

Machine Learning-Based Prediction of Divorce Verdicts Using Posita Data and Imbalanced Data Handling: A Case Study in Padang Sidempuan

Rina Rahmadini¹, Bagus Jati Santoso²

^{1,2}School of Interdisciplinary Management and Technology, Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia

Article Info

Article history:

Received May 29, 2025

Revised Jul 18, 2025

Accepted Aug 19, 2025

Keywords:

Machine Learning,
Court Verdict Prediction,
Divorce Case Analysis,
Posita Data,
Imbalanced Data Handling

ABSTRACT

This study aims to develop a predictive model for divorce verdicts ("Granted" or "Rejected") in the Religious Courts of Indonesia using machine learning techniques. The dataset consists of 2,026 finalized divorce cases from the Religious Court of Padang Sidempuan between 2018 and 2025, incorporating structured variables and posita—narrative texts describing the plaintiff's reasons for divorce. Keyword-based feature extraction was applied to transform these texts into interpretable indicators. To handle class imbalance, Synthetic Minority Over-sampling Technique (SMOTE) was implemented on the training data. Six classical machine learning algorithms were evaluated: Decision Tree, Naïve Bayes, K-Nearest Neighbors, Random Forest, LightGBM, and XGBoost. Performance was measured using accuracy, precision, recall, F1-score, F2-score, and AUC. The results indicate that Naïve Bayes achieved the highest recall (100%) for the "Granted" class, while LightGBM and XGBoost demonstrated the most balanced performance across both classes. Feature importance analysis revealed that mediation outcomes, domestic violence, and economic hardship were among the most influential factors in determining verdicts. The study highlights the applicability of interpretable machine learning in legal decision support and discusses limitations such as the single-court scope and challenges in predicting minority class outcomes. Future work may explore multi-jurisdictional data, deep learning approaches, and domain-specific embeddings for enhanced performance.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Rina Rahmadini,
School of Interdisciplinary Management and Technology,
Institut Teknologi Sepuluh Nopember,
Kampus ITS Tjokroaminoto Jl. Cokroaminoto No. 12A, Tegalsari, Surabaya 60264, Indonesia
Email: 6032231236@student.its.ac.id

1. INTRODUCTION

Divorce cases in Indonesia have shown a significant increase over the years. According to data from the Central Bureau of Statistics (BPS), the national divorce rate reached more than 516,000 cases in 2022—marking the highest figure in the past seven years [1]. This phenomenon reflects the increasingly complex social dynamics of society and the growing workload of judicial institutions, particularly the Religious Courts, which have jurisdiction over family law cases under Islamic law in Indonesia [2].

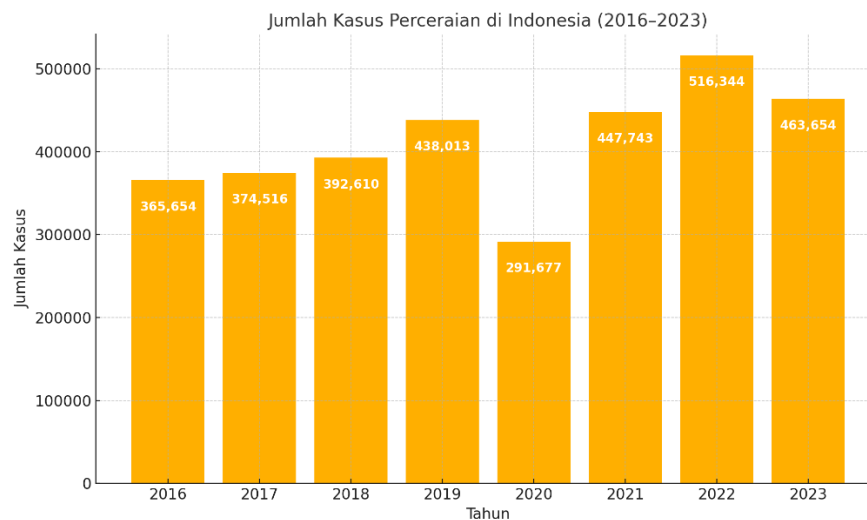


Figure 1. Number of divorce cases in Indonesia from 2016 to 2023 based on data from BPS [1].

The high volume of divorce cases urges the courts to seek innovative solutions to improve decision-making efficiency and consistency. In the era of the Fifth Industrial Revolution (Industry 5.0), there is a growing opportunity to integrate intelligent technologies into the Indonesian judicial system to support decision-making processes [3]. Machine Learning (ML), a subfield of artificial intelligence, enables the identification of patterns from historical data to predict outcomes or classifications without explicit programming [4]. This approach is believed to assist judges by accelerating case analysis and enhancing objectivity in verdicts [5].

Prior studies have demonstrated the effectiveness of ML in various legal contexts. Sohail et al. [6], applied Decision Tree, K-Nearest Neighbor (KNN), and Naïve Bayes in Pakistan to identify key factors influencing divorce rates, showing Naïve Bayes with the best accuracy. Medvedeva et al. [7] applied Random Forest to predict rulings from the European Court of Human Rights (ECHR), achieving high accuracy but focusing on human rights cases rather than family law. Aimran et al. [8], compared Decision Tree, Logistic Regression, and Neural Networks in predicting divorce cases among Malaysian women, with Decision Tree yielding the highest performance. In Indonesia, Bramantoro et al. [9] used Convolutional Neural Networks (CNN) to identify the causes of divorce during the COVID-19 pandemic. More recently, Zhang et al. [10], proposed an event extraction system for judicial assistance in divorce case processing, while Ibrahim et al. [11], explored deep learning models to predict punishment durations in Indonesian court rulings.

While these studies contribute valuable insights, they often rely on structured demographic data and lack focus on real narrative legal texts or the Indonesian Religious Court system. Moreover, very few studies incorporate authentic judicial documents such as *posita*—the plaintiff’s factual narrative—and even fewer evaluate actual verdict outcomes (“Granted” vs. “Rejected”) using such unstructured text. Most notably, prior works often overlook the problem of class imbalance, where rejected verdicts are significantly underrepresented.

Given these gaps, this research offers a more contextual and domain-specific approach to match the judicial environment and data structure of Indonesian Religious Courts. This study adopts a machine learning-based classification approach using real *posita* narratives and final verdict labels collected from the Religious Court of Padang Sidempuan. The *posita* section is a crucial element in divorce proceedings, containing the legal reasoning presented by the plaintiff, and serves as one of the main considerations in judicial decisions [12]. By converting these textual arguments into structured indicators, this research builds predictive models that are both legally relevant and empirically grounded.

Based on the above background, this study aims to:

1. Design and develop a predictive model for divorce verdicts in the Religious Courts using machine learning algorithms based on *posita* data;

2. Measure and evaluate the performance of various algorithms using classification metrics such as accuracy, precision, recall, F1-score, F2-score, and AUC;
3. Analyze and compare the predictive capability of different algorithms to determine the most suitable for court decision support.

This research employs six machine learning algorithms—Decision Tree, Naïve Bayes, KNN, Random Forest, LightGBM, and XGBoost—while addressing severe class imbalance using the SMOTE technique [13].

The novelty of this study lies in its use of authentic posita narratives and final verdict data from real divorce cases in the Indonesian Religious Courts. Unlike prior studies that rely on synthetic or structured demographic data, this research predicts actual court verdicts (“Granted” or “Rejected”) using raw judicial text and handles severe class imbalance through SMOTE. The model is built upon real legal contexts, providing grounded, interpretable, and practically applicable insights for future decision-support systems.

With this approach, the research is expected to serve as an initial foundation for developing a transparent and objective decision support system that enhances legal efficiency in the Religious Courts. As predictive technologies continue to evolve, Religious Courts in Indonesia will be better equipped to integrate these innovations—while upholding judicial independence as mandated by Law No. 48 of 2009 on Judicial Authority [14].

2. RESEARCH METHOD

This research began by collecting divorce case data from the Religious Court of Padang Sidempuan, covering all civil litigation cases that had reached a final verdict (minutated). The data were obtained from the official case-tracking application, Sistem Informasi Penelusuran Perkara (SIPP), with formal permission from the relevant institution and in compliance with ethical principles and privacy protection standards [15]. The collected data then underwent a preprocessing stage, which included data cleaning, variable normalization, and key information extraction from the posita (statement of claims) section.

After preprocessing, the dataset was divided into two subsets: training data and testing data. Due to class imbalance—where the number of “Granted” cases significantly exceeded “Rejected” ones—the training data were balanced using the Synthetic Minority Over-sampling Technique (SMOTE). This technique was employed to improve the model’s sensitivity to the minority class and prevent bias during training [16].

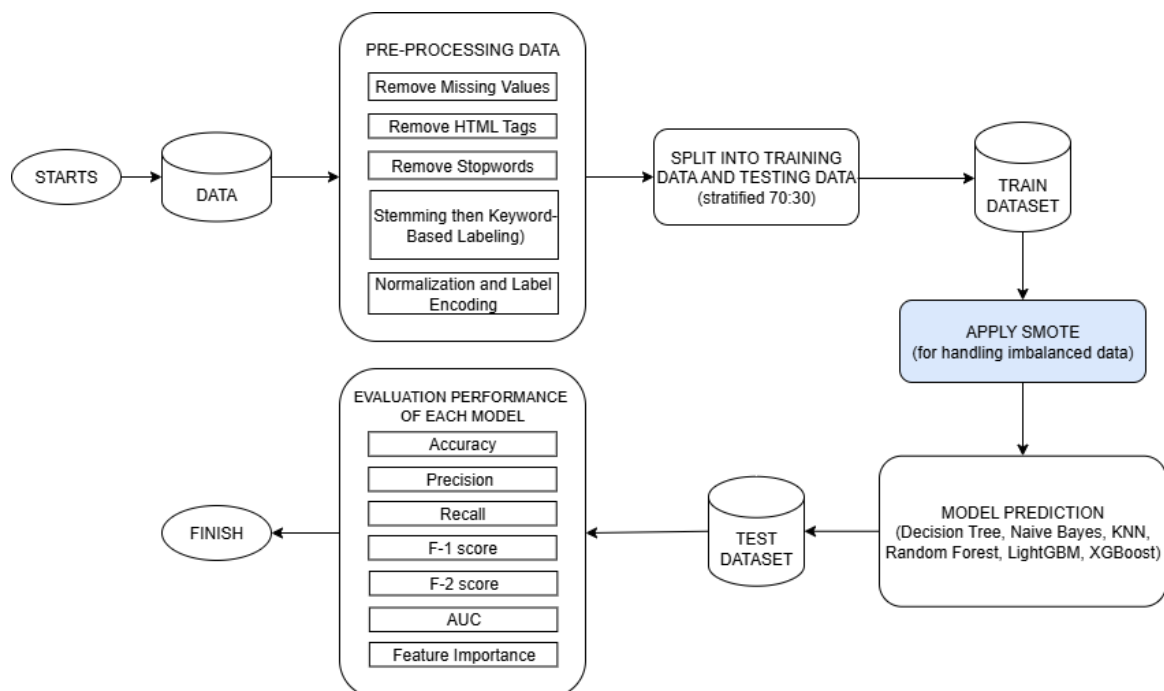


Figure 2. Research Flow Diagram

The predictive model was developed using training data that had been balanced through an oversampling technique and trained using six different machine learning algorithms: Decision Tree, Naïve Bayes, K-Nearest Neighbors (KNN), Random Forest, LightGBM, and XGBoost. Each model was subsequently tested using the original testing dataset, which was not subjected to oversampling. Model performance was evaluated using a range of classification metrics, including accuracy, precision, recall, F1-score, F2-score, and Area Under the Curve (AUC), to assess the model's ability to distinguish between the two target classes: "Granted" and "Rejected" [17]. In addition to performance metrics, a feature importance analysis was conducted to identify the most influential variables contributing to the model's predictions. This evaluation offers deeper insights into the key factors within the *posita* data that most significantly impact court decisions.

This systematic approach was designed to produce a reliable and accurate predictive model while supporting enhanced efficiency and objectivity in the legal decision-making process for divorce cases at the Religious Court of Padang Sidempuan.

2.1. Data Source and Description

As an initial step, data were obtained through a data mining process from the Sistem Informasi Penelusuran Perkara (SIPP) or Case Tracking Information System of the Religious Court of Padang Sidempuan. The dataset comprises 2,026 finalized divorce cases—specifically *cerai gugat* (divorce filed by the wife) and *cerai talak* (divorce filed by the husband)—with verdicts officially recorded as either "granted" or "rejected" between the years 2018 and 2025. This time range was selected to ensure comprehensive coverage and relevance to current socio-legal conditions.

The dataset includes several key features relevant to the characteristics of each case, such as: type of case, age of the plaintiff/applicant, age of the defendant/respondent, duration of the marriage, occupation of both parties, the textual content of the *posita*, and the final court decision. These features served as input variables for the predictive modeling process using machine learning algorithms.

The *posita* section, in particular, is a narrative text containing the plaintiff's account of events and reasons for seeking divorce. This textual content was preprocessed and used to extract legal and social indicators through keyword-based techniques, which were then converted into categorical variables to support classification and prediction of verdict outcomes.

The data used in this study were obtained with permission from the public access section of the Religious Court of Padang Sidempuan. Although SIPP data does not contain personally identifiable information (PII), additional care was taken to remove residual identifiers such as case numbers to ensure compliance with ethical research standards. All records consist solely of legal and demographic attributes relevant for analytical modeling.

This research adopts a case study approach focused on a single jurisdiction, allowing for contextual depth and practical applicability within the Indonesian Religious Court system. However, this approach may limit generalizability to other courts, which is further discussed in the limitation section.

2.2. Data Preprocessing

The data pre-processing stage is a critical step in developing a machine learning model, as the quality of the data significantly affects the accuracy and reliability of prediction outcomes. The first step involved removing rows with missing values in important features. This was done to ensure that only complete and valid data were used for the training and testing processes.

Next, text cleaning was performed on the *isi_posita* column by removing HTML tags and non-alphabetic characters. This column contains divorce reasoning narratives in HTML format, which needed to be cleaned from irrelevant tags and symbols to prepare the text for further analysis. The process then continued with the removal of stopwords—common words that carry minimal semantic value, such as "and", "which", or "for"—to ensure the extracted information from the text is more focused and meaningful.

The next step was stemming and keyword-based extraction of divorce reasons. The text in the *isi_posita* column was analyzed to identify five main categories of divorce reasons: domestic

violence, infidelity, economic issues, conflict, and social involvement. Before keyword matching, the text was processed using stemming to reduce words to their root forms [19]. Each category was then encoded into a binary format: 1 if the reason was found, and 0 otherwise.

This was followed by normalization and label encoding. Categorical variables such as the occupations of the plaintiff and defendant were grouped into six general categories (unemployed, farmer, laborer, entrepreneur, employee, and civil servant), and encoded numerically from 0 to 5. Other categorical features, including case type, gender, and mediation result, were also converted into numeric values using label encoding. For numeric variables such as the ages of the plaintiff, defendant, and length of marriage, Min-Max Scaling was applied to rescale them into the [0, 1] range. Specifically, for the marriage duration feature, a value of 0 represents less than one year, while a value of 1 represents the maximum length of years in the dataset.

Upon completion of the pre-processing stage, the final dataset consists of 15 refined and structured features, which include both engineered and transformed variables. These features are now ready to be used. A detailed summary of each variable, along with its data type and description, is presented in the table below.

Table 1. Description of Final Dataset Features After Pre-processing

No	Variable	Data Type	Feature Description
1	jenis_perkara	Numerical (Integer)	Type of case (1 = Divorce filed by wife, 0 = Divorce filed by husband)
2	jenis_kelamin_p	Numerical (Integer)	Gender of the plaintiff (1 = Female, 0 = Male)
3	usia_pernikahan	Numerical (Float)	Duration of marriage (normalized using Min-Max Scaling).
4	usia_p	Numerical (Float)	Age of the plaintiff at the time of filing for divorce.
5	usia_t	Numerical (Float)	Age of the defendant at the time the divorce was filed.
6	pekerjaan_p	Numerical (Integer)	Occupation type of the plaintiff (normalized and mapped into categories; civil servant, entrepreneur, employee, etc.).
7	pekerjaan_t	Numerical (Integer)	Occupation type of the defendant (normalized and mapped into categories; civil servant, entrepreneur, employee, etc.).
8	jumlah_anak	Numerical (Float)	Number of children in the marriage.
9	mengandung_kekerasan	Numerical (Integer)	Whether the divorce involved domestic violence (1 = Yes, 0 = No).
10	mengandung_perselingkuhan	Numerical (Integer)	Whether infidelity was cited as a reason for divorce (1 = Yes, 0 = No).
11	mengandung_faktorekonomi	Numerical (Integer)	Whether the divorce was influenced by economic/financial issues (1 = Yes, 0 = No).
12	mengandung_pertengkaran	Numerical (Integer)	Whether the divorce was caused by continuous conflict (1 = Yes, 0 = No).
13	mengandung_socialinvolvement	Numerical (Integer)	Whether the divorce was triggered by external social influences (1 = Yes, 0 = No).
14	hasil_mediasi	Numerical (Integer)	Outcome of the mediation session (1 = Successfully reconciled, 0 = Not successful).
15	status_putusan	Numerical (Integer)	Final verdict of the divorce case (1 = Granted, 0 = Rejected).

After all features were cleaned and converted, the dataset was split into training and testing sets using a 70:30 ratio with stratified splitting, to maintain the proportion of class distribution in the target label status_putusan (1 = Granted, 0 = Rejected) [20].

2.3. Handling Imbalanced Data

One of the main challenges in classification modeling is class imbalance in the target label. In this dataset, there is a significant disparity between the number of "Granted" and "Rejected" divorce verdicts, with the former overwhelmingly dominating the class distribution. After the data pre-processing stage, it was found that 1,646 cases were labeled as "Granted," while only 380 cases were labeled as "Rejected." This imbalance can lead to a machine learning model that is biased toward predicting the majority class, thereby reducing its ability to accurately recognize minority class cases.

To address this issue, this research applied the Synthetic Minority Over-sampling Technique (SMOTE) to the training dataset. SMOTE is an oversampling method that generates synthetic

samples for the minority class by interpolating between neighboring instances using the k-nearest neighbors algorithm. This technique increases the number of minority class samples to balance the dataset, allowing the model to learn patterns from both classes more fairly [21] and improving its sensitivity to "Rejected" cases.

The implementation of SMOTE was applied only to the training data to avoid **data leakage**, which could otherwise invalidate the model evaluation. The procedure was carried out as follows:

1. The dataset was initially split into training and testing sets using a 70:30 ratio, with **stratification** to preserve the class distribution in both subsets.
2. In the training data, it was observed that the "Granted" class made up over 80% of the total cases.
3. SMOTE was applied to the training data using the `imblearn.over sampling.SMOTE` library in Python, with
 - `k_neighbors=5`: the number of nearest neighbors used for creating synthetic instances.
 - `sampling_strategy='auto'`: automatically balances the number of instances between classes.
4. The result was a balanced training dataset, where the number of "Granted" and "Rejected" cases became approximately equal.

The SMOTE-augmented dataset was used exclusively as the training set for the machine learning models, while evaluation was performed on the original, unaltered test set. This strategy was adopted to enhance the model's ability to generalize on the minority class and to ensure fair and accurate predictions across both categories of court verdicts.

2.4. Machine Learning Algorithms

This research implements six different machine learning algorithms as classification methods to develop a predictive model for court verdicts in divorce cases. The selected algorithms are: Decision Tree, Naïve Bayes, K-Nearest Neighbor (KNN), Random Forest, Light Gradient Boosting Machine (LightGBM), and Extreme Gradient Boosting (XGBoost). These algorithms were chosen based on prior research findings and their proven effectiveness in handling binary classification tasks, especially those involving structured tabular data and imbalanced classes. Each algorithm brings distinct strengths to the classification process.

- Decision Tree offers high interpretability. Its tree structure allows users to visualize decision paths and understand the logic behind each prediction, making it especially suitable for exploring hidden rules in the dataset and explaining outcomes to non-technical stakeholders [22].
- Naïve Bayes is a probabilistic model that excels in simplicity and speed. It performs well on relatively small datasets and is known for its robustness to irrelevant features, despite its strong assumption of feature independence [23].
- K-Nearest Neighbor (KNN) is a non-parametric algorithm that classifies data points based on the majority class among their nearest neighbors. It is capable of capturing local structures in data, making it useful in cases where patterns depend on proximity in feature space [24].
- Random Forest is an ensemble method that combines the predictions of multiple decision trees to enhance accuracy and reduce the risk of overfitting. By averaging the results from numerous trees built on random subsets of data and features, it provides robust performance across various data types and distributions [25].
- LightGBM is a highly efficient gradient boosting algorithm developed for speed and scalability. It grows trees leaf-wise rather than level-wise, which allows it to converge faster and handle large, complex, and imbalanced datasets with less computational cost [26].
- XGBoost is a powerful gradient boosting framework that incorporates regularization, efficient handling of missing values, and built-in cross-validation. It is widely regarded for

its superior performance in competitions and real-world scenarios involving structured data [27].

The primary reason for employing these classical machine learning algorithms lies in their mature theoretical foundations, efficient implementation, and interpretability. Unlike deep learning models that often require large datasets and significant computational resources, classical algorithms are more practical for moderately sized datasets and still offer competitive performance [28]. They also provide clearer insights into feature importance and decision logic, which is crucial in legal contexts where transparency and explainability are key.

Therefore, these algorithms are considered an appropriate and effective choice for building a reliable and interpretable predictive model in the domain of court decision support systems. All models were implemented using Scikit-learn and Imbalanced-learn in Python 3.10, which offer robust support for classical machine learning workflows and imbalanced data handling.

While transformer-based deep learning models such as BERT and Legal-BERT have demonstrated strong performance in many legal Natural Language Processing (NLP) tasks, this study deliberately excludes such approaches for several methodological reasons. First, there is currently no pretrained Legal-BERT model available for Indonesian legal texts, particularly those rooted in Islamic law as practiced in Religious Courts. Second, this research emphasizes model interpretability and explainability—essential qualities in judicial decision support systems—which are better supported by tree-based algorithms like LightGBM and XGBoost than by transformer models. Third, the computational demands and infrastructure required for training and fine-tuning deep learning models exceed the scope and resources of this study. Thus, classical machine learning algorithms offer a balanced trade-off between performance, transparency, and feasibility for the research objectives and institutional context.

2.5. Evaluation Metrics

To comprehensively assess model performance, this study employs a set of evaluation metrics that provide both general and class-specific insights: Accuracy, Precision, Recall, F1-score, F2-score, the Area Under the Receiver Operating Characteristic Curve (AUC), and feature importance. These metrics were selected to reflect not only the overall correctness of predictions but also the sensitivity of the models in handling class imbalance.

Accuracy represents the ratio of correct predictions to the total number of instances and serves as a general performance indicator. However, in datasets where one class significantly dominates—such as in this study, where “Granted” decisions are far more frequent than “Rejected” ones—accuracy alone can be deceptive, as models may favor the majority class and still achieve high scores.

To address this limitation, Precision and Recall are utilized. Precision measures the proportion of correctly predicted positive instances among all predicted positives, which is essential in minimizing false positives. In contrast, Recall assesses the model’s ability to correctly identify all actual positives, particularly important in scenarios where missing a relevant instance (false negative) could be critical [29].

The F1-score, which is the harmonic mean of Precision and Recall, offers a balanced view of the model’s performance, especially in imbalanced classification problems. Meanwhile, the F2-score gives greater weight to Recall, making it more appropriate in contexts where capturing all relevant instances (e.g., identifying all truly “Rejected” cases) is more important than achieving high Precision.

To evaluate the model’s overall discriminative ability across thresholds, the AUC (Area Under Curve) metric is applied. AUC reflects the probability that the model ranks a randomly chosen positive instance higher than a randomly chosen negative one. A higher AUC value indicates better performance in distinguishing between the two classes across varying thresholds [30].

In addition to these quantitative metrics, the study also employs a feature importance analysis to evaluate the impact of each input variable on the model’s predictions. This analysis is particularly useful for tree-based models such as Random Forest, LightGBM, and XGBoost, which inherently calculate the importance of each feature based on how much it contributes to reducing classification error [31]. The result of this analysis provides insights into which factors—such as

the outcome of mediation, presence of domestic violence, economic issues, or age—are most influential in determining court verdicts. This not only aids in model interpretability but also supports transparency and accountability in decision-support systems within legal domains.

By combining classification performance metrics with feature importance analysis, this study ensures a well-rounded evaluation of the models, enabling not only accurate predictions but also explainable insights that can inform the development of decision-support tools in the legal domain.

Collectively, these metrics provide a holistic view of model performance, enabling a fair comparison of each algorithm's strengths and weaknesses, particularly in light of the dataset's class imbalance.

3. RESULTS AND DISCUSSION

This research implemented six supervised machine learning algorithms—Decision Tree, Naïve Bayes, K-Nearest Neighbors (KNN), Random Forest, LightGBM, and XGBoost—to predict verdict outcomes in divorce cases processed at the Religious Court of Padang Sidempuan. The evaluation employed multiple performance metrics to determine the most accurate and robust model for classifying court verdicts in divorce cases.

3.1. Performance Evaluation of the Models

The performance of each classification algorithm was initially evaluated using the testing set, focusing on their ability to accurately distinguish between “Granted” and “Rejected” outcomes in the context of class imbalance. The dataset used in this research was not equally distributed, with a significantly higher number of cases resulting in “Granted” decisions. This imbalance posed a challenge for classification models, as they tend to favor the majority class during training, potentially leading to misleading overall accuracy if not supported by more detailed evaluation metrics.

To ensure comprehensive assessment, each model was evaluated using Accuracy, Precision, Recall, F1-score, F2-score, and the Area Under the ROC Curve (AUC). Among the tested models, Naïve Bayes achieved the highest recall (100%) for the “Granted” class, indicating exceptional sensitivity in identifying approved divorce requests. This makes it valuable in scenarios where minimizing false negatives is critical. However, the model's lower precision suggests it is more prone to overpredicting positive cases, highlighting the classic trade-off between recall and false positive rates in imbalanced classification tasks.

On the other hand, LightGBM and XGBoost demonstrated the most balanced and stable performance across all evaluation metrics. Notably, both models achieved AUC scores that rounded to 0.79, with LightGBM scoring 0.788 and XGBoost slightly higher at 0.794. These scores indicate strong discriminatory power in distinguishing between the two classes regardless of threshold. Their ability to maintain high precision and recall simultaneously, even under class imbalance, makes them ideal candidates for real-world implementation in legal decision support systems, where both fairness and accuracy are crucial.

3.2. Comparison of Model Performance

After evaluating the performance of each algorithm individually, all six models were compared to determine which one is most suitable for predicting court decisions—particularly considering that the dataset is inherently imbalanced.

The comparative evaluation results of all algorithms are summarized in the following table..

Table 2. Model Performance Comparison on the Test Dataset

Model	Accuracy	AUC	Precision (1)	Recall (1)	F1-score (1)	F2-score (1)	Precision (0)	Recall (0)	F1-score (0)	F2-score (0)
Decision Tree	75.7%	0.721	0.909	0.777	0.838	0.801	0.409	0.667	0.507	0.592
Naïve Bayes	90.6%	0.799	0.897	1.00	0.946	0.977	1.00	0.5	0.667	0.556
KNN	71.6%	0.705	0.878	0.755	0.812	0.777	0.339	0.544	0.418	0.485
Random Forest	84.2%	0.795	0.908	0.897	0.902	0.899	0.575	0.509	0.590	0.599

Light GBM	85.7%	0.788	0.903	0.923	0.913	0.919	0.631	0.57	0.599	0.581
XG Boost	84%	0.794	0.904	0.899	0.902	0.899	0.573	0.588	0.58	0.585

Naïve Bayes achieved the highest accuracy (90.6%) and perfect recall (100%) for the "Granted" class, indicating strong sensitivity to positive outcomes. However, its precision and recall for the "Rejected" class were relatively low, which suggests overprediction of positive cases and potential risks in misclassification.

LightGBM and XGBoost offered the most balanced performance, with F1-scores for both classes remaining high. LightGBM achieved the highest F2-score (0.919) for the "Granted" class and a competitive AUC of 0.788. XGBoost followed closely with an AUC of 0.794 and consistent F1-scores across both classes.

Random Forest also performed competitively, attaining 84.2% accuracy and a strong AUC of 0.795. Although slightly behind LightGBM and XGBoost, it displayed reliable performance in capturing key patterns from the data, especially in F1 and F2 scores for the positive class.

Decision Tree and KNN lagged behind, with lower AUC values (0.721 and 0.705 respectively) and weaker F2-scores for the "Rejected" class. This indicates their limitations in handling class imbalance and capturing complex relationships in the dataset.

Despite the use of SMOTE to balance the class distribution in the training set, most models still demonstrated relatively poor performance in identifying "Rejected" cases (class 0). This is evident from the consistently lower recall and F2-scores for that class across all models. Several factors likely contributed to this outcome:

First, although SMOTE is effective in generating synthetic samples for minority classes, it does not introduce new semantic information. In this study, the "Rejected" class may contain highly diverse or underrepresented legal scenarios—such as cases resolved through reconciliation or those lacking strong evidence—that are not well captured by the available structured features or keyword-based extractions from posita texts.

Second, the effectiveness of SMOTE is algorithm-dependent. For instance, distance-based models like KNN are particularly sensitive to the altered distribution created by SMOTE, which can distort distance relationships between samples. Naïve Bayes may also suffer, as it relies on feature distributions that could be disrupted by synthetic data that do not follow original probabilistic patterns. Conversely, tree-based models like Random Forest, LightGBM, and XGBoost are generally more robust due to their ability to adapt to non-linear relationships and noisy features—though even these models were still challenged by the extreme imbalance and contextual complexity of the minority class.

In essence, SMOTE improved the balance, but not the representation richness of the minority class. Therefore, better performance on "Rejected" cases might require richer and more varied training data, domain-aware feature engineering, or alternative methods such as cost-sensitive learning or contextual embeddings.

3.3. ROC Curve Visualization of Each Model

To support the AUC values presented in the previous section, Figures 3 through 8 display the ROC curves of each machine learning model evaluated in this research. These visualizations illustrate the trade-off between true positive rates and false positive rates at various threshold settings, providing insight into each model's discriminative power.

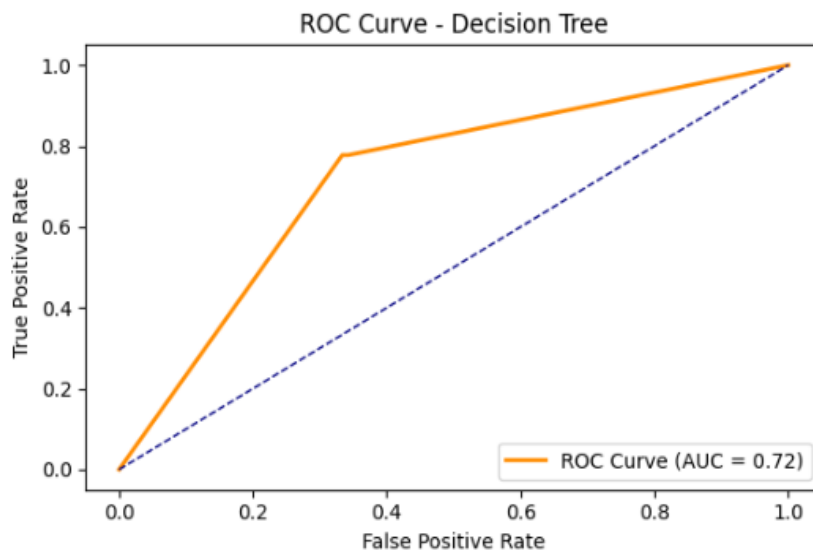


Figure 3. ROC Curve – Decision Tree ($AUC = 0.72$)

Figure 3 shows the ROC curve of the Decision Tree model, which yields an AUC value of 0.72. The curve lies relatively close to the diagonal line, suggesting that the model's ability to discriminate between “Granted” and “Rejected” outcomes is limited. While the model is able to detect general patterns, it struggles to make confident distinctions in class prediction.

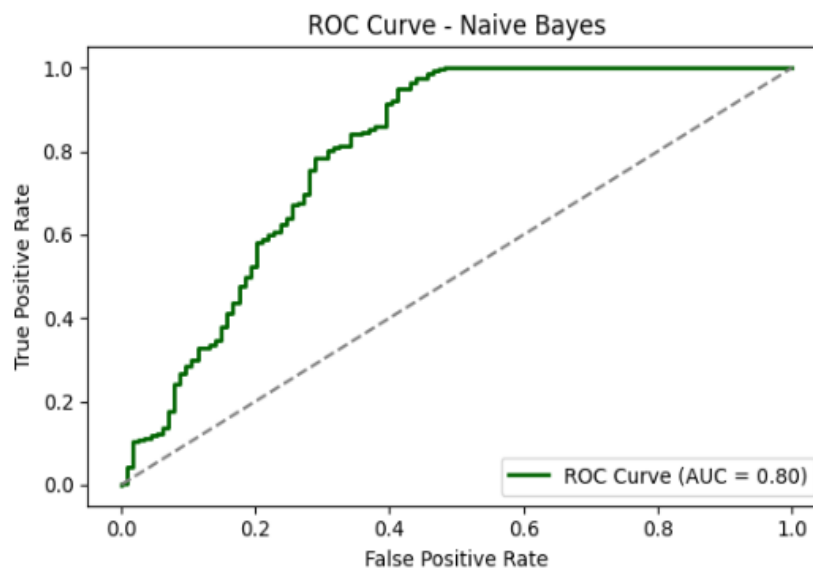


Figure 4. ROC Curve – Naïve Bayes ($AUC = 0.80$)

Figure 4 presents the ROC curve of the Naïve Bayes model, which achieves an AUC of 0.80. This curve rises steeply and levels off at the top, indicating excellent sensitivity for identifying true positives, particularly in the “Granted” class. Despite its simple assumptions, the model performs competitively and offers a strong trade-off between true and false positive rates.

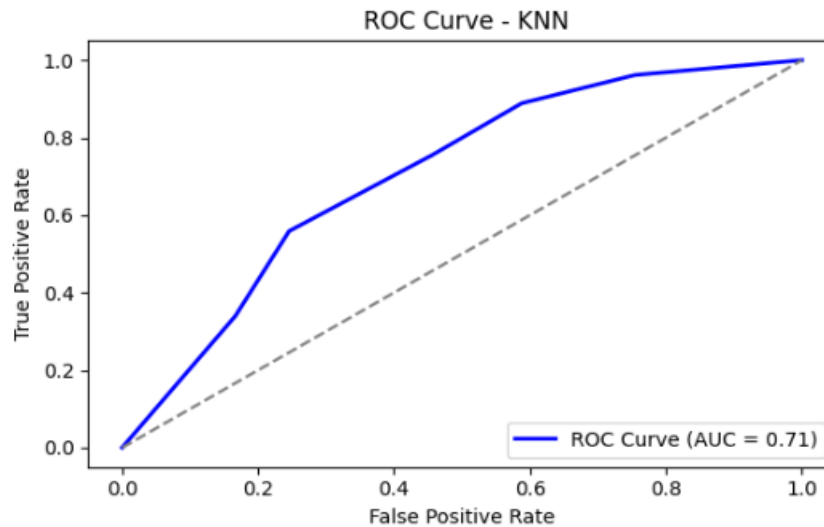


Figure 5. ROC Curve – K-Nearest Neighbor ($AUC = 0.71$)

Figure 5 illustrates the ROC curve for the K-Nearest Neighbors (KNN) model, with an AUC of 0.71. The curve is modest in steepness and remains relatively close to the diagonal. This suggests that KNN has difficulties in clearly separating the two classes, potentially due to sensitivity to data distribution and imbalance.

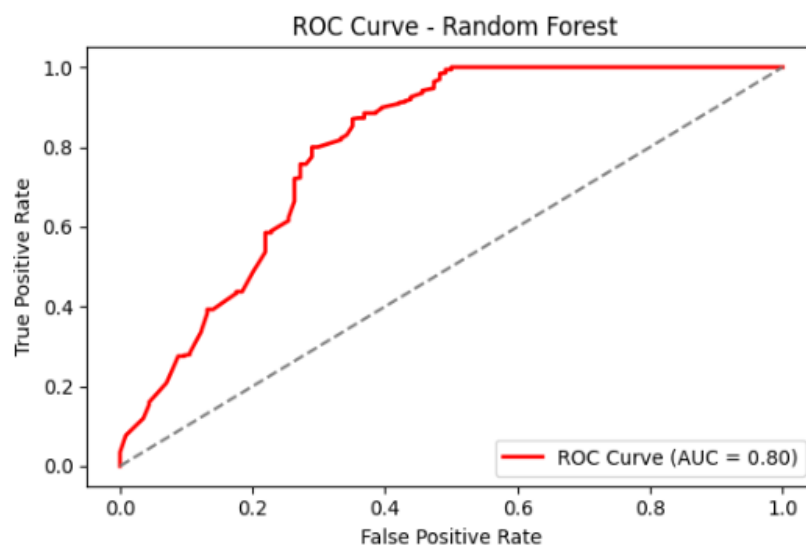


Figure 6. ROC Curve – Random Forest ($AUC = 0.80$)

Figure 6 shows the ROC curve of the Random Forest model, which also achieves an AUC of 0.80. The curve demonstrates a steep rise and a long plateau at the top, indicating strong predictive capability. As an ensemble of decision trees, Random Forest manages to reduce overfitting and capture meaningful relationships between features and outcomes.

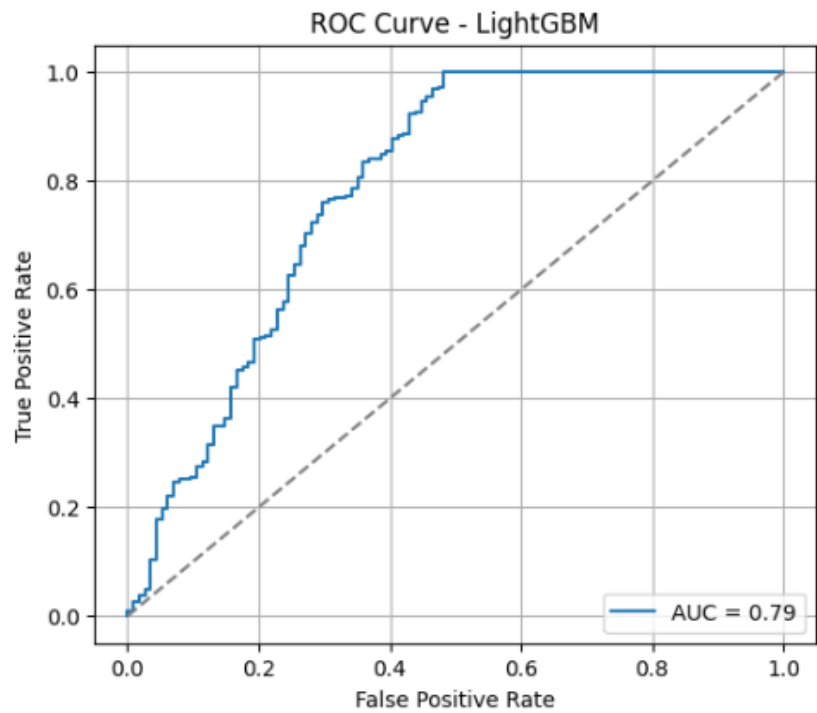


Figure 7. ROC Curve – LightGBM ($AUC = 0.79$)

Figure 7 displays the ROC curve for LightGBM, which attains an AUC of 0.79. The curve shows a strong convex shape, reflecting the model's effectiveness in distinguishing between the outcome classes. LightGBM's gradient boosting structure allows it to efficiently handle large-scale structured data and perform well under imbalance.

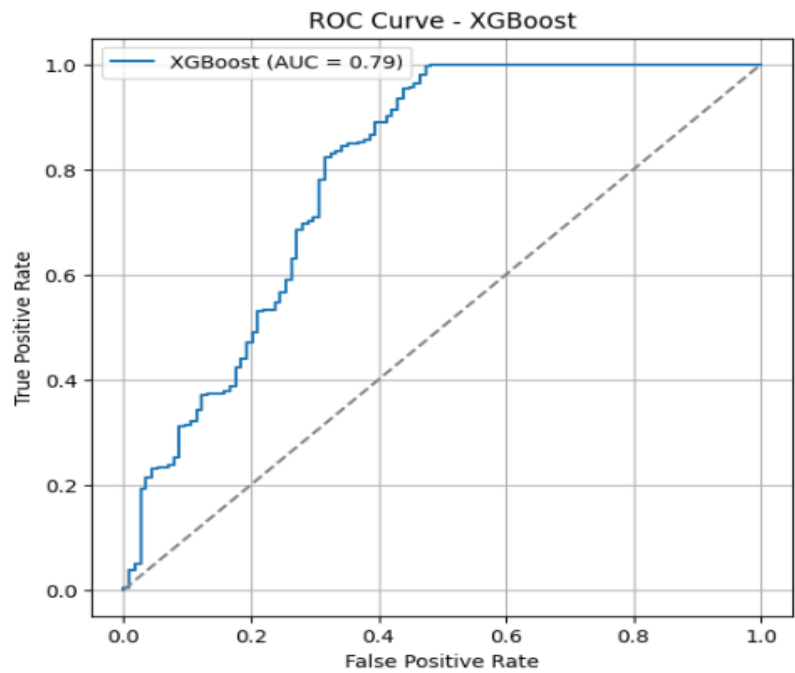


Figure 8. ROC Curve – XGBoost ($AUC = 0.79$)

Figure 8 illustrates the ROC curve for XGBoost, also with an AUC of 0.79. The curve is steep and consistently rises above the baseline, confirming its high capability in class separation. XGBoost offers regularization and scalability, making it suitable for structured datasets like the legal case records used in this research.

A comparative review of the ROC curves reveals that gradient boosting algorithms (LightGBM and XGBoost) consistently exhibit strong classification performance, with each achieving high AUC values of 0.79 and showing steep, convex ROC curves that reflect excellent class separability. Random Forest and Naïve Bayes also demonstrate reliable performance, each obtaining an AUC score of 0.80. Random Forest benefits from its ensemble architecture that captures complex patterns, while Naïve Bayes, despite its simplicity and assumptions of feature independence, proves particularly effective in identifying true positives. In contrast, Decision Tree and KNN show relatively weaker performance, with ROC curves closer to the diagonal line and lower AUC scores of 0.72 and 0.71, respectively. These models exhibit limitations in clearly distinguishing between the "Granted" and "Rejected" classes based on the underlying data.

3.4. Feature importance Analysis

To gain a deeper understanding of how machine learning models make predictions, a feature importance analysis was conducted for each algorithm. This analysis identifies the features that contribute most significantly to classification outcomes, providing insights into how each model learns patterns from the data.

In the Decision Tree model, the most influential feature was `hasil_mediasi` (mediation outcome), followed by `usia_t` (age of the defendant), `usia_p` (age of the plaintiff), and `usia_pernikahan` (length of marriage). These four features consistently appeared near the top levels of the decision tree, indicating their role as primary variables in classification. Legally, the outcome of mediation serves as an initial indicator of whether the dispute can be resolved amicably, making it a crucial factor in judicial decisions.

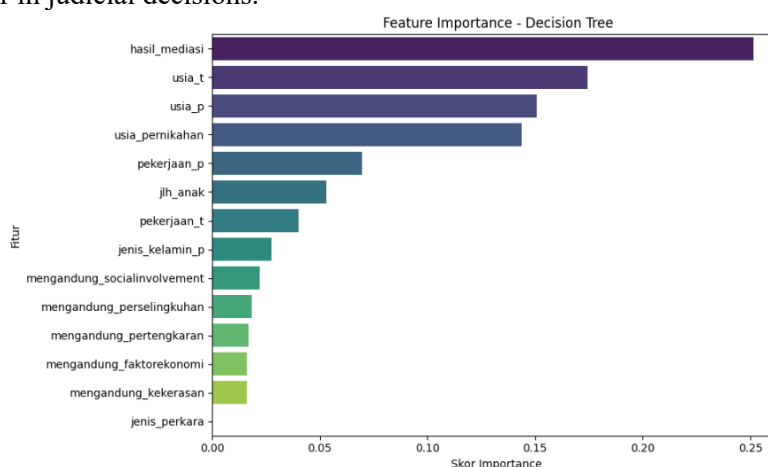


Figure 9. Feature Importance of Decision Tree

The Naïve Bayes model, although it does not present numerical importance in the same way as tree-based models, showed a strong reliance on binary features such as `hasil_mediasi`, `mengandung_kekerasan` (presence of domestic violence), and `jenis_perkara` (case type). Due to its sensitivity to explicit feature distributions, this algorithm is capable of quickly identifying simple yet significant patterns. This aligns with real-world observations, where cases involving domestic violence or divorce petitions filed by wives (*cerai gugat*) often result in "Granted" verdicts.

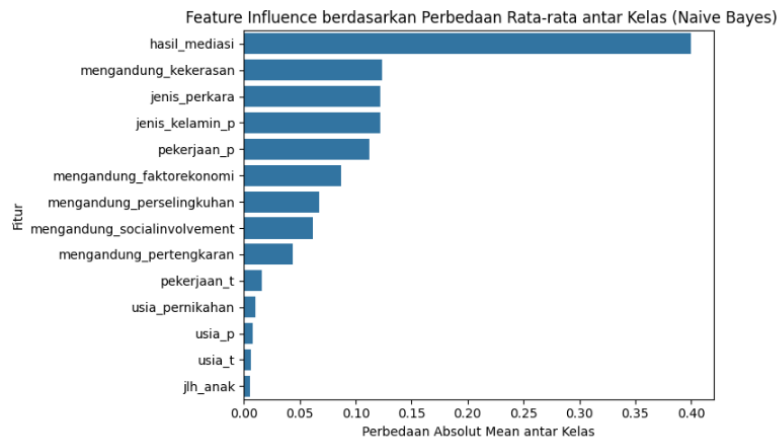


Figure 10. Feature Influence of Naïve Bayes

In the K-Nearest Neighbors (KNN) model, numerical features such as age and mediation outcome were major contributors, as classification is based on data proximity in feature space. The ages of both litigants and the length of marriage served as strong distance indicators for measuring case similarity. However, the model's sensitivity to feature scaling and class imbalance makes it less reliable for complex legal datasets.

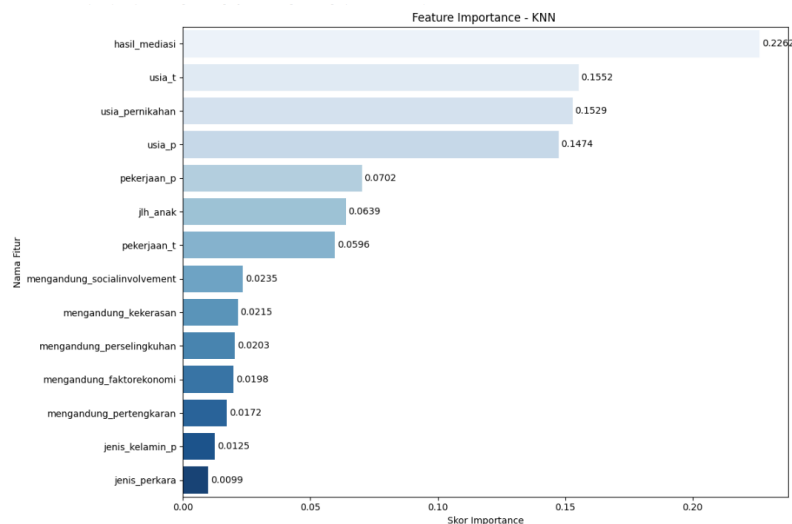


Figure 11. Feature Importance of K-Nearest Neighbors

The Random Forest model, as an ensemble of multiple decision trees, highlighted `hasil_mediasi`, `usia_t`, and `usia_pernikahan` as the most frequently used features in tree splits. It also assigned significant weight to `mengandung_ekonomi` (economic hardship) and `mengandung_kekerasan` (presence of domestic violence), indicating the model's ability to capture non-linear interactions—for instance, correlations between younger age and susceptibility to economic conflict.

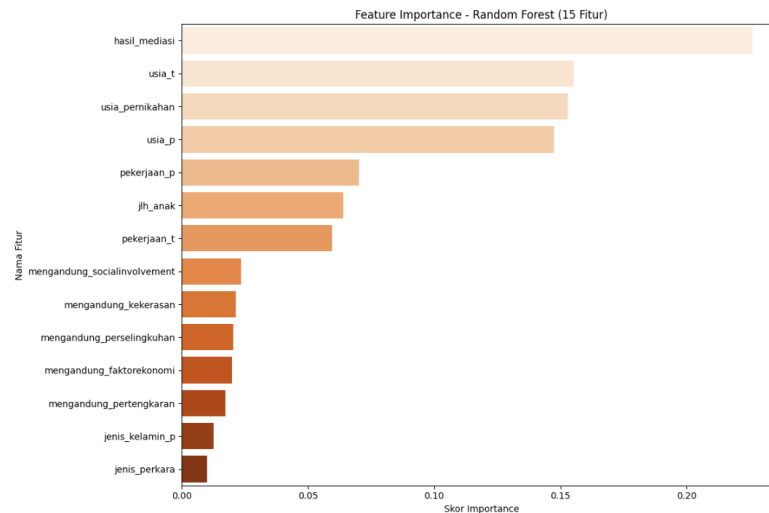


Figure 12. Feature Importance of Random Forest

Meanwhile, the LightGBM model identified *usia_pernikahan* (length of marriage) as the most important feature, followed by *usia_t* (age of the defendant/respondent), *usia_p* (age of the plaintiff/petitioner), and *jenis_pekerjaan* (occupation). Although *hasil_mediasi* (mediation outcome) ranked slightly lower, it remained a meaningful splitting feature in several boosting structures. LightGBM's leaf-wise growth strategy enables sharper data segmentation, particularly when key features exhibit skewed distributions.

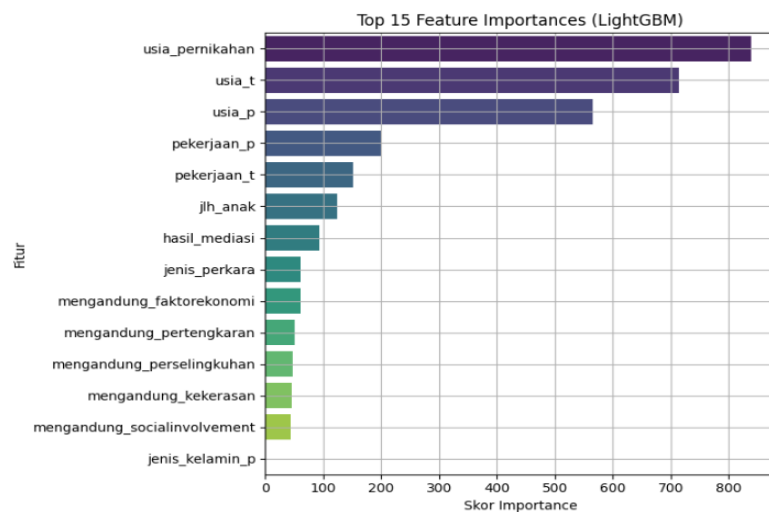


Figure 13. Feature Importance of LightGBM

The XGBoost model emphasized *usia_pernikahan* (length of marriage) and *mengandung_kekerasan* (presence of domestic violence) as the most impactful variables. Other relevant features such as *mengandung_ekonomi* (presence of economic hardship) and *mengandung_perselingkuhan* (presence of infidelity) also contributed significantly to classification accuracy. With its strength in modeling complex feature interactions, XGBoost effectively utilized both structured and textual posita-based features to represent the conditions of household conflict more precisely.



Figure 14. Feature Importance of XGBoost

These results confirm that while each algorithm varies in how it evaluates features, common determinants such as mediation outcome, domestic violence, and infidelity consistently appear across models. These attributes reflect the legal and social dimensions that influence divorce verdicts. Essentially, the trained models successfully captured the pattern that when a divorce case involves domestic violence and infidelity, it is more likely to fail in the mediation process, and consequently, the case proceeds to a final ruling—typically being granted. On the other hand, if mediation is successful, it increases the likelihood that the case will be rejected. This aligns with legal perspectives, where a successful mediation (reconciliation) is considered a valid reason to deem the continuation of divorce proceedings unnecessary. Such findings are consistent with previous studies that demonstrated the predictive power of machine learning models in capturing legal decision patterns.

These findings partially align with previous studies. For instance, Sohail et al. [6] identified economic hardship and domestic conflict as influential factors, which are also reflected in this study's feature importance. Similarly, Medvedeva et al. [7] demonstrated the effectiveness of ensemble methods—supporting the strong performance of Random Forest, LightGBM, and XGBoost in this research. However, unlike Aimran et al. [8] who found Decision Tree most effective, this study shows that gradient boosting methods yield more balanced results, especially under class imbalance conditions. Unlike Bramantoro et al. [9], who focused only on pandemic-related factors, this study offers broader generalizability based on non-pandemic data from multiple years.

3.5. Study Limitations and Future Work

Despite the promising results, this study presents several limitations that should be considered in future research. First, the dataset was sourced solely from the Religious Court of Padang Sidempuan. While this single-jurisdiction case study allows for deep contextual analysis, it limits the generalizability of the model to other regions with different demographic or procedural characteristics. Expanding the dataset to include cases from multiple Religious Courts across Indonesia would enable cross-jurisdictional validation and improve the robustness of the findings.

Second, although SMOTE was employed to address class imbalance, the performance on the minority class ("Rejected") remains relatively low across all models. This indicates the need for more sophisticated imbalance handling techniques, such as cost-sensitive learning or adaptive resampling, to improve recall without sacrificing overall precision.

Third, this study did not apply deep learning models such as LSTM or Legal-BERT, as the focus was on classical machine learning algorithms that are easier to interpret and explain—important considerations in legal contexts. Additionally, deep learning models often require large

volumes of data and computational resources, which were beyond the scope of this research. Future studies are encouraged to explore these advanced methods to capture deeper semantic representations in judicial texts.

Fourth, no statistical significance testing was performed to confirm whether the observed performance differences between models were meaningful. Consequently, while metrics like accuracy, recall, and AUC suggest that certain algorithms outperform others, these findings should be interpreted with caution. Incorporating tests such as McNemar's test or bootstrap-based confidence intervals in future work would strengthen the reliability of model comparisons.

Finally, future research could benefit from integrating legal reasoning components, hybrid rule-based methods, or domain-specific embeddings to enhance model explainability and alignment with judicial logic. Such advancements will be crucial in developing decision support systems that are not only technically sound but also legally and ethically robust.

4. CONCLUSION

This research has successfully addressed three main research questions concerning the development of a predictive model for divorce verdicts at the Religious Court of Padang Sidempuan using machine learning algorithms:

1. The prediction model was built using 2,026 divorce cases from 2018 to 2025. The preprocessing phase involved cleaning the posita texts, extracting divorce reasons using keyword mapping, normalizing age and occupation variables, label encoding, and splitting the dataset into training and testing sets (70:30 ratio). To address class imbalance, the SMOTE technique was applied to the training data before training and testing six machine learning algorithms: Decision Tree, Naïve Bayes, K-Nearest Neighbors, Random Forest, LightGBM, and XGBoost.
2. The evaluation of these algorithms revealed that Naïve Bayes achieved the highest accuracy (90.6%) and perfect recall (100%) for the "Granted" class. This performance was driven by the probabilistic nature of Naïve Bayes, which responded well to dominant features such as violence and economic hardship. However, the model exhibited low performance in identifying "Rejected" cases, indicating bias toward the majority class.
3. LightGBM and XGBoost demonstrated the most stable and balanced results by maintaining high precision and recall for both classes and achieving the highest AUC values (both 0.79). The ROC Curve visualization confirmed their discriminative ability and generalization power. This comparison fulfills the third research aim by identifying LightGBM and XGBoost as the most suitable algorithms for court decision support due to their balanced performance across classes and metrics.

The findings confirm that the developed models not only provide accurate predictions but also offer insights into the critical factors influencing judicial outcomes in divorce cases. This research demonstrates the effectiveness of machine learning in uncovering patterns in legal data and lays the groundwork for building objective, efficient, and transparent decision-support systems within religious courts—while still respecting the independence of judicial authority.

ACKNOWLEDGEMENTS

The author gratefully acknowledges the guidance of the academic supervisor and the support of the Religious Court of Padang Sidempuan for providing access to data and insights that contributed to the completion of this research.

REFERENCES

- [1] Badan Pusat Statistik, *Statistik Indonesia 2023*. Jakarta: BPS, 2023. [Online]. Available: <https://www.bps.go.id>
- [2] Undang-Undang Republik Indonesia Nomor 3 Tahun 2006 tentang Perubahan atas Undang-Undang Nomor 7 Tahun 1989 tentang Peradilan Agama.
- [3] S. Nahavandi, "Industry 5.0—a human-centric solution," *Sustainability*, vol. 11, no. 16, p. 4371, 2019. doi: 10.3390/su11164371

- [4] M. García, J. Rodriguez, A. Sanchez, and L. Torres, "Big data and predictive analytics: A systematic review of applications," *Artificial Intelligence Review*, vol. 57, no. 4, pp. 201–252, 2024, doi: 10.1007/s10462-024-10811-5
- [5] H. A. Almuzaini and A. M. Azmi, "TaSbeeb: A judicial decision support system based on deep learning framework," *Journal of King Saud University – Computer and Information Sciences*, vol. 35, no. 8, p. 101695, 2023, doi: 10.1016/j.jksuci.2023.10169
- [6] A. Sohail, H. Khalid, F. Khalid, dan M. Akram, "Implementation of machine learning algorithm on factors affecting divorce rate," *International Journal of Scientific & Engineering Research*, vol. 9, no. 5, pp. 602–607, 2018
- [7] M. Medvedeva, M. Vols, dan M. Wieling, "Using machine learning to predict decisions of the European Court of Human Rights," *Artificial Intelligence and Law*, vol. 28, no. 2, pp. 237–266, 2020, doi: 10.1007/s10506-019-09255-y
- [8] N. Aimran, R. A. Othman, and A. S. Ahmad, "Prediction of Malaysian women divorce using machine learning techniques," *Malaysian Journal of Computing*, vol. 7, no. 2, pp. 491–502, 2022, doi: 10.24191/mjoc.v7i2.17077.
- [9] A. Bramantoro dan I. Virdyna, "Classification of divorce causes during the COVID-19 pandemic using convolutional neural networks," *PeerJ Computer Science*, vol. 8, p. e998, 2022, doi: 10.7717/peerj-cs.998
- [10] A. B. Dina, R. Sarno, R. N. E. Anggraini, A. T. Haryono, and A. F. Septiyanto, "Comparison of oversampling techniques in prediction judicial decisions of divorce trials in family courts," in *Proc. 2024 Int. Conf. on Information Technology Research and Innovation (ICITRI)*, Surabaya, Indonesia, 2024, pp. 13–18, doi: 10.1109/ICITRI62858.2024.10699016.
- [11] A. S. Ma'arif and D. H. Firmansyah, "Fulfilling children's rights through post-divorce court decisions: A case study of religious court verdicts in Indonesia," *Ahwal: Jurnal Hukum Keluarga Islam*, vol. 16, no. 1, pp. 123–140, 2023, doi: 10.14421/ahwal.2023.16108
- [12] D. Kusnandar and F. Rahma, "Optimizing Legal Protection for Divorce Outside of Court: Study of the Need for Divorce Isbat in the Indonesian Legal System," *Indonesian Journal of Islamic Law*, vol. 6, no. 2, pp. 73–88, Jun. 2023, doi: 10.35719/ijil.v6i1.
- [13] M. Mujahid, Y. T. Shah, and M. Khan, "Machine learning with class imbalance: A review," *Journal of Big Data*, vol. 11, no. 1, art. no. 71, 2024, doi: 10.1186/s40537-024-00943-4
- [14] T. Subrata, "Title and judiciary power based on Law Number 48 of 2009," **Legal Brief**, vol. 11, no. 5, pp. 2875–2881, 2022, doi: 10.35335/legal.
- [15] Direktorat Jenderal Badan Peradilan Agama, "Sistem Informasi Penelusuran Perkara (SIPP)," Mahkamah Agung Republik Indonesia, [Online]. Available: <https://sipp.mahkamahagung.go.id>
- [16] S. Matharaarachchi, M. Domaratzki, and S. Muthukumarana, "Enhancing SMOTE for imbalanced data with abnormal minority instances," *Machine Learning with Applications*, vol. 19, art. no. 100597, 2024, doi: 10.1016/j.mlwa.2024.100597
- [17] M. Owusu-Adjei, J. Ben Hayfron-Acquah, T. Frimpong, and G. Abdul-Salaam, "Imbalanced class distribution and performance evaluation metrics: A systematic review of prediction accuracy for determining model performance in healthcare systems," *PLOS Digital Health*, vol. 2, no. 11, p. e0000290, Nov. 2023, doi: 10.1371/journal.pdig.0000290
- [18] M. Saarela and S. Jauhiainen, "Comparison of feature importance measures as explanations for classification models," *SN Applied Sciences*, vol. 3, no. 6, p. 414, 2021, doi: 10.1007/s42452-021-04148-9
- [19] Y. Zhang, C. Li, Y. Sheng, J. Ge, and B. Luo, "Judicial intelligent assistant system: Extracting events from Chinese divorce cases to detect disputes for the judge," *Expert Systems*, vol. 41, no. 7, e13540, Jan. 2024, doi: 10.1111/exsy.13540
- [20] M. Alshamrani, M. Almaiah, and S. Al-Khalifa, "A machine learning-based approach to predict user satisfaction in mobile government services," *PeerJ Computer Science*, vol. 9, e2131, 2023, doi: 10.7717/peerj-cs.2131

- [21] M. Altalhan, A. Algarni, and M. Turki-Hadj Alouane, "Imbalanced data problem in machine learning: A review," *IEEE Access*, vol. 13, pp. 13398–13424, Jan. 2025, doi: 10.1109/ACCESS.2025.3531662
- [22] I. Domormienye and N. Jere, "A survey of decision trees: Concepts, algorithms, and applications," *IEEE Access*, vol. 12, pp. 86716–86727, 2024, doi: 10.1109/ACCESS.2024.3416838.
- [23] S. S. Prasetyowati and Y. Sibaroni, "Unlocking the potential of Naive Bayes for spatio temporal classification: a novel approach to feature expansion," *Journal of Big Data*, vol. 11, no. 106, 2024, doi: 10.1186/s40537-024-00958-x
- [24] R. K. Halder, M. N. Uddin, M. A. Uddin, S. Aryal, and A. Khraisat, "Enhancing K-nearest neighbor algorithm: a comprehensive review and performance analysis of modifications," *Journal of Big Data*, vol. 11, no. 113, 2024. doi: 10.1186/s40537-024-00973-y
- [25] L. Barreñada, P. Dhiman, D. Timmerman, A.-L. Boulesteix, and B. Van Calster, "Understanding overfitting in random forest for probability estimation: a visualization and simulation study," *Diagnostic and Prognostic Research*, vol. 8, no. 14, 2024, doi: 10.1186/s41512-024-00177-1
- [26] T. O. Ometehinwa, D. O. Oyewola, and E. G. Mounq, "Optimizing the light gradient-boosting machine algorithm for an efficient early detection of coronary heart disease," *Informatics and Health*, vol. 1, p. 70-81, 2024, doi: 10.1016/j.infoh.2024.06.001
- [27] J. Jinbo, L. Yufu, dan M. Haitao, "Handling missing data of using the XGBoost-based multiple imputation by chained equations regression method," *Frontiers in Artificial Intelligence.*, vol. 8, p. 1553220, Apr. 2025, doi: 10.3389/frai.2025.1553220
- [28] X. He, K. Zhao, and X. Chu, "AutoML: A survey of the state-of-the-art," *Knowledge-Based Systems*, vol. 212, p. 106622, 2021, doi:10.1016/j.knosys.2020.106622
- [29] G. M. Foody, "Challenges in the real world use of classification accuracy metrics: From recall and precision to the Matthews correlation coefficient," *PLOS ONE*, 2023, doi: 10.1371/journal.pone.0291908
- [30] J. Li, "Area under the ROC Curve has the most consistent evaluation for binary classification," *PLOS ONE*, 2023, doi: 10.1371/journal.pone.0316019
- [31] A. Kumar, and J. W. Taylor, "Feature importance in the age of explainable AI: Case study of detecting fake news & misinformation via a multi-modal framework," *European Journal of Operational Research*, vol. 317, pp. 401–413, 2023, doi: 10.1016/j.ejor.2023.10.003