PEMBANGUNAN APLIKASI PENDETEKSIAN *FRAUD* PADA PAJAK MENGGUNAKAN *DECISION TREE*

APPLICATION DEVELOPMENT IN TAX FRAUD DETECTION USING DECISION TREE

Natasya Mulyono[1], Shaufiah, S.T., M.T.[2]
[1,2]Program Studi Teknik Informatika, School of Computing, Telkom University, Bandung
[1]natasya.mulyono@gmail.com, [2] shaufiah@telkomuniversity.ac.id

Abstrak

Pajak merupakan kontribusi wajib kepada negara yang terutang oleh orang pribadi atau badan yang bersifat memaksa berdasarkan Undang-Undang, dengan tidak mendapatkan imbalan secara langsung dan digunakan untuk keperluan negara bagi sebesar-besarnya kemakmuran rakyat. Terkadang pada data pajak ditemukannya *fraud.* Dengan semakin berkembangnya perkembangan teknologi informasi maka *fraud* berkembang semakin luas, dikarenakan hal itu perlu dilakukan pencegahan yaitu dengan langkah pendeteksian *fraud.* Sehingga, tugas akhir ini dibuat untuk dapat membantu dalam proses pendeteksian *fraud*

Dalam Tugas Akhir ini dilakukan klasifikasi untuk pendeteksian *fraud* pada data pajak menggunakan *Decision Tree* dengan algoritma C4.5. C4.5 menggunakan konsep Entropy untuk menentukan penyebaran keragaman data dan *gain ratio* untuk pemilihan atribut yang mana attribute dengan *gain ratio* tertinggi akan terpilih sebagai *parent* untuk *node* selanjutnya.

Hasil penelitian menunjukkan bahwa pendeteksian *fraud* menggunakan metode *Decision Tree* dengan algoritma C4.5 menghasilkan rata-rata persentase sebesar 99,51%. Sehingga dapat disimpulkan bahwa algoritma ini cocok melakukan penelitian untuk pendeteksian *fraud* pada data pajak.

Kata Kunci : Pajak*, Decision Tree, Pruning, C4.5, Fraud detection.*

Abstract

Tax is people's sompulsive contribution to State that payable by individual or organization that it's self is forced by law, by not getting the rewards directly and used for the purposes of the state for the greatest prosperity of the people. Sometimes the discovery of fraud in tax data. With the development of information technology, the fraud spread more widely, because it needs to be done, namely the prevention of the fraud detection step. Thus, this thesis is made to be able to assist in the detection of fraud

In this final project is done the classification for the detection of fraud in the use of tax data with an algorithm C4.5 Decision Tree. C4.5 uses the concept of entropy to determine the spread of the diversity of data and gain ratio for the selection of attributes which attribute to gain the highest ratio will be selected as a parent to the next node.

The results showed that the detection of fraud using the C4.5 Decision Tree algorithm generates an average percentage of 99.51%. It can be concluded that the algorithm is suitable to do research for the detection of fraud in the tax data.

Keywords: Tax, Decision Tree, Pruning, C4.5, Fraud detection.

1. Pendahuluan

Pajak merupakan kontribusi wajib kepada negara yang terutang oleh orang pribadi atau badan yang bersifat memaksa berdasarkan Undang-Undang, dengan tidak mendapatkan imbalan secara langsung dan digunakan untuk keperluan negara bagi sebesar-besarnya kemakmuran rakyat [1]. Kedudukan hukum pajak adalah menganut paham imperatif yaitu pelaksanaanya tidak dapat ditunda [2]. Terkadang pada data pajak ditemukannya *fraud*. *Fraud* atau penggelapan dalam KUHP diatur pada Buku II tentang Kejahatan terhadap Harta Kekayaan, yaitu berupa penyerangan terhadap kepentingan hukum orang atas harta benda yang dimilikinya. Dalam konsep *Data Mining, fraud* merupakan sebuah kejadian yang ditemukan pada anomali data dan pattern. Dari keseluruhan *dataset* normal, jumlah *fraud* lebih sedikit dari keseluruhan total *dataset* normal, hal ini menjadi kendala tersendiri dalam pendeteksian, yang menyebabkan *fraud* sulit untuk dideteksi sehingga menyebabkan kerugian cukup besar. Dengan semakin berkembangnya dunia teknologi informasi maka *fraud* semakin berkembang luas sehingga menyebabkan kerugian finansial yang sangat besar. Maka diperlukan *fraud detection*. Adapun teknik-teknik yang digunakan untuk *fraud detection* adalah teknik statistik, kecerdasan buatan(*artificial intelligent*) maupun teknik *Data Mining*.

Data Mining merupakan sebuah proses menemukan pola dari sebuah set data secara otomatis atau semi otomatis di mana pola yang dihasilkan memberikan beberapa keuntungan [3]. Salah satu metode atau teknik *Data Mining* dalam melakukan *fraud detection* yang digunakan adalah *Decision Tree*.

Decision Tree memiliki beberapa kelebihan yaitu, membutuhkan sedikit persiapan data, mampu mengolah data numerik maupun kategorikal, menggunakan mode *whitebox*, memungkinkan untuk memvalidasi model dengan menggunakan uji statistik, dan menghasilkan performansi yang baik jika menggunakan *dataset* yang besar [4]. Oleh karenanya pada tugas akhir ini menggunakan metode *Decision Tree* dan Algoritma C4.5.

Penulis membuat tugas akhir yang berjudul Pembangunan Aplikasi Pendeteksian *Fraud* Pada Pajak Menggunakan *Decision Tree* diharapkan berguna untuk membantu auditor dalam mendeteksi *fraud* yang terjadi pada praktisi perpajakan dan juga membantu bagian perpajakan mendeteksi *fraud* pada data pajak. Pada tugas akhir ini dibuat suatu model *Decision Tree* untuk pendeteksian *fraud*. Untuk membantu klasifikasi data, penulis menggunakan algoritma C4.5. Alasan mengapa penulis menggunakan algoritma C4.5 dikarenakan algoritma ini merupakan salah satu algoritma yang cocok untuk mendeteksi *fraud*, hal ini dibuktikan dari penelitian yang dilakukan oleh Y. Sahin dan E. Duman(2011) [4], yang menyatakan dari hasil pendeteksian penipuan pada kartu kredit bahwa hasil pengklasifikasian oleh *decision tree* lebih bagus dibandingkan hasil pengklasifikasian oleh SVM(*Support Vector Machine).*

2. Landasan Teori

2.1 *Fraud Detection*

Fraud atau penggelapan/penipuan dalam KUHP diatur pada Buku II tentang Kejahatan terhadap Harta Kekayaan, yaitu berupa penyerangan terhadap kepentingan hukum orang atas harta benda yang dimilikinya. *Fraud* bisa melibatkan satu orang atau lebih. Dalam konsep, *fraud* merupakan sebuah kejadian yang ditemukan pada anomaly data dan pattern. Adapun *fraud* yang terjadi pada dataset dapat disebabkan oleh kelalaian petugas ataupun admin dalam proses pemasukkan data, untuk direkap sehingga menhasilkan *fraud-fraud* yang tidak terduga. Dari keseluruhan dataset normal, jumlah *fraud* lebih sedikit dari keseluruhan total dataset normal, hal ini menjadi kendala tersendiri dalam pendeteksian, yang menyebabkan *fraud* sulit untuk dideteksi sehingga menyebabkan kerugian cukup besar. Sehingga dibutuhkan metode pencegahan terjadinya *fraud*, salah satunya adalah dengan sistem *fraud detection* Tujuan dari sistem *fraud detection* adalah untuk memeriksa setiap transaksi untuk kemungkinan menjadi *fraud* terlepas dari mekanisme pencegahan, dan mengidentifikasi *fraud* secepat mungkin setelah penipu sudah mulai melakukan transaksi *fraud*.

Adapun tipe-tipe *fraud* yang dijabarkan berdasarkan kedudukan hukum pajak yang bersifat imperatif, yakni pelaksanaannya tidak dapat ditunda [2], adalah sebagai berikut :

1) Jika Wajib Pajak telat dalam membayar pajak dan juga membayar denda tetapi jumlah yang dibayarkan tidak sesuai dengan jumlah denda yang seharusnya dibayarkan atau tidak membayar denda.
2) Jika Wajib Pajak memiliki tanggungan untuk membayar pajak tetapi tidak membayar pajak
3) Jika Wajib Pajak melakukan pembayaran denda pada suatu bulan tetapi tidak ada tanggungan untuk membayar pajak pada bulan tersebut.

2.2 Pajak

Berdasarkan Buku Ketentuan Umum dan Tata Cara Perpajakan, pajak merupakan kontribusi wajib kepada negara yang terutang oleh orang pribadi atau badan yang bersifat memaksa berdasarkan Undang-Undang, dengan tidak mendapatkan imbalan secara langsung dan digunakan untuk keperluan negara bagi sebesar-besarnya kemakmuran rakyat. Di Indonesia diterapkan 2 jenis pajak yaitu Pajak Pusat dan Pajak Daerah. Berdasarkan Pasal 1 angka 10 Undang-undang Nomor 28 Tahun 2009 Tentang Pajak Daerah dan Retribusi Daerah ("UU No. 28 Tahun 2009"), definisi Pajak Daerah adalah kontribusi wajib kepada daerah yang terutang oleh orang pribadi atau badan yang bersifat memaksa berdasarkan Undang-undang, dengan tidak mendapatkan imbalan secara langsung dan digunakan untuk keperluan daerah bagi sebesar-besarnya kemakmuran rakyat.

2.2.1 Pajak Restaurant

Pajak Restaurant merupakan pajak yang dikenakan atas pelayanan yang telah disediakan oleh restaurant. Sebagaimana yang didefinisikan dalam Peraturan Daerah Kabupaten Situbondo Nomor 4 Tahun 2011 Tentang Pajak Daerah, Restoran adalah fasilitas penyedia makanan dan/atau minuman dengan dipungut bayaran, yang mencakup juga rumah makan, kafetaria, kantin, warung, bar, dan sejenisnya termasuk jasa boga/katering. Dimana Obyek Pajak yang tidak diwajibkan untuk membayar pajak restoran adalah palyanan yang disediakan restoran yang nilai penjualannya kurang dari Rp 75.00,00(Tujuh Puluh Lima Ribu Rupiah) setiap bulannya [5]. Adapun tarif pajak yang dikenakan adalah sebesar 10%(Sepuluh Persen).

Untuk cara perhitungan pajak itu sendiri berdasarkan rumus dibawah ini:

perhitungan pajak=10%*Dasar Pengenaan Pajak (2.1)

2.3 *Data Mining*

Data Mining merupakan sebuah proses dalam menemukan pola dari sebuah set data yang mana proses tersebut harus otomatis atau semi otomatis yang mana pola yang dihasilkan memberikan beberapa keuntungan [3]. *Data Mining* sering disebut juga sebagai *Knowledge Discovery In Database(KDD).*

2.4 Klasifikasi

Klasifikasi merupakan proses yang bertujuan untuk menemukan sebuah himpunan model (atau fungsi) yang menggambarkan dan membedakan kelas-kelas data atau berbagai konsep.

2.5 *Decision Tree*

Decision Tree merupakan salah satu model prediksi yang mana bisa mempresentasikan baik model klasifikasi maupun regresi [5]. Teknik ini terdiri dari kumpulan *decision node*, yang dihubungkan oleh cabang, yang mana bergerak dari root *node* hingga berakhir di leaf *node*.

Decision tree merupakan salah satu model klasifikasi yang cukup terkenal. *Decision tree* merupakan metode diskriminasi nonlinear yang menggunakan sekumpulan variabel independen untuk membagi sampel ke dalam kelompok-kelompok yang lebih kecil secara bertahap.

2.6 Algoritma C4.5

Algoritma C4.5 merupakan merupakan salah satu algoritma dari *Decision tree*. C4.5 merupakan penyempurnaan dari algoritma terdahulu yaitu Algoritma ID3 yang dikembangkan oleh Ross Quinlan pada tahun 1987 [6].

Secara umum algoritma C4.5 membangun *decision tree* dengan tahapan sebagai berikut :

- Pilih atribut sebagai akar
- Buat cabang untuk setiap nilai
- Bagi kasus dalam cabang
- Ulangi proses untuk setiap cabang sampai semua kasus pada cabang memiliki kelas yang sama.

Dalam algoritma C4.5, *gain ratio* digunakan untuk pemilihan atribut yang akan diproses. Yaitu secara heuristik dipilih attribut yang menghasilkan simpul yang paling bersih(purest). Suatu cabang disebut pure, jika cabang pada *decision tree* anggotanya berasal dari satu kelas. Ukuran purity dinyatakan dengan tingkat impurity. Salah satu kriteria *impurity* Kriteria yang digunakan adalah *information gain*. Maka dalam memilih atribut untuk memecah obyek dalam beberapa kelas harus kita pilih atribut yang menghasilkan *gain ratio* paling besar [7]. Ukuran *gain ratio* digunakan untuk memiih attribut uji pada setiap *node* di dalam *tree*. Ukuran ini digunakan untuk memilih attribut atau *node* pada *tree*. Attribut yang memiliki gain ratio tertinggi akan terpilih sebagai *parent* bagi *node* selanjutnya [8].

2.6.1 Entropi

Entropi merupakan suatu variabel untuk mengukur tingkat homogenitas distribusi kelas dari sebuah himpunan data (data set). Sebagai ilustrasi, snipsemakin tinggi tingkat entropi dari sebuah data set maka semakin homogen distribusi kelas pada data set tersebut. Jika distribusi probabilitas dari kelas didefinisikan dengan S = (s1, s2, s3, ..., sk) maka entropi dapat dirumuskan sebagai berikut :

$$Entropy(s_1, s_2, s_3, \dots, s_m) = -\sum_{i=1}^{m} pi * log_2(pi) \quad (2\text{-}1)$$

Keterangan :
- S = himpunan kasus
- s1 = jumlah kasus
- Pi = jumlah sampel untuk kelas i
- M = jumlah nilai pada atribut target

2.6.2 *Information gain*

Setelah melakukan Splitting atau membagi *dataset* berdasarkan sebuah atribut kedalam subset yang lebih kecil, maka entropi dari data akan berubah. Perubaha entropi ini dapat digunakan untuk menentukan kualitas pembagian data yang telah diilakukan. *Information gain* dapat disebut sebagai perubahan entropi. Secara matematis information gain dari suatu attribut A dapat dirumuskan sebagai berikut:

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

Keterangan :
- S = Himpunan Kasus
- A = Atribut
- N = jumlah partisi atribut A
- |Si| = Jumlah kasus pada partisi ke-i
- |S| = Jumlah kasus dalam S

2.6.3 *Gain ratio*

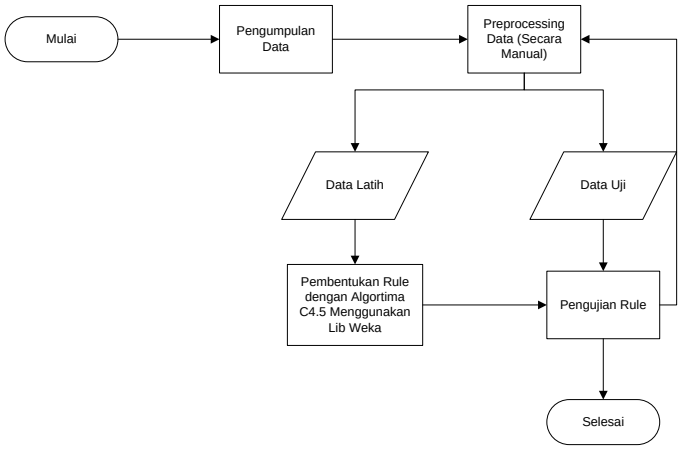
Gain ratio merupakan normalisasi dari *information gain* yang menghitung entropi dari distribusi probabilitas subset setelah dilakukan pembagian. Berikut formula dari *gain ratio* :

$$GainRatio(X) = \frac{Gain(x)}{SplitInfo(x)} \quad (2\text{-}3)$$

3. Perancangan Sistem

3.1 Gambaran Umum Sistem

Pembuatan system ini dilakukan untuk mendeteksi adanya *fraud*. Pada system ini metode yang digunakan pada penelitian ini adalah metode Eksperimen yang tahapannya mengacu pada tahapan Knowledge Discovery in Database (KDD) [9], dan disusun berdasarkan tahapan penelitian Haryanto [8].



Gambar 3 1 Gambaran Umum Sistem

Pada Gambar 3-1 dapat dijelaskan bahwa sistem dimulai dengan melakukan pengumpulan data lalu dilanjutkan ke *preprocessing* data, dari *preprocessing* data data dibagi menjadi data latih dan data uji, dari data latih proses dilanjutkan dengan pembentukan *rule* dengan algoritma C4.5, setelah itu dilakukan pengujian *rule*, dari pengujian data didapat tingkat akurasi data.

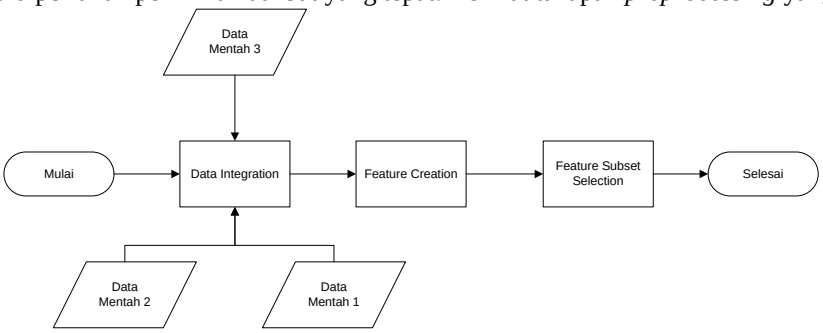
3.2 Data

Data yang digunakan pada Tugas Akhir ini merupakan data pajak restoran dan rumah makan daerah Situbondo periode 2012-2014.

Sebelum digunakan data pajak restoran dan rumah makan melalui tahap processing. Berikut tahap *preprocessing* data yang biasa dilakukan :Sebelum digunakan data pajak restoran dan rumah makan melalui tahap processing. Berikut tahap *preprocessing* data yang biasa dilakukan :

3.3 *Preprocessing* Data

Preprocessing data yang dilakukan pada tugas akhir ini dilakukan diluar sistem yang mana menggunakan software microsoft excel. Hal ini dilakukan dikarenakan data mentah yang didapat tidak dapat langsung dimasukkan sehingga diperlukan pemilihan atribut yang tepat. Berikut tahapan *preprocessing* yang dilakukan:



Gambar 3 1 : Tahap Preprocessing Data

3.3.1 Data Integration

Pada tahap *preprocessing* yang dilakukan pada tugas akhir ini dilakukan integrasi data yang bertujuan untuk menggabungkan data mentah data pajak dan data penerimaan denda pada tahap ini dilakukan cross check antara data yang tertera pada data pembayaran pajak dan penerimaan denda untuk mencocokkan dan memasukkan data denda yang sesuai dengan nama restoran dan bulan pembayaran yang dilakaukan oleh wajib pajak sehingga dihasilkan data yang berisikan atribut denda serta keterangannya.

3.3.2 Feature creation

Salah satu tahap *preprocessing* pada tugas akhir ini adalah feature creation dikarenakan diperlukan beberapa atribut baru yang diperlukan dalam sistem, yang mana atribut-atribut tersebut memiliki kegunaan yaitu berguna untuk menghitung jumlah keterlambatan wajib pajak dalam membayar pajak serta untuk menentukan apakah wajib pajak terlambat dalam membayar pajak, berguna untuk menunjukan apakah wajib pajak memiliki tanggungan pajak, berguna untuk menunjukan apakah wajib pajak ada melakukan pembayaran pajak, berguna untuk menunjukkan apakah wajib pajak memiliki tanggungan untuk membayar denda, berguna untuk menunjukkan apakah jumlah denda yang dibayarkan oleh wajib pajak sesuai dengan jumlah tanggungan denda yang seharusnya, dan berguna untuk menunjukan kelas *fraud* pada data. Atribut-atribut baru tersebut diberi nama JATUH TEMPO, TANGGUNGAN PAJAK, BULAN TELAT, KETERANGAN, PEMBAYARAN PAJAK, TANGGUNGAN DENDA, KESESUAIAN DENDA, dan CLASS.

Pada atribut JATUH TEMPO data pada atribut ini bertipe interval dan berisikan tanggal jatuh tempo untuk pembayaran pajak, yang mana formatnya adalah mmm-yy. Untuk atribut TANGGUNGAN PAJAK diberi label

ADA untuk data yang memiliki tanggungan untuk membayar pajak dan diberi label TIDAK ADA untuk data yang tidak memiliki tagihan untuk membayar pajak. Pada atribut BULAN TELAT merupakan atribut yang menjabarkan seberapa lama wajib pajak terlambat dalam membayarkan pajak yang seharusnya dibayarkan, data dalam atribut ini bertipe numerik. Pada atribut KETERANGAN berisikan label TELAT dan TIDAK TELAT, di mana label TELAT menunjukkan bahwa wajib pajak terlambat dalam melakukan pembayaran pajak dan TIDAK TELAT kebalikannya. Pada atribut PEMBAYARAN PAJAK memiliki label ADA untuk data yang menunjukkan bahwa wajib pajak melakukan pembayaran pajak, dan label TIDAK ADA untuk data yang menunjukkan bahwa wajib pajak tidak meelakukan pembayaran pajak. Atribut TANGGUNGAN DENDA memiliki label ADA untuk data yang mana wajib pajak terlambat dalam membayar pajak dan TIDAK ADA untuk data yang menunjukkan bahwa tidak ada keterlambatan dalam membayar pajak. Atribut KESESUAIAN DENDA memberikan label NO pada data yang menunjukkan bahwa jika wajib pajak terlambat dalam membayar pajak dan denda yang dibayarkan tidak sesusai dengan jumlah denda yang seharusnya dibayarkan maapun wajib pajak yang tidak membayar dendanya, label YES diberikan kepada data yang menunjukkna bahwa wajib pajak tidak melakukan keterlambatan dalam membayar pajaknya maupun membayar denda sesuai dengan denda yang menjadi tanggungannya. Atribut CLASS berisikan label yang menunjukkan kelas *fraud* dari suatu data di mana label CLASS0 menunjukkan bahwa data tersebut tidak *fraud*, label CLASS1 menunjukkan bahwa data tersebut merupakan data *fraud* tipe 1, label CLASS2 menjunjukan bahwa data tersebut merupakan data *fraud* tipe 2, dan label CLASS3 yang menunjukann bahwa data tersebut merupakan data *fraud* tipe 3.

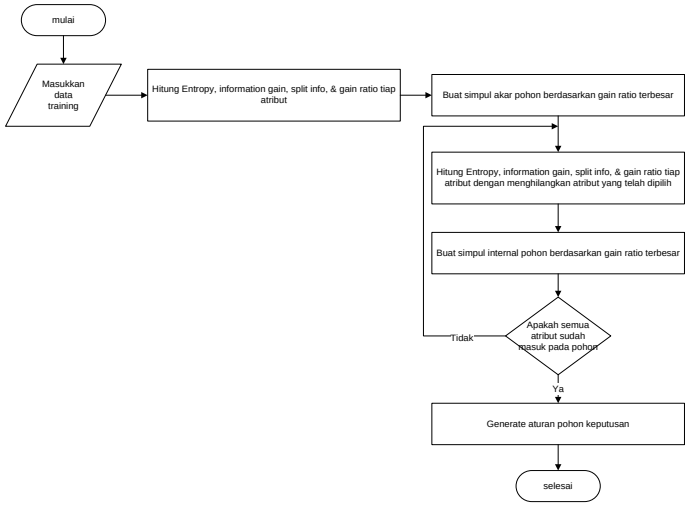
3.3.3 Feature Subset Selection

Pada tugas akhir ini perlu dilakukan *feature subset selection* dikarenakan banyaknya atribut-atribut yang tidak diperlukan sehingga harus dilakukan pemilihan dan pengurangan dimensi data. Teknik *feature subset selection* yang digunakan dalam tugas akhir ini adalah menggunakan pendekatan *filter* dimana *feature* dipilih sebelum algoritma *Data Mining* dijalankan. Berikut ini adalah *irrelevant attributes* yang dihilangkan beserta alasannya

a. Atribut Nomor dihilangkan dikarenakan atribut ini hanya menunjukkan nomor urut sehingga tidak memberikan informasi yang berguna.
b. Atribut Nama dihilangkan dikarenakan menjaga kerahasiaan identitas wajib pajak yang mana datanya digunakan pada tugas akhir ni.
c. Atribut Nomor SKP dihilangkan dikarenakan atribut ini tidak memiliki hubungan untuk *fraud detection*.
d. Atribut Jatuh Tempo dihilangkan karena atribut ini tidak digunakan dalam program.
e. Atribut Nominal dihilangkan karena atribut ini menunjukkan jumlah pajak yang dibayarkan serta realisasinya tidak digunakan dalam program
f. Atribut Lunas BL dihilangkan karena atribut ini tidak dapat digunakan dalam program.
g. Atribut Bulan Telat dihilangkan karena atribut ini tidak dapat digunakan dalam program.
h. Atribut Denda Yang Harus Dibayar dihilangkan karena dari atribut ini diekstrak informasi sehingga menghasilkan atribut baru yang dapat digunakan dalam program, sehingga atribut ini tidak diperlukan lagi.
i. Atribut Denda Yang Dibayar dihilangkan karena dari atribut ini diekstrak informasi sehingga menghasilkan atribut baru yang dapat digunakan dalam program, atribut ini juga bersifat continu sehingga tidak bisa diproses dalam algoritma C4.5 sehingga atribut ini tidak diperlukan lagi.

3.4 Pembentukan *Rule* dengan Algoritma C4.5(Klasifikasi)

C4.5 menghasilkan *tree* dengan jumlah cabang yang berbeda-beda. Pembentukan *rule* yang disusun mengikuti perancangan algoritma Haryanto(2013) serta proses klasifikasi Musyaffa (Musyaffa, 2014) adalah sebagai berikut:

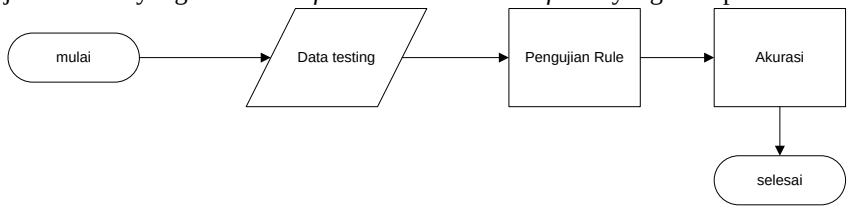


Gambar 3-2: Flowchart Pembentukan *Rule* dengan Algoritma C4.5

Pada tahapan ini pada dimulai dengan menghitung nilai Entropy, lalu hitung nilai *Information gain* menggunakan nilai entropy yang telah dihitung, lalu hitung nilai Split Info dari tiap atribut, dan setelah itu hitung nilai *Gain ratio* menggunakan nilai *information gain* dan split info. Setelah mendapatkan *gain ratio* dilakukan pembuatan simpul akar *tree* berdasarkan nilai *gain ratio* terbesar. Kemudian hilangkan atribut yang dipilih sebelumnya dan ulangi perhitungan nilai entropy, *information gain*, split info, dan *gain ratio*, kemudian atribut yang memiliki atribut yang memiliki *gain ratio* terbesar dijadikan *node*. Ulangi hingga semua atribut *tree* memiliki kelas. Jika semua *tree* memiliki kelas, maka generate *rule decision tree*.

3.5 Pengujian *Rule*

Pada tahapan ini dilakukan pengujian *rule* terhadap *rule* yang telah di generate, dengan menggunakan data testing yang telah disiapkan. Pada tahapan ini didapatkan akurasi dari *rule*, jika tingkat akurasi rendah maka akan dilakukan pruning, jika tingkat akurasi yang didapat tinggi maka akan dilakukan pemunculan interface yang menunjukkan jumlah data yang terindikasi *fraud* dan data bukan *fraud* yang terdapat dalam data testing.



Gambar 3 4 : Flowchart Pengujian Rule

4. PENGUJIAN DAN ANALISIS

4.1 Skenario Pengujian

Dalam tugas akhir ini pengujian dilakukan pada data pajak restoran dan rumah makan daerah situbondo periode 2012-2013, yang telah melewati preprocessing dengan skenario pengujian seperti berikut :

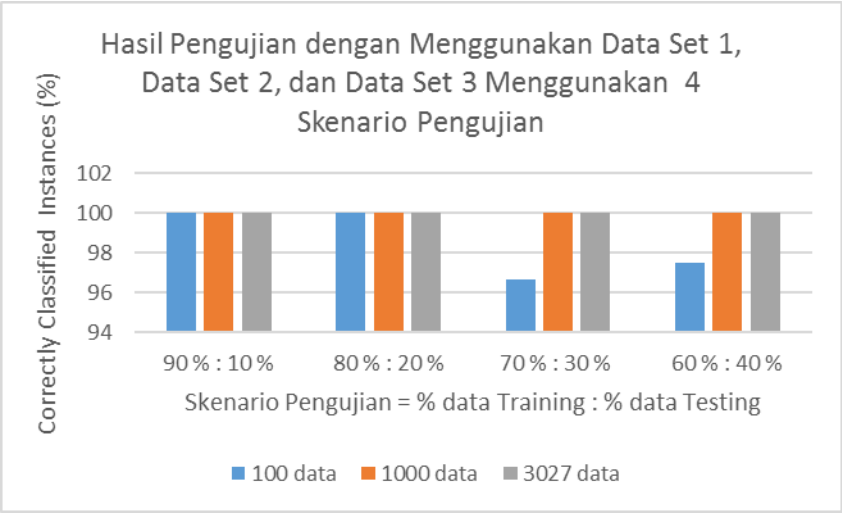
1. Melakukan perhitungan performansi dari rule yang telah terbentuk dengan melakukan perngujian terhadap 3 dataset yang mana telah melewati proses percentage split. Perhitungan performansi dihitung menggunakan Confusion Matrix yang terintegrasi pada weka dimana akan menghasilkan Correctly Classified Instaces . Confusion Matrix merupakan suatu metode yang biasanya digunakan untuk melakukan perhitungan akurasi pada konsep data mining. Adapun skenario pengujian yang dilakukan untuk perhitungan performansi ini menggunakan 4 skenario percentage split yaitu :
   a. Data keseluruhan akan dibagi menjadi 60% data Training dan 40% data Testing. Dimana data Training yang didapat akan digunakan untuk proses pembentukan rule dan data Testing digunakan untuk pengujian rule
   b. Data keseluruhan akan dibagi menjadi 70% data Training dan 30% data Testing Dimana data Training yang didapat akan digunakan untuk proses pembentukan rule dan data Testing digunakan untuk pengujian rule
   c. Data keseluruhan akan dibagi menjadi 80% data Training : 20% data Testing Dimana data Training yang didapat akan digunakan untuk proses pembentukan rule dan data Testing digunakan untuk pengujian rule
   d. Data keseluruhan akan dibagi menjadi 90% data Training : 10% data Testing. Dimana data Training yang didapat akan digunakan untuk proses pembentukan rule dan data Testing digunakan untuk pengujian rule .
2. Melakukan Pendeteksian *Fraud* yang terdiri dari 2 skenario pengujian yaitu :
   a. Menggunakan data yang tidak memiliki label, lalu dilakukan pendeteksian *fraud* dengan menggunakan rule yang telah didapatkan dari perngujian sebelumnya.
   b. Menggunakan data pembayaran pajak, dimana bagian dari data pajak yang perlu dimasukkan adalah No SKP, Omset dari wajib pajak yang melakukan pembayaran pajak, serta jumlah pembayaran pajak yang dilakukannya.

Tabel 4 1 : Spesifikasi Data Yang Diuji Cobakan

Data Set	Jumlah Data
Data Set 1	100
Data Set 2	1000
Data Set 3	3027

4.2 Hasil Pengujian

Dari pengujian pembuktian sistem didapatkan bahwa sistem mampu mengklasifikasikan data yang telah dibagi menjadi 3 dataset dengan menggunakan precentage split yang akan menghasilkan 4 skenario pengujian. Dimana pengujian yang dilakukan menghasilkan grafik yang menunjukkan persentasi dari Correctly Classified Instances yang didapat dari Confusion Matrix untuk setiap dataset dari setiap skenario pengujian.



Gambar 4 1 : Hasil Pengujian dengan Menggunakan Dataset 1, Dataset 2, dan Dataset 3 Menggunakan Skenario Pengujian 90% data Training : 10% data Testing , 80% data Training : 20% data Testing, 70% data Training : 30% data Testing, dan 60% data Training : 40% data Testing

Dari tabel 4.1 didapatkan penjelasan bahwa dilakukan pengujian dengan skenario pengujian menggunakan percentage split dimana data dibagi menjadi 90% data Training dan 10% data Testing, 80% data Training dan 20% data Testing, 70% data Training dan 30% data Testing, dan 60% data Training dan 40% data Testing. Dimana data yang digunakan merupakan dataset 1 yang berjumlah 100 data, dataset 2 yang berjumlah 1000 data, dan dataset 3 yang berjumlah 3027 data. Pada tabel 4.1 menunjukkan perbandingan Correctly Classified Instances yang didapat dari pengujian menggunakan skenario pengujian yang telah dijelaskan pada bab 4.2 menggunakan 3 dataset yang telah disiapkan. Correctly Classified Instances didapat dengan menggunakan Confusion Matrix yang di generate oleh sistem. Dari tabel 4.1 diketahui bahwa hasil pengujian dengan perbandingan data 90% data Training dan 10% data Testing serta 80% data Training dan 20% data Testing, untuk setiap dataset menghasilkan Correctly Classified Instaces terbesar yaitu 100%. Tetapi untuk pengujian dengan perbandingan data 70% data Training dan 30% data Testing serta 60% data Training dan 40% data Testing, terjadi penurunan hasil Correctly Classified Instaces khususnya terhadap dataset 1, yang berkurang menjadi 96,667% pada perbandingan data 70% data Training dan 30% data Testing dan mengalami kenaikan hasil Correctly Classified Instaces yaitu 97,5% pada perbandingan data 60% data Training dan 40% data Testing. Besarnya presentase correctly classified yang didapat, disebabkan beberapa faktor, salah satunya faktor jumlah data training yang banyak menyebabkan kondisi dan rule yang terbentuk dapat lebih menangani variasi dari data testing. Selain itu hal ini juga dapat disebabkan oleh jumlah data testing yang sedikit sehingga membuat presentasenya menjadi sangat besar. Akan tetapi, kualitas data dari data training juga perlu diperhatikan. Karena kualitas data yang baik dapat menghasilkan rule yang baik dan akurasi yang dihasilkan akan semakin bagus.

Dari pengujian yang dilakukan didapatkan Confusion Matrix untuk setiap skenario pengujian yang mana salah satu Confusion Matrix yang akan di analisa pada pembahasan ini adalah Confusion Matrix dengan dataset 2 menggunakan skenario pengujian percentage split 90 % data Training dan 10 % data Testing, maka didapat Confusion Matrix seperti berikut

Tabel 4-1 : Confusion Matrix dari dataset 2 hasil dari skenario pengujian 90 % data *training* : 10 % *data testing*

class0	class1	class2	class3	
74	0	0	0	class0
0	24	0	0	class1
0	0	2	0	class2
0	0	0	0	class3

Dari Confusion Matrix diatas tidak terdapat data yang terklasifikasi dengan label CLASS3. Hal ini dapat disebabkan karena jumlah data yang memiliki label CLASS3 sangat sedikit sehingga kemungkinan data tersebut tidak termasuk kedalam data testing sangat besar.

5. KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan dari hasil pengujian yang telah dilakukan didapatkan kesimpulan sebagai berikut :

1. Metode Decision Tree dengan menggunakan algoritma C4.5 dapat diimplementasikan dalam pembangunan aplikasi klasifikasi *fraud* detection yang mana menghasilkan rata-rata persentase akurasi yaitu 99,51% dari skenario pengujian yang dilakukan menggunakan dataset yang telah disiapkan
2. Data mentah dari pembayaran pajak restoran maupun rumah makan yang direkap tidak dapat digunakan secara langsung untuk melakukan *fraud* detection, dikarenakan tipe data pajak yang bersifat continous dan acak, serta tidak cocok untuk studi kasus yang dijabarkan, sehingga diperlukan preprocessing data terlebih dahulu agar data dapat digunakan dalam proses data mining.

5.2 Saran

Saran dari penulis untuk pengembangan yang dapat dilakukan untuk penelitian pendeteksian *fraud* pada pajak adalah :

1. Perlunya penambahkan data penunjang dalam pendeteksian *fraud* pada data pajak, dimana data yang ditambahkan dapat berupa data rincian pengeluaran dan penghasilan harian dari wajib pajak. Data tersebut selanjutnya diproses lebih dalam untuk mendapatkan atribut baru yang dapat digunakan untuk penelitian lebih lanjut.
2. Menambahkan beberapa atribut lain seperti pembayaran denda, realisasi pajak, tunggakan, dan atribut lain yang dapat membantu dalam pendeteksian *fraud* pada pajak.

Daftar Pustaka

[1] Republik Indonesia, Undang-Undang Nomor 16 tahun 2009 tentang perubahan keempat atas Undang-Undang Nomor 6 tahun 1983 tentang Ketentuan Umum dan Tata Cara Perpajakan pada Pasal 1 ayat 1, Jakarta: Sekretariat Negara, 2009.

[2] Mardiasno, Perpajakan Edisi Revisi 2011, Yogyakarta: Penerbit Andi, 2011.

[3] I. H. Witten, E. Frank and M. A. Hall , Data Mining: Practical Machine Learning Tools and Techniques, Third Edition (The Morgan Kaufmann Series in Data Management Systems), Singapura: Morgan Kaufmann, 2011.

[4] S. Y and D. E, "Detecting Credit Card Fraud by Decision Trees and Support Vector Machines," *Proceeding of the International MultiConference of Engineers and Computer Scientist 2011 Vol I*, 2011.

[5] L. Rokach and O. Maimon, Data Mining With Decision Tree: Theory and Application, Singapura: World Scientific Publishing Co. Pte. Ltd., 2008.

[6] J. R. Quinlan, "Simplifying Decision Trees [catatan penelitian]," *International Journal of Man-Machine Studies*, pp. 221-234, 1987.

[7] B. Santosa, DATA MINING : Teknik Pemanfaatan Data untuk Keperluan Bisnis, Yogyakarta: Graha Ilmu, 2007.

[8] T. Haryanto, Prediksi Penyakin Demam Berdarah dan Typhus dengan Algoritma C5.0, Bandung: Universitas Telkom, 2013.

[9] M. K. Jiawei Han, Data Mining Concept and Tehniques., San Fransisco: Morgan Kauffman, 2006.

[10] P. C. González and J. D. Velásquez, "Characterization and detection of taxpayers with false invoices using data mining techniques," *Expert Systems with Applications*, p. 1427–1436, 2013.

[11] H. Sharma, "Detection of Financial Statement Fraud Using Decision Tree Classifiers," IIT Delhi, New Delhi, 2013.

[12] Suyanto, Artificial Intelligence: Searching, Reasoning, Planning and Learning, Bandung: Penerbit Informatika, 2011.

[13] T. F. Musyaffa, Simulasi Klasifikasi Hujan Wilayah Kota Bandung Dengan Metode Decision Tree Menggunakan Algoritma C4.5, Bandung: Universitas Telkom, 2014.

[14] Kusrini and E. T. Luthfi. , Algoritma Data Mining, Yogyakarta: C.V. Andi Offset, 2009.

[15] Pemerintah Daerah Kabupaten SituBondo, PERATURAN DAERAH KABUPATEN SITUBONDO NOMOR 4 TAHUN 2011 TENTANG PAJAK DAERAH, Situbondo: Sekretariat Daerah Kabupaten Situbondo, 2014.