BAB II

LANDASAN TEORI

2.1 Tinjauan Pustaka

Dalam penelitian ini akan menggunakan dua puluh tinjauan studi yang nantinya dapat mendukung penelitian. Tinjauan studi yang digunakan yaitu sebagai berikut :

1. Sunarti (2019) meneliti tentang Prediksi Promosi Jabatan Karyawan dengan Algoritma C4.5 (Studi Kasus: Apartemen Senayan Jakarta). Tujuan penelitin ini yaitu untuk memprediksi promosi jabatan yang digunakan untuk meningkatkan kinerja karyawan menjadi lebih baik, kreatif, dan bertanggung jawab dengan menggunakan Algortima C4.5 dengan klasifikasi data sesuai kriteria yang ada. Dalam melakukan klasifikasi ini data dianalisis menggunakan software Rapidminer. Dataset yang digunakan pada penelitian ini sebanyak 50 data untuk data training dengan atribut, Pengetahuan pekerjaan, Kompetensi, Kerja tim, Komunikasi, Kualitas pekerjaan, Inovasi, dan Kepemimpinan. Hasil perhitungan klasifikasi dievaluasi dengan confusion matrix menghasikan angka akurasi 78%+-14,00%, precision 83,17% +/-14,67 dan recall 89,17% +/-17,50% serta dengan angka curva ROC 0.867.
2. Permana, Ainiyah and Holle (2021) meneliti tentang Analisis Perbandingan Algoritma Decision Tree, kNN, dan Naive Bayes untuk Prediksi Kesuksesan Start-up. Penelitian ini bertujuan untuk melakukan pengklasifikasian start-up yang sukses dan gagal, sehingga nantinya dapat

digunakan untuk melihat faktor-faktor yang paling mempengaruhi pada keberhasilan start-up. Selain itu penulis menggunakan beberapa metode klasifikasi yang kemudian akan dilakukan perbandingan terhadap nilai performa, sehingga dapat mengetahui algoritma mana yang memiliki tingkat akurasi lebih tinggi. Hasil perbandingan antara algoritma Decision Tree, kNN, dan Naive Bayes menunjukkan algoritma Decision Tree merupakan algoritma yang paling cocok di antara algoritma kNN dan Naive Bayes. Hasil akurasi Decision Tree adalah sebesar 79,29%, sedangkan algoritma kNN sebesar 66,69%, dan Naive Bayes sebesar 64,21%. Untuk nilai presisinya, Decision Tree masil lebih unggul dengan nilai 78,99%, dan algoritma kNN dengan nilai 55,13%, dan Naive Bayes 51,32%. Dari hasil performa recall, algoritma Naive Bayes memberikan hasil sebesar 79,16%, sedangkan Decision Tree 56,27% dan kNN sebesar 40,14%.

3. Kastawan, Wiharta and Sudarma (2018) melakukan penelitian tentang Implementasi Algoritma C5.0 pada Penilaian Kinerja Pegawai Negeri Sipil. Tujuan penelitian ini yaitu untuk melakukan analisa klasifikasi terhadap data penilaian kinerja pegawai negeri sipil dengan menggunakan algoritma C5.0. Data yang digunakan berjumlah 275 data PNS. Dibagi menjadi 2 bagian yaitu 38 untuk data struktural dan 235 untuk data staf. Kemudian dilakukan model pengujian sebanyak 4 kali :
   - 184 data training dan 51 data uji dengan hasil akurasi sebesar 96.08%
   - 118 data training dan 117 data uji dengan hasil akurasi sebesar 89.74%
   - 70 data training dan 165 data uji dengan hasil akurasi sebesar 66.06%

- 32 data training dan 8 data uji dengan hasil akurasi sebesar 100%

Dapat disimpulkan bahwa penggunaan 184 data staf sebagai data training memiliki akurasi 96.08% . Salah satu faktor yang mempengaruhi adalah data tidak memenuhi syarat yang masih kecil. Maka dapat ditingkatkan dengan menambah jumlah data training yang akan digunakan.

4. Zamasi (2021) melakukan penelitian tentang Implementasi Algoritma C5.0 Pada Analisa Data Potensi Pertanian dan Perternakan. Tujuan penelitian ini untuk mengetahui jenis usaha pertanian atau perternakan di wilayah mana yang lebih berpotensi, yaitu dengan mengetahui luas lahannya berapa dan jenis usahanya serta produksi setiap jenis usaha tiap tahunnya di setiap wilayah. Dataset yang digunakan pada penelitian ini sebanyak 30 data. Dengan jumlah 13 data dengan kategori baik, dan 17 data dengan kategori tidak baik. Proses analisis data menggunakan software RapidMiner dengan hasil akurasi yang di dapatkan sebesar 95.67%
5. Fajri, Utami and Maruf (2022) melakukan penelitian tentang Perbandingan Model Pohon Klasifikasi Algoritma C4.5 dan C5.0 untuk Analisis Faktor yang Mempengaruhi Keberhasilan Lelang. Penelitian ini bertujuan untuk memodelkan kasus keberhasilan lelang di KPKNL Palu dengan menggunakan algoritma C4.5 dan algoritma C5.0 serta untuk mengetahui model terbaik yang dihasilkan dengan menggunakan teknik Cross Validation. Dataset yang digunakan berjumlah 984 data pada tahun 2018 dari Kantor KPKNL Palu. Adapun hasil dari model pohon menggunakan algoritma C5.0 menghasilkan ketepatan klasifikasi sebesar 96,43% dan

model pohon dengan menggunakan C4.5 menghasilkan ketepatan sebesar 92.86%. Dengan kesimpulan bahwa algoritma C5.0 memiliki tingkat akurasi yang lebih tinggi daripada algoritma C4.5

6. Padang (2019) melakukan penelitian tentang Implementasi Data Mining Algoritma C5.0 dalam Memprediksi Penerimaan Cleaning Service(CS) pada PT ISS Indonesia Medan. Tujuan penelitian ini yaitu untuk memecahkan masalah penugasan karyawan, seperti karyawan yang tidak tepat dalam penempatan ke bagian yang bukan ahlinya dan karyawan yang berlebihan pada bagian produksi. Serta untuk melakukan prediksi penerimaan cleaning service. Pengolahan data pada penelitian ini menggunakan perangkat lunak RapidMiner 5.3 dengan data karyawan cleaning service pada tahun 2017 dengan atribut pendukung seperti Pendidikan, Tinggi badan, Berat badan, dan Pengalaman. Penelitian ini menghasilkan informasi bersifatklasifikasi yaitu mengubah fakta menjadi pohon keputusan yang mempresentasikan aturan kelayakan prediksi penerima cleaning service (cs).
7. Putri (2016) melakukan penelitian tentang Prediksi Pola Kecelakaan Kerja pada Perusahaan Non Ekstraktif Menggunakan Algoritma Decision Tree C4.5 dan C5.0. Penelitian ini bertujuan untuk membangun suatu perangkat lunak yang mampu menggali pola data dengan menggunakan metode klasifikasi algoritma decision tree. Selain itu, untuk mengetahui perbandingan tingkat keakuratan dari algoritma yaitu C4.5 dan C5.0. Dataset yang digunakan pada penelitian ini berjumlah 200 data kecelakaan kerja yang diambil dari PT. Ajinomoto Indonesia kantor cabang

Mojokerto. Perusahaan ini termasuk industri non ekstraktif yang bergerak dibidang kuliner. Atribut yang digunakan pada proses klasifikasi ini yaitu Id, Jenis kelamin, Umur, Tipe pekerjaan, Lama bekerja, Status, Pelatihan, Kecelakaan kerja. Hasil uji coba dengan algoritma C4.5 dan C5.0 menghasilkan jumlah pola yang berbeda, tergantung pada pembagian data training dan data testing yang digunakan. Semakin banyak data training yang digunakan, maka akan semakin tinggi hasil akurasi yang dihasilkan. Aturan sebagai representasi dari pola yang dihasilkan oleh algoritma C5.0 lebih ringkas daripada C4.5.

8. Suwarno *et al.* (2021) melakukan penelitian tentang Prediksi Pengangkatan Karyawan dengan Metode Klasifikasi Algoritma C5.0 (Studi Kasus CV. T-PICO JAYA MANDIRI). Tujuan penelitian ini yaitu untuk memprediksi permintaan pengambilan keputusan untuk pengangkatan karyawan kontrak menjadi tetap. Jumlah keseluruhan dataset yang digunakan pada penelitian ini adalah 403 data, dengan pembagian jumlah data latih *(training)* sebanyak 303 data, dan 100 data untuk data uji *(testing).* Setelah melakukan perhitungan manual, kemudian data di uji menggunakan tools RapidMiner untuk mengetahui kesesuaian perhitungan manual dengan aplikasi. Hasil penelitian ini dapat disimpulkan bahwa dengan menggunakan klasifikasi algoritma C5.0, dapat memprediksi pegangkatan karyawan, serta dapat berkonstribusi terhadap proses pengambilan keputusan bagi pihak perusahaan. Algoritma decision tree juga memiliki kecepatan dan tingkat keakuratan dalam pengambilan keputusan kepada calon maupun karyawan yang sedang

bekerja. Evaluasi hasil pada prediksi penganggakatan karyawan menggunakan data training dan data testing dengan Confusion Matrix dimana data testing dengan tingkat accuracy sebesar 91.00% dan data training dengan tingkat accuracy 96.75%.

9. Sungkar and Qurohman (2021) melakukan penelitian tentang Penerapan Algoritma C5.0 untuk Prediksi Kelulusan Pembelajaran Mahasiswa pada Matakuliah Arsitektur Sistem Komputer. Penelitian ini bertujuan untuk memprediksi kelulusan mahasiswa pada matakuliah arsitektur sistem computer dengan penerapan data mining menggunakan algoritma C5.0. Proses rediksi dilakukan berdasarkan dengan klasifikasi algoritma C5.0 dengan menggunakan beberapa atribut seperti nilai kehadiran, nilai tugas, nilai UTS dan nilai UAS. Hasil akhir dari proses klasifikasi algoritma C5.0 rule dalam interpretasi pohon keputusan. Kinerja algoritma C5.0 mendapatkan tingkat akurasi yang tinggi sebesar 93,33%
10. Nawangsih *et al.* (2021) melakukan penelitian tentang Prediksi Pengangkatan Karyawan dengan Metode Algoritma C5.0 (Studi Kasus PT. MATARAM CAKRA BUANA AGUNG). Tujuan penelitian ini yaitu untuk mengatasi permasalahan yang ada di perusahaan tersebut dalam menentukan karyawan kontrak menjadi karyawan tetap yang selama ini prosesnya dilakukan dengan cara penyeleksian berkas secara manual, seperti tes lisan, tes fisik, atau tertulis, wawancara dan sebagainya. Dengan penerapan algoritma C5.0 dapat membantu untuk pengambilan keputusan pengangkatan karyawan dan memperoleh informasi dengan cara terkomputerisasi. Jumla dataset yang digunakan pada penelitian ini adalah

403 data untuk data latih *(training)* dan 100 data untuk data uji *(testing)*. Pengujian data dilakukan menggunakan tools *RapidMiner 9.5* dengan hasil model pohon keputusan. Hasil dari klasifikasi dalam prediksi tersebut menggunakan data latih *(training)* dan data uji *(testing)* dengan evaluasi akurasi menggunakan Confusion Matrix dimana data uji *(testing)* sebesar 91.00% dan data latih *(training)* sebesar 96.75%.

11. (Yanti, 2020) melakukan penelitian tentang Prediksi Mahasiswa Berpotensi Non Aktif Menggunakan Data Mining dalam Decision Tree dan Algoritma C4.5. Penelitian ini bertujuan untuk menghindari terjadinya *dropout* mahasiswa secara sepihak. Data yang digunakan berjumlah 107. Hasil dari penelitian ini berupa *rules* atau aturan yang menghasilkan kriteria-kriteria yang tepat dalam menganalisa mahasiswa berpotensi non aktif. Kriterianya yaitu jadwal kuliah, nilai absensi, nilai gagal, dan pembayaran uang kuliah.
12. (Aswan Supriyadi Sunge, 2018) melakukan penelitian tentang Prediksi kompetensi karyawan menggunakan algoritma C4.5 (Studi kasus : PT Hankook Tire Indonesia). Tujuan dari penelitian ini adalah untuk mencari calon karyawan yang sesuai dengan kategori yang diinginkan perusahaan. Data yang dikumpulkan berjumlah 205 data dan kemudian dibagi menjadi data training sebanyak 164, dan data testing sebanyak 41 data. Tingkat akurasi yang dihasilkan dengan data training sebesar 78.64 % dan data testing sebesar 56.00%.
13. (Susanti et al., 2018) melakukan penelitain tentang Prediksi pengangkatan karyawan kontrak menjadi karyawan tetap menggunakan Decision Tree

pada PT. baskara Cipta Pratama dengan tujuan untuk memprediksi karyawan kontrak menjadi karyawan tetap yang dapat membantu perusahaan dalam memilih karyawan yang tepat. Hasil dari penelitian ini membangun sebuah sistem pendukung keputusan menggunakan metode algoritma C.45 guna meningkatkan akurasi dalam penentuan karyawan kontrak menjadi karyawan tetap, Hasil dari pengklasifikasiannya divalidasi dengan k-fold cross validation dengan tingkat akurasi 90.83 %, presisi 91.18% dan recall 62,50 %

14. (Sinaga et al., 2022) melakukan penelitian yang berjudul "Perbandingan akurasi algoritma naive bayes, KNN dan SVM dalam memprediksi penerimaan pegawai." Penelitian ini bertujuan membandingkan performa algoritma naïve bayes, KNN dan SVM dalam memprediksi Data pelamar di Politeknik Bisnis Indonesia dari tahun 2021 sampai Bulan Maret 2022 sebanyak 33. Variabel yang digunakan antara lain pendidikan, Pengalaman Kerja, Jenis Kelamin, TPA, Psikotest, Wawancara dan Usia. Nilai akurasi yang dihasilkan algoritma SVM sebesar 84.9%, presisi 85.1% sedangkan K-NN memiliki nilai akurasi 81.8%, presisi 84.1% dan Naïve Bayes memiliki nilai akurasi 78.8% dan presisi 80.1%.
15. (Rohman & Rufiyanto, 2019) melakukan penelitian tentang Implementasi data mining dengan algoritma decision tree C4.5 untuk prediksi kelulusan mahasiswa di Universitas Pandanaran. Tuuan dari penelitian ini yaitu untuk memprediksi kelulusan mahasiswa terutama jenjang pendidikan D3 di Fakultas Teknik, memiliki data mahasiswa kelas reguler dan mahasiswa kelas karyawan dan kebanyakan statusnya sudah bekerja.

16. (Wahono & Riana, 2020) melakukan riset tentang Prediksi Calon Pendonor Darah Potensial Dengan Algoritma Naïve Bayes, K-Nearest Neighbors dan Decision Tree. Riset ini bertujuan untuk untuk menemukan metode paling tepat dengan menggunakan 3710 dataset donor darah dari PMI Kota Bekasi. Hasil penelitian ini menunjukkan algoritma Decision Tree C4.5 memiliki akurasi yang lebih tinggi yaitu 93.83% dibandingkan algoritma Naïve Bayes yang menghasilkan nilai akurasi sebesar 85.15% dan algoritma K-Nearest Neighbors dengan nilai akurasi sebesar 84.10%.
17. (Rohman & Mujiyono, 2021) melakukan penelitian dengan judul "Permodelan Prediksi Predikat Kelulusan Mahasiswa Menggunakan Decision Tree C4.5". penelitian ini bertujuan untuk memprediksi predikat kelulusan mahasiswa Fakultas Komputer dan Pendidikan Universitas Ngudi Waluyo Ungaran. Data yang digunakan adalah data kelulusan tahun 2021 dalam 2 periode dengan jumlah 35 mahasiswa. Hasil penelitian ini mendapat model pohon keputusan dengan variabel indeks prestasi IP dari semester 1,2,3, yang berpengaruh terhadap predikat kelulusan mahasiswa dengan nilai akurasi 71,67%.
18. (Nuraeni, 2021) melakukan penelitian tentang Klasifikasi Data Mining untuk Prediksi Potensi Nasabah dalam Membuat Deposito Berjangka. Penelitian ini bertujuan untuk dapat memprediksi nasabah mana yang memiliki potensi membuka deposito berjangka dengan melakukan perbandingan nilai akurasi yang dihasilkan dari masing-masing algoritma yang digunakan. hasilnya Decision Tree menjadi algoritma klasifikasi

terbaik dengan nilai akurasi 91.26% dan Naïve Bayes memiliki nilai akurasi sebesar 86.96% dan k-Nearest Neighbor sebesar 90.39%.

19. (Aryanto & Elisa, 2022) melakukan penelitian yang berjudul "Decision Tree technique dalam menentukan penerimaan karyawan supermarket di Lotte Grosir Batam". Penelitian ini bertujuan untuk dapat memilih karyawan yang sesuai dengan kriteria yang diinginkan oleh perusahaan sehingga dapat memajukan perusahaan dan dapat mencapai target penjualan. Data yang digunakan adalah data pelamar kerja dengan 5 variabel antara lain Lulusan terakhir, Kompetensi, Pengalaman kerja, Prestasi yang ada, dan Tes Ujian Masuk. Hasil analisa dengan algoritma C4.5 ini menjelaskan jika variabel yang memiliki gain paling tinggi dan yang dijadikan faktor utama sangat memiliki dampak yang besar pada penentuan penerimaan karyawan. Variable yang memiliki nilai gain tertinggi yaitu Tes ujian masuk, prestasi dan lulusan.
20. (Purba et al., 2022) melakukan penelitian tentang Implementasi algoritma C5.0 dalam menilai kinerja karyawan pada PT SMARTFREN TELCOM TBK. Penelitian ini bertujuan untuk mempermudah penilaian kinerja karyawan untuk memberikan penghargaan, bonus, dan promosi kepada para karyawan. Berdasarkan evaluasi hasil dengan data latih sebanyak 23 data diperoleh tingkat akurasi sebesar 97%, dimana salah satu faktor yang mempengaruhi akurasi adalah data masih kecil, hal ini dapat ditingkatkan dengan menambah jumlah data latih yang digunakan.

Pada penelitian terdahulu yang telah dilakukan oleh beberapa peneliti terdapat perbedaan tingkat akurasi pada masing-masing algoritma tersebut. Seperti

pada penelitian yang dilakukan oleh (Putri, 2016) dengan menggunakan data sebanyak 200 data membuktikan bahwa algoritma C4.5 memiliki tingkat akurasi yang lebih tinggi dibandingkan algoritma C5.0, sedangkan algoritma C5.0 merupakan pengembangan dari algoritma sebelumnya, yakni C4.5. kemudian pada penelitian yang dilakukan oleh (Fajri et al., 2022) dengan menggunakan data sebanyak 984 data memberikan hasil akurasi algoritma C5.0 yang lebih tinggi dibandingkan algoritma C4.5. Pada beberapa penelitian tersebut masih menggunakan dataset yang relatef sedikit. Sehingga, penulis menyimpulkan untuk melakukan penelitian terkait perbandingan algoritma Decision tree jenis C4.5 dan C5.0 dengan tujuan mengetahui tingkat akurasi yang dihasilkan dari masing-masing algoritma tersebut dengan menggunakan dataset yang lebih banyak.

2.2 Data Mining

Data mining merupakan sebuah teknologi yang dapat memproses data menggunakan teknik dan metode tertentu dalam volume besar yang digunakan oleh perusahaan untuk mengubah data mentah menjadi informasi atau pengetahuan yang berguna untuk membuat suatu keputusan bisnis. Data mining disebut juga dengan *Knowledge Discovery in Database* (KDD) yang pada dasarnya mempunyai beberapa teknik yaitu Deskripsi, Klasifikasi, Klasterisasi, Asosiasi, dan Prediksi. Data mining juga dapat diartikan sebagai bidang keilmuan yang menyatukan beberapa teknik dari machine learning (pembelajaran mesin) pengenalan pola, database, statistik, dan visualisasi untuk penggalian informasi yang berguna (T.Larose & D.Larose, 2005). Hasil dari data mining sering kali diintegrasikan dengan *Decision Support System* (DSS) atau sistem pendukung keputusan.

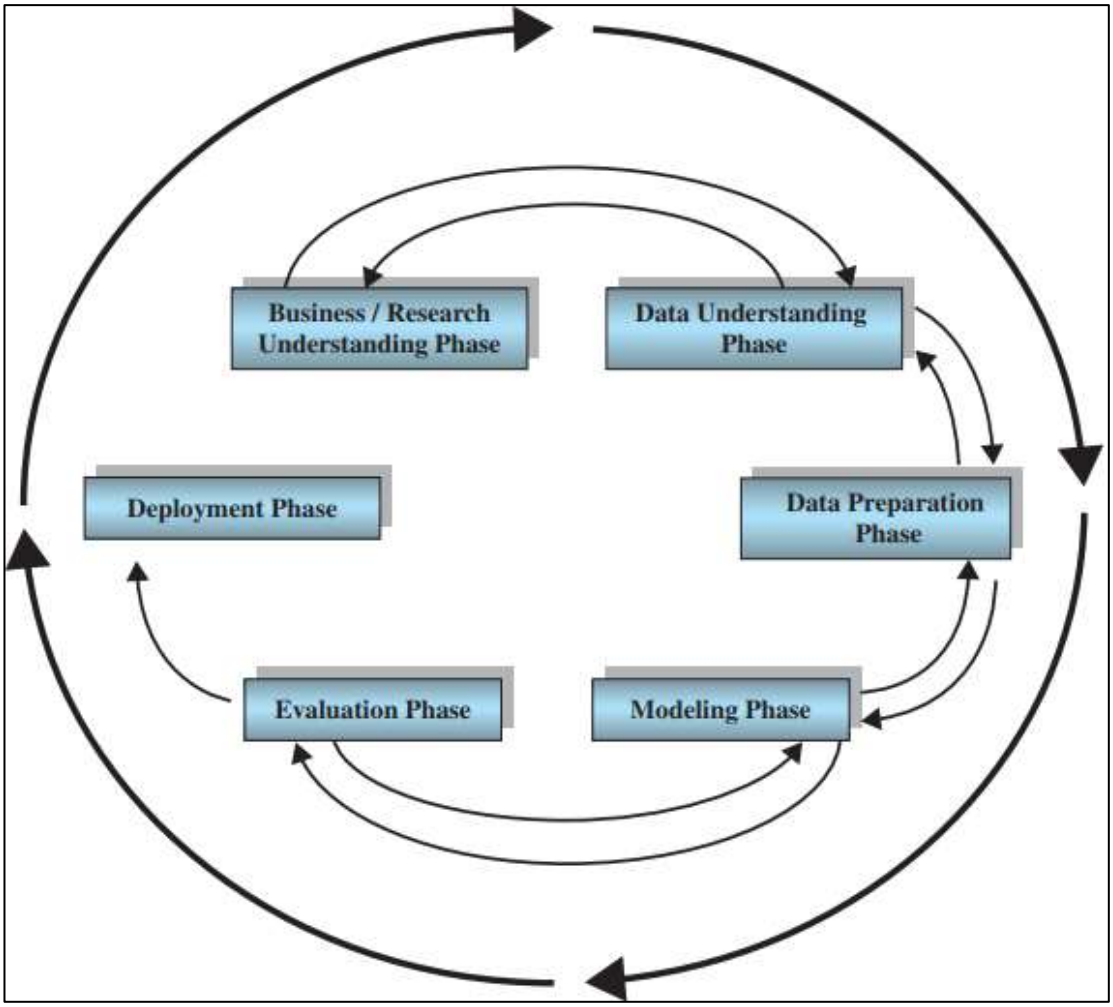
Terdapat dua metode pendekatan pada proses data mining, diantaranya sebagai berikut:

1. Supervised learning, yaitu pembelajaran yang diawasi oleh guru atau supervisor. Guru disini adalah label atau kelas target dari data yang berfungsi untuk melatih data yang digunakan untuk membangun sebuah model. Metode ini biasa dimanfaatkan untuk klasifikasi dan prediksi.
2. Unsupervised learning, yaitu pembelajaran bebas tanpa pengawasan. Metode ini diterapkan tanpa adanya label atau kelas target pada data yang digunakan, sehingga metode ini dapat mengeksplorasi data tanpa adanya latihan. Metode ini biasa diterapkan untuk klastering atau pengelompokan.

2.3 CRISP-DM

Cross Industry Standard Process for Data Mining (CRISP-DM) adalah sebuah proses data mining sebagai standar pemecahan masalah secara umum maupun penelitian. Menurut penelitian terdahulu, metode ini merupakan salah satu metode *data mining* yang sering dipakai. Berdasarkan polling online di KDNuggets pada tahun 2014, 45% responden memilih CRISP-DM sebagai metode utama dalam data analisis, data mining ataupun projek data science lainnya (Azvedo and Santos, 2008). Salah satu keuntungan dari model ini adalah, tahapan fase dari model bukanlah sebuah tahapan yang kaku. Perpindahan maju dan mundur antara tiap fase yang berbeda bisa selalu dilakukan. Hal ini sesuai dengan sifat alami dari data mining itu sendiri, dimana proses data mining tidak selesai saat sebuah hasil di temukan, sebab proses data mining merupakan sebuah proses pembelajaran terus menerus. CRISP-DM memiliki siklus yang terdiri dari enam fase. Tiap tahap berurutan dan bergantung pada tahap sebelumnya. Berikut

ini merupakan proses CRISP-DM yang terlihat pada gambar dibawah ini (T.Larose & D.Larose, 2005) :



Gambar 2.1 Fase CRISP-DM

Berikut ini merupakan enam tahapan siklus pada CRISP-DM :

1. **Tahap Pemahaman Bisnis/Penelitian (*Business/Reaserch Understanding*)**

   Pada tahap awal ini, melakukan pemaparan dari tujuan penelitian dan pengembangan model penelitian secara keseluruhan. Kemudian menyiapkan strategi untuk mencapai tujuan tersebut.

2. **Tahap Pemahaman Data (*Data understanding*)**

   Pada tahap ini dilakukan proses pengumpulan data kemudian melakukan analisis untuk memahami data serta evaluasi kualitas data guna untuk memilih data yang relevan dengan keterkaitan penelitian.

3. **Tahap Persiapan Data (*Data preparation*)**

   Dalam tahap ini mencakup aspek pengecekan dataset yang akan digunakan dengan melakukan seleksi atribut atau variabel yang berpengaruh terhadap proses klasifikasi. Tahap ini sama halnya dengan preprocessing pada KDD yang melakukan proses data cleaning, data transformation, feature selection jika diperlukan.

4. **Tahap pemodelan (*Modelling*)**

   Setelah semua data di processing, pada tahap selanjutnya dilakukan penerapan teknik pemodelan yang sesuai yang akan digunakan.

5. **Tahap Evaluasi (*Evaluation*)**

   Setelah tahap pemodelan dilakukan, model yang dihasilkan harus dievaluasi untuk kualitas dan efektivitasnya untuk menentukan apakah model tersebut sudah mencapai tujuan yang ditetapkan di Tahap 1 (Pemahaman bisnis/penelitian).

6. **Tahap Penerapan (*Deployment*)**

   Pada tahap deployment ini biasanya dilakukan dengan menerapkan model yang terlah dibuat menjadi sebuat website atau sistem. Akan tetapi pada penelitian ini tahap deployment tidak dilakukan, hanya sebatas presentasi hasil dari penelitian yang berupa laporan atau visualisasi dari bentuk model yang dihasilkan berupa rule atau aturan.

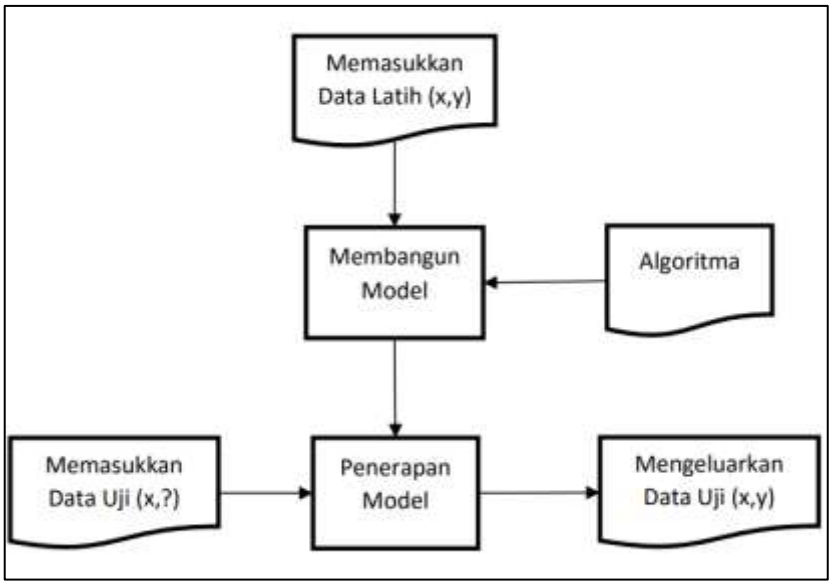
2.4 Dataset

Dalam pengolahan data mining, dataset merupakan salah satu bahan yang sangat dibutuhkan. Menurut terminologi statistik dataset adalah sekumpulan objek yang memiliki atribut atau variabelnya. Sebutan lain dari objek adalah *record, point*, entitas, *pattern, vector, case, event, sample.* Objek ini digambarkan dengan beberapa atribut yang memiliki karakteristik dan sifat. Sementara atribut yaitu baris atau kolom yang manyatakan objek-objek tersebut. Atribut juga biasa disebut dengan *field* atau variabel.

Pada penelitian ini akan menggunakan data karyawan yang diambil dari website www.kaggle.com yang berjumlah 54.808 baris dan 14 atribut. Kemudian 80% data tersebut akan dijadikan data latih (*training)*, dan 20% akan dijadikan data uji (*testing*).

2.5 Klasifikasi

Klasifikasi adalah teknik untuk membedakan kelas atau atribut yang telah memiliki karakteristik yang di telah tentukan. Klasifikasi dapat didefinisikan sebagai suatu proses yang melakukan pelatihan/pembelajaran terhadap fungsi target f yang memetakan setiap vektor (set fitur) x ke dalam satu dari sejumlah label kelas y yang tersedia. Di dalam klasifikasi diberikan sejumlah record yang dinamakan training set, yang terdiri dari beberapa atribut, atribut dapat berupa kontinyu ataupun kategoris, salah satu atribut menunjukkan kelas untuk record (Prasetyo, 2012). Berikut merupakan kerangka kerja klasifikasi :



Gambar 2.2 Kerangka kerja klasifikasi

Klasifikasi merupakan teknik yang digunakan untuk menemukan model agar dapat menentukan atau membedakan konsep atau kelas data, dengan tujuan untuk dapat memperkirakan kelas dari suatu objek. Pada proses klasifikasi, data latih (training) yang digunakan sudah memiliki kelas target/label, sehingga proses ini termasuk bagian dari supervised learning, yaitu untuk menemukan kelas target/label dari data yang ingin di prediksi (Jiawei Han, Micheline Kamber, 2014). Ada banyak algoritma yang biasa digunakan untuk melakukan prediksi atau klasifikasi diantaranya yaitu *Naïve bayes*, *K-Nearest Neighbour*, *Support Vector Mechine*, *Artificial Neural Network*, dan *Decision Tree.*

2.6 Algoritma Decision Tree

Algoritma *Decision Tree* merupakan metode klasifikasi dan prediksi yang cukup populer digunakan karena mudah untuk dipahami oleh manusia. Konsep dasar metode decision tree ini yaitu untuk menggolongkan data berdasarkan atribut-atribut tertentu dan membedakan kelas data atau label ke dalam bentuk

pohon keputusan. Dalam membangun pohon keputusan ini memiliki beberapa tahapan yaitu, memilih atribut sebagai akar (*root node*), dan cabang-cabang (*interna*l) yang memperlihatkan factor yang berpengaruh untuk memberikan keputusan pada daun terakhir (*leaf*). Kelebihan dari motode ini adalah tingkat akurasi yang tinggi dan lebih mudah digunakan dibandingakan dengan algoritma lainnya (Suntoro, 2019)

Ada 3 elemen yang terdapat dalam *Decision Tree*, yaitu :

1. *Root node* (node akar) merupakan node paling atas yang memilih atribut paling berpengaruh.
2. *Internal node* (cabang node) merupakan percabangan yang memiliki input dan output yang menyatakan pengujian.
3. *Leaf node* (daun node) adalah node terakhir yang sudah tidak dapat dipecah lagi, leaf node merupakan hasil akhir yang mempresentasikan prediksi jawaban dari data testing.

Decision Tree memiliki beberapa jenis algoritma diantaranya yaitu algoritma CART (Classification and Regression Tree), ID.3 C.45 dan C5.0 yang ditemukan oleh John Ross Quinlan pada tahun 1986.

2.6.1 Algoritma C4.5

Algoritma C4.5 merupakan salah satu jenis algoritma *Decision Tree* dari perkembangan algoritma ID3. Perbedaan dari algortima C4.5 dengan ID3 adalah dapat menangani data numerik, melakukan pemangkasan (*pruning*), dan penurunan (*deriving*) *rule set.* Penentuan akar node (*root node*) pada algoritma

C4.5 ini menggunakan perhitungan nilai *entrophy* dan gain tertinggi (Kusrini & Taufiq Emha, 2009).

Berikut tahap pembangunan *decision tree* menggunakan algoritma C4.5 :

1. Menyiapkan data training yang akan diklasifikasikan, data training harus berukuran lebih besar daripada data testing.
2. Menentukan akar dari pohon, akar akan diperoleh dari atribut yang terpilih dengan cara menghitung gain dari masing-masing atribut. Nilai gain tertinggi akan dijadikan akar pertama dalam pohon keputusan. Sebelum menghitung nilai gain, terlebih dahulu menghitung *entrophy.*
3. Ulangi langkah 2 hingga semua record terpartisi.
4. Proses partisi akan berhenti saat:
    a. Semua record pada simpul N mendapat kelas yang sama.
    b. Tidak ada atribut didalam record yang akan dipartisi lagi.
    c. Tidak ada record dicabang yang kosong

Berikut rumus *entrophy* yang digunakan pada algoritma C4.5 dihitung menggunakan persamaan 2.1 :

$$Entropy(S) = -\sum_{j=1}^{k} p_j \log_2 p_j \quad (2.1)$$

Keterangan :

S : Himpunan kasus

k : Jumlah kelas pada variable

p_j : Proporsi dari S_j dan S

Selanjutnya untuk mencari nilai *information gain* dihitung menggunakan rumus pada persamaan 2.2:

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^{m} \frac{|S_i|}{|S|} \times Entropy(S_i) \quad (2.2)$$

Keterangan :

S : Himpunan kasus

A : Atribut

m : Jumlah kategori pada atribut A

|S*i*| : Jumlah kasus pada atribut ke-*i*

|S| : Jumlah kasus pada S

Proses ini di ulang untuk setiap cabang, sampai semua cabang memiliki kelasnya masing-masing. Pembentukan akar node (*root node)* dengan atribut yang memiliki nilai *gain* tertinggi. Hasil dari pembangunan tree akan di representasikan kedalam *rules* atau aturan sebagai pola prediksi karyawan yang berpotensi promosi.

2.6.2 Algoritma C5.0

Algoritma C5.0 merupakan perkembangan dari algortima C4.5. Algoritma C5.0 lebih baik daripada C4.5 pada kecepatan, memori, dan efisiensi dan nilai akurasi yang lebih tinggi. Tree yang dihasilkan algoritma C5.0 lebih ringkas dan bervariasi daripada C4.5 oleh karena itu pembangunan tree pada algoritma tersebut lebih cepat. Algoritma C5.0 dapat menangani data numeric, kontinyu dan diskret. Apabila terdapat data yang missing value/kosong, akan secara otomatis dibuang atau diisi dengan *mean* (nilai rata-rata) dari variabel yg bersangkutan. Algoritma C5.0 cocok digunakan untuk kumpulan data besar (Kusrini & Taufiq Emha, 2009).

Tahap dalam pembangunan pohon keputusan (*decision tree*) menggunakan algoritma C5.0 mirip dengan algoritma C4.5. Perbedaan pada algoritma C5.0 ini proses pemilihan atribut sebagai akar (*root node*) di tentukan berdasarkan ukuran *gain ratio* tertinggi sedangkan pada algoritma C4.5 berhenti pada perhitungan *information gain*. Tahapan tersebut antara lain sebagai berikut:

1. Menyiapkan data training yang akan diklasifikasikan, data training harus berukuran lebih besar daripada data testing.
2. Untuk menentukan akar pohon (*root node*) diperoleh dari atribut yang terpilih dengan cara menghitung nilai *gain ratio* dari masing-masing atribut. Nilai *gain ratio* tertinggi akan dijadikan akar pertama dalam pohon keputusan. Untuk memperoleh nilai *gain ratio* digunakan rumus *entrophy*.
3. Ulangi langkah 2 hingga semua record terpartisi.
4. Proses partisi akan berhenti saat:
    a. Semua record pada simpul N mendapat kelas yang sama.
    b. Tidak ada atribut didalam record yang akan dipartisi lagi.
    c. Tidak ada record dicabang yang kosong

Berikut rumus *entrophy* yang digunakan pada algoritma C5.0 dihitung menggunakan persamaan 2.3 :

$$Entropy(S) = -\sum_{j=1}^{k} p_j \log_2 p_j \tag{2.3}$$

Keterangan :

S : Himpunan kasus

k : Jumlah kelas pada variable

p_j : Proporsi dari S_j dan S

Selanjutnya untuk mencari nilai *information gain* dihitung menggunakan rumus pada persamaan 2.4:

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^{m} \frac{|S_i|}{|S|} \times Entropy(S_i) \quad (2.4)$$

Keterangan :

S : Himpunan kasus

A : Atribut

m : Jumlah kategori pada atribut A

|S*i*| : Jumlah kasus pada atribut ke-*i*

|S| : Jumlah kasus pada S

Dan menghitung nilai *split info* menggunakan persmaan 2.5 :

$$SplitInfo(S,A) = -\sum_{i=1}^{m} \frac{Si}{S} log2 \frac{Si}{S} \quad (2.5)$$

Keterangan :

S : Himpunan kasus

A : Atribut

Si : Jumlah kasus untuk atribut i

Selanjutnya menghitung nilai *gain ratio*. Untuk mencari nilai *gain ratio* menggunakan rumus pada persamaan 2.6 :

$$Gain\ Ratio\ (S,A) = \frac{Gain\ (S,A)}{Split\ info(S,A)} \quad (2.6)$$

Keterangan :

Gain (S,A) : Nilai *gain* dari suatu atribut

Split Info (S,A) : Jumlah nilai *split info*

Proses ini di ulang untuk setiap cabang, sampai semua cabang memiliki kelasnya masing-masing.

Dari hasil perhitungan *gain ratio*, akan terpilih *gain ratio* terbesar yang akan dijadikan sebagai akar *node* (*root node*) pada pembangunan tree. Kemudian proses *gain ratio* diulang sampai masing-masing cabang pada semua kelas memiliki kelasnya. Hasil dari pembangunan tree akan di representasikan kedalam *rules* atau aturan sebagai pola prediksi karyawan yang berpotensi promosi.

2.7 RapidMiner

RapidMiner merupakan perangkat lunak *(software)* yang bersifat terbuka *(open source)* yang dioperasikan menggunakan bahasa pemrograman java. *RapidMiner* dapat dijadikan alat (tools) untuk melakukan analisis *data mining*, *text mining*, dan analisis prediksi dengan menggunakan beberapa operator/library yang terdapat dalam *RapidMiner*. Sehingga pengguna dapat melakukan teknik data mining untuk pengambilan keputusan dengan hasil yang baik.

2.8 Python

Python merupakan bahasa pemrograman tingkat tinggi (*high level*) yang diciptakan oleh Guido Rossum yang memiliki berbagai fungi dan kemampuan untuk mengeksekusi berbagai instruksi secara bersamaan. *Syntax* (logika perhitungan) pada bahasa pemrograman ini jelas mudah dipahami bahkan orang awam sekalipun. *Python* juga populer digunakan dalam *data science* dan *machine learning*. Beberapa pustaka (library) pada *python* Pandas, Numpy, Sklearn (Scikitlearn), Matplotlib, dan masih banyak lagi (Wati et al., 2022). *Tools* yang digunakan pada penelitian ini adalah *Google Colabolatory* sebagai alat untuk penyusunan perhitungan dengan bahasa pemrograman *python.*

2.9 Kaggle

Menurut InfoWorld, Kaggle adalah sebuah situs atau wadah yang menyediakan banyak dataset dari berbagai bidang. Kaggle dibentuk oleh Anthony Goldbloom sebagai CEO dan Ben Hamner sebagai CTO pada tahun 2010. Situs ini juga menyediakan banyak perlombaan/kompetisi untuk para data scientist membuat model terbaik mulai dari menganalisa hingga memprediksi suatu dataset yang telah disediakan di kaggle atau dataset yang dibagikan secara open source oleh para pengguna kaggle itu sendiri. Pada penelitian ini dataset yang digunakan adalah data karyawan yang diambil dari website www.kaggle.com pada tahun 2020 yang di sediakan oleh penyelenggara kontes vidhya analytics.

2.10 Cross Validation

Cross Validation adalah teknik untuk evaluasi dan validasi model prediksi serta mengukur kinerja dari metode yang digunakan. Teknik ini membagi data secara acak menjadi *k-fold* dalam setiap bagian *(partisi)* yang akan digunakan. kemudian salah satu *k-fold* tersebut akan dijadikan sebagai data uji (*testing)* dan sisanya dijadikan sebgai data latih *(training).* Setiap record data digunakan di setiap iterasi tepat satu kali untuk training dan satu kali untuk testing.

2.11 Confusion Matrix

Confusion matrix adalah metode yang digunakan untuk mengukur dan menilai *performance* dari algoritma yang digunakan. Hasil Evaluasi dengan menggunakan confusion matrix menghasilkan nilai akurasi, serta laju error. Akurasi menyatakan jumlah data yang benar diklasifikasikan setelah melakukan proses pengujian, sedangkan laju error digunakan untuk menghitung kesalahan identifikasi (Kastawan et al., 2018).

Untuk menghitung akurasi menggunakan persamaan 2.7 :

$$\text{Akurasi} = \frac{TP+TN}{TP+FN+FP+FN} *100 \quad (2.7)$$

Keterangan :

TP : True Positive (jumlah data yang positif diklasifikasi dengan benar)

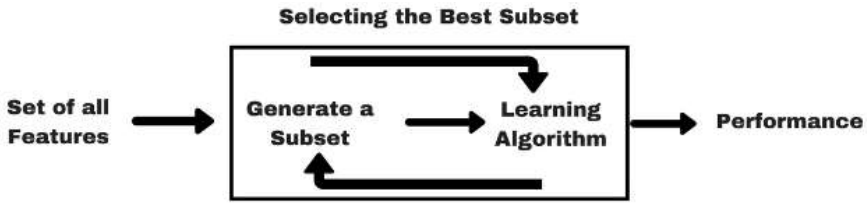
TN : True Negative (jumlah data yang negatif diklasifikasi dengan benar)

FN : False Negative (jumlah data yang negatif dan terklasifikasi salah)

FP : False Positive (jumlah data yang positif dan terklasifikasi salah)

2.12 Backward Elimination

Backward Elimination merupakan metode yang dapat menghilangkan atribut yang tidak signifikan. Berfungsi sebagai seleksi atribut dimana memanfaatkan regresi statistik untuk mengetahui kedekatan setiap kombinasi atribut dengan target. Semakin kecil *significance level*, maka semakin ketat pemilihan atribut yang akan terpilih sehingga semakin sedikit atribut yang terpilih untuk pembentukan model (Amilia et al., 2021). Dibawah ini merupakan tahapan metode *backward elimination*.



Gambar 2.3 Tahapan metode *backward elimination*