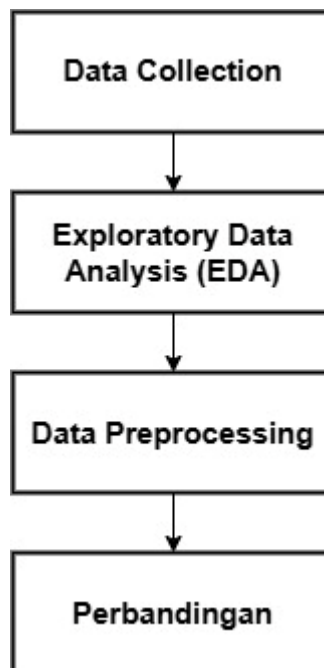


BAB III

METODE PENELITIAN

3.1 Alur Penelitian

Metode penelitian terkait tentang faktor-faktor karyawan meninggalkan perusahaan atau tidak dengan algoritma *Decision Tree*, *Random Forest*, dan *K-Neighbors Classifier (KNN)* yang akan dilakukan dibagi menjadi 4 tahapan, dapat dilihat pada gambar 3.1.



Gambar 3.1 Alur Penelitian

3.1.2 Data Collection

Pada Tahap Penelitian ini menggunakan dataset karyawan yang diperoleh dari repositori *Kaggle*. Dataset yang digunakan berjumlah 4653 data dalam bentuk *comma separated values (csv)*. Pada data tersebut terdiri dari 9 column untuk

predicted dan sebuah target. Kolom sebagai predicted adalah pendidikan, tahun bergabung, kota, tingkat pembayaran, usia, jenis kelamin, pengalaman dibidang saat ini, status pernah menganggur, dan keputusan karyawan untuk meninggalkan atau tidak diperusahaan. Sedangkan, column sebagai target adalah keputusan karyawan untuk meninggalkan atau tidak diperusahaan. Format dataset tersebut dapat dilihat pada tabel 3.1.2

Table 3.1 Meta data

Attribute Data Karyawan		Keterangan	Kategori	Tipe Data
1	Education	Tingkat pendidikan terakhir yang dicapai oleh karyawan	Predictor	Kategorikal
2	JoiningYear	Tahun ketika karyawan mulai bekerja diperusahaan	Predictor	Numerik
3	City	Kota atau lokasi tempat tinggal/kerja karyawan	Predictor	Kategorikal
4	PaymeIntTielr	Tingkat pelmbayaran/gaji yang ditelrma oleh karyawan	Preldictor	Nulmelrik

5	Age	Usia karyawan pada saat data dikumpulkan	Predictor	Nominal
6	Gender	Jenis Kelamin Karyawan	Predictor	Kategorikal
7	EverBenched	Status karyawan pernah melnganggur	Predictor	Kategorikal
8	ExperienceInCurrentDomain	Pengalaman kerja karyawan dibidang atau posisi saat ini	Predictor	Nominal
9	LeaveOrNot	Keputusan karyawan apakah akan meninggalkan perusahaan atau tidak	Target	Nominal

Table 3.1.2 Data Sample

	Education	JoiningYear	City	PaymentTier	Age	Gender	EverBenched	ExperienceInCurrentDomain	LeaveOrNot
0	Bachelors	2017	Bangalore	3	34	Male	No	0	0
1	Bachelors	2013	Pune	1	28	Female	No	3	1
2	Bachelors	2014	New Delhi	3	38	Female	No	2	0
3	Masters	2016	Bangalore	3	27	Male	No	5	1
4	Masters	2017	Pune	3	24	Male	Yes	2	1

3.1.3 Exploratory Data Analysis (EDA)

Pada tahapan ini data yang telah terkumpul perlu dianalisis terlebih dahulu untuk mengidentifikasi karakteristiknya, seperti adanya duplikasi, nilai yang kosong (*null*), jumlah total data, serta tipe data yang mungkin tidak konsisten. Tahap ini sangat penting untuk memastikan kualitas data sebelum melangkah ke analisis lebih lanjut. Melalui EDA, kita dapat lebih memahami struktur data secara menyeluruh dan mengidentifikasi faktor-faktor yang mempengaruhi hasil analisis atau prediksi.

3.1.4 Data Preprocessing

Pada tahapan ini yaitu menghilangkan data yang tidak relevan dengan cara menangani nilai yang hilang (*missing values*), mengatasi data yang berisik (*noise*), serta memperbaiki inkonsistensi dan ketidakrelevanan data pada label prediksi. Proses pre- processing ini dilakukan untuk menambahkan informasi pada atribut yang hilang atau kosong, serta untuk mengubah data yang tidak konsisten menjadi format yang sesuai. Dengan demikian, data yang digunakan dalam analisis akan lebih bersih dan siap untuk mendapatkan hasil yang akurat.

3.1.5 Perbandingan

Setelah data selesai pada tahap pre-processing, selanjutnya adalah melakukan perbandingan dengan algoritma *Decision Tree*, *Random Forest*, *K-Nearest Neighbors (KNN)*. [5]

- a) *Decision Tree* merupakan algoritma yang memiliki struktur menyerupai diagram alir (*flowchart*) yang setiap node merupakan pengujian terhadap variabel atribut. Algoritma ini digunakan dalam analisa klasifikasi dan prediksi dalam bentuk pohon keputusan. [5]

- b) *Random Forest* merupakan salah satu jenis dari algoritma *Decision Tree* yang merupakan kumpulan (ensemble) pohon keputusan yang digunakan sebagai base classifier yang dikombinasikan. Algoritma ini digunakan untuk analisa klasifikasi dan regresi[5]

- c) *K-Nearest Neighbor* adalah metode klasifikasi terhadap suatu dataset berdasarkan jarak data pembelajaran (neighbor) terdekat. Jauh maupun dekatnya data pembelajaran (neighbor) tersebut dihitung dengan jarak Euclidean[5]