**Laporan Tugas Akhir MK. Data Mining (KOM332), Semester Genap 2018/2019**

**Klasifikasi Data *Census Income* oleh Biro Sensus Amerika (*United State Census Bureau*) tahun 1994-1995 dengan Menggunakan Metode Pohon Keputusan dengan Algoritma C4.5**

Muhammad Alif Nurrofli Lubis (G64160016), Nabila Rahmadani (G64160019), Gilang Gunawan (G64160060)

Kelompok 10, Kelas Paralel 2

**Abstrak**

Sensus merupakan proses mendapatkan informasi deskriptif tentang anggota sebuah populasi (tidak hanya populasi [manusia](https://id.wikipedia.org/wiki/Manusia)). Sensus dapat digunakan untuk [demokrasi](https://id.wikipedia.org/wiki/Demokrasi) ([pemilu](https://id.wikipedia.org/wiki/Pemilu)), pengumpulan [pajak](https://id.wikipedia.org/wiki/Pajak), juga digunakan dalam ilmu [ekonomi](https://id.wikipedia.org/wiki/Ekonomi). Salah satu contohnya adalah sensus yang dilakukan oleh Biro Sensus Amerika (*United State Census Bureau*) tentang populasi penduduk di Amerika. Sensus ini bertujuan untuk mendapatkan informasi terkait demografis penduduk dan juga terkait perekonomian di Amerika. Data yang digunakan merupakan data census yang diekstraksi dari Survei Populasi pada tahun 1994 dan 1995 yang dilakukan oleh Biro Sensus Amerika. Data tersebut merupakan data yang dapat digunakan untuk proses klasifikasi. Data ini akan diolah menggunakan metode Pohon Keputusan dengan algoritma C4.5. rpart dan party merupakan *library* yang digunakan dalam proses pengolahan data sampai terbentuk suatu model. Hasil tersebut menunjukan bahwa *library* party memiliki akurasi yang lebih baik yaitu 93% sedangkan jika menggunakan *library* menghasilkan akurasi 92%. Dari kedua *library* tersebut menghasilkan root pada tree yang berbeda diantara keduanya. Jika menggunakan rpart atribut capital\_gain adalah atribut yang paling berpengaruh untuk tingkat pendapatan penduduk Amerika sedangkan saat menggunakan *library* party atribut education merupakan atribut yang paling berpengaruh. Dalam kasus ini *library* rpart lebih baik dibandingkan *library* party karena dapat membuat tree yang lebih ringkas tetapi hanya berbeda 1% saja untuk akurasinya sehingga dapat menyingkat waktu pemrosesan data. Hasil prediksi ini dapat dijadikan acuan pemerintah untuk fokus terhadap peningkatan mutu masyarakat Amerika dalam meningkatkan income sehingga GDP dari negara tersebut dapat meningkat. Selain itu, hasil prediksi ini juga dapat menjadi pelajaran penting untuk masyarakat Amerika jika ingin meningkatkan income dalam setahun.

Kata kunci: *Data Mining*, *Census Income Bureau*, C4.5, Pohon Keputusan, *Classification*, *rpart , party*.

# **PENDAHULUAN**

## **Latar Belakang**

Sensus merupakan proses mendapatkan informasi deskriptif tentang anggota sebuah populasi (tidak hanya populasi [manusia](https://id.wikipedia.org/wiki/Manusia)). Sensus dapat digunakan untuk [demokrasi](https://id.wikipedia.org/wiki/Demokrasi) ([pemilu](https://id.wikipedia.org/wiki/Pemilu)), pengumpulan [pajak](https://id.wikipedia.org/wiki/Pajak), juga digunakan dalam ilmu [ekonomi](https://id.wikipedia.org/wiki/Ekonomi). Salah satu contohnya adalah sensus yang dilakukan oleh Biro Sensus Amerika (*United State Census Bureau*) tentang populasi penduduk di Amerika. Sensus ini bertujuan untuk mendapatkan informasi terkait demografis penduduk dan juga terkait perekonomian di Amerika.

Amerika Serikat merupakan salah satu negara maju di dunia. Salah satu syarat negara maju adalah pendapatan perkapita harus mencapai nilai tertentu. Salah satu cara untuk mempertahankan nilai pendapatan per kapita negara Amerika adalah pemerintah harus dapat mengklasifikasikan pendapatan rakyat sesuai dengan kondisi-kondisi tertentu. Hal ini dilakukan agar pemerintah dengan mudah menentukan kondisi masyarakat seperti apa yang menjadi perhatian utama dalam peningkatan sumber daya manusia sehingga pendapatan per kapita negara ini stabil.

## **Tujuan**

Tujuan yang ingin kami capai dari penelitian tugas akhir ini adalah mengklasifikasi model dari dataset Biro Sensus Amerika (*United State Census Bureau*) metode pohon keputusan dengan algoritma C4.5 dan pengetahuan mengenai tingkat pendapatan Amerika pada tahun 1994 dan 1995.

## **Manfaat**

Manfaat dari penelitian ini adalah membentuk model yang menjadi tolak ukur pemerintah US dalam meningkatkan pendapatan per-kapita negaranya. Selain itu, model ini juga dapat digunakan sebagai tolak ukur masyarakat US dalam menentukan pendidikan, pekerjaan, instansi, dan sebagainya untuk memaksimalkan pendapatan per tahunnya.

## **Ruang Lingkup**

Data ini hanya untuk data penduduk yang tinggal di amerika yang sudah sah menjadi warga amerika saja. Ruang lingkupnya juga hanya untuk penduduk yang memiliki umur lebih dari 14 tahun sesuai referensi yang kami peroleh sesuai dari umur minimum boleh bekerja oleh *U.S Department of Labor.* Lalu kami juga menghilangkan data penduduk yang lahir di luar negeri dan bukan penduduk Amerika untuk memastikan bahwa kami benar menghitung penduduk Amerika dan bukan pendatang.

Teknik data mining yang dilakukan adalah Klasifikasi dengan metode Pohon Keputusan dengan Algoritma C4.5. Peneliti menggunakan R Studio sebagai bahasa pemrograman dan sistem perangkat lunak yang dirancang khusus untuk mengerjakan segala hal terkait komputasi statistik.

## **TINJAUAN PUSTAKA**

**Biro Sensus Amerika**

Biro Sensus Amerika adalah biro di bawah [Departemen Perdagangan Amerika Serikat](https://id.wikipedia.org/wiki/Departemen_Perdagangan_Amerika_Serikat) yang bertugas melakukan [sensus di Amerika Serikat](https://id.wikipedia.org/w/index.php?title=Sensus_di_Amerika_Serikat&action=edit&redlink=1) setiap sepuluh tahun sekali. Data yang dikumpulkan terdiri dari populasi penduduk, perumahan, dan berbagai jenis data perdagangan. Hukum federal menjamin kerahasiaan data yang dikumpulkan dari perorangan, rumah tangga, atau badan, dan data tersebut hanya digunakan untuk keperluan statistik. Biro Sensus Amerika Serikat menjadi badan resmi pelaksana sensus di AS berdasarkan undang-undang Kongres AS tertanggal 6 Maret 1902 . Biro Sensus AS dipimpin seorang direktur yang dibantu seorang wakil direktur, dan staf eksekutif yang terdiri dari beberapa asisten direktur.

Kantor pusat Biro Sensus berada di 4600 Silver Hill Road, [Suitland, Maryland](https://id.wikipedia.org/w/index.php?title=Suitland,_Maryland&action=edit&redlink=1). Kantor perwakilan berlokasi di 12 kota: [Boston](https://id.wikipedia.org/wiki/Boston), [New York City](https://id.wikipedia.org/wiki/New_York_City), [Philadelphia](https://id.wikipedia.org/wiki/Philadelphia), [Detroit](https://id.wikipedia.org/wiki/Detroit), [Chicago](https://id.wikipedia.org/wiki/Chicago), [Kansas City](https://id.wikipedia.org/w/index.php?title=Daerah_Metropolitan_Kansas_City&action=edit&redlink=1), [Seattle](https://id.wikipedia.org/wiki/Seattle), [Charlotte](https://id.wikipedia.org/w/index.php?title=Charlotte&action=edit&redlink=1), [Atlanta](https://id.wikipedia.org/wiki/Atlanta), [Dallas](https://id.wikipedia.org/wiki/Dallas), [Denver](https://id.wikipedia.org/wiki/Denver), dan [Los Angeles](https://id.wikipedia.org/wiki/Los_Angeles). Pusat Pengolahan Data Nasional berada di [Jeffersonville, Indiana](https://id.wikipedia.org/w/index.php?title=Jeffersonville,_Indiana&action=edit&redlink=1), juga terdapat tambahan pusat pengolahan data yang bersifat sementara setiap kali sensus dilaksanakan.

**Data Mining**

*Data mining* merupakan bidang ilmu yang sangat luas penerapanya, *data mining* merupakan cabang ilmu yang termasuk masih baru tetapi telah menghasilkan keuntungan yang cukup besar saat ini. *Data mining* sering dimanfaatkan untuk mendeteksi kejadian-kejadian yang ganjil seperti penyakit tertentu , transaksi yang mencurigakan, hingga mendeteksi telepon orang yang ingin menipu seperti penyalahgunaan kartu kredit.

*Data mining* menurut David Hand, Heika Manila dan Padraic Smyth dari MIT adalah analisis terhadap data (biasanya data yang cukup besar) untuk menemukan hubungan yang jelas serta menyimpulkan yang belum diketahui sebelumnya dan data dapat dipahami dan berguna bagi pemiliknya (Larose, 2005).

*Data mining* mempunyai teori-teori yang banyak dan data mining teorinya pun sudah ada sejak lama seperti *Naive-Bayes, Nearest Neighbor,* Pohon keputusan, Aturan Assoisasi ,*K-Means Cluster dan Text Mining* (Bramer, 2007).

Kemajuan luar biasa yang terus berlanjut dalam bidang *data mining* didorong oleh beberapa faktor, antara lain (Larose, 2005):

* Pertumbuhan yang cepat dalam kumpulan data.
* Penyimpanan data dalam data warehouse, sehingga seluruh perusahaan
* Memiliki akses ke dalam database yang andal.
* Adanya peningkatan akses data melalui navigasi web dan intranet.
* Tekanan kompetisi bisnis untuk meningkatkan penguasaan pasar dalam globalisasi ekonomi.
* Perkembangan teknologi perangkat lunak untuk data mining (ketersediaan teknologi).
* Perkembangan yang hebat dalam kemampuan komputasi dan pengembangan.
* Kapasitas pengembangan media penyimpanan.

**Proses *Data Mining***

Secara skematis, (Gorunescu, 2011) membagi langkah-langkah proses pelaksanan data mining dalam tiga aktivitas yaitu

1. Eksplorasi data, terdiri dari aktivitas pembersihan data, transformasi data, pengurangan dimensi, pemilihan ciri, dan lain-lain.
2. Membuat model dan pengujian validitas model, merupakan pemilihan terhadap model-model yang sudah di kembangkan yang cocok dengan kasus yang dihadapi, dengan kata lain, dilakukan pemilihan model secara kompetitif.
3. Penerapan model dengan data baru untuk menghasilkan perkiraan data kasus ada.

Tahap ini merupakan tahap yang menentukan apakah model yang telah dibangun dapat menjawab permasalahan yang dihadapi.

Untuk menentukan model baik atau buruk, kita memerlukan elemen-elemen kunci antara lain :

1. Akurasi prediksi, untuk menentukan seberapa akurat suatu model dalam memprediksi keluaran.
2. Kecepatan, yang menunjukan seberapa cepat suatu model dalam memproses data masukan.
3. *Robustness*, menggambarkan kemampuan suatu model melakukan prediksi yang akurat, walau dalam kondisi extreme dan banyak gangguan.
4. Skalabilitas, kemampuan suatu proses data ukuran yang besar ataupun kecil.
5. Kesederhanaan, merupakan sifat yang cenderung dipilih untuk menyelesaikan suatu permasalahan.

**Pemodelan Data Mining**

Pemodelan adalah penggunaan prinsip-prinsip atau teknik-teknik tertentu dalam suatu rancangan sistem. Kian kompleksnya masalah-masalah yang dijumpai saat ini membuat proses pemodelan menjadi semakin kompleks.

Sebagai bahan pertimbangan, (Gounescu, 2011) menyarankan tahapan-tahapan sebagai berikut :

1. Identifikasi
2. Estimasi dan pencocokan
3. Pengujian
4. Penerapan praktis
5. Iterasi

**Pohon Keputusan**

Pohon keputusan adalah salah satu metode klasifikasi yang paling populer, karena mudah untuk diinterpretasi oleh manusia. Pohon keputusan adalah model prediksi menggunakan struktur pohon atau struktur berhirarki.

Konsep dari pohon keputusan adalah mengubah data menjadi pohon keputusan dan aturan-aturan keputusan. Manfaat utama dari penggunaan pohon keputusan adalah kemampuannya untuk mem-break down proses pengambilan keputusan yang kompleks menjadi lebih simple, sehingga pengambil keputusan akan lebih menginterpretasikan solusi dari permasalahan.

Nama lain dari pohon keputusan adalah CART (*Classification and Regression Tree*). Dimana metode ini merupakan gabungan dari dua jenis pohon, yaitu classification tree dan juga regression tree. Pohon keputusan juga berguna untuk mengeksplorasi data, menemukan hubungan tersembunyi antara sejumlah calon variabel input dengan sebuah variabel target. Pohon keputusan memadukan antara eksplorasi data dan pemodelan, sehingga sangat bagus sebagai langkah awal dalam proses pemodelan bahkan ketika dijadikan sebagai model akhir dari beberapa teknik lain.

Dalam beberapa aplikasi, akurasi dari sebuah klasifikasi atau prediksi adalah satu-satunya hal yang ditonjolkan dalam metode ini, misalnya sebuah perusahaan direct mail membuat sebuah model yang akurat untuk memprediksi anggota mana yang berpotensi untuk merespon permintaan, tanpa memperhatikan bagaimana atau mengapa model tersebut bekerja.

Kelebihan lain dari metode ini adalah mampu mengeliminasi perhitungan atau data-data yang kiranya tidak diperlukan. Sebab, sampel yang ada biasanya hanya diuji berdasarkan kriteria atau kelas tertentu saja.

Meski memiliki banyak kelebihan, namun bukan berarti metode ini tidak memiliki kekurangan. Pohon keputusan ini bisa terjadi overlap, terutama ketika kelas dan kriteria yang digunakan sangat banyak tentu saja dapat meningkatkan waktu pengambilan keputusan sesuai dengan jumlah memori yang dibutuhkan.

**Algoritma C4.5**

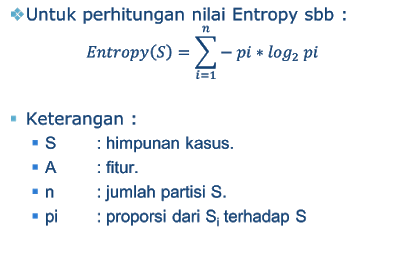
Algoritma C4.5 merupakan algoritma yang digunakan untuk membentuk pohon keputusan (*Decision Tree*). Pohon keputusan merupakan metode klasifikasi dan prediksi yang terkenal. Pohon keputusan berguna untuk mengeksplorasi data, menemukan hubungan tersembunyi antara sejumlah calon variabel input dengan sebuah variabel target.Banyak algoritma yang dapat dipakai dalam pembentukan pohon keputusan, antara lain ID3, CART, dan C4.5. Algoritma C4.5 merupakan pengembangan dari algoritma ID3, Proses pada pohon keputusan adalah mengubah bentuk data (tabel) menjadi model pohon, mengubah model pohon menjadi rule, dan menyederhanakan rule.

Secara umum algoritma C4.5 untuk membangun pohon keputusan adalah sebagai berikut :

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

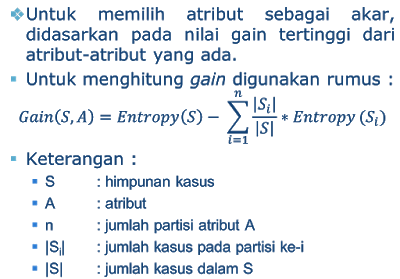
**Konsep Entropy**

* Entropy (S) merupakan jumlah bit yang diperkirakan dibutuhkan untuk dapat mengekstrak suatu kelas (+ atau -) dari sejumlah data acak pada ruang sampel S.
* Entropy dapat dikatakan sebagai kebutuhan bit untuk menyatakan suatu kelas.
* Entropy digunakan untuk mengukur ketidakaslian S.



**Konsep Gain**

* Gain (S,A) merupakan perolehan informasi dari atribut A relative terhadap output data S.
* Perolehan informasi didapat dari output data atau variable dependent S yang dikelompokkan berdasarkan atribut A, dinotasikan dengan gain (S,A).



**METODE**

## **Tentang Dataset**

Data yang dipakai pada proyek kali ini merupakan data census yang diekstraksi dari Survei Populasi pada Tahun 1994 dan 1995 yang dilakukan oleh Biro Sensus Amerika. Karakteristik data merupakan multivariate , data ini terdiri dari 40 atribut. Penjelasan lebih lanjut dapat dilihat di lampiran.

|  |  |  |
| --- | --- | --- |
| Nama Kolom | Deskripsi | Tipe |
| age | Umur dari Penduduk | Numerik |
| class\_worker | Kelas Pekerja | Kategorik |
| det\_ind\_code | Kode Industri | Numerik |
| det\_occ\_code | Kode Pekerjaan | Numerik |
| education | Jenjang Pendidikan | Kategorik |
| wage\_per\_hour | Pendapatan per jam | Numerik |
| hs\_college | Status pendidikan sekarang | Kategorik |
| marital\_stat | Status Pernikahan | Kategorik |
| major\_ind\_code | Major Kode Industri | Kategorik |
| major\_occ\_code | Major Kode Pekerjaan | Kategorik |
| race | Suku | Kategorik |
| hisp\_origin | Berasal dari Spanyol/ Keturunan Spanyol | Kategorik |
| sex | Gender | Kategorik |
| union\_member | Anggota Serikat pekerja | Kategorik |
| unemp\_reason | Alasan tidak bekerja | Kategorik |
| full\_or\_part\_emp | Status pekerjaan tetap/tidak | Kategorik |
| capital\_gains | Capital gains: Keuntungan yang didapat dari Investasi | Numerik |
| capital\_losses | Capital losses: Kerugian yang didapat dari Investasi | Numerik |
| stock\_dividends | Pembagian Laba | Numerik |
| tax\_filer\_stat | Status pengarsipan pajak | Kategorik |
| region\_prev\_res | Regional dari tempat tinggal sebelumnya | Kategorik |
| state\_prev\_res | Negara Bagian dari tempat tinggal sebelumnya | Kategorik |
| det\_hh\_fam\_stat | Detail info status keluarga | Kategorik |
| det\_hh\_summ | Detailed ringkasan keluarga | Kategorik |
| instance\_weight | Menunjukkan jumlah orang dalam populasi yang mewakili setiap catatan karena stratified sampling | Numerik |
| mig\_chg\_msa | Kode Migrasi - berganti di MSA | Kategorik |
| mig\_chg\_reg | Kode Migrasi - Berganti di Regional | Kategorik |
| mig\_move\_reg | Kode Migrasi - Perpindahan Regional | Kategorik |
| mig\_same | Menetap di rumah sekarang lebih dari 1 tahun | Kategorik |
| mig\_prev\_sunbelt | Daerah sebelumnya berada di area sunbelt | Kategorik |
| num\_emp | Orang yang bekerja di perusahaan miliknya | Numerik |
| fam\_under\_18 | Keluarga yang berumur kurang dari 18 tahun | Kategorik |
| country\_father | Negara asal kelahiran ayah | Kategorik |
| country\_mother | Negara asal kelahiran ibu | Kategorik |
| country\_self | Negara asal kelahiran diri sendiri | Kategorik |
| citizenship | Kewarganegaraan | Kategorik |
| own\_or\_self | Keterangan memiliki usaha atau tidak | Numerik |
| vet\_question | Mendapatkan pertanyaan tentang veteran | Kategorik |
| vet\_benefits | Keuntungan veteran yang diterima | Numerik |
| weeks\_worked | Minggu bekerja dalam setahun | Numerik |
| year | Tahun mendapat survey | Numerik |
| class\_income | Pendapatan kurang atau lebih dari $50,000 | Kategorik   * - 50000. * 50000+. |

Data ini dapat diolah menggunakan cara klasifikasi dengan metode pohon keputusan. Data U.S Census ini juga terdapat missing value di dalam yang dituliskan dengan lambang tanda tanya.

**Tahapan Kegiathan**

**Analisis**

Tahapan ini bertujuan untuk memberikan pemahaman mengenai permasalahan yang akan diselesaikan penelitian, untuk mengetahui kesesuaian solusi dengan permasalahan, untuk mengetahui kesesuaian pemilihan konsep atau teori yang digunakan dalam penelitian dan untuk mengetahui kebutuhan data yang akan digunakan dalam penelitian.

**Knowledge Discovery from Data (KDD)**

Berikut merupakan rangkaian proses penemuan pengetahuan yang dilakukan pada penelitian tugas akhir:

1. *Data Praprocessing*

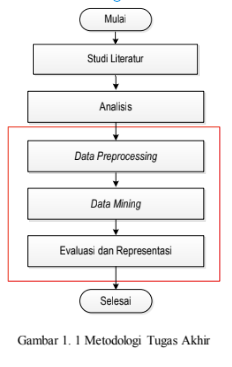
*Data praprocessing* dilakukan dengan tujuan untuk membentuk dataset penelitian. Data preprocessing terdiri dari tahapan pembersihan data, integrasi atau penggabungan data, pemilihan data dan transformasi data. Untuk keperluan validasi, data dibagi menjadi data latih dan data uji dengan perbandingan 20:80.

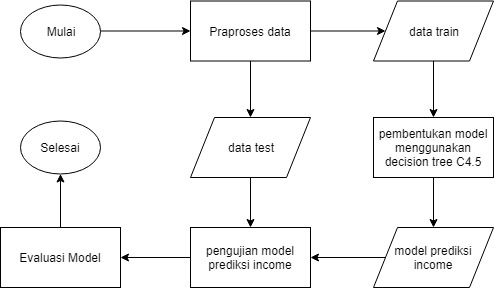
1. *Data Mining*

Tahapan ini dilakukan dengan menerapkan algoritma C4.5 untuk melakukan klasifikasi terhadap data latih dengan bantuan R Studio. Hasil dari proses *data mining* adalah berupa model pohon keputusan.

**Evaluasi dan Representasi**

Tahapan evaluasi yaitu melakukan pengukuran kinerja model pohon keputusan dengan menggunakan *Confusion Matrix* untuk mengetahui seberapa baik model klasifikasi dapat memprediksi data uji.





Gambar 1.2 Penjelasan KDD

## **Lingkungan Pengembangan**

Perangkat Keras :

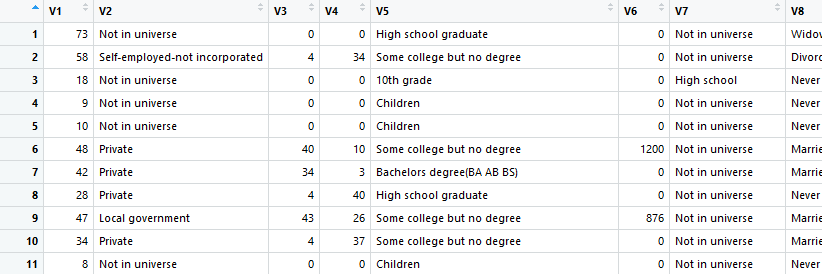
* *Processor* AMD Ryzen 5 2500U
* Memori 8 GB
* *Hard disk* 1000 GB
* Layar 15.6-inch
* *Mouse* dan *Keyboard*

Perangkat Lunak :

* Sistem operasi Windows 10
* RStudio 1.1.463 dan R 3.5.2

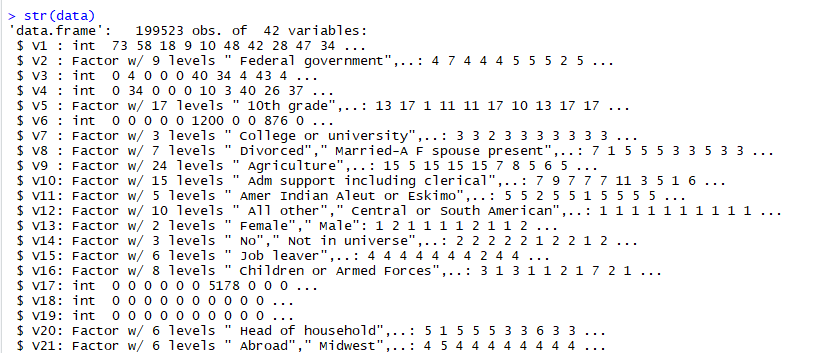
# **HASIL DAN PEMBAHASAN**

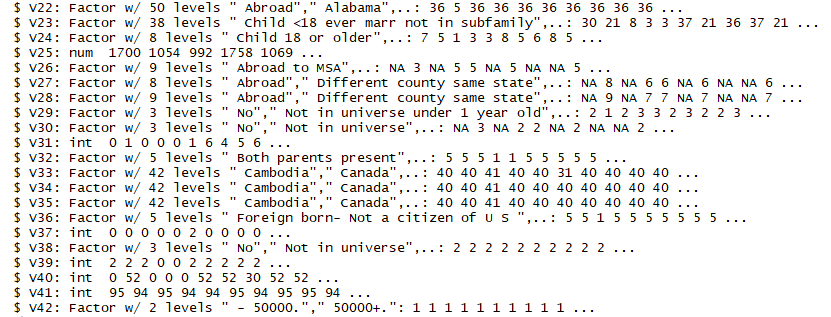
Tahapan awal yang kami lakukan adalah inisiasi *data frame* yang diinput ke dalam variabel “data”. Pada tahap ini kami melakukan pengubahan data dari nilai “?” menjadi *data missing value*



Gambar 2.1 Inisiasi Data Frame

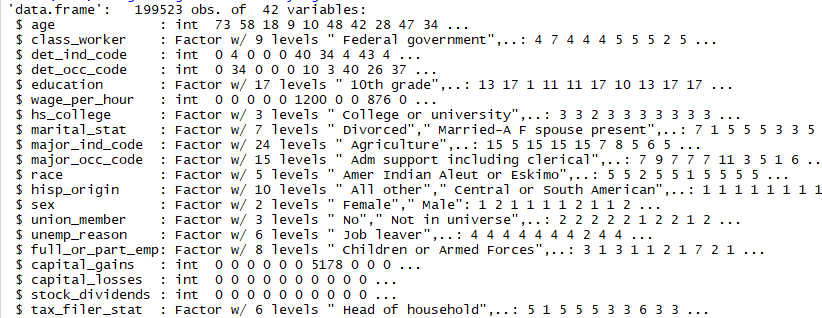
Pada tahapan praproses kami melakukan diskretisasi terhadap atribut​. Hal ini dilakukan agar data yang seharusnya bertipe nominal dan nilainya terbatas tidak mendapatkan perlakuan seperti pada data numerik di dalam perhitungan algoritma ataupun sebaliknya. Berikut adalah tampilan data yang sudah di diskretisasi.

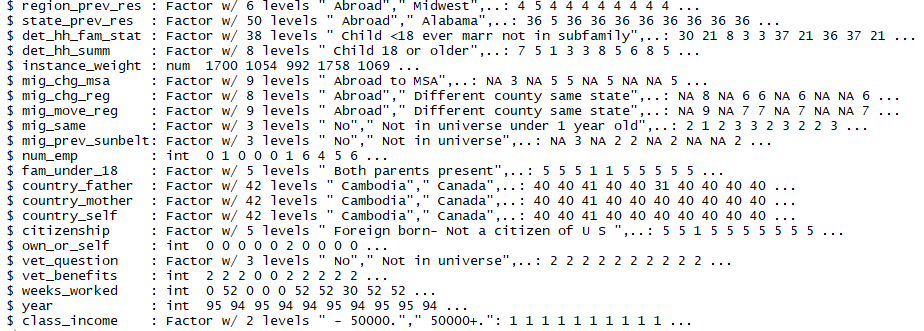




Gambar 2.2 Data Setelah Diskretisasi

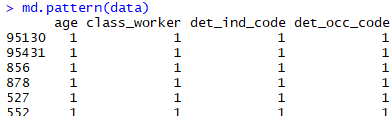
Karena data ini belum memiliki nama atribut di setiap kolomnya, maka tahap selanjutnya adalah penamaan atributnya. Berikut adalah data yang sudah memiliki nama atribut.

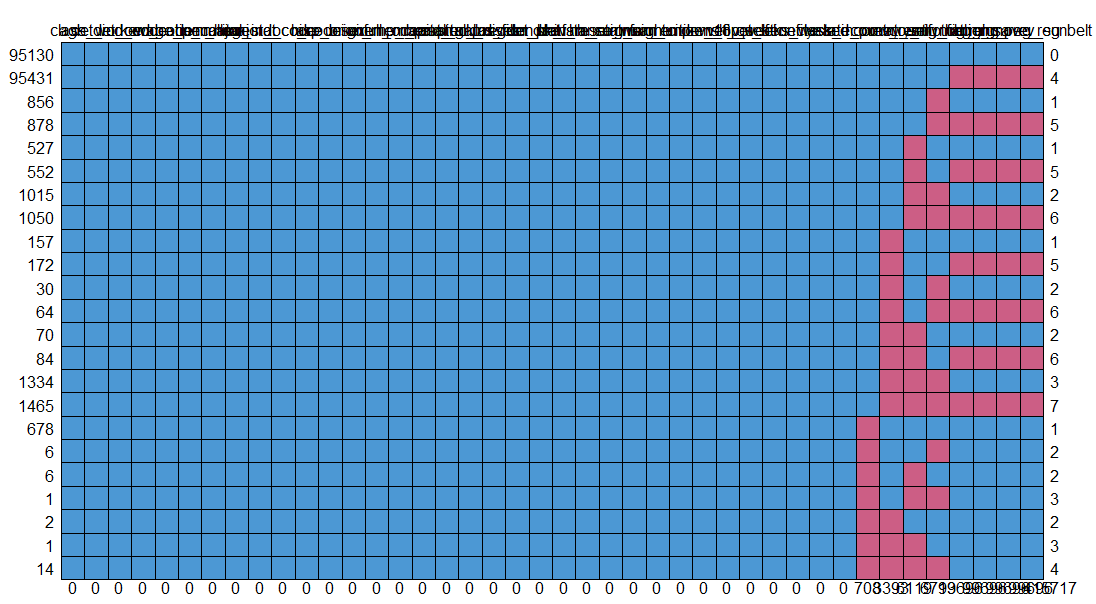




2.3 Data Setelah Ditambahkan Nama Kolom Pada Atribut

Setelah setiap atribut sudah dinamai, tahap selanjutnya pengecekan *missing value.*

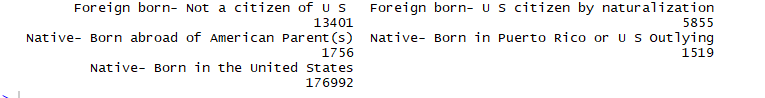




2.4 Pengecekan *Missing Value*

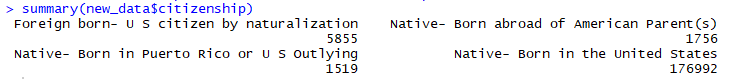
Ternyata terdapat 95130 dari 199523 instance yang tidak terdapat missing value. Itu berarti terdapat 104393 instance yang terdapat setidaknya 1 missing value pada variabelnya.

Tahap pra proses selanjutnya kami ingin menghilangkan data penduduk yang lahir di luar negeri dan bukan penduduk Amerika. Hal ini bisa juga disebut sebagai pengunjung saja, sehingga kami memutuskan untuk menghilangkan *instance* tersebut.



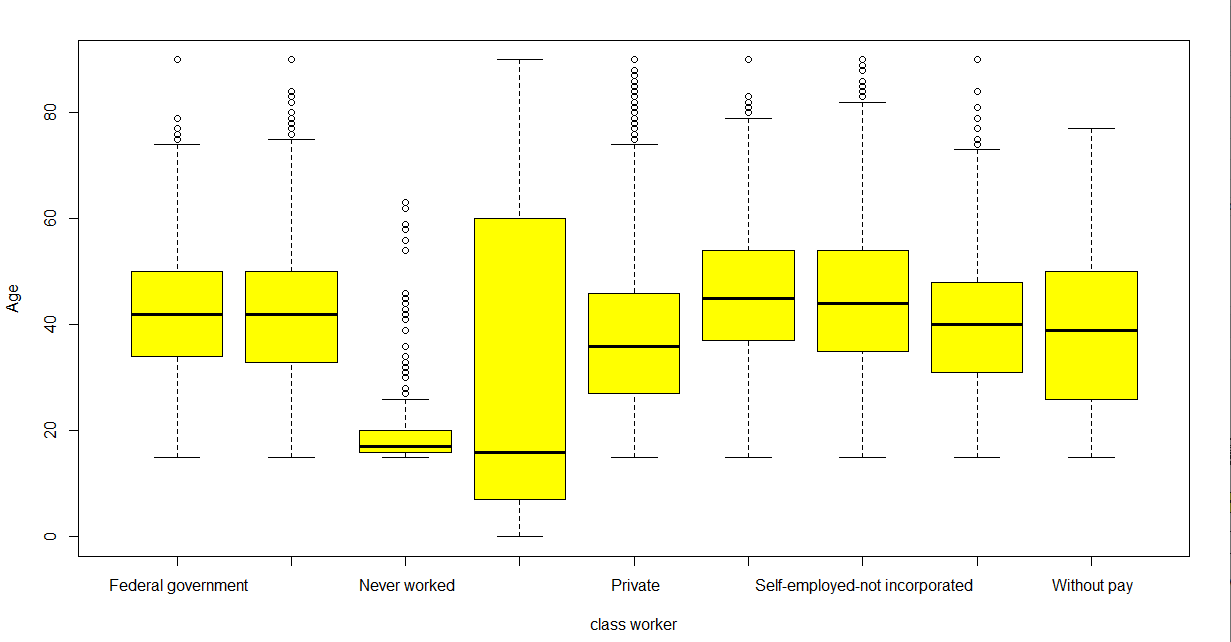
2.5 Data Penduduk Non Amerika

Hasilnya, data yang akan diolah adalah penduduk Amerika yang sekarang berkependudukan Amerika dengan membuang data orang yang bukan penduduk Amerika.

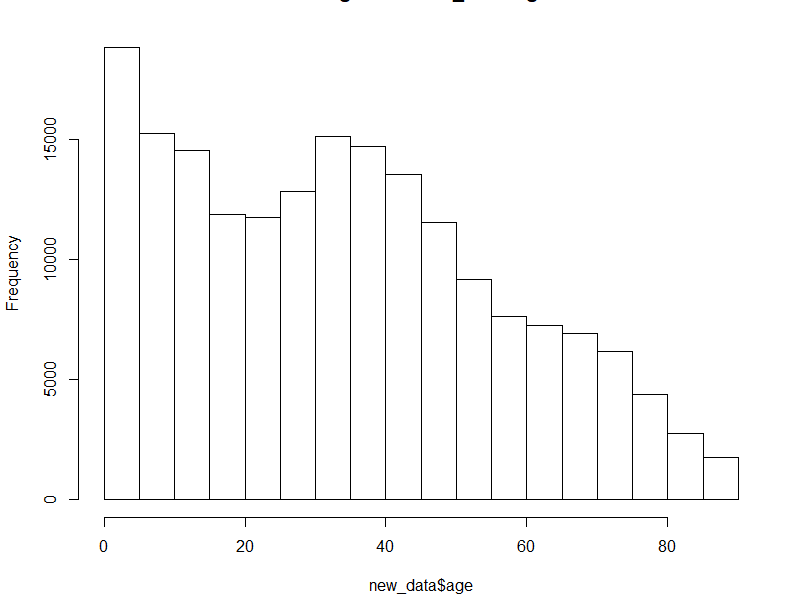


2.6 Data Setelah di Hilangkan Data Penduduk Non Amerika

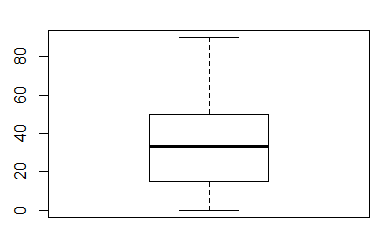
Dikarenakan tujuan akhir kami yaitu mengetahui pemasukan dari penduduk Amerika, maka data penduduk dengan umur 0 - 14 tahun tidak diperlukan. Hal ini disebabkan dari ruang lingkup penelitian kami dari *U.S Department of Labor* bahwa minimum umur pekerja adalah 14 tahun. Selain itu kami juga telah melakukan *plotting* hubungan antara umur dan kelas pekerja. Dimana untuk umur dibawah 20 tahun cenderung memiliki kelas pekerja yaitu “*never worked*” dan “*not in universe*”



2.7 Dendogram Rangkuman Pemasukan Penduduk Amerika

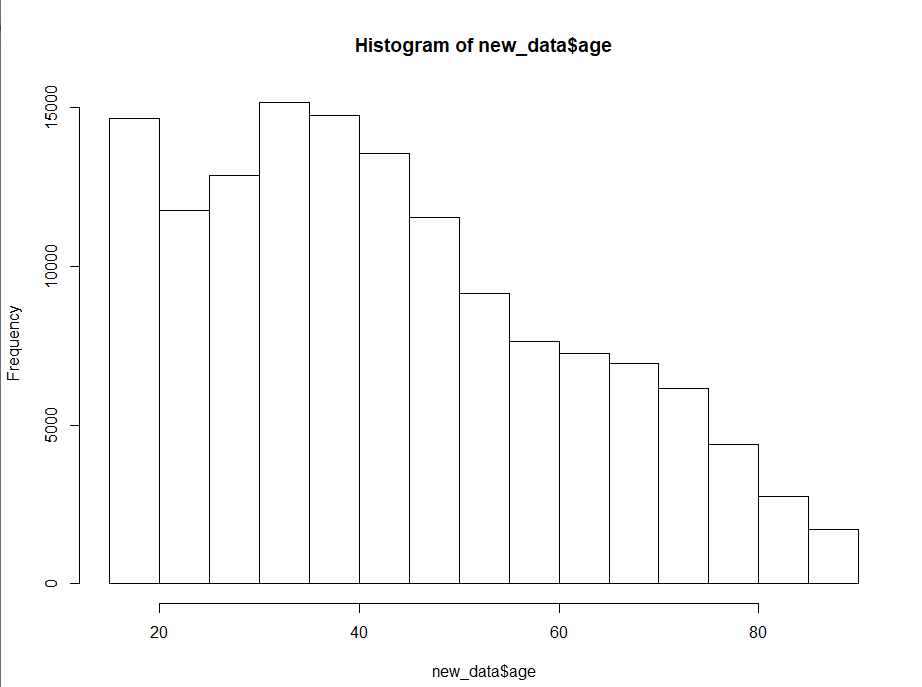


2.8 Frekuensi Persebaran Data



2.9

Setelah dilakukan penghapusan maka histogram dari umur sebagai berikut

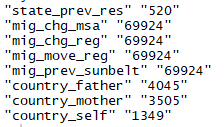


2.10 Histogram Setelah Penghapusan Data



2.11 Ringkasan Statistika Sederhana Setelah Penghapusan Data

Setelah seluruh proses penghapusan *instance* yang tidak diperlukan sudah selesai, tahap selanjutnya yaitu pengisian *missing value*. Karena jumlah *missing value* yang sangat banyak maka *missing value* kami isi dengan level baru yaitu “*unknown*”. Adapun atribut-atribut yang memiliki *missing values* adalah:



2.12 Atribut yang memiliki *missing value*

Karena seluruh atribut yang memiliki *missing value* merupakan atribut kategorikal dan jumlahnya sangat banyak, jadi kami menambahkan nilai baru yaitu “*unknown*”

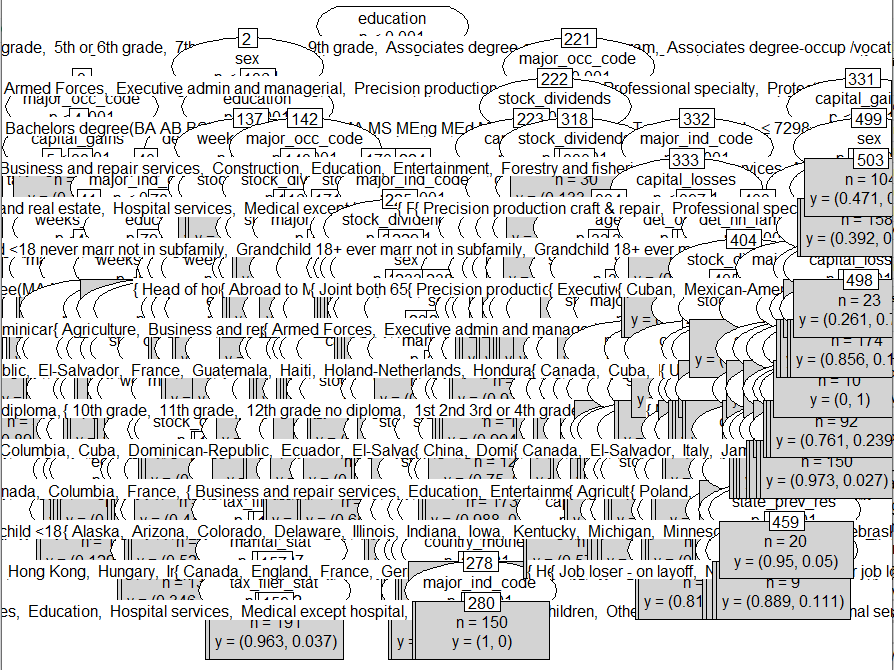
Tahap selanjutnya adalah membagi data ke dalam data training dan data testing. Kami membagi datanya 80% masuk ke data training, dan 20% masuk ke dalam data testing. Berdasarkan prinsip Pareto, pembagian dataset menjadi 80:20 dimaksudkan karena dataset ini merupakan dataset sedang atau tidak terlalu banyak/sedikit, sehingga pembagian sebesar 80:20 menurut kami porsi yang pas. Melalui tahap pembagian dataset ini, data train dan data test yang berhasil terbentuk sebesar 112262 data train dan 27923 data test.



2.13 Banyak Data Train dan Test

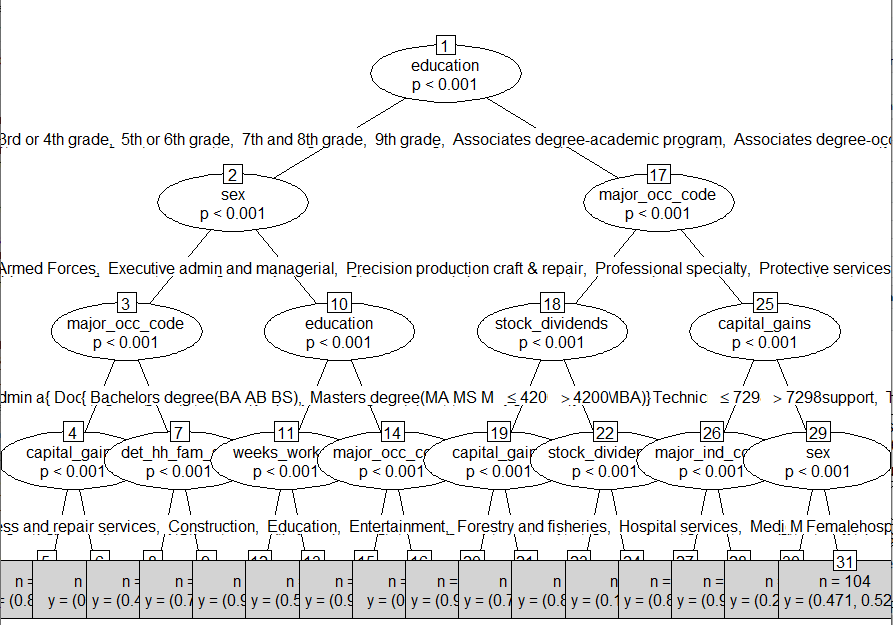
Tahap selanjutnya adalah proses modeling. Kami menggunakan dua package dalam perlakuan dataset kami untuk membentuk pohon keputusan yaitu “rpart” dan “party”. Tujuannya untuk menggunakan kelebihan dari yang dimiliki satu package dan tidak dimiliki oleh package lainnya.

Tahap modeling pertama menggunakan package “party”. Tree yang terbentuk ketika menggunakan package party sangatlah kompleks seperti gambar berikut



2.14 Tree yang Terbentuk

Kami mencoba untuk menyederhanakan tree nya dengan maksimal kedalaman = 4



2.15 Tree yang Terbentuk Dengan Maksimal Kedalaman 4

Setelah mendapatkan model pohon keputusan, dapat dilihat dari tree tersebut atribut tingkat pendidikan dari seseorang di Amerika merupakan atribut yang sangat berpengaruh terhadap tingkat pemasukan dari penduduk Amerika. Setelah itu kami melakukan memprediksi dari data test ke dalam model. Didapatkan akurasi prediksi data sebesar 93.43% dan hasil confusion matriksnya sebagai berikut

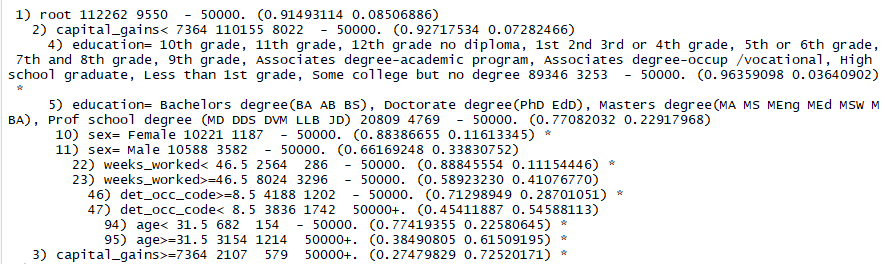


2.16 Confussion Matrix yang Dihasilkan

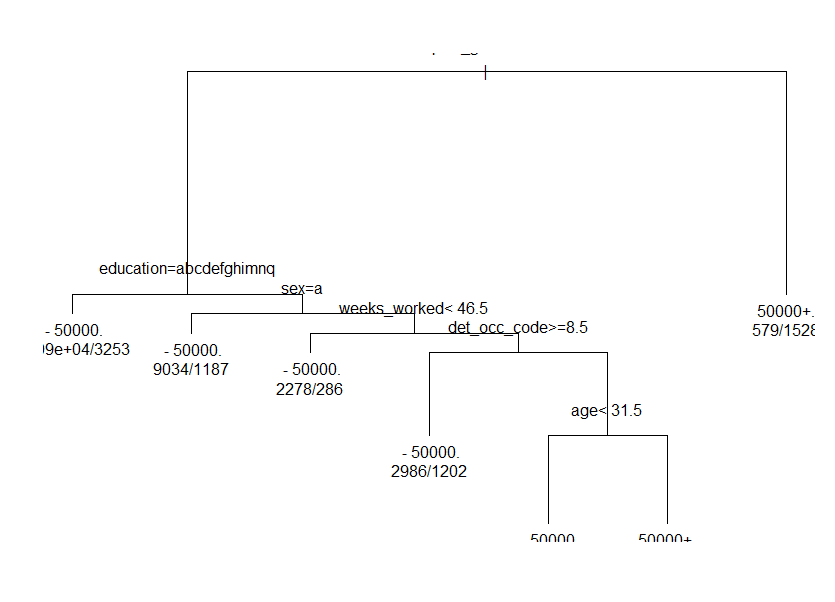


2.17 Akurasi Penelitian

Selanjutnya, kami membuat model dari package yang berbeda, yaitu package “rpart”. Tree yang dihasilkan sebagai berikut

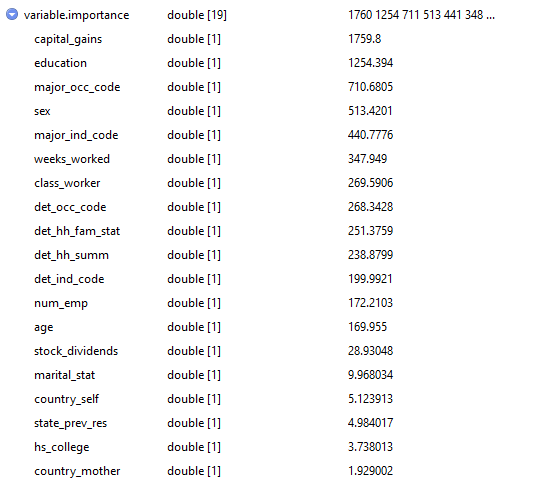


2.18 Model Package rpart



2.19 Decision Tree dari Model Package rpart

Hasil tree yang dihasilkan oleh fungsi rpart berbeda dengan tree yang dihasilkan oleh fungsi party, baik dari kompleksitas dari tree tersebut dan atribut yang paling berpengaruh terhadap tingkat pemasukan penduduk Amerika. Atribut yang paling berpengaruh adalah keuntungan dalam bentuk uang seseorang lalu diikuti dengan atribut tingkat pendidikan dari seseorang tersebut.



2.20 Atribut yang Paling Berpengaruh Kedalam

Tingkat Keuangan Seseorang

Setelah mendapatkan modelnya, kami melakukan pengujian terhadap data test. Didapatkan *confusion matrix* dan akurasi yang dihasilkan sebagai berikut dan dari model tersebut menghasilkan akurasi 92.88%..



2.20 Confussion Matrix Dari Model Package rpart



2.21 Akurasi dari Model Package rpart

**KESIMPULAN DAN SARAN**

Berdasarkan hasil penelitian kami, dapat disimpulkan bahwa algoritma pohon keputusan C4.5 dengan menggunakan fungsi rpart lebih baik jika kita bandingkan dengan menggunakan fungsi party mampu untuk memodelkan tree yang dapat memprediksi income setahun seseorang yang berada di negara Amerika Serikat. Hal ini terbukti algoritma tersebut berhasil memprediksi data dengan akurasi sekitar 92% sedangkan jika menggunakan party memiliki akurasi sebesar 93%, tapi tree yang dihasilkan oleh *library* rpart lebih ringkas dibandingkan dengan menggunakan *library* party dan hanya berbeda 1% saja untuk akurasinya. Menggunakan fungsi rpart dan party juga menghasilkan atribut penting yang berbeda. Hasil prediksi ini dapat dijadikan acuan pemerintah untuk fokus terhadap peningkatan mutu masyarakat Amerika dalam meningkatkan income sehingga GDP dari negara tersebut dapat meningkat. Selain itu, hasil prediksi ini juga dapat menjadi pelajaran penting untuk masyarakat Amerika jika ingin meningkatkan income dalam setahun.

Saran dalam penelitian ini adalah perlu adanya evaluasi kembali terhadap atribut-atribut yang mengandung missing value, karena dalam penelitian ini masih mengisi missing value tersebut dengan nilai “unknown”. Selain itu peneliti juga menyarankan untuk melakukan evaluasi kembali terhadap seluruh atribut yang ada, apakah berkorelasi tinggi terhadap atribut lain sehingga beberapa atribut bisa digabung dan ini akan mempercepat proses pembuatan modelnya.

**DAFTAR PUSTAKA**

“Biro Sensus Amerika Serikat.” *Wikipedia*, Wikimedia Foundation, 11 Nov. 2018, id.wikipedia.org/wiki/Biro\_Sensus\_Amerika\_Serikat.

“Blog.” *GSB*, gsbipb.com/?p=821.

Bramer, M. 2007. Principles of Data Mining: Undergraduate Topics in Computer Science. Springer-Verlag: London.

Gorunescu, Florin. 2011. Data Mining: Concepts, Models, and Techniques. Verlag

Berlin Heidelberg : Springer

Iykra. “Mengenal Decision Tree Dan Manfaatnya.” *Medium*, Iykra, 23 July 2018, medium.com/iykra/mengenal-decision-tree-dan-manfaatnya-b98cf3cf6a8d.

Larose D, T., 2005, Discovering knowledge in data : an introduction to data mining, Jhon Wiley & Sons Inc.

“Pengertian, Fungsi, Proses Dan Tahapan Data Mining.” *Beyonder.asia*, 28 Mar. 2018, beyonder.asia/pengertian-fungsi-proses-dan-tahapan-data-mining/.

US Census Bureau. “Population.” *Census.gov*, 21 Aug. 2018, www.census.gov/topics/population.html.

**Lampiran (Keterangan data Census Income Bureau)**

|  |  |  |
| --- | --- | --- |
| Nama Kolom | Deskripsi | Tipe |
| age | Umur dari Penduduk  Nilai Kontinyu dari 0-90 | Numerik |
| class\_worker | Kelas Pekerja   * Not in universe * Private * Self-employed-not incorporated * Local government * State government * Self-employed-incorporated * Federal government * Never worked * Without pay | Kategorik |
| det\_ind\_code | Kode Industri | Numerik |
| det\_occ\_code | Kode Pekerjaan | Numerik |
| education | Jenjang Pendidikan   * Children * Less than 1st grade * 1st 2nd 3rd or 4th grade * 5th or 6th grade * 7th and 8th grade * 9th grade * 10th grade * 11th grade * 12th grade no diploma * High school graduate * Some college but no degree * Associates degree-academic program * Associates degree-occup /vocational * Bachelors degree(BA AB BS) * Masters degree(MA MS MEng MEd MSW MBA) * Prof school degree (MD DDS DVM LLB JD) * Doctorate degree(PhD EdD) | Kategorik |
| wage\_per\_hour | Pendapatan per jam | Numerik |
| hs\_college | Status pendidikan sekarang   * Not in universe * High school * College or university | Kategorik |
| marital\_stat | Status Pernikahan   * Never married * Married-civilian spouse present * Divorced * Widowed * Separated * Married-spouse absent * Married-A F spouse present | Kategorik |
| major\_ind\_code | Major Kode Industri   * Not in universe or children * Retail trade * Manufacturing-durable goods * Education * Manufacturing-nondurable goods * Finance insurance and real estate * Construction * Business and repair services * Medical except hospital * Public administration * Other professional services * Transportation * Hospital services * Wholesale trade * Agriculture * Personal services except private HH * Social services * Entertainment * Communications * Utilities and sanitary services * Private household services * Mining * Forestry and fisheries * Armed Forces | Kategorik |
| major\_occ\_code | Major Kode Pekerjaan   * Not in universe * Adm support including clerical * Professional specialty * Executive admin and managerial * Other service * Sales * Precision production craft & repair * Machine operators assmblrs & inspctrs * Handlers equip cleaners etc * Transportation and material moving * Farming forestry and fishing * Technicians and related support * Protective services * Private household services * Armed Forces | Kategorik |
| race | Suku   * White * Black * Asian or Pacific Islander * Other * Amer Indian Aleut or Eskimo | Kategorik |
| hisp\_origin | Berasal dari Spanyol/ Keturunan Spanyol   * All other * Mexican-American * Mexican (Mexicano) * Central or South American * Puerto Rican * Other Spanish * Cuban * NA * Do not know * Chicano | Kategorik |
| sex | Gender   * Female * Male | Kategorik |
| union\_member | Anggota Serikat pekerja   * Not in universe * No * Yes | Kategorik |
| unemp\_reason | Alasan tidak bekerja   * Not in universe * Other job loser * Re-entrant * Job loser - on layoff * Job leaver * New entrant | Kategorik |
| full\_or\_part\_emp | Status pekerjaan tetap/tidak   * Children or Armed Forces * Full-time schedules * Not in labor force * PT for non-econ reasons usually FT * Unemployed full-time * PT for econ reasons usually PT * Unemployed part- time * PT for econ reasons usually FT | Kategorik |
| capital\_gains | Capital gains: Keuntungan yang didapat dari Investasi | Numerik |
| capital\_losses | Capital losses: Kerugian yang didapat dari Investasi | Numerik |
| stock\_dividends | Pembagian Laba | Numerik |
| tax\_filer\_stat | Status pengarsipan pajak   * Nonfiler * Joint both under 65 * Single * Joint both 65+ * Head of household * Joint one under 65 & one 65+ | Kategorik |
| region\_prev\_res | Regional dari tempat tinggal sebelumnya   * Not in universe * South * West * Midwest * Northeast * Abroad | Kategorik |
| state\_prev\_res | Negara Bagian dari tempat tinggal sebelumnya   * Not in universe * California * Utah * Florida * North Carolina * ? * Abroad * Oklahoma * Minnesota * Indiana * North Dakota * New Mexico * Michigan * Alaska * Kentucky * Arizona * New Hampshire * Wyoming * Colorado * Oregon * West Virginia * Georgia * Montana * Alabama * Ohio * Texas * Arkansas * Mississippi * Tennessee * Pennsylvania * New York * Louisiana * Vermont * Iowa * Illinois * Nebraska * Missouri * Nevada * Maine * Massachusetts * Kansas * South Dakota * Maryland * Virginia * Connecticut * District of Columbia * Wisconsin * South Carolina * New Jersey * Delaware * Idaho | Kategorik |
| det\_hh\_fam\_stat | Detail info status keluarga   * Householder * Child <18 never marr not in subfamily * Spouse of householder * Nonfamily householder * Child 18+ never marr Not in a subfamily * Secondary individual * Other Rel 18+ ever marr not in subfamily * Grandchild <18 never marr child of subfamily RP * Other Rel 18+ never marr not in subfamily * Grandchild <18 never marr not in subfamily * Child 18+ ever marr Not in a subfamily * Child under 18 of RP of unrel subfamily * RP of unrelated subfamily * Child 18+ ever marr RP of subfamily * Other Rel 18+ ever marr RP of subfamily * Other Rel <18 never marr child of subfamily RP * Other Rel 18+ spouse of subfamily RP * Child 18+ never marr RP of subfamily * Other Rel <18 never marr not in subfamily * Grandchild 18+ never marr not in subfamily * In group quarters * Child 18+ spouse of subfamily RP * Other Rel 18+ never marr RP of subfamily * Child <18 never marr RP of subfamily * Spouse of RP of unrelated subfamily * Child <18 ever marr not in subfamily * Grandchild 18+ ever marr not in subfamily * Grandchild 18+ spouse of subfamily RP * Child <18 ever marr RP of subfamily * Grandchild 18+ ever marr RP of subfamily * Grandchild 18+ never marr RP of subfamily * Other Rel <18 ever marr RP of subfamily * Other Rel <18 never married RP of subfamily * Other Rel <18 spouse of subfamily RP * Child <18 spouse of subfamily RP * Grandchild <18 ever marr not in subfamily * Grandchild <18 never marr RP of subfamily * Other Rel <18 ever marr not in subfamily | Kategorik |
| det\_hh\_summ | Detailed ringkasan keluarga   * Householder * Child under 18 never married * Spouse of householder * Child 18 or older * Other relative of householder * Nonrelative of householder * Group Quarters- Secondary individual * Child under 18 ever married | Kategorik |
| instance\_weight | Menunjukkan jumlah orang dalam populasi yang mewakili setiap catatan karena stratified sampling | Numerik |
| mig\_chg\_msa | Kode Migrasi - berganti di MSA   * Householder * Child under 18 never married * Spouse of householder * Child 18 or older * Other relative of householder * Nonrelative of householder * Group Quarters- Secondary individual * Child under 18 ever married | Kategorik |
| mig\_chg\_reg | Kode Migrasi - Berganti di Regional   * ? * Nonmover * Same county * Different county same state * Not in universe * Different region * Different state same division * Abroad * Different division same region | Kategorik |
| mig\_move\_reg | Kode Migrasi - Perpindahan di Regional   * ? * Nonmover * Same county * Different county same state * Not in universe * Different state in South * Different state in West * Different state in Midwest * Abroad * Different state in Northeast | Kategorik |
| mig\_same | Menetap di rumah sekarang lebih dari 1 tahun   * Not in universe under 1 year old * Yes * No | Kategorik |
| mig\_prev\_sunbelt | Daerah sebelumnya berada di area sunbelt   * ? * Not in universe * No * Yes | Kategorik |
| num\_emp | Orang yang bekerja di perusahan miliknya | Numerik |
| fam\_under\_18 | Keluarga yang berumur kurang dari 18 tahun   * Not in universe * Both parents present * Mother only present * Father only present * Neither parent present | Kategorik |
| country\_father | Negara asal kelahiran ayah   * United-States * Mexico * ? * Puerto-Rico * Italy * Canada * Germany * Dominican-Republic * Poland * Philippines * Cuba * El-Salvador * China * England * Columbia * India * South Korea * Ireland * Jamaica * Vietnam * Guatemala * Japan * Portugal * Ecuador * Haiti * Greece * Peru * Nicaragua * Hungary * Scotland * Iran * Yugoslavia * Taiwan * Cambodia * Honduras * France * Outlying-U S (Guam USVI etc) * Laos * Trinadad&Tobago * Thailand * Hong Kong * Holand-Netherlands * Panama | Kategorik |
| country\_mother | Negara asal kelahiran ibu   * United-States * Mexico * ? * Puerto-Rico * Italy * Canada * Germany * Philippines * Poland * Cuba * El-Salvador * Dominican-Republic * England * China * Columbia * South Korea * Ireland * India * Vietnam * Japan * Jamaica * Guatemala * Ecuador * Peru * Haiti * Portugal * Nicaragua * Hungary * Greece * Scotland * Taiwan * Honduras * France * Iran * Yugoslavia * Cambodia * Outlying-U S (Guam USVI etc) * Laos * Thailand * Hong Kong * Trinadad&Tobago * Holand-Netherlands * Panama | Kategorik |
| country\_self | Kegara asal kelahiran diri sendiri   * United-States * Mexico * ? * Puerto-Rico * Germany * Philippines * Cuba * Canada * Dominican-Republic * El-Salvador * China * South Korea * England * Columbia * Italy * India * Vietnam * Poland * Guatemala * Japan * Jamaica * Peru * Ecuador * Haiti * Nicaragua * Taiwan * Portugal * Iran * Greece * Honduras * Ireland * France * Outlying-U S (Guam USVI etc) * Thailand * Laos * Hong Kong * Cambodia * Hungary * Scotland * Trinadad&Tobago * Yugoslavia * Panama * Holand-Netherlands | Kategorik |
| citizenship | Kewarganegaraan   * Native- Born in the United States * Foreign born- Not a citizen of U S * Foreign born- U S citizen by naturalization * Native- Born abroad of American Parent(s) * Native- Born in Puerto Rico or U S Outlying | Kategorik |
| own\_or\_self | Keterangan memiliki usaha atau tidak | Numerik |
| vet\_question | Mendapatkan pertanyaan tentang veteran   * Not in universe * No * Yes | Kategorik |
| vet\_benefits | Keuntungan veteran yang diterima | Numerik |
| weeks\_worked | Minggu bekerja dalam setahun | Numerik |
| year | Tahun mendapat survey | Numerik |
| income\_50k | Pendapatan kurang atau lebih dari $50,000 | Kategorik   * - 50000. * 50000+. |