e-ISBN 978-623-93343-3-8
Seminar Nasional Mahasiswa Ilmu Komputer dan Aplikasinya (SENAMIKA)
Jakarta-Indonesia, 20 April 2021

Klasifikasi Data Penjualan pada Supermarket dengan Metode *Decision Tree*

Alleyda Irzky Shafarindu[1], Endah Patimah[2], Yohanne Marintan Siahaan[3], Andhika Wisnu Wardhana[4], Ballya Vicky Haekal[5,] Desta Sandya Prasvita, S.Komp., M.Kom.[6]
[1,2,3,4,5,6]Program Studi Informatika, Universitas Pembangunan Nasional Veteran Jakarta
[1,2,3,4,5,6]Jl. RS. Fatmawati Raya, Pd. Labu, Kec. Cilandak, Kota Depok, Jawa Barat 12450
email: [1]alleydais@upnvj.ac.id, [2]endahp@upnvj.ac.id,
[3]yohannems@upnvj.ac.id,[4]andhikawisnu6@gmail.com,[5]vballya12@gmail.com,[6]desta.sandya@gmail.com

Abstrak. Supermarket merupakan tempat pembelanjaan yang menyediakan berbagai kebutuhan sehari-hari. Banyak pembeli yang datang ke supermarket untuk membeli keperluannya. Pertumbuhan supermarket semakin meningkat dan memiliki kompetisi pasar yang tinggi. Supermarket memiliki berbagai macam produk yang berbeda merek, berbagai cabang dan berbagai tipe pelanggan. Penelitian ini bertujuan mencari indeks penjualan dari suatu dataset berisikan data supermarket dengan berbagai informasi terkait supermarket yang ada pada dataset. Pada penelitian ini dilakukan pengembangan dari penelitian terdahulu yang menggunakan objek penelitian yang berbeda serta dengan memodifikasi algoritma yang digunakan.Metode klasifikasi yang digunakan adalah *Decision Tree* dengan algoritma C4.5 dengan normalisasi *MinMax* dan pembagian data dengan metode *K-Fold Cross Validation* dan *Holdout Validation*. Dataset yang digunakan merupakan data dari penjualan pada supermarket. Hasil pengujian dengan metode pembagian data *Hold Out* nilai akurasi regresi memiliki nilai 0.5014285714285714 dan akurasi *decision tree* memiliki nilai 0.5. Sedangkan dengan metode pembagian data *K-Fold* memiliki nilai akurasi regresi memiliki nilai 0.466 dan akurasi *decision tree* memiliki nilai 0.492. Akurasi tertinggi pada kelas *branch* adalah 1.0 dan pada kelas *customer type* ada pada akurasi *decision tree* dengan metode *Hold Out* yang memiliki nilai 0.5. Penelitian ini diharapkan dapat bermanfaat bagi para peneliti untuk kemudian dapat dikembangkan sesuai objek penelitiannya masing-masing.

Kata kunci: Supermarket, Penjualan, Klasifikasi, *Decisision Tree, K-Fold Cross Validation, Holdout Validation.*

I. Pendahuluan

Supermarket merupakan tempat pembelanjaan yang menyediakan berbagai kebutuhan sehari-hari. Banyak pembeli yang datang ke supermarket untuk membeli keperluannya. Pertumbuhan supermarket semakin meningkat dan memiliki kompetisi pasar yang tinggi. Supermarket memiliki berbagai macam produk yang berbeda merek, berbagai cabang dan berbagai tipe pelanggan. Klasifikasi merupakan pembagian atau pengelompokan suatu data. Pada klasifikasi data dikelompokkan menurut kelasnya untuk melakukan prediksi dari suatu objek. Ada berbagai macam Teknik klasifikasi, salah satunya yang sering dipakai adalah *Decision tree.*

Proses klasifikasi data banyak diperlukan untuk menunjang penelitian di berbagai bidang seperti halnya kedokteran, pertanian, bisnis dan terutama pada bidang informatika. Setiap dataset yang diklasifikasikan dengan tujuan tertentu akan berguna baik bagi para penelitinya maupun bagi para pengguna hasil klasifikasinya. Dalam pengerjaannya, dibutuhkan keahlian dalam hal pemrograman serta analisis data yang baik untuk mendapatkan hasil yang terbaik.

II. Tinjauan Pustaka

A. Decision Tree

Decision Tree merupakan metode klasifikasi dan prediksi yang cukup umum digunakan. *Decision tree* adalah metode yang menggunakan struktur pohon untuk menentukan urutan keputusan dan konsekuensi. Tujuannya adalah untuk memprediksi variabel respon atau *output* Y [1]. Secara umum langkah yang diperlukan untuk membangun pohon keputusan dengan *Decision Tree* C4.5 adalah sebagai berikut:

1. Mencari nilai *Entropy* dari setiap kriteria yang ada.
2. Mencari nilai *Gain* dari setiap atribut.

e-ISBN 978-623-93343-3-8
Seminar Nasional Mahasiswa Ilmu Komputer dan Aplikasinya (SENAMIKA)
Jakarta-Indonesia, 20 April 2021

3. Memilih akar berdasarkan *gain* tertinggi.
4. Membentuk cabang berdasarkan masing-masing nilai.
5. Ulangi proses pada setiap cabang.

Entropy merupakan nilai ketidakmurnian (*impurity)* suatu atribut. Semakin kecil nilai *entropy* maka akan semakin baik digunakan untuk mengekstrak suatu kelas. Adapun rumus untuk mencari *entropy* adalah sebagai berikut.

$$Entropy(S) = -\sum_{i=1}^{n} pi * log_2 pi$$

Dimana:
S : himpunan kasus
n : jumlah partisi S
p_i : proporsi dari $S_i$terhadap S

Nilai *entropy* ini akan digunakan untuk menghitung nilai *gain*. Nilai *gain* digunakan untuk memisahkan obyek. Adapun rumus untuk mencari *gain* adalah sebagai berikut.

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

Dimana:

S : himpunan kasus
A : atribut
n : jumlah partisi atribut A
$|S_i|$: jumlah kasus pada partisi ke-1
$|S|$: jumlah kasus dalam S

Atribut yang mempunyai nilai *gain* tertinggi dari atribut lain akan menjadi akar (*node)* dari *decision tree*. *Node* tersebut mempunyai nilai *instance* yang kemudian akan menjadi cabang dari node tersebut. Nilai atribut memiliki nilai *instance* yang berbeda. Nilai *instance* ini kemudian diklasifikasikan lagi agar menjadi lebih sederhana. Setelah itu ulangi proses tersebut untuk setiap cabangnya untuk membentuk pohon keputusan.

B. Normalisasi

Normalisasi adalah salah satu praproses yang digunakan untuk membuat penskalaan nilai atribut dari data. Metode normalisasi yang digunakan pada penelitian ini adalah metode *min-max normalization.* Metode *min-max normalization* adalah metode dimana proses normalisasi dilakukan dengan melakukan transformasi linear pada data asli [2].

C. Pembagian Data

Pembagian data dilakukan untuk membagi dataset menjadi data *training* dan data *testing*. Pada penelitian ini pembagian data menggunakan *k-fold cross validation* dan *holdout validation*. Teknik yang digunakan dalam *k-fold cross validation*iini adalah melakukan pembagian dataset sebanyak K partisi secara acak, lalu akan dilakukan sebanyak k kali percobaan, d mana masing-masing percobaan menggunakan data partisi ke-k sebagai data *testing* dan sisanya sebagai data *training* [3]. Lalu, *holdout validation* adalah pembagian dataset dimana data akan dibagi menjadi data *testing* dan data *training*, misalnya 0.2 maka data tersebut 20% digunakan untuk data testing dan sisanya untuk data *training* [4].

D. Penelitian Terdahulu

Terdapat tiga penelitian terdahulu yang digunakan sebagai acuan untuk mengerjakan penelitian ini, penelitian pertama adalah penelitian dari Eka Pandu Cynthia dan Edi Ismanto dengan judul paper "Metode *Decision Tree* Algoritma C4.5 Dalam Mengklasifikasikan Data Penjualan Bisnis Gerai Makanan Saji." Adapun tujuan penelitian tersebut yaitu untuk mendapatkan informasi hasil klasifikasi data penjualan menu makanan yang digemari dan kurang digemari pelanggan pada bisnis gerai makanan cepat saji. Hasil dari klasifikasi tersebut nantinya dapat dijadikan patokan oleh gerai makanan cepat saji untuk menyediakan menu yang sesuai dengan kegememaran

pelanggannya, sehingga makanan yang dijualnya pun akan laris. Dalam klasifikasi diterapkan teknik data mining dengan tahapan algoritma C4.5 dan metode *decision tree*. Data yang digunakan dalam penelitian adalah data dari penjualan dari menu-menu pada bisnis gerai makanan cepat saji dalam satu bulan. Algoritma yang digunakan dalam klasifikasi ini adalah algoritma C4.5 [5].

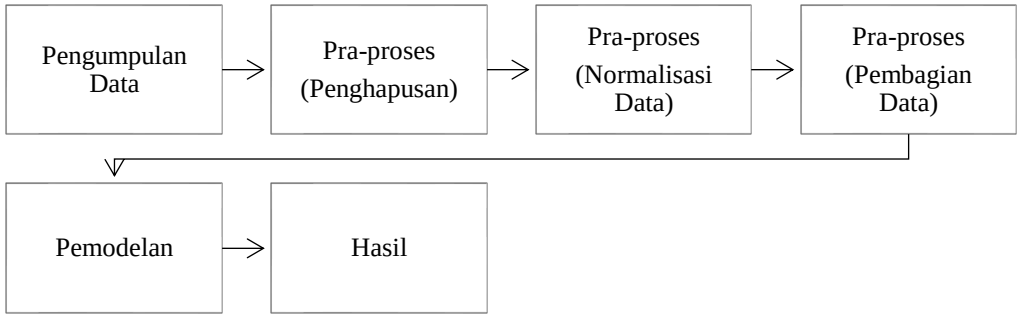
Penelitian kedua adalah penelitian dari Hananda Hafizan dan Anggita Nadia Putri dengan judul *paper* "Penerapan Metode Klasifikasi *Decision Tree* Pada Status Gizi Balita Di Kabupaten Simalungun." Sebagaimana yang dituliskan di judulnya, penelitian ini menggunakan metode *decision tree*. Data yang digunakan dalam penelitian merupakan data balita yang ada di Posyandu Kenanga, kelurahan Wonorejo, kecamatan Kerasaan, kabupaten Simalungun. Adapun dari penelitian tersebut didapatkan hasil bahwa metode klasifikasi C4.5 dapat diterapkan proses klasifikasi status gizi balita dengan bantuan *software* RapidMiner dan hasil perhitungan *software* RapidMiner menunjukkan akurasi menggunakan *tools performance* sebesar 100% [6].

Penelitian ketiga adalah penelitian dari Triuli Novianti dan Iwan Santosa dengan judul "Penentuan Jadwal Kerja Berdasarkan Klasifikasi Data Karyawan Menggunakan Metode *Decision Tree* C4.5 (Studi Kasus Universitas Muhammadiyah Surabaya)." Pada penelitian tersebut, data yang digunakan merupakan data induk Karyawan yang terdiri dari data proses masuk, umur, jenis kelamin serta unit kerja. Hasil dari penelitian tersebut menunjukkan bahwa hasil pengujian keseluruhan data dengan memakai *cross validation* 5 fold mendapat akurasi pengujian sebesar 70% [7].

III. Fokus Pengerjaan

Penelitian ini dilakukan dengan menggunakan metode Teknik klasifikasi yang sama dengan penelitian sebelumnya yaitu metode *decision tree*. Data yang digunakan pada penelitian ini merupakan data dari penjualan supermarket. Dalam penelitian kali ini fokus pengerjaannya adalah untuk mendapatkan informasi tingkat penjualan yang lebih baik dengan akurasi tertinggi dengan menggunakan dua macam tipe kelas yaitu kelas *branch* dan kelas *customer type* atau tipe pelanggan.

IV. Metode Penelitian



Gambar 1. Metode Penelitian

1. Pengumpulan Data
   Data yang digunakan dalam penelitian ini diperoleh dari website Kaggle, dengan link https://www.kaggle.com/aungpyaeap/supermarket-sales?select=supermarket_sales+-+Sheet1.csv.
2. Penghapusan
   Setelah data diperoleh, data akan diolah dengan langkah pertama yaitu menghapus *missing value* yang terdapat pada data. Lalu, menghapus beberapa variabel yang tidak digunakan pada data.
3. Normalisasi Data
   Setelah melakukan penghapusan missing value, lalu langkah selanjutnya adalah melakukan normalisasi dengan menggunakan *min-max*.
4. Pembagian Data
   Setelah melakukan normalisasi data, maka langkah selanjutnya adalah pembagian data dengan menggunakan *k-fold cross validation* atau *holdout validation*.

5. Pemodelan
   Setelah pembagian data, selanjutnya adalah melakukan pemodelan dengan menggunakan *decision tree*.
6. Hasil
   Setelah melakukan pemodelan data, kita akan mengetahui akurasi dari data.

V. Hasil dan Pembahasan

A. Data yang digunakan

Data yang digunakan diambil dari https://www.kaggle.com/aungpyaeap/supermarket-sales?select=supermarket_sales+-+Sheet1.csv. Data tersebut kemudian ditransformasi dengan menghapus beberapa atribut yaitu atribut id, time, dan date. Sehigga, data tersebut memiliki 14 atribut sebagai berikut.

1. *Branch*: Cabang supercenter (3 cabang tersedia diidentifikasi oleh A, B dan C).
2. *City*: Lokasi supercenter
3. *Customer type*: Tipe pelanggan, direkam dari member untuk pelanggan member menggunakan kartu member dan pelanggan normal untuk yang tidak mempunyai kartu member.
4. *Gender*: Jenis kelamin member.
5. *Product line*: Produk umum yang dikategorikan dengan grup, yaitu *Electronic accessories, Fashion accessories, Food and beverages, Health and beauty, Home and lifestyle, Sports* dan *travel*
6. *Unit price*: Harga dari setiap produk dalam $
7. *Quantity*: Jumlah produk yang dibeli oleh pelanggan
8. *Tax*: 5% biaya pajak pada pembelian setiap pelanggan
9. Total: Total harga termasuk pajak
10. *Payment*: Pembayaran yang digunakan oleh pelanggan untuk pembelian (tersedia 3 metode yaitu Tunai, kartu kredit, dan *Ewallet*)
11. *COGS: Cost of goods sold* atau Harga pokok penjualan
12. *Gross margin percentage*: persentase margin kotor
13. *Gross income*: Penghasilan kotor
14. *Rating*: *Rating* stratifikasi pelanggan pada keseluruhan pengalaman berbelanja mereka (dalam skala 1 hingga 10)

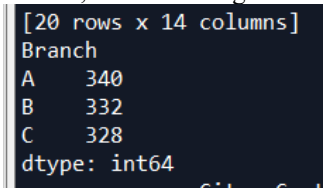
B. Metodologi yang digunakan

Penelitian ini menggunakan metode klasifikasi *decision tree* sesuai dengan penelitian sebelumnya. Lalu *preprocessing* data dengan mengubah data string ke angka. Serta akan dilakukan pembuangan pada data yang duplikat dan menghapus variable yang tidak digunakan pada data. Setelah dilakukan *preprocessing* data, akan dilakukan normalisasi mengunakan *min-max*. Pembagian data *testing* dan *training* dilakukan dengan memakai *Hold Out Estimation* dan *K-Fold Cross Validation*. Penggunaan algoritma C4.5 untuk melakukan pemodelan menggunakan *decision tree* untuk melakukan klasifikasi.

C.Hasil Penelitian

Hasil akan dibagi dua berdasarkan kelas yang digunakan.

1. Berdasarkan *Branch*
   Dari dataset supermarket, berdasarkan *Branch* didapatkan sebanyak 340 orang berbelanja pada *Branch* A, dan 332 orang berbelanja pada *Branch* B, dan 328 orang berberlanja pada *Branch* C.



Gambar 2. Perhitungan jumlah total pelanggan *Branch* A, B, C pada *Spyder*

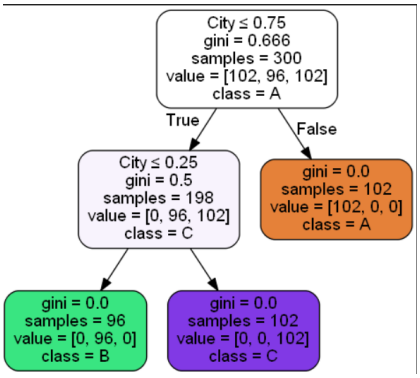
a. Hasil menggunakan *Hold Out Estimation*

Dari hasil percobaan klasifikasi dari *dataset* supermarket dengan pembagian data menggunakan *Hold Out Estimation* dengan data test 70% dan data training 30%, didapatkan akurasi Regresi dan *Decision Tree* sebesar :

```
[8 rows x 13 columns]
akurasi regresi=  1.0
akurasi Decision Tree=  1.0
```

Gambar 3. Akurasi Regresi dan *Decision Tree* yang dihasilkan *Spyder*

Menghasilkan pohon keputusan sebagai berikut :



Gambar 4. Pohon Keputusan yang dihasilkan *Spyder*

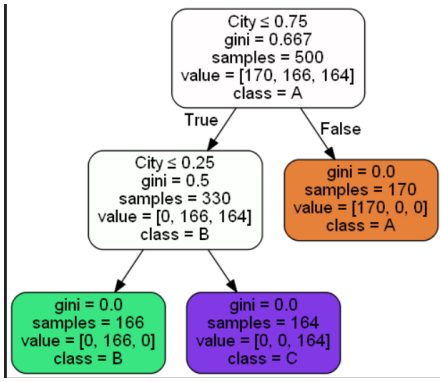
b. Hasil menggunakan *K-Fold Cross Validation*

Dari hasil percobaan klasifikasi dari dataset supermarket dengan pembagian data menggunakan 2 Fold-Cross Validation, didapatkan akurasi Regresi dan *Decision Tree* sebesar :

```
[8 rows x 13 columns]
akurasi regresi=  1.0
akurasi Decision Tree=  1.0
```

Gambar 5. Akurasi Regresi dan *Decision Tree* yang dihasilkan *Spyder*

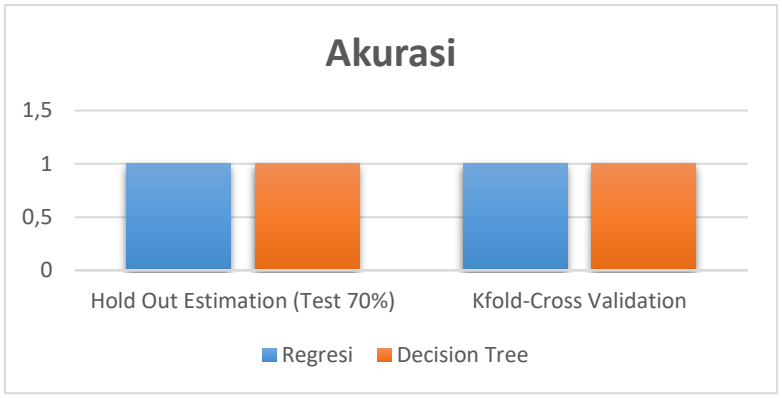
Menghasilkan pohon keputusan sebagai berikut :



Gambar 6. Pohon Keputusan yang dihasilkan *Spyder*

c. Perbandingan hasil *Hold Out Estimation* (70% data test) dan *K-Fold Cross Validation*

Berikut merupakan grafik perbandingan hasil dari kedua pembagian data diatas :



Gambar 7. Perbandingan hasil antara regresi dan *decision tree*

2. Berdasarkan *Customer type*
Dari dataset supermarket, berdasarkan *Customer type* didapatkan sebanyak 501 orang sebagai Member, dan 499 orang sebagai Normal (Bukan Member).

```
[20 rows x 14 columns]
Customertype
Member    501
Normal    499
dtype: int64
```

Gambar 8. Perhitungan jumlah total Member dan Normal pada *Spyder*

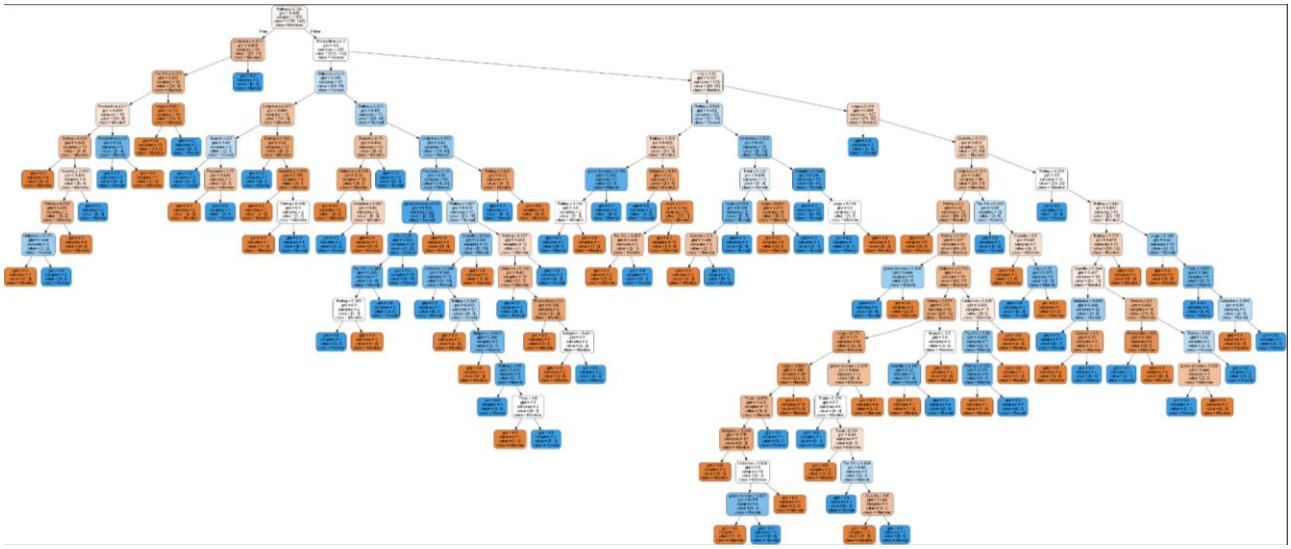
a. Hasil menggunakan *Hold Out Estimation*

Dari hasil percobaan klasifikasi dari *dataset* supermarket dengan pembagian data menggunakan *Hold Out Estimation* dengan *data test* 70% dan *data training* 30%, didapatkan akurasi Regresi dan *Decision Tree* sebesar :

```
[8 rows x 13 columns]
akurasi regresi=  0.5014285714285714
akurasi Decision Tree=  0.5
```

Gambar 9. Akurasi Regresi dan *Decision Tree* yang dihasilkan *Spyder*

Menghasilkan pohon keputusan sebagai berikut :



Gambar 10. Pohon Keputusan yang dihasilkan *Spyder*

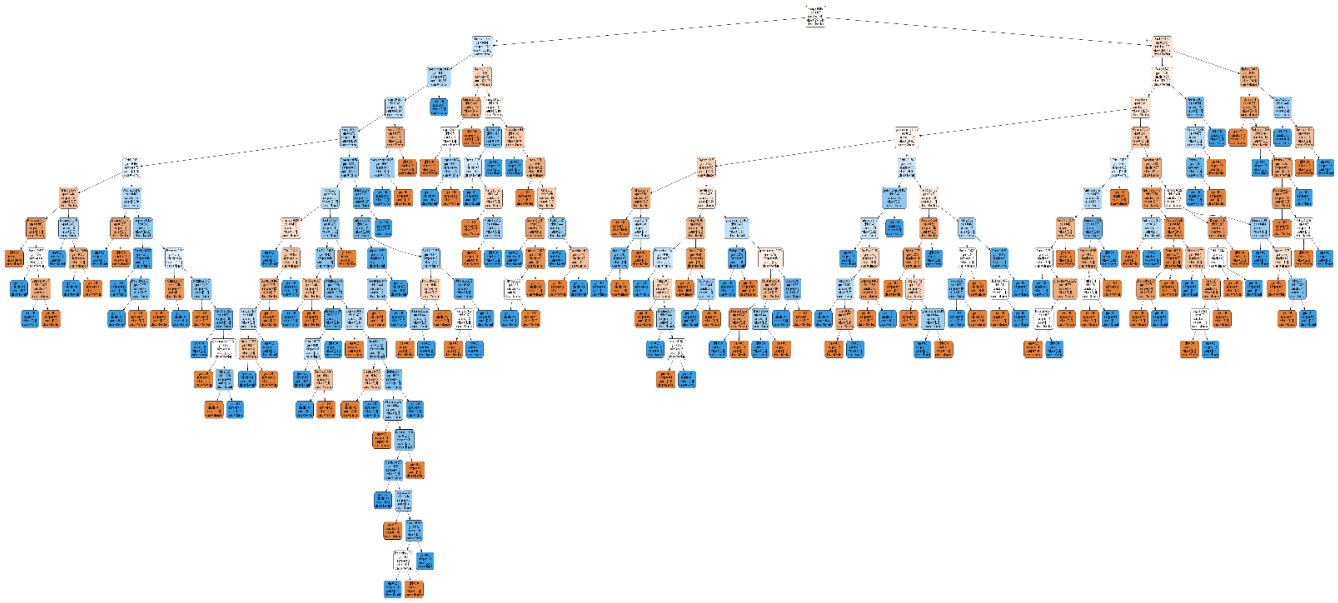
b. Hasil menggunakan *K-Fold Cross Validation*

Dari hasil percobaan klasifikasi dari *dataset* supermarket dengan pembagian data menggunakan 2 Fold-Cross Validation, didapatkan akurasi Regresi dan *Decision Tree* sebesar :

```
akurasi regresi=  0.466
akurasi Decision Tree=  0.492
```

Gambar 11. Akurasi yang dihasilkan *Spyder*

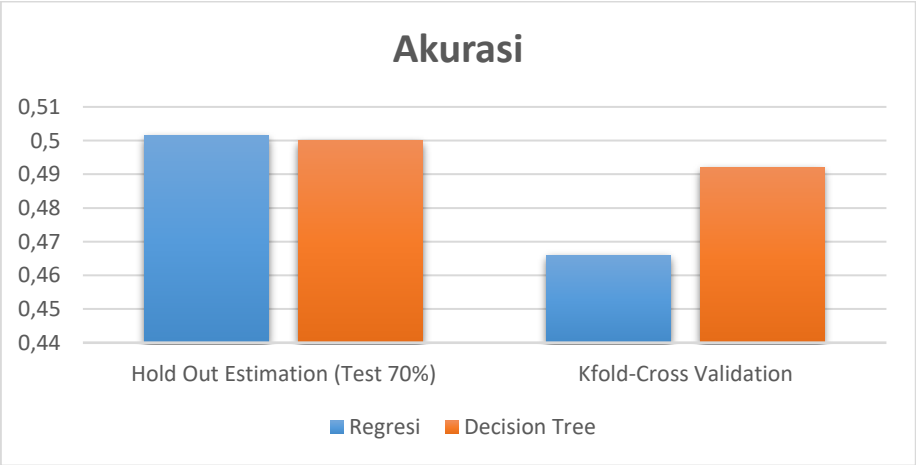
Menghasilkan pohon keputusan sebagai berikut :



Gambar 12. Pohon Keputusan yang dihasilkan *Spyder*

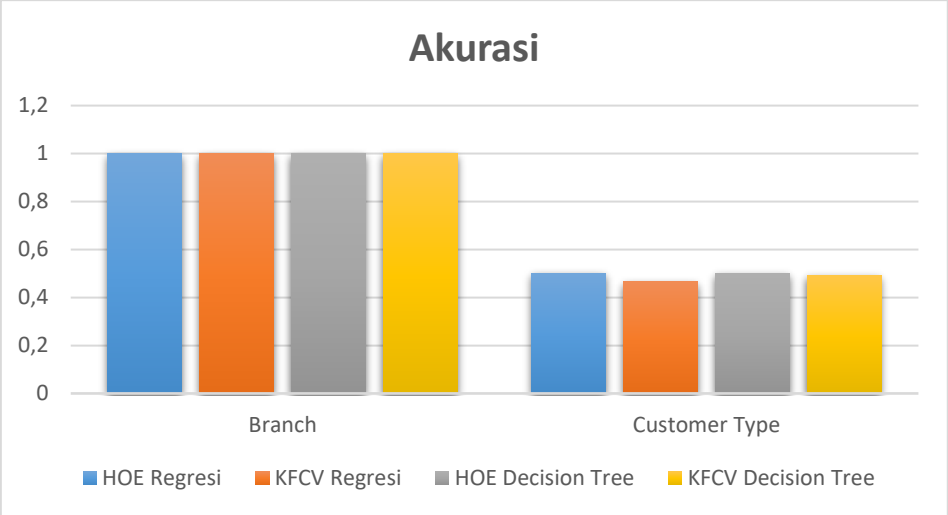
c. Perbandingan hasil *Hold Out Estimation*(70% data test) dan *K-Fold Cross Validation*

Berikut merupakan grafik perbandingan hasil dari kedua pembagian data diatas :



Gambar 13. Hasil perbandingan metode regresi dan *decision tree*

Perbandingan Akurasi Kelas *Branch* dan *Customer Type*



Gambar 14. Hasil perbandingan akurasi class *branch* dan *customer type*

Grafik diatas menunjukkan akurasi terbaik diperoleh dengan menggunakan Class *Branch* dengan *average* akurasi sebesar 1 atau 100%.

VI. Kesimpulan dan Saran

Hasil dari penelitian yang dilakukan akurasi regresi dan *decision tree* pada kelas *branch* memiliki nilai yang sama meskipun dengan metode pembagian data yang berbeda. Dan pada kelas *customer type* nilai akurasi regresi dan *decision tree* berbeda pada kedua metode pembagian data. Dengan metode pembagian data Hold Out nilai akurasi regresi memiliki nilai 0.5014285714285714 dan akurasi *decision tree* memiliki nilai 0.5. Sedangkan dengan metode pembagian data K-Fold memiliki nilai akurasi regresi memiliki nilai 0.466 dan akurasi *decision tree* memiliki nilai 0.492. Akurasi tertinggi pada kelas *branch* adalah 1.0 dan pada kelas *customer type* ada pada akurasi *decision tree* dengan metode *Hold Out* yang memiliki nilai 0.5

Saran masukan yang mungkin dapat diterapkan pada penelitian ini adalah dengan melakukan klasifikasi dengan metode *feed forward backpropagation* untuk klasifikasi selanjutnya.

VII. Daftar Pustaka

[1] Oktafianto. 2016. "*Analisis Kepuasan Mahasiswa Terhadap Pelayanan Akademik Menggunakan Metode Algoritma C4.5 (Studi Kasus: Stmik Pringsewu*)", dalam jurnal: TIM Darmajaya Vol. 02 No. 01.

[2] Nasution, Darnisa Azzahra, et, al. "*Perbandingan Normalisasi Data Untuk Klasifikasi Wine Menggunakan Algoritma KNN*" dalam Journal of Computer Engineering System and Science: Vol. 4, No 1. Jakarta: Universitas Pembangunan Nasional Veteran Jakarta.

[3] Raju, K Srujan, et. al. 2018. "*Support Vector Machine with K-Fold Cross Validation Model for Software Fault Prediction"* dalam International Journal of Pure and Applied Mathematics: Vol 118, No.20.

[4] Giarno, et, all. 2019. *"Suitable Proportion Sample of Holdout validation Validation for Spatial Rainfall Interpolation in Surrounding the Makassar Strait"* dalam Universitas Muhamadiyah Surajrta Online Journals: Vol 32, No. 2.

[5] Cynthia, Eka P , Ismanto, Edi. 2018. Metode *Decision Tree* Algoritma C.45 Dalam Mengklasifikasikan Data Penjualan Bisnis Gerai Makanan Saji. Jurnal Riset Sistem Informasi Dan Teknik Informatika (JURASIK).Volume 3 :1-12.

[6] Hafizan, Hananda., Putri, Anggita Nadia. 2020. Penerapan Metode Klasifikasi *Decision Tree* Pada Status Gizi Balita Di Kabupaten Simalungun. KESATRIA: Jurnal Penerapan Sistem Informasi (Komputer & Manajemen) Vol. 1, No. 2, April (2020), pp. 68-72.

[7] Novianti, Triuli., Santosa, Iwan. 2016. Penentuan Jadwal Kerja Berdasarkan Klasifikasi Data Karyawan Menggunakan Metode *Decision Tree* C4.5 (Studi Kasus Universitas Muhammadiyah Surabaya). Jurnal Komunikasi, Media dan Informatika. Volume 5, No.1.