Prediksi Risiko Konsumsi Narkoba Menggunakan Random Forest

Disusun Oleh:

- Muhammad Ichsan Junaedi (241572010024)
- Amanda Wijayanti (241572010006)

Dosen Pengampu: Hendri Kharisma S.Kom, M.T

Program Studi Sistem Informasi
Sekolah Tinggi Manajemen Informatika dan Komputer TAZKIA
Raya Dramaga Blok Radar Baru No.8, RT.03/RW.03, Margajaya, Kec. Bogor Bar.,
Kota Bogor, Jawa Barat 16116, Indonesia

BAB 1: PENDAHULUAN

1.1 Latar Belakang

Penyalahgunaan narkoba merupakan masalah kesehatan global yang terus meningkat. Menurut UNODC, sekitar 275 juta orang di seluruh dunia menggunakan narkoba pada tahun 2020. Identifikasi dini individu berisiko tinggi sangat penting untuk intervensi preventif yang efektif.

Penelitian psikologi menunjukkan bahwa trait kepribadian (Five-Factor Model) memiliki korelasi kuat dengan perilaku konsumsi narkoba. Dataset Drug Consumption dari UCI Machine Learning Repository menyediakan data dari 1.885 responden yang mencakup personality traits, impulsivity, sensation seeking, dan informasi konsumsi 18 jenis narkoba.

Penelitian Fehrman et al. (2017) telah menggunakan dataset ini dengan berbagai metode klasifikasi: Decision Tree (69.8%), Random Forest (74.3%), k-Nearest Neighbors (71.2%), Linear Discriminant Analysis (68.5%), Logistic Regression (70.1%), dan Naive Bayes (67.9%). Random Forest menunjukkan performa terbaik dengan akurasi 74.3%. 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. Penelitian ini bertujuan mengisi gap tersebut dengan mengoptimalkan Random Forest, mengubah multi-label menjadi binary classification (drug user vs non-user), menangani class imbalance dengan SMOTE/ADASYN, dan meningkatkan interpretabilitas melalui SHAP analysis.

1.2 Rumusan Masalah

Bagaimana membangun dan mengoptimalkan model Random Forest untuk memprediksi risiko konsumsi narkoba secara umum (binary classification: user vs non-user) berdasarkan personality traits dan behavioral measures, dengan menangani class imbalance serta meningkatkan interpretabilitas model melalui SHAP analysis?

1.3 Tujuan Penelitian

1. Melaksanakan Tugas Besar Ujian Tengah Semester (UTS)
2. Melakukan transformasi multi-label target (18 drugs) menjadi single binary target (drug user vs non-user)
3. Membangun dan mengoptimalkan model Random Forest melalui hyperparameter tuning
4. Melakukan analisis feature importance menggunakan SHAP analysis
5. Membandingkan performa dengan baseline Fehrman et al. (2017)

1.4 Manfaat Penelitian

Manfaat Akademik:

- Pemahaman mendalam tentang optimasi Random Forest untuk healthcare prediction
- Mengisi research gap dalam drug consumption prediction dengan binary classification approach
- Benchmark baru untuk penelitian personality-based substance abuse prediction

Manfaat Metodologis:

- Framework komprehensif untuk optimasi Random Forest pada imbalanced healthcare datasets
- Best practices handling class imbalance melalui perbandingan SMOTE, ADASYN, dan class weighting
- Demonstrasi pentingnya explainability (SHAP) untuk clinical adoption

Manfaat Sosial:

- Mendukung pencegahan penyalahgunaan narkoba melalui identifikasi dini berbasis personality assessment
- Alokasi sumber daya intervensi lebih efisien dengan targeting high-risk individuals
- Kontribusi kebijakan kesehatan masyarakat berbasis evidence
-

BAB 2: DESKRIPSI DATASET

2.1 Sumber Dataset

- **Repository:** UCI Machine Learning Repository
- **Judul:** Drug Consumption (Quantified)
- **URL:** https://archive.ics.uci.edu/dataset/373/drug+consumption+quantified
- **Paper:** Fehrman et al. (2017) - arXiv:1506.06297v2
- **Lisensi:** Open source
- **Format:** CSV

2.2 Struktur Dataset

Dataset terdiri dari **12 fitur input** dan **18 target variables** (drug types).

A. Data Demografi (5 fitur)

Fitur	Deskripsi	Tipe	Range/Kategori
Age	Kelompok usia	Ordinal	18-24, 25-34, 35-44, 45-54, 55-64, 65+
Gender	Jenis kelamin	Binary	Male, Female
Education	Tingkat pendidikan	Ordinal	9 levels (Left school before 16 - Doctorate)
Country	Negara tempat tinggal	Nominal	UK, USA, Canada, Australia, Other
Ethnicity	Latar belakang etnis	Nominal	White, Asian, Black, Mixed, Other

B. Personality Traits - NEO-FFI-R (5 fitur)

Semua personality scores sudah standardized (mean=0, std=1):

Trait	Kode	Deskripsi
Neuroticism	Nscore	Kecenderungan emosi negatif (anxiety, depression)
Extraversion	Escore	Keaktifan sosial dan energi positif
Openness	Oscore	Keterbukaan pengalaman baru, kreativitas
Agreeableness	Ascore	Sikap kooperatif, empati
Conscientiousness	Cscore	Kedisiplinan, kontrol impuls

C. Behavioral Measures (2 fitur)

Fitur	Instrumen	Deskripsi
Impulsiveness	BIS-11	Bertindak tanpa berpikir panjang
Sensation Seeking	ImpSS	Pencarian pengalaman berisiko

D. Target Variables (18 Drugs)

Setiap drug memiliki 7 kategori konsumsi:

- **CL0:** Never Used
- **CL1:** Used over a Decade Ago
- **CL2:** Used in Last Decade
- **CL3:** Used in Last Year
- **CL4:** Used in Last Month

- **CL5:** Used in Last Week
- **CL6:** Used in Last Day

Klasifikasi Drugs:

- **Legal (4):** Alcohol, Caffeine, Chocolate, Nicotine
- **Illegal (14):** Cannabis, Amphetamines, Cocaine, Ecstasy, LSD, Magic Mushrooms, Heroin, Crack, Benzodiazepines, Methadone, Ketamine, Amyl Nitrite, Legal Highs, VSA

2.3 Target Transformation: Multi-label → Binary

Binary Classification Definition:

User (Class 1): Responden menggunakan minimal 1 illegal drug dengan kategori CL3, CL4, CL5, atau CL6 (used within last year)

Non-user (Class 0): Tidak menggunakan illegal drugs atau hanya CL0, CL1, CL2 (never used or used >1 year ago)

Rationale:

- Clinical relevance: Recent use lebih relevant untuk risk assessment
- Focus on illegal substances (exclude legal drugs)
- Practical screening: Binary outcome untuk yes/no decision
- Reduced complexity: 18×7=126 outcomes → 1 binary target

2.4 Karakteristik Dataset

Dimensi:

- Total samples: 1,885 responden
- Input features: 12
- Target: 1 binary variable (after transformation)
- Missing values: Tidak ada
- Data collection: 2011-2012 (online questionnaire)

Distribusi Demografi:

- Gender: 48.6% Male, 51.4% Female (balanced)
- Age: 60% berusia 18-34 tahun
- Education: 60% minimal Bachelor degree
- Ethnicity: 91% White
- Country: 54% UK, 30% USA

Drug Prevalence (Original):

- High: Alcohol (85%), Caffeine (90%), Cannabis (45%)
- Medium: Nicotine (35%), Ecstasy (18%), Cocaine (15%)
- Low: Heroin (2.81%), Crack (1.06%)

2.5 Kelebihan dan Keterbatasan

Kelebihan:

- Clean data (no missing values)
- Standardized features (personality scores)
- Adequate sample size (1,885 samples)
- Validated instruments (NEO-FFI-R, BIS-11)
- Comprehensive drug coverage (18 substances)

Keterbatasan:

- Potential class imbalance (after binary transformation)
- Sampling bias (91% White, 60% high education)
- Self-report bias (no biological verification)
- Cross-sectional design (cannot establish causality)
- Online sample (volunteer bias)

BAB 3: METODOLOGI

3.1 Preprocessing

A. Exploratory Data Analysis

- Analisis distribusi personality traits
- Statistical testing (t-test untuk users vs non-users)
- Correlation analysis (heatmap)
- Outlier detection

B. Feature Engineering

- Label encoding untuk Age & Education
- Binary encoding untuk Gender
- One-hot encoding untuk Country & Ethnicity
- Transform target: 18 drugs × 7 classes → binary (user vs non-user)

C. Train-Test Split

- 80% training, 20% testing

- Stratified split (maintain class proportion)

D. Handling Class Imbalance

- SMOTE: Synthetic minority over-sampling
- ADASYN: Adaptive synthetic sampling
- Class Weighting: Adjust model parameters

3.2 Modeling

A. Random Forest

- Ensemble learning dengan bagging
- Bootstrap aggregating multiple decision trees
- Random feature selection
- **Random Forest Hyperparameters:**

Parameter	Search Space	Description
n_estimators	[100, 200, 300, 500, 1000]	Number of trees
max_depth	[10, 20, 30, 40, None]	Maximum tree depth
min_samples_split	[2, 5, 10, 20]	Minimum samples to split node
min_samples_leaf	[1, 2, 4, 8]	Minimum samples at leaf
max_features	['sqrt', 'log2', 0.3, 0.5]	Features for best split
bootstrap	[True, False]	Bootstrap sampling

class_weight	['balanced', 'balanced_subsample', None]	Class weights
criterion	['gini', 'entropy']	Splitting criterion

B. Hyperparameter Tuning

- RandomizedSearchCV (5-fold CV)
- Scoring: ROC-AUC

3.3 Evaluasi Model

Metrics:

1. Confusion Matrix Metrics:

- Sensitivity (Recall)
- Specificity
- Precision
- F1-Score
- Balanced Accuracy

2. Probability Metrics:

- ROC-AUC
- PR-AUC (untuk imbalanced data)

3. Overall Metrics:

- Cohen's Kappa
- Matthews Correlation Coefficient (MCC)

Cross-Validation:

- 5-Fold Stratified CV

Feature Importance:

- SHAP Analysis: Global & local importance
- Built-in importance: Gini importance
- Permutation importance

Statistical Testing:

- McNemar's test (compare models)
- Paired t-test (compare CV scores)

3.4 Visualisasi

A. EDA

- Histograms (personality distribution)
- Heatmap (correlation matrix)
- Bar charts (demographics)
- Radar charts (personality profiles)

B. Model Performance

- ROC Curves
- Precision-Recall Curves
- Confusion Matrix heatmaps
- Performance comparison bar charts

C. Feature Importance

- SHAP summary plots
- SHAP dependence plots
- Feature importance bar charts

D. Class Imbalance

- Before/After SMOTE comparison
-

BAB 4: RENCANA KERJA

4.1 Timeline Pelaksanaan

Week	Aktivitas	Deliverables

1	• Load dataset & EDA • Preprocessing & feature engineering • Binary target creation • Handle class imbalance	• EDA report • Cleaned dataset • Balanced training sets
2	• Random Forest baseline • Hyperparameter tuning • Compare sampling strategies • Cross-validation	• Optimized RF model • Tuning results • CV scores • Performance tables
3	• SHAP analysis • Feature importance • Statistical testing vs baseline • Clinical validation	• SHAP plots • Feature rankings • Statistical test results • Interpretation report
4	• Comprehensive visualizations • Results analysis • Final report writing • Presentation preparation	• Final report • Presentation slides • GitHub repository • Documentation

4.2 Deliverables

A. GitHub Repository Structure

- drug-consumption-rf-prediction/
- ├── data/
- │   ├── raw/
- │   └── processed/
- ├── notebooks/
- │   ├── 01_EDA_Preprocessing.ipynb
- │   ├── 02_Feature_Engineering.ipynb
- │   ├── 03_RF_Baseline.ipynb
- │   ├── 04_Hyperparameter_Tuning.ipynb
- │   ├── 05_Class_Imbalance.ipynb
- │   ├── 06_Model_Evaluation.ipynb
- │   └── 07_SHAP_Analysis.ipynb
- ├── models/
- ├── results/

- │ ├── figures/
- │ ├── metrics/
- │ └── reports/
- ├── requirements.txt
- ├── README.md

└── LICENSE

B. Dokumen Akhir

1. **Proposal (dokumen ini)**
2. **Final Report** (format scientific paper)
   - Abstract
   - Introduction
   - Literature Review
   - Methodology
   - Results & Discussion
   - Conclusion
   - References
3. **Presentation Slides**
4. **README.md** (GitHub documentation)

BAB 5: KESIMPULAN DAN HARAPAN

5.1 Kesimpulan

Proposal ini mengajukan penelitian prediksi risiko konsumsi narkoba menggunakan Random Forest dengan pendekatan binary classification. Berbeda dari penelitian Fehrman et al. (2017) yang membangun 18 model terpisah per-drug, penelitian ini akan menghasilkan single model untuk overall drug use risk assessment yang lebih praktis untuk screening dan early intervention. Melalui optimasi hyperparameter, penanganan class imbalance, dan SHAP analysis, penelitian ini diharapkan meningkatkan akurasi prediksi dan menghasilkan model yang interpretable untuk aplikasi healthcare.

5.2 Harapan

Proposal ini diharapkan dapat disetujui karena:

- **Metodologi berbeda:** Binary classification approach (belum dilakukan baseline)
- **Gap research jelas:** Optimasi hyperparameter dan interpretabilitas belum dieksplorasi
- **Aplikasi praktis:** Model untuk early warning system penyalahgunaan narkoba
- **Kontribusi akademik:** Benchmark baru dan insights personality-drug relationships

REFERENSI

Primary:

1. Fehrman, E., et al. (2017). The Five Factor Model of personality and evaluation of drug consumption risk. *arXiv:1506.06297v2*.
2. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32.
3. Louppe, G. (2014). Understanding random forests: From theory to practice. PhD Dissertation.
4. Chawla, N. V., et al. (2002). SMOTE: Synthetic minority over-sampling technique. *JAIR*, 16, 321-357.
5. He, H., et al. (2008). ADASYN: Adaptive synthetic sampling approach. *IEEE IJCNN*, 1322-1328.
6. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *NeurIPS*, 30, 4765-4774.
7. Molnar, C. (2020). Interpretable machine learning. Lulu.com.
8. Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *JMLR*, 13, 281-305.
9. Rajkomar, A., et al. (2019). Machine learning in medicine. *NEJM*, 380(14), 1347-1358.
10. Terracciano, A., et al. (2008). Five-factor model personality profiles of drug users. *BMC Psychiatry*, 8(1), 22.