Project 5 : Marketing Data

Penerapan Metode *Classification And Regression Trees* (Cart) pada Data Marketing

KELOMPOK 1:	**Muhammad Tibri Syofyan**	**(18337021)**
	Andini Yulianti	**(18337010)**
	Muhamad Fajri	**(18337015)**
	Aprilla Suhada	**(18337031)**
	Ihsanul Fikri	**(18337051)**

A. Pendahuluan

Pesatnya perkembangan *internet* dan *e-commerce* telah membawa tantangan pada metode pemasaran tradisional. Hanya metode pemasaran yang akurat yang dapat membantu perusahaan tidak tersingkir dalam persaingan harga pasar yang ketat. Kegiatan pemasaran tidak dapat dipisahkan dari data. Dalam data pemasaran yang berjumlah besar, terdapat banyak informasi pelanggan yang berharga. Melalui statistik, pengumpulan, dan analasis pada data, presisi pemasaran dapat diwujudkan, yang mana dapat menggurangi biaya pemasaran dan meningkatkan efisiensi pemasaran.

Melalui artikel ini, kami mencoba untuk mengklasifikasi data marketing pada suatu perusahaan menggunakan metode *classification and regression trees* (CART). Proses penganalisisan data marketing diolah menggunakan software RStudio. Data marketing diperoleh dari sebuah situs kaggle.com. Alasan kami menggunakan analisis CART adalah kami ingin mengklasifikasi karakteristik pelanggan dengan kampanye terakhir yang telah diadakan. Sehingga perusahaan dapat meningkatkan kinerja dalam memperkenalkan suatu produk yang dihasilkan. Metode CART efektif bila digunakan pada data dengan pengamatan yang relatif banyak (Du Toit, 1986).

B. Eksplorasi Data Analysis

Data yang digunakan adalah data marketing yang tersedia di kaggle.com dengan url https://www.kaggle.com/jackdaoud/marketing-data. Dataset ini terdiri dari 2240 observasi dan 28 variabel. Data marketing ini berisikan tentang profil pelanggan, jumlah produk yang dikonsumsi selama 2 tahun terakhir, kinerja channel, dan hasil kampanye. Deskripsi variabelnya adalah sebagai berikut:

ID	:	Pengenal unik pelanggan
Year_Birth	:	Tahun lahir pelanggan
Education	:	Tingkat pendidikan pelanggan
Marital_Status	:	Status perkawinan pelanggan
Income	:	Pendapatan rumah tangga tahunan pelanggan
Kidhome	:	Jumlah anak kecil dalam rumah tangga pelanggan
Teenhome	:	Jumlah anak remaja dalam rumah tangga pelanggan
Dt_Customer	:	Tanggal pendaftaran pelanggan dengan perusahaan
Recency	:	Jumlah hari sejak pembelian terakhir pelanggan
MntWines	:	Jumlah yang dihabiskan untuk wine dalam 2 tahun terakhir
MntFruits	:	Jumlah yang dihabiskan untuk buah-buahan dalam 2 tahun terakhir
MntMeatProducts	:	Jumlah yang dihabiskan untuk daging dalam 2 tahun terakhir
MntFishProducts	:	Jumlah yang dihabiskan untuk ikan dalam 2 tahun terakhir
MntSweetProducts	:	Jumlah yang dihabiskan untuk permen dalam 2 tahun terakhir
MntGoldProds	:	Jumlah yang dihabiskan untuk emas dalam 2 tahun terakhir
NumDealsPurchases	:	Jumlah pembelian yang dilakukan melalui diskon
NumWebPurchases	:	Jumlah pembelian yang dilakukan melalui situs web perusahaan
NumCatalogPurchases	:	Jumlah pembelian yang dilakukan melalui katalog
NumStorePurchases	:	Jumlah pembelian yang dilakukan langsung di toko
NumWebVisitsMonth	:	Jumlah kunjungan ke situs web perusahaan dalam sebulan terakhir
AcceptedCmp1	:	1 jika pelanggan menerima tawaran dalam kampanye ke-1, 0 sebaliknya
AcceptedCmp2	:	1 jika pelanggan menerima tawaran dalam kampanye ke-2, 0 sebaliknya
AcceptedCmp3	:	1 jika pelanggan menerima tawaran dalam kampanye ke-3, 0 sebaliknya
AcceptedCmp4	:	1 jika pelanggan menerima tawaran dalam kampanye ke-4, 0 sebaliknya
AcceptedCmp5	:	1 jika pelanggan menerima tawaran dalam kampanye ke-5, 0 sebaliknya
Response	:	1 jika pelanggan menerima tawaran dalam kampanye terakhir, 0 sebaliknya
Complain	:	1 jika pelanggan mengeluh dalam 2 tahun terakhir, 0 sebaliknya
Country	:	Asal pelanggan

Kita akan melihat skala data pada dataset marketing ini (hasil ditampilkan pada Gambar 1). Pada Gambar 1 ditampilkan 16 variabelbertipe data numerik, 10 variabel bertipe data kategorik, 1 variabel bertipe data karakter dan 1 variabel bertipe data tanggal. Pada variabel *education*, kategori "*2n Cycle*" dan "*Master*" merupakan tingkatan pendidikan yang sama. Sehingga kita akan menggabungkan dua kategori ini menjadi kategori "*Master*". Kategori "*Graduation*" merupakan penggunaan kata yang salah untuk tingkat pendidikan, maka kita akan mengansumsikan menjadi "*Undergraduate*". Pada variabel *marital_status*, kategori "*YOLO*","*Alone*", dan "*Absurd*" dapat diasumsikan menjadi kategori "*Single*".

	Variables	Data Type
1	ID	character
2	Year_Birth	numeric
3	Education	factor
4	Marital_Status	factor
5	Income	integer
6	Kidhome	numeric
7	Teenhome	numeric
8	Dt_Customer	Date
9	Recency	numeric
10	MntWines	numeric
11	MntFruits	numeric
12	MntMeatProducts	numeric
13	MntFishProducts	numeric
14	MntSweetProducts	numeric
15	MntGoldProds	numeric
16	NumDealsPurchases	numeric
17	NumWebPurchases	numeric
18	NumCatalogPurchases	numeric
19	NumStorePurchases	numeric
20	NumWebVisitsMonth	numeric
21	AcceptedCmp3	factor
22	AcceptedCmp4	factor
23	AcceptedCmp5	factor
24	AcceptedCmp1	factor
25	AcceptedCmp2	factor
26	Response	factor
27	Complain	factor
28	Country	factor

Gambar 1. Output Skala Data

Kemudian akan dilihat *missing value* pada data yang ditampilkan dengan menggunakan syntax *summary()* pada software RStudio.

```
> summary(mydata)
      ID          Year_Birth          Education     Marital_Status     Income          Kidhome          Teenhome       Dt_Customer            Recency     
 Length:2240        Min.   :1893   Basic      :  54   Divorced:232   Min.   :  1730   Min.   :0.0000   Min.   :0.0000   Min.   :0012-07-30   Min.   : 0.00  
 Class :character   1st Qu.:1959   Master     : 573   Married :864   1st Qu.: 35303   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0013-01-16   1st Qu.:24.00  
 Mode  :character   Median :1970   PhD        : 486   single  :487   Median : 51382   Median :0.0000   Median :0.0000   Median :0013-07-08   Median :49.00  
                    Mean   :1969   Undergraduate:1127   Together:580   Mean   : 52247   Mean   :0.4442   Mean   :0.5062   Mean   :0013-07-10   Mean   :49.11  
                    3rd Qu.:1977                        Widow   : 77   3rd Qu.: 68522   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0013-12-30   3rd Qu.:74.00  
                    Max.   :1996                                       Max.   :666666   Max.   :2.0000   Max.   :2.0000   Max.   :0014-06-29   Max.   :99.00  
                                                                       NA's   :24                                                                             

    MntWines         MntFruits      MntMeatProducts  MntFishProducts  MntSweetProducts  MntGoldProds     NumDealsPurchases NumWebPurchases  NumCatalogPurchases
 Min.   :   0.00   Min.   :  0.0   Min.   :   0.0   Min.   :  0.00   Min.   :  0.00   Min.   :  0.00   Min.   : 0.000    Min.   : 0.000   Min.   : 0.000     
 1st Qu.:  23.75   1st Qu.:  1.0   1st Qu.:  16.0   1st Qu.:  3.00   1st Qu.:  1.00   1st Qu.:  9.00   1st Qu.: 1.000    1st Qu.: 2.000   1st Qu.: 0.000     
 Median : 173.50   Median :  8.0   Median :  67.0   Median : 12.00   Median :  8.00   Median : 24.00   Median : 2.000    Median : 4.000   Median : 2.000     
 Mean   : 303.94   Mean   : 26.3   Mean   : 166.9   Mean   : 37.53   Mean   : 27.06   Mean   : 44.02   Mean   : 2.325    Mean   : 4.085   Mean   : 2.662     
 3rd Qu.: 504.25   3rd Qu.: 33.0   3rd Qu.: 232.0   3rd Qu.: 50.00   3rd Qu.: 33.00   3rd Qu.: 56.00   3rd Qu.: 3.000    3rd Qu.: 6.000   3rd Qu.: 4.000     
 Max.   :1493.00   Max.   :199.0   Max.   :1725.0   Max.   :259.00   Max.   :263.00   Max.   :362.00   Max.   :15.000    Max.   :27.000   Max.   :28.000     

 NumStorePurchases NumWebVisitsMonth AcceptedCmp3 AcceptedCmp4 AcceptedCmp5 AcceptedCmp1 AcceptedCmp2 Response Complain    Country     
 Min.   : 0.00     Min.   : 0.000    0:2077       0:2073       0:2077       0:2096       0:2210       0:1906   0:2219   SP     :1095  
 1st Qu.: 3.00     1st Qu.: 3.000    1: 163       1: 167       1: 163       1: 144       1:  30       1: 334   1:  21   SA     : 337  
 Median : 5.00     Median : 6.000                                                                                        CA     : 268  
 Mean   : 5.79     Mean   : 5.317                                                                                        AUS    : 160  
 3rd Qu.: 8.00     3rd Qu.: 7.000                                                                                        IND    : 148  
 Max.   :13.00     Max.   :20.000                                                                                        GER    : 120  
                                                                                                                         (Other): 112  
```

Gambar 2. Output Syntax Summary()

Berdasarkan Gambar 2, terdapat*missing value* atau data hilang pada variabel *income* sebanyak 24 data. Maka perlu diatasi kasus *missing value* pada dataset ini. Untuk melakukan penanganan *missing value*dapat dilakukan dengan manghapus observasi atau mengganti nilai *missing* dengan suatu nilai rataan atau median. Kita memilih menangani kasus ini dengan mengganti nilai *missing* dengan suatu nilai rataan pada variabel *income*. Dengan syntax yang digunakan ditampilkan pada Gambar 3.

```
mydata1 <- data.frame(impute(mydata$Income,mean))
names(mydata1)[1] <- "imputeincome"
data <- data.frame(c(mydata[-5],mydata1))
summary(data)
```

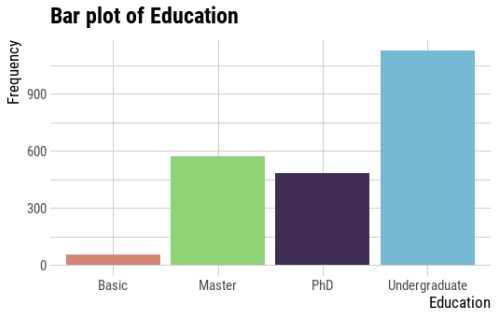
Gambar 3. Syntax Penanganan *Missing Value*

Hasil yang diperolehmenyatakan bahwa dataset marketing data tidak terdapat *missing value* setelah dilakukan imputasi pada variabel *imputeincome*. Hasil dapat dilihat pada Gambar 4.

```
> summary(data)

24 values imputed to 52247.25

       ID              Year_Birth          Education     Marital_Status     Kidhome          Teenhome       Dt_Customer            Recency
 Length:2240        Min.   :1893   Basic      :  54   Divorced:232    Min.   :0.0000   Min.   :0.0000   Min.   :0012-07-30   Min.   : 0.00
 Class :character   1st Qu.:1959   Master     : 573   Married :864    1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0013-01-16   1st Qu.:24.00
 Mode  :character   Median :1970   PhD        : 486   Single  :487    Median :0.0000   Median :0.0000   Median :0013-07-08   Median :49.00
                    Mean   :1969   Undergraduate:1127   Together:580    Mean   :0.4442   Mean   :0.5062   Mean   :0013-07-10   Mean   :49.11
                    3rd Qu.:1977                        Widow   : 77    3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0013-12-30   3rd Qu.:74.00
                    Max.   :1996                                       Max.   :2.0000   Max.   :2.0000   Max.   :0014-06-29   Max.   :99.00

    MntWines         MntFruits       MntMeatProducts  MntFishProducts  MntSweetProducts  MntGoldProds     NumDealsPurchases NumWebPurchases  NumCatalogPurchases
 Min.   :   0.00   Min.   :  0.0   Min.   :   0.0   Min.   :  0.00   Min.   :  0.00   Min.   :  0.00   Min.   : 0.000    Min.   : 0.000   Min.   : 0.000
 1st Qu.:  23.75   1st Qu.:  1.0   1st Qu.:  16.0   1st Qu.:  3.00   1st Qu.:  1.00   1st Qu.:  9.00   1st Qu.: 1.000    1st Qu.: 2.000   1st Qu.: 0.000
 Median : 173.50   Median :  8.0   Median :  67.0   Median : 12.00   Median :  8.00   Median : 24.00   Median : 2.000    Median : 4.000   Median : 2.000
 Mean   : 303.94   Mean   : 26.3   Mean   : 166.9   Mean   : 37.53   Mean   : 27.06   Mean   : 44.02   Mean   : 2.325    Mean   : 4.085   Mean   : 2.662
 3rd Qu.: 504.25   3rd Qu.: 33.0   3rd Qu.: 232.0   3rd Qu.: 50.00   3rd Qu.: 33.00   3rd Qu.: 56.00   3rd Qu.: 3.000    3rd Qu.: 6.000   3rd Qu.: 4.000
 Max.   :1493.00   Max.   :199.0   Max.   :1725.0   Max.   :259.00   Max.   :263.00   Max.   :362.00   Max.   :15.000    Max.   :27.000   Max.   :28.000

 NumStorePurchases NumWebVisitsMonth AcceptedCmp3 AcceptedCmp4 AcceptedCmp5 AcceptedCmp1 AcceptedCmp2 Response Complain    Country      imputeincome
 Min.   : 0.00     Min.   : 0.000    0:2077      0:2073      0:2077      0:2096      0:2210      0:1906   0:2219   SP     :1095   Min.   :  1730
 1st Qu.: 3.00     1st Qu.: 3.000    1: 163      1: 167      1: 163      1: 144      1:  30      1: 334   1:  21   SA     : 337   1st Qu.: 35539
 Median : 5.00     Median : 6.000                                                                                  CA     : 268   Median : 51742
 Mean   : 5.79     Mean   : 5.317                                                                                  AUS    : 160   Mean   : 52247
 3rd Qu.: 8.00     3rd Qu.: 7.000                                                                                  IND    : 148   3rd Qu.: 68290
 Max.   :13.00     Max.   :20.000                                                                                  GER    : 120   Max.   :666666
                                                                                                                  (Other): 112
```

Gambar 4. Ringkasan Data Setelah Penanganan *Missing Value*

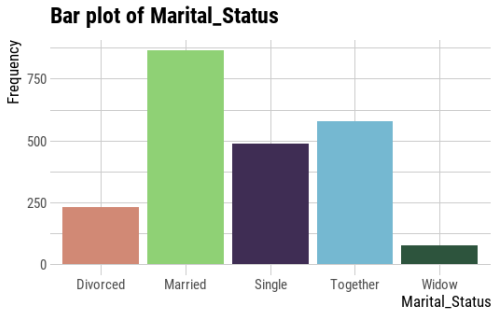
Selanjutnya akan divisualisasikan variabel yang terdapat pada dataset *marketing_data* sebagai berikut



Gambar 5. Diagram Batang Untuk Tingkat Pendidikan

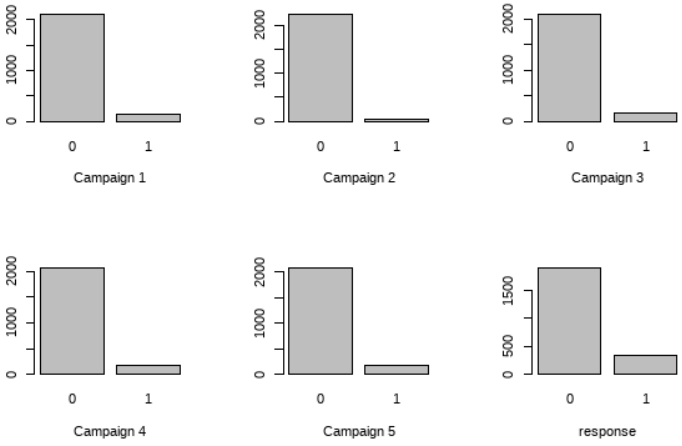
Berdasarkan Gambar 5, menunjukkan bahwa sebagian besar pelanggan merupakan lulusan sarjana. Hal ini dikarenakan lulusan sarjana lebih berpeluang mendapatkan peluang usaha atau pekerjaan dibandingkan dengan yang bukan lulusan sarjana.Pada umumnya, tingkat pendidikan sarjana dijadikan sebagai pendidikan minimal yang harus diselesaikan, dan untuk

melanjutkan ketahap selanjutnya seperti magister dan professor membutuhkan biaya mahal.Sehinggasebagian besar orang memiliki pendidikan terakhir sarjana.



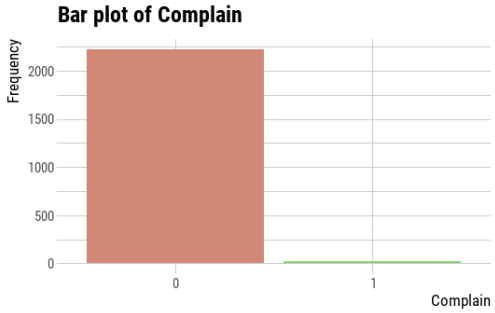
Gambar 6. Diagram Batang Untuk Status

Berdasarkan Gambar 6, menunjukkan bahwa sebagian besar pelanggan telah menikah, diikuti dengan status berpasangan, dan *single*. Hal ini dapat disimpulkan bahwa pelanggan yang telah menikah lebih mementingkan finansial dan sanggup untuk memenuhi kebutuhan berkeluarga, dibandingkan pelanggan yang tidak menikah.



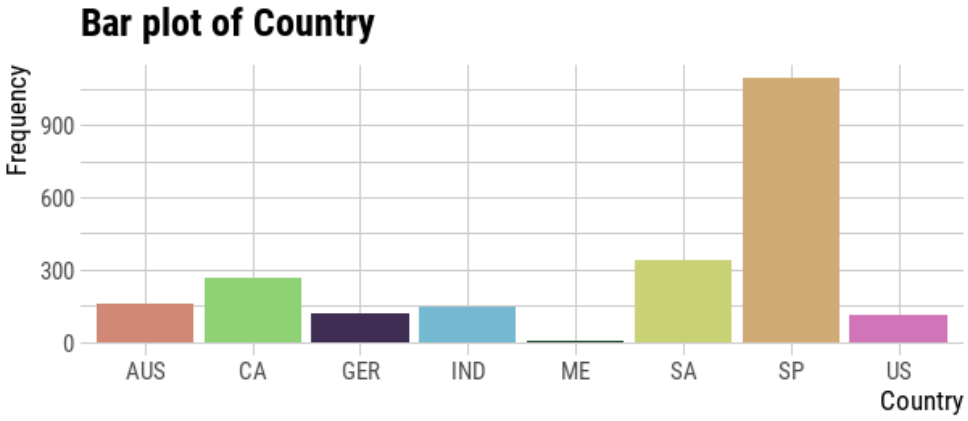
Gambar 7. Diagram Batang Untuk Kampanye

Berdasarkan Gambar 7, dimana kampanyetelah dilakukan sebanyak enam kali, terlihat bahwa kampanye ke-2 memiliki kesuksesan yang lebih tinggi dan lebih menarik pelanggan dibandingkan kampanye lainnya.



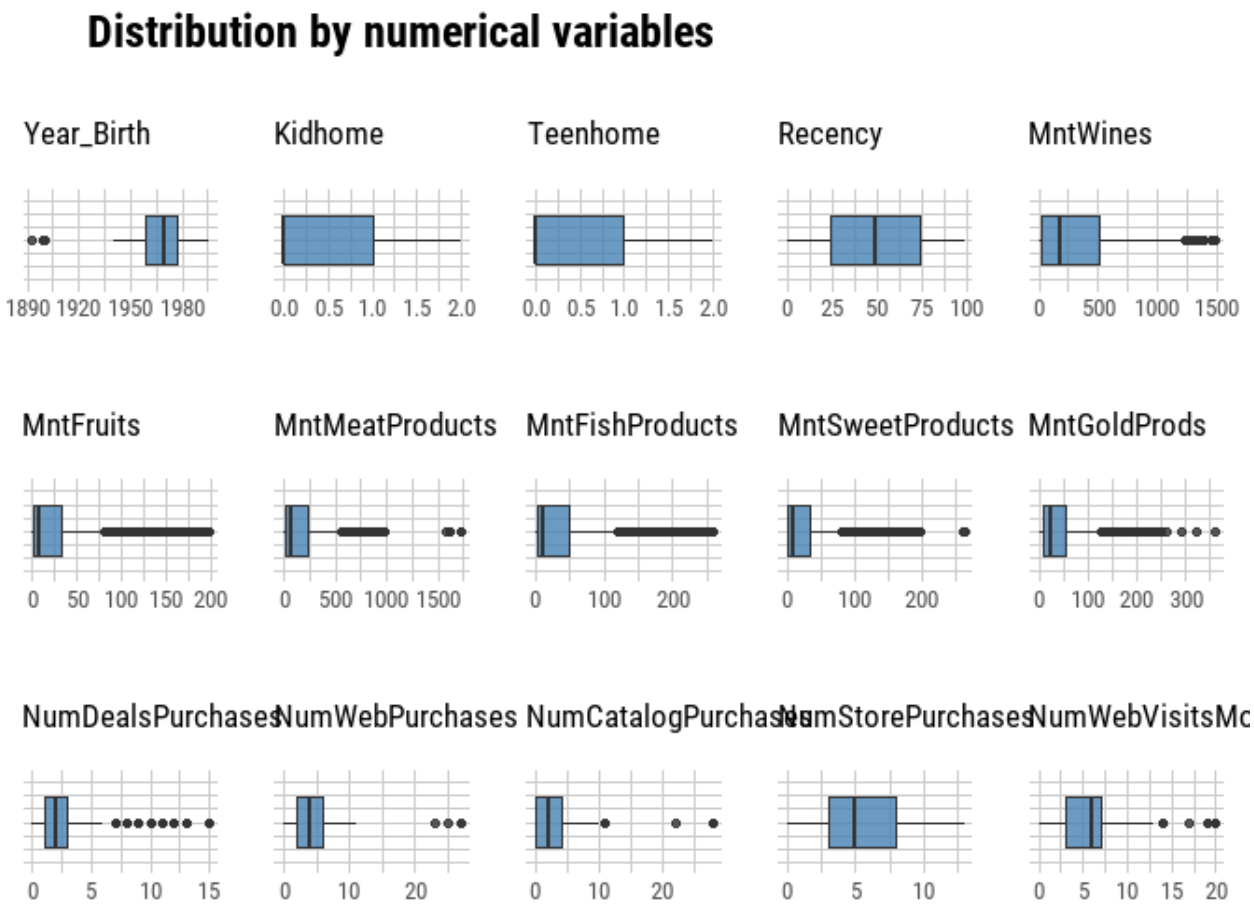
Gambar 8. Diagram Batang Untuk Komplain

Berdasarkan Gambar 8, memperlihatkan bahwa dari 2240 pelanggan yang diobservasi, banyak pelanggan yang melakukan *complain*. Hal ini perlu menjadi perhatian bagi perusahaan untuk menjawab keluhan pelanggan.

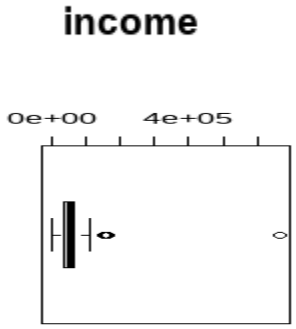


Gambar 9. Diagram Batang Untuk Negara

Berdasarkan Gambar 9, memperlihatkan bahwa sebagian besar pelanggan berasal dari negara Spanyol.



Gambar 10. Boxplot Untuk Masing-Masing Variabel



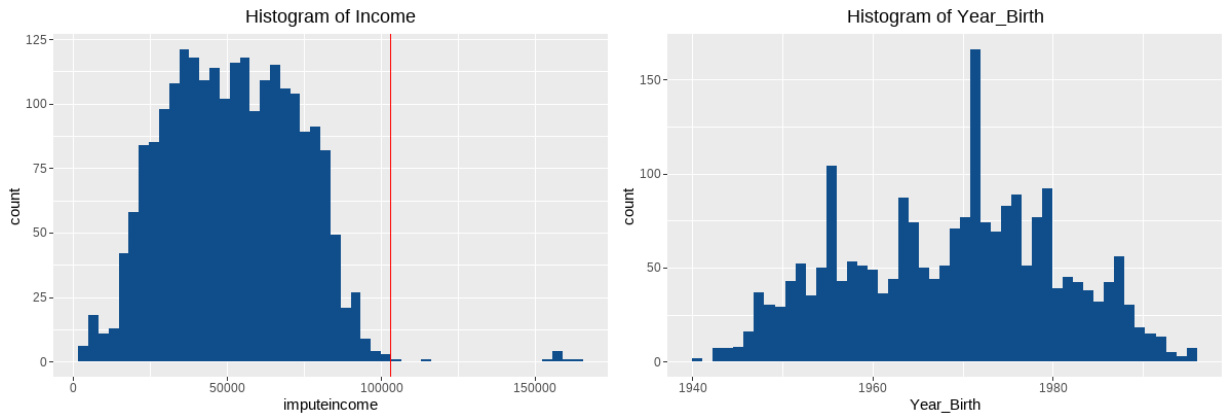
Gambar 11. Boxplot untuk Income

Berdasarkan Gambar 10 dan 11, terlihat bahwa adanya *outlier* pada dataset *marketing_data*. *Outlier* ini terdapat pada variabel *income* dan *year_birth*yang berada jauh di sekitar rataan. Mari kita lihat id pelanggan yang dideteksi sebagai data *outlier* yang ditampilkan pada Tabel 1.

Tabel 1. Data Outlier

ID	Year_Birth	Income
514	1893	60182
528	1977	666666
828	1899	83532
2234	1900	36640

Terdapat 4 pelanggan yang merupakan data *outlier*. Dikarenakan pelanggan dengan ID 528 terdata penghasilannya sebesar $666.666 dan ada 3 pelanggan dengan ID 528, 828, dan 2234 terdata memiliki umur yang lebih dari 100 tahun. Maka kita perlu mengeluarkan data *outlier* pada dataset *marketing_data*.

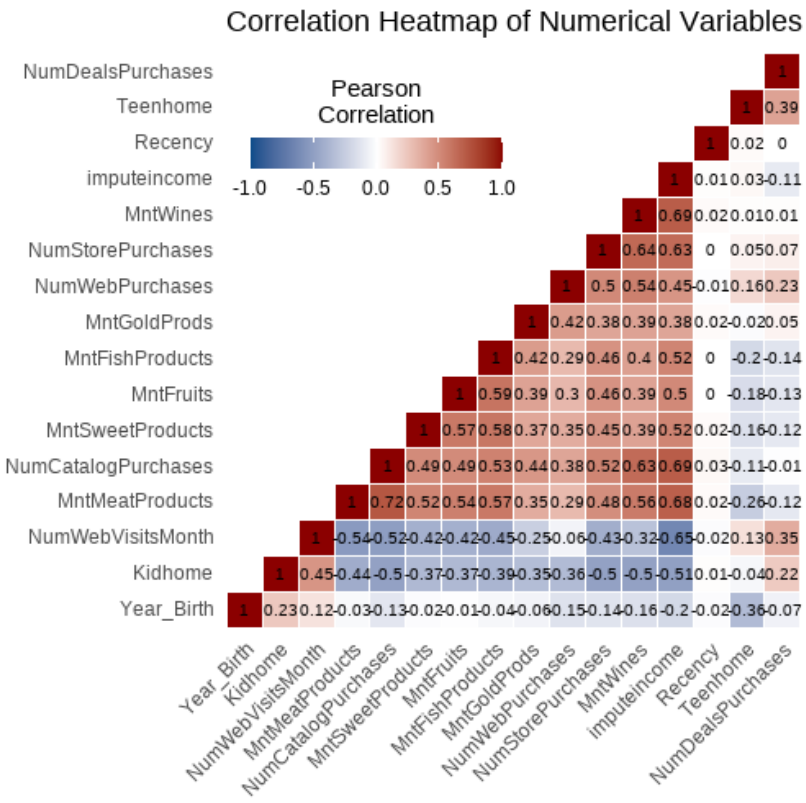


Gambar 12. Histogram dari Income dan Tahun Lahir

Berdasarkan Gambar 12. menunjukkan bahwa tidak ada *outlier* pada data. Pada histogram *income* terdapat beberapa data sedikit menjauh dari kelompok data, hal ini dapat kita anggap bahwa sedikit orang yang berpenghasilan lebih banyak dari kebanyakan orang. Sehingga kita tidak mengganggap data tersebut sebagai data *outlier*.*Outlier* timbul dikarenakan adanya kesalahan dalam pengentrian data, gagal menspesifikasi adanya *missing value* dalam program komputer, *outlier* bukan anggota populasi, dan*outlier* bernilai ekstrim sehingga data tidak berdistribusi normal.

Kemudian akan kita lihat korelasi antar variabel. Semakin merah indikator maka hubungan antara dua variabel semakin berkorelasi positif, dan sebaliknya jika indikator semakin biru maka hubungan antara dua variabel semakin berkorelasi negatif. Korelasi positif artinya bahwa hubungan antara dua variabel itu menunjukkan arah yang sama, jadi apabila salah satu variabel mengalami kenaikan atau pertambahan maka akan diikuti dengan kenaikan pada variabel

satunya. Begitupun dengan korelasi negatif, jika salah satu variabel mengalami kenaikan maka variabel satunya mengalami penuruan.



Gambar 12. Heatmap Korelasi Masing-Masing Variabel

Dari hasil output diatas terlihat bahwa korelasi tertinggi terjadi antara NumCatalogPurchases dengan imputeincome dan MntWines dengan imputeincome, dengan nilai korelasi sebesar 0,69, yang artinya jika imputeincome mengalami kenaikan maka kenaikan akan terjadi pula pada NumCatalogPurchases dan MntWines. Dan korelasi paling rendah terjadi antara NumWebVisitMonth dengan imputeincome dengan nilai korelasi sebesar -0,65.

Berdasarkan dataset ini, kita dapat merekayasa informasi baru antara lain:

1. Dari variabel tahun lahir pelanggan, kita memperoleh umur pelanggan.
2. Dari variabel umur pelanggan, kita dapat memperoleh demografi usia pelanggan, yaitu
    a. *Baby Boomer* adalah pelanggan dengan umur di atas 54 tahun.
    b. Gen X adalah pelanggan dengan umur antara 38 hingga 54 tahun.
    c. Gen Y adalah pelanggan dengan umur antara 18 hingga 38 tahun.
    d. Gen Z adalah pelanggan dengan umur di bawah 18 tahun.

3. Dari variabel *Kidhome* dan *Teenhome*, kita memperoleh total anak pelanggan
4. Dari enam variabel tentang konsumsi produk selama dua tahun terakhir, kita memperoleh jumlah total yang dibelanjakan.
5. Dari tiga variabel tentang banyaknya pembelian pada tiga channel (Web, Store, Catalog), kita dapat memperoleh jumlah pembelian.

C. Big Data Analysis

1. *Clasiffication and Regression Trees* (CART)

CART merupakan salah satu metode atau algoritma dari teknik pohon keputusan (*decision tree*). Metode yang dikembangkan oleh Leo Breiman, Jerome H. Friedman, Richard A. Olshen, dan Charles J. Stone ini merupakan teknik klasifikasi dengan menggunakan algoritma penyekatan rekursif secara biner (*binary recursive partitioning*) (Roger dan Lewis, 2000). Istilah "*binary*" berarti pemilahan dilakukan pada sekelompok data yang terkumpul dalam suatu ruang yang disebut simpul (*node*) menjadi dua kelompok yang disebut simpul anak (*child nodes*). Istilah "*recursive*" berarti prosedur penyekatan secara biner dilakukan secara berulang-ulang.Setiap simpul anak yang diperoleh dari penyekatan simpul awal kemudian bisa dipilah kembali menjadi dua simpul anak lagi, dan begitu seterusnya hingga memenuhi kriteria tertentu. Sedangkan istilah "*partitioning*" memiliki arti bahwa proses klasifikasi dilakukan dengan cara memilah suatu kumpulan data menjadi beberapa bagian atau partisi.

Menurut Breiman (1993), CART akan menghasilkan pohon klasifikasi jika variabel respon mempunyai skala kategorik dan akan menghasilkan pohon regresi jika variabel respon berupa data kontinu. Keuntungan dari penggunaan analisis CART adalah sebagai berikut :

a. Merupakan bentuk statistika non-parametrik, sehingga tidak memerlukan asumsi sebaran dan uji hipotesis.
b. Tidak memerlukan variabel untuk dipilih sebelumnya.
c. Sangat efisien dalam terminologi perhitungan.
d. Dapat menangani dataset dengan struktur yang kompleks.
e. Sangat tangguh dalam menangani outlier, umumnya algoritma pemisahan akan mengisolasikan outlier pada individu node atau beberapa node.
f. Dapat menggunakan sembarang kombinasi data kontinu/numeric dan kategorik.

g. Hasilnya invarian dengan transformasi monoton dan variabel respon, artinya penggantian sembarang variabel dengan algoritmanya atau nilai akar kuadrat, tidak akan menyebabkan struktur pohon berubah.

Kelemahan dari analisis CART sebagai berikut:

a. CART mungkin tidak stabil dalam *decision trees* (pohon keputusan) karena CART sangat sensitif dengan data baru. CART sangat bergantung dengan jumlah sampel. Jika sampel data *learning* dan *testing* berubah maka pohon keputusan yang dihasilkan juga ikut berubah.

b. Tiap pemilihan bergantung pada nilai yang hanya berasal dari satu variabel penjelas.

Algoritma CART melalui tiga tahapan, yaitu pembentukan pohon klasifikasi, pemangkasan pohon klasifikasi dan penentuan pohon klasifikasi optimum.

a. Pembentukan Pohon Klasifikasi

Pembentukan pohon klasifikasi diawali dengan menentukan variabel dan nilai dari variabel tersebut (*threshold*) untuk dijadikan pemilah tiap simpul.Dalam prosesnya, pembentukan pohon klasifikasi dibutuhkan data *training* sampel L yang terdiri atas N pengamatan. Menurut Breiman (1993), proses pembentukan pohon klasifikasi terdiri atas 3 tahapan yaitu sebagai berikut.

1) Pemilihan Pemilah

Pada tahap ini, data yang digunakan adalah sampel data *training* L yang kemudian dipilah berdasarkan aturan pemilahan dan kriteria *goodness of split*. Himpunan bagian yang dihasilkan dari proses pemilahan harus lebih homogen dibandingkan pemilahan sebelumnya. Hal ini dilakukan dengan mendefinisikan keheterogenan simpul *(impurity* atau *i(t)*). Menurut Breiman, et al (1993), fungsi keheterogenan yang sangat mudah dan sesuai diterapkan dalam berbagai kasus adalah Indeks Gini. Indeks Gini akan selalu memisahkan kelas dengan anggota paling besar atau kelas terpenting dalam simpul tersebut terlebih dahulu. Pemilahan yang memberikan nilai penurunan keheterogenan tertinggi merupakan pemilahan terbaik. Fungsi Indeks Gini dituliskan dalam persamaan berikut.

$$i(t) = \sum_{i,j=1} p(j|t)p(i|t), i \neq j$$

Dengan $p(j|t)$ adalah proporsi kelas j pada simpul t dan $p(i|t)$ adalah proporsi kelas i pada simpul t.

Pemilahan simpul dimulai dengan memeriksa nilai–nilai variabel penjelas dan dilakukan secara rekursif pada setiap simpul dengan dua tahapan. Tahapan yang pertama adalah mencari semua kemungkinan pemilah pada variabel penjelas. Menurut Breiman (1993), proses pemilahan simpul menjadi dua simpul anak dilakukan dengan mengikuti aturan sebagai berikut.

a) Setiap pemilahan hanya bergantung pada nilai yang berasal dari satu variabel penjelas saja.

b) Apabila variabel penjelas berskala kontinu, maka pemilahan yang diperbolehkan adalah $x_j \leq c_i$ dan $x_j > c_i$, dengan i = 1, 2, …, n-1dan c_i adalah nilai tengah dari dua nilai amatan sampel berurutan yang berbeda dari variabel X_j. jika suatu ruang sampel berukuran n dan terdapat n nilai amatan sampel yang berbeda pada variabel X_j, maka akan terdapat sebanyak n-1 kemungkinan pemilahan yang berbeda.

c) Apabila variabel penjelas berskala kategorik, maka pemilahan berasal dari semua kemungkinan pemilahan berdasarkan terbentuknya dua simpul yang saling lepas (disjoint). Apabila variabel penjelas berskala nominal bertaraf L, maka akan diperoleh sebanyak 2^{L-1}-1 pemilahan yang mungkin. Akan tetapi, apabila kategori variabel penjelas berskala ordinal bertaraf L, maka akan diperoleh sebanyak L-1 pemilahan yang mungkin.

Pemilahan yang terpilih akan membentuk suatu himpunan kelas yang disebut sebagai simpul. Simpul tersebut akan melakukan pemilahan secara rekursif sampai diperoleh simpul akhir (*terminal nodes*).

Setelah dilakukan pemilahan dari semua kemungkinan pemilah, maka tahapan berikutnya adalah menentukan kriteria *goodness of split* (*ϕ(s,t)*) untuk mengevaluasi pemilah dari pemilah *s* pada simpul t. *Goodness of split* (*ϕ(s,t)*) merupakan penurunan heterogenitas, yaitu:

$$\emptyset(s,t) = \Delta i(s,t) = i(t) - p_L i(t_L) - p_R i(t_R)$$

Dengan $i(t)$: fungsi heterogenitas pada simpul t

p_L : proporsi pengamatan menuju simpul kiri

p_R : proporsi pengamatan menuju simpul kanan

$i(t_L)$: fungsi heterogenitas pada simpul anak kiri

$i(t_R)$: fungsi heterogenitas pada simpul anak kanan

Pemilah yang menghasilkan $\phi(s,t)$ lebih tinggi merupakan pemilah terbaik karena mampu mereduksi heterogenitas lebih tinggi. Pengembangan pohon ini dilakukan dengan pencarian pemilah yang mungkin pada simpul t_1 yang kemudian akan dipilah menjadi t_2 dan t_3 oleh pemilah *s*, dan seterusnya. t_L dan t_R merupakan partisi dari simpul *t* menjadi dua himpunan

bagian saling lepas dimana p_L dan p_R adalah proporsi masing-masing peluang simpul. Karena $tL \cup tR = t$ maka nilai $\Delta i(s,t)$ merepresentasikan perubahan dari kehetoregenan dalam simpul *t* yang semata-semata disebabkan oleh pemilah *s*. Jika simpul yang diperoleh merupakan kelas yang tidak homogen, prosedur yang sama diulangi sampai pohon klasifikasi menjadi suatu konfigurasi dan memenuhi:

$$\Delta i(s^* t_1) = max_{sCS} \Delta i(s, t_1)$$

2) Penentuan Simpul Terminal

Suatu simpul *t*akan menjadi simpul terminal atau tidak, akan dipilah kembali apabila pada simpul *t* tidak terdapat penurunan keheterogenan secara berarti atau adanya batasan minimum sebesar *n* seperti hanya terdapat satu pengamatan pada tiap simpul anak. Menurut Breiman, et al (1993), pengembangan pohon akan berhenti apabila pada simpul terdapat pengamatan berjumlah kurang dari atau sama dengan 5 (n ≤ 5). Selain itu, proses pembentukan pohon juga akan berhenti apabila sudah mencapai batasan jumlah level yang telah ditentukan atau tingkat kedalaman (depth) dalam pohon maksimal.

3) Penandaan Label Kelas

Penentuan label kelas pada simpul terminal berdasarkan aturan jumlah terbanyak, yaitu jika

$$p(j_0|t) = max_j p(j|t) = max_j \frac{N_j(t)}{N(t)}$$

dengan:

p(j|t) : proporsi kelas j pada simpul t

$N_j(t)$: jumlah pengamatan kelas j pada terminal node t

N(t) : jumlah total pengamatan pada terminal node t

Label kelas untuk simpul terminal *t* adalah j_0 yang memberikan nilai dugaan kesalahan pengklasifikasian pada simpul *t* yang paling kecil sebesar $r(t) = 1 - max_j\ p(j|t)$. Proses pembentukan pohon klasifikasi berhenti ketika terdapat hanya satu pengamatan dalam tiap-tiap simpul anak atau adanya batasan minimum *n*, semua pengamatan dalam tiap simpul anak identik, dan adanya batasan jumlah level/kedalaman pohon maksimal. Setelah pembentukan pohon maksimal, tahap selanjutnya adalah pemangkasan pohon untuk mencegah terbentuknya pohon klasifikasi yang berukuran besar dan kompleks.

b. Pemangkasan Pohon Klasifikasi

Pohon yang dibentuk dengan aturan pemilah dan kriteria *goodness of split* berukuran sangat besar karena penghentian pohon berdasarkan banyaknya amatan pada simpul terminal atau besarnya tingkat kehomogenan. Jika semakin banyak pemilahan yang dilakukan, maka dapat mengakibatkan kecilnya tingkat kesalahan prediksi, akan tetapi akibatnya pohon klasifikasi yang dibentuk berukuran besar. Ukuran pohon yang besar dapat dapat memunculkan adanya *overfitting*, akan tetapi apabila pengamatan pohon dibatasi dengan ketepatan batas tertentu, maka dapat terjadi kasus *underfitting*. Oleh karena itu, untuk mendapatkan pohon yang layak, maka perlu dilakukan pemangkasan pruning yaitu suatu penilaian ukuran pohon tanpa mengorbankan akurasi yang berarti.Pemangkasan ini dilakukan dengan melakukan pengurangan simpul pohon sehingga dicapai ukuran pohon yang layak dan tidak terlalu melebar. Menurut Breiman dkk (1993), ukuran pohon yang layak dapat dilakukan dengan pemangkasan pohon dengan ukuran *cost complexity minimum*.

$$R_\alpha(T) = R(T) + \alpha\left|\tilde{T}\right|$$

dengan:

$R_\alpha(T)$ = Resubtitusion Estimate (proporsi kesalahan pada pohon T)

α = kompleksitas parameter (complexity parameter)

$\left|\tilde{T}\right|$= ukuran banyaknya simpul terminal pohon T

$R_\alpha(T)$merupakan kombinasi linear biaya dan kompleksitas pohon yang dibentuk dengan menambahkan *cost penalty* bagi kompleksitas terhadap biaya kesalahan klasifikasi pohon. *Cost complexity pruning* menentukan suatu pohon bagian *T(α)* yang meminimumkan $R_\alpha(T)$ pada seluruh pohon bagian atau untuk setiap nilai α. Selanjutnya, dilakukan pencarian pohon bagian *T(α)<Tmax* yang meminimumkan $R_\alpha(T)$ yaitu

$$R_\alpha(T(\alpha)) = min_{T<Tmax} R_\alpha(T)$$

Pemangkasan pohon dimulai dengan mengambil t_R dan t_L dari T_{max} yang dihasilkan dari simpul induk *t*. jika diperoleh dua simpul anak dan simpul induk yang memenuhi persamaan $R(t)= R(t_R) + R(t_L)$, maka simpul anak t_R dan t_L dipangkas. Sehingga diperoleh pohon T_1 yang memenuhi kriteria $R(T_1) = R(T_{max})$. Jika R(T) digunakan sebagai kriteria penentuan pohon optimal maka akan cenderung akan dipilih pohon terbesar T_i. sebab semakin besar pohon, maka semakin kecil nilai R(T) nya.

c. Penentuan Pohon Klasifikasi

Optimal Pohon klasifikasi yang berukuran besar akan memberikan nilai *cost complexity* yang tinggi karena struktur data yang digambarkan cenderung kompleks sehingga perlu dipilih pohon optimal yang berukuran sederhana tetapi memberikan nilai penduga pengganti yang cukup kecil. Apabila *R(T)* dipilih sebagai penduga terbaik, maka cenderung akan dipilih pohon yang besar, sebab pohon yang semakin besar akan membuat nilai *R(T)* semakin kecil. *R(T)* atau *Resubstitution Estimate*/penduga pengganti merupakan proporsi amatan yang mengalami kesalahan pengklasifikasian.

Penduga pengganti ini sering digunakan apabila pengamatan yang ada tidak cukup besar. Pengamatan dalam L dibagi secara random menjadi V bagian yang saling lepas dengan ukuran kurang lebih sama besar untuk setiap kelas. Pohon $T^{(v)}$ dibentuk dari sampel *training* ke-*v* dengan *v*=1, 2, …, *V*. dimisalkan $d^{(v)}(x)$ adalah hasil pengklasifikasian, maka penduga sampel uji untuk $R(T_t^{(v)})$ adalah sebagai berikut.

$$R\left(T_t^{(v)}\right) = \frac{1}{N_2} \sum_{(x_n, j_n) \in L_v}^{N} X(d^{(v)}(x_n) \neq j_n)$$

Dengan $N_v \cong {}^{N}/_{V}$ adalah jumlah pengamatan dalam L_v.

Selanjutnya dilakukan prosedur yang sama dengan menggunakan semua pengamatan dalam L untuk membentuk deret pohon T_t. Penduga cross validation v-fold untuk $T_t^{(v)}$ adalah.

$$R^{cv}(T_t) = \frac{1}{V} \sum_{v=1_v}^{V} R^{cv}(T_t^{(v)})$$

Pohon klasifikasi yang optimum dipilih T* dengan $R^{cv}(T^*) = \min_t R^{cv}(T_t)$

2. Ukuran *Ketepatan* Klasifikasi

Ketepatan klasifikasi digunakan untuk mengetahui apakah data diklasifikasikan dengan benar atau tidak (Agresti, 2002). Ketepatan klasifikasi merupakan suatu evaluasi untuk melihat peluang kesalahan klasifikasi yang dilakukan oleh suatu fungsi klasifikasi. Beberapa cara yang umum digunakan untuk mengukur ketepatan klasifikasi adalah melalui perhitungan *Apparent Error Rate* (APER), *sensitivity*, *specificity*, *total accuracy rate* (1-APER). APER merupakan proporsi observasi yang diprediksi secara tidak benar (ukuran kesalahan klasifikasi total). *Sensitivity* merupakan ukuran ketepatan dari kejadian yang diinginkan.*Spesificity* merupakan ukuran yang menyatakan persentase kejadian yang tidak diinginkan.*Total accuracy rate*

merupakan proporsi observasi yang diprediksi secara benar (ukuran kesalahan klasifikasi total) (Johnson dan Wichern, 1992). Tabel untuk menghitung ketepatan klasifikasi ditujukkan pada tabel 2.

Tabel 2. Crosstab ketepatan Klasifikasi

Observasi Y	Prediksi Y		Total
	1	2	
1	n_{11}	n_{12}	$N_{1.}$
2	n_{21}	n_{22}	$N_{2.}$
Total	$N_{.1}$	$N_{.2}$	N

Keterangan:

n_{11} : Jumlah subjek dari variabel Y kategori 1 yang tepat diprediksikan sebagai variabel Y kategori 1

n_{12} : Jumlah subjek dari variabel Y kategori 1 yang salah diprediksikan sebagai variabel Y kategori 2

n_{21} : Jumlah subjek dari variabel Y kategori 2 yang salah diprediksikan sebagai variabel Y kategori 1

n_{22} : Jumlah subjek dari variabel Y kategori 2 yang tepat diprediksikan sebagai variabel Y kategori 2

$N_{1.}$: Jumlah observasi dari variabel Y kategori 1

$N_{2.}$: Jumlah observasi dari variabel Y kategori 2

$N_{.1}$: Jumlah prediksi dari variabel Y kategori 1

$N_{.2}$: Jumlah prediksi dari variabel Y kategori 2

N : Jumlah total observasi/prediksi

Rumus untuk menghitung ketepatan klasifikasi antara lain

$$APER = \frac{jumlahprediksisalah}{jumlahtotalprediksi} = \frac{n_{12} + n_{21}}{N}$$

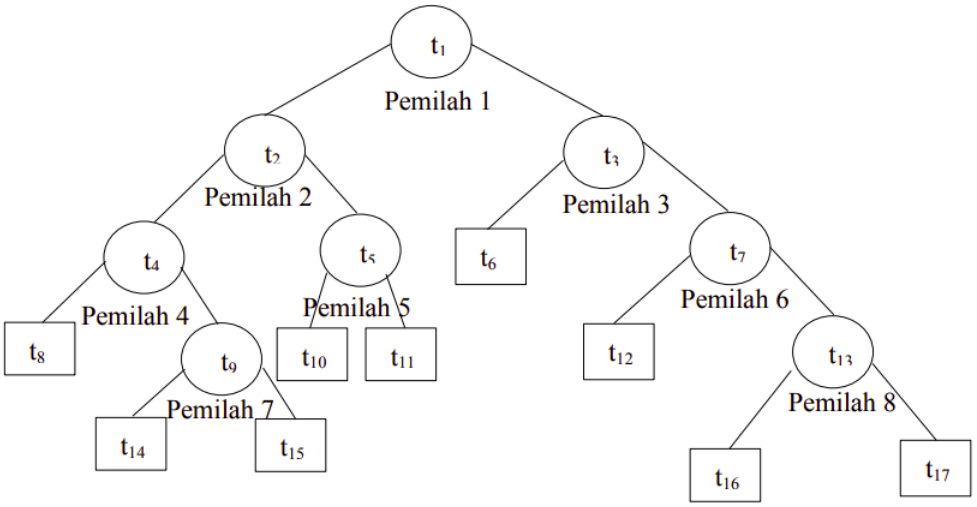
$$Sensitivity = \frac{n_{11}}{N_{1.}}$$

$$Specificity = \frac{n_{22}}{N_{2.}}$$

$$Totalaccuracyrate = 1 - APER = \frac{jumlahprediksisalah}{jumlahtotalprediksi} = \frac{n_{12} + n_{21}}{N}$$

3. Struktur atau Bentuk Pohon Klasifikasi CART

Proses analisis dalam CART digambarkan dalam bentuk atau struktur yang menyerupai sebuah pohon, yaitu pohon klasifikasi berbentuk biner. Biner dalam pohon klasifikasi ini berarti setiap pemecahan parent node menghasilkan 2 child nodes dimana dapat berupa simpul dalam dan simpul akhir. Simpul awal yang disebut parent node dinotasikan dengan t_1, simpul dalam (*internal nodes*) dinotasikan dengan t_2, t_3, t_4, t_7, t_9, dan t_{10}, serta simpul akhir (*terminal node*) yang dinotasikan dengan $t_5, t_6, t_8, t_{11}, t_{12}, t_{13}, t_{14}$ dan t_{15} dimana setelahnya tidak ada lagi pemilahan. Penghitungan *depth* (kedalaman) pohon dimulai dari simpul utama t_1 yang berada pada kedalaman 1, sedangkan t_2 dan t_3 berada pada kedalaman 2 begitu seterusnya hingga t_{12}, t_{13}, t_{14} dan t_{15} yang berada pada kedalaman 5. Selain itu, setiap simpul terminal diberi tanda dengan label kelas. Pohon klasifikasi CART dapat dilihat pada gambar berikut ini:



Sumber: Breiman L., Friedman J.H., Olshen R.A., & Stone C.J. (1993). Clasification and Regression Trees (CART)

Gambar 13. Struktur Pohon Klasifikasi

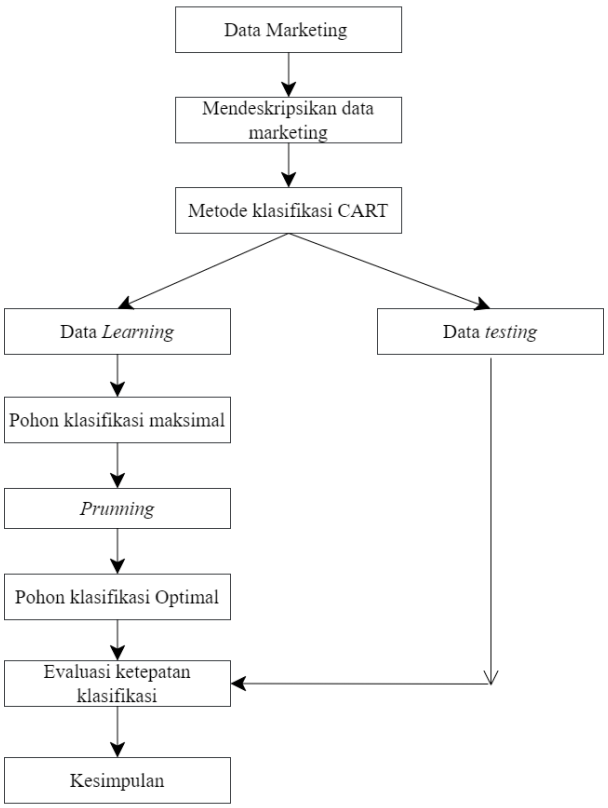
Keterangan:

a. Root node digambarkan dengan lingkaran merupakan nonterminal node paling awal tempat learning sample yang dimiliki.
b. Branch digambarkan dengan 2 garis lurus merupakan cabang dari root node. Branch merupakan tempat pemecahan masing-masing nonterminal node.
c. Nonterminal nodes digambarkan dengan lingkaran merupakan subset atau himpunan bagian dari nonterminal nodediatasnya yang memenuhi criteria pemecahan tertentu.

d. Terminal Nodes, digambarkan dengan persegi merupakan node tempat memprediksi kan sebuah objek pada kelas tertentu (class labeled).

4. DIAGRAM ALUR CART

Berikut diagram alur metode CART yang ditampilkan pada Gambar 14.



Gambar 14. Diagram Alur CART

5. Hasil dan Pembahasan

Variabel yang digunakan sebanyak 12 variabel yaitu *Education, Recency, MntWines, MntFruits, MntMeatProducts, MntFishProducts, MntSweetProducts, MntGoldProds, Response, Income, Age_Demographic, TotalChildren.*Langkah pertama dalam analisis metode CART, kita akan membagi dataset marketing menjadi data training dan data test. Sebelum itu, kita akan mengacak dataset agar saat pemisahaan data menjadi data training dan data test tidak bersifat homogen. Kita dapat menggunakan fungsi sample() pada Software RStudio.

```
#shuffle
shuffle_index <- sample(1:nrow(marketing))
head(shuffle_index)
marketing <- marketing[shuffle_index, ]
head(marketing)
```

Kemudian kita dapat membuat data training dan data test. Data training ini digunakan untuk membentuk model, sedangkan data test digunakan untuk memprediksi model yang telah diperoleh. Pada umumnya, rasio pemisahan data yang digunakan adalah 80/20 dimana 80% dataset berada pada data training, dan 20% sisanya berada pada data test. Kita dapat menggunakan fungsi create_train_test() dengan rumus syntax berikut

```
create_train_test <- function(data, size = 0.8, train = TRUE) {
  n_row = nrow(data)
  total_row = size * n_row
  train_sample <- 1: total_row
  if (train == TRUE) {
    return (data[train_sample, ])
  } else {
    return (data[-train_sample, ])
  }
}
```

Sehingga kita dapat melanjutkan pembuatan data training dan data test menggunakan fungsi diatas

```
data_train <- create_train_test(marketing, size = 0.8, train = TRUE)
data_test <- create_train_test(marketing, size = 0.8, train = FALSE)
```

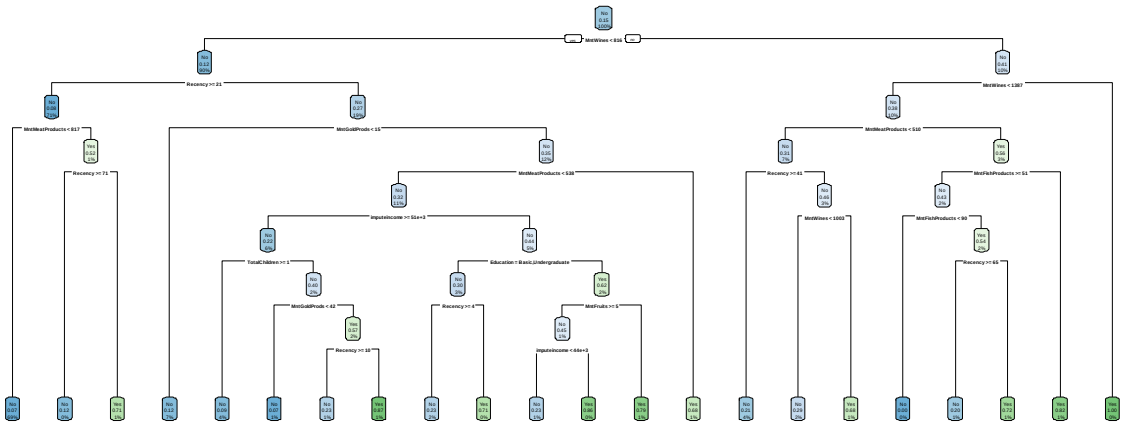
Kita lihat berapa observasi yang ada pada data training dan data test.

```
> dim(data_train)
[1] 1788   12
> dim(data_test)
[1] 448  12
```

Untuk data training kita memakai sebanyak 1788 observasi, sedangkan untuk data test sebanyak 448 observasi. Selanjutnya membentuk model kta dapat menggunakan package rpart.plot dan rpart.

```
#Build Model
library(rpart)
library(rpart.plot)
fit <- rpart(Response~., data = data_train, method = 'class')
rpart.plot(fit, extra = 106)
```

Syntax rpart() berfungsi untuk membentuk model. Syntax rpart.plot berfungsi untuk membentuk plot tree. Pada fungsi rpart() menggunakan indeks Gini. Plot yang dihasilkan ditampilkan pada Gambar 15.



Gambar 15. Pohon Klasifikasi

Berdasarkan Gambar 15 menunjukkan bahwa *terminal nodes* yang dihasilkan pada pohon klasifikasi maksimal sebanyak 3 *terminal nodes*. Berdasarkan perhitungan *Indeks Gini* peubah terbaik yang menjadi pemilah utama, yaitu peubah *MntWines*, dan melibatkan 9 peubah penjelas, yaitu Recency, Education, MntWines, MntFruits, MntMeatProducts, MntGoldProds, imputeincome, TotalChildren.

Pada simpul 1 probabilitas untuk seluruh pelanggan yang menerima penawaran kampanye sebesar 15 persen. Pada simpul 2 sebesar 71 persen proporsi pelanggan yang menghabiskan wine lebih sedikit dari 816 berkemungkinan menerima penawaran kampanye sebesar 8 persen. Selanjutnya pada simpul ke 3 pelanggan yang menghabiskan daging lebih kecil dari 817, berkemungkinan menerima penawaran sebesar 7 persen. Karakteristik pelanggan yang berpeluang untuk menerima tawaran kampanye terakhir salah satunya adalah pelanggan yang mengkonsumsi wine lebih dari 1387 barang dalam 2 tahun terakhir.

Kita dapat memprediksi menggunakan syntax *predict()* pada data test. Kita ingin memprediksi pelanggan mana yang lebih menerima tawaran kampanye terakhir pada data test.

```
predict_unseen <-predict(fit, data_test, type = 'class')
table_mat <- table(data_test$Response, predict_unseen)
table_mat
```

Output ditampilkan pada Tabel 3.

Tabel 3. Prediksi Model

	No	Yes
No	355	21
Yes	58	14

Berdasarkan Tabel 3 menunjukkan bahwa model memprediksi dengan benar bahwa 335 pelanggan tidak menerima tawaran kampanye terakhir namun mengklasifikasikan pelanggan yang menerima tawaran sebagai tidak menerima tawaran kampanye terakhir.

Kita dapat menghitung ukuran akurasi untuk mengklasifikasi menggunakan *confusion matrix*. *Confusion matrix* merupakan suatu pilihan yang lebih baik untuk mengevaluasi kinerja model. Sehingga dari *confusion matrix* ini, kita dapat menghitung akurasi model.

```
accuracy_Test <- sum(diag(table_mat)) / sum(table_mat)
print(paste('Accuracy for test', accuracy_Test))
```

Sehingga diperoleh hasil akurasi untuk memprediksi karakteristik pelanggan terhadap kampanye terakhir yang telah dilaksanakan perusahaan sebesar 82,37%.

D. Kesimpulan

Pohon klasifikasi terbaik menghasilkan sebanyak 3 simpul terminal dengan melibatkan 9 peubah penjelas, yaitu Recency, Education, MntWines, MntFruits, MntMeatProducts, MntGoldProds, imputeincome, TotalChildren. Probabilitas untuk seluruh pelanggan yang menerima penawaran kampanye sebesar 15 persen. Sebesar 71 persen proporsi pelanggan yang menghabiskan wine lebih sedikit dari 816 berkemungkinan menerima penawaran kampanye sebesar 8 persen. Pelanggan yang menghabiskan daging lebih kecil dari 817, berkemungkinan menerima penawaran sebesar 7 persen. Karakteristik umum pelanggan terhadap kampanye terakhir yang telah dilaksanakan perusahaan memiliki akurasi sebesar 82,37% pelanggan tidak menerima tawaran kampanye terakhir namun mengklasifikasikan pelanggan yang menerima tawaran sebagai tidak menerima tawaran kampanye terakhir.

6. DAFTARPUSTAKA

Agresti, A. (2002). *Categorical Data Analysis*. John Wiley and Sons, Inc., New York

Breiman L., Friedman J.H., Olshen R.A., & Stone C.J. (1993). *Classification And Regression Trees*. New York: Chapman And Hall

Du Toit SHC, Steyn AGW & Stumph RH.(1986). *Graphical Exploratory Data Analysis*. New York : Springer-Verlag.

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining Concepts and Techniques Third Edition.* USA: Elsevier Inc.

Johnson, Richard A and Wichern, Dean W. (1992).*Applied Multivariate Statistical Analysis*. America: Pearson International Edition.

Lewis, M.D dan Roger, J. (2000).*An Introduction to Classification and Regression Tree (CART) Analysis.*Presented at the 2000 Anual Meeting of Society For Academy Emergency Medicine in San Fransisco, California [Online]. Avaible: diakses tanggal 28 September 2013.

Zheng, Y. (2020). Decision Tree Algorithm for Precision Marketing via Network Channel. *International Journal of Computer Systems Science & Engineering*, 35(4): 293-298.