

# **BUSINESS REQUIREMENT DOCUMENT**

*Studi Komparatif Multinomial Logistic Regression dan CatBoost Classifier Berbasis Analisis SHAP dalam Memprediksi Tingkat Risiko Attrition Karyawan sebagai Dasar Penentuan Strategi Mode Bekerja*

---

## 1. Latar Belakang

Perusahaan/organisasi menghadapi tantangan menjaga retensi karyawan dalam dinamika pola kerja (WFO/Hybrid/WFH). Tingginya attrition meningkatkan biaya rekrutmen, pelatihan, dan hilangnya produktivitas. Karena itu dibutuhkan pendekatan berbasis data untuk memprediksi kecenderungan karyawan resign (attrition), memahami faktor dominan yang mempengaruhi attrition, dan mengubah hasil prediksi menjadi segmentasi risiko (Rendah/Sedang/Tinggi) untuk membantu penentuan strategi mode bekerja yang tepat.

## 2. Tujuan Projek

Adapun tujuan projek dalam penelitian ini meliputi :

1. Membangun model prediksi attrition dengan target biner :  
Active = 0 dan Resigned = 1
2. Membandingkan performa model:  
Logistic Regression (sebagai baseline)  
CatBoost Classifier
3. Menggunakan SHAP untuk interpretasi dan identifikasi faktor dominan.
4. Menghasilkan probabilitas resign (risk score) dan memetakannya menjadi Rendah/Sedang/Tinggi.
5. Menyusun rekomendasi strategi WFO/Hybrid/WFH berbasis segmentasi risiko.

## 3. Ruang Lingkup (Scope)

Adapun yang termasuk ke dalam ruang lingkup penelitian ini meliputi :

1. In scope

Prediksi attrition berbasis 3 variabel utama yaitu Experience\_Years (lama bekerja), Salary\_INR (gaji) dan Job\_Title (jabatan), Membuat model dan

evaluasi kinerja (Logistic Regression vs CatBoost), Interpretabilitas model dengan SHAP, Segmentasi risiko dari probabilitas prediksi dan Rekomendasi strategi mode bekerja (WFO/Hybrid/WFH) berdasarkan risiko.

## 2. Out of scope

Penambahan variabel lain (mis. usia, gender, departemen, performa, dll) bila tidak tersedia/di luar batasan penelitian, Implementasi aplikasi produksi/HRIS penuh (jika belum diminta) dan Kebijakan HR final yang mengikat (hasil berupa rekomendasi analitis).

## 4. Kebutuhan Fungsional

### FR-01 Pengolahan Data

Sistem/analisis harus mampu:

- Memuat dataset dan melakukan pembersihan data (missing value, duplikasi, encoding kategorikal).
- Menetapkan target attrition menjadi biner (Active/Resigned).

### FR-02 Pemodelan

- Membangun model Logistic Regression dan CatBoost Classifier.
- Menghasilkan output prediksi kelas dan probabilitas Resigned untuk tiap karyawan.

### FR-03 Evaluasi Model

- Menghitung metrik evaluasi klasifikasi.
- Menyediakan confusion matrix dan ringkasan perbandingan antar model.

### FR-04 Interpretabilitas (XAI)

Menghasilkan analisis SHAP:

- feature importance global
- kontribusi fitur terhadap prediksi (lokal/individual bila diperlukan)

### FR-05 Segmentasi Risiko

Mengonversi probabilitas Resigned (risk score) menjadi 3 kategori (Rendah/Sedang/Tinggi) dengan aturan threshold yang didefinisikan dalam analisis/penelitian)

#### FR-06 Rekomendasi Strategi Mode Bekerja

Menghasilkan rekomendasi WFO/Hybrid/WFH berdasarkan kategori risiko.

### 5. Kebutuhan Non-Fungsional

- Reproducibility: proses analisis dapat dijalankan ulang dengan hasil konsisten (pipeline jelas).
- Transparansi: model dapat dijelaskan (via SHAP + ringkasan aturan segmentasi).
- Keterbacaan output: hasil disajikan dalam bentuk tabel/grafik yang mudah dipahami HR/Manajemen.
- Maintainability: struktur kode/notebook rapi (bisa dikembangkan bila variabel bertambah).

### 6. Analisis SWOT

#### a. Strengths

Prediksi berbasis data mengurangi keputusan subjektif.

Ada interpretabilitas (SHAP) → lebih mudah diterima manajemen.

#### b. Weaknesses

Batasan hanya 3 variabel bisa membatasi akurasi/kedalaman insight.

Ketergantungan pada kualitas data historis.

#### c. Opportunities

Bisa dikembangkan dengan variabel HR lain dan diintegrasikan ke dashboard.

Segmentasi risiko dapat membantu program retensi yang lebih tepat sasaran.

#### d. Threats

Perubahan kebijakan kerja/perusahaan bisa mengubah pola data

Bias data historis dapat memengaruhi rekomendasi jika tidak dikontrol.

## 7. Kebutuhan Data

### a. Atribut Data Minimum

Experience\_Years (numerik)

Salary\_INR (numerik)

Job\_Title (kategorikal)

Attrition/Status (Active/Resigned) sebagai target

### b. Aturan Pra-pemrosesan (ringkas)

Missing value ditangani (drop/impute sesuai konteks).

Encoding untuk Job\_Title (mis. one-hot/target encoding sesuai metode).

Split data latih-uji (mis. train-test split).

## 8. Indikator Evaluasi

Metrik utama untuk membandingkan model :

- Accuracy
- Precision
- Recall
- F1-Score
- ROC-AUC (untuk klasifikasi biner)
- Confusion Matrix
- SHAP Feature Importance Ranking (interpretasi, bukan “skor akurasi”)

Catatan: Precision/Recall penting karena konteks HR sering lebih sensitif terhadap false negative/false positive (mis. karyawan berisiko tinggi tapi tidak terdeteksi).

## 9. Business Impact & Deliverables

Proyek ini membantu HR memprediksi risiko attrition lebih dini sehingga intervensi retensi dapat diprioritaskan secara tepat sasaran dan biaya akibat pergantian karyawan bisa ditekan. Hasil prediksi dibuat lebih transparan melalui interpretasi SHAP, lalu probabilitas resign dipetakan menjadi risiko Rendah/Sedang/Tinggi sebagai dasar rekomendasi mode kerja (WFO/Hybrid/WFH). Deliverables mencakup data yang sudah diproses, perbandingan performa Logistic Regression vs CatBoost beserta

metrik evaluasi dan visualisasi pendukung, analisis SHAP, segmentasi risiko, serta rekomendasi strategi mode bekerja.

## 10. Business Impact & Deliverables

Pelaksanaan dimulai dari pengumpulan dan pembersihan data, dilanjutkan EDA. Setelah itu dibangun model Logistic Regression sebagai baseline dan CatBoost sebagai pembanding, kemudian dievaluasi dengan metrik yang sama. Tahap berikutnya adalah analisis SHAP untuk interpretasi, pembuatan segmentasi risiko dari probabilitas resign, dan penyusunan rekomendasi mode kerja. Terakhir, seluruh hasil dirangkum dalam laporan akhir.