JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol XII No 2, Maret 2026, hlm. 263 –270 ISSN 2550-0201 (Online)
DOI: https://doi.org/10.33330/jurteksi.v12i2.4432
Available online at https://jurnal.stmikroyal.ac.id/index.php/jurteksi

RANDOM FOREST BASED SYSTEM FOR PREDICTING AND RECOMMENDING INMATE REHABILITATION PROGRAMS

Syahrul Farhan[1*], Nurul Rahmadani[1], Mardalius[1]
[1]Information System, Universitas Royal
email: *sahrulfarhan052@gmail.com

Abstract: Rehabilitation programs are essential in correctional systems to equip inmates with the skills and behavioral readiness required for social reintegration. However, rehabilitation program assignment in many correctional institutions remains dependent on manual and subjective assessments, which may result in inconsistent decisions. This study develops a Random Forest–based prediction system to support objective and data-driven rehabilitation program determination. A quantitative approach was applied using historical inmate data from January 2023 to January 2025, comprising 2,023 records. The research process included data preprocessing, an 80:20 training–testing split, model training, and performance evaluation using accuracy, precision, recall, and F1-score metrics. The results show that the model achieved an accuracy of 86.17% during training in Google Colab and 68.83% when deployed within the application system. This performance gap reflects real-world deployment and computational constraints rather than model failure. The proposed system provides consistent and objective rehabilitation program recommendations, thereby supporting more effective rehabilitation planning and decision-making in correctional institutions.

Keywords: correctional institutions; inmate rehabilitation programs; machine learning; random forest; prediction system

Abstrak: Program pembinaan narapidana memiliki peran penting dalam sistem pemasyarakatan untuk membekali warga binaan dengan keterampilan serta kesiapan perilaku dalam proses reintegrasi ke masyarakat. Namun, pada banyak lembaga pemasyarakatan, penentuan program pembinaan masih bergantung pada penilaian manual yang bersifat subjektif, sehingga berpotensi menimbulkan ketidakkonsistenan dalam pengambilan keputusan. Penelitian ini mengembangkan sistem prediksi program pembinaan narapidana berbasis algoritma Random Forest guna mendukung pengambilan keputusan yang objektif dan berbasis data. Pendekatan kuantitatif diterapkan menggunakan data historis narapidana periode Januari 2023 hingga Januari 2025 sebanyak 2.023 data. Tahapan penelitian meliputi prapemrosesan data, pembagian data latih dan uji dengan rasio 80:20, pelatihan model, serta evaluasi performa menggunakan metrik akurasi, precision, recall, dan F1-score. Hasil penelitian menunjukkan bahwa model mencapai akurasi sebesar 86,17% pada tahap pelatihan di Google Colab dan 68,83% saat diimplementasikan pada sistem aplikasi. Perbedaan performa tersebut mencerminkan keterbatasan lingkungan operasional, bukan kegagalan model. Secara keseluruhan, sistem yang dikembangkan mampu memberikan rekomendasi program pembinaan yang lebih objektif dan konsisten, sehingga mendukung perencanaan pembinaan yang lebih efektif.

Kata kunci: mesin pembelajaran; program pembinaan narapidana; random Forest; sistem pemasyarakatan; sistem prediksi

INTRODUCTION

Inmate rehabilitation programs constitute an essential component of correctional systems, as they are intended to develop inmates' skills, behavioral competencies, and social preparedness prior to their reintegration into society. The effectiveness of rehabilitation initiatives is not determined solely by their availability, but also by the accuracy of assigning appropriate rehabilitation programs that correspond to each inmate's individual profile and rehabilitation needs. Inappropriate program placement may disrupt the rehabilitation process and reduce the achievement of broader correctional objectives [1].

In many correctional facilities, including the Correctional Institution Class IIB Tanjungbalai which serves as the object of this research, various types of rehabilitation programs are implemented, such as personality development programs, self-reliance training, vocational skill development, and social or religious guidance activities. The determination of which rehabilitation program should be assigned to an inmate is still predominantly conducted through manual assessment based on officers' observations and subjective judgment. This process is often affected by limited time, differences in officers' experience, and the high number of inmates under supervision, leading to inconsistent decisions, particularly among inmates with similar characteristics.

Alongside these challenges, correctional institutions now maintain comprehensive inmate databases containing diverse information, including demographic attributes, criminal offense types, sentence duration, behavioral records during incarceration, disciplinary history, and participation in previous rehabilitation programs. These data attributes represent important features that can be utilized to predict suitable rehabilitation program assignments. However, in practice, such data are primarily used for administrative purposes and have not been optimally exploited to support structured and predictive decision-making.

Recent advancements in machine learning provide effective approaches to address these limitations by enabling the analysis of large and complex datasets. By utilizing historical inmate data and relevant feature variables, machine learning techniques can identify hidden patterns and relationships that are difficult to detect through conventional manual evaluation. As a result, predictive systems can be developed to generate rehabilitation program recommendations that are more objective, consistent, and reproducible [2].

Previous studies have applied classification algorithms such as Naive Bayes and Support Vector Machine (SVM) in decision support and classification tasks [3]. While these methods have demonstrated satisfactory performance, they often rely on specific data distribution assumptions or are sensitive to parameter tuning, which may limit their robustness when applied to heterogeneous real-world datasets. In contrast, the Random Forest algorithm employs an ensemble-based approach by integrating multiple decision trees, allowing it to model complex interactions among features while maintaining stability and reducing the risk of overfitting. Moreover, Random Forest can effectively handle both numerical and categorical variables, making it particularly suitable for inmate data with diverse characteristics.

Although Random Forest has been widely used in various predictive applications, its implementation for predicting inmate rehabilitation programs especially in application systems designed for direct use by correctional officers remains relatively limited. This indicates a research gap that warrants further investigation.

Therefore, this study develops a Random Forest-based prediction system for inmate rehabilitation program assignment at the Correctional Institution Class IIB Tanjungbalai. The proposed system aims to predict the most appropriate type of rehabilitation program based on inmate data features, thereby supporting correctional officers in decision-making processes and improving the effectiveness and accuracy of rehabilitation program implementation within correctional institutions.

METHOD

This research adopts a quantitative approach by applying machine learning techniques to develop a predictive system for inmate rehabilitation program assignment [4]. The methodological design is structured to ensure reproducibility, encompassing sequential stages starting from data acquisition and preparation, data preprocessing, model construction, performance evaluation, and concluding with model deployment within an application environment.

Research Workflow

The overall research process consists of several interconnected phases, including data collection, preprocessing, dataset division, Random Forest model development, performance assessment, and implementation of the prediction system within an application framework. This structured workflow is intended to ensure systematic model development and reliable evaluation results.

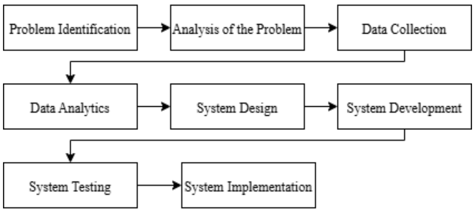


Image 1. Research Workflow Diagram

Dataset and Data Collection

The dataset utilized in this study was obtained from a correctional institution and spans a two-year period from January 2023 to January 2025. A total of 2,023 inmate records were included, representing individual characteristics as well as historical participation in rehabilitation programs. The dataset contains multiple attributes related to inmates, such as demographic information, type of criminal offense, sentence duration, behavioral records during incarceration, and the rehabilitation program category assigned or followed by the inmate.

Table 1. Description of Inmate Data Attributes

No	Attribute	Data Type	Descriptions
1	Inmate code	Categorical	Unique code used to identify each inmate
2	Age	Numerical	Age of the inmate in years
3	Gender	Categorical	Gender of the inmate
4	Education level	Categorical	Highest education level attained by the inmate
5	Type of criminal offense	Categorical	Type of criminal offense committed by the inmate
6	Sentence length	Numerical	Length of sentence in months
7	Number of violations	Numerical	Number of offenses committed by the inmate

No	Attribute	Data Type	Descriptions
	during incarceration		during the incarceration period
8	Rehabilitation program	Categorical	Type of rehabilitation program attended or recommended for the inmate

These attributes were selected because they reflect real operational conditions in the rehabilitation process and provide relevant information for constructing a predictive model based on historical inmate data.

Data Preprocessing and Data Partioning

Prior to model training, data preprocessing was conducted to ensure data quality and suitability for analysis. This stage involved identifying and handling missing values, eliminating duplicate entries, and standardizing data formats. Categorical variables were transformed into numerical representations using appropriate encoding techniques to enable processing by the Random Forest algorithm.

These preprocessing steps were designed to minimize noise and improve model performance. Data normalization was not applied, as Random Forest is not sensitive to feature scale differences. Once preprocessing was completed, the dataset was divided into training and testing subsets. An 80:20 split ratio was applied, where 80% of the data was used for model training and the remaining 20% for testing. The training subset was utilized to build the predictive model, while the testing subset served to evaluate its generalization capability on unseen data. This ratio is commonly adopted in predictive modeling to maintain a balance between learning and evaluation.

Random Forest Model and Hyperparameter Settings

The predictive model in this study was developed using the Random Forest Classifier, an ensemble learning algorithm that constructs multiple decision trees using bootstrap sampling and aggregates their outputs through majority voting [5]. Each decision tree generates a classification result, and the final model prediction is determined based on the class receiving the highest number of votes.

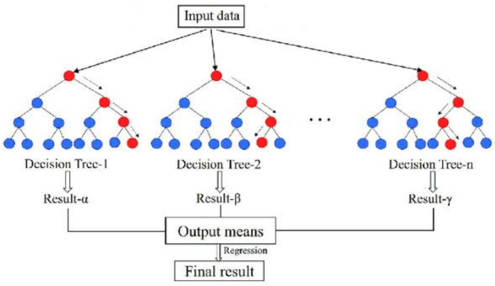


Image 2. Random Forest Model Architecture

To ensure model stability and optimal performance, several hyperparameters were defined prior to training. The number of decision trees (n_estimators) was set to 100 to balance predictive accuracy and computational efficiency. The maximum tree depth (max_depth) was limited to prevent excessive model complexity and reduce the risk of overfitting. Additionally, the minimum number of samples required to split an internal node (min_samples_split) and the minimum number of samples required at a leaf node (min_samples_leaf) were specified to control tree growth. Feature selection at each split was determined using the square root of the total number of features (max_features = sqrt), which is commonly used in classification tasks.

$$\hat{y}=\text{mode}\{h_1(x),h_2(x),\dots,h_T(x)\} \quad (1)$$

DOI: https://doi.org/10.33330/jurteksi.v12i2.4432
Available online at https://jurnal.stmikroyal.ac.id/index.php/jurteksi

The Random Forest algorithm was selected due to its ability to handle both numerical and categorical attributes, its robustness against overfitting, and its effectiveness in modeling complex relationships within heterogeneous inmate data [6].

$$\mathrm{Gini(D)}=1-\sum_{i=1}^{C} p_i^2 \quad (2)$$

During tree construction, attribute selection at each node was guided by the Gini impurity criterion, which measures node impurity and helps identify optimal splits [7]. Lower Gini values indicate higher data homogeneity within a node, leading to more informative decision tree structures.

Model Evaluation

The performance of the model was assessed using multiple metrics, including accuracy, precision, recall, F1-score, and the confusion matrix [8]. Employing a variety of evaluation metrics aims to provide a more comprehensive understanding of the model's performance, particularly in cases where the class distribution is imbalanced. The model's accuracy is represented by Equation (3), which measures the extent to which the model's predictions match the actual data.

$$\mathrm{Accuracy}=\frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

Precision is measured using Equation (4), which indicates the proportion of positive predictions made by the model that are correctly classified as actual positive cases.

$$\mathrm{Precision}=\frac{TP}{TP+FP} \quad (4)$$

Recall is calculated using Equation (5), which represents the proportion of actual positive cases that are correctly identified by the model.

$$\mathrm{Recall}=\frac{TP}{TP+FN} \quad (5)$$

The F1-score is calculated using Equation (6), which represents the harmonic mean of precision and recall. This metric provides a balance between the model's ability to avoid false positives and its capacity to correctly identify actual positive cases.

$$\mathrm{F1\text{-}Score}=2\times\frac{\mathrm{Precision}\times\mathrm{Recall}}{\mathrm{Precision}+\mathrm{Recall}} \quad (6)$$

RESULT AND DISCUSSION

Data Input

This research utilized a dataset of 2023 inmate records collected between January 2023 and January 2025. The dataset includes various attributes representing inmate characteristics, such as demographic information, criminal history, and participation in rehabilitation programs. The combination of numerical and categorical attributes enables the model to learn complex patterns for providing appropriate recommendations for inmate rehabilitation programs.

Table 2. Sample Initial Data

No.	Inmate Code	Age	Gender	Education Level N	Type Of Criminal Offense	Sentence Length	Number Of Violations During Incarceration	Types Of Crimes	Rehabilitation Program
1.	NAPI0001	33	L	D3	Violence	23	1	Personality	Psychological Counseling
2.	NAPI0002	24	L	SMP	Narcotics	33	0	Narcotics Rehabilitation	Medical Rehabilitation
...									
2003	NAPI2023	45	P	D3	Narcotics	20	1	Narcotics Rehabilitation	Medical Rehabilitation

From the data snippet, it can be observed that the dataset contains a diverse range of values and data types. This characteristic aligns with previous studies that utilized heterogeneous data in developing Random Forest-based predictive systems.

Data Pre-processing

Data preprocessing was performed to prepare the dataset for model construction by handling missing values, removing duplicate records, and standardizing attribute formats. Categorical variables were encoded into numerical values to ensure compatibility with the Random Forest algorithm.

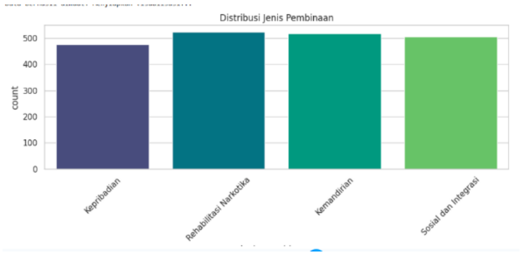


Image 3. Data Distribution Chart

An analysis of class distribution indicated that the rehabilitation program classes were relatively balanced, so resampling methods such as SMOTE were not required. These preprocessing steps were applied to improve data quality and minimize potential errors during model training

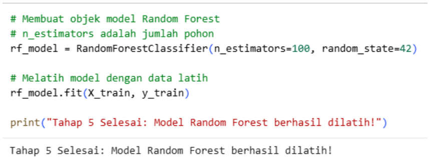
Data Splitting

Once the data was cleaned and ready for use, the dataset was split into two subsets: 80% for training and 20% for testing. This partitioning aims to assess the model's ability to generalize to new, unseen data.

Table 3. Data Partition

Dataset	Total Records
Training Data	1,618
Testing Data	405
Total	2,023

Model Training



```
# Membuat objek model Random Forest
# n_estimators adalah jumlah pohon
rf_model = RandomForestClassifier(n_estimators=100, random_state=42)

# Melatih model dengan data latih
rf_model.fit(X_train, y_train)

print("Tahap 5 Selesai: Model Random Forest berhasil dilatih!")
```

```
Tahap 5 Selesai: Model Random Forest berhasil dilatih!
```

Image 4. Model Training

This ensemble approach helps the model reduce the risk of overfitting while enhancing prediction stability, particularly when applied to inmate datasets with complex and diverse characteristics.

Model Evaluation

The model's performance was evaluated using multiple metrics, including accuracy, precision, recall, F1-score, and the confusion matrix. Employing a variety of metrics provides a more comprehensive assessment of the model's performance, not only in terms of overall prediction accuracy but also its ability to correctly classify each individual class.

Table 4. Accuration Model

Accura tion(%)	Preci sion(%)	Recall (%)	F1 Score (%)
86.17	86.0	83.0	85.0

The evaluation results indicate that the Random Forest model performs well in predicting inmate rehabilitation programs. The high accuracy value signifies that the majority of the data were correctly classified, while the consistent precision, recall, and F1-score values demonstrate that the model performs balanced classification across all classes.

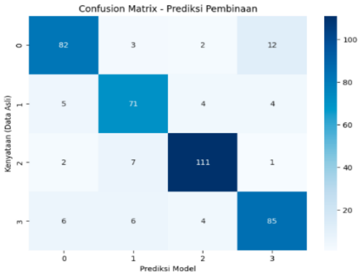


Image 5. Confusion Matrix of Model Training

Application Implementation

The Random Forest model developed in this study was implemented within a decision support system application to assist in classifying inmate rehabilitation programs. The application was designed with a clear separation between the training and testing processes to ensure transparency, reproducibility, and objective evaluation.

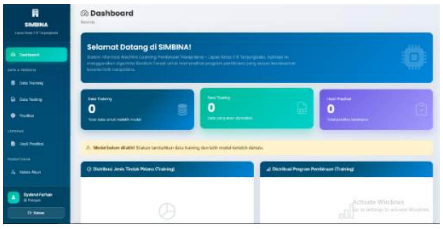

SIMBINA
Dashboard
Selamat Datang di SIMBINA!
Data Training
Data Testing
Prediksi
Hasil Prediksi
0
0
0
Distribusi Jenis Tindak Pidana (Training)
Distribusi Program Pembinaan (Training)
Activate Windows

Image 6. Full Application Interface

The Data Training module trains the Random Forest model using preprocessed data with an 80% training portion, while automatically presenting evaluation metrics such as accuracy, precision, recall, and F1-score. The Data Testing module evaluates the trained model using unseen data to assess its generalization capability and generate rehabilitation program predictions. Model performance was compared across two environments: Google Colab and the application system. Training in Google Colab, supported by automatic data partitioning and more flexible computational resources, achieved an accuracy of 86.17%.

In contrast, the application system applied a manual 80:20 split from 2,023 inmate records (1,618 training and 405 testing samples), resulting in an accuracy of 68.83%. This performance difference is influenced not only by resource and parameter constraints, but also by variations in data preprocessing procedures and differences in Scikit-Learn library versions between the two environments. Despite the lower accuracy, the implemented system demonstrates stable prediction behavior, reflecting practical deployment conditions and supporting its use as a data-driven decision support tool.

CONCLUSION

This research confirms that the Random Forest algorithm can be effectively used to predict inmate rehabilitation programs based on historical data from 2023 to 2025, consisting of 2,023 records. The model demonstrates stable classification performance by utilizing inmate characteristics to generate consistent predictions. Evaluation results indicate that model performance varies across training environments, with higher accuracy achieved in Google Colab than in the application system. Nevertheless, the deployed model maintains reliable performance under real-world conditions, supporting its role as a data-driven decision support tool. Performance assessment using accuracy, precision, recall, and F1-score further confirms the robustness of the proposed approach.

Future research may focus on improving predictive accuracy through hyperparameter optimization, incorporating additional relevant features, and conducting comparisons with other machine learning algorithms. Expanding the evaluation to multiple correctional institutions and standardizing preprocessing procedures may also enhance the model's generalizability

BIBLIOGRAPHY

[1] Y. R. Pratama and W. C. Warih2, "Evaluasi Kegiatan Pembinaan Kepribadian Narapidana Di Lembaga Pemasyarakatan Kelas Iia Bojonegoro Dengan Menggunakan Model Evaluasi Cipp (Context, Input, Process, ProducT)," Jurnal Ilmu Sosial & Hukum, Dec. 2025, doi: 10.61104/alz.v3i4.1883.

[2] Y. N. Primadani, G. Gamaputra, P. Studi, S. Terapan, A. Negara, and F. Vokasi, "Evaluasi Program Pembinaan Narapidana Lembaga Pemasyarakatan Kelas Iia Sidoarjo (Studi Kasus Pada Narapidana Narkoba) Evaluation Of The Guidance Program For Inpatients At Class Iia Correctional Institution In Sidoarjo (Case Study On Drug Inpatients)," Jurnal Ilmu Sosial & Hukum, vol. 3, no. 3, pp. 2025–2065, 2025.

[3] M. Diqi, M. E. Hiswati, Hamzah, I. W. Ordiyasa, S. H. Mulyani, N. Wijaya, and P. Wanda, "Optimizing Breast Cancer Detection: A Comparative Study of SVM and Naive Bayes Performance," *Applied Technology and Computing Science Journal*, vol. 7, no. 1, pp. 80–88, Jun. 2024, doi: 10.33086/atcsj.v7i1.6336.

[4] T. Tambunan, M. Yohanna, and A. P. Silalahi, "Penerapan Metode Random Forest Dalam Mendeteksi Berita Hoax," METHOMIKA Jurnal Manajemen Informatika dan Komputerisasi Akuntansi, vol. 7, no. 2, pp. 301–306, Dec. 2023, doi: 10.46880/jmika.Vol7No2.pp301-306.

[5] Z. A. Dwiyanti and C. Prianto, "Prediksi Cuaca Kota Jakarta Menggunakan Metode Random Forest," Jurnal Tekno Insentif, vol. 17, no. 2, pp. 127–137, Oct. 2023, doi: 10.36787/jti.v17i2.1136.

[6] M. Putri, "Prediksi Penyakit Stroke Menggunakan Machine Learning Dengan Algoritma Random Forest," Jurnal Infomedia: Teknik Informatika, Multimedia & Jaringan Vol. 9 No. 2 Maret 2024| P-ISSN: 2527-9858 E-ISSN: 2548-1180, 2024.

[7] A. D. Harahap, D. Juardi, and A. S. Y. Irawan, "Rancang Bangun Sistem Pendeteksi Link Phishing Menggunakan Algoritma Random Forest Berbasis Web," Jurnal Informatika dan Teknik Elektro Terapan, vol. 12, no. 3, Aug. 2024, doi: 10.23960/jitet.v12i3.4858.

[8] S. Sathyanarayanan, "Confusion Matrix-Based Performance Evaluation Metrics," African Journal of Biomedical Research, pp. 4023–4031, Nov. 2024, doi: 10.53555/ajbr.v27i4s.4345.