



**IMPLEMENTASI PERBANDINGAN METRIK JARAK PADA
ALGORITMA KNN DENGAN PENERAPAN SMOTE UNTUK
PREDIKSI CACAT *SOFTWARE***

Skripsi

**Untuk Memenuhi Persyaratan Dalam Menyelesaikan Sarjana Strata-
1 Ilmu Komputer**

Oleh

KHUSNUL RAHMI MAULIDHA

NIM. 2011016220011

**PROGRAM STUDI S1 ILMU KOMPUTER
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMBUNG MANGKURAT BANJARBARU
JULI 2024**

SKRIPSI

IMPLEMENTASI PERBANDINGAN METRIK JARAK PADA ALGORITMA KNN DENGAN PENERAPAN SMOTE UNTUK PREDIKSI CACAT *SOFTWARE*

Oleh:

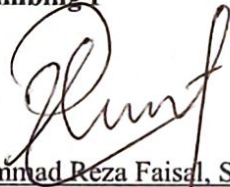
Khusnul Rahmi Maulidha

2011016220011

Telah dipertahankan di depan Dosen Penguji pada tanggal 22 Juli 2024

Susunan Dosen penguji:

Pembimbing I



Mohammad Reza Faisal, S.T, M.T, Ph.D.

NIP. 197612202008121001

Penguji I



Friska Abadi, S.Kom., M.Kom.

NIP. 198809132023211010

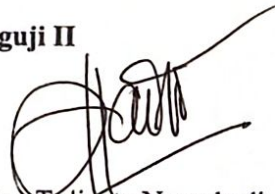
Pembimbing II



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

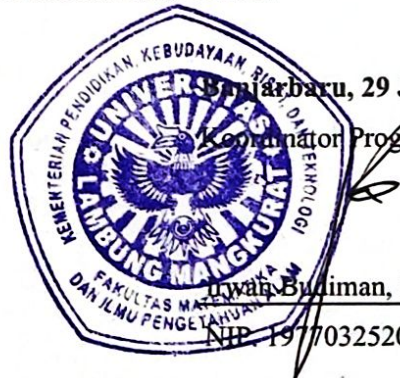
NIP. 198808072023211027

Penguji II



Dodon Turiyanto Nugrahadi, S.Kom, M.Eng.

NIP. 198001122009121002



Barbaru, 29 Juli 2024

Koordinator Program Studi Ilmu Komputer

Irwah Budiman, S.T M.Kom.

NIP. 197703252008121001

PERNYATAAN

Dengan ini saya menyatakan bahwa dalam skripsi 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, 10 Juli 2024



Khusnul Rahmi Maulidha

NIM. 2011016220011

ABSTRAK

IMPLEMENTASI PERBANDINGAN METRIK JARAK PADA ALGORITMA KNN DENGAN PENERAPAN SMOTE UNTUK PREDIKSI CACAT *SOFTWARE*

(Oleh: Khusnul Rahmi Maulidha; Pembimbing: Mohammad Reza Faisal, S.T, M.T, Ph.D dan Setyo Wahyu Saputro, S.Kom., M.Kom; 2024; 64 halaman)

Seiring meningkatnya kompleksitas dan skala proyek, muncul tantangan baru terkait penanganan cacat pada *software*. Pendekatan untuk mengatasi masalah ini adalah menggunakan teknik prediksi cacat *software* dengan metode *machine learning*. Salah satu contoh algoritma klasifikasi *machine learning* adalah KNN. Kinerja KNN bergantung pada ukuran jarak/kemiripan dan aturan mayoritas dalam menentukan hasil akhir klasifikasinya. Metode klasifikasi berbasis mayoritas ini memperburuk kinerja KNN ketika diterapkan pada data yang memiliki masalah ketidakseimbangan. Dalam penelitian ini, peneliti membandingkan kinerja metrik jarak euclidian, hamming, cosine, dan canberra pada KNN sebelum dan sesudah penerapan SMOTE. Hasil penelitian menunjukkan nilai AUC dan F-1 Measure terbaik pada dataset EQ diperoleh oleh jarak euclidian sebesar 0,7571 dan 0,7079 (tanpa SMOTE) serta 0,7752 dan 0,7311 (dengan SMOTE), dataset JDT pada jarak euclidian sebesar 0,7331 dan 0,6010 (tanpa SMOTE) serta jarak canberra 0,7707 dan 0,6342 (dengan SMOTE), dataset LC pada jarak cosine sebesar 0,6161 dan 0,3029 (tanpa SMOTE) serta jarak canberra 0,6752 dan 0,3733 (dengan SMOTE), dataset ML pada jarak canberra sebesar 0,6435 dan 0,4006 (tanpa SMOTE) serta 0,6845 dan 0,4261 (dengan SMOTE), dataset PDE pada jarak canberra sebesar 0,5995 dan 0,3139 (tanpa SMOTE) serta 0,6580 dan 0,3957 (dengan SMOTE). Penerapan SMOTE secara signifikan dapat meningkatkan kinerja AUC dan F-1 Measure pada KNN dengan P value sebesar 0,0001.

Kata kunci: Prediksi Cacat *Software*, Euclidian, Hamming, Cosine, Canberra, SMOTE, Algoritma KNN

ABSTRAC

IMPLEMENTATION OF DISTANCE METRIC COMPARISON IN KNN ALGORITHM WITH SMOTE APPLICATION FOR SOFTWARE DEFECT PREDICTION

(By: Khusnul Rahmi Maulidha; Supervisor: Mohammad Reza Faisal, S.T, M.T, Ph.D and Setyo Wahyu Saputro, S.Kom,. M.Kom; 2024; 64 pages)

As the complexity and scale of projects increase, new challenges arise related to handling software defects. An approach to overcome this problem is to use software defect prediction techniques with machine learning methods. One example of a machine learning classification algorithm is KNN. The performance of KNN depends on the distance/similarity measure and the majority rule in determining the final classification result. This majority vote-based classification method worsens the performance of KNN when applied to data that has imbalance problems. In this research, researcher compared the performance of euclidian, hamming, cosine, and canberra distance metrics on KNN before and after the application of SMOTE. The result shows the AUC and F-1 Measure values on EQ dataset obtained by euclidian distance of 0.7571 and 0.7079 (without SMOTE) and 0.7752 and 0.7311 (with SMOTE), JDT dataset on euclidian distance of 0.7331 and 0.6010 (without SMOTE) and canberra distance of 0.7707 and 0.6342 (with SMOTE), LC dataset on cosine distance of 0.6161 and 0.3029 (without SMOTE) and canberra distance of 0.6752 and 0.3733 (with SMOTE), ML dataset on canberra distance of 0.6435 and 0.4006 (without SMOTE) and 0.6845 and 0.4261 (with SMOTE), PDE dataset on canberra distance of 0.5995 and 0.3139 (without SMOTE) and 0.6580 and 0.3957 (with SMOTE). The application of SMOTE can significantly improve the performance of the AUC and F-1 measures on KNN with a P value of 0.0001.

Keyword:Software Defect Prediction, Euclidian, Hamming, Cosine, Camberra, SMOTE, KNN Algorithm

PRAKATA

Puji Syukur penulis panjatkan kepada Tuhan Yang Maha Esa, karena atas berkat rahmat dan karunia-Nya penulis dapat menyelesaikan skripsi yang berjudul “Implementasi Perbandingan Metrik Jarak pada Algoritma KNN dengan Penerapan SMOTE untuk Prediksi Cacat *Software*” untuk memenuhi syarat dalam menyelesaikan Pendidikan program S1 Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Lambung Mangkurat. Tak lupa shalawat serta salam penulis tuturkan kepada Nabi Muhammad SAW, yang telah membawa umat manusia dari alam kegelapan menuju cahaya iman dan ilmu pengetahuan.

Pada lembar ini, penulis ingin menyampaikan ucapan terima kasih kepada pihak-pihak yang sangat mendukung penulis dalam pembuatan dan penyusunan skripsi ini. Adapun orang yang dimaksud adalah sebagai berikut:

1. Orang tua dan keluarga yang selalu memberikan semangat, dukungan, dan doa dari awal perkuliahan hingga proses penyelesaian skripsi ini.
2. Bapak Mohammad Reza Faisal, S.T, M.T, Ph.D selaku dosen pembimbing utama yang turut serta membantu dan meluangkan waktu demi kelancaran penyelesaian skripsi ini.
3. Bapak Setyo Wahyu Saputro, S.Kom M.Kom. dosen pembimbing pendamping yang turut serta membantu dan meluangkan waktu demi kelancaran penyelesaian skripsi ini.
4. Bapak Irwan Budiman, S.T M.Kom. selaku Koordinator Program Studi Ilmu Komputer FMIPA ULM, atas bantuan dan izin beliau skripsi ini dapat diselesaikan.
5. Seluruh dosen dan staff Program Studi Ilmu Komputer atas ilmu dan bantuan bermanfaat yang diberikan selama ini.
6. Teman-teman dan sahabat Ilmu Komputer Angkatan 2020 yang senantiasa memberikan dukungan dan bantuan selama perkuliahan dan penyelesaian skripsi ini.
7. Semua pihak yang tidak bisa disebutkan satu persatu yang juga telah turut membantu dalam penyelesaian skripsi ini.

Akhir kata, penulis menyadari bahwa penulisan ini masih jauh dari sempurna. Namun, penulis mengharapkan bantuan berupa kritik dan saran yang membangun dari semua pihak demi kesempurnaan mutu penulisan skripsi ini. Semoga tulisan ini dapat bermanfaat bagi ilmu pengetahuan dan pembaca khususnya, serta mendapatkan keridhaan Allah SWT.

Banjarbaru, 10 Juli 2024



Khusnul Rahmi Maulidha

DAFTAR ISI

HALAMAN JUDUL	i
LEMBAR PENGESAHAN	ii
PERNYATAAN.....	iii
ABSTRAK	iv
ABSTRAC.....	v
PRAKATA.....	vi
DAFTAR ISI.....	viii
DAFTAR TABEL.....	x
DAFTAR GAMBAR.....	xii
DAFTAR LAMPIRAN	xiii
BAB I PENDAHULUAN.....	1
1.1 Latar Belakang	1
1.2 Rumusan Masalah	2
1.3 Batasan Masalah.....	2
1.4 Tujuan Penelitian	3
1.5 Manfaat Penelitian	3
BAB II TINJAUAN PUSTAKA.....	4
2.1 Kajian terdahulu	4
2.2 Prediksi Cacat <i>Software</i>	8
2.3 Dataset AEEEM.....	9
2.4 SMOTE	10
2.5 K-NN.....	11

2.6 Metrik Jarak	12
2.7 K-Fold Cross Validation	13
2.8 Evaluasi	13
2.9 T-Test	15
BAB III METODE PENELITIAN	17
3.1 Alat Penelitian	17
3.2 Bahan Penelitian	17
3.3 Variabel Penelitian	17
3.4 Prosedur Penelitian	17
BAB IV HASIL DAN PEMBAHASAN.....	20
4.1 Hasil	20
4.1.1 Pengumpulan Dataset	20
4.1.2 Pembagian Dataset	21
4.1.3 <i>Resampling Data</i>	24
4.1.4 Klasifikasi Tanpa SMOTE	28
4.1.5 Klasifikasi Menggunakan SMOTE	36
4.1.6 Perbandingan Hasil Kinerja KNN Menggunakan SMOTE dan Tanpa SMOTE	44
4.2 Pembahasan	49
BAB V PENUTUP	59
5.1 Kesimpulan	59
5.2 Saran	60
DAFTAR PUSTAKA	61
LAMPIRAN.....	65
RIWAYAT HIDUP	72

DAFTAR TABEL

Tabel 1. Keaslian Penelitian.....	7
Tabel 2. Perancangan Penelitian	8
Tabel 3. Spesifikasi Dataset AEEEM	9
Tabel 4. Keakuratan hasil klasifikasi berdasarkan AUC	14
Tabel 5. Confusion Matrix	15
Tabel 6. Spesifikasi Dataset AEEEM	20
Tabel 7. Pembagian Dataset EQ.....	21
Tabel 8. Pembagian Dataset JDT	22
Tabel 9. Pembagian Dataset LC.....	22
Tabel 10. Pembagian Dataset ML.....	23
Tabel 11. Pembagian Dataset PDE	24
Tabel 12. Dataset EQ Sebelum dan Sesudah <i>Resampling</i>	25
Tabel 13. Dataset JDT Sebelum dan Sesudah <i>Resampling</i>	25
Tabel 14. Dataset LC Sebelum dan Sesudah <i>Resampling</i>	26
Tabel 15. Dataset ML Sebelum dan Sesudah <i>Resampling</i>	27
Tabel 16. Dataset ML Sebelum dan Sesudah <i>Resampling</i>	28
Tabel 17. Hasil Klasifikasi Dataset EQ Tanpa SMOTE	29
Tabel 18. Hasil Klasifikasi Dataset JDT Tanpa SMOTE	30
Tabel 19. Hasil Klasifikasi Dataset LC Tanpa SMOTE	32
Tabel 20. Hasil Klasifikasi Dataset ML Tanpa SMOTE	33
Tabel 21. Hasil Klasifikasi Dataset PDE Tanpa SMOTE.....	35
Tabel 22. Hasil Klasifikasi Dataset EQ dengan SMOTE	37
Tabel 23. Hasil Klasifikasi Dataset JDT dengan SMOTE.....	38
Tabel 24. Hasil Klasifikasi Dataset LC dengan SMOTE.....	40
Tabel 25. Hasil Klasifikasi Dataset ML dengan SMOTE.....	41
Tabel 26. Hasil Klasifikasi Dataset PDE dengan SMOTE	43
Tabel 27. Nilai AUC terbaik pada metrik jarak Euclidian.....	50
Tabel 28. Nilai F-1 Measure terbaik pada metrik jarak Euclidian.....	50
Tabel 29. Nilai AUC terbaik pada metrik jarak Hamming	51

Tabel 30. Nilai F-1 Measure terbaik pada metrik jarak Hamming	52
Tabel 31. Nilai AUC terbaik pada metrik jarak Cosine	53
Tabel 32. Nilai F-1 Measure terbaik pada metrik jarak Cosine	53
Tabel 33. Nilai AUC terbaik pada metrik jarak Canberra	54
Tabel 34. Nilai F-1 Measure terbaik pada metrik jarak Canberra.....	54
Tabel 35. Hasil AUC dan F-1 Measure terbaik tiap dataset	56

DAFTAR GAMBAR

Gambar 1. Grafik ROC	13
Gambar 2. Diagram Alur Penelitian.....	18
Gambar 3. Perbandingan AUC Dataset EQ	45
Gambar 4. Perbandingan F-1 Measure Dataset EQ	45
Gambar 5. Perbandingan AUC Dataset JDT.....	46
Gambar 6. Perbandingan F-1 Measure Dataset JDT.....	46
Gambar 7. Perbandingan AUC Dataset LC	46
Gambar 8. Perbandingan F-1 Measure Dataset LC	47
Gambar 9. Perbandingan AUC Dataset ML	48
Gambar 10. Perbandingan F-1 Measure Dataset ML.....	48
Gambar 11. Perbandingan AUC Dataset PDE.....	49
Gambar 12. Perbandingan F-1 Measure Dataset PDE.....	49
Gambar 13. Keseluruhan nilai AUC dan F-1 Measure Metrik Jarak Euclidian.....	51
Gambar 14. Keseluruhan nilai AUC dan F-1 Measure Metrik Jarak Hamming.....	52
Gambar 15. Keseluruhan nilai AUC dan F-1 Measure Metrik Jarak Cosine	54
Gambar 16. Keseluruhan nilai AUC dan F-1 Measure Metrik Jarak Canberra	55
Gambar 17. Hasil <i>T-Test</i> AUC.....	57
Gambar 18. Hasil <i>T-Test</i> F-1 Measure.....	57

DAFTAR LAMPIRAN

Lampiran 1. Source Code.....	66
------------------------------	----