INTISARI

Explainable AI: Pembangkitan Penjelasan Kontrafaktual pada Model Prediksi Atrisi Karyawan dengan Metode CFProto berbasis Algoritma Genetika dan *Determinantal Point Process* (DPP)

Oleh

Ahmad Wildan Jauharul Fuad

22/504334/PA/21690

Kemajuan *machine learning* menghasilkan model prediktif yang akurat, namun seringkali bersifat *black-box*. Dalam konteks pengambilan keputusan bernilai tinggi seperti manajemen atrisi karyawan, ketidaktransparanan ini menimbulkan potensi bias dan risiko etika. *Explainable Artificial Intelligence* (XAI) hadir untuk menjawab tantangan ini, dengan *counterfactual explanation* (CFE) sebagai salah satu pendekatannya. CFE mencari perubahan minimal pada fitur input sehingga prediksi model berubah ke kelas yang diinginkan, menghasilkan skenario *what-if* yang dapat dimanfaatkan sebagai dasar rekomendasi tindakan bagi pengambil keputusan. Namun, metode CFE umumnya belum mampu menjamin CFE yang sekaligus plausibel dan beragam. Penelitian ini mengusulkan CFProto-GA, adaptasi dari CFProto yang menggantikan optimasi berbasis gradien dengan algoritma genetika untuk menangani data bertipe campuran dan model *non-differentiable*. Plausibilitas diukur melalui *reconstruction error* autoencoder, sementara *Determinantal Point Process* (DPP) diterapkan secara *post-hoc* untuk memilih subset CFE yang memaksimalkan kualitas dan keberagaman secara bersamaan.

Eksperimen pada IBM HR Analytics Dataset dengan model AdaBoost menunjukkan bahwa integrasi DPP meningkatkan validitas menjadi sempurna 1,0000 dan keberagaman antar-CFE sebesar 590%, dengan konsekuensi peningkatan jarak (*proximity*) yang moderat. Analisis agregat mengidentifikasi kompensasi dan pengurangan beban kerja sebagai rekomendasi paling konsisten, mengonfirmasi bahwa CFE yang dihasilkan dapat memandu keputusan retensi secara individual.

Kata Kunci: *explainable AI*, penjelasan kontrafaktual, atrisi karyawan, CFProto, algoritma genetika, *Determinantal Point Process*.

ABSTRACT

Explainable AI: Generation of Counterfactual Explanation for Employee Turnover Predictive Model via the CFProto Method based on Genetic Algorithm and Determinantal point process (DPP)

By

Ahmad Wildan Jauharul Fuad

22/504334/PA/21690

Recent advances in machine learning have produced highly accurate predictive models, yet these models are often black-box in nature. In high-stakes contexts such as employee attrition management, this lack of transparency introduces potential bias and ethical risks. Explainable Artificial Intelligence (XAI) has emerged to address this challenge, with counterfactual explanation (CFE) as one of its key approaches. CFE identifies minimal input changes that shift a model's prediction to a desired class, generating what-if scenarios that can serve as actionable recommendations for decision-makers.

However, existing CFE methods generally struggle to simultaneously guarantee plausibility and diversity. This study proposes CFProto-GA, an adaptation of CFProto that replaces gradient-based optimization with a genetic algorithm to handle mixed-type data and non-differentiable models. Plausibility is assessed via autoencoder reconstruction error, while a Determinantal Point Process (DPP) is applied post-hoc to select a CFE subset that jointly maximizes quality and diversity.

Experiments on the IBM HR Analytics Dataset with an AdaBoost model show that DPP integration improves validity to a perfect 1.0000 and inter-CFE diversity by 590%, with only a moderate increase in proximity as a trade-off. Aggregate analysis identifies compensation and workload reduction as the most consistently recommended changes, confirming that the generated CFEs can guide retention decisions at the individual level.

Keywords: explainable AI, counterfactual explanation, employee attrition, CFProto, genetic algorithm, Determinantal Point Process.