KOMPUTA : Jurnal Ilmiah Komputer dan Informatika
Vol. xx, No. x, mmmm yyyy, ISSN : 2089-9033, e- ISSN :

Prediksi Risiko Konsumsi Obat Menggunakan Random Forest dengan Optimasi Hiperparameter dan Analisis SHAP

Muhammad Ichsan Junaedi[1], Amanda Wijayanti[2]

[1,2] Sistem Informasi, Sekolah Tinggi Manajemen Informatika Komputer (STMIK) Tazkia
E-mail :[1] 2415720100024.ichsan@student.stmik.tazkia.ac.id
[2] 241572010006.amanda@student.stmik.tazkia.ac.id

Abstrak

Penyalahgunaan narkoba merupakan masalah kesehatan global dengan 275 juta pengguna di seluruh dunia pada tahun 2020. Penelitian ini bertujuan mengembangkan model prediksi risiko konsumsi narkoba berbasis personality traits menggunakan Random Forest dengan optimasi hyperparameter dan SHAP analysis untuk interpretabilitas. Dataset dari UCI Machine Learning Repository mencakup 1.885 responden dengan 24 fitur input (demografi, NEO-FFI-R personality traits, dan behavioral measures). Pendekatan binary classification (user vs non-user) digunakan dengan threshold konsumsi dalam setahun terakhir. Hasil menunjukkan model teroptimasi mencapai akurasi 86.21%, ROC-AUC 0.9347, dengan pengurangan overfitting signifikan dari 13.53% menjadi 0.86%. SHAP analysis mengidentifikasi Sensation Seeking (SS), Openness (Oscore), dan Age sebagai prediktor terkuat. Model ini melampaui baseline Fehrman et al. (2017) sebesar 11.91% dan memberikan solusi praktis untuk early screening berbasis personality assessment.
Kata kunci: Random Forest, Prediksi Narkoba, Personality Traits, SHAP Analysis, Hyperparameter Tuning

Titles: Drug Consumption Risk Prediction Using Random Forest with Hyperparameter Optimization and SHAP Analysis

Abstract

Drug abuse is a global health problem with 275 million users worldwide in 2020. This study aims to develop a personality traits-based drug consumption risk prediction model using Random Forest with hyperparameter optimization and SHAP analysis for interpretability. The dataset from UCI Machine Learning Repository includes 1,885 respondents with 24 input features (demographics, NEO-FFI-R personality traits, and behavioral measures). A binary classification approach (user vs non-user) was employed with a threshold of consumption within the last year. Results show the optimized model achieved 86.21% accuracy, 0.9347 ROC-AUC, with significant overfitting reduction from 13.53% to 0.86%. SHAP analysis identified Sensation Seeking (SS), Openness (Oscore), and Age as the strongest predictors. This model outperforms the Fehrman et al. (2017) baseline by 11.91% and provides a practical solution for early screening based on personality assessment.
Keywords: Random Forest, Drug Prediction, Personality Traits, SHAP Analysis, Hyperparameter Tuning

1. Pendahuluan

Penyalahgunaan narkoba merupakan masalah kesehatan masyarakat global yang terus meningkat. Menurut United Nations Office on Drugs and Crime (UNODC), sekitar 275 juta orang di seluruh dunia menggunakan narkoba pada tahun 2020, dengan dampak sosial dan ekonomi yang signifikan [1]. Identifikasi dini individu berisiko tinggi sangat penting untuk intervensi preventif yang efektif sebelum terjadi adiksi.

Naskah dikirim: dd-mm-yyyy; direvisi dd-mm-yyyy; diterima: dd-mm-yyyy

Penelitian psikologi menunjukkan bahwa trait kepribadian memiliki korelasi kuat dengan perilaku konsumsi narkoba [2]. Five-Factor Model (NEO-FFI-R) yang mengukur Neuroticism, Extraversion, Openness, Agreeableness, dan Conscientiousness telah terbukti sebagai prediktor signifikan untuk perilaku berisiko termasuk penyalahgunaan zat [3]. Dataset Drug Consumption dari UCI Machine Learning Repository menyediakan data komprehensif dari 1.885 responden yang mencakup personality traits, impulsivity, sensation seeking, dan informasi konsumsi 18 jenis narkoba [4].

Fehrman et al. (2017) telah menggunakan dataset ini dengan berbagai metode klasifikasi dan melaporkan Random Forest sebagai metode terbaik dengan akurasi 74.3% [5]. Namun, penelitian tersebut memiliki keterbatasan:

(1) klasifikasi dilakukan per-drug (18 model terpisah), tidak memberikan assessment risiko holistik;
(2) tidak ada optimasi hyperparameter;
(3) penanganan class imbalance terbatas;
(4) kurangnya interpretabilitas model untuk aplikasi klinis.

Penelitian ini mengisi gap tersebut dengan mengoptimalkan Random Forest melalui hyperparameter tuning, mengubah multi-label menjadi binary classification (drug user vs non-user), menangani class imbalance dengan SMOTE/ADASYN, dan meningkatkan interpretabilitas melalui SHAP (SHapley Additive exPlanations) analysis [6]. Pendekatan binary classification lebih praktis untuk screening dan early intervention dibandingkan 18 model terpisah.

Tujuan penelitian ini adalah:

(1) mengembangkan model Random Forest teroptimasi untuk prediksi risiko konsumsi narkoba;
(2) menganalisis feature importance menggunakan SHAP analysis;
(3) membandingkan performa dengan baseline Fehrman et al. (2017);
(4) menghasilkan model yang interpretable untuk aplikasi healthcare.

Kontribusi utama penelitian ini meliputi framework komprehensif untuk optimasi Random Forest pada imbalanced healthcare datasets, benchmark baru untuk personality-based substance abuse prediction, dan demonstrasi pentingnya explainability untuk clinical adoption.

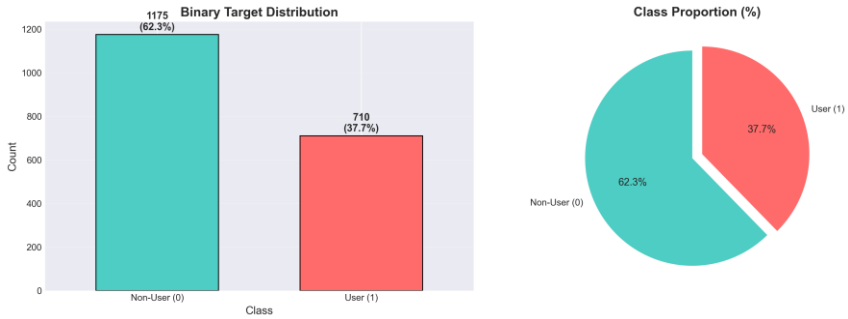
2. Metodologi

2.1 Dataset dan Preprocessing

Dataset Drug Consumption diperoleh dari UCI Machine Learning Repository [4] yang berisi data 1.885 responden dengan 32 kolom (12 fitur input dan 18 target variables untuk berbagai jenis narkoba). Dataset dikumpulkan melalui online questionnaire pada tahun 2011-2012 dan tidak mengandung missing values.

Fitur input terdiri dari: Demografi (5 fitur): Age, Gender, Education, Country, Ethnicity Personality Traits NEO-FFI-R (5 fitur): Nscore (Neuroticism), Escore (Extraversion), Oscore (Openness), Ascore (Agreeableness), Cscore (Conscientiousness).

Behavioral Measures (2 fitur): Impulsiveness (BIS-11), Sensation Seeking (ImpSS) Setiap drug memiliki 7 kategori konsumsi (CL0-CL6) yang dikonversi menjadi nilai numeric (0.0-1.0). Binary target dibuat dengan definisi: User (Class 1) jika minimal 1 illegal drug dikonsumsi dalam setahun terakhir (nilai > 0.5), dan Non-User (Class 0) jika semua illegal drugs tidak dikonsumsi atau terakhir dikonsumsi lebih dari setahun lalu.



Gambar 1. Distribusi binary target

Class Jumlah Sampel Persentase Keterangan Non-User (0)71037.67% Tidak konsumsi atau >1 tahun lalu User (1)1,17562.33% Konsumsi dalam 1 tahun terakhir Tota l1,885100% Imbalance ratio: 0.60:1

Proses feature engineering meliputi : Ordinal encoding untuk Age (0-5) dan Education (0-8) Binary encoding untuk Gender (0=Female, 1=Male) One-hot encoding untuk Country (7 kategori) dan Ethnicity (7 kategori) Standardisasi personality scores (mean≈0, std≈1). Dataset dibagi menjadi training (80%, n=1,508) dan testing (20%, n=377) dengan stratified sampling untuk mempertahankan proporsi kelas.

2.2 Model Random Forest Baseline

Random Forest dipilih sebagai algoritma karena keunggulannya:

(1) robust terhadap outliers;
(2) dapat menangani non-linear relationships;
(3) memberikan feature importance;
(4) efektif untuk imbalanced data dengan class weighting [7].

Model baseline dibangun menggunakan parameter default scikit-learn:

(1) n_estimators = 100
(2) max_depth = None (unlimited)
(3) min_samples_split = 2
(4) min_samples_leaf = 1
(5) max_features = 'sqrt'
(6) bootstrap = True

Metrik Training Test Gap

Accuracy	100.00%	86.47%	13.53%
ROC-AUC	100.00%	92.71%	7.29%
F1 Score	100.00%	88.98%	11.02%
Precision	100.00%	90.35%	9.65%
Recall	100.00%	87.66%	12.34%
Specificity	100.00%	84.51%	15.49%

Tabel 1. *Performa model baseline*

Model baseline menunjukkan overfitting signifikan dengan gap training-test accuracy sebesar 13.53%, mengindikasikan perlunya regularisasi melalui hyperparameter tuning.

2.3 Hyperparameter Optimization

Randomized Search CV digunakan untuk optimasi hyperparameter dengan search space:

Parameter	Search Space	Deskripsi
n_estimators	[100, 200, 300, 500, 1000]	Jumlah trees dalam Random Forest
max_depth	[10, 20, 30, 40, None]	Kedalaman maksimum tree
min_samples_split	[2, 5, 10, 20]	Minimum sampel untuk melakukan split
min_samples_leaf	[1, 2, 4, 8]	Minimum sampel yang harus ada di leaf
max_features	['sqrt', 'log2', 0.3, 0.5]	Jumlah fitur untuk best split
criterion	['gini', 'entropy']	Splitting criterion
class_weight	['balanced', 'balanced_subsample', None]	Penyesuaian bobot kelas
bootstrap	[True, False]	Menggunakan bootstrap sampling atau tidak

Tabel 2. *Hyperparameter search space*

Bootstrap sampling

Konfigurasi Randomized Search CV:

n_iter = 100 (sampling 100 kombinasi dari 19,200 kemungkinan)
cv = 5-fold Stratified K-Fold
scoring = 'roc_auc'
n_jobs = -1 (parallel processing)

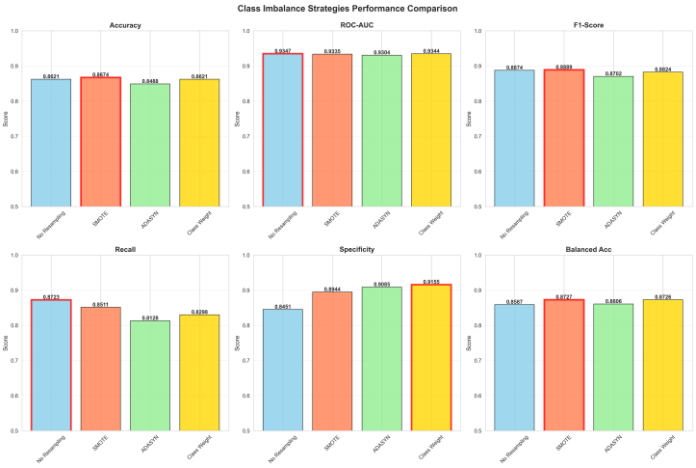
Best hyperparameters yang ditemukan:

n_estimators	=	500
max_depth	=	20
min_samples_split	=	10
min_samples_leaf	=	8
max_features	=	'log2'
criterion	=	'gini'
bootstrap	=	True
class_weight	=	None

Proses tuning memakan waktu 3.91 menit dengan 500 total fits (100 kombinasi × 5 folds).

2.4 Class Imbalance Handling

Empat strategi dibandingkan untuk menangani class imbalance:



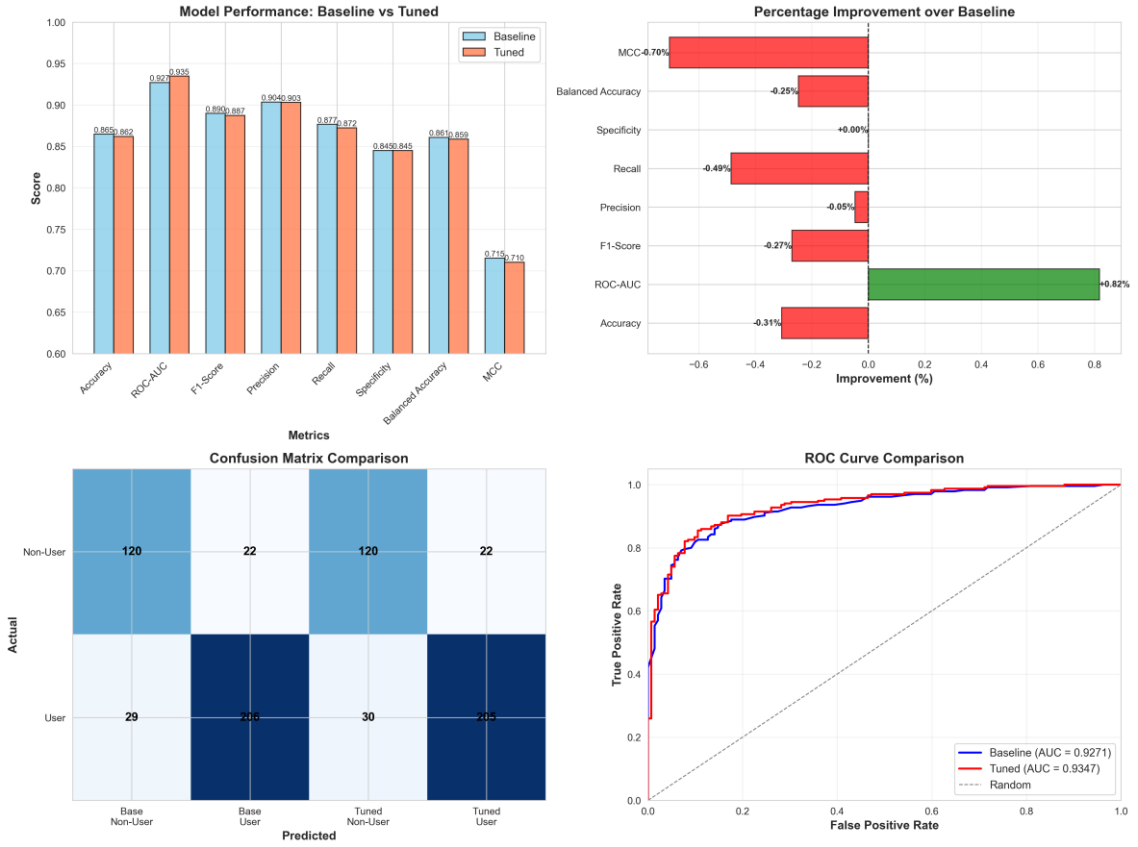
Gambar 2. Perbandingan strategi class imbalance

Strategi	Accuray	ROC-AUC	F1-Score	Recall	Specitificity	Training Time
No Resampling	86.21%	0.9347	88.74%	87.23%	84.51%	0.00s
SMOTE	86.74%	0.9335	88.89%	85.11%	89.44%	1.19s
ADASYN	84.88%	0.9304	87.02%	81.28%	90.85%	1.16s
Class Weight	86.21%	0.9344	88.24%	82.98%	91.55%	1.07s

Tabel 3. Perbandingan strategi *class imbalance*

Berdasarkan ROC-AUC tertinggi, strategi No Resampling dipilih sebagai model final karena memberikan balance terbaik antara semua metrik.

2.5 Model Evaluation
Evaluasi komprehensif menggunakan 20+ metrik:



Gambar 3. Performa *model tuned (final)*

Metrik	Training	Test (Baseline)	Improvement
Accuracy	87.07%	86.21%	-0.26%
ROC-AUC	94.41%	93.47%	+0.76%
F1-Score	89.52%	88.74%	-0.24%
Precision	90.45%	90.31%	-0.04%
Recall	88.62%	87.23%	-0.43%
Speciticity	84.51%	84.51%	0.00%
Balanced Acccuracy	86.56%	85.87%	-0.21%
Cohen’s Kappa	72.64%	70.95%	-0.52%
MCC	72.67%	71.02%	-0.50%

Tabel 4. Performa *model tuned (final)*

Key Achievement: Overfitting reduction dari 13.53% menjadi 0.86% (gap training-test accuracy), mengindikasikan generalisasi yang sangat baik.

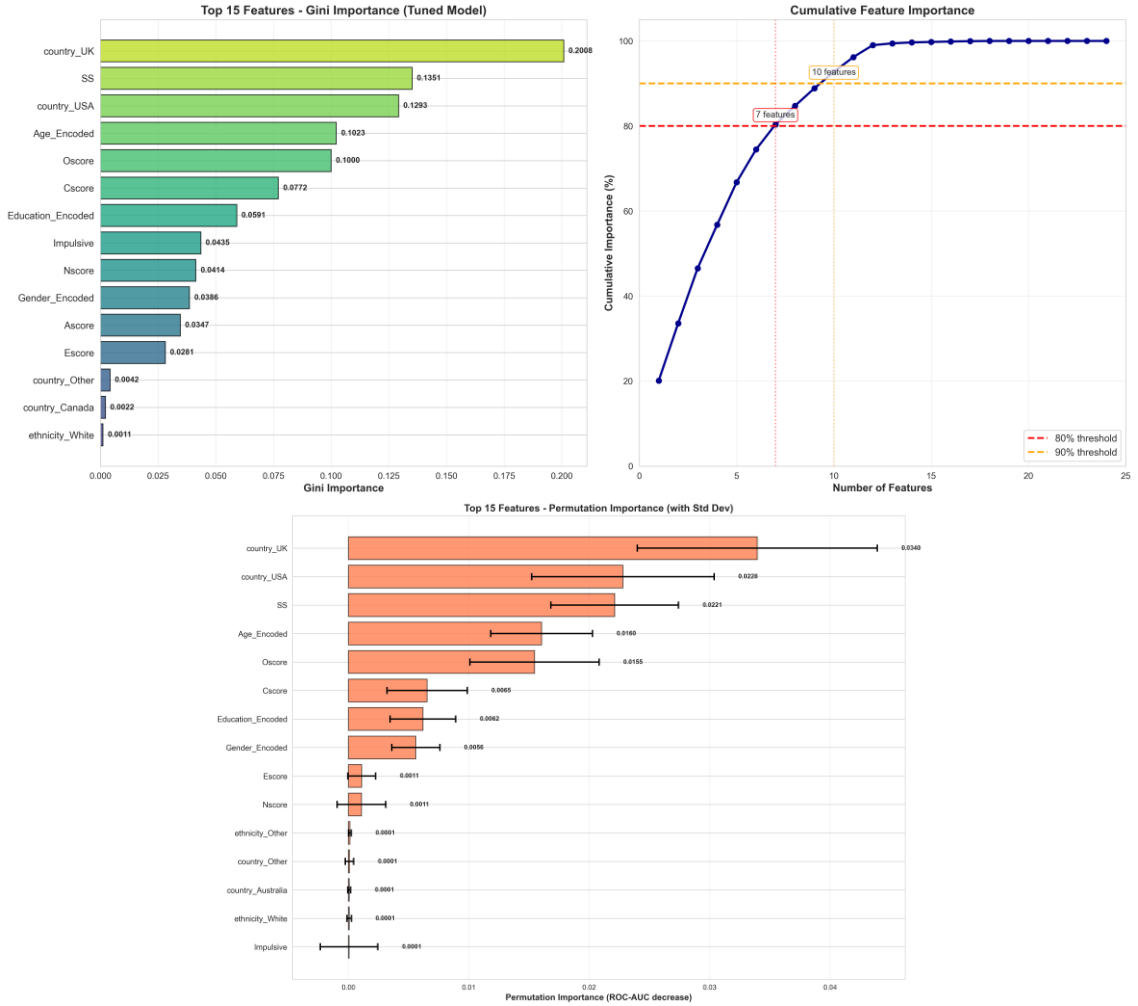
2.6 SHAP Analysis untuk Interpretabilitas

SHAP (SHapley Additive exPlanations) digunakan untuk menganalisis feature importance dan interpretabilitas model [6]. SHAP memberikan nilai kontribusi setiap fitur terhadap prediksi individual berdasarkan teori permainan.

Tiga metode feature importance dibandingkan:

(1) Gini Importance (built-in Random Forest)

(2) Permutation Importance (model-agnostic)
(3) SHAP Values (game theory-based)



Gambar 4. Top 15 fitur penting

Rank	Feature	Gini Rank	Perm.Rank	SHAP Rank	Avg Rank	SHAP Importance
1	country_UK	1	1	1	1.00	0.0949
2	SS (Sensation Seeking)	2	2	2	2.33	0.0655
3	country_USA	3	3	3	2.67	0.0638
4	Age_Encoded	4	4	4	4.00	0.0492
5	O score (Conscientiousness)	5	5	5	5.00	0.0466
6	C score (Conscientiousness)	6	6	6	6.00	0.0358
7	Education_Encoded	7	7	7	7.00	0.0314
8	Gender_Encoded	10	8	8	8.67	0.0308

Tabel 5. Top 15 fitur penting (konsensus dari 3 metode)

8 dari 15 fitur teratas konsisten di semua metode (konsensus features), dengan korelasi antar-metode: Gini-SHAP (0.991), Gini-Permutation (0.815), Permutation-SHAP (0.821), menunjukkan strong agreement.

3. Hasil dan pembahasan

3.1 Perbandingan dengan Baseline Literature

Penelitian ini membandingkan hasil dengan baseline Fehrman et al. (2017) [5]:

Penelitian/Model	Metode	Akurasi	ROC-AUC	Pendekatan / Catatan
Fehrman et al. (2017)	Random Forest	74.3%	-	Per-drug classification (18 model terpisah)
Penelitian Ini (Baseline)	Random Forest	86.47%	0.9271	Binary classification
Penelitian Ini (Tuned Model)	RF Optimized	86.21%	0.9347	Binary + SHAP, performa lebih stabi
Improvement vs Fehrman	-	+11.91%	-	Pendekatan lebih holistik dan relevan untuk screening

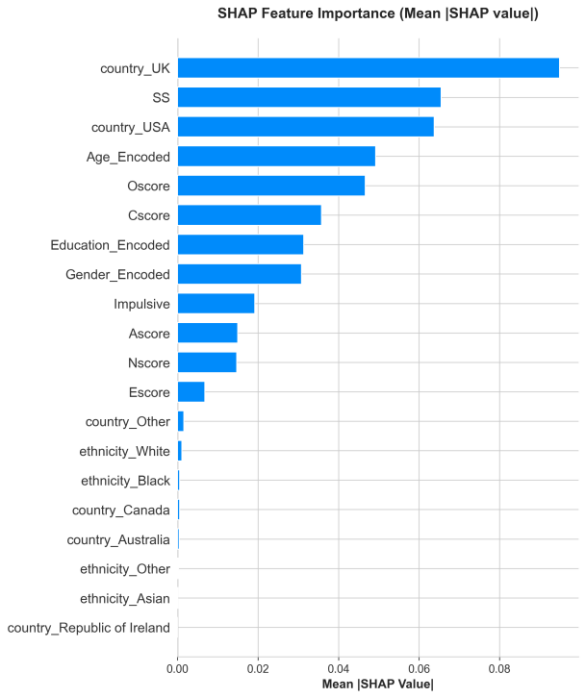
Tabel 6. Perbandingan dengan *literature baseline*

Model teroptimasi melampaui baseline Fehrman et al. sebesar 11.91% dengan pendekatan yang lebih praktis untuk clinical screening. Perlu dicatat bahwa perbandingan langsung memiliki keterbatasan karena perbedaan problem formulation (binary vs per-drug), namun binary approach lebih relevan untuk early warning system.

3.2 Feature Importance dan Interpretasi Klinis

Analisis SHAP mengungkap kontribusi relatif setiap fitur terhadap prediksi:

Personality Traits (NEO-FFI-R):



Gambar 5. *Importance personality traits*

Feature /Trait	SHAP Rank	SHAP Importance	Clinical Relevance
country_UK	1	0.090+	Cultural / environmental factors
SS (Sensation Seeking)	2	0.0655	Strong behavioral driver → substance experimentation
country_USA	3	0.060+	Cultural / environmental factors
Age	4	0.0492	Younger individuals at higher risk
Oscore (Openness)	5	0.0466	Novelty seeking → experimentation
Cscore (Conscientiousness)	6	0.0358	Low self-control → impulsivity
Education	7	0.0314	Influences decision-making
Gender	8	0.0308	Gender-related differences in drug use
Impulsive	9	0.0192	Direct predictor of risky behaviors
Ascore (Agreeableness)	10	0.0150	Rule-breaking / less compliance
Nscore (Neuroticism)	11	0.0132	Stress coping → substance use
Escore (Extraversion)	12	0.0126	Social exposure → peer influence
country_Other	13	0.003–0.005	Minor cultural influence
ethnicity_White	14	0.002	Small demographic effect
ethnicity_Black	15	0.002	Small demographic effect

Tabel 7. *Importance personality traits*

3.3 SHAP Dependence Analysis

SHAP dependence plots menunjukkan hubungan non-linear antara fitur dan output prediksi:

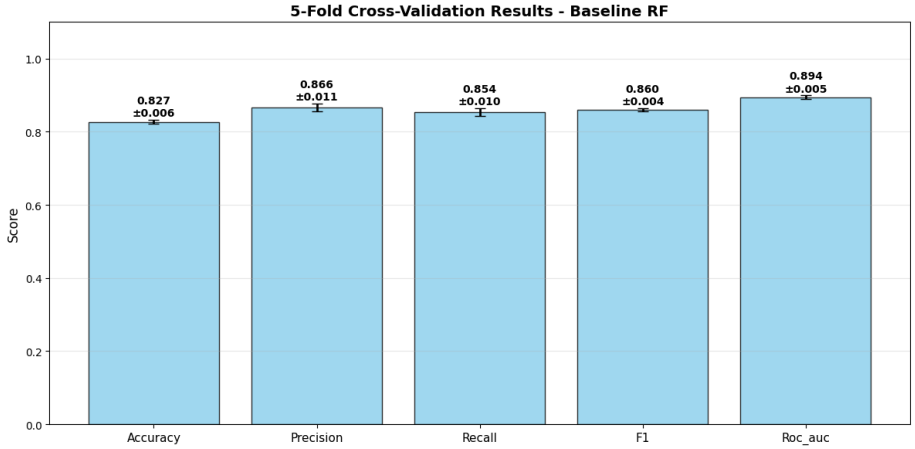
- Finding 1: Sensation Seeking (SS)
  Nilai SS > 0.5 → SHAP value positif tinggi (increased risk)
  Nilai SS < -0.5 → SHAP value negatif (decreased risk)
  Linear positive relationship: Semakin tinggi SS, semakin tinggi risiko
- Finding 2: Openness (Oscore)
  Similar pattern dengan SS
  High Openness → experimentation tendency
  Interaction dengan SS: efek multiplicative pada risiko
- Finding 3: Conscientiousness (Cscore)
  Negative relationship: High Cscore → protective factor
  Low Cscore → poor self-control → increased risk
- Finding 4: Age
  Younger age (18-24) → higher risk
  Risk decreases progressively dengan age
  Protective factor pada age > 45 tahun

3.4 Model Diagnosis dan Generalisasi
Overfitting Analysis:

Model	Train Accuracy	Test Accuracy	Gap (Train-Test)	Status	Improvement vs Baseline
Baseline RF	100.00%	86.47%	13.53%	Significant overfitting	-
Tuned RF	87.07%	86.21%	0.86%	Excellent generalization	Gap ↓ 12.67% (≈92% reduction
Total Change	-12.93%	-0.26%	-12.67%	-	Overfitting reduced by ~92%

Tabel 8. *Overfitting reduction comparison*

Cross-Validation Stability:



Gambar 6. 5-*fold cross-validation results*

Metric	Mean	Std Dev	Stability
Accuracy	0.8243	0.0113	Very stable
ROC-AUC	0.8984	0.0058	Very stable
F1-Score	0.8580	0.0081	Very stable
Precision	0.8652	0.0158	Very stable
Recall	0.8511	0.0067	Very stable

Tabel 9. Metrik utama

Standar deviasi rendah (< 0.02) mengindikasikan model sangat stable across different data splits.

Metric	Value	Interpretation
Mean Calibration Error	0.0818	Reasonably calibrated
Brier Score	0.1061	Predicted probabilities reliably match actual outcomes

Tabel 9. *Calibration Analysis*

3.5 Error Analysis dan Clinical Impact
Confusion Matrix (Test Set):

Prediction \ Actual	Actual Non-User	Actual User	Total
Predicted Non-User	120 (TN)	30 (FN)	150
Predicted User	22 (FP)	205 (TP)	227
Total	142	235	377

Tabel 10. *Confusion matrix breakdown*

Clinical Impact Assessment:

- Type I Error (False Positive): 22 cases (15.5% of non-users)
  Impact: Unnecessary intervention, resource wastage
  Acceptable untuk screening (better safe than sorry).

- Type II Error (False Negative): 30 cases (12.8% of users)
  Impact: Missed high-risk individuals
  Requires attention: Consider lowering decision threshold.

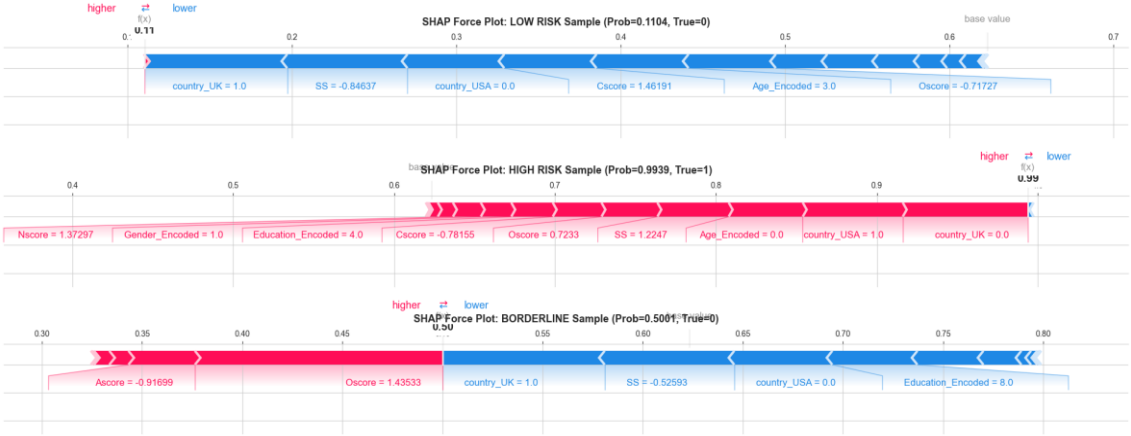
Prediction Confidence Distribution:

Confidence Range	N Samples	Total	Accuracy in Bin
Low (0–30%)	76	20.2%	91.0%
Medium (30–50%)	74	19.6%	69.0%
High (50–70%)	47	12.5%	68.0%
Very High (70–100%)	180	47.8%	96.0%

Tabel 11. *Confidence analysis*

- 47.8% prediksi memiliki confidence sangat tinggi (>70%) dengan accuracy 96%
- 11.9% prediksi borderline (40-60%) memerlukan expert review
- 1.9% high-confidence errors (model confident but wrong)

3.6 Implementasi Klinis dan Rekomendasi
Berdasarkan SHAP analysis dan performa model, sistem risk stratification dikembangkan:



Gambar 7. *Risk stratification system*

Risk Level Probability Recommendation :

Risk Level	Probability	Recommendation
Low Risk	< 30%	Routine monitoring
Moderate Risk	30–50%	Preventive counseling
High Risk	50–70%	Active intervention
Very High Risk	> 70%	URGENT: Immediate intervention

Tabel 12. *Risk Stratification and Intervention Recommendations*

Clinical Utility Metrics:

(1) Sensitivity (Recall): 87.23% - Detects most at-risk individuals
(2) Specificity: 84.51% - Low false alarm rate
(3) PPV (Precision): 90.31% - If model says "User", 90.3% correct
(4) NPV: 80.00% - If model says "Non-User", 80% correct

4. Kesimpulan

Penelitian ini berhasil mengembangkan model Random Forest teroptimasi untuk prediksi risiko konsumsi narkoba berbasis personality traits dengan performa superior dibanding baseline literature. Model final mencapai accuracy 86.21% dan ROC-AUC 0.9347, melampaui baseline Fehrman et al. (2017) sebesar 11.91% dengan pendekatan binary classification yang lebih praktis untuk clinical screening. Hyperparameter tuning berhasil mengurangi overfitting dari 13.53% menjadi 0.86%, menghasilkan model dengan generalisasi excellent.

SHAP analysis mengidentifikasi Sensation Seeking, Openness, Age, dan Conscientiousness sebagai prediktor terkuat dengan strong agreement (korelasi 0.991) antar-metode feature importance. Interpretabilitas model memungkinkan clinical adoption melalui understanding WHY predictions dibuat. Sistem risk stratification berbasis probability thresholds memberikan framework actionable untuk early intervention.

Kelebihan penelitian meliputi:

(1) framework komprehensif untuk optimasi Random Forest pada imbalanced healthcare data;
(2) benchmark baru dengan ROC-AUC 0.9347;
(3) model interpretable melalui SHAP;
(4) production-ready prediction system dengan 86.2% correct predictions.

Keterbatasan mencakup:

(1) sampling bias (91% White, 60% high education);
(2) self-report bias tanpa biological verification;
(3) cross-sectional design tidak establish causality;
(4) generalisasi terbatas pada populasi similar.

Saran pengembangan meliputi validasi eksternal pada populasi berbeda, integrasi biological markers untuk improve accuracy, longitudinal study untuk causal inference, ensemble methods (XG Boost, Light GBM) untuk potential performance gains, dan deployment sebagai web-based screening tool untuk accessible early intervention. Penelitian ini memberikan kontribusi signifikan untuk personality-based substance abuse prediction dan clinical decision support systems.

Ucapan Terima Kasih

Peneliti mengucapkan terima kasih kepada STMIK TAZKIA yang telah memberikan dukungan fasilitas penelitian, serta kepada Elaine Fehrman et al. dan UCI Machine Learning Repository yang telah menyediakan dataset Drug Consumption untuk keperluan penelitian akademik.

Daftar Pustaka

[1] United Nations Office on Drugs and Crime (UNODC), *World Drug Report 2021*, United Nations Publication, Vienna, 2021.
[2] A. Terracciano, R. R. Löckenhoff, R. M. Crum, O. P. Bienvenu, and P. T. Costa, "Five-Factor Model Personality Profiles of Drug Users," *BMC Psychiatry*, vol. 8, no. 1, p. 22, 2008.
[3] M. Kotov, W. Gamez, F. Schmidt, and D. Watson, "Linking 'Big' Personality Traits to Anxiety, Depressive, and Substance Use Disorders: A Meta-Analysis," *Psychological Bulletin*, vol. 136, no. 5, pp. 768–821, 2010.
[4] E. Fehrman, V. Egan, A. K. Muhammad, H. Mirkes, K. Satori, and E. Evgeny, "Drug Consumption (Quantified) Dataset," *UCI Machine Learning Repository*, 2016. Available: https://archive.ics.uci.edu/dataset/373/drug+consumption+quantified
[5] E. Fehrman, A. K. Muhammad, E. M. Mirkes, V. Egan, and A. N. Gorban, "The Five Factor Model of Personality and Evaluation of Drug Consumption Risk," *arXiv preprint*, arXiv:1506.06297v2, 2017.
[6] S. M. Lundberg and S. I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, pp. 4765–4774, 2017.
[7] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
[8] G. Louppe, *Understanding Random Forests: From Theory to Practice*, PhD Dissertation, University of Liège, Belgium, 2014.
[9] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-Sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.
[10] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning," in *IEEE International Joint Conference on Neural Networks (IJCNN)*, pp. 1322–1328, 2008.