Jurnal Ilmiah JNIUS : Jurnal Informatika UPS
Vol. 1 N.o 1 (2022)

PENERAPAN ALGORITMA GRADIENT BOOSTED DECISION TREES PADA ADABOOST UNTUK KLASIFIKASI STATUS DESA

Hasbi Firmansyah[*1], Zainal Abidin[2]
[1]Prodi Studi Informatika Fakultas Teknik dan Ilmu Komputer, UPS Tegal,
[2] Sekolah Tinggi Teknik Pati, Indonesia
Email : [*1]hasbifirmansyah@upstegal.ac.id, [2]zainal.frsd@yahoo.co.id

Abstrak

Sistem informasi masuk di desa, harus bekerjasama dengan pihak terkait salah satunya dengan Kementerian Desa PDTT bekerjasama dengan Badan Perencanaan Pembangunan Nasional dan Badan Pusat Statistik. Salah satu metode paling populer untuk menyelesaikan masalah ketidakseimbangan kelas adalah pengambilan. Namun, metode under-sampling dan over-sampling yang paling banyak digunakan mengubah distribusi kelas asli dari data yang tidak seimbang dengan menghilangkan instance kelas mayoritas atau meningkatkan instance kelas minoritas. Gradient Boosted Decision Tree (GBDT) adalah algoritma ensemble dari model regresi atau model pohon klasifikasi yang menggabungkan fungsi-fungsi parameter sederhana dengan hasil yang 'buruk' (tingginya kesalahan prediksi) untuk menciptakan prediksi yang sangat akurat. algoritma yang diusulkan memiliki kinerja terbaik dengan nilai rata-rata adalah 88.91% dibandingkan dengan algoritma lainnya di mana masing-masing nilai accuracy nya adalah 63.26%, 40.16%, 78.13%, 84.38%. Algoritma unggul kedua, ketiga dan keempat adalah metode Adabost+GBDT,GBDT,DT dan NB. Pada pengukuran Classification Error, algoritma yang diusulkan memiliki nilai rata-rata classification error pertama paling kecil dibanding algoritma lainnya, nilai masing-masing adalah 36.74% (DT), 59.84% (NB), 21.87% (GBDT), 15.62% (Adaboost+GBDT) dan 11.09%(SMOTE+Adaboost+GBDT metode usulan).

***Kata Kunci**: Data Desa, adboost, GBDT*

Abstract

The incoming information system in the village must cooperate with related parties, one of which is the Ministry of Villages, PDTT in collaboration with the National Development Planning Agency and the Central Statistics Agency. One of the most popular methods for solving class imbalance problems is retrieval. However, the most widely used under-sampling and over-sampling methods alter the original class distribution of the unbalanced data by eliminating majority class instances or increasing minority class instances. Gradient Boosted Decision Tree (GBDT) is an ensemble algorithm of regression models or classification tree models that combines simple parameter functions with 'poor' results (high prediction error) to create highly accurate predictions. The proposed algorithm has the best performance with an average value of 88.91% compared to other algorithms where each accuracy value is 63.26%, 40.16%, 78.13%, 84.38%. The second, third and fourth superior algorithms are the Adabost+GBDT, GBDT, DT and NB methods. In the Classification Error measurement, the proposed algorithm has the smallest average value of the first classification error compared to other algorithms, the respective values are 36.74% (DT), 59.84% (NB), 21.87% (GBDT), 15.62% (Adaboost+). GBDT) and 11.09%(SMOTE+Adaboost+GBDT proposed meth.

Keywords: *Village Data Addaboots GBDT*

PENDAHULUAN

Data Podes 2014 merupakan cara pengukuran yang disusun berdasarkan tingkat perkembangan desa di Indonesia yang menjadikan desa sebagai unit analisis dengan mengacu pada Undang Undang Nomor 6 Tahun 2014 tentang desa, yang dimaksudkan untuk memotret tingkat perkembangan desa di Indonesia dan dapat digunakan sebagai acuan untuk penyusunan perencanaan kebijakan dan pengawasan pembangunan desa [1]. Dalam system informasi diperlukan metode klasifikasi dalam bidang kajian data mining sangat popular digunakan di berbagai bidang dengan berbagai tujuan untuk mendukung sebuah keputusan perencanaan kebijakan yang bersumber dari analisis klasifikasi [2].

Ada beberapa teknik Esemble yang sangat populer yang akan disampaikan secara rinci seperti Bagging dan Boosting. Bagging singkatan dari *Booststrap Aggregation. Booststrap* adalah teknik pengambilan sample set data yang berbeda dari set pelatihan tertentu dengan menggunakan penggantian. Setelah melakukan *Booststrap* pada set data, model pada semua set yang berbeda dan menggabungkan hasilnya, Teknik ini dikenal sebagai Agregasi Bootstrap atau Bagging. Bagging dapat juga diartikan sebagai jenis teknik esemble dimana algoritma pelatihan tunggal digunakan pada subset berbeda dari data pelatihan dimana pengambilan sample subset dilakukan dengan penggantian (Bootstrap). Setelah agoritma dilatih pada semua subset, kemudian bagging memprediksi dengan menggabungkan semua prediksi yang dibuat oleh algoritma pada subset yang berbeda. Boosting merupakan ide memfilter atau membobot data yang digunakan untuk melatih tim yang lemah, sehingga dapat memberi bobot lebih atau hanya dengan observasi yang telah diklasifikasikan dengan buruk. Boosting mengacu pada algoritma yang mengubah dari yang lemah menjadi yang kuat. Boosting adalah metode esembel untuk meningkatkan prediksi model dari algoritma pembelajaran apapun. Selain itu, metode ensemble berbasis Bagging, dan Boosting [3][4] adalah metode lain yang banyak digunakan untuk menangani masalah yang tidak seimbang [5], [6].

Algoritma GBDT (*Gradient Boosted Decision Tree*) dikenalkan oleh [5][7] untuk membuat area pekerjaan terbaru untuk meningkatkan performa algoritma. Menggunakan koneksi antara peningkatan dan optimasi, pekerjaan baru ini mengusulkan Gradient Boosting Machine. Dalam masalah estimasi fungsi apa pun untuk menemukan fungsi regresi, yang meminimalkan beberapa ekspektasi hilangnya fungsi dan ketidak keseimbangan kelas (class imbalance).

METODE

Desain Penelitian

Dalam penelitian ini menggunakan metode eksperimen. Metode eksperimen yaitu fokus untuk melakukan investigasi pada beberapa variabel yang dipengaruhi oleh kondisi eksperimen yang kami lakukan. Selanjutnya, melakukan percobaan untuk memverifikasi metode-metode yang sudah ada, bertujuan untuk mengisolasi dan melakukan kontrol setiap kondisi-kondisi yang relevan dengan situasi yang diteliti serta melakukan pengamatan terhadap efek atau pengaruh ketika kondisi-kondisi tersebut dimanipulasi. Gambar 0.1 menunjukkan rancangan penelitian secara sistematis serta mendefinisikan masalah ilmiah dalam bentuk eksperimen dengan tahapan penelitian. Tahapan penelitian dapat dilihat pada skema dibawah ini.

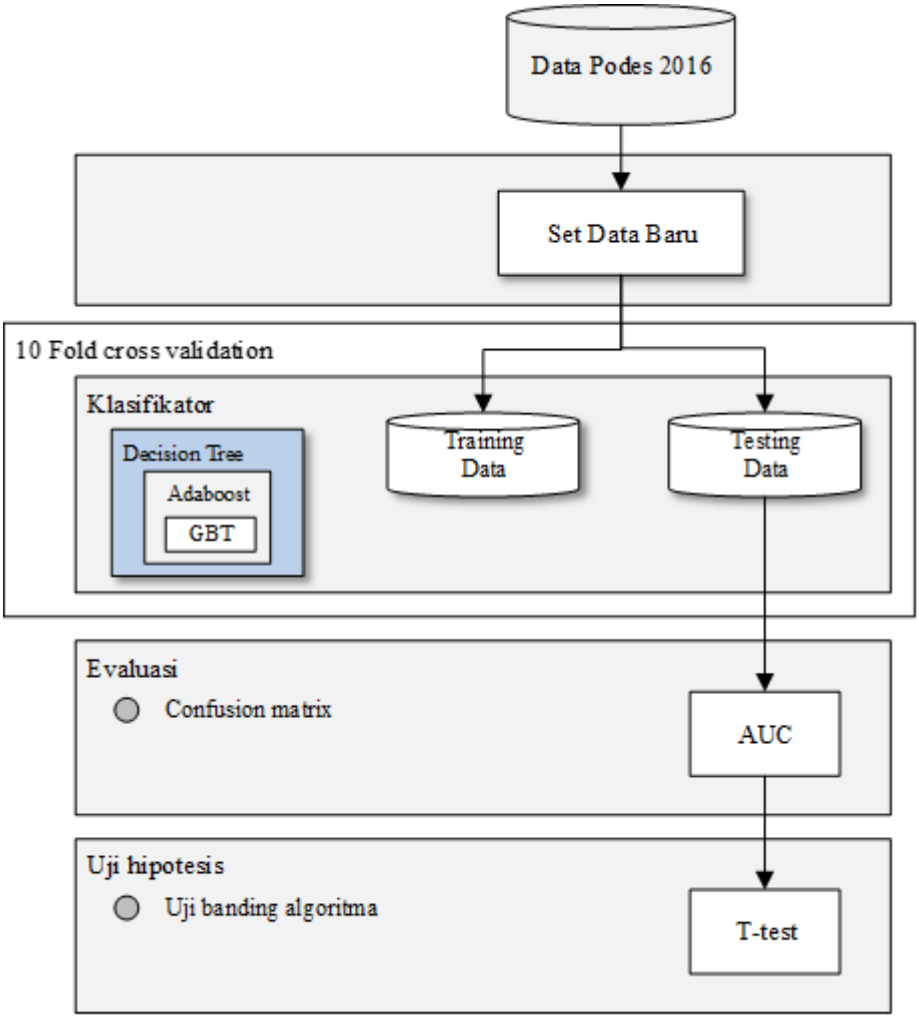
Gambar: Desain Penelitian

Menunjukkan tahapan penelitian dan aktifitas yang dilakukan pada tiap tahapan, secara detail tahapan penelitian diuraikan sebagai berikut:

1. Analisis Masalah dan Studi Literatur
2. Pengumpulan Data
3. Algoritma yang Diusulkan
4. Experimen dan Pengujian Algoritma
5. Evaluasi dan Validasi Hasil

Algoritma yang Diusulkan

Metode *Gradient Boosted Decision Tree* (GBDT) terbukti efisien [8], cepat dan mudah di implementasikan dalam menangani klasifikasi pada penggolongan status wilayah desa/kota, akan tetapi ada situasi di mana *Gradient Boosted Decision Tree* (GBDT) memiliki kelemahan dalam fungsi dan ketidak keseimbangan kelas [5] pada multi kelas data [9].



Gambar *Flowdiagram* dari Metode yang Diusulkan

GBDT menggunakan pohon keputusan untuk mempelajari fungsi dari ruang input X^s ke ruang gradien G [10]. Misalkan kita memiliki satu set pelatihan dengan n i.i.d. instans $\{x_1, \dots, x_n\}$, di mana setiap x_i adalah vektor dengan dimensi s dalam ruang X^s. Dalam setiap iterasi peningkatan gradien, gradien negatif dari fungsi kehilangan sehubungan dengan output model dilambangkan sebagai $\{g_1, \dots, g_n\}$. Model pohon keputusan membagi setiap node pada fitur yang paling informatif (dengan perolehan informasi terbesar). Untuk GBDT, penguatan informasi biasanya diukur dengan varians setelah pemisahan, yang didefinisikan sebagai berikut.

Diketahui $\{x_i, y_i\}_{i=1}^{n}$ menggambarkan dataset. GBDT digunakan untuk memastikan konvergensi pebelajaran. Langkah-langkah GBDT [10, 11] disajikan sebagai berikut:

Tahap 1. Inisialisasi nilai konstanta β

$$f_0(x) = \arg\min_{\beta} \sum_{i=1}^{N} L(y_i, \beta)$$

Tahap 2. Untuk jumlah iterasi $m = 1: M$ (M adalah waktu iterasi), arah gradien residual dihitung dengan:

$$y_i^* = -\left[\frac{\partial L\big(y_i, F(x_i)\big)}{\partial L(x_i)}\right]_{F(x)-F_{m-1(x)}} , i = \{1, 2, \dots, N\}$$

Tahap 3. Pengklasifikasi dasar digunakan untuk mencocokkan data sampel dan mendapatkan model awal, dihitung dengan

$$a_m = \arg\min_{\alpha\beta} \sum_{i=1}^{N} [y_i^* - \beta h(x_i; a)]^2$$

Tahap 4. Fungsi kerugian diminimalkan dihitung dengan:

$$\beta_m = \arg\min_{\alpha\beta} \sum_{i=1}^{N} L(y_i, F_{m-1}(x) + \beta h(x_i; a))$$

Tahap 5. Model diperbaharui

$$F_m(x) = F_{m-1} + \beta_m h(x_i; a)$$

HASIL DAN PEMBAHASAN

Pada tahap ini akan dilaporkan hasil penelitian dari metode yang telah diusulkan. Metode yang diusulkan adalah penerapan SMOTE pada algoritma Adaboost menggunakan algoritma Gradient Boosted Decision Tree untuk klasifikasi status desa di Indonesia. Data Podes 2016 digunakan untuk menguji metode yang diusulkan dengan metode pembanding, Decision Tree (DT), Naïve Bayes (NB), Gradient Boosted decision Tree (GBDT) dan Boosted decision Tree (GBDT+adaboost), yang bertujuan untuk mengamati perilaku metode sehubungan dengan dataset dengan karakteristik data.

Hasil Evaluasi Confussion Matrix

Pada Tabel 4.6 dapat dilihat hasil evaluasi algoritma menggunakan confussion matrix pada semua algoritma yaitu akurasi, jumlah dari: kenyataan salah prediksi salah, kenyataan benar prediksi salah, kenyataan salah prediksi benar dan kenyataan benar prediksi benar.

Tabel 0.1 Hasil evaluasi Akurasi pada Semua Algoritma

DT					
accuracy: 63.26% +/- 2.16% (micro average: 63.26%)					
	true Tertinggal	true Berkembang	true Sangat_tert	true Maju	true Mandiri
pred. tertinggal	779	421	74	18	0
pred. berkem.	286	1220	6	300	2
Pred. sangat_tert	43	15	133	1	0
pred. maju	11	181	1	310	51
pred. mandiri	0	1	0	17	17
NB					
accuracy: 40.16% +/- 3.16% (micro average: 40.16%)					
	true Tertinggal	true Berkembang	true Sangat_tert	true Maju	true Mandiri
pred. tertinggal	247	166	10	1	1
pred. berkem.	142	993	1	71	0
Pred. sangat_tert	729	108	203	0	0
pred. maju	1	83	0	56	7
pred. mandiri	0	488	0	518	62
GBDT					
accuracy: 78.13% +/- 2.14% (micro average: 78.13%)					
	true Tertinggal	true Berkembang	true Sangat_tert	true Maju	true Mandiri
pred. tertinggal	853	125	88	0	0
pred. berkem.	246	1628	1	234	1
Pred. sangat_tert	19	0	125	0	0
pred. maju	1	85	0	409	47
pred. mandiri	0	0	0	3	22
Adaboost+GBDT					
accuracy: **84.38**% +/- 1.14% (micro average: 84.38%)					
	true Tertinggal	true Berkembang	true Sangat_tert	true Maju	true Mandiri
pred. tertinggal	941	81	64	0	0
pred. berkem.	157	1673	0	149	0
Pred. sangat_tert	20	0	150	0	0
pred. maju	1	83	0	484	38
pred. mandiri	0	1	0	13	32
SMOTE+Adaboost+GBDT(Usulan)					

accuracy: **88.91%** +/- 1.07% (micro average: 88.91%)					
	true Tertinggal	true Berkembang	true Sangat_tert	true Maju	true Mandiri
pred. tertinggal	927	99	71	0	0
pred. berkem.	169	1662	0	143	0
Pred. sangat_tert	23	0	143	0	0
pred. maju	0	76	0	473	15
pred. mandiri	0	1	0	30	1823

Seperti yang dilihat pada Tabel 4.6 menunjukkan algoritma yang diusulkan memiliki akurasi yang baik di mana rata-rata akurasi yang dihasilkan juga sangatlah baik masing-masing diantaranya adalah akurasi 88,91% dengan rata-rata keselahan klasifikasi adalah 1.07% di mana kesalahn ini paling kecil dibandingkan algoritma pembanding lainnya.

Tabel Hasil Rangkuman Evaluasi pada Semua Algoritma

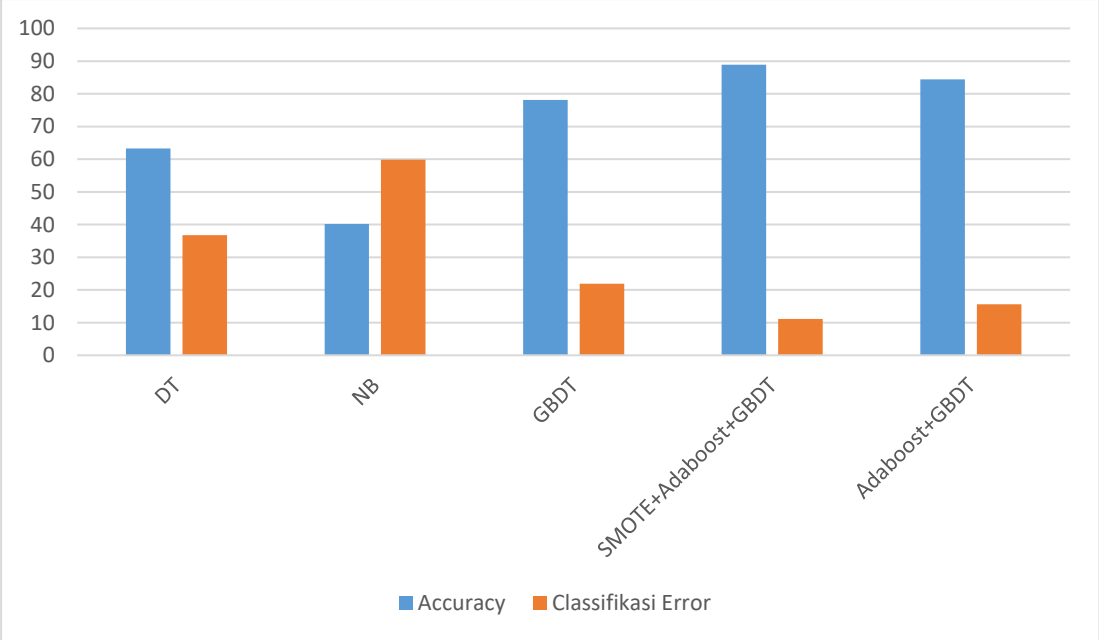
	DT	NB	GBDT	Adaboost+GBDT	SMOTE+Adaboost+GBDT
Accuracy	63.26	40.16	78.13	84.38	**88.91**
Kappa	0.441	0.248	0.658	0.760	**0.848**
Classi. Error %	36.74	59.84	21.87	15.62	**11.09**
RMSE	0.551	0.754	0.456	0.368	**0.309**
AUC	0.633	0.402	0.781	0.844	**0.889**

PEMBAHASAN

Pada pembahasan ini akan menjelaskan perbandingan kinerja algoritma, antara algoritma yang diusulkan, dengan algoritma pembanding yang digunakan dalam penelitian ini. Perbandingan kinerja algoritma berdasarkan pada kasus klasifikasi status desa di Indonesia. Algoritma yang diusulkan yaitu SMOTE Adaboost GBDT akan dibandingkan dengan algoritma DT, NB, GBDT dan Adaboost + GBDT

Tabel Hasil Accuracy dan Classification Error pada Algoritma Pembanding dan Algoritma Usulan

	DT	NB	GBDT	SMOTE+Adaboost+GBDT	Adaboost+GBDT
Accuracy	63.26	40.16	78.13	**88.91**	**84.38**
Classif. Error	36.74	59.84	21.87	**11.09**	**15.62**



Gambar diagram akurasi dan klasifikasi

KESIMPULAN

Dalam penelitian ini, penambahan SMOTE pada algoritma yang diusulkan untuk meningkatkan kinerja algoritma GBDT untuk klasifikasi status desa di Indonesia. Pengujian dilakukan dengan menggunakan data Podes 2016. Algoritma yang diusulkan tersebut kemudian dibandingkan dengan penelitian terkait sebelumnya yaitu Decision Tree (DT), Naïve Bayes (NB), Gradient Boosting Decision Tree (GBDT), Adaboost+GBDT. Kelima algoritma tersebut diukur evaluasinya menggunakan Accurasy, Classificaion Error, dan Koeficient Kappa, nilai evaluasi tertinggi dari Accuracy dan Koeficien Kappa adalah algoritma terbaik, sedangkan nilai evaluasi terkecil dari Classification Error adalah algoritma terbaik. Dilakukan juga uji beda menggunakan t-test pada algoritma pembanding dan algoritma usulan.

Pada pengukuran accuracy, algoritma yang diusulkan memiliki kinerja terbaik dengan nilai rata-rata adalah 88.91% dibandingkan dengan algoritma lainnya di mana masing-masing nilai accuracy nya adalah 63.26%, 40.16%, 78.13%, 84.38%. Algoritma unggul kedua, ketiga dan keempat adalah metode Adabost+GBDT,GBDT,DT dan NB. Pada pengukuran Classification Error, algoritma yang diusulkan memiliki nilai rata-rata classification error pertama paling kecil dibanding algoritma lainnya, nilai masing-masing adalah 36.74% (DT), 59.84% (NB), 21.87% (GBDT), 15.62% (Adaboost+GBDT) dan 11.09%(SMOTE+Adaboost+GBDT metode usulan). Terakhir, pada pengukuran Kappa, algoritma usulan juga mempati posisi pertama paling unggul disbanding algoritma lainnya.

DAFTAR PUSTAKA

[1] Y. Religia, “Metode Manhattan , Euclidean Dan Chebyshev Pada Algoritma K-Means Untuk Pengelompokan Status Desa,” 2016.

[2] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining Third Edition*. MK Morgan Kaufman, 2011.

[3] H. Rao *et al.*, "Feature selection based on artificial bee colony and gradient boosting decision tree," *Appl. Soft Comput. J.*, vol. 74, pp. 634–642, 2019.

[4] Y. Freund, R. E. Schapire, and M. Hill, "Experiments with a New Boosting Algorithm," *Proc. Int. Conf. Mach. Learn.*, pp. 148–156, 1996.

[5] N. V Chawla, A. Lazarevic, L. O. Hall, and K. W. Bowyer, "SMOTEBoost: Improving Prediction of the Minority Class in Boosting," *Eur. Conf. Princ. Data Min. Knowl. Discov.*, vol. 2838, pp. 107–119, 2003.

[6] C. Seiffert, T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano, "RUSBoost: A hybrid approach to alleviating class imbalance," *IEEE Trans. Syst. Man, Cybern. Part ASystems Humans*, vol. 40, no. 1, pp. 185–197, 2010.

[7] M. J. M. Saberian and N. Vasconcelos, "Multiclass Boosting: Theory and Algorithms," *Adv. Neural Inf. Process. Syst. 24 25th Annu. Conf. Neural Inf. Process. Syst. 2011. Proc. a Meet. held 12-14 December 2011, Granada, Spain.*, pp. 2124–2132, 2011.

[8] L. Guelman, "Gradient boosting trees for auto insurance loss cost modeling and prediction," *Expert Syst. Appl.*, vol. 39, no. 3, pp. 3659–3667, 2012.

[9] J. Friedman, T. Hastie, and R. Tibshirani, "ADDITIVE LOGISTIC REGRESSION: A STATISTICAL VIEW OF BOOSTING," *Ann. Stat.*, vol. 28, no. 2, pp. 337–407, 2000.

[10] Y. Ganjisaffar and R. Caruana, "Bagging Gradient-Boosted Trees for High Precision , Low Variance Ranking Models Categories and Subject Descriptors," *Proc. 34th Int. ACM SIGIR Conf. Res. Dev. Inf. - SIGIR '11*, no. c, pp. 85–94, 2011.

[11] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining Third Edition*. Elsevier, 2011.