TUGAS PROYEK

PENGANTAR DATA MINING

"Klasifikasi *House Price* Data Menggunakan *Decision Tree* dan Aturan Asosiasi Pada Data Penjualan *Grocery Store*"

OLEH

KELOMPOK 1 "DATA CLEANING"

FADLYANSYAH NOER NASRUDDIN	H051171315
INTAN PERMATASARI	H051181328
SAPRIADI RASYID	H051191003

PROGRAM STUDI STATISTIKA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS HASANUDDIN

MAKASSAR

2022

KATA PENGANTAR

Pertama-tama marilah kita panjatkan puji syukur atas berkat rahmat Tuhan Yang Maha Kuasa, sehingga kami dapat menyelesaikan Tugas Proyek: Pengantar Data Mining. Kami juga berterima kasih kepada dosen kami, Sri Astuti Thamrin, S.Si., M.Stat., Ph.D. yang telah membimbing sehingga kami dapat menyelesaikan Tugas Proyek: Pengantar Data Mining sebagai salah satu tuntutan dalam mata kuliah Pengantar Data Mining dengan baik.

Tugas Proyek: Pengantar Data Mining telah kami susun dengan maksimal dan telah mendapatkan bantuan dari berbagai pihak sehingga dapat memperlancar pembuatan Tugas Proyek: Pengantar Data Mining. Untuk itu kami menyampaikan banyak terima kasih kepada semua pihak yang telah berkontribusi dalam Tugas Proyek: Pengantar Data Mining. Terlepas dari semua itu, kami menyadari sepenuhnya bahwa masih ada kekurangan baik dari segi susunan kalimat maupun tata bahasanya.

Oleh karena itu dengan tangan terbuka kami menerima segala saran dan kritik dari pembaca agar kami dapat memperbaiki Tugas Proyek: Pengantar Data Mining ini menjadi jauh lebih baik. Akhir kata kami berharap semoga Tugas Proyek: Pengantar Data Mining ini dapat memberikan manfaat maupun inspirasi terhadap pembaca.

Makassar, 6 April 2022

KELOMPOK 1

DAFTAR ISI

KATA PENGANTAR i
DAFTAR ISI ii
DAFTAR GAMBAR iii
DAFTAR TABEL iv
BAB 1 PENDAHULUAN 1
1.1 Latar Belakang 1
1.2 Rumusan Masalah 2
1.3 Tujuan Penelitian 2
BAB 2 TINJAUAN PUSTAKA 3
2.1 Pengantar Data Mining 3
2.2 Dataset 4
2.3 *Software* R 4
2.4 *Decision Tree* 4
2.5 Evaluasi Kinerja Pengklasifikasian 5
2.6 *Assosiasion Rules* 5
BAB 3 METODE PENELITIAN 6
3.1 Jenis, Waktu dan Tempat Pelaksanaan 6
3.2 Variabel dan Instrument 6
3.3 Tahapan Rancangan Percobaan 6
3.4 Data dan Sumber Data 6
BAB 4 ANALISIS DATA 7
4.1 Deskripsi Data 7
4.2 *Pre-Processing* Data 8
4.3 Klasifikasi Data *House Price* dengan *Decision Tree* 9
4.4 *Assosiasion Rules* 11
BAB 5 PENUTUP 12
5.1 Kesimpulan 12
5.2 Saran 12
DAFTAR PUSTAKA 13
LAMPIRAN-LAMPIRAN .....
Lampiran 1. Tampilan Kaggle House Prediction di csv 14
Lampiran 2. Tampilan Kaggle Grocery Store di csv 15
Lampiran 3. *Pre-Procesing* 16
Lampiran 4. Pemodelan Klasifikasi Harga Rumah dengan *Decision Tree* 17
Lampiran 5. Model aturan Asosiasi pada Data *Grocery Store* 18
Lampiran 6. Upload Data Github 19

DAFTAR GAMBAR

Gambar 1 Tabel 5 data teratas dari *House Price* 7
Gambar 2. Plot Rumah terhadap Harga ($) 7
Gambar 3. Data dari *Grocery Store* 8
Gambar 4. Tabel hasil dari *per-processing* 8
Gambar 5. Model *Decision Tree* Klasifikasi Harga Rumah 9
Gambar 6. Scatterplot dari 78 Aturan asosiasi yang terbentuk.....11

DAFTAR TABEL

Tabel 4.1 Matriks Kelas Prediksi dan Kelas Aktual 9
Tabel 4.2 Aturan Asosiasi dari Perbelanjaan Pelanggan di *Grocery Store* 11

BAB 1

PENDAHULUAN

1.1 Latar Belakang

Dalam kehidupan sehari-hari, manusia tak akan lepas dari data. Bahkan, seiring berjalannya waktu, maka semakin banyak pula data yang dapat dihasilkan manusia. Salah satunya yaitu, struk belanja.Tanpa disadari, data belanja yang dilakukan itu terekam oleh sistem. Pramana, dkk. (2018) menyatakan bahwa perkembangan teknologi digital yang cukup pesat meningkatkan volume dan kecepatan data yang dihasilkan dari berbagai internet melalui smartphones dan GPS (internet of things).

Big data menjadi sebuah tantangan bagi ilmuwan data (data scientist) untuk mampu menyediakan informasi, ringkasan, visualisasi, prediks dan pola dari data tersebut. Kehadiran "Tsunami Data" membuat kita tenggelam dalam data namun masih sangat kurang akan informasi yang diperoleh (Pramana, dkk., 2018). Dengan demikian, diperlukan proses pengolahan data. Salah satu proses pengolahan data agar menjadi sebuah informasi dapat diterapkan pada data mining.

Data mining merupakan langkah analisis terhadap proses penemuan pengetahuan di dalam basis data atau *knowledge discovery in database* yang di singkat menjadi KDD (Suyanto, 2019). Data mining merupakan analisis yang cocok di gunakan untuk skala data yang besar. Data mining merupakan proses menggunakan teknik statistika, matematika, kecerdasan buatan dan *machine learning*. Bahkan, data mining merupakan kegiatan mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai database yang besar (Jollyta, dkk., 2020).

Data mining mengalami pertumbuhan yang luar biasa, sebanding dengan kecepatan peningkatan urutan data di era big data saat ini, dengan volume data sudah mencapai satu trilyun gigabyte (zettabyte). Pada dasarbya, data mining digunakan untuk enam fungsi yaitu: klasifikasi, klasterisasi, regresi, deteksi anomaly, pembelajaran aturan asosiasi, pemodelan kebergantungan dan perangkuman (Suyanto, 2019).

Berdasarkan uraian tersebut, maka data mining sangat diperlukan untuk dapat mengolah data menjadi sebuah informasi. Sehingga, sangat diperlukan *soft skill* data mining. Dengan demikian, kami membuat makalah untuk menambah wawasan mengenai data mining yang berjudul "Klasifikasi *House Price* Data Menggunakan *Decision Tree* dan Aturan Asosiasi Pada Data Penjualan *Grocery Store*".

1.2 Perumusan Masalah

Berdasarkan judul dan latar belakang tersebut, permasalahan yang akan dikaji dalam tugas proyek ini yaitu:

1. Bagaimana evaluasi kinerja pengklasifikasi *House Price* Data yang terdapat pada dataset Kaggle Menggunakan *Decision Tree*?
2. Bagaimana Aturan Asosiasi Pada Data Penjualan *Grocery Store* yang terdapat pada dataset Kaggle?

1.3 Tujuan

Berdasarkan judul dan latar belakang tersebut, permasalahan yang akan dikaji dalam tugas proyek ini yaitu:

1. Untuk mengevaluasi kinerja pengklasifikasi *House Price* Data yang terdapat pada dataset Kaggle Menggunakan *Decision Tree.*
2. Untuk mengetahui Aturan Asosiasi Pada Data Penjualan *Grocery Store* yang terdapat pada dataset Kaggle.

BAB 2

TINJAUAN PUSTAKA

2.1 Pengantar Data Mining

Menurut Gartner Group, data mining adalah proses menemukan hubungan baru yang mempunyai arti, pola dan kebiasaan dengan memilah-milah sebagian besardata yang disimpan dalam media penyimpanan dengan menggunakan teknologi pengenalan pola seperti teknik statistic dan matematika. *Data mining* merupakan gabungan dari beberapa disiplin ilmu yang menyatukan teknik dari pembelajaran mesin, pengenalan pola, statistik, *database*, dan visualisasi untuk penanganan permasalahan pengambilan informasi dari *database* yang besar.

Data mining menurut David Hand, Heikki Mannila, dan Padhraic Smyth dari MIT adalah analisa terhadap data (biasanya data yang berukuran besar) untuk menemukan hubungan yang jelas serta menyimpulkannya yang belum diketahui sebelumnya dengan cara terkini dipahami dan berguna bagi pemilik data tersebut. (Bramer, 2007) Data mining merupakan analisis yang cocok di gunakan untuk skala data yang besar. Data mining merupakan proses menggunakan teknik statistika, matematika, kecerdasan buatan dan *machine learning* (Jollyta, dkk., 2020).

Data mining adalah proses yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan *machine learning* untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai *database besar*. Data mining merupakan serangkaian proses untuk menggali nilai tambah dari suatu kumpulan data berupa pengetahuan yang selama ini tidak diketahui secara manual (Han, 2011).

Dari definisi-definisi yang telah disampaikan, hal penting yang terkait dengan *data mining*:

1. *Data mining* merupakan suatu proses otomatis terhadap data yang sudah ada.
2. Data yang akan diproses merupakan data yang sangat besar.
3. Tujuan *data mining* adalah mendapatkan hubungan atau pola yang mungkin memberikan indikasi yang bermanfaat.

Data mining bukan suatu bidang yang baru. Salah satu kesulitan untuk mendefinisikan *data mining* adalah kenyataan bahwa *data mining* mewarisi banyak aspek dan teknik dari bidang-bidang ilmu yang dulu sudah mapan terlebih dulu. *Data mining* memiliki akar yang panjang dari bidang ilmu yang berbeda seperti kecerdasan buatan *(artificial intelligent)*, *machine learning*, statistik, *database*, dan juga *information retrieval*. (Larose, 2005)

2.2 Dataset

Dataset atau himpunan data merupakan kumpulan objek dan atributnya. Nama lain dari objek yang sering digunakan diantaranya *record, point, vector, pattern, event*, *observation, case*, *sample*, *instance*, entitas. Objek digambarkan dengan sejumlah atribut yang menerangkan sifat atau karakteristik dari objek tersebut. Atribut juga sering disebut variabel, *field*, fitur, atau dimensi. Atribut adalah sifat/properti/karakteristik objek yang nilainya bisa bermacam-macam dari satu objek dengan objek lainnya, dari satu waktu ke waktu yang lainnya. Sebagai contoh seorang pelanggan merupakan objek, dimana objek pelanggan tersebut memiliki beberapa atribut seperti id pelanggan, nama, alamat dan lain-lain. Setiap pelanggan memungkinkan memiliki nilai atribut yang berbeda dengan pelanggan lainnya, serta memungkinkan perubahan nilai atribut dari waktu ke waktu. (Kusrini, 2009)

2.3 *Software* R

Dalam menyelesaikan kasus *Data mining*, perlu digunakan *software*. Salah satu *software* yang sering di gunakan yaitu *software* R. *Software* R adalah salah satu dari program sumber terbuka yang dapat diunduh secara gratis. R diluncurkan pertama kali tahun 1997, dan versi terakhir ketika tulisan diluncurkan ini dibuat tahun 1997, dan versi terakhir ketika tulisan ini dibuat adalah 3.1.0. R studio adalah salah satu GUI untuk R, salah satu keunggulan pada R studio ini adalah dapat dijalankan pada browser sehingga dijalankan diatas browser, maka pengguna tidak memerlukan lagi installasi R, kecuali paket (*package*) pemrograman sesuai dengan kebutuhan pengguna. (Muharom dkk. 2016).

2.4 Decision Tree

Algoritma *Decision Tree* merupakan salah satu algoritma untuk metode data mining yang sering diterapkan sebagai solusi untuk mengklasifikasikan sebuah masalah. *Decision Tree* juga merupakan salah satu klasifikasi pada data mining. Pembuatan *Decision Tree* sendiri menggunakan metode *supervised machine learning* yaitu proses pembelajaran dimana data baru diklasifikasikan berdasarkan training samples yang sudah ada. *Decision Tree* terdiri dari *root*, internal *node* dan *leaf*. Konsep yang sering digunakan untuk penentuan *root*, internal *node* dan juga *leaf* pada *Decision Tree* juga berupa konsep *entropy* dan konsep Gini (Ananto, 2017).

Decision Tree dapat diterapkan untuk mempelajari klasifikasi dan memprediksi pola dari data dan menggambarkan relasi dari variabel attribut x dan variabel target y dalam bentuk pohon. D*ecision Tree* merupakan struktur yang menyerupai *flowchart* dimana untuk setiap internal node merupakan pengujian terhadap variabel atribut, tiap cabangnya adalah hasil dari pengujian tersebut, sedangkan *node* terluar yaitu *leaf* menjadi labelnya (Ananto, 2017).

2.5 Evaluasi Kinerja Pengklasifikasian

Klasifikasi merupakan teknik *data mining* yang banyak digunakan untuk data di bidang pembelajaran. Banyak penelitian memakai data mining pada bidang pembelajaran khususnya untuk Teknik klasifikasi. Klasifikasi merupakan suatu proses analisa pada data yang menciptakan model-model untuk menggambarkan kelas-kelas yang ada dari data tersebut Model tersebut kerap disebut klasifikasi. Sehingga klasifikasi inilah yang hendak digunakan guna menyusun kelas-kelas yang tercantum dalam data, misalnya untuk metode *Decision Tree* maka kelas-kelas tersebut ditafsirkan dalam wujud pohon berakar (Ananto dkk., 2017).

Proses klasifikasi data didasarkan oleh 4 komponen mendasar antara lain yaitu kelas, prediktor, training set, serta pengujian dataset. Di antara sebagian model klasifikasi yang sangat popular yaitu metode *Decision/Classification Trees*, *Bayesian Classifiers*/*Naïve Bayes Classifiers*, *Neural Networks*, *Statistical Analysis*, *Genetic Algorithms*, *Rough Sets*, *K-Nearest Neighbor Classifier*, *Rule-based Methods*, *Memory Based Reasoning*, *Support Vector Machines*. Klasifikasi merupakan suatu proses guna menghasilkan fungsi ataupun model menerangkan kelas pada data ataupun konsep guna untuk memprediksi kelas dari suatu objek yang labelnya belum didapatkan. (Ananto dkk., 2017)

2.6 Association Rules

Association rules adalah salah satu task data mining deskriptif yang bertujan untuk menemukan aturan asosiasif antara item data. Langkah utama yang perlu dalam *association rules* adalah mengetahui seberapa sering kombinasi item muncul dalam database, yang disebut sebagai *frequent patterns*. Pramudiono menyatakan bahwa penting tidaknya suatu aturan asosiatif dapat diketahui dengan dua parameter yakni support yaitu sebuah persentase kombinasi item dalam *database* dan *confidence* yaitu kuatnya hubungan antar item-item dalam aturan asosiatif sendiri (Arafah, 2015).

BAB 3

METODOLOGI PENELITIAN

3.1 Jenis, Waktu dan Tempat Pelaksanaan

Penelitian ini merupakan Tugas Proyek yang telah di diskusikan dan di kerjakan mulai pada tanggal 18 Maret 2022 sampai 8 April 2022. Tugas Proyek telah dilakukan di tempat masing-masing dengan berbagi tugas dan tanggung jawab. Diskusi dilakukan melalui perantara jaringan sosial media yaitu *WhatsApp.*

3.2 Variabel dan Instrument

Dalam Tugas Proyek ini, instrument yang kami gunakan yaitu:

- Dataset pada Kaggle untuk mengambil data yang akan diolah.
- *Software* R sebagai alat bantu dalam pengolahan data.
- *WhatsApp* sebagai alat bantu diskusi
- Sikola dan github account sebagai media kumpul tugas

Adapun variabel yang akan digunakan yaitu

- Pada dataset *House Price* variable yang digunakan yaitu, *price* atau harga rumah sebagai variabel respon dan terdapat 14 variabel prediktor (beberapa variabel prediktor yang dianggap tidak penting sudah melalui proses *cleaning*).
- Pada dataset *Grocery Store* variable yang digunakan yaitu barang belanja yang terdiri dari 11 jenis yakni: Biscuit, Bournvita, Bread, Cock, Coffee, Cornflakes, Jam, Maggi, Milk, Suger dan Tea.

3.3 Tahap Pelaksanaan

Dalam tugas proyek ini, adapun prosedur yang kami gunakan yaitu:

1. Membuat Grup diskusi menggunakan *WhatsApp* dan membagi tugas dan tanggung-jawab masing-masing.
2. Mengunduh dataset pada Kaggle dan mendeskripsi data yang ada
3. Melakukan *pra-procesing* sebelum mengolah data lanjut.
4. Mengklasifikasikan Data *House Price* dengan *Decision Tree*
5. Menggunakan *Association Rules*
6. Menginterpretasikan hasil data dan menyatukan makalah.

3.4 Data dan Sumber Data

Data yang digunakan merupakan data sekunder yang di ambil dari dataset pada Kaggle (sesuai petunjuk soal). Adapun dataset yang digunakan yaitu: *House Price,* data dapat dilihat pada Lampiran 1 dan *Grocery Store,* data dapat dilihat pada Lampiran 2.

BAB 4

ANALISIS DATA

4.1 Deskripsi Data

Data yang digunakan bersumber dari dataset yang ada di Kaggle (sesuai petunjuk soal). Dari petunjuk soal, terdapat dua data yang digunakan yaitu:

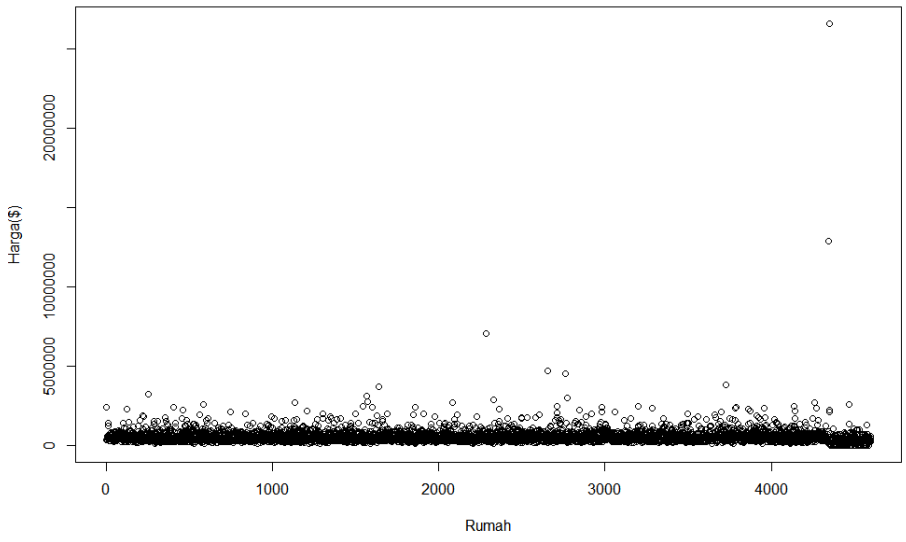
4.1.1 House Price

Pada data *House Price*, terdapat 18 variabel dimana, akan digunakan untuk mengklasifikasikan Harga Rumah yang gambaran data awal disajikan pada Gambar 1. Tabel 5 data teratas dari *House Price* dibawah ini.

date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	sqft_above	sqft_basem	yr_built	yr_renovati	street	city	statezip	country
5/2/2014 0:00	313000	3	1.5	1340	7912	1.5	0	0	3	1340	0	1955	2005	18810 Den	Shoreline	WA 98133	USA
5/2/2014 0:00	2384000	5	2.5	3650	9050	2	0	4	5	3370	280	1921	0	709 W Blair	Seattle	WA 98119	USA
5/2/2014 0:00	342000	3	2	1930	11947	1	0	0	4	1930	0	1966	0	26206-262	Kent	WA 98042	USA
5/2/2014 0:00	420000	3	2.25	2000	8030	1	0	0	4	1000	1000	1963	0	857 170th	Bellevue	WA 98008	USA
5/2/2014 0:00	550000	4	2.5	1940	10500	1	0	0	4	1140	800	1976	1992	9105 170th	Redmond	WA 98052	USA
5/2/2014 0:00	490000	2	1	880	6380	1	0	0	3	880	0	1938	1994	522 NE 88t	Seattle	WA 98115	USA

Gambar 1. Tabel 5 data teratas dari *House Price*

Terlihat pada Gambar 1. Variabel-variabel yang tidak relevan dalam mengklasifikasikan *price* (Harga Rumah) adalah variabel-variabel yang berupa *date*, variabel yang memiliki nilai yang sama seperti *country* dan variabel yang berupa identitas seperti *statezip*, *street* dan *city*. Harga Rumah yang akan diklasifikasikan, juga perlu ditelusuri apakah terdapat anomali atau *outlier*, yang secara visual dapat dilihat pada *scatterplot* dibawah ini.



Gambar 2. Plot Rumah terhadap Harga ($)

Terlihat pada Gambar 2. Terdapat Harga rumah yang tidak masuk akal seperti $ 0 dan beberapa harga rumah yang lebih dari $ 1.000.000 yang dimana, sebelum masuk ke pengolahan data terlebih dahulu akan dibersihkan (*cleaning*), tahap tersebut adalah tahap *preprocessing*.

4.1.2 Grocery Store

Pada data *Grocery Store*, terdapat 20 transaksi pelanggan dengan minimum transaksi sebanyak 2 jenis barang dari 11 jenis barang yang tersedia, yakni *Biscuit, Bournvita, Bread, Cock, Coffee, Cornflakes, Jam, Maggi, Milk, Suger* dan *Tea*.

1	MILK,BREAD,BISCUIT
2	BREAD,MILK,BISCUIT,CORNFLAKES
3	BREAD,TEA,BOURNVITA
4	JAM,MAGGI,BREAD,MILK
5	MAGGI,TEA,BISCUIT
6	BREAD,TEA,BOURNVITA
7	MAGGI,TEA,CORNFLAKES
8	MAGGI,BREAD,TEA,BISCUIT
9	JAM,MAGGI,BREAD,TEA
10	BREAD,MILK
11	COFFEE,COCK,BISCUIT,CORNFLAKES
12	COFFEE,COCK,BISCUIT,CORNFLAKES
13	COFFEE,SUGER,BOURNVITA
14	BREAD,COFFEE,COCK
15	BREAD,SUGER,BISCUIT
16	COFFEE,SUGER,CORNFLAKES
17	BREAD,SUGER,BOURNVITA
18	BREAD,COFFEE,SUGER
19	BREAD,COFFEE,SUGER
20	TEA,MILK,COFFEE,CORNFLAKES

Gambar 3. Data dari *Grocery Store*

4.2 Pre-Proscessing

Pada tahap *preprocessing*, Harga rumah yang termasuk anomali akan dihilangkan karena bisa berdampak pada model yang akan dibangun. Karena Harga rumah ingin diklasifikasi, maka Harga rumah yang datanya berupa numerik akan ditransformasi menjadi kategorik ke dalam dua kategori, yaitu kategori dibawah rata-rata dan kategori diatas rata-rata. Beberapa variabel juga perlu dikategorikan seperti *basement*, *yr_renov* dan *yr_build*, dimana proses tersebut terdapat pada Lampiran 3. dan hasilnya terdapat pada Gambar dibawah ini

price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	sqft_above	sqft_basem	yr_built	yr_renovated	s
0	3	1.5	1340	7912	1.5	0	0	3	1340	0	0	1	
1	5	2.5	3650	9050	2	0	4	5	3370	1	0	0	
0	3	2	1930	11947	1	0	0	4	1930	0	0	0	
0	3	2.25	2000	8030	1	0	0	4	1000	1	0	0	
0	4	2.5	1940	10500	1	0	0	4	1140	1	1	1	

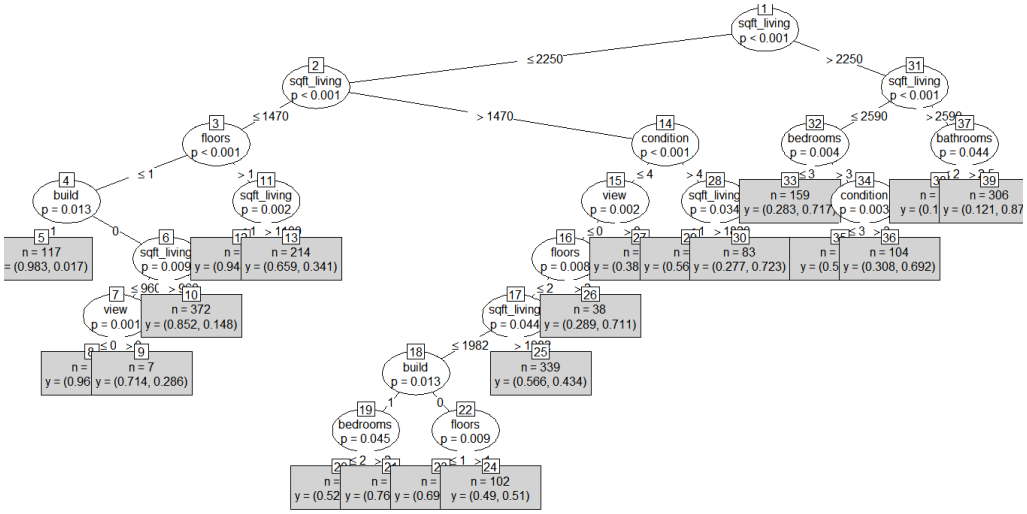
Gambar 4. Tabel hasil dari *preprocessing*

Data dari hasil dari *preprocessing* akan dilanjutkan ke tahap *modeling* yaitu *Decision Tree* dengan *price* sebagai variabel respon dan 14 variabel lainya sebagai prediktor.

4.3 Klasifikasi Data House Price dengan Decision Tree

Sebelum data diklasifikasan, data terlebih dahulu dibagi menjadi dua bagian yaitu bagian data *train* dan bagian data *test*. Data *train* akan digunakan untuk membangun model, sedangkan data *test* digunakan sebagai evaluasi model. Data *train* dibagi menjadi 80% atau 3336 observasi dari data, sedangkan data *test* adalah 20% atau 874 observasi dari data. Setelah data dibagi, maka dilakukan pemodelan *Decision Tree* untuk mengklasifikasikan Harga Rumah dan proses keseluruhan pembagian dan proses modeling terdapat pada Lampiran 4.

4.3.1 Interpretasi Model



Gambar 5. Model *Decision Tree* klasifikasi Harga rumah

Terlihat Gambar 3. Model dari Klasifikasi dari *Decision Tree* didapatkan, bahwa variabel paling berpengaruh adalah *sqt_living* (luas ruang tamu) terhadap Harga jual dari Rumah di Amerika, karena memiliki nilai *gini* terbesar diantara atribut yang lain, sehingga menjadi akar dari model yang terbentuk.

Misalkan seseorang ingin membeli rumah dengan kriteria luas ruang tamu sebesar 1300, lantai 1 dan rumah dengan kondisi baru, maka berdasarkan model diprediksi harga jual dari rumah tersebut adalah dibawah rata-rata atau dibawah $473.500.

4.3.2 Evaluasi Kinerja Klasifikasi

Setelah dilakukan modeling, maka dilanjutkan ke tahap evaluasi untuk mengukur kinerja dari sebuah klasifikasi. Kinerja tersebut akan diuji dengan data *testing* dan hasil prediksi dari model yang telah dibangun. Hasil prediksi dan nilai aktual disajikan pada tabel 4.1.

Tabel 4.1 Matriks kelas prediksi dan kelas aktual

Aktual	Prediksi	
	dibawah rata-rata	diatas rata-rata
dibawah rata-rata	418	73
diatas rata-rata	136	247

Akurasi

$$Accuracy = \frac{tp + tn}{(tp + tn + fp + fn)} = \frac{418 + 247}{(418 + 247 + 136 + 73)} = 0.7608$$

Jadi besar akurasi pengkalsifikasian dengan metode *decision tree* adalah sebesar 0.7608 atau 76,08%

Presisi

$$Precision = \frac{tp}{tp + fp} = \frac{418}{418 + 136} = 0.7545$$

Jadi besar ketepatan atau kualitas dari metode *decision tree* adalah sebesar 0.7545 atau 75.45%

Recall

$$recall = \frac{tp}{tp + fn} = \frac{418}{418 + 73} = 0.8513$$

Jadi besar kelengkapan atau kuantitas adalah sebesar 0.8513 atau 85.13%

F measure

$$F = \frac{2(precision)(recall)}{precision + recall} = \frac{2(0.7545)(0.8513)}{0.7545 + 0.8513} = 0.7999$$

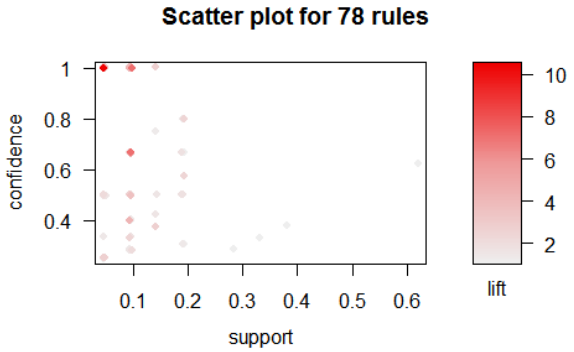
4.4 Association Rules

Pada data *Grocery Store*, asosiasi rule dapat diterapkan dengan melihat pola-pola atau aturan yang dibuat untuk melihat apakah seorang pelanggan memiliki pola dalam berbelanja. Berdasarkan Lampiran 5. Didapatkan dengan nilai *support* minimum 0.005 dan *confidence* 0.25 adalah total 168 aturan. Aturan yang terbentuk juga akan dipangkas melalui fungsi *redudant* pada R diperoleh, dari 168 rule ada total 90 rule yang perlu dipangkas karena 90 aturan tersebut adalah aturan yang berulang atau *redundant*, sehingga didapatkan 78 rule hasil pangkasan. Aturan yang terbentuk dapat disajikan 10 aturan teratas berdasarkan nilai *lift* terbesar pada Tabel 4.2 dibawah ini

Tabel 4.2 Aturan asosiasi dari perbelanjaan pelanggan di *Grocery Store*

lhs	Rhs			support	confidence	coverage	lift	count
[1]	{MAGGI,MILK}	=>	{JAM}	0.047619	1	0.047619	10.5	1
[2]	{BREAD,MAGGI}	=>	{JAM}	0.095238	0.666667	0.142857	7	2
[3]	{BISCUIT,COFFEE}	=>	{COCK}	0.095238	1	0.095238	7	2
[4]	{BISCUIT,CORNFLAKES}	=>	{COCK}	0.095238	0.666667	0.142857	4.666667	2
[5]	{JAM}	=>	{MAGGI}	0.095238	1	0.095238	4.2	2
[6]	{MAGGI}	=>	{JAM}	0.095238	0.4	0.238095	4.2	2
[7]	{BREAD,CORNFLAKES}	=>	{MILK}	0.047619	1	0.047619	4.2	1
[8]	{COFFEE,TEA}	=>	{MILK}	0.047619	1	0.047619	4.2	1
[9]	{BISCUIT,TEA}	=>	{MAGGI}	0.095238	1	0.095238	4.2	2
[10]	{BISCUIT,COCK}	=>	{CORNFLAKES}	0.095238	1	0.095238	3.5	2

Terlihat pada Tabel 4.2 Aturan yang memiliki nilai *lift* terbesar adalah aturan yang dimana jika seorang pelanggan membeli sebuah *Maggi* (mie instan) dan *Milk* (susu), maka pelanggan tersebut juga akan membeli sebuah *Jam* (Selai), dengan nilai *support* atau persentase transaksi yang dilakukan sebesar 4,76 % dan nilai *confidence* atau kepercayaan sebesar 100%. Secara garis besar, semua aturan yang terbentuk dapat dilihat dari visualisasi *Scatter Plot*.



Gambar 6. Scatterplot dari 78 Aturan asoisasi yang terbentuk

BAB 5

PENUTUP

5.1 Kesimpulan

Berdasarkan analisis data, maka kesimpulannya yaitu:

1. Dari hasil pengklasifikasian dengan metode *decision tree*, didapatkan pengaruh terbesar Harga Rumah di Amerika adalah Luas Ruang Tamu dengan dengan evaluasi kinerja pengkalsifikasian, yaitu nilai *accuracy* pengklasifikasian sebesar 76.08% sedangkan nilai *precision* sebesar 75.45% dan nilai *recall* 85.13% .
2. Penerapan aturan asosiasi pada data *Grocery Store* didapatkan banyaknya pola-pola seorang pelanggan dalam berbelanja adalah sebanyak 78 pola, yang dimana salah satu polanya adalah "Jika seorang pelanggan membeli sebuah *Maggi* (mie instan) dan *Milk* (susu), maka pelanggan tersebut juga akan membeli sebuah *Jam* (Selai)".

5.2 Saran

Berdasarkan kesimpulan, Adapun saran yaitu:

1. Bagi peneliti selanjutnya untuk menerapkan pengklasifikasian selanjutnya juga dapat menggunakan metode naïve bayes ataupun regresi logistik biner.

DAFTAR PUSTAKA

Ananto, R. P., Purwanto, Y., & Novianty, A. (2017). Deteksi Jenis Serangan pada Distributed Denial of Service Berbasis Clustering dan Classification menggunakan Algoritma Minkowski Weighted KMeans dan Decision Tree. EProceeding of Engineering, 4(1), 879–886.

Arafah, A.A. & Mukhlash, I., (2015), "The Application of Fuzzy Association Rule on Co-Movement Analyze of Indonesian Stock Price", *Procedia Computer Science* 59 (2015) 235-243.

Bramer, Max. (2007) *Principles of Data Mining*, Springer Science

Han, J., Kamber, M., & Pei, J., (2011), *Data Mining: Concepts and Techniques Third Edition*, Morgan Kaufmann, USA.

Jollyta, D., Ramadhan, W. dan Zarlis, M. 2020. *Konsep Data Mining Dan Penerapan*. Yogyakarta: Deepublish.

Kusrini & Taufiz, E.L. (2009). Algoritma Data Mining. Yogyakarta: Andi

Larose, Daniel T. (2005) *Discovering Knowledge in Data Mining An Introduction to Data Mining*, Wiley Interscience.

Muharom, L.A., Hadi, A.F., & Anggraeni, D., (2016), "Rancang Bangun Data *Warehouse* dan R Studio Serta Pemanfaatanya dalam Peramalan Pola Konsumsi Masyarakat di Kabupaten Jember", Jember, Indonesia.

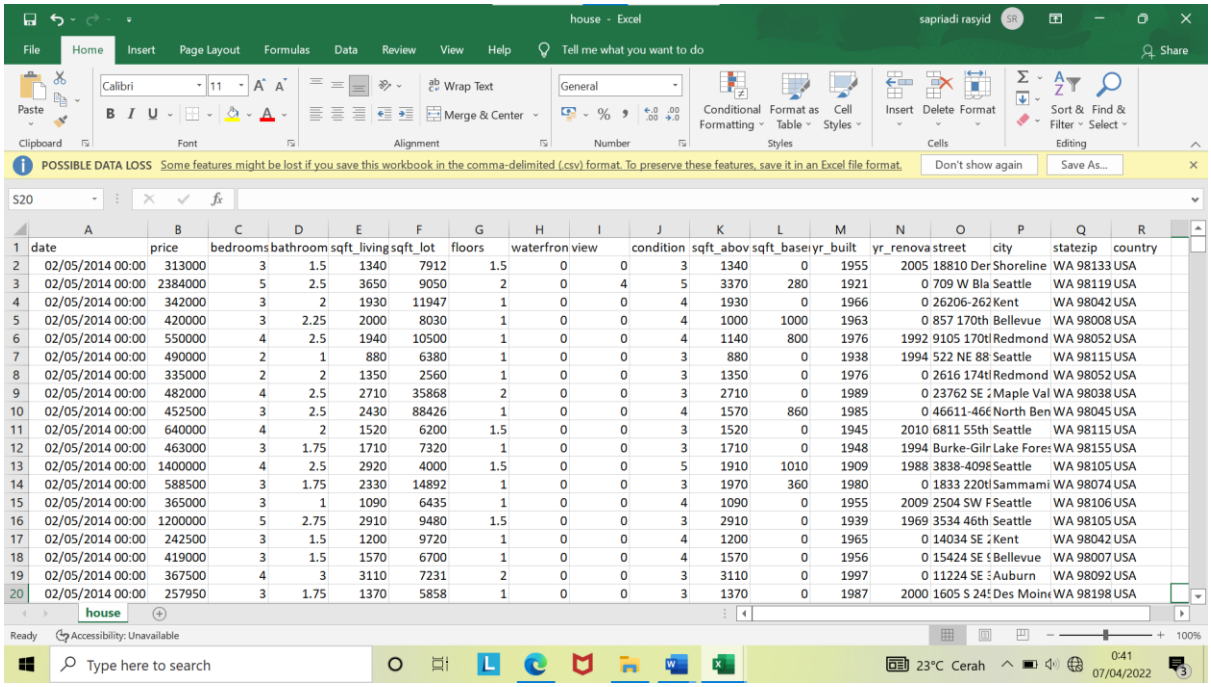
Pramana, S., dkk. 2018. Data Mining dengan R Konsep serta Implementasi. Bogor: In Media.

Suyanto.2019. *Data Mining Untuk Klasifikasi dan Klasterisasi Data, Edisi Revisi*. Bandung: Informatika Bandung.

.

LAMPIRAN

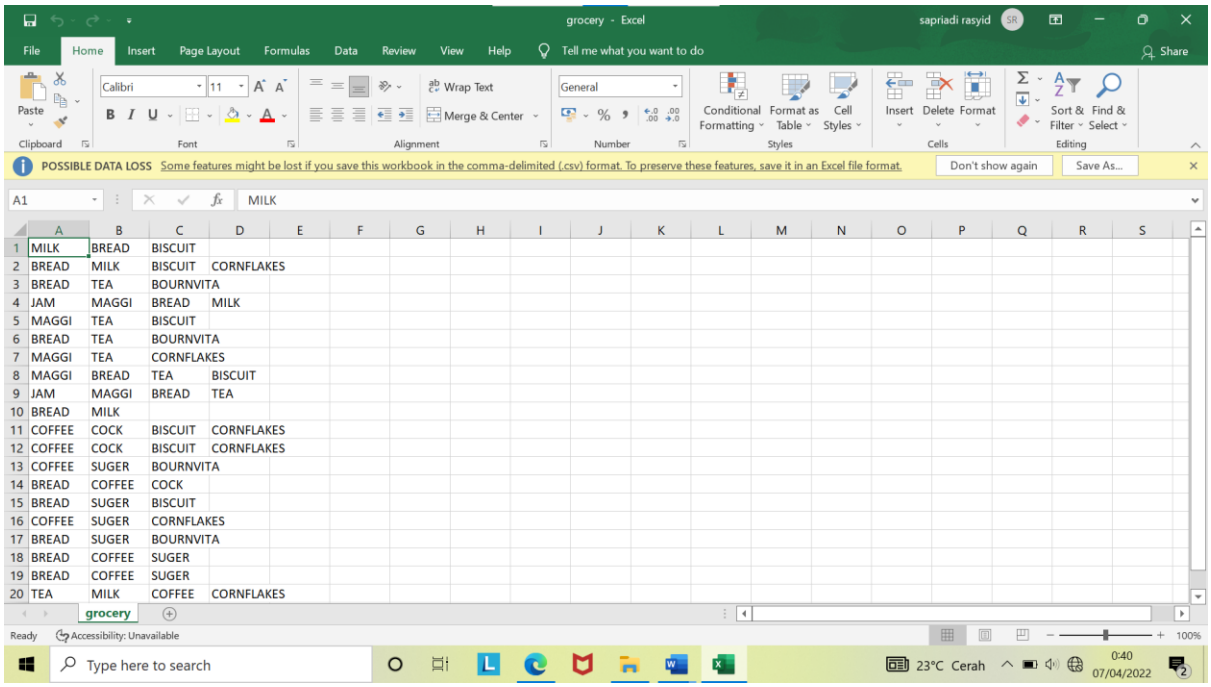
Lampiran 1. Tampilan Kaggle House Prediction di csv



house - Excel

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	date	price	bedrooms	bathroom	sqft_living	sqft_lot	floors	waterfron	view	condition	sqft_abov	sqft_base	yr_built	yr_renova	street	city	statezip	country
2	02/05/2014 00:00	313000	3	1.5	1340	7912	1.5	0	0	3	1340	0	1955	2005	18810 Der	Shoreline	WA 98133	USA
3	02/05/2014 00:00	2384000	5	2.5	3650	9050	2	0	4	5	3370	280	1921	0	709 W Bla	Seattle	WA 98119	USA
4	02/05/2014 00:00	342000	3	2	1930	11947	1	0	0	4	1930	0	1966	0	26206-262	Kent	WA 98042	USA
5	02/05/2014 00:00	420000	3	2.25	2000	8030	1	0	0	4	1000	1000	1963	0	857 170th	Bellevue	WA 98008	USA
6	02/05/2014 00:00	550000	4	2.5	1940	10500	1	0	0	4	1140	800	1976	1992	9105 170t	Redmond	WA 98052	USA
7	02/05/2014 00:00	490000	2	1	880	6380	1	0	0	3	880	0	1938	1994	522 NE 88	Seattle	WA 98115	USA
8	02/05/2014 00:00	335000	2	2	1350	2560	1	0	0	3	1350	0	1976	0	2616 174t	Redmond	WA 98052	USA
9	02/05/2014 00:00	482000	4	2.5	2710	35868	2	0	0	3	2710	0	1989	0	23762 SE ź	Maple Val	WA 98038	USA
10	02/05/2014 00:00	452500	3	2.5	2430	88426	1	0	0	4	1570	860	1985	0	46611-466	North Ben	WA 98045	USA
11	02/05/2014 00:00	640000	4	2	1520	6200	1.5	0	0	3	1520	0	1945	2010	6811 55th	Seattle	WA 98115	USA
12	02/05/2014 00:00	463000	3	1.75	1710	7320	1	0	0	3	1710	0	1948	1994	Burke-Gilr	Lake Fores	WA 98155	USA
13	02/05/2014 00:00	1400000	4	2.5	2920	4000	1.5	0	0	5	1910	1010	1909	1988	3838-4098	Seattle	WA 98105	USA
14	02/05/2014 00:00	588500	3	1.75	2330	14892	1	0	0	3	1970	360	1980	0	1833 220t	Sammami	WA 98074	USA
15	02/05/2014 00:00	365000	3	1	1090	6435	1	0	0	4	1090	0	1955	2009	2504 SW F	Seattle	WA 98106	USA
16	02/05/2014 00:00	1200000	5	2.75	2910	9480	1.5	0	0	3	2910	0	1939	1969	3534 46th	Seattle	WA 98105	USA
17	02/05/2014 00:00	242500	3	1.5	1200	9720	1	0	0	4	1200	0	1965	0	14034 SE ź	Kent	WA 98042	USA
18	02/05/2014 00:00	419000	3	1.5	1570	6700	1	0	0	4	1570	0	1956	0	15424 SE ź	Bellevue	WA 98007	USA
19	02/05/2014 00:00	367500	4	3	3110	7231	2	0	0	3	3110	0	1997	0	11224 SE ź	Auburn	WA 98092	USA
20	02/05/2014 00:00	257950	3	1.75	1370	5858	1	0	0	3	1370	0	1987	2000	1605 S 24ź	Des Moine	WA 98198	USA

Lampiran 2. Tampilan Kaggle Grocery Store di csv



	A	B	C	D
1	MILK	BREAD	BISCUIT	
2	BREAD	MILK	BISCUIT	CORNFLAKES
3	BREAD	TEA	BOURNVITA	
4	JAM	MAGGI	BREAD	MILK
5	MAGGI	TEA	BISCUIT	
6	BREAD	TEA	BOURNVITA	
7	MAGGI	TEA	CORNFLAKES	
8	MAGGI	BREAD	TEA	BISCUIT
9	JAM	MAGGI	BREAD	TEA
10	BREAD	MILK		
11	COFFEE	COCK	BISCUIT	CORNFLAKES
12	COFFEE	COCK	BISCUIT	CORNFLAKES
13	COFFEE	SUGER	BOURNVITA	
14	BREAD	COFFEE	COCK	
15	BREAD	SUGER	BISCUIT	
16	COFFEE	SUGER	CORNFLAKES	
17	BREAD	SUGER	BOURNVITA	
18	BREAD	COFFEE	SUGER	
19	BREAD	COFFEE	SUGER	
20	TEA	MILK	COFFEE	CORNFLAKES

Lampiran 3. *Preprocessing*

```
df<-read.csv(file ="C:/Users/lenovo/UNHAS/SEMESTER 6/DATA MINING/house.csv",
        header=TRUE, sep =";")
deta <- df[-c(1,15,16,17,18)]
deta$price <- ifelse(deta$price == 0, NA, deta$price)
deta$price <- ifelse(deta$price > 1000000, NA, deta$price)
deta <- deta[complete.cases(deta), ]

deta$price2 <- ifelse(deta$price > mean(deta$price), 1, 0)
deta$renovated <- ifelse(deta$yr_renovated > 0, 1, 0)
deta$build <- ifelse(deta$yr_built> 1970, 1, 0)
deta$basement <- ifelse(deta$sqft_basement > 0, 1, 0)

deta$basement <- as.factor(deta$basement)
deta$build <- as.factor(deta$build)
deta$renovated <- as.factor(deta$renovated)
deta$price2 <- as.factor(deta$price2)
deta<-deta[,-c(1,11,12,13)]
s <- sample(2, nrow(deta), replace = T, prob = c(0.8, 0.2))
train_TP1 <- deta[s == 1,]
test_TP1 <- deta[s == 2,]
```

Lampiran 4. Pemodelan Klasifikasi Harga Rumah dengan *Decision Tree*

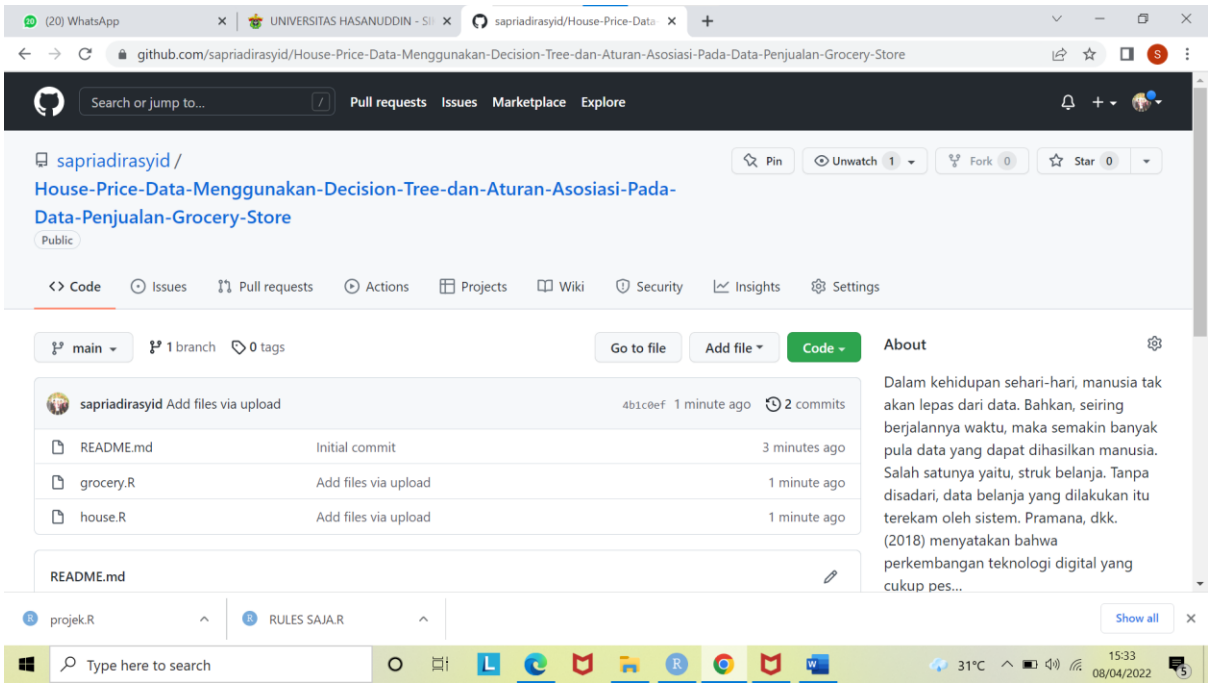
```
library(party)
library(rpart)
library(caret)
pohon <-ctree(price2~.,data=train_TP1)
plot(pohon)

predict_model<-predict(pohon, test_TP1)
predict_model
m_at <- table(test_TP1$price2, predict_model)
m_at
```

Lampiran 5. Model aturan Asosiasi pada Data *Grocery Store*

```
grocery <- read.csv('GroceryStoreDataSet.csv', header=F)
colnames(grocery) <- c('Items')
grocery <- data.frame(grocery)
write.csv(grocery,file="market_basket_transactions.csv", quote = FALSE, row.names = FALSE)
df <- read.transactions("market_basket_transactions.csv", format = 'basket', sep=',')
summary(df)
df
rules = apriori(data = df, parameter = list(support = 0.005, confidence = 0.25))
inspect(rules)
summary(rules)
sortir <- sort(rules, by="lift")
redundant <- is.redundant(sortir)
rules.pruned <- sortir[!redundant]
summary(rules.pruned)
inspect(rules.pruned)
plot(rules.pruned)
```

Lampiran 6. Upload Data Github



Dapat diakses pada link berikut: https://github.com/sapriadirasyid/House-Price-Data-Menggunakan-Decision-Tree-dan-Aturan-Asosiasi-Pada-Data-Penjualan-Grocery-Store.git