

# ***DRAFT PROPOSAL***

*Studi Komparatif Multinomial Logistic Regression dan CatBoost Classifier Berbasis Analisis SHAP untuk Memprediksi Retensi Karyawan Berdasarkan Lama Bekerja, Gaji dan Jabatan Sebagai Dasar Penentuan Strategi Mode Bekerja*

---

## **BAB I PENDAHULUAN**

### **1.1 Latar Belakang**

Perkembangan teknologi digital telah membawa perubahan besar dalam dunia kerja modern. Organisasi kini tidak hanya dituntut untuk merekrut tenaga kerja yang berkualitas, tetapi juga harus mampu mempertahankan karyawan agar tetap loyal dan produktif. Fenomena turnover karyawan yang tinggi menjadi tantangan strategis karena berdampak langsung pada stabilitas organisasi dan biaya operasional ([Handayani 2023](#)). Dalam era data-driven, pendekatan analitik berbasis machine learning telah terbukti mampu membantu perusahaan memahami perilaku karyawan secara lebih akurat.

Menurut ([Musa et al. 2023](#)), Human Resource Analytics (HRA) memungkinkan organisasi untuk melakukan analisis prediktif terhadap faktor-faktor yang memengaruhi retensi dan kinerja karyawan. Penerapan model prediksi ini memungkinkan HR untuk mengidentifikasi risiko turnover sejak dini, sekaligus merancang strategi intervensi yang tepat. Salah satu pendekatan statistik yang paling banyak digunakan dalam analisis perilaku karyawan adalah Multinomial Logistic Regression (MLR).

Model ini efektif untuk mengklasifikasikan keputusan multi-kelas, misalnya antara karyawan yang memilih bertahan, berpindah, atau berpotensi keluar ([Dewi et al. 2022](#)). Namun, penelitian oleh ([Handayani 2023](#)) menunjukkan bahwa MLR memiliki keterbatasan dalam menangani data nonlinier dan variabel kategorikal yang kompleks kondisi yang sering muncul dalam data SDM. Sebagai alternatif, algoritma modern seperti CatBoost Classifier muncul sebagai solusi yang lebih adaptif. ([Zhang et al. 2024](#)) menjelaskan bahwa CatBoost mampu menangani data kategorikal tanpa proses encoding manual, serta memiliki kemampuan gradient boosting yang efisien. Ketika digabungkan

dengan analisis SHAP (Shapley Additive Explanations), model ini tidak hanya memberikan akurasi tinggi tetapi juga menghasilkan interpretasi yang transparan dan mudah dipahami ([Chen et al. 2025](#)).

Perbandingan antara MLR dan CatBoost menjadi penting karena keduanya merepresentasikan dua paradigma analisis berbeda model statistik klasik yang interpretatif dan model machine learning modern yang adaptif. Menurut ([Baydili and Tasci 2025](#)), integrasi Explainable AI (XAI) dengan model prediksi karyawan dapat meningkatkan kepercayaan manajemen terhadap hasil analisis, karena mampu menjelaskan mengapa dan bagaimana model membuat keputusan tertentu. Dalam konteks penelitian ini, fokus diarahkan pada tiga variabel utama yang secara empiris berpengaruh terhadap retensi karyawan, yaitu lama bekerja, gaji, dan jabatan. Menurut ([Adeniyi 2024](#)), lama bekerja berhubungan erat dengan loyalitas karena karyawan yang telah lama bekerja cenderung memiliki komitmen emosional terhadap organisasi. Sementara ([Handayani 2023](#)) menegaskan bahwa gaji merupakan faktor kompensasi paling signifikan yang memengaruhi niat bertahan. Selain itu, posisi atau jabatan juga memiliki pengaruh penting karena menentukan tingkat tanggung jawab dan peluang pengembangan karier ([Ismail and Rahman 2023](#)).

Ketiga variabel tersebut dipilih karena mewakili dimensi utama dalam manajemen SDM pengalaman kerja, kesejahteraan finansial, dan posisi karir yang secara bersama-sama membentuk dasar pengambilan keputusan karyawan untuk bertahan atau keluar. Penelitian ini akan membandingkan kemampuan MLR dan CatBoost dalam memprediksi retensi berdasarkan tiga variabel tersebut, serta mengidentifikasi faktor dominan yang berpengaruh menggunakan analisis SHAP. Dengan demikian, hasil penelitian diharapkan tidak hanya memberikan model prediksi yang akurat, tetapi juga rekomendasi strategi mode kerja (WFH, WFO, Hybrid) yang dapat mendukung peningkatan retensi karyawan di era pasca pandemi.

## **1.2 Rumusan Masalah**

Rumusan masalah berfungsi sebagai panduan utama dalam menentukan fokus analisis serta batasan ruang lingkup penelitian. Berdasarkan latar belakang yang telah dijelaskan sebelumnya, penelitian ini merumuskan beberapa masalah yaitu :

1. Bagaimana penerapan algoritma Multinomial Logistic Regression dan CatBoost Classifier dalam memprediksi retensi karyawan?
2. Bagaimana hasil perbandingan performa kedua model berdasarkan metrik evaluasi (Accuracy, F1-Score, dan AUC)?
3. Faktor apa yang paling berpengaruh terhadap retensi karyawan berdasarkan analisis SHAP?
4. Bagaimana rekomendasi strategi mode kerja optimal (WFH, WFO, atau Hybrid) berdasarkan hasil analisis?

### **1.3 Tujuan Penelitian**

Tujuan penelitian menggambarkan hasil akhir yang ingin dicapai melalui proses analisis dan pengolahan data. Tujuan ini disusun berdasarkan rumusan masalah dan diarahkan untuk memberikan kontribusi nyata, baik dalam aspek teoritis maupun praktis. Berikut merupakan detail penjelasannya yaitu :

1. Menerapkan dan membandingkan algoritma Multinomial Logistic Regression dan CatBoost Classifier dalam memprediksi retensi karyawan.
2. Mengidentifikasi faktor dominan yang memengaruhi keputusan retensi melalui analisis SHAP.
3. Memberikan rekomendasi mode kerja yang optimal berdasarkan hasil interpretasi model prediktif.

### **1.4 Manfaat Penelitian**

Manfaat penelitian menjelaskan nilai tambah yang dihasilkan dari pelaksanaan studi ini, baik bagi pengembangan ilmu pengetahuan maupun penerapannya di dunia kerja. Penelitian ini diharapkan tidak hanya memberikan kontribusi akademis dalam bidang Explainable Artificial Intelligence (XAI), tetapi juga menghasilkan rekomendasi praktis bagi perusahaan dalam mengelola dan mempertahankan karyawan di tengah dinamika sistem kerja modern.

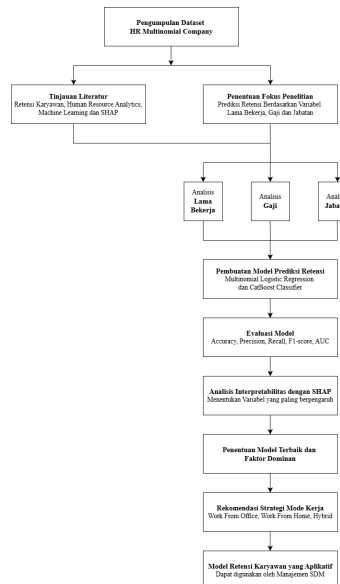
### **1.5 Batasan Masalah**

Agar penelitian berjalan secara terarah dan tidak melebar dari fokus utama, diperlukan pembatasan ruang lingkup pembahasan. Batasan masalah ini disusun untuk memperjelas area penelitian yang dikaji serta asumsi yang digunakan. Berikut penjelasannya yaitu :

1. Penelitian ini menggunakan tiga variabel utama, yaitu lama bekerja, gaji, dan jabatan.
2. Data yang digunakan merupakan dataset HR publik atau data simulasi yang menggambarkan kondisi umum karyawan di Indonesia.
3. Model yang dibandingkan terbatas pada Multinomial Logistic Regression dan CatBoost Classifier dengan analisis interpretasi menggunakan SHAP.

## 1.6 Kerangka Berpikir

Kerangka berpikir penelitian ini menggambarkan alur sistematis dari identifikasi masalah hingga solusi strategis yang diusulkan. Penelitian dimulai dengan pengumpulan data HR yang berisi variabel gaji, lama bekerja, dan jabatan, yang diasumsikan memengaruhi keputusan retensi karyawan. Data kemudian diproses dan dianalisis menggunakan dua algoritma machine learning, yaitu Multinomial Logistic Regression dan CatBoost Classifier. Kedua model ini dievaluasi menggunakan metrik Accuracy, F1-Score, dan AUC (Area Under Curve) untuk menentukan model dengan performa terbaik. Hasil model kemudian dianalisis menggunakan Analisis SHAP (Shapley Additive Explanations) untuk mengidentifikasi variabel paling berpengaruh terhadap keputusan karyawan bertahan atau resign. Berdasarkan hasil interpretasi tersebut, penelitian ini menyusun solusi strategis dalam bentuk rekomendasi mode kerja (WFH, WFO, atau Hybrid) sebagai strategi peningkatan retensi karyawan.



*Gambar 1. Kerangka Berpikir*

## **1.7 Sistematika Penulisan**

Sistematika penulisan dalam penelitian ini disusun untuk memberikan gambaran secara terstruktur mengenai langkah langkah penelitian yang dilakukan. Berikut merupakan detail sistematika penulisan pada penelitian ini yaitu :

### **BAB I PENDAHULUAN**

Bab ini berisi uraian mengenai latar belakang penelitian, rumusan masalah, tujuan penelitian, manfaat penelitian, batasan masalah, kerangka berpikir, serta sistematika penulisan. Bab ini memberikan gambaran umum mengenai alasan dilakukannya penelitian, fokus utama yang ingin dicapai, serta arah solusi strategis berupa rekomendasi mode kerja (Work From Home, Work From Office, atau Hybrid) berdasarkan hasil analisis prediksi retensi karyawan.

### **BAB II TINJAUAN PUSTAKA**

Bab ini memuat teori-teori yang mendukung penelitian, seperti konsep retensi karyawan, manajemen sumber daya manusia, serta pendekatan machine learning yang digunakan yaitu Multinomial Logistic Regression dan CatBoost Classifier. Selain itu, bab ini juga menjelaskan konsep Explainable Artificial Intelligence (XAI) melalui analisis SHAP serta meninjau beberapa penelitian terdahulu yang relevan sebagai dasar penguatan teoritis penelitian.

### **BAB III METODOLOGI PENELITIAN**

Bab ini menjelaskan pendekatan penelitian, jenis dan sumber data, variabel yang digunakan, serta tahapan analisis data. Proses penelitian meliputi pengumpulan dan preprocessing data HR, penerapan algoritma Multinomial Logistic Regression dan CatBoost Classifier, serta evaluasi performa model menggunakan metrik Accuracy, F1-Score, dan AUC (Area Under Curve). Di bagian akhir, dijelaskan pula proses interpretasi model menggunakan analisis SHAP untuk mengidentifikasi faktor-faktor dominan yang memengaruhi retensi karyawan.

## **BAB IV        HASIL DAN PEMBAHASAN**

Bab ini menyajikan hasil implementasi model prediksi retensi karyawan serta perbandingan performa antara kedua algoritma yang digunakan. Selain itu, ditampilkan hasil interpretasi SHAP terhadap variabel-variabel penting serta pembahasan mengenai rekomendasi strategi mode kerja (WFH, WFO, Hybrid) sebagai solusi untuk meningkatkan retensi dan kepuasan kerja karyawan. Bab ini juga dilengkapi dengan visualisasi hasil analisis dan interpretasi dari setiap model yang digunakan.

## **BAB V        PENUTUP**

Bab ini memuat kesimpulan yang diperoleh dari hasil penelitian serta saran untuk penelitian lanjutan. Kesimpulan difokuskan pada hasil perbandingan performa model, identifikasi faktor dominan yang memengaruhi retensi karyawan, serta rekomendasi kebijakan mode kerja yang paling sesuai. Saran diberikan untuk pengembangan riset lanjutan dalam penerapan strategi data driven decision making di bidang manajemen sumber daya manusia.

## **BAB II TINJAUAN PUSTAKA**

### **2.1 Landasan Teori**

#### **2.1.1 Human Resource Analytics (HRA)**

Human Resource Analytics (HRA) merupakan pendekatan ilmiah dalam pengambilan keputusan berbasis data pada bidang manajemen sumber daya manusia. Menurut [\(Wu et al. 2024\)](#), analisis berbasis algoritma memungkinkan HR untuk menilai kinerja, memprediksi perilaku, dan meningkatkan produktivitas karyawan melalui proses yang terukur dan objektif. Dengan analitik prediktif, organisasi dapat mengenali pola yang memengaruhi loyalitas, retensi, dan tingkat turnover. [\(Kasaie and Rajendran 2023\)](#) menambahkan bahwa integrasi machine learning dengan prinsip analitik SDM membuka peluang untuk menghasilkan keputusan yang lebih adaptif dan presisi tinggi. Dalam konteks penelitian ini, HRA digunakan untuk membangun model prediksi retensi karyawan yang memanfaatkan dua pendekatan berbeda Multinomial Logistic Regression sebagai model statistik interpretatif dan CatBoost Classifier sebagai model machine learning berbasis boosting.

#### **2.1.2 Retensi dan Turnover Karyawan**

Retensi karyawan mengacu pada kemampuan organisasi dalam mempertahankan karyawan potensial agar tetap bekerja secara berkelanjutan. Menurut [\(Chaudhary et al. 2025\)](#), retensi yang baik berkaitan erat dengan kepuasan kerja, kejelasan karier, dan kompensasi yang adil. Sementara itu, [\(Ahammod Bin Atique et al. 2025\)](#) menemukan bahwa organisasi yang mampu mengidentifikasi faktor-faktor penyebab turnover melalui analisis prediktif akan memiliki keunggulan kompetitif dalam mengurangi kehilangan tenaga kerja berkualitas. Dalam penelitian ini, teori retensi menjadi dasar pemilihan tiga variabel utama, yaitu lama bekerja, gaji, dan jabatan. Ketiganya dianggap sebagai indikator penting yang menggambarkan loyalitas, kesejahteraan, dan posisi karir karyawan. Dengan pendekatan prediktif berbasis machine learning, penelitian ini berupaya memperkirakan kecenderungan retensi dari setiap karyawan berdasarkan ketiga faktor tersebut.

### **2.1.3 Mode Bekerja dan Retensi Karyawan**

Perubahan model kerja dari konvensional ke sistem fleksibel (WFH, WFO, dan Hybrid) telah mengubah cara organisasi mengelola tenaga kerja. Menurut ([Agrawal et al. 2025](#)), fleksibilitas mode kerja berkontribusi positif terhadap peningkatan kepuasan dan loyalitas, terutama bila didukung dengan sistem manajemen kinerja berbasis data. Namun, penelitian oleh ([Chaudhary et al. 2025](#)) menunjukkan bahwa pengaruh mode kerja terhadap retensi tidak bersifat universal, melainkan bergantung pada jabatan, pengalaman kerja, serta tingkat kompensasi. Dalam konteks penelitian ini, hasil prediksi retensi berdasarkan lama bekerja, gaji, dan jabatan akan menjadi dasar dalam merumuskan rekomendasi strategi mode kerja yang optimal. Pendekatan ini diharapkan membantu organisasi memilih model kerja yang paling sesuai untuk mempertahankan karyawan berpotensi tinggi.

## **2.2 Teori dan Metode Pendukung**

### **2.2.1 Machine Learning dan Explainable Artificial Intelligence (XAI)**

Machine Learning (ML) merupakan cabang kecerdasan buatan yang memungkinkan sistem komputer belajar dari data tanpa pemrograman eksplisit. Menurut ([Ahammod Bin Atique et al. 2025](#)), ML mampu mengenali pola kompleks antar variabel yang sulit dijelaskan dengan metode statistik tradisional, sehingga cocok untuk membangun model prediktif seperti retensi karyawan. Dalam konteks ini, dua pendekatan yang digunakan adalah Multinomial Logistic Regression (MLR) sebagai model interpretatif, dan CatBoost Classifier yang unggul dalam menangani data kategorikal serta hubungan non linier. Salah satu tantangan utama ML adalah kurangnya transparansi model.

Untuk itu digunakan Explainable Artificial Intelligence (XAI), yang bertujuan menjelaskan proses dan hasil prediksi. Menurut Agrawal, Singh, & Patel (2025), XAI khususnya metode SHAP (Shapley Additive Explanations) dapat menunjukkan kontribusi setiap variabel terhadap hasil prediksi, sehingga meningkatkan kepercayaan pengguna terhadap model. Penelitian ([Chaudhary et al. 2025](#)) dan ([Wu et al. 2024](#)) menegaskan bahwa penerapan XAI dalam sistem prediktif SDM meningkatkan pemahaman manajerial terhadap faktor penyebab turnover dan kinerja, menjadikan hasil analisis ML lebih transparan dan aplikatif dalam pengambilan keputusan strategis.

### **2.2.2 Konsep Dasar Prediksi**

Prediksi adalah proses memperkirakan suatu hasil berdasarkan pola data masa lalu. [\(Abasi et al. 2025\)](#) menjelaskan bahwa model prediktif dalam Machine learning bekerja melalui dua tahap, yaitu training dan testing, yang digunakan untuk memvalidasi kemampuan model dalam mengenali pola data. Dalam penelitian ini, konsep prediksi digunakan untuk memperkirakan tingkat retensi karyawan berdasarkan variabel lama bekerja, gaji, dan jabatan. Hasil prediksi akan membantu HR dalam menetapkan strategi retensi yang berbasis bukti.

### **2.2.3 Multinomial Logistic Regression**

Multinomial Logistic Regression (MLR) merupakan metode statistik yang digunakan untuk memodelkan hubungan antara beberapa variabel independen dan variabel dependen kategorikal dengan lebih dari dua kelas. Menurut [\(Kasaie and Rajendran 2023\)](#), Multinomial Logistic Regression (MLR) memberikan interpretasi yang mudah dipahami karena menghasilkan koefisien yang menjelaskan arah pengaruh variabel terhadap peluang suatu kategori. Dalam penelitian ini, Multinomial Logistic Regression (MLR) digunakan sebagai baseline model untuk memprediksi retensi karyawan. Keunggulan Multinomial Logistic Regression (MLR) terletak pada kemampuannya menjelaskan hubungan variabel lama bekerja, gaji, dan jabatan secara eksplisit terhadap kecenderungan bertahan.

### **2.2.4 Catboost Classifier**

CatBoost adalah algoritma gradient boosting berbasis decision tree yang dirancang untuk menangani data kategorikal dengan efisien. Menurut [\(Ahammod Bin Atique et al. 2025\)](#), CatBoost memiliki kemampuan ordered boosting yang dapat mengurangi overfitting dan meningkatkan akurasi pada dataset berukuran kecil hingga menengah. Dalam penelitian ini, CatBoost digunakan sebagai model pembanding terhadap Multinomial Logistic Regression (MLR). Fokus utama adalah mengukur seberapa baik CatBoost mampu memprediksi retensi dibandingkan Multinomial Logistic Regression (MLR), serta mengevaluasi sejauh mana hasil prediksinya dapat diinterpretasikan melalui analisis SHAP.

### 2.2.5 Evaluasi Model Klasifikasi

Evaluasi model bertujuan menilai performa algoritma dalam menghasilkan prediksi yang akurat. ([Abasi et al. 2025](#)) menyebutkan beberapa metrik yang umum digunakan, yaitu Accuracy, Precision, Recall, F1-Score, dan AUC (Area Under Curve). Dalam penelitian ini, metrik tersebut digunakan untuk membandingkan performa MLR dan CatBoost. Model dengan nilai AUC tertinggi dan keseimbangan metrik terbaik akan dipilih sebagai model optimal untuk interpretasi selanjutnya.

### 2.2.6 Analisis SHAP

SHAP merupakan metode Explainable Artificial Intelligence (XAI) yang memberikan penjelasan matematis terhadap kontribusi setiap fitur dalam hasil prediksi. ([Agrawal et al. 2025](#)) menyebutkan bahwa SHAP menggunakan prinsip teori permainan (Game Theory) untuk menentukan pengaruh relatif dari tiap fitur terhadap output model. Dalam penelitian ini, SHAP digunakan untuk menganalisis seberapa besar pengaruh lama bekerja, gaji, dan jabatan terhadap probabilitas retensi karyawan. Hasil interpretasi SHAP membantu manajemen SDM memahami faktor utama yang memengaruhi keputusan bertahan karyawan.

## 2.3 Penelitian Terdahulu

Berikut merupakan ringkasan penelitian terdahulu yang menjadi dasar dalam perumusan metodologi penelitian ini :

*Tabel 1. Penelitian Terdahulu*

No	Peneliti & Tahun	Judul Penelitian	Metode atau Algoritma	Fokus dan Hasil Utama	Relevansi dengan Penelitian Ini

1	Kasaie & Rajendran (2023)	<i>Integrating Machine Learning Algorithms and Explainable Artificial Intelligence Approach for Predicting Patient Unpunctuality in Psychiatric Clinics</i>	Multinomial Logistic Regression & SHAP	Menunjukkan transparansi model prediksi berbasis XAI	Memberi dasar metodologis untuk integrasi MLR & SHAP
2	Chaudhary et al. (2025)	<i>An Integrated Model to Evaluate the Transparency in Predicting Employee Churn Using Explainable AI</i>	Random Forest & SHAP	XAI meningkatkan kepercayaan hasil prediksi churn	Relevan untuk retensi karyawan dengan pendekatan XAI
3	Ahammod Bin Atique et al. (2025)	<i>Enhancing Employee Turnover Prediction: An Advanced Feature Engineering Analysis with CatBoost</i>	CatBoost	Akurasi tinggi dengan <i>feature engineering</i>	Menjadi acuan performa CatBoost dalam data HR
4	Agrawal et al. (2025)	<i>Fostering Trust and Interpretability: Integrating Explainable AI with ML</i>	XAI + Machine Learning	SHAP meningkatkan transparansi keputusan model	Digunakan untuk bagian interpretabilitas model

5	Varkiani et al. (2025)	<i>Predicting Employee Attrition XAI-Power Models for Managerial Decision Making</i>	CatBoost + SHAP	Identifikasi faktor utama turnover melalui XAI	Relevan untuk analisis variabel lama bekerja, gaji, jabatan
6	Fatemi Aghda et al. (2025)	<i>Machine Learning Models for Reinjury Risk Prediction Using CPET Data</i>	Ensemble + AUC	Meningkatkan akurasi prediksi hingga 93%	Digunakan sebagai acuan evaluasi model prediktif
7	Wu et al. (2024)	<i>Unbiased Employee Performance Evaluation Using Machine Learning</i>	Random Forest	Evaluasi kinerja HR berbasis data	Memberi dasar teoritis HRA dan analitik SDM
8	Fadilah & Rahmawati (2023)	<i>Data Analytics for Optimizing and Predicting Employee Performance</i>	Gradient Boosting	Hubungan faktor kerja dengan performa SDM	Relevan untuk konteks HR dan retensi karyawan

## 2.4 Analisis Penelitian Terdahulu

Berdasarkan telaah literatur di atas, sebagian besar penelitian terdahulu telah menggunakan algoritma machine learning untuk menganalisis turnover dan performa karyawan. Namun, mayoritas studi tersebut masih berfokus pada aspek akurasi model dan belum mengintegrasikan explainable AI secara mendalam dalam konteks retensi karyawan di Indonesia. Penelitian oleh ([Kasaie and Rajendran 2023](#)) dan ([Chaudhary et al. 2025](#)) menyoroti pentingnya interpretabilitas model, namun belum membahas perbandingan langsung antara pendekatan statistik klasik seperti MLR dengan model boosting modern seperti CatBoost. Sementara itu, penelitian ([Ahammod Bin Atique et al. 2025](#)) lebih menitikberatkan pada optimasi performa tanpa analisis kontekstual terhadap

faktor SDM. Oleh karena itu, penelitian ini mengisi celah (research gap) tersebut dengan membandingkan dua pendekatan Multinomial Logistic Regression dan CatBoost Classifier untuk memprediksi retensi karyawan berdasarkan lama bekerja, gaji, dan jabatan. Lebih lanjut, integrasi analisis SHAP diharapkan memberikan nilai tambah berupa transparansi hasil dan rekomendasi kebijakan mode kerja (WFH, WFO, Hybrid) yang berbasis bukti.

## **BAB III METODOLOGI PENELITIAN**

### **3.1 Jenis dan Pendekatan Penelitian**

Penelitian ini menggunakan pendekatan kuantitatif dengan metode komparatif prediktif berbasis machine learning. Pendekatan ini digunakan untuk membandingkan dua algoritma utama, yaitu Multinomial Logistic Regression (MLR) dan CatBoost Classifier, dalam memprediksi retensi karyawan berdasarkan variabel lama bekerja, gaji, dan jabatan. Tujuan dari pendekatan ini bukan hanya untuk melihat performa prediksi yang paling optimal, tetapi juga untuk mengidentifikasi variabel yang paling berpengaruh terhadap keputusan karyawan untuk tetap bekerja atau keluar dari perusahaan. Selain itu, metode Explainable Artificial Intelligence (XAI) melalui analisis SHAP (Shapley Additive Explanations) digunakan untuk menjelaskan hasil prediksi secara interpretatif agar dapat dimanfaatkan dalam kebijakan sumber daya manusia.

### **3.2 Sumber dan Karakteristik Data**

Dataset yang digunakan bersumber dari Kaggle berjudul HR Data MNC (Rohit Grewal, 2023). Dataset ini terdiri atas 15 kolom (atribut) dan 1.500 baris data karyawan yang mencakup variabel demografis (umur, jenis kelamin, status pernikahan), profesional (departemen, lama bekerja, gaji), dan perilaku (evaluasi kerja, absensi, promosi). Dataset ini diterbitkan oleh pengguna Rohit Grewal di bawah lisensi terbuka Creative Commons Public Domain (CC0) dan dapat diakses melalui tautan:

Link Dataset : <https://www.kaggle.com/datasets/rohitgrewal/hr-data-mnc>

Menurut dokumentasi Kaggle, dataset ini digunakan untuk penelitian analitik SDM di sektor korporasi multinasional (MNC) yang bertujuan mengidentifikasi faktor-faktor penyebab employee turnover dan retensi kerja. Variabel-variabel yang digunakan akan disesuaikan dengan indikator penelitian ini, yaitu retensi dan kinerja karyawan.

### **3.3 Tahapan Penelitian**

Metodologi penelitian mengacu pada tahapan umum machine learning pipeline sebagaimana diuraikan oleh ([Permatasari 2022](#)) dan ([Nawawi 2024](#)) :

1. Unduh data,

Data diunduh dari Kaggle dan dibersihkan dari *missing values* serta duplikasi. Validasi tipe data dilakukan untuk memastikan setiap variabel sesuai dengan tipe numerik atau kategorikal yang diperlukan.

2. Pre-processing dan Feature Engineering,

Dilakukan proses:

- Handling missing data menggunakan teknik imputasi median.
- Label encoding pada variabel kategorikal seperti gender dan departemen.
- Feature scaling dengan metode Min-Max Normalization untuk menjaga keseimbangan distribusi nilai antar variabel.

Tahap ini penting karena menurut [\(Franseda 2020\)](#), ketidakseimbangan data dapat menurunkan akurasi model klasifikasi

3. Pembagian Data (Data Splitting),

Dataset dibagi menjadi 80% data pelatihan (training) dan 20% data pengujian (testing) menggunakan teknik Stratified Sampling agar proporsi kelas target tetap terjaga. Pembagian ini mengikuti praktik standar yang digunakan oleh [\(Sukmawati 2022\)](#) pada penelitian prediksi karir berbasis machine learning.

4. Pemodelan Algoritma,

Tiga model utama diuji:

- CatBoost Classifier

Dipilih karena kemampuannya menangani data kategorikal tanpa perlu one-hot encoding, serta menghasilkan interpretasi dengan Shapley Additive Explanations (SHAP).

- Gradient Boosting Classifier

Digunakan untuk menguji stabilitas model ensemble terhadap variasi data numerik.

- Multinomial Logistic Regression

Sebagai pembanding tradisional yang memiliki interpretasi kuat terhadap pengaruh variabel prediktor.

## 5. Evaluasi Model

Evaluasi dilakukan menggunakan metrik:

- Akurasi (Accuracy)
- Presisi (Precision)
- Recall (Sensitivitas)
- F1-Score

Penggunaan metrik majemuk ini mengacu pada [\(Handayani et al. 2024\)](#) yang menunjukkan bahwa penggunaan satu metrik tunggal tidak cukup mewakili performa model kompleks.

### 3.4 Validasi dan Interpretasi Model

Validasi dilakukan dengan membandingkan hasil antar model menggunakan analisis Receiver Operating Characteristic (ROC) dan Area Under Curve (AUC). Model dengan nilai AUC tertinggi dianggap paling optimal dalam memprediksi kemungkinan karyawan bertahan atau keluar dari perusahaan. Selain performa numerik, interpretasi hasil dilakukan menggunakan analisis SHAP yang memungkinkan pembimbing memahami pengaruh tiap fitur misal gaji, lama bekerja, performa evaluasi terhadap probabilitas retensi karyawan. Pendekatan interpretatif ini direkomendasikan oleh [\(Ibrahim 2020\)](#) dan [\(Iang and Wang 2021\)](#) pada studi serupa tentang prediksi sosial-ekonomi berbasis CatBoost.

## **BAB IV HASIL DAN PEMBAHASAN**

### **4.1 Gambaran Umum Proses Analisis**

Bagian ini akan berisi deskripsi tahapan analisis yang dilakukan pada data, dimulai dari data preprocessing, pelatihan model, hingga interpretasi hasil dengan SHAP. Hasil yang disajikan akan meliputi:

- Tabel perbandingan performa model MLR dan CatBoost
- Visualisasi confusion matrix dan ROC curve
- Analisis interpretasi variabel dengan SHAP summary plot

### **4.2 Analisis Model**

Pada tahap ini akan dilakukan pembahasan mengenai:

- Hasil uji performa model berdasarkan metrik evaluasi.
- Perbandingan hasil antara MLR dan CatBoost dalam hal akurasi dan kemampuan klasifikasi.
- Pengaruh masing-masing variabel terhadap probabilitas retensi berdasarkan hasil analisis SHAP.

Penjelasan akan disusun secara deskriptif dan interpretatif agar mudah dipahami oleh pembimbing dan pihak manajemen SDM.

### **4.3 Pembahasan Temuan (Rencana)**

Pembahasan di bab ini akan menghubungkan hasil prediksi dengan teori retensi karyawan, sistem kompensasi, dan strategi mode kerja (WFO, WFH, Hybrid).

Analisis ini diharapkan dapat menjawab pertanyaan utama yaitu : “Bagaimana variabel lama bekerja, gaji, dan jabatan memengaruhi tingkat retensi karyawan, dan mode kerja apa yang paling efektif untuk mempertahankannya?”

## **BAB V KESIMPULAN**

### **5.1 Kesimpulan (Rencana)**

Bab ini akan berisi kesimpulan yang diharapkan berdasarkan analisis data dan perbandingan model. Fokus utama kesimpulan adalah:

- Model mana yang paling efektif dalam memprediksi retensi.
- Variabel mana yang paling dominan memengaruhi keputusan retensi.
- Implikasi hasil untuk strategi mode kerja di perusahaan.