BAB III
METODE PENELITIAN

3.1 Objek Penelitian

Objek penelitian ini adalah proses prediksi risiko *attrition* karyawan pada sistem *Human Resource* (HR) PT Inti Ganda Perdana (PT IGP). Penelitian berfokus pada pembuatan model *machine learning* yang mampu mengidentifikasi karyawan dengan potensi keluar (*attrition*) berdasarkan data karakteristik dan riwayat kerja mereka.

Secara umum, objek penelitian mencakup:

1. Karyawan PT IGP sebagai subjek yang dianalisis berdasarkan usia, lama bekerja, jabatan, gaji, performa, dan kepuasan kerja.
2. Data HR perusahaan, seperti data demografis dan data pekerjaan, yang digunakan untuk menghasilkan model prediksi.
3. Proses bisnis HR, terutama bagaimana perusahaan menangani *turnover* dan retensi karyawan.

Tabel 3.1 Contoh Data

No	Nama Variabel	Data	Keterangan
1	*Age*	24 tahun	Usia karyawan memengaruhi stabilitas kerja dan keputusan bertahan.
2	*Gender*	Laki-Laki	Digunakan sebagai variabel demografis pendukung analisis.
3	*Department*	Produksi	Divisi tempat karyawan bekerja.
4	*JobRole*	Operator Mesin	Posisi pekerjaan dapat memengaruhi tingkat stres atau beban kerja.
5	*Salary*	4.500.000	Tingkat gaji menjadi salah satu faktor utama retensi karyawan.
6	*Overtime*	Yes	Lembur sering menjadi pemicu *attrition* karena kelelahan kerja.
7	*JobSatisfaction*	2	Kepuasan kerja rendah meningkatkan risiko karyawan keluar.
8	*WorkLife Balance*	2	Keseimbangan kerja–hidup yang buruk berpengaruh pada keputusan resign.

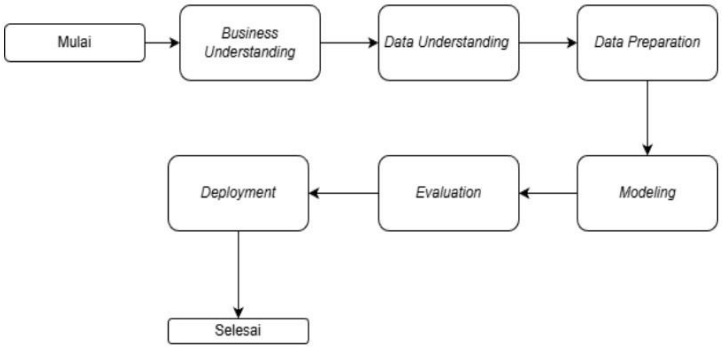
No	Nama Variabel	Data	Keterangan
9	*PerformanceRating*	3	Rating performa dapat menjadi indikator komitmen karyawan.
10	*YearsAtCompany*	1 tahun	Lama bekerja yang pendek biasanya lebih berisiko untuk turnover.
11	*Attrition*	1 = Keluar	Variabel target yang diprediksi oleh model *machine learning*.

3.2 Prosedur Penelitian

Prosedur penelitian pada penelitian ini menggunakan metode penelitian *framework Cross-Industry Standard Process for Data Mining (CRISP-DM)* untuk memastikan tahapan pengembangan model *machine learning* dilakukan secara terstruktur, sistematis, dan dapat direplikasi oleh peneliti lain. *Framework CRISP-DM* dipilih karena mampu menggambarkan alur penelitian secara komprehensif, mulai dari pemahaman permasalahan hingga evaluasi dan pemanfaatan hasil model.

Tahapan penelitian dalam *framework CRISP-DM* meliputi *Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation,* dan *Deployment*. Pada tahap *Business Understanding*, dilakukan identifikasi permasalahan *employee attrition* pada sistem *Human Resource* PT Inti Ganda Perdana serta penentuan tujuan penelitian, yaitu memprediksi risiko *attrition* karyawan. Tahap *Data Understanding* mencakup pemahaman terhadap karakteristik data karyawan yang diperoleh dari bagian HR, termasuk identifikasi variabel dan struktur data yang digunakan.

Selanjutnya, pada tahap *Data Preparation* dilakukan proses pengolahan dan prapemrosesan data, yang meliputi pembersihan data, pengkodean variabel kategorikal, normalisasi atau standarisasi data numerik, serta pembagian dataset



menjadi data pelatihan (*training data*) dan data pengujian (*testing data*). Tahap *Modeling* dilakukan dengan membangun model prediksi menggunakan algoritma *Logistic Regression* dan *Random Forest*, yang dipilih berdasarkan kesesuaian algoritma terhadap permasalahan klasifikasi biner.

3.2.1 Bussines Understanding

Pada tahap *Business Understanding*, dilakukan identifikasi dan analisis terhadap permasalahan bisnis yang terjadi di PT Inti Ganda Perdana (IGP), khususnya yang berkaitan dengan tingkat *attrition* karyawan. Berdasarkan hasil koordinasi dengan pihak *Human Resource* (HR), diketahui bahwa tingkat *attrition* karyawan berada pada kisaran 15–20% per tahun, yang berdampak langsung terhadap stabilitas operasional perusahaan.

Tingginya tingkat attrition tersebut menimbulkan beberapa permasalahan, antara lain ketidakpastian ketersediaan tenaga kerja, terutama pada lini produksi *rear axle* dan *propeller shaft* yang menggunakan sistem *make-to-order*. Kondisi ini menyebabkan potensi kekurangan tenaga kerja pada shift tertentu, khususnya shift malam, sehingga dapat menghambat pencapaian target produksi yang telah ditetapkan. Selain itu, tingginya tingkat keluar masuk karyawan juga meningkatkan beban biaya perusahaan dalam proses rekrutmen, pelatihan, dan adaptasi karyawan baru.

Berdasarkan kondisi tersebut, perusahaan membutuhkan suatu pendekatan yang lebih efektif dalam mengelola risiko *attrition*. Selama ini, penanganan *attrition* masih bersifat reaktif, yaitu dengan melakukan penggantian tenaga kerja setelah karyawan keluar, sehingga kurang optimal dalam menjaga stabilitas tenaga kerja. Oleh karena itu, diperlukan suatu sistem yang mampu melakukan deteksi dini terhadap karyawan yang berpotensi keluar, sehingga perusahaan dapat melakukan tindakan preventif.

Sebagai solusi, penelitian ini mengusulkan penerapan pendekatan *machine learning* untuk membangun model prediksi risiko *attrition* karyawan. Model ini diharapkan mampu mengidentifikasi pola-pola yang berkaitan dengan keputusan karyawan untuk keluar berdasarkan data historis yang dimiliki perusahaan. Dengan demikian, hasil prediksi dapat digunakan sebagai sistem pendukung keputusan bagi

pihak HR dalam merancang strategi retensi karyawan secara lebih proaktif, efektif, dan berbasis data.

3.2.2 Data Undestanding

Pada tahap *Data Understanding*, dilakukan proses pengumpulan, eksplorasi, dan pemahaman terhadap dataset yang digunakan dalam penelitian. Data yang digunakan merupakan data karyawan PT Inti Ganda Perdana yang diperoleh dari sistem *Human Resources* (HR), dengan fokus pada informasi yang berkaitan dengan karakteristik karyawan serta status keberlanjutan kerja (attrition).

Dataset yang digunakan terdiri dari 58.810 data karyawan, di mana setiap baris merepresentasikan satu individu karyawan. Variabel yang digunakan meliputi usia (*Age*), jenis kelamin (*Gender*), departemen (*Department*), jabatan (*JobRole*), gaji (*Salary*), status lembur (*Overtime*), tingkat kepuasan kerja (*JobSatisfaction*), keseimbangan kehidupan kerja (*WorkLifeBalance*), penilaian kinerja (*PerformanceRating*), lama bekerja (*YearsAtCompany*), serta variabel target *Attrition* yang menunjukkan apakah karyawan masih bekerja atau telah keluar dari perusahaan.

Selain memahami struktur data, pada tahap ini juga dilakukan eksplorasi data awal (*exploratory data analysis*) untuk mengidentifikasi karakteristik dan kualitas data. Proses ini mencakup pemeriksaan distribusi data, identifikasi *missing value*, deteksi *outlier*, serta analisis hubungan awal antar variabel. Dari hasil eksplorasi, diketahui bahwa terdapat ketidakseimbangan distribusi pada variabel target, di mana jumlah karyawan yang bertahan lebih dominan dibandingkan karyawan yang keluar, sehingga data bersifat *imbalanced*.

Selain itu, beberapa variabel kategorikal seperti *Department*, *JobRole*, dan *Overtime* memerlukan transformasi ke bentuk numerik agar dapat diproses oleh algoritma *machine learning*. Variabel numerik seperti *Age*, *Salary*, dan *YearsAtCompany* juga dianalisis untuk melihat sebaran data dan potensi adanya nilai ekstrem (*outlier*) yang dapat memengaruhi performa model.

Hasil dari tahap *Data Understanding* ini memberikan gambaran awal mengenai kondisi dataset serta permasalahan yang ada, sehingga menjadi dasar dalam menentukan langkah-langkah pada tahap *Data Preparation*, seperti proses

pembersihan data, transformasi variabel, serta penanganan ketidakseimbangan data menggunakan teknik SMOTE.

3.2.3 Data Preparation

Pada tahap *Data Preparation*, dilakukan serangkaian proses pra-pemrosesan terhadap dataset karyawan PT Inti Ganda Perdana yang berjumlah 58.810 data agar siap digunakan dalam proses pemodelan. Tahap ini bertujuan untuk meningkatkan kualitas data serta memastikan bahwa data yang digunakan telah sesuai dengan kebutuhan algoritma *machine learning*.

Adapun langkah-langkah yang dilakukan pada tahap ini adalah sebagai berikut:

1. *Data Cleaning*

   Pada tahap ini dilakukan pembersihan data untuk mengatasi permasalahan kualitas data. Proses yang dilakukan meliputi:

   a. Menghapus data duplikat untuk menghindari bias dalam pelatihan model
   b. Menangani *missing value* dengan metode imputasi atau penghapusan data yang tidak lengkap
   c. Memastikan konsistensi format data pada setiap atribut

2. *Transformasi Data (Encoding)*

   Variabel kategorikal seperti *Gender*, *Department*, *JobRole*, dan *Overtime* diubah ke dalam bentuk numerik menggunakan teknik *encoding*, seperti *label encoding* atau *one-hot encoding*. Transformasi ini diperlukan agar data dapat diproses oleh algoritma machine learning yang hanya menerima input numerik.

3. *Feature Engineering*

   Dilakukan pengolahan fitur untuk meningkatkan kualitas informasi yang digunakan dalam model. Proses ini meliputi:

   a. Pengelompokan variabel *YearsAtCompany* ke dalam kategori masa kerja (misalnya: < 2 tahun, 2–5 tahun, > 5 tahun)
   b. Pengelompokan variabel *Salary* ke dalam kategori tertentu untuk mempermudah analisis pola
   c. Pemilihan fitur (*feature selection*) yang relevan terhadap variabel target

4. Normalisasi Data

Variabel numerik seperti *Age*, *Salary*, dan *YearsAtCompany* dinormalisasi menggunakan metode seperti *Min-Max Scaling* agar seluruh fitur memiliki skala yang seragam. Hal ini bertujuan untuk meningkatkan performa model, khususnya pada algoritma yang sensitif terhadap skala data seperti *Logistic Regression*.

5. *Handling Imbalanced Data* (SMOTE)

   Berdasarkan hasil analisis pada tahap sebelumnya, diketahui bahwa distribusi data bersifat tidak seimbang. Oleh karena itu, diterapkan metode SMOTE (*Synthetic Minority Over-sampling Technique*) pada data pelatihan untuk menyeimbangkan jumlah data antara kelas *attrition* dan *non-attrition*. Teknik ini dilakukan dengan menghasilkan data sintetis pada kelas minoritas sehingga model dapat belajar secara lebih optimal.

6. *Data Splitting*

   Dataset yang telah diproses kemudian dibagi menjadi data latih (*training data*) sebesar 80% dan data uji (*testing data*) sebesar 20%. Data latih digunakan untuk membangun model, sedangkan data uji digunakan untuk mengevaluasi performa model.

   Hasil dari tahap *Data Preparation* ini adalah dataset yang telah bersih, terstruktur, dan seimbang, sehingga siap digunakan dalam tahap *Modeling* untuk membangun model prediksi risiko *attrition* karyawan.

3.2.4 Modeling

Pada tahap *Modeling*, dilakukan proses pembangunan model klasifikasi untuk memprediksi risiko *attrition* karyawan menggunakan dua algoritma *machine learning*, yaitu *Logistic Regression* dan *Random Forest*. Pemilihan kedua algoritma ini didasarkan pada perbedaan karakteristiknya, di mana *Logistic Regression* digunakan untuk memodelkan hubungan linear yang mudah diinterpretasikan, sedangkan *Random Forest* digunakan untuk menangkap pola hubungan yang lebih kompleks dan *non-linear*.

Proses pemodelan dilakukan menggunakan bahasa pemrograman *Python* dengan bantuan *library* seperti *scikit-learn*. Dataset yang digunakan pada tahap ini adalah data latih (*training data*) yang telah melalui proses *data preparation*,

termasuk normalisasi dan penanganan ketidakseimbangan data menggunakan SMOTE.

Langkah-langkah pada tahap modeling adalah sebagai berikut:

1. Inisialisasi Model

   Mendefinisikan algoritma yang akan digunakan, yaitu *Logistic Regression* dan *Random Forest*, beserta parameter awal (*default parameter*) yang digunakan dalam proses pelatihan.

2. Pelatihan Model (*Training*)

   Model dilatih menggunakan *data training* dengan *variabel independen* sebagai fitur *input* dan *variabel Attrition* sebagai target. Pada tahap ini, algoritma akan mempelajari pola hubungan antara karakteristik karyawan dengan kemungkinan terjadinya *attrition*.

3. Proses Prediksi

   Setelah model dilatih, dilakukan proses prediksi terhadap data uji (testing data) untuk mengetahui hasil klasifikasi dari masing-masing model terhadap data yang belum pernah dilihat sebelumnya.

4. Pengaturan Parameter (Opsional)

   Untuk meningkatkan performa model, dilakukan pengaturan parameter (*hyperparameter tuning*) seperti jumlah pohon (*n_estimators*) pada *Random Forest* atau parameter regularisasi pada *Logistic Regression*, sehingga model dapat menghasilkan prediksi yang lebih optimal.

5. *Output Model*

   Hasil dari tahap ini adalah dua model terlatih yang menghasilkan prediksi berupa klasifikasi risiko *attrition* (keluar atau bertahan), serta probabilitas dari masing-masing prediksi yang akan digunakan pada tahap evaluasi.

Dengan adanya dua model yang dibangun pada tahap ini, penelitian dapat melakukan perbandingan performa secara objektif untuk menentukan algoritma yang paling optimal dalam memprediksi risiko *attrition* karyawan.

3.2.5 Evaluation

Pada tahap *Evaluation*, dilakukan proses penilaian dan perbandingan kinerja model *Logistic Regression* dan *Random Forest* yang telah dibangun pada tahap sebelumnya. Evaluasi dilakukan menggunakan data uji (*testing data*) yang tidak

terlibat dalam proses pelatihan model, sehingga hasil yang diperoleh dapat merepresentasikan kemampuan model dalam melakukan generalisasi terhadap data baru.

Proses evaluasi dilakukan dengan menggunakan *confusion matrix* sebagai dasar perhitungan metrik performa. *Confusion matrix* membagi hasil prediksi ke dalam empat kategori, yaitu *True Positive (TP), True Negative (TN), False Positive (FP),* dan *False Negative (FN),* yang kemudian digunakan untuk menghitung beberapa metrik evaluasi.

Adapun metrik yang digunakan dalam penelitian ini adalah sebagai berikut:

1. *Accuracy*

   Accuracy digunakan untuk mengukur tingkat ketepatan model dalam mengklasifikasikan data secara keseluruhan. Metrik ini menunjukkan proporsi jumlah prediksi yang benar dibandingkan dengan total data. Namun, pada kondisi data yang tidak seimbang, *accuracy* tidak selalu mencerminkan performa model secara akurat karena dapat dipengaruhi oleh dominasi kelas mayoritas.

2. Precision

   Precision digunakan untuk mengukur ketepatan model dalam memprediksi kelas positif (*attrition*). Nilai *precision* yang tinggi menunjukkan bahwa sebagian besar karyawan yang diprediksi akan keluar memang benar-benar keluar, sehingga dapat mengurangi kesalahan dalam pengambilan keputusan oleh pihak HR.

3. *Recall*

   Recall digunakan untuk mengukur kemampuan model dalam mendeteksi seluruh data yang benar-benar termasuk dalam kelas *attrition*. Dalam penelitian ini, *recall* menjadi metrik yang sangat penting karena berkaitan langsung dengan kemampuan model dalam mengidentifikasi karyawan yang berisiko keluar. Nilai *recall* yang tinggi menunjukkan bahwa model mampu mendeteksi sebagian besar kasus *attrition*, sehingga perusahaan dapat melakukan *intervensi* lebih awal.

4. *F1-Score*

   F1-score merupakan rata-rata harmonis antara *precision* dan *recall* yang digunakan untuk memberikan keseimbangan antara kedua metrik tersebut. Metrik ini digunakan untuk mengevaluasi performa model secara keseluruhan, terutama pada kondisi data yang tidak seimbang.

Hasil evaluasi dari kedua model kemudian dibandingkan untuk menentukan model dengan performa terbaik. Model yang dipilih adalah model yang memiliki keseimbangan terbaik antara *accuracy*, *precision*, *recall*, dan *F1-score*, serta memiliki kemampuan paling baik dalam mendeteksi risiko *attrition* karyawan. Dengan demikian, tahap evaluasi ini menjadi dasar dalam menentukan model yang paling optimal untuk digunakan pada tahap *deployment*.

3.2.6 Deployment

Pada tahap *Deployment*, dilakukan penerapan model terbaik yang diperoleh dari tahap evaluasi ke dalam konteks penggunaan nyata pada sistem *Human Resources* (HR) PT Inti Ganda Perdana. Model yang dipilih, yaitu model dengan performa terbaik berdasarkan metrik evaluasi, digunakan untuk memprediksi risiko attrition karyawan berdasarkan data karakteristik yang dimiliki.

Implementasi model dilakukan dalam bentuk sistem pendukung keputusan (*decision support system*), di mana data karyawan yang tersedia akan diproses oleh model untuk menghasilkan prediksi berupa status risiko *attrition*. *Output* yang dihasilkan tidak hanya berupa klasifikasi (berisiko keluar atau tidak), tetapi juga nilai probabilitas yang menunjukkan tingkat kemungkinan karyawan untuk keluar dari perusahaan.

Hasil prediksi tersebut kemudian dapat disajikan dalam bentuk kategori tingkat risiko, seperti risiko rendah, sedang, dan tinggi, sehingga memudahkan pihak HR dalam melakukan interpretasi dan pengambilan keputusan. Informasi ini dapat digunakan untuk mengidentifikasi karyawan yang memiliki potensi tinggi untuk keluar, sehingga perusahaan dapat melakukan tindakan preventif seperti peningkatan kepuasan kerja, penyesuaian beban kerja, atau perbaikan sistem kompensasi.

Selain itu, hasil prediksi juga dapat divisualisasikan dalam bentuk *dashboard* sederhana yang menampilkan daftar karyawan beserta tingkat risiko *attrition*,

sehingga memudahkan monitoring secara berkala. Implementasi ini memungkinkan perusahaan untuk beralih dari pendekatan reaktif menjadi pendekatan proaktif dalam mengelola attrition karyawan.

Dengan adanya tahap *deployment* ini, model *machine learning* yang dibangun tidak hanya berhenti pada tahap analisis, tetapi juga memberikan manfaat nyata dalam mendukung pengambilan keputusan strategis di bidang sumber daya manusia.