ISBN: 978-602-60280-1-3

Klasifikasi Data Mahasiswa Menggunakan Metode *Decision Tree Algoritma C4.5* divisualisasikan dalam GIS (Studi Kasus: Universitas PGRI Banyuwangi)

Sony Panca Budiarto
Program Studi Teknik Informatika
Sekolah Tinggi Ilmu Komputer PGRI Banyuwangi
Email : sonystikombanyuwangi@gmail.com

Abstract*— Permasalahan dalam data mahasiswa di Universitas PGRI Banyuwangi yaitu jumlah mahasiswa yang terus bertambah menghasilkan basis data mahasiswa dengan berbagai macam variabel dan atribut yang terlampau besar dan panjang, mengakibatkan kompleksitas dari data sehingga sulit untuk di mengerti dan dipahami dengan cara biasa. Metode klasifikasi menggunakan algoritma C4.5 divisualisasikan dalam GIS/Peta Interaktif dimanfaatkan untuk menambang pengetahuan dari data mahasiswa untuk (1) menemukan atribut dari data mahasiswa yang menyertai keberadaan mahasiswa di lokasi kecamatan asal mahasiswa pada wilayah/jarak ≤ 20 km dari UNIBA (dekat), > 20km - < 45km dari UNIBA (sedang) dan ≥ 45km dari UNIBA (jauh) divisualisasikan dalam GIS, (2) mendapatkan pasar potensial calon mahasiswa baru berdasarkan kedekatan asal Desa/Kelurahan mahasiswa di tiap kecamatan. Pengumpulan data dilakukan dengan metode studi pustaka (*library research*) dan metode pengumpulan data lapangan (*field research*) dengan menyebarkan Angket profil data diri mahasiswa UNIBA. Adapun hasil dari penelitian ini adalah informasi klasifikasi data mahasiswa dalam bentuk *decision tree*/pohon keputusan, digunakan sebagai sebagai alat untuk penentu keterangan visualisasi Pemodelan Deskriptif mendefinisikan lokasi kecamatan asal mahasiswa pada Peta Interaktif Kabupaten Banyuwangi dengan ketepatan 100%, dilengkapi dengan legenda informasi hasil klasifikasi. Prediksi pasar potensial mahasiswa baru menggunakan record data training mahasiswa angkatan 2012, 2013 dan 2014 diuji menggunakan data mahasiswa angkatan 2015 memiliki tingkat akurasi (61%).

Keywords— data mahasiswa, data mining, klasifikasi, algoritma C4.5, decision tree, GIS

I. PENDAHULUAN

A. Latar Belakang

Mahasiswa Universitas PGRI Banyuwangi adalah peserta didik yang telah terdaftar dan memenuhi persyaratan lain yang ditetapkan oleh Universitas PGRI Banyuwangi (UNIBA). Mahasiswa Universitas PGRI Banyuwangi berasal dari berbagai macam daerah dan sekolah yang tersebar di seluruh Kabupaten Banyuwangi pada jarak yang berbeda-beda, dengan status sosial ekonomi, latar belakang pendidikan keluarga, informasi tentang Universitas PGRI Banyuwangi yang berbeda-beda. Setiap mahasiswa yang terdaftar di Universitas PGRI Banyuwangi memiliki profil data mahasiswa yang terdiri dari Nama Mahasiswa, Nomor Induk Mahasiswa (NIM), Program Studi, Jenis Kelamin, Tempat/Tgl. lahir, Agama, Alamat Asal Mahasiswa, Asal Sekolah, Angkatan, Nama Ayah, Nama Ibu, Pendidikan Terakhir Orang Tua, dan Penghasilan Orang Tua Rp/bln. Semakin banyak mahasiswa UNIBA, data mahasiswa yang tersimpan di Sistem Informasi Akademik UNIBA semakin besar juga. Data mahasiswa yang terlampau besar dengan berbagai macam atribut yang ada jarang dilihat karena terlalu panjang dan membosankan serta tidak menarik.

Dari pemaparan diatas semakin banyak mahasiswa semakin besar data mahasiswa yang terkumpul, kecenderungan untuk mengamati dan mengolah data mahasiswa yang berasal dari suatu wilayah tertentu dengan berbagai macam atribut yang menyertainya semakin rumit untuk dimengerti dan dipahami dengan cara biasa. Ketika memilih perguruan tinggi, karakteristik area bukan merupakan hal penting, jarak memainkan peran lebih penting ketika memilih sebuah perguruan tinggi [1]. Kecenderungan mahasiswa di suatu universitas berasal dari satu wilayah/daerah tertentu pada jarak ke universitas yang berbeda-beda, dengan status sosial ekonomi yang berbeda-beda, latar belakang pendidikan keluarga yang berbeda-beda [2].

Algoritma C4.5 merupakan salah satu algoritma dalam teknik klasifikasi data yang menghasilkan satu pohon keputusan. Dari pohon keputusan yang dihasilkan maka akan terdapat beberapa rule atau pengetahuan untuk suatu kasus. Pohon Keputusan (Decision Tree) berguna untuk mengeksplorasi data, menemukan hubungan tersembunyi antara sejumlah calon variabel input dengan variabel target. Pohon keputusan mamadukan antara eksplorasi data dan pemodelan, pohon keputusan sangat bagus sebagai langkah awal dalam proses pemodelan ataupun ketika dijadikan sebagai model akhir dari beberapa teknik lain. Output dari algoritma C4.5 adalah pohon klasifikasi yang menunjukkan bagaimana suatu benda mungkin ditempatkan ke kelas yang sesuai atas dasar nilai-nilai atribut hasil klasifikasi [3]. Tetapi hasil dari klasifikasi metode decision tree algoritma C4.5 belum mampu memperhitungkan lokasi keberadaan dari objek dan hubungan antara objek dengan objek disekitarnya [4]. Didukung oleh penelitian sebelumnya yang menyatakan output dari algoritma C4.5 adalah pohon klasifikasi yang menunjukkan bagaimana suatu benda mungkin ditempatkan ke kelas yang sesuai atas dasar nilai-nilai atribut hasil klasifikasi [5].

Sistem Informasi Geografis (GIS) merupakan sebuah metode system informasi khusus yang mengelola data yang memiliki informasi spasial. Spasial adalah aspek keruangan suatu objek atau kejadian yang mencakup lokasi, letak dan

ISBN: 978-602-60280-1-3

posisinya. (UU No.4 tahun 2011, Tentang Informasi Geospasial). Pada penelitian ini sebuah peta interaktif direncanakan dengan mengintegrasikan teknik klasifikasi data mahasiswa universitas PGRI Banyuwangi menggunakan metode decision tree algoritma C4.5 divisualisasikan dalam GIS/Peta Interaktif. Sehingga dapat membantu memberikan tambahan informasi kepada penentu kebijakan di UNIBA terkait hasil klasifikasi data mahasiswa dengan atribut yang paling berpengaruh dan pola klasifikasi data mahasiswa beserta atribut apa saja dalam data mahasiswa yang menyertai keberadaan mahasiswa dari suatu daerah tertentu. Dalam hal ini yang divisualisasikan adalah model basis data spasial yang menggambarkan lokasi keberadaan mahasiswa di tiap kecamatan dilengkapi dengan keterangan atribut yang menyertai hasil klasifikasi data mahasiswa pada jarak dekat, sedang dan jauh dari UNIBA.

1.1 Tujuan

Menambang pengetahuan dari data mahasiswa Universitas PGRI Banyuwangi menggunakan *Decision Tree* Algoritma C4.5 divisualisasikan dalam GIS, untuk (1) menemukan atribut dari data mahasiswa yang menyertai keberadaan mahasiswa di titik lokasi (kecamatan) asal mahasiswa pada wilayah/jarak ≤ 20 km dari UNIBA (dekat), > 20 km - < 45 km dari UNIBA (sedang) dan ≥ 45 km dari UNIBA (jauh) divisualisasikan dalam GIS, (2) mendapatkan pasar potensial calon mahasiswa baru berdasarkan kedekatan asal Desa/Kelurahan mahasiswa di tiap kecamatan.

II. TINJAUAN PUSTAKA

2.1 Penelitian Terkait

Andrienko, G., & Andrienko, N. (Sept.5,1999 to Sept 6,1999), dengan judul "*Data Mining with C4.5 and Interactive Cartographic Visualization*" menggunakan proses KDD metode data mining klasifikasi Algoritma C4.5 dalam penerapan teknik penemuan pengetahuan database di suatu wilayah atau kota di Negara eropa, didukung oleh tampilan peta interaktif. Data mengacu pada unit pembagian wilayah, seperti kejadian beberapa penyakit dikota suatu negara. Analis perlu mendeteksi cluster territorial dengan kejadian yang sama, pada objek dan studi yang sama dalam hal atribut seperti pendapatan, usia, kondisi lingkungan, dll. Objek spasial yang ditunjukkan pada peta dikelompokkan ke dalam kelas yang berbeda dibagi menjadi subinterval disesuaikan dengan nilai-nilai atribut yang ada di peta. Misalnya peta diberi warna yang sama untuk data yang berada dikelas interval yang sama.

Pada penelitian ini peneliti berusaha mencapai sinergi dua pendekatan eksplorasi data spasial dan analisis visual dengan menampilkan katografi atau peta interaktif yang memberi gambaran hasil klasifikasi database menggunakan algoritma C4.5. Penggunaan pengetahuan yang ditemukan yaitu dengan memasukkan pengetahuan kedalam system lain untuk ditindak lanjuti.

Sebuah pohon keputusan yang dihasilkan dari Algoritma C4.5 tidak mengandung informasi tentang fitur spasial dari objek yang diklasifikasikan. Untuk memudahkan menganalisa hasil klasifikasi yaitu, menggabungkan hasil pohon keputusan dengan peta interaktif untuk divisualisasi, dengan cara menyorot di peta tempat/lokasi suatu kejadian atau benda yang mempunyai kelas yang sama hasil klasifikasi dalam setiap node yang dipilih dari pohon keputusan.

Gambar 1.Negara-negara dengan tingkat kelahiran rendah dan angka harapan hidup yang tinggi dalam kaitannya dengan tingkat kesuburan

Das, S., Dahiya, S., & Bharadwaj, A. "*An Online Software for Decission Tree Clasification and Visualization using C4.5 Algorithm (ODTC)*". Klasifikasi abstrak adalah penting dan banyak dilakukan dalam pengolahan data mining. Ini adalah tugas pemodelan prediktif yang didefinisikan sebagai membangun model untuk variabel target sebagai fungsi dari variabel penjelas. Pohon keputusan adalah alat pendukung keputusan yang menggunakan pohon seperti grafik atau model keputusan yang dibuat dari hasil klasifikasi menggunakan algoritma C4.5 [6].

Dalam penelitian ini sebuah perangkat lunak berbasis web menggunakan decision tree algoritma C4.5 digunakan untuk menghasilkan rule/aturan klasifikasi. Rule/aturan klasifikasi divisualisasi dalam bentuk struktur pohon keputusan agar mudah dipahami. ODTC merupakan fasilitas online yang memvisualisasikan hasil dari pohon keputusan menjadi rule/aturan. Pengguna dapat mengekspor hasil di excel, pilihan mencetak hasil. Sistem ini digunakan untuk mengekstraksi pola tersembunyi yang ada didataset besar menggunakan algoritma C4.5. Sistem ini digunakan sebagai perangkat lunak generasi pengetahuan yang dapat mengekstrak dan memberikan pengetahuan tertentu yang ada didataset dalam bentuk aturan-aturan dan pohon keputusan. Software akan berguna bagi akademisi, peneliti dan siswa yang bekerja di bidang data mining, pertanian dan bidang lain yang menghasilkan data dalam jumlah besar.

Andrienko, G. L., & Andrienko, N. V. (n.d.). "*GIS Visualization Support to the C4.5 Classification Algorithm of KDD*". Menunjukkan sinergi KDD dan visualisasi kartografi yang dicapai dengan integrasi dari Sistem KDD Kepler dan DESCARTES untuk menampilkan peta otomatis yang dihasilkan dari proses KDD metode data mining algoritma C4.5 [7].

Pada penelitian ini data diperoleh berdasarkan unit pembagian wilayah, ekonomi dan demografi tentang negara-negara Eropa. Target atribut klasifikasi adalah produk nasional perkapita, kisaran nilai telah dibagi menjadi 4 subinterval, yang menghasilkan nilai-nilai '<4000', '4.000-10.500', '10500-19000 ', '> 19000'; Klasifikasi Tingkat Kesuburan, dan Usia. Hasil klasifikasi menunjukkan bagaimana kelas-kelas ini dapat dibedakan atas dasar nilai-nilai lainnya atribut (Tingkat kelahiran, angka harapan

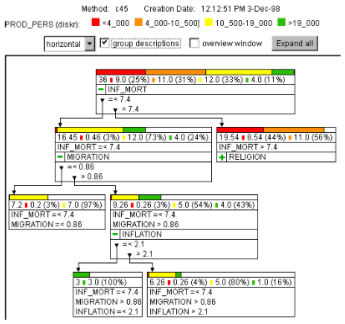
ISBN: 978-602-60280-1-3

hidup, kematian bayi, tingkat migrasi, inflasi, agama yang dominan, dll).

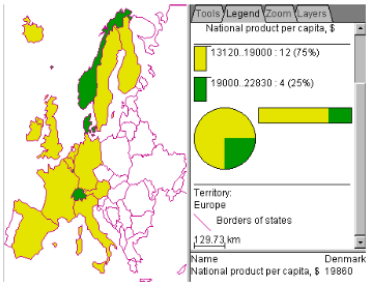
Langkah pertama analis memberikan tanda ke objek spasial menurut sifat spasial mereka (lokasi, ukuran, dll) atau sekitarnya. Tanda kemudian menjadi nilai-nilai atribut yang dihasilkan. Misalnya, kabupaten kota dapat dibagi menjadi 'pusat' dan 'pinggiran', kota dan kota-kota dapat ditandai sebagai berada di kawasan hutan, di daerah pertanian, atau dalam jumlah besar aglomerasi kota, dan sebagainya. Langkah kedua membagi wilayah yang ditunjukkan pada peta ke daerah yang diberi nama. Data atribut diklasifikasikan sesuai dengan kecocokan wilayah atribut berasal. Langkah ketiga menyesuaikan objek yang diklasifikasikan menurut subinterval nilai-nilai terkait. Hasil klasifikasi ditampilkan pada peta, misalnya dengan memberi warna pada interval dan daerah objek dalam warna-warna yang berbeda. Interpretasi hasil KDD:

- Pohon klasifikasi
- Aturan
- Kelompok data/atribut sama

Hasil klasifikasi pohon keputusan seperti ditunjukkan pada gambar 2. Ketika pengguna mengklik node akar (root) yang berisi kondisi ‘INF_MORT ≤ 7.4 (angka kematian bayi kurang dari atau sama dengan 7,4 kematian per 1.000 kelahiran hidup), peta interaktif akan tampil seperti terlihat pada gambar 3. Menunjukkan bahwa negara bagian di Eropa Barat memiliki nilai pendapatan produk nasional bruto perkapita yang sangat tinggi (minimum adalah 13.120 $).



Gambar 2. Pohon keputusan yang dihasilkan dari table dengan data ekonomi dan demografi negara-negara Eropa



Gambar 3. Tampilan distribusi geografis dari negara-negara sesuai dengan simpul tingkat kedua sebelah kiri pohon keputusan.

Ketika pengguna mengklik node lain, peta baru akan tampil mewakili node terakhir yang dipilih menggantikan peta

2.2 Teori Penunjang

Data Mining adalah proses yang mempekerjakan satu atau lebih teknik pembelajaran computer (mechine learning) untuk menganalisis dan mengekstraksi pengetahuan (knowledge) secara otomatis. *Knowledge Discovery in Database* (KDD) adalah penerapan metode saintifik pada data mining. Dalam konteks ini data mining merupakan satu langkah dari proses KDD [8]. Alasan mengapa melakukan data mining dari sudut pandang komersial karena (1) meledaknya volume data yang dihimpun dan disimpan dalam data warehouse seperti data web, e-commerce, penjualan di departemen store, transaksi bank/credit card, (2) proses kommputasi yang dapat diupayakan, (3) kuatnya tekanan kompetitif untuk dapat menyediakan layanan-layanan yang lebih baik customnisasi dan informasi sedang menjadi produk yang berarti. Alasan dari sudut pandang keilmuan adalah (1) kecepatan data yang dihimpun dan disimpan (GB/hour) seperti pada remote sensor pada satelit, telescope untuk memindai langit dan simulasi saintifik yang membangkitkan data dalam ukuran terabytes, (2) teknik-teknik tradisional tidak fisibel untuk mengolah data mentah, (3) data mining untuk reduksi data untuk klasifikasi dan segmentasi data serta membantu ilmuwan dalam melakukan formasi hipotesis.

Klasifikasi adalah penempatan objek-objek ke salah satu dari beberapa kategori yang telah ditetapkan sebelumnya. Klasifikasi merupakan proses yang terdiri dari dua tahap, yaitu tahap pembelajaran dan tahap pengklasifikasian. Pada tahap pembelajaran, sebuah algoritma klasifikasi akan membangun sebuah model klasifikasi dengan cara menganalisis training data. Tahap pembelajaran dapat juga dipandang sebagai tahap pembentukan fungsi atau pemetaan Y=F(X) diman Y adalah kelas hasil prediksi dan X adalah *tuple* yang ingin diprediksi kelasnya. Selanjutnya pada tahap pengklasifikasian, model yang telah dihasilkan akan digunakan untuk melakukan klasifikasi. Model klasifikasi berguna untuk keperluan berikut [9]:

A. Pemodelan Deskriptif. Model klasifikasi dapat bertindak sebagai alat penjelas untuk membedakan objek-objek dari kelas-kelas yang berbeda. Sebagai contoh, untuk para ahli Biologi, model deskriptif yang meringkas data berguna untuk menjelaskan fitur-fitur apakah yang mendefinisikan vertebrata sebagai mammal, bird, fish, reptile atau amphibian.
B. Pemodelan Prediktif. Model klasifikasi juga dapat digunakan untuk memprediksi label kelas dari record yang tidak diketahui. Sebuah model klasisfikasi dapat dipandang sebagai kotak hitam yang secara otomatis memberikan sebuah label kelas ketika dipresentasikan dengan himpunan atribut dari record yang tidak diketahui.

Algoritma C4.5 merupakan pengembangan dari metode Iterative Dichotomiser 3 (ID3), dimana untuk pemilihan split atribut menggunakan metode Info Gain Ratio (IGR) menggantikan Info Gain (IG). C4.5 yang diperkenalkan dapat bekerja pada variabel kontinyu dan missing value. Info Gain Ratio memiliki sifat : 1) bernilai besar bila data menyebar rata; 2) bernilai kecil bila semua

ISBN: 978-602-60280-1-3

data masuk dalam satu cabang. Rumus persamaan Info Gain Ratio (IGR) seperti berikut [10]:

1. Cari atribut sebagai akar.

Data dikelompokkan terlebih dahulu dengan menghitung jumlah kasus, jumlah kasus untuk anggota yang berada pada jarak dekat, sedang, dan jauh. Selanjutnya dihitung nilai Entrophy, Gain dari dan Info Gain Ratio (IGR) semua atribut.

Nilai *entropy* dapat diperoleh dari rumus:

$$Entrophy(\text{S}) \text{ atau Info } (S) = \sum_{i=1}^{a} -Pi.Log_2\,(Pi)$$

Dimana Log_2 Pi,dapat diartikan dalam rumus 2:

$$\text{Log2(x)} = \frac{\text{Ln(x)}}{\text{Ln(2)}}$$

Keterangan:
S : Himpunan kasus
A : Fitur
n : Jumlah partisiS
p_i : Proporsi dari $S_i$terhadap S

Nilai *Entrophy (Si)* atau $Info_x$(T) untuk setiap kasus data dapat dilihat pada rumus dibawah ini

$$Entrophy(Si) \text{ atau } Info_x(T) = \sum_{i=1}^{n} \frac{|Si|}{|S|}\ x\ Entrophy(Si)$$

Nilai *Gain* diperoleh dengan rumus:

$$Gain\ (S,A) = Entrophy(S) - \sum_{i=1}^{n} \frac{|Si|}{|S|}\ x\ Entrophy(Si)$$

Keterangan:
S : Himpunan kasus
A : Atribut
n : Jumlah partisiatribut A
$|S_i|$: Jumlah kasus padapartisi ke-i
|S| : Jumlah kasus dalam S

Nilai Info Gain Ratio (IGR) dapat diperoleh dari rumus:

$$Gain\ Ratio\ (X) = \frac{Gain\ (X)}{Split\ Info\ (X)}$$

Dimana *split info(x)* diperoleh dengan rumus:

$$Split\ Info\ (x) = -\sum_{i=1}^{n} \frac{|\text{T1}|}{|\text{T}|} \text{x Log}_2 \left(\frac{|\text{T1}|}{|\text{T}|}\right)$$

2. Buat cabang untuk masing – masing nilai.
3. Bagi kasus dalam cabang.
4. Ulang proses untuk masing-masing cabang sampai semua kasus pada cabang memiliki kelas yang sama. Untuk menentukan akar diperoleh dengan cara menghitung nilai gain masing-masing atribut, kemudian dipilih nilai info gain tertinggi, maka atribut dengan nilai info gain tertinggi itulah yang akan menjadi akarnya.

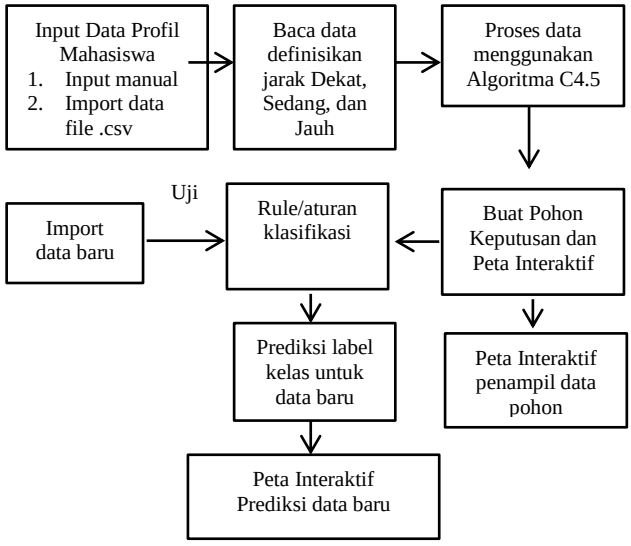
III. METODE PENELITIAN

Jumlah mahasiswa yang mencapai 1314 di UNIBA saat ini adalah peserta didik yang telah terdaftar dan memenuhi persyaratan lain yang ditetapkan oleh UNIBA, sehingga dibuatkan data profil diri atau identitas pribadi mahasiswa yang tersimpan di buku induk ataupun system informasi akademik kampus (SIAKAD). Data yang besar dengan berbagai macam atribut yang ada sangat sulit untuk dimengerti dan dipahami dengan cara biasa. Untuk mendapatkan pola klasifikasi serta informasi/ pengetahuan tersembunyi yang ada dalam database mahasiswa tersebut, maka perlu dibuatkan sistem yang mampu mengeksplorasi data, menemukan hubungan tersembunyi antara sejumlah variabel input. Penggunaan GIS/Peta Interaktif untuk visualisasi pola klasifikasi dari pohon keputusan ditujukan untuk memberi kemudahan mendifinisikan jarak dekat, sedang dan jauh lokasi kecamatan asal mahasiswa.

Konsep jarak dalam penelitian ini digunakan untuk mengetahui keterkaitan antara lokasi Universitas PGRI Banyuwangi (UNIBA) di kecamatan Banyuwangi dengan kecamatan-kecamatan lain yang ada di seluruh Kabupaten Banyuwangi. Pemilihan atribut jarak tempat tinggal mahasiswa ke UNIBA yang berada pada wilayah/jarak ≤ 20 km dari UNIBA (dekat), > 20 km - < 45 km dari UNIBA (sedang), ≥ 45 km dari UNIBA (jauh) sebagai target atribut dimaksudkan untuk mempermudah proses penggalian data/informasi mahasiswa dari database mahasiswa.

Angket data profil data mahasiswa yang akan digunakan sebagai data training dan data test ketika dibuat menggunakan program microsoft excel harus di simpan dengan format .CSV (comma delimeted). Pada penelitian ini program hanya bisa membaca dan menyimpan file yang memiliki format data .CSV.

Pada penelitian ini sistem dirancang dengan cara membagi program menjadi dua bagian, yaitu: program antarmuka dan program perhitungan. Program antarmuka dibuat menggunakan bahasa Pascal dengan IDE Lazarus 1.6, sementara program perhitungan dibuat menggunakan bahasa Fortran dengan IDE Force 2.0.



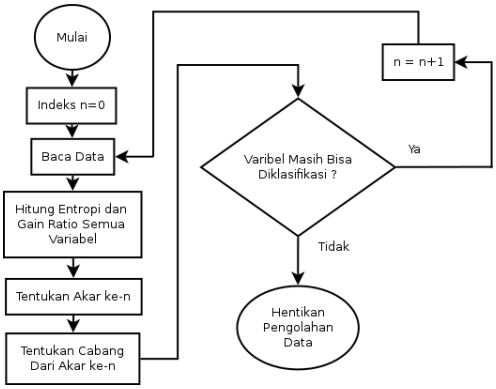
Gambar 4. Diagram Alir Perancangan Sistem

ISBN: 978-602-60280-1-3

Preprocessing file input data training awalnya berbentuk microsoft excel dalam format .CSV. Sistem yang dibuat pada penelitian ini memproses data yang berformat .txt. Supaya file input data training yang berformat .CSV dapat diproses oleh sistem maka data harus dirubah kedalam format .txt. Data dalam format .txt yang diproses oleh sistem adalah: a_pend_ortu.txt; a_pengh_ortu.txt; a_prodi.txt; a_info_kampus.txt; a_jarak.txt; a_minat.txt; a_nama.txt; a_pekerj_ortu.txt; a_asal_sklh.txt; a_daerah_asl.txt; datafile.txt, file data ini dibuat dengan menggunakan bahasa fortran.

Proses data menggunakan algoritma C4.5 yang dibuat menggunakan bahasa Fortran dengan IDE Force 2.0. Memiliki 3 bagian penting yang bertujuan untuk membuat file data dalam format .txt dan .csv yang nantinya akan digunakan untuk ditampilkan pada pohon keputusan, peta dan rule. Program dimulai dengan mendeklarasikan semua variabel data input dengan tujuan untuk membaca file .txt yang digunakan sebagai input program yang sudah dibuat. Tidak semua variabel dari input data training akan digunakan sebagai data prediktor untuk diolah menggunakan Algoritma C4.5. Hanya 6 variabel yang akan dihitung menggunakan algoritma C4.5 antara lain: program studi, pekerjaan orang tua, penghasilan orang tua, pendidikan orang tua, info kampus, faktor minat. Variabel lain seperti nama, asal sekolah, dan daerah asal mahasiswa digunakan sebagai data reffrensi.

Selanjutnya menentukan variabel yang akan diolah menggunakan algoritma C4.5. Variabel yang akan diolah antara lain: program studi, pekerjaan orang tua, penghasilan orang tua, pendidikan orang tua, info kampus, dan faktor minat yang didapatkan dari input data training. Diagram alir perhitungan klasifikasi data menggunakan algoritma C4.5 dapat dilihat pada gambar 5 dibawah ini.



Gambar 5. Diagram Alir Perhitungan Klasifikasi Algoritma C4.5

Adapun langkah-langkah proses pembentukan pohon keputusan dengan Algoritma C4.5 antara lain, yaitu:

1. Dimulai menginput data training
2. Definisikan input data training ke jarak dekat, sedang, jauh
3. Lakukan perhitungan pertama untuk mencari variabel yang akan menjadi internal root (indeks n=0)
4. Baca data semua variabel yang akan diklasifikasi
5. Hitung Entropi dan Gain Ratio semua variabel
6. Dari hasil perhitungan pertama tentukan variabel dengan nilai Gain Ratio tertinggi yang kemudian akan menjadi root
7. Jabarkan cabang dari variabel yang menjadi root
8. Jika variabel masih bisa diklasifikasi maka pehitungan dilanjutkan untuk akar berikutnya (n=n+1). Jika tidak maka perhitungan dihentikan.
9. Selesai

Program akan menghasilkan file data dalam format .txt dan .csv setelah input data training dideklarasikan kemudian dihitung entropi dan gain ratio menggunakan algoritma C4.5.

Langkah-langkah yang ditempuh dalam melakukan penelitian ini, yaitu:

1. Analisa masalah dan analisa kebutuhan: Dalam tahap ini dilakukan pencarian masalah yang ada pada data mahasiswa di Universitas PGRI Banyuwangi dan mengidentifikasinya. Dari masalah yang ditemukan kemudian dianalisa untuk dicarikan solusi atas permasalahan yang ada sesuai dengan kebutuhan pengguna.
2. Pengumpulan data: Pengumpulan data dilakukan dengan metode studi pustaka (*library research*) dan metode pengumpulan data lapangan (*field research*) dengan menyebarkan Angket profil data diri mahasiswa UNIBA pada saat Ujian Akhir Semester (UAS) berlangsung.
3. Pengujian sistem sebelum di implementasikan: Pada tahap ini dilakukan dua kali pengujian pada sistem yang dibuat dengan cara, yaitu : 1) membandingkan nilai hitung Gain Ratio sistem yang telah dibuat dengan nilai gain ratio dari sistem aplikasi datamining Weka. 2) membandingkan tingkat akurasi prediksi yang didapat menggunakan test option teknik use training set dari sistem yang telah dibuat dengan sistem aplikasi datamining Weka.
4. Implementasi: Sistem yang telah diuji nilai gain ratio dan tingkat akurasinya akan digunakan untuk:
   a. Pemodelan Deskriptif: pohon keputusan yang dihasilkan berguna untuk menjelaskan fitur-fitur apakah yang mendefinisikan keberadaan mahasiswa pada jarak dekat, sedang dan jauh divisualisasikan dalam Peta Interaktif.
   b. Pemodelan Prediktif : memprediksi keberadaan mahasiswa Universitas PGRI Banyuwangi pada jarak dekat, sedang, dan jauh. Data Uji yang di “Prediksi Benar” menunjukkan jarak atau lokasi keberadaan mahasiswa (kecamatan) yang memiliki pola klasifikasi relevan dengan rule klasifikasi dari data training. Peta Prediksi yang dibuat berdasarkan data uji ”Prediksi Benar” akan dijadikan rujukan/acuan basis keberadaan mahasiswa UNIBA saat ini.
5. Evaluasi: Untuk mengukur apakah sistem yang dibuat berhasil atau tidak adalah dengan melakukan evaluasi. Evaluasi digunakan untuk mengukur keakuratan hasil yang dicapai sistem. Evaluasi kinerja dari sistem dapat diketahui dari banyaknya (count) dataset/record yang diprediksi secara benar dan tidak benar berdasarkan pola klasifikasi yang dihasilkan oleh sistem yang dibuat.

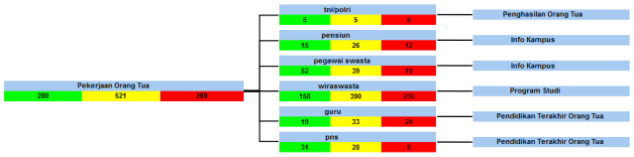
ISBN: 978-602-60280-1-3

Confusion Matrix digunakan untuk mentabulasikan hasil prediksi

IV. HASIL DAN PEMBAHASAN

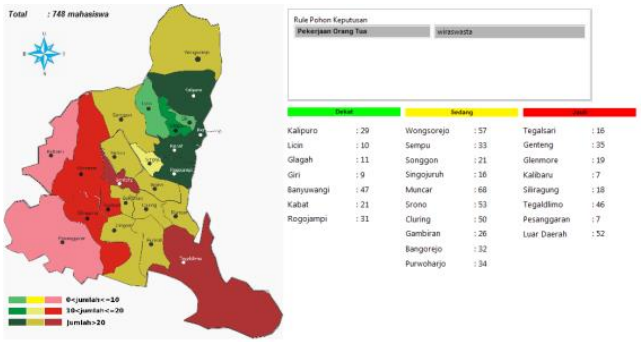
A. Implementasi

Implementasi pohon keputusan untuk visualisasi peta pemodelan deskriftif bertujuan untuk menganalisa data klasifikasi melalui pohon keputusan yang terbentuk. Pembentukan pohon keputusan dilakukan melalui proses pengolahan input data training menggunakan algoritma C4.5. Data training yang digunakan berasal dari data mahasiswa angkatan 2012, 2013 dan 2014 sebanyak 1070 mahasiswa.



Gambar 6. Pohon Keputusan Yang Dibuat Menggunakan 1070 Data Training

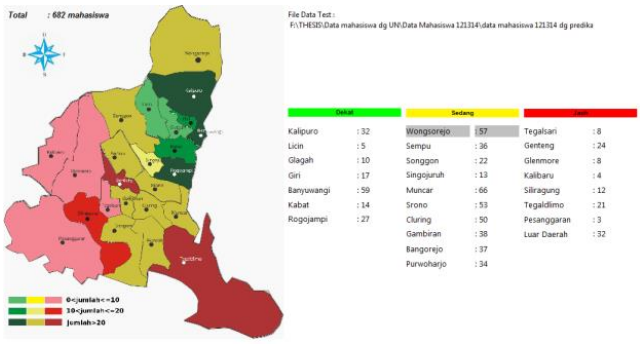
Dari pohon keputusan yang dihasilkan peneliti dapat menganalisa sebaran jumlah mahasiswa pada jarak dekat, sedang dan jauh dari kampus UNIBA. Faktor yang paling berpengaruh pada data mahasiswa untuk memilih kuliah di UNIBA. Dibawah ini dicontohkan peta sebaran mahasiswa berdasarkan pekerjaan orang tuanya wiraswasta.



Gambar 7. Peta Sebaran Mahasiswa Yang Pekerjaan Orang Tua Wiraswasta

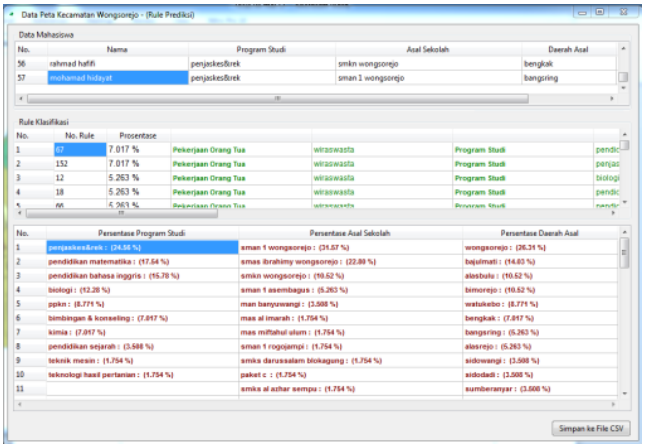
Pada gambar 7 peta sebaran mahasiswa yang pekerjaan orang tua wiraswasta. Peta tersebut mendeskripsikan sebaran jumlah mahasiswa pada tiap kecamatan pada jarak dekat, sedang atau jauh dari lokasi kampus UNIBA. Gradasi warna yang diperlihatkan pada peta menunjukkan besarnya jumlah mahasiswa yang berada di suatu kecamatan. Semakin gelap warna peta kecamatan menunjukkan semakin banyak mahasiswa hasil klasifikasi yang berada di kecamatan tersebut. Proses penggalian pengetahuan terhadap hasil klasifikasi lebih detail dapat dilakukan dengan cara menekan internal node dan leaf node dari tiap cabang pada pohon keputusan

Implementasi program untuk penggalian informasi prediksi pasar potensial mahasiswa baru dilakukan dengan cara penggalian pengetahuan tersembunyi yang ada pada data training kumpulan data mahasiswa angkatan 2012, 2013 dan data mahasiswa angkatan 2014 sebanyak 1070 data mahasiswa. Data tersebut diklasifikasikan menggunakan program yang sudah dibuat. Dari proses penggalian informasi tersembunyi yang ada pada data training diketahui bawasannya sebanyak 682 (63,73%) mahasiswa dari 1070 mahasiswa mempunyai pola klasifikasi dan atribut target jarak yang sama.



Gambar 8. Peta Prediksi Pasar Potensial Mahasiswa Baru

Berdasarkan peta prediksi pasar potensial mahasiswa baru, peneliti mencoba menggali informasi dan pengetahuan tersembunyi yang ada pada hasil klasifikasi di Kecamatan Wongsorejo, dari legenda peta seperti ditunjukkan pada gambar 9 dibawah ini.



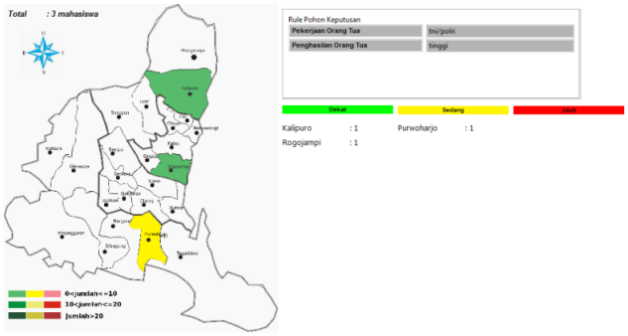
Gambar 9. Legenda Informasi Prediksi Pasar Potensial Mahasiswa Baru di Wongsorejo

Berdasarkan data hasil klasifikasi yang ada pada legenda diperoleh informasi rule klasifikasi yang berlaku di suatu kecamatan, prosentase program studi, prosentase asal sekolah mahasiswa, dan prosentase daerah asal (desa/kelurahan) yang ada di suatu kecamatan yang akan digunakan untuk prediksi pasar potensial mahasiswa baru.

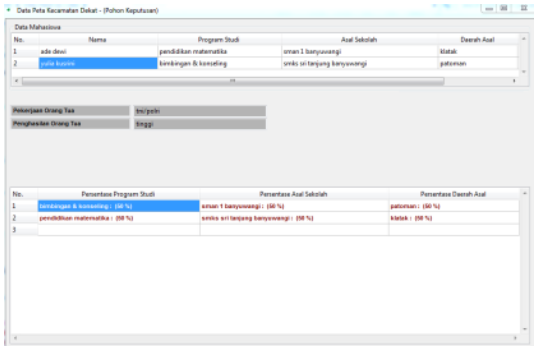
B. Uji Coba

Peta sebaran mahasiswa dan legenda peta yang dihasilkan dengan cara klik rule/aturan klasifikasi diuji kebenarannya dengan data asli yang diklasifikasi menggunakan teknik filter dengan program excel. Metode pengujian yang dilakukan adalah dengan cara: Pengujian pertama menggunakan rule no 7, rule yang digunakan adalah Pekerjaan Orang Tua tni/polri-Penghasilan Orang Tua tinggi-dekat, Visualisasi Peta ditampilkan seperti pada gambar 10, dibawah ini.

ISBN: 978-602-60280-1-3

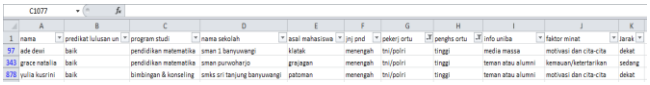


Gambar 10. Pekerjaan Orang Tua tni/polri-Penghasilan Orang Tua tinggi



Gambar 11. Legenda Dari Peta Pekerjaan Orang Tua tni/polri-Penghasilan Orang Tua tinggi-Dekat

Data visualisasi peta sebaran mahasiswa dan informasi hasil klasifikasi yang ditampilkan pada legenda diuji/dibandingkan dengan data hasil klasifikasi menggunakan program excel.



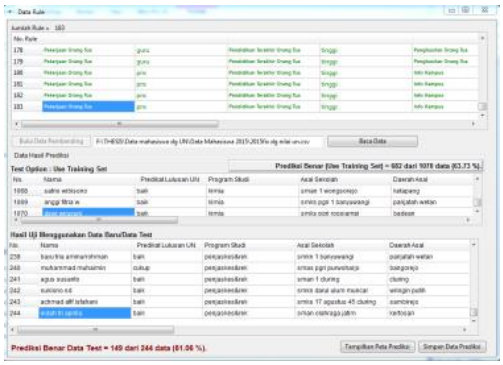
Gambar 12. Pengujian Rule Pekerjaan Orang Tua tni/polri-Penghasilan Orang Tua tinggi-Dekat Menggunakan Excel

Dengan cara yang sama pengujian dilakukan menggunakan 25 macam rule, didapatkan hasil tingkat akurasi ketepatan visualisasi dan informasi data hasil klasifikasi pada legenda sebesar 100%.

Uji coba tingkat akurasi program untuk prediksi pasar potensial mahasiswa baru menggunakan data mahasiswa angkatan 2015. Peneliti berpendapat besarnya tingkat akurasi prediksi benar yang terjadi pada hasil klasifikasi data mahasiswa angkatan 2015, menunjukkan bahwa informasi yang ditampilkan program dari proses penggalian data training (data mahasiswa angkatan 2012, 2013 dan 2014) dapat digunakan untuk memprediksi pasar potensial calon mahasiswa baru angkatan selanjutnya.

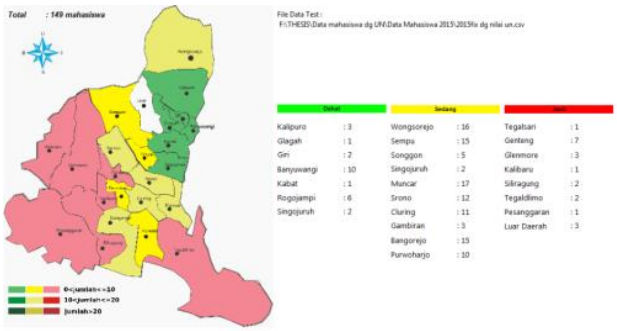
Uji Coba Klasifikasi Data Mahasiswa Angkatan 2015.

Data Training : Data mahasiswa angkatan 2012 + 2013+2014
Jumlah data : 1070 record/data mahasiswa
Data Test : Data mahasiswa angkatan 2015
Jumlah data : 244 record/data mahasiswa

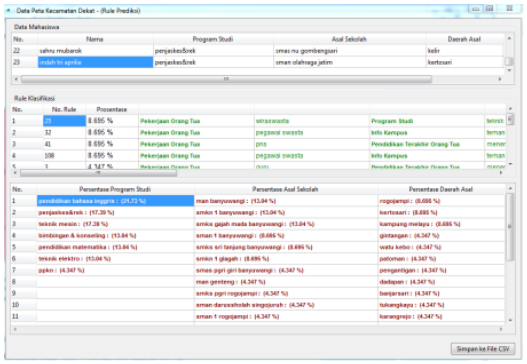


Gambar 13. Antarmuka Program Untuk Pengujian Klasifikasi Data Baru

Gambar 13 antarmuka program untuk pengujian prediksi data baru (data mahasiswa angkatan 2015). Pada bagian atas antarmuka ditampilkan jumlah rule dan pola/aturan klasifikasi yang dihasilkan dari pemprosesan input data training menggunakan algoritma C4.5. Pada bagian tengah ditampilkan data hasil prediksi menggunakan test option teknik use training set serta tingkat akurasinya. Pada bagian bawah digunakan untuk import record data test/data uji dan ditampilkan hasil klasifikasinya serta jumlah data yang diprediksi benar dan tingkat akurasinya.



Gambar 14. Peta Prediksi Benar Lokasi Mahasiswa Angkatan 2015



Gambar 15. Legenda Data Peta Kecamatan Hasil Klasifikasi Prediksi Benar Pada Jarak Dekat

Dari keseluruhan data test yang diklasifikasi dengan prediksi benar pada jarak dekat, sedang dan jauh memiliki tingkat akurasi sebesar (61,06%) artinya hasil klasifikasi data baru yang diprediksi dengan benar sebanyak 149 (61,06%) mahasiswa dari 244 mahasiswa yang diujikan memiliki pola klasifikasi dan target atribut jarak yang sama dengan mahasiswa angkatan sebelumnya.(Data Mahasiswa angkatan 2012,2013, dan 2014 yang dijadikan data training).

ISBN: 978-602-60280-1-3

Evaluasi sistem menggunakan metode Confusion Matrix untuk perhitungan nilai Precision, Recall dan Akurasi pada klasifikasi data mahasiswa angkatan 2015.

Tabel 1. Confusion Matrix data 2015

	Dekat	Sedang	Jauh	Jumlah
Dekat	23	20	12	55
Sedang	11	106	14	131
Jauh	12	26	20	58
Jumlah	46	152	46	244

Perhitungan Nilai Precision: Perhitungan Nilai Recall :

a. Dekat : 23/46 = 0,5 Dekat: 23/55 = 0,42
b. Sedang: 106/152 = 0,7 Sedang:106/131= 0,8
c. Jauh : 20/46 = 0,43 Jauh: 20/58 = 0,34

C. Analisa Pembahasan

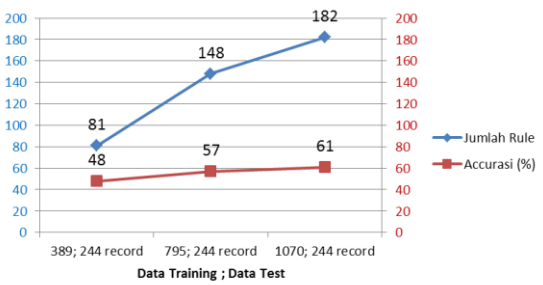
Pohon keputusan yang dihasilkan oleh program berjalan sesuai dengan yang diharapkan dalam hipotesis yaitu sebagai alat untuk penentu keterangan visualisasi Pemodelan Deskriptif mendefinisikan lokasi kecamatan asal mahasiswa pada Peta Interaktif Kabupaten Banyuwangi dengan ketepatan 100%, dilengkapi dengan legenda informasi data hasil klasifikasi.

Prediksi pasar potensial mahasiswa baru berdasarkan pola klasifikasi yang dihasilkan dari record data training mahasiswa angkatan 2012, 2013 dan 2014 diuji cobakan terhadap data mahasiswa angkatan 2015 memiliki tingkat akurasi (61%) dapat diartikan bahwa (61%) mahasiswa angkatan 2015 berada di wilayah prediksi dekat, sedang dan jauh dengan pola klasifikasi yang sama dengan data mahasiswa angkatan 2012, 2013 dan 2014.

Pohon keputusan, Rule dan Peta yang dibuat pada program ini bersifat dinamis artinya setiap terjadi perubahan pada input data training (penambahan data maupun pengurangan data training) maka secara otomatis terjadi perubahan pohon keputusan, rule dan peta yang dihasilkan program.

V. KESIMPULAN

1. Integrasi teknik klasifikasi data mahasiswa Universitas PGRI Banyuwangi menggunakan metode decision tree algoritma C4.5 dengan GIS/Peta Interaktif menghasilkan pohon keputusan yang dapat bekerja sesuai dengan yang diharapkan dalam hipotesis yaitu sebagai alat untuk penentu keterangan Visualisasi Pemodelan Deskriptif mendefinisikan lokasi kecamatan asal mahasiswa pada Peta Interaktif Kabupaten Banyuwangi.
2. Semakin banyak data training yang digunakan semakin banyak rule yang dihasilkan semakin besar tingkat akurasi yang didapatkan.
3. Tingkat akurasi prediksi pasar potensial mahasiswa baru sebesar (61%) diperoleh dari klasifikasi data mahasiswa angkatan 2015 sebanyak 149 dari 244 mahasiswa angkatan 2015 diprediksi benar berada di jarak dekat, sedang dan jauh dari lokasi kampus UNIBA sesuai dengan rule/aturan klasifikasi.
4. Kekurangan Algoritma C4.5 data yang dihasilkan harus dibaca sesuai rule, tidak bisa difilter hanya pada atribut yang dipilih.



DAFTAR PUSTAKA

[1] Veloutsou, C., Lewis, J., & Paton, R. (2004). University selection information requirements and importance. *The International Journal of Education Management* , Vol. 18, pp. 160-71.

[2] Tang, H., & McDonald, S. (2002). Integrating GIS and Spatial Data Mining Techique For Target Marketing Of University Courses. *Symposium on Geospatial Theory, Processing and Applications,Symposium sur la théorie, les traitements et les applications des données Géospatiales.* Ottawa: ISPRS, IGU, CIG, SIPT, UCI, ACSG.

[3] Quinlan, J. R. (1986). Introduction of Decision Tree. *Machine Learning*, 81-106.

[4] Andrienko, G., & Andrienko, N. (Sept.5,1999 to Sept 6,1999). Data Mining with C4.5 and Interactive Cartographic Visualization. *IEEE* , 162.

[5] Andrienko, G.L., & Andrienko, N "Interactive Maps for Visual Data Exploration", *International Journal of Geographical Information Science*, 13 (4), 1999, pp.355-374.

[6] Das, S., Dahiya, S., & Bharadwaj, A. (2014). An Online Software for Decission Tree Clasification and Visualization using C4.5 Algorithm (ODTC). *Konferensi Internasional tentang Komputasi Berkelanjutan Global Development (IndiaCom)* (hal. 978-93-80544-12-0/14). India: IEEE.

[7] Andrienko, G. L., & Andrienko, N. V. (n.d.). GIS Visualization Support to the C4.5 Classification Algorithm of KDD. *GMD - German National Research Center for Information Technology Schloss Birlinghoven, Sankt-Augustin, D-53754 Germany* .

[8] Maimon, R. (2005). Data Mining and Knowledge Discovery Handbook. *Springer*.

[9] Ph.D, Sani Susanto., & S.T., M.S, Suryadi Dedy. (2010). *Pengantar Data Mining Menggali Pengetahuan dari Bongkahan Data.* Bandung: Andi Yogyakarta.

[10] Quinlan, J.R. (1993). C4.5: *Program for Machine Learning*. The Morgan Kaufmann Publishers.