LAPORAN PROYEK DATA MINING

Predicting Employee's Performance with Apriori Algorithm

Disusun oleh:

1. 12S17007 – Ernike N. Manurung
2. 12S17021 – Inggrit S. Purba
3. 12S17024 – Yohana V. Aritonang

PROGRAM STUDI SARJANA SISTEM INFORMASI

FAKULTAS INFORMATIKA DAN TEKNIK ELEKTRO

INSTITUT TEKNOLOGI DEL

JANUARI 2021

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

BAB 1
BUSINESS UNDERSTANDING

Tahap pertama pada metodologi CRISP-DM untuk melakukan prediksi kinerja karyawan adalah *business understanding*. Pada bab ini akan dijelaskan pemahaman mengenai substansi dari aktivitas *data mining* yang akan dilaksanakan serta kebutuhan dari perspektif bisnis. Aktivitasnya antara lain menentukan sasaran bisnis, memahami situasi bisnis, menerjemahkan tujuan atau sasaran bisnis ke dalam tujuan *data mining.*

1.1 *Determine Bussiness Objective*

Saat ini, banyak organisasi atau perusahaan yang berupaya untuk meningkatkan kualitas dari perusahaannya. Organisasi mencoba untuk mengenali faktor-faktor apa saja yang menjadi pengaruh kemajuan perusahaan mereka. Salah satu faktor penting yang menunjang kemajuan suatu perusahaan adalah sumber daya manusia ataupun karyawan yang dipekerjakan. Hal ini dapat dilihat pada berbagai *test* yang harus diselesaikan oleh pelamar kerja untuk mendapatkan suatu posisi atau jabatan dalam perusahaan. Dengan kata lain, perusahaan mencoba untuk menemukan karyawan yang tepat untuk membantu mereka dalam mengembangkan bisnis atau perusahaan[1]. Dalam menangani kualitas sumber daya manusia, umumnya perusahaan memiliki *Human Resource Departement* (HRD). HRD bertugas untuk merekrut dan menyeleksi pelamar kerja serta memastikan para karyawan yang bekerja mematuhi seluruh kebijakan yang ditetapkan oleh perusahaan. Selain itu, HRD juga berperan dalam melakukan evaluasi terhadap kinerja dari setiap karyawan yang bertujuan untuk memberikan penghargaan bagi pekerja, mengetahui kinerja dari setiap karyawan untuk selanjutnya digunakan dalam pengembangan kualitas sumber daya manusia[2].

HRD sebagai bagian yang berperan dalam memanajemen sumber daya manusia, perlu memperhatikan faktor apa saja yang berpengaruh terhadap kinerja dari karyawan perusahaan. Dalam mengevaluasi kinerja karyawan ada berbagai faktor yang terlibat, baik itu faktor dari karyawan maupun faktor dari perusahaan tersebut. Dari berbagai faktor tersebut, HRD perlu mengetahui faktor-faktor apa saja yang sangat berpengaruh dalam menentukan kinerja karyawan tersebut sehingga perusahaan mampu meningkatkan kemajuannya.

HRD perlu menganalisis semua data terkait karyawan yang terdapat pada perusahaan tersebut, baik itu karyawan yang masih bekerja ataupu yang telah keluar dari perusahaan tersebut. Dalam kurun waktu yang lama setelah perusahaan berdiri, tentunya akan semakin banyak data karyawan yang perlu dianalisis untuk mengetahui faktor yang berpengaruh terhadap kinerja karyawan. Oleh karena itu, perlu dilakukan sebuah metode untuk menganalisis hal-hal apa saja

yang berdampak pada kinerja karyawan tersebut, seberapa baikkah atau seberapa burukkah hal tersebut mempengaruhi kinerja dari karyawan. Dengan mengetahui hal tersebut, HRD dapat melakukan evaluasi terhadap hal-hal yang berpengaruh untuk ke depannya dapat meningkatkan kinerja dari setiap karyawan perusahaan tersebut.

Sehingga objektif yang akan dicapai dari proyek ini adalah:

- Menemukan faktor-faktor apa saja yang berpengaruh terhadap kinerja karwayan
- Meningkatkan kinerja karyawan dengan mengevaluasi faktor yang berpengaruh terhadap kinerja

Proyek tersebut akan dikatakan sukses jika:

- Ditemukan faktor-faktor yang berpengaruh terhadap kinerja karyawan
- Proyek selesai dikerjakan dengan *on time* dan *under budget*.

1.2 *Situation Assesment*

Dalam pengerjaan proyek ini, perlu ditentukan ketersediaan sumber daya mencakup *hardware, data sources,* dan *personel*. *Hardware* yang digunakan untuk mengerjakan project tersebut adalah Laptop Lenovo Ideapad 310 dengan i5-7200 *microprocessor* dan 8GB RAM. *Dataset* yang akan dianalisis pada pengerjaan proyek ini adalah data *employee* yang diperoleh dari link: https://www.kaggle.com/shivan118/hranalysis?select=train.csv. Data tersebut merupakan data yang bersifat statis karena dalam format file CSV (*Comma Separated Values*) dan tidak ada data lebih lanjut yang akan dikumpulkan pada dataset tersebut dan metadata untuk dataset tersebut tidak tersedia. Selain itu, terkait personel yang dibutuhkan dalam pengerjaan proyek ini adalah 3 orang mahasiswa yang terlibat dalam setiap proses dalam proyek ini, baik itu dalam proses *business understanding, data understanding, data preparation, modeling, evaluation,* dan *deployment*.

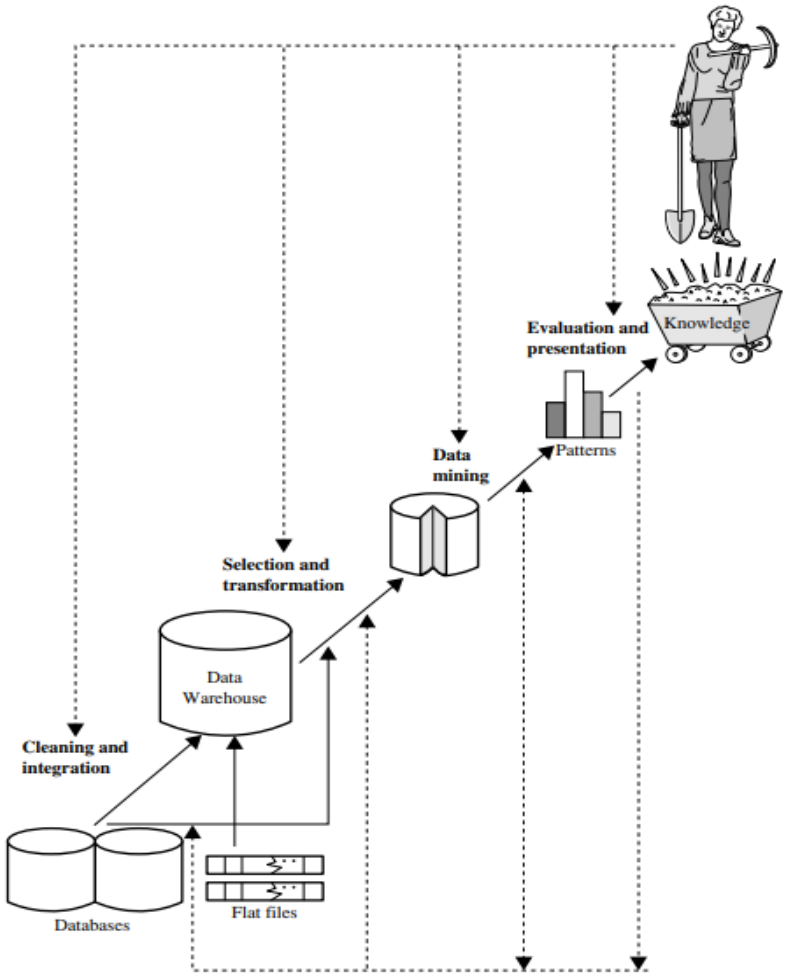
Selain itu, asumsi dalam pengerjaan proyek ini adalah bahwa team management hanya perlu memahami hasil yang akan diperoleh dengan kata lain tidak perlu memahami model yang digunakan untuk pengklasifikasian. Untuk risiko dalam pengerjaan proyek ini tidak terlalu signifikan. Projek akan dikerjakan secara daring oleh personel selama kurang lebih 4 minggu pengerjaan. Oleh karena itu, personel perlu berusaha dengan baik untuk memaksimalkan waktu dan sumber daya yang tersedia dalam pengerjaan proyek. Risiko lainnya yang mungkin terjadi adalah terkait masalah data, masalah teknologi. Namun hal tersebut masih memungkinkan untuk diatasi dengan baik. *Cost/Benefit analysis* juga perlu diperhatikan dalam pengerjaan proyek ini. Estimasi *cost* yang diperlukan mencakup pengumpulan data dan biaya operasi (berupa biaya akses internet). Sementara itu *benefit* yang diperoleh adalah mampu mencapai

objektif utama, menambah wawasan yang diperoleh dari pemahaman data maupun pemahaman dalam tahapan pengerjaan proyek.

1.3 *Determine Data Mining Goal*

Tujuan bisnis pada penelitian ini untuk memprediksi bagaimana kinerja sumber daya manusia (karyawan) adalah menemukan faktor apa saja yang mempengaruhi kinerja karwayan serta melakukan evaluasi terhadap hubungan beberapa faktor yang berpengaruh terhadap kinerja karyawan. Manfaat dari penelitian ini yakni dapat menentukan karyawan yang berhak mendapatkan penghargaan, memberikan motivasi kepada karyawan, mengembangkan kualitas kerja dan potensi diri karyawan serta mengambil keputusan untuk menghentikan hubungan kerja terhadap karyawan. Sehingga penilaian kinerja karyawan dapat menjadi data acuan untuk evaluasi perusahaan di masa mendatang dalam mencapai tujuan bisnis perusahaan. Dalam memprediksi tingkat kinerja karyawan diperlukan teknis penilaian yang efektif untuk mendapatkan keputusan secara subjektif. Maka dalam penelitian ini diterapkan suatu model yang dapat mengidentifikasi pola hubungan antara data. Model tersebut dapat dibentuk dengan metode *data mining.*

Data mining merupakan suatu proses untuk menemukan pola yang menarik dari data yang berjumlah besar[3]. *Data mining task* dikelompokkan menjadi *description*, *estimation*, *prediction*, *classification*, *clustering*, dan *association* [4]. Berdasarkan definisi *data mining* tersebut, *data mining* yang diterapkan pada penelitian ini merupakan proses menambang atau mengumpulkan data penilaian kinerja karyawan, lalu menemukan gambaran dan aturan yang diterapkan di data tersebut untuk mendapatkan informasi baru yang berguna. Ketika melakukan proses *data mining*, harus dilakukan beberapa tahapan antara lain, pembersihan data, integrasi data, pemilihan data, transformasi data, penemuan pola, evaluasi pola dan presentasi pengetahuan.



Gambar 1. *Data Mining* sebagai Tahapan dalam Proses *Knowledge Discovery*

Sumber: Data mining concepts and techniques (H. Jiawei, M. Kamber, and P. Jian,2014)

Dalam menemukan faktor apa saja yang berpengaruh terhadap kinerja karwayan serta melakukan evaluasi terhadap hubungan beberapa faktor yang berpengaruh terhadap kinerja karyawan diperlukan digunakan *data mining task* dengan teknik asosiasi. *Association rule mining* adalah metode pembelajaran mesin berbasis aturan untuk menemukan hubungan yang menarik antara variabel dalam data yang berjumlah besar[5]. *Association Rule* merupakan *data mining task* yang bertujuan untuk mencari pola yang sering muncul pada banyak transaksi, dimana pada setiap transaksinya terdiri dari beberapa item [6]. Penerapan *data mining* dengan *association rules* bertujuan menemukan informasi item-item yang saling berhubungan dalam bentuk aturan/*rule*. *Association rules* adalah teknik *data mining* untuk menemukan aturan asosiasi antara suatu kombinasi item [7]. Dengan menerapkan *Association rules mining* pada dataset karyawan pada penelitian ini, akan mudah untuk mengidentifikasi atribut-atribut yang berhubungan dengan kinerja karyawan dan memprediksi atribut apa saja yang mempengaruhi kinerja karyawan.

Dalam teknik *Association rules mining* ini dibutuhkan algoritma untuk mencari kandidat aturan asosiasi. Salah satu algoritma yang digunakan untuk teknik *association rules* adalah algoritma apriori. Kelebihan *association rules* dengan algoritma apriori ini adalah dapat menangani data yang berskala besar. Sedangkan algoritma lainnya memiliki kelemahan dalam penggunaan memori saat jumlah data besar. Hal tersebut tentu berpengaruh terhadap banyaknya item yang diproses. *Association rules* dapat diketahui dengan 2 parameter, minimum *support* (persen jumlah dari kombinasi item dalam *database*) dan minimum *confidence* (kuatnya hubungan antar item dalam aturan asosiatif), dimana kedua parameter tersebut ditentukan oleh *user* [3]. Penggunaan algoritma apriori pada penelitian ini digunakan untuk menemukan *association rules* dalam prediksi kinerja karyawan yaitu menambang keterhubungan antara item-item yang terkandung dalam data karyawan.

Pada penelitian yang dilakukan oleh M. Afdal dan Muhammad Rosadi, pada tahun 2019 yang berjudul "Penerapan *Association Rule Mining* Untuk Analisis Penempatan Tata Letak Buku Di Perpustakaan Menggunakan Algoritma Apriori". Penelitian tersebut membuat model *apriori algorithm* menggunakan *Association Rule Mining* untuk mengatur penempatan buku dengan memperhatikan tingkat keseringan pengunjung dalam meminjam buku. Data yang digunakan adalah 11.550 transaksi peminjaman buku selama 3 tahun yang telah diproses menghasilkan 4 *rules* dengan kombinasi item terbesar adalah kategori buku agama dan ilmu sosial sering dipinjam secara bersamaan dengan nilai *support* 11,71% dan *confidence* 41,43%. Serta kategori buku teknologi dan ilmu sosial sering dipinjam secara bersamaan dengan nilai *support* 13,8% dan *confidence* 40,75% [8].

Penelitian lainnya terkait penerapan *Association Rule Mining* yaitu "*Association Rule Mining with Permutation for Estimating Students Performance and Its Smart Education System*" oleh Nongnuch Ketui, dkk. pada tahun 2019. Pada penelitian tersebut, menerapkan *Association Rule Mining (*ARM*)* untuk menemukan pola yang menarik antara berbagai prestasi akademik dalam dataset yang diperoleh. menggunakan 17.875 prestasi akademik dari 483 siswa. Hasil percobaan menunjukkan bahwa 248 *rules* pada *confidence* 0,2 dan *support* 0,7 sementara kinerja *rules* dengan set yang baru yaitu 76.00 % [5].

Berdasarkan penelitian-penelitian *Association Rule Mining* tersebut, algoritma Apriori dapat menangani hubungan antar item untuk melakukan prediksi kinerja karyawan, maka penelitian ini akan dilakukan dengan menerapkan *Association Rule Mining* dengan algoritma Apriori yang diharapkan menghasilkan prediksi yang akurat terhadap penelitian "*Predicting*

Employee's Performance with Apriori Algorithm" ini. Hasil keluaran dari implementasi teknik *Association Rule Mining* menggunakan algoritma apriori ini diharapkan dapat memenuhi minimum *support* 10 % dan minimum *confidence* 60%.

1.4 *Produce Project Plan*

Tahap perencanaan yang dilakukan untuk mencapai tujuan *data mining* dan mencapai tujuan bisnis pada penelitian "*Predicting Employee's Performance with Apriori Algorithm*" ini adalah sebagai berikut:

Tabel 1. Jadwal Pelaksanaan Proyek

Tahapan	Waktu	Sumber daya yang dibutuhkan	Kegiatan	Ketergantungan
Bussiness Understanding	3 hari	Semua *analysts*	Menentukan tujuan utama bisnis, melakukan penilaian terhadap situasi, menentukan tujuan *data mining*, dan membuat *project plan*.	Perkembangan penerapan konsep *data mining*
Data understanding	4 hari	Semua *analysts*	Mengumpulkan data yang akan digunakan, mendeskripsikan data, melakukan eksplorasi data, dan memverifikasi kualitas data.	Masalah data dan teknologi
Data preparation	2 minggu	*Data mining consultant*, beberapa *database analyst time*	Memilih data yang akan digunakan, membersihkan data dari *noise* atau *outlier*, membangun data, menggabungkan data, dan membuat format data.	Masalah data dan teknologi
Modelling	1 minggu	*Data mining consultant*, beberapa *database analyst time*	Memilih teknik pemodelan, membuat *Test Design*, membangun model, dan menilai model	Ketidakmampuan menemukan model yang tepat
Evaluation	3 hari	Semua *analysts*	Mengevaluasi hasil, meninjau proses, dan menentukan tahapan selanjutnya	Ketidakmampuan untuk menerapkan hasil, terjadi kesalahan pada proses pengerjaan proyek, perkembangan penerapan konsep *data mining*
Deployment	1 minggu	*Data mining consultant*, beberapa *database analyst time*	Membuat *plan deployment, Monitoring and Maintenance Plan* dan meninjau proyek	Ketidakmampuan untuk menerapkan hasil, perkembangan penerapan konsep *data mining*

Dalam pelaksanaan proyek dalam penelitian ini, diperlukan *tools data mining* yang mendukung metode untuk berbagai tahapan proses. *Tools* dan teknik yang digunakan dapat mempengaruhi keseluruhan proyek. *Tools* yang digunakan dalam mengerjakan proyek ini adalah *python*. *Python* adalah bahasa pemrograman berorientasi objek yang digunakan dalam pengembangan perangkat lunak maupun dalam analisis dan *data science*. *Python* memiliki berbagai *library* yang menyediakan fungsi untuk melakukan analisis data, memproses data, memvisualisasikan data, dll.

Python menyediakan *library* seperti *scikit-learn*, Keras, TensorFlow untuk membantu dalam pembuatan model *data mining* dengan cepat. Selain itu, terdapat juga *library* yang dapat digunakan untuk membagi *dataset* menjadi data *training* dan data *test*, misalnya menggunakan *cross-validation*. Metode atau algoritma yang akan digunakan dalam proyek ini adalah algoritma Apriori. Apriori merupakan algoritma yang menemukan *frequent itemset* (memenuhi *minimum support*) yang kemudian akan di-*generate* untuk mendapatkan *rule* yang memenuhi *minimum confidence* dari *frequent itemset* sebelumnya. Dalam pengimplementasian model dengan algoritma Apriori akan menggunakan *library* Apriori yang telah disediakan *python*.

BAB 2
DATA UNDERSTANDING

Tahap kedua pada metodologi CRISP-DM setelah *business understanding* untuk melakukan prediksi kinerja karyawan adalah data *understanding*. Pada bab ini akan dijelaskan mengenai pengumpulan data, pendeskripsian data untuk dapat memahami data yang akan digunakan dalam penelitian serta pengidentifikasian masalah yang berhubungan dengan data.

2.1 *Collect Initial Data*

Hal pertama yang dilakukan pada tahap data understanding adalah pengumpulan data yang merupakan langkah persiapan untuk menemukan data awal. Data berasal dari beberapa sumber yang relevan dengan tujuan proyek. *Dataset* yang akan digunakan untuk memprediksi kinerja dari *employee* adalah data *employee* yang dapat diakses pada link https://www.kaggle.com/shivan118/hranalysis?select=train.csv. *Dataset* tersebut memiliki format file CSV (*Comma Separated Values*) sehingga datanya bersifat statis.

2.2 *Describe Data*

Dataset yang digunakan untuk memprediksi kinerja dari karyawan menggunakan algoritma apriori adalah predicting-employee-performance dataset. Dataset ini berisi 14 atribut atau variabel dan memuat 54808 record. Dataset yang akan dianalisis pada proyek ini terdiri dari 8 atribut nominal, 8 atribut ordinal, 3 atribut binary, dan 9 atribut numerik. Berikut tabel untuk menjelaskan setiap atributnya. Berikut tabel untuk menjelaskan setiap atributnya.

Tabel 2. Tabel Atribut pada *Dataset*

No.	Nama atribut (variabel)	Tipe atribut	Deskripsi
1.	*employee_id*	Nominal	ID dari dari setiap karyawan (unik)
2.	*department*	Nominal	Nama departemen karyawan
3.	*region*	Nominal	Asal wilayah kerja
4.	*education*	Nominal	Tingkat pendidikan
5.	*gender*	Binary	Jenis kelamin karyawan
6.	*recruitment_channel*	Nominal	Kanal rekrutmen untuk karyawan
7.	*no_of_trainings*	Numerik	Total jumlah pelatihan yang diselesaikan pada tahun sebelumnya (seperti *soft kill*, *technical skills*, dll)
8.	*age*	Nominal	Usia karyawan
9.	*rating*	Ordinal	Peringkat karyawan

No.	Nama atribut (variabel)	Tipe atribut	Deskripsi
10.	*length_of_service*	Numerik	Lama bekerja dalam beberapa tahun
11.	*KPIs_met >8No%*	Binary	Jika persen KPI (Key Performance Indicators) lebih dari 80% => Yes, jika tidak => No
12.	*awards_won?*	Binary	Jika karyawan memenangkan penghargaan pada tahun sebelumnya => Yes, jika tidak => No
13.	*avg_training_score*	Numerik	Skor rata-rata dalam evaluasi pelatihan
14.	*is_promoted*	Binary	Jika karyawan direkomendasikan untuk promosi => Yes, jika tidak => No

2.3 *Explore Data*

Exploratory Data Analysis (EDA) diperlukan sebagai sebuah pendekatan dalam menganalisis dataset untuk meringkas karakteristik utama *dataset*. Biasanya dilakukan dengan menggunakan metode visual. EDA digunakan untuk memahami data, mendapatkan konteks data, memahami variabel dan hubungan di antara variabel, dan merumuskan hipotesis yang berguna dalam membangun model prediksi. Atribut atau fitur pada dataset tidak semua diperlukan dalam mengananalisis. Fitur atau atribut yang digunakan merupakan atribut yang relevan dan sesuai dengan tujuan proyek yaitu fitur yang berpengaruh pada kinerja karyawan serta faktor-faktor lainnya yang dapat meningkatkan kinerja karyawan. Variabel yang relevan terkait kinerja karyawan pada proyek ini adalah variabel *performance_rating* dan hubungannya dengan variabel lain yang relevan dalam dataset akan dipelajari secara mendalam. Terdapat beberapa variabel terkait yang diprediksi berpengaruh terhadap kinerja dari karyawan. Hipotesis-hipotesis tersebut dirumuskan sebagai berikut:

1) *Education* berpengaruh terhadap kinerja karyawan. Karyawan dengan *education* 'Master's & above' akan memiliki kinerja yang lebih baik.
2) *No_of_trainings* berpengaruh terhadap kinerja karyawan. Karyawan dengan jumlah training (*no_of_trainings*) yang lebih banyak akan memiliki kinerja yang lebih baik.
3) KPIs_met>80% berpengaruh terhadap kinerja karyawan. Karyawan dengan KPIs lebih dari 80% akan memiliki kinerja yang lebih baik.
4) *Awards_won* berpengaruh terhadap kinerja karyawan, karyawan yang memenangkan hadiah akan memiliki kinerja yang lebih baik.

5) Rata-rata nilai karyawan selama pelatihan berpengaruh terhadap kinerja karyawan. Karyawan dengan *avg_training_score* yang tinggi akan memiliki kinerja yang lebih baik juga.

Dari beberapa hipotesis yang dirumuskan, terdapat beberapa atribut yang relevan yang selanjutnya akan digunakan untuk menentukan kinerja dari karyawannya, yaitu variabel *education, no_of_trainings, KPIs_met >80%, awards_won, avg_training_score* berpengaruh terhadap kinerja dari karyawan tersebut.

Setelah dilakukan eksplorasi pada data tersebut, ditemukan karakteristik-karakteristik baru yang berpengaruh terhadap hipotesis yang sebelumnya telah dirumuskan. Hal ini membuat perlu mengidentifikasi kembali subset data yang relevan untuk digunakan pada tahapan selanjutnya yang sesuai dengan tujuan data mining task pada proyek ini, yaitu untuk melakukan prediksi terhadap kinerja karyawan dalam perusahaan.

2.4 *Verify Data Quality*

Tahapan ini berisi evaluasi dan kualitas data dan kelengkapan data yang digunakan. Terjadinya *error* maupun kesalahan ketika *input* data mengakibatkan terjadinya *missing value* maupun *noise* pada data. Pada tahapan ini dilakukan pemeriksaan atribut yang hilang atau kosong. *Data cleaning* diperlukan untuk menjaga kekonsistenan dan menghilangkan data tidak relevan. *Data cleaning* pada proses *data mining* dapat mengurangi jumlah dan kompleksitas data. Memperkirakan apakah semua *value* dan ejaan nilai-nilai rasional serta apakah fitur dengan *value* yang berbeda memiliki pengertian yang sama. Hasil penelusuran yang dilakukan menemukan:

a. Terdapat atribut yang memiliki nilai *null* atau kosong, seperti pada atribut “education” dan “rating”, terdapat *cell* yang memiliki nilai *null* atau kosong.
b. Format tipe data pada atribut “rating” dalam bentuk integer dan “is_promoted” yang dibuat dalam *binary* yang disajikan dalam angka akan menyebabkan kesulitan dalam memahami data untuk menemukan asosiasi.

BAB 3
DATA PREPARATION

Data preparation merupakan tahap setelah dilakukan pengumpulan data awal yang telah dilakukan pada fase crisp-dm sebelumnya, yaitu *bussiness understanding*. Pada tahap *data preparation* ini, dilakukan proses menyiapkan data awal, memilih variabel yang akan dianalisis dan membersihkan data. Dalam pengerjaan proyek, bahasa pemrograman yang digunakan adalah pemrograman *python* dengan *software* pengolah data **Jupyter Notebook.**

3.1 *Package*

Untuk dapat menjalankan beberapa kode program yang akan dijalankan, dibutuhkan beberapa *package* yang harus diinstal, yaitu:

1. **Pandas**, untuk memuat sebuah file ke dalam tabel virtual seperti *spreadsheet*, mengumpulkan data, dan mengolahnya.
2. **Numpy**, untuk operasi verktor dan matriks serta analisis data.
3. **Mlxtend,** *library* untuk *machine learning* dalam menganalisis data.
4. **Matplotlib**, untuk menyajikan data ke dalam visual yang lebih menarik dan rapi.
5. **Seaborn**, untuk menyajikan data ke dalam visualisasi data secara statistik (dibangun di atas matpotlib.
6. **Apriori**, untuk memuat algoritma apriori.
7. **Association_rules,** untuk membangun sebuah variabel yang memiliki *association rules* (aturan asosiasi).
8. **Image**, untuk menampilkan gambar.

3.2 *Dataset Description*

Pada fase ini, *dataset* akan dideskripsikan dengan memanfaatkan bahasa pemrograman python. Berikut beberapa fungsi yang dijalankan dalam mendeskripsikan *dataset* tersebut:

1. **.describe()**, untuk menampilkan berbagai ringkasan atau deskripsi statistik data, seperti jumlah data di setiap kolom (*count*), rata-rata nilai per kolom (*mean*), standard deviasi (*std*), nilai minimun (*min*), nilai maksimum (*max*), serta batas nilai dari masing-masing kuartil (25%, 50%, 75%). Berikut beberapa ringkasan atau deskripsi statistik data pada atribut yang bertipe data numerik.

```
1 employee.describe()
```

	no_of_trainings	age	length_of_service	avg_training_score
count	54808.000000	54808.000000	54808.000000	54808.000000
mean	1.195282	34.803915	5.865512	63.386750
std	0.556788	7.660169	4.265094	13.371559
min	1.000000	20.000000	1.000000	39.000000
25%	1.000000	29.000000	3.000000	51.000000
50%	1.000000	33.000000	5.000000	60.000000
75%	1.000000	39.000000	7.000000	76.000000
max	10.000000	60.000000	37.000000	99.000000

***Gambar 2.* Fungsi *.describe()* pada Atribut bertipe data Numerik**

Jika atribut yang ingin diketahui informasi deskriptif data statistiknya tidak bertipe data numerik, maka informasi yang diberikan adalah *count* (jumlah data setiap kolom), *unique* (jumlah data yang unik), *top* (data dengan nilai paling banyak atau paling besar), dan *freq* (jumlah data yang paling banyak atau yang paling besar). Berikut penerapan fungsi *.describe()* pada atribut *performance_rating*.

```
1 employee['performance_rating'].describe()
```

```
count      50684
unique         5
top        Great
freq       18618
Name: performance_rating, dtype: object
```

Gambar 3. Fungsi *.describe()* pada Atribut *performance_rating*

2. **.shape()**, untuk menampilkan dimensi data (jumlah baris dan kolom). Berikut keluaran data yang diperoleh dari fungsi .shape tersebut.

```
1 employee.shape
```

```
(54808, 13)
```

Gambar 4. Fungsi *.shape()*

Pada gambar 4 diatas menunjukkan bahwa *dataset* yang digunakan memiliki 54808 baris dan 13 kolom.

3. **.info()**, untuk menampilkan gambaran mengenai *dataset*.

```
employee.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 54808 entries, 65438 to 51526
Data columns (total 13 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   department            54808 non-null  object
 1   region                54808 non-null  object
 2   education             52410 non-null  object
 3   gender                54808 non-null  object
 4   recruitment_channel   54808 non-null  object
 5   no_of_trainings       54808 non-null  int64
 6   age                   54808 non-null  int64
 7   performance_rating    50684 non-null  object
 8   length_of_service     54808 non-null  int64
 9   KPIs_met >80%         54808 non-null  object
 10  awards_won?           54808 non-null  object
 11  avg_training_score    54808 non-null  int64
 12  is_promoted           54808 non-null  object
dtypes: int64(4), object(9)
memory usage: 5.9+ MB
```

Gambar 5. Fungsi *.info()*

4. **.head()**, untuk melihat 5 sampel data teratas.

```
employee.head()
```

	employee_id	education	no_of_trainings	performance_rating	length_of_service	KPIs_met >80%	awards_won?	avg_training_score
0	65438	Master's & above	1	Outstanding	8	Yes	No	49
1	65141	Bachelor's	1	Outstanding	4	Yes	No	60
2	7513	Bachelor's	1	Great	7	No	No	50
3	2542	Below Secondary	2	Low	10	No	No	50
4	48945	Bachelor's	1	Great	2	No	No	73

Gambar 6. Fungsi *.head()*

3.3 *Clean Data*

Pada fase ini dilakukan pembersihan data. *Data cleaning* yang dilakukan adalah dengan cara menghapus objek data yang tidak mengandung nilai (*missing value*).

```
#check missing value

employee.isna()
```

	employee_id	department	region	education	gender	recruitment_channel	no_of_trainings	age	performance_rating	length_of_service	KPIs_met >80%	awa
0	False	False	False	False	False	False	False	False	False	False	False	
1	False	False	False	False	False	False	False	False	False	False	False	
2	False	False	False	False	False	False	False	False	False	False	False	
3	False	False	False	False	False	False	False	False	False	False	False	
4	False	False	False	False	False	False	False	False	False	False	False	
...	...	...	...	...	...	...	...	...	...	...	...	
54803	False	False	False	False	False	False	False	False	False	False	False	
54804	False	False	False	False	False	False	False	False	False	False	False	
54805	False	False	False	False	False	False	False	False	False	False	False	
54806	False	False	False	True	False	False	False	False	False	False	False	
54807	False	False	False	False	False	False	False	False	False	False	False	

54808 rows × 14 columns

Gambar 7. Fungsi *.isna()*

Fungsi **isna()** berguna untuk mengembalikan nilai *boolean* (True dan False). Jika *cell* berisi *value* "False", maka artinya *cell* tersebut tidak mengandung *missing value* dan sebaliknya, jika *cell* berisi *value* "True", maka *cell* tersebut mengandung *missing value*.

Untuk memudahkan dalam memahami data, maka perlu dilakukan agregasi data dengan fungsi **sum().** Dengan menggunakan fungsi **sum()**, maka akan diketahui berapa jumlah data yang *missing value* dan berasal dari atribut apa.

```
# Agregasi data untuk mengetahui jumlah cell yang hilang dan pada lokasi atribut
employee.isna().sum()
```

```
employee_id              0
department               0
region                   0
education             2398
gender                   0
recruitment_channel      0
no_of_trainings          0
age                      0
performance_rating    4124
length_of_service        0
KPIs_met >80%            0
awards_won?              0
avg_training_score       0
is_promoted              0
dtype: int64
```

Gambar 8. Fungsi *.isna().sum()*

Dari gambar 8 diatas menunjukkan bahwa terdapat *missing value* pada atribut *education* (sebanyak 2398 data) dan atribut *performance_rating* (sebanyak 4124 data).

Selanjutnya menggunakan fungsi .**dropna()** untuk menghilangkan data yang hilang. Setelah fungsi .**dropna()** dijalankan, maka data yang mengandung *missing value* terhapus. Pada gambar 9 menunjukkan bahwa baris (*record*) pada *dataset* telah berkurang.

```
[ ]:
1  # drop missing value
2  employee_dropna = employee.dropna()
3  print(employee_dropna.shape)
```

```
(48671, 14)
```

Gambar 9. Fungsi *.dropna()*

```
1  #recheck missing value
2  employee_dropna.isna().sum()
```

```
employee_id           0
department            0
region                0
education             0
gender                0
recruitment_channel   0
no_of_trainings       0
age                   0
performance_rating    0
length_of_service     0
KPIs_met >80%         0
awards_won?           0
avg_training_score    0
is_promoted           0
dtype: int64
```

Gambar 10. Pengecekkan Kembali *Missing Value* setelah Fungsi .dropna() dijalankan

3.4 *Select Data (Data Reduction)*

Untuk meningkatkan efisiensi *data mining*, maka perlu memilih beberapa bagian dari *dataset* yang diperlukan sehingga *dataset* yang diproses lebih sedikit. Pada fase *select data* ini, yang dilakukan adalah mengurangi atribut yang tidak relevan. Atribut yang tidak relevan adalah atribut *department*, *region*, *gender*, *recruiment_channel*, dan *age*.

```
1  # Drop irrelevant attributes
2  employee = employee.drop('department', axis=1)
3  employee = employee.drop('region', axis=1)
4  employee = employee.drop('gender', axis=1)
5  employee = employee.drop('recruitment_channel', axis=1)
6  employee = employee.drop('age', axis=1)
```

Gambar 11. Penghapusan Atribut yang Tidak Relevan

BAB 4
MODELLING

Tahap keempat pada metodologi CRISP-DM untuk melakukan prediksi kinerja karyawan adalah *modeling*. Pada bab ini akan dijelaskan mengenai pemilihan teknik modelling, menghasilkan *test design*, membangun model, dan menilai model yang telah dibangun.

4.1 *Select Modeling Technique*

Teknik pemodelan yang digunakan didorong oleh tujuan penambangan data yang ingin dicapai oleh perusahaan. Secara sederhana, hal tersebut melibatkan penerapan *association rule* untuk mendapatkan keterkaitan antar variabel dalam menentukan kinerja karyawan dalam suatu perusahaan. Data karyawan dalam kurun waktu tertentu akan ditambahkan untuk mendapatkan hasil yang lebih baik. Penerapan *association rule* dengan algoritma apriori cocok digunakan untuk jenis pemodelan tersebut. Algoritma Apriori akan menggenerate *association rule* dengan menemukan *frequent item set* yang memenuhi minimum *support* (persentase kombinasi item dalam *dataset*) dan minimum *confidence* (keterkaitan antar item) yang diinisialisasi oleh penambang data untuk menentukan faktor-faktor yang berpengaruh dalam menentukan kinerja karyawan.

Dalam menggenerate *association rule*, algoritma apriori tidak memerlukan data untuk di-*split* ke dalam *test sets* dan *training sets*. Oleh karena itu, penambangan data hanya dilakukan dengan sebuah *dataset* yang memiliki data yang cukup untuk memperoleh hasil yang dapat dipercaya terhadap model yang digunakan. Hal ini berarti, *dataset* menggambarkan data secara keseluruhan dengan baik sehingga algoritma apriori nantinya akan dapat menghasilkan *rule* yang sesuai dalam menentukan kinerja karyawan.

Dalam teknik pemodelan dengan *association rule* memerlukan asumsi spesifik terhadap data, yaitu semua atribut memiliki distribusi yang sama, tidak ada *missing value*. Untuk atribut yang tidak *categorical* (*nominal*) maka akan dilakukan pembuatan bin terlebih dahulu sebelum dilakukan penerapan algoritma apriori pada data tersebut.

4.2 *Generate Test Design*

Sebelum melakukan pembangunan model, perlu dilakukan perancangan terhadap bagaimana model akan diuji. Terdapat 2 cara untuk menghasilkan *test design* yang komprehensif:

- Menggambarkan kriteria "*goodness*" dari sebuah model
- Mendefinisikan data dengan kriteria yang akan diuji.

“goodness” algoritma apriori dapat diukur dengan menggunakan *interestingness measure* (ukuran ketertarikan) yang didapatkan dari hasil pengolahan data dengan perhitungan tertentu. Beberapa ukuran tersebut adalah:

1) *Support*

   Support adalah nilai penunjang atau persentase kombinasi sebuah item dalam *dataset*.

2) *Confidence*

   Confidence adalah nilai kepastian yaitu kuatnya hubungan antar-item dalam *association rule*. *Confidence* dapat dicari setelah pola frekuensi munculnya sebuah item ditemukan.

3) *Lift*

   Lift adalah nilai yang mengukur besarnya hubungan antara *antecedent* dan *consequent* yang tidak saling bergantung (*independent*). *Lift* memiliki range mulai dari 0 sampai dengan ∞. Nilai yang mendekati 1 mengindikasikan bahwa *antecedent* dan *consequent* tidak memiliki ketergantungan. Nilai yang jauh dari 1 mengindikasikan bahwa *antecedent* menyediakan informasi tentang *consequent*.

4) *Conviction*

   Conviction adalah nilai yang mengukur tingkat implikasi dari suatu aturan. *Conviction* sangat memperhatikan arah dari suatu *association rule*. *Conviction* mengindikasikan bahwa *conviction*$(A \rightarrow B) \neq$ *conviction*$(B \rightarrow A)$

5) *Leverage*

   Leverage adalah nilai yang mengukur banyaknya item *antecedent* dan *consequence* yang dijual secara bersamaan dalam suatu *dataset* yang lebih dari yang diharapkan. Nilai 0 menunjukkan *antecedent* dan *consequent independent*. *Leverage* memiliki jangkauan nilai dari -0,25 sampai dengan 0,25.

Kriteria model yang dinilai bergantung pada *data mining goals* pada model yang akan dibangun. Tidak ada cara objektif untuk menilai model sampai disajikan pada HRD secara langsung. Namun HRD memerlukan aturan yang menghasilkan prediksi terhadap kinerja karyawan. Berikut adalah *test design* yang dibuat untuk model yang dibangun.

(1) Melakukan pengukuran terhadap ukuran ketertarikan dari *association rule* yang telah ditemukan.
(2) Menentukan keakuratan *rule* yang dihasilkan berdasarkan ukuran ketertarikan yang diperoleh dengan menyesuaikan terhadap nilai parameter yang dinyatakan berhasil.

(Misalnya, *minimum confidence* = 0.5, *lift* 1 yang menandakan bahwa *antecedents* dan *consequents* berkolerasi positif, dan lain-lain).

4.3 *Build Model*

Dalam proses pembangunan model, terdapat 3 informasi yang digunakan dalam pengambilan keputusan *data mining*, yaitu:

- *Parameter settings*, merupakan parameter yang digunakan pada model yang memberikan hasil yang baik.
- Model dihasilkan.
- Deskripsi dari hasil pemodelan, termasuk *performance* dan *data issued* yang terjadi selama pengeksekusian model dan eksplorasi hasilnya.

Association rule dengan algoritma apriori dibangun pada bahasa pemrograman *python* dengan memanfaatkan *library python* untuk algoritma apriori, yaitu mlxtend. Mlxtend adalah *library python* yang mengimplementasikan algoritma dalam *machine learning*. Dari *library* mlxtend menggunakan *frequent patterns* untuk mengimport apriori dan *association rule* yang digunakan untuk menggenerate *association rule* dengan algoritma apriori dari *frequent itemset* yang dihasilkan. Dalam pembangunan model tersebut menggunakan *dataset* yang telah melalui tahapan *preprocessing* seperti yang dijelaskan pada asumsi di bagian 4.1 sebelumnya. Dimana atribut dengan tipe bukan kategorikal akan dibuat ke dalam beberapa bin. Seperti variabel *length_of_service*, *avg_training_score* akan dibagi ke dalam 4 bin.

```
#create bins for length_of_service
pd.qcut(employee['length_of_service'], q=4)
```

```
0          (5.0, 8.0]
1          (3.0, 5.0]
2          (5.0, 8.0]
3         (8.0, 37.0]
4        (0.999, 3.0]
             ...
54802    (0.999, 3.0]
54803     (8.0, 37.0]
54804      (5.0, 8.0]
54805    (0.999, 3.0]
54807      (3.0, 5.0]
Name: length_of_service, Length: 48671, dtype: category
Categories (4, interval[float64]): [(0.999, 3.0] < (3.0, 5.
0] < (5.0, 8.0] < (8.0, 37.0]]
```

Gambar 12. Pembuatan *Bin* untuk Atribut *length_of_service*

```
#create bins for avg_training_score
pd.qcut(employee['avg_training_score'], q=4)
```

```
0        (38.999, 51.0]
1          (51.0, 60.0]
2        (38.999, 51.0]
3        (38.999, 51.0]
4          (60.0, 76.0]
              ...
54802    (38.999, 51.0]
54803      (76.0, 99.0]
54804      (51.0, 60.0]
54805      (76.0, 99.0]
54807    (38.999, 51.0]
Name: avg_training_score, Length: 48671, dtype: category
Categories (4, interval[float64]): [(38.999, 51.0] < (51.0,
60.0] < (60.0, 76.0] < (76.0, 99.0]]
```

Gambar 13. Pembuatan Bin untuk Atribut *avg_training_score*

Setiap nilai pada kolom *length_of_service* dan *avg_training_score* akan diganti dengan nilai bin yang telah dibuat. Berikut adalah tampilannya sebelum dilakukan pembuatan *bin* dan setelah pembuatan *bin* dilakukan.

length_of_service	KPIs_met >80%	awards_won?	avg_training_score	is_promoted
8	Yes	No	49	No
4	Yes	No	60	No
7	No	No	50	No
10	No	No	50	No
2	No	No	73	No

Gambar 14. Nilai Atribut *length_of_service* dan *avg_training_score* Sebelum Dibuat ke Dalam Bin

aining_score	is_promoted	length_of_service_range	avg_training_score_range
49	No	>5 <=8	<=51
60	No	>3 <=5	>51 <=60
50	No	>5 <=8	<=51
50	No	>8	<=51
73	No	<=8	>60 <=76

Gambar 15. Nilai Atribut *length_of_service* dan *avg_training_score* Setelah Dibuat ke Dalam Bin

Dalam penerapan algoritma apriori terdapat beberapa parameter yang perlu ditentukan. Variabel-variabel yang akan dilihat keterkaitannya satu sama lain perlu diinisialisasi, sehingga dalam penerapan algoritma apriori hanya akan menggunakan kolom-kolom yang telah diinisialisasi untuk menggenerate *association rule*. Adapun dalam projek ini menggunakan 7 kolom dari 14 kolom yang ada di *dataset* yaitu, *performance_rating, education, no_of_trainings, length_of_service_range, KPIs_met > 80%, awards_won?,* dan *avg_training_score_range*. Selain itu, terdapat beberapa parameter lainnya yang perlu ditentukan sebelum algoritma apriori dijalankan, yaitu:

- *Minimum support* = 0.1
- *Minimum confidence* = 0.6
- Maksimum *frequent itemset* yang dihasilkan = 3

Paramater-parameter tersebut diperoleh berdasarkan ukuran ketertarikan dari setiap *association rule* yang dihasilkan dengan algoritma apriori.

Setelah parameter ditentukan, maka akan dilakukan penerapan algoritma apriori dengan menentukan targetnya sebagai *performance rating* (*low, good, great, excellent, outstanding*).

(i) Berikut adalah penerapan algoritma apriori dengan targetnya adalah performance_rating_Low.

```
#for performance_rating Low

#Apriori min support
min_support = 0.1

#Max lenght of apriori n-grams
max_len = 3

frequent_items = apriori(employee, use_colnames=True,
                         min_support=min_support, max_len=max_len + 1)
rules = association_rules(frequent_items, metric='lift', min_threshold=1)

target = '{\'performance_rating_Low\'}'

results_performance_rating_low = rules[rules['consequents'].astype(str).str.contains(target, na=False)].
                                    sort_values(by='confidence', ascending=False)
results_performance_rating_low
```

Gambar 16. Kode Program Penerapan Algoritma Apriori untuk *Performance Rating Low*

Model tersebut menghasilkan 4 *association rule* yang memenuhi *minimum support* dan *minimum confidence*, yaitu:

Tabel 3. ***Association Rule Performance Rating Low***

Antecedents	***Consequents***
KPIs_met > 80%_No, awards_won?_No, education_Below Secondary	Performance_rating_Low
KPIs_met > 80%_No, education_Below Secondary	Performance_rating_Low
awards_won?_No, education_Below Secondary	Performance_rating_Low
education_Below Secondary	Performance_rating_Low

(ii) Berikut adalah penerapan algoritma apriori dengan targetnya adalah performance_rating_Good.

```
#for performance_rating Good

#Apriori min support
min_support = 0.1

#Max lenght of apriori n-grams
max_len = 3

frequent_items = apriori(employee, use_colnames=True,
                         min_support=min_support, max_len=max_len + 1)
rules = association_rules(frequent_items, metric='lift', min_threshold=1)

target = '{\'performance_rating_Good\'}'

results_performance_rating_good = rules[rules['consequents'].astype(str).str.contains(target, na=False)]
                                  .sort_values(by='confidence', ascending=False)
results_performance_rating_good
```

Gambar 17. Kode Program Penerapan Algoritma Apriori untuk *Performance Rating Good*

Model tersebut tidak menghasilkan *rule* yang memenuhi *minimum support* dan *minimum confidence*. Oleh karena itu, berikut ditampilakn *rule* yang dihasilkan ketika *minimum support* diubah menjadi 0.01.

```
#for performance_rating Good

#Apriori min support
min_support = 0.01

#Max lenght of apriori n-grams
max_len = 3

frequent_items = apriori(employee, use_colnames=True,
                         min_support=min_support, max_len=max_len + 1)
rules = association_rules(frequent_items, metric='lift', min_threshold=1)

target = '{\'performance_rating_Good\'}'

results_performance_rating_good = rules[rules['consequents'].astype(str).str.contains(target, na=False)]
                                  .sort_values(by='confidence', ascending=False)
results_performance_rating_good
```

Gambar 18. Kode Program Penerapan Algoritma Apriori untuk *Performance Rating Good* (*support* diubah)

Model tersebut menghasilkan beberapa *association rule*, diantaranya yaitu :

Tabel 4. *Association Rule Performace Rating Good*

Antecedents	***Consequents***
KPIs_met > 80%_No, awards_won?_No, education_Master's & above	performance_rating_Good
KPIs_met > 80%_No, education_Master's & above	performance_rating_Good
KPIs_met > 80%_No, no_of_trainings_1, education_Master's & Above	performance_rating_Good

Namun, dari semua *rule* tersebut tidak memenuhi *minimum support* dan *minimum confidence* yang telah didefinisikan.

(iii) Berikut adalah penerapan algoritma apriori dengan targetnya adalah performance_rating_Great.

```
#for performance_rating Great

#Apriori min support
min_support = 0.1

#Max lenght of apriori n-grams
max_len = 3

frequent_items = apriori(employee, use_colnames=True,
                         min_support=min_support, max_len=max_len + 1)
rules = association_rules(frequent_items, metric='lift', min_threshold=1)

target = '{\'performance_rating_Great\'}'

results_performance_rating_great = rules[rules['consequents'].astype(str).str.contains(target, na=False)]
                                   .sort_values(by='confidence', ascending=False)
results_performance_rating_great.head(15)
```

Gambar 19. Kode Program Penerapan Algoritma Apriori untuk *Performance Rating Great*

Model tersebut menghasilkan 10 *association rule* yang memenuhi *minimum support* dan *minimum confidence*, yaitu:

Tabel 5. *Association Rule Performance Rating Great*

Antecedents	***Consequents***
Education_Bachelor's, KPIs_met > 80%_No, avg_training_score > 60 <=76	performance_rating_Great
Awards_won?_No, education_Bachelor's, KPIs_met > 80%_No	performance_rating_Great
Education_Bachelor's, KPIs_met > 80%_No	performance_rating_Great
No_of_trainings_1, education_Bachelor's, KPIs_met > 80%_No	performance_rating_Great
Awards_won?_No, KPIs_met > 80%_No, avg_training_score > 60 <=76	performance_rating_Great
KPIs_met > 80%_No, avg_training_score > 60 <=76	performance_rating_Great
Awards_won?_No, KPIs_met > 80%_No	performance_rating_Great
KPIs_met > 80%_No	performance_rating_Great
No_of_trainings_1, Awards_won?_No, KPIs_met > 80%_No	performance_rating_Great
No_of_trainings_1, KPIs_met > 80%_No	performance_rating_Great

(iv) Berikut adalah penerapan algoritma apriori dengan targetnya adalah performance_rating_Excellent.

```
#for performance_rating Excellent

#Apriori min support
min_support = 0.1

#Max lenght of apriori n-grams
max_len = 3

frequent_items = apriori(employee, use_colnames=True,
                         min_support=min_support, max_len=max_len + 1)
rules = association_rules(frequent_items, metric='lift', min_threshold=1)

target = '{\'performance_rating_Excellent\'}'

results_performance_rating_excellent = rules[rules['consequents'].astype(str).str.contains(target, na=False)]
                                       .sort_values(by='confidence', ascending=False)
results_performance_rating_excellent
```

Gambar 20. Kode Program Penerapan Algoritma Apriori untuk *Performance Rating Excellent*

Model tersebut menghasilkan 7 *association rule* yang memenuhi *minimum support* dan *minimum confidence*, yaitu:

Tabel 6. *Association Rule Performance Rating Excellent*

Antecedents	***Consequents***
Awards_won?_Yes, KPIs_met>80%_Yes, no_of_trainings_1	performance_rating_Excellent
Education_Bachelor's, awards_won?_Yes, KPIs_met > 80%_Yes	performance_rating_Excellent
Awards_won?_Yes, KPIs_met > 80%_Yes	performance_rating_Excellent
Education_Bachelor's, awards_won?_Yes, no_of_trainings_1	performance_rating_Excellent
awards_won?_Yes, no_of_trainings_1	performance_rating_Excellent
Education_Bachelor's, awards_won?_Yes	performance_rating_Excellent
Awards_won?_Yes	performance_rating_Excellent

(v) Berikut adalah penerapan algoritma apriori dengan targetnya adalah performance_rating_Outstanding.

```
#for performance_rating Outstanding

#Apriori min support
min_support = 0.1

#Max lenght of apriori n-grams
max_len = 3

frequent_items = apriori(employee, use_colnames=True,
                         min_support=min_support, max_len=max_len + 1)
rules = association_rules(frequent_items, metric='lift', min_threshold=1)

target = '{\'performance_rating_Outstanding\'}'

results_performance_rating_outstanding = rules[rules['consequents'].astype(str).str.contains(target, na=False)]
                                         .sort_values(by='confidence', ascending=False)
results_performance_rating_outstanding
```

Gambar 21. Kode Program Penerapan Algoritma Apriori untuk *Performance Rating Outstanding*

Model tersebut menghasilkan 3 *association rule* yang memenuhi *minimum support* dan *minimum confidence*, yaitu:

Tabel 7. ***Association Rule Performance Rating Outstanding***

Antecedents	***Consequents***
KPIs_met>80%_Yes, awards_won?_No, education Bachelor's	performance_rating_outstanding
No_of_trainigs_1, KPIs_met>80%_Yes, awards_won?_No	performance_rating_outstanding
KPIs_met>80%_Yes, awards_won?_No	performance_rating_outstanding

4.4 *Assess Model*

Assess model merupakan tahapan yang dilakukan untuk menilai kesesuaian model yang telah dibangun dengan kriteria sukses yang telah didefinisikan. Secara umum, hasil yang diperoleh dari pembangunan model dengan menggunakan algoritma apriori telah menghasilkan *rule* yang baik yang dinilai berdasarkan kesesuaiannya dengan minimum support dan minimum confidence. Diperoleh beberapa *rule* yang berkaitan dalam menentukan kinerja karyawan (*performance rating*) *low*, *great*, *excellent*, dan *outstanding*. Namun untuk *performance rating* dengan kategori *good* tidak diperoleh *rule* yang memenuhi *minimum support* dan *minimum confidence*.

Berikut adalah rangkuman jumlah *rule* yang diperoleh dari setiap kategori *performance rating*, serta rentang *confidence* dari *rule* yang dihasilkan tersebut.

Tabel 8. Rangkuman Jumlah *Rule*

Performance rating	**Jumlah *rule***	***Range confidence***
Low	4	0.95 – 0.96
Great	10	0.62 – 0.86
Excellent	7	0.88 – 0.95
Outstanding	3	0.70 – 0.74

Berdasarkan Tabel 8 diatas dapat dilihat bahwa model yang dibangun telah cukup baik dalam menggenerate *association rule* dalam menentukan kinerja karyawan berdasarkan *performance rating*.

BAB 5
EVALUATION

Pada tahap *Evaluation* (Evaluasi), akan dijelaskan mengenai evaluasi terhadap model untuk memprediksi kinerja karyawan yang dihasilkan dengan menggunakan algoritma Apriori. Evaluasi adalah fase interpretasi terhadap hasil *data mining*. Evaluasi dilakukan secara mendalam dengan tujuan agar hasil pada tahap *modelling* sesuai dengan sasaran yang ingin dicapai dalam tahap *business understanding*.

5.1 Evaluate Results

Tahap ini dilakukan untuk mengetahui performa *association rule* dengan menggunakan dataset yang diperoleh serta menghitung *confidence*-nya. Dari pemodelan yang dilakukan pada tahap sebelumnya, dilakukan implementasi menggunakan bahasa pemograman python. Hasil berdasarkan *rules* yang terbentuk dari *antecedents* dan *consequents* relatif berbeda. Berikut hasil *rules* yang memenuhi *minimum support* = 0.1, *minimum confidence* = 0.6 dan maksimum *frequent itemset* yang dihasilkan = 3.

Tabel 9. *Rules* yang memenuhi Parameter

Performance rating	**Jumlah *rule***	***Rules***	***Support***	***Confidence***
Low	4	[KPIs_met > 80%_No, awards_won?_No, education_Below Secondary] => [performance_rating_low]	0.104128	0.966807
		[KPIs_met > 80%_No, education_Below Secondary]=> [performance_rating_low]	0.104128	0.966622
		[awards_won?_No, education_Below Secondary]=> [performance_rating_low]	0.117853	0.959037
		education_Below Secondary	0.117894	0.956493
Great	10	[Education_Bachelor's, avg_training_score > 60 <=76] => [performance_rating_great]	0.101806	0.867320
		[Awards_won?_No, education_Bachelor's, KPIs_met > 80%_No] => [performance_rating_great]	0.354955	0.842937
		[Education_Bachelor's, KPIs_met > 80%_No] => [performance_rating_great]	0.362865	0.835430

		[No_of_trainings_1, education_Bachelor's, KPIs_met > 80%_No] => [performance_rating_great]	0.283310	0.825293
		[Awards_won?_No, KPIs_met > 80%_No, avg_training_score > 60 <=76] => [performance_rating_great]	0.100902	0.701471
		[KPIs_met > 80%_No, avg_training_score > 60 <=76] => [performance_rating_great]	0.102936	0.699037
		[Awards_won?_No, KPIs_met > 80%_No] => [performance_rating_great]	0.358098	0.632769
		[KPIs_met > 80%_No] => [performance_rating_great]	0.366009	0.631143
		[No_of_trainings_1, Awards_won?_No, KPIs_met > 80%_No] => [performance_rating_great]	0.279592	0.623220
		[No_of_trainings_1, KPIs_met > 80%_No] => [performance_rating_great]	0.285673	0.620382
Excellent	7	[Awards_won?_Yes, KPIs_met>80%_Yes, no_of_trainings_1] => [performance_rating_excellent]	0.109655	0.957309
		[Education_Bachelor's, awards_won?_Yes, KPIs_met > 80%_Yes] => [performance_rating_excellent]	0.109346	0.956506
		[Awards_won?_Yes, KPIs_met > 80%_Yes] => [performance_rating_excellent]	0.114442	0.955075
		[Education_Bachelor's, awards_won?_Yes, no_of_trainings_1] => [performance_rating_excellent]	0.108257	0.897768
		[awards_won?_Yes, no_of_trainings_1] => [performance_rating_excellent]	0.113353	0.896782
		[Education_Bachelor's, awards_won?_Yes] => [performance_rating_excellent]	0.112901	0.885006
		[Awards_won?_Yes] => [performance_rating_excellent]	0.118222	0.883464
Outstanding	3	[KPIs_met>80%_Yes, awards_won?_No, education Bachelor's] => [performance_rating_outstanding]	0.206283	0.740140

		[No_of_trainigs_1, KPIs_met>80%_Yes, awards_won?_No] => [performance_rating_outstanding]	0.201927	0.706745
		[KPIs_met>80%_Yes, awards_won?_No] => [performance_rating_outstanding]	0.210331	0.700493

Berdasarkan hasil yang diperoleh dari pembangunan model dengan menggunakan algoritma apriori telah menghasilkan *rule* yang baik yang dapat dinilai berdasarkan kesesuaiannya dengan parameter yang telah ditentukan termasuk *minimum support* dan *minimum confidence.* Beberapa *rule* yang diperoleh terkait dalam memprediksi kinerja karyawan (*performance rating*), *rule* yang memenuhi parameter yang diberikan yaitu *performance rating* dengan kategori *low*, *great*, *excellent*, dan *outstanding* sedangkan *performance rating* dengan kategori *good* tidak memiliki *rule* yang memenuhi *minimum support* dan *minimum confidence*. Dengan demikian, model yang dibangun telah cukup baik dalam menerapkan algoritma Apriori untuk memprediksi kinerja karyawan.

5.2 Evaluate Process

Tahap ini memeriksa kembali tahapan dari awal untuk memastikan bahwa tidak ada faktor penting dalam proses tersebut yang terabaikan atau terlewati. Berdasarkan hasil peninjauan proses awal proyek *data mining* dengan metodologi CRISP-DM, maka dapat dipahami bahwa:

- Proses eksplorasi data akan membantu dalam memilih atribut yang berkaitan dengan memprediksi kinerja karyawan (*performance rating*).
- *Data Preparation*, khususnya pada proses *data cleaning* dan *data reduction,* sehingga data yang diperoleh dapat menghasilkan model yang baik.
- Sangat penting untuk tetap fokus pada masalah bisnis yang dihadapi, karena setelah data siap di analisis, maka akan dilakukan tahap pemodelan. *Business understanding* sangat penting dalam memutuskan bagaimana menerapkan hasil yang diperlukan dalam memprediksi kinerja karyawan (*performance rating*).

5.3 Determine Next Steps

Tahapan ini menentukan langkah apa yang akan diambil selanjutnya. Berdasarkan hasil evaluasi terhadap model yang digunakan dengan algoritma Apriori, dengan *rules* yang diperoleh yang memenuhi parameter yang telah ditentukan termasuk *minimum support* dan *minimum confidence* ketika men-*generate association rules* dalam memprediksi kinerja

karyawan (*performance rating*), maka diputuskan pengerjaan proyek akan dilanjutkan ke tahap akhir yakni *deployment*.

BAB 6
DEPLOYMENT

Tahap keenam pada metodologi CRISP-DM untuk melakukan prediksi kinerja karyawan adalah *deployment*. Pada bab ini akan dijelaskan mengenai perencanaan fase penyebaran atau penggunaan model yang sudah dihasilkan, perencanaan pemantauan dan pemeliharan serta

6.1 *Plan Deployment*

Pada fase *plan deployment* ini, model yang telah terbentuk pada fase *modelling* akan digunakan sesuai dengan tujuan *data mining* yang dibutuhkan. Penggunaan model yang telah dihasilkan akan memerlukan *dataset* yang sesuai dengan tujuan penggunaannya. Pada kasus proyek ini algoritma apriori akan digunakan sesuai data karyawan yang sudah diperbaharui secara *real time*. Data yang sudah diperbaharui tersebut akan digunakan untuk memprediksi kinerja karyawan menggunakan model yang sudah dirancang. Namun, jika *dataset* yang akan digunakan masih kotor atau terdapat *record* yang tidak memiliki nilai (*missing value*) serta terdapat beberapa variabel yang tidak dibutuhkan untuk memprediksi kinerja karyawan, maka *dataset* tersebut harus dibersihkan terlebih dahulu (*data preprocessing*) sesuai penjelasan pada bab 3. Sehingga proses pemodelan nantinya akan berjalan dengan baik dengan spesifik atribut atau parameter yang memiliki distribusi yang sesuai. Selanjutnya *dataset* tersebut akan diproses sesuai dengan jenis tipe datanya dan akan diproses menggunakan model yang telah dihasilkan. Dari penggunaan model tersebut, maka akan dihasilkan beberapa *rule* sesuai dengan kebutuhan objek (HR) yang dibutuhkan.

6.2 *Plan Monitoring and Maintenance*

Dalam *monitoring* dan *maintenance* adalah untuk menentukan apakah prediksi yang digunakan dengan teknik *association rules* benar-benar berfungsi. Artinya apakah *association rules* yang di-*generate* menghasilkan *rules* yang dapat memenuhi parameter yang telah ditentukan termasuk *minimum support* dan *minimum confidence.* Apakah atribut yang digunakan tepat sehingga memenuhi parameter yang telah ditentukan termasuk *minimum support* dan *minimum confidence.* Dikarenakan proyek yang dilakukan di masa depan dapat menghasilkan model yang lebih kompleks, maka *monitoring* akan ditingkatkan. Alternatif yang memungkinkan adalah dengan mencoba pembuatan model untuk prediksi dengan tepat dan akurat yang sangat dibutuhkan dalam pengerjaan proyek *data mining.*

6.3 *Produce Final Report*

Pada akhir proyek, tim proyek membuat laporan akhir dari penambangan data yang telah dilakukan. *Report* tersebut mencakup ringkasan dari proyek yang dilakukan, *deliverables* yang dihasilkan dari proyek, dan mengorganisir hasil yang diperoleh untuk disampaikan kepada *audience*. Dalam proyek ini, *final report* yang dimaksud mencakup dokumen pengerjaan proyek yang mengikuti metode CRISP-DM, file presentasi mencakup langkah-langkah pengerjaan yang dilakukan sesuai metode CRISP-DM, poster, dan video presentasi untuk menyampaikan tahapan dan hasil yang diperoleh.

6.4 *Review Project*

Review project digunakan untuk menilai baik, buruknya projek yang telah dibangun, apa yang telah selesai dan yang perlu dilakukan perbaikan ke depannya. Dalam pengerjaan proyek ini, tim proyek terlibat dalam pengerjaan proyek dari awal hingga akhir sehingga mampu mendapatkan pemahaman lebih detail mengenai eksplorasi data pada dataset yang digunakan, tahapan pemrosesan data untuk mendapatkan data yang siap digunakan pada penerapan algoritma apriori, *association rule* dengan algoritma apriori dalam menentukan *performance rating*, serta melakukan evaluasi terhadap *rule* yang diperoleh. Selain itu, tim proyek juga mendapatkan pemahaman dengan menerapkan secara langsung bagaimana penerapan metode CRISP-DM dalam melakukan *data mining task*. Namun, proyek ini masih memiliki beberapa kekurangan dalam pengimplementasiannya, karena proyek ini masih hanya menemukan *pattern* atau pola-pola dengan *association rule* untuk menentukan kinerja karyawan (*performance rating*). Oleh karena itu, kedepannya perlu dilakukan perbaikan agar *association rule* yang telah ditemukan dapat diintegrasikan dengan sistem. Selain itu, untuk data yang digunakan dalam pembangunan *association rule* juga sebaiknya diperbaharui kedepannya untuk mendapatkan hasil yang lebih akurat.

Referensi

[1] Q. A. Al-Radaideh and E. Al Nagi, "Using Data Mining Techniques to Build a Classification Model for Predicting Employees Performance," *Int. J. Adv. Comput. Sci. Appl.*, vol. 3, no. 2, 2012.

[2] M. A. KAREEM and I. J. HUSSEIN, "The Impact of Human Resource Development on Employee Performance and Organizational Effectiveness," *Manag. Dyn. Knowl. Econ.*, vol. 7, no. 3, pp. 307–322, 2019.

[3] H. Jiawei, M. Kamber, and P. Jian, *Data mining: Data mining concepts and techniques*. 2014.

[4] D. T. Larose, *Discovering Knowledge in Data: An Introduction to Data Mining: Second Edition*, vol. 9780470908. 2014.

[5] N. Ketui, W. Wisomka, and K. Homjun, "Association Rule Mining with Permutation for Estimating Students Performance and Its Smart Education System," *J. Comput.*, vol. 30, no. 2, pp. 93–102, 2019.

[6] B. S. Hasugian, "Penerapan Metode Association Rule Untuk Menganalisa Pola Pemakaian Bahan Kimia Di Laboratorium Menggunakan Algoritma FP-Growth (Studi Kasus di Laboratorium Kimia PT. PLN (Persero) Sektor Pembangkitan Belawan Medan)," *Algoritm. J. Ilmu Komput. dan Inform.*, vol. 3, no. 2, pp. 56–69, 2019.

[7] N. R. Ardani and N. Fitriana, "Sistem Rekomendasi Pemesanan Sparepart Dengan Algoritma FP-Growth," *Semin. Nas. Apl. Teknol. Inf. Dan Multimed.*, vol. 5, no. 1, pp. 6–7, 2016.

[8] M. Afdal and M. Rosadi, "PENERAPAN ASSOCIATION RULE MINING UNTUK ANALISIS PENEMPATAN TATA LETAK BUKU DI PERPUSTAKAAN MENGGUNAKAN ALGORITMA APRIORI," *J. Ilm. Rekayasa dan Manaj. Sist. Inf.*, vol. 5, no. 1, pp. 99–108, 2019.