# IMPLEMENTASI METODE OVER SAMPLING UNTUK MENINGKATKAN PERFORMA RANDOM FOREST CLASSIFICATION PADA DATA YANG TIDAK SEIMBANG

**SKRIPSI**

Diajukan untuk memenuhi salah satu syarat mencapai derajat Sarjana

Program Studi S1 Informatika


disusun oleh

**CHERFLY KAOPE**

**19.11.3012**


Kepada


**FAKULTAS ILMU KOMPUTER**

**UNIVERSITAS AMIKOM YOGYAKARTA**

**YOGYAKARTA**


**2022**

# IMPLEMENTASI METODE OVER SAMPLING UNTUK MENINGKATKAN PERFORMA RANDOM FOREST CLASSIFICATION PADA DATA YANG TIDAK SEIMBANG

**SKRIPSI**

untuk memenuhi salah satu syarat mencapai derajat Sarjana

Program Studi S1 Informatika


disusun oleh

**CHERFLY KAOPE**

**19.11.3012**


Kepada


**FAKULTAS ILMU KOMPUTER**

**UNIVERSITAS AMIKOM YOGYAKARTA**

**YOGYAKARTA**


**2022**

# HALAMAN PERSETUJUAN

## SKRIPSI

## IMPLEMENTASI METODE OVER SAMPLING UNTUK MENINGKATKAN PERFORMA RANDOM FOREST CLASSIFICATION PADA DATA YANG TIDAK SEIMBANG

yang disusun dan diajukan oleh

**Cherfly Kaope**

**19.11.3012**

telah disetujui oleh Dosen Pembimbing Skripsi
pada tanggal 24 November 2022

**Dosen Pembimbing,**

**Anna Baita, M.Kom.**
**NIK. 190302290**

# HALAMAN PENGESAHAN

**SKRIPSI**

**IMPLEMENTASI METODE OVER SAMPLING UNTUK MENINGKATKAN PERFORMA RANDOM FOREST CLASSIFICATION PADA DATA YANG TIDAK SEIMBANG**

yang disusun dan diajukan oleh

**Cherfly Kaope**

**19.11.3012**

Telah dipertahankan di depan Dewan Penguji
pada tanggal 24 November 2022

**Susunan Dewan Penguji**

| Nama Penguji | Tanda Tangan |
|---|---|
| **Mardhiya Hayaty, S.T., M.Kom.**<br>**NIK. 190302108** | _____ |
| **Yuli Astuti, M.Kom**<br>**NIK. 190302146** | _____ |
| **Anna Baita, M.Kom.**<br>**NIK. 190302290** | _____ |

Skripsi ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Sarjana Komputer
Tanggal 24 November 2022

**DEKAN FAKULTAS ILMU KOMPUTER**

**Hanif Al Fatta, S.Kom., M.Kom.**
**NIK. 190302096**

# HALAMAN PERNYATAAN KEASLIAN SKRIPSI

## HALAMAN PERNYATAAN KEASLIAN SKRIPSI

Yang bertandatangan di bawah ini,

**Nama mahasiswa : Cherfly Kaope**
**NIM : 19.11.3012**

Menyatakan bahwa Skripsi dengan judul berikut:

**IMPLEMENTASI METODE OVER SAMPLING UNTUK MENINGKATKAN PERFORMA RANDOM FOREST CLASSIFICATION PADA DATA YANG TIDAK SEIMBANG**

Dosen Pembimbing : Anna Baita, M.Kom

1. Karya tulis ini adalah benar-benar ASLI dan BELUM PERNAH diajukan untuk mendapatkan gelar akademik, baik di Universitas AMIKOM Yogyakarta maupun di Perguruan Tinggi lainnya.
2. Karya tulis ini merupakan gagasan, rumusan dan penelitian SAYA sendiri, tanpa bantuan pihak lain kecuali arahan dari Dosen Pembimbing.
3. Dalam karya tulis ini tidak terdapat karya atau pendapat orang lain, kecuali secara tertulis dengan jelas dicantumkan sebagai acuan dalam naskah dengan disebutkan nama pengarang dan disebutkan dalam Daftar Pustaka pada karya tulis ini.
4. Perangkat lunak yang digunakan dalam penelitian ini sepenuhnya menjadi tanggung jawab SAYA, bukan tanggung jawab Universitas AMIKOM Yogyakarta.
5. Pernyataan ini SAYA buat dengan sesungguhnya, apabila di kemudian hari terdapat penyimpangan dan ketidakbenaran dalam pernyataan ini, maka SAYA bersedia menerima SANKSI AKADEMIK dengan pencabutan gelar yang sudah diperoleh, serta sanksi lainnya sesuai dengan norma yang berlaku di Perguruan Tinggi.

Yogyakarta, 24 November 2022

Yang Menyatakan,

Cherfly Kaope

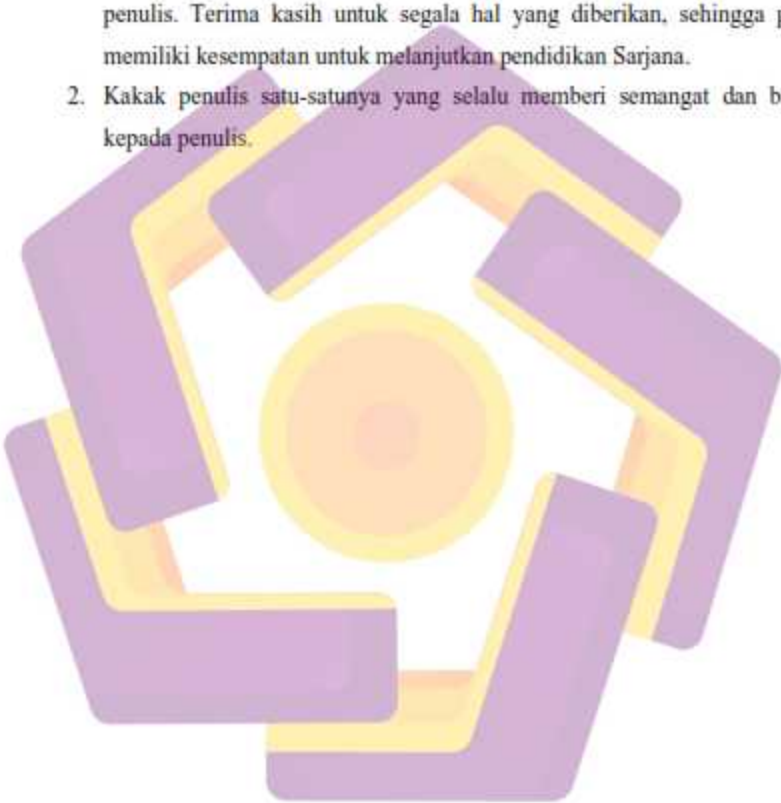


v

# HALAMAN PERSEMBAHAN

Segala puji syukur ke hadirat Tuhan Yang Maha Esa yang telah memberikan berkat dan pertolongan-Nya, sehingga penulis dapat menyelesaikan skripsi ini dengan baik. Skripsi ini penulis persembahkan kepada:

1. Ayah dan ibu tercinta yang senantiasa memberikan dukungan dan doa kepada penulis. Terima kasih untuk segala hal yang diberikan, sehingga penulis memiliki kesempatan untuk melanjutkan pendidikan Sarjana.
2. Kakak penulis satu-satunya yang selalu memberi semangat dan bantuan kepada penulis.



# KATA PENGANTAR

Segala puji syukur ke hadirat Tuhan Yang Maha Esa atas kasih dan karunia-Nya sehingga penulis dapat menyelesaikan skripsi ini dengan judul "Implementasi Metode Over Sampling untuk Meningkatkan Performa Random Forest Classification pada Data yang Tidak Seimbang".

Dalam menyusun skripsi ini, penulis memperoleh banyak arahan, saran, dan kritik, sehingga penulis dapat menyelesaikan skripsi ini dengan baik. Oleh karena itu, penulis ingin mengucapkan terima kasih kepada:

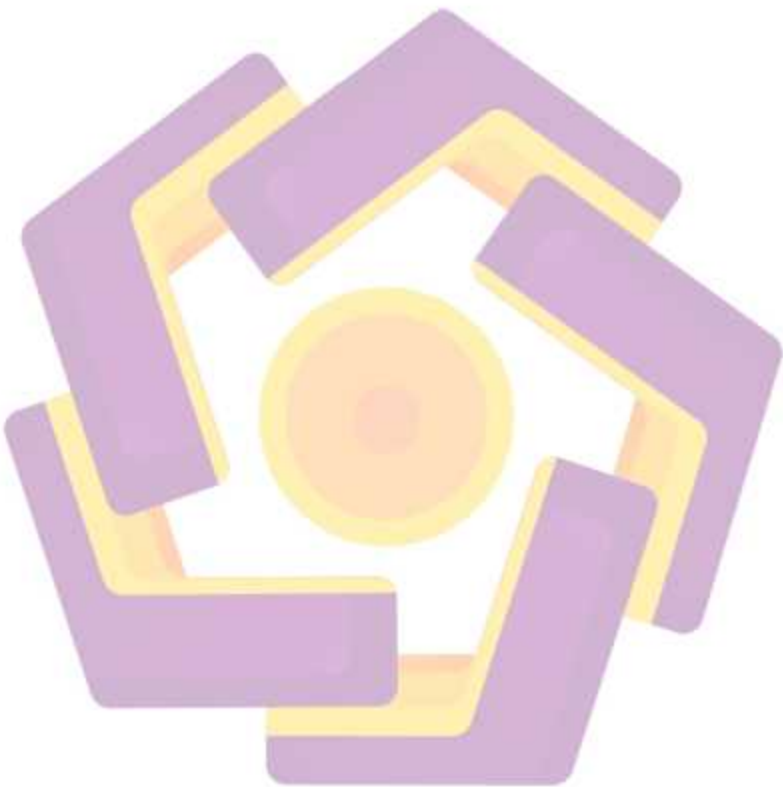
1. Ibu Anna Baita, M. Kom. selaku dosen pembimbing atas waktu, saran, dan bimbingannya dalam penyusunan skripsi ini.
2. Seluruh dosen penguji atas saran yang diberikan, sehingga skripsi ini menjadi lebih baik.
3. Keluarga penulis khususnya orang tua dan kakak penulis yang selalu memberikan dukungan dan doa kepada penulis.
4. Rekan-rekan penulis kelas Informatika 07 angkatan 2019.
5. Seluruh pihak yang telah membantu penulis dan tidak bisa penulis sebutkan satu persatu.

Penulis menyadari skripsi ini masih memiliki keterbatasan. Untuk itu penulis menerima segala kritik dan saran yang bersifat membangun. Penulis berharap skripsi ini dapat memberi manfaat untuk pembaca dan khususnya sebagai referensi untuk penelitian terkait.

Yogyakarta, 24 November 2022

Penulis

# DAFTAR ISI



# DAFTAR TABEL

# DAFTAR GAMBAR

# INTISARI

Ketidakseimbangan data merupakan tantangan dalam pembelajaran mesin. Masalah data yang tidak seimbang menyebabkan pengklasifikasi bias terhadap kelas mayoritas dalam melakukan proses klasifikasi, sehingga hasil klasifikasi tidak dapat dipercaya.

Pada penelitian ini, metode SMOTE, ADASYN, dan Borderline-SMOTE diimplementasikan untuk meningkatkan jumlah sampel pada kelas minoritas, sehingga kelas dalam keadaan yang seimbang. Kemudian, data yang dihasilkan dari proses oversampling digunakan untuk membangun model klasifikasi. Fokus penelitian ini adalah untuk meningkatkan kinerja klasifikasi Random Forest dengan menerapkan beberapa metode oversampling pada data yang tidak seimbang dan menemukan metode oversampling terbaik untuk mengatasi masalah ketidakseimbangan data dalam klasifikasi Random Forest.

Kinerja klasifikasi diukur menggunakan *accuracy, recall, f1-score,* dan *g-mean* dengan dua skenario pembagian data. Hasil penelitian menunjukkan metode oversampling dapat membantu meningkatkan performa klasifikasi Random Forest dengan rata-rata peningkatan *recall* pada metode SMOTE 42,57%, ADASYN 50,6%, dan Borderline-SMOTE 53,28%. Kemudian, peningkatan *g-mean* pada metode SMOTE 19,08%, ADASYN 20,04%, dan Borderline-SMOTE 21,21%. Berdasarkan pengujian kinerja pada keempat dataset, metode yang menghasilkan performa klasifikasi paling baik ditunjukkan oleh integrasi metode Borderline-SMOTE dan Random Forest, dengan hasil klasifikasi yang lebih tinggi dibandingkan dengan metode SMOTE dan ADASYN.

**Kata Kunci**: Ketidakseimbangan Data, Klasifikasi, Oversampling, Random Forest

# ABSTRACT

*Data imbalances are a challenge in machine learning. The problem of imbalanced data causes the classifier to be biased towards the majority class in carrying out the classification process, so the classification results are not reliable.*

*In this study, the SMOTE, ADASYN, and Borderline-SMOTE methods were implemented to increase the number of samples in the minority class, so that the classes were in a balanced state. Then, the data generated from the oversampling process is used to build the classification model. The focus of this study is to improve the classification performance of the Random Forest by applying several oversampling methods to imbalanced data and finding the best oversampling method to deal with data imbalance problems in the Random Forest classification.*

*Classification performance was measured using accuracy, recall, f1-score, and g-mean in two data splitting scenarios. The results showed that the oversampling method could help improve the performance of Random Forest classification with an average increase in recall using the SMOTE method of 42.57%, ADASYN 50.6%, and Borderline-SMOTE 53.28%. Then, the g-mean increase in the SMOTE method was 19.08%, ADASYN 20.04%, and Borderline-SMOTE 21.21%. Based on performance testing on four datasets, the method that produces the best classification performance is demonstrated by the integration of Borderline-SMOTE and Random Forest methods, with higher classification results compared to the SMOTE and ADASYN methods.*

***Keyword**: Imbalanced data, Classification, Oversampling, Random Forest*