MODEL *DATA MINING* UNTUK MEMPREDIKSI PERTUMBUHAN PENDUDUK DI KOTA BATAM DENGAN MENGGUNAKAN TEKNIK *DECISION TREE*

SKRIPSI

Oleh:

Deni Lugi Wiyono
140210040

PROGRAM STUDI TEKNIK INFORMATIKA
UNIVERSITAS PUTERA BATAM
2019

MODEL *DATA MINING* UNTUK MEMPREDIKSI PERTUMBUHAN PENDUDUK DI KOTA BATAM DENGAN MENGGUNAKAN TEKNIK *DECISION TREE*

SKRIPSI

Untuk memenuhi salah satu syarat
guna memperoleh gelar Sarjana

Oleh
Deni Lugi Wiyono
140210040

PROGRAM STUDI TEKNIK INFORMATIKA
UNIVERSITAS PUTERA BATAM
2019

PERNYATAAN

Dengan ini saya menyatakan bahwa:

1. Skripsi ini adalah asli dan belum pernah diajukan untuk mendapatkan gelar akademik (sarjana, dan/atau magister), baik di Universitas Putera Batam maupun di perguruan tinggi lain.
2. Skripsi ini adalah murni gagasan, rumusan, dan penelitian saya sendiri, tanpa bantuan pihak lain, kecuali arahan pembimbing.
3. Dalam skripsi ini tidak terdapat karya atau pendapat yang telah ditulis atau dipublikasikan orang lain, kecuali secara tertulis dengan jelas dicantumkan sebagai acuan dalam naskah dengan disebutkan nama pengarang dan dicantumkan dalam daftar pustaka.
4. Pernyataan ini saya buat dengan sesungguhnya dan apabila di kemudian hari terdapat penyimpangan dan ketidakbenaran dalam pernyataan ini, maka saya bersedia menerima sanksi akademik berupa pencabutan gelar yang telah diperoleh, serta sanksi lainnya sesuai dengan norma yang berlaku di perguruan tinggi.

Batam, 9 Agustus 2019

Yang membuat pernyataan,

Materai Rp 6.000,00

Deni Lugi Wiyono
140210040

MODEL *DATA MINING* UNTUK MEMPREDIKSI PERTUMBUHAN PENDUDUK DI KOTA BATAM DENGAN MENGGUNAKAN TEKNIK *DECISION TREE*

Oleh

Deni Lugi Wiyono

140210040

SKRIPSI

Untuk memenuhi salah satu syarat

guna memperoleh gelar Sarjana

Telah disetujui oleh Pembimbing pada tanggal

seperti tertera di bawah ini

Batam, 9 Agustus 2019

Yulia, S.Kom., M.Kom.

ABSTRAK

Pertumbuhan penduduk merupakan keadaan di mana suatu wilayah mengalami peningkatan jumlah populasinya. Penduduk adalah orang yang tinggal dalam suatu wilayah dalam kurun waktu tertentu. Penelitian ini dilakukan degan tujuan untuk memprediksi pertumbuhan penduduk di kota Batam dengan pemanfaatan ilmu *data mining* menggunakan *algoritma* C4.5. Metode yang digunakan adalah klasifikasi, teknik ini mempelajari sekumpulan data sehingga menghasilkan aturan yang bisa mengklasifikasikan atau menggali data-data baru yang belum pernah dipelajari. Klasifikasi merupakan proses membagi data menjadi suatu anggota suatu kategori atau kelas. *Software* yang dipakai untuk menguji adalah Weka. *Algoritma* C4.5 merupakan salah satu *algoritma* yang ada di *decision tree* dan *algoritma* yang banyak digunakan untuk menghasilkan pohon keputusan. *Algoritma* C4.5 merupakan pengembangan ID3. Pembentukan *tree* pada *algoritma* C4.5 menganut pendekatan *top-down* di mana *tree* dibentuk dari *root* menuju *leaf*. Variabel yang digunakan dalam penelitian ini adalah jumlah kelahiran, jumlah kematian, jumlah pindah dan jumlah datang. Data dari Dinas Kependudukan dan Pencatatan Sipil Kota Batam (Disduk Capil) dari 2014 – 2018. Data diklasifikasikan menjadi tiga yaitu tinggi, sedang dan rendah. Hasil dari penelitian ini adalah berupa pohon keputusan, yang bisa digunakan untuk mengetahui bagaimana proses pertumbuhan penduduk yang terjadi di kota Batam. Berdasarkan uji akurasi menunjukan nilai 0,89 dan uji *error rate* bernilai 0,10 dari hasil pengujian ini menunjukan *algoritma* C4.5 merupakan *algoritma* yang baik dalam penggunaanya.

Kata Kunci: Populasi, *Data Mining, Algoritma C4.5, Decision Tree, Error Rate*.

ABSTRACT

Population growth is a condition where an area has an increasing population. Residents are people who live in an area within a certain period. This research was conducted with the aim of predicting population growth in the city of Batam *with the use of data mining using the C4.5 algorithm. The method used is classification, this technique studies a set of data so as to produce rules that can classify or explore new data that has never been studied. Classification is the process of dividing data into members of a category or class. The software used to test is Weka. C4.5 algorithm is one of the algorithms in the decision tree and the algorithm is widely used to produce a decision tree. C4.5 algorithm is the development of ID3. The tree formation in C4.5 algorithm follows the top-down approach where the tree is formed from root to leaf. The variables used in this study are number of births, number of deaths, number of transfers and number of arrivals. Data from the* Batam *City Population and Civil Registry Office (*Disduk Capil*) from 2014-2018. The data are classified into three namely high, medium and low. The results of this study are in the form of a decision tree, which can be used to find out how the population growth process that occurs in the city of* Batam. *Based on the accuracy test shows the value of 0,89 and the error rate test value of 0,10 from the results of this test shows the C4.5 algorithm is a good algorithm in its use.*

Keywords:.*Population, Data Mining, C4.5Algorithm,* Decision Tree, *Error Rate.*

KATA PENGANTAR

Segala puji dan syukur kehadirat Allah SWT yang telah melimpahkan segala rahmat dan karuniaNya, sehingga penulis dapat menyelesaikan laporan tugas akhir yang merupakan salah satu persyaratan untuk menyelesaikan program studi strata satu (S1) pada Program Studi Teknik Informatika Universitas Putera Batam.

Penulis menyadari bahwa skripsi ini masih jauh dari sempurna. Karena itu, kritik dan saran akan senantiasa penulis terima dengan senang hati.

Dengan segala keterbatasan, penulis menyadari pula bahwa skripsi ini takkan terwujud tanpa bantuan, bimbingan, dan dorongan dari berbagai pihak. Untuk itu, dengan segala kerendahan hati, penulis menyampaikan ucapan terima kasih kepada:

1. Rektor Universitas Putera Batam.
2. Bapak Andi Maslan, S.T., M.SI. Selaku Ketua Program Studi Teknik Informatika Universitas Putera Batam.
3. Ibu Yulia, S.Kom., M.Kom. Selaku pembimbing Skripsi pada Program Studi Teknik Informatika Universitas Putera Batam.
4. Ibu Sestri Novia Rizki, S.Kom., M.Kom. Selaku pembimbing akademik yang Tak bosennya selalu menggingatkan penulis untuk rajin belajar.
5. Kedua orang tua beserta kakak dan adik yang telah memberikan doa dan dukungan selama proses pembuatan skripsi.
6. Teman-teman satu angkatan 2014 yang telah memberikan dukungan dalam Proses pembuatan skripsi.

4. Dosen dan Staff Universitas Putera Batam
5. Dinas Kependudukan dan Pencatatan Sipil (Disduk) Kota Batam yang telah memberi kesempatan untuk mengambil penelitian disana.

Penulis mohon maaf atas segala kesalahan yang pernah dilakukan. Semoga Allah SWT yang membalas kebaikan dan selalu mencurahkan hidayah serta taufikNya, Aamiin ya rabbal'alamiin.

Batam, 7 Agustus 2019

Penulis

Deni Lugi Wiyono

DAFTAR ISI

DAFTAR LAMPIRAN

DAFTAR TABEL

Halaman

DAFTAR GAMBAR

Halaman

DAFTAR RUMUS

Halaman

DAFTAR LAMPIRAN

BAB I

PENDAHULUAN

1.1 Latar Belakang

Kota Batam adalah sebuah kota yang berada di Provinsi Kepulauan Riau, Indonesia. Wilayah kota Batam terdiri dari Pulau Batam, Pulau Rempang dan Pulau Galang dan pulau-pulau kecil lainnya di kawasan Selat Singapura dan Selat Malaka. Selain berada di jalur pelayaran Internasional, kota Batam juga dijuluki sebagai kota industri dan pariwisata, karena begitu banyaknya kawasan industri dan juga tempat-tempat wisata. Dari segi transportasi yang ada juga sangat memadai seperti halnya pelabuhan logistik dan pelabuhan penumpang yang sangat menunjang akses bagi pertumbuhan ekonomi dan pertumbuhan penduduk. Pertumbuhan suatu wilayah tidak terlepas dari adanya penduduk. Penduduk merupakan orang yang berada di suatu wilayah dengan lama waktu tertentu dan tercatat di daerah tersebut. Menurut (Ruslan, 2016) penduduk adalah orang-orang yang berada di dalam suatu wilayah yang terikat oleh aturan–aturan yang berlaku dan saling berinteraksi satu sama lain secara terus menerus/kontinyu.

Masalah pertumbuhan penduduk merupakan hal yang tidak bisa dihindari di kota Batam, karena kota Batam adalah kota industri di mana industri membutuhkan banyak tenaga kerja. Salah satu faktor yang mempengaruhi pertumbuhan penduduk adalah kelahiran, faktor kelahiran bersifat menambah

jumlah penduduk. Tingkat kelahiran sendiri dipengaruhi oleh Pernikahan di usia dini yang terjadi pada perempuan usia 15 tahun mempunyai masa reproduksi jauh lebih panjang dibanding mereka yang menikah di atas usia 25 tahun di mana masa reproduksi yang lama maka kemungkinan untuk melahirkan semakin besar sehingga bisa saja mempunyai anak lebih dari dua bahkan lebih dari lima (Normalasari, Gani, & Amalia, 2018).

Migrasi juga merupakan salah satu faktor yang menyebabkan pertumbuhan penduduk yang besar di kota Batam. Ada beberapa hal yang mendorong terjadinya migrasi yaitu menyempitnya lapangan pekerjaan di tempat asal, kesempatan mendapatkan pekerjaan yang lebih baik karena hal itulah banyak orang merantau ke kota Batam untuk mencari kehidupan yang lebih baik. Berdasarkan data dari Dinas Kependudukan dan Pencatatan Sipil Kota Batam jumlah penduduk datang pada tahun 2014 berjumlah 15.685, tahun 2015 berjumlah 46.385, tahun 2016 berjumlah 63.309, tahun 2017 berjumlah 48.078, tahun 2018 berjumlah 14.875. Selain kelahiran dan migrasi masuk sebagai faktor yang menambah jumlah pertumbuhan penduduk faktor migrasi keluar dan kematian juga berbengaruh terhadap jumlah pertumbuhan penduduk di kota Batam. Menurut data dari Dinas Kependudukan dan Pencatatan Sipil Kota Batam jumlah penduduk yang melakukan pindah dari tahun 2014 berjumlah 9.644, tahun 2015 berjumlah 26.009, tahun 2016 berjumlah 39.220, tahun 2017 berjumlah 46.943, pada tahun 2018 berjumlah 17.654. Dari data tersebut menunjukan angka perpindahan penduduk dari luar ke Batam lebih tinggi jika dibandingkan dengan

perpindahan penduduk dari Batam ke luar, hal inilah yang menjadi salah satu faktor meningkatnya pertumbuhan penduduk di kota Batam.

Data mining menggunakan teknik maupun metode tertentu yang ada di *data mining*. *Data mining* merupakan ilmu yang yang menghasilkan informasi berupa pola. *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 (Yuli, 2017). Metode yang peneliti gunakan adalah *decision tree* dengan algoritma C4.5. (Yulia & Nurul Azwanti, 2018) Algoritma C4.5 yaitu sebuah algoritma yang digunakan untuk membangun *decision tree* (pengambilan keputusan). Metode ini adalah metode yang sangat populer, dengan pohon keputusan yang menggubah fakta besar menjadi pohon keputusan. Dengan pemanfaatan *data mining* ini diharapkan dapat memberikan alternatif lain dalam memperkirakan dan memprediksi pertumbuhan penduduk di kota Batam pada setiap tahunnya. Dari permasalahan tersebut, maka peneliti ingin mengangkat judul penelitian yaitu **"Model *Data Mining* Untuk Memprediksi Pertumbuhan Penduduk Di Kota Batam Dengan Menggunakan Teknik *Decision Tree* "**.

1.2 Identifikasi Masalah

Berdasarkan latar belakang yang telah dipaparkan, identifikasi masalah yang di ambil sebagai berikut:

1. Tingginya angka kelahiran di kota Batam.
2. Perpindahan penduduk dari luar ke kota Batam tiap tahunnya meningkat.
3. Penduduk yang keluar Batam lebih sedikit jika dibandingkan dengan yang datang dari luar kota Batam.
4. Jumlah angka kematian lebih sedikit dibandingkan angka kelahiran.

1.3 Pembatasan Masalah

Agar penelitian ini lebih fokus pada inti penelitian, maka berdasarkan latar belakang peneliti melakukan pembatasan masalah sebagai berikut:

1. Data yang diolah adalah data penduduk kota Batam tahun 2014 – 2018.
2. Dengan pemanfaatan ilmu *data mining* dengan metode *decision tree* dan algoritma C4.5.
3. Penelitian ini dilakukan di Dinas Kependudukan dan Pencatatan Sipil Kota Batam.
4. *Software* yang dipakai adalah Weka 3.8.3.

1.4 Rumusan Masalah

Berdasarkan uraian di atas, maka dapat dirumuskan beberapa permasalahan yang muncul, diantaranya adalah:

1. Bagaimana cara menentukan pertumbuhan penduduk yang terjadi di kota Batam ?
2. Bagaimana membuat model *data mining* dengan metode *decision tree* dan *algoritma* C4.5 ?
3. Bagaimana mengimplementasikan model *data mining* yang sudah dibuat untuk menganalisis penyebab pertumbuhan penduduk dan tingkat validitas dengan menggunakan model *data mining* tersebut ?

1.5 Tujuan Penelitian

Berdasarkan rumusan masalah di atas, maka tujuan penelitian ini adalah:

1. Untuk mengetahui prediksi jumlah pertumbuhan penduduk di kota Batam.
2. Untuk mengetahui bagaimana model *data mining* digunakan dengan metode *decision tree* dalam prediksi tingkat pertumbuhan penduduk.
3. Untuk menguji tingkat validitas pertumbuhan penduduk yang telah ditetapkan di atas dengan model *data mining.*

1.6 Manfaat Penelitian

Penelitian ini diharapkan memberikan manfaat bagi perkembangan ilmu pengetahuan, bagi universitas, bagi instansi terkait yang diteliti maupun bagi peneliti sendiri, sebagai berikut:

1. Bagi Perkembangan Ilmu Pengetahuan
    a. Penelitian ini bisa menambah wawasan dan pengetahuan tentang *data mining*
    b. Memberikan gambaran tentang ilmu *data mining* dalam memprediksi pertumbuhan penduduk menggunakan aplikasi Weka 3.8.3
    c. Penelitian ini diharapkan dapat digunakan untuk menambah referensi penelitian tentang *data mining* pada masa yang akan datang

2. Bagi Universitas
    a. Menambah arsip penelitian tentang *data mining*
    b. Dapat dijadikan referensi bagi mahasiswa yang ingin melakukan penelitian tentang *data mining*

3. Bagi Dinas Terkait
    a. Hasil dari penelitian ini diharapkan bisa dipakai bahan pertimbangan pemerintah dalam pemgambilan kebijakan maupun keputusan pada masa yang akan datang

b. Dinas terkait bisa mengadopsi ilmu *data mining* dan menerapkanya dalam prediksi pertumbuhan penduduk

4. Bagi Peneliti

a. Sebagai salah satu syarat untuk mendapatkan gelar Sarjana

b. Bisa merapkan ilmu yang selama ini diperolah selama masa kuliah dan mempraktekanya lewat penelitian ini

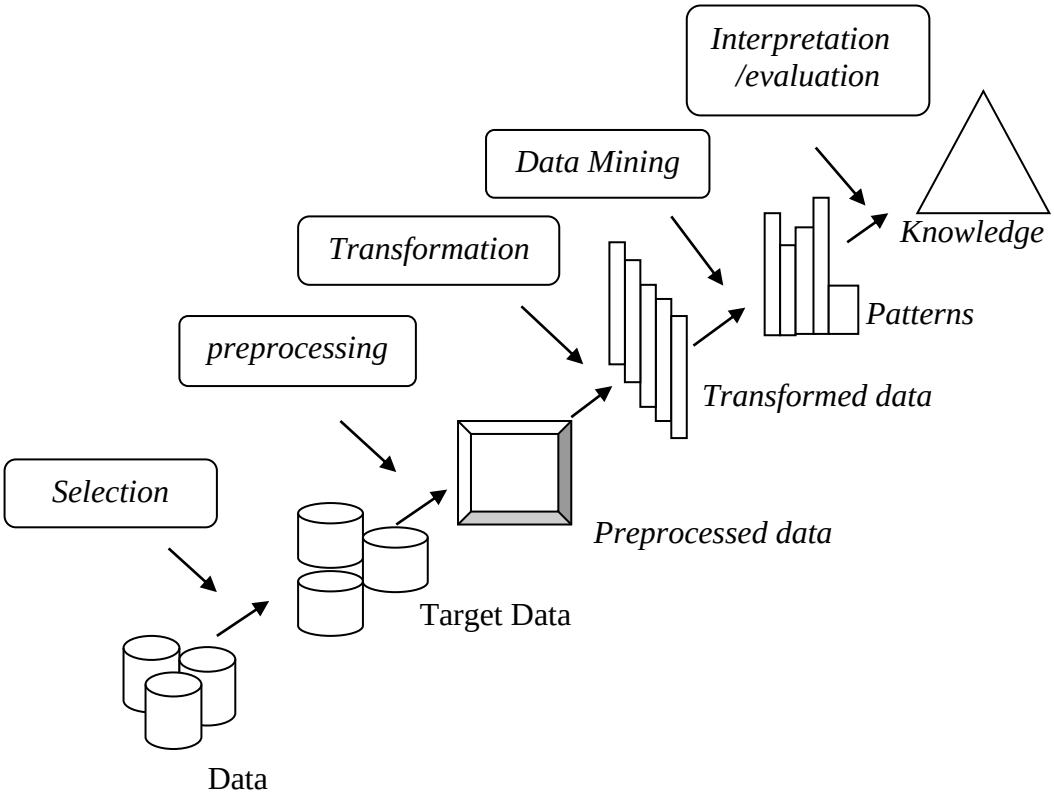
BAB II

KAJIAN PUSTAKA

2.1 *Knowladge Discovery in Database* (KDD)

Dalam aplikasinya, *Data mining* sebenarnya merupakan salah satu bagian dari proses *Knowledge Discovery in Database* (KDD) yang bertugas untuk mengekstrak pola atau model dari data dengan menggunakan suatu *algoritma* yang spesifik. Informasi yang dihasilkan diperoleh dengan cara mengekstraksi dan mengenali pola yang penting atau menarik dari data yang terdapat pada basis data. *Data mining* terutama digunakan untuk mencari pengetahuan yang terdapat dalam basis data yang besar sehingga sering disebut *knowledge Discovery in Database* (Retno Tri Vulandari, 2017).

KDD merupakan keseluruhan proses pencarian pola atau informasi dalam *database*, dimulai dari pemilihan dan persiapan data sampai representasi pola yang ditemukan dalam bentuk yang mudah di mengerti oleh pihak yang berkepentingan. *Data mining* merupakan salah satu komponen dalam KDD yang difokuskan pada pengalian pola tersembunyi dalam *database.*



Gambar 2.1 Tahapan Proses KDD
Sumber: (Retno Tri Vulandari, 2017)

Adapun penjelasan proses KDD sebagai berikut:

1. *Data Selection*: merupakan pemilihan data dari sekumpulan data operasional yang perlu dilakukan sebelum tahap penggalian informasi dalam KDD dimulai. Dari hasil seleksi yang akan digunakan untuk proses *data mining*, disimpan dalam suatu berkas, terpisah dari data operasional.
2. *Pre-processing/Cleaning*: sebelum proses *data mining* dapat dilaksanakan, perlu dilakukan proses *cleaning* pada data yang menjadi fokus KDD. Dengan tujuan untuk membuang duplikasi data, memeriksa data yang inkonsisten, dan memperbaiki kesalahan pada data, seperti kesalahan cetak (*tipografi*). Juga dilakukan proses *enrichment*, yaitu proses "memperkaya"

data yang sudah ada dengan data atau informasi yang relefan dan diperlukan untuk KDD, seperti data atau informasi *eksternal*.

3. *Transformation*: yaitu proses *coding* pada data yang telah dipilih, sehingga data tersebut sesuai untuk proses *data mining*. Proses *coding* dalam KDD merupakan proses kreatif dan sangat tergantung pada jenis atau pola informasi yang akan dicari dalam *database*.
4. *Data Mining*: proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu. Teknik, metode atau *algoritma* dalam *data mining* sangat bervariasi. Pemilihan metode atau *algoritma* yang tepat sangat bergantung pada tujuan dan proses KDD secara keseluruhan.
5. *Interpretation/Evaluation*: pola informasi yang dihasilkan dari proses *data mining* perlu ditampilkan dalam bentuk yang berkepentingan. Tahap ini merupakan bagian dari proses KDD yang disebut dengan *interpretation.*

Tahap ini mencakup pemeriksaan apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesa yang ada sebelumnya. Dalam proses KDD yang sesungguhnya dapat saja terjadi iterasi atau penggulangan pada tahap-tahap tertentu. Pada setiap tahap dalam proses KDD, seorang peneliti dapat saja kembali ke tahap sebelumnya. Sebagai contoh pada saat *coding* atau *data mining*, peneliti menyadari proses *cleaning* belum dilakukan dengan sempurna atau mungkin saja peneliti menemukan data atau informasi baru untuk "memperkaya" data yang sudah ada.

2.2 *Data Mining*

Data mining merupakan salah satu solusi yang baik untuk pengalian informasi di *database* dengan ukuran yang besar. Dalam suatu organisasi, perusahaan atau institusi yang mempunyai banyak data, tidak menutup kemungkinan terdapat banyak informasi yang bisa diperolah dengan pemanfatan *data mining. Data mining* membahas penggalian atau pengumpulan informasi yang berguna dari sekumpulan data. Informasi yang biasanya dikumpulkan adalah pola-pola tersembunyi pada data, hubungan antar elemen-elemen data, ataupun pembuatan model untuk keperluan peramalan (Adinugroho & Sari, 2018).

Informasi yang biasanya dikumpulkan merupakan pola-pola yang tersembunyi pada data atau pembuatan model untuk keperluan peramalan data. Perkembangan basis data dan industri manajemen data memiliki beberapa fungsi penting di antaranya adalah data *collection and database creation,* data *management* serta *advance* data *analysis.* Dalam proses *data mining,* peneliti perlu menemukan pengetahuan dalam bentuk pola yang nantinya akan diekstrak menjadi informasi yang akan bermanfaat untuk selanjutnya dilakukan interpretasi terhadap data tersebut. *Data mining* adalah sebuah bidang ilmu yang berupaya menemukan pola, kaidah-kaidah, aturan, dan informasi berharga yang menarik dan belum diketahui sebelumnya dari sekumpulan besar data (Indrawan, 2016).

Berikut definisi dari *data mining* yang dikenal di antaranya:

1. *Data mining* adalah serangkaian proses untuk mengali nilai tambah dari suatu kumpulan data berupa pengetahuan yang selama ini tidak diketahui secara manual.
2. *Data mining* adalah analisa otomatis dari data yang berjumlah besar atau kompleks dengan tujuan untuk menemukan pola atau kecenderungan yang penting yang biasa tidak disadari keberadaanya.
3. *Data Mining* atau *Knowledge Discovery in Database* (KDD) adalah pengambilan informasi yang tersembunyi, di mana informasi tersebut sebelumnya tidak dikenal dan berpotensi bermanfaat.

2.2.1 Manfaat *Data Mining*

Ketersediaan data yang melimpah, kebutuhan akan informasi atau pengetahuan sebagai pendukung pengambilan keputusan untuk membuat solusi bisnis, dan dukungan insfrastruktur dibidang teknologi informasi merupakan cikal-bakal dari lahirnya teknologi *data mining*. Ketersediaan data transaksi dalam jumlah yang besar, dalam bidang-bidang industri yang memiliki data transaksi dalam dalam jumlah besar seperti jaringan ritel, telekomunikasi, perbankan dan lain-lain. Selain itu data yang melimpah juga terdapat pada organisasi politik, institusi pemerintahan dan juga pada bidang pendidikan. Sistem manajemen transaksi pada industri maupun organisasi tersebut menyimpan informasi–informasi rinci yang di perlukan dalam proses bisnis mereka.

Data mining merupakan teknologi baru yang berguna untuk membantu perusahaan-perusahaan menemukan informasi yang sangat penting dari gudang data mereka. Kakas *data mining* meramalkan *trend* dan sifat-sifat perilaku bisnis yang sangat berguna untuk mendukung pengambilan keputusan penting. Analisis yang diotomatisasi oleh *data mining* melebihi yang dilakukan oleh sistem pendukung keputusan tradisional yang sudah banyak digunakan. *Data mining* dapat menjawab pertanyaan-pertanyaan bisnis yang dengan cara tradisional memerlukan waktu untuk menjawabnya. *Data mining* mengeksplorasi basis data untuk menemukan pola-pola yang tersembunyi, mencari informasi prediksi yang mungkin saja terlupakan oleh para pelaku bisnis karena terletak di luar ekspektasi mereka.

Banyak perusahaan yang sudah meluncurkan aplikasi *data mining* dan telah mendapat keuntungan. Teknologi ini tidak hanya cocok untuk digunakan oleh industri-indutri yang mengelola informasi secara intensif seperti perbankan, tetapi juga perusahaan apa saja yang ingin memanfaatkan gudang data untuk manajemen *customer* dengan lebih baik dan pemahaman atau identifikasi yang baik terhadap proses bisnis dimana *data mining* akan diaplikasikan. Beberapa contoh bidang-bidang bisnis yang telah berhasil menerapkan aplikasi *data mining* adalah:

1. Perusahaan farmasi dapat menganalisis aktivitas penjualan terkininya dan menggunakan hasilnya untuk menargetkan dokter-dokter yang berpotensi menggunakan produknya dan menentukan aktivitas pemasaran yang paling efektif untuk beberapa bulan mendatang.

2. Perusahaan kartu kredit dapat memanfaatkan data transaksi pelangan-pelangannya untuk merancang produk kredit baru yang akan menarik minat para pelangan tersebut.
3. Perusahaan transportasi yang menyediakan berbagai jenis pelayanan. *Data mining* dapat digunakan untuk mengidentifikasi prospek-prospek pelayanan yang menjanjijkan pelayanan keuntungan.
4. Perusahaan produk makanan atau kebutuhan sehari-hari. *Data mining* dapat dimanfaatkan untuk meningkatkan penjualan produk ke para pengecer (*retailer*). Data pelangan, pengiriman, aktivitas kompetitor dapat digunakan untuk menganalisis sebab-sebab *customer* perpindah ke produk merek lain. Kemudian, hasilnya dapat digunakan untuk menyusun strategi pemasaran yang lebih efektif.

Pemanfaatan *data mining* pada sudut padang komersial dan keilmuan sebagai berikut: Dilihat dari sudut pandang pemanfaatan *data mining* secara komersial dapat digunakan untuk menangani meledaknya volume data, dengan memakai komputasi dapat digunakan untuk menghasilkan informasi-informasi yang dibutuhkan, yang mana merupakan aset yang dapat meningkatkan daya saing suatu perusahaan (Retno Tri Vulandari, 2017).

Pemanfaatan *data mining* pada sudut pandang komersial, sebagai berikut:

1. Bagaimana mengetahui hilangnya pelangan karena pesaing.
2. Bagaimana mengetahui produk atau konsumen yang memiliki kesamaan karakteristik.

3. Bagaimana mengidentifikasi produk-produk yang terjual bersamaan dengan produk lain.
4. Bagaimana memprediksi keuntungan dan kerugian.
5. Bagaimana melihat resiko dalam menentukan jumlah produksi suatu item.

Dilihat dari sudut pandang keilmuan, *data mining* dapat digunakan untuk *mencapture*, menganalisis serta menyimpan data yang bersifat *real time* dan sangat besar, misalnya:

1. *Remote* sensor yang ditempatkan pada suatu satelit.
2. Prediksi jumlah manusia di bumi yang akan datang.
3. Prediksi *simulasi saintifik* yang membangkitkan data dalam ukuran *terabytes.*

2.2.2 Kategori *Data Mining*

Data mining dibagi manjadi dua bagian utama menurut *Han* dan *Kamber*, 2006 dalam buku (Retno Tri Vulandari, 2017) sebagai berikut:

1. Prediktif, Tujuan dari tugas prediktif adalah untuk memprediksi nilai dari atribut tertentu berdasarkan pada nilai atribut-atribut lain. Atribut yang diprediksi umumnya dikenal sebagai target atau variabel tak bebas, sedangakan atribut-atribut yang digunakan untuk membuat prediksi dikenal sebagai *explanatory* atau variabel bebas.
2. Deskriptif, Tujuan dari tugas deskriptif adalah untuk menurunkan pola-pola (*korelasi*, *trend*, *cluster*, *teritori* dan *anomali*) yang meringkas hubungan yang pokok dalam data. Tugas *data mining* deskriptif sering digunakan

dalam penyelidikan dan seringkali memerlukan teknik *post-processing* untuk validasi dan penjelasan hasil.

2.2.3 Teknik *Data Mining*

Ada beberapa teknik *data mining* yang dapat digunakan untuk menemukan "penemuan" (*discovery*) dan "pembelajaran" (*learning)* yang terbagi dalam tiga metode utama pembelajaran (Retno Tri Vulandari, 2017) yaitu:

1. *Supervised Learning*

   Supervised learning adalah teknik yang paling banyak dipakai. Teknik ini sama dengan "*programming by example*". Teknik ini melibatkan fase pelatihan di mana data pelatihan historis yang karakter-karakternya dipetakan ke hasil-hasil yang telah diketahui diolah dalam algoritma *data mining*. Proses ini melatih *algoritma* untuk mengenali variabel-variabel dan nilai-nilai kunci yang nantinya akan digunakan sebagai dasar dalam membuat perkiraan-perkiraan ketika diberikan data baru.

2. *Unsupervised Learning*

   Teknik pembelajaran ini tidak melibatkan fase pelatihan seperti yang terdapat pada *supervised learning.* Teknik ini tergantung penggunaan algoritma yang mendeteksi semua pola, seperti *associations* dan *sequences,* yang muncul pada kriteria penting yang spesifik dalam data masukan. Pendekatan ini mengarah pada pembuatan banyak aturan (*rules*) yang mengkarakteristikan penemuan *associations, cluster* dan *segments.* Aturan-aturan ini kemudian dianalisis untuk menemukan hal-hal yang penting.

3. *Reinforcement Learning*

Teknik pembelajaran ini jarang digunakan dibandingakan dengan dua teknik lainnya, namun memiliki penerapan-penerapan yang terus dioptimalkan dari waktu ke waktu dan memiliki kontrol adaptif. Teknik ini sangat menyerupai teknik kehidupan nyata yaitu seperti "*on-job-training*", di mana seorang pekerja diberikan sekumpulan tugas yang membutuhkan keputusan-keputusan. Pada beberapa titik waktu kelak diberikan penilaian atas *performance* pekerja tersebut kemudian pekerja diminta mengevaluasi keputusan-keputusan yang telah dibuatnya sehubungan dengan hasil *performance* pekerja tersebut. *Reinforcement learning* sangat tepat digunakan untuk menyelesaikan masalah-masalah yang sulit yang bergantung pada waktu.

2.3 Metode *Data Mining*

Salah satu metode yang ada di *data mining* yaitu klasifikasi. Metode klasifikasi merupakan teknik mempelajari sekumpulan data sehingga menghasilkan aturan yang bisa mengklasifikasi atau mengali data-data baru yang belum pernah dipelajari. Klasifikasi banyak digunakan dalam berbagai aplikasi, di antaranya deteksi kecurangan (*fraud detection*), pengelolaan pelanggan, diagnosis medis, prediksi penjualan, dan sebagainya (Suyanto, 2017). Klasifikasi merupakan salah satu teknik yang sering digunakan dalam *data mining.* Klasifikasi merupakan proses membagi sekumpulan data sehingga setiap data menjadi anggota suatu kategori atau kelas. Keanggotaan data pada setiap kelas

bersifat *mutually exhaustive* dan *mutually exclusive* dimana setiap data hanya dapat menjadi anggota sebuah kelas saja. Sebuah data tidak boleh menjadi anggota lebih dari satu kelas atau tidak menjadi anggota suatu kelas sama sekali.

Klasifikasi dapat dibangun berdasarkan pengetahuan seorang pakar (ahli). Mengingat himpunan data yang sangat besar, model klasifikasi lebih sering dibangun menggunakan teknik pembelajaran dalam bidang *machine learning.* Proses pembelajaran secara otomatis terhadap suatu himpunan data mampu menghasilkan model klasifikasi (fungsi target) yang memetakan objek data *x* (*input*) ke salah satu kelas *y* yang telah didefinisikan sebelumnya. Proses pembelajaran memerlukan masukan (*input*) berupa himpunan data latih (*training set*) yang berlabel (memiliki atribut kelas) dan mengeluarkan *output* yang berupa sebuah model klasifikasi. Terdapat banyak teknik klasifikasi yang telah diusulkan para ahli, yang dapat dikelompokan kedalam dua kategori yaitu teknik klasifikasi global (memperhitungkan semua data latih) dan teknik klasifikasi lokal (hanya memperhitungkan sebagian data latih).

2.3.1 *Decision Tree*

Decision tree atau pohon keputusan adalah pohon yang digunakan sebagai prosedur penalaran untuk mendapatkan jawaban dari permasalahan yang dimasukan. Konsep dari pohon keputusan adalah mengubah data menjadi pohon keputusan dan aturan-aturan keputusan (Retno Tri Vulandari, 2017). Dengan menggunakan pohon keputusan, permasalahan yang ada bisa diidentifikasi dan

bisa dilihat antara hubungan faktor-faktor yang mempengaruhi suatu masalah dan bisa dicari penyelesaian terbaik dengan memperhitungkan faktor-faktor tersebut.

Decision tree adalah salah satu metode klasifikasi yang populer dan banyak digunakan secara praktis (Suyanto, 2017). Pohon yang dibentuk tidak selalu pohon *biner*. Jika semua fitur dalam data set menggunakan 2 macam nilai kategorial maka bentuk pohon yang didapatkan berupa pohon *biner*. Jika dalam fitur berisi lebih dari 2 macam nilai kategorial atau menggunakan tipe *numerik* maka bentuk pohon yang didapatkan biasanya tidak berupa pohon *biner*. Pohon keputusan merupakan salah satu metode klasifikasi yang terkenal. Pohon keputusan adalah salah satu metode klasifikasi yang paling populer karena mudah untuk diinterpretasikan oleh manusia (Retno Tri Vulandari, 2017).

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

Decision tree menggunakan struktur data *tree* sebagai model dalam proses penentuan kelas dari suatu data. Terdapat tiga jenis *node* pada *decision tree*:

1. *Root node,* merupakan *node* yang tidak memiliki *edge* masukan dan memiliki nol atau lebih *edge* keluaran.
2. *Internal node*, memiliki tepat satu *edge* masukan dan memiliki dua atau lebih *edge* keluaran.

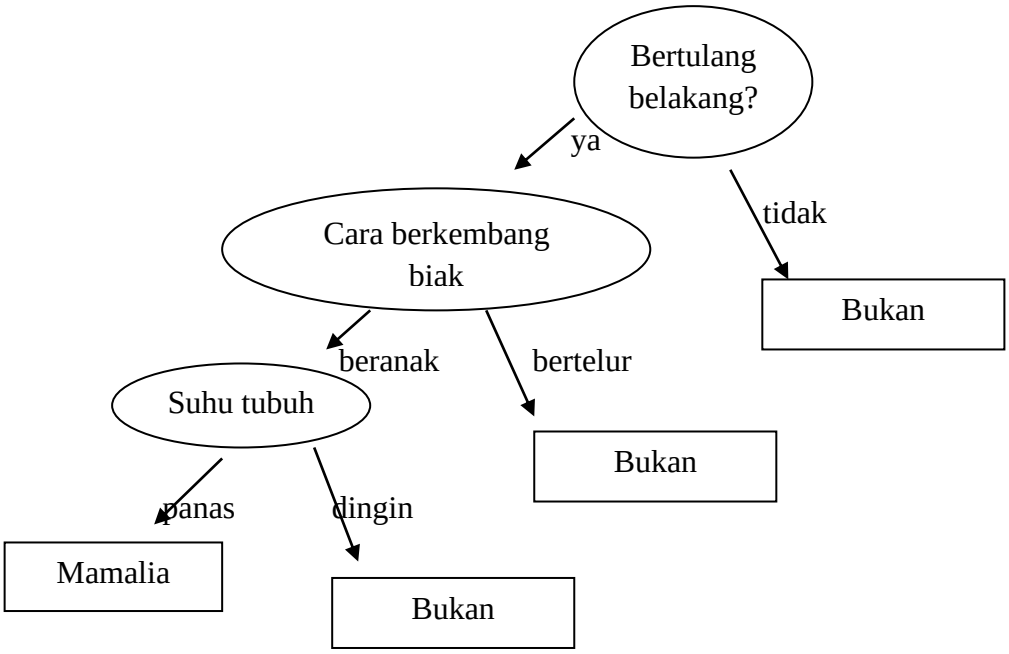
3. *Leaf* atau *terminal node*, mempunyai tepat satu *edge* masukan dan tidak mempunyai *edge* keluaran.

Pada sebuah *decision tree*, setiap *leaf* memiliki sebuah nama kelas. *Root node* dan *internal node* berisi aturan kondisional yang digunakan untuk memisahkan data yang memiliki karateristik berbeda. Penyusunan suatu *decision tree* sederhana dapat dilakukan dengan memisahkan karakteristik yang berbeda dari suatu kelas terhadap kelas lainnya.

Contoh karakteristik suatu hewan berjenis mamalia adalah sebagai berikut:

1. Memiliki tulang belakang.
2. Berkembang biak dengan beranak.
3. Berdarah panas.

Berdasarkan karateristik tersebut, sebuah *decision tree* seperti pada gambar di bawah ini, dapat disusun untuk mengklasifikasikan apakah sebuah binatang termasuk kedalam kelas mamalia atau non-mamalia. Pembentukan *decision tree* dengan memisahkan kelas secara manual berdasarkan karakteristiknya mungkin tidak dapat dilakukan pada semua kasus. Jika karakteristik kelas-kelas pada data belum diketahui maka pembuatan *decision tree* secara manual sulit dilakukan.



Gambar 2.2 *Decision Tree* Untuk Klasifikasi Hewan
Sumber: (Adinugroho & Sari, 2018)

1. Kelebihan pohon keputusan

   Kelebihan dari pohon keputusan adalah sebagai berikut:

   a. Daerah pengambilan keputusan yang sebelumnya kompleks dan sangat global, dapat diubah menjadi lebih *simple* dan spesifik.
   b. Eliminasi perhitungan-perhitungan yang tidak diperlukan, karena ketika menggunakan metode pohon keputusan maka sampel diuji hanya berdasarkan kriteria atau kelas tertentu.
   c. Fleksibel untuk memilih fitur dari internal *node* yang berbeda, fitur yang dipilih akan membedakan suatu kriteria dibandingakan kriteria yang lain dalam *node* yang sama.

d. Dalam analisis *multivariat*, dengan kriteria dan kelas yang jumlahnya sangat banyak, seorang penguji biasanya perlu untuk mengestimasikan baik itu distribusi dimensi tinggi ataupun parameter tertentu dari distribusi kelas tersebut.

2. Kekurangan pohon keputusan
   a. Terjadi *overlap* terutama ketika kelas-kelas dan kriteria yang digunakan jumlahnya sangat banyak. Hal tersebut juga dapat menyebabkan meningkatnya waktu pengambilan keputusan dan jumlah memori yang diperlukan.
   b. Pengakumulasian jumlah *eror* dari setiap tingkat dalam sebuah pohon keputusan yang besar.
   c. Kesulitan dalam mendesain pohon keputusan yang optimal.
   d. Hasil kualitas keputusan yang didapatkan dari metode pohon keputusan sangat tergantung pada bagaimana pohon tersebut didesain.

2.3.1.1 Algoritma C4.5

Algoritma C4.5 merupakan salah satu algoritma yang ada di *decision tree* dan algoritma yang banyak digunakan untuk menghasilkan *decision tree*. Dikembangkan oleh Ross Quinlan sebagai pengembangan dari algoritma ID3. Pembentukan *tree* pada algoritma C4.5 menganut pendekatan *top-down* di mana *tree* dibentuk dari *root* menuju *leaf*, algoritme C4.5 bersifat rekursif, dengan tahapan sebagai berikut (Adinugroho & Sari, 2018):

1. Cek apakah ada kondisi berhenti yang terpenuhi.

2. Carilah variabel yang paling optimal untuk membagi data C4.5 menggunakan parameter *gain ratio* untuk memilih variabel yang digunakan untuk membagi data.
3. Bagilah data latih S menjadi *S1,S2,S3,…* menggunakan variabel yang telah dipilih sebelumnya.
4. Ulangi langkah No.1 untuk setiap *S1,S2,S3,…*
5. Jika semua data telah berada pada *decision tree,* lakukan *pruning.*

Kondisi berhenti pada pembuatan *decision tree* adalah:

1. Semua data pada data latih memiliki kelas yang sama. Pada kasus ini, bentuklah sebuah *leaf* yang berisi kelas tersebut.
2. Data latih kosong (tidak memiliki isi). Jika pada kondisi ini penbentukan *tree* dihentikan.
3. Suatu variabel tidak memiliki nilai (kosong). Jika kondisi ini terjadi, buatlah sebuah *leaf* yang berisi kelas yang paling banyak muncul.

Algoritma C4.5 menggunakan parameter *gain ratio* untuk memilih variabel mana yang akan digunakan untuk membentuk cabang pada *decision tree*. *Gain ratio* dapat dihitung dengan persamaan berikut:

$$Gain\ ratio_{split} = \frac{Gain_{split}}{SplitInfo}$$

Rumus 2.1 *Gain rasio*$_{split}$

Dengan:

$$Gain_{split} = Entropy(p) - \left(\sum_i^k \frac{n_i}{n} Entropy(i)\right)$$

Rumus 2.2 *gain split*

Keterangan:

p = *parent node* yang dibagi menjadi *K* partisi

n_i = jumlah data pada partisi ke-*i*

Entropy pada *node t* dapat dihitung menggunakan persamaan berikut:

$$Entropy(t) = -\sum_j p(j|t) log_2 p(j|t)$$

Rumus 2.3 *Entropy*

Keterangan:

p p(*j*|*t*) = frekuensi relatif kelas *j* pada *node t* penghitungan *SplitInfo* dilakukan menggunakan persamaan sebagai berikut:

$$SplitInfo = -\sum_{i=1}^{k} \frac{n_i}{n} log_2 \frac{n_i}{n}$$

Rumus 2.4 *SplitInfo*

Keterangan:

P = *parent node* yang dibagi menjadi *k* partisi

N_i = jumlah data pada partisi ke – *i*

Berikut ini contoh perhitungan menggunakan dataset *weather.nominal.arff*. Data tersebut merupakan data bawaan yang tersedia ketika Weka di *instal*, karakteristik data tersebut ditunjukkan seperti tabel berikut:

Tabel 2.1 Struktur Data *Weather.Nominal.Arff*

No	Nama Variabel	Tipe Variabel	Nilai
1	*outlook*	*nominal*	{*sunny, overcast, rainy*}
2	temperatur	*nominal*	{*hot, mild, cool*}
3	*humidity*	*nominal*	{*high,* normal}
4	*windy*	*nominal*	{*TRUE, FALSE*}
5	*paly*	*Nominal, label*	{*yes, no*}

Tabel 2.2 Contoh Isi Dataset *Weather.Nominal.Arff*

No	***Outlook***	***Temperature***	***Humidity***	***Windy***	***Play***
1	*sunny*	*hot*	*high*	*FALSE*	*no*
2	*sunny*	*hot*	*high*	*TRUE*	*no*
3	*overcast*	*hot*	*high*	*FALSE*	*no*
4	*rainy*	*mild*	*high*	*FALSE*	*no*
5	*rainy*	*cool*	*normal*	*FALSE*	*no*
6	*rainy*	*cool*	*normal*	*TRUE*	*no*
7	*overcast*	*cool*	*normal*	*TRUE*	*no*
8	*sunny*	*mild*	*high*	*FALSE*	*no*
9	*sunny*	*cool*	*normal*	*FALSE*	*yes*
10	*rainy*	*mild*	*normal*	*FALSE*	*yes*
11	*sunny*	*mild*	*normal*	*TRUE*	*yes*

Tabel 2.2 lanjutan

12	*overcast*	*mild*	*high*	*TRUE*	*yes*
13	*overcast*	*hot*	*normal*	*FALSE*	*yes*
14	*rainy*	*mild*	*high*	*TRUE*	*no*

Dataset *weather.nominal.arff* memiliki empat variabel yang mempengaruhi nilai variabel *play*. Oleh karena itu, terdapat empat kemungkinan pembentukan cabang pertama pada *decision tree* menggunakan *algoritma* C4.5, yaitu percabangan menggunakan variabel *outlook, temperature, humidity,* atau *windy*. Pemilihan percabangan yang paling tepat ditentukan oleh nilai *gain ratio* terbesar dari keempat variabel yang ada. Perhitungan *gain ratio* pada variabel adalah sebagai berikut:

1. *Gain ratio* variabel *outlook*

   a) *Entropy parent*

   Entropy parent dihitung berdasarkan peluang dari setiap nilai variabel *outlook.* Peluang masing-masing nilai *outlook* adalah:

   - *P* (*sunny*) = $\frac{5}{14}$
   - *P (overcast)* = $\frac{4}{14}$
   - *P (rainy)* = $\frac{5}{14}$

   Nilai *entropy parent* diperoleh dari persamaan:

   Entropy (parent) = − p(sunny)log₂p(sunny) – P(overcast)log₂p(overcast) –

   P(rainy)log₂p(rainy)

Entropy (parent) = $-\frac{5}{14}log_2\frac{5}{14}-\frac{4}{14}log_2\frac{4}{14}-\frac{5}{14}log_2\frac{5}{14}$ = 1.5774

b) *Entropy (outlook = sunny)*

Entropy pada *child* dengan *outlook* = *sunny* diperoleh berdasarkan persamaan *p (play|outlook = sunny)* sebagai berikut:

- *P (play = yes|outlook = sunny)* = $\frac{2}{5}$
- *P (play) = no|outlook = sunny)* = $\frac{3}{5}$

Entropy (outlook = sunny) = $-\frac{2}{5}log_2\frac{2}{5}-\frac{3}{5}log_2\frac{3}{5}=0{,}9710$

c) *Entropy (outlook = overcast)*

- *P (play = yes|outlook = overcast)* = $\frac{4}{4}$
- *P (play = no|outlook = overcast) = 0*

Entropy (outlook = overcast) = $-\frac{4}{4}log_2\frac{4}{4}=0$

d) *Entropy (outlook = rainy)*

- *P (play = yes|outlook = rainy)* = $\frac{3}{5}$
- *P (play = no|outlook = rainy)* = $\frac{2}{5}$

Entropy (outlook = rainy) = $-\frac{3}{5}log_2\frac{3}{5}-\frac{2}{5}log_2\frac{2}{5}=0{,}9710$

e) *Gain (outlook)*

$Gain_{outlook}$ = *entropy (p)* $(\sum_i^k \frac{n_i}{n}Entropy(i))$

= $1.5774-\frac{5}{14}0.9710-\frac{4}{14}0-\frac{5}{14}\ 0.9710$ = 0.8838

f) *SplitInfo (outlook)*

$$SplitInfo = -\sum_{i-1}^{k} \frac{n_i}{n} log \frac{n_i}{n}$$

$$= -\frac{n_{sunny}}{n} log_2 \frac{n_{sunny}}{n} - \frac{n_{overcast}}{n} log_2 \frac{n_{overcast}}{n} - \frac{n_{rainy}}{n} log_2 \frac{n_{rainy}}{n}$$

$$= -\frac{5}{14} \log_2 \frac{5}{14} - \frac{4}{14} \log_2 \frac{4}{14} - \frac{5}{14} \log_2 \frac{5}{14} = 1.5774$$

g) *Gain Ratio (outlook)*

$$GainRatio_{outlook} = \frac{Gain_{outlook}}{SplitInfo}$$

$$= \frac{0.8838}{0.5603} = 0.5603$$

2. *Gain Ratio* Variabel *Temperature*

a) Entropy parent

Entropy parent dihitung berdasarkan peluang dari setiap nilai variabel *temperature*. Peluang masing-masing nilai *temperature* adalah:

- *p (hot)* = $\frac{4}{14}$
- *p (mild)* = $\frac{6}{14}$
- *P (cool)* = $\frac{4}{14}$

Nilai *Entropy parent* dari persamaan:

Entropy (parent) = – p(hot)log₂p(hot) – P(mild)log₂p(mild) – P(cool)log₂p(cool)

$$Entropy\ (parent) = -\frac{4}{14} \log_2 \frac{4}{14} - \frac{6}{14} \log_2 \frac{6}{14} - \frac{4}{14} \log_2 \frac{4}{14} = 1.5567$$

b) *Entropy (temperature = hot)*

Entropy pada *child* dengan diperoleh berdasarkan persamaan *p(play|temperature = hot)* sebagai berikut:

- *P (play = yes|temperature = hot)* = $\frac{2}{4}$
- *P (play = no|temperature = hot)* = $\frac{2}{4}$

Entropy (temperature = hot) = $-\frac{2}{4}\log_2\frac{2}{4}-\frac{2}{4}\log_2\frac{2}{4}=1$

c) *Entropy (temperature = mild)*

- *P (play = yes|temperature = mild)* = $\frac{4}{6}$
- *P (play = no|temperature = mild)* = $\frac{2}{6}$

Entropy (temperature = mild) = $-\frac{4}{6}\log_2\frac{4}{6}-\frac{2}{6}\log_2\frac{2}{6}=0.9183$

d) *Entropy (temperature = cool)*

- *P (play = yes|temperature = cool)* = $\frac{3}{4}$
- *P (play = no|temperature = cool)* = $\frac{1}{4}$

Entropy (temperature = cool) = $-\frac{3}{4}\log_2\frac{3}{4}-\frac{1}{4}\log_2\frac{1}{4}=0.8113$

e) *Gain (temperature)*

$= 1.5567 - \frac{4}{14}\ 1\ \frac{6}{14}0.9183 - \frac{4}{14}\ 0.8113 = 1.5567$

f) *SplitInfo (temperature)*

$= -\frac{n_{hot}}{n}\log_2\frac{n_{hot}}{n}-\frac{n_{mild}}{n}\log_2\frac{n_{mild}}{n}-\frac{n_{cool}}{n}\log_2\frac{n_{cool}}{n}$

$= -\frac{4}{14}\log_2\frac{4}{14} - \frac{6}{14}\log_2\frac{6}{14} - \frac{4}{14}\log_2\frac{4}{14} = 1.5567$

g) GainRatio (temperature)

$GainRatio_{temperature} = \frac{Gain_{temperature}}{SplitInfo}$

$= \frac{0.6151}{1.5567} = 0.3957$

3. *Gain Ratio* variabel *humidity*

a) Entopy parent

Entropy parent dihitung berdasarkan peluang dari setiap nilai variabel *humidity.* Peluang masing-masing adalah:

- *P (high)* $= \frac{7}{14}$
- *P (normal)* $= \frac{7}{14}$

Nilai *entropy parent* diperoleh dari persamaan:

Entropy (parent) = − (high)log₂p(high) – P(normal)log₂p(normal)

Entropy (parent) $= -\frac{7}{14}\log_2\frac{7}{14} - \frac{7}{14}\log_2\frac{7}{14} = 1$

b) *Entopy (humidity = high)*

- *P (play = yes|humidity = high)* $= \frac{3}{7}$
- *P (play = no|humidity = high)* $= \frac{4}{7}$

Entropy (humidity = high) $= -\frac{3}{7}\log_2\frac{3}{7} - \frac{4}{7}\log_2\frac{4}{7} = 0.9852$

c) *Entopy (humidity = normal)*

- *P (play = yes|humidity = normal)* = $\frac{6}{7}$
- *P (play = no|humidity = normal)* = $\frac{1}{7}$

Entropy (humidity = normal) = $-\frac{6}{7}\log_2\frac{6}{7}-\frac{1}{7}\log_2\frac{1}{7}$ = 0.5917

d) Gain (humidity)

= $1-\frac{7}{14}$ 0.9852 $-\frac{7}{14}$0.5917 = 0.2116

e) *SplitInfo (humidity)*

= $-\frac{n_{high}}{n}\log_2\frac{n_{high}}{n}-\frac{n_{high}}{n}\log_2\frac{n_{high}}{n}$

= $-\frac{7}{14}\log_2\frac{7}{14}-\frac{7}{14}\log_2\frac{7}{14}$ = 1

f) *GainRatio (humidity)*

GainRatio$_{humidity}$ = $\frac{Gain_{humidity}}{SplitInfo}$

= $\frac{0.2116}{1}$ = 0.2116

4. *Gain Ratio* variabel *windy*

a) Entropy parent

Entropy parent dihitung berdasarkan peluang dari setiap nilai variabel *windy*. Peluang masing-masing nilai *windy* adalah:

- *P (TRUE)* = $\frac{6}{14}$
- *P (FALSE)* = $\frac{8}{14}$

Nilai *entropy parent* diperoleh dari persamaan:

Entropy (parent) = − p(TRUE)log₂p(TRUE) − p(FALSE)log₂p(FALSE)

Entropy (parent) = $-\frac{6}{14}\log_2\frac{6}{14}-\frac{8}{14}\log_2\frac{8}{14}$ = 0.9852

b) *Entropy (windy = TRUE)*

- *P (play = yes|windy = TRUE)* = $\frac{3}{6}$
- *P (play = no|windy = TRUE)* = $\frac{3}{6}$

Entropy (windy = TRUE) = $-\frac{3}{6}\log_2\frac{3}{6}-\frac{3}{6}\log_2\frac{3}{6}$ = 1

c) *Entropy (windy = FALSE)*

- *P (play = yes|windy = FALSE)* = $\frac{6}{8}$
- *P (play = no|windy = FALSE)* = $\frac{2}{8}$

Entropy (windy = FALSE) = $-\frac{6}{8}\log_2\frac{6}{8}-\frac{2}{8}\log_2\frac{2}{8}$ = 0.8113

d) *Gain (windy)*

= $0.9852-\frac{6}{14}1-\frac{8}{14}0.8113$

e) *SplitInfo* (*windy*)

= $-\frac{n_{TRUE}}{n}\log_2\frac{n_{TRUE}}{n}-\frac{n_{FALSE}}{n}\log_2\frac{n_{FALSE}}{\mathrm{n}}$

= $-\frac{6}{14}\log_2\frac{6}{14}-\frac{8}{14}\log_2\frac{8}{14}$ = 0.9852

f) *GainRatio (windy)*

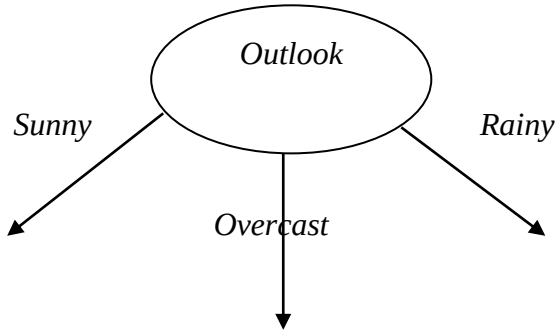
$GainRatio_{windy} = \frac{Gain_{windy}}{SplitInfo}$

$= \frac{0.0930}{0.9854} = 0.0944$

Nilai *Gain ratio* pada setiap variabel ditampilkan pada tabel 2.3 di bawah ini. Nilai *Gain ratio* terbesar pada variabel *Outlook*, sehingga percabangan pertama pada proses pembentukan *tree* dilakukan pada variabel *Outlook.* Proses pembentukan cabang kedua dan seterusnya menggunakan proses yang sama, yaitu berdasarkan variabel dangan nilai *Gain ratio* terbesar. Pada pembentukan cabang kedua, nilai *Gain ratio* dari variabel *Outlook* tidak lagi dihitung.

Tabel 2.3 Nilai *Gain Ratio* Pada Setiap Variabel

No	**Variabel**	***GainRatio***
1.	*Outlook*	0.5603
2.	*Temperature*	0.3951
3.	*Humidity*	0.2116
4.	*Windy*	0.0944



Gambar 2.3 Pembentukan Cabang Pertama Pada *Tree*
Sumber: (Adinugroho & Sari, 2018)

2.3.1.2 *Entropy*

Entropy adalah keberbedaan atau keberagaman. Dalam *data mining*, *entropy* didefinisikan sebagai suatu parameter untuk mengukur *heterogenitas* (keberagaman) dalam suatu himpunan data. Semakin heterogen suatu himpunan data, semakin besar pula nilai *entropyny*-nya, secara sistematis, *entropy* dirumuskan sebagai berikut:

$$Entropy(S) \equiv \sum_{i}^{c} -p_i log_2 p_i$$

Rumus 2.5 *Entropy*

Di mana *c* adalah jumlah nilai yang terdapat pada atribut target (jumlah kelas). Sedangkan p_i menyatakan porsi atau *rasio* antara jumlah sampel di kelas *I* dengan jumlah semua sampel pada himpunan data. Berdasarkan formula di atas, himpunan data yang memiliki dua kelas dengan jumlah sampel di kelas pertama sama persis dengan jumlah sampel di kelas kedua akan memiliki *Entropy* yang

maksimum (yaitu sama dengan 1). Artinya, himpunan data tersebut memiliki keberagaman maksimum. Sebaliknya, himpunan data yang memiliki dua kelas dengan jumlah sampel pada salah satu kelas adalah 0 akan memiliki *Entropy* yang minimum (yaitu sama dengan 0). Artinya, himpunan data tersebut memiliki keberagaman minimum.

2.3.1.3 *Information Gain*

Information Gain adalah perolehan informasi. Dalam *data mining*, *information gain* didefinisikan sebagai ukuran efektifitas suatu atribut dalam mengklasifikasikan data secara sistematis, *information gain* dari suatu atribut *A*, dituliskan sebagai berikut:

$$Gains(S, A) \equiv Entropy(S) - \sum_{v \in values(A)} \frac{|s_v|}{|s|} Entropy(s_v)$$

Rumus 2.6 *Information gain*

Keterangan:

A : atribut

V : menyatakan suatu nilai yang mungkin untuk atribut A

Value (A) : himpunan nilai-nilai yang mungkin untuk atribut

$|S_v|$: jumlah sampel untuk nilai *v*

$|S|$: jumlah seluruh sampel data

Entropy (S*v*) : *entropy* untuk sampel-sampel yang memiliki nilai *v*

2.3.1.4 Algoritma ID3

ID3 didesain untuk himpunan data dengan atribut yang bertipe kategorial. Untuk himpunan data dengan atribut yang bernilai *numerik*, anda bisa menggunakan teknik diskritisasi data untuk mengubah nilai *numerik* menjadi kategorial. Kemudian melatihkan data semua atributnya sudah bertipe kategorial ke dalam *algoritma* ID3. Cara lainnya, anda dapat menggunakan data *numeric*, salah satunya adalah *algoritma* C4.5. *Algoritma* ini merupakan *decision tree learning* yang melakukan pancarian secara rakus (*greedy*) sehingga belum tentu optimal (Suyanto, 2017).

***Function** ID3* (Kumpulan Sampel, Atribut Target, Kumpulan Atribut)

1. Buat simpul *Root*
2. **if** semua sampel adalah kelas *i* ***then return*** pohon satu simpul *Root* dengan
3. label = *i*
4. **if** KumpulanAtribut kosong ***then return*** pohon satu simpul *Root* dengan

   label = nilai atribut label target yang paling umum (yang paling sering muncul)

   Else

   A ← Atribut yang merupakan *the best classifer* (dengan *information gain* terbesar) Atribut keputusan untuk *Root* ← A

   For v_i (setiap nilai pada A)

   Tambahkan suatu cabang di bawah *Root* sesuai dengan

nilai v_i buat suatu variabel, misalnya *Sampel* $_{vi,}$ sebagai himpunan bagian (subset) dari KumpulanSampel yang bernilai v_i pada atribut A

If sampel$_{vi}$ kosong

Then dibawah cabang ini tambahkan suatu simpul daun

(*leaf node*, simpul yang tidak punya anak di bawahnya) dengan label = nilai

atribut target yang paling umum (yang paling sering muncul)

Else di bawah cabang ini tambahkan *subtree* dengan memanggil

fungsi ID3(sampel$_{vi}$, atributTarget, Atribut-{A})

End

End

End

5. **Return** *Root*

2.4 *Software* Pendukung

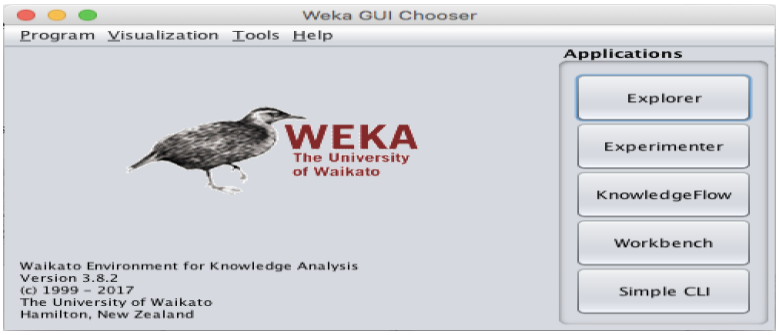
Software yang dipakai dalam penelitian ini adalah *software* Weka 3.8.3. Weka merupakan *software* terintegrasi yang berisi implementasi dari metode-metode *data mining*. Weka dikembangkan oleh Universitas Waikato, Selandia Baru menggunakan bahasa pemprograman java. Weka merupakan singkatan dari *Waikato Environment For Knowledge Analysis.*

Dengan mengadopsi konsep *open source software,* menjadikan Weka dapat digunakan dan dimodifikasi siapapun secara gratis. Weka memiliki keunggulan jika dibandingkan dengan perangkat lunak *data mining* lainnya. Penggunaan Weka murah karena aplikasi tersebut berlisensi GNU *General Public License*, yang artinya dapat digunakan secara gratis. Penggunaan bahasa java dalam pengembangan Weka menyebabkan Weka dapat di *instal* pada hampir semua sistem operasi modern, sepanjang sistem operasi tersebut mendukung *Java Virtual Machine.*

Weka adalah sebuah paket *tools maching learning* praktis (Retno Tri Vulandari, 2017). Berbagai macam algoritme *data mining,* mulai dari pemprosesan awal sampai dengan permodelan data, telah disertakan dalam Weka sehingga memudahkan pengguna dalam menganalisis data. Apabila algoritme yang akan digunakan tidak tersedia pada Weka, pengguna dapat menambahkan algoritme tersebut melalui melalui bahasa pemprograman java. Penggunaan Weka tergolong mudah karena telah dibekali dengan antarmuka grafis (*Graphical User Interface*) sehingga pengguna dapat menggunakan tanpa perlu menulis satu baris kode pun.

Kehadiran antarmuka grafis memudahkan pengguna dalam berinteraksi dengan Weka. Weka menyediakan tiga antarmuka grafis untuk mengolah data. Tampilan pertama adalah *Explorer* yang menyediakan akses ke semua fungsi-fungsi pada Weka melalui menu-menu yang urut dan mudah digunakan (Adinugroho & Sari, 2018). Fungi-fungsi yang mudah digunakan melalui *Explorer* adalah:

1. *Preprocessing,* merupakan panel yang digunakan untuk memilih data yang akan diproses. Data dapat berbentuk berkas ARFF, *database,* maupun data buatan yang dibangkitkan dengan pola tertentu. Pemilihan variabel atau fitur yang akan digunakan juga dapat dilakukan pada panel ini.
2. *Classifier*, merupakan panel untuk menggunakan algoritme klasifikasi atau regresi terhadap data yang telah dimasukan di panel *preprocessing*. Terhadap bermacam-macam algoritme klasifikasi dan regresi yang telah disediakan dan siap digunakan. Terdapat pula sejumlah metode evaluasi yang dapat digunakan untuk mengevaluasi hasil klasifikasi.
3. *Cluster,* merupakan panel yang memberikan akses keberbagai metode *clustering* yang disediakan oleh Weka. *Panel* ini memiliki beberapa opsi pembagian data menjadi data latih dan data uji untuk keperluan pengujian hasil *clustering*.
4. *Associate,* menyediakan algoritme untuk *association rule* yang berguna untuk mengenali relasi antar item pada data.
5. *Select attribute,* menyediakan berbagai algoritme untuk memilih atribut atau variabel yang paling relevan untuk suatu permasalahan.



Gambar 2.4 Gambar Antarmuka Weka
Sumber: (Adinugroho & Sari, 2018)

Antarmuka *KnowledgeFlow* menyuguhkan fungsi yang sama dengan *Explorer* yaitu menggunakan berbagai algoritme *data mining* yang telah disediakan di Weka dengan data yang dimasukan oleh pengguna. Bedanya proses yang dilakukan secara visual. Pengguna dapat memilih komponen-komponen yang tersedia dan meletakannya pada kanvas. Selanjutnya setiap komponen disambungkan dengan menggunakan *graph* berarah untuk membentuk rangkaian proses dan analisis data. *KnowledgeFlow* sangat berguna untuk memvisualisasikan aliran data dan proses.

Experimenter merupakan *interface* kedua pada Weka GUI yang didesain untuk melakukan eksperimen. Eksperimen yang di maksud di sini adalah membandingkan *algoritme-algoritme* klasifikasi yang tersedia di Weka pada data yang berbeda. Kinerja *algoritme* yang dibandingkan dapat dipilih, misalnya akurasi, nilai, *Kappa*, *Mean absolute error*, waktu eksekusi dan lain-lain. Dengan melakukan perbandingan, *algoritme* yang paling sesuai untuk data dapat diketahui.

Gambar 2.5 Gambar Antarmuka *KnowledgeFlow*
Sumber: (Adinugroho & Sari, 2018)

Format data yang digunakan dalam Weka adalah *flat*, ARFF karena Weka perlu mengetahui beberapa informasi tentang tiap atribut yang tidak dapat disimpulkan secara otomatis dari nilai-nilainya. *File* ARFF (*Attribute-Relation File Format*) adalah sebuah *file* teks ASCII yang berisi daftar *instances* dalam sekumpulan atribut. Cara mengubah data ke format ARFF:

1. Apabila data awal berformat .xls buka data tersebut dari *Microsoft Excel* dan simpan sebagai .csv.
2. Buka *file* tersebut dari *Microsoft Word, notepad* atau *editor* teks lainnya dan data sudah berubah dalam format *comma-separated.*
3. Sesuaikan data tersebut dangan menambahkan informasi awal dan data tersebut sudah bisa digunakan sebagai inputan dalam Weka.

2.5 Penelitian Terdahulu

Penelitian terdahulu ini menjadi salah satu acuan penulis dalam melakukan penelitian, sehingga penulis dapat memperkaya teori yang digunakan dalam mengkaji penelitian yang dilakukan. Dari penelitian terdahulu, peneliti tidak menemukan penelitian dengan judul yang sama seperti judul peneliti angkat. Namun peneliti mengangkat beberapa penelitian sebagai referensi dalam memperkaya bahan kajian pada penelitian ini. Berikut merupakan penelitian terdahulu berupa beberapa jurnal terkait penelitian yang dilakukan.

1. Menurut penelitian yang dilakukan oleh (Ruslan, 2016). Dengan judul "**Prediksi Jumlah Penduduk Provinsi Kalimantan Selatan Menggunakan Metode Averange**", ISSN: 2461 0690. Penelitian ini menggunakan metode Semi *Average* sebagai metode penghitungan untuk mengetahui nilai-nilai prediksi. Dari hasil penelitian maka dapat mengambil kesimpulan bahwa analisis prediksi menggunakan Metode *Semi Average* dapat dipergunakan untuk memprediksi perolehan jumlah penduduk periode yang akan datang berdasarkan data penduduk tahun sebelumnya karena menghasilkan hasil yang mendekati kebenaran. Dari hasil uji coba 3 tahun terakhir menunjukkan validitas Metode *Semi Average* adalah 98,34% sehingga dinyatakan valid.
2. Menurut penelitian yang dilakukan oleh (Jayanti, 2017). Dengan judul "**Hubungan Pertumbuhan Penduduk Dengan Tujuan Pembangunan Berkelanjutan di Sumatera**", ISSN: 2549-8355. Penelitian ini

menggunakan metode deskriptif dan kualitatif untuk menganalisa hubungan pertumbuhan penduduk dengan pembangunan berkelanjutan. Alat analisis yang digunakan adalah pendekatan Model *Panel Square* (PLS), pada *fixed effect* Model. Penelitian ini menggunakan pertumbuhan penduduk terdiri dari tingkat kelahiran alamiah dan pertumbuhan migrasi, sementara pembangunan berkelanjutan berkaitan dengan variabel degradasi lingkungan, seperti penutupan lahan, penyusutan hutan, polusi air, ketersediaan air bersih. Hasil dari penelitian ini menunjukkan bahwa secara simultan variabel dependen berpengaruh terhadap semua variabel independen.

3. Penelitian yang dilakukan oleh (Harahap, 2014). Yang berjudul "**Analisis Pertumbuhan dan Persebaran Penduduk Provinsi Sumatera Utara Berdasarkan Hasil Sensus Penduduk Tahun 2010**", ISSN: 2407-7429. Metode yang digunakan adalah analisa data sekunder yang bersifat deskriptif dengan pendekatan keruangan. Sebagai objek penelitian adalah jumlah penduduk, pertumbuhan penduduk, dan persebaran penduduk Sumatera Utara hasil Sensus penduduk tahun 2010. Hasil penelitian menunjukkan bahwa pertumbuhan penduduk Provinsi Sumatera Utara dari tahun 2000 sampai dengan tahun 2010 sebesar 1,22% per tahun.
4. Penelitian yang dilakukan oleh (Suartha, 2016). Yang berjudul "**Faktor-Faktor Yang Mempengaruhi Tingginya Laju Pertumbuhan dan Implementasi Kebijakan Penduduk di Provinsi Bali**", ISSN: 1907-3275. Faktor-faktor yang mempengaruhi laju pertumbuhan penduduk setelah

diadakan analisis dengan pendekatan penelitian diskriptif kuantitatif dan kualitatif dengan analisis statistik dan teori kebijakan publik dari Edward III; dimana obyek penelitian di seluruh Kabupaten/Kota di Provinsi Bali. Laju pertumbuhan penduduk ditentukan oleh perubahan dinamika kependudukan seperti kelahiran, kematian, dan migrasi di Provinsi Bali.

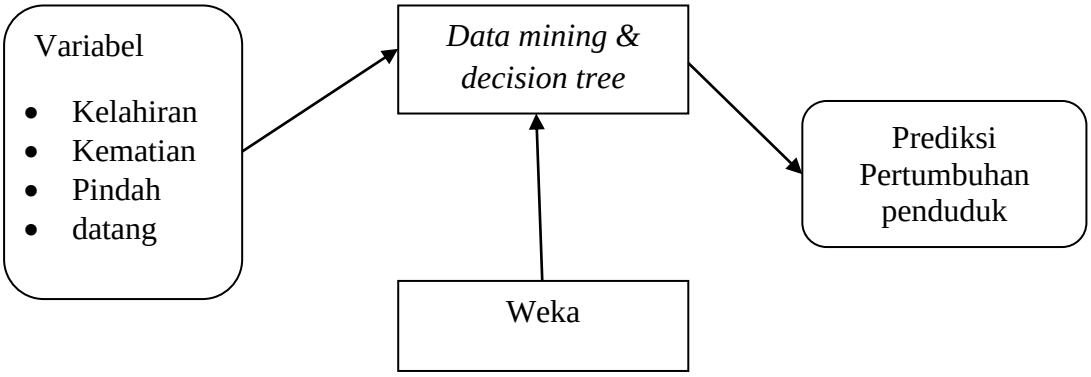
5. Penelitian yang dilakukan oleh (Kamagi & Hansun, 2018) yang berjudul "**Implementasi *Data Mining* Dengan Algoritma C4.5 Untuk Memprediksi Tingkat Kelulusan Mahasiswa**", ISSN: 2085-4552. Data training yang akan digunakan oleh peneliti adalah data alumni mahasiswa program studi Teknik Informatika Universitas Multimedia Nusantara angkatan 2007 dan 2008, sedangkan untuk data testing akan digunakan data alumni angkatan 2009. Dari kumpulan data training dan data testing, dapat diketahui informasi kelulusan yang dapat mempengaruhi beberapa keputusan program studi menggunakan data mining dengan algoritma C4.5. Data mining dengan algoritma C4.5 dapat diimplementasikan untuk memprediksi tingkat kelulusan mahasiswa dengan empat kategori yaitu lulus cepat, lulus tepat, lulus terlambat dan drop out. Attribute yang paling berpengaruh dalam hasil prediksi adalah IPS semester enam.
6. Penelitian yang dilakukan oleh (Tenan, Tavecchia, Oro, & Pradel, 2019). Yang berjudul "***Assesing the Effect Of Density On Population Growth When Modeling Individual Encounter Data***" *https://doi.org/10.6084/m9. The relative role of density-dependent and density-independent variation in vital rates and population size remains largely unsolved. Despite its*

importance to the theory and application of population ecology, and to conservation biology, quantifying the role and strength of density dependence is particularly challenging. A measure of relative population size is built in the model and serves to detect density dependence directly on population growth rate. More generally, we use this modeling framework along with simulated and empirical data to show the value of including density dependence when modeling individual encounter data without the need for auxiliary data.

2.6 Kerangka Pemikiran

Dalam proses membuat kerangka pemikiran, peneliti akan menentukan beberapa indikator yang akan dipakai sebagai masukan (*input*) untuk diproses dengan aplikasi Weka. Beberapa indikator peneliti gunakan untuk penelitian ini yaitu, data penduduk berupa data kelahiran, migrasi masuk, migrasi keluar dan data kematian dari tahun 2014 sampai 2018. Jika peneliti melakukan sebuah penelitian dengan melibatkan sebuah masukan, maka hasil yang peneliti sebut adalah *output.* Dengan pemanfaatan ilmu *data mining* dan latar belakang masalah yang dikemukakan, maka peneliti memiliki pemikiran bahwa dengan mengetahui berapa jumlah pertumbuhan penduduk yang akan datang pemerintah maupun badan terkait bisa mempersiapkan hal-hal yang mungkin menjadi kebutuhan dari pertumbuhan penduduk tadi.

Kerangka pemikiran merupakan dasar dari penelitian bagaimana fakta data, wawancara dan kajian kepustakaan diolah dengan baik dan benar untuk untuk mendapatkan hasil yang bermanfaat. Uraian kerangka pemikiran menjelaskan bagaimana hubungan maupun keterkaitan antar variabel penelitian. Pada gambar di bawah ini peneliti membuat gambaran bagaimana kerangka penelitian yang peneliti lakukan.



Gambar 2.6 Kerangka Pemikiran
Sumber: Data Olahan Peneliti (2019)

2.7 Hipotesis

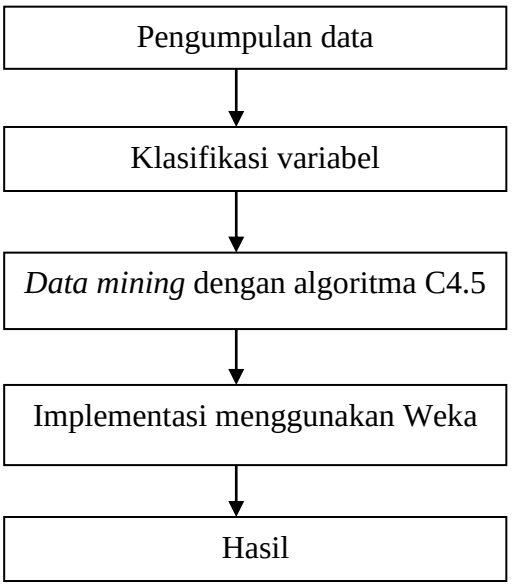
Data mining menggunakan *Algoritma* C4.5 untuk memprediksi pertumbuhan penduduk di kota Batam diharapkan dapat memberikan keputusan yang tepat.

BAB III

METODE PENELITIAN

3.1 Desain Penelitian

Dalam melakukan penelitian sangat perlu dilakukan perancangan penelitian supaya penelitian yang dilakukan dapat berjalan dengan baik dan sistematis. Dalam suatu proses penelitian, terlebih dahulu perlu dibuat desain penelitian. Hal ini bertujuan untuk memberikan kemudahan penelitian yang lebih lanjut. Untuk mempermudah dalam pengambilan data penelitian yang digambarkan melalui desain penelitian masalah berikut:



Gambar 3.1 Desain Penelitian
Sumber: Olahan Peneliti (2019)

Berdasarkan gambar 3.1 peneliti bisa menjabarkan urutan-urutan langkah penelitian, sebagai barikut:

1. Pengumpulan data

   Pengambilan data pada penelitian *data mining* untuk memprediksi pertumbuhan penduduk kota Batam dengan metode *decision tree* ini berupa data kelahiran, kematian, pindah dan datang di Dinas Kependudukan dan Pencatatan Sipil Kota Batam, dari tahun 2014 sampai 2018.

2. Klasifikasi variabel

   Dari data yang terkumpul dilakukan pengklasifikasian pada variabel agar mudah diolah, Beberapa variabel peneliti gunakan untuk penelitian ini adalah data kelahiran, kematian, pindah dan datang.

3. *Data mining* dengan algoritma C4.5

   Kemudian, dilakukan analisa dengan teknik algoritma C4.5 yang akan membentuk pohon keputusan.

4. Implementasi menggunakan Weka

   Setelah penelitian ini dianalisa, untuk mengetahui hasil yang tepat dan akurat, maka peneliti menggunakan aplikasi Weka untuk menentukan hasil yang lebih tepat.

5. Hasil

   Dari semua langkah penelitian yang telah dilakukan oleh peneliti, tahap terakhir adalah mengeluarkan hasil yang sudah diolah dengan algoritma C4.5 dan diimplementasikan dengan Weka.

3.2 Teknik Pengumpulan Data

Metode pengumpulan data adalah teknik atau cara-cara yang dapat digunakan oleh peneliti untuk pengumpulan data. Pengumpulan data merupakan hal yang harus dilakukan untuk melakukan sebuah penelitian. Ada suatu hubungan antara metode pengumpulan data dengan masalah penelitian yang ingin di pecahkan. Masalah akan memberi arah dan mempengaruhi metode pengumpulan data.

Data yang dikumpulkan peneliti menggunakan berbagai teknik yang ada, yaitu:

1. Wawancara

   Wawancara biasa dilakukan secara langsung maupun tidak langsung (seperti, lewat telepon, *cha*t maupun e-mail). Wawancara merupakan komunikasi dua arah antara pewawancara dan responden untuk menggali informasi yang relevan dengan tujuan penelitian. Pewawancara akan meminta responden memberikan informasi dalam bentuk fakta, opini, sikap, sehingga pembicara bisa leluasa memberikan lebih banyak informasi yang dibutuhkan.

2. Studi kepustakaan

   Studi kepustakaan merupakan teknik pengumpulan data dengan mengadakan penelaahan terhadap buku-buku, literatur-literatur, catatan-catatan, dan laporan yang ada hubungannya dengan masalah yang ingin dipecahkan.

3.3 Operasional Variabel

Operasional variabel merupakan proses menguraikan variabel penelitian kedalam sub variabel, dimensi, indikator sub variabel dan pengukuran. Syarat penguraian operasional dilakukan bila dasar konsep dan indikator masing-masing variabel sudah jelas, apabila belum jelas secara konseptual maka perlu dilakukan analisis faktor.

Untuk menguji hipotesis yang diajukan, maka variabel-variabel yang akan diteliti memerlukan indikator sebagai berikut:

1. Variabel independen

   Variabel independen yaitu variabel yang tidak terpengaruh variabel lain dan variabel yang bisa juga mempengaruhi variabel terikat. Variabel yang peneliti gunakan dalam penelitian ini adalah kelahiran, kematian pindah dan datang.

2. Variabel dependen

   Variabel dependen yaitu variabel terikat yang dipengaruhi oleh variabel bebas. Variabel dependen pada penelitian ini adalah pertumbuhan penduduk.

Tabel 3.2 Pengklasifikasian Data

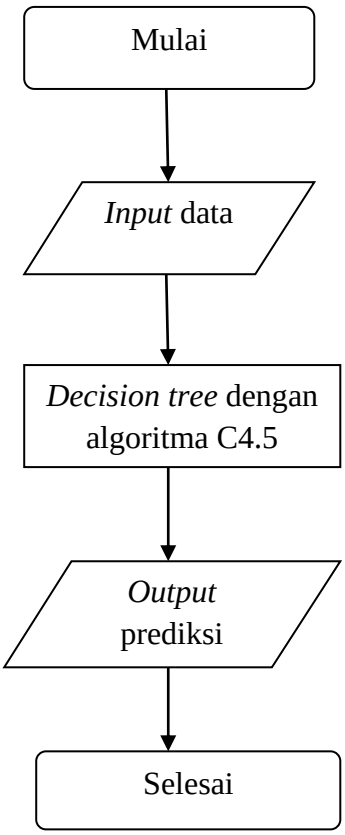
No	Variabel	Kategori	*Range*
1	Kelahiran	Rendah	0 - 2.281
		Sedang	2.282 - 2.909
		Tinggi	2.910 - 5.000
2	Kematian	Rendah	0 - 199
		Sedang	200 - 348
		Tinggi	349 - 700
3	Pindah	Rendah	0 - 2.438
		Sedang	2.438 - 3.367
		Tinggi	3.368 - 7.000
4	Datang	Rendah	0 - 3.487
		Sedang	3.488 - 5.888
		Tinggi	5.889 - 9.000
5	Total Pertumbuhan	Rendah	0 - 3.322
		Tinggi	3.323 - 7.000

Sumber: Data Olahan Peneliti (2019)

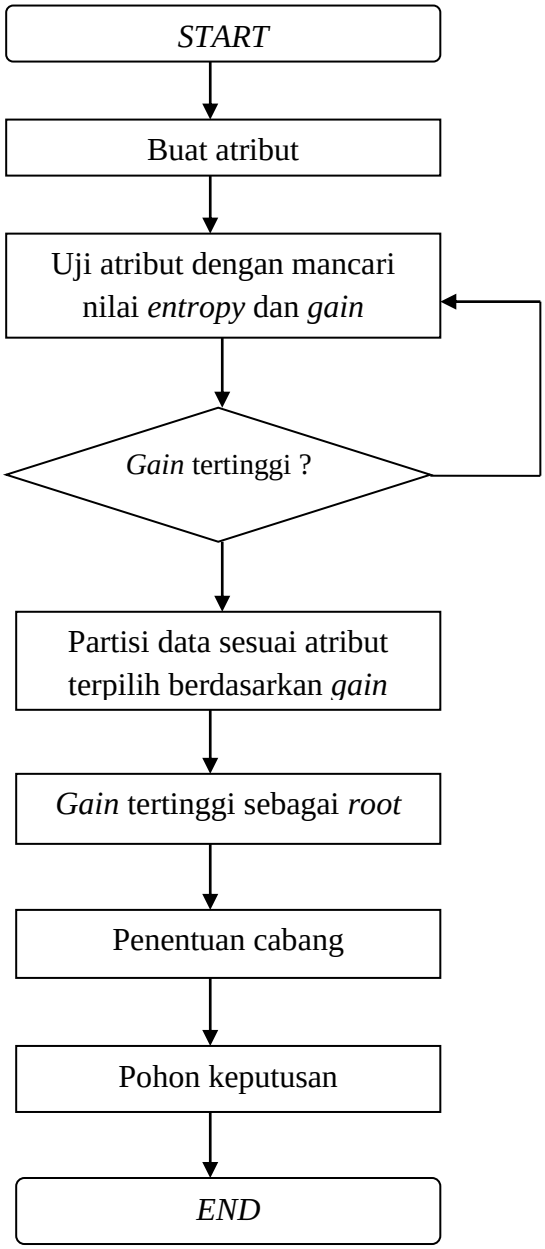
Berdasarkan tabel 3.1 di atas dapat dijelaskan bahwa data akan terklasifikasi berdasarkan *range*. Seperti variabel kelahiran, angka kelahiran dikatakan rendah apabila jumlah kelahiran terjadi 1-2.281 kelahiran, dikatakan sedang apabila angka kelahiran 2.282-2.909 kelahiran, dikatakan tinggi apabila angka kelahiran 2.910–5.000.

3.4 Metode Analisa Rancangan Sistem

Pada tahap ini adalah membuat suatu rancangan untuk membangun suatu sistem dalam bentuk *flowchart*, sebagai berikut:



Gambar 3.3 *flowchart* Sistem
Sumber: Hasil Olahan Peneliti (2019)



Gambar 3.4 *Flowchart* Algoritma C4.5
Sumber: Olahan Peneliti (2019)

Sebagaimana gambar *flowchart* di atas dapat diketahui bahwa alur algoritma C4.5 yang digunakan. Pada persiapan awal ditentukan atribut yang digunakan

kemudian melakukan uji atribut dengan mencari nilai *Gain* tertinggi berdasarkan perhitungan *entropy* dari masing-masing atribut. Apabila ditemukan *gain* tertinggi maka *gain* tersebut akan menjadi *root* awal. Selanjutnya dilakukan penentuan cabang dengan cara yang sama dengan melihat *gain* tertinggi dari tiap hasil pertisi.

3.5 Lokasi dan Jadwal Penelitian

Penelitian dilakukan di kantor Dinas Kependudukan dan Pencatan Sipil Kota Batam dan jadwal penelitian sampai selesai peneliti uraikan di bawah ini:

3.5.1 Lokasi Penelitian

Dalam melakukan penelitian ini, peneliti mengambil data di Dinas Kependudukan dan Pencatatan Sipil Kota Batam yang beralamat di jalan Ir. Sutami Telp.(0778) 321249 Sekupang Batam.

3.5.2 Jadwal Penelitian

Peneliti melakukan penelitian ini pada rentang waktu bulan Maret 2019 sampai dengan bulan Juli 2019.

Tabel 3.5 Jadwal Penelitian

No	Kegiatan	Pelaksanaan Kegiatan																		
		Maret 2019	April 2019				Mei 2019				Juni 2019				Juli 2019				Agustus 2019	
		4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2
1	Pengumpulan data																			
2	Bab I																			
3	Bab II																			
4	Bab III																			
5	Bab IV																			
6	Bab V																			
8	Pengumpulan Skripsi																			

Sumber: Hasil Olahan Peneliti (2019)

Tabel 3.5 di atas merupakan gambaran waktu peneliti melaksanakan penelitian sampai pengumpulan skripsi. Terlihat bahwa diawal, peneliti melakukan pengambilan data terlebih dahulu. Pada minggu ke tiga bulan mei peneliti memulai bab tiga dan selesai pada bulan juni minggu ke tiga, selain mengerjakan bab tiga peneliti juga mulai mengerjakan bab empat pada bulan minggu pertama bulan juni selesai pada bulan juli minggu ke tiga dan merupakan waktu terlama dalam pengerjaan bab. Bulan ke empat juni minggu ke 4 peneliti mengerjakan bab lima dan minggu berikutnya penyelesaian skripsi dan dilakukan penyerahan di Baak setelah di tanda tangani pembimbing.