Contemporary Mathematics and Applications
Vol. 4, No. 1, 2022, pp. 51-71.

Pembentukan Model Pohon Keputusan pada Database *Car Evaluation* Menggunakan Statistik *Chi-square*

Retno Maharesi

Departemen Teknik Informatika, Fakultas Teknologi Industri, Universitas Gunadarma

Corresponding author: rmaharesi@staff.gunadarma.ac.id

Abstract. The study discusses problems related to the formation of a decision tree based on a collection of evaluation data records obtained from a number of car buyers. This secondary data was obtained from the UCL machine learning website. The purpose of this research is to produce a prototype algorithm for obtaining an inductive decision tree based on *Chi-square* statistics. An inductive decision tree formation method based on the *Chi-square* contingency test was compared with a decision tree obtained using a machine learning algorithm which was done using RapidMiner5 software. The work to produce an inductive decision tree was carried out by first processing data using Microsoft excel and next processed using SPSS software, on the crosstabs descriptive menu. The results of the two methods provide some kind of similar rules, in terms of the order of priority of the variables that most influencing people's decision to accept an automotive product. The formation of the decision tree uses a random sampling of size 300 data records among 1729 respondent data records in the car evaluation database. The resulting decision tree should have a minimal structure like a binary tree. This is possible because its formation is based on the statistical inferential method, so it does not require a separate pruning process as an addition step in the C4.5 algorithm, which actually this algorithm also considers aspects of the statistical significance.

Keywords:*C4.5 algorithm, Chi-square Statistics, data mining, Entropy, inductive decision tree.*

1 Pendahuluan

Studi mengenai pohon keputusan banyak dilakukan oleh berbagai kalangan masyarakat untuk membantu proses pengambilan keputusan di berbagai bidang kehidupan. Pembentukan pohon keputusan ditujukan untuk mend**a**patkan informasi yang belum terungkap secara jelas, yang terdapat dalam sebuah data berbentuk tabulasi *record* data. Salah satu algoritma pembentuk pohon keputusan semisal C4.5 yang menggunakan konsep *Entropy* dalam pembentukan cabang-cabangnya, banyak diaplikasikan untuk membangun berbagai sistem informasi sebagai contoh adalah, penilaian kredit perbankan [1], identifikasi faktor pemicu kecelakaan kerja [2], deteksi website apakah melakukan *phising* kepada para pengunjungnya [3]. Algoritma dengan konsep agak berbeda yaitu *Clasification and Regression Tree* atau CART digunakan untuk membangun pohon keputusan menggunakan sampel kecil dan lima variable [4], sampel pada penelitian tersebut merupakan *subset*dari *Car evaluation database*. CART mengelompokkan data menjadi dua bagian berdasarkan tebakan awal kandidat cabang awal kelompok kiri dan

kanan. Tebakan awal tersebut diperbarui menggunakan ukuran kesesuaian (*goodness*) berdasarkan nilai formula *distance measure* sederhana secara rekursif sebanyak level dari pohon keputusan yang dihasilkan.

Isu dari pemodelan data adalah bagaimana menghasilkan model yang tidak berlebihan (*overestimate*) atau terlampau sederhana (*underestimate*). Karena model yang *overestimate* terkendala dengan masalah pengintepretasian, sedang model yang *underestimate* memberikan tingkat akursi kurang baik. Untuk menjawab permasalahan tersebut, suatu solusi rasional adalah meggunakan komputasi Statistika. Sebagai contoh, [5] dalam penelitiannya menggunakan pendekatan *Chi-square* untuk melakukan pemangkasan simpul dari pohon keputusan yang dibangun menggunakan algoritma ID3, yaitu algoritma pembentukan pohon keputusan berbasis konsep *Entropy* sebelum algoritma C4.5.

Konsep *Entropy* dalam Fisika sering digunakan membangun struktur cabang dari model pohon keputusan. Sebagai contoh, algoritma C4.5 merupakan pengembangan dari beberapa varianalgoritma pembangun pohon keputusan sebelumnya yang sama-sama berbasis konsep *Entropy*, seperti *ID3* [6]. Pertanyaan yang muncul adalah, apakah tidak lebih beralasan jika konsep yang menjelaskan dinamika kumpulan materi (variabel) dapat disubstitusi dengan konsep keterkaitan antar variabel, seperti pada contoh kasus penerimaan konsumen pada suatu produk kendaraan, dapat dipahami aspek keamaan lebih diutamakan dari pada kenyamanan. Sehingga prioritas pertama adalah memberikan input pada variabel yang mengukur aspek keamanan, setelah itu dapat diikuti oleh aspek kenyamanan, kemudian ukuran atau kapasitas kendaraan, kemudian biaya perawatan dan mungkin yang terakhir adalah harga beli kendaraan. Berdasarkan penjelasan ini maka maka uji independensi/ dependensi antar variabel yang menggunakan statistic *Chi-square* dapat dimanfaatkan dalam rangka membangun rasionalitas suatu keputusan melalui pembentukan sebuah pohon keputusan.

Keputusan yang rasional dilandasi oleh sejumlah kondisi (nilai variabel-variabel) yang tingkat kepentingannya berhirarki untuk mewakili proporsionalitas dalam berargumentasi. Berdasarkan alasan tersebut, cukuplah beralasan untuk meyakini bahwa terdapat dependensi antara variabel input yang tingkat kontribusinya dalam menghasilkan suatu keputusan mengikuti suatu model hirarki, dengan struktur pohon (*tree*). Dalam hal penentuan urutan hirarki menggunakan konsep dependensi antar variabel input dan target (keputusan), dapat digunakan nilai koefisien kontingensi sebagai ukuran tingkat keeratan hubungan antar variabel. Variabel dengan nilai koefisien kontingensi terbesar dipilih sebagai prioritas utama dalam membangun kandidat semua premis yang memberikan kesimpulan berupa nilai suatu keputusan. Premis berikutnya yang menjadi bawahan premis utama dipilih secara rekursif menggunakan proses serupa. Selanjutnya informasi tersebut diperoleh dengan cara melibatkan metode pengujian hipotesis statistika dalam proses penentuan simpul-simpul yang membangun sebuah pohon keputusan. Apabila

ketersediaan data mencukupi, maka penemuan informasi yang mewakili populasi aktualnya menjadi tidak terkendala. Oleh karena *database Car evaluation* memiliki jumlah data yang besar, maka dapatlah dikatakan mewakili data populasi responden pengguna kendaraan.

Tujuan penelitian ini adalah membuat sebuah prototype algoritma yang dapat menghasilkan pohon keputusan induktif dengan struktur pohon yang seminimal mungkin di mana konsep *Entropy* dalam penentuan tingkatan keterkaitan antara atribut-atributnya atau variabel-variabel input yang memberikan nilai variabel keputusan disubstitusi dengan statistic *Chi-square*.

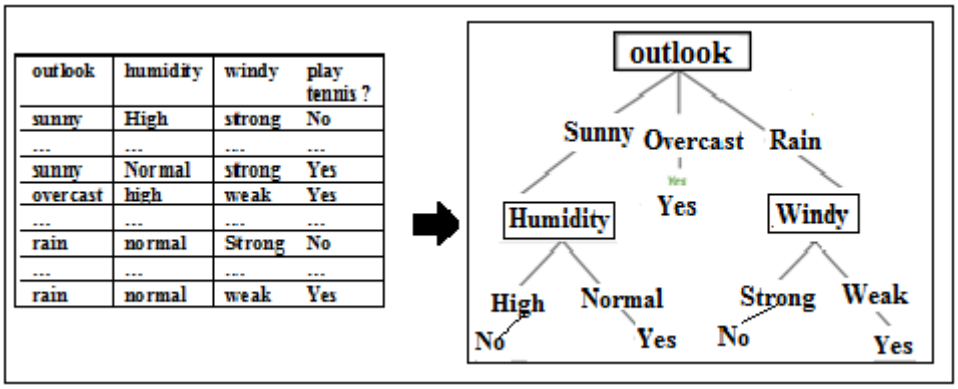
Lingkup pembahasan penelitian adalah pertama menghasilkan suatu algoritma pembentukan pohon keputusan sebagaimana pohon keputusan berdasarkan dengan konsep *Entropy* informasi untuk menentukan percabangannya. Namun dalam hal ini proses tersebut disubtitusi dengan analisis tabulasi silang (*crosstabs*) yang terdapat dalam menu *software* Statistika SPSS. Sebelum diekspor ke SPSS, terlebih dahulu data sampel sebanyak n = 300 yang dipilih secara acak menggunakan *uniform random number generator* dari data *Car Evaluation* yang memuat 1729 data responden diolah menggunakan Ms.excel. Kedua membandingkan hasil pohon keputusan yang diperoleh dari proses manual di atas dengan pohon keputusan yang dihasilkan dari software data mining seperti RapidMiner 5 menggunakan parameter yang ditetapkan sesuai default programnya.

2 Tinjuan Pustaka

Pada bagian ini dijelaskan beberapa tinjauan metode yang digunakan untuk penyelesaian permasalahan.

2.1 Pohon Keputusan

Data Mining merupakan metode untuk mengidentifikasi adanya pola keteraturan dan suatu bentuk hubungan dalam kumpulan data berukuran besar. Data yang dapat diproses dengan *data mining* diantaranya adalah *database* dan *data warehouse*. Bagaimana pola keteraturan yang terdapat dalam suatu data berukuran besar dapat ditemukan adalah dengan cara merepresentasikan data tersebut ke suatu struktur hirarki berupa pohon keputusan. Struktur berbentuk pohon keputusan dapat memfasilitasi analisis pengambilan keputusan yang kompleks menjadi lebih sederhana secara hirarkis berdasarkan urutan prioritas faktor yang berperan dalam menentukan nilai dari variabel keputusan. Pada gilirannya, pohon keputusan tersebut digunakan untuk menemukan hubungan tersembunyi antara sejumlahkan di data variabel input terhadap sebuah variabel target atau keputusan. Pada pohon keputusan untuk model prediksi,struktur pohon berbentuk berhirarki seperti pada Gambar 1.



Gambar 1 Model Pohon Keputusan (keputusan main tennis atau tidak) berdasarkan data kasus

Pada diagram pohon di atas simpul akar (*out look*) maupun internal {*humidity*, *wind*} berfungsi sebagai kondisi atau syarat. Masing-masing simpul bertindak sebagai variabel masukan $\{x_1, x_2, \ldots, x_m\}$ yang mempunyai nilai berupa status dari simpul-simpul tersebut. Jenis data yang mengisi nilai variabel masukannya dapat berupa nilai kategorik (nominal dan ordinal) atau juga numerik. Syarat tersebut harus dipenuhi sehingga sebuah jalur yang menghubungkan dimulai dari akar, kemudian simpul internal dan berakhir disimpul eksternal dapat merepresentasikan suatu logika implikasi. Simpul eksternal merepresentasikan nilai keputusan. dari variabel target *Y*. Variabel *Y* biasanya data berskala nominal, lebih sering berupa nilai binomial, sebagaimana pada contoh adalah {*yes*, *no*}. Berdasarkan Gambar 1,dapat dibuat pernyataan bahwa jika cuaca terlihat cerah dan kelembaban udara normal maka waktu yang baik untuk main tennis. Data untuk pembentukan sebuah pohon keputusan biasanya sudah dalam bentuk tabulasi, di mana kolom menunjukkan atribut atau variabel input dan variabel target sedangkan baris menyatakan unit pengamatan. Proses komputasi pembentukan pohon keputusan dilakukan dengan menemukan keterkaitan berupa hubungan hirarkis dalam tabel data menggunakan model diagram pohon yang selanjutnya diperoleh sejumlah aturan (*rule*).

Tahap pertama dari pembentukan pohon keputusan adalah menentukan simpul awal, berupa variabel yang paling berperan dalam menentukan output keputusan yang tersimpan dalam variabel keputusan *Y* [7], metode pemilihan simpul akar yang paling banyak digunakan menggunakan konsep *Entropy*

$$Entropy(S) = -\sum_{i=1}^{l} p_i \log_2 p_i \qquad (1)$$

dengan p_i adalah proporsi kasus *i= 1, 2, 3, …, l,* dengan *l* adalah banyaknya variabel input dalam ruang contoh *S*.

2.2 Algoritma C4.5

Secara umum, untuk membangun pohon keputusan, algoritma C4.5 membutuhkan empat langkah [8], yaitu; pertama, proses pemilihana tribute sebagai simpul awal (akar), kedua, pembentukan cabang dari simpul akar menurut banyaknya macam nilai dari atributnya, ketiga, mempartisi data sampel sesuai komposisi menurut macam dari nilai atribut, dan keempat, iterasi proses komputasi untuk masing-masing cabang sampai semua kasus pada cabang memiliki nilai variabel keputusan.

Algoritma C4.5 menggunakan gain ratio yang merupakan normalisasi dari *information Gain* menggunakan *split information* yang bertujuan untuk memperbaiki *information gain*, dengan rumus *GainRatio*:

$$GainRatio(S,A) = \frac{Gain(S,A)}{SplitInfo(S,A)} \quad , \qquad (2)$$

dimana S = ruang contoh yang digunakan untuk training, $Gain(S,A)$ = capaian informasi dari atribut A,

$$Gain(S,A) = Entropy(S) - \sum_{i=1}^{l} \left| \frac{S_i}{S} \right| Entropy(S_i) \qquad (3)$$

Split Info(S, A) adalah partisi informasi pada atribut A yang menyatakan *Entropy* atau informasi potensial yang dinyatakan sebagai Persamaan (4).

$$SplitInfo(S,A) = \sum_{i=1}^{l} -\frac{S_i}{S} \log_2 \left| \frac{S_i}{S} \right|, \qquad (4)$$

dengan S = ruang contoh yang digunakan untuk *training*, S_i = jumlah sampel untuk atribut i dan proporsi p_i disubstitusi dengan perbandingaan jumlah kasus S_i dalam ruang contoh S. Atribut dengan nilai *Gain Ratio* tertinggi menjadi prioritas untuk menjadi simpul akar dalam hirarki pohon keputusan, dengan gain berupa *information gain***.** Pendekatan menggunakan skala prioritas pada simpul akar untuk dilakukan pengujian pembentukan pohon keputusan melalui proses *training* seperti validasi silang dilakukan secara *heuristic*, karena dari hasil penelitian sebelumnya diketahui bahwa pembentukan suatu pohon keputusan optimal di mana semua simpulnya baik simpul akar, internal dan eksternal yang signifikan secara statistik merupakan *NP Complete problem* [7]. Karena alasan komplesitas komputasi yang bertumbuh secara eksponensial seiring dengan bertambahnya jumlah variabel, maka hanya pendekatan secara *heuristic* seperti *top down* yang bisa dilakukan dalam membangun sebuah pohon keputusan. Pendekatan *top down* dimulai dengan menentukan variabel atau simpul untuk menjadi akar sebuah pohon keputusan berdasarkan kriteria yang ditetapkan menggunakan konsep *Entropy* (algoritma C4.5) atau Koefisien Kontingensi pada seperti dalam penelitian ini. Algoritma C4.5 dikenal memiliki tiga fitur penting bagi sebuah pohonkeputusan, yaitu pemangkasan ranting, pembentukan aturan-aturan dan juga evaluasi aturan yang dihasilkan [7].

Pemangkasan pohon dilakukan dengan terlebih dahulu mengenali untuk kemudian menghapus cabang-cabang. Hal ini bertujuan untuk mengurangi informasi yang cenderung berlebihan atau *overfitting*. Pohon yang cabang-cabangnya dipangkas akan menjadi lebih mudah dipahami dan pemrosesan informasi klasifikasi menjadi lebih cepat. Dua metode pemangkasan dalam pembentukan pohon keputusan, yaitu:

1. *Prepruning* yang menghentikan pembangunan suatu *subtree* lebih awal, dengan cara menghentikan pembentukan cabang jika partisi yang akan dibuat tidak signifikan.
2. *Postpruning* dengan cara membuang beberapa cabang *subtree* setelah pohon selesai dibangun.

2.3 RapidMiner

Pengembangan perangkat lunak *open source* RapidMiner yang di mulai di Universitas Dortmund pada tahun 2001, dilanjutkan oleh Rapid-I GmbH sejak 2007 didasari oleh ketiga fitur penting pada algoritma *C4.5* yang disebut dibagian sebelumnya. Perangkat lunak untuk *data* mining digunakan secara meluas mengikuti fitur yang baik pada algoritma C4.5 dan juga karena antar muka bagi pengguna yang nyaman karena disain rancangan yang modular [9]. Kenyamanan ini yang memudahkan berinteraksi dengan system aplikasi lain seperti Ms. Excel dan SPSS (perangkat lunak untuk analisis data statistik). RapidMiner menyediakan fasilitas *model training* dan *testing* karena menerapkan prosedur validasi silang. Proses training untuk membentuk pohon keputusan misalnya menggunakan prosedur validasi silang 9:1. Prosedur validasi silang 9:1 untuk pembentukan pohon keputusan artinya dilakukan 10 kali proses training menggunakan 9 bagian data sampel dan pengujian menggunakan 1 bagian data sampel yang sudah diacak. Akurasi hasil keputusan dihitung secara rerata dari hasil *testing* terhadap 10 pohon keputusan. Dengan demikian dapat dikatakan galat hasil estimasinya lebih minimal mengingat galat berbanding terbalik dengan ukuran sampel [10].

2.4 Kaitan antara Information dan Uji Statistik *Chi-square*

Dikutip dari Attneave (1959), [7] menyatakan bahwa likelihood dari *InformationGain* dapat dituliskan sebagai berikut

$$G^2(A_i, S) = 2.\ln(2).|S|.InformationGain(A_i, S). \qquad (5)$$

Formulasi tersebut dapat digunakan untuk mengukur signifikansi dari kriteria *Information Gain* secara statistik. Jika variabel input atau atribut A_i untuk $i = 1,2,3,\ldots m$ dan variabel target Y independen, maka nilai statistik ujinya menyebar mengikuti *Chi-square* dengan derajat bebas bersesuaian dengan domain nilai atribut *Ai*-1 dan domain nilai variabel target *Y*-1 diruang sampel *S* yang merepresentasikan himpunan kasus. Berdasarkan pernyataan tersebut maka menjadi beralasan jika statistik uji *Chi-square*

dapat secara langsung digunakan untuk membangun sebuah pohon keputusan yang bersesuaian dengan himpunan kasus *S* yang biasanya tersedia dalam bentuk tabulasi *database*.

2.5 Statistik *Chi-square* untuk Pembentukan Pohon Keputusan Induktif

Uji hipothesis statistika mengenai adanya keterkaitan antara dua variabel (*X*, *Y*) menggunakan hipotesis nol, bahwa jika kejadian *X* dan *Y* independen maka peluangnya dapat dihitung sebagai $P(X \cap Y) = P(X).P(Y)$, tetapi jika terdapat keterkaitan antara kejadian *X* dan *Y* maka $P(X \cap Y) = P(X).P(Y|X)$ dengan *P*(*X*/*Y*) adalah peluang kejadian *X* bersyarat *Y*, sehingga penulisan uji hipotesis statistiknya menjadi:

$$H_0 : P(X \cap Y) = P(X).P(Y) \qquad vs \qquad H_1 : P(X \cap Y) \neq P(X).P(Y) \qquad (6)$$

Dari penulisan hipotesis ini, jika variabel X dan Y saling bebas maka jumlah kuadrat beda nilai frekuensi observasi O_{ij} dan frekuensi harapan f_{ij} mendekati nol, sehingga semakin besar nilai satistik X^2 maka hipotesis nul dapat ditolak, yang artinya kedua variabel berkaitan.

Untuk mengukur adanya dependensi diantara dua variabel tersebut digunakan statistik uji *Chi-square*:

$$\chi^2 = \sum_{I,J} \frac{(o_{ij} - f_{ij})^2}{f_{ij}} \ , \qquad (7)$$

dengan frekuensi observasi O_{ij} dan frekuensi harapan f_{ij} dihitung berdasarkan banyaknya kejadian di dalam ruang contoh *S* atau himpunan kasus. Tabel kontingensi diekstraksi dari tabel *database S* untuk menghasilkan sebuah tabel yang memuat sel (*i*,*j*) sebanyak *r*.*c*, dengan $i = 1,2,...,r$ dan $j=1,2,\ldots,c$. Dalam hal ini, nilai *r* dan *c* merepresentasikan banyaknya macam nilai dari variabel atau atribut yang diuji independensinya. Kemudian frekuensi harapan yang bersesuaian dengan sel (*i*, *j*), didapat dengan formula:

$$f_{ij} = P(X)P(Y)n = \frac{\sum_i r_i \sum_j c_j}{n}, \qquad (8)$$

dengan r_i, c_j adalah frekuensi terobservasi pada baris *i* dan kolom *j* pada table kontingensi. Setelah mengetahui terdapat hubungan di antara pasangan variabel yang signifikan secara statistik, maka koefisien kontingensi Cramer [10] dapat digunakan untuk mengukur tingkat keeratan hubungannya yang dinyatakan oleh Persamaan (9)

$$C = \sqrt{\frac{\chi^2_{(r-1)(c-1)}}{n(l-1)}} \quad , \tag{9}$$

dimana *n* adalah ukuran sampel, $\chi^2_{(r-1)(c-1)}$ adalah statistik *Chi-square* dengan derajat bebas (*r-1*)(*c-1*) dan $l = min\{r, c\}$.

Dalam simpulan atas penelitian yang dilandasi hasil [11] berkaitan dengan pertanyaan apakah penambahan suatu simpul (representasi suatu variabel masukan) memberikan keputusan penolakan hipotesis null yang signifikan secara statistik, [5] menyatakan pendekatan *Chi-square*yang digunakan untuk mengetahui apakah pemangkasan suatu simpul dalam pohon keputusan yang didapat menggunakan algoritma ID3 (generasi sebelum algoritma C4.5) tidak memengaruhi sebaran data kasus.

3 Metode Penelitian

Langkah dalam penelitian adalah:Pertama, kajian pustaka yang relevan dengan topik pembentukan pohon keputusan, berupa sejumlah algoritma klasifikasi *data mining* berbasis konsep *Entropy* seperti ID3 dan C4.5 maupun partisi biner seperti algoritma CART dari sejumlah jurnal dan buku-buku. Kedua, eksplorasi data dan mengimplementasikan algoritma pembangun pohon keputusan berdasarkan statistik *Chi-square* untuk data sampel dari *Car Evaluataion database*. Berikut adalah uraian bagian kedua dari metode yang digunakan:

1. Pengolahan data sampel yang semula terdiri atas 1729 kemudian dipilih secara acak sebanyak 300 responden menggunakan program pembangkit bilangan acak yang terdapat di Ms. excel. Dari hasil olah data tersebut, dilanjutkan dengan penggunaan software RapidMiner 5 untuk mendapatkan pohon keputusan. Hal ini dilakukan dengan cara mengimpor data dalam file Ms.excel, kemudian menetapkan parameter pemodelan pohon keputusan menggunakan *default* program RapidMiner.
2. Penggunaan *software* SPSS untuk menentukan simpul awal pohon keputusan dengan cara menjalankan sebanyak enam kali prosedur analisis deskriptif *crosstabs* untuk melakukan pencarian nilai *Chi-square* terbesar dan atau signifikan diantara enam pasang uji keterkaitan variabel terikat *Y* yang merepresentasikan penerimaan masyarakat terhadap produk otomotif mobil dengan enam riabel bebas {*buying, maint, doors, persons, lug_boot, safety*}.
3. Berdasarkan hasil pada langkah 3, data sampel dipartisi sesuai variabilitas dari nilai variabelnya untuk dibuatkan cabang bermula dari simpul awal. Data yang semula berukuran 300 diseleksi sesuai pengelompokan nilai variabel yang menghasilkan percabangan pohon keputusan. Prosedur ini dikerjakan menggunakan Ms. excel.

4. Berdasarkan hasil pada langkah 3, kembali digunakan *software* SPSS untuk menentukan simpul internal secara berurutan guna melakukan pencarian nilai *Chi-square* terbesar dan atau signifikan untuk menentukan simpul internal maupun eksternal.
5. Pengulangan langkah 3 dan 4 sampai tidak ada lagi penambahan simpul internal maupun eksternal.
6. Pembentukan diagram pohon keputusan dan pemangkasan simpul untuk menyederhanakan *rule* yang dihasilkan. Pada tahapan ini jumlah *rule* dapat direduksi dengan hanya mempertimbangkan simpul dengan nilai *Chi-square* yang signifikan secara statistik.
7. Melakukan perbandingan atas hasil pohon keputuan dari software RapidMiner 5 dan metode usulan berbasis *Chi-square*.

4 Data Penelitian

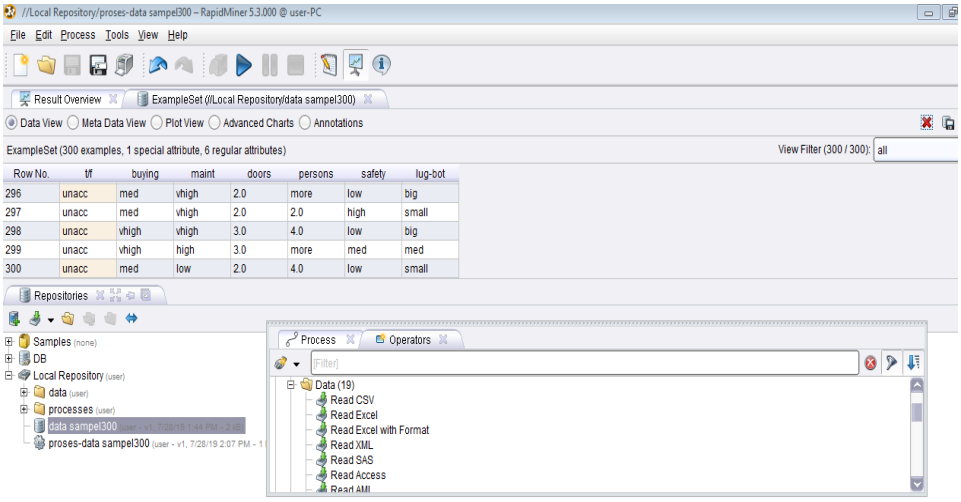
Data sekunder dalam penelitian ini adalah *Car Evaluation database* [11] yang semula oleh DEX, M. Bohanec, V. Rajkovic digunakan untuk membangun sebuah system pakar untuk penentuan keputusan berstruktur hirarki. Model untuk evaluasi kendaraan tersebut dirancang berdasarkan konsep struktur hirarki faktor atau atribut. Dapat dipahami bahwa penerimaan masyarakat terhadap produk otomotif mobil dipengaruhi oleh faktor biaya yang meliputi harga jual (*buying*) dan harga pemeliharaan (*maint*), karakteristik teknis yang mewujudkan kenyamanan menaiki mobil, yaitu jumlah pintu mobil (door), kapasitas penumpang (*person*), ukuran bagasi mobil (*lug-boot*) dan tingkat keamanan pengoperasian kendaraan (*safety*), shingga database memuat enam variabel bebas *buying, maint, doors, persons, lug_boot, safety* dan satu variabel terikat yang bertipe binomial yaitu *accepted* (gabungan dari kriteria *accepted, good, very good*) dan *unaccepted*.

Berikut adalah keragaman nilai dari masing masing variabel terikat dan bebas: label *acc* dan *unacc*, atribut *buying* meliputi: *vhigh, high, med* dan *low*, atribut *maint*: *vhigh, high, med, low*, atribut *doors*: *2, 3, 4, more*, atribut *persons*: 2, *4, more*, atribut *lug_boot*: *small, med, big*, atribut *safety*: *low, med, high*. Jumlah data sampel secara keseluruhan terdiri atas 1729 observasi dan untuk penelitian ini digunakan sampel berukuran 300 yang diambil secara random diantara 1729 data responden terkait penerimaan produk otomotif mobil berdasarkan enam variabel atribut dngan konsep harga, kenyamanan dan keamanan pegoperasian mobil.

5 Hasil Implementasi Software RapidMiner 5 dan Metode Usulan Berbasis Statistik *Chi-square*

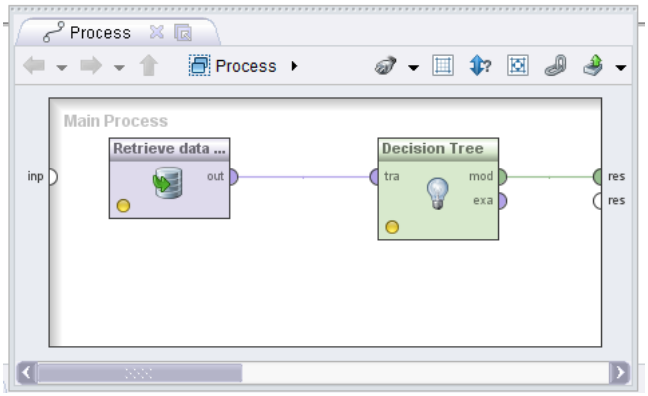
RapidMiner adalah suatu aplikasi yang khusus untuk menangani berbagai permasalahan dalam *data mining*, dalam hal ini permasalahan klasifikasi data evaluasi penerimaan konsumen terhadap produk kendaran. Importasi data dari Ms.Excel ke sistem RapidMiner dapat langsung dilakukan dengan mudah, serupa dengan mengimpor data ke software

SPSS untuk melakukan analisis Statistik. Gambar 2 adalah pohon keputusan hasil penerapan data sampel menggunakan software RapidMiner 5. Gambar-gambar berikut adalah *screenshoot* yang diambil dari proses pada RapidMiner sampai dihasilkan reperesentasi grafis pohon keputusan yang ditampilkan pada Gambar 2.



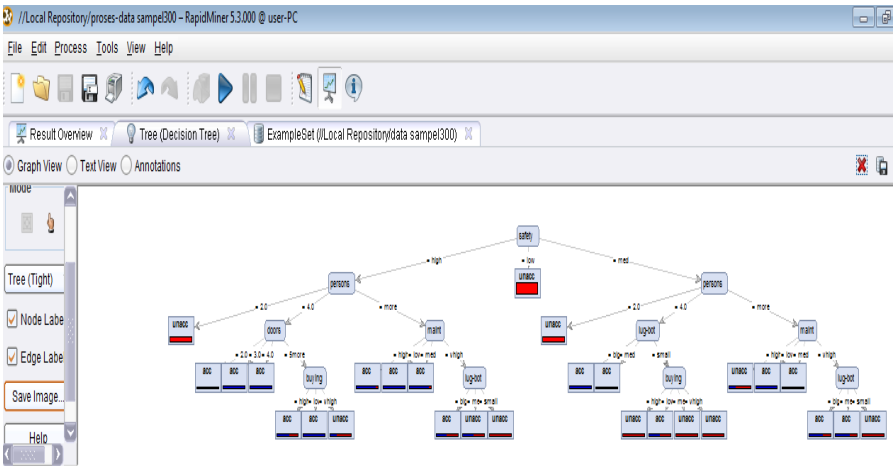
Gambar 2 Jendela yang menampilkan 300 data sampel *Car Evaluation dataset*

Tampilan pada Gambar 2 adalah tangkapan layar dari jendela RapidMiner dari lima *record* terakhir diantara *300* data sampel setelah data berhasil diimpor dari Ms. Excel.Setelah data tersimpan dalam RapidMiner, proses analisis dimulai dengan mengambil data dari *repository* dengan menu *drag* and *drop* pada software yang memvisualisasikan proses yang dipilih. Di sini digunakan visualisasi, proses *Retrieve data*, yang dilanjutkan dengan analisis data, dalam hal ini *Decision Tree* menggunakan parameter default. Gambar 3 adalah tangkapan layar untuk dua proses tersebut.



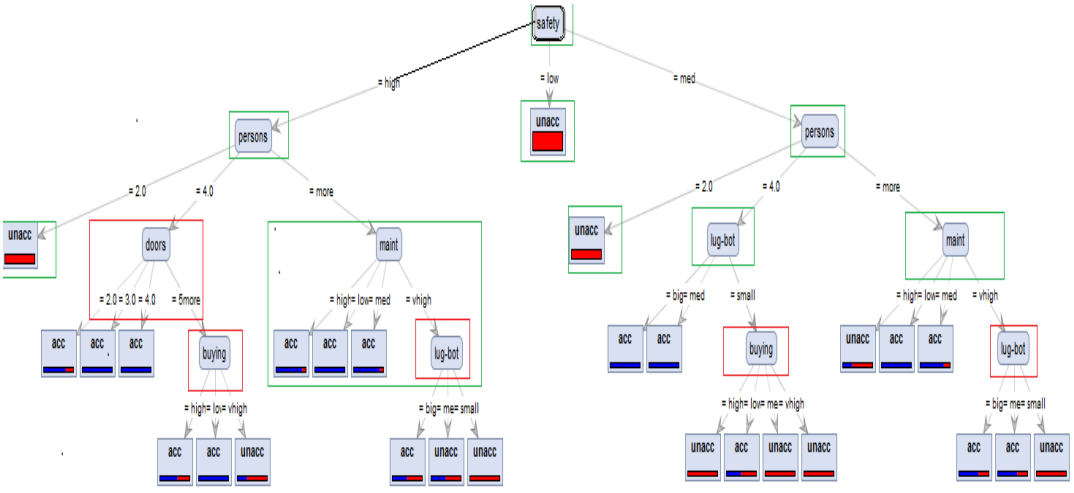
Gambar 3 Tangkapan layar jendela yang menampilkan dua proses sederhana di *RapidMiner*

Hasil dari proses analisis menggunakan Decision Tree berupa visualisasi grafis sebagaimana ditunjukkan oleh Gambar 4 atau visualisasi pohon keputusan dalam bentuk data text.



Gambar 4 Gambar jendela yang menampilkan visualisasi grafis (*tight*) dari pohon keputusan di RapidMiner

Langkah selanjutnya menentukan simpul untuk ditandai sebagai simpul eksternal yang signifikan kotak hijau dan simpul tidak signifikan secara statistik (kotak merah) secara manual.



Gambar 5 Pohon keputusan dengan RapidMiner 5dengan kotak hijau/ merah untuk simpul yang signifikan/tidak signifikan

Dengan menggunakan SPSS, pohon keputusan yang dihasilkan *software RapidMiner* dianalisis menggunakan menu statistika Deskriptif pada opsi pilihan *crosstabs* dengan terlebih dahulu memprosesnya di Ms. excel. Sebagai satu contoh adalah: Proses penghitungan nilai *Chi-square* untuk mengetahui signifikansi kebergantungan antara variabel input dengan kriteria {*safety =high, person = 4, buying* = high, med, low} dan variabel keputusan *Y* {*acc, unacc*} dilakukan menggunakan Ms. excel menggunakan formula *COUNTIFS* untuk pencacahan frekuensi *i* macam kejadian teramati, $O_{i=1,2,..}$. Hasilnya adalah sebuah tabel kontingensi dengan jumlah baris tiga dan kolom dua seperti ditunjukkan Tabel 1.

Tabel 1 Contoh tabel kontingensi level 3

	acc	*unacc*
safety =high, person = 4, buying = vhigh	o*safety =high, person = 4, buying* = vhigh, acc =*true*	o*safety =high, person = 4, buying* = high, acc = *false*
safety =high, person = 4, buying = high	o*safety =high, person = 4, buying* = high, acc =*true*	o*safety =high, person = 4, buying* = high, acc = *false*
safety =high, person = 4, buying = med	o *safety =high, person = 4, buying = med*, acc =true	o*safety=high, person = 4, buying = med*, acc =*false*
safety =high, person = 4, buying = low	o *safety =high, person = 4, buying = low, acc =true*	o*safety=high, person = 4, buying = low, acc = false*

Untuk simpul yang lain, dilakukan cara serupa pada contoh tersebut, yang selanjutnya dapat dihitung nilai *Chi-square* hitungnya menggunakan Persamaan (7-8) untuk dibandingkan dengan nilai tabel pada derajat bebas(*r*-1) dengan tingkat signifikan α. Hasil pengujian disajikan oleh Gambar 5 dengan kotak merah menunjukkan simpul yang tidak signifikan sedangkan kotak hijau menunjukkan simpul yang signifikan pada taraf nyata *5*%. Pada Tabel 2 disajikan *rule* atau aturan yang signifikan dan tidak signifikan.

Tabel 2 Beberapa aturan yang dihasilkan oleh pohon keputusan dengan parameter default *RapidMiner*

No	8 diantara 11 Rule yang signifikan	8 diantara 27 Rule yang tidak signifikan
1	Jika *safety* rendah maka ditolak	Jika *safety* tinggi penumpangnya 4 dan jumlah pintu = 2 sampai 4 maka diterima
2	Jika *safety* tinggi dan penumpangnya 2 maka ditolak	Jika *safety* tinggi penumpangnya = 4 dan jumlah pintu > 5 dan harga jual rendah sampai tinggi maka diterima
3	Jika *safety* tinggi penumpang > 4 dan biaya pemeliharaan rendah atau sedang atau mahal maka diterima	jika *safety* tinggi penumpangnya 4 dan jumlah pintu > 5 dan harga jual sangat tinggi maka ditolak.

No	8 diantara 11 Rule yang signifikan	8 diantara 27 Rule yang tidak signifikan
4	Jika *safety* tinggi penumpang >4 dan biaya pemeliharaan rendah atau sedang atau mahal maka diterima	Jika *safety* tinggi penumpang >4 dan biaya pemelihraan sangat mahal tetapi bagasi besar maka diterima sedangkan jika bagasi kecil atau sedang maka ditolak
5	Jika *safety* sedang dan penumpangnya 2 maka ditolak	Jika *safety* sedang dan penumpangnya 4 dengan ukuran bagasi kecil dan harganya sedang sampai sangat tinggi maka ditolak
6	Jika *safety* sedang dan penumpangnya 4 dan ukuran bagasi sedang atau besar maka diterima	Jika *safety* sedang dan penumpangnya 4 dan ukuran bagasi kecil dan harganya rendah maka diterima.
7	Jika *safety* sedang dan penumpangnya >4 dan biaya pemeliharaanya tinggi maka ditolak	Jika *safety* sedang dan penumpangnya >4 dan biaya pemeliharaanya sangat tinggi dan bagasinya sedang atau besar maka diterima
8	Jika *safety* sedang dan penumpangnya >4 dan biaya pemeliharanya rendah atau sedang maka diterima.	Jika *safety* sedang dan penumpangnya >4 dan biaya pemeliharaanya sangat tinggi dan bagasinya kecil maka ditolak

Selanjutnya akan dijelaskan proses pembentukan pohon keputusan menggunakan software SPSS 17. Persiapan data sampel mengunakan Ms. excel dilakukan dan selanjutnya *file* data diimpor ke software SPSS 17 untuk mendapatkan uji kontingensi setiap pasang variabel input dan variabel keputusan, (X_i, Y). Volume perhitungan secara manual sekalipun sudah menggunakan Ms. excel dan SPSS masih tergolong tinggi. Sebagai gambaran pada simpul akar diperlukan pemrosesan analisis *crosstab*s sebanyak jumlah variabel input, dan semakin meningkat sejalan dengan pertumbuhan cabang sebagai hasil variabilitas nilai variabel input. Banyaknya macam nilai varibel input berkisar antara 3 sampai 4. Pada kasus dalam penelitian ini digunakan 6 variabel input X_1=*buying*, X_2= *maint*, X_3 =*doors*, X_4 =*persons*, X_5=*lug_boot*, X_6 =*safety* dan satu variabel terikat Y yang bertipe binomial yaitu *accepted* (gabungan dari kriteria *accepted*, *good*, *very good*) dan*unaccepted*. Berikut ini adalah urutan proses untuk membangun sebuah pohon keputusan menggunakan uji independensi *Chi-square*:

1. Mempersiapkan data dalam bentuk Ms. *excel spread sheet* dalam bentuk tabel dengan kolom berupa variabel input dan variabel keputusan yang bertipe kategorik atau nominal untuk digunakan di program SPSS.
2. Melakukan uji kontingensi untuk setiap pasang variabel input dan variabel keputusan guna memilih variabel atribut dengan nilai *Chi-square* terbesar dan atau paling signifikan di tingkat signifikan α sebagai simpul akar.

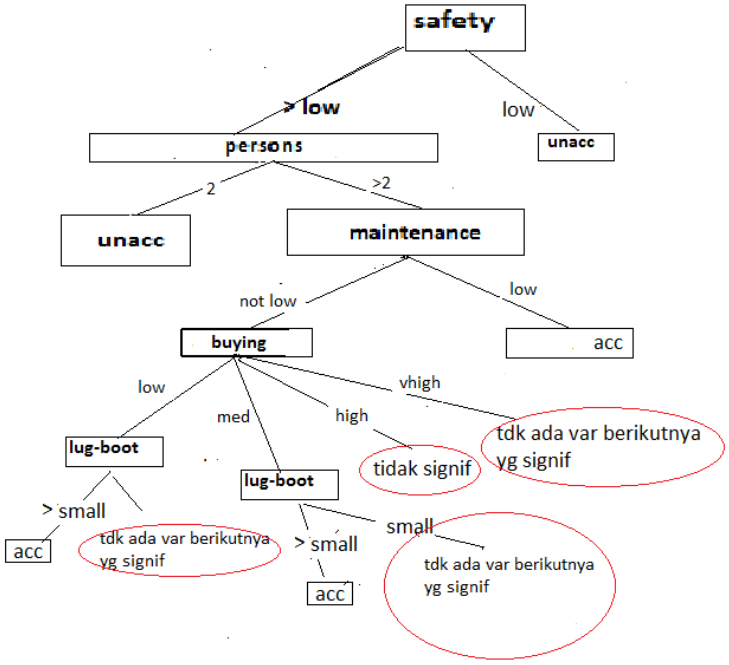
3. Berdasarkan hasil dari langkah 2 yaitu nilai *Chi-square* tertinggi, data sampe dipartisi berdasarkan banyaknya macam nilai variabel input untuk membangun cabang di level 2 sambil diidentifikasi apakah dalam tahap ini diperoleh simpul eksternal berupa nilai variabel keputusan.
4. Melakukan olah data di Ms. excel untuk mendapat data sampel guna membangun level pohon keputusan di level berikutnya
5. Pengulangan langkah 2-3 pada cabang yang memuat variabel keputusan yang tidak seragam nilainya terhadap variabel input yang tersisa saat olah data dengan Ms excel untuk membentuk level pohon berikutnya, yaitu level 3, 4 dan 5. Pengulangan berhenti sampai program SPSS tidak dapat menghasilkan nilai *Chi-square* karena sudah tidak ada cabang yang memuat nilai keputusan yang tidak seragam pada prosedur tabulasi silang yang dilakukan terhadap variabel input yang tersisa.

Tabel 3 memperlihatkan hasil analisis dari tabulasi silang yang signifikan pada level *5*% untuk setiap pasang variabel atribut dan keputusan.

Tabel 3 Koefisien kontingensi pasangan variabel target dan atribut

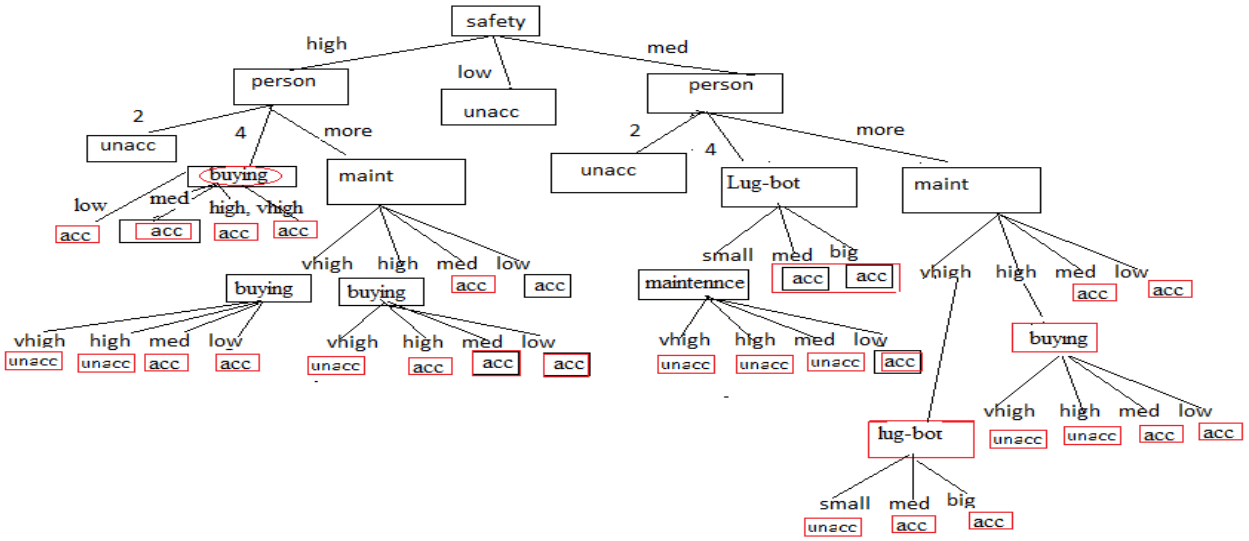
No	Variabel	*Chi-square contingensi*
1	*safety*	68.896
2	person	59.047
3	maint	14.555
4	doors	2.42
5	Lug-bot	3.770
6	buying	6.077

Untuk dapat melakukan uji independensi antara variabel atribut dan variabel keputusan guna membangun percabangan pada level yang lebih tinggi, terlebih dahulu data diolah menggunakan Ms. excel membentuk kolom kolom yang merepresentasikan kriteria dari variabel atribut dan variabel keputusan sesuai level dari pohon keputusan. Sebagai gambaran pada level 5 dihasilkan pohon keputusan sebagaimana diperlihatkan Gambar 6.



Gambar 6 Pembentukan pohon keputusan dimulai dari nilai atribut safety>low (*not low)*

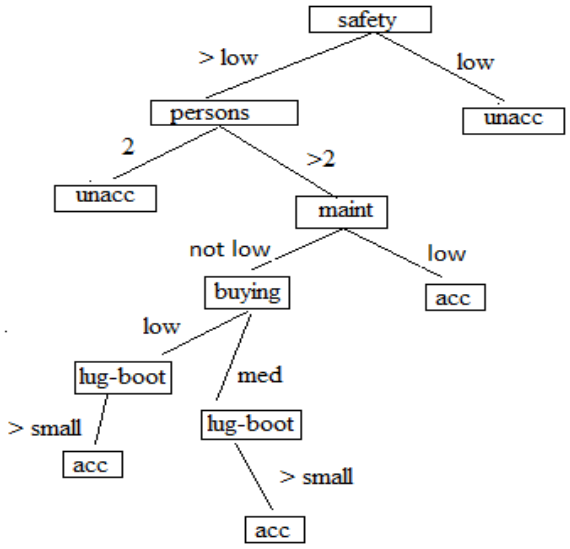
Gambar 7 menyajikan pohon keputusan secara utuh, terlihat simpul keputusan berupa persegi empat warna merah yang menunjukkan keputusan tidak signifikan pada level 5% sedangkan kotak hitam menunjukkan simpul yang signifikan pada tingkat 5%. Gambar 7 adalah hasil implementasi dari urutan langkah-langkah pembuatan pohon keputusan berbasis uji kontingensi antara variabel keputusan *Y*{*acc*, *unacc*} dan himpunan variabel atribut{$X_1, X_2, X_3, X_4, X_5, X_6$}. Pada pohon keputusan berbasis statistik *Chi-square*, semakin tinggi level percabangannya maka semakin banyak atribut atau variabel input yang dilibatkan dalam menentukan nilai keputusan dari variabel *Y*. Sebagai *contoh pada* Gambar 7, *simpul akar* (variabel *safety*) diikuti simpul internal (variabel *person*) pada cabang *high*. Hal ini berarti terdapat hubungan atau dependensi antara kedua variabel input yaitu *safety* dan *person* terhadap variabel keputusan *Y* yang signifikan secara statistik. Sehingga untuk level yang lebih dalam, yaitu dengan menyertakan variabel *buying* di level berikutnya, maka terdapat dependensi antara variabel; {*safety*, *person*, *buying*} pada cabang *high* terhadap variabel keputusan *Y*{*acc*, *unacc*}yang signifikan secara statistik.



Gambar 7 Pohon keputusan dengan percabangan tidak minimal berdasarkan statistik *Chi-square*, warna kotak merah menunjukkan simpul eksternal yang tidak signifikan

Aturan-aturan dalam bentuk *if*-then diturunkan dari pohon keputusan dengan melakukan penelusuran dari simpul akar sampai kesimpul eksternal. Banyaknya *rule* yang dihasilkan sebuah pohon keputusan sesuai deengan banyaknya simpul eksternal. Tata cara pembentukan rule adalah kata "*if*" diikuti dengan simpul akar kemudian diikuti simpul internal yang terletak dalam satu jalur menuju suatu simpul eksternal yang ditulis setelah kata "then", simpul eksternal merepresentasikan suatu keputusan. Prosedur *Chi-square* dengan data *training* sebanyak 300, memberikan 4 rule signifikan secara statistik diantara total 32 rule yang dihasilkan. *Rule* yang sigifikan menggunakan statistik *Chi-square* adalah:

1) Jika *safety* tinggi dan jumlah penumpang 2 maka ditolak.
2) Jika *safety* tinggi dan person >4 dan biaya perawatan rendah maka diterima.
3) Jika *safety* rendah maka ditolak.
4) Jika *safety* menengah dan jumlah penumpang *2* maka ditolak.



Gambar 8 Pohon keputusan dengan strukur biner sebagai hasil pengujian tabulasi silang dari pohon pada gambar 6

Pohon keputusan pada Gambar 8 didapat dengan mengusahakan agar jumlah percabanganminimal, tetapi tidak menutup kemungkinan menggunakan lebih dari dua percabangan, asalkan pembentukan cabang bersesuaian dengan operator logika seperti (komplemen, >, <, rentang nilai $a<x<b$). *Rule* yang dihasilkan dengan metode statistik *Chi-square* di mana semuanya signifikan pada level *5*% adalah sebagai berikut:

1) Jika *safety* rendah maka ditolak.
2) Jika *safety* sedang sampai tinggi dan penumpangnya 2 maka ditolak
3) Jika *safety* sedang sampai tinggi dan penumpangnya >2 dan biaya perawatan rendah maka diterima
4) Jika *safety* sedang sampai tinggi dan penumpangnya >2 dan biaya perawatan tidak rendah dan harga jualnya rendah serta ruang bagasi tidak kecil maka diterima.
5) Jika *safety* sedang sampai tinggi dan penumpangnya >2 dan biaya perawatan tidak rendah dan harga jualnya menengah serta ruang bagasi tidak kecil maka diterima.

6 Diskusi

Dua penelitian dengan topik serupa dengan penelitian ini yaitu [12] dan [13] menggunakan sumber data sekunder yang sama-sama diambil dari sebuah *web repository machine learning* [11]. Pembentukan pohon keputusan pada kedua penelitian tersebut berdasarkan pada algoritma *C4.5* dan *Clasification and Regresion Tree* (CART), di mana kedua algoritma sama-sama menggunakan konsep *Entropy*.

Pada bagian ini perbandingan struktur pohon keputusan yang dihasilkan melalui konsep *Entropy* yang melandasi proses dalam perangkat lunak RapidMiner dan pohon keputusan induktif berbasis statistik uji *Chi-square*, yaitu uji dependensi variabel atribut dan keputusan yang akan menempati posisi hirarki dalam sebuah pohon keputusan dijelaskan. Penentuan simpul akar pohon keputusan induktif dalam penelitian ini, berdasarkan hasil uji statistik *Chi-square* untuk setiap pasang atribut dalam himpunan data kasus *S*. Hasil uji yang signifikan pada taraf nyata 5% (berupa nilai maksimum koefisien kontingensi) menentukan atribut yang menjadi akar pohon keputusan. Pembentukan level pohon keputusan induktif serupa dengan proses pembentukan percabangan dengan konsep berbasis *Entropy* seperti di algoritma C4.5. Dalam artian, proses pembentukan mengikuti alur penempatan simpul secara *heuristic top down* (mulai dari simpul akar) dan rekursif ke bawah untuk memilih atribut di level berikutnya dalam suatu struktur pohon keputusan. Simpul akar ditentukan berdasarkan partisi optimal berdasarkan nilai atributnya. Proses penempatan simpul internal pada satu atau lebih cabang optimal beserta percabangannya diulang serupa dengan proses penentuan simpul akar beserta percabangannya, sampai pemisahan lebih lanjut tidak dapat dilakukan [14], [8].

Tabel 4 menyajikan hasil perbandingan dirangkum berdasarkan pembahasan pada bagian 5. Penilaian signifikansi dari *rule* yang dihasilkan dengan program aplikasi RapidMiner dilakukan dengan cara sebagai berikut: Pertama adalah menentukan semua *rule* yang dihasilkan, dalam hal ini terdapat 15 *rule*. Kedua menelusuri data yang relevan untuk pengujian berdasarkan jalur pada pohon keputusan yang membentuk masing-masing *rule*. Ketiga mengidentifikasi data olahan yang sesuai dengan jalur tersebut di mana data induk yang disimpan dalam format file Ms. excel perlu dilakukan proses seleksi data terlebih dahulu agar sesuai dengan jalur dari *rule* yang dihasilkan *RapidMiner* 5 pada *default parameter*. Keempat melakukan pengujian *crosstabs* data yang dihasilkan pada langkah tiga terhadap variabel keputusan *Y*.

Tabel 4 Perbandingan hasil pohon keputusan dengan tiga skema berbeda pada data sampel *n*=300

Metode dan atau model{A, B, C}	Uraian yang mengacu dari Gambar 5, 6 dan 7 berdasarkan kriteria: ada tidaknya kasus overlap pada *rule* yang dihasilkan, jumlah *rule* signifikan secara statistik yang dihasilkan, *rule of thumb* penentuan urutan/ hirarki variabel input yang menghasilkan simpul keputusan
A.RapidMiner 5 dengan parameter default	1. Banyak keputusan yang *overlap*, yaitu data training tidak murni memberikan nilai output dari salah satu variabel keputusan. Hal ini ditandai dengan prosentase warna berbeda yaitu merah (*acc*) dan biru (*unacc*) untuk satu keputusan. 2. Dari total 27 *rule* yang dihasilkan, hampir 20 *rule* tidak signifikan secara statistik pada level 5%. 3. Dari 27 *rule* tersisa 7 *rule* yang signifikan tanpa *overla*p.

	4. Urutan tingkat kontribusi variabel input dengan model A yang signifikan, yaitu *safety, persons, maint* dan *lug-boot*.
B. Model *Chi-square* dengan struktur pohon tidak dibatasi	1. Tidak banyak keputusan yang *overlap*, yaitu data *training* tidak murni memberikan output nilai dari salah satu variabel keputusan. 2. Jumlah *rule* sebanding dengan yang dihasilkan dengan *RapidMiner* dengan parameter default. 3. Metode *Chi-square* memberikan 4 rule signifikan secara statistik diantara total 32 rule yang dihasilkan. 4. Urutan tingkat kontribusi variabel input tidak berbeda jauh dengan model A, yaitu *safety*, *persons*, *maint* dan *lug-boot* sedang variabel *buying* dan *doors* tidak signifikan secara statistik.
C. Model *Chi-square* dengan struktur pohon dioptimalkan serendah mungkin jumlah percabangannya	1. Jumlah *rule* yang dihasilkan hanya 5, sehingga jumlah *rule* yang dihasilkan paling sedikit dibandingkan dua model lainnya, yaitu 27 dan 32. Namun semua *rule* signifikan secara statistik. 2. Urutan tingkat kontribusi variabel input yaitu *safety*, *persons*, *maint*, *buying* dan *lug-boot* yang semuanya signifikan secara statistik. 3. Model C mengisyaratkan bahwa jumlah pintu mobil bukan merupakan pertimbangan penting bagi para kustomer produk otomotif mobil.

Output dari RapidMiner memberikan 11 rule signifikan dengan 3-4 *rule* diantaranya *overlap* dari keseluruhan 27 *rule* yang dihasilkan. Lebih tingginya jumlah *rule* yang signifikan pada RapidMiner karena proses training menggunakan prosedur validasi silang 9:1. Hal ini berarti melakukan 10 kali proses *training* yang memungkinkan untuk memilih hasil terbaik diantara 10 *output* jika dibandingkan dengan prosedur *Chi-square* yang proses training hanya dilakukan satu kali. Perbandingan antara pohon keputusan yang dihasilkan oleh algoritma berbasis *Entropy* seperti dalam RapidMiner dan *Chi-square* dalam kasus data *Car evaluation*, terlihat dalam penambahan variabel *doors* (RapidMiner) dan *buying* (*Chi-square*) semata karena di satu pihak tidak menggunakan aspek pengujian signifikansi dari statistik dan di lain pihak menggunakan aspek tersebut. Dengan kata lain, atribut *buying* mempunyai signifikansi uji dependensi terhadap variabel keputusan yang signifikan dibanding atribut *doors*.

Dibandingkan dengan RapidMiner, pohon keputusan berdasarkan metode *Chi-square* memberikan jumlah *rule* signifikan yang lebih sedikit sekalipun semua variabel internal yang disertakan dalam pohon semuanya signifikan. Untuk mengatasi masalah ini sebaiknya digunakan sampel >300 misalkan semua *record* dalam database dijadikan sampel. Alasannya, semua suku pembilang pada nilai hitung *Chi-square* dapat dipandang sebagai kuadrat deviasi antara frekuensi teramati terhadap frekuensi harapan-nya, sehingga jika jumlah datanya diperbesar maka nilai pembilangnya ikut membesar dan

pada gilirannya menaikkan nilai hitung statistik *Chi-square*. Di mana hal ini berarti memudahkan hipotesis nol (independensi antara variabel atribut dan keputusan) untuk ditolak pada suatu taraf nyata yang diinginkan.

7 Simpulan dan Saran

Pembentukan pohon keputusan menggunakan metode *Chi-square* mempunyai hasil yang tidak berbeda secara signifikan dengan yang dihasillkan oleh RapidMiner 5 pada kondisi parameter *default*. Kedua metode memberikan jumlah *rule* jika-maka (*if-then rule*) yang kurang lebih sama banyak dan tidak semuanya signifikan secara statistik. Kemudian bila digunakan struktur pohon yang lebih sederhana yang cenderung berupa pohon biner, *rule* yang dihasilkan jumlahnya lebih sedikit namun mendukung konsep pohon keputusan yang induktif, yaitu pohon keputusan yang dapat mewakili data populasi. Satu kelebihan yang dapat diidentifikasi dari hasil penelitian ini, adalah tidak seperti metode pembentukan pohon keputusan berbasis konsep *Entropy*, metode uji statistik *Chi-square* sudah dengan sendirinya menyediakan proses pemangkasan. Sebagaimana diuraikan dalam bagian studi pustaka, algoritma C4.5 yang berbasis *Entropy*-pun melakukan proses pemangkasan untuk menghindarkan *overfitting decision tree* yang dihasilkannya.

Pada estimasi pohon keputusan menggunakan validasi silang 9:1, tentunya galat pendugaannya lebih kecil karena merupakan rerata dari 10 kali hasil estimasi dibandingkan dengan yang menggunakan satu kali pemilihan sampel acak seperti dalam penelitian ini. Namun untuk tingkat akurasi prediksi, baik pada model pohon keputusan yang diperoleh menggunakan proses *sampling* acak berulang seperti dalam skema validasi silang maupun satu kali sampel acak tidak dapat ditentukan akurasinya, karena nilai prediksi tidak terdapat dalam data sampel. Penelitian belum memuat tingkat akurasi dari hasil prediksi nilai variabel target berdasarkan masukan nilai variabel input atau atribut, hal ini merupakan tahapan yang dapat dilakukan untuk penelitian selanjutnya, baik menggunakan skema validasi silang maupun sampel acak tunggal.

8 Daftar Pustaka

[1] Hermanto B., Azhari S.N., 2017, Klasifikasi Nilai Kelayakan Calon Debitur Baru Menggunakan *Decision Tree* C4.5, *IJCCS*, 11, 43-54.

[2] Elisa E., 2017, Analisa dan Penerapan Algoritma C4.5 Dalam Data Mining Untuk Mengidentifikasi Faktor-Faktor Penyebab Kecelakaan Kerja Kontruksi PT.Arupadhatu Adisesanti, *Jurnal Online Informatika*, 2, 36-41.

[3] Salim T., dan Giap Y.C., 2017, Data Mining Identifikasi Website Phising Menggunakan Algoritma C4.5, *Jurnal TAM (Technology Acceptance Model)*, 8, 130-135.

[4] Astuti R., 2018, Data Mining Untuk Klasifikasi Dengan Algoritma CART (*Classification and Regression Trees*), *Media Informatika*, 17, 114-124.

[5] Patel N. dan Upadhyay S., 2012, Study of Various Decision Tree Pruning Methods

with their Empirical Comparison in WEKA, *International Journal of Computer Applications, 60, 09*75 –8887.
[6] Berry, M.W dan M. Browne,2006, Lecturer Note in Data Mining Singapore, World Scientific.
[7] Rokach L., Maimon O., 2010, Book Chapter 9: Data Mining and Knowledge Handbook.
[8] Jiawei H.,Kamber, Micheline dan Jian P., 2012, *Data Mining Concepts and Techniques Third Edition*. Elsevier Inc; Amsterdam.
[9] Sebastian L.dan Simon, F., 2012, *RapidMiner5*.Rapid-IGmbH; Dortmund.
[10] Walpole E., Myers R. H., Myers S. L., dan Ye K., 2012, Probability and Statistics for Engineers & Scientists, 9th Edition Pearson Education, Inc.
[11] https://archive.ics.uci.edu/ml/datasets/Car+Evaluation, diunduh tanggal 30 Mei 2019.
[12] Sugiatna E., Ibrahim A.M, Hadi I.A, 2019, Implementasi Algoritma Klasifikasi C4.5 Untuk Memprediksi Kelayakan Pembelian Kendaraan**,** *Jurnal Teknologi Informasi dan Multimedia*,1, 124-132.
[13] Astuti R., 2018, Data Mining Untuk Klasifikasi Dengan Algoritma CART (*Classification and Regression Trees*), *Media Informatika*, 17, 114-124.
[14] Gorunescu, F., 2011, Data Mining Concepts, Model and Technique. Berlin: Springer.