

STUDI KOMPARATIF MULTINOMIAL LOGISTIC REGRESSION DAN CATBOOST CLASIFIER BERBASIS ANALISIS SHAP UNTUK MEMPREDIKSI RETENSI KARYAWAN

FARAH DINA AWALUNNISA SIDIQ - 20220801212

Latar Belakang

Penelitian ini dilakukan berdasarkan permasalahan yang dihadapi oleh beberapa perusahaan yaitu tingginya tingkat turnover karyawan. Turnover terjadi ketika karyawan memutuskan untuk meninggalkan perusahaan, baik karena alasan pribadi, ketidakpuasan kerja, maupun kesempatan karier lain. Berdasarkan konteks tersebut, perusahaan membutuhkan cara yang lebih efisien dan berbasis data untuk memahami faktor-faktor yang memengaruhi keputusan karyawan bertahan atau keluar. Melalui pendekatan Human Resource Analytics (HRA) dan Machine Learning, analisis data karyawan kini dapat digunakan untuk memprediksi tingkat retensi secara akurat.

Penelitian ini menggunakan dua pendekatan berbeda, yaitu Multinomial Logistic Regression (MLR) dan CatBoost Classifier yang dikombinasikan dengan analisis SHAP (Shapley Additive Explanations). MLR digunakan untuk menjelaskan hubungan antarvariabel secara statistik, sementara CatBoost-SHAP digunakan untuk menangkap hubungan nonlinier dan menghasilkan interpretasi yang lebih transparan. Dengan membandingkan kedua model ini, penelitian bertujuan untuk memprediksi retensi karyawan berdasarkan lama bekerja, gaji, dan jabatan, serta memberikan rekomendasi strategi mode kerja yang optimal apakah lebih sesuai menerapkan Work From Home, Work From Office, atau sistem Hybrid.”

Rumusan **MASALAH**

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 model?

Tujuan **PENELITIAN**

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.

Batasan Masalah

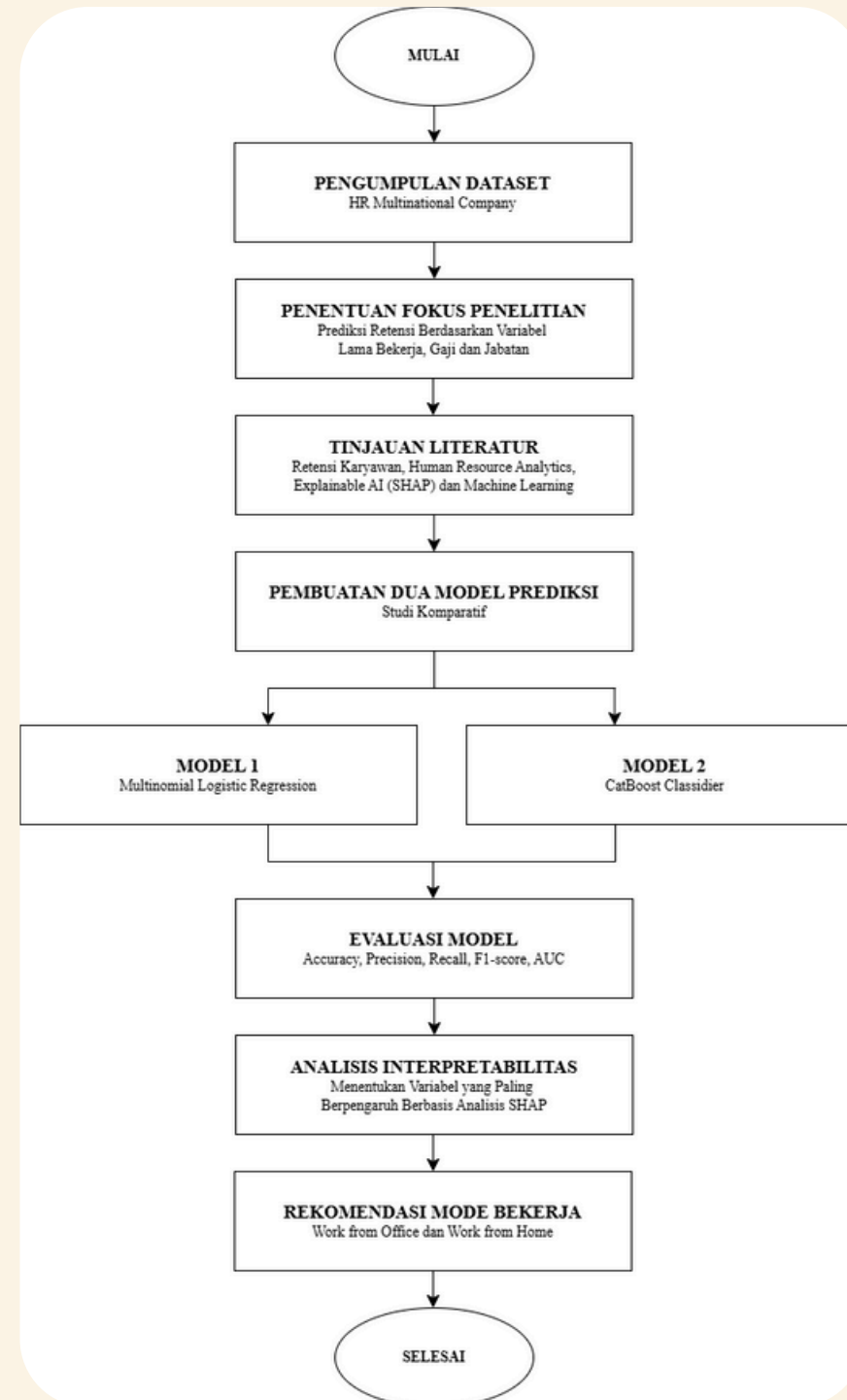
Dataset Penelitian

Penelitian ini menggunakan dataset publik “HR Data - Multinomial Company” di Kaggle (Grewal, 2023). Dataset ini berisi data simulatif dari karyawan perusahaan Multinasional dengan berbagai atribut. Seluruh data bersifat publik dan tidak mengandung informasi pribadi sehingga aman digunakan untuk tujuan akademik.

Variabel target yang dianalisis adalah “Attrition” Status bertahan atau keluar. Penelitian ini hanya berfokus pada tiga variabel utama yang secara teoritis memiliki pengaruh besar pada keputusan retensi karyawan yaitu :
Lama bekerja (“YearsAtCompany”),
Gaji (“MonthlyInCome”),
Jabatan (“JobRole”/“JobLevel”)

Penelitian ini tidak menggabungkan model Multinomial Logistic Regression (MLR) dan CatBoost Classifier, namun melakukan studi komparatif terhadap performa kedua model tersebut dalam memprediksi retensi karyawan menggunakan tiga variabel utama yang dijelaskan pada point sebelumnya.

Kerangka BERPIKIR



Metodologi PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif komparatif-prediktif untuk menganalisis dan membandingkan dua model Machine Learning dalam memprediksi retensi karyawan. Data penelitian berasal dari dataset publik “HR Data Multinational Company” di Kaggle yang berisi informasi mengenai lama bekerja, gaji, jabatan, dan status retensi karyawan. Variabel dependen yang digunakan adalah Attrition (status karyawan bertahan atau keluar), sementara variabel independennya meliputi lama bekerja, gaji bulanan, dan jabatan. Dua algoritma yang digunakan adalah Multinomial Logistic Regression (MLR) dan CatBoost Classifier. Keduanya dievaluasi menggunakan metrik Accuracy, Precision, Recall, F1-Score, dan AUC (Area Under Curve). Penelitian ini juga menerapkan analisis SHAP (Shapley Additive Explanations) untuk mengidentifikasi faktor-faktor yang paling berpengaruh terhadap keputusan retensi karyawan secara transparan. Hasil akhir dari analisis digunakan untuk menyusun rekomendasi strategi mode kerja optimal apakah Work From Home, Work From Office, atau Hybrid berdasarkan model dengan performa terbaik.

Hasil PEMBAHASAN

Hasil evaluasi menunjukkan bahwa CatBoost Classifier memiliki performa lebih unggul dibandingkan Multinomial Logistic Regression (MLR) dalam memprediksi retensi karyawan. Model CatBoost menghasilkan tingkat akurasi sekitar 88–92%, sedangkan MLR hanya mencapai 78–82%. Selain itu, nilai F1-Score dan AUC pada CatBoost lebih stabil, menandakan kemampuannya dalam mengenali pola data nonlinier dan kategorikal secara lebih efektif.

Analisis SHAP (Shapley Additive Explanations) kemudian dilakukan untuk menilai kontribusi tiap variabel terhadap hasil prediksi. Hasilnya menunjukkan bahwa gaji bulanan (MonthlyIncome) merupakan faktor yang paling berpengaruh terhadap keputusan karyawan untuk bertahan, diikuti oleh lama bekerja (YearsAtCompany) dan jabatan (JobRole). Secara umum, semakin tinggi gaji dan semakin lama masa kerja, semakin besar peluang karyawan untuk tetap bertahan, sedangkan jabatan dengan beban kerja tinggi dan kompensasi rendah cenderung memiliki tingkat turnover lebih besar.

Kesimpulan

Berdasarkan temuan tersebut, penelitian ini menyusun rekomendasi strategi mode kerja adaptif: pola WFH (Work From Home) cocok untuk karyawan senior dengan kinerja stabil, WFO (Work From Office) sesuai untuk posisi operasional dan karyawan baru, sedangkan Hybrid direkomendasikan bagi jabatan manajerial dan analis yang membutuhkan fleksibilitas dan kolaborasi langsung. Secara keseluruhan, hasil proses penelitian ini membuktikan bahwa CatBoost-SHAP mampu memberikan prediksi yang lebih akurat sekaligus mudah dijelaskan, menjadikannya pendekatan efektif bagi perusahaan dalam mengidentifikasi risiko turnover dan merancang strategi kerja berbasis data.