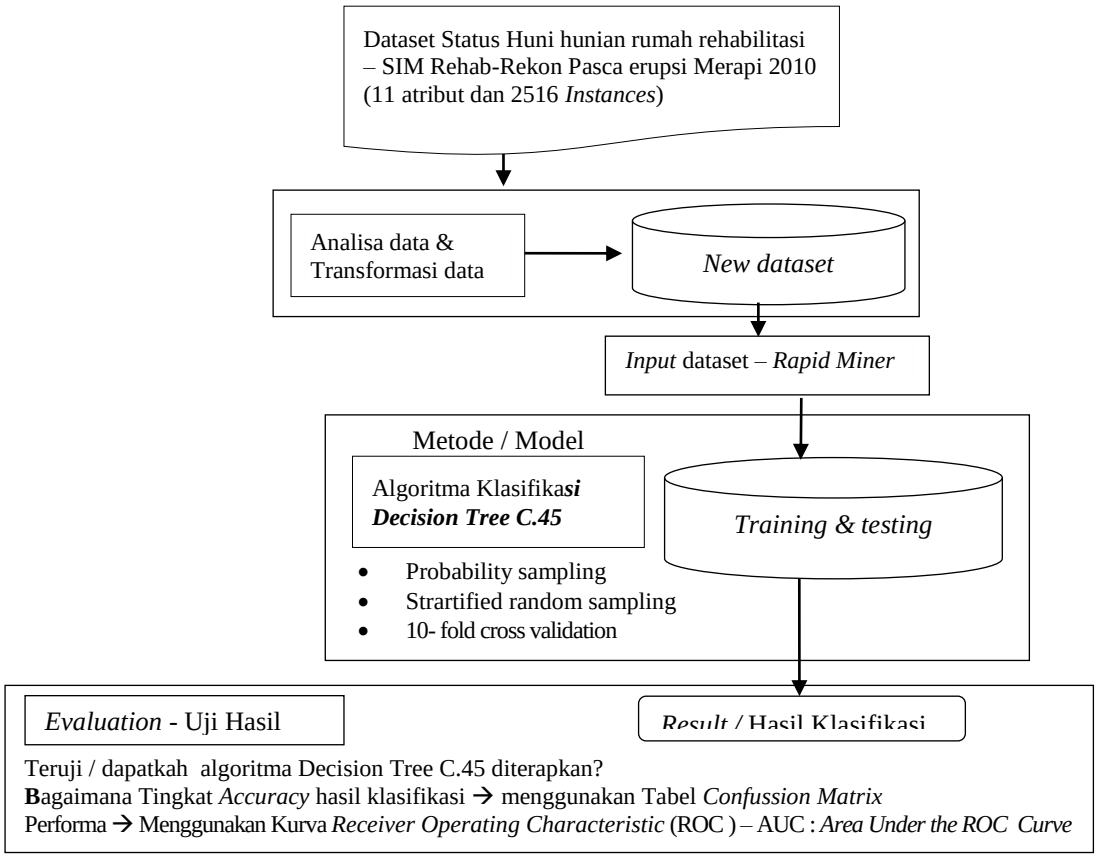
2.5 ROC (*Receiver Operating Characteristic*) *Curve*

Kurva ROC terdapat dalam dua dimensi, dimana tingkat TP diplot pada sumbu Y dan tingkat FP diplot pada sumbu X. Namun untuk merepresentasikan grafis yang menentukan klasifikasi mana yang lebih baik, digunakan metode yang mengbitung luas daerah di bawah kurva ROC yang disebut AUC (*Area Under the ROC curve*) yang diartikan sebagai probabilitas [7]. AUC mengukur kinerja diskriminatif dengan memperkirakan probabilitas *output* dari sampel yang dipliih secara acak dari populasi positif atau negatif, semakin besar AUC, semakin kuat klasifikasi yang digunakan. Karena AUC adalah bagian dari daerah unit persegi nilainya akan selalu antara 0,0 dan 1,0.

2.6 *Cross Validation*

Cross Validation merupakan teknik validasi data dengan cara membagi data secara acak ke dalam *k* bagian dan masing-masing bagian akan dilakukan proses klasifikasi. *Cross validation* akan melakukan percobaan sebanyak *k*. Secara umum pengujian nilai *k* dilakukan sebanyak 10 kali untuk memperkirakan akurasi estimasi [7]. Pada penelitian ini validasi data digunakan *Cross Validation*.

2.7 Metodologi/Rancangan skema penelitian



Gambar 1. *Flowchart* skema penelitian

3. PEMBAHASAN

Seperti terlihat pada gambar 1. *flowchart* skema penelitian, dataset yang digunakan pada penelitian ini adalah sebanyak 2516 *record/instances* dan 11 attribut, validasi dataset digunakan validasi data *10-fold cross validation*, dimana *10-fold cross validation* memiliki nilai bias varian relatif rendah. Pada *10-fold cross validation*, data dibagi menjadi 10 bagian terlebih dahulu secara acak dengan perbandingan yang sama, lalu hitung *error rate* setiap bagian, setelah itu diperoleh error rate secara keseluruhan dari menghitung rata-rata *error rate* dari semua bagian data [4]. Untuk menguji data dengan record sebanyak 2516 dengan 11 atribut digunakan *tools* pengolah data (*framework*) *RapidMiner*. Hasil akurasi dan performa klasifikasi dapat ditunjukkan ke dalam tabel 2 berikut:

Tabel 2. Hasil akurasi dan performa algoritma klasifikasi *Decision Tree* C.45 status huni rumah rehabilitasi pasca erupsi Merapi

Nilai Akurasi	Performa ROC - AUC
91,34%	0,739

Berikutnya hasil nilai akurasi klasifikasi pada tabel 2 diuji menggunakan tabel *confusion matrix*. Hasil tabel *confusion matrix* ditunjukkan pada tabel 3 berikut:

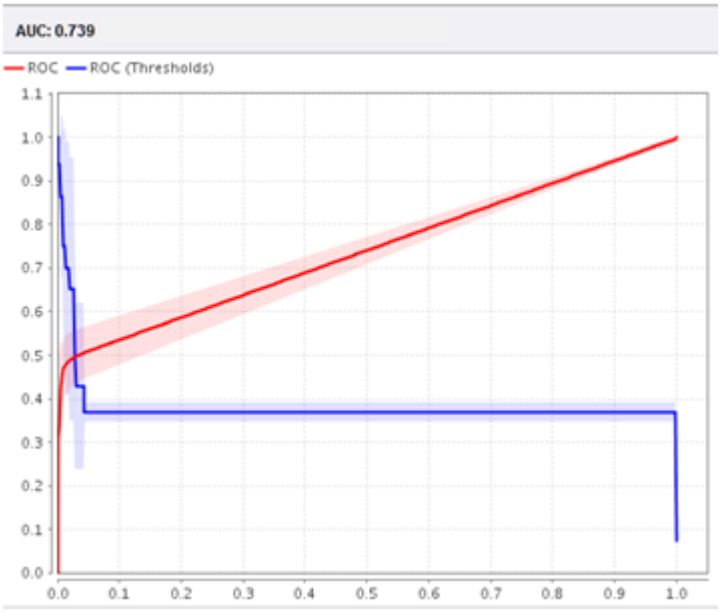
Tabel 3. *Confusion matrix table* klasifikasi data status huni rumah rehabilitasi pasca erupsi Merapi dengan Algoritma *Decision Tree* C.45

	***True* Sudah**	***True* Belum**
Pred.Sudah	*a*	*b*
	2140	212
Pred.Belum	*c*	*D*
	6	158

Selanjutnya dilakukan pengujian dengan menghitung nilai akurasi berdasarkan persamaan (3). Hasil perhitungan :

$$Akurasi = \left(\frac{2140+158}{2053+212+6+158} \right) = \left(\frac{2298}{2516} \right) x100\% = 91,34\%$$

Pada hasil perhitungan akurasi secara manual yang dapat terlihat di atas, perhitungan persentase tingkat akurasi klasifikasi dengan algoritma *Decision Tree* C.45 dicapai nilai sebesar 91,34%. Pengujian performa hasil klasifikasi Algoritma Naive Bayes dilakukan dengan menggunakan *tools Rapidmine*r dengan cara melihat hasil keluaran berupa kurva *Receiver Operating Characteristic curve* (ROC) dengan *Area Under the ROC Curve* (AUC) yang dapat ditunjukkan pada visualisasi gambar 2 berikut:



Gambar 2. Kurva ROC – AUC klasifikasi data status huni rumah rehabilitasi pasca erupsi Merapi dengan algoritma *Decision Tree C.45*

Nilai Performa AUC mencapai 0739. Hasil performa klasifikasi dengan nilai 0.739 tergolong ke dalam klasifikasi berperforma *fair classification*. Nilai di antara 0.70 sampai dengan 0.80 = klasifikasi cukup baik atau *fair classification* [7].

4. KESIMPULAN

Algoritma *decision tree* C.45 dapat diterapkan untuk mengklasifikasikan data status huni rumah rehabilitasi pasca erupsi Gunung Merapi. Penerapan hasil klasifikasi data status huni hunian rumah rehabilitasi pasca erupsi Gunung Merapi menggunakan Algoritma *decision tree* C.45 ini diperoleh nilai akurasi mencapai angka 91.34% dan performa klasifikasi diperoleh nilai AUC (*Area Under the ROC curve*) sebesar 0.739. Berdasarkan hasil nilai akurasi klasifikasi berikut nilai performa klasifikasi, maka Algoritma *decision tree* C.45 dapat diterapkan untuk

data status huni rumah rehabilitasi pasca erupsi Gunung Merapi dengan kategori cukup baik (*fair classification*).

DAFTAR PUSTAKA

[1] Bekti, S., 2013. *Pendampingan yang mencerahkan*. Cetakan Pertama. Kementrian Pekerjaan Umum, Direktorat Jenderal Cipta Karya Republik Indonesia

[2] KPI,.2014. *Key Performance Indicators* : *Community-based Settlements Rehabilitation and Reconstruction Project* (Rekompak), KPI Rekompak JRF-PSF Status November 2014. tersedia di http://www.rekompakciptakarya.org/KPI [diakses : 18 Mei 2016].

[3] Bramer, M., 2011. *Principles of data mining*, London, Springer

[4] Han J., dan Kamber M., 2011. *Data Mining: Concepts, Models, and Techniques*, Verlag Berlin Heidelberg, Springer

[5] Kusrini, dan Luthfi T.E., 2009. *Algoritma Data Mining*. P. Theresia Ari, Ed. Yogyakarta, Indonesia: Andi Offset.

[6] Bustami., 2014 . *Penerapan Algoritma Naive Bayes untuk mengklasifikasi data nasabah asuransi*. Jurnal Informatika Vol. 8, No 1, hlm. 884-898.

[7] Gorunescu, F., 2011. *Data Mining Concept, Models and Techniques,12th ed.,Prof. Lakhmi C. Jain Prof. Janusz Kacprzyk,Ed.* Craiova, Romania, Springer

[8] Polczynski's D.R.L., Lecture. 2010. WEKA *Classification Using Decision Trees.* *Computer science* : college of engginering & Applied sciences

[9] Satuan Kerja Rehabilitasi/Rekonstruksi Rumah Pasca Gempa Bumi DIY & Jateng, November 2014 "*Pembangunan Permukiman Lestari : Layak huni dan berkelanjutan*," Ditjen Cipta Karya, Kementrian Pekerjaan Umum.

[10] Sunjana, 2010. *Aplikasi Mining Data Mahasiswa Dengan Metode Klasifikasi Decision Tree*. Seminar Nasional Aplikasi Teknologi Informasi 2010. Snati 2010. , 24-29.

[11] Andriani, Anik. 2013. "*Sistem Prediksi Penyakit Diabetes Berbasis Decision Tree*". Jurnal Informatika. Vol.1.No.1

[12] Nurhadi W. , 2018. *Penerapan Algoritma Klasifikasi Naive Bayes Untuk Data Status Huni Rumah Bantuan Dana Rehabilitasi Dan Rekonstruksi Pasca Erupsi Gunung Merapi 2010.* Prosiding seminar nasional, Pendekatan Multidisiplin Ilmu Dalam Manajemen Bencana. Universitas Respati Yogyakarta. Daerah Istimewa Yogyakarta. Vol 1, No 1, ISSN 2657-2397