



**IMPLEMENTASI KLASIFIKASI *CATBOOST* DENGAN MENGGUNAKAN
HYPER-PARAMETER TUNING BAYESIAN SEARCH UNTUK
MEMPREDIKSI PENYAKIT DIABETES**

Skripsi

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

Oleh

**ARIF DARMAWAN
NIM 171101621002**

**PROGRAM STUDI S-1 ILMU KOMPUTER
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMBUNG MANGKURAT
BANJARBARU
SEPTEMBER 2023**



**IMPLEMENTASI KLASIFIKASI *CATBOOST* DENGAN MENGGUNAKAN
HYPER-PARAMETER TUNING BAYESIAN SEARCH UNTUK
MEMPREDIKSI PENYAKIT DIABETES**

Skripsi

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

Oleh

ARIF DARMAWAN

NIM 1711016210002

**PROGRAM STUDI S-1 ILMU KOMPUTER
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMBUNG MANGKURAT
BANJARBARU
SEPTEMBER 2023**

SKRIPSI

**IMPLEMENTASI KLASIFIKASI *CATBOOST* DENGAN MENGGUNAKAN
HYPER-PARAMETER TUNING BAYESIAN SEARCH UNTUK
MEMPREDIKSI PENYAKIT DIABETES**

Oleh:
ARIF DARMAWAN
NIM 1711016210002

Telah dipertahankan di depan Dosen Penguji pada tanggal 06 September 2023.

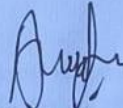
Susunan Dosen Penguji:

Pembimbing I

Dosen Penguji I



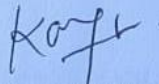
Muliadi, S.Kom., M.Cs.
NIP. 197804222010121002



Triando Hamonangan Saragih, S.Kom., M.Kom.
NIP. 199308242019031012

Pembimbing II

Dosen Penguji II



Dwi Kartini, S.Kom., M.Kom
NIP. 198704212012122003



Raditvo Adi Nugroho, S.T., M.Kom.
NIP. 198212042008011006

Banjarnegara 06 September 2023
Koordinator Program Studi Ilmu Komputer



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, Juli 2022

Yang Menyatakan,



Arif Darmawan

NIM.1711016210002

ABSTRAK

IMPLEMENTASI KLASIFIKASI *CATBOOST* DENGAN MENGGUNAKAN *HYPER-PARAMETER TUNING BAYESIAN SEARCH* UNTUK MEMPREDIKSI PENYAKIT DIABETES

(Oleh : Arif Darmawan; Pembimbing: Muliadi S.Kom, M.Cs dan Dwi Kartini S.Kom., M.Kom.; 2022; 77 halaman)

Diabetes merupakan masalah kesehatan masyarakat dunia dengan prevalensi yang selalu meningkat setiap tahun. Penyakit Diabetes ini perlu didiagnosis sejak dini menggunakan algoritma klasifikasi. Dataset yang digunakan yaitu *PIMA Indians Diabetes Database* dari *Kaggle* dengan 768 data dan 8 fitur. Metode pengklasifikasi yang digunakan yaitu *CatBoost*. Klasifikasi *CatBoost* dapat bekerja baik dalam menangani ketidak seimbangan data, namun kinerja algoritma ini masih bisa ditingkatkan lagi. Untuk mengatasi permasalahan tersebut peneliti menggunakan solusi *Hyper-parameter tuning*. *CatBoost* memiliki beberapa *Hyper-parameter* yang dapat dikonfigurasi untuk meningkatkan kinerja dari model. Untuk mengidentifikasi nilai yang baik untuk *Hyper-parameter* disebut *Hyper-parameter tuning*. Metode *Hyper-parameter tuning* yang digunakan yaitu *Bayesian Search* yang kemudian divalidasi menggunakan *10-Fold Cross Validation* sebanyak 10 iterasi. *Hyper-parameter CatBoost* yang dikonfigurasi antara lain *depth* (1,9), *learning_rate* (0.01, 1.00) dan *Iterations* (1,500). Penelitian ini bertujuan untuk mengetahui AUC dan presisi pada masing masing model. Pengujian pada *CatBoost* tanpa *Hyper-parameter tuning* memperoleh nilai AUC sebesar 0,859 dan presisi sebesar 62,5%. Untuk pengujian *CatBoost* dengan *Hyper-parameter tuning* memperoleh AUC sebesar 0,906 dan presisi sebesar 63,46%. Menambahkan *Hyper-parameter tuning Bayesian Search* pada metode klasifikasi *CatBoost* dapat meningkatkan hasil nilai AUC dan nilai presisi.

Kata Kunci : Klasifikasi; *CatBoost*; AUC; *Hyper-parameter Tuning*; *Bayesian Search*

ABSTRACT

IMPLEMENTATION OF CATBOOST CLASSIFICATION USING BAYESIAN SEARCH HYPER-PARAMETER TUNING TO PREDICT DIABETES

(By : Arif Darmawan; Supervisor: Muliadi S.Kom, M.Cs and Dwi Kartini S.Kom., M.Kom.; 2022; 77 pages)

Diabetes is a global public health problem with a continuously increasing prevalence every year. Early diagnosis of this disease using classification algorithms is crucial. The dataset used is the PIMA Indians Diabetes Database from Kaggle, which consists of 768 data points and 8 features. The classification method employed is *CatBoost*. *CatBoost* classification performs well in handling imbalanced data, but its algorithm's performance can still be further improved. To address this issue, researchers utilized Hyper-parameter tuning as a solution. *CatBoost* has several Hyper-parameters that can be configured to enhance the model's performance. Determining good values for these Hyper-parameters is known as Hyper-parameter tuning. *Bayesian Search* was the Hyper-parameter tuning method employed, which was then validated using 10-Fold Cross Validation over 10 iterations. The configured Hyper-parameters for *CatBoost* included depth (1,9), learning_rate (0.01, 1.00), and Iterations (1,500). The aim of this study was to evaluate the AUC, accuracy, and precision for each model. Testing *CatBoost* without Hyper-parameter tuning resulted in an AUC of 0.859, an accuracy of 71.43%, and a precision of 62.5%. On the other hand, testing *CatBoost* with Hyper-parameter tuning yielded an AUC of 0.906, an accuracy of 74.03%, and a precision of 63.46%. Adding *Bayesian Search* Hyper-parameter tuning to the *CatBoost* classification method can improve the AUC value, accuracy, and precision

Kata Kunci : *Classification; CatBoost; Accuracy; AUC; Hyper-parameter Tuning; Bayesian Search*

PRAKATA

Puji syukur penulis panjatkan ke Tuhan kita Yang Maha Esa karena atas berkat rahmat dan karunia-Nya penulis dapat menyelesaikan skripsi yang berjudul **“IMPLEMENTASI KLASIFIKASI *CATBOOST* DENGAN MENGGUNAKAN *HYPER-PARAMETER TUNING BAYESIAN SEARCH* UNTUK MEMPREDIKSI PENYAKIT DIABETES”** 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 terimakasih kepada pihak-pihak yang sangat mendukung penulis dalam pembuatan dan penyusunan skripsi ini, adapun yang dimaksud adalah sebagai berikut :

1. Keluarga yang selalu memberikan bantuan, semangat, doa dan dukungan dalam proses penyelesaian skripsi ini, terutama untuk Bapak, Ibu, Dea dan Viona.
2. Bapak Muliadi, S.Kom, M.Cs. selaku dosen pembimbing utama yang turut serta membantu dan meluangkan waktu demi kelancaran dalam penyelesaian skripsi ini.
3. Ibu Dwi Kartini, S.Kom, M.Kom. selaku dosen pembimbing pendamping yang turut serta membantu dan meluangkan waktu demi kelancaran dalam penyelesaian skripsi ini.
4. Bapak Irwan Budiman S.T., M.Kom selaku Ketua Program Studi Ilmu Komputer FMIPA ULM, atas bantuan dan izin beliau skripsi ini dapat diselesaikan.
5. Seluruh Dosen dan staf Program Studi Ilmu Komputer FMIPA ULM atas ilmu dan bantuan yang diberikan selama ini yang sangat bermanfaat.
6. Teman-teman dan sahabat-sahabat keluarga Ilmu Komputer angkatan 2017 yang memberikan dukungan dan selalu mengingatkan serta mendoakan dalam proses mengerjakan skripsi.
7. Ucapan terima kasih kepada Intan, Ryan, Barok, Shega, Bowo, Aldo, Irvan, Ubai, Ilham, Zamzam, Yudhit, Fauzi, Haikal, Said, Eddy, Azmi, Luthfi, Didit, Rizqon, Reza,

Muflih, Noval, Apri, Iastri, yang memberikan dukungan dalam proses mengerjakan skripsi.

8. Semua pihak yang tidak dapat disebutkan satu persatu yang telah turut membantu dalam penyelesaian skripsi ini.

Akhir kata penulis menyadari sepenuhnya bahwa penulisan ini jauh dari sempurna, namun penulis mengharapkan bantuan serupa berupa saran dan kritik yang membangun dari semua pihak demi kesempurnaan dan mutu penulisan skripsi ini.

Semoga tulisan ini dapat bermanfaat bagi ilmu pengetahuan dan pembacakhususnya serta mendapat keridhaan Allah SWT.

Banjarbaru, 06 September 2023



Arif Darmawan

DAFTAR ISI

ABSTRAK	4
ABSTRACT	5
PRAKATA	6
DAFTAR ISI	8
BAB I PENDAHULUAN	
1.1 Latar Belakang	1
1.2 Rumusan Masalah	3
1.3 Batasan Masalah	3
1.4 Tujuan	4
1.5 Manfaat Penelitian	4
BAB II TINJAUAN PUSTAKA	
2.1 Kajian Terdahulu	5
2.2 Landasan Teori	12
2.2.1. Diabetes Melitus	12
2.2.2. Dataset PIMA Indians Diabetes	12
2.2.3. Ketidakseimbangan Kelas	12
2.2.4. Outlier Removal	13
2.2.6. Decision Tree	14
2.2.7. Boosting	15
2.2.8. Categorical Boosting	15
2.2.9. Bayesian Search	17
2.2.11. Hyper-parameters Tuning	19
2.2.12. Cross Validation	19
2.2.13. Area Under Curve	20

2.2.14. Confusion Matrix.....	21
BAB III METODE PENELITIAN	
3.1 Bahan Penelitian	23
3.2 Alat Penelitian.....	23
3.3 Variabel Penelitian.....	23
3.4 Prosedur Penelitian	24
BAB IV HASIL DAN PEMBAHASAN	
4.1 Hasil	27
4.1.1 Pengumpulan Dataset	27
4.1.2 Preprocessing Data	28
4.1.2.1 <i>Outlier Removal</i>	28
4.1.2.2 Pembagian Data <i>Training</i> dan Data <i>Testing</i>	36
4.1.3 Hyper-parameter Tuning	36
4.1.4 Klasifikasi	37
4.1.4.1 Klasifikasi <i>CatBoost</i>	37
4.1.5 Evaluasi Klasifikasi CatBoost Tanpa Tuning.....	53
4.1.6 Evaluasi Klasifikasi CatBoost Menggunakan Bayesian Search....	54
4.2 Pembahasan.....	56
BAB V PENUTUP	
5.1 Kesimpulan	58
5.2 Saran	58
DAFTAR PUSTAKA	

DAFTAR TABEL

Tabel 1. Dataset Hasil Penelitian	5
Tabel 2. perbandingan hasil penelitian ibrahim A.	7
Tabel 3 Hasil data CatBoost dengan <i>Hyperparameter</i>	8
Tabel 4. Keaslian Penelitian.....	9
Tabel 5. Atribut dan Outcome.....	12
Tabel 6. Kelebihan dan Kekurangan Decision Tree.	15
Tabel 7. Parameter CatBoost	19
Table 8. Stratified 10-Fold Cross Validation	20
Tabel 9. Confusion matrix.....	21
Tabel 10. Dataset Diabetes.....	25
Tabel 11. Nama Fitur	27
Tabel 12. Contoh PIMA Indians Diabetes Database	27
Tabel 13. Jumlah Distribusi Kelas	28
Tabel 14. Nilai <i>Hyperparameter</i> Default CatBoost.....	37
Tabel 15 <i>Confusion Matrix</i> Klasifikasi <i>CatBoost</i>	37
Tabel 16. Fitur yang sudah terurut	43
Table 17. setelah penambahan residual.....	43
Tabel 18. Domain pencarian <i>Hyperparameter</i> CatBoost dengan Bayesian Search... 46	
Tabel 19. Kandidat <i>Hyperparameter</i> Bayesian Search.....	47
Tabel 20 <i>Confusion Matrix</i> Klasifikasi <i>CatBoost</i> dengan <i>Bayesian Search</i>	47
Tabel 21. Hasil evaluasi parameter Iteration.....	54
Tabel 22. Hasil evaluasi parameter Depth	55
Tabel 23. Hasil evaluasi parameter Learning Rate	55
Tabel 24. Perbedaan Konfigurasi <i>Hyperparameter</i>	59

DAFTAR TABEL

Table 25. Hasil AUC dan presisi.....	60
--------------------------------------	----

DAFTAR GAMBAR

Gambar 1. Struktur Decission Tree.....	14
Gambar 2. Contoh Oblivious Decision Tree.....	16
Gambar 3. Algoritma Pembuatan Pohon di CatBoost	17
Gambar 4. Ilustrasi Bayesian	18
Gambar 5. Contoh Kurva AUC.....	21
Gambar 6. Alur Peneliti	24
Gambar 7. Perbandingan Presentase Pasien Diabetes dan Bukan Pasien Diabetes ...	28
Gambar 8. <i>Outlier removal</i> pada fitur pregnancies sebelum dihapus	29
Gambar 9. <i>Outlier removal</i> pada fitur pregnancies setelah dihapus	29
Gambar 10. <i>Outlier removal</i> pada fitur Glucose sebelum dihapus	30
Gambar 11. <i>Outlier Removal</i> pada fitur Glucose setelah dihapus	30
Gambar 12. <i>Outlier Removal</i> pada fitur BloodPressure sebelum dihapus.....	30
Gambar 13. <i>Outlier Removal</i> pada fitur BloodPressure setelah dihapus.....	31
Gambar 14. <i>Outlier removal</i> pada fitur SkinThickness sebelum dihapus.....	31
Gambar 15. <i>Outlier removal</i> pada fitur SkinThickness setelah dihapus	31
Gambar 16. <i>Outlier removal</i> pada fitur Insulin sebelum dihapus	32
Gambar 17. <i>Outlier removal</i> pada fitur Insulin setelah dihapus	32
Gambar 18. <i>Outlier removal</i> pada fitur BMI sebelum dihapus.....	33
Gambar 19. <i>Outlier removal</i> pada fitur BMI setelah dihapus	33
Gambar 20. <i>Outlier removal</i> pada fitur DiabetesPedigreeFunction sebelum dihapus	33
Gambar 21. <i>Outlier removal</i> pada fitur DiabetesPedigreeFunction setelah dihapus .	34
Gambar 22. <i>Outlier removal</i> pada fitur Age sebelum dihapus.....	34
Gambar 23. <i>Outlier removal</i> pada fitur Age setelah dihapus.....	34
Gambar 24. Contoh Penentuan Q1 dan Q3 yang sudah di urutkan	35

DAFTAR GAMBAR

Gambar 25. Ilustrasi Pembagian Dataset	36
Gambar 26 Pohon Keputusan <i>CatBoost</i> dengan Root Pregnancies tanpa tuning.	38
Gambar 27 Pohon Keputusan <i>CatBoost</i> dengan Root SkinThickness tanpa tuning. .	39
Gambar 28. Pohon Keputusan <i>CatBoost</i> dengan Root BloodPressure tanpa tuning. 40	
Gambar 29. Pohon Keputusan <i>CatBoost</i> dengan Root Age tanpa tuning.	41
Gambar 30. Pohon Keputusan <i>CatBoost</i> dengan Root Insulin tanpa tuning.	42
Gambar 31. Proses masuknya nilai residual ke leaf.....	44
Gambar 32. Penambahan nilai residual yang ke dua	44
Gambar 33. Penambahan nilai residual yang ke tiga	44
Gambar 34. Penambahan nilai residual yang ke empat	45
Gambar 35. Penambahan nilai residual yang ke lima.....	45
Gambar 36. Proses penambahan nilai residual selanjutnya	45
Gambar 37 Pohon Keputusan <i>CatBoost</i> dengan Root BMI menggunakan tuning Bayesian.	48
Gambar 38. Pohon Keputusan <i>CatBoost</i> dengan Root SkinThickness menggunakan tuning Bayesian.	49
Gambar 39. Pohon Keputusan <i>CatBoost</i> dengan Root Pregnancies menggunakan tuning Bayesian.	50
Gambar 40. Pohon Keputusan <i>CatBoost</i> dengan Root Insulin menggunakan tuning Bayesian.	51
Gambar 41 . Pohon Keputusan <i>CatBoost</i> dengan Root BloodPressure menggunakan tuning Bayesian.	52
Gambar 42. Grafik ROC AUC Model <i>CatBoost</i>	53
Gambar 43. Grafik ROC AUC Model <i>CatBoost</i> dengan Bayesian Search	56

DAFTAR LAMPIRAN

Lampiran

Lampiran 1 Source Code Library *CatBoost*

Lampiran 2 Source Code Rumus Outlier

Lampiran 3 Source Code Penghapusan Outlier

Lampiran 4 Source Code Pembagian Data

Lampiran 5 Source Code *CatBoost*

Lampiran 6 Source Code *CatBoost* dengan Menggunakan *Bayesian Search*