

## IMPLEMENTASI CATBOOST UNTUK KLASIFIKASI RISIKO KEKERASAN ANAK BERDASARKAN PROFIL SOSIAL EKONOMI DI KABUPATEN JEMBER

### IMPLEMENTATION CATBOOST FOR CLASSIFYING CHILD VIOLENCE RISK ON SOCIO-ECONOMIC PROFILES IN JEMBER

Alvio Damayanti<sup>1</sup>, Priza Pandunata<sup>2</sup>

E-mail: [alviodynamayanti@gmail.com](mailto:alviodynamayanti@gmail.com)<sup>1</sup>, [priza@unej.ac.id](mailto:priza@unej.ac.id)<sup>2</sup>

Program Studi Informatika, Fakultas Ilmu Komputer, Universitas Jember

#### Abstrak

Kekerasan terhadap anak merupakan permasalahan sosial yang berdampak pada perkembangan fisik, psikologis, dan sosial anak serta dapat memengaruhi kualitas hidup anak dalam jangka panjang. Tingginya kasus kekerasan anak menunjukkan pentingnya pendekatan berbasis data untuk membantu mengidentifikasi wilayah dengan tingkat risiko yang lebih tinggi. Penelitian ini bertujuan mengklasifikasikan tingkat risiko kekerasan anak berdasarkan profil sosial ekonomi rumah tangga di Kabupaten Jember menggunakan algoritma *CatBoost*. Penelitian menggunakan pendekatan kuantitatif non-eksperimental dengan data sosial ekonomi dari 31 kecamatan yang melalui tahap *preprocessing* berupa pembersihan dan transformasi data. Pelabelan risiko dilakukan menggunakan metode kuantil 33-66, sedangkan validasi model menggunakan *Leave-One-Out Cross Validation (LOOCV)*. Selain itu, penelitian ini juga menerapkan *RandomizedSearchCV* untuk memperoleh kombinasi hyperparameter optimal pada model *CatBoost*. Hasil penelitian menunjukkan bahwa model *CatBoost* memperoleh akurasi sebesar 97% dan ROC-AUC sebesar 0,991. Selain itu, rata-rata *validation accuracy* sebesar  $0,968 \pm 0,177$  menunjukkan bahwa performa model relatif stabil pada setiap iterasi validasi. Hasil klasifikasi menunjukkan 12 kecamatan termasuk kategori risiko tinggi, 9 kecamatan risiko sedang, dan 10 kecamatan risiko rendah. Analisis *SHAP* menunjukkan bahwa tingkat pendidikan dan jenis pekerjaan menjadi variabel yang memiliki kontribusi dominan terhadap hasil klasifikasi. Penelitian ini menunjukkan bahwa *CatBoost* mampu memberikan performa klasifikasi yang baik dalam mengidentifikasi risiko kekerasan anak berbasis data sosial ekonomi serta dapat digunakan sebagai pendukung pengambilan keputusan dalam penentuan prioritas wilayah intervensi perlindungan anak.

**Kata kunci:** kekerasan anak, klasifikasi risiko, *CatBoost*, sosial ekonomi, *machine learning*

#### Abstract

Violence against children is a social problem that affects children's physical, psychological, and social development and may influence their quality of life in the long term. The increasing number of child violence cases highlights the importance of a data-driven approach to help identify areas with higher levels of risk. This study aims to classify the level of child violence risk based on household socioeconomic profiles in Jember Regency using the *CatBoost* algorithm. The study employed a quantitative non-experimental approach using socioeconomic data from 31 districts, which underwent preprocessing stages including data cleaning and transformation. Risk labeling was carried out using the 33–66 quantile method, while model validation was conducted using *Leave-One-Out Cross Validation (LOOCV)*. In addition, this study applied *RandomizedSearchCV* to obtain the optimal hyperparameter combination for the *CatBoost* model. The results showed that the *CatBoost* model achieved an accuracy of 97% and a ROC-AUC score of 0.991. Furthermore, the average validation accuracy of  $0.968 \pm 0.177$  indicated that the model performance was relatively stable across validation iterations. The classification results showed that 12 districts were categorized as high risk, 9 districts as medium risk, and 10 districts as low risk. *SHAP* analysis revealed that education level and occupation type were the dominant variables contributing to the classification results. This

*study demonstrates that CatBoost provides strong classification performance in identifying child violence risk based on socioeconomic data and can support decision-making in determining priority areas for child protection interventions.*

**Keywords:** *child violence, risk classification, CatBoost, socio-economic factors, machine learning*

*Submitted : 22 Mei 2026 Revised : 2026 Juli 22 Accepted : 2026 Juli 28*

## 1. PENDAHULUAN

Kekerasan terhadap anak masih menjadi permasalahan sosial yang serius di Indonesia karena berdampak pada perkembangan fisik, psikologis, dan sosial anak. Berdasarkan data Sistem Informasi Online Perlindungan Perempuan dan Anak (SIMFONI-PPA), sebanyak 62,6% dari total korban kekerasan di Indonesia pada tahun 2024 merupakan anak. Jawa Timur termasuk salah satu provinsi dengan jumlah kasus kekerasan anak yang tinggi, termasuk Kabupaten Jember. Data Dinas Pemberdayaan Perempuan, Perlindungan Anak, dan Keluarga Berencana (DP3AKB) Kabupaten Jember menunjukkan terdapat 264 kasus kekerasan anak pada tahun 2024 dan 139 kasus hingga Juni 2025. Kondisi tersebut menunjukkan bahwa kasus kekerasan anak masih menjadi permasalahan yang perlu mendapat perhatian meskipun Kabupaten Jember telah memperoleh predikat Kabupaten Layak Anak (KLA) tingkat Nindya sejak tahun 2022.

Faktor sosial ekonomi rumah tangga diketahui memiliki hubungan dengan tingkat kerentanan anak terhadap kekerasan. Kondisi seperti rendahnya tingkat pendidikan, ketidakstabilan pekerjaan, serta keterbatasan ekonomi rumah tangga dapat meningkatkan risiko terjadinya kekerasan terhadap anak. Oleh karena itu, data sosial ekonomi dapat dimanfaatkan sebagai indikator untuk mengidentifikasi wilayah yang memiliki tingkat kerentanan lebih tinggi sehingga upaya mitigasi, pencegahan, dan penyusunan program perlindungan anak dapat dilakukan secara lebih dini sebelum kasus kekerasan meningkat.

Perkembangan *machine learning* memungkinkan pemanfaatan data sosial ekonomi untuk melakukan klasifikasi tingkat risiko kekerasan anak. Salah satu algoritma yang memiliki performa baik dalam menangani data kategorikal dan numerik adalah *CatBoost*. Algoritma ini mampu menghasilkan prediksi yang stabil serta mengurangi risiko *overfitting* pada data dengan jumlah observasi yang terbatas. Penelitian sebelumnya juga menunjukkan bahwa *CatBoost* memiliki performa klasifikasi yang tinggi pada berbagai permasalahan sosial. Di sisi lain, penelitian mengenai kekerasan anak masih didominasi oleh pendekatan statistik atau clustering untuk memetakan wilayah rawan. Pendekatan tersebut belum menghasilkan model klasifikasi risiko yang memanfaatkan profil sosial ekonomi wilayah sebagai dasar mitigasi.

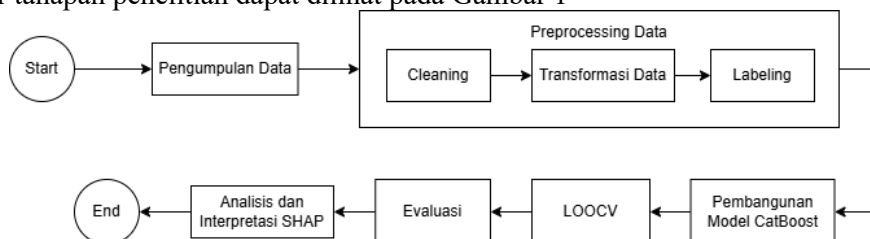
Berdasarkan gap tersebut, penelitian ini menerapkan algoritma *CatBoost* untuk mengklasifikasikan tingkat risiko kekerasan anak berdasarkan profil sosial ekonomi pada tingkat kecamatan di Kabupaten Jember serta menggunakan *SHAP* untuk menginterpretasikan kontribusi setiap variabel terhadap hasil klasifikasi. Pendekatan ini memberikan informasi tidak hanya mengenai tingkat risiko setiap wilayah, tetapi juga faktor sosial ekonomi yang paling berpengaruh sehingga dapat mendukung penyusunan strategi mitigasi risiko kekerasan anak secara lebih tepat sasaran. Penelitian ini menggunakan data sosial ekonomi dari 31 kecamatan di Kabupaten Jember untuk mengklasifikasikan tingkat risiko kekerasan anak ke dalam kategori rendah, sedang, dan tinggi.

## 2. METODOLOGI

Penelitian ini menggunakan pendekatan kuantitatif non-eksperimental berbasis *machine learning* untuk mengklasifikasikan tingkat risiko kekerasan anak di Kabupaten Jember berdasarkan kondisi sosial ekonomi wilayah. Pendekatan ini digunakan untuk menganalisis hubungan antara variabel sosial ekonomi dengan tingkat risiko kekerasan anak melalui proses klasifikasi data. Data penelitian diperoleh dari Dinas Pemberdayaan Perempuan, Perlindungan

Anak, dan Keluarga Berencana (DP3AKB) Kabupaten Jember, Dinas Sosial, serta Dinas Kependudukan dan Pencatatan Sipil (Dispendukcapil) Kabupaten Jember.

Data kekerasan anak digunakan sebagai variabel target yang mencakup kasus kekerasan fisik, psikis, seksual, dan penelantaran. Sementara itu, variabel prediktor yang digunakan meliputi tingkat pendidikan kepala keluarga, jenis pekerjaan kepala keluarga, jumlah anggota rumah tangga, jumlah penerima bantuan sosial Program Keluarga Harapan (PKH), serta jumlah rumah tangga miskin berdasarkan Data Tunggal Sosial Ekonomi Nasional (DTSEN). Penelitian ini menggunakan algoritma *CatBoost* untuk mengklasifikasikan tingkat risiko kekerasan anak ke dalam kategori rendah, sedang, dan tinggi berdasarkan karakteristik sosial ekonomi wilayah. Alur tahapan penelitian dapat dilihat pada Gambar 1



Gambar 1 Tahapan penelitian

## 2.1 Pengumpulan Data

Data penelitian diperoleh dari Dinas Pemberdayaan Perempuan, Perlindungan Anak, dan Keluarga Berencana (DP3AKB), Dinas Sosial, serta Dinas Kependudukan dan Pencatatan Sipil (Dispendukcapil) Kabupaten Jember. Data kasus kekerasan anak tahun 2022-2025 mencakup kekerasan fisik, psikis, seksual, dan penelantaran. Data tahun 2022-2024 tersedia pada tingkat kecamatan, sedangkan data tahun 2025 tersedia pada tingkat desa/kelurahan di 31 kecamatan Kabupaten Jember. Penelitian ini juga menggunakan data sosial ekonomi rumah tangga tahun 2025 yang mencakup 248 desa/kelurahan di Kabupaten Jember. Variabel yang dipilih didasarkan hasil diskusi dengan DP3AKB Kabupaten Jember yang mengacu pada Buku Saku Sistem Perlindungan Anak (Kementerian PPN/Bappenas, 2022) dan pedoman Kementerian PPPA (2012). Variabel penelitian yang digunakan dapat dilihat pada Tabel 1

Tabel 1 Variabel dataset sosial-ekonomi

| Kode | Variabel                                    |
|------|---------------------------------------------|
| X1   | Tingkat Pendidikan Terakhir Kepala Keluarga |
| X2   | Jenis Pekerjaan Utama Kepala Keluarga       |
| X3   | Jumlah Anggota Rumah Tangga (KK)            |
| X4   | Jumlah Penerima Bantuan Sosial (PKH)        |
| X5   | Jumlah Rumah Tangga Miskin (DTSEN)          |

## 2.2 Preprocessing Data

Tahap *preprocessing* dilakukan untuk menyiapkan data sebelum proses pemodelan menggunakan algoritma *CatBoost*. Tahapan *preprocessing* meliputi pembersihan data (*cleaning*), transformasi data, dan *labeling* data.

### 2.2.1 Pembersihan data (*Cleaning*)

Pada tahap ini dilakukan pemeriksaan kualitas data untuk mengidentifikasi data duplikat maupun data yang tidak konsisten. Data duplikat yang ditemukan kemudian dihapus agar tidak mempengaruhi hasil analisis dan proses pemodelan yang dilakukan.

### 2.2.2 Transformasi Data

Transformasi data dilakukan untuk menyesuaikan variabel agar dapat diproses oleh algoritma *CatBoost*. Variabel pendidikan dikelompokkan menjadi tiga kategori, yaitu pendidikan rendah, menengah, dan tinggi berdasarkan jenjang pendidikan formal di Indonesia. Pendidikan rendah mencakup SD/MI dan SMP/MTs, pendidikan menengah mencakup

SMA/SMK/MA, sedangkan pendidikan tinggi mencakup Diploma, Sarjana, dan jenjang pendidikan di atasnya. Pengelompokan dilakukan berdasarkan Undang-undang Nomor 20 Tahun 2003 tentang Sistem Pendidikan Nasional.

Variabel pekerjaan juga dikelompokkan menjadi kategori formal, semi-formal, dan informal berdasarkan karakteristik hubungan kerja dan stabilitas pekerjaan. Selain itu, variabel tingkat kesejahteraan berdasarkan data Data Tunggal Sosial Ekonomi Nasional (DTSEN) dikelompokkan menjadi enam kategori, yaitu desil 1, desil 2, desil 3, desil 4, desil 5, serta desil 6–10. Selain itu, karena data sosial ekonomi tahun 2025 tersedia pada tingkat desa, sedangkan data kekerasan anak tahun 2022-2024 tersedia pada tingkat kecamatan, dilakukan proses agregasi data desa ke tingkat kecamatan menggunakan metode penjumlahan total agar seluruh variabel memiliki unit analisis yang konsisten.

### 2.2.3 Labeling

*Labeling* dilakukan menggunakan pendekatan *quantile-based labeling* dengan kuantil 33% dan 66% untuk mengelompokkan tingkat risiko kekerasan anak menjadi tiga kategori, yaitu rendah, sedang, dan tinggi. Nilai kuantil tersebut dipilih karena membagi data ke dalam tiga kelompok (*tertile*) dengan jumlah observasi yang relatif seimbang sehingga distribusi kelas menjadi lebih proporsional. Pendekatan ini digunakan untuk membentuk batas klasifikasi berdasarkan distribusi jumlah kasus kekerasan anak pada setiap kecamatan.

*Rendah, jika  $P \leq Q_{0.33}$*   
*Sedang, jika  $Q_{0.33} < P \leq Q_{0.66}$*   
*Tinggi, jika  $P > Q_{0.66}$*

- a.  $P$  adalah jumlah total kasus kekerasan terhadap anak di setiap kecamatan
- b.  $Q_{(u)}$  merupakan fungsi kuantil pada persentase  $u$
- c.  $Q_{0.33}$  dan  $Q_{0.66}$  masing-masing adalah kuantil 33% dan 66% yang diperoleh dari distribusi jumlah kasus kekerasan anak

### 2.3 Pembangunan Model CatBoost

Setelah proses *labeling*, tahap selanjutnya adalah membangun model klasifikasi menggunakan algoritma *CatBoost*. Model ini digunakan untuk mengklasifikasikan tingkat risiko kekerasan anak ke dalam tiga kategori, yaitu rendah, sedang, dan tinggi berdasarkan variabel sosial ekonomi yang telah diproses sebelumnya. Variabel sosial ekonomi digunakan sebagai fitur input, sedangkan label tingkat risiko digunakan sebagai variabel target dalam proses pelatihan model. *CatBoost* merupakan algoritma berbasis *gradient boosting* yang mampu menangani data kategorikal dan numerik dengan baik. Algoritma ini juga memiliki kemampuan mengurangi risiko *overfitting* pada data dengan jumlah observasi yang terbatas. Oleh karena itu, *CatBoost* dipilih karena sesuai untuk karakteristik dataset pada penelitian ini.

Proses validasi model dilakukan menggunakan metode *Leave-One-Out Cross Validation (LOOCV)* karena jumlah sampel penelitian relatif terbatas, yaitu sebanyak 31 kecamatan. Pada setiap iterasi, satu kecamatan digunakan sebagai data uji dan 30 kecamatan lainnya digunakan sebagai data latih hingga seluruh data memperoleh kesempatan menjadi data uji. Penentuan parameter *CatBoost* dilakukan menggunakan *RandomizedSearchCV* untuk memperoleh kombinasi *hyperparameter* yang menghasilkan performa klasifikasi terbaik. Berdasarkan hasil tuning diperoleh best accuracy sebesar 0,969 dengan kombinasi parameter optimal yang digunakan pada model sebagaimana ditunjukkan pada Tabel 2.

**Tabel 2 Hasil RandomizedSearchCV**

| Parameter     | Nilai Terbaik |
|---------------|---------------|
| learning_rate | 0.1           |
| l2_leaf_reg   | 3             |
| iterations    | 400           |
| Depth         | 4             |

## 2.4 Validasi Model dengan LOOCV (Leave-One-Out Cross-Validation)

Setelah model *CatBoost* dibangun, tahap selanjutnya adalah melakukan validasi model menggunakan metode *Leave-One-Out Cross Validation (LOOCV)*. Metode ini dipilih karena jumlah sampel penelitian relatif terbatas, yaitu sebanyak 31 kecamatan. Melalui *LOOCV*, setiap sampel dapat dimanfaatkan secara optimal dalam proses pelatihan maupun pengujian model. *LOOCV* merupakan bentuk khusus dari *cross validation* di mana jumlah *fold* sama dengan jumlah sampel pada dataset. Pada setiap iterasi, satu sampel digunakan sebagai data uji, sedangkan sampel lainnya digunakan sebagai data latih. Dengan demikian, proses validasi dilakukan sebanyak 31 iterasi sehingga setiap kecamatan memiliki kesempatan menjadi data uji sebanyak satu kali.

## 2.5 Evaluasi Model

Tahap evaluasi dilakukan untuk menilai performa model *CatBoost* yang telah dioptimasi dan divalidasi. Evaluasi dilakukan dengan menggunakan *confusion matrix* gabungan, metrik akurasi, presisi, recall, F1-score, dan *ROC-AUC* berdasarkan keseluruhan prediksi yang dihasilkan selama proses validasi.

## 2.6 Analisis dan Interpretasi Model SHAP

Analisis dan interpretasi model dilakukan menggunakan metode *SHAP (SHapley Additive exPlanations)* untuk mengetahui kontribusi masing-masing variabel terhadap hasil prediksi model. Metode *SHAP* digunakan karena mampu memberikan interpretasi model secara kuantitatif berdasarkan kontribusi setiap fitur terhadap hasil klasifikasi. Dalam penelitian ini, *SHAP* digunakan untuk mengidentifikasi faktor sosial ekonomi yang memiliki pengaruh terhadap tingkat risiko kekerasan anak di Kabupaten Jember. Nilai *SHAP* positif menunjukkan kontribusi terhadap peningkatan risiko, sedangkan nilai negatif menunjukkan kontribusi terhadap penurunan risiko. Hasil interpretasi *SHAP* digunakan untuk memahami variabel sosial ekonomi yang paling dominan dalam model klasifikasi *CatBoost*.

# 3. HASIL DAN PEMBAHASAN

## 3.1 Hasil Implementasi CatBoost

Penelitian ini menggunakan algoritma *CatBoost* untuk mengklasifikasikan tingkat risiko kekerasan anak berdasarkan kondisi sosial ekonomi pada 31 kecamatan di Kabupaten Jember. Sebelum proses pemodelan, dilakukan tahap *preprocessing* berupa *cleaning*, transformasi data, agregasi data, dan *labeling* menggunakan kuantil 33% dan 66% untuk membagi tingkat risiko menjadi kategori rendah, sedang, dan tinggi. Hasil pelabelan menggunakan metode kuantil 33-66 dapat dilihat pada tabel 3.

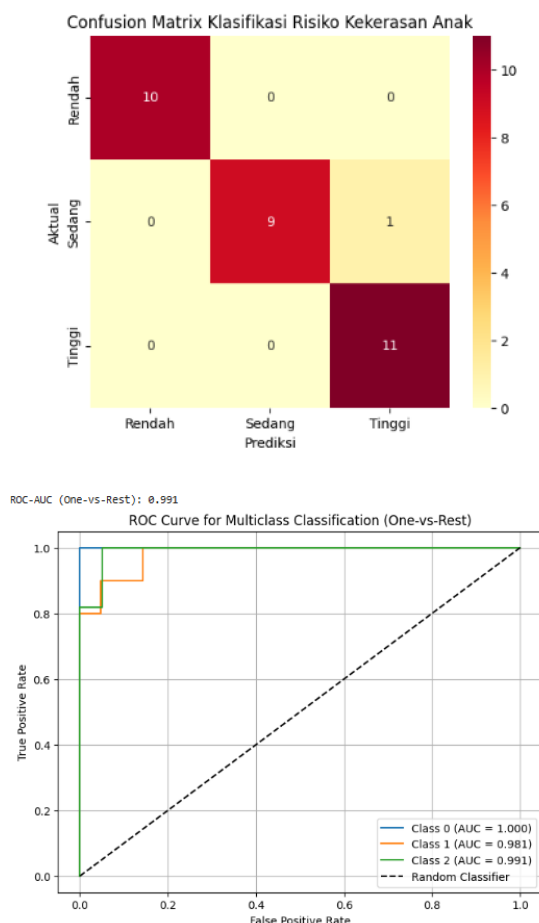
**Tabel 3 Pembagian kelas risiko menggunakan kuantil 33-66**

| Kelas Risiko  | Kode Kelas | Jumlah Kecamatan |
|---------------|------------|------------------|
| Risiko Rendah | 0          | 10               |
| Risiko Sedang | 1          | 10               |
| Risiko Tinggi | 2          | 11               |

Hasil klasifikasi menunjukkan bahwa model *CatBoost* mampu menghasilkan akurasi sebesar 0,968 dengan nilai *ROC-AUC* sebesar 0,991. Selain itu, hasil validasi menggunakan *LOOCV* menunjukkan rata-rata *validation accuracy* sebesar  $0,968 \pm 0,177$  sehingga menunjukkan performa model relatif stabil pada setiap iterasi validasi meskipun jumlah sampel penelitian terbatas. Berdasarkan hasil klasifikasi, terdapat 12 kecamatan kategori risiko tinggi, 9 kecamatan kategori risiko sedang, dan 10 kecamatan kategori risiko rendah. Nilai *precision*, *recall*, dan *F1-score* pada masing-masing kelas juga menunjukkan bahwa model memiliki kemampuan yang baik dalam membedakan tingkat risiko kekerasan anak antar wilayah. Detail hasil evaluasi model dapat dilihat pada Tabel 4 dan Gambar 2.

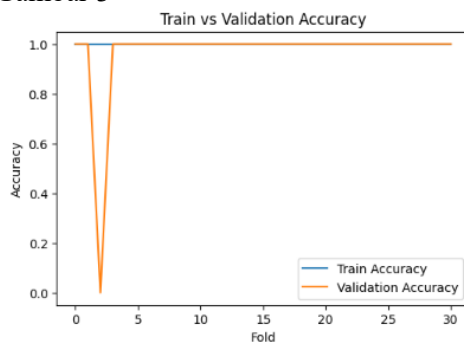
**Tabel 4 Evaluasi kinerja model per kelas**

| Kategori Risiko | <i>Precision</i> | <i>Recall</i> | <i>F1-Score</i> |
|-----------------|------------------|---------------|-----------------|
| Risiko Rendah   | 1.00             | 1.00          | 1.00            |
| Risiko Sedang   | 1.00             | 0.90          | 0.95            |
| Risiko Tinggi   | 0.92             | 1.00          | 0.96            |



**Gambar 2 Evaluasi Model**

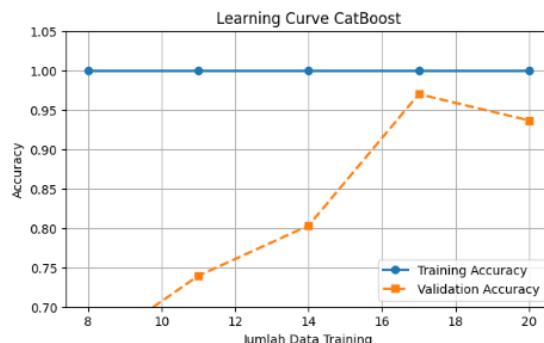
Analisis performa *training* dan *validation* pada setiap iterasi *LOOCV* menunjukkan bahwa model *CatBoost* tidak mengalami indikasi *overfitting* yang signifikan. Hal ini ditunjukkan oleh rata-rata *training accuracy* sebesar 1,00 dan *validation accuracy* sebesar 0,968 yang relatif konsisten. Meskipun terdapat beberapa fold dengan kesalahan klasifikasi pada data validasi, performa model secara umum tetap stabil dalam mengklasifikasikan tingkat risiko kekerasan anak. Visualisasi perbandingan performa *training* dan *validation* pada setiap fold *LOOCV* dapat dilihat pada Gambar 3



**Gambar 3 Train vs Validation akurasi**



Selain itu, berdasarkan *learning curve* pada Gambar 4, *validation accuracy* cenderung meningkat seiring bertambahnya jumlah data training, sedangkan *training accuracy* tetap berada pada nilai yang tinggi dan stabil. Kondisi tersebut menunjukkan bahwa model memiliki kemampuan generalisasi yang cukup baik serta tidak menunjukkan indikasi *overfitting* yang ekstrem meskipun jumlah sampel penelitian relatif terbatas.

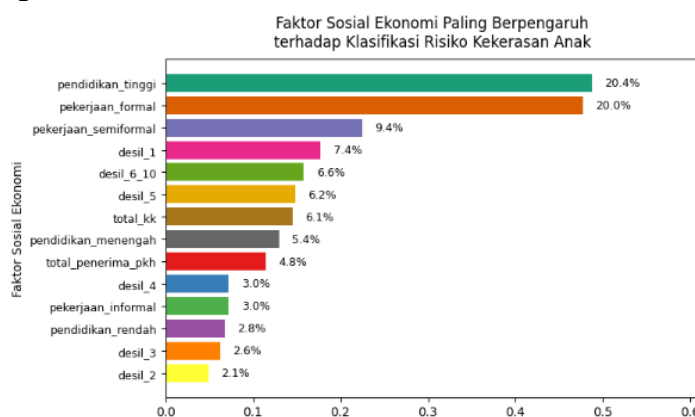


Gambar 4 Learning curve

Model *CatBoost* mampu mengklasifikasikan kelas risiko rendah dan risiko tinggi dengan sangat baik tanpa kesalahan prediksi pada kedua kelas tersebut. Kesalahan klasifikasi hanya terjadi pada satu data pada kelas risiko sedang yang diprediksi sebagai risiko tinggi. Hal ini menunjukkan bahwa karakteristik sosial ekonomi pada data tersebut lebih mendekati karakteristik kelas risiko tinggi dibandingkan kelas risiko rendah, sehingga model cenderung mengelompokkannya ke dalam kategori risiko tinggi. Kategorisasi tingkat risiko dalam penelitian ini dibentuk berdasarkan distribusi statistik menggunakan pendekatan kuantil dan bukan merupakan diagnosis risiko aktual pada setiap wilayah. Oleh karena itu, terdapat kemungkinan beberapa kecamatan dengan jumlah kasus yang relatif berdekatan masuk ke dalam kategori risiko yang berbeda akibat batas kuantil yang digunakan.

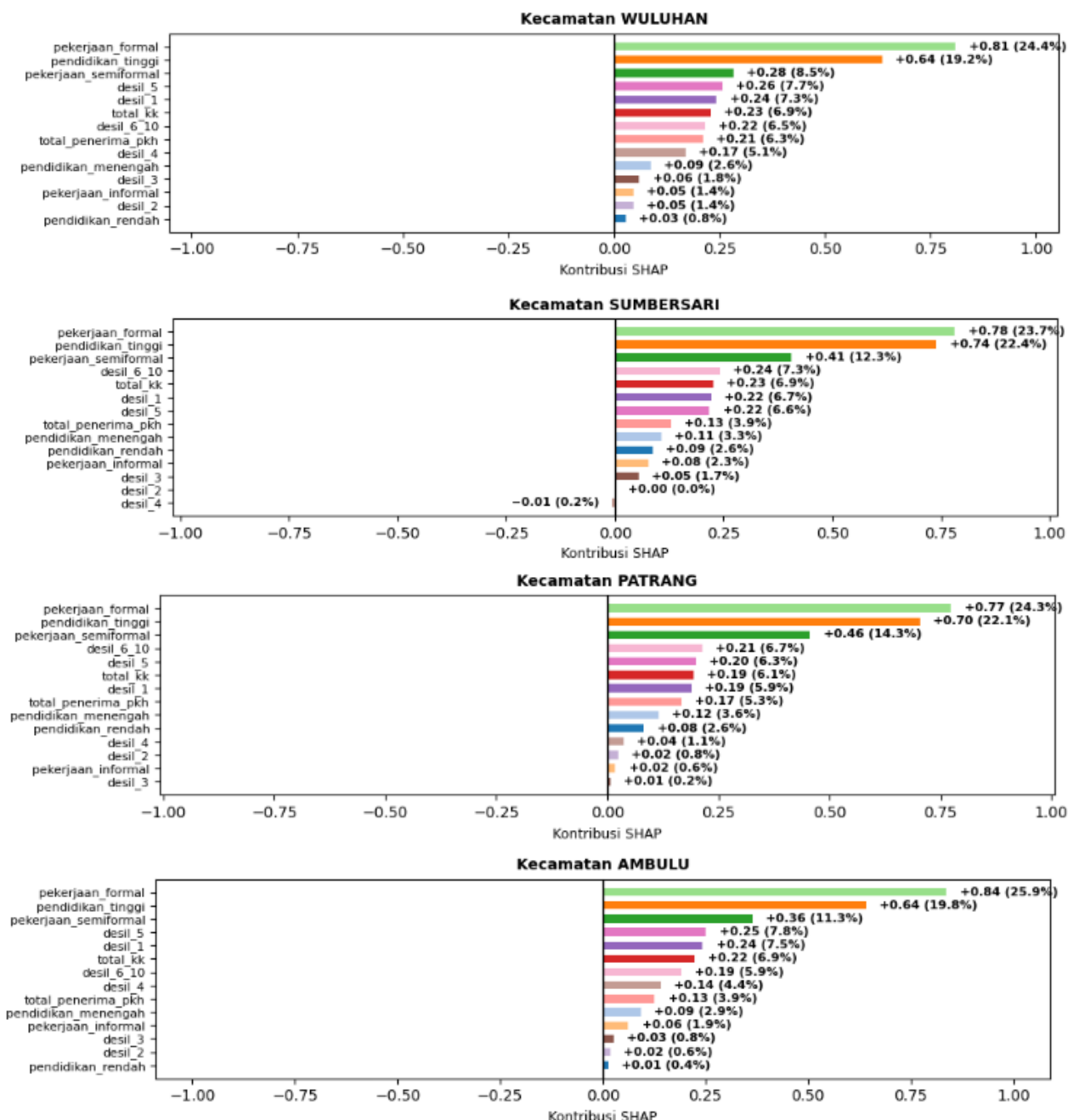
### 3.2 Analisis Faktor Sosial Ekonomi Menggunakan SHAP

Hasil analisis menunjukkan bahwa variabel pendidikan tinggi memiliki kontribusi terbesar terhadap prediksi model sebesar 20,4%, diikuti pekerjaan formal sebesar 20,0%, dan pekerjaan semi-formal sebesar 9,4%. Hasil tersebut menunjukkan bahwa struktur pendidikan dan pekerjaan pada suatu wilayah menjadi faktor dominan dalam klasifikasi risiko kekerasan anak. Selain itu, variabel pendidikan menengah juga memberikan kontribusi sebesar 5,4% terhadap hasil prediksi model. Sementara itu, variabel yang merepresentasikan kondisi kemiskinan seperti desil 1, desil 2, dan desil 3 memiliki kontribusi yang relatif lebih rendah dibandingkan variabel struktural lainnya. Hasil tersebut menunjukkan bahwa risiko kekerasan anak pada tingkat wilayah tidak hanya dipengaruhi oleh faktor kemiskinan, tetapi juga dipengaruhi oleh kombinasi faktor sosial ekonomi yang lebih kompleks. Detail hasil kontribusi dapat dilihat pada gambar 5



Gambar 5 Faktor sosial ekonomi yang berpengaruh

Secara keseluruhan, hasil analisis *SHAP* menunjukkan bahwa model lebih sensitif terhadap faktor struktural dan komposisi sosial ekonomi wilayah dibandingkan indikator kemiskinan semata. Wilayah dengan kategori risiko tinggi berdasarkan hasil prediksi model meliputi Kecamatan Wuluhan, Summersari, Patrang, Ambulu, Jenggawah, Ajung, Kaliwates, Silo, Arjasa, Kencong, Balung, dan Kalisat. Contoh detail hasil analisis tiap kecamatan dengan risiko tinggi dapat dilihat pada gambar 6

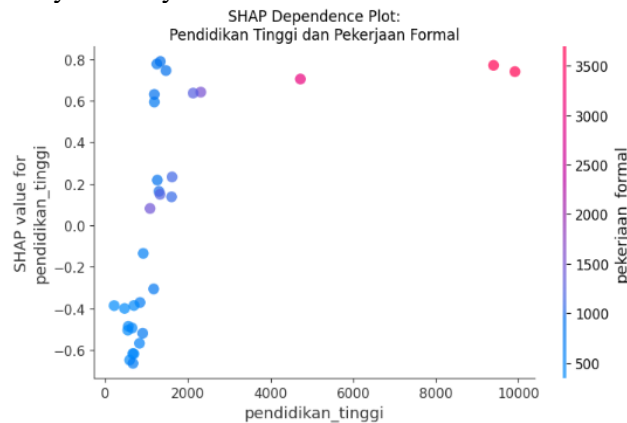


**Gambar 6 Contoh faktor sosial ekonomi berpengaruh pada kecamatan risiko tinggi**

Berdasarkan *SHAP dependence plot* pada Gambar 7 di bawah, variabel pendidikan tinggi menunjukkan pola kontribusi yang berbeda terhadap hasil klasifikasi risiko kekerasan anak. Nilai *SHAP* positif menunjukkan kontribusi yang lebih besar terhadap prediksi risiko tinggi, sedangkan nilai negatif menunjukkan kontribusi terhadap risiko yang lebih rendah. Selain itu, variasi warna pada plot menunjukkan adanya interaksi antara pendidikan tinggi dan pekerjaan formal dalam proses prediksi model. Kecamatan dengan nilai pendidikan tinggi dan



pekerjaan formal yang lebih besar cenderung memiliki kontribusi *SHAP* positif yang lebih tinggi dibandingkan wilayah lainnya.



**Gambar 7 SHAP dependence plot**

Hasil interpretasi *SHAP* tidak menunjukkan hubungan sebab-akibat secara langsung, melainkan menggambarkan pola kontribusi variabel sosial ekonomi terhadap hasil klasifikasi model. Oleh karena itu, variabel dengan kontribusi tinggi tidak dapat diartikan sebagai penyebab utama kekerasan anak, tetapi sebagai indikator statistik yang membantu model dalam mengidentifikasi wilayah dengan tingkat risiko yang lebih tinggi dibandingkan wilayah lainnya. Kondisi tersebut juga dapat dipengaruhi oleh faktor lain seperti tingginya akses pelaporan kasus, tingkat kesadaran masyarakat, kepadatan penduduk, serta kemudahan akses layanan perlindungan anak pada wilayah dengan karakteristik sosial ekonomi yang lebih maju.

Pemanfaatan *SHAP* pada penelitian ini konsisten dengan penelitian Subianto *et al.* (2023) yang menggunakan *SHAP* untuk mengidentifikasi variabel penting pada model *CatBoost*. Perbedaannya, penelitian ini menginterpretasikan faktor-faktor sosial ekonomi yang berkontribusi terhadap klasifikasi risiko kekerasan anak di Kabupaten Jember.

Penelitian ini masih memiliki beberapa keterbatasan, di antaranya jumlah sampel yang relatif kecil yaitu 31 kecamatan, penggunaan pelabelan berbasis kuantil yang bergantung pada distribusi data, serta cakupan data yang masih terbatas pada satu kabupaten sehingga generalisasi model ke wilayah lain masih terbatas. Selain itu, penggunaan metode *Leave-One-Out Cross Validation (LOOCV)* pada dataset berukuran kecil berpotensi menghasilkan variasi performa model yang cukup tinggi. Penelitian selanjutnya dapat mengembangkan pendekatan analisis spasial, pemetaan geospasial risiko, maupun temporal *forecasting* untuk menganalisis pola kekerasan anak secara lebih komprehensif. Selain itu, penggunaan metode validasi lain seperti *repeated stratified k-fold cross validation* dan *bootstrap validation*, dapat dilakukan untuk meningkatkan stabilitas performa serta interpretabilitas model.

#### 4. KESIMPULAN

Penelitian ini berhasil mengklasifikasikan tingkat risiko kekerasan anak di Kabupaten Jember menggunakan algoritma *CatBoost* berdasarkan faktor sosial ekonomi. Hasil klasifikasi menunjukkan 12 kecamatan kategori risiko tinggi, 9 kecamatan kategori risiko sedang, dan 10 kecamatan kategori risiko rendah. Model *CatBoost* mampu menghasilkan akurasi sebesar 0,968 dengan nilai *ROC-AUC* 0,991. Selain itu penggunaan *LOOCV validation accuracy* sebesar  $0,968 \pm 0,177$  yang menandakan performa lebih stabil meskipun jumlah sampel terbatas. Hasil interpretasi menggunakan *SHAP*, variabel pendidikan tinggi, pekerjaan formal, dan pekerjaan semi-formal menjadi faktor sosial ekonomi yang memiliki kontribusi dominan terhadap hasil prediksi model. Namun, faktor tersebut tidak dapat diartikan sebagai penyebab langsung terjadinya kekerasan anak, melainkan sebagai indikator kerentanan wilayah yang membantu model dalam mengidentifikasi pola risiko kekerasan anak berdasarkan karakteristik sosial ekonomi kecamatan.

## 5. DAFTAR RUJUKAN

- [1] Alfina, O., & Safii, M. (2025). Klasifikasi pola konsumsi energi rumah tangga menggunakan algoritma machine learning untuk mendukung implementasi smart city. *METHOMIKA: Jurnal Manajemen Informatika & Komputerisasi Akuntansi*, 9(2), 300-306. <https://doi.org/10.46880/jmika.Vol9No2.pp300-306>
- [2] Amalia, D. O., Sabarinah, S., Siregar, K. N., & Hadi, E. N. (2025). Childhood violence exposure and its contributing factors in Indonesia: A secondary data analysis of the National Survey on Children and Adolescents' Life Experience. *BMJ Open*, 15(1), e090618. <https://doi.org/10.1136/bmjopen-2024-090618>  
<https://doi.org/10.1136/bmjopen-2024-090618>  
PMid:39832971 PMCID:PMC11751955
- [3] Chadaga, K., Prabhu, S., Sampathila, N., Chadaga, R., Bairy, M., & Swathi, K. S. (2024). An explainable framework to predict child sexual abuse awareness in people using supervised machine learning models. *Journal of Technology in Behavioral Science*, 9, 346-362. <https://doi.org/10.1007/s41347-023-00343-0>  
<https://doi.org/10.1007/s41347-023-00343-0>
- [4] Han, J., Kamber, M., & Pei, J. (2023). *Data Mining: Concepts and Techniques* (4th ed.). Morgan Kaufmann.
- [5] Hancock, J. T., & Khoshgoftaar, T. M. (2020). CatBoost for big data: an interdisciplinary review. *Journal of big data*, 7(1), 94. <https://doi.org/10.1186/s40537-020-00369-8>  
PMid:33169094
- [6] Kementerian Pemberdayaan Perempuan dan Perlindungan Anak Republik Indonesia. (2012). Peraturan Menteri Negara Pemberdayaan Perempuan dan Perlindungan Anak Republik Indonesia Nomor 8 Tahun 2012 tentang Panduan Penguatan Kelompok Dasawisma untuk Pencegahan dan Penanganan Dini Tindak Kekerasan terhadap Anak.
- [7] Kementerian Perencanaan Pembangunan Nasional/Badan Perencanaan Pembangunan Nasional. (2022). *Buku saku sistem perlindungan anak*.
- [8] Khairunnisa, S., & Herbawani, C. K. (2023). Determinan Kekerasan Terhadap Anak oleh Orang Tua di Indonesia: Studi Literatur. *Jurnal Kesehatan dan Pendidikan*, 11(2). <https://doi.org/10.32922/jkp.v11i2.817>  
<https://doi.org/10.32922/jkp.v11i2.817>
- [9] Mawardi, A. B., Pradini, R. S., & Haris, M. S. (2024). Komparasi algoritma boosting untuk prediksi gangguan tidur. *JITET (Jurnal Informatika dan Teknik Elektro Terapan)*, 13(3), 1377-1386. <https://doi.org/10.23960/jitet.v13i3.7281>  
<https://doi.org/10.23960/jitet.v13i3.7281>
- [10] Naziah, H. R., Sunarti, E., & Riany, Y. E. (2023). Pengaruh manajemen stres dan status pekerjaan ibu terhadap pengasuhan anak usia sekolah saat pandemi COVID-19. *Jurnal Ilmu Keluarga dan Konsumen*, 16(1), 23-36.

<https://doi.org/10.24156/jikk.2023.16.1.23>

<https://doi.org/10.24156/jikk.2023.16.1.23>

[11] Redivo, E., Viroli, C., & Farcomeni, A. (2023). Quantile-distribution functions and their use for classification, with application to naïve Bayes classifiers. *Statistics and Computing*, 33, 55. <https://link.springer.com/article/10.1007/s11222-023-10224-4>

<https://doi.org/10.1007/s11222-023-10224-4>

[12] Rohadatul'aisy, N., Hardiawan, D., & Sihaloho, E. D. (2025). The impact of household poverty on child abuse in Indonesia. *Signifikan: Jurnal Ilmu Ekonomi*, 14(1), 131-148.

<https://doi.org/10.15408/sjie.v14i1.45142>

<https://doi.org/10.15408/sjie.v14i1.45142>

[13] Sihotang, A. S., Bektı, R. D., & Jatipaningrum, M. T. (2024). Pengelompokan Kabupaten/Kota di Provinsi Sumatera Utara berdasarkan indikator jenis kekerasan terhadap anak menggunakan metode K-Medoids dan Fuzzy C-Means. *Jurnal Statistika Industri dan Komputasi*, 9(2), 11-19. <https://doi.org/10.34151/statistika.v9i2.4850>

<https://doi.org/10.34151/statistika.v9i2.4850>