IMPACT OF A SYNTHETIC DATA VAULT FOR IMBALANCED CLASS IN CROSS-PROJECT DEFECT PREDICTION

Skripsi

Untuk Memenuhi Persyaratan
Dalam Menyelesaikan Strata-1 Ilmu Komputer

Oleh
PUTRI NABELLA
NIM 2011016220026

PROGRAM STUDI S-1 ILMU KOMPUTER
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMBUNG MANGKURAT
BANJARBARU
APRIL 2024

IMPACT OF A SYNTHETIC DATA VAULT FOR IMBALANCED CLASS IN CROSS-PROJECT DEFECT PREDICTION

Skripsi

Untuk Memenuhi Persyaratan
Dalam Menyelesaikan Strata-1 Ilmu Komputer

Oleh
PUTRI NABELLA
NIM 2011016220026

PROGRAM STUDI S-1 ILMU KOMPUTER
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMBUNG MANGKURAT
BANJARBARU
APRIL 2024

SKRIPSI

IMPACT OF A SYNTHETIC DATA VAULT FOR IMBALANCED CLASS IN CROSS-PROJECT DEFECT PREDICTION

Oleh:

PUTRI NABELLA

NIM. 2011016220026

Telah dipertahankan di depan Dosen Penguji pada tanggal 19 April 2024.

Susunan Penguji:

Pembimbing Utama

Rudy Herteno, S.Kom., M.Kom
NIP. 198809252022031003

Penguji I

Mohammad Reza Faisal S.Si., S.T., M.T., PhD
NIP. 197612202008121001

Pembimbing Pendamping

Setyo Wahyu Saputro, S.Kom., M.Kom
NIP. 198808072023211027

Penguji II

Friska Abadi S.Kom., M.Kom
NIP. 198809132023211010

Banjarbaru, 19 April 2024
Ketua Program Studi Ilmu Komputer

Irwan Budiman, S.T., M.Kom
NIP. 197703252008121001

KATA PENGANTAR

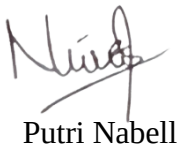
Puji syukur saya panjatkan pada Allah SWT karena berkat kasih, rahmat dan karunia-Nya penulis dapat menyelesaikan jurnal yang berjudul "*Impact of a Synthetic Data Vault for Imbalanced Class in Cross-Project Defect Prediction*" untuk memenuhi syarat dalam menyelesaikan pendidikan program S1 Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Lambung Mangkurat.

Pada lembar ini penulis ingin menyampaikan ucapan terima kasih kepada semua pihak yang sangat mendukung penulis dalam pembuatan dan penyusunan jurnal ini, adapun yang dimaksud adalah sebagai berikut:

1. Kedua orang tua, Samsi dan Rusinah, yang selalu menguatkan, memberikan doa, semangat, dukungan dan optimis bahwa jurnal ini pasti dapat selesai dengan baik.
2. Terima kasih kepada adik-adik saya, Putri Salsabila, Hana Aida Sahila, dan Budi Firmansyah Habibi, atas dukungan dan semangat yang mereka berikan.
3. Bapak Rudy Herteno, S.Kom., M.Kom. selaku dosen pembimbing utama dan Bapak Setyo Wahyu Saputro, S.Kom., M.Kom. selaku dosen pembimbing pendamping yang turut serta membantu dan meluangkan waktu demi kelancaran dalam penyelesaian jurnal ini.
4. Bapak Irwan Budiman, S. T., M. Kom. selaku Ketua Program Studi Ilmu Komputer FMIPA ULM, serta seluruh Dosen.
5. Staf Prodi Ilmu Komputer yaitu Ka Azizah atas bantuan dan arahan dalam pemberkasan dan hal-hal lainnya.
6. Muhammad Yoga Adha Pratama yang telah memberikan semangat, dukungan, serta membantu dari awal penelitian sehingga jurnal ini bisa terselesaikan dengan baik.
7. Teman-teman Garda Terdepan yang telah menjadi teman seperjuangan sejak awal kuliah hingga saat ini.
8. Serta semua pihak yang tidak dapat disebutkan satu persatu yang telah turut membantu dalam penyelesaian jurnal ini.

Akhir kata penulis menyadari sepenuhnya bahwa penulisan ini jauh dari sempurna. Meski demikian, diharapkan tulisan ini dapat memberikan manfaat bagi ilmu pengetahuan dan pembaca, serta mendapatkan berkah yang melimpah dari Tuhan Yang Maha Esa.

Banjarbaru, 22 April 2024



Putri Nabella

ABSTRAK

PENGARUH DARI SYNTHETIC DATA VAULT UNTUK MENANGANI KELAS TIDAK SEIMBANG PADA CORSS-PROJECT DEFECT PREDICTION

(Oleh: Putri Nabella; Pembimbing : Rudy Herteno, S.Kom., M.Kom. dan Setyo Wahyu Saputro, S.Kom., M.Kom.; 2024; halaman)

Software Defect Prediction (SDP) sangat penting untuk memastikan kualitas perangkat lunak. Namun, ketidakseimbangan kelas (CI) menimbulkan tantangan yang signifikan dalam pemodelan prediktif. Penelitian ini memperkenalkan pendekatan baru dengan menggunakan *Synthetic Data Vault* (SDV) untuk menangani CI dalam *Cross-Project Defect Prediction* (CPDP). Secara metodologis, penelitian ini membahas CI di beberapa set data (ReLink, MDP, dan PROMISE) dengan memanfaatkan SDV untuk menambah kelas minoritas. Klasifikasi menggunakan *Decision Tree* (DT), *Logistic Regression* (LR*), K-Nearest Neighbors* (KNN), *Naive Bayes* (NB), *dan Random Forest* (RF), serta kinerja model dievaluasi menggunakan AUC dan *t-Test*. Hasilnya secara konsisten menunjukkan bahwa SDV memiliki kinerja yang lebih baik daripada SMOTE dan teknik lainnya dalam berbagai proyek. Keunggulan ini terbukti melalui peningkatan yang signifikan secara statistik. Dominasi KNN pada hasil rata-rata AUC, dengan nilai 0.695, 0.704, dan 0.750. Pada ReLink, KNN menunjukkan peningkatan 16,06% dibandingkan *imbalanced* dan 12,84% dibandingkan SMOTE. Demikian pula, pada MDP, KNN menunjukkan peningkatan 20,71% dibandingkan dengan *imbalanced* dan 10,16% dibandingkan dengan SMOTE. Selain itu, pada PROMISE, KNN 13,55% lebih baik dari *imbalanced* dan 7,01% dari SMOTE. RF menunjukkan kinerja yang moderat, diikuti oleh LR dan DT, sementara NB tertinggal di belakang. Secara keseluruhan, SDV mendapatkan peningkatan 10,10% dari *imbalanced*, dan 7,54% dari SMOTE. Signifikansi statistik dari temuan ini dikonfirmasi oleh *t-Test*, semuanya di bawah ambang batas 0,05. Implikasi praktis dari penerapan SDV untuk deteksi cacat dan mitigasi CI terletak pada efektivitasnya yang telah terbukti, terutama dengan KNN sebagai algoritma klasifikasi terbaik, yang menunjukkan potensi yang menjanjikan untuk meningkatkan kualitas perangkat lunak dengan mengatasi CI dan meningkatkan hasil pemodelan prediktif.

Kata kunci: *Class Imbalance, Cross Project Defect Prediction, Machine Learning, Software Defect Prediction, Synthetic Data Vault*

ABSTRACT

IMPACT OF A SYNTHETIC DATA VAULT FOR IMBALANCED CLASS IN CROSS-PROJECT DEFECT PREDICTION

(By: Putri Nabella; Supervisors: Rudy Herteno, S.Kom., M.Kom. and Setyo Wahyu Saputro, S.Kom., M.Kom.; 2024; page)

Software Defect Prediction (SDP) is crucial for ensuring software quality. However, class imbalance (CI) poses a significant challenge in predictive modeling. This study introduces a novel approach by employing the Synthetic Data Vault (SDV) to tackle CI within Cross-Project Defect Prediction (CPDP). Methodologically, the study addresses CI across multiple datasets (ReLink, MDP, and PROMISE) by leveraging SDV to augment minority classes. Classification utilizing Decision Tree (DT), Logistic Regression (LR), K-Nearest Neighbors (KNN), Naive Bayes (NB), and Random Forest (RF), also model performance is evaluated using AUC and t-Test. The results consistently show that SDV performs better than SMOTE and other techniques in various projects. This superiority is evident through statistically significant improvements. KNN dominance in average AUC results, with values 0.695, 0.704, and 0.750. On ReLink, KNN show 16.06% improvement over the imbalanced and 12.84% over SMOTE. Similarly, on MDP, KNN 20.71% improvement over the imbalanced and a 10.16% over SMOTE. Moreover, on PROMISE, KNN 13.55% improvement over the imbalanced and 7.01% over SMOTE. RF displays moderate performance, closely followed by LR and DT, while NB lags behind. Overall, SDV got an improvement of 10.10% from imbalanced, and 7.54% from SMOTE. The statistical significance of these findings is confirmed by t-Test, all below the 0.05 threshold. The practical implication of adopting SDV for defect detection and CI mitigation lies in its demonstrated effectiveness, particularly with KNN as the best classification algorithm, showcasing promising potential to enhance software quality by addressing CI and improving predictive modeling outcomes.

Keywords: *Class Imbalance, Cross Project Defect Prediction, Machine Learning, Software Defect Prediction, Synthetic Data Vault*

SURAT PERNYATAAN

Dengan ini saya menyatakan bahwa dalam jurnal ini tidak terdapat karya yang pernah diajukan untuk memperoleh gelar kesarjanaan di suatu Perguruan Tinggi, dan sepanjang pengetahuan saya juga tidak terdapat karya atau pendapat yang pernah ditulis atau diterbitkan oleh orang lain, kecuali yang secara tertulis diacu dalam naskah ini dan disebutkan dalam Daftar Pustaka.

Banjarbaru, 22 April 2024
Yang Menyatakan,

Putri Nabella
NIM. 2011016220026