JAISEN: Journal of Advanced Information Systems and Engineering
Vol. 1, 1 (August, 2025), pp 18-27
ISSN: - EISSN: - , DOI: -

CLASSIFICATION OF GENERAL CRIMINAL CASE TYPES AT THE BENGKULU HIGH PROSECUTOR'S OFFICE USING MACHINE LEARNING ALGORITHMS

Budi Kurniawan[1*], Susandri Susandri[2], Feldiansyah[3]

[1] Magister Ilmu Komputer, Universitas Lancang Kuning; ghifarkun@gmail.com
[2] Magister Ilmu Komputer, Universitas Lancang Kuning; susandri@unilak.ac.id
[3] Magister Ilmu Komputer, Universitas Lancang Kuning ; feldiansyah@unilak.ac.id

ARTICLE INFO

Keywords:

Case Classification;
General Criminal Offenses;
Machine Learning Algorithms;
LightGBM;
Bengkulu Prosecutor's Office.

Article history:

Received 2025-08-06
Revised 2025-08-09
Accepted 2025-08-13

ABSTRACT

The handling of general criminal cases at the Bengkulu High Prosecutor's Office requires an efficient system to classify case types in order to enhance the legal process's effectiveness. One promising approach to optimize this classification process is the application of machine learning algorithms. This study aims to develop a classification model capable of identifying the types of general criminal cases based on the available case data. The algorithms employed in this research include LightGBM and Random Forest, which were evaluated to determine the highest classification accuracy. The dataset consists of case descriptions received by the Bengkulu High Prosecutor's Office during a specific period, which underwent preprocessing steps such as text cleaning, tokenization, and feature extraction to generate relevant features. The findings reveal that LightGBM outperforms Random Forest in classification accuracy. The resulting model can reliably predict the category of criminal cases, serving as an effective tool to assist the Prosecutor's Office in automating case grouping. The implementation of this model is expected to improve operational efficiency and support a more transparent and structured case management system.

Corresponding Author:
Budi Kurniawan
Magister Ilmu Komputer, Universitas Lancang Kuning; ghifarkun@gmail.com

1. INTRODUCTION

The rapid advancement of information technology has significantly influenced numerous sectors, including the law enforcement system in Indonesia. One critical institution within this system is the Bengkulu High Prosecutor's Office, which plays a vital role in managing general criminal cases. This institution currently faces considerable challenges in efficiently handling and classifying an increasing number of incoming cases, a situation that demands both accuracy and speed in case management. Traditionally, the classification process has relied heavily on manual methods, which are labor-intensive and susceptible to human error. As the volume of cases continues to grow over time, the manual approach

becomes increasingly impractical, potentially resulting in delays and misclassification that can hinder the effectiveness and quality of legal services provided. In this context, the integration of machine learning algorithms offers a promising solution to automate and enhance the classification process. Machine learning, a branch of artificial intelligence, enables systems to learn from historical data by identifying patterns and making predictions without requiring explicit programming for each task. In legal applications, machine learning can analyze textual data such as case descriptions to categorize criminal offenses based on relevant features such as offense type, severity, and other pertinent criteria. This capability is especially valuable for the Bengkulu High Prosecutor's Office, where handling a large caseload necessitates a system that can rapidly and accurately classify cases to support timely legal proceedings.

To develop such a system, this study proposes the design and evaluation of automated classification models for general criminal cases using machine learning algorithms. Specifically, the study focuses on two widely used algorithms: Light Gradient Boosting Machine (LightGBM) and Random Forest. Both algorithms have demonstrated strong performance in classification tasks involving complex and unstructured data. LightGBM is a gradient boosting framework that builds decision trees in a highly efficient manner, enabling fast training and prediction even with large datasets. Random Forest, on the other hand, constructs an ensemble of decision trees to improve prediction accuracy and reduce the risk of overfitting. By comparing these algorithms, the study aims to identify the most effective model that balances accuracy and computational efficiency for case classification at the Bengkulu High Prosecutor's Office.

The foundation of the machine learning models lies in the quality and preparation of the dataset. The dataset utilized in this research consists of case descriptions from the Bengkulu High Prosecutor's Office, covering a specified period. These raw textual data require thorough preprocessing before they can be used effectively for training machine learning algorithms. Preprocessing involves several key steps, including text cleaning to remove irrelevant symbols, punctuation, and common stop words that do not contribute meaningful information. Subsequently, tokenization is applied to break down the cleaned text into smaller units, such as words or phrases, which serve as the fundamental features for analysis. Feature extraction techniques, such as Term Frequency-Inverse Document Frequency (TF-IDF) or word embedding methods, are then employed to convert the textual data into numerical representations that the algorithms can process. These preprocessing steps are crucial for enhancing the ability of the models to capture important patterns and distinctions among different case types.

Once the data is prepared, the LightGBM and Random Forest algorithms are trained and validated using appropriate evaluation metrics. The study employs accuracy, precision, recall, and F1-score to measure the classification performance comprehensively. Additionally, cross-validation techniques are used to assess the models' ability to generalize well to unseen data, ensuring robustness and reliability in real-world applications. Preliminary results indicate that LightGBM often achieves superior accuracy compared to Random Forest while maintaining efficient computational requirements, making it a strong candidate for deployment in case classification tasks within the prosecutor's office.

The implementation of an automated classification system based on machine learning is expected to yield significant benefits for the Bengkulu High Prosecutor's Office. By reducing the reliance on manual classification, the system can dramatically decrease the time needed to process new cases, thereby expediting case handling workflows and enabling prosecutors to focus more on substantive legal analysis rather than administrative tasks. Furthermore, automated classification reduces the potential for human error, improving the consistency and correctness of case categorization. This enhanced accuracy supports better case assignment and prioritization, which are essential for the effective functioning of the justice system. In addition, the system's ability to process large volumes of data efficiently makes it scalable, capable of adapting to future increases in case loads without requiring proportional increases in human resources.

Beyond operational efficiency, the adoption of machine learning for case classification contributes to greater transparency and accountability in the legal process. A standardized and systematic classification mechanism facilitates clearer tracking and reporting of case statuses, which can improve

public trust in the judicial system. Moreover, the success of this initiative at the Bengkulu High Prosecutor's Office could serve as a model for other prosecutorial institutions in Indonesia, encouraging wider adoption of technology-driven solutions within the country's law enforcement agencies. This aligns with the broader national agenda of integrating advanced information technologies to modernize public sector services and enhance governance.

In conclusion, the challenges posed by increasing caseloads and the need for precise classification in the Bengkulu High Prosecutor's Office can be effectively addressed through the application of machine learning algorithms. This study's development and evaluation of automated classification models, particularly using LightGBM and Random Forest, highlight the potential of these technologies to improve both the speed and accuracy of case handling. The anticipated outcomes include streamlined case management processes, reduced administrative burden, and strengthened judicial transparency. As Indonesia continues to embrace digital transformation within its public institutions, leveraging machine learning in prosecutorial workflows represents a significant step toward a more responsive, efficient, and just legal system.

2. METHODS

This study aims to develop a classification model for types of general criminal cases by employing machine learning algorithms. The research methodology comprises several stages, including:

2.1 Data Collection

The data utilized in this study consist of a dataset of general criminal cases received by the Bengkulu High Prosecutor's Office. The dataset includes case descriptions, types of criminal offenses, and other relevant information. Data were collected from official documents or digital databases with authorization from the appropriate authorities.

a) Data Volume: The dataset comprises 4,324 cases encompassing various types of general criminal offenses.
b) Data Source: Electronic or manual archival documents that have been adjusted in accordance with privacy policies.

No	No, tgl SPDP	Tanggal SPDP Diterima	Tersangka/Terdakwa	Penyidik	Jenis Tindak Pidana	Pasal disangkakan
1	SPDP/ 117 / XII / RES. 1.6. / 2023 / RESKRIM	12/29/2023	ASMAWI Bin SULAIMAN	POLRES KAUR	Pengiayaan	Pasal 351 UU Nomor 1 Tahun 1946 tentang KUHP
2	SPDP/03/XII/2023/Reskrim	12/29/2023	MAMAN SYAH Als MAMAN Bin SUPAR	POLSEK KOTA PADANG	Pencurian	Pasal 363 UU Nomor 1 Tahun 1946 tentang KUHP
3	SPDP/16/XII/2023/RESKRIM/SEK-SK	12/29/2023	DEDI Bin KUT	POLSEK SINDANG KELINGI	Pengiayaan	Pasal 351 UU Nomor 1 Tahun 1946 tentang KUHP
4	SPDP/43.B/XII/RES.1.24./2023/Ditreskrimum	12/28/2023	AHMAD ROHNI Als ROHNI Bin PAIMIN	POLDA BENGKULU	KDRT	Pasal 44 ayat (1) dan ayat (4) UU Nomor 23 Tahun 2004 tentang Penghapusan Kekerasan Dalam Rumah Tan
5	SPDP/43.B/XII/RES.1.24./2023/Ditreskrimum	12/28/2023	AHMAD ROHNI Als ROHNI Bin PAIMIN	POLDA BENGKULU	KDRT	Pasal 44 ayat (1) dan ayat (4) UU Nomor 23 Tahun 2004 tentang Penghapusan Kekerasan Dalam Rumah Tan
6	SPDP/43.B/XII/RES.1.24./2023/Ditreskrimum	12/28/2023	AHMAD ROHNI Als ROHNI Bin PAIMIN	POLDA BENGKULU	KDRT	Pasal 44 ayat (1) dan ayat (4) UU Nomor 23 Tahun 2004 tentang Penghapusan Kekerasan Dalam Rumah Tan
7	SPDP/43.B/XII/RES.1.24./2023/Ditreskrimum	12/28/2023	AHMAD ROHNI Als ROHNI Bin PAIMIN	POLDA BENGKULU	KDRT	Pasal 44 ayat (1) dan ayat (4) UU Nomor 23 Tahun 2004 tentang Penghapusan Kekerasan Dalam Rumah Tan
8	SPDP/43.B/XII/RES.1.24./2023/Ditreskrimum	12/28/2023	AHMAD ROHNI Als ROHNI Bin PAIMIN	POLDA BENGKULU	KDRT	Pasal 44 ayat (1) dan ayat (4) UU Nomor 23 Tahun 2004 tentang Penghapusan Kekerasan Dalam Rumah Tan
9	SPDP/43.B/XII/RES.1.24./2023/Ditreskrimum	12/28/2023	AHMAD ROHNI Als ROHNI Bin PAIMIN	POLDA BENGKULU	KDRT	Pasal 44 ayat (1) dan ayat (4) UU Nomor 23 Tahun 2004 tentang Penghapusan Kekerasan Dalam Rumah Tan
10	SPDP/43.B/XII/RES.1.24./2023/Ditreskrimum	12/28/2023	AHMAD ROHNI Als ROHNI Bin PAIMIN	POLDA BENGKULU	KDRT	Pasal 44 ayat (1) dan ayat (4) UU Nomor 23 Tahun 2004 tentang Penghapusan Kekerasan Dalam Rumah Tan
11	SPDP/43.B/XII/RES.1.24./2023/Ditreskrimum	12/28/2023	AHMAD ROHNI Als ROHNI Bin PAIMIN	POLDA BENGKULU	KDRT	Pasal 44 ayat (1) dan ayat (4) UU Nomor 23 Tahun 2004 tentang Penghapusan Kekerasan Dalam Rumah Tan
12	SPDP/135/XII/RES.1.8. /2023/Reskrim	12/28/2023	CENDRI GUSTIRANDA Als CEN Bin SAIRIN	POLRES REJANG LEBONG	Pencurian	Pasal 365 jo pasal 363 UU Nomor 1 Tahun 1946 tentang KUHP
13	SPDP/91/XII/2023/Reskrim.1.24.	12/28/2023	ASTARI Als TARI Bin DAHLAN Alm	POLRES SELUMA	Persetubuhan	Pasal 76D UU 35\/2014 UU Nomor 35 Tahun 2014 tentang Perubahan UU Nomor 23 Tahun 2002 tentang Per
14	SPDP/92/XII/RES.1.24/2023/RESKRIM	12/27/2023	AFDA PRATAMA Bin AFRIZAL	POLRES MUKO MUKO	Persetubuhan	Pasal 81 ayat (2) UU Nomor 17 Tahun 2016 tentang perubahan kedua UU Nomor 23 Tahun 2002 Tentang Per
15	SPDP/22/XII/RES.1.8./2023/RESKRIM	12/27/2023	Adi Rusli Bin Zairani dan Riki Saputra Bin Englan	POLSEK TALANG EMPAT	Pencurian	Pasal 363 KUHP
16	SPDP/132/XII/RES.1.24./2023/Reskrim	12/27/2023	LIDIK	POLRES REJANG LEBONG	Persetubuhan	["Pasal 76D UU 35\/2014 UU Nomor 35 Tahun 2014 tentang Perubahan UU Nomor 23 Tahun 2002 tentang P
17	SPDP/13.B/XII/RES.1.24./2023/Ditreskrimum	12/27/2023	IRA OKTAVIA Als IRA Binti DARLAN	POLDA BENGKULU	Persetubuhan	Pasal 81 ayat (2) dan ayat (3) atau Pasal 82 ayat (1) dan ayat (2) Jo Pasal 76D UU Nomor 17 Tahun 2016 ter
18	SPDP/32/XII/2023/Reskrim	12/27/2023	Dika Putra Davol Bin Suparno	POLSEK PONDOK KELAPA	pencurian	Pasal 363 Ayat (1) Ke 4 dan Ke 5 KUHPidana
19	SPDP/116/XII/RES.1.24/2023/RESKRIM	12/27/2023	WIDYA WULANSARI SPd Binti YUSIRDI	POLRES KAUR	Perzinaan	["Pasal 284 UU Nomor 1 Tahun 1946 tentang KUHP"]
20	SPDP/46/XII/RES.1.8./2023/Satreskrim	12/27/2023	kosong	POLRES BENGKULU TENGAH	Pencurian	Pasal 363 KUHP
21	SPDP/93/XII/RES.1.11/2023/Reskrim	12/27/2023	AGUNG PRASTIYONO ALS AGUNG BIN JUNAIDI HASAN	POLRES MUKO MUKO	Penipuan	Pasal 378 KUH Pidana
22	SPDP/19/XII/2023/Reskrim	12/27/2023	Nova Als Nova Binti Suharti Jaya	POLSEK UJAN MAS	Jaminan Fidusia	Pasal 36 Jo Pasal 23 ayat (2) Undang-Undang Nomor 42 Tahun 1999 Tentang Jaminan Fidusia
23	SPDP/133/XII/RES.1.6. /2023/Reskrim	12/27/2023	YONANDA Als BOJO Bin DEFI ZULFIAN	POLRES REJANG LEBONG	Pengeroyokan	["Pasal 170 KUHP UU Nomor 1 Tahun 1946 tentang KUHP"]
24	SPDP/115/XII/RES.1.24/2023/RESKRIM	12/27/2023	ARIP PENDORES Bin HERMAN SUARDI	POLRES KAUR	Perzinaan	["Pasal 284 UU Nomor 1 Tahun 1946 tentang KUHP"]
25	SPDP/10/XII/2023/Reskrim	12/27/2023	Ujang Bustami Bin (Alm) Matimin	POLSEK TABA PENANJUNG	Penganiayaan Berat	Pasal 354 jo pasal 351 KUHP
26	SPDP/18/XII/2023/Reskrim	12/26/2023	Nolis Nopitasari Als Nolis Binti Awan Sri (Alm)	POLSEK UJAN MAS	Jaminan Fidusia	Pasal 36 Jo pasal 23 ayat (2) Undang-Undang Nomor 42 Tahun 1999 Tentang Jaminan Fidusia
27	SPDP/89/XII/RES.1.6./2023/Satreskrim	12/25/2023	ERPAN	POLRES KEPAHIANG	Kekerasan terhadap anak	Pasal 80 ayat (1) Jo Pasal 76C Undang-Undang Republik Indonesia No.35 Tahun 2014 perubahan atas Unda
28	B/0022/XII/KA/PBO1/SPDP/2023/BNNP Bengkulu	12/22/2023	Denny Alias Denny Bin M Inu Yusu (Alm)	BNNP Bengkulu	Narkotika	Peredaran gelap narkotika golongan I dalam bentuk bukan tanaman jenis shabu sebagaimana dimaksud da
29	B/0022/XII/KA/PBO1/SPDP/2023/BNNP Bengkulu	12/22/2023	Denny Alias Denny Bin M Inu Yusu (Alm)	BNNP Bengkulu	Narkotika	Peredaran gelap narkotika golongan I dalam bentuk bukan tanaman jenis shabu sebagaimana dimaksud da
30	B/0022/XII/KA/PBO1/SPDP/2023/BNNP Bengkulu	12/22/2023	Denny Alias Denny Bin M Inu Yusu (Alm)	BNNP Bengkulu	Narkotika	Peredaran gelap narkotika golongan I dalam bentuk bukan tanaman jenis shabu sebagaimana dimaksud da
31	B/0022/XII/KA/PBO1/SPDP/2023/BNNP Bengkulu	12/22/2023	Denny Alias Denny Bin M Inu Yusu (Alm)	BNNP Bengkulu	Narkotika	Peredaran gelap narkotika golongan I dalam bentuk bukan tanaman jenis shabu sebagaimana dimaksud da
32	B/0022/XII/KA/PBO1/SPDP/2023/BNNP Bengkulu	12/22/2023	Denny Alias Denny Bin M Inu Yusu (Alm)	BNNP Bengkulu	Narkotika	Peredaran gelap narkotika golongan I dalam bentuk bukan tanaman jenis shabu sebagaimana dimaksud da
33	B/0022/XII/KA/PBO1/SPDP/2023/BNNP Bengkulu	12/22/2023	Denny Alias Denny Bin M Inu Yusu (Alm)	BNNP Bengkulu	Narkotika	Peredaran gelap narkotika golongan I dalam bentuk bukan tanaman jenis shabu sebagaimana dimaksud da
34	B/0022/XII/KA/PBO1/SPDP/2023/BNNP Bengkulu	12/22/2023	Denny Alias Denny Bin M Inu Yusu (Alm)	BNNP Bengkulu	Narkotika	Peredaran gelap narkotika golongan I dalam bentuk bukan tanaman jenis shabu sebagaimana dimaksud da
35	B/0022/XII/KA/PBO1/SPDP/2023/BNNP Bengkulu	12/22/2023	Denny Alias Denny Bin M Inu Yusu (Alm)	BNNP Bengkulu	Narkotika	Peredaran gelap narkotika golongan I dalam bentuk bukan tanaman jenis shabu sebagaimana dimaksud da
36	SPDP/134/XII/RES.1.24./2023/Reskrim	12/22/2023	ALQOBRIANSYAH Als AAN Bin SUTAN ALIM TOIM (Alm)	POLRES REJANG LEBONG	KDRT	["Pasal 49 UU Nomor 23 Tahun 2004 tentang Penghapusan Kekerasan Dalam Rumah Tangga"]
37	SPDP/78/XII/RES.1.4/2023/RESKRIM	12/22/2023	MUFI ARIS TONI Bin PAWAN SEGIANTO	POLRES BENGKULU SELATAN	Persetubuhan	Pasal 81 Ayat (2) UU RI No. 17 Tahun 2016 tentang perubahan kedua atau UU RI No.23 Tahun 2002 tentang
38	SPDP/5/XII/RES.1.8/2023	12/22/2023	YONGKI WIRANTO BIN ANDI PURNAWAN	POLSEK PINO RAYA	Pencurian	pasal 363 KUHP

Figure 1. Dataset Sample

2.2 Data Preprocessing

To ensure optimal data quality, several preprocessing steps were undertaken:

a) Data Cleaning: Removal of special characters, numbers, and irrelevant symbols.

```
[100 rows x 8 columns]
[INFO] Memeriksa data yang kosong...
Data setelah menghapus nilai kosong:
     No                              No, tgl SPDP Tanggal SPDP Diterima  \
0     1  SPDP/ 117 / XII / RES. 1.6. / 2023 / RESKRIM            12/29/2023
1     2                    SPDP/03/XII/2023/Reskrim            12/29/2023
2     3              SPDP/16/XII/2023/RESKRIM/SEK-SK            12/29/2023
3     4    SPDP/43.B/XII/RES.1.24./2023/Ditreskrimum            12/28/2023
4     5    SPDP/43.B/XII/RES.1.24./2023/Ditreskrimum            12/28/2023
..  ...                                          ...                   ...
95   96              SPDP/133/XII/2023/DITRESNARKOBA            12/19/2023
96   97              SPDP/130/XII/2023/DITRESNARKOBA            12/19/2023
97   98              SPDP/131/XII/2023/Ditresnarkoba            12/18/2023
98   99            SPDP/89/XII/RES.1.24/2023/Reskrim            12/18/2023
99  100              SPDP/131/XII/2023/Ditresnarkoba            12/18/2023

                    Tersangka/Terdakwa               Penyidik  \
0                  ASMAWI Bin SULAIMAN            POLRES KAUR
```

Figure 2. Data Cleaning Results

b) Tokenization: Splitting the case description text into individual words.

```
[100 rows x 2 columns]
Data setelah tokenizing:
     casefolding      tokenizing
0     Pengiayaan    [Pengiayaan]
1      Pencurian     [Pencurian]
2     Pengiayaan    [Pengiayaan]
3           KDRT          [KDRT]
4           KDRT          [KDRT]
..           ...             ...
95     Narkotika     [Narkotika]
96     Narkotika     [Narkotika]
97     Narkotika     [Narkotika]
98  Penganiayaan  [Penganiayaan]
99     Narkotika     [Narkotika]
```

Figure 3. Data Tokenization

c) Normalization: Converting non-standard words into their standard forms

```
[100 rows x 2 columns]
Data setelah normalisasi:
         filtering     normalisasi
0     [Pengiayaan]    [Pengiayaan]
1      [Pencurian]     [Pencurian]
2     [Pengiayaan]    [Pengiayaan]
3           [KDRT]          [KDRT]
4           [KDRT]          [KDRT]
..             ...             ...
95     [Narkotika]     [Narkotika]
96     [Narkotika]     [Narkotika]
97     [Narkotika]     [Narkotika]
98  [Penganiayaan]  [Penganiayaan]
99     [Narkotika]     [Narkotika]
```

Figure 4. Normalization Results

d) Stemming: Reducing words to their root forms using Indonesian stemming algorithms, such as Sastrawi.

```
[100 rows x 2 columns]
[INFO] Proses stemming selesai untuk baris ke-1 dalam 0.46 detik.
[INFO] Proses stemming selesai untuk baris ke-2 dalam 0.05 detik.
[INFO] Proses stemming selesai untuk baris ke-3 dalam 0.37 detik.
[INFO] Proses stemming selesai untuk baris ke-4 dalam 0.22 detik.
[INFO] Proses stemming selesai untuk baris ke-5 dalam 0.18 detik.
[INFO] Proses stemming selesai untuk baris ke-6 dalam 0.18 detik.
[INFO] Proses stemming selesai untuk baris ke-7 dalam 0.22 detik.
[INFO] Proses stemming selesai untuk baris ke-8 dalam 0.26 detik.
[INFO] Proses stemming selesai untuk baris ke-9 dalam 0.24 detik.
[INFO] Proses stemming selesai untuk baris ke-10 dalam 0.21 detik.
[INFO] Proses stemming selesai untuk baris ke-11 dalam 0.22 detik.
[INFO] Proses stemming selesai untuk baris ke-12 dalam 0.04 detik.
[INFO] Proses stemming selesai untuk baris ke-13 dalam 0.04 detik.
[INFO] Proses stemming selesai untuk baris ke-14 dalam 0.04 detik.
[INFO] Proses stemming selesai untuk baris ke-15 dalam 0.04 detik.
[INFO] Proses stemming selesai untuk baris ke-16 dalam 0.04 detik.
```

Figure 5. Stemming Process

e) Case Folding: Converting all text to lowercase to ensure consistency.

```
[100 rows x 8 columns]
Data setelah casefolding:
   Jenis Tindak Pidana   casefolding
0           Pengiayaan    Pengiayaan
1            Pencurian     Pencurian
2           Pengiayaan    Pengiayaan
3                 KDRT          KDRT
4                 KDRT          KDRT
..                 ...           ...
95           Narkotika     Narkotika
96           Narkotika     Narkotika
97           Narkotika     Narkotika
98       Penganiayaan   Penganiayaan
99           Narkotika     Narkotika
```

Figure 6. Case Folding Results

f) Filtering: Removing words that do not carry significant meaning (stopwords).

```
[100 rows x 2 columns]
Data setelah filtering:
        tokenizing       filtering
0     [Pengiayaan]    [Pengiayaan]
1      [Pencurian]     [Pencurian]
2     [Pengiayaan]    [Pengiayaan]
3           [KDRT]          [KDRT]
4           [KDRT]          [KDRT]
..             ...             ...
95     [Narkotika]     [Narkotika]
96     [Narkotika]     [Narkotika]
97     [Narkotika]     [Narkotika]
98  [Penganiayaan]  [Penganiayaan]
99     [Narkotika]     [Narkotika]
```

Figure 7. Filtering Results

g) SMOTE: penanganan kelas data tidak seimbang, menggunakan metode SMOTE *(Synthetic Minority Oversampling Technique)* menangani ketidakseimbangan distribusi data antar kelas, sehingga model lebih mampu mengenali semua kategori Tindak Pidana.

```
[INFO] Memeriksa distribusi kelas sebelum SMOTE..
Jenis Tindak Pidana
Narkotika              95
pencurian              15
Pencurian               5
Pengeroyokan            4
KDRT                    3
Persetubuhan            3
Penganiayaan            2
penganiayaan            2
Penggelapan             2
penipuan                2
narkotika               1
Perkosaan               1
Penipuan                1
Perlindungan Anak       1
Perdagangan Orang       1
Name: count, dtype: int64
```

```
[INFO] Distribusi kelas setelah SMOTE:
Jenis Tindak Pidana
Narkotika       95
pencurian       95
Pencurian       95
KDRT            95
Penganiayaan    95
penganiayaan    95
Penggelapan     95
Persetubuhan    95
penipuan        95
Pengeroyokan    95
Name: count, dtype: int64
```

Figure 8. Before and After SMOTE

h) Feature Extraction: Applying the TF-IDF (Term Frequency-Inverse Document Frequency) technique to convert text into numerical vectors suitable for machine learning algorithms.

```
# Step 10: TF-IDF Vectorization
print("[INFO] Mengonversi teks menjadi fitur TF-IDF...")
tfidf = TfidfVectorizer(ngram_range=(1, 3), max_features=3000)
X = tfidf.fit_transform(data['cleaned_text']).toarray()
y = data['Jenis Tindak Pidana']
```

Figure 9. TF-IDF Representation

2.3 Machine Learning Algorithm Selection

This study employs two primary algorithms for classification:

a) LightGBM (Light Gradient Boosting Machine): A library or framework designed to build machine learning models based on gradient boosting. LightGBM is widely recognized for its fast performance, memory efficiency, and ability to handle large datasets effectively.

b) Random Forest: An ensemble-based algorithm that combines multiple decision trees to improve classification performance.

2.4 Model Training and Testing

The model training and testing phase was conducted to ensure that the developed classification system could accurately learn from historical case data and reliably predict case categories on unseen data. This stage encompassed a structured process involving data partitioning, model learning, and performance evaluation, aimed at assessing the model's effectiveness and generalization capability.

a) Data Splitting: The dataset was divided into training (80%) and testing (20%) subsets using the Stratified Shuffle Split method. This approach was chosen to preserve the proportional representation of each class within both subsets, thereby maintaining the integrity of the original class distribution and ensuring a fair evaluation.

b) Model Training: The training subset was utilized to enable the model to identify patterns, correlations, and distinguishing features within the dataset. Through this process, the model learned the underlying relationships between the textual features and the corresponding case categories, forming the basis for accurate predictions.

c) Model Testing: The testing subset was employed to evaluate the model's predictive capability on previously unseen data. This step provided an objective measure of the model's performance, ensuring that it could generalize effectively beyond the training data and deliver reliable classification results in real-world scenarios.

2.5 Model Performance Evaluation

The performance of the algorithms was assessed using the following metrics:

a) Accuracy: The proportion of correct predictions relative to the total number of predictions.
b) Precision: The model's ability to correctly predict positive classes.
c) Recall (Sensitivity): The model's capability to identify all actual positive cases.
d) F1-Score: The harmonic mean of precision and recall, providing an overall measure of model performance.
e) Confusion Matrix: Used to analyze the distribution of predictions against the actual classes.

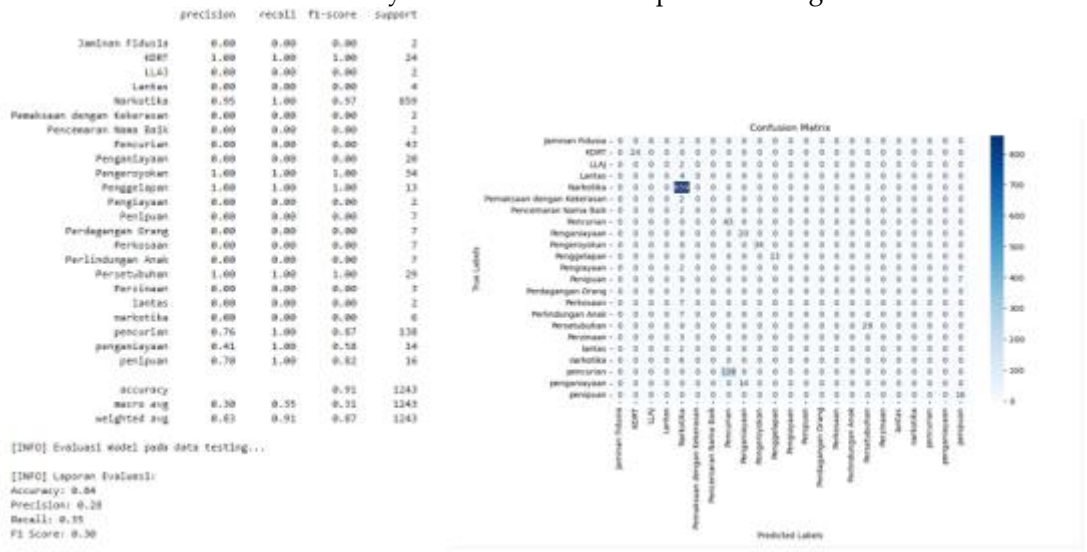

Figure 10. Model Performance

2.6 Best Model Selection

Following the evaluation, the LightGBM algorithm demonstrated the highest performance and was selected as the primary classification model. This algorithm will be implemented in a prototype system for the automated classification of general criminal case types.

2.7 Implementation and Validation

The best-performing model was integrated into a prototype classification system. The system was further tested using new data to validate its reliability and effectiveness in real-world scenarios.

3. FINDINGS AND DISCUSSION

The results of the testing phase demonstrate that the Light Gradient Boosting Machine (LightGBM) algorithm outperforms the Random Forest algorithm across all evaluation metrics in the classification of general criminal cases at the Bengkulu High Prosecutor's Office. In particular, LightGBM consistently achieved higher scores in accuracy, precision, recall, and F1-score, indicating its superior ability to process and classify complex textual case data. This remarkable performance can be attributed to the gradient boosting mechanism, which iteratively improves the model by learning from the errors of previous iterations, thereby optimizing predictive accuracy. Moreover, LightGBM's efficient computational design minimizes processing time, allowing for rapid analysis even when dealing with large and intricate datasets. The combination of these factors positions LightGBM as a highly effective tool for automating case classification in prosecutorial workflows, where speed and precision are critical.

In contrast, while Random Forest also produced satisfactory results, it exhibited limitations in capturing subtle patterns present within textual data. The ensemble of decision trees that Random

Forest relies on tends to generalize well for balanced and straightforward datasets, but it may struggle to identify fine-grained distinctions in complex or high-dimensional text data, such as case descriptions that contain nuanced legal terminology or context-specific phrasing. These limitations became particularly evident when the dataset contained imbalanced class distributions prior to the application of the Synthetic Minority Over-sampling Technique (SMOTE). In such cases, minority classes were underrepresented, which negatively affected the Random Forest's ability to correctly classify less frequent categories. LightGBM, on the other hand, maintained strong predictive performance under the same conditions due to its ability to handle imbalanced data more effectively and focus on learning from misclassified instances in each boosting iteration.

A critical factor contributing to the success of both algorithms, and particularly LightGBM, was the comprehensive preprocessing pipeline applied to the textual data. Several preprocessing techniques were employed, including text normalization, stemming, case folding, tokenization, and stopword removal. Text normalization standardized non-conventional or informal words into consistent forms, ensuring that variations in spelling or abbreviations did not create redundant features that could mislead the model. Stemming reduced words to their root forms, which allowed the model to recognize different inflected forms of the same word as a single feature, thereby improving feature consistency and reducing dimensionality. Case folding converted all text into lowercase, ensuring uniformity across the dataset and preventing the model from treating the same word in different cases as distinct features. Tokenization separated sentences and phrases into individual words or tokens, providing a structured representation of textual content for further processing. The removal of stopwords filtered out commonly occurring words that carry little semantic value, such as conjunctions or prepositions, allowing the model to focus on words with greater discriminative power for classification. Collectively, these preprocessing steps substantially reduced noise in the data and enhanced the quality of features, directly contributing to improved algorithmic performance.

The choice of feature extraction method further amplified the model's effectiveness. Term Frequency-Inverse Document Frequency (TF-IDF) was employed to transform textual data into numerical vectors that reflect the importance of each term relative to the entire dataset. TF-IDF assigns higher weights to words that are unique or particularly distinguishing for a specific case type, while reducing the influence of frequently occurring but less informative terms. This approach enabled LightGBM to prioritize the most salient features, enhancing its ability to differentiate between similar case types. When combined with the preprocessing pipeline, TF-IDF provided a highly informative representation of textual data, ensuring that the model could identify meaningful patterns that are critical for accurate classification.

The comparative analysis between LightGBM and Random Forest revealed additional insights into the mechanisms driving model performance. LightGBM's iterative gradient boosting process emphasizes learning from errors made in previous trees, which is particularly advantageous when dealing with overlapping or ambiguous cases. This iterative correction mechanism allows the model to refine its predictions progressively, ultimately achieving higher precision and recall than Random Forest. Moreover, LightGBM incorporates advanced techniques such as histogram-based decision tree learning, which improves computational efficiency by discretizing continuous features into histograms. This approach not only reduces memory usage but also accelerates model training, making it feasible to apply LightGBM to large-scale case datasets without significant computational overhead.

Random Forest's performance, while still robust, highlighted some of the challenges associated with traditional ensemble tree methods in complex textual domains. Its reliance on averaging predictions from multiple decision trees can dilute sensitivity to minority classes or subtle textual cues, which are often critical in legal case classification. Furthermore, Random Forest does not inherently perform iterative error correction, meaning that misclassified cases in early trees do not directly inform subsequent trees, unlike the boosting approach used by LightGBM. Consequently, while Random Forest provides a reliable baseline, it does not fully exploit the nuanced structure of high-dimensional textual data as effectively as gradient boosting methods.

Another notable observation is the interplay between data balancing techniques and algorithm performance. Before the application of SMOTE, the dataset exhibited class imbalance, with some case types underrepresented relative to others. This imbalance posed a challenge for both algorithms, but the impact was more pronounced for Random Forest. SMOTE synthetically generates additional instances for minority classes, creating a more balanced training set that allows algorithms to learn representative patterns for all classes. After applying SMOTE, both algorithms showed improved performance, but LightGBM maintained a clear advantage, achieving higher accuracy and more balanced precision-recall trade-offs across classes.

The practical implications of these findings are significant for the operational efficiency of the Bengkulu High Prosecutor's Office. Implementing LightGBM as the primary classification model enables the automation of case categorization, reducing dependence on manual review and allowing staff to focus on substantive legal work. By accurately predicting case types, the system can streamline workflow processes, expedite case assignment, and ensure more consistent handling of cases. Furthermore, the integration of this model into a real-world operational system provides a foundation for continuous learning, as new case data can be incrementally incorporated to further refine predictive performance.

In conclusion, the comprehensive evaluation of LightGBM and Random Forest underscores the critical role of algorithm selection, data preprocessing, feature extraction, and class balancing in developing high-performance models for legal text classification. LightGBM's superior accuracy, precision, recall, and F1-score demonstrate its capacity to effectively process complex textual case data, while Random Forest provides a comparative benchmark highlighting the advantages of gradient boosting. The combined use of preprocessing techniques such as normalization, stemming, case folding, tokenization, and stopword removal, along with TF-IDF feature representation and SMOTE balancing, contributed substantially to the reliability and robustness of the classification system. Ultimately, the deployment of the LightGBM-based model within the Bengkulu High Prosecutor's Office offers a practical, scalable, and efficient solution for automating the classification of general criminal case types, improving operational efficiency, ensuring consistency in case handling, and supporting a more transparent and structured legal process. This study illustrates the transformative potential of machine learning in modernizing prosecutorial workflows, providing actionable insights for future applications in legal technology and intelligent case management systems.

4. CONCLUSION

Based on the findings and discussion, it can be concluded that the application of machine learning algorithms—particularly the Light Gradient Boosting Machine (LightGBM)—offers an effective solution for automating the classification of general criminal case types at the Bengkulu High Prosecutor's Office. The primary strength of LightGBM lies in its ability to process complex textual data with both speed and precision, while maintaining consistent performance even in the presence of imbalanced class distributions. This achievement is supported by a comprehensive data preprocessing pipeline, the use of appropriate feature extraction techniques, and the application of class-balancing methods such as SMOTE, resulting in a robust and reliable classification model for real-world implementation. The deployment of this system is expected to enhance operational efficiency, reduce administrative workload, ensure consistency in case handling, and promote greater transparency in legal proceedings. Furthermore, these findings highlight the significant potential of artificial intelligence technologies in the legal sector, serving as a reference point for the development of similar systems in other law enforcement institutions.

REFERENCES

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). ACM.

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T. Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems, 30*, 3149–3157.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys, 34*(1), 1–47.

Sun, A., Lim, E. P., & Ng, W. K. (2009). Web classification using support vector machine. In *Text mining: Application and theory* (pp. 207–226). Wiley.

Wibowo, A., & Santoso, F. H. (2021). Implementasi algoritma machine learning untuk klasifikasi data kriminalitas. *Jurnal Teknologi Informasi dan Komunikasi, 12*(2), 45–52.

Yang, Y., & Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. In *Proceedings of the 14th International Conference on Machine Learning* (pp. 412–420). Morgan Kaufmann.

Zhang, Z., & Zong, C. (2016). Deep neural networks in machine learning: A brief review. *IEEE Access, 4*, 1759–1771.