BAB II

LANDASAN TEORI

2.1 Pelatihan (Training), Data Mining, dan Machine Learning

2.1.1 Pelatihan (Training)

Untuk meningkatkan kompetensi atau kemampuan karyawan dalam menjalankan pekerjaan di perusahaan maka diperlukan adanya *training* karyawan. Berdasarkan pasal 1 ayat 9 Undang-undang No.13 tahun 2013 tentang ketenagakerjaan, *training* atau pelatihan kerja adalah "keseluruhan kegiatan untuk memberi, memeroleh, meningkatkan, serta mengembangkan kompetensi kerja, produktivitas, disiplin, sikap dan etos kerja pada tingkat keterampilan dan keahlian tertentu sesuai dengan jenjang dan kualifikasi jabatan dan pekerjaan". Berdasarkan definisi tersebut maka dapat disimpulkan bahwa *Training* karyawan diartikan sebagai suatu bentuk intervensi yang diberikan oleh perusahaan yang dirancang untuk meningkatkan kinerja individu atau tim melalui perubahan pengetahuan, keterampilan, atau sikap [22]. Kegiatan *training* karyawan dapat dilakukan secara *online* maupun *offline*. Pelatihan *online* dapat menjadi alternatif untuk melatih karyawan, terutama dalam hal mengembangkan keterampilan teknologi informasi dan komunikasi, sedangkan pelatihan *offline* dapat memberikan manfaat langsung dan interaksi sosial di antara karyawan [23]. *Training* karyawan secara *online* dapat memanfaatkan berbagai layanan penyedia pembelajaran *online* seperti Udemy, Percipio, dan juga Linkedin Learning.

2.1.1.1 Linkedin Learning

Linkedin Learning merupakan salah satu layanan penyedia *online learning* untuk kebutuhan *training* karyawan. Linkedin Learning adalah platform belajar *online* yang menjadi bagian

dari Linkedin sebagai media bagi para profesional yang dirilis pada tahun 2003 [24]. Linkedin Learning menawarkan ribuan *course* dan pelatihan video berkualitas tinggi dalam berbagai topik, termasuk keterampilan teknis, manajemen, pemasaran, dan banyak topik lain yang tersedia. Pengguna Linkedin Learning dapat berupa perorangan maupun perusahaan yang dapat memilih dari berbagai kursus yang disusun dalam modul pelajaran dan memiliki kemampuan untuk belajar secara mandiri, berkolaborasi dengan rekan kerja, dan memperoleh sertifikat untuk menunjukkan keberhasilan dalam memperoleh kompetensi yang diajarkan dalam *course*.

2.1.2 Data Mining

Data mining merupakan sebuah proses untuk mengekstraksi informasi yang berharga dari data dengan menggunakan metode statistik, matematika, dan teknik komputasi lainnya [25]. Dalam proses data *mining*, data yang telah diolah dapat memberikan informasi yang berguna dalam pengambilan keputusan atau prediksi yang lebih akurat. Oleh karena itu, data *mining* menjadi sebuah teknik penting dalam dunia bisnis, industri, dan ilmu pengetahuan. Beberapa konsep penting dalam data *mining* meliputi *preprocessing* data, *exploratory* data *analysis*, pembangunan model, evaluasi model, dan interpretasi hasil [26]. *Preprocessing* data melibatkan pengolahan data mentah untuk menghilangkan data yang tidak penting atau data yang rusak. *Exploratory* data *analysis* dilakukan untuk mengidentifikasi pola-pola pada data yang tidak dapat dilihat secara langsung. Pembangunan model melibatkan proses untuk menghasilkan model yang dapat digunakan untuk membuat prediksi atau mengambil keputusan. Evaluasi model dilakukan untuk mengukur seberapa baik model tersebut dalam melakukan prediksi atau mengambil keputusan. Interpretasi hasil adalah proses untuk menjelaskan hasil dari model kepada pengguna atau

stakeholder. Beberapa teknik yang sering digunakan dalam data *mining* adalah regresi, klasifikasi, pengelompokan (*clustering*), asosiasi, dan visualisasi data [27].

2.1.2.1 Visualisasi Data

Visualisasi data adalah teknik untuk mengubah data menjadi bentuk visual atau grafis yang mudah dimengerti dan dapat memberikan wawasan baru dalam analisis data. Visualisasi data dapat membantu pemahaman data dan pemrosesan informasi yang kompleks, serta memberikan kejelasan dalam komunikasi dan eksplorasi data [28]. Visualisasi data yang efektif dapat meningkatkan keterlibatan pengguna, menghasilkan pemahaman yang lebih baik, dan memungkinkan pengambilan keputusan yang lebih baik [29]. Visualisasi data dapat dilakukan dengan menggunakan berbagai *tools* mulai dari yang sederhana seperti Microsoft Excel hingga *tools* yang lebih spesifik untuk visualisasi data seperti Tableau dan Power BI. Visualisasi data harus disesuaikan dengan hasil identifikasi masalah agar informasi yang ditemukan dari data yang ada dapat akurat dan bermanfaat.

2.1.2.2 Klasifikasi Data

Klasifikasi data adalah proses pengelompokan objek ke dalam kelas atau kategori yang telah ditentukan berdasarkan pada atribut atau fitur yang dimiliki. Klasifikasi data sangat penting dalam berbagai bidang, seperti ilmu pengetahuan, teknik, dan bisnis, dan sering digunakan dalam pengambilan keputusan, pengenalan pola, dan prediksi. Klasifikasi data dapat digunakan untuk memperoleh pemahaman yang lebih baik tentang data, membuat prediksi yang lebih akurat, dan membantu dalam pengambilan keputusan yang lebih baik [30].

Contoh dari klasifikasi data dalam dunia bisnis adalah sebagai berikut [27]:

1) Menentukan transaksi suatu kartu kredit termasuk kecurangan atau tidak.
2) Menentukan kualitas konsumen dalam pengajuan kartu kredit menjadi suatu kredit yang baik atau buruk.
3) Melakukan diagnosis penyakit dari seorang pasien berdasarkan ciri-ciri yang dimiliki.

Melalui berbagai contoh tersebut, maka dapat disimpulkan bahwa klasifikasi data memiliki manfaat dalam melakukan identifikasi suatu masalah berdasarkan kriteria yang dimiliki oleh data tersebut.

2.1.3 Machine Learning

Machine learning adalah suatu bidang kecerdasan buatan yang memungkinkan sistem komputer untuk belajar secara otomatis dari data, tanpa harus secara eksplisit diprogram untuk tugas tertentu. Konsep *machine learning* pertama kali ditemukan pada tahun 1952 untuk melakukan identifikasi permainan catur untuk disimpan gerakannya. Identifikasi permainan catur tersebut dilakukan oleh Arthur Samuel ke dalam sebuah program pada komputer IBM [31]. *Machine learning* memungkinkan komputer untuk mempelajari pola dan aturan dalam data dengan menggunakan algoritma dan model statistik [32]. Dalam beberapa tahun terakhir, *machine learning* telah digunakan dalam berbagai aplikasi, termasuk pengenalan wajah, analisis teks, kendaraan otonom, dan sebagainya. Jadi dapat disimpulkan bahwa *machine learning* berkaitan dengan data dan algoritma yang disesuaikan dengan tujuan penelitian. Terdapat berbagai algoritma yang dapat dipelajari untuk digunakan dalam melakukan tugas, contohnya untuk klasifikasi dengan algoritma *decision tree*, analisis regresi dengan *logistic*

regression, maupun *image processing* dengan *Convolutional Neural Network* (CNN).

2.2 Decision Tree, Random Forest, CRISP-DM, dan SEMMA

2.2.1 Decision Tree (DT)

DT adalah model *machine learning* yang digunakan untuk melakukan klasifikasi suatu target dengan membangun pohon keputusan dari data *training*. DT merupakan algoritma yang dikembangkan pertama kali oleh Ross Quinlan J. pada tahun 1979 yaitu dengan nama algoritma ID3 sebagai pembentuk pohon keputusan. Pohon keputusan ini terdiri dari simpul (*node*) dan tepi (*edge*), dimulai dari simpul akar (*root node*) dan mengarah ke simpul daun (*leaf node*). Setiap simpul dalam pohon keputusan mewakili fitur data dan setiap cabang mewakili keputusan yang diambil berdasarkan nilai fitur tersebut. *Decision tree* bekerja dengan mengidentifikasi fitur terbaik yang membagi data menjadi sub set yang lebih kecil dengan varian minimum. Proses ini terus berlanjut hingga keseluruhan data yang diuji telah diklasifikasi [33].

Terdapat dua tahapan yang perlu dilakukan DT yaitu *learning* dan *classification*. Pada tahap *learning*, algoritma DT menggunakan data yang telah disertakan dengan hasil klasifikasinya yang akan dibuatkan pohon keputusannya. Pohon keputusan tersebut kemudian menjadi model yang akan digunakan dalam klasifikasi data yang belum diklasifikasikan. DT memiliki langkah kerja yang mudah dipahami dan cepat untuk diimplementasikan [33]. Terdapat perhitungan yang dapat dilakukan untuk mengukur kualitas suatu *node* yaitu dengan menggunakan *Gini index* yang didefinisikan seperti pada rumus 2.1.

$$G = \sum_{k=1}^{C} \hat{p}mk\,(1 - \hat{p}mk)$$

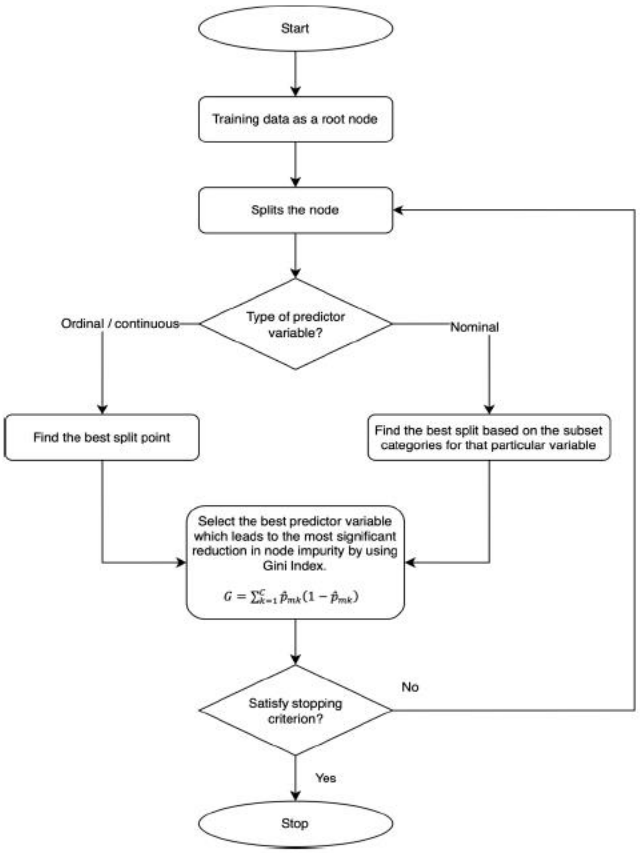
Rumus 2.1 Gini Index
Sumber: [33]

dan *entropy* yang didefinisikan seperti pada rumus 2.2.

$$D = \sum_{k=1}^{C} -\hat{p}mk\,log_2\,\hat{p}mk$$

Rumus 2.2 Entropy
Sumber: [33]

dengan $\hat{p}_{mk}$ merupakan proporsi dari observasi dalam m data *subset* yang berasal dari k *class* [35]. Kedua kalkulasi tersebut akan menghasilkan hasil minimum jika suatu *node* didominasi oleh klasifikasi dari suatu *class* saja yang menunjukkan sifat homogen. Sebaliknya, jika kedua kalkulasi tersebut maksimal maka *node* tersebut tidak homogen dan dapat menghasilkan beberapa *class* dari data. DT dapat diterapkan dengan beberapa algoritma seperti CART, ID3, dan C4.5 [36]. CART merupakan algoritma yang paling populer dan bahkan mudah diaplikasikan dengan bahasa pemrograman Python menggunakan *library* Scikit-learn. CART memiliki alur kerja yang digambarkan pada gambar 2.1.



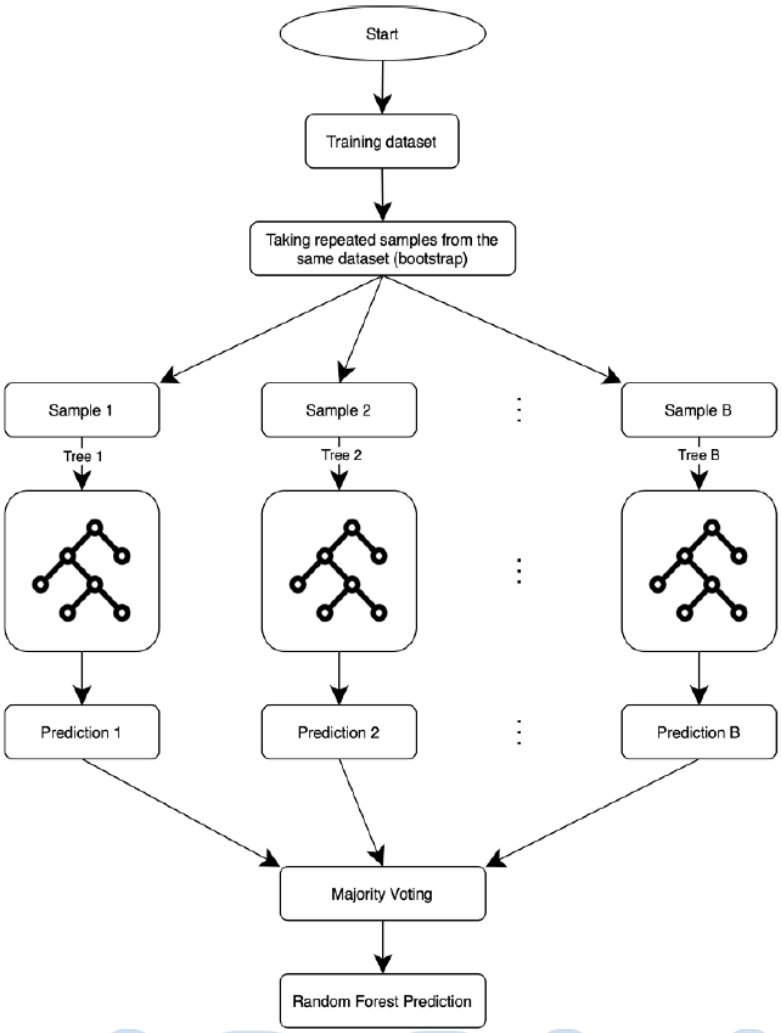
Gambar 2.1 Alur *Decision Tree* CART
Sumber: [33]

2.2.2 Random Forest (RF)

Algoritma RF merupakan salah satu algoritma yang paling umum digunakan untuk menangani masalah multi klasifikasi dan prediksi. Algoritma RF merupakan algoritma yang diperkenalkan oleh Breiman pada tahun 2001 sebagai kombinasi dari pembelajaran *ensemble bagging* dan metode *subspace* acak [36]. *Ensemble bagging* merupakan teknik pembelajaran dengan cara kerja penggabungan yang digunakan untuk meningkatkan kinerja model prediksi atau klasifikasi dengan menggabungkan beberapa model yang lebih sederhana dan terpisah menjadi satu kesatuan yang lebih kuat. *Subspace* yang acak berarti teknik pembelajaran yang digunakan untuk meningkatkan kinerja model prediksi atau klasifikasi dengan memilih hanya sebagian fitur atau atribut

dari set data pelatihan dan mengabaikan yang lainnya. *Ensemble bagging* dan *subspace* acak dilakukan pada RF sebagai pengembangan dari DT yang tidak menerapkannya, sehingga RF dapat dikatakan sebagai penyempurnaan dari DT [37]. RF membuat banyak *bagging tree* sesuai dengan namanya yaitu hutan (*forest*) yang terdiri dari banyak pohon (*tree*) dan menggunakan data dari data yang sama yang disebut *bootstrap*. DT yang dibentuk menggunakan data serta variabel yang acak dan dilakukan secara berulang [35].

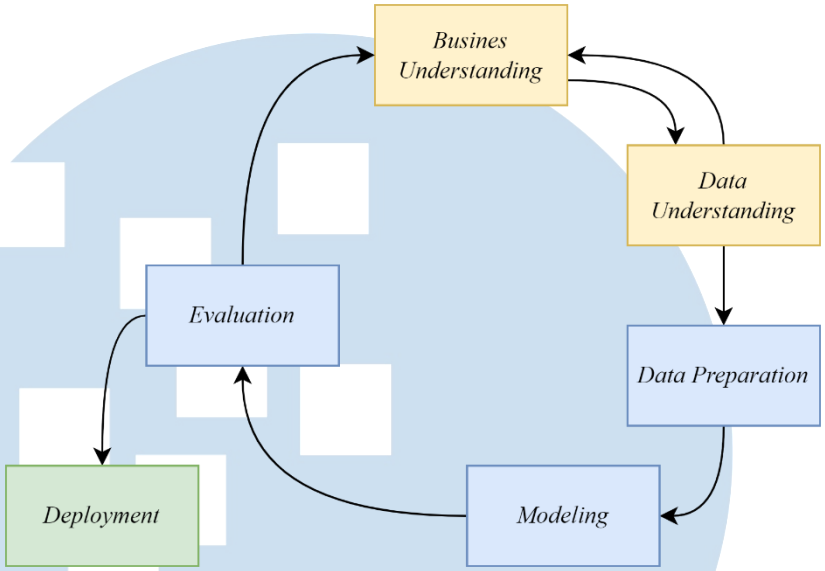
Dalam pembuatan banyak *bagging tree* di dalam RF, tidak selamanya variabel dari data yang digunakan selalu acak melainkan jika ada variabel yang sangat mempengaruhi model, maka kumpulan *bagging tree* akan tidak jauh berbeda [35]. Perbedaan antar pohon dapat terjadi atas pemilihan berbagai variabel dependen secara acak. Setelah terbentuk banyak *bagging tree* tersebut dilakukan voting hasil akhir yang menjadi prediksi RF [34]. Algoritma RF memiliki sensitivitas rendah terhadap multikolinearitas dan hasilnya relatif stabil dalam hal data yang hilang dan tidak seimbang [12]. Alur kerja dari RF dapat digambarkan pada gambar 2.2.



Gambar 2.2 Alur *Random Forest*
Sumber: [33]

2.2.3 CRISP-DM

CRISP-DM (*CRoss-Industry Standard Process for Data Mining*) adalah sebuah metode atau proses yang secara *de-facto* menjadi standar yang digunakan oleh untuk proyek data *mining* untuk mengembangkan solusi dari suatu permasalahan atau proyek data *mining* [38]. CRISP-DM pertama kali dirilis pada tahun 2000 dan tetap digunakan hingga saat ini sebagai sebuah *industry-independent process model* untuk data *mining* [39]. Metode ini memiliki enam tahap utama yang dapat digambarkan pada gambar 2.3.



Gambar 2.3 Metode CRISP-DM

Sumber: [38]

2.2.6.1 Business Understanding

Pada tahap ini harus dilakukan identifikasi situasi bisnis agar dapat memahami sumber daya yang ada dan diperlukan. Tujuan dari melakukan data *mining* juga harus ditentukan pada tahap ini, misalnya melakukan klasifikasi atau analisis regresi. Keberhasilan dari proses data *mining* yang akan dilakukan juga perlu ditentukan seperti berdasarkan nilai akurasi atau presisi. Rencana terhadap penelitian data *mining* harus disusun dengan jelas.

2.2.6.2 Data Understanding

Tahap ini merupakan tahap untuk memahami data yang akan digunakan. Hal tersebut diperoleh dengan mengumpulkan data yang tersedia, mengeksplorasi dan menganalisis data tersebut untuk memahami karakteristik data dan menemukan kemungkinan kekurangan, kelemahan atau kesalahan pada data. Untuk mendalami tahap ini dapat dilakukan analisis statistika

dalam memeriksa hubungan antar variabel pada data seperti korelasinya. Langkah ini perlu dilakukan dengan teliti agar informasi yang ingin diperoleh dari data dapat ditemukan dengan baik. Ringkasan (*summary*) terhadap data dapat dilakukan untuk pemahaman atas data secara menyeluruh.

2.2.6.3 Data Preparation

Tahap ini meliputi persiapan data untuk pengolahan lebih lanjut dengan membersihkan data dari *noise*, menghilangkan data yang tidak relevan atau duplikat, dan melakukan penggabungan data dari beberapa sumber. Pemilihan data juga dapat dilakukan pada tahap ini seperti menentukan variabel independen dan dependen yang akan digunakan dalam pembangunan suatu model *machine learning*. Beberapa hal umum yang dapat dilakukan dalam tahap data *preparation* meliputi [38]:

1) Menentukan aksi yang akan dilakukan terhadap data yang hilang atau kosong. Data yang hilang atau kosong dapat mengganggu terpenuhinya tujuan dari suatu proses data *mining*. Data yang hilang dapat ditindak lanjuti dengan melakukan beberapa cara seperti mengisi nilai kosong dengan rata-rata data lain atau dibiarkan saja. Hal tersebut perlu disesuaikan dengan alasan data tersebut kosong karena data yang kosong dapat memiliki suatu arti ataupun tidak memiliki arti.
2) Melakukan pemeriksaan terhadap kebenaran data. Kebenaran data dapat dilakukan dengan melakukan pemeriksaan terhadap adanya duplikasi data. Data yang duplikat dan memiliki tujuan yang sama dapat dihapus untuk

data duplikatnya karena dapat mengganggu tujuan data *mining*.

2.2.6.4 Modelling

Tahap ini merupakan tahap untuk membangun model dengan algoritma yang telah ditentukan dalam menjawab beberapa tahapan sebelumnya. *Modelling* dilakukan sesuai dengan permasalahan dan data yang dimiliki yang dapat dilakukan dengan penerapan teknik dan algoritma data *mining* dengan bantuan aplikasi atau *software* pendukung. Jika terdapat beberapa algoritma untuk menjawab permasalahan, maka penilaian model algoritma yang akan digunakan dapat dievaluasi. Selain itu, pemilihan model tersebut juga harus dapat dijelaskan alasannya. Dalam tahap *modelling*, jika kebutuhan atas data masih kurang, maka dapat kembali ke langkah selanjutnya untuk pengolahan data yang akan digunakan.

2.2.6.5 Evaluation

Pada tahap ini dilakukan pemeriksaan apakah solusi yang dibuatkan pada tahapan sebelumnya dapat menjawab permasalahan yang telah ditentukan di awal. Tindakan selanjutnya dari solusi yang dibuat juga harus ditentukan pada tahap ini. Tahap *evaluation* terhadap model *machine learning* dapat dilakukan dengan pengujian kualitas model. Kualitas suatu model dapat diukur dengan menggunakan perhitungan akurasi, *precision*, *recall*, dan *F1 score*. Akurasi dari suatu model merupakan tingkat keberhasilan prediksi suatu target berdasarkan fitur yang digunakan dari data *training* dan data *testing* yang dibandingkan. Untuk *precision, recall,* dan *F1 score* dapat dihitung melalui TP (*True Positive*), FP (*False Positive*), FN (*False Negative*), dan TN (*Total Negative*).

TP adalah jumlah prediksi *true positive*. FP adalah jumlah prediksi *false negative*. FN adalah jumlah prediksi *false negative*. TN adalah jumlah prediksi *true negative*. *Precision*, *recall*, dan *F1 score* memanfaatkan keempat perhitungan tersebut. *Precision* digambarkan dengan rumus 2.3, *recall* ditunjukkan dengan rumus 2.4, dan *F1 score* ditunjukkan oleh rumus 2.5.

$$Precision = \frac{TP}{(TP + FP)}$$

Rumus 2.3 *Precision*
Sumber: [38]

$$Recall = \frac{TP}{(TP + FN)}$$

Rumus 2.4 *Recall*
Sumber: [38]

$$F1\ Score = \frac{2\ .(Precision\ .Recall)}{(Precision + Recall)}$$

Rumus 2.5 *F1 Score*
Sumber: [38]

Berdasarkan perhitungan tersebut, *precision* digunakan untuk mengukur kemampuan model dalam melakukan identifikasi atas sampel yang positif. *Recall* digunakan untuk mengukur kemampuan model dalam menemukan semua sampel positif. *F1 score* digunakan untuk mengukur akurasi model secara keseluruhan akan kemapuannya dalam memprediksi sampel positif.

2.2.6.6 Deployment

Tahap ini memerlukan prosedur yang ditentukan oleh pengguna solusi yang dibangun. Tahap *deployment* dapat beraneka ragam seperti pembentukan suatu *software* maupun laporan. Pihak yang berperan dalam tahap ini harus mengatur tahap dari deployment, *monitoring*, hingga *maintanance*.

Metode CRISP-DM didasarkan pada prinsip bahwa data *mining* adalah sebuah proses iteratif yang memerlukan kolaborasi antara tim data *mining* dengan pemilik bisnis atau *end user*. Proses ini juga menekankan pentingnya tahap evaluasi dan validasi untuk memastikan keberhasilan model dan mengidentifikasi area yang dapat diperbaiki pada iterasi berikutnya [39].

2.2.4 SEMMA

SEMMA merupakan metode yang dapat digunakan dalam proyek data *mining* sebagai alternatif dari CRISP-DM. SEMMA dikembangkan oleh SAS Institute dan dapat terintegrasi dengan aplikasi SAS Enterprise Miner [40]. SEMMA merupakan singkatan dari kelima tahapannya yang meliputi *sample*, *explore*, *modify*, *model*, dan *access*. *Sample* adalah tahap dalam melakukan pengumpulan dan pemilahan data yang diperlukan dalam mencapai tujuan data *mining*. *Explore* merupakan proses menentukan *class attribute* dari data yang akan digunakan. *Modify* merupakan tahap *pre-processing* terhadap data sebelum diolah. Model adalah tahap penerapan algoritma *machine learning* terhadap data yang telah diolah sebelumnya. Terakhir, *access* adalah tahap melakukan evaluasi terhadap model yang telah dibentuk sebelumnya [41].

2.3 Tools Penelitian

2.3.1 Tableau

Tableau adalah sebuah perangkat lunak yang digunakan untuk membuat visualisasi data interaktif dan mudah dipahami [42]. Tableau pertama kali dirilis pada tahun 2003 oleh perusahaan bernama Tableau Software yang berbasis di Seattle, Washington, Amerika Serikat dan terus berkembang sampai saat ini. Tableau juga dapat digunakan untuk mengintegrasikan berbagai sumber data, termasuk data terstruktur dan tidak terstruktur, serta sumber data *online* dan *offline*. Dengan menggunakan Tableau, pengguna dapat melakukan analisis data secara *real-time*, melakukan visualisasi data dengan berbagai jenis grafik, dan berbagi hasil analisis dengan mudah kepada pihak lain dalam format yang dapat diakses oleh banyak orang. Dengan kemampuan-kemampuan ini, Tableau telah menjadi salah satu alat yang paling populer dan efektif dalam analisis data di berbagai sektor, termasuk bisnis, pemerintahan, dan akademik [43].

2.3.2 Power BI

Power BI merupakan aplikasi untuk membuat berbagai bentuk laporan maupun grafik yang dikembangkan oleh Microsoft dan dirilis pada tahun 2011. Power BI dapat menggunakan data yang berasal dari berbagai sumber baik data lokal maupun yang tersimpan di dalam *cloud*. Power BI memberikan kemudahan untuk melakukan visualisasi data dan sekaligus membagikannya kepada pihak di dalam ataupun di luar organisasi. Power BI memiliki beberapa layanan yang tersedia seperti Power BI Desktop, Power BI *online*, dan Power BI *mobile* yang tersedia untuk *platform* Windows, Android, dan iOS [44]. Kemampuan Power BI tersebut dapat diakses secara gratis yang menjadikan Power BI sebagai aplikasi yang dapat mengelola data menjadi wawasan yang koheren secara visual dan interaktif yang digemari.

2.3.3 Google Colaboratory

Google Colaboratory atau biasa disingkat Google Colab memungkinkan pengguna untuk membuat dan menjalankan kode Python, serta memanipulasi data secara interaktif melalui browser web, tanpa perlu mengunduh perangkat lunak apa pun di komputer lokal [45]. Ini memungkinkan pengguna untuk mengakses sumber daya komputasi yang kuat, seperti GPU dan TPU, untuk melatih model *deep learning* dengan lebih cepat. Sejak dirilis pada tahun 2007, Google Colab terus berkembang dan marak digunakan oleh berbagai kalangan. Selain itu, Google Colab juga memungkinkan pengguna untuk berkolaborasi dalam proyek secara *online* dengan tim, dengan mudah berbagi kode dan catatan di seluruh *platform* sistem operasi. Kemudahan dikarenakan Google Colab merupakan penyedia layanan pengembangan dan pelatihan *machine learning* oleh Google yang terintegrasi dengan produk Google lainnya seperti Google Drive dan Google Sheets. Semua ini menjadikan Google Colab menjadi alat yang populer dan berguna untuk pelatihan dan pengembangan *machine learning* dan AI (*Artificial Intelligence*) [46].

2.3.4 Jupyter Notebook

Jupyter Notebook merupakan aplikasi yang dapat digunakan untuk melakukan *scientific computation* seperti data *science* dan pembangunan model *machine learning* menggunakan bahasa Python sejak dirilis untuk publik pada 1 Februari 2010. Jupyter Notebook memiliki tampilan yang sederhana dan menggabungkan antara *code* sebagai *input* dengan *output* sehingga terlihat dinamis [47]. Untuk menggunakan Jupyter Notebook, maka pengguna harus memiliki *platform* Anaconda terlebih dahulu. Jupyter Notebook menjadi bagian dari *platform* Anaconda dan dapat berjalan melalui halaman *browser*. Sebagai *platform* dari Jupyter Notebook, Anaconda merupakan *platform open source* untuk bahasa

Python dan R. Anaconda dapat memudahkan dalam mengelola bahasa Python yang telah terpasang pada sistem operasi [48].

2.3.5 Python

Seperti yang telah dijelaskan sebelumnya bahwa Google Colab menggunakan bahasa Python. Python adalah bahasa pemrograman tingkat tinggi, interpretatif, dan multi-paradigma yang dapat digunakan untuk berbagai macam aplikasi, termasuk pengembangan web, pengolahan data, pembelajaran mesin, dan kecerdasan buatan [49]. Python dirancang dengan fokus pada kemudahan penggunaan dan membaca kode yang bersih dan mudah dipahami. Python sangat populer di kalangan pengembang perangkat lunak dan data *scientist* karena mudah dipelajari, fleksibel, dan mempunyai dukungan komunitas yang besar.

Sejak kehadirannya pada tahun 1994 oleh penciptanya yaitu Guido van Rossum, Python memiliki berbagai modul dan pustaka yang memungkinkan pengguna untuk memperluas fungsionalitas dan kegunaan bahasa pemrograman ini. Beberapa pustaka populer untuk pengolahan data dan pembelajaran mesin adalah NumPy, Pandas, Scikit-learn, Keras, dan TensorFlow [50]. Python juga memiliki berbagai jenis pengembangan dan pengujian kode, termasuk pengembangan berbasis *notebook* seperti Jupyter Notebook dan Google Colaboratory, serta berbagai lingkungan pengembangan terintegrasi seperti PyCharm dan Visual Studio Code.

2.3.6 R

R adalah bahasa pemrograman statistika yang dapat digunakan untuk proses manipulasi data dan membentuk grafik sejak kehadirannya pada tahun 1995. Nama R diciptakan dari nama penemunya yaitu Ross Ihaka dan Robert Gentleman. R juga merupakan implementasi dari bahasa program S yang difokuskan untuk analisis data, statistik data, dan

pembuatan model grafis. R populer dalam statistik karena merupakan bahasa yang dapat diakses secara gratis. R kini dikembangkan oleh R Core Team [49]. R memiliki komunitas yang besar dan memiliki berbagai paket yang mendukung suatu fungsi melalui *cran*-R yang dapat diunduh secara gratis. Editor dalam bahasa R yang paling populer adalah R Studio. R Studio dapat berperan selain sebagai editor, juga sebagai IDE bahasa R [51].

2.4 Penelitian Terdahulu

Penelitian terdahulu yang menjadi acuan dalam penelitian ini diuraikan pada tabel 2.1.

Tabel 2.1 Penelitian Terdahulu

PENELITIAN 1	
Penulis (tahun)	Ghosh, Abhishek, & Maiti, Ramkrishna (2021)
Judul	"*Soil erosion susceptibility assessment using logistic regression, decision tree and random forest: study on the Mayurakshi river basin of Eastern India*" [52]
Jurnal	Environmental Earth Sciences, vol. 80, no. 8, 2021
Metode	*Logistic regression, decision tree,* dan *random forest*
Hasil	Penelitian ini berhasil menilai kerentanan erosi tanah di daerah cekungan sungai Mayurakshi di India Timur menggunakan ketiga metode dengan hasil akurasi *decision tree* dan *random forest* sebesar 87,8% yang lebih tinggi dari *logistic regression* yaitu 85,6%.
PENELITIAN 2	
Penulis (tahun)	Dabiri, Hamed, Farhangi, Visar, Moradi, Mohammad Javad, Zadehmohamad, Mehdi, & Karakouzian, Moses (2022)
Judul	"*Applications of Decision Tree and Random Forest as Tree-Based Machine Learning Techniques for Analyzing the Ultimate Strain of Spliced and Non-Spliced Reinforcement Bars*" [53]
Jurnal	Applied Sciences, vol. 12, no. 10, p. 1-13, 2021
Metode	*Decision tree* dan *random forest*
Hasil	Penelitian ini menghasilkan model *machine learning* untuk memprediksi *reinforcement bar* baik sambungan maupun tidak sebelum diaplikasikan ke dalam konstruksi dengan

	metode *decision tree* dan *random forest* yang memiliki $R^2 \geq$ 85%
PENELITIAN 3	
Penulis (tahun)	Zhou, Xiaoyi, Lu, Pan, Zheng, Zijian, Tolliver, Denver, & Keramati, Amin (2020)
Judul	"*Accident Prediction Accuracy Assessment for Highway-Rail Grade Crossings Using Random Forest Algorithm Compared with Decision Tree*" [54]
Jurnal	Reliability Engineering & System Safety, vol. 200, p. 106931, Jan. 2020
Metode	*Random forest* dan *decision tree*
Hasil	Penelitian ini membuktikan bahwa algoritma *random forest* menghasilkan prediksi lebih akurat dibandingkan *decision tree* dalam memprediksi kecelakaan *Highway-Rail Grade Crossings*
PENELITIAN 4	
Penulis (tahun)	T. Prasandy, K. Nurkhasanah, M. P. Sari, and T. R. Fazry (2020)
Judul	"Perbandingan Hasil Penggunaan Metode Decision Tree dan Random Tree Pada Data Training Aplikasi Pencarian Tukang" [55]
Jurnal	Ultima InfoSys : Jurnal Ilmu Sistem Informasi, vol 10, no. 2, p. 93-97, 2020
Metode	*Decision tree* dan *random tree*
Hasil	Penelitian ini membandingkan penggunaan metode *decision tree* dan *random tree* dengan hasil metode *decision tree* direkomendasikan untuk aplikasi pencarian tukang karena memiliki parameter dan kemungkinan lebih banyak.
PENELITIAN 5	
Penulis (tahun)	Aryanti, Dessy Setiawan, Johan (2019)
Judul	"Visualisasi Data Penjualan dan Produksi PT Nitto Alam Indonesia Periode 2014-2018" [56]
Jurnal	Ultima InfoSys : Jurnal Ilmu Sistem Informasi, vol 9, no. 2, p. 86-91, 2019
Metode	Tableau
Hasil	Penelitian berhasil melakukan visualisasi data penjualan dan produksi PT Nitto Alam Indonesia periode 2014-2018. Hasilnya menunjukkan tren peningkatan produksi dari tahun ke tahun dan fluktuasi penjualan selama periode yang sama.

PENELITIAN 6	
Penulis (tahun)	Oetama, Raymond Sunardi Heng, Tan Thing Tjahjana, David (2020)
Judul	"Sebuah Pola Cluster Geospatial Eksplorasi Kejahatan Narkoba di DKI Jakarta" [57]
Jurnal	Ultima InfoSys : Jurnal Ilmu Sistem Informasi, vol 11, no. 1, p. 57-62, 2020
Metode	Tableau
Hasil	Penelitian berhasil membuat visualisasi *geospatial* dengan menggunakan analisis *clustering* untuk kejahatan narkoba di DKI Jakarta. Terdapat tiga *cluster* yaitu *cluster* angka kejahatan tinggi, *cluster* angka kejahatan sedang, dan *cluster* angka kejahatan rendah.
PENELITIAN 7	
Penulis (tahun)	P. Afikah, I. R. Affandi, and F. N. Hasan (2022)
Judul	"Implementasi Business Intelligence Untuk Menganalisis Data Kasus Virus Corona di Indonesia Menggunakan Platform Tableau" [58]
Jurnal	*Pseudocode*, vol. 9, no. 1, pp. 25–32
Metode	Tableau
Hasil	Penelitian berhasil membuat *dashboard* yang berisi jumlah kasus terkonfirmasi, kematian dan kesembuhan dari kasus Virus Corona di berbagai provinsi di Indonesia yang dapat digunakan untuk mendukung sebuah pengambilan keputusan.
PENELITIAN 8	
Penulis (tahun)	Goh, Clearance (2022)
Judul	"*Data Dashboarding in Accounting using Tableau*" [59]
Jurnal	Economics and Business Quarterly Reviews, 6(1), 269-275.
Metode	Tableau
Hasil	Penelitian ini membahas peran data dashboard dalam akuntansi dan mengilustrasikan, dengan menggunakan contoh yang telah dikerjakan, bagaimana sebuah data dashboard dapat dibangun menggunakan Tableau.

PENELITIAN 9	
Penulis (tahun)	Schröer, Christoph Kruse, Felix Gómez, Jorge Marx (2019)
Judul	"*A systematic literature review on applying CRISP-DM process model*" [60]
Jurnal	Procedia Computer Science, vol. 181, p. 526-534
Metode	*Systematic literature review*
Hasil	Penelitian melakukan sistematisasi literatur terhadap penggunaan model CRISP-DM dalam berbagai penelitian dan menyimpulkan bahwa CRISP-DM adalah proses standar dalam melakukan proyek data *mining* secara *de-facto* oleh mayoritas *author* di dunia.
PENELITIAN 10	
Penulis (tahun)	S. J. Saleh, S. Q. Ali, & A. M. Zeki (2022)
Judul	"*Random Forest vs. SVM vs. KNN in classifying Smartphone and Smartwatch sensor data using CRISP-DM*" [61]
Jurnal	2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy, ICDABI 2020, pp. 28–31
Metode	CRISP-DM
Hasil	Metode CRISP-DM digunakan dalam membandingkan penggunaan algoritma Random Forest, SVM, dan KNN untuk mengklasifikasikan data sensor smartphone dan smartwatch

Berdasarkan penelitian terdahulu yang ditunjukkan pada tabel 2.1, topik penelitian untuk membuat klasifikasi *assigned learning* yang tepat secara *personalized learning* berdasarkan klasifikasi preferensi karyawan dalam topik pembelajaran ketika mengerjakan *online training* serta topik untuk visualisasi data *training* karyawan dapat dilakukan. Algoritma yang dipilih adalah *decision tree* dan *random forest* sebagai hasil perbandingan dari penelitian terdahulu dengan akurasi yang baik [52], [53]. Penelitian ini melakukan perbandingan akurasi antara algoritma *random forest* dan *decision tree* karena masing-masing unggul dalam penelitian terdahulu yang pernah dilakukan [54], [55]. Dalam penelitian ini juga dilakukan visualisasi data *training* karyawan

dengan metode menggunakan aplikasi Tableau sebagai aplikasi yang telah berhasil dimanfaatkan untuk visualisasi berbagai tujuan dari analisis data seperti melihat tren bahkan visualisasi *geospatial* dengan *clustering* [56]-[59].

Kebaruan penelitian ini dibandingkan pada penelitian sebelumnya adalah objek penelitian terkait kegiatan *training* karyawan. Belum terdapat penelitian yang secara khusus melakukan klasifikasi *course* berdasarkan objek penelitian tersebut. Selain itu, penelitian ini juga menghasilkan dua buah luaran yang dicapai yaitu visualisasi data dan model *machine learning* menjadi kebaruan penelitian dari seluruh penelitian terdahulu dalam penelitian ini. Sejumlah penelitian terdahulu tersebut belum menerapkan metode penelitian CRISP-DM yang digunakan dalam penelitian ini. CRISP-DM dipilih berdasarkan hasil penelitian yang menunjukkan metode tersebut merupakan standar dalam proyek data *mining* secara *de-facto*, mudah digunakan, dan *reliable* [17], [18], [60], [61]. CRISP-DM telah menjawab berbagai permasalahan mulai dari bidang kesehatan, pendidikan dan penelitian, teknik dan produksi, pemerintahan, manajemen, teknik informatika, makanan, keuangan, hingga hiburan dalam pengerjaan proyek data *mining* [60].