

BUSINESS REQUIREMENT DOCUMENT

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

1. Latar Belakang

Organisasi modern menghadapi tantangan tinggi dalam mempertahankan karyawan di tengah perubahan pola kerja hybrid (WFH, WFO, dan kombinasi). Tingginya tingkat turnover menyebabkan biaya rekrutmen, pelatihan, dan hilangnya produktivitas. Diperlukan sistem berbasis data untuk memprediksi retensi karyawan secara akurat dan menjelaskan faktor penyebabnya. Model statistik tradisional (seperti Multinomial Logistic Regression) mampu memberikan interpretasi, tetapi akurasinya sering terbatas. Sebaliknya, algoritma modern (CatBoost Classifier dengan SHAP Analysis) mampu memberikan hasil prediktif yang lebih akurat sekaligus menjelaskan faktor dominan.

2. Tujuan Proyek

Membangun model komparatif berbasis machine learning dan statistical learning untuk :

- Memprediksi kemungkinan retensi karyawan berdasarkan lama bekerja, gaji, dan jabatan.
 - Menganalisis faktor dominan yang memengaruhi keputusan bertahan atau keluar menggunakan analisis SHAP.
 - Memberikan rekomendasi strategis mode bekerja (WFH/WFO/Hybrid) berbasis hasil model.
-

3. Ruang Lingkup (Scope)

Aspek	Deskripsi
Data Input	Dataset SDM yang mencakup: lama bekerja (tahun), gaji bulanan, jabatan, status kerja, dan variabel demografis opsional.
Metode yang Digunakan	Multinomial Logistic Regression (MLR) dan CatBoost Classifier (dengan SHAP interpretability).
Tujuan Analitik	Prediksi status retensi: bertahan, berisiko keluar, keluar.
Output Sistem	Model komparatif dengan nilai akurasi dan interpretasi fitur.- Visualisasi feature importance (via SHAP). Rekomendasi strategi mode bekerja per kategori jabatan.
Lingkup Eksperimen	Data perusahaan atau dataset publik (karyawan) selama 3–5 tahun terakhir.

4. Kebutuhan Fungsional (Functional Requirements)

No	Kebutuhan	Deskripsi

F1	Data Preparation Module	Melakukan data cleaning, encoding variabel kategorikal, dan normalisasi gaji.
F2	Model Building (MLR)	Melatih model Multinomial Logistic Regression untuk baseline interpretasi.
F3	Model Building (CatBoost)	Melatih model CatBoost Classifier dengan optimasi hyperparameter (GridSearchCV).
F4	Explainability Layer (SHAP)	Menampilkan hasil feature importance, dependence plots, dan force plot untuk interpretasi faktor retensi.
F5	Model Comparison	Menghasilkan tabel perbandingan akurasi, F1-score, dan ROC-AUC antara MLR dan CatBoost.
F6	Recommendation Engine	Memberikan saran mode bekerja optimal (WFH/WFO/Hybrid) berdasarkan skor retensi per jabatan.
F7	Reporting Dashboard	Menyajikan hasil visual dalam bentuk grafik dan tabel untuk manajemen HR.

5. Kebutuhan Non-Fungsional (*Non-Functional Requirements*)

Aspek	Kriteria

Akurasi Model	Minimal 85% (dengan perbedaan $\leq 5\%$ antara training & test).
Interpretabilitas	SHAP values wajib dihasilkan untuk tiap prediksi individual.
Waktu Komputasi	Maksimum 5 menit per model (dataset ≤ 10.000 baris).
Konsistensi Data	Tidak ada missing values; variabel distandarisasi.
Reusability	Model dapat diimplementasikan ulang untuk dataset SDM lain.
Keamanan Data	Data pribadi karyawan dianonimkan (tanpa nama/NIK).

6. Analisis SWOT dari Perspektif Bisnis

Aspek	Uraian
Strengths	Memberikan wawasan kuantitatif terhadap faktor retensi. Menggunakan model modern yang dapat dijelaskan (Explainable AI).
Weaknesses	Ketergantungan pada data internal HR yang sering tidak lengkap Membutuhkan pemahaman teknis tinggi untuk SHAP interpretasi.
Opportunities	Dapat digunakan sebagai HR Decision Support System.

	Potensi implementasi ke sistem rekomendasi mode kerja adaptif.
Threats	Risiko bias model akibat data tidak seimbang. Potensi kesalahan interpretasi SHAP jika tidak dikontekstualisasikan.

7. Kebutuhan Data (*Data Requirements*)

Jenis Data	Contoh Atribut	Sumber	Format
Data Demografi Karyawan	Usia, Jenis Kelamin, Pendidikan	HR Database	CSV / SQL
Data Jabatan	Jabatan, Level, Divisi	Struktur Organisasi	Excel / CSV
Data Keuangan	Gaji, Bonus	Payroll	CSV
Data Retensi	Lama bekerja, Status (bertahan/keluar)	HR Exit Data	CSV

8. Indikator Evaluasi (*Evaluation Metrics*)

1. Akurasi (Accuracy)
2. F1-Score (macro average)

-
3. ROC-AUC (One-vs-Rest)
 4. Confusion Matrix
 5. SHAP Feature Importance Ranking
-

9. Business Impact & Deliverables

<i>Dampak</i>	<i>Deskripsi</i>
<i>Efisiensi HR</i>	<i>Mengurangi waktu analisis penyebab turnover melalui model otomatis.</i>
<i>Strategi SDM Adaptif</i>	<i>Menentukan pola kerja (WFH/WFO/Hybrid) berbasis data.</i>
<i>Decision Transparency</i>	<i>Model dapat dijelaskan kepada manajemen non-teknis (via SHAP).</i>
<i>Deliverables Utama</i>	<i>Model MLR dan CatBoost terlatih Dashboard visual hasil SHAP Laporan komparatif performa model Rekomendasi mode kerja per jabatan.</i>

10. Timeline Implementasi (Estimasi)

Fase	Aktivitas	Durasi

1	Pengumpulan & Pembersihan Data	2 minggu
2	Eksperimen MLR & CatBoost	3 minggu
3	Evaluasi & Validasi Model	2 minggu
4	Analisis SHAP & Visualisasi	2 minggu
5	Penyusunan Rekomendasi & Laporan	1 minggu
Total Durasi	10 minggu ($\pm 2,5$ bulan)	

11. Kesimpulan BRD

Dokumen BRD ini menetapkan kebutuhan fungsional, teknis, dan bisnis untuk mengembangkan sistem prediksi retensi karyawan berbasis kombinasi model statistik dan machine learning. Hasil akhir diharapkan tidak hanya menghasilkan model dengan akurasi tinggi, tetapi juga dapat dijelaskan secara transparan, sehingga membantu organisasi dalam merancang strategi mode bekerja yang optimal dan adaptif terhadap pola kerja masa kini.